Statistical and computational phase transitions in spiked tensor estimation

Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, Lenka Zdeborová

Statistical and computational phase transitions in spiked tensor estimation

Thibault Lesieur[†], Léo Miolane[◊], Marc Lelarge[◊], Florent Krzakala[⋆] & Lenka Zdeborová[†]

Abstract

We consider tensor factorization using a generative model and a Bayesian approach. We compute rigorously the mutual information, the Minimal Mean Squared Error (MMSE), and unveil information-theoretic phase transitions. In addition, we study the performance of Approximate Message Passing (AMP) and show that it achieves the MMSE for a large set of parameters, and that factorization is algorithmically "easy" in a much wider region than previously believed. It exists, however, a "hard" region where AMP fails to reach the MMSE and we conjecture that no polynomial algorithm will improve on AMP.

This study inscribes into the line of research on low-rank tensor decomposition, a problem with many applications ranging from signal processing to machine learning [2, 6, 25]. We consider the model of [24] where the tensor is a noisy version of a r-dimensional randomly generated spike and analyze the Bayes-optimal inference of the spike, compute the associated mutual information and the minimum mean-squared error (MMSE). We also investigate whether the MMSE is achievable with some known efficient algorithms, and most particularly by approximate message passing (AMP).

1 The spiked tensor model

One observes an order-p tensor $Y \in \bigotimes^p \mathbb{R}^N$ created as

$$Y = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{k=1}^{r} (X_k^0)^{\otimes p} + V \,, \tag{1}$$

where $X_1^0, \dots, X_r^0 \in \mathbb{R}^N$ are r unknown vectors to be inferred from Y, and $V \in \bigotimes^p \mathbb{R}^N$ is a symmetric tensor accounting for noise. We denote by X the $N \times r$ matrix that collects the r vectors X_k. The observed tensor Y can thus be seen as a rank r perturbation of a random symmetric tensor V. Consider now the setting where the X^0 is generated at random from a known prior distribution. The core question considered in this paper is: What is the best possible reconstruction of X^0 one can hope for?

In fact, we can look at even more general noise than just additive one as in (1). Denote for $i = 1, \dots, N$, $x_i = (x_{i,1}, \dots, x_{i,r}) \in \mathbb{R}^r$ the r-dimensional vector created by aggregating the i^{th} coordinates of the r vectors

This paper was presented at the IEEE International Symposium on Information Theory (ISIT) 2017 in Aachen, Germany.

† Institut de Physique Théorique, CNRS & CEA & Université Paris-Saclay, Saclay, France.
⋆ Laboratoire de Physique Statistique, CNRS & Université Pierre et Marie Curie & École Normale Supérieure & PSL Université, Paris, France.
◊ Département d'Informatique de l'ENS, École Normale Supérieure & CNRS & PSL Research University & Inria, Paris, France.

X_k. Assume that for $1 \leq i \leq N$ the x_i^0 are generated independently from a probability distribution P_X over $\mathbb{R}^r$. We denote, for $(i_1, i_2, \ldots, i_p) \in \{1, \ldots, N\}^p$

$$W^0_{i_1,i_2,\ldots,i_p} = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{k=1}^{r} x^0_{i_1,k} x^0_{i_2,k} \cdots x^0_{i_p,k} \,. \tag{2}$$

For simplicity, we will assume to only observe the extra-diagonal elements of Y, i.e. the coefficients $Y_{i_1,i_2,\ldots,i_p}$ for $1 \leq i_1 < \cdots < i_p \leq N$. The case where all coefficients are observed can be directly deduced from this case. The observed tensor Y is generated from W^0 using a noisy component-wise output channel P_{out} so that

$$P(Y|X^0) = \prod_{i_1<i_2<\cdots<i_p} P_{\text{out}}\left(Y_{i_1,i_2,\ldots,i_p} \middle| W^0_{i_1,i_2,\ldots,i_p}\right) . \tag{3}$$

The simplest situation corresponds to eq. (1) with additive white Gaussian noise (AWGN), i.e. $P_{\text{out}}(\,\cdot\,|\,w) = \mathcal{N}(w, \Delta)$.

Given the above generative model and assuming that both the prior distribution P_X and the output channel P_{out} are known we can write the Bayes-optimal estimator of X^0 as marginalization of the following posterior distribution

$$P(X|Y) = \frac{1}{\mathcal{Z}_N} \prod_{i=1}^{N} P_X(x_i) \prod_{i_1<i_2<\cdots<i_p} P_{\text{out}}\left(Y_{i_1,i_2,\ldots,i_p} \middle| W_{i_1,i_2,\ldots,i_p}\right) , \tag{4}$$

where $\mathcal{Z}_N$ is a normalization constant depending of the observed tensor Y, $W_{i_1,i_2,\ldots i_p}$ is defined analogously to (2) (with X instead of X^0).

We will study this tensor estimation problem in the limit where the dimension $N \to \infty$ while the rank r remains constant. The factor $N^{\frac{p-1}{2}}$ is here to ensure that information-theoretically the inference problem is neither trivially hard nor trivially easy when one deals with signals such that $\|x_i\|$ and the noise magnitude are of order 1. The factor $\sqrt{(p-1)!}$ is used for convenient rescaling of the signal-to-noise ratio.

2 Related work and summary of results

Recently there have been numerous works on the matrix ($p = 2$) version of the above setting. In particular an explicit single-letter characterization of the mutual information and of the optimal Bayesian reconstruction error have been rigorously established [4, 9, 12–14]. A large part of these results rely on the approximate message passing (AMP) algorithm. For the rank-one matrix estimation problems AMP has been introduced by [23], who also derived the state evolution (SE) formula to analyze its performance, generalizing techniques developed by [5]. In [15–17] the generalization to larger rank, and general output channel, was considered. Following the theorem proven in [4, 9, 14], we know that indeed AMP is Bayes-optimal and achieves the minimum mean-squared error (MMSE) for a large set of parameters of the problem. There, however, might exist a region denoted as *hard*, where this is not the case, and polynomial algorithms improving on AMP are not known.

In comparison, there has been much less work on Bayesian low-rank tensor estimation. In statistical physics, the measure (4) was considered for Y with random i.i.d. components. For a Gaussian P_X, it is called the spherical p-spin glass [8], while for Rademacher P_X it is the Ising p-spin glass [19]. AMP for tensor estimation is actually equivalent to the so-called Thouless-Anderson-Palmer equations in spin glass theory [7, 12, 27]. In the context of tensor PCA these equations have been studied by Richard and Montanari [24] for the maximum likelihood estimation. Interestingly, they showed that the *hard* phase was particularly large in the tensor estimation case and that, with side information (such that for each component $x_{i,k}$ we have its *direct* noisy

observation), the estimation problem becomes easier. However, such a kind of component-wise side information is very strong and rarely available in applications. The tight statistical limits for the present tensor model were also studied in [12] for the special case of the Rademacher (Ising) prior. For more generic priors only upper and lower bounds are known rigorously [22].

Summary of results: In this contribution, we aim to bridge the gap between what is known for the general $r, P_x, P_{\rm out}$ Bayesian estimation for low-rank matrices and what is known for low-rank tensors. We present the following contributions:

(A) The AMP algorithm and its state evolution analysis for the Bayes-optimal tensor estimation, see sections 3 and 4.

(B) The so-called *channel universality* result that allows us to map any generic channel $P_{\rm out}$ on a model with additive Gaussian noise, see section 3.

(C) Rigorous formula for the asymptotic mutual information and the MMSE, thus generalizing the matrix results of [4, 14], see section 5.

(D) The identification of statistical and computational phase transitions. In fact, we show that as soon as the effect of a non-zero-mean prior is taken properly into account, the hard region shrinks considerably, making the tensor decomposition problem much easier than hitherto believed, at least for algorithms that do take the prior information into account. Having a reliable prior information on the distribution of x_i (not on each of the components as in [24]) is rather realistic in applications, for instance when constraints of negativity or membership to clusters are imposed. This is presented in sections 4 and 6.

3 AMP algorithm & channel universality

We discuss in this section the Approximate Message Passing (AMP) algorithm for the Bayesian version of the problem. This is a relatively straightforward generalization of what has been done for the low-rank matrix estimation in e.g. [17, 18, 23], i.e. $p = 2$ case of the present setting. In general, AMP is derived from belief propagation by taking into account that every variable in the corresponding graphical model has a large number of neighbors. Since the incoming messages are considered independent one can use the central limit theorem and represent each message as a Gaussian with a given mean $\hat{x}_i \in \mathbb{R}^r$ and covariance $\sigma_i \in \mathbb{R}^{r\times r}$.

A crucial property, called *channel universality*, that the tensor-AMP shares with the low-rank matrix estimation, allows to drastically simplify the problem of tensor estimation with generic output channel $P_{\rm out}$. The justification of this property follows closely the low-rank matrix estimation case, and we refer the reader to [13, 15, 17]. First, we define the Fisher score tensor S associated to the output channel $P_{\rm out}$ and its Fisher information Δ as

$$S \equiv \left.\frac{\partial \log P_{\rm out}(Y,w)}{\partial w}\right|_{w=0}, \tag{5}$$

$$\frac{1}{\Delta} \equiv \mathbb{E}_{Y\sim P_{\rm out}(\cdot\,|\,0)}\left[\left(\frac{\partial \log P_{\rm out}(Y,w)}{\partial w}\right)^2_{w=0}\right]. \tag{6}$$

where it is understood in (5) that the function $y \mapsto \left.\frac{\partial \log P_{\rm out}(y,w)}{\partial w}\right|_{w=0}$ acts component-wise on Y. Informally speaking, the channel universality property states that the mutual information of the problem defined by the output channel $P_{\rm out}$ is the same as the one of a AWGN (1) with variance Δ, and that the AMP algorithm written for the Bayes-optimal inference of low-rank tensors then depends on the data tensor Y and the output channel $P_{\rm out}$ only trough the tensor S and the effective noise Δ.

AMP involves an auxiliary function that depends explicitly on the prior as follows. Define the probability distribution

$$\mathcal{M}(x) = \frac{1}{\mathcal{Z}_X(A,B)} P_X(x) e^{B^\top x - \frac{x^\top A x}{2}}, \tag{7}$$

where $\mathcal{Z}_X(A,B)$ is a normalization factor. Then AMP uses the function $f_{\text{in}}(A,B) \in \mathbb{R}^r, A \in \mathbb{R}^{r\times r}, B \in \mathbb{R}^r$ defined by the expectation $f_{\text{in}}(A,B) = \mathbb{E}_{\mathcal{M}(x)}[x]$ as well as the covariance matrix $\partial_B f_{\text{in}}(A,B)$. We shall denote the *overlap* of $u = (u_1, \dots u_N), v = (v_1, \dots v_N) \in (\mathbb{R}^r)^N$ by

$$u \cdot v = \frac{1}{N} \sum_{j=1}^{n} u_j v_j^\top \in \mathbb{R}^{r\times r}.$$

AMP is then written as an iterative update procedure on the estimates of the posterior means and co-variances $\hat{x}_i$ and σ_i that uses auxiliary variables $B_i \in \mathbb{R}^r$ and $A \in \mathbb{R}^{r\times r}$:

$$B_i^t = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{i_2<i_3<\dots<i_p} S_{i,i_2,i_3\dots i_p} \hat{x}_{i_2}^t \circ \hat{x}_{i_3}^t \circ \dots \circ \hat{x}_{i_p}^t - \frac{(p-1)}{\Delta} \left[\frac{1}{N} \sum_{j=1}^{N} \sigma_j^t \circ (\hat{x}^t \cdot \hat{x}^{t-1})^{\circ(p-2)} \right] \hat{x}_i^{t-1} \tag{8}$$

$$A^t = \frac{1}{\Delta} (\hat{x}^t \cdot \hat{x}^t)^{\circ(p-1)} \tag{9}$$

$$\hat{x}_i^{t+1} = f_{\text{in}}(A^t, B_i^t) \tag{10}$$

$$\sigma_i^{t+1} = \partial_B f_{\text{in}}(A^t, B_i^t), \tag{11}$$

where $\circ$ denotes a component-wise (Hadamard) product of matrices, and $x^{\circ p}$ the corresponding component-wise power.

4 Theoretical analysis

4.1 State evolution of AMP

The evolution of the AMP algorithm in the limit of large systems $N \to \infty$ can be tracked via a low-dimensional set of equations called the *state evolution* (SE). For maximum-likelihood estimation the state evolution have been used in [24]. Its heuristic derivation for the present case of general rank r, prior P_X, and output P_{out} follows line by line the matrix estimation case detailed in [17].

For the Bayes-optimal inference, SE is written in terms of an order parameter $M^t \in \mathbb{R}^{r\times r}$ describing the overlap between $\hat{x}^t$ (the AMP estimator at iteration t) and the ground truth x^0 defined as $M^t = \hat{x}^t \cdot x^0$, and reads

$$M^{t+1} = \mathbb{E}_{Z,x_0} \left[f_{\text{in}} \left(\widehat{M}^t, \widehat{M}^t x_0 + \left(\widehat{M}^t\right)^{1/2} Z \right) x_0^\top \right], \tag{12}$$

$$\widehat{M}^t = (M^t)^{\circ(p-1)}/\Delta, \tag{13}$$

where $Z \sim \mathcal{N}(0, I_r)$ and $x_0 \sim P_X$ are independent. $M^{\circ(n)}$ is again the n-th Hadamard power of a matrix M.

We shall not present a rigorous proof of the SE for tensor estimation and rely instead on standard arguments from statistical physics. The performance of the AMP algorithm can be understood by initializing the SE at $M^{t=0} = 0$. Or when $M = 0$ is a fixed point of SE we initialize as $M^{t=0} = \epsilon$, an infinitesimally small number (accounting for the fact that a random initialization of AMP will —due to finite size fluctuations— be infinitesimally correlated with the ground truth). We denote M_{AMP} the fixed point of the state evolution resulting from iterations of (12-13) from this initialization. The mean-squared error achieved by tensor-AMP

is then

$$\mathrm{MSE}_{\mathrm{AMP}} = \mathrm{Tr}\left[\Sigma_X - M_{\mathrm{AMP}}\right] . \tag{14}$$

where $\Sigma_X = \mathbb{E}_x[xx^\top]$. When P_X has zero mean, this is the covariance matrix of P_X.

4.2 Information-theoretically optimal inference

Our next goal is to analyze the performance of (possibly intractable) Bayes-optimal inference that evaluates the marginals of the posterior probability distribution (4). The error achieved by this procedure will be denoted the minimum mean-squared error (MMSE) and is formally defined as

$$\mathrm{MMSE}_N = \inf_{\hat{\theta}} \left\{ \frac{1}{N}\mathbb{E}\left[\left\|X^0 - \hat{\theta}(Y)\right\|^2\right]\right\} = \frac{1}{N}\mathbb{E}\left[\left\|X^0 - \mathbb{E}[X^0|Y]\right\|^2\right] ,$$

where the infimum is taken over all measurable functions $\hat{\theta}$ of Y. In order to compute the MMSE it is instrumental to compute the mutual information $I(X^0;Y)$. This quantity is related to the free energy from statistical physics (see section 5 and [13]). To compute the limit of such quantities, one traditionally applies the replica method stemming from statistical physics [19]. We take advantage of the fact that for the Bayes-optimal inference the so-called *replica symmetric* version of this method yields the correct free energy [28]. The replica method yields

$$\frac{1}{N}I(X^0;Y) \xrightarrow[N\to\infty]{} \frac{1}{2p\Delta}\sum_{k,k'=1}^{r}(\Sigma_X)^p_{k,k'} - \sup_{M\in S_r^+}\phi_{\mathrm{RS}}(M) , \tag{15}$$

$$\phi_{\mathrm{RS}}(M) = \mathop{\mathbb{E}}_{Z,x_0}\left[\log \mathcal{Z}_X\left(\widehat{M}, \widehat{M}x_0 + \left(\widehat{M}\right)^{1/2}Z\right)\right] - \frac{p-1}{2p\Delta}\sum_{k,k'=1}^{r} M^p_{kk'} \tag{16}$$

where $\widehat{M} = M^{\circ(p-1)}/\Delta$, $\mathcal{Z}_X(A,B)$ is defined in eq. (7), $x_0 \sim P_X$ and $Z \sim \mathcal{N}(0, I_r)$ are independent random variables. S_r^+ denotes the set of $r \times r$ symmetric positive semi-definite matrices. In section 5 we prove this result for the rank-one case ($r = 1$).

The replica free energy (16) not only provides the limit of the mutual information $I(X^0;Y)$, but thanks to an "I-MMSE Theorem" (similar to [11]) it yields the value of the MMSE for tensor estimation, see sec. 5. Denoting $M^* = \mathrm{argmax}_M \phi_{\mathrm{RS}}(M)$ we get

$$\mathrm{MMSE} = \lim_{N\to\infty}\mathrm{MMSE}_N = \mathrm{Tr}\left[\Sigma_X - M^*\right] . \tag{17}$$

We proved (17) rigorously, but only in the rank-one case and for odd values of p, see again sec. 5. Notice that when $r \geq 2$ the estimation problem is symmetric under permutations of the r columns of X^0: (17) is not expected to be true without further assumptions.

4.3 Statistical and computational trade-off

By evaluation of the derivative of (16) with respect to M one can check that critical points of (16) are fixed points of the state evolution equations (12-13) allowing all the results to be read of the curve $\phi_{\mathrm{RS}}(M)$: The global maximum of (16) gives the MMSE while the (possibly local) maximum reached by iteration of (12-13) from the uninformative initialization yields the $\mathrm{MSE}_{\mathrm{AMP}}$.

We now discuss the interplay between the MMSE and $\mathrm{MSE}_{\mathrm{AMP}}$. The working hypothesis in this paper is that AMP yields lowest MSE among known polynomial algorithms. Depending on the parameters of model (4), i.e. the order of the tensor p, rank r, prior distribution P_X, and output channel P_{out} that appears in the SE

only via its Fisher information Δ, we can distinguish between two cases: the **easy** phase where asymptotically AMP is Bayes optimal so that $\mathrm{MMSE} = \mathrm{MSE}_{\mathrm{AMP}}$, and the **hard** phase where $\mathrm{MMSE} < \mathrm{MSE}_{\mathrm{AMP}}$.

Given both the MMSE and $\mathrm{MSE}_{\mathrm{AMP}}$ are non-decreasing in Δ we denote the borders of the hard phase (when it exists) as follows: **Information theoretic threshold** Δ_{IT} as the (limsup of the) highest Δ for which $\mathrm{MMSE} < \mathrm{MSE}_{\mathrm{AMP}}$. **Algorithmic threshold** Δ_{Alg} as the (liminf of the) lowest Δ for which $\mathrm{MMSE} < \mathrm{MSE}_{\mathrm{AMP}}$. Another threshold used in this paper is that of a critical value Δ_c defined as smallest Δ such that for $\Delta > \Delta_c$ one has $M_{\mathrm{AMP}} = M^*(\Delta = +\infty)$ (the estimate one can do when the noise is infinite), and for $\Delta < \Delta_c$ one has $M_{\mathrm{AMP}} > M^*(\Delta = +\infty)$. Note that from the definition we must have $\Delta_c \geq \Delta_{\mathrm{Alg}}$. In cases where the hard phase does not exist, but $\Delta_c < \infty$ we will consider that $\Delta_c = \Delta_{\mathrm{IT}} = \Delta_{\mathrm{Alg}}$.

Existing results on maximum likelihood estimation [24] suggest that for tensor decomposition $p \geq 3$ we have $\Delta_{\mathrm{Alg}} = \Delta_c = 0$ in the limit $N \to 0$ considered in this paper. This means that the spiked model of low-rank tensor decomposition is algorithmically very hard, compared to matrix $p = 2$ case. The authors of [24] give a good account on how Δ needs to scale with N for known polynomial algorithms to work.

For the Bayes-optimal estimation the situation seems at first sight similar. Indeed, whenever the prior P_X has a zero mean, for $p \geq 3$ we get $\Delta_{\mathrm{Alg}} = \Delta_c = 0$ and the hard phase is consequently huge. This can be seen as follows. Indeed if the mean of the prior P_X is zero then the state evolution equations (12-13) have a fixed point $M = 0$. Expanding the state evolution around this fixed point we find

$$M^{t+1} = \frac{1}{\Delta} \Sigma_X \left[(M^t)^{\circ(p-1)} \right] \Sigma_X \,. \tag{18}$$

Whenever $p \geq 3$ the fixed point $M = 0$ is stable for all $\Delta > 0$. Hence $\Delta_{\mathrm{Alg}} = \Delta_c = 0$ for priors of zero mean.

A closer look, however, shows that the situation is not so pessimistic. Indeed, as soon as the mean of the prior P_X is non-zero, $M = 0$ is no longer a fixed point of the state evolution and once we solve the state evolution equations we observe either $\Delta_{\mathrm{Alg}} > 0$ (with AMP performing optimally for $\Delta < \Delta_{\mathrm{Alg}}$) or the hard phase is completely absent and AMP has information-theoretically optimal performance for all Δ. We give examples of such priors in section 6.

5 Rigorous results

We present in this section rigorous results for the rank-one case ($r = 1$). As mentioned above, the universality property [13, 15] reduces the computation of the mutual information to the case of additive white Gaussian noise.

Consider a probability distribution P_X over $\mathbb{R}$ that admits a finite second moment Σ_X. The observation model (1) reduces in the rank-one case to

$$Y_{i_1,\dots,i_p} = \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} x^0_{i_1} \dots x^0_{i_p} + V_{i_1,\dots,i_p} \quad \text{for } 1 \leq i_1 < \dots < i_p \leq N,$$

where $X^0 = (x^0_1, \dots, x^0_N) \overset{\text{i.i.d.}}{\sim} P_X$ and $(V_{i_1,\dots,i_p})_{i_1<\dots<i_p} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Delta)$ are independent. We define the Hamiltonian

$$H_N(X) = \Delta^{-1} \sum_{i_1<\dots<i_p} \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} Y_{i_1,\dots,i_p} x_{i_1} \dots x_{i_p} - \frac{(p-1)!}{2N^{p-1}} \left(x_{i_1} \dots x_{i_p}\right)^2 , \tag{19}$$

for $X = (x_1, \dots, x_N) \in \mathbb{R}^N$. We also write $dP_X(X) = \prod_{i=1}^N dP_X(x_i)$. The posterior distribution (4) of X^0

given Y reads then:

$$dP(X^0 = X|Y) = \frac{1}{\mathcal{Z}_N} dP_X(X) e^{H_N(X)}\,, \tag{20}$$

where $\mathcal{Z}_N$ is the appropriate normalizing factor. Then the free energy is defined as (minus) the logarithm of $\mathcal{Z}_N$ of the Boltzmann probability divided by N and averaged over Y. This is of particular interest since it is related to the mutual information (see [13]):

$$I(X^0; Y) = \frac{N}{2p\Delta}\Sigma_X^p - \mathbb{E}[\log \mathcal{Z}_N] + O(1)\,.$$

In the rank-one case, the expression (16) of $\phi_{\rm RS}$ simplifies, so that we will use in this section

$$\phi_{\rm RS} : m \geq 0 \mapsto \mathbb{E}\left[\log \int dP_X(x) \exp\left(\sqrt{\frac{m^{p-1}}{\Delta}} Z x + \frac{m^{p-1}}{\Delta} x x^0 - \frac{m^{p-1}}{2\Delta} x^2\right)\right] - \frac{p-1}{2p\Delta} m^p\,, \tag{21}$$

where $\mathbb{E}$ is the expectation with respect to the independent random variables $x_0 \sim P_X$ and $Z \sim \mathcal{N}(0,1)$. The proof of (15) reduces then to the following Theorem.

Theorem 1 (Replica-symmetric formula for the free energy)**.** *Let P_X be a probability distribution over $\mathbb{R}$, with finite second moment. Then, for all $\Delta > 0$*

$$F_N \equiv \frac{1}{N}\mathbb{E}\left[\log \mathcal{Z}_N\right] \xrightarrow[N\to\infty]{} \sup_{m\geq 0} \phi_{\rm RS}(m) \equiv F_{\rm RS}(\Delta)\,. \tag{22}$$

We now define the tensor-MMSE, T-MMSE$_N$ by

$$\text{T-MMSE}_N(\Delta) = \inf_{\hat\theta}\left\{\frac{p!}{N^p}\sum_{i_1<\cdots<i_p}\left(x^0_{i_1}\dots x^0_{i_p} - \hat\theta(Y)_{i_1\dots i_p}\right)^2\right\}\,,$$

where the infimum is taken over all measurable functions $\hat\theta$ of the observations Y.

Let us write $\lambda = \frac{1}{\Delta}$. Using an "I-MMSE Theorem" (see [11]) and the fact that the tensor MMSE is achieved by the posterior mean of $(X^0)^{\otimes p}$ given Y, it is not difficult to verify that

$$\frac{\partial F_N}{\partial \lambda} = \frac{N(N-1)\dots(N-p+1)}{2pN^p}\left(\Sigma_X^p - \text{T-MMSE}_N(\Delta)\right).$$

The arguments are the same than in the matrix ($p = 2$) case, see [14] Corollary 17. T-MMSE(Δ) increases with the noise level Δ, so that $\frac{\partial}{\partial\lambda}F_N$ is a non-decreasing function of λ. F_N is thus a convex function of λ, and so is $F_{\rm RS}$ its pointwise limit. Consequently, $\frac{\partial}{\partial\lambda}F_N \to \frac{\partial}{\partial\lambda}F_{\rm RS}$ at all values of λ at which $F_{\rm RS}$ is differentiable, that is for almost every $\Delta > 0$. For these values of Δ, one can also verify that the maximizer m^* of $\phi_{\rm RS}$ is unique: we refer to [14] for a detailed proof in the matrix case $p = 2$. We thus obtain the following theorem:

Theorem 2 (Tensor-MMSE)**.** *For almost every $\Delta > 0$, $\phi_{\rm RS}$ admits a unique maximizer $m^*(\Delta)$ over $\mathbb{R}_+$ and*

$$\textit{T-MMSE}_N \xrightarrow[N\to\infty]{} \Sigma_X^p - m^*(\Delta)^p\,.$$

The information-theoretic threshold $\Delta_{\rm IT}$ is the maximal value of Δ such that $\lim \text{T-MMSE}_N < \Sigma_X^p - \mathbb{E}_{P_X}[x]^{2p}$ (which is the asymptotic performance achieved by random guess). We obtain thus the precise location of the information-theoretic threshold:

$$\Delta_{\rm IT} = \sup\left\{\Delta > 0 \,\middle|\, m^*(\Delta) > \mathbb{E}_{P_X}[x]^2\right\}\,.$$

Let $X = (x_1, \dots, x_N)$ be a sample from the posterior (4), independently of everything else. An extension of Theorem 2 of [14] (that was derived for priors P_X with bounded support) to the tensor case, gives that for almost every $\Delta > 0$,

$$\mathbb{E}\left| \left(\frac{1}{N} \sum_{i=1}^{N} x_i^0 x_i \right)^p - m^*(\Delta)^p \right| \xrightarrow[N\to\infty]{} 0\,, \tag{23}$$

i.e. the p^{th}-power of the overlap $X \cdot X^0$ concentrates around m^*. This leads to

Theorem 3 (Vector-MMSE for odd p). *Suppose that P_X has a bounded support. If p is odd, then for almost every $\Delta > 0$*

$$\mathit{MMSE}_N \xrightarrow[N\to\infty]{} \Sigma_X - m^*(\Delta).$$

Before showing how (23) implies Theorem 3 we need to introduce a fundamental property of Bayesian inference: the Nishimori identity.

Proposition 1 (Nishimori identity). *Let (X, Y) be a couple of random variables on a polish space. Let $k \geq 1$ and let $X^{(1)}, \dots, X^{(k)}$ be k i.i.d. samples (given Y) from the distribution $P(X = \cdot \,|\, Y)$, independently of every other random variables. Let us denote $\langle \cdot \rangle$ the expectation with respect to $P(X = \cdot \,|\, Y)$ and $\mathbb{E}$ the expectation with respect to (X, Y). Then, for all continuous bounded function f*

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

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

We will now use Proposition 1 to prove Theorem 3.

Proof of Theorem 3. Let $\langle \cdot \rangle$ denote the expectation with respect to the posterior distribution $P(X^0 = \cdot \,|\, Y)$, and let X be a sample from this distribution, independently of everything else. The best estimator of X^0 in term of mean-squared error is the posterior mean $\langle X \rangle = (\langle x_1 \rangle, \dots, \langle x_N \rangle)$. Therefore

$$\begin{aligned} \mathrm{MMSE}_N &= \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} (x_i^0 - \langle x_i \rangle)^2 \right] = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} (x_i^0)^2 + \langle x_i \rangle^2 - 2\langle x_i^0 x_i \rangle \right] \\ &= \Sigma_X + \mathbb{E}\langle X \cdot X' \rangle - 2\mathbb{E}\langle X^0 \cdot X \rangle\,, \end{aligned}$$

where X' is another sample from $\langle \cdot \rangle$, independently of everything else. We apply now the Nishimori identity (Proposition 1) to obtain $\mathbb{E}\langle X \cdot X' \rangle = \mathbb{E}\langle X^0 \cdot X \rangle$. This gives

$$\mathrm{MMSE}_N = \Sigma_X - \mathbb{E}\langle X \cdot X^0 \rangle\,.$$

We then deduce from (23) that $\mathbb{E}\langle X \cdot X^0 \rangle \xrightarrow[N\to\infty]{} m^*$, because p is here supposed to be odd. This concludes the proof. □

We will now prove Theorem 1. For the matrix case ($p = 2$), this has been proved in [4, 13, 14] and we explain here how this can be adapted to the case $p \geq 2$. To prove the limit (22), one shows successively an upper bound on $\limsup F_N$ and the matching lower bound on $\liminf F_N$. As shown in [14] (Section 6.2.2) one only need to prove Theorem 1 for input distributions P_X with finite support S. We now assume to be in this situation.

5.1 Adding a small perturbation

One of the key ingredient of the proof is the introduction of a small perturbation of our model, that takes the form of a small amount of side information. This kind of techniques are frequently used for the study of spin glasses, where these small perturbations forces the Gibbs measure to verify some crucial identities, see [21]. In our context of Bayesian inference, we will see that small quantities of side information "breaks" the correlations of the signal variables under the posterior distribution.

Let us fix $\epsilon \in [0,1]$, and suppose we have access to the additional information, for $1 \leq i \leq N$

$$Y_i' = \begin{cases} x_i^0 & \text{if } L_i = 1, \\ * & \text{if } L_i = 0, \end{cases} \tag{24}$$

where $L_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(\epsilon)$ and $*$ is a symbol that does not belong to $\mathbb{R}$. The posterior distribution of X is now

$$P(X|Y,Y') = \frac{1}{\mathcal{Z}_{N,\epsilon}} \left(\prod_{i|Y_i' \neq *} 1(x_i = Y_i') \right) \left(\prod_{i|Y_i'=*} P_X(x_i) \right) e^{H_N(X)},$$

where $\mathcal{Z}_{N,\epsilon}$ is the appropriate normalization constant. For $X = (x_1, \dots, x_N) \in \mathbb{R}^N$ we will use the notation

$$\bar{X} = (L_1 x_1^0 + (1-L_1)x_1, \dots, L_N x_N^0 + (1-L_N)x_N). \tag{25}$$

$\bar{X}$ is thus obtained by replacing the coordinates of X that are revealed by Y' by their revealed values. The notation $\bar{X}$ will allow us to obtain a convenient expression for the free energy of the perturbed model

$$F_{N,\epsilon} = \frac{1}{N}\mathbb{E}\log \mathcal{Z}_{N,\epsilon} = \frac{1}{N}\mathbb{E}\Big[\log \sum_{X \in S^N} P_X(X) e^{H_N(\bar{X})}\Big].$$

The next lemma shows that the perturbation does not change the free energy up to the order of ϵ. Recall that we supposed the support S of P_X to be finite, so we can find a constant K such that $S \subset [-K, K]$.

Lemma 1. *For all $n \geq 1$ and all $\epsilon, \epsilon' \in [0,1]$, we have*

$$|F_{N,\epsilon} - F_{N,\epsilon'}| \leq \frac{K^{2p}}{\Delta}|\epsilon - \epsilon'|.$$

Lemma 1 follows from a direct adaptation of Proposition 23 from [14] to the tensor case. Consequently, if we suppose $\epsilon \sim \mathcal{U}([0,1])$ and define $\epsilon_N = N^{-1/2}\epsilon$ and $L_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(\epsilon_N)$, independently of everything else, we have

$$|F_N - \mathbb{E}_\epsilon[F_{N,\epsilon_N}]| \xrightarrow[N\to\infty]{} 0, \tag{26}$$

where $\mathbb{E}_\epsilon$ denotes the expectation with respect to ϵ only. It remains therefore to compute the limit of the free energy under a small perturbation. As shown in [20], the perturbation (24) forces the correlations to vanish asymptotically.

Lemma 2 (Lemma 3.1 from [20])**.**

$$\mathbb{E}_\epsilon \left[\frac{1}{N^2} \sum_{1 \leq i,j \leq N} I(x_i^0; x_j^0 \,|\, Y, Y') \right] \leq \frac{2H(P_X)}{\sqrt{N}}.$$

Let us write $\langle \cdot \rangle$ the expectation with respect to $P(X = \cdot \mid Y, Y')$, and let $X^{(1)}, X^{(2)}$ be two independents samples from $P(X = \cdot \mid Y, Y')$, independently of everything else. We define $Q = \langle X^{(1)} \cdot X^{(2)} \rangle$. Notice

that Q is a non-negative random variable. As a consequence of Lemma 2, the overlaps under the posterior distribution concentrates around Q:

Lemma 3.

$$\mathbb{E}\left\langle \left(X^{(1)}\cdot X^{(2)}-Q\right)^2\right\rangle \xrightarrow[N\to\infty]{} 0 \quad \textit{and} \quad \mathbb{E}\left\langle \left(X^{(1)}\cdot X^{0}-Q\right)^2\right\rangle \xrightarrow[N\to\infty]{} 0\,, \tag{27}$$

where $\mathbb{E}$ denotes the expectation with respect all random variables.

Lemma 3 follows from the arguments of section 4.4 from [14].

The arguments presented in this section are robust and apply to a large class of Hamiltonians. In particular, we will be able to apply in the sequel Lemmas 1 and 3 to other Hamiltonians and posterior distributions (and corresponding free energies).

5.2 Guerra's interpolation scheme

The lower bound is obtained by extending the bound derived for $p=2$ in [13], using a Guerra-type interpolation [10] as was already done for tensors by Korada and Macris in [12] (who consider tensors in the special case of Rademacher P_X).

Lemma 4.

$$\liminf_{N\to\infty} F_N \geq \sup_{m\geq 0} \phi_{\rm RS}(m)\,.$$

Proof. We use a Guerra-type interpolation [10]: Let $0\leq t\leq 1$ and $m\in\mathbb{R}_+$. We suppose to observe Y and $\tilde{Y}$ given by

$$\begin{cases} Y_{i_1,\dots,i_p} = \dfrac{\sqrt{t(p-1)!}}{N^{(p-1)/2}} x^0_{i_1}\dots x^0_{i_p} + V_{i_1,\dots,i_p} & \text{for } 1\leq i_1<\dots<i_p\leq N \\ \tilde{Y}_j \quad = \sqrt{(1-t)m^{p-1}}\, x^0_j + \tilde{V}_j & \text{for } 1\leq j\leq N \end{cases}$$

where the variables $V_{i_1,\dots,i_p}$ and $\tilde{V}_j$ are i.i.d. $\mathcal{N}(0,\Delta)$ random variables. We define the interpolating Hamiltonian

$$\begin{aligned} H_{N,t}(X) = \Delta^{-1}\sum_{i_1<\dots<i_p} \frac{\sqrt{t(p-1)!}}{N^{(p-1)/2}} Y_{i_1,\dots,i_p} x_{i_1}\dots x_{i_p} - \frac{t(p-1)!}{2N^{p-1}}(x_{i_1}\dots x_{i_p})^2 \\ + \Delta^{-1}\sum_{j=1}^{N}\sqrt{(1-t)m^{p-1}}\tilde{Y}_j x_j - \frac{1}{2}(1-t)m^{p-1}x_j^2. \end{aligned}$$

Then, the posterior distribution of X^0 given Y and $\tilde{Y}$ reads

$$P(X^0=X|Y,\tilde{Y}) = \frac{1}{\mathcal{Z}_{N,t}} P_X(X)\exp(H_{N,t}(X))\,, \tag{28}$$

where $\mathcal{Z}_{N,t}$ is the appropriate normalization. Let $\psi_N(t)=\frac{1}{N}\mathbb{E}[\log\mathcal{Z}_{N,t}]$ be the corresponding free energy. Notice that

$$\begin{cases} \psi_N(1) &= F_N\,, \\ \psi_N(0) &= \phi_{\rm RS}(m) - \frac{(1-p)m^p}{2\Delta p}\,. \end{cases}$$

Let $\langle\cdot\rangle_t$ denote the expectation with respect to the posterior (28) and let X be a sampled from (28), independently of everything else.

Using Gaussian integration by parts and the Nishimori identity of Proposition 1 one can show (see [13,14]) that for all $0 \leq t \leq 1$

$$\psi_N'(t) = \frac{1}{2\Delta p}\mathbb{E}\left\langle \left(X \cdot X^0\right)^p - pm^{p-1}\left(X \cdot X^0\right)\right\rangle_t + o_N(1)\,.$$

By convexity of the function $a \mapsto a^p$ on $\mathbb{R}_+$ we have, for all $a, b \geq 0$: $a^p - pb^{p-1}a \geq (1-p)b^p$. We would like to use this inequality with $a = X \cdot X^0$ and $b = m$ to obtain that $\psi_N'(t) \geq \frac{(1-p)m^p}{2\Delta p}$. This would conclude the proof of the lower bound because

$$\begin{aligned}\liminf_{N\to\infty} F_N = \liminf_{N\to\infty} \psi_N(1) &= \liminf_{N\to\infty}\left[\psi_N(0) + \int_0^1 \psi_N'(t)dt\right] \\ &\geq \phi_{\rm RS}(m)\,.\end{aligned}$$

However, we do not know that $X \cdot X^0 \geq 0$ almost surely. To bypass this issue we can add, as in sec. 5.1, a small perturbation (24) that forces $X \cdot X^0$ concentrates around a non-negative value (Lemma 3), without affecting the "interpolating free energy" $\psi_N(t)$ in the $N \to \infty$ limit, see (26). The arguments are the same than in sec. 5.1, so we omit the details and the rewriting of the previous calculations with the perturbation term. This concludes the proof. □

5.3 Proving the upper-bound: Aizenman-Sims-Starr scheme

We are now going to show how the arguments of [14] for the upper bound —using cavity computations with an Aizenman-Sims-Starr approach [1]— can be extended to the tensor case.

Lemma 5.

$$\limsup_{N\to\infty} F_N \leq \sup_{m\geq 0} \phi_{\rm RS}(m)\,.$$

Proof. We are going to compare the system with N variables to the system with $N+1$ variables. Define $A_N = \mathbb{E}[\log \mathcal{Z}_{N+1}] - \mathbb{E}[\log \mathcal{Z}_N]$. Consequently, $F_N = \frac{1}{N}\sum_{k=0}^{N-1} A_k$ and $\limsup F_N \leq \limsup A_N$.

We are thus going to upper-bound A_N. Let $X \in S^N$ be the N-first variables and $\sigma \in S$ the $(N+1)^{\rm th}$ variable. We decompose $H_{N+1}(X,\sigma) = H_N'(X) + \sigma z(X) + \sigma^2 s(X)$ where

$$\begin{aligned}H_N'(X) &= \sum_{i_1<\cdots<i_p} \frac{\Delta^{-1}\sqrt{(p-1)!}}{(N+1)^{(p-1)/2}} Y_{i_1\dots i_p} x_{i_1}\dots x_{i_p} - \frac{\Delta^{-1}(p-1)!}{2(N+1)^{p-1}}(x_{i_1}\dots x_{i_p})^2\,,\\ z(X) &= \Delta^{-1}\sum_{i_1<\cdots<i_{p-1}\leq n} \frac{\sqrt{(p-1)!}}{(N+1)^{(p-1)/2}} Y_{i_1\dots i_{p-1},n+1} x_{i_1}\dots x_{i_{p-1}}\,,\\ s(X) &= -\Delta^{-1}\sum_{i_1<\cdots<i_{p-1}\leq n} \frac{(p-1)!}{2(N+1)^{p-1}}(x_{i_1}\dots x_{i_{p-1}})^2\,.\end{aligned}$$

One can also decompose $H_N(X) = H_N'(X) + y(X)$ in law, where

$$\begin{aligned}y(X) = \Delta^{-1}\sum_{i_1<\cdots<i_p} &\sqrt{(p-1)!}\left(\frac{p-1}{N^p} + r_n\right)^{1/2} V'_{i_1\dots i_p} x_{i_1}\dots x_{i_p} \\ &+ (p-1)!\left(\frac{p-1}{N^p} + r_n\right)\left(x^0_{i_1}\dots x^0_{i_p} x_{i_1}\dots x_{i_p} - \frac{1}{2}(x_{i_1}\dots x_{i_p})^2\right)\,.\end{aligned}$$

In the above definition, the V' are i.i.d. $\mathcal{N}(0,\Delta)$ random variables, independent of everything else, and $r_n = o(N^{-p})$. If we denote by $\langle\cdot\rangle'$ the Gibbs measure on S^N corresponding to the Hamiltonian $\log P_X + H_N'$ we

can rewrite

$$A_N = \mathbb{E}\log\left\langle \sum_{\sigma\in S} P_X(\sigma) e^{\sigma z(X)+\sigma^2 s(X)} \right\rangle' - \mathbb{E}\log\left\langle e^{y(X)}\right\rangle', \tag{29}$$

where X is a sample from $\langle\cdot\rangle'$, independently of everything else. A_N is thus a difference of two terms that will correspond exactly to the terms of (21). As in sec. 5.1, we can show that under a small perturbation of the system, the overlap $X \cdot X^0$ with the planted configuration concentrates around a non-negative value Q'. This leads to simplifications in (29):

$$\limsup_{N\to\infty} A_N \leq \limsup_{N\to\infty} \mathbb{E}[\phi_{\rm RS}(Q')] \leq F_{\rm RS}\,. \tag{30}$$

For a precise derivation of (30), the reader is invited to report to the matrix case (see [14], sec. 4.6), since there is no major difference with the tensor case on this point. The arguments presented there are commonly used in the study of spin glasses and are the analog of cavity computations in the SK model developed in [26], sec. 1.5. This concludes the proof. □

6 Examples of phase transitions

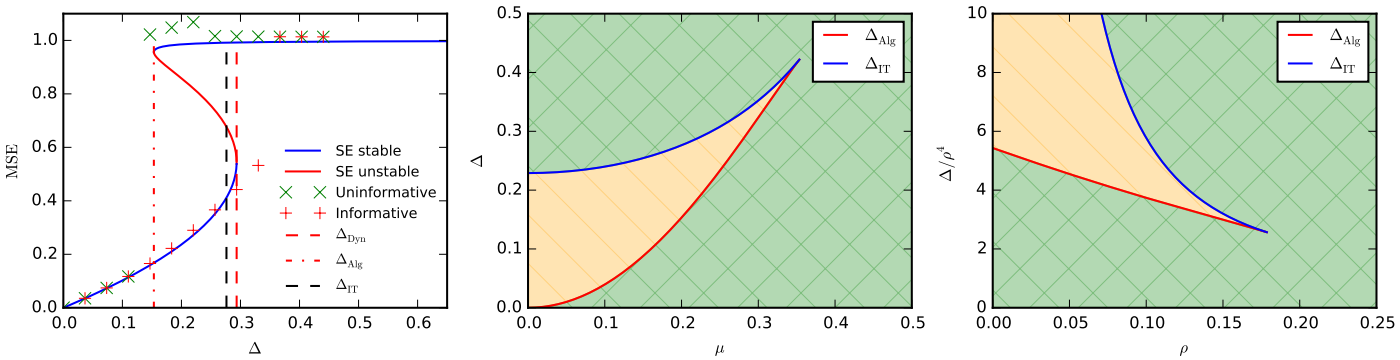

Figure 1: **Left panel:** Comparison between the AMP fixed point reached from uninformative (marked with crosses) or informative (i.e. strongly correlated with the ground truth, marked with pluses) initialization and the fixed point of the SE equations (stable fixed point in blue, unstable in red). The data are for the Gaussian prior with mean $\mu = 0.2$, unit variance, $p = 3$, $r = 1$. The AMP runs are done on a system of size $N = 1000$. **Central panel:** Phase diagram for the order $p = 3$ tensor factorization, rank $r = 1$, Gaussian prior of mean μ (x-axes) and unit variance. In the green-shaded zone AMP matches the information-theoretically optimal performance, $\text{MMSE} = \text{MSE}_{\text{AMP}}$. In the orange-shaded zone $\text{MMSE} < \text{MSE}_{\text{AMP}}$. The tri-critical point is located at $\mu_{\text{Tri}} = (p-2)/(2\sqrt{p-1})$ and $\Delta_{\text{Tri}} = x_{\text{Tri}}^{p-2}/(1+x_{\text{Tri}})^{p-1}$ where $x_{\text{Tri}} = (p-2)(3p-4)/p^2$. **Right panel:** Phase diagram for the order $p = 3$ tensor factorization, rank $r = 1$, the Bernoulli prior as a function of ρ and Δ/ρ^4. The tri-critical point is located at $\rho_{\text{Tri}} = 0.178$ and $\Delta_{\text{Tri}}/\rho^4 = 2.60$. As $\rho \to 0$ we observed $\Delta_{\text{Alg}}/\rho^4 \to 2e$. Compare to Fig. 5 in [17] where the same phase diagram is presented for the matrix factorization $p = 2$ case.

We used the state evolution eqs. (12-13), and the free energy (16), to compute the values of the thresholds Δ_c, Δ_{IT} and Δ_{Alg} for several examples of the prior distributions: Gaussian (spherical spins), $P_X(x) = \mathcal{N}(\mu, 1)$; Rademacher (Ising spins), $P_X(x) = \frac{1}{2}[\delta(x-1) + \delta(x+1)]$; Bernoulli (sub-tensor localization), $P_X(x) = \rho\delta(x-1) + (1-\rho)\delta(x)$; and clustering (tensor stochastic block model), $P_X(x) = \frac{1}{r}\sum_{k=1}^{r}\delta(x-\vec{e}_k)$, where $\vec{e}_k \in \mathbb{R}^r$ is a vector with a 1 at coordinate k and 0 elsewhere. Examples of values of the thresholds for the above priors are given in Table 1. For the zero mean Gaussian and the Rademacher prior our results for Δ_{IT} indeed agree with those presented in [12, 22]. Central and right part of Fig. 1 present the thresholds for

the Gaussian and Bernoulli prior as a function of the mean μ and density ρ, respectively. Left part of Fig. 1 illustrates that indeed the fixed points of the state evolution agree with the fixed points of the AMP algorithm.

Prior / p	Gaussian $\mathcal{N}(0,1)$		Rademacher		Bernoulli $\rho = 0.1$		3 clusters	
	$\Delta_{\rm IT} p\log(p)$	$\Delta_{\rm Alg}$	$\Delta_{\rm IT}$	$\Delta_{\rm Alg}$	$\Delta_{\rm IT}\rho^{-p}$	$\Delta_{\rm Alg}\rho^{-2p+2}$	$\frac{\Delta_{\rm IT}}{\Delta_{\rm Alg}}$	$\frac{\Delta_{\rm Alg} r^{2p-2}}{p-1}$
2	$2\log 2$	1	1	1	–	–	1	1
3	0.754	0	0.2828	0	0.577	3.738	1	1
4	0.701	0	0.1902	0	0.398	6.017	1.18	1
5	0.685	0	0.1473	0	0.311	8.251	1.62	1
10	0.677	0	0.07216	0	0.154	19.30	6.59	1

Table 1: Examples of the information-theoretic $\Delta_{\rm IT}$ and algorithmic $\Delta_{\rm Alg}$ thresholds for order-p tensor decomposition for different priors on the factors. For the Gaussian case $\Delta_{\rm IT} p\log(p)$ converges to 1 at large p. For the Bernoulli case the rescaling in power of ρ is for convenience to present quantities of order one, we did not check if it describes the large p limit.

6.1 Results for Gaussian prior

In this section we detail the analysis of the state evolution for rank $r = 1$ Gaussian prior of mean μ and variance 1.

$$P_X^{\rm Gauss} = \mathcal{N}(\mu, 1) \ . \tag{31}$$

Using (12) one gets for the SE equation

$$M^{t+1} = \frac{\Delta\mu^2 + (M^t)^{p-1}(1+\mu^2)}{\Delta + (M^t)^{p-1}} \ , \tag{32}$$

where M is a scalar, and Δ is the inverse Fisher information of the output channel. It turns out that as soon as $p \geq 3$ the SE equation exhibits multiple stable fixed points.

For the zero mean $\mu = 0$ case one gets

$$M^{t+1} = \frac{(M^t)^{p-1}}{\Delta + (M^t)^{p-1}} \ . \tag{33}$$

Here the fixed point $M = 0$ is stable whatever the noise $\Delta > 0$ and therefore AMP will not achieve performance better than random guessing for any $\Delta > 0$. Ref. [24] studies the scaling of Δ with N for which AMP and other algorithms succeed.

For positive mean $\mu > 0$, however, the AMP algorithm is able to recover the signal for values of $\Delta < \Delta_{\rm Alg}$ with

$$\Delta_{\rm Alg}(\mu) = \frac{x_{\rm Alg}^{p-2}}{(1+x_{\rm Alg})^{p-1}}, \quad \Delta_{\rm Dyn}(\mu) = \frac{x_{\rm Dyn}^{p-2}}{(1+x_{\rm Dyn})^{p-1}}, \tag{34}$$

$$x_{\rm Alg}(\mu) = \frac{p-2+2\mu^2 - \sqrt{(p-2)^2 - 4\mu^2(p-1)}}{2(1+\mu^2)}, \tag{35}$$

$$x_{\rm Dyn}(\mu) = \frac{p-2+2\mu^2 + \sqrt{(p-2)^2 - 4\mu^2(p-1)}}{2(1+\mu^2)}, \tag{36}$$

where we defined a new threshold $\Delta_{\rm Dyn}$ as the smallest such that for $\Delta > \Delta_{\rm Dyn}$ the state evolution has a

unique fixed point. We know of no analytical formula for $\Delta_{\rm IT}$ and for Figure 1 we computed it numerically. The tri-critical point where all these curve meet is located at

$$\mu_{\rm Tri} = \frac{p-2}{2\sqrt{p-1}}\,. \tag{37}$$

Using the above expressions we derive that

$$\Delta_{\rm Dyn}(\mu=0) \quad = \quad \frac{1}{p-2}\left(\frac{p-2}{p-1}\right)^{p-1} \underset{p\to\infty}{\sim} \frac{1}{ep}\,, \tag{38}$$

$$\Delta_{\rm Alg}(\mu) \quad \underset{\mu\to 0}{\sim} \quad \left(\frac{\mu^2}{p-2}\right)^{p-2}\,. \tag{39}$$

We can also compute the limit of the $\Delta_{\rm IT}(\mu=0,p)$ as $p\to\infty$ and get

$$\Delta_{\rm IT}(\mu=0,p) \underset{p\to\infty}{\sim} \frac{1}{p\log(p)}\,. \tag{40}$$

This scaling agrees with the large p behavior derived in [24] and [22].

6.2 Results for clustering prior

An interesting example of the prior for rank r tensor estimation is

$$P_X^{\rm Clusters}(x) = \frac{1}{r}\sum_{1\le k\le r}\delta(x-\vec{e}_k)\,. \tag{41}$$

This describes a model of r non-overlapping clusters. Due to the channel universality, this prior also describes the stochastic block model on dense hyper-graphs as considered for sparse hyper-graph in e.g. [3]. This model was considered in detail for $p=2$ in [17].

The above clustering prior has non-zero mean, and it also exhibits the transition Δ_c from a phase where recovery of clusters better than chance is not possible, to a phase where it is.

To analyze the SE equations we first notice that the stable fixed point will be of the form

$$M = \frac{bI_r}{r} + \frac{(1-b)J_r}{r^2} \in \mathbb{R}^{r\times r}, b\in[0;1]\,, \tag{42}$$

where I_r is the identity matrix and J_r is a matrix filled with ones. $b=0$ means that the estimate of the marginals does not carry any information. $b=1$ means perfect reconstruction. The state evolution now becomes

$$b^{t+1} = \mathcal{M}_r\left(r\frac{\left(\frac{b^t}{r}+\frac{1-b^t}{r^2}\right)^{p-1}-\left(\frac{1-b^t}{r^2}\right)^{p-1}}{\Delta}\right)\,, \tag{43}$$

where $\mathcal{M}_r$ is a function that was defined and studied in [15]. Its Taylor expansion is

$$\mathcal{M}_r(x) = \frac{x}{r^2} + x^2\frac{r-4}{2r^4} + O\left(x^3\right)\,. \tag{44}$$

We further notice that $b=0$ is always a fixed point of (43). By expanding (43) to first order one gets

$$b^{t+1} = b^t\frac{p-1}{\Delta r^{2p-2}} + O\left(b^{t2}\right)\,. \tag{45}$$

This fixed point therefore becomes unstable when

$$\Delta < \Delta_c \equiv \frac{p-1}{r^{2p-2}}\,. \tag{46}$$

By analyzing eq. (43) further we can prove that $\forall x \in \mathbb{R}^+$

$$m(x) \;\; = \;\; \mathcal{M}_r(x) \tag{47}$$

$$\Delta(x) = r\frac{\left(\frac{\mathcal{M}_r(x)}{r} + \frac{1-\mathcal{M}_r(x)}{r^2}\right)^{p-1} - \left(\frac{1-\mathcal{M}_r(x)}{r^2}\right)^{p-1}}{x}. \tag{48}$$

m is a fixed point of (43) when $m = m(x)$ and $\Delta = \Delta(x)$. Rather than finding the fixed point iteratively, the above equations allow us to draw all the fixed point of (43), be it stable or unstable We have that $m(x)$ is a stable fixed point of (43) if and only if

$$\frac{\partial \Delta(x)}{\partial x} < 0\,. \tag{49}$$

The next question is whether there is a first or second order phase transition at Δ_c. To answer this, one needs to analyze whether the fixed point close to $b = 0$ is stable or unstable. For this we to compute $\frac{\partial \Delta(x)}{\partial x}$ at $x = 0$ to get using (44) that

$$\frac{\partial \Delta(x)}{\partial x} = \frac{p-1}{2r^{2p}}\left(-2p - r + pr\right)\,. \tag{50}$$

Therefore if $-2p - r + pr > 0$ there will be no stable fixed point close to $b = 0$ and the system must have a first order phase transition (discontinuity in the $\mathrm{MSE}_{\mathrm{AMP}}$) at $\Delta_c = \Delta_{\mathrm{Alg}}$.

For two clusters $r = 2$, there is a second order phase transition at Δ_c for all $p \geq 2$. However, analyzing the state evolution numerically we observed that for $p \geq 5$ there is a discontinuity later at some $\Delta_{\mathrm{Alg}} < \Delta_c$. For three and more clusters $r \geq 3$ we always have $\Delta_{\mathrm{Alg}} = \Delta_c$, and for $-2p - r + pr \leq 0$ we have not detected any other discontinuities. Values of Δ_{IT} for three clusters and some values of p are given in Table 1.

Acknowledgment

This work has been supported by the ERC under the European Union's FP7 Grant Agreement 307087-SPARCS.

References

[1] Michael Aizenman, Robert Sims, and Shannon L Starr. Extended variational principle for the Sherrington-Kirkpatrick spin-glass model. *Physical Review B*, 68(21):214403, 2003.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[3] Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborová. Spectral detection on sparse hypergraphs. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing*, page 66, 2015.

[4] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Inf. Proc. Systems*, page 424, 2016.

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

[6] Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163, 2015.

[7] A Crisanti and H-J Sommers. Thouless-anderson-palmer approach to the spherical p-spin spin glass model. *J. de Phys. I*, 5:805, 1995.

[8] Andrea Crisanti and H-J Sommers. The spherical p-spin interaction spin glass model: The statics. *Zeitschrift für Physik B Condensed Matter*, 87(3):341–354, 1992.

[9] Y. Deshpande and A. Montanari. Information-theoretically optimal sparse PCA. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201, 2014.

[10] Francesco Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Commun. Math. Phys.*, 233:1, 2003.

[11] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

[12] Satish Babu Korada and Nicolas Macris. Exact solution of the gauge symmetric p-spin glass model on a complete graph. *Journal of Statistical Physics*, 136(2):205–230, 2009.

[13] F. Krzakala, J. Xu, and L. Zdeborová. Mutual information in rank-one matrix estimation. In *IEEE Inf. Theory Workshop*, pages 71–75, 2016.

[14] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *arXiv:1611.03888*, 2016.

[15] T. Lesieur, F. Krzakala, and L. Zdeborová. MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *53rd Annual Allerton Conf. on Comm., Control, and Comp.*, 2015.

[16] T. Lesieur, F. Krzakala, and L. Zdeborová. Phase transitions in sparse PCA. In *IEEE Int. Symp. on Inf. Theory*, page 1635, 2015.

[17] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.

[18] Ryosuke Matsushita and Toshiyuki Tanaka. Low-rank matrix reconstruction and clustering via approximate message passing. In *Advances in Neural Information Processing Systems*, page 917. 2013.

[19] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin-Glass Theory and Beyond*, volume 9. World Scientific, 1987.

[20] Andrea Montanari. Estimating random variables from random sparse observations. *Trans. Emerging Tel. Tech.*, 19(4):385, 2008.

[21] Dmitry Panchenko. *The Sherrington-Kirkpatrick model*. Springer Science & Business Media, 2013.

[22] Amelia Perry, Alexander S Wein, and Afonso S Bandeira. Statistical limits of spiked tensor models. *arXiv preprint arXiv:1612.07728*, 2016.

[23] Sundeep Rangan and Alyson K Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *IEEE Int. Symp. on Inf. Theory*, pages 1246–1250, 2012.

[24] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, page 2897, 2014.

[25] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

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

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

[28] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65:453–552, 2016.