Constrained Low-rank Matrix Estimation: Phase Transitions, Approximate Message Passing and Applications

Thibault Lesieur, Florent Krzakala, Lenka Zdeborová

Constrained Low-rank Matrix Estimation: Phase Transitions, Approximate Message Passing and Applications.

Thibault Lesieur[1], Florent Krzakala[2], and Lenka Zdeborová[1,*]
[1] *Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, F-91191, Gif-sur-Yvette, France.*
[2] *Laboratoire de Physique Statistique, Ecole Normale Supérieure, & Université Pierre et Marie Curie, Sorbonne Universités.*
[*] *To whom correspondence shall be sent: lenka.zdeborova@cea.fr*

(Dated: January 5, 2017)

This article is an extended version of previous work of the authors [40, 41] on low-rank matrix estimation in the presence of constraints on the factors into which the matrix is factorized. Low-rank matrix factorization is one of the basic methods used in data analysis for unsupervised learning of relevant features and other types of dimensionality reduction. We present a framework to study the constrained low-rank matrix estimation for a general prior on the factors, and a general output channel through which the matrix is observed. We draw a paralel with the study of vector-spin glass models – presenting a unifying way to study a number of problems considered previously in separate statistical physics works. We present a number of applications for the problem in data analysis. We derive in detail a general form of the low-rank approximate message passing (Low-RAMP) algorithm, that is known in statistical physics as the TAP equations. We thus unify the derivation of the TAP equations for models as different as the Sherrington-Kirkpatrick model, the restricted Boltzmann machine, the Hopfield model or vector (xy, Heisenberg and other) spin glasses. The state evolution of the Low-RAMP algorithm is also derived, and is equivalent to the replica symmetric solution for the large class of vector-spin glass models. In the section devoted to result we study in detail phase diagrams and phase transitions for the Bayes-optimal inference in low-rank matrix estimation. We present a typology of phase transitions and their relation to performance of algorithms such as the Low-RAMP or commonly used spectral methods.

CONTENTS

I. INTRODUCTION

A. Problem Setting

In this paper we study a generic class of statistical physics models having Boltzmann probability measure that can be written in one of the two following forms:

- **Symmetric vector-spin glass model**:

$$P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\le i\le N} P_X(x_i) \prod_{1\le i<j\le N} e^{g(Y_{ij}, x_i^\top x_j/\sqrt{N})}\,. \tag{1}$$

  Here $Y_{ij} \in \mathbb{R}^{N\times N}$ and $X \in \mathbb{R}^{N\times r}$ are real valued matrices. In this case Y_{ij} is a symmetric matrix. In statistical physics Y is called the quenched disorder. In the whole paper we denote by $x_i \in \mathbb{R}^r$ the *vector-spin* i (r-dimensional column vector) that collects the elements of the ith row of the matrix X, $x_i^\top x_j$ is the scalar product of the two vectors. $Z_X(Y)$ is the corresponding partition function playing role of the normalization.

- **Bipartite vector-spin glass model**:

$$P(U,V|Y) = \frac{1}{Z_{UV}(Y)} \prod_{1\le i\le N} P_U(u_i) \prod_{1\le j\le M} P_V(v_j) \prod_{1\le i\le N, 1\le j\le M} e^{g(Y_{ij}, u_i^\top v_j/\sqrt{N})}\,. \tag{2}$$

  Defined as above, this time $Y_{ij} \in \mathbb{R}^{N\times M}$ and $U \in \mathbb{R}^{N\times r}$, $V \in \mathbb{R}^{M\times r}$. Again we denote by u_i, v_j the *vector-spins* of dimension r that collect rows of matrices U, V. In this case the graph of interactions between spins is bipartite.

The main motivation on this work is twofold. On the one hand, the above mentioned probability measures are posterior probability measures of an important class of high-dimensional inference problems known as constrained low-rank matrix estimation. In what follows we give examples of applications of these matrix estimation problems in data processing and statistics. On the other hand, our motivation from the physics point of view is to present a unified formalism providing the (replica symmetric) solution for a large class of mean-field vectorial spin models with disorder.

The general nature of the present work stems from the fact that the probability distributions P_X, P_U, P_V and the function g are very generic (assumptions are summarize in section I C). These functions can even depend on the node i or edge ij. For simplicity we will treat site-independent functions P_X, P_U, P_V and g, but the theory developed here generalizes very straightforwardly to the site or edge dependent case. From a statistical physics point of view the terms P_X, P_U, P_V play a role of generic local magnetic fields acting on the individual spins. Distributions P_X, P_U, P_V describe the nature of the vector-spin variables and the fields that act on them. The simplest example is the widely studied Ising spins for which $r = 1$ and $P_X(x) = \rho\delta(x-1) + (1-\rho)\delta(x+1)$, where ρ here would be related to the usual magnetic field h and inverse temperature β as $\rho = e^{\beta h}/(2\cosh\beta h)$. In this paper we treat a range of other examples with $r \ge 1$ and elements of x being both discrete or continuous. This involves for instance spherical spin modeld with P_X being Gaussian, or Heisenberg spins where $r = 3$ and each x is confined to the surface of a sphere.

Denoting

$$w_{ij} = x_i^\top x_j/\sqrt{N}\,, \quad \text{or} \quad w_{ij} = u_i^\top v_j/\sqrt{N} \tag{3}$$

according to symmetric or bipartite context, the terms $g(Y,w)$ are then interactions between pairs of spins that depend only on the scalar product between the corresponding vectors. The most commonly considered form of interaction in statistical physics is simply

$$g(Y,w) = \beta Y w \tag{4}$$

with β being a constant called inverse temperature, leading to a range of widely considered models with pair-wise interactions. We will refer to this form of function as the *conventional Hamiltonian*.

In order to complete the setting of the problem we need to specify how is the quenched disorder Y chosen. We will consider two main cases of the quenched disorder defined below. We note that even for problems where the matrix Y is not generated by either of the below our approach might still be relevant, e.g. for the restricted Boltzmann machine that is equivalent to the above bipartite model with Y that were learned from data (see for instance [20, 64] and the discussion below).

- **Randomly quenched disorder:** In this case the matrix elements of Y are chosen independently at random from some probability distribution $P(Y_{ij})$. In this paper we will consider this distribution to be independent of N and in later parts for simplicity we will restrict its mean to be zero. This case of randomly quenched disorder will encompass many well known and studied spin glass models such as the Ising spin glass, Heisenberg spin glass or the spherical spin glass or the Hopfield model.
- **Planted models:** Concerning applications in data science this is the more interesting case and most of this paper will focus on it. In this case we consider that there is some *ground truth* value of $X_0 \in \mathbb{R}^{N\times r_0}$ (or $U_0 \in \mathbb{R}^{N\times r_0}$, $V_0 \in \mathbb{R}^{M\times r_0}$) with rows that are generated independently at random from some probability distribution P_{X_0} (or P_{U_0}, P_{V_0}). Then the disorder Y is generated element-wise as a noisy observation of the product $w^0_{ij} = x^{0,\top}_i x^0_j/\sqrt{N}$ (or $w^0_{ij} = u^{0,\top}_i v^0_j/\sqrt{N}$) via an output *channel* characterized by the output probability distribution $P_{\rm out}(Y_{ij}|w^0_{ij})$.

B. Preliminaries on the planted setting

1. Bayes optimal inference

Many applications, listed and analyzed below, in which the planted setting is relevant, concern problems where we aim to infer some ground truth matrices X_0, U_0, V_0 from the observed data Y and from the information we have about the distributions P_{X_0}, P_{U_0}, P_{V_0} and $P_{\rm out}$. The information-theoretically optimal way of doing inference if we know how the data Y and how the ground-truth matrices were generated is to follow the Bayesian inference and compute marginals of the corresponding posterior probability distribution. According to the Bayes formula, the posterior probability distribution for the symmetric case is

$$P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\le i\le N} P_{X_0}(x_i) \prod_{1\le i<j\le N} P_{\rm out}\left(Y_{ij}\middle|\frac{x_i^\top x_j}{\sqrt{N}}\right). \tag{5}$$

For the bipartite case it is

$$P(U,V|Y) = \frac{1}{Z_{UV}(Y)} \prod_{1\le i\le N} P_{U_0}(u_i) \prod_{1\le j\le M} P_{V_0}(v_j) \prod_{1\le i\le N, 1\le j\le M} P_{\rm out}\left(Y_{ij}\middle|\frac{u_i^\top v_j}{\sqrt{N}}\right). \tag{6}$$

Making link with the Boltzmann probability measures (1) and (2) we see that the *Bayes optimal inference* of the planted configuration is equivalent to the statistical physics of the above vector-spin models with

$$P_{X_0} = P_X\,, \quad P_{U_0} = P_U\,, \quad P_{V_0} = P_V\,, \quad P_{\rm out}(Y|w) = e^{g(Y,w)}\,. \tag{7}$$

This approach is optimal in the sense that the statistical estimator $\hat{X}$ computed from the data Y that minimizes the expected mean-squared error between the estimator $\hat{X}$ and the ground truth X_0 is given by the mean of the marginal of variable x_i in the probability distribution (5)

$$\hat{x}_i(Y) = \int {\rm d}x\, x\, \mu_i(x)\,, \quad \text{where} \quad \mu_i(x) = \int P(X|Y) \prod_{\{x_j\}_{j\neq i}} {\rm d}x_j\,. \tag{8}$$

Analogously for the bipartite case.

In the Bayes-optimal setting defined by conditions (7) the statistical physics analysis of the problem presents important simplifications known as the Nishimori conditions [50, 70], which will be largely used in the present paper. These conditions can be proven and stated without the usage of the methodology developed below, they are a direct consequence of the Bayesian formula for conditional probability and basic properties of probability distributions.

Assume Bayes-optimality of the output channel, that is $P_{\rm out} = e^{g(Y,w)}$. First let us notice that every probability distribution has to be normalized

$$\forall w, \int {\rm d}Y P_{\rm out}(Y|w) = 1\,. \tag{9}$$

By deriving the above equation with respect to w one gets.

$$\forall w, \int {\rm d}Y P_{\rm out}(Y|w) \frac{\partial g(Y,w)}{\partial w} = \mathbb{E}_{P_{\rm out}(Y|w)}\left[\frac{\partial g(Y,w)}{\partial w}\right] = 0\,, \tag{10}$$

$$\forall w, \int {\rm d}Y P_{\rm out}(Y|w) \left[\left(\frac{\partial g(Y,w)}{\partial w}\right)^2 + \frac{\partial^2 g(Y,w)}{\partial w^2}\right] = \mathbb{E}_{P_{\rm out}(Y,w)}\left[\left(\frac{\partial g(Y,w)}{\partial w}\right)^2 + \frac{\partial^2 g(Y,w)}{\partial w^2}\right] = 0\,. \tag{11}$$

Anticipating the derivation in the following we also define the inverse Fisher information of an output channel $P_{\rm out}$ at $w = 0$ as

$$\frac{1}{\Delta} = \mathbb{E}_{P_{\rm out}(Y|w=0)}\left[\left(\frac{\partial g}{\partial w}\right)^2_{Y,w=0}\right], \tag{12}$$

Let us now assume Bayes-optimality also for the prior distributions, i.e. $r = r_0$ and $P_{X_0} = P_X$, $P_{U_0} = P_U$, $P_{V_0} = P_V$. Then under averages over the posterior distribution (5) or (6) we can replace the random variables X, U, V for the ground truth X_0, U_0, V_0 and vice versa. To prove this let us consider the symmetric case, the generalization to the bipartite case is straightforward. Take X, X_1, X_2 to be three independent samples from the posterior probability distribution $P(X|Y)$, eq. (5). We then consider some function $f(A, B)$ of two configurations of the variables A, B. Consider the following two expectations

$$\mathbb{E}\left[f(X_1, X_2)\right] = \int f(X_1, X_2)P(Y)P(X_1|Y)P(X_2|Y)\mathrm{d}X_1\,\mathrm{d}X_2\,\mathrm{d}Y\,, \tag{13}$$

$$\begin{aligned}\mathbb{E}\left[f(X_0, X)\right] &= \int f(X_0, X)P(X_0, X)\mathrm{d}X\,\mathrm{d}X_0 = \int f(X_0, X)P(X_0, X, Y)\mathrm{d}X\,\mathrm{d}X_0\mathrm{d}Y \\ &= \int f(X_0, X)P(X|Y, X_0)P_{\rm out}(Y|X_0)P_0(X_0)\mathrm{d}X\,\mathrm{d}X_0\mathrm{d}Y\,. \end{aligned} \tag{14}$$

where we used the Bayes formula. We further observe that $P(X|Y, X_0) = P(X|Y)$ because X is independent of X_0 when conditioned on Y. In the Bayes optimal case, i.e. when $P(X) = P_0(X)$ and $P(Y|X) = P_{\rm out}(Y|X)$, we then obtain

$$\text{Bayes optimal}: \qquad \mathbb{E}\left[f(X_1, X_2)\right] = \mathbb{E}\left[f(X_0, X)\right]\,, \tag{15}$$

meaning that under expectations there is no statistical difference between the ground truth assignment of variables X_0 and an assignment sampled uniformly at random from the posterior probability distribution (5). This is a simple yet important property that will lead to numerous simplifications in the Bayes optimal case and it will be used in several places of this paper, under the name *Nishimori condition*.

From the point of view of statistical physics of disordered systems the most striking property of systems that verify the Nishimori conditions is that there cannot be any *replica symmetry breaking* in the equilibrium solution of these systems [50, 51, 70]. This simplifies considerably the analysis of the Bayes-optimal inference. Note, however, that metastable (out-of-equilibrium) properties of Bayes-optimal inference do not have to satisfy the Nishimori conditions and replica symmetry breaking might be needed for their correct description (this will be relevant in the cases of first order phase transition described in section IV C).

2. Principal component analysis

In the above probabilistic inference setting the Bayesian approach of computing the marginals is optimal. However, in a majority of the interesting cases it is computationally intractable (NP hard) to compute the marginals exactly. In all low-rank estimation problems the method that (for the bipartite case) comes to mind as first when we look for a low-rank matrix close to an observe matrix Y is the *singular value decomposition* (SVD) where a rank r matrix $\tilde{Y}$ that minimizes the squared difference in computed

$$\mathrm{argmin}_{\tilde{Y}}\left[\sum_{i,j}(Y_{ij} - \tilde{Y}_{ij})^2\right] = \sum_{s=1}^{r} u_s\lambda_s v_s^\top\,, \tag{16}$$

where λ_s is the sth largest singular value of Y, and $u_s \in \mathbb{R}^N$, $v_s \in \mathbb{R}^M$ are the corresponding left-singular and right-singular vectors of the matrix Y. The above property is know as the Eckart-Young-Mirsky theorem [18]. In the symmetric case $Y_{ij} = Y_{ji}$ we simply replace the singular values by eigenvalues and the singular vectors by the eigenvectors. The above unconstrained low-rank approximation of the matrix Y, eq. (16), is also often referred to as *principal component analysis* (PCA), because indeed when the matrix Y is interpreted as N samples of M-dimensional data then the right-singular vectors v_s are directions of greatest variability in the data.

PCA and SVD are methods of choice when the measure of performance is the sum of square differences between the observe and estimated matrices, and when there are no particular requirements on the elements of the matrices U, V or X.

The methodology developed in this paper for the planted probabilistic models, generalizes to arbitrary cost function that can be expressed as a product of element-wise terms $e^{g(Y_{ij},w_{ij})}$ and to arbitrary constraints on the rows of the matrices U, V, X as long as they can be described by row-wise probability distributions P_U, P_V, P_X. Systematically comparing our results to the performance of PCA is useful because PCA is well known and many researcher have good intuition about what are its strengths and limitations in various settings.

C. The large size limit, assumptions and channel universality

In this article we focus on the thermodynamic limit where $N, M \to \infty$ whereas $r = O(1)$, and $\alpha \equiv M/N = O(1)$ and all the elements of Y, X, U and V are of order 1. The functions P_X, P_U, P_V and g do not depend on N explicitly. In the planted model also the distribution P_{X_0}, P_{U_0}, P_{V_0} and $P_{\rm out}$ do not depend on N explicitly. The only other requirement we impose on the distributions P_X, P_U, P_V and P_{X_0}, P_{U_0}, P_{V_0} is that they all have a finite second moment.

The factor $1/\sqrt{N}$ in the second argument of the function g ensures that the behaviour of the above models is non-trivial and that there is an interesting competition between the number $O(N)$ of local magnetic fields P_X, P_U, P_V and the number of $O(N^2)$ interactions. To physics readership familiar with the Sherrington-Kirkpatrick (SK) model this $1/\sqrt{N}$ factor will be familiar because in the SK model the interaction between the Ising spins that lead to extensive free energy are also of this order (with mean that is of order $1/N$). This is compared to the ferromagnetic Ising model on a fully connected lattice for which the interactions leading to extensive free energy scale as $1/N$.

For readers interested in inference problems, i.e. the planted setting, the $1/\sqrt{N}$ factor is the scaling of the signal-to-noise ratio for which inference of $O(N)$ unknown from $O(N^2)$ measurements is neither trivially easy nor trivially impossible. In the planted setting Y can be viewed as a random matrix with a rank-r perturbation. The regime where the eigenvalues of dense random matrices with low-rank perturbations split from the bulk of the spectral density is precisely when the strength of the perturbation is $O(1/\sqrt{N})$, see e.g. [2].

We are therefore looking at statistical physics models with $O(N^2)$ pairwise interactions where each of the interactions depend only weakly on the configuration of the vector-spins. As a consequence, properties of the system in the thermodynamic limit $N \to \infty$ depend only weakly on the details of the interaction function $g(Y_{ij}, w_{ij})$ with w_{ij} given by (3). The results of this paper hold for every function g for which the following Taylor expansion is well defined

$$e^{g(Y_{ij},w_{ij})} = e^{g(Y_{ij},0)} \left\{ 1 + \frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} w_{ij} + \left[\left(\frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} \right)^2 + \frac{\partial^2 g(Y_{ij},w)}{\partial w^2}\Big|_{w=0} \right] \frac{w_{ij}^2}{2} + O(w_{ij}^3) \right\} . \tag{17}$$

In order to simplify the notation in the following we denote

$$S_{ij} \equiv \frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} , \tag{18}$$

$$R_{ij} \equiv \left(\frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} \right)^2 + \frac{\partial^2 g(Y_{ij},w)}{\partial w^2}\Big|_{w=0} . \tag{19}$$

We will refer to the matrix S as the Fisher score matrix. The above expansion can now be written in a more compact way

$$e^{g(Y_{ij},w_{ij})} = e^{g(Y_{ij},0)} \left[1 + S_{ij} w_{ij} + \frac{R_{ij} w_{ij}^2}{2} + O(w_{ij}^3) \right] = e^{g(Y_{ij},0) + S_{ij} w_{ij} + \frac{1}{2}(R_{ij} - S_{ij}^2) w_{ij}^2 + O(w_{ij}^3)} . \tag{20}$$

Let us now analyze the orders in this expansion. In the Boltzmann measure (1) and (2) the terms $e^{g(Y_{ij},w_{ij})}$ appears in a product over $O(N^2)$ terms and $w = O(1/\sqrt{N})$. At the same time only terms of order $O(N)$ in the exponent of the Boltzmann measure influence the leading order (in N) of the marginals (local magnetizations), therefore all the terms that depend on 3rd and higher order of w are negligible. This means that the leading order of the marginals depend on the function $g(Y, w)$ only trough the matrices of its first and second derivatives at $w = 0$, denoted S and R (18-19). This means in particular that in order to understand the phase diagram of a model with general $g(Y, w)$ we only need to consider one more order than in the conventional Hamiltonian considered in statistical physics (4).

In the sake of specificity let us state here the two examples of the output channels $g(Y, w)$ considered most prominently in this paper and their corresponding matrices S and R. The first example corresponds to observations of low-rank matrices with additive Gaussian noise, we will refer to this as the Gaussian output channel

$$\textbf{Gaussian Noise Channel}: \qquad g(Y,w) = \frac{-(Y-w)^2}{2\Delta} - \frac{1}{2}\log 2\pi\Delta , \qquad S_{ij} = \frac{Y_{ij}}{\Delta} , \qquad R_{ij} = \frac{Y_{ij}^2}{\Delta^2} - \frac{1}{\Delta} , \tag{21}$$

where Δ is the variance of the noise, for the specific case of the Gaussian output channel Δ is also the Fisher information as defined in eq. (12). The second example is the one most often considered in physics given by eq. (4)

$$\textbf{Conventional Hamiltonian}: \qquad g(Y,w) = \beta Y w\,, \qquad S_{ij} = \beta Y_{ij}\,, \qquad R_{ij} = \beta^2 Y_{ij}^2\,. \tag{22}$$

with β being a constant called inverse temperature. Another quite different example of the output channel will be given to model community detection in networks in section I D 2 d.

D. Examples and applications

The Boltzmann measures (1) and (2) together with the model for the disorder Y can be used to describe a range of problems of practical and scientific interest studied previously in physics and/or in data sciences. In this section we list several examples and applications for each of the four categories – the symmetric and bipartite case, and the randomly quenched and planted disorder.

1. Examples with randomly quenched disorder

a. **Sherrington-Kirkpatrick (SK) model.** The SK model [61] stands at the roots of the theory of spin glasses. It can be described by the symmetric Boltzmann measure (1) with the conventional Hamiltonian $g(Y,w) = \beta Y w$.

The x_i are Ising spins, i.e. $x_i \in \{\pm 1\}$, with distribution

$$P_X(x_i) = \rho\delta(x_i - 1) + (1-\rho)\delta(x_i + 1). \tag{23}$$

The parameter ρ is related to the inverse temperature β and an external magnetic field h as $\rho = e^{\beta h}/(2\cosh\beta h)$. Note that the parameter ρ could also be site-dependent and our approach would generalize, but in this paper we work with site independent functions P_X.

The elements of the (symmetric) matrix Y_{ij} are the quenched random disorder, i.e. they are generated independently at random from some probability distribution. Most usually considered distributions of disorder would be the normal distribution $Y_{ij} \sim \mathcal{N}(0,1)$, or binary $Y_{ij} = 1$ with probability $1/2$ and $Y_{ij} = -1$ otherwise.

The algorithm developed in this paper for the general case corresponds to the Thouless-Anderson-Palmer [63] equations for the SK model. The theory developed here correspond to the replica symmetric solution of [61]. Famously this solution is wrong below certain temperature where effects of replica symmetry breaking (RSB) have to be taken into account. In this paper we focus on the replica symmetric solution, that leads to exact and novel phase diagrams for the planted models. The RSB solution in the present generic setting will be presented elsewhere. We present the form of the TAP equations in the general case encompassing a range of existing works.

b. **Spherical spin glass.** Next to the SK model, the spherical spin glass model [32] stands behind large fraction of our understanding about spin glass. Mathematically much simpler than the SK model this model stands as a prominent case in the development in mathematical physics. The spherical spin glass is formulated via the symmetric Boltzmann measure (1) with the conventional Hamiltonian $q(Y,w) = \beta Y w$. The function $P_X(x_i) = e^{-x_i^2}$ with $x_i \in \mathbb{R}$ enforces (canonically) the spherical constraint $\sum_i x_i^2 = N$. External magnetic field can also be included in $P_X(x_i)$.

The disorder Y_{ij} is most commonly randomly quenched in physics studies of the spherical spin glass model.

c. **Heisenberg spin glass.** In Heisenberg spin glass [62] the Hamiltonian is again the conventional symmetric one with randomly quenched disorder. The spins are 3-dimensional vectors, $x_i \in \mathbb{R}^3$, of unit length, $x_i^\top x_i = 1$. Magnetic field influences the direction of the spin so that

$$P_X(x_i) = e^{\beta h^\top x_i}\,, \tag{24}$$

where $h \in \mathbb{R}^3$.

d. **Restricted Boltzmann Machine.** Restricted Boltzmann machines (RBMs) are one of the triggers of the recent revolution in machine learning called deep learning [24, 25, 38]. The way RBMs are used in machine learning is that one considers the bipartite Boltzmann measure (2). In the training phase one searches a matrix Y_{ij} such that the data represented as a set of configurations of the u-variable have high likelihood (low energy). The v-variable are called the hidden units and columns of the matrix Y_{ij} (each corresponding to one hidden unit) are often interpreted as features that are useful to explain the structure of the data.

The RBM is most commonly considered for the conventional Hamiltonian $g(Y,w) = Yw$ and for binary variables $u_i \in \{0,1\}$ and $v_i \in \{0,1\}$. But other distributions for both the data-variables u_i and the hidden variables v_i were considered in the literature and the approach of the present paper applies to all of them.

We note that the disorder Y_{ij} that was obtained for an RBM trained on real datasets does not belong to the classes for which the theory developed in this paper is valid (training introduces involved correlations). However, it was shown recently that the Low-RAMP equations as studied in the present paper can be used efficiently for training of the RBM [20, 31].

The RBM with Gaussian hidden variables is related to the well known Hopfield model of associative memory [26]. Therefore the properties of the bipartite Boltzmann measure (2) with a randomly quenched disorder Y_{ij} are in one-to-one correspondence with the properties of the Hopfield model. This relation in the view of the TAP equations was studied recently in [47].

2. Examples with planted disorder

So far we covered examples where the disorder was randomly quenched (or more complicated as in the RBM). The next set of examples involves the planted disorder that is more relevant for applications in signal processing or statistics, where the variables X, U, V represent some signal we aim to recover from its measurements Y. Sometimes it is the low-rank matrix w_{ij} that we aim to recover from its noisy measurements Y. In the literature the general planted problem can be called low-rank matrix factorization, matrix recovery, matrix denoising or matrix estimation.

a. **Gaussian estimation** The most elementary examples of the planted case is when the measurement channel is Gaussian as in eq. (21), and the distributions P_X, P_U and P_V are also Gaussian i.e.

$$P_X(x_i) = \frac{1}{\sqrt{\mathrm{Det}(2\pi\sigma_X)}} e^{-\frac{1}{2}(x_i-\mu_X)^\top \sigma_X^{-1}(x_i-\mu_X)}\,, \tag{25}$$

$$P_U(u_i) = \frac{1}{\sqrt{\mathrm{Det}(2\pi\sigma_U)}} e^{-\frac{1}{2}(u_i-\mu_U)^\top \sigma_U^{-1}(u_i-\mu_U)}\,, \tag{26}$$

$$P_V(v_i) = \frac{1}{\sqrt{\mathrm{Det}(2\pi\sigma_V)}} e^{-\frac{1}{2}(v_i-\mu_V)^\top \sigma_V^{-1}(v_i-\mu_V)}\,, \tag{27}$$

where $\mu_X, \mu_U, \mu_V \in \mathbb{R}^r$ are the means of the distributions, $\sigma_X, \sigma_U, \sigma_V \in \mathbb{R}^{r\times r}$ are the covariance matrices, $Y_{ij}, w_{ij} \in \mathbb{R}$ with w_{ij} being given by (3).

We speak about the estimation problem as Bayes-optimal Gaussian estimation if the disorder Y_{ij} was generated according to

$$P_{\rm out}(Y_{ij}|w^0_{ij}) = e^{g(Y_{ij},w^0_{ij})}\,, \tag{28}$$

where $g(Y,w)$ is given by eq. (21), and

$$w^0_{ij} = {x^0}_i^\top x^0_j/\sqrt{N}\,, \quad \text{or} \quad w^0_{ij} = {u^0_i}^\top v^0_j/\sqrt{N}\,. \tag{29}$$

with X_0, U_0, and V_0 being generated from probability distributions $P_{X_0} = P_X$, $P_{U_0} = P_U$, $P_{V_0} = P_V$. The goal is to estimate matrices X_0, U_0, and V_0 from Y.

b. **Gaussian Mixture Clustering** Another example belonging to the class of problems discussed in this paper is the model for Gaussian mixture clustering. In this case the spin variables u_i are such that

$$P_U(u_i) = \sum_{s=1}^{r} n_s \delta(u_i - e_s)\,, \tag{30}$$

where r is the number of clusters, and e_s is a unit r-dimensional vector with all components except s equal to zero, and the s-component equal to 1, e.g. for $r = 3$ we have $e_1 = (1,0,0)^\top$, $e_2 = (0,1,0)^\top$, $e_s = (0,0,1)^\top$. Having $u_i = e_s$ is interpreted as data points i belongs to cluster s. We have N data points.

The columns of the matrix V then represent centroids of each of the r clusters in the M-dimensional space. The distribution P_V can as an example take the Gaussian form (27) with the covariance σ_V being an identity and the mean μ_V being zero. The output channel is Gaussian as in (21). All together this means the Y_{ij} collects positions of N points in M dimensional space that are organized in r Gaussian clusters. The goal is to estimate the centers of the clusters and the cluster membership from Y.

Standard algorithms for data clustering incluse those based on the spectral decomposition of the matrix Y such as principal component analysis [23, 67], or Loyd's k-means [42]. Works on Gaussian mixtures that are methodologically closely related to the present paper include application of the replica method to the case of two clusters $r = 2$ in [5, 8, 68] or the AMP algorithm of [45]. Note that for two clusters with the two centers being symmetric around the

origin, the resulting Boltzmann measure of the case with randomly quenched disorder is equivalent to the Hopfield model as treated e.g. in [47].

Note also that there are interesting variants of the Gaussian mixture clustering such as subspace clustering [53] where only some of the M directions are relevant for the clustering. This can be modeled by a prior on the vectors v_i that have a non-zero weight of v_i being the null vector.

The approach described in the present paper on the Bayes-optimal inference in the Gaussian mixture clustering problem has been used in a work of the authors with other collaborators in the work presented in [39].

c. **Sparse PCA** Sparse Principal Component Analysis (PCA) [29, 71] is a dimensionality reduction technique where one seeks a low-rank representation of a data matrix with additional sparsity constraints on the obtained representation. The motivation is that the sparsity helps to interpret the resulting representation. Formulated within the above setting sparse PCA corresponds to the bipartite case (2). The variables U is considered unconstrained, as an example one often considers a Gaussian prior on u_i (26). The variables V are such that many of the matrix-elements are zero.

In the literature the sparse PCA problem was mainly considered in the rank-one case $r = 1$ for which a series of intriguing results was derived. The authors of [29] suggested an efficient algorithm called diagonal thresholding that solves the sparse PCA problem (i.e. estimates correctly the position and the value of the non-zero elements of U) whenever the number of data samples is $N > CK^2 \log M$ [1], where K is the number of non-zeros and C is some constant. More recent works show existence of efficient algorithm that only need $N > \hat{C}K^2$ samples [15]. For very sparse systems, i.e. small K, this is a striking improvement over the conventional PCA that would need $O(M)$ samples. This is why sparsity linked together with PCA brought excitement into data processing methods. At the same time, this result is not as positive as it may seem, because by searching exhaustively over all positions of the non-zeros the correct support can be discovered with high probability with number of samples $N > \tilde{C}K \log M$.

Naively one might think that polynomial algorithms that need less thank $O(K^2)$ samples might exist and a considerable amount of work was devoted to their search without success. However, some works suggest that perhaps polynomial algorithm that solve the sparse PCA problems for number of samples $N < O(K^2)$ do not exist. Among the most remarkable one is the work [33] showing that the SDP algorithm, that is otherwise considered rather powerful, fails in this regime. The work of [7] goes even further showing that if sparse PCA can be solved for $N < O(K^2)$ then also a problem known as the planted clique problem can be solved in a regime that is considered as algorithmically hard for already several decades.

The problem of sparse PCA is hence one of the first examples of a relatively simple to state problem that currently presents a wide barrier between computational and statistical tractability. Deeper description of the origin of this barrier is likely to shed light on our understanding of typical algorithmic complexity in a broader scope.

The above works consider the scaling when $N \to \infty$ and K is fixed (or growing slower than $O(N)$). A regime natural to many applications is when $K = \rho N$ where $\rho = O(1)$. This regime was considered in [14] where it was shown that for $\rho > \rho_0$ an efficient algorithm that achieves the information theoretical performance exists. This immediately bring a question of what exactly happens for $\rho < \rho_0$ and how does the barrier described above appear for $K \ll N$? This question was illuminated in a work by the present authors [41] and will be developed further in this paper.

We consider sparse PCA in the bipartite case, with Gaussian U (27) and sparse V

$$P_V(v_i) = \rho\delta(v_i - 1) + (1 - \rho)\delta(v_i)\,, \tag{31}$$

as corresponds to the formulation of [1, 7, 29, 71] and others. This probabilistic setting of sparse PCA was referred to as *spiked Wishart model* in [14], notation that we will adopt in the present paper. This model is also equivalent to the one studied recently in [48] where the authors simply integrate over the Gaussian variables.

In [14] the authors also considered a symmetric variant of the sparse PCA, and refer to it as the *spiked Wigner model*. The spiked Wigner model is closer to the planted clique problem, that can be formulated using (1) with X having many zero elements. In the present work we will consider several models for the prior distribution P_X. The *Bernoulli model* as in [14] where

$$\textbf{Bernoulli model}: \qquad P_X(x_i) = \rho\delta(x_i - 1) + (1 - \rho)\delta(x_i)\,. \tag{32}$$

The spiked Bernoulli model can also be interpreted as a problem of submatrix localization where a submatrix of size $\rho N \times \rho N$ of the matrix Y has a larger mean than a randomly chosen submatrix. The submatrix localization is also relevant in the bipartite case, where it has many potential applications. The most striking ones being in gene expression where large-mean submatrices of the matrix of gene expressions of different patients may correspond to groups of patients having the same type of disease [10, 43].

In this paper we will also consider the *spiked Rademacher-Bernoulli model* with

$$\textbf{Rademacher} - \textbf{Bernoulli model}: \qquad P_X(x_i) = \frac{\rho}{2}\left[\delta(x_i - 1) + \delta(x_i + 1)\right] + (1 - \rho)\delta(x_i)\,, \tag{33}$$

as well as the *spiked Gauss-Bernoulli*

$$\textbf{Gauss}-\textbf{Bernoulli model}: \qquad P_X(x_i) = \rho\mathcal{N}(x_i, 0, 1) + (1-\rho)\delta(x_i)\,, \tag{34}$$

where $\mathcal{N}(x_i, 0, 1)$ is the Gaussian distribution with zero mean and unit variance.

So far we discussed the sparse PCA problem in the case of rank one, $r = 1$, but the case with larger rank is also interesting, especially in the view of the question of how does the algorithmic barrier depend on the rank. To investigate this question in [41] we also considered the jointly-sparse PCA, where the whole r-dimensional lines of X are zero at once, the non-zeros are Gaussians of mean $\vec{0}$ and covariance being the identity. Mathematically, $x_i \in \mathbb{R}^r$ with

$$P_X(x_i) = \frac{\rho}{(2\pi)^{r/2}} e^{\frac{-x^\top x}{2}} + (1-\rho)\delta(x_i)\,. \tag{35}$$

Another example to consider is the independently-sparse PCA where each of the r components of the lines in X is taken independently from the Gauss-Bernoulli distribution, for $x_i \in \mathbb{R}^r$ we have then

$$P_X(x_i) = \prod_{1\le k\le r}\left[\frac{\rho}{\sqrt{2\pi}} e^{\frac{-x_{ik}^2}{2}} + (1-\rho)\delta(x_{ik})\right]. \tag{36}$$

d. **Community detection** Detection of communities in networks is often modeled by the stochastic block model (SBM) where pairs of nodes get connected with probability that depends on the indices of groups to which the two nodes depend. Community detection is a widely studied problem, see e.g. the review [19]. Studies of statistical and computationally barriers in the SBM recently became very active in mathematics and statistics starting perhaps with a series of statistical physics conjectures about the existence of phase transitions in the problem [11, 12]. The particular interest of those works is that they are set in the sparse regime where every node has $O(1)$ neighbors.

Also the dense regime where every node has $O(N)$ neighbors is theoretically interesting when the difference between probabilities to connect depends only weakly on the group membership. The relevant scaling is the same as in the Potts glass model studied in [22]. In fact the dense community detection is exactly the planted version of this Potts glass model. In the setting of the present model (1) the community detection was already considered in [13] for two symmetric groups, in [40], and in [4]. In the present paper we detail the results reported briefly in [40] and in [4]. We consider the case with a general number of equal sized groups, the symmetric case. And also a case with two groups of different sizes, but such that the average degree in each of the groups is the same.

To set up the dense community detection problem we consider a network with N nodes. Each node i belongs to a community indexed by $t_i \in \{1, \cdots, r\}$. For each pair (i, j) we create an edge with probability $C_{t_i t_j}$. Where C is an $r \times r$ matrix called the connectivity matrix. In the examples of this paper we will consider two special cases of the community detection problem.

One example with r symmetric equally sized groups where for each pair of nodes (i, j) we create an edge between the two nodes with probability $p_{\rm in}$ if they are in the same group and with probability $p_{\rm out}$ if not:

$$C = \begin{pmatrix} p_{\rm in} & p_{\rm out} & \cdots & p_{\rm out} \\ p_{\rm out} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & p_{\rm out} \\ p_{\rm out} & \cdots & p_{\rm out} & p_{\rm in} \end{pmatrix} = p_{\rm in} I_r + p_{\rm out}(J_r - I_r)\,, \tag{37}$$

where I_r is a r-dimensional identity matrix, and J_r is a r-dimensional matrix filled with unit elements. The scaling we consider here is

$$p_{\rm out} = O(1)\,, \quad p_{\rm in} = O(1)\,, \tag{38}$$

$$|p_{\rm in} - p_{\rm out}| = \frac{\mu}{\sqrt{N}}\,, \quad \mu = O(1)\,, \tag{39}$$

so that the average degree in the graph is extensive. Note, however, that by rather generic correspondence between diluted and dense models, that has been made rigorous recently [13, 44], the results derived in this case hold even for average degrees that diverge only very mildly with n. The goal is to infer the group index to which each node belongs purely from the adjacency matrix of the network (up to a permutation of the indices). This problem is transformed into the low-rank matrix factorization problem through the use of the following prior probability distribution

$$P_X(x_i) = \frac{1}{r}\sum_{s=1}^{r}\delta(x_i - e_s)\,. \tag{40}$$

where $e_s \in \mathbb{R}^r$ is the vector with 0 everywhere except a 1 at position s. Eq. (40) is just a special case of (30). The output channel that describes the process of creation of the graph is

$$P_{\rm out}(Y_{ij} = 1|x_i^\top x_j/\sqrt{N}) = p_{\rm out} + \frac{\mu x_i^\top x_j}{\sqrt{N}}\,, \tag{41}$$

$$P_{\rm out}(Y_{ij} = 0|x_i^\top x_j/\sqrt{N}) = 1 - p_{\rm out} - \frac{\mu x_i^\top x_j}{\sqrt{N}}\,. \tag{42}$$

Next to the conventional Hamiltonian (22) and the Gaussian noise (21), the SBM output (41-42) is a third example of an output channel that we consider in this article. It will be used to illustrate the simplicity that arrises due to the channel universality, as also considered in [13] and [40]. Here, we obtain for the output matrices

$$S_{ij}(Y_{ij} = 1) = \tfrac{\mu}{p_{\rm out}}, \quad S_{ij}(Y_{ij} = 0) = \frac{-\mu}{1 - p_{\rm out}}\,, \tag{43}$$

$$R_{ij}(Y_{ij} = 1) \ = 0, \quad R_{ij}(Y_{ij} = 0) = \ 0\,. \tag{44}$$

Here μ is parameter that can be used to fix the signal to noise ratio.

Another example of community detection is the one with two balanced communities, i.e. having different size but the same average degree. In that setting there are two communities of size ρn and $(1-\rho)n$ with $\rho \in [0; 1]$. The connectivity matrix of this model is given by

$$C = \begin{pmatrix} p_{\rm out} & p_{\rm out} \\ p_{\rm out} & p_{\rm out} \end{pmatrix} + \frac{\mu}{\sqrt{N}} \begin{pmatrix} \frac{1-\rho}{\rho} & -1 \\ -1 & \frac{\rho}{1-\rho} \end{pmatrix}\,. \tag{45}$$

This can be modelled at the symmetric matrix factorization with rank $r = 1$ and the prior given as

$$P_X(x) = \rho\delta\left(x - \sqrt{\frac{1-\rho}{\rho}}\right) + (1-\rho)\delta\left(x + \sqrt{\frac{\rho}{1-\rho}}\right)\,. \tag{46}$$

The values in C are chosen so that each community has an average degree of $p_{\rm out}N$. The fact that in both of these cases each community has the same average degree means that one can not hope to just use the histogram of degrees to make the distinction between the communities. The output channel here is identical to the one given in (41-42)

A third example of community detection is locating one denser community in a dense network, as considered in [49] (specifically the large degree limit considered in that paper). We note that thanks to the output channel universality (Sec. I C) this case is equivalent to the spiked Bernoulli model of symmetric sparse PCA.

As a side remark we note that the community detection setting as considered here is also relevant in the bipartite matrix factorization setting where it becomes the problem of biclustering [10, 43]. The analysis developed in this paper can be straightforwardly extended to the bipartite case.

E. Main results and related work

The present paper is built upon two previous shorter papers by the same authors [40, 41]. it focuses on the study of the general type of models described by probability measures (1) and (2). On the one hand, these represent Boltzmann measures of vectorial-spin systems on fully connected symmetric or bipartite graphs. Examples of previously studied physical models that are special cases of the setting considered here would be the Sherrington-Kirkpatrick model [61], the Hopfield model [26, 47], the inference (not learning) in the restricted Boltzmann machine [20, 47, 65]. Our work hence provides a unified replica symmetric solution and TAP equations for generic class of prior distributions and Hamiltonians.

On the other hand, these probability distributions (1) and (2) represent the posterior probability distribution of a low-rank matrix estimation problem that finds a wide range of applications in high-dimensional statistics and data analysis. The thermodynamic limit $M/N = \alpha$ with $\alpha = O(1)$ whereas $N, M \to \infty$ was widely considered in the studies of spin glasses, in the context of low-rank matrix estimation this limit correspond to the challenging high-dimensional regime, whereas traditional statistics considers the case where $M/N \gg 1$. We focus on the analysis of the phase diagrams and phase transitions of low-rank matrix estimation corresponding to the Bayes optimal matrix estimation. We note that because we assume the data Y were generated from random factors U, V, X we obtain much tighter, including the constant factors, control of the high-dimensional behaviour $N, M \to \infty$, $\alpha = M/N = O(1)$, than some traditional bounds in statistics that aim not to assume any generative model but instead craft proofs under verifiable conditions of the observed data Y.

We note at this point that methodologically closely related series of work on matrix factorization is concerned with the case of high rank, i.e. when $r/N = O(1)$ [30, 35, 52]. While that case also has a set of important application (among then the learning of overcomplete dictionaries) it is different from the low-rank case considered here. The theory developed for the high rank case requires the prior distribution to be separable component-wise. The high rank case also does not present any known output channel universality, the details of the channel enter explicitly the resulting state evolution. Whereas the low-rank case can be viewed as a generalization of a spin model with pairwise interaction, in the graphical model for the high-rank case the interactions involve $O(N)$ variables.

No attempt is made at mathematical rigor in the present article. Contrary to the approach traditionally taken in statistics most of the analysis in this paper is non-rigorous, and rather relies of the accepted assumptions of the replica and cavity methods. It it worth, however, mentioning that for the case of Bayes-optimal inference, a large part of the results of this paper were proven rigorously in a recent series of works [4, 13, 14, 28, 37, 44, 58]. These proofs include the mutual information (related to the replica free energy) in the Bayes-optimal setting and the corresponding minimum mean-squared-error (MMSE), and the rigorous establishment that the state evolution is indeed describing asymptotic evolution of the Low-RAMP algorithm. The study out of the Bayes-optimal conditions (without the Nishimori conditions) are move involved.

It has become a tradition in related literature [70] to conjecture that the performance of the Low-RAMP algorithm cannot be improved by other polynomial algorithms. We do analyze here in detail the cases where Low-RAMP does not achieve the MMSE, and we remark that since effects of replica symmetry breaking need to be taken into account when evaluating the performance of the best polynomial algorithms, the conjecture of the Low-RAMP optimality among the polynomial algorithms deserves further detailed investigation.

This section gives a brief summary of our main results and their relation to existing work.

- **Approximate Message Passing** for Low-Rank matrix estimation (**Low-RAMP**): In section II we derive and detail the approximate message passing algorithm to estimate marginal probabilities of the probability measures (1) and (2) for general prior distribution, rank and Hamiltonian (output channel). We describe various special case of these equations that arrise due to the Nishimori conditions or due to self-averaging. In the physics literature this would be the TAP equations [63] generalized to vectorial spins with general local magnetic fields and generic type of pairwise interactions. The Low-RAMP equations encompass as a special case the original TAP equations for the Sherrington-Kirkpatrick model, TAP equations for the Hopfield model [46, 47], or the restricted Boltzmann machine [20, 47, 64, 65]. Within the context of low-rank matrix estimation, the AMP equations were discussed in [13, 14, 39–41, 45, 58]. Recently the Low-RAMP algorithm was even generalized to spin-variables that are not real vectors but live on compact groups [54].

  AMP type of algorithm is a promising alternative to gradient descent type of methods to minimize the likelihood. One of the main advantage of AMP is that it provides estimates of uncertainty which is crucial for accessing reliability and interpretability of the result. Compared to other Bayesian algorithms, AMP tends to be faster than Monte-Carlo based algorithms and more precise than variational mean-field based algorithms.

  We distribute two open-source Julia and Matlab versions of LowRAMP at http://krzakala.github.io/LowRAMP/. We strongly encourage the reader to download, modify, and improve on it.

- In physics, message passing equations are always closely linked with the **Bethe free energy** who's stationary points are the message passing fixed point equations. In the presence of multiple fixed points it is the value of the Bethe free energy that decides which of the fixed points is the correct one. In section II E and appendix B we derive the Bethe free energy on a single instance of the low-rank matrix estimation problem. The form of free energy that we derive has the convenience to be variational in the sense that in order to find the fixed point we are looking for a maximum of the free energy, not for a saddle. Corresponding free energy for the compressed sensing problem was derived in [34, 59] and can be used to guide the iterations of the Low-RAMP algorithm [66].

- In section III we derive the general form of the **state evolution** (under the replica symmetric assumption) of the Low-RAMP algorithm, generalizing previous works, e.g. [14, 40, 41, 58]. We present simplifications for the Bayes-optimal inference and for the conventional form of the Hamiltonian. We also give the corresponding expression for the free energy. We derive the state evolution and the free energy using both the cavity and the replica method.

  For the Bayes-optimal setting the replica Bethe free energy is up to a simple term related to the mutual information from which one can deduce the value of the **minimum information-theoretically achievable mean squared error**. Specifically, the MMSE correspond to the global maximum of the replica free energy (defined here with the opposite sign than in physics), the performance of Low-RAMP correspond to the maximum of the replica free energy that has the highest MSE.

We stress here that Low-RAMP algorithm belongs to the same class of approximate Bayesian inference algorithms as generic type of Monte Carlo Markov chains of variational mean-field methods. Yet Low-RAMP is very particular compared to these other two because of the fact that on a class of random models considered here its performance can be analyzed exactly via the state evolution and (out of the hard region) Low-RAMP asymptotically matches the performance of the Bayes-optimal estimator. Study of AMP-type of algorithms hence opens a way to put the variational mean field algorithms into more theoretical framework.

- We discuss the **output channel universality** as known in special cases in statistical physics (replica solution of the SK model depends only on the mean and variance of the quenched disorder not on other details of the distribution) and statistics [13] (for the two group stochastic block model). The general form of this universality was first put into light for the Bayes-optimal estimation in [40], proven in [37], in this paper we discuss this universality out of the Bayes-optimal setting.

- In section IV A we show that the state evolution with a Gaussian prior can be use to analyse the asymptotic **performance of spectral algorithms** such as PCA (symmetric case) or SVD (bipartite case) and derive the corresponding spectral mean-squared errors and phase transitions as studied in the random matrix literature [2]. For a recent closely related discussion see [55].

- In section IV B and IV C we discuss the **typology of phase transition and phases** that arise in Bayes-optimal low-rank matrix estimation. We provide sufficient criteria for existence of phases where estimation better than random guesses from the prior distribution is not information-theoretically possible. We analyze linear stability of the fixed point of the Low-RAMP algorithm related to this phase of undetectability, to conclude that the threshold where Low-RAMP algorithm starts to have better performance than randomly guessing from the prior agrees with the spectral threshold known in the literature on low-rank perturbations of random matrices. We also provide sufficient criteria for existence of first order phase transitions related to existence of phases where information-theoretically optimal estimation is not achievable by existing algorithms. And analyze the three thresholds $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ related to the first order phase transition.

- In section V A we give a number of examples of **phase diagrams** for the following models: rank-one $r = 1$ symmetric $XX^\top$ case with prior distribution being Bernoulli, Rademacher-Bernoulli, Gauss-Bernoulli, corresponding to balanced 2-groups. For generic rank $r \geq 1$ we give the phase diagram for the symmetric $XX^\top$ case for the jointly-sparse PCA, and for the symmetric community detection.

- Section V B and appendix D is devoted to **small ρ analysis** of the above models. This is motivated by the fact that in most existing literature with sparsity constraints the number of non-zeros is usually considered to be a vanishing fraction of the system size. In our analysis the number of non-zeros is a finite fraction ρ of the system size, we call the the regime of *linear sparsity*. We investigate whether the $\rho \to 0$ limit corresponds to previously studied cases. What concerns the information-theoretically optimal performance and related threshold $\Delta_{\rm IT}$ our small ρ limit agrees with the results known for sub-linear sparsity. Concerning the performance of efficient algorithms, from our analysis we conclude that for linear sparsity in the leading order the existing algorithms do not beat the threshold of the naive spectral method. This correspond to the known results for the planted dense subgraph problem (even when the size of the planted subgraph is sub-linear). However, for the sparse PCA problem with sub-linear sparsity algorithms such as covariance thresholding are known to beat the naive spectral threshold [15]. In the regime of linear sparsity we do not recover such behaviour, suggesting that for linear sparsity, $\rho = O(1)$, efficient algorithms that take advantage of the sparsity do not exist.

- In section V C and appendix E we discuss analytical results that we can obtain for the community detection, and joint-sparse PCA models in the **limit of large rank** r. These large rank results are matching the rigorous bounds derived for these problems in [3].

II. LOW-RANK APPROXIMATE MESSAGE PASSING, LOW-RAMP

In this section we derive the approximate message passing for the low-rank matrix estimation problems, i.e. estimation of marginals of probability distributions (1) and (2). We detail the derivation for the symmetric case (1) and state the resulting equations for the bipartite case (2). We derive the algorithm in several stages. We aim to derive a tractable algorithm that under certain conditions (replica symmetry, and absence of phase coexistence related to a first order phase transitions) estimates marginal probabilities of the probability distribution (1) exactly in the large size limit. We start with the belief propagation (BP) algorithm, as well explained e.g. in [69].

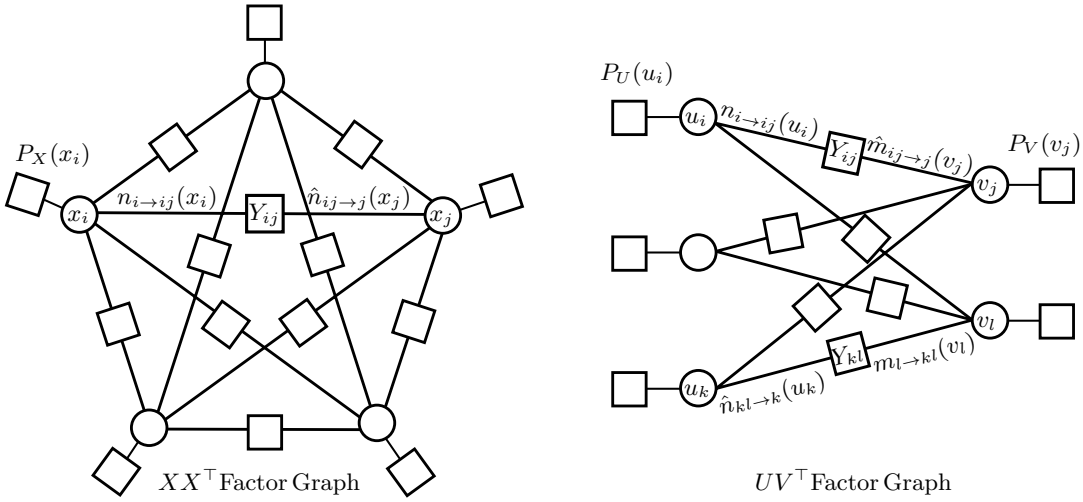

FIG. 1. This is the factor graph in the symmetric, $XX^\top$, and bipartite, $UV^\top$, matrix factorization. The squares are factors (or interaction terms), the circles represent variables. This factor graph allows us to introduce messages $n_{i\to ij}(x_i)$ and $\hat{n}_{ij\to j}(x_j)$ for the $XX^\top$ case. These are messages from variables to factors and from factors to variables. For the $UV^\top$ we introduce the four kinds of messages. $\hat{n}_{ij\to i}(u_i)$, $\hat{m}_{ij\to j}(v_j)$, $\hat{n}_{kl\to k}(u_k)$ and $m_{l\to kl}(v_l)$.

Dealing with continuous variables makes the BP algorithm quite unsuitable to implementation, as one would need to represent whole probability distributions. One takes advantage of the fact that BP assumes incoming messages to be statistically independent and since there are many incoming messages due to central limit theorem only means and variances of the products give the desired result called *relaxed-belief propagation* (in analogy to similar algorithm for sparse linear estimation [57]).

Next step is analogous to what is done for the TAP equations in the SK model [63] where one realizes that the algorithm can be further simplified and instead of dealing with one message for every directed edge, we can work only with node-dependent quantities. This gives rise to the so-called Onsager terms. Keeping track of the correct time indices under iteration [70] in order to preserve convergence of the iterative scheme, we end up with the Low-RAMP algorithm that was derived in various special forms in several previous works. In the early literature on the Hopfield model these equations were written in various forms [46] without paying close attention to the time iteration indices that cause convergence problems [70]. For more recent revival of interest in this approach that was motivated mainly by application of similar algorithmic scheme to the problem of linear estimation [6, 17] and by the need of new algorithms for constrained low-rank estimation see [14, 41, 45, 47, 58]. We finish this section with simplifications that can be done due to self-averaging, or due to the Nishimori condition in the Bayes-optimal case of due to the particular form of the Hamiltonian.

We note that the Low-RAMP algorithm is related but different from the variational mean field equations for this problem. To make the difference apparent, we state the the variational mean field algorithms in the notation of this paper in Appendix A.

A. From BP to relaxed BP

To write the BP equations for the probability measure (1) we represent it by a fully connected factor graph, Fig. 1, where every node corresponds to a variables node x_i, every edge (ij) corresponds to a pair-wise factor node $e^{g(Y_{ij},x_i^\top x_j/\sqrt{N})}$, and every node is related to a single site factor node $P_X(x_i)$.

We introduce the messages $n_{i\to ij}(x_i)$, $\tilde{n}_{ij\to i}(x_i)$ respectively as the r-dimensional messages from a variable node to a factor node and from a factor node to a variable node. The belief propagation equations then are

$$\tilde{n}_{ki\to i}(x_i) = \frac{1}{Z_{ki\to i}} \int \mathrm{d}x_k n_{k\to ki}(x_k) e^{g\left(Y_{ki},\frac{x_k^\top x_i}{\sqrt{N}}\right)}\,, \tag{47}$$

$$n_{i\to ij}(x_i) = \frac{P_X(x_i)}{Z_{i\to ij}} \prod_{1\le k\le N, k\neq i,j} \tilde{n}_{ki\to i}(x_i)\,. \tag{48}$$

The factor graph from which these messages are derived is given in Fig. 1. The most important assumption made in the BP equations is that the messages $\tilde{n}_{ki\to i}(x_i)$ are conditionally on the value x_i independent of each other thus allowing to write the product in eq. (48).

The message in (47) can be expanded as in (20) around $w = 0$ thanks to the $1/\sqrt{N}$ term. One gets

$$\tilde{n}_{ki\to i}(x_i) = \frac{e^{g(Y_{ki},0)}}{Z_{ki\to i}} \int \mathrm{d}x_k n_{k\to ki}(x_k) \left[1 + S_{ki}\frac{x_k^\top x_i}{\sqrt{N}} + \frac{x_i^\top x_k x_k^\top x_i}{2N} R_{ki} + O\left(\frac{1}{N^{3/2}}\right)\right], \tag{49}$$

where matrices S_{ij} and R_{ij} were defined in (18-19). One then defines the mean (r-dimensional vector) and covariances matrix (of size $r \times r$) of the message $n_{k\to ki}$ as

$$\hat{x}_{k\to ki} = \int \mathrm{d}x_k n_{k\to ki}(x_k) x_k^\top , \tag{50}$$

$$\sigma_{x,k\to ki} = \int \mathrm{d}x_k n_{k\to ki}(x_k) x_k x_k^\top - \hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top . \tag{51}$$

The mean with respect to $n_{k\to ki}(x_k)$ is then taken in (49) and one gets

$$\tilde{n}_{ki\to i}(x_i) = \frac{1}{Z_{ki\to i}} \exp\left[g(Y_{ki},0) + S_{ki}\frac{\hat{x}_k^\top x_i}{\sqrt{N}} - \frac{x_i^\top \hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top x_i}{2N} S_{ki}^2 + \frac{x_i^\top(\hat{x}_{k\to ik}\hat{x}_{k\to ik}^\top + \sigma_{x,k\to ik})x_i}{2N} R_{ki} + O\left(\frac{1}{N^{3/2}}\right)\right]. \tag{52}$$

Eqs. (48) and (52) are combined to get

$$n_{i\to ij}(x_i) = \frac{P_X(x_i)}{Z_{i\to ij}} \exp\left(B_{X,i\to ij}^\top x_i - \frac{x_i^\top A_{X,i\to ij} x_i}{2}\right), \tag{53}$$

where the r-dimensional vector $B_{i\to ij}$ and the $r \times r$ matrix $A_{i\to ij}$ are defined as

$$B_{X,i\to ij} = \frac{1}{\sqrt{N}} \sum_{1\le k\le N, k\neq j} S_{ki}\hat{x}_{k\to ki} , \tag{54}$$

$$A_{X,i\to ij} = \frac{1}{N} \sum_{1\le k\le N, k\neq j} \left[S_{ki}^2 \hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top - R_{ki}\left(\hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top + \sigma_{x,k\to ki}\right)\right] . \tag{55}$$

The new mean and variance of the message (53) then needs to be computed. For this we define the function $f_{\rm in}^x$

$$f_{\rm in}^x(A,B) \equiv \frac{\partial}{\partial B}\log\left(\int \mathrm{d}x P_X(x)\exp\left(B^\top x - \frac{x^\top A x}{2}\right)\right) = \frac{\int \mathrm{d}x P_X(x)\exp\left(B^\top x - \frac{x^\top A x}{2}\right) x}{\int \mathrm{d}x P_X(x)\exp\left(B^\top x - \frac{x^\top A x}{2}\right)} \tag{56}$$

as the mean of the normalized density probability

$$\mathcal{W}(x,A,B) = \frac{1}{Z_x(A,B)} P_X(x)\exp\left(B^\top x - \frac{x^\top A x}{2}\right) . \tag{57}$$

The variance of the message (53) can be computed by writing the derivative of $f_{\rm in}^x(A,B)$ with respect to B and getting

$$\frac{\partial f_{\rm in}^x(A,B)}{\partial B} = \mathbb{E}_{\mathcal{W}(\mathcal{A},\mathcal{B})}(xx^\top) - f_{\rm in}^x(A,B) f_{\rm in}^{x\,\top}(A,B) . \tag{58}$$

This expression is the covariance matrix of distribution (57). Also it is worth nothing that

$$\frac{\partial Z_x(A,B)}{\partial B} = f_{\rm in}^x(A,B) . \tag{59}$$

Adding the time indexes to clarify how these equations are iterated we get the following *relaxed BP* algorithm for the symmetric low-rank matrix estimation

$$B^t_{X,i\to ij} = \frac{1}{\sqrt{N}} \sum_{1\le k\le N, k\neq j} S_{ki}\hat{x}^t_{k\to ki}\,, \tag{60}$$

$$A^t_{X,i\to ij} = \frac{1}{N} \sum_{1\le k\le N, k\neq j} \left[S^2_{ki}\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} - R_{ki}(\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} + \sigma^t_{x,k\to ki})\right]\,, \tag{61}$$

$$\hat{x}^{t+1}_{i\to ij} = f^x_{\rm in}(A^t_{i\to ij}, B^t_{i\to ij})\,, \tag{62}$$

$$\sigma^{t+1}_{x,i\to ij} = \frac{\partial f^x_{\rm in}}{\partial B}(A^t_{i\to ij}, B^t_{i\to ij})\,. \tag{63}$$

B. Low-RAMP: TAPyfication and Onsager terms

The above relaxed BP algorithm uses $O(N^2)$ messages which can be memory demanding. But all the messages depend only weakly on the target node, and hence the algorithm can be reformulated using only $O(N)$ messages and collecting correcting terms called the Onsager terms in order to get estimators of the marginals that in the large size limit are equivalent to the previous ones. We call this formulation the TAPyfication, because of its analogy to the work of [63]. We present the derivation in the case of symmetric low-rank matrix estimation. We notice that the variables $B_{i\to ij}$ and $A_{i\to ij}$ depend only weakly on the target node j. One can use this fact to close the equations on the marginals of the system. In order to do that we introduce the variables $B_{X,i}$ and $A_{X,i}$ as

$$B^t_{X,i} = \frac{1}{\sqrt{N}} \sum_{1\le k\le N} S_{ki}\hat{x}^t_{k\to ki}\,, \tag{64}$$

$$A^t_{X,i} = \frac{1}{N} \sum_{1\le k\le N} \left[S^2_{ki}\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} - R_{ki}\left(\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} + \sigma^t_{x,k\to ki}\right)\right]\,. \tag{65}$$

We also define the variables $\hat{x}^t_i$ and $\sigma^t_{x,i}$ as the estimators of the mean and covariance matrix of x_i, reading

$$\hat{x}^{t+1}_i = f^x_{\rm in}(A^t_{X,i}, B^t_{X,i})\,, \tag{66}$$

$$\sigma^{t+1}_{x,i} = \frac{\partial f^x_{\rm in}}{\partial B}(A^t_{X,i}, B^t_{X,i})\,. \tag{67}$$

In order to close the equations we need to write the $B^t_{X,i}$ and $A^t_{X,i}$ as a function of the estimators $\hat{x}^t_i$ and $\sigma^t_{x,i}$.

From the definition of the parameters A and B, we have that $\forall j$, $B^t_{X,i} - B^t_{X,i\to ij} = \frac{S_{ij}}{\sqrt{N}}\hat{x}^t_{j\to ij} = O\left(\frac{1}{\sqrt{N}}\right)$. $A_{X,i} - A_{X,i\to ij} = O\left(\frac{1}{N}\right)$. One deduces from (62) and (63), (66) and (67), and the Taylor expansion, that in the leading order the difference between the messages and the node estimators is

$$\hat{x}^t_{k\to ki} - \hat{x}^t_k = f(A^{t-1}_{X,k\to ki}, B^{t-1}_{X,k\to ki}) - f(A^{t-1}_{X,k}, B^{t-1}_{X,k}) = -\frac{S_{ki}}{\sqrt{N}}\sigma^t_{x,k}\hat{x}^{t-1}_{i\to ki} + O\left(\frac{1}{N}\right) = -\frac{S_{ki}}{\sqrt{N}}\sigma^t_{x,k}\hat{x}^{t-1}_i + O\left(\frac{1}{N}\right)\,. \tag{68}$$

By plugging (68) into (64) one gets in the leading order

$$B^t_{X,i} = \frac{1}{\sqrt{N}}\sum_{k=1}^N S_{ki}\hat{x}^t_k - \hat{x}^{t-1}_i \frac{1}{N}\sum_{k=1}^N S^2_{ki}\sigma^t_{x,k}\,, \tag{69}$$

$$A^t_{X,i} = \frac{1}{N}\sum_{k=1}^N \left[S^2_{ki}\hat{x}^t_k\hat{x}^{t,\top}_k - R_{ki}\left(\hat{x}^t_k\hat{x}^{t,\top}_k + \sigma^t_{x,k}\right)\right]\,. \tag{70}$$

These two equations, together with eqs. (66-67) give us the low-rank approximate message passing algorithm (Low-RAMP) with $O(N)$ variables. The second term in the equation (69) is called the Onsager reaction term. Notice the iteration index $t-1$ which is non-intuitive on the first sight and was often misplaced until recently, see e.g. discussion in [70]. Note also that there is no Onsager reaction term in the expression for the covariance $A_{X,i}$, that is because the individual terms in a sum in A are of order $O(1/N)$ and not $O(1/\sqrt{N})$. This is a common pattern in AMP-type algorithm, the Onsager terms appear only in the terms that estimate means, not in the variances.

The Low-RAMP algorithm is related in spirit to the AMP algorithm for linear sparse estimation [17], for instance the function $f_{\rm in}$ is the same thresholding function as in the linear-estimation AMP. However, the linear estimation AMP is more involved for a generic output channel and the structure of the two algorithms are quite different, stemming from the fact that in the present case all interactions are pairwise whereas for the linear estimation each interaction involves all the variables, giving rise to non-trivial terms that do not appear in Low-RAMP.

The following pseudocode summarizes our implementation of the Low-RAMP algorithm:

LOW-RAMP SYMMETRIC($S_{ij}, H_{ij}, r, f^x_{\rm in}, \lambda, \epsilon_{\rm criterium}, t_{\max}, \hat{x}^{\rm init}$)

Initialize each $\hat{x}_i \in \mathbb{R}^{r\times 1}$ vector using $\hat{x}^{\rm init}$: $\forall i$, $\hat{x}_i \leftarrow \hat{x}^{\rm init}_i$.
Initialize each $\hat{x}^{\rm old}_i \in \mathbb{R}^r$ vector to zero: $\forall i$, $\hat{x}^{\rm old}_i \leftarrow 0$.
Initialize each vector $B_{X,i} \in \mathbb{R}^{r\times 1}$ to zero, $B_{X,i} \leftarrow 0$.
Initialize each N $r \times 1$ vector $B^{\rm old}_{X,i}$ to zero, $\forall i$, $B^{\rm old}_{X,i} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A_{X,i}$ with, $A_{X,i} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A^{\rm old}_{X,i}$ with, $A^{\rm old}_{X,i} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $\sigma_{X,i}$ with, $\sigma_{X,i} \leftarrow 0$.
while conv $* \lambda > \epsilon_{\rm criterium}$ and $t < t_{\max}$:
    do $t \leftarrow t+1$;
        $\forall i$, $B^{\rm new}_{X,i} \leftarrow$ Update with equation (69) or (73).
        $\forall i$, $A^{\rm new}_{X,i} \leftarrow$ Update with equation (70) or (74).
        $\forall i$, $B_{X,i} \leftarrow \lambda B^{\rm new}_{X,i} + (1-\lambda) B^{\rm old}_{X,i}$,
        $\forall i$, $A_{X,i} \leftarrow \lambda A^{\rm new}_{X,i} + (1-\lambda) A^{\rm old}_{X,i}$,
        $\forall i$, $\hat{x}^{\rm old}_i \leftarrow \hat{x}_i$, $\hat{x}_i \leftarrow f^x_{\rm in}(A_X, B_{X,i})$,
        $\forall i$, $\sigma_{X,i} \leftarrow \frac{\partial f^x_{\rm in}}{\partial B}(A_X, B_{X,i})$,
conv $\leftarrow \frac{1}{N}\sum \|\hat{x}_i - \hat{x}^{\rm old}_i\|$.
return signal components $\mathbf{x}$.

The canonical initialization we use is

$$\forall i \in [1; N],\ \ \hat{x}^{\rm init}_i \leftarrow 10^{-3}\mathcal{N}(0, I_r)\,. \tag{71}$$

The constant 10^{-3} here can be changed, but it is a bad idea to initialize exactly at zero since $\hat{x}_i = 0$ could be exactly a fixed point of the equations. In order to analyze the algorithm for a specific problem it is instrumental to initialize in the solution:

$$\forall i \in [1; N],\ \ \hat{x}^{\rm init}_i \leftarrow x^0_i\,, \tag{72}$$

where x^0_i is the planted (ground truth) configuration.

In the above pseudocode the damping factor λ is chosen constant for the whole duration of the algorithm. It is possible to choose λ dynamically in order to improve the convergence. Using the fact that the Low-RAMP algorithm finds a stationary fixed point of the Bethe free energy given in (II E) one can choose the damping factor λ so that at each step so that the Bethe-free-energy increases, this is described in [66]. Another way to choose λ is by ensuring that $\sum \|\hat{x}^t - \hat{x}^{t+1}\|$ does not oscillate too much. If one sees too much oscillations one increases the damping and decreases it otherwise. Further way to improve convergence is related to randomization of the update scheme as argued for the related compressed sensing problem in [9].

C. Simplifications

The above generic Low-RAMP equations simplify further under certain conditions.

1. *Advantage of self-averaging*

We can further simplify these equations by noticing that in all the expressions where S^2_{ij} appears we can replace it by its mean without changing the leading order of the quantities. This follows from the assumption made in the BP equations, that states that the messages incoming to a node are independent conditionally on the value of the node. Consequently the sums in eqs. (60-61) are sum of $O(N)$ independent variables and can hence in the leading order be replaced by their means.

This allows us to write the Low-RAMP equations (69-70) in an even simpler form

$$B^t_{X,i} = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} S_{ki} \hat{x}^t_k - \left(\frac{1}{N\widetilde{\Delta}} \sum_{k=1}^{N} \sigma^t_{x,k} \right) \hat{x}^{t-1}_i \,, \tag{73}$$

$$A^t_X = \frac{1}{N\widetilde{\Delta}} \sum_{k=1}^{N} \hat{x}^t_k \hat{x}^{t,\top}_k - \overline{R} \frac{1}{N} \sum_{k=1}^{N} \left(\hat{x}^t_k \hat{x}^{t,\top}_k + \sigma^t_{x,k} \right) \,, \tag{74}$$

where we defined

$$\frac{1}{\widetilde{\Delta}} \equiv \frac{1}{2N^2} \sum_{1 \le i < j \le N} S^2_{ij} \,, \tag{75}$$

$$\overline{R} \equiv \frac{1}{2N^2} \sum_{1 \le i < j \le N} R_{ij} \,. \tag{76}$$

Whereas $\widetilde{\Delta}$ is always positive, $\overline{R}$ can be positive or negative. Together with eqs. (66-67) the above two expressions are a closed set of equations. Note in particular that in (74) the dependence on index i disappeared in the leading order.

2. Bayes-optimal case

In the Bayes-optimal inference case we derived expression (11). Putting it together with the definition of the matrix R_{ij} in eq. (19) and realizing that the average over sites i, j and the average over $P(Y|w)$ act the same way we get that for $\overline{R}$ as defined in (76) we have $\overline{R} = 0$. This property belongs to the class of properties called the Nishimori conditions. In the Bayes optimal case the expression for A^t_X simplified further into

$$A^t_X = \frac{1}{N\Delta} \sum_{k=1}^{N} \hat{x}^t_k \hat{x}^{t,\top}_k \,, \tag{77}$$

where we used the Bayes-optimality once more to realize that $\widetilde{\Delta} = \Delta$ as defined in eq. (12). The convenient property of the Bayes-optimality is that the quantity A^t_X now has to be non-negative.

3. Conventional Hamiltonian : SK model

Another case that is worth specifying is the conventional Hamiltonian where g is given by (22). One gets

$$A^t_X = -\overline{R} \frac{1}{N} \sum_{k=1}^{N} \sigma^t_{x,k} \,. \tag{78}$$

It is slightly counter-intuitive that this variance-like term is negative, but it is always used only in the function f^x_{in} defined in (56) where it gets multiplied by the $P_X(x_i)$ therefore if P_X is decaying fast enough or has a bounded support, the corresponding integrals exist and are finite.

This is a convenient point where we can make the link between the Low-RAMP algorithm and the TAP equations for the SK model. For the Ising spins (23) the function f^x_{in} becomes

$$f^x_{\text{in}}(A, B) = \tanh\left(\beta h + B\right) \,, \quad \frac{\partial f^x_{\text{in}}(A, B)}{\partial B} = 1 - \tanh^2\left(\beta h + B\right) \,. \tag{79}$$

Notice the independence on the parameter A. The conventional Hamiltonian of the SK model corresponds to $g(Y, w) = \beta Y w$ so that $S = \beta Y$. Which gives us for the update of the parameter B eq. (69)

$$B^t_{X,i} = \frac{\beta}{\sqrt{N}} \sum_{k=1}^{N} Y_{ki} \hat{x}^t_k - \hat{x}^{t-1}_i \frac{\beta^2}{N} \sum_{k=1}^{N} Y^2_{ki} (1 - \hat{x}^{2,t}_k) \,. \tag{80}$$

Together with (79) we get the well known TAP equations [63]

$$\hat{x}^{t+1}_i = \tanh\left[\beta h + \frac{\beta}{\sqrt{N}} \sum_{k=1}^{N} Y_{ki} \hat{x}^t_k - \hat{x}^{t-1}_i \frac{\beta^2}{N} \sum_{k=1}^{N} Y^2_{ki} (1 - \hat{x}^{2,t}_k) \right] . \tag{81}$$

D. Summary of Low-RAMP for the bipartite low-rank estimation

The derivation for the bipartite case $UV^\top$ is completely analogous. The *relaxed BP* equations read

$$B^t_{U,i\to ij} = \frac{1}{\sqrt{N}} \sum_{1\le l\le M, l\neq j} S_{il}\hat{v}^t_{l\to il}\,, \tag{82}$$

$$A^t_{U,i\to ij} = \frac{1}{N} \sum_{1\le l\le M, l\neq j} \left[S^2_{il}\hat{v}^t_{l\to il}\hat{v}^{t,\top}_{l\to il} - R_{il}\left(\hat{v}^t_{l\to il}\hat{v}^{t,\top}_{l\to il} + \sigma^t_{v,l\to il}\right)\right]\,, \tag{83}$$

$$\hat{u}^t_{i\to ij} = f^u_{\rm in}(A^t_{U,i\to ij}, B^t_{U,i\to ij})\,, \tag{84}$$

$$\sigma^t_{u,i\to ij} = \frac{\partial f^u_{\rm in}}{\partial B}(A^t_{U,i\to ij}, B^t_{U,i\to ij})\,, \tag{85}$$

$$B^t_{V,j\to ij} = \frac{1}{\sqrt{N}} \sum_{1\le k\le N, k\neq j} S_{kj}\hat{u}^t_{k\to kj}\,, \tag{86}$$

$$A^t_{V,j\to ij} = \frac{1}{N} \sum_{1\le k\le N, k\neq i} \left[S^2_{kj}\hat{u}^t_{k\to kj}\hat{u}^{t,\top}_{k\to kj} - R_{kj}\left(\hat{u}^t_{k\to kj}\hat{u}^{t,\top}_{k\to kj} + \sigma^t_{u,k\to kj}\right)\right]\,, \tag{87}$$

$$\hat{v}^{t+1}_{j\to ij} = f^v_{\rm in}(A^t_{V,j\to ij}, B^t_{V,j\to ij})\,, \tag{88}$$

$$\sigma^{t+1}_{v,j\to ij} = \frac{\partial f^v_{\rm in}}{\partial B}(A^t_{V,j\to ij}, B^t_{V,j\to ij})\,. \tag{89}$$

Note that here we broke the symmetry between U and V by choosing an order in the update. We first update estimators of U without increasing the time index and only then estimators of V while increasing the time index by one.

The Low-RAMP equations with their Onsager terms for the bipartite low-rank matrix estimation read

$$B^t_{U,i} = \frac{1}{\sqrt{N}} \sum_{l=1}^{M} S_{il}\hat{v}^t_l - \left(\frac{1}{N}\sum_{l=1}^{M} S^2_{il}\sigma^t_{v,l}\right)\hat{u}^{t-1}_i\,, \tag{90}$$

$$A^t_{U,i} = \frac{1}{N} \sum_{l=1}^{M} \left[S^2_{il}\hat{v}^t_l\hat{v}^{t,\top}_l - R_{il}\left(\hat{v}^t_l\hat{v}^{t,\top}_l + \sigma^t_{v,l}\right)\right]\,, \tag{91}$$

$$\hat{u}^t_i = f^u_{\rm in}(A^t_{U,i}, B^t_{U,i})\,, \tag{92}$$

$$\sigma^t_{u,i} = \left(\frac{\partial f^u_{\rm in}}{\partial B}\right)(A^t_{U,i}, B^t_{U,i})\,, \tag{93}$$

$$B^t_{V,j} = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} S_{kj}\hat{u}^t_k - \left(\frac{1}{N}\sum_{k=1}^{N} S^2_{kj}\sigma^t_{u,k}\right)\hat{v}^t_j\,, \tag{94}$$

$$A^t_{V,j} = \frac{1}{N} \sum_{k=1}^{N} \left[S^2_{kj}\hat{u}^t_k\hat{u}^{t,\top}_k - R_{kj}\left(\hat{u}^t_k\hat{u}^{t,\top}_k + \sigma^t_{u,k}\right)\right]\,, \tag{95}$$

$$\hat{v}^{t+1}_j = f^v_{\rm in}(A^t_{V,j}, B^t_{V,j})\,, \tag{96}$$

$$\sigma^{t+1}_{v,j} = \left(\frac{\partial f^v_{\rm in}}{\partial B}\right)(A^t_{V,j}, B^t_{V,j})\,. \tag{97}$$

Due to independence between messages assumed in the BP algorithm we can simplify the Low-RAMP algorithm introducing (75) and (76) also for the bipartite case. As well as take advantage of the Bayes optimality and the Nishimori conditions or of the particular form of the conventional Hamiltonian.

LowRAMP-bipartite($S_{ij}, H_{ij}, \Delta, r, f^u_{\rm in}, f^v_{\rm in}, \lambda, \epsilon_{\rm criterium}, t_{\max}, \hat{u}^{\rm init}, \hat{v}^{\rm init}$)

Initialize each N, $r \times 1$ vector $\hat{u}_i$ using $\hat{u}^{\rm init}$, $\forall i$, $\hat{u}_i \leftarrow \hat{u}^{\rm init}_i$.
Initialize each M, $r \times 1$ vector $\hat{v}_j$ using $\hat{v}^{\rm init}$, $\forall j$, $\hat{v}_j \leftarrow \hat{v}^{\rm init}_j$.
Initialize each M, $r \times 1$ vector $\hat{u}^{\rm old}_i$ to 0, $\forall i$, $\hat{u}^{\rm old}_i \leftarrow 0$.
Initialize each M, $r \times 1$ vector $\hat{v}^{\rm old}_j$ to 0, $\forall j$, $\hat{v}^{\rm old}_j \leftarrow 0$.
Initialize to zero each N, $r \times 1$ vector $B_{U,i}$ to zero, $\forall i$, $B_{U,i} \leftarrow 0$.
Initialize to zero each M, $r \times 1$ vector $B_{V,j}$ to zero, $\forall j$, $B_{V,j} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A_{U,i}$ with, $\forall j$, $A_{U,i} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A^{\rm old}_{U,j}$ with, $\forall i$, $A^{\rm old}_{U,i} \leftarrow 0$.
Initialize to zero each M matrix, $r \times r$ matrix $A_{V,i}$ with, $\forall j$, $A_{V,j} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A^{\rm old}_{V,j}$ with, $\forall j$, $A^{\rm old}_{V,j} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $\sigma_{U,i}$ with, $\sigma_{U,i} \leftarrow 0$.
Initialize to zero each M matrix, $r \times r$ matrix $\sigma_{V,j}$ with, $\sigma_{V,j} \leftarrow 0$.
while conv $* \lambda > \epsilon_{\rm criterium}$ and $t < t_{\max}$:
    do $t \leftarrow t + 1$;
        Update variables U
        $\forall i,\ B^{\rm new}_{U,i} \leftarrow$ Update with equation (90).
        $\forall i,\ A^{\rm new}_{U,i} \leftarrow$ Update with equation (91).
        $\forall i,\ B_{U,i} \leftarrow \lambda B^{\rm new}_{U,i} + (1-\lambda) B^{\rm old}_{U,i}$,
        $\forall i,\ A_{U,i} \leftarrow \lambda A^{\rm new}_{U,i} + (1-\lambda) A^{\rm old}_{U,i}$,
        $\forall i,\ \hat{u}^{\rm old}_i \leftarrow \hat{u}_i,\ \hat{u}_i \leftarrow f^u_{\rm in}(A_{U,i}, B_{U,i})$,
        $\forall i,\ \sigma_{U,i} \leftarrow \frac{\partial f^u_{\rm in}}{\partial B}(A_{U,i}, B_{U,i})$.
        Update variables V
        $\forall j,\ sB^{\rm new}_{V,j} \leftarrow$ Update with equation (94).
        $\forall j,\ A^{\rm new}_{V,j} \leftarrow$ Update with equation (95).
        $\forall j,\ B_{V,j} \leftarrow \lambda B^{\rm new}_{V,j} + (1-\lambda) B^{\rm old}_{V,j}$,
        $\forall j,\ A_{V,j} \leftarrow \lambda A^{\rm new}_{V,j} + (1-\lambda) A^{\rm old}_{V,j}$,
        $\forall j,\ \hat{v}^{\rm old}_j \leftarrow \hat{u}_i,\ \hat{u}_i \leftarrow f^v_{\rm in}(A_{V,j}, B_{V,j})$,
        $\forall j,\ \sigma_{V,j} \leftarrow \frac{\partial f^v_{\rm in}}{\partial B}(A_{V,j}, B_{V,j})$.
        Compute distance with previous iteration
        conv $\leftarrow \frac{1}{N}\sum \|\hat{u}_i - \hat{u}^{\rm old}_i\| + \frac{1}{M}\sum \|\hat{v}_j - \hat{v}^{\rm old}_j\|$.
return signal components **x**

The initialization and damping factor λ are chosen similarly as for the symmetric case as discussed in section II B.

1. *Example The Hopfield model*

In order to illustrate the form of the Low-RAMP equations on an example of a bipartite matrix factorization problem we choose the Hopfield model [26], a well known model of associative memory. It turns out that the Hopfield model can be mapped to the bipartite low-rank matrix factorisation problem. This is analyzed with greater care recently in [47]. The Hopfield model is defined by the following Boltzmann distribution

$$P_{\rm Hopfield}(U) \sim \exp\left(\beta \sum_{1\le i\le j\le N} J_{ij} u_i u_j\right), \tag{98}$$

$$J_{ij} = \frac{1}{N}\sum_{1\le \mu\le M} Y_{i\mu} Y_{j\mu}\,, \tag{99}$$

where the u_i are ± 1 spin variables, and $Y_{i\mu}$ is $N \times M$ matrix collecting the M of N-dimensional patterns to remember. The elements of the matrix $Y_{i\mu}$ are most commonly considered as independent ± 1 variables. The relation (99) is then known as the Hebbs rule of associative memory.

Using a Gaussian integral we can map the Hopfield model to the following probability distribution

$$P(U,V|Y) = \frac{1}{Z_{UV}(Y)} \exp\left(\sum_{1\le\mu\le M}\left[-\frac{\beta v_\mu^2}{2} + v_\mu \sum_{1\le j\le N} \frac{\beta Y_{j\mu} u_j}{\sqrt{N}}\right]\right) \prod_{1\le i\le N} P_U(u_i)\,, \tag{100}$$

where $P_U(u) = [\delta(u-1) + \delta(u+1)]/2$. Indeed, we can check that integrating expression (100) we recover the Hopfield Boltzmann distribution (98). The Hopfield model is hence equivalent to the bipartite matrix factorization with Gaussian spins on one side, Ising spins on the other side, the conventional Hamiltonian and randomly quenched disorder $Y_{\mu j}$.

In order to write the Low-RAMP equations (90-97), known as the TAP equations in the context of the Hopfield model, we remind that the S and R matrices for the conventional Hamiltonian are given by eq. (22). Further, we remind that for Ising spins u, eq. (23), with no magnetic field ($\rho = 1/2$) the thresholding function $f^u_{\text{in}}(A, B)$ is given by

$$f^u_{\text{in}}(A, B) = \tanh(B)\,, \qquad \frac{\partial f^u_{\text{in}}(A, B)}{\partial B} = 1 - \tanh^2 B\,. \tag{101}$$

It follows for instance that $\sigma^t_{u,i} = 1 - (\hat{u}^t_i)^2$. Finally for Gaussian spins with zero mean and variance $1/\beta$ the thresholding function $f^v_{\text{in}}(A, B)$ is given by

$$f^v_{\text{in}}(A, B) = \frac{B}{A+\beta}\,, \qquad \frac{\partial f^v_{\text{in}}(A, B)}{\partial B} = \frac{1}{A+\beta}\,. \tag{102}$$

We insert these expressions into equations (90-97) to get

$$\hat{v}^{t+1}_\mu = \frac{\frac{1}{\beta\sqrt{N}} \sum_{i=1}^{N} Y_{i\mu}\hat{u}^t_i - (1-q^t)\hat{v}^t_\mu}{C^t}\,, \tag{103}$$

$$\hat{u}^{t+1}_i = \tanh\left[\frac{\beta}{\sqrt{N}} \sum_{\mu=1}^{M} Y_{i\mu}\hat{v}^{t+1}_\mu - \frac{\alpha\hat{u}^t_i}{C^t}\right]\,, \tag{104}$$

where

$$q^t = \frac{1}{N} \sum_{1\le i\le N} (\hat{u}^t_i)^2\,, \tag{105}$$

$$C^t = \frac{1}{\beta} - (1-q^t)\,. \tag{106}$$

It is possible to close the equations on the v^t_j by injecting (103) in (104), and then noticing that the term $\beta\sum_{\mu=1}^{M} Y_{i\mu}\hat{v}^t_\mu/\sqrt{N}$ can be replaced by $\tanh^{-1}(\hat{u}^t_i) + \frac{\alpha\hat{u}_i^{t-1}}{C^t}$ (104)

$$\hat{u}^{t+1}_i = \tanh\left[\frac{1}{C^t}\sum_{j=1}^{N} J_{ij}\hat{u}^t_j - \frac{\alpha u^t_i}{C^t} - \frac{1-q^t}{C^t}\left(\tanh^{-1}(\hat{u}^t_i) + \frac{\alpha\hat{u}_i^{t-1}}{C^t}\right)\right]\,. \tag{107}$$

These are the same equations as the TAP equations for the Hopfield model discussed in [47] eqs. (48,49,50), except that here they are expressed in term of the magnetization and not the fields $\tanh^{-1}(u^t_i)$.

E. Bethe Free Energy

We define the free energy of a given probability measure as the logarithm of its normalization (in physics it is usually the negative logarithm, but in this article we adopt the definition without the minus sign). Notably for the symmetric vector-spin glass model (1) we define

$$\Phi_{XX^\intercal}(Y) = \log\left(Z_X(Y)\right) - \sum_{1\le i<j\le N} g(Y_{ij}, 0)\,, \tag{108}$$

where $Z_X(Y)$ is the normalization, i.e. the partition function, in (1). We subtract the constant term on the right hand side for convenience. For the bipartite vector-spin glass model (2) analogously

$$\Phi_{UV^\intercal}(Y) = \log\left(Z_{UV}(Y)\right) - \sum_{1\le i<N, 1\le j\le M} g(Y_{ij}, 0)\,. \tag{109}$$

We remove the constant term $g(Y_{ij},0)$ so that the Free energy $\Phi_{XX^\top}$ and $\Phi_{UV^\top}$ will be $O(N)$ and self averaging in the large N limit. The exact free energies are intractable to compute, therefore we use approximations equivalent to those used for derivation of the Low-RAMP algorithm to derive the so-called *Bethe free energy*. Under the assumption of replica symmetry this free energy is exact in the leading order in N and with high probability over the ensemble of instances. To derive the Bethe free energy we use the Plefka expansion, later extended by Georges and Yedidia [21, 56]. The derivation is presented in the appendix B.

The Bethe free energy for the symmetric vector-spin glass case, $XX^\top$, is

$$\begin{aligned}
\Phi_{\text{Bethe},XX^\top} &= \max_{\{A_{X,i}\},\{B_{X,i}\}} \Phi_{\text{Bethe},XX^\top}(\{A_{X,i}\},\{B_{X,i}\}) \\
\Phi_{\text{Bethe},XX^\top}(\{A_{X,i}\},\{B_{X,i}\}) &= \sum_{1\le i\le N} \log(Z_x(A_{X,i},B_{X,i})) - B_{X,i}^\top \hat{x}_i + \frac{1}{2}\text{Tr}\left[A_{X,i}(\hat{x}_i\hat{x}_i^\top + \sigma_{x,i})\right] \\
&+ \frac{1}{2}\sum_{1\le i,j\le N}\left[\frac{1}{\sqrt{N}}S_{ij}\hat{x}_i^\top\hat{x}_j + \frac{R_{ij}}{2N}\text{Tr}\left[(\hat{x}_i\hat{x}_i^\top+\sigma_{x,i})(\hat{x}_j\hat{x}_j^\top+\sigma_{x,j})\right] - \frac{S_{ij}^2}{2N}\text{Tr}\left[\hat{x}_i\hat{x}_i^\top\hat{x}_j\hat{x}_j^\top\right] - \frac{1}{N}S_{ij}^2\text{Tr}\left[\sigma_{x,i}\sigma_{x,j}\right]\right], \quad (110)
\end{aligned}$$

where $\hat{x}_i = f_{\text{in}}^x(A_{X,i},B_{X,i})$ and $\sigma_{x,i} = \partial_B f_{\text{in}}^x(A_{X,i},B_{X,i})$ are considered as explicit functions of $A_{X,i}$ and $B_{X,i}$, where the function f_{in}^x depends on the prior probability distribution P_X via eq. (56).

For the bipartite vector-spin glass case, $UV^\top$, the Bethe free energy is

$$\begin{aligned}
\Phi_{\text{Bethe},UV^\top} &= \max_{\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}} \Phi_{\text{Bethe},UV^\top}(\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}), \\
\Phi_{\text{Bethe},UV^\top}(\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}) &= \sum_{1\le i\le N} \log(Z_u(A_{U,i},B_{U,i})) - {B_{U,i}}^\top \hat{u}_i + \frac{1}{2}\text{Tr}\left[A_{U,i}(\hat{u}_i\hat{u}_i^\top + \sigma_{u,i})\right] \\
&+ \sum_{1\le j\le M} \log(Z_v(A_{V,j},B_{V,j})) - {B_{V,j}}^\top \hat{v}_j + \frac{1}{2}\text{Tr}\left[A_{V,j}(\hat{v}_j\hat{v}_j^\top + \sigma_{v,j})\right] \\
&+ \sum_{1\le i\le N,1\le j\le M}\left[\frac{1}{\sqrt{N}}S_{ij}\hat{u}_i^\top\hat{v}_j + \frac{1}{2N}R_{ij}\text{Tr}\left[(\hat{u}_i\hat{u}_i^\top+\sigma_{u,i})(\hat{v}_j\hat{v}_j^\top+\sigma_{v,j})\right] - \frac{S_{ij}^2\text{Tr}\left(\hat{u}_i\hat{u}_i^\top\hat{v}_j\hat{v}_j^\top\right)}{2N} - \frac{1}{N}S_{ij}^2\text{Tr}\left[\sigma_{u,i}\sigma_{v,j}\right]\right], \quad (111)
\end{aligned}$$

where the $\hat{u}_i = f_{\text{in}}^u(A_{U,i},B_{U,i})$, $\hat{v}_j = f_{\text{in}}^v(A_{V,j},B_{V,j})$, $\sigma_{u,i} = \partial_B f_{\text{in}}^u(A_{U,i},B_{U,i})$, $\sigma_{v,j} = \partial_B f_{\text{in}}^v(A_{V,j},B_{V,j})$, again seen as a function of variables A, and B. Note that fixed points of the Low-RAMP algorithm are stationary points of the Bethe free energy as can be checked explicitly by taking the derivatives of the formulas.

The main usage of the free energy is when there exist multiple fixed points of the Low-RAMP equations then the one that corresponds to the best achievable mean squared error is the one for which the free energy is the largest. Another way to use the free energy is in order to help the convergence of the Low-RAMP equations, the adaptive damping is used [59, 66] and relies on the knowledge of the above expression for the Bethe free energy.

To conclude, we recall that Low-RAMP is distributed in Matlab and Julia at http://krzakala.github.io/LowRAMP/, in a version that include the use of the Bethe free energy as a guide to increase convergence, as in [66].

III. STATE EVOLUTION

An appealing property of the Low-RAMP algorithm is that its large-system-size behaviour can be analyzed via the so called state evolution (or single letter characterization in information theory). In the statistical physics context the state evolution is the cavity method [46] thanks to which one can derive the replica symmetric solution from the TAP equations, taking properly into account the distribution of the disorder (random or quenched). Mathematically, at least in the Bayes optimal setting, the state evolution for the present systems is a rigorous statement about the asymptotic behaviour of the Low-RAMP algorithm [28].

Here we present derivation of the state evolution for the symmetric matrix factorization and state it for the bipartite case. The main idea of state evolution is to describe the current state of the algorithm using a small number of variables – called order parameters in physics. We then compute how the order parameters evolve as the number of iterations increases.

A. Derivation for the symmetric low-rank estimation

To derive the state evolution we assume that all updates are done in parallel with no damping (the state evolution does depend on the update strategy). A distinction will be made between r the rank assumed in the posterior distribution and r_0 the true rank of the planted solution. Let us introduce the order parameters that will be of relevance here

$$M_x^t = \frac{1}{N}\sum_{1\le i\le N} \hat{x}_i^t x_i^{0,\top} \in \mathbb{R}^{r\times r_0}, \tag{112}$$

$$Q_x^t = \frac{1}{N}\sum_{1\le i\le N} \hat{x}_i^t \hat{x}_i^{t,\top} \in \mathbb{R}^{r\times r}, \tag{113}$$

$$\Sigma_x^t = \frac{1}{N}\sum_{1\le i\le N} \sigma_{x,i}^t \in \mathbb{R}^{r\times r}\,, \tag{114}$$

where M_x^t is a matrix of size $r\times r_0$, while Q_x^t and Σ_x^t are $r\times r$ matrices. The interpretation of these order parameters is the following

- M_x^t measures how much the current estimate of the mean is correlated with the planted solution x_i^0. Physicist would call this the magnetization of the system.
- Q_x^t is called the self-overlap.
- Σ_x^t is the mean variance of variables.

In this section we will not assume the Bayes optimal setting, and distinguish between the prior $P_X(x_i)$ and the distribution $P_{X_0}(x_i^0)$ from which the planted configuration x_i^0 was drawn. Similarly, we will assume $g(Y,w)$ in the posterior distribution, but the data matrix Y was created from the planted configuration via $P_{\rm out}(Y|w)$. In general $P_X \neq P_{X_0}$ and $g(Y,w) \neq \log P_{\rm out}(Y|w)$.

We do assume, however, that (10) holds for our choice of $g(Y,w)$ and $P_{\rm out}(Y|w)$ even when $g(Y,w) \neq \log P_{\rm out}(Y|w)$. This will indeed hold in all examples presented in this paper. Using self-averaging arguments such as in Sec. II C 1 the averages over the quenched randomness $P_{\rm out}(Y|w)$ and over elements (ij) are interchangeable. Eq. (10) then in practice means that, in order for the state evolution as derived in this section to be valid, the Fisher score matrix S should have an empirical mean of elements of order $o\left(1/\sqrt{N}\right)$. If this assumption is not met, i.e. we have $\mathbb{E}(S) = a/\sqrt{N}$ with $a \gg 1$, it means that the matrix $S/\sqrt{N}$ will have an eigenvalue of order a, while the eigenvalues corresponding to the planted signal will be $O(1)$. This means that for $a \gg 1$ the eigenvalues corresponding to the signal will be subdominant and this would require additional terms in the state evolution.

We know that $\hat{x}_i^{t+1} = f_{\rm in}(A_{X,i}^t, B_{X,i}^t)$ and $\sigma_{x,i}^{t+1} = \frac{\partial f_{\rm in}}{\partial B}(A_{X,i}^t, B_{X,i}^t)$, eqs. (66-67). Therefore to compute the updated order parameters in the large N limit one needs to compute the probability distribution of $B_{X,i}^t$ and $A_{X,i}^t$

$$P(B_{X,i}^t | x_i^0, Q_x^t, M_x^t, \Sigma_x^t)\,, \tag{115}$$

$$P(A_{X,i}^t | x_i^0, Q_x^t, M_x^t, \Sigma_x^t)\,. \tag{116}$$

Quantities $B_{X,i}^t$ and $A_{X,i}^t$ are defined by eq. (64) and (65). Notably, by the assumptions of belief propagation the terms in the sums on the right hand side of eqs. (64) and (65) are independent. By the central limit theorem $B_{X,i}^t$ and $A_{X,i}^t$ then behave as Gaussian random variables.

Therefore all one needs to compute is their mean and variance with respect to the output channel. Using eq. (64) we get

$$\mathbb{E}(B_{X,i}^t) = \frac{1}{\sqrt{N}}\sum_{1\le k\le N}\int {\rm d}Y_{ki} P_{\rm out}\left(Y_{ki}\middle|\frac{x_k^{0,\top}x_i^0}{\sqrt{N}}\right)\left(\frac{\partial g}{\partial w}\right)_{Y_{ki},0}\hat{x}_{k\to ki}^t\,. \tag{117}$$

Let us now expand $P_{\rm out}$ around 0

$$\mathbb{E}(B_{X,i}^t) = \frac{1}{\sqrt{N}}\sum_{1\le k\le N}\int {\rm d}Y_{ki} P_{\rm out}\left(Y_{ki}|0\right)\left[1 + \frac{x_k^{0,\top}x_i^0}{\sqrt{N}}\left(\frac{\partial \log(P_{\rm out}(Y_{ki},w)}{\partial w}\right)_{Y_{ki},0} + O\left(\frac{1}{N}\right)\right]\left(\frac{\partial g}{\partial w}\right)_{Y_{ki},0}\hat{x}_{k\to ki}^t\,. \tag{118}$$

Using the above stated assumption of validity of eq. (10) we can simplify into

$$\mathbb{E}(B^t_{X,i}) = \frac{1}{N\widehat{\Delta}} \sum_{1\leq k\leq N} \hat{x}^t_k x^{0,\top}_k x^0_i = \frac{M^t_x}{\widehat{\Delta}} x^0_i \,, \tag{119}$$

where we used the definition (112) of the order parameter M^t and where we defined $\widehat{\Delta}$ via

$$\frac{1}{\widehat{\Delta}} \equiv \mathbb{E}_{P_{\rm out}} \left[\left(\frac{\partial g(Y,w)}{\partial w} \right)_{Y,0} \left(\frac{\partial \log(P_{\rm out}(Y|w))}{\partial w} \right)_{Y,0} \right] . \tag{120}$$

Let us now compute the variance of $B^t_{X,i}$. Using the assumption of belief propagation that messages incoming to a variable are independent in the leading order we get that the covariance of $B^t_{X,i}$ is the sum of all the covariance matrices of the terms in the sum defining $B^t_{X,i}$.

$$\mathrm{Cov}(B^t_{X,i}) = \frac{1}{N} \sum_{1\leq i\leq N} \mathrm{Cov}(S_{ki}\hat{x}^t_{k\to ki}) \,. \tag{121}$$

Doing a similar computation as for the mean one gets in the leading order

$$\mathrm{Cov}(B^t_{X,i}) = \frac{1}{N\widetilde{\Delta}} \sum_{1\leq i\leq N} \hat{x}^t_{k\to ki} \hat{x}^{t,\top}_{k\to ki} = \frac{Q^t_x}{\widetilde{\Delta}} \,, \tag{122}$$

where $\widetilde{\Delta}$ was introduced in (75) and thanks to self-averaging it also equals

$$\frac{1}{\widetilde{\Delta}} = \mathbb{E}_{P_{\rm out}(Y|w)} \left[\left(\frac{\partial g}{\partial w} \right)^2_{Y,0} \right] . \tag{123}$$

Here one did not even have to expand $P_{\rm out}$ to second order, the first order was enough to get the leading order of the variance.

The distribution of the $A^t_{X,i}$ now needs to be computed. Using the definition of $A^t_{X,i}$ eq. (65) and the self-averaging of section II C 1 we obtain directly that

$$\mathbb{E}(A^t_{X,i}) = \frac{Q^t_x}{\widetilde{\Delta}} - \overline{R} \left(Q^t_x + \Sigma^t_x \right) \,, \tag{124}$$

where $\overline{R}$ is defined in (76) and also equals

$$\overline{R} = \mathbb{E}_{P_{\rm out}(Y|w)} \left[\left(\frac{\partial g}{\partial w} \right)^2_{Y,0} + \left(\frac{\partial^2 g}{\partial w^2} \right)^2_{Y,0} \right] . \tag{125}$$

Here things are simpler then for $B^t_{X,i}$ since $A^t_{X,i}$ concentrates around its mean, its variance is of smaller order.

Overall, using (66) and (67) one gets for the state evolution equations

$$M^{t+1}_x = \mathbb{E}_{x_0,W} \left[f^x_{\rm in} \left(\frac{Q^t_x}{\widetilde{\Delta}} - \overline{R}(Q^t_x + \Sigma^t_x), \frac{M^t_x}{\widehat{\Delta}} x_0 + \sqrt{\frac{Q^t_x}{\widetilde{\Delta}}} W \right) x_0^\top \right] , \tag{126}$$

$$Q^{t+1}_x = \mathbb{E}_{x_0,W} \left[f^x_{\rm in} \left(\frac{Q^t_x}{\widetilde{\Delta}} - \overline{R}(Q^t_x + \Sigma^t_x), \frac{M^t_x}{\widehat{\Delta}} x_0 + \sqrt{\frac{Q^t_x}{\widetilde{\Delta}}} W \right) f^x_{\rm in}(\cdots,\cdots)^\top \right] , \tag{127}$$

$$\Sigma^{t+1}_x = \mathbb{E}_{x_0,W} \left[\frac{\partial f^x_{\rm in}}{\partial B} \left(\frac{Q^t_x}{\widetilde{\Delta}} - \overline{R}(Q^t_x + \Sigma^t_x), \frac{M_x}{\widehat{\Delta}} x_0 + \sqrt{\frac{Q^t_x}{\widetilde{\Delta}}} W \right) \right] , \tag{128}$$

where W and x_0 are two independent random variables. W is a Gaussian noise of mean 0 and covariance matrix I_r and x_0 is a random variable of probability distribution P_{X_0}. The thresholding function $f^x_{\rm in}$ is defined in eq. (56). This also allows us to know that at a fixed point the marginals x_i will be distributed according to

$$\hat{x_i}^{t=+\infty} = f^x_{\rm in} \left(\frac{Q_x}{\widetilde{\Delta}} - \overline{R}(Q_x + \Sigma_x), \frac{M_x}{\widehat{\Delta}} x_0 + \sqrt{\frac{Q_x}{\widetilde{\Delta}}} W \right) , \tag{129}$$

where W is a Gaussian variable and mean 0 and covariance I_r and x_0 is taken with respect to P_{X_0}. Let us state that the large N limit of the $\text{MSE}_X = \sum_{i=1\dots N} ||\hat{x}_i - x_i^0||_2^2/N$ is computed from the order parameters as

$$\text{MSE}_X = \text{Tr}\left[\langle x^0 x^{0,\top}\rangle - 2M_x + Q_x\right] , \tag{130}$$

where $\langle x^0 x^{0,\top}\rangle = \mathbb{E}_{P_{X_0}}(x^0 x^{0,\top})$ is the average with respect to the distribution P_{X_0}.

Note that the state evolution equations only depend on the assumed and truth noise channels through three variables $\widetilde{\Delta}$, $\widehat{\Delta}$ and $\overline{R}$. In the Bayes-optimal case these equations will simplify even further and the noise channel will be described through one parameter Δ, the Fisher information, this is derived in section III D 1. This universality with respect to the output channel has been observed elsewhere in a special case of the present problem [16] (see e.g. their remark 2.5) in the study of detection of a small hidden clique with approximate message passing.

Finally one additional assumption made in this whole section is that no Replica Symmetry Breaking (RSB) appears. It is known that RSB does appear for some regimes of parameters out of the equilibrium Bayes-optimal case. We let the investigation of RSB in the context of low-rank matrix estimation for future work, in the examples analyzed in this paper we will restrict ourselves to the Bayes-optimal case where RSB at equilibrium cannot happen [70].

B. Summary for the bipartite low-rank matrix factorization

State evolution can also be written similarly for the $UV^\top$ case. In that case there are six order parameters

$$M_u^t = \frac{1}{N}\sum_{1\le i\le N} u_i^t u_i^{0,\top}, \quad Q_u^t = \frac{1}{N}\sum_{1\le i\le N} u_i^t u_i^{t,\top}, \quad \Sigma_u^t = \frac{1}{N}\sum_{1\le i\le N} \sigma_{u,i}^t , \tag{131}$$

$$M_v^t = \frac{1}{M}\sum_{1\le j\le M} v_j^t v_j^{0,\top}, \quad Q_v^t = \frac{1}{M}\sum_{1\le j\le M} v_j^t v_j^{t,\top}, \quad \Sigma_v^t = \frac{1}{M}\sum_{1\le j\le M} \sigma_{v,j}^t . \tag{132}$$

These order parameters are updated according to the following state evolution equations

$$M_u^t = \mathbb{E}_{u_0,W}\left[f_{\text{in}}^u\left(\frac{\alpha Q_v^t}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v^t+\Sigma_v^t), \alpha\frac{M_v^t}{\widehat{\Delta}}u_0 + \sqrt{\frac{\alpha Q_v^t}{\widetilde{\Delta}}}W\right) u_0^\top\right] , \tag{133}$$

$$Q_u^t = \mathbb{E}_{u_0,W}\left[f_{\text{in}}^u\left(\alpha\frac{Q_v^t}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v), \alpha\frac{M_v^t}{\widehat{\Delta}}u_0 + \sqrt{\frac{\alpha Q_v^t}{\widetilde{\Delta}}}W_v\right) f_{\text{in}}^u(\cdots,\cdots)^\top\right] , \tag{134}$$

$$\Sigma_u^t = \mathbb{E}_{u_0,W}\left[\frac{\partial f_{\text{in}}^u}{\partial B}\left(\alpha\frac{Q_v^t}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v^t+\Sigma_v^t), \alpha\frac{M_v^t}{\widehat{\Delta}}u_0 + \sqrt{\frac{\alpha Q_v^t}{\widetilde{\Delta}}}W\right)\right] , \tag{135}$$

$$M_v^{t+1} = \mathbb{E}_{v_0,W}\left[f_{\text{in}}^v\left(\frac{Q_u^t}{\widetilde{\Delta}} - \overline{R}(Q_u^t+\Sigma_u^t), \frac{M_u^t}{\widehat{\Delta}}v_0 + \sqrt{\frac{Q_u^t}{\widetilde{\Delta}}}W\right) v_0^\top\right] , \tag{136}$$

$$Q_v^{t+1} = \mathbb{E}_{v_0,W}\left[f_{\text{in}}^v\left(\frac{Q_u^t}{\widetilde{\Delta}} - \overline{R}(Q_u^t+\Sigma_u^t), \frac{M_u^t}{\widehat{\Delta}}v_0 + \sqrt{\frac{Q_u^t}{\widetilde{\Delta}}}W\right) f_{\text{in}}^v(\cdots,\cdots)^\top\right] , \tag{137}$$

$$\Sigma_v^{t+1} = \mathbb{E}_{v_0,W}\left[\frac{\partial f_{\text{in}}^v}{\partial B}\left(\frac{Q_u^t}{\widetilde{\Delta}} - \overline{R}(Q_u^t+\Sigma_u^t), \frac{M_u^t}{\widehat{\Delta}}v_0 + \sqrt{\frac{Q_u^t}{\widetilde{\Delta}}}W\right)\right] . \tag{138}$$

In these equations W, u_0 and v_0 are independent random variables, W is r dimensional Gaussian variable of mean $\vec{0}$ and covariance matrix I_r, u_0 and v_0 are sampled from density probability P_{U_0} and P_{V_0} respectively.

The $\text{MSE}_U = \sum_{i=1\dots N} ||\hat{u}_i - u_i^0||_2^2/N$ and $\text{MSE}_V = \sum_{j=1\dots M} ||\hat{v}_j - v_j^0||_2^2/M$ can be computed from the order parameters as

$$\text{MSE}_U = \text{Tr}\left[\mathbb{E}_{P_{U_0}}(U_0 U_0^\top) - 2M_u + Q_u\right] , \tag{139}$$

$$\text{MSE}_V = \text{Tr}\left[\mathbb{E}_{P_{V_0}}(V_0 V_0^\top) - 2M_v + Q_v\right] . \tag{140}$$

C. Replica Free Energy

State evolution can also be used to derive large size limit of the Bethe free energy (110) and (111) defined as

$$\Phi_{\mathrm{RS},XX^\top} \equiv \lim_{N\to+\infty} \frac{1}{N} \left\langle \log(Z_X(Y)) - \sum_{1\le i<j\le N} g(Y_{ij},0) \right\rangle , \tag{141}$$

$$\Phi_{\mathrm{RS},UV^\top} \equiv \lim_{N\to+\infty} \frac{1}{N} \left\langle \log(Z_{UV}(Y)) - \sum_{1\le i\le N,1\le j\le M} g(Y_{ij},0) \right\rangle , \tag{142}$$

where the average is taken with respect to density probability (1), or (2), the $Z_X(Y)$ and $Z_{UV}(Y)$ are the corresponding partition functions. We subtract the constant $g(Y_{ij},0)$ for convenience in order to get a quantity that is self averaging in the large N limit.

Alternatively to the state evolution, the average free energy can be derived using the replica method as we summarize in Appendix C. The resulting replica free energy for the symmetric $XX^\top$ case is (assuming the replica symmetric ansatz to hold)

$$\Phi_{\mathrm{RS},XX^\top} = \max \left\{ \phi_{\mathrm{RS}}(M_x,Q_x,\Sigma_x), \frac{\partial\phi_{\mathrm{RS}}}{\partial M_x} = \frac{\partial\phi_{\mathrm{RS}}}{\partial Q_x} = \frac{\partial\phi_{\mathrm{RS}}}{\partial\Sigma_x} = 0 \right\} , \tag{143}$$

where

$$\begin{aligned}\phi_{\mathrm{RS}}(M_x,Q_x,\Sigma_x) = \frac{\mathrm{Tr}(Q_xQ_x^\top)}{4\widetilde{\Delta}} - \frac{\mathrm{Tr}(M_xM_x^\top)}{2\widehat{\Delta}} - \frac{\overline{R}}{2}\mathrm{Tr}((Q_x+\Sigma_x)(Q_x+\Sigma_x)^\top) \\ + \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(\frac{Q_x}{\widetilde{\Delta}} - \overline{R}(Q_x+\Sigma_x), \frac{M_x}{\widehat{\Delta}}x_0 + \sqrt{\frac{Q_x}{\widetilde{\Delta}}}W\right)\right)\right] ,\end{aligned} \tag{144}$$

where the function $Z_x(A,B)$ is defined as the normalization in eq. (57).

For the bipartite $UV^\top$ case we have analogously for the replica free energy

$$\Phi_{\mathrm{RS},UV^\top} = \max \left\{ \phi_{\mathrm{RS}}(M_u,Q_u,\Sigma_u,M_v,Q_v,\Sigma_v), \frac{\partial\phi_{\mathrm{RS}}}{\partial M_u} = \frac{\partial\phi_{\mathrm{RS}}}{\partial Q_u} = \frac{\partial\phi_{\mathrm{RS}}}{\partial\Sigma_u} = \frac{\partial\phi_{\mathrm{RS}}}{\partial M_v} = \frac{\partial\phi_{\mathrm{RS}}}{\partial Q_v} = \frac{\partial\phi_{\mathrm{RS}}}{\partial\Sigma_v} = 0 \right\} , \tag{145}$$

where

$$\begin{aligned}\phi_{\mathrm{RS}}(M_u,Q_u,\Sigma_u,M_v,Q_v,\Sigma_v) = \frac{\alpha\mathrm{Tr}(Q_vQ_u^\top)}{2\widetilde{\Delta}} - \frac{\alpha\mathrm{Tr}(M_vM_u^\top)}{\widehat{\Delta}} - \alpha\overline{R}\mathrm{Tr}((Q_v+\Sigma_v)(Q_u+\Sigma_u)^\top) \\ + \mathbb{E}_{W,u_0}\left[\log\left(Z_u\left(\frac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v), \frac{\alpha M_v}{\widehat{\Delta}}u_0 + W\sqrt{\frac{\alpha Q_v}{\widetilde{\Delta}}}\right)\right)\right] \\ + \alpha\mathbb{E}_{W,v_0}\left[\log\left(Z_v\left(\frac{Q_u}{\widetilde{\Delta}} - \overline{R}(Q_u+\Sigma_u), \frac{M_u}{\widehat{\Delta}}v_0 + \sqrt{\frac{Q_u}{\widetilde{\Delta}}}W\right)\right)\right] ,\end{aligned} \tag{146}$$

with the function $Z_u(A,B)$ and $Z_v(A,B)$ also defined as the normalization in eq. (57).

It is worth noting that there is a close link between the expression of the replica free energy and the state evolution equations. Namely fixed points of the state evolution equations are stationary points of the replica free energy and vice-versa. Therefore, by looking for a stationary point of these equations one finds back the state evolution equations (126-128,133-138).

D. Simplification of the SE equations

1. *Simplification in the Bayes optimal setting*

The state evolution equations simplify considerably when we restrict ourselves to the Bayes-optimal setting defined in Sec. I B 1 by eq. (7).

From the definitions of $\widehat{\Delta}$ in eq. (120) and $\widetilde{\Delta}$ in eq. (75) and using the identity (7) defining the Bayes optimal setting we obtain

$$\frac{1}{\widehat{\Delta}} = \frac{1}{\widetilde{\Delta}} = \frac{1}{\Delta} = \mathbb{E}_{P_{\rm out}(Y,w=0)} \left[\left(\frac{\partial g}{\partial w} \right)^2_{Y,w=0} \right] , \tag{147}$$

where Δ is the Fisher information of the output channel defined in eq. (12). Note for instance that for the Gaussian input channel (21), Δ is simply the variance of the Gaussian noise. The bigger the Δ the harder the inference problem becomes. The smaller the Δ the easier the inference is.

Further consequence of having Bayes optimality (7) is that $\overline{R} = \mathbb{E}(R_{ij}) = 0$ as proven in equation (11). This simplifies greatly the state evolution equations into

$$M_x^{t+1} = \mathbb{E}_{x_0,W} \left[f_{\rm in}^x \left(\frac{Q_x^t}{\Delta}, \frac{M_x^t}{\Delta} x_0 + \sqrt{\frac{Q_x^t}{\Delta}} W \right) x_0^\top \right] , \tag{148}$$

$$Q_x^{t+1} = \mathbb{E}_{x_0,W} \left[f_{\rm in}^x \left(\frac{Q_x^t}{\Delta}, \frac{M_x^t}{\Delta} x_0 + \sqrt{\frac{Q_x^t}{\Delta}} W \right) f_{\rm in}^x(\cdots,\cdots)^\top \right] , \tag{149}$$

where x_0 and W are as before independent random variables, W is Gaussian of mean 0 and covariance matrix I_r, and x_0 has probability distribution P_{X_0}.

Another property that arises in the Bayes optimal setting (7) and follows from the Nishimori condition (15) and the definition of the order parameters M_x, Q_x and Σ_x^t in (112-114) is that

$$Q_x^t = M_x^t = M_x^\top , \qquad \Sigma_x^t = \Sigma_x = \langle x^0 x^{0,\top} \rangle_{P_X} . \tag{150}$$

Enforcing $Q_x^t = M_x^t$ simplifies the state evolution equations further so that for the symmetric matrix factorization one gets

$$M_x^{t+1} = \mathbb{E}_{x_0,W} \left[f_{\rm in}^x \left(\frac{M_x^t}{\Delta}, \frac{M_x^t}{\Delta} x_0 + \sqrt{\frac{M_x^t}{\Delta}} W \right) x_0^\top \right] , \tag{151}$$

where x_0 and W are independent random variables distributed as above. For the rest of the article we define the Bayes-optimal state evolution function $f_{P_X}^{\rm SE}$ for prior P_X

$$M_x^{t+1} = f_{P_X}^{\rm SE} \left(\frac{M_x^t}{\Delta} \right) , \tag{152}$$

Let us comment on the output channel universality as discussed in Sec. I C. In the Bayes-optimal setting the channel universality becomes particularly simple and striking. For an arbitrary output channel $P_{\rm out}(Y|w)$ (for which the expansion done in section I C is meaningful) we have the following

- The Low-RAMP algorithm in the Bayes optimal setting depends on the noise channel only through the Fisher score matrix S as defined in eq. (18). This is specified in section (II C 2).
- The state evolution in the Bayes-optimal setting depends on the output channel through the Fisher information of the channel Δ (12) as described in section III D 1. As a consequence the minimal achievable error, the minimal efficiently achievable error and all other quantities that can be obtained from the state evolution depend on the output channel only trough the Fisher information Δ.

The replica free energy (144) in the Bayes-optimal case becomes

$$\phi_{\rm RS}(M_x) = \mathbb{E}_{W,x_0} \left[\log \left(Z_x \left(\frac{M_x}{\Delta}, \frac{M_x}{\Delta} x_0 + \sqrt{\frac{M_x}{\Delta}} W \right) \right) \right] - \frac{{\rm Tr}(M_x M_x^\top)}{4\Delta} . \tag{153}$$

This was first derived in [40], and proven for a special case in [13], and later in full generality in [37, 44]. We also remind that the order parameter M_x used in the state evolution is related to the mean-squared error as

$${\rm MSE}_X = {\rm Tr} \left[\langle x^0 x^{0,\top} \rangle - M_x \right] , \tag{154}$$

For the bipartite vector spin models, $UV^\top$ case, the state evolution in the Bayes optimal setting reads

$$M_u^t = \mathbb{E}_{u_0,W}\left[f_{\rm in}^u\left(\frac{\alpha M_v^t}{\Delta}, \alpha\frac{M_v^t}{\Delta}u_0 + \sqrt{\frac{\alpha M_v^t}{\Delta}}W\right)u_0^\top\right], \tag{155}$$

$$M_v^{t+1} = \mathbb{E}_{v_0,W}\left[f_{\rm in}^v\left(\frac{Q_u^t}{\Delta}, \frac{M_u^t}{\Delta}v_0 + \sqrt{\frac{Q_u^t}{\Delta}}W\right)v_0^\top\right]. \tag{156}$$

The replica free energy (146) in the Bayes optimal setting becomes

$$\phi_{\rm RS}(M_u, M_v) = \mathbb{E}_{W,u_0}\left[\log\left(Z_u\left(\frac{\alpha M_v}{\Delta}, \frac{\alpha M_v}{\Delta}u_0 + W\sqrt{\frac{\alpha M_v}{\Delta}}\right)\right)\right] + \alpha\mathbb{E}_{W,v_0}\left[\log\left(Z_v\left(\frac{M_u}{\Delta}, \frac{M_u}{\Delta}v_0 + \sqrt{\frac{M_u}{\Delta}}W\right)\right)\right] - \frac{\alpha{\rm Tr}(M_v M_u^\top)}{2\Delta}, \tag{157}$$

where once again $M_u = M_u^\top$ and $M_v = M_v^\top$. The global maximum of the free energy is asymptotically the equilibrium free energy, the value of M_u and M_v at this maximum is related to the MMSE via

$${\rm MSE}_U = {\rm Tr}\left[\mathbb{E}_{P_{U_0}}(U_0U_0^\top) - M_u\right], \tag{158}$$

$${\rm MSE}_V = {\rm Tr}\left[\mathbb{E}_{P_{V_0}}(V_0V_0^\top) - M_v\right]. \tag{159}$$

Performance of the Low-RAMP in the limit of large system sizes is given by the fixed point of the state evolution reached with initialization where both M_u and M_v are close to zero.

2. *Simplification for the conventional Hamiltonian and randomly quenched disorder*

Another illustrative example of the state evolution we give in this section is for the conventional Hamiltonian (4) with randomly quenched disorder, as this is the case most commonly considered in the existing physics literature. In that case the model (1) corresponds to a generic vectorial spin glass model. To take into account that the disorder is not planted, but random, we plug into the generic state evolution

$$P_{\rm out}(Y, w) = \frac{1}{\sqrt{2\pi J^2}}\exp\left(-\frac{Y^2}{2J^2}\right) \tag{160}$$

such that $P_{\rm out}(Y, w)$ does not depend on w, meaning that the disorder Y is chosen independently, there is no planting. For the conventional Hamiltonian (4) and output channel (160) we obtain

$$\overline{R} = \frac{1}{\widetilde{\Delta}} = \mathbb{E}\left[Y^2\right] = J^2, \tag{161}$$

$$\frac{1}{\widehat{\Delta}} = 0. \tag{162}$$

The state evolution (126) and (127) then becomes

$$M_x^{t+1} = 0, \tag{163}$$

$$Q_x^{t+1} = \mathbb{E}_W\left[f_{\rm in}^x\left(-J^2\Sigma_x^t, J\sqrt{Q_x^t}W\right)f_{\rm in}^x(\cdots,\cdots)^\top\right], \tag{164}$$

$$\Sigma_x^{t+1} = \mathbb{E}_W\left[\frac{\partial f_{\rm in}^x}{\partial B}\left(-J^2\Sigma_x^t, J\sqrt{Q_x^t}W\right)\right]. \tag{165}$$

The free energy (144) is given by

$$\phi_{\rm RS}(M_x, Q_x, \Sigma_x) = \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(-J^2\Sigma_x, J\sqrt{Q_x}W\right)\right)\right] + \frac{J^2{\rm Tr}(Q_xQ_x^\top)}{4} - \frac{J^2}{2}{\rm Tr}((Q_x+\Sigma_x)(Q_x+\Sigma_x)^\top). \tag{166}$$

Specifically, for the Sherrington-Kirkpatrick model [61], where the rank is one and the spins are Ising eq. (23) with $\rho = 1/2$ the $f_{\rm in}^x(A, B)$ is given by (101), the state evolution becomes

$$M_x^{t+1} = 0\,, \tag{167}$$

$$Q_x^{t+1} = \mathbb{E}_W \left[\tanh\left(J\sqrt{Q_x^t} W \right)^2 \right], \tag{168}$$

$$\Sigma_x^{t+1} = 1 - Q_x^t\,. \tag{169}$$

where W is a Gaussian random variables of zero mean and unit variance. With a free energy (144) given by

$$\phi_{\rm RS}(Q_x) = \frac{J^2(1-Q_x)^2 - J^2}{4} + \mathbb{E}_W \left[\log\left(\cosh\left(J\sqrt{Q_x} W \right) \right) \right]. \tag{170}$$

The reader will notice that these are just the replica symmetric equations of the Sherrington-Kirkpatrick solution [61].

IV. GENERAL RESULTS ABOUT LOW-RANK MATRIX ESTIMATION

A. Analyzis of the performance of PCA

In this section we analyze the performance of a maximum likelihood algorithm by estimating the behavior of the (replica symmetric) state evolution in the limit where the interactions are given by $\exp(\beta g(Y, w))$ with $\beta \to +\infty$, and the prior does not contain hard constraints and is independent of β. Note that PCA and related spectral methods correspond to taking $g(Y, w) = -(Y - w)^2/2$. The presented method allows us to analyze the property of the generalized spectral method where $g(Y, w)$ can be taken to be any function including for instance $g(Y, w) = -(D(Y) - w)^2/2$ which would correspond to performing PCA on an elementwise function D of the matrix Y_{ij}, this can be for instance the Fisher score matrix S.

1. *Maximum likelihood for the symmetric $XX^\top$ case*

Maximum likelihood can be seen within the Bayesian approach as analyzing the following posterior for $\beta \to \infty$

$$P(X|Y) = \frac{1}{Z(Y)} \prod_{1 \le i \le N} \frac{\exp(-\|x_i\|_2^2/2)}{\sqrt{2\pi}^r} \prod_{1 \le i < j \le N} \exp\left(\beta g\left(Y, \frac{x_i^\top x_j}{\sqrt{N}} \right) \right). \tag{171}$$

The function $g(Y, w)$ defines the parameters $\widehat{\Delta}$ (120), $\widetilde{\Delta}$ (75), and $\overline{R}$ (76). We want to analyse the overlap between $\hat{X}$ and X_0 in the limit $N \to \infty$, $\beta \to \infty$ since then the posterior will be dominated by the likelihood terms $g(Y, w)$. We put here a prior P_X Gaussian to ensure that $Z(Y) < +\infty$. We could have chosen any β-independent prior $P_X(x)$ as long as the support of $P_X(x)$ is the whole $\mathbb{R}^r$. As $\beta, N \to \infty$ the details of $P_X(x)$ will be washed away. One can write the state evolution equations (126-128)

$$M_x^{t+1} = \beta \Sigma_x^{t+1} \frac{M_x^t \Sigma_0}{\widehat{\Delta}}\,, \tag{172}$$

$$Q_x^{t+1} = \beta \Sigma_x^{t+1} \left[\frac{M_x^t \Sigma_0 M^t}{\widehat{\Delta}^2} + \frac{Q_x^t}{\widetilde{\Delta}} \right] \beta \Sigma_x^{t+1}\,, \tag{173}$$

$$\beta \Sigma_x^{t+1} = \left[\frac{1}{\beta} + Q_x^t \left(\frac{1}{\widetilde{\Delta}} - \overline{R} \right) - \Sigma_x^t \left(\frac{(\beta - 1)}{\widetilde{\Delta}} + \overline{R} \right) \right]^{-1}, \tag{174}$$

where $\Sigma_0 = \langle x_0 x_0^\top \rangle_{P_{X_0}} \in \mathbb{R}^{r \times r}$. We take the limit of $\beta \to \infty$ to get

$$M_x^{t+1} = \Sigma'^{t+1} \frac{M_x^t \Sigma_0}{\widehat{\Delta}}\,, \tag{175}$$

$$Q_x^{t+1} = \Sigma_x'^{t+1} \left[\frac{M_x^t \Sigma_0 M_x^t}{\widehat{\Delta}^2} + \frac{Q_x^t}{\widetilde{\Delta}} \right] \Sigma_x'^{t+1}\,, \tag{176}$$

$$\Sigma_x'^{t+1} = \left[Q_x^t \left(\frac{1}{\widetilde{\Delta}} - \overline{R} \right) - \frac{\Sigma_x'^t}{\widetilde{\Delta}} \right]^{-1}, \tag{177}$$

where $\Sigma'^t = \lim_{\beta\to+\infty} \beta\Sigma^t$.

In general, effects of replica symmetry breaking have to be taken into account in the analysis of maximum likelihood. One exception are the spectral methods for which we take $g(Y,w) = (D(Y)-w)^2/2$, where D is some element-wise function. In that case the maximum likelihood reduces to computation of the spectrum of the matrix $D(Y)$. Obtaining the spectrum is a polynomial problems which is a sign that no replica symmetry breaking is needed to analyze the performance of the spectral methods on element-wise functions of the matrix Y.

Following the derivation of the state evolution, eq. (129), we get that at the fixed point of the state evolution the spectral estimator $\hat{x}_i$ is distributed according to

$$\hat{x_i} = \Sigma'_x \left[\frac{M_x}{\widehat{\Delta}} x_i^0 + \sqrt{\frac{Q_x}{\widetilde{\Delta}}} W_i \right] , \tag{178}$$

where x_i^0 is the planted signal or the rank-one perturbation, and the W_i are independent (in the leading order in N) Gaussian variables of mean 0 and covariance matrix I_r.

2. MSE achieved by the spectral methods

When one uses spectral methods to solve a low rank matrix estimation problem one computes the r leading eigenvalues of the corresponding matrix and then one is left with the problem of what to do with the eigenvectors. Depending on what problem one tries to solve one can for instance cluster the eigenvectors using the k-means algorithm. A more systematic way is the following: We know by (178) that the elements of the eigenvectors $\hat{x}_i$ can be written as a random variable distributed as

$$\hat{x}_i = \hat{M} x_i^0 + \sqrt{\hat{Q}} W_i \,, \tag{179}$$

with $\hat{M} = \Sigma'_x M_x/\widehat{\Delta}$, and $\hat{Q} = Q_x(\Sigma'_x)^2/\widetilde{\Delta}$, where M_x, Q_x and Σ'_x are fixed points of the state evolution equations (175-177). This formula allows us to approach the problem as a low-dimensional Bayesian denoising problem. Writing

$$P(\hat{x}_i, x_i^0) = P(\hat{x}_i|x_i^0) P_{X_0}(x_i^0) = P_{X_0}(x_i^0) \frac{1}{\sqrt{\mathrm{Det}\left(2\pi\hat{Q}\right)}} \exp\left(\left(\hat{x}_i^\top - x_i^{0,\top}\hat{M}^\top\right) \hat{Q}^{-1} \left(\hat{x}_i - \hat{M} x_i^0\right)\right) , \tag{180}$$

one gets

$$P(x_i^0|\hat{x}_i) = \frac{P(\hat{x}_i|x_i^0)P_X(x_i^0)}{P(\hat{x}_i)} = \frac{1}{P(\hat{x}_i)} P_{X_0}(x_i^0) \frac{1}{\sqrt{\mathrm{Det}\left(2\pi\hat{Q}\right)}} \exp\left(-\frac{\left(\hat{x}_i^\top - x_i^{0,\top}\hat{M}^\top\right) \hat{Q}^{-1} \left(\hat{x}_i - \hat{M} x_i^0\right)}{2} \right) . \tag{181}$$

By taking the average with respect to the posterior probability distribution one gets for the spectral estimator

$$\mathbb{E}_{P(x_i^0|\hat{x}_i)}\left[x_i^0\right] = f_{\mathrm{in}}^x \left(\hat{M}^\top \hat{Q}^{-1} \hat{M}, \hat{M}^\top \hat{Q}^{-1} \hat{x}_i \right) . \tag{182}$$

By combining (182) with (179) one gets that the mean-squared error achieved by the spectral method is given by

$$\mathrm{MSE}_{\mathrm{PCA}} = \mathbb{E}_{x^0,W} \left\{ \left[x^0 - f_{\mathrm{in}}^x \left(\hat{M}^\top \hat{Q}^{-1} \hat{M}, \hat{M}^\top \hat{Q}^{-1} \hat{M} x^0 + \sqrt{\hat{M}^\top \hat{Q}^{-1} \hat{M}} W \right) \right]^2 \right\} , \tag{183}$$

where the W are once again Gaussian variables of zero mean and unit covariance, and the variable x^0 is distributed according to P_{X_0}. In the figures presented in subsequent sections we evaluate the performance of PCA via (183).

B. Zero-mean priors, uniform fixed point and relation to spectral thresholds

This section summarizes properties of problems for which the prior distribution P_x has zero mean. We will see that in those cases a particularly simple fixed point of both the Low-RAMP and its state evolution exists. We analyze the stability of this fixed points, and note that linearization around it leads to a spectral algorithm on the Fisher

score matrix. As a result we observe equivalence between the corresponding spectral phase transition and a phase transition beyond which Low-RAMP performs better than a random guess based on the prior.

We do stress, however, that the results of this section hold only when the prior has zero mean, and do not hold for generic priors of non-zero mean. So that the spectral phase transitions known in the literature are in general not related to the physically meaningful phase transitions we observe in the performance of Low-RAMP or in the information theoretically best performance.

1. *Linearization around the uniform fixed point*

From the definition of the thresholding function (56), it follows that that $\hat{x}_i = 0$, $\forall 1 \leq i \leq n$ is a fixed point of the self-averaged low-RAMP equations (66,67,73,74) whenever

$$\int \mathrm{d}x P_X(x) x = \langle x \rangle_{P_X} = 0, \quad \text{and} \quad \overline{R} = 0\,. \tag{184}$$

We will call $\hat{x}_i = 0$, $\forall 1 \leq i \leq n$ the *uniform fixed point*. The interpretation of this fixed point is that according to it there is no information about the planted configuration X_0 in the observed values Y_{ij} and the estimator giving the lowest error is the one that simply sets every variable to zero. When this is the stable fixed point with highest free energy then this is indeed the Bayes-optimal estimator.

In previous work on inference and message passing algorithms [36, 70] we learned that when a uniform fixed point of the message passing update exists it is instrumental to expand around it and investigate the spectral algorithm to which such a procedure leads. We follow this strategy here and expand the Low-RAMP equations around the uniform fixed point. In the linear order in $\hat{x}$, the term A_X^t is negligible, from the definition of the thresholding function (56) one gets

$$\hat{X}^{t+1} = \left(\frac{S}{\sqrt{N}} \hat{X}^t - \hat{X}^{t-1} \frac{\Sigma_x}{\widetilde{\Delta}} \right) \langle x x^\top \rangle_{P_X}\,. \tag{185}$$

Where Σ_x is the average value of the variance of the estimators, as defined in (114), at the uniform fixed point. In the Bayes optimal setting we remind from equation (150) that we moreover have $\Sigma_x = \langle x x^\top \rangle_{P_X}$.

If we consider this last equation as a fixed point equation for $\hat{X}$ we see that columns of $\hat{X}$ are related to the eigenvectors of the Fisher score matrix S. Expanding around the uniform fixed point the Low-RAMP equations thus yields a spectral algorithm that is essentially PCA applied to the matrix S (not the original dataset Y).

For the bipartite model ($UV^\top$ case) the situation is analogous. The self-averaged Low-RAMP equations have a uniform fixed point $\hat{u}_i = 0\ \forall i$, $\hat{v}_j = 0\ \forall j$ when $\overline{R} = 0$ and when the priors P_U and P_V have zero mean. Expanding around this uniform fixed point gives a linear operator whose singular vectors are related to the left and right singular vectors of the Fisher score matrix S.

2. *Example of the spectral decomposition of the Fisher score matrix*

Spectral method always come to mind when thinking about estimation of low-rank matrices. Analysis of the linearized Low-RAMP suggests that in cases where we have some guess about the form of the output channel $P_{\text{out}}(Y|w)$ then the optimal spectral algorithm should not be ran on the data matrix Y_{ij} but instead on the Fisher score matrix S_{ij} defined by (18). This was derived in [40] and further studied in [55]. In this section we give an example of a case where the spectrum of Y_{ij} does not carry any information for some region of parameter, but the one of S_{ij} does.

Consider as an example the output channel to be

$$P_{\text{out}}(Y|w) = \frac{1}{2} \exp\left(-|Y - w|\right)\,. \tag{186}$$

This is just an additive exponential noise. The Fisher score matrix S_{ij} for this channel is

$$S_{ij} = \ \text{sign}(Y_{ij})\,. \tag{187}$$

Consider the rank one case when the true signal distribution P_{X_0} is Gaussian of zero mean and variance σ.

Now let us look at the spectrum of both Y and S in Fig. 2. We plot the spectrum of S and Y for $\sigma = 1.4$. For this value of variance we see that an eigenvalue associated with an eigenvector that carries information about the planted

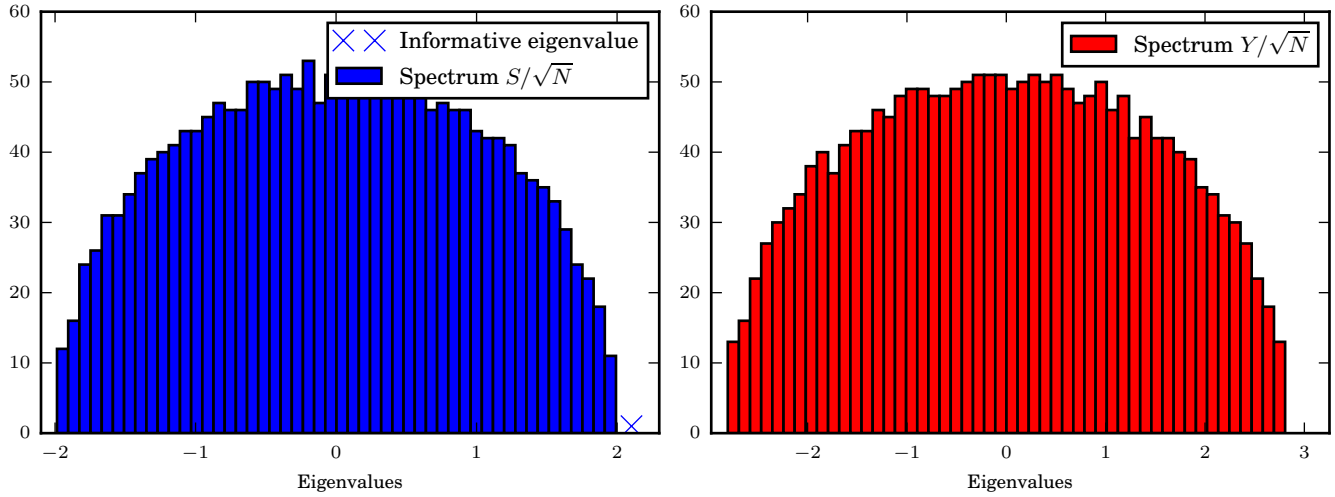

FIG. 2. Spectrum of the Fisher score matrix $S/\sqrt{N}$ and of the data matrix $Y/\sqrt{2N}$ for the same instance of a problem with exponential output noise in the rank-one symmetric $XX^\top$ case. The planted configuration is generated from a Gaussian of zero mean and variance 1.4. We see that an eigenvalue is out of the bulk for S but not for Y. The data were generated on a system of size $N = 2000$.

configuration gets out of the bulk of S but not of Y. Even though some information on the signal was encoded into Y in that specific case one had to take the absolute value of Y to be able to recover an informative eigenvalue.

This situation can be quantified using the spectral analysis of section IV A applied to two different noise channels $g_1(Y, w) = -\beta(\,\mathrm{sign}(Y) - w)^2/2$ and $g_2(Y, w) = -\beta(Y - w)^2/4$ and taking the limit $\beta \to \infty$ as in (171). For these noise channels a theoretical analysis of the top eigenvectors of S and Y can be performed using theory presented in section IV A. Taking square of (175) and dividing by (176) one can show that the state evolution equation describing the overlap of the top positive eigenvectors of Y and S is given by the only stable fixed point of the following update equation

$$\frac{\left(m_x^{t+1}\right)^2}{q_x^{t+1}} = \frac{\frac{{m_x^t}^2}{q_x^t}\frac{\sigma^2\widetilde{\Delta}}{\widehat{\Delta}^2}}{1 + \frac{{m_x^t}^2}{q_x^t}\frac{\sigma\widetilde{\Delta}}{\widehat{\Delta}^2}}\,, \tag{188}$$

$$\widehat{\Delta} = \widetilde{\Delta} = 1 : \text{for } g_1(Y, w)\,, \tag{189}$$

$$\widehat{\Delta} = \widetilde{\Delta} = 2 : \text{for } g_2(Y, w)\,, \tag{190}$$

where $\widehat{\Delta}$ and $\widetilde{\Delta}$ are computed when $\beta = 1$. The trivial fixed point $\frac{{m_x^t}^2}{q_x^t} = 0$ of this equation is unstable as soon as

$$\sigma \geq \frac{\widehat{\Delta}^2}{\widetilde{\Delta}}\,. \tag{191}$$

This analysis tells us that the top eigenvectors are correlated with the planted solution x^0 when $\sigma > 1$ for S and $\sigma > 2$ for Y. Therefor for $\sigma = 1.4$ the Fisher score matrix has informative leading eigenvector, while the top eigenvector of Y does not contain any information on the planted solution.

3. Stability of the uniform fixed point in Bayes-optimal setting

In this section we restrict for simplicity to the Bayes optimal setting defined by eq. (7). As we derived in section III the evolution of the Low-RAMP algorithm can be tracked using the state evolution equations. In the Bayes optimal case we have $\overline{R} = 0$ therefore the (sufficient) condition for the existence of the uniform fixed point is to have prior P_X (or both P_U and P_V) of zero mean. The existence of a uniform fixed point of the Low-RAMP algorithm translates into the existence of a fixed point of the state evolution with $M_x^t = 0$ for the symmetric case (or $M_u^t = M_v^t = 0$ for the bipartite case).

The stability of this fixed point is analyzed by expanding in linear order the state evolution equation (151) around the uniform fixed point, taking into account the definition of the thresholding function (56). For the $XX^\top$ case this gives

$$M_x^{t+1} = \frac{\Sigma_x M_x^t \Sigma_x}{\Delta}\,, \tag{192}$$

where Σ_x in the Bayes-optimal case is the covariance matrix of the signal (and prior) distribution as given by eq. (150). Calling $\lambda_{\max}^x$ the largest eigenvalue of the covariance of the distribution of the signal-elements, Σ_x, we obtain a simple criterion for the stability of the uniform fixed point

$$\begin{cases} \Delta_c = (\lambda_{\max}^x)^2 < \Delta \Rightarrow \text{stable} \\ \Delta < \Delta_c = (\lambda_{\max}^x)^2 \Rightarrow \text{unstable} \end{cases}. \tag{193}$$

It is useful to specify that for the rank-one case where both Σ_x and M_x are scalars we get $\Sigma_x = \langle x_0^2 \rangle$ to be the variance of the prior distribution P_x. For the rank one, $r = 1$, case the stability criterium becomes

$$\begin{cases} \Delta_c = \langle x_0^2 \rangle^2 < \Delta \Rightarrow \text{stable} \\ \Delta < \Delta_c = \langle x_0^2 \rangle^2 \Rightarrow \text{unstable} \end{cases}. \tag{194}$$

Interestingly, the criteria (193) is the same as the criteria for the spectral phase transition of the Fisher score matrix S. When the uniform fixed point is not stable the Fisher score matrix has an eigenvalue going out of the bulk [2, 27]. We stress here that this analysis is particular to signals of zero mean. If the mean is non-zero the spectral threshold does not change, but the Bayes optimal performance gets better and hence superior to PCA.

The critical value of Δ_c separates two parts of the phase diagram

- For $\Delta > \Delta_c$ inference is algorithmically hard or impossible. The Low-RAMP algorithm (and sometimes it is conjectured that all other polynomial algorithms) will not be able to get a better MSE than corresponding to random guessing from the prior distribution.

- For $\Delta < \Delta_c$ inference better than random guessing is algorithmically efficiently tractable. The Low-RAMP and also PCA give an MSE strictly better that random guessing from the prior.

For the bipartite case, $UV^\top$, the stability analysis is a tiny bit more complicated since there are two order parameters M_u^t and M_v^t. Linearization of the state evolutions leads to

$$M_u^t = \alpha \frac{\Sigma_u M_v^t \Sigma_u}{\Delta}\,, \tag{195}$$

$$M_v^{t+1} = \frac{\Sigma_v M_u^t \Sigma_v}{\Delta}\,, \tag{196}$$

where in the Bayes-optimal case the Σ_u and Σ_v are simply the covariances of the prior distribution P_U and P_V, i.e. $\Sigma_u = \langle uu^\top \rangle_{P_U(u)}$ and $\Sigma_v = \langle vv^\top \rangle_{P_V(v)}$. By replacing (196) in (195) one gets

$$M_u^{t+1} = \left(\frac{\sqrt{\alpha}\Sigma_u \Sigma_v}{\Delta}\right) M_u^t \left(\frac{\sqrt{\alpha}\Sigma_u \Sigma_v}{\Delta}\right)^\top. \tag{197}$$

Calling $\lambda_{\max}^{uv}$ the largest eigenvalue of the matrix $\Sigma_u \Sigma_v$ gives us the stability criteria of the uniform fixed point in the bipartite case as

$$\begin{cases} \Delta_c = \sqrt{\alpha}\lambda_{\max}^{uv} < \Delta \Rightarrow \text{Stable} \\ \Delta < \Delta_c = \sqrt{\alpha}\lambda_{\max}^{uv} \Rightarrow \text{Unstable} \end{cases}. \tag{198}$$

Also this criteria agrees with the criteria for spectral phase transition for the Fisher score matrix and also here Δ_c separates two parts of the phase diagram, one where estimating the signal betten than randomly sampling from the prior distribution is not posible with Low-RAMP (and conjecturally with no other polynomial algorithm), and another where the MSE provided by Low-RAMP or PCA is strictly better than the one achieved by guessing at random.

C. Multiple stable fixed points: First order phase transitions

The narrative of this paper is to transform the high-dimensional problem of low-rank matrix factorization to analysis of stable fixed point of the Low-RAMP algorithms and correspondingly of the state evolution equations. A case that deserves detailed discussion is when there exists more than one stable fixed point. The present section is devoted to this discussion in the Bayes-optimal setting, where the replica symmetric assumption is fully justified and hence a complete picture can be obtained.

We encounter two types of situations with multiple stable fixed points

- **Equivalent fixed points due to symmetry:** The less interesting type of multiple fixed points arrises when there is an underlying symmetry in the definition of the problem, then both the state evolution and and Low-RAMP equations have multiple fixed points equivalent under the symmetry. All these fixed points have the same Bethe and replica free energy. One example of such a symmetry is a Gaussian prior of zero mean and isotropic covariance matrix. Then there is a global rotational symmetry present. Another example of a symmetry is given by the community detection problem defined in section I D 2 d. In the case of r symmetric equally sized groups with connectivity matrix (37) there is a permutational symmetry between communities so that any fixed of the state evolution or Low-RAMP equations exists in $r!$ versions.

- **Non-equivalent fixed points**. More interesting case is when Low-RAMP and the state evolution equations have multiple stable fixed points, not related via any symmetry, having in general different free-energies. This is then related to phenomenon that is in physics called the *first order phase transition*. This is the type that we will discuss in detail in the present section.

In general, it is the fixed point with highest free energy that provides the true marginals of the posterior distribution (we remind change of sign in our definition of the free energy w.r.t. the standard physics definition). From an algorithmic perspective, it is the fixed point achieved from uninformed initialization (see below) that gives a way to compute the error achievable by the Low-RAMP algorithm. A conjecture that appears in a number of papers analyzing Bayes optimal inference on random instances is that the error reached by Low-RAMP is the best achievable with a polynomial algorithm. Note, however, that replica symmetry breaking effects might play a role out of the equilibrium solution and hence may influence the properties of the best achievable mean-squared-error.

1. *Typical first order phase transition: Algorithmic interpretation*

The concept of a first order phase transition is best explained on a specific example. For the sake of the explanation we will consider the symmetric $XX^\top$ case in the Bayes optimal setting. For the purpose of giving a specific example we consider the Gaussian output channel with variance of the noise Δ, the signal is drawn from the spiked (i.e. $r = 1$) Rademacher-Bernoulli model with fraction of non-zeros being $\rho = 0.08$.

In Fig. 3 we plot all the fixed points of the state evolution equation and the corresponding value of the replica free energy (153) as a function of Δ/ρ^2. The equations for the state evolution specific to the spiked Rademacher-Bernoulli model are given by (214). For this model the uniform fixed point exists and is stable down to $\Delta_c = \rho^2$, eq. (193). The numerically stable fixed points are drawn in blue, the unstable ones in red. We focus on the interval of Δ where more than one stable fixed point of the state evolution (151) exists. We use the example of the spiked Rademacher-Bernoulli model for the purpose of being specific in figure 3. The definitions and properties defined in the rest of this section are generic and apply to all the settings considered in this paper, not only to the spiked Rademacher-Bernoulli model.

Let us define two (possibly equal) stable fixed points of the state evolution as follows:

- $M_{\text{Uninformative}}(\Delta)$ is the fixed point of the state evolution one reaches when initialising the $M^{t=0} = \epsilon I_r$ (with ϵ very small and positive, I_r being the identity matrix). We call this the uninformative initialisation.

- $M_{\text{Informative}}(\Delta)$ is the fixed point of the state evolution one reaches when initialising the $M^{t=0} = \langle xx^\top \rangle_{P_X}$. This is the informative initialisation where we start as if the planted configuration was known.

In principle there could be other stable fixed points apart of $M_{\text{Uninformative}}(\Delta)$ and $M_{\text{Informative}}(\Delta)$, but among all the examples that we analyzed in this paper we have not observed any such case. In this paper we hence discuss only the case with at most two stable fixed points, keeping in mind that if more stable fixed points exist than the theory does apply straightforwardly as well (the physical fixed point is still the one of highest free energy), and there could be several first order phase transitions following each other as Δ is changed.

With two different stable fixed points existing for some values of Δ we observe three critical values defined as follows

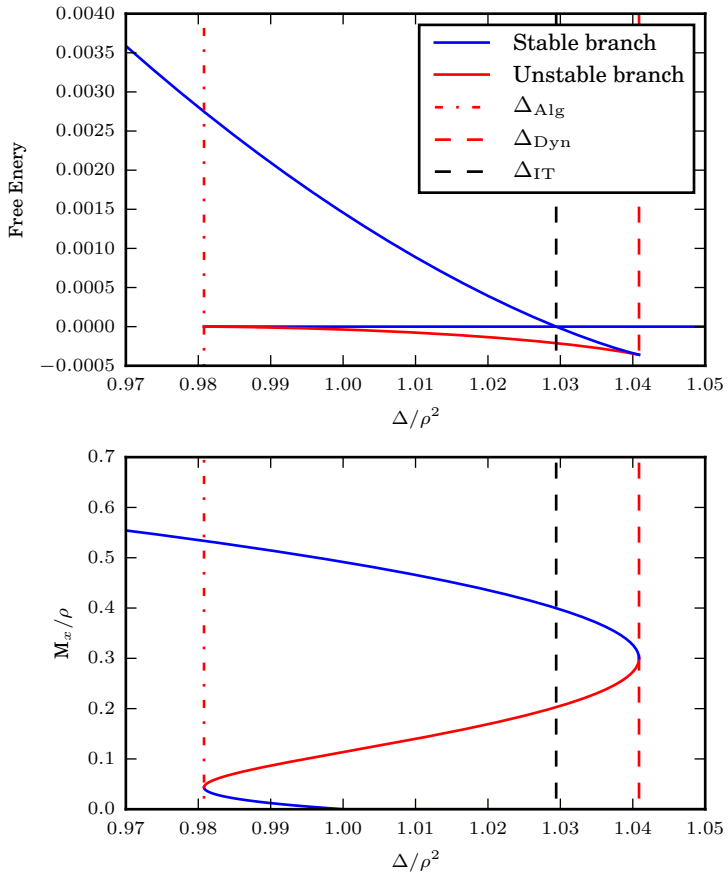

FIG. 3. In the lower pannel, we plot as a function of Δ/ρ^2 all the fixed points M_x/ρ of state evolution equations (151) for the spiked Rademacher-Bernoulli model with fraction of non-zero being $\rho = 0.08$. Numerically stable fixed points are in blue, unstable in red. We remind that the order parameter M_x is related to the mean-squared error as $\text{MSE}_X = \text{Tr}\left[\langle x^0 x^{0,\top}\rangle - M_x\right]$. In the upper pannel, we plot the replica free energy (153) corresponding to these fixed points, again as a function of Δ/ρ^2. When there are multiple stable fixed points to the SE equations the one that corresponds to performance of the Bayes optimal estimation is the one with the largest free energy (we remind that with respect to the most common physics notation we defined the free energy as the positive logarithm of the partition function). The Δ for which these two branches cross in free-energy is called the information theoretic phase transition, $\Delta_{\rm IT}$. The two spinodal transitions $\Delta_{\rm Alg}$ and $\Delta_{\rm Dyn}$ are where the lower MSE an higher MSE stable fixed point disappears. The $\Delta_c/\rho^2 = 1$ corresponds to the spectral transition at which the uniform fixed point become unstable.

- $\Delta_{\rm Alg}$ called the *algorithmic spinodal transition*, is the value of Δ at which the fixed point $M_{\rm Uninformative}$ stops existing and becomes equal to $M_{\rm Informative}$.

- $\Delta_{\rm IT}$ called the *information theoretic phase transition*, is the value of Δ at which the two fixed points $M_{\rm Uninformative} \neq M_{\rm Uninformative}$ exist and have the same replica free energy (153).

- $\Delta_{\rm Dyn}$ called the *dynamic spinodal transition*, is the value of Δ at which the fixed point $M_{\rm Informative}$ stops existing and becomes equal to $M_{\rm Uninformative}$.

In Fig. 3 these three transition are marked by vertical dashed lines, the information theoretic transition in black and the two spinodal transition in red.

We recall that for the priors of zero mean, where the uniform fixed points discussed in section IV B 1 exists, the stability point of the uniform fixed point Δ_c (193) is in general unrelated to the $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$. In the example presented in Fig. 3 of spiked Rademacher-Bernoulli model at $\rho = 0.08$ we have $\Delta_{\rm Alg} < \Delta_c < \Delta_{\rm IT}$. In general the position of Δ_c with respect to $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ can be arbitrary, in the following sections we will observe several examples. A notable situation is when $\Delta_c = \Delta_{\rm Alg}$, cases where this happen are discussed in section IV C 3. It should be noted that this is the case in the community detection problem, and since this is a well known and studied example

it is sometimes presented in the literature as the generic case. But from the numerous examples presented in this paper we see that cases where $\Delta_c \neq \Delta_{\rm Alg}$ are also very common.

Phase transitions are loved and cherished in physics, in the context of statistical inference the most intriguing properties related to phase transitions is their implications in terms of average computational complexity. Notably in the setting of Bayes-optimal low-rank matrix factorization as studied in this paper we distinguish two different regions

- **Phase where Low-RAMP is asymptotically Bayes-optimal:** For $\Delta \geq \Delta_{\rm IT}$ and for $\Delta \leq \Delta_{\rm Alg}$ the Low-RAMP algorithm in the limit of large system sizes gives the information theoretically optimal performance. This is either because the fixed point it reaches is unique (up to symmetries) or because it is the one with larger free energy.
- **The hard phase:** For $\Delta_{\rm Alg} < \Delta < \Delta_{\rm IT}$ the estimation error achieved by the Low-RAMP algorithm is strictly larger that the lowest information-theoretically achievable error. On the other hand, and in line with other statistical physics works on inference problems in the Bayes optimal setting, we conjecture that in this region no other polynomial algorithm that would achieve better error than Low-RAMP exists. This conjecture could be slightly modified by the fact that the branch of stable fixed points that do not correspond to the MMSE could present aspects of replica symmetry breaking which could modify its position. Investigation of this is left for future work.

From a mathematically rigorous point of view the results of this paper divide into three parts:

- (a) Those that are rigorous, known from existing literature that is not related to statistical physics considerations. An example is given by the performance of the spectral methods that is better that random guessing for $\Delta < \Delta_c$ [2].
- (b) A second part regroups the results directly following from the analysis of the Bayes-optimal Low-RAMP and the MMSE that are not presented rigorously in this paper, but were made rigorous in a series of recent works [4, 37, 44]. Most of these results are proven and although some cases are still missing a rigorous proof, it is safe to assume that it is a question of time that researchers will fill the corresponding gaps and weaken the corresponding assumptions. For instance the state evolution of Low-RAMP [14, 28, 58], its Bayes-Optimality in the easy phase (at least for problem where the paramagnetic fixed point is not symmetric) and the value of the information theoretically optimal MMSE [4, 13, 37, 44] are all proven rigorously.
- (c) The third kind of claims are purely conjectures. For instance, the claim that among all polynomial algorithms the performance of Low-RAMP cannot be improved (so that the hard phase is indeed hard). Of course proving this in full generality would imply that P≠NP and so we cannot expect that such a proof would be easy to find. At the same time from a broad perspective of understanding average computational complexity this is the most intriguing claim and is worth detailed investigation and constant aim to find a counter-example.

2. Note about computation of the first order phase transitions

In this section we discuss how to solve efficiently the SE equations in the $XX^\top$ Bayes optimal case for rank one. We notice that for a given prior distribution P_X the only way the symmetric Bayes optimal state evolution, eq. (151), depends on the noise parameter Δ is via the ratio m^t/Δ. One can write the SE equations in the form (152)

$$m^{t+1} = f_{P_X}^{\rm SE}\left(\frac{m^t}{\Delta}\right) . \tag{199}$$

Let us further define fixed points of (199) in a parametric way

$$\Delta = \frac{f_{P_X}^{\rm SE}(x)}{x} , \tag{200}$$

$$m = f_{P_X}^{\rm SE}(x) . \tag{201}$$

To get a fixed point (m, Δ) of (199) we choose a value of x and compute $(f(x), f(x)/x)$. We observe from the form of the state evolution (rank-one symmetric Bayes optimal case) that $f(x)$ is a non-decreasing function of x. We further observe that (m, Δ) is a stable fixed points if and only if $\partial_x \Delta(x) < 0$. The two spinodal thresholds $\Delta_{\rm Alg}$ and $\Delta_{\rm Dyn}$ are defined by loss of existence of corresponding stable fixed points and they can hence be computed as

$$\Delta_{\rm Alg}, \Delta_{\rm Dyn} \in \left\{ \Delta(x), x \in \mathbb{R}^+, \frac{\partial \Delta(x)}{\partial x} = 0 \right\} . \tag{202}$$

The information theoretic transition $\Delta_{\rm IT}$ relies on computation of the replica free energy (153). Using the fact that

$$\frac{\partial\phi(m,\Delta)}{\partial m} = \frac{1}{2\Delta}\left(f_{P_X}^{\rm SE}\left(\frac{m}{\Delta}\right) - m\right) \tag{203}$$

allows us to compute the difference in energy between a fixed point m, Δ and the uniform fixed point $m = 0$ as

$$\phi(m(x),\Delta(x)) - \phi(0,\Delta(x)) = \frac{1}{2}\left[\int_0^x {\rm d}u\, f_{P_X}^{\rm SE}(u) - \frac{x f^{\rm SE}(x)}{2}\right] . \tag{204}$$

The $x_{\rm IT}$ for which (204) is zero then gives the information theoretic phase transition $\Delta_{\rm IT} = f(x_{\rm IT})/x_{\rm IT}$.

3. *Sufficient criterium for existence of the hard phase*

This section is specific to the Bayes-optimal cases when the prior P_X has a zero mean and hence the uniform fixed point of the Low-RAMP and the state evolution exists. In section IV B 3 we derived that the uniform fixed point is stable at $\Delta > \Delta_c$ and unstable for $\Delta < \Delta_c$. It follows from the theory of bifurcations that the critical point where a fixed-point changes stability must be associated with an onset of another close-by fixed point. In general there are two possibilities

- **2nd order bifurcation.** If the fixed point close to the uniform fixed point departs in the direction of smaller $\Delta < \Delta_c$, where the uniform fixed point is unstable, then this close-by fixed point is stable. This case corresponds to Fig. 3. Behaviour in the vicinity of the uniform fixed point then does not let us distinguish between (a) existence of a first order phase transition at lower Δ (as in Fig. 3), or (b) continuity on the MMSE down to $\Delta = 0$ with no algorithmically hard phase existing in that case.
- **1st order bifurcation.** If the fixed point close to the uniform fixed point departs in the direction of larger $\Delta > \Delta_c$, where the uniform fixed point is stable, then this close-by fixed point is unstable. In that case, the fixed point that is stable from $\Delta < \Delta_c$ is not close to the uniform fixed point and this case forces existence of a first order phase transition with $\Delta_c = \Delta_{\rm Alg}$.

Expansion of the state evolution (151) around the uniform fixed point gives us a closed form criteria to distinguish whether the phase transition happening at Δ_c is a 1st or 2nd order bifurcation. In case of a 2nd order bifurcation, this expansion allows us to compute the mean squared error obtained with Low-RAMP close to Δ_c.

For specificity we consider the rank-one, $r = 1$ case, and expand eq. (151) to 2nd order to get

$$m^{t+1} = m^t \frac{\langle x_0^2\rangle^2}{\Delta} - \frac{(m^t)^2}{\Delta^2}\left(\langle x_0^2\rangle^3 - \frac{\langle x_0^3\rangle^2}{2}\right) + O\left((m^t)^3\right) , \tag{205}$$

where the mean $\langle\cdots\rangle$ of x_0 are taken with respect to $P_{X_0} = P_X$. This is done by expanding $f_{\rm in}^x(A,B)$ to order 4 in B and 2 in A. All the derivatives

$$\forall i,j \quad \frac{\partial^{i+j} f_{\rm in}^x}{\partial A^i \partial B^j}(A=0,B=0) \tag{206}$$

are linked to moments of the density probability P_X.

The stability criteria $\Delta_c = \langle x_0^2\rangle^2$ appears once again as in section IV B 3. Below $\Delta < \Delta_c$ the uniform fixed point $m = 0$ is unstable and m^t will converge towards another fixed point different from $m = 0$. For $\Delta > \Delta_c = \langle x_0^2\rangle$ the uniform fixed point $m = 0$ is stable. Using expansion (205) near $\Delta \simeq \Delta_c = \langle x_0^2\rangle$ we can write what is the other fixed point next to $m_{\rm uniform} = 0$, we get

$$m_{\rm close-by} = \frac{\Delta_c(\Delta_c - \Delta)}{\Delta_c^{3/2} - \frac{\langle x_0^3\rangle^2}{2}} + O\left((\Delta - \Delta_c)^2\right) . \tag{207}$$

By definition of the order parameters in the Bayes-optimal setting we must have at a fixed point $m \geq 0$, therefore we distinguish two cases

- If $\langle x_0^3\rangle^2 < 2\langle x_0^2\rangle^3$, eq. (207) is a stable fixed point in the region $\Delta < \Delta_c$. Eq. (207) then gives the expansion of this fixed point. This situation corresponds to Fig. 3 where the Rademacher-Bernoulli prior has zero 3rd moment. This is the 2nd order bifurcation at Δ_c.

- $\langle x_0^3\rangle^2 > 2\langle x_0^2\rangle^3$ eq. (207) is an unstable (and hence irrelevant) fixed point in the region $\Delta > \Delta_c$. But also in this case there must be a stable fixed point for $\Delta < \Delta_c$, but this fixed point cannot have a small values of m. The only way a stable fixed point can appear in this case is by a discontinuous (1st order) transition at Δ_c. This is the 1st order bifurcation at Δ_c.

To summarize, we obtained a simple (sufficient) criteria for the existence of a first order phase transition with $\Delta_{\rm Alg} = \Delta_c$. Notably there is a 1st order phase transition when

$$\begin{cases} \langle x_0\rangle_{P_X} = 0 \\ \langle x_0^3\rangle^2_{P_X} > 2\langle x_0^2\rangle^3_{P_X} \end{cases} . \tag{208}$$

The more skewed the prior distribution is the easier the problem is. Till a point where if the skewness of the signal is bigger than $\sqrt{2}$ then a first order phenomena will appear in the system. In this case the MSE achieved by the Low-RAMP algorithm becomes discontinuously better than $\text{MSE}_{\text{uniform}} = \langle x_0^2\rangle_{P_X}$

On the other hand when the criteria (208) is not met, then for $\Delta < \Delta_c$ the MSE achieved by the Low-RAMP algorithm is in first approximation equal to

$$\text{MSE}(\Delta) = \langle x_0^2\rangle^2_{P_X} - \frac{\Delta_c(\Delta_c-\Delta)}{\Delta_c^{3/2} - \frac{\langle x_0^3\rangle_{P_X}{}^2}{2}} + O\left((\Delta-\Delta_c)^2\right) . \tag{209}$$

So the MSE obtained by an Low-RAMP algorithm is linear near the transition in $\Delta - \Delta_c$.

Interestingly spectral method also give an MSE linear in $\Delta - \Delta_c$. As derived in section IV A the MSE one achieves using the eigenvectors of matrix S is

$$\text{MSE}_{\text{Spectral}}(\Delta) = \langle x_0^2\rangle_{P_X} - \frac{\Delta_c - \Delta}{\sqrt{\Delta_c}} . \tag{210}$$

From the coefficient of linearity in the error we observe that the error achieved by PCA is always worse than the error achieved by Low-RAMP.

Let us remind that uniform fixed points and 1st order phase transition (at $\Delta_{\rm Alg} = \Delta_c$ or elsewhere) can exist even if the criteria derived above are not met, there are sufficient, not necessary conditions. Examples are included in subsequent sections.

V. PHASE DIAGRAMS FOR BAYES-OPTIMAL LOW-RANK MATRIX ESTIMATION

From now on we restrict our analysis to the Bayes optimal setting as defined in section I B 1, eq. (7). The motivation is to investigate performance of the Bayes-optimal and the Low-RAMP estimators for a set of benchmark problems. We investigate phase diagrams stemming from the state evolution equations and from the corresponding replica free energies summarized in section III D 1.

A. Examples of phase diagram

In this section we present example of phase diagrams for the symmetric low-rank matrix estimation. The first three examples are for rank one, the last two examples are for general rank.

1. *Spiked Bernoulli model*

The spiked Bernoulli model is defined by prior (32) with density ρ of ones, and $1-\rho$ of zeros. This prior has a positive mean and consequently the state evolution does not have the uniform fixed point. This is a problem where one tries to recover a submatrix of size $\rho N \times \rho N$ with mean of elements equal to 1, in a $N \times N$ matrix of lower mean. Using the Bayes-optimal state evolution (126) one gets,

$$m^t = f^{\rm SE}_{\rm Bernoulli}\left(\frac{m^t}{\Delta}\right) , \tag{211}$$

$$f^{\rm SE}_{\rm Bernoulli}(x) = \rho \mathbb{E}_W \left[\frac{\rho}{\rho + (1-\rho)\exp\left(\frac{-x}{2} + W\sqrt{x}\right)}\right] , \tag{212}$$

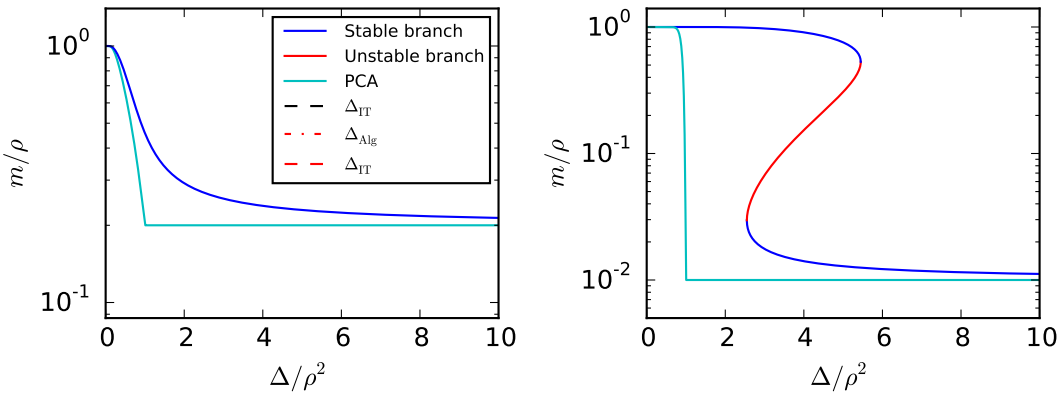

FIG. 4. Plots of the fixed points of the state evolution for the spiked Bernoulli model, MSE being given by MSE $= \rho - m$ (154). The performance of PCA as analyzed in section IV A is plotted for comparison. These two plots are made for $\rho = 0.2$ (left) and $\rho = 0.01$ (right).

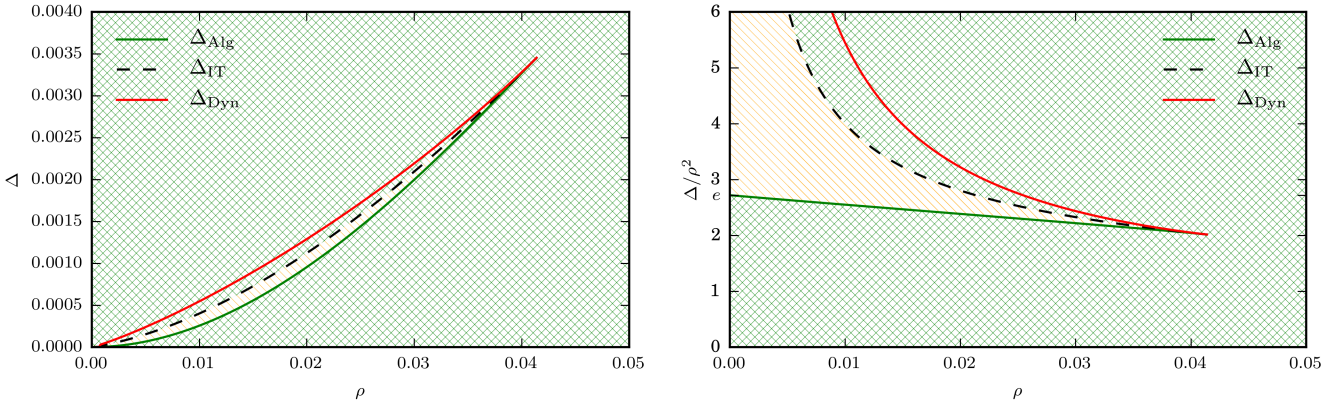

FIG. 5. The phase diagram of the spiked Bernoulli model (32) as a function of the density ρ and effective noise Δ (left) or Δ/ρ^2 (right). There is no phase transition in the system for $\rho > 0.04139$ and a 1st order phase transition for $\rho < 0.04139$. The lower green curve is the algorithmic spinodal $\Delta_{\rm Alg}$ curve, that converges to $\Delta_{\rm Alg} =_{\rho\to 0} e\rho^2$. The dashed black line is the information theoretic threshold $\Delta_{\rm IT}$. The upper red curve is the dynamical spinodal $\Delta_{\rm Dyn}$. The orange hashed zone is the hard region in which Low-RAMP does not reach the Bayes-optimal error. In the rest of the phase diagram (green hashed) the Low-RAMP provides in the large size limit the Bayes-optimal error. Note that this is exactly the same phase diagram as presented in [49] (Fig. 5) for the problem of finding one dense subgraph, this is thanks to output channel universality and the fact that large degree sparse graphs have upon rescaling the same phase diagram as dense graphs.

where W is (here and from now on) a Gaussian variables of zero mean and unit variance.

Depending on the value of ρ there are two kinds of behaviour of the fixed points of these equations as a function of the effective noise Δ. We plot the two cases in Fig. 4. For larger values of ρ there is a unique fixed point corresponding to the MMSE that is asymptotically achieved by the Low-RAMP algorithm, this is the regime in which the proof of [14] applies. For small enough values of ρ we do observe a region of $\Delta_{\rm Alg} < \Delta < \Delta_{\rm Dyn}$ where there are 3 fixed points, two stable and one unstable. The replica free energy associated to the these fixed points crosses at $\Delta_{\rm IT}$ so that the higher fixed point in the relevant one at $\Delta < \Delta_{\rm IT}$, and the lower fixed point at $\Delta > \Delta_{\rm IT}$. These phase transitions were defined in section IV C 1. In Fig. 5 we plot the phase transitions $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ as a function of ρ. The y-axes in the left pannel is simply the effective noise parameter Δ, on the right pannel the same data are plotted with Δ/ρ^2 on the y-axes. We observe that the $\Delta_{\rm Alg} =_{\rho\to 0} e\rho^2$, with e being the Euler number, this is the same asymptotic behaviour as obtained previously in [49] (Fig. 5).

2. Rademacher-Bernoulli and Gauss-Bernoulli

Next we analyze the spiked Rademacher-Bernoulli and Gauss-Bernoulli priors defined by (33) and (36). The first thing we notice is that both these distribution have zero mean and variance ρ. This means according to (IV B 1) that there is a uniform fixed point of the SE equations that is stable for $\Delta > \rho^2$. The skewness of both these distribution is 0, which means that at Δ_c there is no discontinuity of the MSE (IV C 3). The SE equations for these models are

$$m^{t+1} = f^{\rm SE}_{\rm Rademacher-Bernoulli}\left(\frac{m^t}{\Delta}\right), \tag{213}$$

$$f^{\rm SE}_{\rm Rademacher-Bernoulli}(x) = \rho \mathbb{E}_W\left[\tanh\left(x + W\sqrt{x}\right)\frac{\rho}{(1-\rho)\frac{\exp(x/2)}{\cosh\left(x+W\sqrt{x}\right)}+\rho}\right], \tag{214}$$

$$m^{t+1} = f^{\rm SE}_{\rm Gaussian-Bernoulli}\left(\frac{m^t}{\Delta}\right), \tag{215}$$

$$f^{\rm SE}_{\rm Gauss-Bernoulli}(x)/\rho = \frac{x}{1+x}\mathbb{E}_W\left[W^2\hat{\rho}(x, W\sqrt{x^2+x})\right], \tag{216}$$

where $\hat{\rho}$ is

$$\hat{\rho}(a, b^2) = \frac{\rho}{(1-\rho)\exp\left(\frac{-b^2}{2(1+a)}\right)(1+a)^{\frac{r}{2}}+\rho}. \tag{217}$$

Both these models have similar phase diagram.

We first illustrate in Fig. 6 the different types of phase transition that we observe for the spiked Rademacher-Bernoulli model as the density ρ is varied. We plot all the fixed points of equation (214) for several values of ρ as a function of the effective noise Δ/ρ^2. The four observed case are the following

- $\rho = 0.097$ example: For ρ large enough (in the present case $\rho > \rho_{\rm tri} = 0.0964$) whatever the Δ there is only one stable fixed point.
- For small enough $\rho < \rho_{\rm tri} = 0.0964$ three different fixed points exist in a range of $\Delta_{\rm Alg}(\rho) < \Delta < \Delta_{\rm Dyn}(\rho)$, where the thresholds $\Delta_{\rm Alg}$, and $\Delta_{\rm Dyn}$ are defined by the limits of existence of the three fixed points. The information theoretic threshold where the free energy corresponding to the two stable fixed points crosses is $\Delta_{\rm IT}$. Depending on the values of ρ we observed 3 possible scenarios of how $\Delta_c = \rho^2$ is placed w.r.t. the other thresholds.
  - $\rho = 0.0909$ example where $\Delta_{\rm Dyn} < \Delta_c$.
  - $\rho = 0.0863$ example where $\Delta_{\rm IT} < \Delta_c < \Delta_{\rm Dyn}$.
  - $\rho = 0.08$ examples where $\Delta_{\rm Alg} < \Delta_c < \Delta_{\rm IT}$.

Finally Fig. 8 presents the four thresholds $\Delta_{\rm Dyn}$, $\Delta_{\rm IT}$, $\Delta_{\rm Alg}$ and Δ_c as a function of the density ρ.

In Fig. 9 we present for completness the comparison between the fixed points on the state evolution and the fixed points of the Low-RAMP algorithm for the Gauss-Bernoulli model, with rank one, Bayes optimal case. The experiment is done on one random instance of size $N = 2 \times 10^4$ and we see the agreement is very good, finite size effect are not very considerable. The data are for the Gauss-Bernoulli model at $\rho = 0.1$, that is in a region where $\Delta_{\rm Alg}$ is so close to Δ_c that in this figure the difference is unnoticeable.

We also compare to the MSE reached by the PCA spectral algorithm and from its analysis eq. (183). We see that whereas both Low-RAMP and PCA work better than random guesses below Δ_c, the MSE reached by Low-RAMP is considerably smaller.

3. Two balanced groups

The next example of phase diagram we present is for community detection with two balanced (i.e. one group is smaller ρN, but both have the same average degree) groups as defined in eqs. (45-46). This is an example of a system where the bifurcation at Δ_c is of a first order, with $\Delta_c = \Delta_{\rm Alg}$. In this problem the prior given by eq. (46), with $\langle x^0\rangle_{P_X} = 0$, $\langle (x^0)^2\rangle_{P_X} = 1$. The output channel is of the stochastic block model type eq. (41-42), leading to effective noise parameter

$$\Delta = \frac{p_{\rm out}(1-p_{\rm out})}{\mu^2}, \tag{218}$$

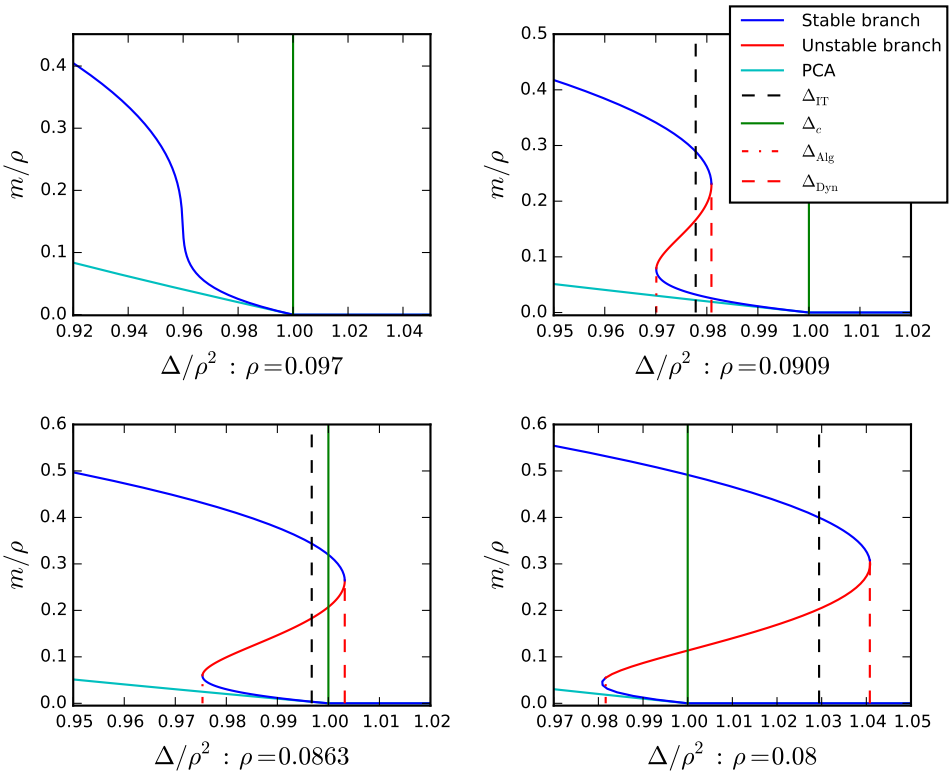

FIG. 6. We plot all the fixed points of the state evolution equations for the spiked Rademacher-Bernoulli model in the Bayes optimal setting as a function of Δ/ρ^2 for four representative values of ρ. The blue curves are stable fixed points, red are the unstable ones. MSE is given by (154). The vertical lines mark the two spinodal transition (dynamical $\Delta_{\rm Dyn}$ and algorithmic $\Delta_{\rm Alg}$, red dashed), the information theoretic transition $\Delta_{\rm IT}$. The vertical green full line mark the stability point of the uniform fixed point $\Delta_c = \rho^2$. We remind that the error MSE $= \rho - m$ achieved by the Bayes optimal estimator corresponds to the upper branch for $\Delta < \Delta_{\rm IT}$, and to the lower branch for $\Delta > \Delta_{\rm IT}$. Error achieved by the Low-RAMP algorithm always correspond to the lower branch (larger error). Note that in the three panels where multiple fixed points exists, the only element that changes is the position of the (spectral) stability threshold Δ_c with respect to the other thresholds.

where μ and $p_{\rm out}$ are the parameters from (41-42). Therefore the uniform fixed point becomes unstable when $\Delta < 1$.

Using eqs. (126) and (46) we get for the state evolution for community detection with two balanced groups

$$m^{t+1} = f^{\rm SE}_{\rm TwoBalanced}\left(\frac{m^t}{\Delta}\right), \forall t, m^t \in [0;1]\,, \tag{219}$$

where

$$f^{\rm SE}_{\rm TwoBalanced}(x) = \int_{-\infty}^{+\infty} \frac{e^{\frac{-u^2}{2}}}{\sqrt{2\pi}} \frac{2\rho(1-\rho)\sinh\left(\frac{x}{2\rho(1-\rho)} + u\sqrt{\frac{x}{\rho(1-\rho)}}\right)}{1+2\rho(1-\rho)\left(\cosh\left(\frac{x}{2\rho(1-\rho)} + u\sqrt{\frac{x}{\rho(1-\rho)}}\right) - 1\right)}\,. \tag{220}$$

To investigate whether $\Delta_c = 1$ is a 1st or 2nd order bifurcation we compute the second order expansion of the state evolution equations as we have done in (205). We find that the expansion of the state evolution up to second order for the two balanced communities is

$$m^{t+1} = f_\rho\left(\frac{m^t}{\Delta}\right) = \frac{m^t}{\Delta} + \left(\frac{m^t}{\Delta}\right)^2 \frac{1-6\rho(1-\rho)}{2\rho(1-\rho)}\,. \tag{221}$$

In a similar fashion as in section (IV C 3) it is the sign of the second order terms that decides between 1st or 2nd order bifurcation at $\Delta_c = 1$. We find that if $\rho(1-\rho) < 1/6$ then the second order derivative of (220) is positive leading

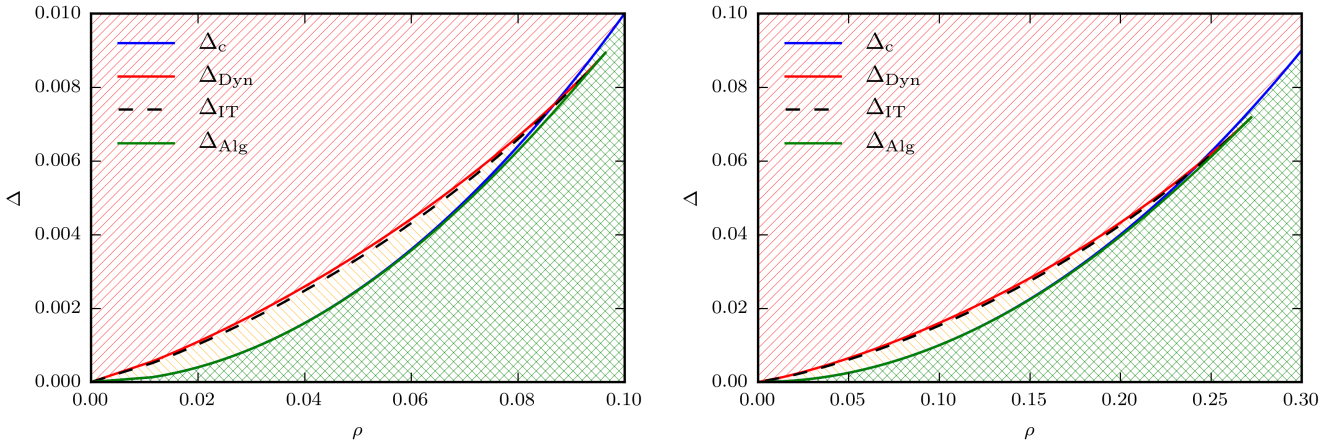

FIG. 7. Phase diagram of the spiked Rademacher-Bernoulli (left hand side pannel) and spiked Gauss-Bernoulli (right hand side pannel) models. We plot Δ as a function of ρ. The local stability threshold of the uniform fixed point $\Delta_c/\rho^2 = 1$ is in blue. The algorithmic spinodal $\Delta_{\rm Alg}$ (green), the dynamical spinodal $\Delta_{\rm Dyn}$ (red) and the information theoretic transition $\Delta_{\rm IT}$ (black dashed) all join into a tri-critical point located at $(\Delta_{\rm tri} = 0.008935, \Delta_{\rm tri}/\rho^2_{\rm tri} = 0.9612, \rho_{\rm tri} = 0.09641)$ for the Rademacher-Bernoulli model (left pannel), and at $(\Delta_{\rm tri} = 0.07182, \Delta_{\rm tri}/\rho^2_{\rm tri} = 0.9693, \rho_{\rm tri} = 0.2722)$ for the Gauss-Bernoulli model (right pannel). The hash materializes the different phases. The easy phase where the Low-RAMP algorithm is Bayes-optimal and achieves better error than random guessing is hashed in green crossed lines, the hard phase where Low-RAMP is suboptimal is hashed in yellow \\, and the impossible phase where even the best achievable error is as bad as random guessing is hashed in red //.

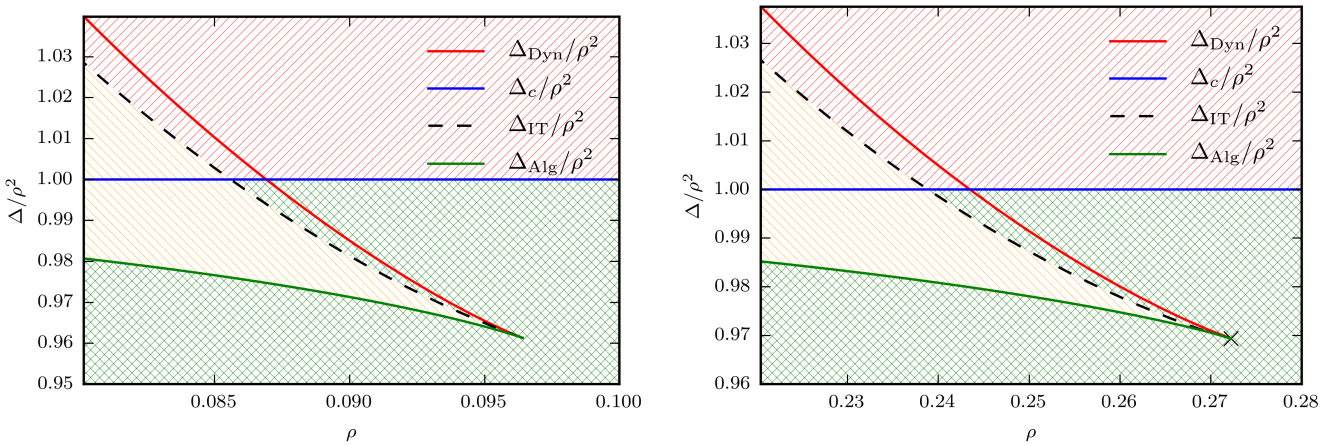

FIG. 8. The same plot as in Fig. 7 zoomed into the region of the tri-critical point with y-axes rescaled by ρ^2. The spiked Rademacher-Bernoulli mode on left hand side pannel, the spiked Gauss-Bernoulli model on the right hand side.

to a jump in MSE when Δ crosses the value $\Delta = 1$ which means that there will be first order phase transition for all

$$\rho \in \left[0; \frac{1}{2} - \frac{1}{\sqrt{12}} \approx 0.21\right] \cup \left[\frac{1}{2} + \frac{1}{\sqrt{12}} \approx 0.89; 1\right] . \tag{222}$$

It turns out that for the two balanced groups this criteria is both sufficient and necessary. Out of the interval (222) the phase transition at Δ_c is of second order, with no discontinuities. Defining the phase transitions $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$, $\Delta_{\rm Dyn}$ as before in section IV C 1, we have $\Delta_{\rm Alg} = \Delta_c$ and we plot the three phase transitions for community detection for two balanced groups in the phase diagram Fig. 10.

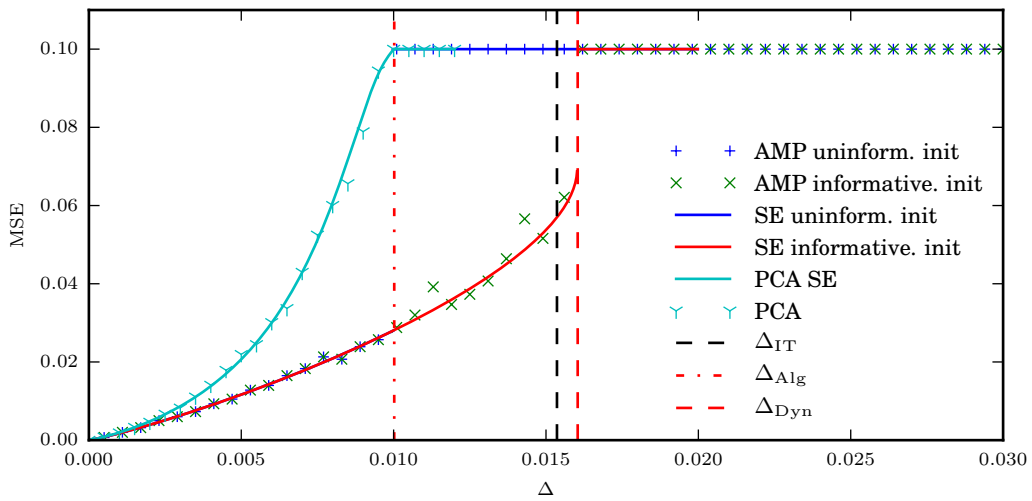

FIG. 9. Comparison between the state evolution and the fixed point of the Low-RAMP algorithm, for the spiked Gauss-Bernoulli model of sparse PCA with rank one and density $\rho = 0.1$. The phase transitions stemming from state evolution are $\Delta_{\rm Alg} \approx \Delta_{\rm c} = 0.01, \Delta_{\rm IT} = 0.0153, \Delta_{\rm Dyn} = 0.0161$. The points are the fixed points of the Low-RAMP algorithm run on one typical instance of the problem of size $N = 20000$. Blue pluses is the MSE reached from an uninformative initialization of the algorithm. Green crosses is the MSE reached from the informative initialization of the algorithm.

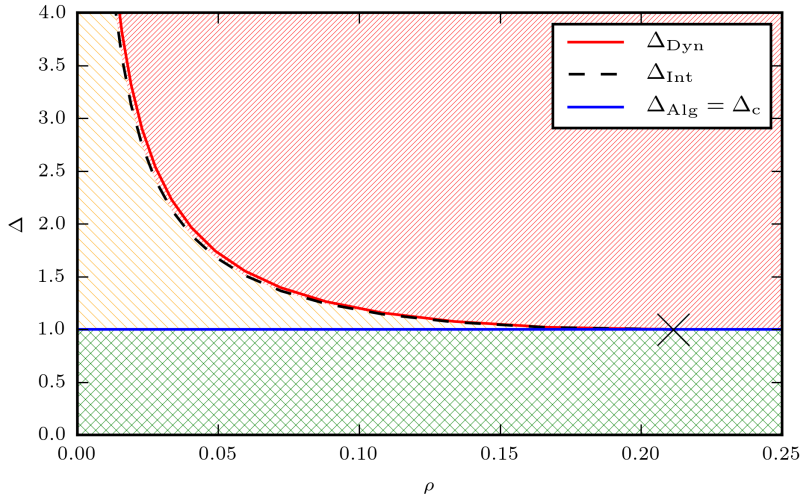

FIG. 10. We plot here $\Delta_{\rm Alg} = \Delta_c$, $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ as a function of ρ for the community detection with two balanced groups. All the curves merge at $\rho = \frac{1}{2} - \frac{1}{\sqrt{12}}$. The hashed regions have the same meaning as in previous figures, red is the impossible inference phase, green is easy and yellow is hard inference.

4. Jointly-sparse PCA generic rank

In this section we discuss analysis of the Gauss-Bernoulli jointly sparse PCA as defined by the prior distribution (35) for a generic rank r. This just means that each vector $x_i \in \mathbb{R}^r$ is either $\vec{0}$ with probability $1-\rho$ or is taken from Gaussian density probability of mean zero and covariance matrix I_r with probability ρ. The prior distribution (35) has a zero mean, therefore the uniform fixed point exist and according to (193) is stable down to $\Delta_c = \rho^2$.

In order to deal with the r-dimensional integrals and $r \times r$ dimensional order parameter M^t we notice that there is a rotational SO(r) symmetry in the problem, which in the Bayes optimal setting implies that

$$M^t = m^t I_r, \tag{223}$$

with m^t being a scalar parameter. The problem is hence greatly simplified, one can then treat the r dimensional

integral in (151). After integration by parts and integration on the sphere one gets

$$m^{t+1} = f^{\rm SE}_{\rm Joint-GB}\left(\frac{m^t}{\Delta}\right), \tag{224}$$

$$f^{\rm SE}_{\rm Joint-GB}(x) = \frac{\rho x}{1+x}\int {\rm d}u \frac{1}{(2\pi)^{\frac{r}{2}}}\exp\left(\frac{-u^2}{2}\right) S_r u^{r-1}\left\{1+\frac{xu^2\left[1-\hat\rho(x,(x^2+x)u^2)\right]}{r}\right\}\hat\rho(x,(x^2+x)u^2)\,, \tag{225}$$

where S_r is the surface of a unit sphere in r dimensions and where $\hat\rho$ is the posterior probability that a vector is equal to $\vec{0}$.

$$\hat\rho(a,b^2) = \frac{\rho}{(1-\rho)\exp\left(\frac{-b^2}{2(1+a)}\right)(1+a)^{\frac{r}{2}}+\rho}\,. \tag{226}$$

An expansion of (224) around $m^t = 0$ yields

$$m^{t+1} = \frac{\rho^2 m^t}{\Delta} - \rho^3\left(\frac{m^t}{\Delta}\right)^2 + O\left((m^t)^3\right)\,. \tag{227}$$

Analogously to the conclusions we reached when studying expansion (205), we conclude that since the second term is negative there is always a 2nd order bifurcation at $\Delta_c = \rho^2$ with a stable fixed point for $\Delta < \Delta_c$ that stays close to the uniform fixed point. At the same time this close-by fixed point typically exists only in a very small interval of $(\Delta_{\rm Alg}, \Delta_c)$, similarly as in Fig. 9.

5. *Community detection with symmetric groups*

In this section we discuss the phase diagram of the symmetric communities detection model as defined in section I D 2 d. The corresponding prior distribution is (40) which leads to the function $f^x_{\rm in}$

$$\forall k \in [1:r], f^x_{\rm in}(A,B)_k = \frac{\exp\left(B_k - A_{kk}/2\right)}{\sum\limits_{k'=1\cdots r}\exp\left(B_{k'} - A_{k'k'}/2\right)}\,. \tag{228}$$

The corresponding output is given by (41-42), corresponding to the effective noise

$$\Delta = \frac{p_{\rm out}(1-p_{\rm out})}{\mu^2} = \frac{p_{\rm out}(1-p_{\rm out})}{N(p_{\rm in}-p_{\rm out})^2}\,. \tag{229}$$

Once again we study the phase diagram by analyzing the state evolution (151). We can verify that the following form of the order parameter in invariant under iterations of the Bayes-optimal state evolution

$$M^t = b^t\frac{I_r}{r} + \frac{(1-b^t)J}{r^2}\,, \tag{230}$$

where J is the matrix filled with 1. The order parameter at time $t+1$ will be of the same form with a new b^{t+1}.

- Having $b^t = 0$ is equivalent to having all the variables x_i saying that they have an equal probability to be in every community. This corresponds to initializing the estimators of the algorithm to be $\hat{x}_i^{t=0} = \left(\frac{1}{r},\cdots,\frac{1}{r}\right)$
- Having $b^t = 1$ means that the communities have been perfectly reconstructed. This corresponds to initializing the algorithm in the planted solution.

Using (151) the state evolution equations for b^t can be written.

$$b^{t+1} = \mathcal{M}_r\left(\frac{b^t}{\Delta}\right)\,, \tag{231}$$

where

$$\mathcal{M}_r(x) = \frac{1}{r-1}\left[r\int\frac{\exp\left(\frac{x}{r}+u_1\sqrt{\frac{x}{r}}\right)}{\exp\left(\frac{x}{r}+u_1\sqrt{\frac{x}{r}}\right)+\sum\limits_{i=2}^{r}\exp\left(u_i\sqrt{\frac{x}{r}}\right)}\prod_{i=1}^{r}{\rm d}u_i\frac{\exp\left(\frac{-u_i^2}{2}\right)}{\sqrt{2\pi}} - 1\right]\,. \tag{232}$$

This can be proven by computing M_{11}^{t+1} using (151) and (228) which yields

$$b^{t+1}\left(\frac{1}{r}-\frac{1}{r^2}\right)=\frac{1}{r}\left[\int \frac{\exp\left(\frac{x}{r}+u_1\sqrt{\frac{x}{r}}+u_0\sqrt{\frac{1-b^t}{r^2\Delta}}\right)}{\exp\left(\frac{x}{r}+u_1\sqrt{\frac{x}{r}}+u_0\sqrt{\frac{1-b^t}{r^2\Delta}}\right)+\sum_{i=2}^{r}\exp\left(u_i\sqrt{\frac{x}{r}}+u_0\sqrt{\frac{1-b^t}{r^2\Delta}}\right)}\prod_{i=0}^{r}\mathrm{d}u_i\frac{\exp\left(\frac{-u_i^2}{2}\right)}{\sqrt{2\pi}}\right]. \quad (233)$$

Here we have separated the noise W into two sources W_{I_r} and W_{J_r} (the sum of two independent Gaussian is still a Gaussian) of covariance matrices $\frac{b^t I_r}{r\Delta}$ and $\frac{(1-b^t)J_r}{r^2\Delta}$. The first term corresponds to $u_k, 1\le k\le n$ and the last term to u_0.

One observes that eq. (231) has always the uniform fixed point $b^t=0$. This is an example of a non-zero mean prior for which nevertheless there is a uniform fixed point because other kind of symmetry is present in the model. Let us expand (231) around $b^t=0$ to determine the stability of this fixed point, one gets

$$b^{t+1}=\frac{b^t}{\Delta r^2}+\frac{r-4}{2\Delta^2 r^4}{b^t}^2+O\left({b^t}^3\right). \quad (234)$$

The uniform fixed point hence becomes unstable for $\Delta>\Delta_c=\frac{1}{r^2}$. Translated back into the parameters of the stochastic block model this gives the easy/hard phase transition at $|p_{\rm in}-p_{\rm out}|=r\sqrt{p_{\rm out}(1-p_{\rm out})/N}$ well known in the sparse case where $p=\mathrm{const}/N$ from [11].

In terms of the type of phase transitions there are two cases

- 2nd order for $r\le 3$. The second term in (231) is negative, this means that there will be a fixed point close-by to the uniform one for $\Delta<\Delta_c$. The transition is of second order.
- 1st order for $r\ge 5$. The second term in (231) is positive, this means that there will be a jump in the order parameter at the transition $\Delta_c=\Delta_{\rm Alg}$. This is the signature a first order phase transition.

Rank $r=4$ is a marginal case in which we observed by directly solving the state evolution equations that the transition is continuous. We have checked numerically that no first order phase transition exists for the symmetric community detection problem for $r\in\{2,3,4\}$ meaning that first order phenomena appear only for $r\ge 5$.

In Fig. 11 left pannel, we illustrate the first order phase transition in the state evolution and in the behaviour of the Low-RAMP algorithms for $r=15$ groups.

To compute the values of $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$, we write $b^{t+1}=\mathcal{M}_r(b^t/\Delta)$ and carry a similar analysis as in section IV C 2, as detailed in appendix E. Fig. 11 (right pannel) summarizes the values in a scaling that anticipated the large rank expansion done is section V C 2.

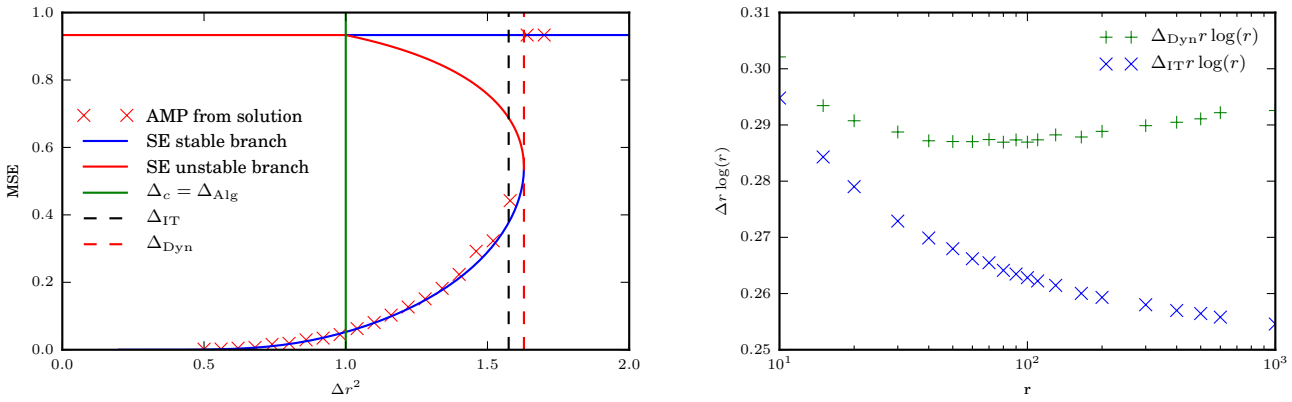

FIG. 11. Left: We plot MSE deduced from state evolution (lines) and from Low-RAMP algorithm (marks) for $r=15$ groups and $N=20000$ as a function of Δr^2. The vertical full green line is $\Delta_c r^2=1$. The vertical dashed black line is $\Delta_{\rm IT}$ and the full lines correspond to the MSE obtained from the informative initialization and have discontinuities at $\Delta_{\rm Dyn}$. Note that the MSE from does not go to zero at finite positive Δ instead at small noise one has MSE $\sim\exp(-\mathrm{const.}/\Delta)$. Right: We plot $\Delta r\log r$ for the information theoretic $\Delta_{\rm IT}$ and dynamical spinodal $\Delta_{\rm Dyn}$ phase transitions obtained from the state evolution using the protocol described in Appendix E. We rescale the Δ in this way to compare with the large rank expansion in (266) and (267).

B. Large sparsity (small ρ) expansions

In existing literature the sparse PCA was mostly studied for sparsity levels that are much smaller than a finite fraction of the system size (as considered in this paper). In order to compare with existing results we hence devote this section to the study of small density ρ expansions of the results obtained from the state evolution for some of the models studied.

1. *Spiked Bernoulli, Rademacher-Bernoulli, and Gauss-Bernoulli models*

For both the Rademacher-Bernoulli, and Gauss-Bernoulli models we have the uniform fixed point stable above $\Delta_c = \rho^2$, and in the leading order in $1/\rho$ we have $\Delta_{\rm alg} \sim_{\rho\to 0} \Delta_c = \rho^2$.

The small ρ limit behaviour of the information theoretic $\Delta_{\rm IT}$ threshold and the dynamical spinodal threshold $\Delta_{\rm Dyn}$ are given by

Bernoulli and Rademacher-Bernoulli

$$\Delta_{\rm Dyn}(\rho) \sim_{\rho\to 0} \frac{-\rho}{2\log(\rho)}\,, \tag{235}$$

$$\Delta_{\rm IT}(\rho) \sim_{\rho\to 0} \frac{-\rho}{4\log(\rho)}\,, \tag{236}$$

Gaussian-Bernoulli

$$\Delta_{\rm Dyn}(\rho) \sim_{\rho\to 0} \frac{-\rho}{\log(\rho)} \max\left\{ \frac{\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)}{\beta}, \beta\in\mathbb{R}^+ \right\} \sim 0.595 \frac{-\rho}{\log(\rho)}\,, \tag{237}$$

$$\Delta_{\rm IT}(\rho) \sim_{\rho\to 0} \frac{-\rho}{\log(\rho)} \max\left\{ \frac{\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)}{\beta}, \right.$$
$$\left. \int_0^\beta \mathrm{d}u\, \frac{2\exp\left(\frac{-1}{u}\right)}{\sqrt{\pi u}} + \left(\frac{1}{\sqrt{u}}\right) = \frac{1}{2}\beta\left[\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)\right]; \beta\in\mathbb{R}^+ \right\} \sim 0.528\frac{-\rho}{\log(\rho)}\,. \tag{238}$$

The information theoretic transitions for all these 3 models scale like $O\left(\frac{-\rho}{\log(\rho)}\right)$ while the algorithmic transition scales like $O(\rho^2)$. This means that for small ρ there is a large gap between what is information theoretically and algorithmically achievable. Note at this point that the bounds derived in [3] for sparse PCA have the same leading order behaviour when ρ is small as (235) and (236).

To derive the above small ρ expressions we combine (202) and (204) and the following small ρ limit of the state-evolution functions $f^{\rm SE}(x)$

$$\forall\beta\in\mathbb{R}^+, \lim_{\rho\to 0} \frac{f^{\rm SE}(-\beta\log(\rho))}{\rho} = 1(\beta>2), \textbf{ Bernoulli} \tag{239}$$

$$\forall\beta\in\mathbb{R}^+, \lim_{\rho\to 0} \frac{f^{\rm SE}(-\beta\log(\rho))}{\rho} = 1(\beta>2), \textbf{ Rademacher-Bernoulli} \tag{240}$$

$$\forall\beta\in\mathbb{R}^+, \lim_{\rho\to 0} \frac{f^{\rm SE}(-\beta\log(\rho))}{\rho} = \frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\beta\pi}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right), \textbf{ Gaussian-Bernoulli} \tag{241}$$

Here the functions $f^{\rm SE}$ are the state-evolution update functions stated in (212), (214) and (216) for the different models (when the model is clear from context we will omit the lower index specifying the model). The above is proven by deriving the state evolution equations for each of these models. The computation is done in appendix D.

2. Two balanced groups, limit of small planted subgraph

In this section we analyze the small ρ limit for the two balanced groups of section V A 3. From the definition of the function f_ρ (221), and a computation done in appendix D we get

$$\lim_{\rho\to 0} f_\rho(-\beta\rho(1-\rho)\log(\rho(1-\rho))) = 1(\beta > 2)\,. \tag{242}$$

By combining this with (202) and (204) one gets

$$\Delta_{\rm Dyn}(\rho) \sim_{\rho\to 0} \frac{1}{-2\rho(1-\rho)\log(\rho(1-\rho))}\,, \tag{243}$$

$$\Delta_{\rm IT}(\rho) \sim_{\rho\to 0} \frac{1}{-4\rho(1-\rho)\log(\rho(1-\rho))}\,. \tag{244}$$

The derivations of these limits is done in the appendix D.

Note that the limit of small ρ in the two balanced groups model is closely related to the problem of planted clique. However, in the planted clique problem the size of the clique to recover is much smaller that the size of the graph, typically $O(\sqrt{N})$ for efficient recovery and $O(\log N)$ for information theoretically possible recovery. Whereas our equation were derives when $\rho = O(1)$, we can try to see what would the above scaling imply for $k = \rho N = o(N)$. In the planted clique problem the first (smaller) group is fully connected, this means that the entry C_{11} of the connectivity matrix C (37) is equal to one. Therefore for $\rho \ll 1$ one has

$$\mu = (1 - p_{\rm out})\rho\sqrt{N}\,. \tag{245}$$

Note that in the canonical definition of the planted clique problem the average degree of the nodes belonging to the clique are slightly larger than the average degree of the rest of the graph. The present case of balanced groups corresponds to a version of the planted clique problem where next to planting a clique a corresponding number of edges is added to the rest of the graph to ensure that the average degree of every node is the same.

Recalling the definition of Δ (218) for the community detection output channel, and using $\Delta_c = 1$ for the spectral threshold, and $\Delta_{\rm IT} = -1/(4\rho\log\rho)$ for the information theoretic threshold at small ρ we get

$$k_c = \sqrt{N}\sqrt{\frac{p_{\rm out}}{1-p_{\rm out}}}\,, \tag{246}$$

$$k_{\rm IT} = \log(N)\frac{4p_{\rm out}}{1-p_{\rm out}}\,. \tag{247}$$

We indeed recover the scaling known from the planted clique problem, see e.g. [16]. The $p_{\rm out}$-dependent constant are indeed the tight constants (for efficient and information theoretically optimal recovery) for the balanced version of the planted clique where the expected degree of every node is the same independently of the fact in the node is in the clique or not.

3. Sparse PCA at small density ρ

We investigate here the small ρ limit of the bipartite $UV^\top$ spiked Gaussian-Bernoulli, and Rademacher-Bernoulli model. We remind that $U \in \mathbb{R}^N$, $V \in \mathbb{R}^M$, while $\alpha = M/N$. In the model we consider that elements of U are Gaussian of zero mean and unit variance, while P_V is given by (33) for the Rademacher-Bernoulli model, and by (36) for the Gauss-Bernoulli model. The state evolution equations then read

$$m_u^t = \frac{\alpha m_v^t}{\Delta + \alpha m_v^t}\,, \tag{248}$$

$$m_v^{t+1} = f^{\rm SE}_{\rm Gauss-Bernoulli}\left(\frac{m_u^t}{\Delta}\right)\,, \tag{249}$$

$$m_v^{t+1} = f^{\rm SE}_{\rm Rademacher-Bernoulli}\left(\frac{m_u^t}{\Delta}\right)\,, \tag{250}$$

where $f^{\rm SE}_{\rm Gauss-Bernoulli}$ and $f^{\rm SE}_{\rm Rademacher-Bernoulli}$ are defined in (214) and (216) respectively. By combining these equations one gets

$$m_v^{t+1} = f^{\rm SE}_{\rm Gauss-Bernoulli}\left(\frac{\alpha m_v^t}{\Delta^2+\alpha\Delta m_v^t}\right), \tag{251}$$

$$m_v^{t+1} = f^{\rm SE}_{\rm Rademacher-Bernoulli}\left(\frac{\alpha m_v^t}{\Delta^2+\alpha\Delta m_v^t}\right). \tag{252}$$

Because in both these cases P_U and P_V have zero mean the state evolution equations will have the uniform fixed point at $(m_u, m_v) = (0, 0)$. This fixed point becomes unstable when

$$\frac{\rho^2\alpha}{\Delta^2} > 1, \text{or}, \Delta < \Delta_c = \rho\sqrt{\alpha}\,. \tag{253}$$

Also in this case the stability transition Δ_c corresponds to the spectral transition where one sees informative eigenvalues get out of the bulk of the matrix S, as is known in the theory of low-rank perturbations of random matrices [2] (these methods are known not to take advantage of the sparsity). For ρ small enough there will be again a first order phase transitions with $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$, $\Delta_{\rm Dyn}$ defined as in section IV C 1. The asymptotic behaviour of these thresholds is

Rademacher-Bernoulli

$$\Delta_{\rm Dyn}(\rho) \underset{\rho\to 0}{\sim} \sqrt{\frac{-\rho\alpha}{2\log(\rho)}}\,, \tag{254}$$

$$\Delta_{\rm IT}(\rho) \underset{\rho\to 0}{\sim} \sqrt{\frac{-\rho\alpha}{4\alpha\log(\rho)}}\,, \tag{255}$$

Gaussian-Bernoulli

$$\Delta_{\rm Dyn}(\rho) \underset{\rho\to 0}{\sim} \sqrt{\frac{-\rho\alpha}{\log(\rho)}\max\left\{\frac{\operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)}{\beta}, \beta\in\mathbb{R}^+\right\}} \sim 0.771\sqrt{\frac{-\rho\alpha}{\log(\rho)}}\,, \tag{256}$$

$$\Delta_{\rm IT}(\rho) \underset{\rho\to 0}{\sim} \sqrt{\frac{-\rho\alpha}{\log(\rho)}}$$
$$\sqrt{\max\left\{\frac{\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}}+\operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)}{\beta}, \int_0^\beta \mathrm{d}u\,\frac{2\exp\left(\frac{-1}{u}\right)}{\sqrt{\pi u}} + \left(\frac{1}{\sqrt{u}}\right) = \frac{1}{2}\beta\left[\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}}+\operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)\right]; \beta\in\mathbb{R}^+\right\}}$$
$$\sim 0.726\sqrt{\frac{-\rho\alpha}{\log(\rho)}}\,. \tag{257}$$

Reminding that the value of $\Delta_{\rm Alg}$ scales in the same way as Δ_c (253), we see that a large hard phase opens as $\rho \to 0$.

To put the above results in relation to existing literature on sparse PCA [1, 15]. The main difference is that the regime considered in existing literature is that the number of non-zero element in the matrix V, $\rho N = o(1)$. The information theoretic threshold found in eq. (255) correspond (up to a constant) to information theoretic bounds found in [1]. However, the algorithmic performance that is found in the case of very small sparsity, by e.g. the covariance thresholding of [15], is not reproduced in our analysis of linear sparsity $\rho = O(1)$. This suggest that in case when the sparsity is small but linear in N, efficient algorithms that take advantage of the sparsity might not exist. This regime should be investigated further.

To derive the small ρ limit we follow a similar strategy as we did in the symmetric $XX^\top$ case. The main idea is to find the value of Δ for which $m_v = f^{\rm SE}_{P_X}(x)$ is a fixed point of the SE equations (252) or (251).

$$\Delta(x) = \frac{\alpha f^{\rm SE}(x) + \sqrt{\alpha^2 {f^{\rm SE}}^2(x) + 4\alpha\frac{f^{\rm SE}(x)}{x}}}{2}\,, \tag{258}$$

This means that

$$(m_u, m_v) = (x\Delta(x), f^{\rm SE}(x)) \tag{259}$$

is a fixed point of the state evolution equations for $\Delta = \Delta(x)$. The free-energy is computed as

$$\begin{aligned}\phi(m_u = x\Delta(x), m_v = f(x), \Delta = \Delta(x)) - \phi(m_u = 0, m_v = 0, \Delta = \Delta(x)) = \\ = \alpha \int_0^x \mathrm{d}u f^{\mathrm{SE}}(u) + \int_0^{\alpha f^{\mathrm{SE}}(x)/\Delta(x)} \mathrm{d}u \frac{u}{1+u} - x f^{\mathrm{SE}}(x)\,. \end{aligned} \tag{260}$$

Combining this with (241) and (240) one gets the above asymptotic behaviour.

C. Large rank expansions

Another limit that can be worked out analytically is the large rank r limit. This section summarizes the results.

1. *Large-rank limit for jointly-sparse PCA*

We analyze the large r limit of jointly-sparse PCA for which the state evolution equation is given by (224). We notice that u^2 will have mean r and standard deviation $\sqrt{2r}$. This essentially means that to get the large r limit of the density evolution equations one need to replace u^2 by r everywhere it appears. Expanding in large r then gives

$$m^{t+1} = \frac{\rho m^t}{m^t + \Delta}\left(1 + \frac{m^t}{\Delta}(1-\hat{\rho})\right)\hat{\rho}\,, \tag{261}$$

where

$$\hat{\rho} = \lim_{r\to+\infty} \frac{\rho}{(1-\rho)\exp\left(\frac{r}{2}\left[\frac{m^t}{\Delta} + \log(1+\frac{m^t}{\Delta})\right] + \rho\right.} = \begin{cases} 1 \text{ if } r{m^t}^2 \to 0 \\ \rho \text{ if } r{m^t}^2 \to +\infty \end{cases}\,. \tag{262}$$

This means that for any $m^t \gg \frac{1}{\sqrt{r}}$ the update equations will be approximately

$$m^{t+1} = \frac{m^t \rho}{m^t + \Delta} + o(1)\,. \tag{263}$$

This update equation can be easily analyzed, it only has one stable fixed point located at

$$\max(\rho - \Delta, 0)\,. \tag{264}$$

Analogous expansion of the replicated free energy leads to the result that in the large rank limit we have $\Delta_{\mathrm{Dyn}} = \Delta_{\mathrm{IT}} = \rho$ whereas $\Delta_{\mathrm{Alg}} = \Delta_c = \rho^2$. This is plotted in the large rank phase diagram (12).

2. *Community detection*

The large rank limit is analyzed for the problem of symmetric community detection in appendix E. The asymptotic behavior of $\Delta_{\mathrm{IT}}(r)$ and $\Delta_{\mathrm{Dyn}}(r)$ as $r \to +\infty$ are

$$\Delta_c = \frac{1}{r^2}\,, \tag{265}$$

$$\Delta_{\mathrm{Dyn}} = \frac{1}{2r\ln(r)}[1 + o_r(1)]\,, \tag{266}$$

$$\Delta_{\mathrm{IT}} = \frac{1}{4r\ln(r)}[1 + o_r(1)]\,. \tag{267}$$

We see that a large gap opens between Δ_c and Δ_{IT} as r grows. The behavior Δ_{IT} and Δ_{Dyn} for moderately large r is illustrated in figure 11 and we see that the above limit is reached very slowly. Using eq. (229) this translates into the large r limit phase transition in terms of parameters of the stochastic block model as discussed in [40], and proven in [3].

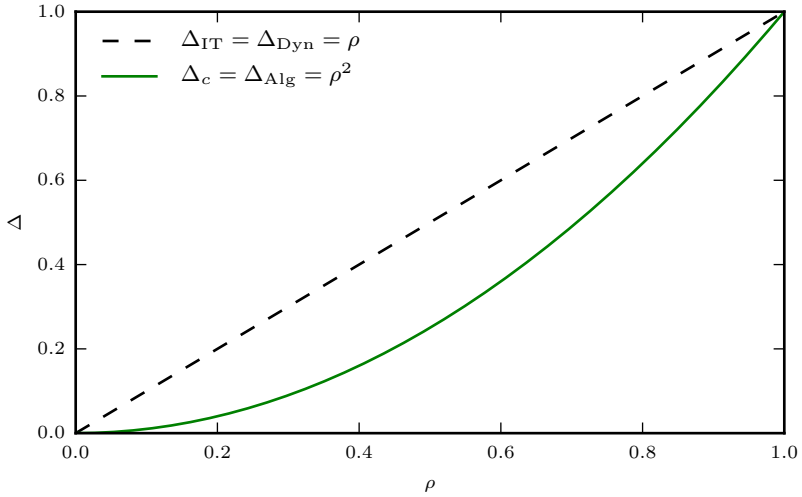

FIG. 12. Phase diagram of the jointly-sparse PCA model at large rank r. In the large r limit the algorithmic spinodal merges with Δ_c, the dynamical spinodal and the information theoretic one converge toward the line $\Delta = \rho$.

VI. DISCUSSION AND PERSPECTIVE

The main results of this paper are summarised in section I E. This concluding section is devoted to open directions and questions.

A large part of results reported in this paper concerns the Bayes-optimal setting for which the mutual information and MMSE was recently established rigorously for the rank one symmetric $XX^\top$ case [4, 44]. The proof for generic rank r and for the bipartite $UV^\top$ case seems well within the reach of the existing methods as reported by the authors of [44].

A formidable open question concerns the nature of the hard phase, that is the region of parameters for which Low-RAMP does not reach the MMSE. Such regions have been identified in a considerable number of inference problems and it remains a great challenge to study whether efficient algorithms that perform better than Low-RAMP exist or to provide further evidence towards their non-existence. One notable direction is that whereas the MMSE in Bayes-optimal setting is correctly described by the replica symmetric solution, the nature of the metastable branch describing the behaviour of Low-RAMP in the hard phase may involve effect of replica symmetry breaking, this should be investigated in future work.

The results presented in this paper focused on the Bayes-optimal inference, the theory and methodology presented is not limited to this case and phase diagrams involving mismatch between the teacher and student models or parameters could provide further insight into issues such as model and parameter selection and learning, overfitting, or performance of minimisation-based algorithms. Effects of replica symmetry breaking are expected to play a role in this case. This is another interesting direction for future work.

In the present paper we focused on pair-wise interaction corresponding to matrix factorizations. In physics as well as in statistics a generalization to p-spin interaction have been studied. In statistics this type of problems is termed tensor PCA [60]. Detailed study of tensor PCA is also an interesting extension.

In the present paper we treated real-valued vectors spins. The generalization to spin-values that live on more involved spaces such as compact groups was recently considered in [54] and many interesting questions in this direction remain open.

Finally in the present work we concentrated on the theoretical development and phase diagrams for randomly generated data. The Low-RAMP algorithm (that is distributed in Matlab and Julia at http://krzakala.github.io/LowRAMP/) can be successfully applied to real-world data. In that case the set of theoretical guarantees on its performance is limited, but this is the case common to many other practically used approaches. Investigating performance of the Low-RAMP algorithm on data coming from applications of interest is another formidably rich and important direction for future work.

VII. ACKNOWLEDGEMENT

FK acknowledges funding from the EU (FP/2007-2013/ERC grant agreement 307087-SPARCS). LZ acknowledges funding for the project SaMURai from LabEx PALM (Ref : ANR-10-LABX-0039-PALM). We would like to thank Marc Mézard, and Cristopher Moore for insightful discussions related to this work, Christian Schmidt for discussions pointing errors, and Alexis Bozio for proofreading of the manuscript.

[1] Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *IEEE International Symposium on Information Theory*, pages 2454–2458. IEEE, 2008.
[2] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005.
[3] Jess Banks, Cristopher Moore, Roman Vershynin, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *arXiv preprint arXiv:1607.05222*, 2016.
[4] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborov'a. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances In Neural Information Processing Systems*, pages 424–432, 2016.
[5] N Barkai and Haim Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. *Physical Review E*, 50(3):1766, 1994.
[6] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.
[7] Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse PCA. *arXiv preprint arXiv:1304.0828*, 2013.
[8] Michael Biehl and Andreas Mietzner. Statistical mechanics of unsupervised structure recognition. *Journal of Physics A: Mathematical and General*, 27(6):1885, 1994.
[9] Francesco Caltagirone, Lenka Zdeborová, and Florent Krzakala. On convergence of approximate message passing. In *IEEE International Symposium on Information Theory*, pages 1812–1816. IEEE, 2014.
[10] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.
[11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.
[12] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, 2011.
[13] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the binary stochastic block model. In *IEEE International Symposium on Information Theory (ISIT)*, pages 185–189. IEEE, 2016.
[14] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2197–2201. IEEE, 2014.
[15] Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–342, 2014.
[16] Yash Deshpande and Andrea Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, pages 1–60, 2015.
[17] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914–18919, 2009.
[18] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
[19] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
[20] Marylou Gabrié, Eric W Tramel, and Florent Krzakala. Training restricted Boltzmann machine via the Thouless-Anderson-Palmer free energy. In *Advances in Neural Information Processing Systems*, pages 640–648, 2015.
[21] A Georges and J S Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9):2173, 1991.
[22] D. J. Gross, I. Kanter, and H. Sompolinsky. Mean-field theory of the potts glass. *Phys. Rev. Lett.*, 55(3):304–307, Jul 1985.
[23] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.
[24] Geoffrey Hinton. A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1):926, 2010.
[25] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.
[26] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
[27] David C Hoyle and Magnus Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E*, 69(2):026124, 2004.

[28] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference*, page iat004, 2013.

[29] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *Unpublished manuscript*, 7, 2004.

[30] Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265, 2016.

[31] Hilbert J Kappen and FB Rodriguez. Boltzmann machine learning using mean field theory and linear response correction. *Advances in neural information processing systems*, pages 280–286, 1998.

[32] JM Kosterlitz, DJ Thouless, and Raymund C Jones. Spherical model of a spin-glass. *Physical Review Letters*, 36(20):1217, 1976.

[33] Robert Krauthgamer, Boaz Nadler, Dan Vilenchik, et al. Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 2015.

[34] Florent Krzakala, Andre Manoel, Eric W Tramel, and Lenka Zdeborová. Variational free energies for compressed sensing. In *IEEE International Symposium on Information Theory*, pages 1499–1503. IEEE, 2014.

[35] Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Phase diagram and approximate message passing for blind calibration and dictionary learning. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 659–663. IEEE, 2013.

[36] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[37] Florent Krzakala, Jiaming Xu, and Lenka Zdeborová. Mutual information in rank-one matrix estimation. *to appear in ITW 2016, preprint arXiv:1603.08447*, 2016.

[38] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[39] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering. *to appear in Allerton 2016, preprint arXiv:1610.02918*, 2016.

[40] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 680–687. IEEE, 2015.

[41] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse PCA. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1635–1639, 2015.

[42] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[43] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45, 2004.

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

[45] Ryosuke Matsushita and Toshiyuki Tanaka. Low-rank matrix reconstruction and clustering via approximate message passing. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 917–925. Curran Associates, Inc., 2013.

[46] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin-Glass Theory and Beyond*, volume 9 of *Lecture Notes in Physics*. World Scientific, Singapore, 1987.

[47] Marc Mézard. Mean-field message-passing equations in the Hopfield model and its generalizations. *arXiv preprint arXiv:1608.01558*, 2016.

[48] Rémi Monasson and Dario Villamaina. Estimating the principal components of correlation matrices from all their empirical eigenvectors. *EPL (Europhysics Letters)*, 112(5):50001, 2015.

[49] Andrea Montanari. Finding one community in a sparse graph. *Journal of Statistical Physics*, 161(2):273–299, 2015.

[50] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, Oxford, UK, 2001.

[51] Hidetoshi Nishimori and David Sherrington. Absence of replica symmetry breaking in a region of the phase diagram of the ising spin glass. In *American Institute of Physics Conference Series*, volume 553, pages 67–72, 2001.

[52] Jason T Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing part I: Derivation. *IEEE Transactions on Signal Processing*, 62(22):5839–5853, 2014.

[53] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

[54] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Message-passing algorithms for synchronization problems over compact groups. *arXiv preprint arXiv:1610.04583*, 2016.

[55] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of PCA for spiked random matrices and synchronization. *arXiv preprint arXiv:1609.05573*, 2016.

[56] T Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971, 1982.

[57] Sundeep Rangan. Estimation with random linear mixing, belief propagation and compressed sensing. In *44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2010.

[58] Sundeep Rangan and Alyson K Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1246–1250. IEEE, 2012.

[59] Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson Fletcher, and Volkan Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. In *IEEE International Symposium on Information Theory Proceedings*

(ISIT), pages 664–668. IEEE, 2013.
[60] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.
[61] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35:1792–1796, 1975.
[62] Hans-Juergen Sommers. Theory of a Heisenberg spin glass. *Journal of Magnetism and Magnetic Materials*, 22(3):267–270, 1981.
[63] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.
[64] Eric W Tramel, Andre Manoel, Francesco Caltagirone, Marylou Gabrié, and Florent Krzakala. Inferring sparsity: Compressed sensing using generalized restricted Boltzmann machines. In *IEEE Information Theory Workshop (ITW)*, pages 265–269. IEEE, 2016.
[65] Jérôme Tubiana and Rémi Monasson. Emergence of compositional representations in restricted Boltzmann machines. *arXiv preprint arXiv:1611.06759*, 2016.
[66] Jeremy Vila, Philip Schniter, Sundeep Rangan, Florent Krzakala, and Lenka Zdeborová. Adaptive damping and mean removal for the generalized approximate message passing algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2021–2025. IEEE, 2015.
[67] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.
[68] TLH Watkin and J-P Nadal. Optimal unsupervised learning. *Journal of Physics A: Mathematical and General*, 27(6):1899, 1994.
[69] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Exploring artificial intelligence in the new millennium. chapter Understanding Belief Propagation and Its Generalizations, pages 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.
[70] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65:453–552, 2016.
[71] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Appendices

A. MEAN FIELD EQUATIONS

In order to compare the Low-RAMP algorithm with the commonly used variational mean field inference we write here the variational mean field equations in the same notation we used for Low-RAMP. We also write the mean field free energy.

For the symmetric vector-spin glass the naive mean field equations read

$$B^t_{X,i} = \sum_{k=1}^{N} \frac{1}{\sqrt{N}} S_{ki} \hat{x}^t_k \,, \tag{268}$$

$$A^t_{X,i} = \frac{1}{N} \sum_{k=1}^{N} (S^2_{ki} - R_{ki}) \left(\hat{x}^t_k \hat{x}^{t,\top}_k + \sigma^t_{x,k} \right) \,, \tag{269}$$

$$\hat{x}^{t+1}_i = f^x_{\rm in}(A^t_{X,i}, B^t_{X,i}) \,, \tag{270}$$

$$\sigma^{t+1}_{x,i} = \frac{\partial f^x_{\rm in}}{\partial B}(A^t_{X,i}, B^t_{X,i}) \,. \tag{271}$$

The mean field free energy for the symmetric $XX^\top$ case reads

$$F^{\rm MF}_{XX^\top}(\{A_{X,i}\},\{B_{X,i}\}) = \sum_{1\le i\le N} \log(\mathcal{Z}_x(A_{X,i},B_{X,i})) - B^\top_{X,i}\hat{x}_i + \frac{1}{2}{\rm Tr}\left[A_{X,i}(\hat{x}_i\hat{x}_i^\top + \sigma_{x,i})\right]$$
$$+ \frac{1}{2} \sum_{1\le i,j\le N} \left[\frac{1}{\sqrt{N}} S_{ij} \hat{x}_i^\top \hat{x}_j + \frac{(R_{ij} - S^2_{ij})}{2N} {\rm Tr}\left[(\hat{x}_i\hat{x}_i^\top + \sigma_{x,i})(\hat{x}_j\hat{x}_j^\top + \sigma_{x,j})\right] \right] . \tag{272}$$

For the bipartite $IUV^\top$ case the mean field equations read

$$B^t_{U,i} = \frac{1}{\sqrt{N}} \sum_{l=1}^{M} S_{il} \hat{v}^t_l \,, \tag{273}$$

$$A^t_U = \frac{1}{N} \sum_{l=1}^{M} \left(S^2_{il} - R_{il}\right) \left(\hat{v}^t_l \hat{v}^{t,\top}_l + \sigma^t_{u,l}\right) \,, \tag{274}$$

$$\hat{u}^t_i = f^u_{\rm in}(A^t_U, B^t_{U,i}) \,, \tag{275}$$

$$\sigma^t_{u,i} = \left(\frac{\partial f^u_{\rm in}}{\partial B}\right)(A^t_U, B^t_{U,i}) \,, \tag{276}$$

$$B^t_{V,j} = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} S_{kj} \hat{u}^t_k \,, \tag{277}$$

$$A^t_{V,j} = \frac{1}{N} \sum_{k=1}^{N} (S^2_{kj} - R_{kj}) \left(\hat{u}^t_k \hat{u}^{t,\top}_k + \sigma^t_{u,k}\right) \,, \tag{278}$$

$$\hat{v}^{t+1}_j = f^v_{\rm in}(A^t_{V,j}, B^t_{V,j}) \,, \tag{279}$$

$$\sigma^{t+1}_{v,j} = \left(\frac{\partial f^v_{\rm in}}{\partial B}\right)(A^t_{V,j}, B^t_{V,j}) \,. \tag{280}$$

The mean field free energy for the bipartite case reads

$$\begin{aligned} F^{\rm MF}_{UV^\top}(\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}) = & \sum_{1\le i\le N} \log(\mathcal{Z}_u(A_{U,i}, B_{U,i})) - {B_{U,i}}^\top \hat{u}_i + \frac{1}{2}{\rm Tr}\left[A_{U,i}(\hat{u}_i\hat{u}_i^\top + \sigma_{u,i})\right] \\ & + \sum_{1\le j\le M} \log(\mathcal{Z}_v(A_{V,j}, B_{V,j})) - {B_{V,j}}^\top \hat{v}_j + \frac{1}{2}{\rm Tr}\left[A_{V,j}(\hat{v}_j\hat{v}_j^\top + \sigma_{v,j})\right] \\ & + \sum_{1\le i\le N, 1\le j\le M} \left[\frac{1}{\sqrt{N}} S_{ij} \hat{u}_i^\top \hat{v}_j + \frac{1}{2N}(R_{ij} - S^2_{ij}){\rm Tr}\left[(\hat{u}_i\hat{u}_i^\top + \sigma_{u,i})(\hat{v}_j\hat{v}_j^\top + \sigma_{v,j})\right]\right] . \end{aligned} \tag{281}$$

The difference between the mean field equations and the Low-RAMP equations from section II B and II D can be seen for both variables A and B.

B. BETHE FREE ENERGY DERIVED BY THE PLEFKA EXPANSION

We present here a way to compute the Bethe free energy for relatively generic class of Hamiltonians. The main hypothesis for this computation to be exact in the large N limit is that the Fisher score matrix S_{ij} can be well approximated as having elements with weak correlation of order $\frac{1}{N}$ one from another. If this is not true then the following equations do not apply.

As it turns out the mean-field free energy from Appendix A is just the first order term in the high temperature expansion of the free-energy. as was derived in [21, 56] first by Plefka and then by Georges and Yedidia. The Bethe free energy is then the 2nd order expansion. Suppose one is given a general system with with N classical variables x_i that weakly interact. The Hamiltonian H defined the free energy

$$\Phi = \log\left({\rm Tr}\left[\exp(\beta H(x_1, \cdots, x_N))\right]\right) \,. \tag{282}$$

We define a new Hamiltonian that fixes the marginal by

$$\beta H_{\rm Field}(\{\lambda_i\}) = \beta H + \sum_{1\le i\le N} \lambda_i(x_i, \beta) \,, \tag{283}$$

$$\Phi_{\rm Field}(\beta, \{\lambda_i\}) = \log\left({\rm Tr}\left[\exp(\beta H_{\rm Field}(\{\lambda_i\}))\right]\right) \,. \tag{284}$$

By taking the Legendre transform one gets

$$\Phi_{\text{Legendre}}(\beta,\{\mu_i(x_i)\}) = \min\left\{\Phi_{\text{Field}}(\beta,\{\lambda_i\}) - \sum_{1\leq i\leq n}\int \mathrm{d}x\mu_i(x)\lambda_i(x)\right\}, \tag{285}$$

$$\{\Lambda_i(x_i)\} = \operatorname{argmin}\left\{\Phi_{\text{Field}}(\beta,\{\lambda_i\}) - \sum_{1\leq i\leq n}\int \mathrm{d}x\mu_i(x)\lambda_i(x)\right\}, \tag{286}$$

here the minimization is done over the fields $\lambda_i(x_i)$. The $\mu_i(x_i)$ are marginal density probabilities that one aims to impose on the system. The $\Lambda_i(x_i)$ are the fields one uses to fix the marginals equal to $\mu_i(x_i)$. The $\Lambda_i(x_i)$ depend on the problem H, β and the $\mu_i(x_i)$. Because the $\Lambda_i(x_i)$ are defined up to a constant

$$\int \mathrm{d}x\mu_i(x)\Lambda_i(x) = 0 \tag{287}$$

is imposed. According to the definition of the Legendre transformation on has

$$\Phi = \max\left\{\Phi_{\text{Legendre}}(\beta,\{\mu_i(x_i)\})\right\}. \tag{288}$$

To compute $\Phi_{\text{Legendre}}(\beta,\{\mu_i(x_i)$ we resort to a high temperature expansion. This procedure is called the Plefka expansion [56], it relies on the following high temperature expansion (we stop at order 2)

$$\Phi_{\text{Bethe}} = \Phi_{\text{Frustrated}}(\beta=0) + \beta\left(\frac{\partial\Phi_{\text{Legendre}}}{\partial\beta}\right)(\beta=0) + \frac{\beta^2}{2}\left(\frac{\partial^2\Phi_{\text{Legendre}}}{\partial\beta^2}\right)(\beta=0). \tag{289}$$

The expansion was explained in detail in [21], here we remind the main steps. Let us introduce the following operator

$$U = H - \langle H\rangle - \sum_{1\leq i\leq N}\frac{\partial\Lambda_i(x_i)}{\partial\beta}, \tag{290}$$

where the average is taken with respect to probability distribution induced by (283). One can show that for all observables O one has

$$\forall\beta, \frac{\partial\langle O\rangle}{\partial\beta} = \left\langle\frac{\partial O}{\partial\beta}\right\rangle - \langle OU\rangle. \tag{291}$$

According to the definition of $F_{\text{Frustrated}}$ one can prove that.

$$\forall\beta, \frac{\partial\Phi_{\text{Legendre}}}{\partial\beta} = \langle H\rangle. \tag{292}$$

Using (292) and (291) we get

$$\forall\beta, \frac{\partial^2\Phi_{\text{Legendre}}}{\partial\beta^2} = \frac{\langle H\rangle}{\partial\beta} = -\langle HU\rangle. \tag{293}$$

Therefore

$$\Phi_{\text{Bethe}} = \Phi_{\text{Legendre}}(\beta=0) + \beta\langle H\rangle(\beta=0) - \frac{\beta^2}{2}\langle HU\rangle(\beta=0). \tag{294}$$

We still need to compute $\frac{\partial\Lambda_i(x_i)}{\partial\beta}$ at $\beta=0$. This can be done by computing the derivative of the marginals with respect to β and noticing that they have to be zero.

$$\left(\frac{\partial\langle\delta(x_i-\hat{x}_i)\rangle}{\partial\beta}\right) = \left(\frac{\partial\mu_i(\hat{x}_i)}{\partial\beta}\right) = 0 = \langle U\delta(x_i-\hat{x}_i)\rangle = \left\langle\left(H - < H > - \sum_{1\leq i\leq N}\frac{\partial\Lambda_i(x_i)}{\partial\beta} + \int \mathrm{d}\hat{x}_i\mu_i(\hat{x}_i)\frac{\partial\Lambda_i(\hat{x}_i)}{\partial\beta}\right)\delta(x_i-\hat{x}_i)\right\rangle. \tag{295}$$

From this one deduces

$$\left\langle\frac{\partial\Lambda_i(x_i)}{\partial\beta}\delta(x_i-\hat{x}_i) + C(i,\beta)\delta(x_i-\hat{x}_i)s\right\rangle = \langle H\delta(x_i-\hat{x}_i) - \langle H\rangle\delta(x_i-\hat{x}_i)\rangle, \tag{296}$$

where $C(i,\beta)$ is

$$C(i,\beta) = \int \mathrm{d}\hat{x}_i \mu_i(\hat{x}_i) \frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \,. \tag{297}$$

By definition

$$\langle \delta(x_i - \hat{x}_i) \rangle = \mu_i(\hat{x}_i) \,, \tag{298}$$
$$\langle H \delta(x_i - \hat{x}_i) \rangle = \mu_i(x_i) \langle H \rangle_{x_i = \hat{x}_i} \,, \tag{299}$$

where $\langle H \rangle_{x_i=\hat{x}_i}$ is the average energy conditioned on the fact that $x_i = \hat{x}_i$. Using (296) we get

$$\mu_i(\hat{x_i}) \left(\frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \right) = \mu_i(\hat{x_i}) \langle H \rangle_{x_i=\hat{x}_i} - \mu_i(\hat{x_i}) \langle H \rangle + \mu_i(\hat{x_i}) C(i,\beta) \,, \tag{300}$$

$$\left(\frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \right) = \langle H \rangle_{x_i=\hat{x}_i} - \langle H \rangle + C(i,\beta) \,. \tag{301}$$

We can see from (283) that $\Lambda_i(\hat{x}_i, \beta)$ is defined up to a constant we fix that constant by having

$$\int \mathrm{d}\hat{x}_i \mu_i(\hat{x}_i) \Lambda_i(\hat{x}_i) = 0 \,. \tag{302}$$

Therefore we get

$$\forall \beta, \left(\frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \right) = \langle H \rangle_{x_i=\hat{x}_i} - \langle H \rangle \ = \frac{1}{\mu_i(\hat{x}_i)} \langle H \delta(x_i - \hat{x}_i) \rangle - \langle H \rangle \,, \tag{303}$$

where once again $\langle H \rangle_{x_i=\hat{x}_i}$ is the average of the energy where we have conditioned on the event $x_i = \hat{x_i}$. Since we do all expansion around $\beta = 0$ one is able to compute all the means present in this formula since at $\beta = 0$ the density probability of the system at $\beta = 0$ is just

$$P_{\text{Factorised}} = \prod_{i=1\cdots N} \mu_i(x_i) \,. \tag{304}$$

By using (289) and (303) around $\beta = 0$ one gets

$$\Phi_{\text{Bethe}} = \sum_{1 \le i \le N} S_{\mu_i} + \beta \langle H \rangle + \frac{\beta^2}{2} \left[\langle H^2 \rangle - \langle H \rangle^2 - \sum_{1 \le i \le N} \int \mathrm{d}x_i \mu_i(\hat{x}_i) (\langle H \rangle_{x_i=\hat{x}_i} - \langle H \rangle)^2 \right] , \tag{305}$$

where here S_{μ_i} is here the entropy of the density probability $\mu_i(x_i)$. In this setting all terms of order 3 and beyond will give a zero contribution in the large N limit. This is due to the fact that the matrix Y has components largely uncorrelated since they were taken independently from one another (the signal induces correlation that are too small to be significant). One reason why this approach might fail is that even though order higher than 3 might disappears in the large N limit their sum might be non convergent. This is the sign of a replica symmetry breaking in the system.

From this expression we deduce the Bethe free energy for the $XX^\top$ case as a function of the variables $A_{X,i}$ and $B_{X,i}$. The marginals that will be enforced are

$$P_{X,\text{Marginal}}(x_i) = \frac{1}{\mathcal{Z}_x(A_{X,i}, B_{X,i})} P_X(x_i) \exp\left(B_{X,i}^\top x_i - \frac{x_i^\top A_{X,i} x_i}{2} \right) , \tag{306}$$

$$P_{U,\text{Marginal}}(u_i) = \frac{1}{\mathcal{Z}_u(A_{U,i}, B_{U,i})} P_U(u_i) \exp\left(B_{U,i}^\top u_i - \frac{u_i^\top A_{U,i} u_i}{2} \right) , \tag{307}$$

$$P_{V,\text{Marginal}}(v_j) = \frac{1}{\mathcal{Z}_v(A_{V,j}, B_{V,j})} P_X(v_j) \exp\left(B_{V,i}^\top v_j - \frac{v_j^\top A_{V,j} v_j}{2} \right) . \tag{308}$$

The Hamiltonians are

$$H_{XX^\top} = \sum_{i=1\cdots N} \log(P_X(x_i)) + \sum_{1 \le i < j \le N} \left[g\left(Y_{ij}, \frac{x_i^\top x_j}{\sqrt{N}} \right) - g(Y_{ij}, 0) \right] , \tag{309}$$

and

$$H_{UV^\top} = \sum_{i=1\cdots N} \log(P_U(u_i)) + \sum_{j=1\cdots M} \log(P_V(v_j)) + \sum_{1\leq i\leq N, 1\leq j\leq M} \left[g\left(Y_{ij}, \frac{u_i^\top v_j}{\sqrt{N}}\right) - g(Y_{ij}, 0)\right]. \tag{310}$$

Using (306,307,308) and (309,310) in (305) one gets for the symmetric $XX^\top$ case

$$\Phi_{XX^\top}(\{A_{X,i}\},\{B_{X,i}\}) = \sum_{1\leq i\leq N}\left[\log(\mathcal{Z}_x(A_{X,i},B_{X,i})) - B_{X,i}^\top \hat{x}_i + \frac{1}{2}\mathrm{Tr}\left[A_{X,i}(\hat{x}_i\hat{x}_i^\top + \sigma_{x,i})\right]\right]$$
$$+\frac{1}{2}\sum_{1\leq i,j\leq N}\left[\frac{1}{\sqrt{N}}S_{ij}\hat{x}_i^\top\hat{x}_j + \frac{R_{ij}}{2N}\mathrm{Tr}\left[(\hat{x}_i\hat{x}_i^\top+\sigma_{x,i})(\hat{x}_j\hat{x}_j^\top+\sigma_{x,j})\right] - \frac{S_{ij}^2}{2N}\mathrm{Tr}\left[\hat{x}_i\hat{x}_i^\top\hat{x}_j\hat{x}_j^\top\right] - \frac{1}{N}S_{ij}^2\mathrm{Tr}\left[\sigma_{x,i}\sigma_{x,j}\right]\right]. \tag{311}$$

For the $UV^\top$ case

$$\Phi_{\mathrm{Bethe},UV^\top}(\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}) = \sum_{1\leq i\leq N}\left[\log(\mathcal{Z}_u(A_{U,i},B_{U,i})) - {B_{U,i}}^\top\hat{u}_i + \frac{1}{2}\mathrm{Tr}\left[A_{U,i}(\hat{u}_i\hat{u}_i^\top+\sigma_{u,i})\right]\right]$$
$$+\sum_{1\leq j\leq M}\left[\log(\mathcal{Z}_v(A_{V,j},B_{V,j})) - {B_{V,j}}^\top\hat{v}_j + \frac{1}{2}\mathrm{Tr}\left[A_{V,j}(\hat{v}_j\hat{v}_j^\top+\sigma_{v,j})\right]\right]$$
$$+\sum_{1\leq i\leq N,1\leq j\leq M}\left[\frac{1}{\sqrt{N}}S_{ij}\hat{u}_i^\top\hat{v}_j + \frac{1}{2N}R_{ij}\mathrm{Tr}\left[(\hat{u}_i\hat{u}_i^\top+\sigma_{u,i})(\hat{v}_j\hat{v}_j^\top+\sigma_{v,j})\right] - \frac{S_{ij}^2\mathrm{Tr}\left(\hat{u}_i\hat{u}_i^\top\hat{v}_j\hat{v}_j^\top\right)}{2N} - \frac{1}{N}S_{ij}^2\mathrm{Tr}\left[\sigma_{u,i}\sigma_{v,j}\right]\right], \tag{312}$$

where the $\hat{x}_i, \hat{u}_i, \hat{v}_j, \sigma_{x,i}, \sigma_{u,i}, \sigma_{v,j}$ are the mean of the marginals we want to enforce. These are all function of variables B_i, and A_i. One should try and minimize these function. This is what Low-RAMP does. Fixed points of the Low-RAMP algorithm are stationary points of the of the Bethe free energy.

C. REPLICA COMPUTATION $UV^\top$ CASE.

In this appendix we present the derivation replica free-energy in the case of the $UV^\top$ case. In the coming computation the indices i and k will go from 1 to N. And j and l will go from 1 to M.

$$\mathcal{Z}(\{Y_{ij}\}) = \int \prod_i \mathrm{d}u_i P_U(u_i) \prod_j \mathrm{d}v_j P_U(v_j) \prod_{ij} \exp\left(g\left(Y_{ij}, \frac{u_i^\top v_j}{\sqrt{N}}\right) - g\left(Y_{ij}, 0\right)\right). \tag{313}$$

One would like to compute the average of $\langle\log(\mathcal{Z}(\{Y_{ij}\}))\rangle$. This can be computed using the replica trick

$$\langle\log(\mathcal{Z}(\{Y_{ij}\}))\rangle_Y = \lim_{n\to 0}\frac{\langle\mathcal{Z}^n\rangle - 1}{n} = \lim_{n\to 0}\frac{\log(\langle\mathcal{Z}^n\rangle)}{n}. \tag{314}$$

These can be hopefully be computed for any $n\in\mathbb{N}$. We will then compute this function as $n\to 0$. We start with evaluating

$$\mathcal{Z}^n(\{Y_{ij}\}) = \int \prod_{a=1\cdots n}\prod_{i=1\cdots N}\mathrm{d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M}\mathrm{d}v_j^a P_V^a(v_j^a) \exp \prod_{i=1\cdots N, j=1\cdots M} \exp\left(g\left(Y_{ij}, \frac{u_i v_j}{\sqrt{N}}\right) - g\left(Y_{ij}, 0\right)\right). \tag{315}$$

Therefore one has

$$\mathbb{E}(\mathcal{Z}^n) = \int \prod_{i=1\cdots N, j=1\cdots M}\mathrm{d}Y_{ij} P_{\mathrm{out}}\left(Y_{ij}, w=0\right) \prod_{a=0\cdots n}\left(\left[\prod_{i=1\cdots N}\mathrm{d}u_i^a P_U^a(u_i^a)\prod_{j=1\cdots M}\mathrm{d}v_j^a P_V^a(v_j^a)\right]\right.$$
$$\left.\left[\prod_{i=1\cdots N, j=1\cdots M}\exp\left(\sum_{a=0\cdots n} g^a\left(Y_{ij}, \frac{u_i^a v_j^a}{\sqrt{N}}\right) - g^a\left(Y_{ij}, 0\right)\right)\right]\right), \tag{316}$$

where

- if $a=0$ then $g^a = g^0 = \log(P_{\rm out}(Y,w))$, $P_U^a(u) = P_{U_0}(u)$ and $P_V^a(v) = P_{V_0}(v)$
- if $a\neq 0$ then $g^a = g$, $P_U^a(u) = P_U(u)$ and $P_V^a(v) = P_V(v)$

We expand the function g^a to order 2 and get

$$\mathbb{E}(\mathcal{Z}^n) = \int \prod_{i=1\cdots N, j=1\cdots M} {\rm d}Y_{ij} P_{\rm out}\left(Y_{ij}, w=0\right) \prod_{a=0\cdots n} \left(\left[\prod_{i=1\cdots N} {\rm d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} {\rm d}v_j^a P_V^a(v_j^a) \right] \right.$$
$$\left. \left[\prod_{i=1\cdots N, j=1\cdots M} \exp\left(\sum_{a=0\cdots n} \left(\frac{\partial g^a}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a}{\sqrt{N}} + \left(\frac{\partial^2 g^a}{\partial w^2}\right)_{Y_{ij},0} \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^1.5}\right)\right)\right]\right) . \quad (317)$$

By expanding the exponential to order two one gets

$$\mathbb{E}(\mathcal{Z}^n) = \int \prod_{i=1\cdots N, j=1\cdots M} {\rm d}Y_{ij} P_{\rm out}\left(Y_{ij}, w=0\right) \prod_{a=0\cdots n} \left[\prod_{i=1\cdots N} {\rm d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} {\rm d}v_j^a P_V^a(v_j^a) \right]$$
$$\prod_{i=1\cdots N, j=1\cdots M} \left[1 + \sum_{a=0\cdots n} \left(\frac{\partial g^a}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a}{\sqrt{N}} + \right.$$
$$\sum_{a=1\cdots n} \left(\frac{\partial g}{\partial w}\right)_{Y_{ij},0} \left(\frac{\partial g^0}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a u_i^0 v_j^0}{N} + \sum_{1\leq a<b\leq n} \left(\frac{\partial g}{\partial w}\right)_{Y_{ij},0} \left(\frac{\partial g}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a u_i^b v_j^b}{N} +$$
$$\left. \sum_{a=0\cdots n} \left[\left(\frac{\partial^2 g^a}{\partial w^2}\right)_{Y_{ij},0} + \left(\frac{\partial^2 g^a}{\partial w^2}\right)^2_{Y_{ij},0} \right] \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^{1.5}}\right) \right] . \quad (318)$$

By averaging with respect to the Y_{ij} one gets

$$\mathbb{E}(\mathcal{Z}^n) = \int \prod_{a=0\cdots n} \left[\prod_{i=1\cdots N} {\rm d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} {\rm d}v_j^a P_V^a(v_j^a) \right]$$
$$\prod_{i=1\cdots N, j=1\cdots M} \left[1 + \sum_{a=1\cdots n} \frac{u_i^a v_j^a u_i^0 v_j^0}{N\widehat{\Delta}} + \sum_{1\leq a<b\leq n} \frac{u_i^a v_j^a u_i^b v_j^b}{N\widetilde{\Delta}} + \sum_{a=1\cdots n} \overline{R} \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^{1.5}}\right) \right] .$$
$$= \int \left[\prod_{a=0\cdots n} \prod_{i=1\cdots N} {\rm d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} {\rm d}v_j^a P_V^a(v_j^a) \right]$$
$$\exp\left(\sum_{i=1\cdots N, j=1\cdots M} \sum_{a=1\cdots n} \frac{u_i^a v_j^a u_i^0 v_j^0}{N\widehat{\Delta}} + \sum_{1\leq a<b\leq n} \frac{u_i^a v_j^a u_i^b v_j^b}{N\widetilde{\Delta}} + \sum_{a=1\cdots n} \overline{R} \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^{1.5}}\right) \right) . \quad (319)$$

We now introduce the order parameters

$$q_{ab}^u = \frac{1}{N} \sum_{i=1\cdots N} u_i^a u_i^b , \quad (320)$$
$$q_{ab}^v = \frac{1}{M} \sum_{j=1\cdots M} v_j^a v_j^b . \quad (321)$$

This leads to

$$\mathbb{E}(\mathcal{Z}^n) = NM \int \prod_{0\leq a\leq b\leq n} {\rm d}q_{ab}^u {\rm d}q_{ab}^v \exp\left(\frac{N\alpha}{\widehat{\Delta}} \sum_{a=1\cdots n} q_{a0}^u q_{a0}^v + \frac{N\alpha}{\widetilde{\Delta}} \sum_{1\leq a<b\leq n} q_{ab}^u q_{ab}^v + \frac{N\alpha}{2} \sum_{a=\cdots n} \overline{R} q_{aa}^u q_{aa}^v \right.$$
$$\left. + N I_u(\{q_{ab}^u\}) + N\alpha I_v(\{q_{ab}^v\}) \right) , \quad (322)$$

where

$$\hat{I}_u(\{q^u_{ab}\}) = \frac{1}{N}\log\left[\int\left(\prod_{a=1\cdots n}\prod_{i=1\cdots N}\mathrm{d}u_i^a P_V^a(u_i^a)\right)\prod_{0\le a\le b\le n}\delta\left(\sum_{i=1\cdots N}u_i^a u_i^b - N q^u_{ab}\right)\right], \tag{323}$$

and

$$\hat{I}_v(\{q^v_{ab}\}) = \frac{1}{M}\log\left[\int\left(\prod_{a=1\cdots n}\prod_{j=1\cdots M}\mathrm{d}v_j^a P_V^a(v_j^a)\right)\prod_{0\le a\le b\le n}\delta\left(\sum_{i=1\cdots N}v_j^a v_j^b - M q^v_{ab}\right)\right]. \tag{324}$$

Here $\hat{I}_u(\{q^u_{ab}\})$ and $\hat{I}_v(\{q^v_{ab}\})$ are the entropy costs one pays in order for the order parameters to take one specific value. We can treat this constraint by going to Fourier space and then rotating the path of integration

$$I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\}) = \log\left(\int\prod_{a=0\cdots n}P_U^a(u_a)\mathrm{d}u^a\exp\left(\sum_{0\le a\le b\le n}\hat{q}^u_{ab}u^{a\top}u^b\right)\right) - \sum_{0\le a\le b\le n}\hat{q}^u_{ab}q^u_{ab}\,, \tag{325}$$

$$I_v(\{q^v_{ab}\},\{\hat{q}^v_{ab}\}) = \log\left(\int\prod_{a=0\cdots n}P_U^a(v_a)\mathrm{d}v^a\exp\left(\sum_{0\le a\le b\le n}\hat{q}^v_{ab}v^{a\top}v^b\right)\right) - \sum_{0\le a\le b\le n}\hat{q}^v_{ab}q^v_{ab}\,. \tag{326}$$

Therefore

$$\mathbb{E}(\mathcal{Z}^n) = NM\int\prod_{0\le a\le b\le n}\mathrm{d}q^u_{ab}\mathrm{d}\hat{q}^u_{ab}\mathrm{d}q^v_{ab}\mathrm{d}\hat{q}^v_{ab}\exp\left(\frac{N\alpha}{\widehat{\Delta}}\sum_{a=1\cdots n}q^u_{a0}q^v_{a0} + \frac{N\alpha}{\widetilde{\Delta}}\sum_{1\le a<b\le n}q^u_{ab}q^v_{ab} + \frac{N\alpha}{2}\sum_{a=\cdots n}\overline{R}q^u_{aa}q^v_{aa}\right.$$
$$\left. + NI_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\})) + N\alpha I_v(\{q^v_{ab}\},\{\hat{q}^v_{ab}\}))\right) \tag{327}$$

We need to extremize this function with respect to all variables. By taking the derivative equal to 0 with respect to variables q^u_{ab} and q^v_{ab} we get

$$\hat{q}^u_{a0} = \frac{\alpha\hat{q}^v_{a0}}{\widehat{\Delta}}, \qquad \hat{q}^v_{a0} = \frac{\hat{q}^u_{a0}}{\widehat{\Delta}}\,, \tag{328}$$

$$\forall 1\le a<b\le n, \hat{q}^u_{ab} = \frac{\alpha\hat{q}^v_{ab}}{\widetilde{\Delta}}, \qquad \hat{q}^v_{ab} = \frac{\hat{q}^u_{ab}}{\widetilde{\Delta}}\,, \tag{329}$$

$$\hat{q}^u_{aa} = \overline{R}\alpha\hat{q}^v_{ab}, \qquad \hat{q}^v_{ab} = \overline{R}\hat{q}^u_{ab}\,. \tag{330}$$

We now assume the replica symmetric ansatz

$$\forall a,b>1, \quad q^u_{ab} = \delta^a_b(\Sigma_u + Q_u) + (1-\delta^a_b)Q_u\,, \tag{331}$$

$$\forall a>1, \quad q^u_{a0} = M_u\,, \tag{332}$$

$$\forall a,b>1, \quad q^v_{ab} = \delta^a_b(\Sigma_v + Q_v) + (1-\delta^a_b)Q_v\,, \tag{333}$$

$$\forall a>1, \quad q^v_{a0} = M_v\,. \tag{334}$$

We can then express the free energy. Let us compute I_u and I_v in that RS ansatz

$$\exp\left(I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\}) + \sum_{0\le a\le b\le n}\hat{q}^u_{ab}q^u_{ab}\right) = \int\prod_{a=0\cdots n}P_U^a(u_a)\mathrm{d}u^a\exp\left(\sum_{0\le a\le b\le n}\hat{q}^u_{ab}u^{a\top}u^b\right). \tag{335}$$

By using Hubbard-Stratonovich identity we get

$$\exp\left(I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\}) + \sum_{0\le a\le b\le n}\hat{q}^u_{ab}q^u_{ab}\right) = \tag{336}$$

$$= \int\mathcal{D}W P_{U_0}(u_0)\mathrm{d}u^0\left[\int\mathrm{d}u P_U(u)\exp\left(\tfrac{\alpha M_v}{\widehat{\Delta}}uu_0 + W\sqrt{\tfrac{\alpha Q_v}{\widetilde{\Delta}}}u - \left(\tfrac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)\right)\tfrac{u^2}{2}\right)\right]^n. \tag{337}$$

We can now compute the limit as $n \to 0$.

$$\lim_{n\to 0} \frac{I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\})}{N} = \frac{\alpha Q_v Q_u}{2\widetilde{\Delta}} - \frac{\alpha M_v M_u}{\widehat{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)(Q_u+\Sigma_u) + \mathbb{E}_{W,u_0}\left[\log\left[\int \mathrm{d}u P_U(u)\exp\left(\frac{\alpha M_v}{\widehat{\Delta}}uu_0 + W\sqrt{\frac{\alpha Q_v}{\widetilde{\Delta}}}u - \left(\frac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)\right)\frac{u^2}{2}\right)\right]\right]$$

$$\lim_{n\to 0} \frac{I_v(\{q^v_{ab}\},\{\hat{q}^v_{ab}\})}{N} = \frac{Q_v Q_u}{2\widetilde{\Delta}} - \frac{M_v M_u}{\widehat{\Delta}} - \overline{R}(Q_v+\Sigma_v)(Q_u+\Sigma_u) + \mathbb{E}_{W,v_0}\left[\log\left[\int \mathrm{d}v P_V(v)\exp\left(\frac{M_u}{\widehat{\Delta}}vv_0 + W\sqrt{\frac{Q_u}{\widetilde{\Delta}}}u - \left(\frac{Q_u}{\widetilde{\Delta}} - \overline{R}(Q_u+\Sigma_u)\right)\frac{u^2}{2}\right)\right]\right]. \quad (338)$$

This finally gives us the replica free energy.

$$\Phi_{\mathrm{RS}}(M_u,Q_u,\Sigma_u,M_v,Q_v,\Sigma_v) = \frac{\alpha Q_v Q_u}{2\widetilde{\Delta}} - \frac{\alpha M_v M_u}{\widehat{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)(Q_u+\Sigma_u) + \mathbb{E}_{W,u_0}\left[\mathcal{Z}_u\left(\frac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v), \frac{\alpha M_v}{\widehat{\Delta}}u_0 + W\sqrt{\frac{\alpha Q_v}{\widetilde{\Delta}}}\right)\right] + \alpha\mathbb{E}_{W,v_0}\left[\mathcal{Z}_v\left(\frac{Q_u}{\widetilde{\Delta}} - \overline{R}(Q_u+\Sigma_u), \frac{M_u}{\widehat{\Delta}}v_0 + Ws\sqrt{\frac{Q_u}{\widetilde{\Delta}}}\right)\right]. \quad (339)$$

This can also be computed with vectorial notations in both the $XX^\top$ and $UV^\top$ setting leading to eqs. (146) and (144).

D. SMALL ρ EXPANSION

In this appendix we give the small-ρ limits of the state evolution update functions for the Rademacher-Bernoulli, Gauss-Bernoulli, 2 balanced groups and Bernoulli models.

Rademacher-Bernoulli model. We want to compute $\forall\beta > 0$ the limit $f_{\mathrm{Rademacher-Bernoulli}}(-\beta\log(\rho))/\rho$ (214) as $\rho \to 0$,

$$\frac{f^{\mathrm{SE}}_{\mathrm{Rademacher-Bernoulli}}(-\beta\log(\rho))}{\rho} = \mathbb{E}_W\left[\frac{\rho\tanh\left(-\beta\log(\rho) + W\sqrt{-\beta\log(\rho)}\right)}{(1-\rho)\frac{\exp(-\beta\log(\rho)/2)}{\cosh\left(-\beta\log(\rho)+W\sqrt{-\beta\log(\rho)}\right)} + \rho}\right]. \quad (340)$$

Taking the small ρ limit here we get

$$\lim_{\rho=0}\frac{f^{\mathrm{SE}}_{\mathrm{Rademacher-Bernoulli}}(-\beta\log(\rho))}{\rho} = \lim_{\rho=0}\frac{1}{\rho^{\beta/2-1}+1} = 1(\beta>2), \quad (341)$$

where we used the fact that the noise $W\sqrt{-\beta\log(\rho)}$ is negligible compared to $\log(\rho)$ when $\rho \to 0$.

Gauss-Bernoulli model. We want to compute $\forall\beta > 0$ the limit $f_{\mathrm{Gauss-Bernoulli}}(-\beta\log(\rho))/\rho$ (216) as $\rho \to 0$ and $r = 1$.

$$f^{\mathrm{SE}}_{\mathrm{Gauss-Bernoulli}}(x)/\rho = \mathbb{E}_{W,x_0}\left[f^{\mathrm{Gauss-Bernoulli}}_{\mathrm{in}}\left(x, xx_0+\sqrt{x}W\right)x_0\right], \quad (342)$$

$$= \mathbb{E}_W\left[\int \frac{\exp\left(-x_0^2/2\right)}{\sqrt{2\pi}}x_0 f^{\mathrm{Gauss-Bernoulli}}_{\mathrm{in}}\left(x, xx_0+\sqrt{x}W\right)\right], \quad (343)$$

By intregrating by part one gets

$$= x\mathbb{E}_W\left[\int \frac{\exp\left(-x_0^2/2\right)}{\sqrt{2\pi}}\frac{\partial f^{\mathrm{Gauss-Bernoulli}}_{\mathrm{in}}}{\partial B}\left(x, xx_0+\sqrt{x}W\right)\right], \quad (344)$$

Here $xx_0 + \sqrt{x}W$ is a Gaussian random variable of mean 0 and variance $x + x^2$. Therefore one has

$$f^{\rm SE}_{\rm Gauss-Bernoulli}(x)/\rho = x\mathbb{E}_W\left[\frac{\partial f^{\rm Gauss-Bernoulli}_{\rm in}}{\partial B}\left(x, \sqrt{x^2+x}W\right)\right], \tag{345}$$

By making another integration by part one gets

$$f^{\rm SE}_{\rm Gauss-Bernoulli}(x)/\rho = \frac{x}{\sqrt{x^2+x}}\mathbb{E}_W\left[f^{\rm Gauss-Bernoulli}_{\rm in}\left(x, \sqrt{x^2+x}W\right)W\right] = \frac{x}{1+x}\mathbb{E}_W\left[W^2\hat{\rho}(x, W\sqrt{x^2+xs}\right], \tag{346}$$

where

$$\hat{\rho}(x, W^2(x^2+x)) = \frac{\rho}{(1-\rho)\exp\left(\frac{-xW^2}{2}\right)\sqrt{1+x}+\rho}\,. \tag{347}$$

By writing $x = -\beta\log(\rho)$ Depending on the value of W, $\hat{\rho}$ will either go to zero or one as $\rho \to 0$. This will appear in the limit of $\hat{\rho}$. By taking $x = -\beta\log(\rho)$ one gets

$$\lim_{\rho\to 0}\hat{\rho}(x, W^2(x^2+x)) = \lim_{\rho\to 0}\frac{1}{(1-\rho)\rho^{\frac{\beta W^2}{2}-1}\sqrt{1-\beta\log(\rho)}+1} = 1(\beta W^2 > 2). \tag{348}$$

From this we deduce that

$$\lim_{\rho\to 0}\frac{f^{\rm SE}_{\rm Gauss-Bernoulli}(-\beta\log(\rho))}{\rho} = \lim_{\rho\to 0}\frac{-\beta\log(\rho)}{1-\beta\log(\rho)}\mathbb{E}_W\left[W^2 1\left(|W| > \sqrt{\frac{2}{\beta}}\right)\right] = \frac{2\exp(-1/\beta)}{\sqrt{\beta\pi}} + \text{erfc}\left(\frac{1}{\sqrt{\beta}}\right). \tag{349}$$

Here we used again the fact that the noise $W\sqrt{-\beta\log(\rho)}$ is negligible compared to $\log(\rho)$ when $\rho \to 0$.

2 balanced groups. We want to compute $\forall\beta > 0$ the limit $f_{\rm Balanced}(-\beta\rho(1-\rho)\log(\rho(1-\rho)))$ (220) as $\rho \to 0$.

$$f^{\rm SE}_{\rm Balanced}(-\beta\rho(1-\rho)\log(\rho(1-\rho))) = \mathbb{E}_W\left[\frac{2\rho(1-\rho)\sinh\left(\frac{-\beta\log(\rho(1-\rho))}{2} + W\sqrt{-\beta\log(\rho(1-\rho))}\right)}{1+2\rho(1-\rho)\left(\cosh\left(\frac{-\beta\log(\rho(1-\rho))}{2} + W\sqrt{-\beta\log(\rho(1-\rho))}\right)-1\right)}\right],$$

$$\lim_{\rho\to 0} f^{\rm SE}_{\rm Balanced}(-\beta\rho(1-\rho)\log(\rho(1-\rho))) = \lim_{\rho\to 0}\frac{2\left[\rho(1-\rho)\right]^{1-\beta/2}}{1+2\left[\rho(1-\rho)\right]^{1-\beta/2}} = 1(\beta > 2)\,. \tag{350}$$

Here we used the fact that the noise $W\sqrt{-\beta\log(\rho(1-\rho))}$ is negligible compared to $\log(\rho(1-\rho))$ when $\rho \to 0$.

Spiked Bernoulli model The state evolution update function in this model is given by (212). Once again we set $x = -\beta\log(\rho)$ to get

$$\frac{f^{\rm SE}_{\rm Bernoulli}(-\beta\log(\rho))}{\rho} = \mathbb{E}_W\left[\frac{1}{1+(1-\rho)\exp\left((\beta/2-1)\log(\rho) + W\sqrt{-\beta\log(\rho)}\right)}\right], \tag{351}$$

$$\lim_{\rho\to 0}\frac{f^{\rm SE}_{\rm Bernoulli}(-\beta\log(\rho))}{\rho} = \lim_{\rho\to 0}\frac{1}{1+(1-\rho)\rho^{\beta/2-1}} = 1(\beta > 2)\,. \tag{352}$$

Here we used the fact that the noise $W\sqrt{-\beta\log(\rho)}$ is negligible compared to $\log(\rho)$ when $\rho \to 0$.

To compute the limiting behaviour of the phase transitions we analyzed the functions $f_{\rm Rademacher-Bernoulli}$,$f_{\rm Gauss-Bernoulli}$ or $f_{\rm Bernoulli}$ (that we will call f_ρ) as follows. We remind

$$\Delta_{\rm Dyn}(\rho) = \max_{x\in\mathbb{R}^+}\frac{f_\rho(x)}{x} = \frac{\rho}{-\log(\rho)}\max_{\beta\in\mathbb{R}^+}\frac{f_\rho(-\beta\log(\rho))}{\rho\beta}\,. \tag{353}$$

In the small ρ limit

$$\Delta_{\rm Dyn}(\rho) \sim_{\rho\to 0}\frac{\rho}{-\log(\rho)}\max_{\beta\in\mathbb{R}^+}\frac{1(\beta > 2)}{\beta} = \frac{\rho}{-2\log(\rho)}\,. \tag{354}$$

The information-theoretic phase transition in the small ρ limit is computed as follows

$$\Delta_{\rm IT}(\rho) \quad = \quad \max_{x\in\mathbb{R}^+}\left\{\Delta(x), \int_0^x {\rm d}m f_\rho\left(u\right) = \frac{x f_{\rho(x)}}{2}\right\}, \tag{355}$$

$$\Delta_{\rm IT}(\rho) \sim_{\rho\to 0} \max_{\beta\in\mathbb{R}^+}\left\{\Delta(-\beta\log(\rho)), \int_0^{-\beta\log(\rho)} {\rm d}u f_\rho\left(u\right) = \frac{-\beta\log(\rho) f_\rho(-\beta\log(\rho))}{2}\right\}, \tag{356}$$

$$\Delta_{\rm IT}(\rho) \sim_{\rho\to 0} \frac{\rho}{-\log(\rho)} \max_{\beta\in\mathbb{R}^+}\left\{\frac{1(\beta>2)}{\beta}, \int_0^{\beta} {\rm d}u 1(u>2) = \frac{\beta 1(\beta>2)}{2}\right\}, \tag{357}$$

$$\Delta_{\rm IT}(\rho) \sim_{\rho\to 0} \frac{\rho}{-4\log(\rho)}\,. \tag{358}$$

E. LARGE RANK BEHAVIOR FOR THE SYMMETRIC COMMUNITY DETECTION

To locate the phase transitions $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ in the symmetric community detection case we make a couple of remarks about the state evolution. First we remark that for $\forall x \in \mathbb{R}^+$ $b = \mathcal{M}_r(x)$ is a fixed point of (231) for $\Delta = \mathcal{M}_r(x)/x$. By definition $\Delta_{\rm Dyn}$ is the greatest Δ for which a fixed point exists

$$\Delta_{\rm Dyn}(r) = \max_{x\in\mathbb{R}^+}\left\{\frac{\mathcal{M}_r(x)}{x}\right\}. \tag{359}$$

To find the $\Delta_{\rm IT}$ we notice that by taking the derivative with respect to Q and M of (144) one finds a combination of (126) and (127). Therefore we have

$$\phi(b_2,\Delta) - \phi(b_1,\Delta) = \frac{r-1}{2r^2\Delta}\int_{b_1}^{b_2} {\rm d}u \mathcal{M}_r\left(\frac{u}{\Delta}\right) - u\,. \tag{360}$$

We deduce a way to compute $\Delta_{\rm IT}$ as

$$\Delta_{\rm IT}(r) = \max_{x\in\mathbb{R}^+}\left\{\frac{\mathcal{M}_r(x)}{x}, \text{s.t.} \int_0^x {\rm d}u \mathcal{M}_r(u) = \frac{x\mathcal{M}_r(x)}{2}\right\}. \tag{361}$$

To compute $\Delta_{\rm IT}$ we evaluate $\mathcal{M}_r$ on a whole interval, then for each x draw a line between point $(0,0)$ and $(x,H(x))$. We then compute the area between $\mathcal{M}_r$ and this line. When this area is zero then $\mathcal{M}_r(x)/x$ gives us $\Delta_{\rm IT}$.

In order to compute $\mathcal{M}_r$ we study the function $\mathcal{M}_r(x)$ where we take $x = \beta r\log(r)$, with $\beta = \Omega(1)$. The important part of $\mathcal{M}_r$ is the integral

$$\int \frac{\exp\left(\frac{x}{r} + \sqrt{\frac{x}{r}}u_1\right)}{\exp\left(\frac{x}{r} + \sqrt{\frac{x}{r}}u_1\right) + \sum_{i=2}^r \exp\left(\sqrt{\frac{x}{r}}u_i\right)} \prod_{i=1}^r \mathcal{D}u_i\,. \tag{362}$$

The important variables to look at are (when taking $x = \beta r\log(r)$)

$$F_1 = \exp\left(\frac{x}{r} + \sqrt{\frac{x}{r}}u_1\right) = \exp\left(\beta\log(r) + \sqrt{\beta\log(r)}u_1\right), \tag{363}$$

$$F_2 = \sum_{i=2}^r \exp\left(\sqrt{\frac{x}{r}}u_i\right) = \sum_{i=2}^r \exp\left(\sqrt{\beta\log(r)}u_i\right). \tag{364}$$

If the typical value of F_1 dominates F_2 as $r\to+\infty$ then $\mathcal{M}_r = 1$, otherwise if F_2 dominates F_1 then $\mathcal{M}_r = 0$.

To estimate F_1, F_2 let us notice that with high probability the maximum value of the u_i will be of order $\sqrt{2\log(r)}$. This is a general property of Gaussian variables that the maximum of r independent Gaussian variables of variance σ^2 is of order $\sigma\sqrt{2\log(r)}$. We can therefore compute the mean of F_2 while conditioning on the fact that all of the u_i

are smaller than $\sqrt{2\log(r)}$. This allows us to compute the typical value of F_1 as $F_1 \sim r^{\beta}$. For F_2 we obtain: when $\beta < 2$ then $F_2 \sim r^{\frac{\beta}{2}+1}$, and when if $\beta > 2$ then $F_2 \sim r^{\sqrt{2\beta}}$. We have to look at which of the F_1 or F_2 has a higher exponent. $\beta = 2$ is the value at which these two exponent cross. Therefore we have

$$\mathcal{M}_r(\beta r \log(r)) = 1\,(\beta > 2)\ . \tag{365}$$

Now let us remind (359) while keeping $x = \beta r \log(r)$ to get

$$\Delta_{\mathrm{Dyn}}(r) \sim_{r\to\infty} \max\left\{\frac{1\,(\beta > 2)}{\beta r \log(r)}, \beta \in \mathbb{R}^+\right\} = \frac{1}{2r\log(r)}\,. \tag{366}$$

To get the information-theoretic transition we use (361). Let us find the β that satisfies equation (361) we get

$$\beta r \log(r)\left[1 - \frac{2}{\beta}\right] = \frac{\beta r \log(r)}{2}\,. \tag{367}$$

We deduce $\beta = 4$ and therefore

$$\Delta_{\mathrm{IT}}(r) \sim_{r\to\infty} \frac{1}{4r\log(r)}\,. \tag{368}$$