Phase transitions and sample complexity in Bayes-optimal matrix factorization

Yoshiyuki Kabashima[1], Florent Krzakala[2], Marc Mézard[3], Ayaka Sakata[1], and Lenka Zdeborová[4,*]

[1] *Department of Computational Intelligence & Systems Science, Tokyo Institute of Technology, Yokohama 226-8502, Japan*

[2] *Laboratoire de Physique Statistique, CNRS UMR 8550 Ecole Normale Supérieure, & Université Pierre et Marie Curie, Paris*

[3] *Univ. Paris-Sud & CNRS, LPTMS, UMR8626, Bât. 100, 91405 Orsay, France.*

[4] *Institut de Physique Théorique, CEA Saclay, and URA 2306, CNRS, 91191 Gif-sur-Yvette, France.*

[*] *To whom correspondence shall be sent: lenka.zdeborova@cea.fr*

We analyse the matrix factorization problem. Given a noisy measurement of a product of two matrices, the problem is to estimate back the original matrices. It arises in many applications such as dictionary learning, blind matrix calibration, sparse principal component analysis, blind source separation, low rank matrix completion, robust principal component analysis or factor analysis. We use the tools of statistical mechanics – the cavity and replica methods – to analyze the achievability and tractability of the inference problems in the setting of Bayes-optimal inference, which amounts to assuming that the two matrices have random independent elements generated from some known distribution, and this information is available to the inference algorithm. In this setting, we compute the minimal mean-squared-error achievable in principle in any computational time, and the error that can be achieved by an efficient approximate message passing algorithm. The computation is based on the asymptotic state-evolution analysis of the algorithm. The performance that our analysis predicts, both in terms of the achieved mean-squared-error, and in terms of sample complexity, is extremely promising and motivating for a further development of the algorithm.

CONTENTS

I. INTRODUCTION

We study in this paper a variety of questions which all deal with the general problem of matrix factorization. Generically, this problem is stated as follows: Given a $M \times P$ dimensional matrix Y, that was obtained from element-wise measurements of a matrix Z, one seeks a factorization $Z = FX$, where the $M \times N$ dimensional matrix F and the $N \times P$ dimensional matrix X must satisfy some specific requirements like sparsity or non-negativity. This type of problem appears in many applications, for instance dictionary learning [1–3], sparse principal component analysis [4], blind source separation [5], low rank matrix completion [6, 7] or robust principal component analysis [8], that will be described below.

Theoretical limits on when matrix factorization is possible and tractable are still rather poorly understood. In this work we make a step towards this understanding by determining the limits of matrix factorization and its algorithmic tractability when Z is created using randomly generated matrices F and X, and measured element-wise via a known noisy output channel $P_{\rm out}(Y|Z)$. Our results are valid in the limit where $N, M, P \to \infty$ with fixed ratios $M/N = \alpha$, $P/N = \pi$.

The methods that we use in this paper are based on a generalization of approximate message passing (AMP) [9] to the matrix factorization problem, and on its asymptotic analysis which is known in statistical physics as the cavity method [10, 11], and has been called state evolution in the context of compressed sensing [9]. We also use the replica method whose results are equivalent to those of the cavity method [10]. These methods are widely believed to be exact in the context of theoretical statistical physics, but most of the results that we shall obtain in the present work are not rigorously established. This work builds upon some previous steps that we described in earlier reports [12, 13]. The message passing algorithm related to our analysis was first presented in [13] and is very closely related to the Big-AMP algorithm developed and tested in [14, 15]; relations and differences with Big-AMP will be mentioned in several places throughout the paper. Our main focus here, beside the detailed derivation of the algorithm, is the asymptotic analysis and phase diagrams which were not studied in [14, 15].

Our general method provides a unifying framework for the study of tractability and identifiability of various matrix factorization problem: the phase diagrams that we shall derive establishes thresholds in the plane $\alpha - \pi$ where the problem of matrix factorization ceases to be solvable in principle, and thresholds where AMP ceases to find the best solution. In most existing works the computation of phase transitions was treated separately for each of the various problems. For instance, redundant dictionaries for sparse representations and low rankness are usually thought as two different kinds of dimensional reduction. Interestingly, in our work these two reductions (and others) are described within a unified formalism, which is theoretically interesting in the context of recent developments [16].

A. Setting of the problem

In a general matrix factorization problem one measures some information about matrix elements of the product of two unknown matrices $F \in \mathbb{R}^{M\times N}$ and $X \in \mathbb{R}^{N\times P}$, whose matrix elements will be denoted $F_{\mu i}$ and x_{il}. Let us denote the product $Z = FX \in \mathbb{R}^{M\times P}$, with elements $z_{\mu l}$. The element-wise measurement $y_{\mu l}$ of $z_{\mu l}$ is then specified by some known probability distribution function $P_{\rm out}(y_{\mu l}|z_{\mu l})$, so that:

$$P_{\rm out}(Y|Z) = \prod_{\mu,l} P_{\rm out}^{\mu l}(y_{\mu l}|z_{\mu l})\,. \tag{1}$$

The goal of matrix factorization is to estimate both matrices F and X from the measurements Y.

In this paper we will treat this problem in the framework of Bayesian inference. In particular we will assume that the matrices F and X were both generated from a known separable probability distribution

$$P_F(F) = \prod_{\mu=1}^{M}\prod_{i=1}^{N} P_F^{\mu i}(F_{\mu i})\,, \tag{2}$$

$$P_X(X) = \prod_{i=1}^{N}\prod_{l=1}^{P} P_X^{il}(x_{il})\,. \tag{3}$$

In the following we shall mostly study the case where the distributions $P_F^{\mu i}$ are all identical (there is a single distribution $P_F(F_{\mu i})$), and the distributions P_X^{il} for various il (as well as $P_{\rm out}^{\mu l}$ for various μl) are also all identical. Our approach can be generalized to the case where $P_F^{\mu i}$, P_X^{il}, $P_{\rm out}^{\mu l}$ depend in a known way on the indices μi, il, and μl, provided that this dependence is through parameters that themselves are taken from separable probability

distributions. Examples of such dependence include the blind matrix calibration or the factor analysis. On the other hand our theory does not cover the case of an arbitrary matrix $\tilde{F}_{\mu i}$ for which we would have $P_F^{\mu i} = \delta(F_{\mu i} - \tilde{F}_{\mu i})$.

The posterior distribution of F and X given the measurements Y is written as

$$P(F, X|Y) = \frac{1}{\mathcal{Z}(Y)} P_F(F) P_X(X) P_{\rm out}(Y|FX) = \frac{1}{\mathcal{Z}(Y)} \prod_{\mu,i} P_F(F_{\mu i}) \prod_{i,l} P_X(x_{il}) \prod_{\mu,l} P_{\rm out}\left(y_{\mu l} | \sum_i F_{\mu i} x_{il}\right) , \quad (4)$$

where $\mathcal{Z}(Y)$ is the normalization constant, known as the partition function in statistical physics. Notice that, while the original problem of finding F and X, given the measurements Y, is not well determined (because of the possibility to obtain, from a given solution, an infinity of other solutions through the transformation $F \to FU^{-1}$ and $X \to UX$, where U is any $N \times N$ nonsingular matrix), the fact of using well defined priors $P_F^{\mu i}$ and P_X^{il} actually lifts the degeneracy and the problem of finding the most probable F, X given the measurements and the priors is well defined. In case the priors $P_F^{\mu i}$ and P_X^{il} do not depend on the indices μl and il we are left with a permutational symmetry between the N column of F and N rows of X. Both in the algorithm and the asymptotic analysis this symmetry is broken and one of the N solutions is chosen at random, but we need to keep this fact in mind when evaluating the mean squared error of our results.

Typically, in most applications, the distributions $P_{\rm out}$, P_F and P_X will depend on a set of parameters (such as the mean, variance, sparsity, noise strength, etc.) that we usually will not write explicitly in the general case, in order to simplify the notations. The prior knowledge of these parameters is not necessarily required in our approach: these parameters can be learned via an expectation maximization algorithm that we will discuss briefly below.

The main purpose of this paper is to present the asymptotic analysis of the Bayesian inference approach to matrix factorization. For this purpose we shall study the "thermodynamic limit" defined as $N, M, P \to \infty$ with fixed ratios $M/N = \alpha$, $P/N = \pi$. When sparsity is involved we consider that a finite *fraction* of matrix elements are non-zero. Similarly, when we treat matrices with low ranks we consider again the ranks to be a finite fraction of the total dimension. This definition of the thermodynamic limit, where all matrix sizes, and the level of sparsity or the rank, all scale similarly, is in contrast to many works where the number of non-zero elements (or the rank) is a fixed constant whereas the dimensions of the matrices go to infinity. All the identifiability thresholds and corresponding algorithmic tractability barriers discussed in our work are valid in this thermodynamic limit.

For the purpose of asymptotic analysis we will also assume that the elements of matrices X, Y, and Z are of order $O(1)$, whereas the elements of F scale in a consistent way, meaning that the mean of each $F_{\mu i}$ is of order $O(1/N)$ and its variance is also of order $O(1/N)$.

B. Bayes-optimal inference

The general case considered in the theoretical parts of this paper is that the matrices X, F and Y are generated with some assumed distributions P_X, P_F and $P_{\rm out}(Y|Z)$, whereas the in reality the matrices were generated using some other distributions P_{X^0}, P_{F^0} and $P^0_{\rm out}(Y|Z)$. The message passing algorithm and its asymptotic evolution will be derived in this case.

However, our most important results concern the Bayes-optimal setting, i.e. when we assume that

$$P_{X^0} = P_X, \qquad P_{F^0} = P_F, \qquad P^0_{\rm out}(Y|Z) = P_{\rm out}(Y|Z) \, . \quad (5)$$

In this case, an estimator $X^\star$ that minimizes the mean-squared error (MSE) with respect to the original signal X^0, defined as

$$\mathrm{MSE}(X|Y) = \int \mathrm{d}F^0 \, \mathrm{d}X^0 \left[\frac{1}{PN} \sum_{il} (x_{il} - x^0_{il})^2 \right] P(F^0, X^0|Y) \, , \quad (6)$$

is obtained from marginals of x_{il} with respect to the posterior probability measure $P(F, X|Y)$, i.e.,

$$x^\star_{il} = \int \mathrm{d}x_{il} \, x_{il} \, \nu_{il}(x_{il}) \, , \qquad \text{where} \qquad \nu_{il}(x_{il}) \equiv \int_{\{F_{\mu j}\}} \int_{\{x_{jn}\}_{jn \neq il}} P(F, X|Y) \, , \quad (7)$$

is the marginal probability distribution of the variable il.

A similar result holds for the estimator of F^0 that minimizes the mean-squared error

$$\mathrm{MSE}(F|Y) = \int \mathrm{d}F^0 \, \mathrm{d}X^0 \left[\frac{1}{M} \sum_{\mu i} (F_{\mu i} - F^0_{\mu i})^2 \right] P(F^0, X^0|Y) \, , \quad (8)$$

which is obtained from the mean of $F_{\mu i}$ with respect to the posterior probability measure $P(F, X|Y)$. In the reminder of this article we will be using these estimators.

1. The Nishimori identities

There are important identities that hold for the Bayes-optimal inference and that simplify many of the calculations that follow. In the physics of disordered systems these identities are known as the Nishimori conditions [17–19]. Basically, the Nishimori identities follow from the fact that the true signal F^0, X^0 is an equilibrium configuration with respect to the Boltzmann measure $P(F, X|Y)$ (4). Hence many properties of the true signal F^0, X^0 can be computed by using averages over the distribution $P(F, X|Y)$ even if one does not know F^0, X^0 precisely (4).

In order to derive the Nishimori identities, we need to define three types of averages: the thermodynamic average, the double thermodynamic average, and the disorder average.

- Consider a function $A(F, X)$ depending on a "trial" configuration F and X. We define the "thermodynamic average" of A given Y as:

$$\langle A(F,X)\rangle_{F,X|Y} \equiv \int \mathrm{d}X\, \mathrm{d}F\, A(F,X)\, P(F,X|Y)\,, \tag{9}$$

  where $P(F, X|Y)$ is given by Eq. (4). The thermodynamic average $\langle A(F, X)\rangle_{F,X|Y}$ is a function of Y.

- Similarly, for a function $A(F_1, X_1, F_2, X_2)$ that depends on *two* trial configurations X_1, F_1 and X_2, F_2, we define the "double thermodynamic average" of A given Y as:

$$\langle\langle A(F_1,X_1,F_2,X_2)\rangle\rangle_{F_1,X_1,F_2,X_2|Y} \equiv \int \mathrm{d}X_1\, \mathrm{d}F_1\, \mathrm{d}X_2\, \mathrm{d}F_2\, A(F_1,X_1,F_2,X_2)\, P(F_1,X_1|Y)P(F_2,X_2|Y)\,; \tag{10}$$

  this is again a function of Y.

- For a function B that depends on the measurement Y and on the true signal F^0, X^0, we define the "disorder average" as

$$[B(F^0,X^0,Y)]_{F^0,X^0,Y} \equiv \int \mathrm{d}Y\, \mathrm{d}X^0\, \mathrm{d}F^0\, P_X(X^0)\, P_F(F^0)\, P_{\mathrm{out}}(Y|F^0X^0)\, B(F^0,X^0,Y)\,. \tag{11}$$

  Note that if the quantity B depends only on Y, then we have

$$[B(Y)]_Y \equiv \int \mathrm{d}Y\, \mathcal{Z}(Y) B(Y)\,. \tag{12}$$

  This is simply because the partition function $\mathcal{Z}(Y)$ is $\mathcal{Z}(Y) = \int \mathrm{d}X^0\, \mathrm{d}F^0\, P_X(X^0)\, P_F(F^0)\, P_{\mathrm{out}}(Y|F^0X^0)$.

Let us now derive the Nishimori identities. We consider a function $A(F, X, F^0, X^0)$ that depends on the trial configuration F, X and on the true signal F^0, X^0. Its thermodynamic average $\langle A(F, X, F^0, X^0)\rangle_{F,X|Y}$ is a function of the measurement Y and the true signal F^0, X^0. The disorder average of this quantity can be written as

$$\begin{aligned}
&[\langle A(F,X,F^0,X^0)\rangle_{F,X|Y}]_{F^0,X^0,Y} \\
&= \int \mathrm{d}F^0\ \mathrm{d}X^0\, \mathrm{d}Y\ P_X(X^0)\, P_F(F^0)\, P_{\mathrm{out}}(Y|F^0X^0) \int \mathrm{d}F\, \mathrm{d}X\, A(F,X,F^0,X^0)\, P(F,X|Y) \\
&= \int \mathrm{d}Y\, \mathcal{Z}(Y) \int \mathrm{d}F^0\ \mathrm{d}X^0 \mathrm{d}F\, \mathrm{d}X\, A(F^0,X^0,F,X) \frac{P_X(X^0)P_F(F^0)P_{\mathrm{out}}(Y|F^0X^0)}{\mathcal{Z}(Y)} P(F,X|Y)\ .
\end{aligned} \tag{13}$$

In this last expression, renaming F, X to F_1, X_1 and F^0, X^0 to F_2, X_2, we see that the average over F, X, F^0, X^0 is nothing but the double thermodynamic average:

$$\begin{aligned}
[\langle A(F,X,F^0,X^0)\rangle_{F,X|Y}]_{F^0,X^0,Y} &= \int \mathrm{d}Y\, \mathcal{Z}(Y) \int \mathrm{d}F_1\, \mathrm{d}X_1\, \mathrm{d}F_2\, \mathrm{d}X_2\, A(F_1,X_1,F_2,X_2) P(F_1,X_1|Y) P(F_2,X_2|Y) \\
&= [\langle\langle A(F_1,X_1,F_2,X_2)\rangle\rangle_{F_1,X_1,F_2,X_2|Y}]_Y\ ,
\end{aligned} \tag{14}$$

where in the last step we have used the form (12) of the disorder average.

The identity

$$[\langle A(F,X,F^0,X^0)\rangle_{F,X|Y}]_{F^0,X^0,Y} = [\langle\langle A(F_1,X_1,F_2,X_2)\rangle\rangle_{F_1,X_1,F_2,X_2|Y}]_Y \tag{15}$$

is the general form of the Nishimori identity. Written in this way, it holds for many inference problems where the model for signal generation is known.

It is well known (although it may be sometimes hard to prove) that in the thermodynamic limit the "self-averaging" property holds for many quantities. This means that the distribution of $\langle A(F,X,F^0,X^0)\rangle_{F,X|Y}$ concentrates with respect to the realization of disorder Y,F^0,X^0, i.e. for large system sizes the quantity $\langle A(F,X,F^0,X^0)\rangle_{F,X|Y}$ converges with probability one to its average over disorder, $[\langle A(F,X,F^0,X^0)\rangle_{F,X|Y}]_{F^0,X^0,Y}$. This self-averageness property makes the Nishimori identity very useful in practice.

To give one particularly useful example, let us define $m_x = (1/(PN))\sum_{il} x^0_{il}x_{il}$ and $q_X = (1/(PN))\sum_{il} x^1_{il}x^2_{il}$. The Nishimori condition states that $[\langle m_x\rangle_{F,X|Y}]_{F^0,X^0,Y} = [\langle\langle q_X\rangle\rangle]_{F_1,X_1,F_2,X_2|Y}]_Y$. Assuming that the quantities m_x and q_X are self-averaging, we obtain in the thermodynamic limit, for almost all Y: $\langle m_x\rangle_{F,X|Y} = \langle\langle q_X\rangle\rangle_{F_1,X_1,F_2,X_2|Y}$. Explicitly, this gives:

$$(1/(PN))\sum_{il} x^0_{il}\langle x_{il}\rangle_{F,X|Y} = (1/(PN))\sum_{il}\left(\langle x_{il}\rangle_{F,X|Y}\right)^2 \ . \tag{16}$$

This is a remarkable identity concerning the mean of x_{il} with the posterior distribution $\nu_{il}(x_{il})$. The left-hand side measures the overlap between this mean and the seeked true value x^0_{il}. The right-hand side measures the self overlap of the mean, which can be estimated without any knowledge of the true value x^0_{il}, by generating two independent samples from $P(F,X|Y)$.

By symmetry, all these examples apply also to averages and functions of the matrix F:

$$(1/M)\sum_{\mu l} F^0_{\mu l}\langle F_{\mu l}\rangle_{F,X|Y} = (1/M)\sum_{il}\left(\langle F_{\mu l}\rangle_{F,X|Y}\right)^2 \ . \tag{17}$$

2. Absence of replica symmetry breaking in Bayes-optimal inference

The study of the Bayes-optimal inference, i.e. the case when the prior probability distribution matches the true distribution with which the data have been generated, is fundamentally simpler than the general case. The fundamental reason for this simplicity is that, in the Bayes-optimal case, a major complication referred to as *static replica symmetry breaking* (RSB) in statistical physics literature [11] does not appear. RSB is a source of computational complications in many optimization and constraint satisfaction problems such as the random K-satisfiability [20] or models of spin glasses [10].

In statistical mechanics the matrix factorization problem where elements of F and X are generated independently at random is a mean field system. In such systems, if there is no static RSB, the approximation made when deriving the belief propagation equations become negligible in the thermodynamic limit, and therefore the belief propagation equations are asymptotically valid. In the case of Bayes-optimal inference, it follows from the Nishimori conditions that the two-point correlations between variables are such that the static RSB cannot happen for the posterior measure $P(F,X|Y)$ [21]. This situation then provides a great simplification because it means that as long as elements of X and F are random independent variables then there is a fixed point of GAMP and a fixed point of the state evolution that describes asymptotically exactly the performance of the (otherwise intractable) optimal Bayes inference. If the GAMP or the state evolution has multiple fixed points one needs to choose the one that is maximizing the likelihood $\mathcal{Z}(Y)$. We shall not detail here the long discussion relating the Nishimori conditions to the absence of RSB, we refer the reader to the literature [19].

3. Expectation maximization learning of parameters

In practice, even if one knows the probability distribution that approximates well the matrix elements of F and X, one often does not have the full information about the parameters of this distribution (typically its mean and variance). In the same manner, one might know the nature of the measurement noise, but without a precise knowledge of its amplitude.

In order to learn such hyper-parameters, a common technique in statistical inference is the expectation maximization [22], where one updates iteratively the hyper-parameters in order to maximize the posterior likelihood, i.e. the normalization $\mathcal{Z}(Y)$ in the posterior probability measure. In the context of approximate message passing the expectation maximization has been derived and implemented e.g. in [23, 24]. It turns out that the expectation maximization

update of hyper-parameters is analogous to the update where one imposes the Nishimori conditions. In other words one updates the hyper-parameters in such a way that they correspond to the values of the currently estimated signal. For instance the mean and variance of the distribution P_X should correspond to the empirical mean and variance of the estimators of elements x_{il}. The variance of the noise in a Gaussian output channel should be the same as the squared difference between the observed matrix Y and the estimator of the product FX, etc. Therefore, imposing the Nishimori conditions is a way to perform expectation maximization learning of hyperparameters and can be easily implemented within the GAMP code.

It is interesting to note that, as the Bayes-optimal setting has several nice properties in terms of simplicity of the analytical approach and in terms of convergence of the message-passing algorithms, bringing the iterations back on the Nishimori line by doing expectation maximization improves quite generically the convergence of the algorithm. This is one of the properties observed in [15].

C. Applications of matrix factorization

Several important problems which have received a lot of attention recently are special cases of the matrix factorization problem as set above. In this paper we will analyse the following ones.

a. Dictionary learning Many signals of interest are sparse in some basis, this fact is widely used in data compression and more recently in compressed sensing. A lot of work has been devoted to analyzing bases in which different data are sparse. The goal of dictionary learning is to infer a basis in which the data are sparse based purely on a large number of samples from the data. The $M \times P$ matrix Y then represents the P samples of M-dimensional data. The goal is to decompose $Y = FX + W$ into a $M \times N$ matrix F, and a $N \times P$ sparse matrix X, W is the noise.

In this paper we will analyse the following teacher-student scenario of dictionary learning. We will generate a random Gaussian matrix F^0 with iid elements of zero mean and variance $1/N$, and a random Gauss-Bernoulli matrix X^0 with fraction $0 < \rho < 1$ of non-zero elements. The non-zero elements of X^0 will be iid Gaussian with mean $\overline{x}$ and variance σ. The noise W^0 with elements $w^0_{\mu l}$ is also iid Gaussian with zero mean and variance Δ. We hence have

$$P_F(F_{\mu i}) = P_{F^0}(F_{\mu i}) = \frac{1}{\sqrt{2\pi/N}} e^{-\frac{N F_{\mu i}^2}{2}} , \tag{18}$$

$$P_X(x_{il}) = P_{X^0}(x_{il}) = (1-\rho)\delta(x_{il}) + \frac{\rho}{\sqrt{2\pi\sigma}} e^{-\frac{(x_{il}-\overline{x})^2}{2\sigma}} , \tag{19}$$

$$P_{\rm out}(y_{\mu l}|z_{\mu l}) = \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{(y_{\mu l}-z_{\mu l})^2}{2\Delta}} , \tag{20}$$

where $\delta(x)$ is the Dirac delta function. The goal is to infer F^0 and X^0 from the knowledge of Y with the smallest possible number of samples P.

In the noiseless case, $\Delta = 0$, exact reconstruction might be possible only when the observed information is larger that the information that we want to infer. This provides a simple counting bound on the number of needed samples P:

$$P \geq \frac{\alpha}{\alpha - \rho} N . \tag{21}$$

Note that the above assumptions on P_F, P_X and $P_{\rm out}$ likely do not hold in any realistic problem. However, it is of theoretical interest to analyze the average theoretical performance and tractability of such a problem, as it gives a well defined benchmark. Moreover, we anticipate that if we develop an algorithm working well in the above case it might also work well in many real applications where the above assumptions are not satisfied, in the same spirit as the approximated message passing algorithm derived for compressed sensing with zero mean Gaussian measurement matrices [9] works also for other kinds of matrices.

Typically, we would look for an invertible basis F^0 with $N = M$, in that case we speak of a *square* dictionary. However, with the compressed-sensing application in mind, it is also very interesting to consider that Y might be under-sampled measurements of the actual signal, corresponding then to $\alpha = M/N < 1$. Hence we will be interested in the whole range $0 < \alpha \leq 1$. The regime of $\alpha < 1$ corresponds to an *overcomplete* dictionary, in which case each of the P measurements $\vec{y}_l$ is a sparse linear combination of the columns (atoms) of the dictionary.

We remind that the N columns of the matrix F^0 can always be permuted arbitrarily and multiplied by ± 1. This is also true for rows of the matrix X^0: all these operations do not change Y, nor the posterior probability distribution. This is hence an intrinsic freedom in the dictionary learning problems that we have to keep in mind. Note that many works consider the dictionary to be column normalized, which lifts part of the degeneracy in some optimization formulations of the problem. In our setting the equivalent of column normalization is asymptotically determined by the properties of the prior distribution.

b. Blind matrix calibration In dictionary learning one does not have any specific information about the elements $F^0_{\mu i}$. In some applications of compressed sensing one might have an approximate knowledge of the measurement matrix: it is often possible to use known samples of the signal x in order to calibrate the matrix in a supervised manner (i.e. using known training samples of the signal). Sometimes, however, the known training samples are not available and hence the only way to calibrate is to measure a number of unknown samples and perform their reconstruction and calibration of the matrix at the same time, such a scenario is called the *blind calibration*.

In blind matrix calibration, the properties of the signal and the output function are the same as in dictionary learning, eqs. (19-20). As for the matrix elements $F^0_{\mu i}$ one knows a noisy estimation $F'_{\mu i}$. In this work we will assume that this estimation was obtained from $F^0_{\mu i}$ as follows

$$F'_{\mu i} = \frac{F^0_{\mu i} + \sqrt{\eta}\xi_{\mu i}}{\sqrt{1+\eta}}\,, \tag{22}$$

where $\xi_{\mu i}$ is a Gaussian random variables with zero mean and variance $1/N$. This way, if the matrix elements $F^0_{\mu i}$ have zero mean and variance $1/N$, then the same is true for the elements $F'_{\mu i}$.

The control parameter η is then quantifying how well one knows the measurement matrix. It provides a way to interpolate between the pure compressed sensing $\eta = 0$, where one knows the measurement matrix F^0, and the dictionary learning problems $\eta \to \infty$. Explicitly, the prior distribution of a given element of the matrix F is

$$P_F(F_{\mu i}) = \mathcal{N}\left(\frac{F'_{\mu i}}{\sqrt{1+\eta}}, \frac{\eta}{N(1+\eta)}\right)\,, \tag{23}$$

where $\mathcal{N}(a,b)$ is a Gaussian distribution with mean a and variance b.

c. Low-rank matrix completion Another special case of matrix factorization that is often studied is the low-rank matrix completion. In that case one "knows" only a small (but in our case finite when $N \to \infty$) fraction ϵ of elements of the $M \times P$ matrix Y. Also one knows which elements are known and which are not; let us call $\mathcal{M}$ the set on which elements are known, it is a set of size ϵMP. In this case the output function is:

$$\begin{aligned} P_{\rm out}(y_{\mu l}|z_{\mu l}) &= \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{(y_{\mu l}-z_{\mu l})^2}{2\Delta}} \quad \text{if} \quad \mu l \in \mathcal{M}\,, \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2_{\mu l}}{2}} \quad \text{if} \quad \mu l \notin \mathcal{M}\,. \end{aligned} \tag{24}$$

The precise choice on the function on the second line is arbitrary as long as it does not depend on the $z_{\mu l}$. In what follows we will assume that the ϵMP known elements were chosen uniformly at random.

In low rank matrix completion N is small compared to M and P, hence both π and α are relatively large. Note, however, that the limit we analyse in this paper keeps $\pi = O(1)$ and $\alpha = O(1)$ while $N \to \infty$, whereas in many previous works on low-rank matrix completion the rank was considered or $O(1)$ and hence α and π of order $O(N)$. Compared to those works the analysis here applies to "not-so-low-rank" matrix completion. The question is what fraction ϵ of elements of Y needs to be known in order to be able to reconstruct the two matrices F^0 and X^0.

For negligible measurement noise, $\Delta = 0$, a simple counting bound gives that the fraction of known elements we need for reconstruction is at least

$$\epsilon \geq \frac{\alpha+\pi}{\alpha\pi}\,. \tag{25}$$

Again we will study the student-teacher scenario when a low-rank Z is generated from X^0 and F^0 having iid elements distributed according to eq. (18) and (19) with $\rho = 1$ (no sparsity). To construct Y we keep a random fraction ϵ of elements of Z, and the goal is to reconstruct X^0 and F^0 from that knowledge.

d. Sparse PCA and blind source separation Principal component analysis (PCA) is a well-known tool for dimensional reduction. One usually considers the singular value decomposition (SVD) of a given matrix and keeps a given number of largest values, thus minimizing the mean square error between the original matrix and its low-rank approximation. The SVD is computationally tractable, and provides the minimization of the mean square error between the original matrix and its low-rank approximation. However, with additional constraints there is no general tractable approach.

A variant of PCA that is relevant for a number practical application requires that one of the low-rank components is sparse. The goal is then to approximate a matrix Y by a product FX where F is a tall matrix, and X a wide sparse matrix. The teacher-student scenario for sparse PCA that we will analyse in this paper uses eq. (18-20) and the matrix dimensions are such that $\pi = P/N$ and $\alpha = M/N$ are both large, but still of order $O(1)$, and comparable one

to the other. Hence it is only the region of interest for α and π that makes this problem different from the dictionary learning. Note that, in the same way as for the matrix completion, many works in the literature consider $N = O(1)$ whereas here we have $N \to \infty$ in such a way that $\pi = P/N = O(1)$ and $\alpha = M/N = O(1)$. Hence we work with low rank, but not as low as most of the existing literature.

In the zero measurement noise case, $\Delta = 0$, the simple counting bound gives that the rank N for which the reconstruction problem may be solvable needs to be smaller than

$$N \leq \frac{MP}{M + \rho P}\,. \tag{26}$$

One important application where sparse PCA is relevant is the blind source separation problem. Given N ρ-sparse (in some known basis) source-signals of dimension P, they are mixed in an unknown way via a matrix F into M channel measurements Y. In blind source separation typically both the number of sources N and the number of channels (sensors) M are small compared to P. When $N < M$ we obtain an overdetermined problem which may be solvable even for $\rho = 1$. More interesting is the undetermined case with the number of sensors smaller than the number of sources, $M < N$, which would not be solvable unless the signal is sparse $\rho < 1$ (in some basis), in that case the bound (26) applies.

e. Robust PCA Another variant of PCA that arises often in practice is the robust PCA, where the matrix Y is very close to a low rank matrix FX plus a sparse full rank matrix. The interpretation is that Y was created as low rank but then a small fraction of elements was distorted by a large additive noise. The resulting Y is hence not low rank.

In this paper we will analyse a case of robust PCA when F and X are generated from eq. (18-19) with $\rho = 1$ and the output function is

$$P_{\rm out}(y_{\mu l}|z_{\mu l}) = \epsilon \frac{1}{\sqrt{2\pi\Delta_s}} e^{-\frac{(y_{\mu l}-z_{\mu l})^2}{2\Delta_s}} + (1-\epsilon)\frac{1}{\sqrt{2\pi\Delta_l}} e^{-\frac{(y_{\mu l}-z_{\mu l})^2}{2\Delta_l}}\,, \tag{27}$$

where ϵ is the fraction of elements that were not largely distorted, $\Delta_s \ll 1$ is the small measurement noise on the non-distorted elements, Δ_l is the large measurement noise on the distorted elements. We will require $\Delta_l \approx \overline{x}^2 + \sigma$ to be comparable to the variance of $z_{\mu l}$ such that there is no reliable way to tell which elements were distorted by simply looking at the distribution of $y_{\mu l}$. The parameters regime we are interested in here is π and α both relatively large and comparable one to another.

In robust PCA in the zero measurement noise case, $\Delta_s = 0$, the simple counting bound gives that the fraction of non-distorted elements ϵ under which the reconstruction may still be solvable needs to satisfy the same bound as for matrix completion (25). Indeed this counting bound does not distinguish between the case of matrix completion when the position of known elements of Y is known and the case of RPCA when their positions are unknown.

f. Factor analysis One major objective of multivariate data analysis is to infer an appropriate mechanism from which observed high-dimensional data are generated. Factor analysis (FA) is a representative methodology for this purpose. Let us suppose that a set of M-dimensional vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_P$ is given and its mean is set to zero by a pre-processing. Under such a setting, FA assumes that each observed vector $\mathbf{y}_l$ is generated by $N(\leq M)$-dimensional *common factor* $\mathbf{x}_l$ and M-dimensional *unique factor* $\mathbf{w}_l$ as $\mathbf{y}_l = F\mathbf{x}_l + \mathbf{w}_l$, where $F \in \mathbb{R}^{M\times N}$ is termed the *loading matrix*. The goal is to determine the entire set of F, $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P)$, and $W = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_P)$ from only $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_P)$. Therefore, FA is also expressed as a factorization problem of the form of $Y = FX + W$ in matrix terms.

The characteristic feature of FA is to take into account the site dependence of the output function as

$$P_{\rm out}(y_{\mu l}|z_{\mu l}, \psi_\mu) = \frac{1}{\sqrt{2\pi\psi_\mu}} e^{-\frac{(y_{\mu l}-z_{\mu l})^2}{2\psi_\mu}}\,, \tag{28}$$

where ψ_μ denotes the variance of the μ-th component of the unique factor. In addition, it is normally assumed that $\mathbf{x}_l$ $(l = 1, 2, \ldots, P)$ independently obeys a zero mean multivariate Gaussian distribution, the variance of which is set to the identity matrix in the basic case. These assumptions make it possible to express the log-likelihood for Y in a compact manner as $\log P(Y|F,\Psi) = \log\left(\int \prod_{\mu=1,l=1}^{M,P} P_{\rm out}(y_{\mu l}|z_{\mu l},\psi_\mu) \prod_{i=1,l=1}^{N,P} P_X(x_{il}) dX\right) = -\frac{P}{2}\log\left(FF^{\rm T}+\Psi\right) - \frac{1}{2}\sum_{\mu=1}^{P} \mathbf{y}^{\rm T}\left(FF^{\rm T}+\Psi\right)\mathbf{y} + const$, where $\Psi = (\psi_\mu \delta_{\mu\nu}) \in \mathbb{R}^{M\times M}$. In a standard scheme, F and Ψ, which parameterize the generative mechanism of the observed data Y, are determined by maximizing the log-likelihood function [25]. After obtaining these, the posterior distribution $P(X|Y, F^{\rm ML}, \Psi^{\rm ML})$ is used to estimate the common factor X, where $F^{\rm ML}$ and $\Psi^{\rm ML}$ are the maximum likelihood estimators of F and Ψ, respectively. Finally, unique factor W is determined from the relation $Y = FX + W$. Other heuristics to minimize certain discrepancies between the sample variance-covariance matrix $P^{-1}YY^{\rm T}$ and $FF^{\rm T} + \Psi$ are also conventionally used for determining F and Ψ. As an alternative approach, we employ the Bayesian inference for FA.

g. Non-negative matrix factorization In many application of matrix factorization it is known that both the coefficient of the signals X and the coefficients of the dictionary F must be non-negative. In our setting this can be taken into account easily by imposing the distributions P_X and P_F to have non-negative support. In the student-teacher scenario of this paper we can hence consider the nonzero elements of X to be $P_X(x) = 0$ for $x < 0$ and $P_X(x) = N(\overline{x}, \sigma)$ for $x > 0$, and analogously for P_F.

D. Related work and positioning of our contribution

Matrix factorization and its special cases as mentioned above are well studied problems with extensive theoretical and algorithmic literature that we are not able to cover fully here. We will hence only give examples of relevant works and will try to be exhaustive only concerning papers that are very closely related to our work (i.e. papers using message passing algorithms or analyzing the same phase transitions in the same scaling limits).

The dictionary learning problem was identified in the work of [1, 2] in the context of image representation in the visual cortex, and the problem was studied extensively since [3]. Learning of overcomplete dictionaries for sparse representations of data has many applications, see e.g. [26]. One of the principal algorithms that is used is based on K-SVD [27]. Several authors studied the identifiability of the dictionary under various (in general weaker than ours) assumptions, e.g. [28–34]. An interesting view on the place of sparse and redundant representations in todays signal processing is given in [16].

The closely related problems of sparse principal component analysis or blind source separation is also explored in a number of works, see e.g. [35–39]. A short survey on the topic with relevant references can be found in [40].

Matrix completion is another problem that belongs to the class treated in this paper. Again, many important works were devoted to this problem giving theoretical guarantees, algorithm and applications, see e.g. [6, 7, 41, 42].

Another related problem is the robust principal component analysis that was also studied by many authors; algorithms and theoretical limits were analyzed in [8, 43–45].

Our work differs from the mainstream of existing literature in its general positioning. Let us mention here some of the main differences with most other works.

- Most existing works concentrate on finding theoretical guarantees and algorithms that work in the worst possible case of matrices F and X. In our work we analyze the typical cases when elements of F and X are generated at random. Arguably a worst case analysis is useful for practical applications in that it provides some guarantee. On the other hand, in some large size applications one can be confronted in practice with a situation which is closer to the typical case that we study here. Our typical-case analysis provides results that are much tighter than those usually obtained in literature in terms of both achievability and tractability. For instance our results imply much smaller sample complexity, or much smaller mean-squared error for a given signal-to-noise ratio, etc.

- Our main contribution is the asymptotic phase diagram of Bayes-optimal inference in matrix factorization. Special cases of our result cover important problems in signal processing and machine learning. In the present work we do not concentrate on validation of the associated approximate message passing algorithm on practical examples, nor on its comparison with existing algorithms. A contribution in this direction can be found in [13, 15].

- A large part of existing machine-learning and signal-processing literature provides rigorously proven theorems. Our work is based on statistical physics methods that are conjectured to give exact results. While many results obtained with these methods on a variety of problems have been indeed proven later on, a general proof is not known yet.

- Many existing algorithms are based on convex relaxations of the corresponding problems. In this paper we analyze the Bayes-optimal setting. Not surprisingly, this setting gives much better performance than convex relaxation. The reason why it is less studied is that it is often considered as hopelessly complicated from the algorithmic perspective; however some recent results using message-passing algorithms which stand at the core of our analysis have shown very good performance in the Bayes-optimal problem. Besides, the Bayes-optimal offers an optimal benchmark for testing algorithmic methods.

- When treating low rank matrices we consider the rank to be a finite fraction of the total dimension, whereas most of existing literature considers the rank to be a finite constant.

- When treating sparsity we consider the number of non-zeros is a finite fraction of the total dimension, whereas existing literature often considers a constant number of non-zeros.

The paper is organized as follows: First, in Sec. II we give a detailed derivation of the approximate message passing algorithm for the matrix factorization problem. This algorithm is equivalent to maximization of the Bethe free entropy, whose expression is discussed in Sec. III. These first two sections thus state our algorithmic approach to matrix factorization. The asymptotic performance of the AMP algorithm and the Bayes-optimal MMSE are analyzed in Sec. IV using two technics: i) the state evolution of the AMP algorithm and ii) the replica method. As usual, the two methods are found to agree and to give the same predictions. We then use these results in order to study some exemples of matrix factorization problems in Sec. V. In particular we derive the phase diagram, the MMSE and the sample complexity of dictionary learning, blind matrix calibration, sparse PCA, blind source separation, low rank matrix completion, robust PCA and factor analysis. Our results are summarized and discussed in the conclusion in Sec. VI.

II. APPROXIMATE MESSAGE PASSING FOR MATRIX FACTORIZATION

A. Approximate belief propagation for matrix factorization

Bayes inference amounts to the computation of marginals of the posterior probability (4). In order to make it tractable we have to resort to approximations. In compressed sensing, the Bayesian approach combined with a belief propagation (BP) reconstruction algorithm leads to the so-called approximate message passing (AMP) algorithm. It was first derived in [9] for the minimization of ℓ_1, and subsequently generalized in [46, 47]. We shall now adapt the same strategy to the case of matrix factorization.

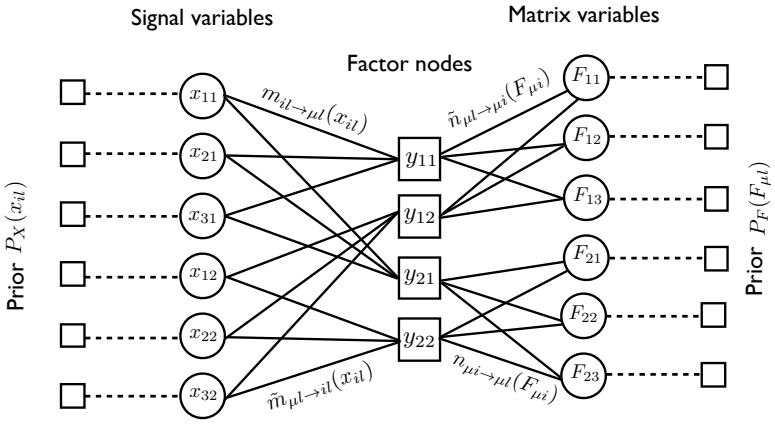

FIG. 1. Factor graph used for the belief propagation inference, here drawn using $N = 3$, $P = 2$ and $M = 2$. The factor nodes are associated to the probability $P_{\rm out}(y_{\mu l}|\sum_i F_{\mu i}x_{il})$.

The factor graph corresponding to the posterior probability (4) is depicted in Fig. 1. The canonical BP iterative equations are written using messages $m_{il\to\mu l}(x_{il})$, $n_{\mu i\to\mu l}(F_{\mu i})$ from variables to factors, and using messages $\tilde{m}_{\mu l\to il}(x_{il})$, $\tilde{n}_{\mu l\to\mu i}(F_{\mu i})$ from factors to variables. On tree graphical models the messages are defined as marginal probabilities of their arguments conditioned to the fact that the variable/factor to which the message is incoming is not present in the graph. The following BP equations provide the exact values for these conditional marginals on trees

$$m_{il\to\mu l}(t+1,x_{il}) = \frac{1}{\mathcal{Z}_{il\to\mu l}} P_X(x_{il}) \prod_{\nu(\neq\mu)}^{M} \tilde{m}_{\nu l\to il}(t,x_{il})\,, \tag{29}$$

$$n_{\mu i\to\mu l}(t+1,F_{\mu i}) = \frac{1}{\mathcal{Z}_{\mu i\to\mu l}} P_F(F_{\mu i}) \prod_{n(\neq l)}^{P} \tilde{n}_{\mu n\to\mu i}(t,F_{\mu i})\,, \tag{30}$$

$$\tilde{m}_{\mu l\to il}(t,x_{il}) = \frac{1}{\mathcal{Z}_{\mu l\to il}} \int \prod_{j(\neq i)}^{N} dx_{jl} \prod_{k}^{N} dF_{\mu k}\, P_{\rm out}(y_{\mu l}|\sum_{k}^{N} F_{\mu k}x_{kl}) \prod_{k}^{N} n_{\mu k\to\mu l}(t,F_{\mu k}) \prod_{j(\neq i)}^{N} m_{jl\to\mu l}(t,x_{jl})\,, \tag{31}$$

$$\tilde{n}_{\mu l\to\mu i}(t,F_{\mu i}) = \frac{1}{\mathcal{Z}_{\mu l\to\mu i}} \int \prod_{j}^{N} dx_{jl} \prod_{k(\neq i)}^{N} dF_{\mu k}\, P_{\rm out}(y_{\mu l}|\sum_{k}^{N} F_{\mu k}x_{kl}) \prod_{k(\neq i)}^{N} n_{\mu k\to\mu l}(t,F_{\mu k}) \prod_{j}^{N} m_{jl\to\mu l}(t,x_{jl})\,, \tag{32}$$

where $\mathcal{Z}_{il\to\mu l}$, $\mathcal{Z}_{\mu i\to\mu l}$, $\mathcal{Z}_{\mu l\to il}$, $\mathcal{Z}_{\mu l\to\mu i}$ are normalization constants ensuring that all the messages are probability distributions, $t \in \mathbb{N}$ is denoting the iteration time-step, and the notation $\prod_{\nu(\neq\mu)}^{M}$ means a product over all integer values of ν in $\{1,\ldots,M\}$, except the value μ.

Of course, the factor graph of matrix factorization, shown in Fig. 1, is extremely far from a tree. The above BP equations can, however, still be asymptotically exact if the dependence between the incoming messages is negligible in the leading order. This indeed happens in compressed sensing (where the matrix F is a known matrix, generated randomly with zero mean), as follows from the rigorously proven success of approximate message passing [9, 48]. In the present case of matrix factorization, we do not have any rigorous proof yet, but based on our experience from studies of mean field spin glass systems [10, 11], we conjecture that the fixed points of the above belief propagation equations describe asymptotically exactly (in the same sense as for compressed sensing) the performance of Bayes-optimal inference (note that a crucial point for this statement is that we deal with a situation where there can be no static RSB, as can be seen from Nishimori identities). Hence the analysis of the fixed points of the above equations leads to the understanding of information-theoretic limitations for matrix factorization. The associated phase transitions describe possible algorithmic barriers. This analysis is the main goal and result of the present paper.

The above BP iterative equations are written for probability distributions over real values variables and the $2N-1$-uple integrals from the r.h.s. are intractable in this form. We now define means and variances of the variable-to-factor messages as

$$a_{il\to\mu l}(t) = \int \mathrm{d}x_{il}\, m_{il\to\mu l}(t,x_{il})\, x_{il}\,, \tag{33}$$

$$c_{il\to\mu l}(t) + a^2_{il\to\mu l}(t) = \int \mathrm{d}x_{il}\, m_{il\to\mu l}(t,x_{il})\, x^2_{il}\,, \tag{34}$$

$$r_{\mu i\to\mu l}(t) = \sqrt{N}\int \mathrm{d}F_{\mu i}\, n_{\mu i\to\mu l}(t,F_{\mu i})\, F_{\mu i}\,, \tag{35}$$

$$s_{\mu i\to\mu l}(t) + r^2_{\mu i\to\mu l}(t) = N\int \mathrm{d}F_{il}\, n_{\mu i\to\mu l}(t,F_{\mu i})\, F^2_{\mu i}\,. \tag{36}$$

Notice that the factors $\sqrt{N}$ in the definition of r, and N in the definition of s, have been introduced in order to ensure that all the messages a, c, r, s are of order $O(1)$ in the thermodynamic limit.

Using this scaling, we shall now show that the BP equations can be simplified in the thermodynamic limit, and that they can actually be written as a closed set of equations involving only messages a, c, r, s. Our general aim is to design an algorithm which in some region of parameters will asymptotically match the performance of the exact (intractable) Bayes-optimal inference. Belief propagation provides such an algorithm, but in order to make it really tractable in practice, writing it in terms of the messages a, c, r, s is crucial, provided one is careful not to loose any terms in the asymptotic analysis of the thermodynamic limit to leading order.

Let us define the Fourrier transform of the output function

$$\hat{P}_{\mathrm{out}}(y,\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} P_{\mathrm{out}}(y|z)e^{-i\xi z}\mathrm{d}z\,, \tag{37}$$

to rewrite the update equation for message $\tilde{m}_{\mu l\to il}(t,x_{il})$ as

$$\tilde{m}_{\mu l\to il}(t,x_{il}) = \frac{1}{\sqrt{2\pi}\mathcal{Z}_{\mu l\to il}}\int \mathrm{d}\xi\, \hat{P}_{\mathrm{out}}(y_{\mu l},\xi)\int \mathrm{d}F_{\mu i}\, e^{i\xi F_{\mu i}x_{il}} n_{\mu i\to\mu l}(t,F_{\mu i}) \prod_{j(\neq i)}^{N}\left[\int \mathrm{d}x_{jl}\mathrm{d}F_{\mu j}\, e^{i\xi F_{\mu j}x_{jl}} n_{\mu j\to\mu l}(t,F_{\mu k}) m_{jl\to\mu l}(t,x_{jl})\right]. \tag{38}$$

In order to perform the integral in the square-bracket we recall that the elements of matrix $F = O(1/\sqrt{N})$ and hence we can expand the exponential to second order, use definitions (68-71) and re-exponentiate the result without loosing any leading order terms in $\tilde{m}_{\mu l\to il}(t,x_{il})$. The whole square bracket then becomes

$$\exp\left\{i\frac{\xi}{\sqrt{N}}r_{\mu j\to\mu l}(t)a_{jl\to\mu l}(t) - \frac{\xi^2}{2N}\left[s_{\mu j\to\mu l}(t)c_{jl\to\mu l}(t) + s_{\mu j\to\mu l}(t)a^2_{jl\to\mu l}(t) + r^2_{\mu j\to\mu l}(t)c_{jl\to\mu l}(t)\right]\right\}. \tag{39}$$

Next we perform the integral over variable ξ which is simply a Gaussian integral. This gives for the message

$$\tilde{m}_{\mu l\to il}(t,x_{il}) = \frac{1}{\mathcal{Z}_{\mu l\to il}}\int \mathrm{d}F_{\mu i}\, n_{\mu i\to\mu l}(t,F_{\mu i})\frac{1}{\sqrt{2\pi V_{\mu il}}}\int \mathrm{d}z\, P_{\mathrm{out}}(y_{\mu l}|z)e^{-\frac{(z-F_{\mu i}x_{il}-\omega_{\mu il})^2}{2V_{\mu il}}}\,, \tag{40}$$

where we introduced auxiliary variables that are both of order $O(1)$

$$V_{\mu i l} \equiv \frac{1}{N}\sum_{j(\neq i)}^{N}\left[s_{\mu j\to\mu l}(t)c_{jl\to\mu l}(t)+s_{\mu j\to\mu l}(t)a^2_{jl\to\mu l}(t)+r^2_{\mu j\to\mu l}(t)c_{jl\to\mu l}(t)\right], \tag{41}$$

$$\omega_{\mu i l} \equiv \frac{1}{\sqrt{N}}\sum_{j(\neq i)}^{N} r_{\mu j\to\mu l}(t)a_{jl\to\mu l}(t)\,. \tag{42}$$

The last integral to be performed in (40) is the one over the matrix element $F_{\mu i}$. Using again the fact that $F_{\mu i} = O(1/\sqrt{N})$, we expand the exponential in which $F_{\mu i}$ appears to second order and perform the integration to obtain

$$\tilde{m}_{\mu l\to il}(t,x_{il}) = \frac{1}{\mathcal{Z}_{\mu l\to il}}\frac{1}{\sqrt{2\pi V_{\mu il}}}\int \mathrm{d}z\, P_{\rm out}(y_{\mu l}|z)e^{-\frac{(z-\omega_{\mu il})^2}{2V_{\mu il}}} \left\{1+\frac{z-\omega_{\mu il}}{\sqrt{N}V_{\mu il}}r_{\mu i\to\mu l}(t)x_{il}-\frac{x^2_{il}}{2N}\left[\frac{1}{V_{\mu il}}-\frac{(z-\omega_{\mu il})^2}{V^2_{\mu il}}\right]\left[r^2_{\mu i\to\mu l}(t)+s_{\mu i\to\mu l}(t)\right]\right\}. \tag{43}$$

Following the notation of [47] we now define the output-function as

$$g_{\rm out}(\omega,y,V) \equiv \frac{\int \mathrm{d}z P_{\rm out}(y|z)\,(z-\omega)\,e^{-\frac{(z-\omega)^2}{2V}}}{V\int \mathrm{d}z P_{\rm out}(y|z)e^{-\frac{(z-\omega)^2}{2V}}}\,. \tag{44}$$

The following useful identity holds for the average of $(z-\omega)^2/V^2$ in the above measure

$$\frac{\int \mathrm{d}z P_{\rm out}(y|z)\,(z-\omega)^2\,e^{-\frac{(z-\omega)^2}{2V}}}{V^2\int \mathrm{d}z P_{\rm out}(y|z)e^{-\frac{(z-\omega)^2}{2V}}} = \frac{1}{V}+\partial_\omega g_{\rm out}(\omega,y,V)+g^2_{\rm out}(\omega,y,V)\,. \tag{45}$$

With this definition, and re-exponentiating the x_{il}-dependent terms in (43) while keeping all the leading order terms, we obtain finally that $\tilde{m}_{\mu l\to il}(t,x_{il})$ is a Gaussian probability distribution

$$\tilde{m}_{\mu l\to il}(t,x_{il}) = \sqrt{\frac{A^t_{\mu l\to il}}{2\pi N}}e^{-\frac{x^2_{il}}{2N}A^t_{\mu l\to il}+B^t_{\mu l\to il}\frac{x_{il}}{\sqrt{N}}-\frac{(B^t_{\mu l\to il})^2}{2A^t_{\mu l\to il}}} \tag{46}$$

with

$$B^t_{\mu l\to il} = g_{\rm out}(\omega_{\mu il},y_{\mu l},V_{\mu il})\,r_{\mu i\to\mu l}(t)\,, \tag{47}$$

$$A^t_{\mu l\to il} = -\partial_\omega g_{\rm out}(\omega_{\mu il},y_{\mu l},V_{\mu il})\,\left[r^2_{\mu i\to\mu l}(t)+s_{\mu i\to\mu l}(t)\right]-g^2_{\rm out}(\omega_{\mu il},y_{\mu l},V_{\mu il})\,s_{\mu i\to\mu l}(t)\,. \tag{48}$$

In a completely analogous way we obtain that the message $\tilde{n}_{\mu l\to\mu i}(t,F_{\mu i})$ is also a Gaussian distribution

$$\tilde{n}_{\mu l\to\mu i}(t,F_{\mu i}) = \sqrt{\frac{S^t_{\mu l\to\mu i}}{2\pi}}e^{-\frac{F^2_{\mu i}}{2}S^t_{\mu l\to\mu i}+R^t_{\mu l\to\mu i}F_{\mu i}-\frac{(R^t_{\mu l\to\mu i})^2}{2S^t_{\mu l\to\mu i}}} \tag{49}$$

with

$$R^t_{\mu l\to\mu i} = g_{\rm out}(\omega_{\mu il},y_{\mu l},V_{\mu il})\,a_{il\to\mu l}(t)\,, \tag{50}$$

$$S^t_{\mu l\to\mu i} = -\partial_\omega g_{\rm out}(\omega_{\mu il},y_{\mu l},V_{\mu il})\,\left[c_{il\to\mu l}(t)+a^2_{il\to\mu l}(t)\right]-g^2_{\rm out}(\omega_{\mu il},y_{\mu l},V_{\mu il})\,c_{il\to\mu l}(t)\,. \tag{51}$$

At this point we follow closely the derivation of AMP from [23] and define the probability distributions

$$\mathcal{M}_X(\Sigma,T,x) = \frac{1}{\hat{Z}_X(\Sigma,T)}P_X(x)\frac{1}{\sqrt{2\pi\Sigma}}e^{-\frac{(x-T)^2}{2\Sigma}}\,, \tag{52}$$

$$\mathcal{M}_F(Z,W,F) = \frac{1}{\hat{Z}_F(Z,W)}P_F(F)\frac{1}{\sqrt{2\pi Z}}e^{-\frac{(\sqrt{N}F-W)^2}{2Z}}\,, \tag{53}$$

where $\hat{Z}_X(\Sigma,T)$, and $\hat{Z}_F(Z,W)$ are normalizations. We define the average and variance of $\mathcal{M}_X$ and $\mathcal{M}_F$ as

$$f_a(\Sigma,T) \equiv \int \mathrm{d}x\, x\, \mathcal{M}_X(\Sigma,T,x)\,, \qquad f_c(\Sigma,T) \equiv \int \mathrm{d}x\, x^2\, \mathcal{M}_X(\Sigma,T,x) - f_a^2(\Sigma,T)\,, \tag{54}$$

$$f_r(Z,W) \equiv \sqrt{N}\int \mathrm{d}F\, F\, \mathcal{M}_F(Z,W,F)\,, \qquad f_s(Z,W) \equiv N\int \mathrm{d}F\, F^2\, \mathcal{M}_F(Z,W,F) - f_r^2(Z,W)\,. \tag{55}$$

These are the input auxiliary function of [47]. It is instrumental to notice that

$$f_a(\Sigma,T) = T + \Sigma\frac{\mathrm{d}}{\mathrm{d}T}\log\hat{Z}_X(\Sigma,T)\,, \qquad f_c(\Sigma,T) = \Sigma\frac{\mathrm{d}}{\mathrm{d}T}f_a(\Sigma,T)\,. \tag{56}$$

and analogously for f_r and f_s

$$f_r(Z,W) = W + Z\frac{\mathrm{d}}{\mathrm{d}W}\log\hat{Z}_F(Z,W)\,, \qquad f_s(Z,W) = Z\frac{\mathrm{d}}{\mathrm{d}W}f_r(Z,W)\,. \tag{57}$$

With these definition we obtain from (29,30) using (46,49) that

$$a_{il\to\mu l}(t+1) = f_a\left(\frac{N}{\sum_{\nu(\neq\mu)} A^t_{\nu l\to il}}, \frac{\sqrt{N}\sum_{\nu(\neq\mu)} B^t_{\nu l\to il}}{\sum_{\nu(\neq\mu)} A^t_{\nu l\to il}}\right), \tag{58}$$

$$c_{il\to\mu l}(t+1) = f_c\left(\frac{N}{\sum_{\nu(\neq\mu)} A^t_{\nu l\to il}}, \frac{\sqrt{N}\sum_{\nu(\neq\mu)} B^t_{\nu l\to il}}{\sum_{\nu(\neq\mu)} A^t_{\nu l\to il}}\right), \tag{59}$$

$$r_{\mu i\to\mu l}(t+1) = f_r\left(\frac{N}{(\sum_{n(\neq l)} S^t_{\mu n\to\mu i}}, \frac{\sqrt{N}\sum_{n(\neq l)} R^t_{\mu n\to\mu i}}{\sum_{n(\neq l)} S^t_{\mu n\to\mu i}}\right), \tag{60}$$

$$s_{\mu i\to\mu l}(t+1) = f_s\left(\frac{N}{\sum_{n(\neq l)} S^t_{\mu n\to\mu i}}, \frac{\sqrt{N}\sum_{n(\neq l)} R^t_{\mu n\to\mu i}}{\sum_{n(\neq l)} S^t_{\mu n\to\mu i}}\right), \tag{61}$$

It is clear from the above expressions that all the messages a, c, r, s and A, C, R, S scale as $O(1)$ in the thermodynamic limit. For instance, as A are positive, the quantity $\sum_{\nu(\neq\mu)} A^t_{\nu l\to il}$ is $O(1)$. On the other hand, the message $r_{\mu i\to\mu l}(t)$ is an estimate of $\sqrt{N}F_{\mu i}$; this estimate is $O(1)$, but a sum like $(1/\sqrt{N})\sum_{\nu(\neq\mu)} r_{\nu i\to\nu l}(t)$ is an estimate of $\sum_\nu F_{\nu i}$. As P_F has mean variance of order $O(1/N)$, this sum is actually of $O(1)$. The same argument suggests that $(1/\sqrt{N})\sum_{\nu(\neq\mu)} B^t_{\nu l\to il}$ is $O(1)$. Recalling eqs. (47,48) and (50,51), we have derived that in the thermodynamic limit the general belief propagation equations simplify into a closed set of equations in the messages which are the means and variances a, r, c, s defined in (33-36). To iterate this message passing algorithm we initialize as

$$a_{il\to\mu l}(0) = \int \mathrm{d}x\, x P_X^{il}(x)\,, \tag{62}$$

$$c_{il\to\mu l}(0) = \int \mathrm{d}x\, x^2 P_X^{il}(x) - a^2_{il\to\mu l}(0)\,, \tag{63}$$

$$r_{\mu i\to\mu l}(0) = \sqrt{N}\int \mathrm{d}F\, F P_F^{\mu l}(F)\,, \tag{64}$$

$$s_{\mu i\to\mu l}(0) = N\int \mathrm{d}F\, F^2 P_F^{\mu l}(F) - r^2_{\mu i\to\mu l}(0)\,, \tag{65}$$

then we compute $V_{\mu il}$ and $\omega_{\mu il}$ from (41-42), then we compute B^t, A^t, R^t and S^t according to (47-48) and (50-51) using definition of $g_{\rm out}$ (44). Finally we update the messages according to (58-61) and iterate. Notice, however, that we work with $O(N^3)$ messages, each of them takes N steps to update, and hence the computational complexity of this algorithm is relatively high. In the next section we will write a simplification that reduces this complexity.

From the fixed point of the belief propagation equations one can also compute the approximated marginal probabilities of the posterior, defined as

$$m_{il}(t+1, x_{il}) = \frac{1}{\mathcal{Z}_{il}} P_X(x_{il}) \prod_\nu^M \tilde{m}_{\nu l\to il}(t, x_{il})\,, \tag{66}$$

$$n_{\mu i}(t+1, F_{\mu i}) = \frac{1}{\mathcal{Z}_{\mu i}} P_F(F_{\mu i}) \prod_n^P \tilde{n}_{\mu n\to\mu i}(t, F_{\mu i})\,, \tag{67}$$

One again defines the mean and variance of these two messages, $a_{il}(t+1)$, $c_{il}(t+1)$, $r_{\mu i}(t+1)$ and $s_{\mu i}(t+1)$ analogously to (68-71):

$$a_{il}(t+1) = \int \mathrm{d}x_{il}\, m_{il}(t+1, x_{il})\, x_{il}\,, \tag{68}$$

$$c_{il}(t+1) + a_{il}^2(t+1) = \int \mathrm{d}x_{il}\, m_{il}(t+1, x_{il})\, x_{il}^2\,, \tag{69}$$

$$r_{\mu i}(t+1) = \sqrt{N} \int \mathrm{d}F_{\mu i}\, n_{\mu i}(t+1, F_{\mu i})\, F_{\mu i}\,, \tag{70}$$

$$s_{\mu i}(t+1) + r_{\mu i}^2(t+1) = N \int \mathrm{d}F_{il}\, n_{\mu i}(t+1, F_{\mu i})\, F_{\mu i}^2\,. \tag{71}$$

Those quantities are then expressed as

$$a_{il}(t+1) = f_a\left(\frac{N}{\sum_\nu A^t_{\nu l\to il}}, \frac{\sqrt{N}\sum_\nu B^t_{\nu l\to il}}{\sum_\nu A^t_{\nu l\to il}}\right), \tag{72}$$

$$c_{il}(t+1) = f_c\left(\frac{N}{\sum_\nu A^t_{\nu l\to il}}, \frac{\sqrt{N}\sum_\nu B^t_{\nu l\to il}}{\sum_\nu A^t_{\nu l\to il}}\right), \tag{73}$$

$$r_{\mu i}(t+1) = f_r\left(\frac{N}{\sum_n S^t_{\mu n\to \mu i}}, \frac{\sqrt{N}\sum_n R^t_{\mu n\to \mu i}}{\sum_n S^t_{\mu n\to \mu i}}\right), \tag{74}$$

$$s_{\mu i}(t+1) = f_s\left(\frac{N}{\sum_n S^t_{\mu n\to \mu i}}, \frac{\sqrt{N}\sum_n R^t_{\mu n\to \mu i}}{\sum_n S^t_{\mu n\to \mu i}}\right). \tag{75}$$

B. GAMP for matrix factorization

The message-passing form the AMP algorithm for matrix factorization derived in the previous section uses $2NMP$ messages, one between each variable component (il) and (μi) and each measurement (μl), in each iteration. In fact, exploiting again the simplifications which take place in the thermodynamic limit, always within the assumption that the elements of the matrix F scale as $O(1/\sqrt{N})$, it is possible to rewrite and close the BP equations in terms of only $2N(P+M)$ messages. In statistical physics terms, the resulting equations correspond to the Thouless-Anderson-Palmer equations (TAP) [49] used in the study of spin glasses. In the thermodynamic limit, these are asymptotically equivalent to the BP equations. Going from BP to TAP is, in the compressed sensing literature, the step to go from the rBP [50] to the AMP [9] algorithm. Let us now show how to take this step for the present problem of matrix factorization.

In the thermodynamic limit, it is clear from (58-61) that the messages $a_{il\to\mu l}$, $v_{il\to\mu l}$ and $r_{\mu i\to\mu l}$, $s_{\mu i\to\mu l}$ are nearly independent of μl. For instance in the equation giving $a_{il\to\mu l}$, the only dependence on μ is through the fact that the sum over ν avoids the value $\nu = \mu$. But this is one term in M, and therefore one might expect that this term is negligible. However, one must be careful to keep the leading correction, called in spin-glass theory the "Onsager reaction terms". This is what we now explain.

Let us define the following variables all of order $O(1)$ on which we will close the equations

$$T^t_{il} = \frac{\sqrt{N}\sum_\nu B^t_{\nu l\to il}}{\sum_\nu A^t_{\nu l\to il}}\,, \qquad \Sigma^t_{il} = \frac{N}{\sum_\nu A^t_{\nu l\to il}}\,, \tag{76}$$

$$W^t_{\mu i} = \frac{\sqrt{N}\sum_n R^t_{\mu n\to \mu i}}{\sum_n S^t_{\mu n\to \mu i}}\,, \qquad Z^t_{\mu i} = \frac{N}{\sum_n S^t_{\mu n\to \mu i}}\,, \tag{77}$$

$$V^t_{\mu l} = \frac{1}{N}\sum_j [c_{jl\to\mu l}(t) s_{\mu j\to\mu l}(t) + c_{jl\to\mu l}(t) r^2_{\mu j\to\mu l}(t) + a^2_{jl\to\mu l}(t) s_{\mu j\to\mu l}(t)]\,, \tag{78}$$

$$\omega^t_{\mu l} = \frac{1}{\sqrt{N}}\sum_j a_{jl\to\mu l}(t) r_{\mu j\to\mu l}(t)\,. \tag{79}$$

To keep track of all the Onsager terms that will influence the leading order of the final equations we notice that

$$\begin{aligned} a_{il\to\mu l}(t+1) &= f_a\left(\frac{N}{\sum_\nu A^t_{\nu l\to il} - A^t_{\mu l\to il}}, \frac{\sqrt{N}\sum_\nu B^t_{\nu l\to il} - \sqrt{N}B^t_{\mu l\to il}}{\sum_\nu A^t_{\nu l\to il} - A^t_{\mu l\to il}}\right), \\ &= a_{il}(t+1) - \frac{1}{\sqrt{N}} B^t_{\mu l\to il}\Sigma^t_{il}\frac{\partial f_a}{\partial T}\left(\Sigma^t_{il}, T^t_{il}\right) + O\left(1/N\right), \\ &= a_{il}(t+1) - \frac{1}{\sqrt{N}} g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) r_{\mu i}(t) c_{il}(t+1) + O\left(1/N\right). \end{aligned} \tag{80}$$

Similarly:

$$r_{\mu i\to\mu l}(t+1) = r_{\mu i}(t+1) - \frac{1}{\sqrt{N}} g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) a_{il}(t) s_{\mu i}(t+1) + O\left(1/N\right), \tag{81}$$

$$c_{il\to\mu l}(t+1) = c_{il}(t+1) + O\left(1/\sqrt{N}\right), \qquad s_{\mu i\to\mu l}(t+1) = s_{\mu i}(t+1) + O\left(1/\sqrt{N}\right), \tag{82}$$

$$g_{\rm out}(\omega^t_{\mu i l}, y_{\mu l}, V^t_{\mu i l}) = g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) - \frac{1}{\sqrt{N}} r_{\mu i}(t) a_{il}(t) \partial_\omega g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) + O(1/N), \tag{83}$$

From these expansions we obtain the GAMP algorithm for matrix factorization

$$V^t_{\mu l} = \frac{1}{N}\sum_j \left[c_{jl}(t) s_{\mu j}(t) + c_{jl}(t) r^2_{\mu j}(t) + a^2_{jl}(t) s_{\mu j}(t)\right], \tag{84}$$

$$\omega^t_{\mu l} = \frac{1}{\sqrt{N}}\sum_j a_{jl}(t) r_{\mu j}(t) - g_{\rm out}(\omega^{t-1}_{\mu l}, y_{\mu l}, V^{t-1}_{\mu l}) \frac{1}{N}\sum_j \left[r_{\mu j}(t) r_{\mu j}(t-1) c_{jl}(t) + a_{jl}(t) a_{jl}(t-1) s_{\mu j}(t)\right], \tag{85}$$

$$(\Sigma^t_{il})^{-1} = \frac{1}{N}\sum_\mu \left\{-\partial_\omega g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) \left[r^2_{\mu i}(t) + s_{\mu i}(t)\right] - g^2_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) s_{\mu i}(t)\right\}, \tag{86}$$

$$\begin{aligned} T^t_{il} = \Sigma^t_{il}\Big\{ &\frac{1}{\sqrt{N}}\sum_\mu g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) r_{\mu i}(t) - a_{il}(t)\frac{1}{N}\sum_\mu r^2_{\mu i}(t)\partial_\omega g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) \\ &- a_{il}(t-1)\frac{1}{N}\sum_\mu s_{\mu i}(t) g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) g_{\rm out}(\omega^{t-1}_{\mu l}, y_{\mu l}, V^{t-1}_{\mu l})\Big\}, \end{aligned} \tag{87}$$

$$(Z^t_{\mu i})^{-1} = \frac{1}{N}\sum_l \left\{-\partial_\omega g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) \left[a^2_{il}(t) + c_{il}(t)\right] - g^2_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) c_{il}(t)\right\}, \tag{88}$$

$$\begin{aligned} W^t_{\mu i} = Z^t_{il}\Big\{ &\frac{1}{\sqrt{N}}\sum_l g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) a_{il}(t) - r_{\mu i}(t)\frac{1}{N}\sum_l a^2_{il}(t)\partial_\omega g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) \\ &- r_{\mu i}(t-1)\frac{1}{N}\sum_l c_{il}(t) g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) g_{\rm out}(\omega^{t-1}_{\mu l}, y_{\mu l}, V^{t-1}_{\mu l})\Big\}, \end{aligned} \tag{89}$$

$$a_{il}(t+1) = f_a(\Sigma^t_{il}, T^t_{il}), \qquad c_{il}(t+1) = f_c(\Sigma^t_{il}, T^t_{il}), \tag{90}$$

$$r_{\mu i}(t+1) = f_r(Z^t_{\mu i}, W^t_{\mu i}), \qquad s_{\mu i}(t+1) = f_s(Z^t_{\mu i}, W^t_{i\mu}). \tag{91}$$

The initial condition for iterations are

$$a_{il}(t=0) = \int {\rm d}x\, x P^{il}_X(x), \qquad c_{il}(t=0) = \int {\rm d}x\, x^2 P^{il}_X(x) - a^2_{il\to\mu l}(t=0), \tag{92}$$

$$r_{\mu i}(t=0) = \sqrt{N}\int {\rm d}F\, F P^{\mu i}_F(F), \qquad s_{\mu i}(t=0) = N\int {\rm d}F\, F^2 P^{\mu i}_F(F) - r^2_{\mu i}(t=0). \tag{93}$$

In order to compute $\omega^{t=0}$, $T^{t=0}$ and $W^{t=0}$ use the above equations as if $r_{\mu i}(-1) = 0$ and $a_{il}(-1) = 0$.

The interpretation of the terms in the GAMP for matrix factorization is the following: $\omega^t_{\mu l}$ is the mean of the current estimate of $z_{\mu l} = \sum_i F_{\mu i} x_{il}$ and $V^t_{\mu l}$ is the variance of that estimate; T_{il} and Σ_{il} is the mean and variance of the current estimate of x_{il} without taking into account the prior information of x_{il}; the parameters a_{il} and c_{il} are then the mean and variance of the current estimate of x_{il} with the prior information taken into account. Analogously for $W_{\mu i}$ and $Z_{\mu i}$ being the mean and variance of the estimate for $F_{\mu i}$ before the prior is taken into account, and $r_{\mu i}$ with $s_{\mu i}$ are the mean and variance once the prior information was accounted for.

A reader familiar with the AMP and GAMP algorithm for compressed sensing [9, 23, 47] will recognize that the above equations indeed reduce to the compressed sensing GAMP of [47] when one sets $r_{\mu i}(t) = \sqrt{N} F_{\mu i}$ and $s_{\mu i} = 0$.

The above algorithm is closely related to the BiG-AMP of [15]. There are however three differences between our algorithm and BiG-AMP:

1. We find a $s_{\mu i}$-dependent term in the expression (86) for Σ_{il} which is not present in BiG-AMP.
2. Similarly, we find a c_{il}-dependent term in the expression (88) for $Z_{\mu i}$ which is not present in BiG-AMP.
3. The time indices are slightly different.

Considering the last point, the fact of having different time indices during the iterations does not influence the fixed points, in which we are mainly interested. However, the use of correct time indices is crucial for the assumptions leading to the density evolution of this algorithm (that we derive in section IV A) to hold.

As for the missing terms in the BiG-AMP expressions of [15] for Σ_{il} and $Z_{\mu i}$, they have a more serious effect as they can change the fixed point. To the best of our understanding, these terms have been neglected in [15], while they should be kept. It seems to us that some of the leading order terms are missing in eqs. (15-16) from [15].

C. Simplifications due to the Nishimori conditions

In the previous section we derived the GAMP algorithm for matrix factorization, eqs. (84-91). This algorithm can in principle be used for any set of matrices F and X. If iterated in the form derived in Section II B it often shows problems of convergence.

There are ways to slightly improve the convergence of the above algorithm in a wide range of applications by a number of empirical methods suggested in [15]. We will focus here mainly on the particular case when matrices X and F were indeed generated from the separable probability distributions $P_F(F)$ and $P_X(X)$ eqs. (2-3). In this case the belief propagation is a proxy for the optimal Bayes inference algorithms and a number of properties described in section I B hold. In analogy with fundamental works on spin glasses [18, 19] we called these properties the Nishimori conditions.

The Nishimori conditions/identities hold and the system is *on the Nishimori line* of parameters when the matrices X and F are indeed generated from the assumed probability distributions (2-3). This means that one is using the correct priors on F and X in the reconstruction process. In the limit $N \to \infty$ and thanks to self-averaging we then have on the Nishimori line at every iteration step t

$$\frac{1}{PM} \sum_{\mu,l} \frac{\int \mathrm{d}z P_{\rm out}(y_{\mu l}|z)(z-\omega_{\mu l}^t)^2 e^{-\frac{(z-\omega_{\mu l}^t)^2}{2V_{\mu l}^t}}}{\int \mathrm{d}z P_{\rm out}(y_{\mu l}|z) e^{-\frac{(z-\omega_{\mu l}^t)^2}{2V_{\mu l}^t}}} = \frac{1}{MP} \sum_{\mu,l} V_{\mu l}^t \,. \tag{94}$$

The meaning of this condition is that the mean squared error of the current estimate of $Z = FX$ computed from the current estimates of variances $V_{\mu l}^t$ is equal to the mean squared difference between the true Z and its current estimate $\omega_{\mu l}^t$. Using the above expression and eq. (45) we obtain an identity

$$-\frac{1}{MP} \sum_{\mu,l} \partial_\omega g_{\rm out}(\omega_{\mu l}^t, y_{\mu l}, V_{\mu l}^t) = \frac{1}{MP} \sum_{\mu,l} g_{\rm out}^2(\omega_{\mu l}^t, y_{\mu l}, V_{\mu l}^t) \,. \tag{95}$$

The above identity holds also if the sum is only over μ or only over l. Finally using the conditional independence assumed in BP between the incoming messages we get also

$$-\frac{1}{M} \sum_{\mu} \partial_\omega g_{\rm out}(\omega_{\mu l}^t, y_{\mu l}, V_{\mu l}^t) s_{\mu i}(t) = \frac{1}{M} \sum_{\mu} g_{\rm out}^2(\omega_{\mu l}^t, y_{\mu l}, V_{\mu l}^t) s_{\mu i}(t) \,. \tag{96}$$

Under this condition we can simplify considerably the expressions for Σ_{il}^t and $Z_{\mu i}^t$ and get

$$(\Sigma_{il}^t)^{-1} = -\frac{1}{N} \sum_{\mu} \partial_\omega g_{\rm out}(\omega_{\mu l}^t, y_{\mu l}, V_{\mu l}^t)\, r_{\mu i}^2(t) \,, \tag{97}$$

$$(Z_{\mu i}^t)^{-1} = -\frac{1}{N} \sum_{l} \partial_\omega g_{\rm out}(\omega_{\mu l}^t, y_{\mu l}, V_{\mu l}^t)\, a_{il}^2(t) \,. \tag{98}$$

Note that the r.h.s. of the two above equations is always strictly positive, which is reassuring given these expressions play the role of a variance of a probability distribution. Note also that the BiG-AMP algorithm of [15] uses expressions (97,98) instead of (86,88), however, without mentioning the reason.

D. Simplification for matrices with random entries

Relying on the definitions of order parameters (146-147) and using part of the results on section IV A we can write a version of the GAMP for matrix factorization that is in the leading order equivalent to (84-91) for matrices X and F having iid elements.

Let us define the analog of $\hat{\chi}^t$ and $\hat{q}^t$ (155-156) as empirical means of the corresponding functions

$$\tilde{\chi}^t \equiv -\frac{1}{MP}\sum_{\mu,l}\partial_\omega g_{\rm out}(\omega^t_{\mu i l}, y_{\mu l}, V^t_{\mu i l})\,, \tag{99}$$

$$\tilde{q}^t \equiv \frac{1}{MP}\sum_{\mu,l} g^2_{\rm out}(\omega^t_{\mu i l}, y_{\mu l}, V^t_{\mu i l})\,. \tag{100}$$

Using the same reasoning as in section IV A we realize that in the leading order quantities V_{il}, Σ_{il} and $Z_{\mu i}$ do not depend on their indices. We have

$$V^t = Q^t_F Q^t_x - q^t_F q^t_x\,, \tag{101}$$

$$(\Sigma^t)^{-1} = \alpha Q^t_F \tilde{\chi}^t - \alpha(Q^t_F - q^t_F)\tilde{q}^t\,, \tag{102}$$

$$(Z^t)^{-1} = \pi Q^t_x \tilde{\chi}^t - \pi(Q^t_x - q^t_x)\tilde{q}^t\,. \tag{103}$$

These three equations can hence replace (84), (86) and (88) in GAMP. Furthermore, if we focus on the fixed point and hence disregard some of the time indices eqs. (85), (87) and (89) can be simplified as

$$\omega^t_{\mu l} = \frac{1}{\sqrt{N}}\sum_j a_{jl}(t) r_{\mu j}(t) - g_{\rm out}(\omega^{t-1}_{\mu l}, y_{\mu l}, V^{t-1})(Q^t_x q^t_F + q^t_F Q^t_x - 2q^t_F q^t_x)\,, \tag{104}$$

$$T^t_{il} = \Sigma^t\left\{\frac{1}{\sqrt{N}}\sum_\mu g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t) r_{\mu i}(t) + \alpha a_{il}(t)\tilde{\chi}^t\, q^t_F - \alpha a_{il}(t)(Q^t_F - q^t_F)\tilde{q}^t\right\}\,, \tag{105}$$

$$W^t_{\mu i} = Z^t\left\{\frac{1}{\sqrt{N}}\sum_l g_{\rm out}(\omega^t_{\mu l}, y_{\mu l}, V^t) a_{il}(t) + \pi r_{\mu i}(t)\tilde{\chi}^t\, q^t_x - \pi r_{\mu i}(t)(Q^t_x - q^t_x)\tilde{q}^t\right\}\,. \tag{106}$$

Under the Nishimori condition from previous section (95) we have moreover $\tilde{\chi}^t = \tilde{q}^t$ and hence we can use only one of those parameters computed either from (99) or from (100). The equations then further simplify to strictly positive expressions for the variance-parameters

$$(\Sigma^t)^{-1} = \alpha q^t_F \tilde{q}^t\,, \qquad (Z^t)^{-1} = \pi q^t_x \tilde{q}^t\,. \tag{107}$$

The set of eqs. (90-91), (101), (104-107) was presented for the simple output channel with white noise in [13].

We want to stress here that all these simplifications take place for any output channel $P_{\rm out}(y|z)$. In contrast with the "uniform variance" approximation of [15] the above result does not mean that the variances c_{il} and $s_{\mu i}$ are independent in the leading order on their indices. On the contrary, these variances depend on their indices even in the simplest case of GAMP when the matrix $F_{\mu i}$ is known, i.e. for the compressed sensing problem.

III. THE BETHE FREE ENTROPY

The fixed point of the belief propagation equations or its AMP version can be used to estimate the posterior likelihood, i.e. the normalization $\mathcal{Z}(Y)$ of the posterior probability (4). The logarithm of this normalization is called the Bethe free entropy in statistical physics [51]. Negative logarithm of the normalization is called the free energy, in physics there is usually a temperature associated to the free energy. Bethe free entropy is computed from the fixed point of the BP equations (29-32) as [11, 51]:

$$\Phi^{\rm Bethe} = \sum_{\mu l}\log\mathcal{Z}^{\mu l} + \sum_{\mu i}\log\mathcal{Z}^{\mu i} + \sum_{il}\log\mathcal{Z}^{il} - \sum_{\mu i l}\log\mathcal{Z}^{il,\mu l} - \sum_{\mu i l}\log\mathcal{Z}^{\mu i,\mu l}\,, \tag{108}$$

where the five contributions are

$$\mathcal{Z}^{\mu i} = \int \mathrm{d}F_{\mu i}\, P_F(F_{\mu i}) \prod_{l=1}^{P} \tilde{n}_{\mu l\to\mu i}(F_{\mu i})\,, \tag{109}$$

$$\mathcal{Z}^{il} = \int \mathrm{d}x_{il}\, P_X(x_{il}) \prod_{\mu=1}^{M} \tilde{m}_{\mu l\to il}(x_{il})\,, \tag{110}$$

$$\mathcal{Z}^{\mu l} = \int \prod_{i=1}^{N} \mathrm{d}x_{il} \prod_{k=1}^{N} \mathrm{d}F_{\mu k} P_{\rm out}(y_{\mu l}|\sum_{k=1}^{N} F_{\mu k}x_{kl}) \prod_{i=1}^{N} m_{il\to\mu l}(x_{il}) \prod_{k=1}^{N} n_{\mu k\to\mu l}(x_{il})\,, \tag{111}$$

$$\mathcal{Z}^{il,\mu l} = \int \mathrm{d}x_{il}\, m_{il\to\mu l}(x_{il})\, \tilde{m}_{\mu l\to il}(x_{il})\,, \tag{112}$$

$$\mathcal{Z}^{\mu i,\mu l} = \int \mathrm{d}F_{\mu i}\, n_{\mu i\to\mu l}(F_{\mu i})\, \tilde{n}_{\mu l\to\mu i}(F_{\mu i})\,. \tag{113}$$

The derivatives of this expression for $\Phi^{\rm Bethe}$ with respect to the messages give back the full BP equations of (29-32). In this general form, the computation of $\Phi^{\rm Bethe}$ for the present problem is not of practical interest, and it is thus very useful to carry out the same steps that we did in Section II A in order to obtain a more tractable form of $\Phi^{\rm Bethe}$ that is asymptotically equivalent to (108) in the thermodynamic limit, using the set of AMP message passing equations (41-42), (47-48), (50-51), and (58-61). The result is:

$$\Phi^{\rm Bethe} = \sum_{\mu l} \log \mathcal{Z}^{\mu l} + \sum_{\mu i} \log \mathcal{X}^{\mu i} + \sum_{il} \log \mathcal{X}^{il} + \sum_{\mu il} \log \frac{\mathcal{X}^{\mu i\to\mu l}}{\mathcal{X}^{\mu i}} + \sum_{\mu il} \log \frac{\mathcal{X}^{il\to\mu l}}{\mathcal{X}^{il}}\,, \tag{114}$$

with

$$\mathcal{Z}^{\mu l} = \int \mathrm{d}z \frac{e^{-\frac{(\omega_{\mu l}-z)^2}{2V_{\mu l}}}}{\sqrt{2\pi V_{\mu l}}} P_{\rm out}(y_{\mu l}|z)\,, \tag{115}$$

$$\mathcal{X}^{\mu i} = \int \mathrm{d}F_{\mu i} P_F(F_{\mu i}) e^{-\frac{NF_{\mu i}^2}{2\mathcal{Z}_{\mu i}} + \sqrt{N}F_{\mu i}\frac{W_{\mu i}}{\mathcal{Z}_{\mu i}}}\,, \tag{116}$$

$$\mathcal{X}^{il} = \int \mathrm{d}x_{il} P_X(x_{il}) e^{-\frac{x_{il}^2}{2\Sigma_{il}} + x_{il}\frac{T_{il}}{\Sigma_{il}}}\,, \tag{117}$$

$$\mathcal{X}^{\mu i\to\mu l} = \int \mathrm{d}F_{\mu i} P_F(F_{\mu i}) e^{-\frac{F_{\mu i}^2}{2}\sum_{n\neq l} S_{\mu n\to\mu i} + F_{\mu i}\sum_{n\neq l} R_{\mu n\to\mu i}}\,, \tag{118}$$

$$\mathcal{X}^{il\to\mu l} = \int \mathrm{d}x_{il} P_X(x_{il}) e^{-\frac{x_{il}^2}{2N}\sum_{\nu\neq\mu} A_{\nu l\to il} + \frac{x_{il}}{\sqrt{N}}\sum_{\nu\neq\mu} B_{\nu l\to il}}\,. \tag{119}$$

Finally we might want to express the free entropy using the fixed point of the GAMP eqs. (84-91). In order to do this we need to rewrite the last two terms in (114). Using an expansion in $1/N$ and keeping the leading order terms we get

$$\sum_{l} \log \frac{\mathcal{X}^{\mu i\to\mu l}}{\mathcal{X}^{\mu i}} = -\frac{W_{\mu i}}{\mathcal{Z}_{\mu i}} r_{\mu i} + \frac{1}{2\mathcal{Z}_{\mu i}}(s_{\mu i} + r_{\mu i}^2) + \frac{1}{2N} s_{\mu i} \sum_{l=1}^{P} g_{\rm out}^2(\omega_{\mu l}, y_{\mu l}, V_{\mu l}) a_{il}^2\,, \tag{120}$$

$$\sum_{\mu} \log \frac{\mathcal{X}^{il\to\mu l}}{\mathcal{X}^{il}} = -\frac{T_{il}}{\Sigma_{il}} a_{il} + \frac{1}{2\Sigma_{il}}(c_{il} + a_{il}^2) + \frac{1}{2N} c_{il} \sum_{\mu=1}^{M} g_{\rm out}^2(\omega_{\mu l}, y_{\mu l}, V_{\mu l}) r_{\mu i}^2\,. \tag{121}$$

We remind that the above expressions give the posterior likelihood given a *fixed point* on the GAMP equations.

To write the final formula in a more easily interpretable form we use the probability distributions $\mathcal{M}_F$ and $\mathcal{M}_X$ defined in (52-53) with normalizations

$$\hat{\mathcal{X}}^{\mu i} = \hat{Z}_F(\mathcal{Z}_{\mu i}, W_{\mu i})\sqrt{2\pi Z_{\mu i}} = \int \mathrm{d}F_{\mu i} P_F(F_{\mu i}) e^{-\frac{(\sqrt{N}F_{\mu i} - W_{\mu i})^2}{2\mathcal{Z}_{\mu i}}}\,, \tag{122}$$

$$\hat{\mathcal{X}}^{il} = \hat{Z}_X(\Sigma_{il}, T_{il})\sqrt{2\pi\Sigma_{il}} = \int \mathrm{d}x_{il} P_X(x_{il}) e^{-\frac{(x_{il}-T_{il})^2}{2\Sigma_{il}}}\,. \tag{123}$$

Putting all pieces together we find:

$$\Phi^{\rm Bethe} = \sum_{il}\left[\log\hat{\mathcal{X}}^{il}\left(T_{il},\Sigma_{il}\right)+\frac{c_{il}+(a_{il}-T_{il})^2}{2\Sigma_{il}}\right]+\sum_{\mu i}\left[\log\hat{\mathcal{X}}^{\mu i}\left(W_{\mu i},Z_{\mu i}\right)+\frac{s_{\mu i}+(r_{\mu i}-W_{\mu i})^2}{2Z_{\mu i}}\right]$$
$$+\sum_{\mu l}\left[\log Z^{\mu l}\left(\omega_{\mu l},V_{\mu l}\right)+\frac{1}{2N}\sum_{\mu i l}g_{\rm out}^2\left(\omega_{\mu l},V_{\mu l}\right)\left(s_{\mu i}a_{il}^2+r_{\mu i}^2c_{il}\right)\right]. \tag{124}$$

The above expression evaluated at the fixed point of the AMP algorithm hence gives the Bethe approximation to the log-likelihood. It is mainly use to decide which fixed point of AMP is better. Indeed, there are cases where there exist more than one AMP fixed point and it is the one with the largest Bethe entropy that corresponds asymptotically to the optimal Bayesian inference.

A. Fixed-point generating Bethe free entropy

Since the free entropy has a meaning only at the fixed point, we can transform it by using any of the fixed point identities verified by the BP messages. It is convenient to turn the free entropy into a form that allows to generate the fixed-point BP equations as a stationary point. This can be achieved by writing the Bethe free entropy eq. (124) as

$$\Phi^{\rm Bethe}_{\rm AMP}\left(\{T_{il}\},\{\Sigma_{il}\},\{W_{\mu i}\},\{Z_{\mu i}\},\{\omega_{\mu l}\},\{a_{il}\},\{c_{il}\},\{r_{\mu i}\},\{s_{\mu i}\}\right)=\sum_{il}\left[\log\hat{\mathcal{X}}^{il}\left(T_{il},\Sigma_{il}\right)+\frac{c_{il}+(a_{il}-T_{il})^2}{2\Sigma_{il}}\right]$$
$$+\sum_{\mu i}\left[\log\hat{\mathcal{X}}^{\mu i}\left(W_{\mu i},Z_{\mu i}\right)+\frac{s_{\mu i}+(r_{\mu i}-W_{\mu i})^2}{2Z_{\mu i}}\right]+\sum_{\mu l}\left[\log Z^{\mu l}\left(\omega_{\mu l},V_{\mu l}\right)+\frac{1}{2}\sum_{\mu l}\frac{(\omega_{\mu l}-\sum_i r_{\mu i}a_{il}/\sqrt{N})^2}{V_{\mu l}-\sum_i s_{\mu i}c_{il}/N}\right] \tag{125}$$
$$\text{with}\quad V_{\mu l}=\frac{1}{N}\sum_j[c_{jl}s_{\mu j}+c_{jl}r_{\mu j}^2+a_{jl}^2s_{\mu j}]\,.$$

In order to derive (125) from (124), we have substituted $g_{\rm out}^2$ by its fixed point expression, and imposed the values of the variance V. Under the present form, the Bethe free entropy satisfies the following theorem:

Theorem. (Bethe/AMP correspondance) *The fixed point of the AMP equations eqs. (84-91) are the stationary points of the cost function $\Phi^{\rm Bethe}_{\rm AMP}$ eq. (125).*

Proof. This can be checked explicitly by setting to zero the derivatives of $\Phi^{\rm Bethe}_{\rm AMP}$. Indeed, the derivatives with respect to T,Σ,W and Z yield

$$a_{il}=T_{il}+\Sigma_{il}\frac{\partial}{\partial T}\log\hat{\mathcal{X}}^{il}=f_a(\Sigma_{il},T_{il})\,, \tag{126}$$
$$c_{il}=\Sigma_{il}^2\frac{\partial}{\partial\Sigma_{il}}\log\hat{\mathcal{X}}^{il}-(a_{il}-T_{il})^2=f_c(\Sigma_{il},T_{il})\,, \tag{127}$$
$$r_{\mu i}=W_{\mu i}+Z_{\mu i}\frac{\partial}{\partial W}\log\hat{\mathcal{X}}^{\mu i}=f_r(Z_{\mu i},W_{\mu i})\,, \tag{128}$$
$$s_{\mu i}=Z_{\mu i}^2\frac{\partial}{\partial Z_{\mu i}}\log\hat{\mathcal{X}}^{\mu i}-(r_{\mu i}-W_{\mu i})^2=f_s(Z_{\mu i},W_{\mu i})\,. \tag{129}$$

Then, the stationarity with respect to ω can be expressed easily by noting that $\frac{\partial}{\partial\omega_{\mu z}}\log Z^{\mu l}=g_{\rm out}$ (a consequence of Eq. (115):

$$g_{\rm out}(V_{\mu l})+\frac{(\omega_{\mu l}-\sum_i r_{\mu i}a_{il}/\sqrt{N})}{V_{\mu l}-\sum_i s_{\mu i}c_{il}/N}=0\,, \tag{130}$$

which is nothing but the fixed point equation for ω.

It is convenient to compute the derivative with respect to V (even though this quantity is eventually a function of r,s,a and c) using $\frac{\partial}{\partial V_{\mu z}}\log Z^{\mu l}=\frac{1}{2}\left(g_{\rm out}^2+\partial_\omega g_{\rm out}\right)$ so that at the fixed point, when eq. (130) is satisfied, we have

$$\frac{\partial\Phi^{\rm Bethe}_{\rm AMP}}{\partial V_{\mu i}}=\frac{1}{2}\left(g_{\rm out}^2+\partial_\omega g_{\rm out}\right)-\frac{1}{2}\frac{(\omega_{\mu l}-\sum_i r_{\mu i}a_{il}/\sqrt{N})^2}{(V_{\mu l}-\sum_i s_{\mu i}c_{il}/N)^2}=\frac{1}{2}\partial_\omega g_{\rm out}\,. \tag{131}$$

Using this equation, one can finally check explicitly that deriving with respect to a, c, r and s yields the remaning AMP equations for T, Σ, W and Z. This concludes the proof. □

B. The variational Bethe free entropy

We have shown that the fixed points of the approximate message passing equations are extrema of $\Phi^{\rm Bethe}_{\rm AMP}$. However they are in general saddle points of this function, and it is very useful to derive an alternative "variational" free entropy, the *maxima* of which are the fixed points. In particular, this will allow us to find these fixed points by alternative methods which do not rely on iterating the equations. This variational free entropy can also be used not only at the maximum, but for each possible values of the parameters, as the current estimate of the quality of reconstruction.

1. Generic output channel

In order to derive the variational Bethe free entropy, we impose the fixed point conditions, and express the free entropy only as a function of the parameters of our trial distributions for the two matrices. Then, we simply have

$$\Phi^{\rm Bethe}_{\rm var}\left(\{T_{il}\},\{\Sigma_{il}\},\{W_{\mu i}\},\{Z_{\mu i}\}\right) = \Phi^{\rm Bethe}_{\rm AMP}\left(\{T_{il}\},\{\Sigma_{il}\},\{W_{\mu i}\},\{Z_{\mu i}\},\{\omega^*_{\mu l}\},\{a^*_{il}\},\{c^*_{il}\},\{r^*_{\mu i}\},\{s^*_{\mu i}\}\right) \tag{132}$$

where a^*, c^*, r^*, s^* are given in terms of the Eqs. (126-129) by: $a^* = f_a(\Sigma_{il}, T_{il})$, $c^* = f_c(\Sigma_{il}, T_{il})$, $r^* = f_r(Z_{\mu i}, W_{\mu i})$, $s^* = f_s(Z_{\mu i}, W_{\mu i})$, and ω^* is the solution of (85).

In order to write this variational expression in a nicer form, let us notice that the Kullback-Leibler divergences between $\mathcal{M}_X$, $\mathcal{M}_F$ (52-53) and the prior distribution are

$$-D_{\rm KL}(\mathcal{M}_F||P_F) = \log\hat{\mathcal{X}}^{\mu i} + \frac{f_s(Z_{\mu i},W_{\mu i}) + (f_r(Z_{\mu i},W_{\mu i})-W_{\mu i})^2}{2Z_{\mu i}}\,, \tag{133}$$

$$-D_{\rm KL}(\mathcal{M}_X||P_X) = \log\hat{\mathcal{X}}^{il} + \frac{f_c(\Sigma_{il},T_{il}) + (f_a(\Sigma_{il},T_{il})-T_{il})^2}{2\Sigma_{il}}\,. \tag{134}$$

Let us define an additional distribution

$$\mathcal{M}_{\rm out}(\omega_{\mu l}, V_{\mu l}, z) = \frac{1}{\mathcal{Z}^{\mu l}} P_{\rm out}(y_{\mu l}|z)\frac{1}{\sqrt{2\pi V_{\mu l}}} e^{-\frac{(z-\omega_{\mu l})^2}{2V_{\mu l}}}\,, \tag{135}$$

$$\tag{136}$$

where $\mathcal{Z}^{\mu l}$ is given by (115). Then one has

$$-D_{\rm KL}(\mathcal{M}_{\rm out}||P_{\rm out}) = \log\mathcal{Z}^{\mu l} + \frac{1}{2}\log 2\pi V_{\mu l} + \frac{1}{2}\left(1 + V_{\mu l}\partial_\omega g_{\rm out} + V_{\mu l} g^2_{\rm out}\right)\,. \tag{137}$$

Starting from (124), we find:

$$\begin{aligned}
&\Phi^{\rm Bethe}_{\rm var}\left(\{T_{il}\},\{\Sigma_{il}\},\{W_{\mu i}\},\{Z_{\mu i}\}\right) = -\sum_{\mu i} D_{\rm KL}(\mathcal{M}_F(Z_{\mu i},W_{\mu i})||P_F) - \sum_{il} D_{\rm KL}(\mathcal{M}_X(\Sigma_{il},T_{il})||P_X)\\
&+\sum_{\mu l}\left[\log Z^{\mu l}\left(\omega^*_{\mu l}, V^*_{\mu l}\right) + \frac{g^2_{\rm out}}{2}(V^*_{\mu l} - \sum_j s^*_{\mu j}c^*_{jl}/N)\right]\,,\\
&= -\sum_{\mu i} D_{\rm KL}(\mathcal{M}_F(Z_{\mu i},W_{\mu i})||P_F) - \sum_{il} D_{\rm KL}(\mathcal{M}_X(\Sigma_{il},T_{il})||P_X)\\
&-\sum_{\mu l}\left[D_{\rm KL}(\mathcal{M}_{\rm out}(\omega^*_{\mu l},V^*_{\mu l})||P_{\rm out}) - \frac{1}{2}\left(\log 2\pi V^*_{\mu l} + 1 + V^*_{\mu l}\partial_\omega g_{\rm out} + \frac{g^2_{\rm out}}{N}\sum_j s^*_{\mu j}c^*_{jl}\right)\right]\,,
\end{aligned} \tag{138}$$

with V^* and ω^* satisfying eqs. (101) and (104). Note that this expression has the same form as the one used in [52] for the simpler case of GAMP for compressed sensing and for the generalized linear problem. Our expression thus generalizes the formula of [52] to the bi-linear case.

2. The AWGN output channel

In the case of the additive white Gaussian noise output channel (20) the function $g_{\rm out}$ takes the simple form:

$$g_{\rm out}(\omega_{\mu l}, y_{\mu l}, V_{\mu l}) = \frac{y_{\mu l} - \omega_{\mu l}}{\Delta + V_{\mu l}} \ , \tag{139}$$

hence $\partial_\omega g_{\rm out}$ does not depend on the variable $\omega_{\mu l}$. The only explicit dependence on $\omega_{\mu l}$ in the free entropy is through eq. (115) which becomes for the AWGN output channel

$$\mathcal{Z}^{\mu l} = \frac{1}{\sqrt{2\pi(\Delta + V_{\mu l})}} e^{-\frac{(y_{\mu l} - \omega_{\mu l})^2}{2(\Delta + V_{\mu l})}} \ . \tag{140}$$

The free entropy is defined only at the fixed point of the GAMP equations. Given a fixed point we can express from (85) for the AWGN channel

$$\frac{y_{\mu l} - \omega_{\mu l}}{\Delta + V_{\mu l}} = \frac{y_{\mu l} - \frac{1}{\sqrt{N}} \sum_j a_{jl} r_{\mu j}}{\Delta + \frac{1}{N} \sum_j c_{jl} s_{\mu j}} \, . \tag{141}$$

We plug this last expression into (140) to obtain

$$\mathcal{Z}^{\mu l} = \frac{1}{\sqrt{2\pi(\Delta + V_{\mu l})}} e^{-\frac{(y_{\mu l} - \frac{1}{\sqrt{N}} \sum_j a_{jl} r_{\mu j})^2}{2(\Delta + \frac{1}{N} \sum_j c_{jl} s_{\mu j})^2} (\Delta + V_{\mu l})} \, . \tag{142}$$

Simplifying the last two terms of eq. (138) we obtain for the AWGN channel:

$$\begin{aligned} \Phi^{\rm Bethe}_{\rm AWGN}(\{T_{il}\}, \{\Sigma_{il}\}, \{W_{\mu i}\}, \{Z_{\mu i}\}) = & - \sum_{\mu i} D_{\rm KL}(\mathcal{M}_F(Z_{\mu i}, W_{\mu i}) || P_F) - \sum_{il} D_{\rm KL}(\mathcal{M}_X(\Sigma_{il}, T_{il}) || P_X) \\ & - \sum_{\mu l} \frac{(y_{\mu l} - \frac{1}{\sqrt{N}} \sum_j a^*_{jl} r^*_{\mu j})^2}{2(\Delta + \frac{1}{N} \sum_j c^*_{jl} s^*_{\mu j})} - \frac{1}{2} \sum_{\mu l} \log \left[2\pi (\Delta + V^*_{\mu l}) \right] . \end{aligned} \tag{143}$$

The first three terms of this free entropy are clearly negative and the last term cannot be larger than $-MP \log(2\pi\Delta)/2$. Hence the free entropy (143) is bounded from above. This is consistent with its interpretation as a variational expression. The stationary points of (143) are the fixed points of the GAMP algorithm and hence the fixed points corresponding to the maximum likelihood could also be found by direct maximization of the expression (143). This offers an interesting algorithmic alternative to the iterative AMP algorithm that was explored for the compressed sensing problem in [53].

Another use of the expression (143) is that during the iteration of the GAMP algorithm its value should be increasing, hence we can adaptively choose the step-size of the iterations to ensure this increase. Such an adaptive dumping was implemented in [15] using a different form of the free entropy that does not correspond to the Bethe free entropy but to the variational mean field (VMF) free entropy which reads

$$\begin{aligned} \Phi^{\rm VMF}_{\rm AWGN}(\{T_{il}\}, \{\Sigma_{il}\}, \{W_{\mu i}\}, \{Z_{\mu i}\}) = & - \sum_{\mu i} D_{\rm KL}(\mathcal{M}_F(Z_{\mu i}, W_{\mu i}) || P_F) - \sum_{il} D_{\rm KL}(\mathcal{M}_X(\Sigma_{il}, T_{il}) || P_X) \\ & - \frac{1}{2\Delta} \sum_{\mu l} \left[(y_{\mu l} - \frac{1}{\sqrt{N}} \sum_i r_{\mu i} a_{il})^2 + V_{\mu l} \right] - \frac{MP}{2} \log 2\pi\Delta \, . \end{aligned} \tag{144}$$

It is easy to check that $\Phi^{\rm VMF}_{\rm AWGN}(T_{il}, \Sigma_{il}, W_{\mu i}, Z_{\mu i}) < \Phi^{\rm Bethe}_{\rm AWGN}(T_{il}, \Sigma_{il}, W_{\mu i}, Z_{\mu i})$, which could be expected since the Bethe expression, which is asymptotically exact, should be a better approximation than the mean field approximation.

IV. ASYMPTOTIC ANALYSIS

A. State evolution

In this section we derive the asymptotic ($N \to \infty$) evolution of the GAMP iterations for matrix factorization. This asymptotic analysis holds as long as all the elements of the true matrix F are iid random variables generated from a distribution P_{F^0}, and all elements of the true matrix X are iid random variables generated from a distribution P_{X^0}. In general we will not assume $P_{F^0} = P_F$ and $P_{X^0} = P_X$: this special case of Bayes-optimal analysis will be treated in the next section.

In the present section we will also distinguish between the true output channel characterized by the conditional probability distribution $P^0_{\rm out}(y_{\mu l}|z^0_{\mu l})$ and the output channel that is being used in the GAMP algorithm $P_{\rm out}(y_{\mu l}|z_{\mu l})$. We remind that $z^0_{\mu l} = \sum_{i=1}^N F^0_{\mu i} x^0_{il}$, where $F^0_{\mu i}$ and x^0_{il} are the elements of the actual matrices that we do not know and aim to recover, and $z_{\mu l} = \sum_{i=1}^N F_{\mu i} x_{il}$. Again the special case of $P^0_{\rm out} = P_{\rm out}$ will be treated the next section.

We will assume that at least one of the probability distributions P_{F^0} and P_{X^0} (and also at least one of P_F and P_X) has zero mean, otherwise there would be additional terms in this asymptotic analysis, as e.g. in [54].

Let us first define the order parameters; all of them are finite, of order $O(1)$, in the thermodynamic limit:

$$m^t_x \equiv \frac{1}{NP}\sum_{jl} a_{jl}(t) x^0_{jl}\,, \qquad m^t_F \equiv \frac{1}{M\sqrt{N}}\sum_{\mu i} r_{\mu i}(t) F^0_{\mu i}\,, \tag{145}$$

$$q^t_x \equiv \frac{1}{NP}\sum_{jl} a^2_{jl}(t)\,, \qquad q^t_F \equiv \frac{1}{NM}\sum_{\mu i} r^2_{\mu i}(t)\,, \tag{146}$$

$$Q^t_x \equiv q^t_x + \frac{1}{NP}\sum_{jl} c_{jl}(t)\,, \qquad Q^t_F \equiv q^t_F + \frac{1}{NM}\sum_{\mu i} s_{\mu i}(t)\,. \tag{147}$$

Note also that the above sums over a pair of indices could also be sums over only one index (and adjusted normalization) and the order parameters would not change in the leading order: for instance, we expect that in the thermodynamic limit, $\frac{1}{N}\sum_j a_{jl}(t) x^0_{jl}$ will go to the same limit as m^t_x defined in (145).

So far we were characterizing the output channel by the conditional distribution $P^0_{\rm out}(y|z^0)$ or $P_{\rm out}(y|z)$. In this section it will be instrumental to think of the output as a deterministic function h of z and a random variables w, i.e. $y_{\mu l} = h(z_{\mu l}, w_{\mu l}) = h^0(z^0_{\mu l}, w^0_{\mu l})$. The random variables w and w^0 are specified by their probability distributions $P(w)$ and $P_0(w^0)$. We can relate $P_{\rm out}$ to h as follows

$$P_{\rm out}(y|z) = \int P(w)\,{\rm d}w\,\delta[y - h(z,w)]\,, \tag{148}$$

$$P^0_{\rm out}(y|z^0) = \int P_0(w^0)\,{\rm d}w^0\,\delta[y - h^0(z^0,w^0)]\,, \tag{149}$$

In what follows we will be averaging over the realizations of X^0, F^0 and w^0.

First let us compute the average over realizations of X^0, F^0 and w^0 of the quantity $V^t_{\mu l}$ defined by eq. (78). By the assumptions of the belief propagation equations (31-32), the terms in the product in (78) are statistically independent and we can hence write for the average, to the leading order

$$V^t = Q^t_F Q^t_x - q^t_F q^t_x\,. \tag{150}$$

Further, we realize that the variance of this quantity (again over the realizations of X^0, F^0 and w^0) is

$$\mathbb{E}[(V^t_{\mu l} - V^t)^2] = \mathbb{E}\left[\left\{\frac{1}{N}\sum_i [c_{il\to\mu l}(t) - \frac{1}{N}\sum_k c_{kl}(t)] s_{\mu i\to\mu l}() + ... + ...\right\}^2\right] = O(1/N)\,. \tag{151}$$

In order to derive this result, we expand the square and obtain a double sum over i and j. Because of the conditional independence between incoming messages assumed in belief propagation, the terms with $i \neq j$ average exactly to zero. As for the terms with $i = j$, they add up to a contribution of order $O(1/N)$. From this we can conclude that, to leading order in the thermodynamic limit, the quantity $V^t_{\mu l} = V^t$ does not depend on its indices.

Further we are interested in the average Σ^t of the quantity Σ^t_{il} over the realization of F^0, X^0, and w^0. Using the definition of Σ^t_{il} eq. (76) and the expression for $A^t_{\mu l\to il}$ eq. (48) we obtain

$$(\Sigma^t_{il})^{-1} = -\frac{1}{N}\sum_\mu \left\{\partial_\omega g_{\rm out}(\omega^t_{\mu il}, y_{\mu l}, V^t_{\mu il})[r^2_{\mu i\to\mu l}(t) + s_{\mu i\to\mu l}(t)] + g^2_{\rm out}(\omega^t_{\mu il}, y_{\mu l}, V^t_{\mu il}) s_{\mu i\to\mu l}(t)\right\}\,. \tag{152}$$

We proceed analogously for $Z^t_{\mu i}$. Using again the conditional independence between incoming messages assumed in BP equations we obtain

$$(\Sigma^t)^{-1} = \alpha Q_F^t \hat{\chi}^t - \alpha(Q_F^t - q_F^t)\hat{q}^t \,, \tag{153}$$
$$(Z^t)^{-1} = \pi Q_x^t \hat{\chi}^t - \pi(Q_x^t - q_x^t)\hat{q}^t \,, \tag{154}$$

where we introduced new parameters

$$\hat{\chi}^t = -\frac{1}{M}\mathbb{E}_{F^0,X^0,w^0}\left[\sum_\mu \partial_\omega g_{\rm out}(\omega^t_{\mu i l}, y_{\mu l}, V^t_{\mu i l})\right] , \tag{155}$$

$$\hat{q}^t = \frac{1}{M}\mathbb{E}_{F^0,X^0,w^0}\left[\sum_\mu g^2_{\rm out}(\omega^t_{\mu i l}, y_{\mu l}, V^t_{\mu i l})\right] . \tag{156}$$

We use $y_{\mu l} = h(z^0_{\mu l}, w^0)$, and we remind that, to leading order, $V^t_{\mu i l} = V^t$. The function $g_{\rm out}$ above hence depends on two correlated fluctuating variables $\omega^t_{\mu i l}$ and $z^0_{\mu l}$, and on w^0. Both the variables $\omega^t_{\mu i l}$ and $z^0_{\mu l}$ are sums over many independent (for z this is by construction, for ω by the BP assumptions) terms. Hence, according to the central limit theorem, they are Gaussian random variables. Their mean is zero when at least one of the distributions P_X and P_F (and one of the P_{X^0} and P_{F^0}) have zero mean (which we assume in this section). The covariance matrix between the variables $\omega^t_{\mu i l}$ and $z^0_{\mu l}$ is

$$\mathbb{E}(\omega^2_{\mu i l}) = \frac{1}{N}\sum_{j(\neq i)}\mathbb{E}(r^2_{\mu j\to\mu l}a^2_{jl\to\mu l}) + \frac{1}{N}\sum_{j(\neq i),k(\neq i,j)}\mathbb{E}(r_{\mu j\to\mu l}r_{\mu k\to\mu l}a_{jl\to\mu l}a_{kl\to\mu l}) = q_F q_x \,, \tag{157}$$

$$\mathbb{E}(\omega_{\mu i l}z^0_{\mu l}) = \frac{1}{N}\sum_{j(\neq i)}\mathbb{E}(F^0_{\mu j}r_{\mu j\to\mu l}x^0_{jl}a_{jl\to\mu l}) + \frac{1}{N}\sum_{k,j(\neq i,k)}\mathbb{E}(F^0_{\mu k}r_{\mu j\to\mu l}x^0_{kl}a_{jl\to\mu l}) = m_F m_x \,, \tag{158}$$

where we again used the BP assumption of independence between the incoming messages, but also between $F^0_{\mu i}$ and the message $a_{il\to\mu l}$, and between x^0_{il} and $r_{\mu i\to\mu l}$. As for the variance of z^0, we denote it by:

$$\mathbb{E}[(z^0_{\mu l})^2] = N\,\mathbb{E}[(F^0_{\mu i})^2]\,\mathbb{E}[(x^0_{il})^2] = \langle (z^0)^2\rangle \,. \tag{159}$$

Altogether this gives for $\hat{\chi}$ and $\hat{q}$

$$\hat{\chi}^t = -\int {\rm d}w\, P_0(w)\int {\rm d}p\,{\rm d}z\,\mathcal{N}[p,z;q^t_F q^t_x, \langle (z^0)^2\rangle, m^t_F m^t_x]\,\partial_p g_{\rm out}[p,h(z,w),Q^t_F Q^t_x - q^t_F q^t_x] \,, \tag{160}$$

$$\hat{q}^t = \int {\rm d}w\, P_0(w)\int {\rm d}p\,{\rm d}z\,\mathcal{N}[p,z;q^t_F q^t_x, \langle (z^0)^2\rangle, m^t_F m^t_x]\, g^2_{\rm out}[p,h(z,w),Q^t_F Q^t_x - q^t_F q^t_x] \,. \tag{161}$$

where $\mathcal{N}[p,z;\langle p^2\rangle,\langle z^2\rangle,\langle pz\rangle]$ is a joint Gaussian distribution of variables p and z with zero means and variances given in the argument. From the above analysis it also follows that in the leading order the quantities Σ^t_{il} and $Z^t_{\mu i}$ do not depend on their indices il and μl.

We now study the asymptotic behavior of T^t_{il} defined by eq. (76)

$$T^t_{il}/\Sigma^t_{il} = \frac{1}{\sqrt{N}}\sum_\mu B^t_{\mu l\to\imath l} = \frac{1}{\sqrt{N}}\sum_\mu r_{\mu i\to\mu l}(t) g_{\rm out}(\omega^t_{\mu i l}, y_{\mu l}, V^t_{\mu i l}) \,, \tag{162}$$

where we used definition of $B^t_{\mu l\to\imath l}$ in eq. (47). The message $r_{\mu i\to\mu l}(t)$ are uncorrelated with all the other incoming messages and also with all the $F^0_{\mu j}$ for $j\neq i$. It is, however, correlated with $F^0_{\mu i}$ and the dependence on $F^0_{\mu i}$ has to be hence treated separately. After an expansion in the leading order we obtain

$$T^t_{il}/\Sigma^t_{il} = x^0_{il}\frac{1}{\sqrt{N}}\sum_\mu r_{\mu i\to\mu l}(t)\,F^0_{\mu i}\,\partial_z g_{\rm out}(\omega^t_{\mu i l}, h(z_{\mu l},w_{\mu l}), V^t_{\mu i l}) + \frac{1}{\sqrt{N}}\sum_\mu r_{\mu i\to\mu l}(t)\,g_{\rm out}(\omega^t_{\mu i l}, h(\sum_{j(\neq i)}F^0_{\mu j}x^0_{jl}, w), V^t_{\mu i l})$$
$$\approx \alpha x^0_{il} m^t_F \hat{m}^t + \mathcal{N}(0,1)\sqrt{\alpha q^t_F \hat{q}^t} \,, \tag{163}$$

where in the first term we defined the new parameter $\hat{m}$ as

$$\hat{m}^t = \frac{1}{M}\mathbb{E}_{F^0,X^0,w^0}\left[\sum_\mu \partial_z g_{\rm out}(\omega^t_{\mu i l}, h(z_{\mu l},w_{\mu l}), V^t_{\mu i l})\right] . \tag{164}$$

Using the same kind of analysis as we did for $\hat{q}$ and $\hat{\chi}$, we find that

$$\hat{m}^t = \int \mathrm{d}w\, P_0(w) \int \mathrm{d}p\, \mathrm{d}z\, \mathcal{N}[p, z; q_F^t q_x^t, \langle (z^0)^2 \rangle, m_F^t m_x^t]\, \partial_z g_{\rm out}[p, h(z, w), Q_F^t Q_x^t - q_F^t q_x^t]\,, \tag{165}$$

The second term of (163) when averaged over realization of F^0, X^0 and w^0 behaves as a Gaussian random variable. In (163) we moreover assumed that

$$\mathbb{E}_{F^0, X^0, w^0} \left[\frac{1}{M} \sum_\mu g_{\rm out}(\omega_{\mu il}^t, h(\sum_{j(\neq i)} F_{\mu j}^0 x_{jl}^0, w), V_{\mu il}^t) \right] = 0\,, \tag{166}$$

which is true in all the special cases analysed in this paper, and is also true in general under the Bayes-optimal inference as detailed in the next section. If the zero-mean assumption (166) did not hold the density evolution equations would contain additional terms (similarly as if both F and X had non-zero means), see e.g. the state evolution in compressed sensing for non-zero mean matrices [54]. Under the zero-mean assumption, the variance of the Gaussian variable is $\alpha q_F^t \hat{q}^t$ with $\hat{q}^t$ given by (161).

Analogously we have

$$W_{\mu i}^t / Z_{\mu i}^t \approx \pi \sqrt{N} F_{\mu i}^0 m_x^t \hat{m}^t + \mathcal{N}(0, 1) \sqrt{\pi q_x^t \hat{q}^t}\ . \tag{167}$$

With the use of (153-153), (163), (163), and the expressions for messages (90-91) we obtain

$$Q_x^{t+1} - q_x^{t+1} = \int \mathrm{d}x^0 P_{X^0}(x^0) \int \mathcal{D}\xi\ f_c \left[\frac{1}{\alpha Q_F^t \hat{\chi}^t - \alpha (Q_F^t - q_F^t) \hat{q}^t}, \frac{\alpha m_F^t \hat{m}^t\, x^0 + \xi \sqrt{\alpha q_F^t \hat{q}^t}}{\alpha Q_F^t \hat{\chi}^t - \alpha (Q_F^t - q_F^t) \hat{q}^t} \right], \tag{168}$$

$$q_x^{t+1} = \int \mathrm{d}x^0 P_{X^0}(x^0) \int \mathcal{D}\xi\ f_a^2 \left[\frac{1}{\alpha Q_F^t \hat{\chi}^t - \alpha (Q_F^t - q_F^t) \hat{q}^t}, \frac{\alpha m_F^t \hat{m}^t\, x^0 + \xi \sqrt{\alpha q_F^t \hat{q}^t}}{\alpha Q_F^t \hat{\chi}^t - \alpha (Q_F^t - q_F^t) \hat{q}^t} \right], \tag{169}$$

$$m_x^{t+1} = \int \mathrm{d}x^0 P_{X^0}(x^0) \int \mathcal{D}\xi\ x^0 f_a \left[\frac{1}{\alpha Q_F^t \hat{\chi}^t - \alpha (Q_F^t - q_F^t) \hat{q}^t}, \frac{\alpha m_F^t \hat{m}^t\, x^0 + \xi \sqrt{\alpha q_F^t \hat{q}^t}}{\alpha Q_F^t \hat{\chi}^t - \alpha (Q_F^t - q_F^t) \hat{q}^t} \right], \tag{170}$$

where $\mathcal{D}\xi = \mathrm{d}\xi e^{-\xi^2/2}/\sqrt{2\pi}$ is a Gaussian integration measure. Analogously we have for the F-related order parameters

$$Q_F^{t+1} - q_F^{t+1} = \int \mathrm{d}F^0 P_{F^0}(F^0) \int \mathcal{D}\xi\ f_s \left[\frac{1}{\pi Q_x^t \hat{\chi}^t - \pi (Q_x^t - q_x^t) \hat{q}^t}, \frac{\pi m_x^t \hat{m}^t\, \sqrt{N} F^0 + \xi \sqrt{\pi q_x^t \hat{q}^t}}{\pi Q_x^t \hat{\chi}^t - \pi (Q_x^t - q_x^t) \hat{q}^t} \right], \tag{171}$$

$$q_F^{t+1} = \int \mathrm{d}F^0 P_{F^0}(F^0) \int \mathcal{D}\xi\ f_r^2 \left[\frac{1}{\pi Q_x^t \hat{\chi}^t - \pi (Q_x^t - q_x^t) \hat{q}^t}, \frac{\pi m_x^t \hat{m}^t\, \sqrt{N} F^0 + \xi \sqrt{\pi q_x^t \hat{q}^t}}{\pi Q_x^t \hat{\chi}^t - \pi (Q_x^t - q_x^t) \hat{q}^t} \right], \tag{172}$$

$$m_F^{t+1} = \int \mathrm{d}F^0 P_{F^0}(F^0) \int \mathcal{D}\xi\ \sqrt{N} F^0 f_r \left[\frac{1}{\pi Q_x^t \hat{\chi}^t - \pi (Q_x^t - q_x^t) \hat{q}^t}, \frac{\pi m_x^t \hat{m}^t\, \sqrt{N} F^0 + \xi \sqrt{\pi q_x^t \hat{q}^t}}{\pi Q_x^t \hat{\chi}^t - \pi (Q_x^t - q_x^t) \hat{q}^t} \right]. \tag{173}$$

The six equations (168-170) together with (160-165) are the general form of density evolution for GAMP in the general case of matrix factorization. We remind that these equations describe the asymptotic evolution of the algorithm in the "thermodynamic" limit of large sizes, as long as the matrices X, X^0, F, F^0 were generated with iid elements, at least one of the random variables X and F, and at least one of the X^0 and F^0 has zero mean, and the output channel satisfies the condition (166). The first of these condition is absolutely essential for our approach; the restriction to zero means is here for convenience, the non-zero means and generic form of output function can be treated with the same formalism that we have used here, with additional terms.

1. *State evolution of the Bayes-optimal inference*

To satisfy the Nishimori conditions we have to suppose that all the prior distributions are matching the true distributions from which the signal and noise were generated, i.e.

$$P_X(x) = P_{X^0}(x)\,, \qquad P_F(F) = P_{F^0}(F)\,, \qquad P_{\rm out}^0(y|z) = P_{\rm out}(y|z)\,. \tag{174}$$

The third of these equalities is equivalent to $P_0(w) = P(w)$ and $h(z,w) = h^0(z,w)$. Under these conditions the asymptotic analysis simplifies considerably since we have

$$q_x^t = m_x^t\,, \qquad q_F^t = m_F^t\,, \tag{175}$$

$$Q_x^t = \langle (x^0)^2 \rangle\,, \qquad Q_F^t = N\langle (F^0)^2 \rangle\,, \tag{176}$$

$$\hat{\chi}^t = \hat{q}^t = \hat{m}^t\,. \tag{177}$$

To justify the above statement we need to prove that if (174) is satisfied and (175-177) hold up to iteration t then (175-177) hold also in iteration $t+1$. This is done in the next two subsections.

Under the Nishimori conditions the state evolution simplifies into

$$m_x^{t+1} = \int \mathrm{d}x P_X(x) \int \mathcal{D}\xi\ f_a^2\left[\frac{1}{\alpha m_F^t \hat{m}^t}, \frac{\alpha m_F^t \hat{m}^t\, x + \xi\sqrt{\alpha m_F^t \hat{m}^t}}{\alpha m_F^t \hat{m}^t}\right], \tag{178}$$

$$m_F^{t+1} = \int \mathrm{d}F P_F(F) \int \mathcal{D}\xi\ f_r^2\left[\frac{1}{\pi m_x^t \hat{m}^t}, \frac{\pi m_x^t \hat{m}^t\, \sqrt{N}F + \xi\sqrt{\pi m_x^t \hat{m}^t}}{\pi m_x^t \hat{m}^t}\right], \tag{179}$$

$$\hat{m}^t = -\int \mathrm{d}w\, P(w) \int \mathrm{d}p\, \mathrm{d}z \frac{e^{-\frac{p^2}{2m_F^t m_x^t}} e^{-\frac{(z-p)^2}{2[\langle (z^0)^2\rangle - m_F^t m_x^t]}}}{2\pi\sqrt{m_F^t m_x^t(\langle (z^0)^2\rangle - m_F^t m_x^t)}}\, \partial_p g_{\mathrm{out}}(p, h(z,w), \langle (z^0)^2\rangle - m_F^t m_x^t)\,, \tag{180}$$

where $\mathcal{D}\xi$ is a Gaussian integral $\mathrm{d}\xi e^{-\xi^2/2}/\sqrt{2\pi}$. We remind $\langle (z^0)^2\rangle = N\langle (F^0)^2\rangle\langle (x^0)^2\rangle$. Here we chose to use the expression coming from eq. (169), (172) and (160), but we could have used any of the other expressions that are equivalent on the Nishimori line. Where m_F and m_x are initialized as squares of the means of the corresponding prior distributions

$$m_F^{t=0} = N\left[\int \mathrm{d}F\, F P(F)\right]^2, \qquad m_x^{t=0} = \left[\int \mathrm{d}x\, x P(x)\right]^2. \tag{181}$$

In case the prior distribution depends on another random variables, e.g. in case of matrix calibration, we take additional average with respect to that variable. If the above initialization gives $m_F^{t=0} = 0$ and $m_x^{t=0} = 0$ then this is a fixed point of the state evolution. This is due to the permutational symmetry between the columns of matrix F and rows of matrix X. To obtain a nontrivial fixed point we initialize at $m_F^{t=0} = \eta$ for some very small η, corresponding to an infinitesimal prior information about the matrix elements of the matrix F. Note that this is needed only in the state evolution, the algorithm breaks the permutational symmetry spontaneously. The same situation appears in an Ising ferromagnet at low temperature where zero magnetization is a fixed point of the equilibrium equations, but the physically correct solution to which dynamical procedures converge had large magnetization in absolute value.

Our general strategy in the asymptotic analysis of optimal Bayesian inference and related phase transition is that the corresponding fixed points must satisfy the Nishimori conditions, hence we will restrict our search for fixed points to parameters lying on the Nishimori line, i.e. satisfying the conditions (175-177). When these conditions are not imposed the iterations of the state evolution equations are not always converging to fixed points on the Nishimori line. This is also reflected in problems with convergence in the GAMP algorithm for matrix factorization. In the algorithm the Nishimori conditions are unfortunately not straightforward to impose.

2. The input Nishimori conditions

Assume that in the state evolution the Nishimori conditions (175-177) hold for all iteration times smaller or equal to t. Out aim is to show that then m_x^{t+1} and q_x^{t+1} computed from (170) and (169) are equal. Recall that

$$m_x^{t+1} = \int \mathrm{d}x_0 P_X(x_0) \int \mathcal{D}\xi\ x_0 f_a\left[\frac{1}{\alpha m_F^t \hat{m}^t}, \frac{\alpha m_F^t \hat{m}^t\, x_0 + \xi\sqrt{\alpha m_F^t \hat{m}^t}}{\alpha m_F^t \hat{m}^t}\right], \tag{182}$$

$$q_x^{t+1} = \int \mathrm{d}x_0 P_X(x_0) \int \mathcal{D}\xi\ f_a^2\left[\frac{1}{\alpha m_F^t \hat{m}^t}, \frac{\alpha m_F^t \hat{m}^t\, x_0 + \xi\sqrt{\alpha m_F^t \hat{m}^t}}{\alpha m_F^t \hat{m}^t}\right], \tag{183}$$

$$\tag{184}$$

and the definition of

$$f_a(\Sigma^2, R) = \frac{\int \mathrm{d}x\, e^{-\frac{(x-R)^2}{2\Sigma^2}} x\, P_X(x)}{\int \mathrm{d}x\, e^{-\frac{(x-R)^2}{2\Sigma^2}} P_X(x)} = \frac{\int \mathrm{d}x\, e^{-\frac{x^2}{2\Sigma^2} + \frac{R}{\Sigma^2}x} x\, P_X(x)}{\int \mathrm{d}x\, e^{-\frac{x^2}{2\Sigma^2} + \frac{R}{\Sigma^2}x} P_X(x)}\,. \tag{185}$$

Denoting $\tilde{q} = \alpha m_F^t \hat{m}^t$ we have:

$$m_x^{t+1} = \int \mathrm{d}x_0\, P_X(x_0) \int \mathcal{D}\xi\; x_0 \frac{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + (\tilde{q}x_0 + \xi\sqrt{\tilde{q}})x} x\, P_X(x)}{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + (\tilde{q}x_0 + \xi\sqrt{\tilde{q}})x} P_X(x)}\,, \tag{186}$$

$$q_x^{t+1} = \int \mathrm{d}x_0 P_X(x_0) \int \mathcal{D}\xi\; \left[\frac{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + (\tilde{q}x_0 + \xi\sqrt{\tilde{q}})x} x\, P_X(x)}{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + (\tilde{q}x_0 + \xi\sqrt{\tilde{q}})x} P_X(x)}\right]^2. \tag{187}$$

Performing the change of variables $\xi\sqrt{\tilde{q}} + \tilde{q}x_0 \to \xi\sqrt{\tilde{q}}$ the Gaussian measure become $\mathcal{D}\xi e^{-\frac{\tilde{q}x_0^2}{2} + \xi x_0\sqrt{\tilde{q}}}$ so that

$$m_x^{t+1} = \int \mathcal{D}\xi \int \mathrm{d}x_0 P_X(x_0) e^{-\frac{\tilde{q}x_0^2}{2} + \xi x_0\sqrt{\tilde{q}}}\, x_0\, \frac{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + \xi\sqrt{\tilde{q}}x} x P_X(x)}{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + \xi\sqrt{\tilde{q}}x} P_X(x)} = \int \mathcal{D}\xi \frac{\left[\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + \xi\sqrt{\tilde{q}}x} x P_X(x)\right]^2}{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + \xi\sqrt{\tilde{q}}x} P_X(x)}\,. \tag{188}$$

Analogously we obtain

$$q_x^{t+1} = \int \mathcal{D}\xi \int \mathrm{d}x_0 P_X(x_0) e^{-\frac{\tilde{q}x_0^2}{2} + \xi\sqrt{\tilde{q}}x_0} \left[\frac{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + \xi\sqrt{\tilde{q}}x} x\, P_X(x)}{\int \mathrm{d}x\, e^{-\frac{\tilde{q}x^2}{2} + \xi\sqrt{\tilde{q}}x} P_X(x)}\right]^2 = m_x^{t+1}\,. \tag{189}$$

The proof of $m_F^{t+1} = q_F^{t+1}$ is exactly the same.

The next identity we want to prove is $Q_x^{t+1} = \langle (x^0)^2 \rangle$. From the general state evolution equation (168) we get under conditions (175-177) that

$$Q_x^{t+1} - q_x^{t+1} = \int \mathrm{d}x_0 P_X(x_0) \int \mathcal{D}\xi\; f_c\left[\frac{1}{\alpha m_F^t \hat{m}^t}, \frac{\alpha m_F^t \hat{m}^t\, x_0 + \xi\sqrt{\alpha m_F^t \hat{m}^t}}{\alpha m_F^t \hat{m}^t}\right]. \tag{190}$$

Using the definition of the function $f_c(\Sigma, R)$ (54), the same change of variables, and resulting cancelations as above we get

$$Q_x^{t+1} = \int \mathcal{D}\xi \int \mathrm{d}x_0\, x_0^2\, P_X(x_0) e^{-\frac{\tilde{q}x_0^2}{2} + \xi\sqrt{\tilde{q}}x_0} = \langle (x^0)^2 \rangle\,. \tag{191}$$

And analogously for $Q_F^{t+1} = N\langle (F^0)^2 \rangle$.

3. The output Nishimori conditions

Let us now assume that the input Nishimori conditions (175-176) are satisfied and we want to show that (177) and (166) hold.

We depart from the general expressions (160-165). We notice that for $q_F^t = m_F^t$ and $q_x^t = m_x^t$ the joint Gaussian measure for variables p and z in (157-159) can be written as a product of two Gaussian measures. We have in that case $\langle \omega_{\mu l}(\omega_{\mu l} - z_{\mu l}^0) \rangle = 0$, hence one of the Gaussian has zero mean and variance $m_F^t m_x^t$ and the other one mean p and variance $V^t = \langle (z^0)^2 \rangle - m_F^t m_x^t$. Furthermore, performing the integration over variable z by parts in eq. (165) and then using the relation between $h(z, w)$ and $P_{\rm out}(y|z)$ from eq. (148) and the definition of $g_{\rm out}$ from eq. (44) we get

$$\hat{m}^t = \int \mathrm{d}y\, \mathrm{d}p\, \mathrm{d}z\, P_{\rm out}(y|z)\, \mathcal{N}(p, z) \frac{z - p}{V^t} \frac{\int \mathrm{d}z' P_{\rm out}(y|z')(z' - p) e^{-\frac{(z'-p)^2}{2V^t}}}{V^t \int \mathrm{d}z'' P_{\rm out}(y|z'') e^{-\frac{(z''-p)^2}{2V^t}}}\,. \tag{192}$$

Doing analogous manipulations of expliciting the Gaussian measure and using eq. (148) and the definition of $g_{\rm out}$ in equation (161) we obtain

$$\hat{q}^t = \hat{m}^t\,. \tag{193}$$

For $\hat{\chi}^t$ we do integration with respect to p by parts and using steps as in the above we obtain

$$\hat{\chi}^t = -\frac{1}{m_F^t m_x^t} \int \mathrm{d}y\, \mathrm{d}p\, \mathrm{d}z\, P_{\rm out}(y|z)\, \mathcal{N}(p, z)\, p \frac{\int \mathrm{d}z' P_{\rm out}(y|z')(z' - p) e^{-\frac{(z'-p)^2}{2V^t}}}{V^t \int \mathrm{d}z'' P_{\rm out}(y|z'') e^{-\frac{(z''-p)^2}{2V^t}}} + \hat{q}^t = \hat{q}^t\,. \tag{194}$$

Thanks to a cancelation between the integrals over variables z and z'' we can perform explicitly the integral over y (keeping in mind that $P_{\rm out}(y|z')$ is a normalized probability distribution). The remaining Gaussian integral is then zero.

In a analogous manner we prove eq. (166) by noticing that the expectation with respect to F^0, X^0 and w^0 is exactly the integral $\int \mathrm{d}y\,\mathrm{d}p\,\mathrm{d}z\,P_{\rm out}(y|z)\mathcal{N}(p,z)$.

To conclude, conditions (177) and (166) hold in the limit $N\to\infty$. However, as common in statistical physics we can recall the self-averaging property under which quantities on almost every large ($N\to\infty$) instance are equal to their averages aver randomness (disorder) of F^0, X^0 and w^0. This self-averaging then for instance justifies the use of eq. (95) on large single instances of the matrix factorization problem.

B. Replica method

The replica method is known as a non-rigorous approach to evaluate the typical performance of various Bayesian inference problems. We here show how this can be employed for the matrix factorization problem.

1. *Moment assessment for $n\in\mathbb{N}$*

The expression of the partition function

$$Z(Y)=\int \mathrm{d}F\mathrm{d}X\prod_{\mu,i}P_F(F_{\mu i})\prod_{i,l}P_X(X_{il})\prod_{\mu,l}P_{\rm out}\left(y_{\mu l}|\sum_i F_{\mu i}x_{il}\right) \tag{195}$$

constitutes the basis of our analysis. In statistical mechanics, one can generally examine properties of systems via evaluation of the free entropy $\log Z(Y)$, which statistically fluctuates depending on the realization of Y in the current case. However, as $N,M,P\to\infty$, one can expect that the self-averaging property holds and, therefore, the free entropy density $N^{-2}\log Z(Y)$ converges to its typical value $\phi\equiv N^{-2}\left[\log Z(Y)\right]_Y$ with probability of unity. This is also expected to hold for other macroscopic quantities relevant to the performance of the matrix factorization. Therefore, assessment of ϕ is the central issue in our analysis.

This can be systematically carried out by the replica method. For this, we first evaluate the n-th moment of $Z(Y)$, $\left[Z^n(Y)\right]_Y=\int \mathrm{d}Y P_0(Y)Z^n(Y)$, for $n\in\mathbb{N}$ utilizing an identity

$$Z^n(Y)=\int\prod_{a=1}^n\{\mathrm{d}F^a\mathrm{d}X^a\prod_{\mu,i}P_F(F^a_{\mu i})\prod_{i,l}P_X(X^a_{il})\}\times\{\prod_{\mu,l}\prod_{a=1}^n P_{\rm out}(y_{\mu l}|\sum_i F^a_{\mu i}x^a_{il})\}, \tag{196}$$

with respect to the generative distribution of Y

$$P_0(Y)=\int \mathrm{d}F^0\mathrm{d}X^0\prod_{\mu,i}P_{F^0}(F^0_{\mu i})\prod_{i,l}P_{X^0}(X^0_{il})\prod_{\mu,l}P^0_{\rm out}(y_{\mu l}|\sum_i F^0_{\mu i}x^0_{il}), \tag{197}$$

where we assumed that the functional forms of $P_{F^0}(F_{\mu i})$, $P_{X^0}(x_{il})$ and $P^0_{\rm out}(y_{\mu l}|\sum_i F_{\mu i}x_{il})$ may be different from those of the assumed model $P_F(F_{\mu i})$, $P_X(x_{il})$ and $P_{\rm out}(y_{\mu l}|\sum_i F_{\mu i}x_{il})$ for generality. When they are equal, which correspond to the Nishimori condition, $P_0(Y)=Z(Y)$ holds.

In performing the integrals of $2(n+1)$ matrices $(F^0,\{F^a\}_{a=1}^n)$, $(X^0,\{X^a\}_{a=1}^n)$ and Y that come out in evaluating $\left[Z^n(Y)\right]_Y$, we insert trivial identities with respect to all combinations of replica indices $a\le b=0,1,2,\ldots,n$

$$1=M\int \mathrm{d}q_F^{ab}\delta\left(F^a\cdot F^b-Mq_F^{ab}\right) \tag{198}$$

and

$$1=NP\int \mathrm{d}q_x^{ab}\delta\left(X^a\cdot X^b-NPq_x^{ab}\right) \tag{199}$$

to the integrand, where $F^a\cdot F^b\equiv\sum_{\mu,i}F^a_{\mu i}F^b_{\mu i}$ and similarly for $X^a\cdot X^b$. Let us denote $\mathcal{Q}_F\equiv(q_F^{ab})$ and $\mathcal{Q}_x\equiv(q_x^{ab})$, and introduce two joint distributions

$$P_F(\{F^a\};\mathcal{Q}_F)=\frac{1}{V_F(\mathcal{Q}_F)}\prod_{\mu,i}\left(P_{F^0}(F^0_{\mu i})\prod_{a=1}^n P_F(F^a_{\mu i})\right)\prod_{a\le b}\delta\left(F^a\cdot F^b-Mq_F^{ab}\right) \tag{200}$$

and

$$P_X(\{X^a\};\mathcal{Q}_X)=\frac{1}{V_x(\mathcal{Q}_x)}\prod_{i,l}\left(P_{X^0}(x^0_{il})\prod_{a=1}^{n}P_X(x^a_{il})\right)\prod_{a\le b}\delta\left(X^a\cdot X^b-NPq_x^{ab}\right), \tag{201}$$

where $V_F(\mathcal{Q}_F)$ and $V_x(\mathcal{Q}_x)$ are the normalization constants. These yield an expression of $[Z^n(Y)]_Y$ as

$$[Z^n(Y)]_Y=\int \mathrm{d}Y\mathrm{d}(M\mathcal{Q}_F)\mathrm{d}(NP\mathcal{Q}_x)\left\{V_F(\mathcal{Q}_F)V_x(\mathcal{Q}_x)\right.$$
$$\left.\times\left[\prod_{\mu,l}\left(P^0_{\mathrm{out}}(y_{\mu l}|\sum_i F^0_{\mu i}x^0_{il})\prod_{a=1}^{n}P_{\mathrm{out}}(y_{\mu l}|\sum_i F^a_{\mu i}x^a_{il})\right)\right]_{\mathcal{Q}_F,\mathcal{Q}_x}\right\}, \tag{202}$$

where $[\cdots]_{\mathcal{Q}_F,\mathcal{Q}_x}$ denotes the average with respect to (200) and (201). In computing $[\cdots]_{\mathcal{Q}_F,\mathcal{Q}_x}$, it is noteworthy that $\{F^a\}$ and $\{X^a\}$ follow statistically independent distributions, and either of them has zero mean and both of them have finite variances from our assumption. These allow us to handle $z^a_{\mu l}\equiv\sum_i F^a_{\mu i}X^a_{\mu i}$ $(a=0,1,2,\ldots,n;\mu=1,2,\ldots,M;l=1,2,\ldots,P)$ as multivariate Gaussian random variables whose distribution is given by

$$P_Z(\{z^a_{\mu l}\}|\mathcal{Q}_F,\mathcal{Q}_x)=\prod_{\mu,l}\frac{1}{\sqrt{(2\pi)^{n+1}\det\mathcal{T}}}\exp\left(-\frac{1}{2}\sum_{a,b}z^a_{\mu l}\left(\mathcal{T}^{-1}\right)z^b_{\mu l}\right), \tag{203}$$

where $\mathcal{T}=(q_F^{ab}q_X^{ab})\in\mathbb{R}^{(n+1)\times(n+1)}$. Employing this and evaluating the integrals of $\mathcal{Q}_F$ and $\mathcal{Q}_x$ by means of the saddle point method yield an expression

$$\frac{1}{N^2}\log\left[Z^n(Y)\right]_Y=\operatorname*{extr}_{\mathcal{Q}_F,\mathcal{Q}_x}\left\{\alpha\pi\mathcal{I}_{Fx}(\mathcal{Q}_F,\mathcal{Q}_x)+\alpha\mathcal{I}_F(\mathcal{Q}_F)+\pi\mathcal{I}_x(\mathcal{Q}_x)\right\}, \tag{204}$$

where

$$\mathcal{I}_{Fx}(\mathcal{Q}_F,\mathcal{Q}_x)=\log\left(\int\left(\int P_Z(\{z^a\}|\mathcal{Q}_F,\mathcal{Q}_x)\left(P^0_{\mathrm{out}}(y|z^0)\prod_{a=1}^{n}P_{\mathrm{out}}(y|z^a)\right)\prod_{a=0}^{n}\mathrm{d}z^a\right)\mathrm{d}y\right), \tag{205}$$

$$\begin{aligned}\mathcal{I}_F(\mathcal{Q}_F)&=\frac{1}{NM}\log V_F(\mathcal{Q}_F)\\&=\operatorname*{extr}_{\hat{\mathcal{Q}}_F}\left\{\frac{1}{2}\mathrm{Tr}\hat{\mathcal{Q}}_F\mathcal{Q}_F+\log\left(\int P_{F^0}(F^0)\prod_{a=1}^{n}P_F(F^a)\exp\left(-\frac{N}{2}\mathbf{F}^{\mathrm{T}}\hat{\mathcal{Q}}_F\mathbf{F}\right)\prod_{a=0}^{n}\mathrm{d}F^a\right)\right\},\end{aligned} \tag{206}$$

$$\begin{aligned}\mathcal{I}_x(\mathcal{Q}_x)&=\frac{1}{NP}\log V_x(\mathcal{Q}_x)\\&=\operatorname*{extr}_{\hat{\mathcal{Q}}_x}\left\{\frac{1}{2}\mathrm{Tr}\hat{\mathcal{Q}}_x\mathcal{Q}_x+\log\left(\int P_{X^0}(x^0)\prod_{a=1}^{n}P_X(x^a)\exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\hat{\mathcal{Q}}_x\mathbf{x}\right)\prod_{a=0}^{n}\mathrm{d}x^a\right)\right\},\end{aligned} \tag{207}$$

and $\operatorname{extr}_\Theta\{\cdots\}$ denotes the operation of extremization with respect to Θ. $\hat{\mathcal{Q}}_F=(\hat{q}_F^{ab})\in\mathbb{R}^{(n+1)\times(n+1)}$ are introduced for the saddle point evaluation of $V_F(\mathcal{Q}_F)$ and $\mathbf{F}=(F^a)\in\mathbb{R}^{n+1}$, and similarly for $\hat{\mathcal{Q}}_x\in\mathbb{R}^{(n+1)\times(n+1)}$ and $\mathbf{x}=(x^a)\in\mathbb{R}^{n+1}$.

2. Replica symmetric free entropy

Literally evaluating (204) should yield the correct leading order estimate of $N^{-2}\log[Z^n(Y)]_Y$ for each $n\in\mathbb{N}$. However, we here restrict the candidate of the dominant saddle point to that of the replica symmetric form as

$$(q_F^{ab},q_x^{ab},\hat{q}_F^{ab},\hat{q}_x^{ab})=\begin{cases}(Q_F^0,Q_x^0,\hat{Q}_F^0,\hat{Q}_x^0), & a=b=0,\\(Q_F,Q_x,\hat{Q}_F,\hat{Q}_x), & a=b\ (a,b\neq 0),\\(q_F,q_x,-\hat{q}_F,-\hat{q}_x), & a\neq b\ (a,b\neq 0),\\(m_F,m_x,-\hat{m}_F,-\hat{m}_x), & a=0,b\neq 0,\end{cases} \tag{208}$$

in order to obtain an analytic expression with respect to $n \in \mathbb{R}$. After some algebra utilizing a formula $\exp(A\sum_{a\leq b}u^a u^b) = \exp\left(-\frac{A}{2}\sum_a (u^a)^2\right)\int \mathcal{D}\xi \exp\left(\sqrt{A}\xi\sum_a u^a\right)$ for $A \geq 0$, this provides

$$\mathcal{I}_{Fx}(\mathcal{Q}_F,\mathcal{Q}_x) = \log\left(\int \mathrm{d}y\mathcal{D}\xi\left(\int \mathcal{D}u^0 P^0_{\rm out}\left(y|\sqrt{Q^0_F Q^0_x - \frac{m^2_F m^2_x}{q_F q_x}}u^0 + \frac{m_F m_x}{\sqrt{q_F q_x}}\xi\right)\right.\right.$$
$$\left.\left.\times\left(\int \mathcal{D}u P_{\rm out}\left(y|\sqrt{Q_F Q_x - q_F q_x}u + \sqrt{q_F q_x}\xi\right)\right)^n\right)\right), \tag{209}$$

$$\mathcal{I}_F(\mathcal{Q}_F) = \underset{\hat{Q}^0_F,\hat{Q}_F,\hat{q}_F,\hat{m}_F}{\rm extr}\left\{\frac{\hat{Q}^0_F Q^0_F}{2} + \frac{n\hat{Q}_F Q_F}{2} - \frac{n(n-1)\hat{q}_F q_F}{2} - n\hat{m}_F m_F\right.$$
$$\left. + \log\left(\int \mathcal{D}\xi \mathrm{d}F^0 e^{-\frac{N\hat{Q}^0_F}{2}(F^0)^2} P_{F^0}(F^0)\left(\int \mathrm{d}F e^{-\frac{N(\hat{Q}_F+\hat{q}_F)}{2}F^2 + (\sqrt{N\hat{q}_F}\xi + N\hat{m}_F F^0)F} P_F(F)\right)^n\right)\right\}, \tag{210}$$

and

$$\mathcal{I}_x(\mathcal{Q}_x) = \underset{\hat{Q}^0_x,\hat{Q}_x,\hat{q}_x,\hat{m}_x}{\rm extr}\left\{\frac{\hat{Q}^0_x Q^0_x}{2} + \frac{n\hat{Q}_x Q_x}{2} - \frac{n(n-1)\hat{q}_x q_x}{2} - n\hat{m}_x m_x\right.$$
$$\left. + \log\left(\int \mathcal{D}\xi \mathrm{d}x^0 e^{-\frac{\hat{Q}^0_x}{2}(x^0)^2} P_{X^0}(x^0)\left(\int \mathrm{d}x e^{-\frac{\hat{Q}_x+\hat{q}_x}{2}x^2 + (\sqrt{\hat{q}_x}\xi + \hat{m}_x x^0)x} P_X(x)\right)^n\right)\right\}, \tag{211}$$

all of which are analytic with respect to $n \in \mathbb{R}$. Substituting these into an identity $N^{-2}\left[\log Z(Y)\right]_Y = \lim_{n\to 0}\frac{\partial}{\partial n}N^{-2}\log\left[Z^n(Y)\right]_Y$ leads to the general expression of the free entropy of the matrix factorization problems. In this, $Q^0_F = N\int dF^0 (F^0)^2 P_{F^0}(F^0) = N\langle (F^0)^2\rangle$, $Q^0_x = \int dx^0 (x^0)^2 P_{X^0}(x^0) = \langle (x^0)^2\rangle$, $\hat{Q}^0_F = 0$, and $\hat{Q}^0_x = 0$ are enforced for the consistency of $\lim_{n\to 0}\left[Z^n(Y)\right]_Y = 1$. After taking these into account, the expression becomes as

$$\begin{aligned}
\phi &= \frac{1}{N^2}\left[\log Z(Y)\right]_Y \\
&= \mathrm{extr}\left\{\alpha\pi\int \mathrm{d}y\,\mathcal{D}\xi\,\mathcal{D}u^0 P^0_{\rm out}\left(y|\sqrt{Q^0_F Q^0_x - \frac{m^2_F m^2_x}{q_F q_x}}u^0 + \frac{m_F m_x}{\sqrt{q_F q_x}}\xi\right)\log\left(\int \mathcal{D}u\, P_{\rm out}\left(y|\sqrt{Q_F Q_x - q_F q_x}u + \sqrt{q_F q_x}\xi\right)\right)\right. \\
&+ \alpha\left(\frac{\hat{Q}_F Q_F}{2} + \frac{\hat{q}_F q_F}{2} - \hat{m}_F m_F + \int \mathcal{D}\xi\,\mathrm{d}F^0 P_{F^0}(F^0)\log\left(\int \mathrm{d}F e^{-\frac{N(\hat{Q}_F+\hat{q}_F)}{2}F^2 + (\sqrt{N\hat{q}_F}\xi + N\hat{m}_F F^0)F} P_F(F)\right)\right) \\
&\left.+\pi\left(\frac{\hat{Q}_x Q_x}{2} + \frac{\hat{q}_x q_x}{2} - \hat{m}_x m_x + \int \mathcal{D}\xi\,\mathrm{d}x^0 P_{X^0}(x^0)\log\left(\int \mathrm{d}x e^{-\frac{\hat{Q}_x+\hat{q}_x}{2}x^2 + (\sqrt{\hat{q}_x}\xi + \hat{m}_x x^0)x} P_X(x)\right)\right)\right\},
\end{aligned} \tag{212}$$

where the extremization with respect to $Q_F, Q_x, q_F, q_x, m_F, m_x$, and their conjugate variables leads to the same equations we obtained in the state evolution. Particularly extremization w.r.t. the six conjugate variables gives eqs. (168-173). Extremization with respect to $Q_F, Q_x, q_F, q_x, m_F, m_x$ gives $\hat{Q}_F = \pi Q_x(\hat{\chi} - \hat{q})$, $\hat{Q}_x = \alpha Q_F(\hat{\chi} - \hat{q})$, $\hat{q}_F = \pi q_x \hat{q}$, $\hat{q}_x = \alpha q_F \hat{q}$, $\hat{m}_F = \pi m_x \hat{m}$, $\hat{m}_x = \alpha m_F \hat{m}$ where $\hat{\chi}$, $\hat{q}$, and $\hat{m}$ are given byt eqs. (160-161) and (165).

3. Simplification on the Nishimori condition

One can generally evaluate thermodynamically dominant values of $Q_F, Q_x, q_F, q_x, m_F, m_x$ by solving the extremization problem of (212), which is involved with twelve variables including the conjugate variables and therefore is rather complicated to handle. However, the problem is significantly simplified when the Nishimori condition holds.

This is because the Nishimori condition allows us to handle $F^0 = (F^0_{\mu i})$ and $X^0 = (x^0_{il})$ as if they were the $n+1$-st replica variables added to the n-replicated system composed of $\{F^a\}_{a=1}^n = \{(F^a_{\mu i})\}_{a=1}^n$ and $\{X^a\}_{a=1}^n = \{(x^a_{il})\}_{a=1}^n$ in the computation of the moment $\left[Z^n(Y)\right]_Y = \int dY P_0(Y) Z^n(Y) = \int dY P_0^{n+1}(Y)$. The replica symmetry among $a = 0, 1, 2, \ldots, n$ ensures the following properties:

- $m_F = q_F$, $m_x = q_x$, $\hat{m}_F = \hat{q}_F$, and $\hat{m}_x = \hat{q}_x$ are satisfied.
- $Q_F = Q^0_F$, $Q_x = Q^0_x$, $\hat{Q}_F = 0$, and $\hat{Q}_x = 0$ hold for $n \to 0$.

Substituting these into (212) yields a simplified expression of the free entropy, which is involved with *only* four macroscopic variables m_F, m_x, $\hat{m}_F$, and $\hat{m}_x$, as

$$\begin{aligned}\phi &= \frac{1}{N^2}\left[\log Z(Y)\right]_Y = \frac{1}{N^2}\int dY P_0(Y)\log P_0(Y)\\ &= \mathrm{extr}\left\{\alpha\pi\int \mathrm{d}y\,\mathcal{D}\xi\,\mathcal{D}u^0 P_{\rm out}\left(y|\sqrt{Q_F^0Q_x^0-m_Fm_x}u^0+\sqrt{m_Fm_x}\xi\right)\log\left(\int \mathcal{D}u\,P_{\rm out}\left(y|\sqrt{Q_F^0Q_x^0-m_Fm_x}u+\sqrt{m_Fm_x}\xi\right)\right)\right.\\ &\quad +\alpha\left(-\frac{\hat{m}_Fm_F}{2}+\int\mathcal{D}\xi\,\mathrm{d}F^0 e^{-\frac{N\hat{m}_F}{2}(F^0)^2+\sqrt{N\hat{m}_F}\xi F^0}P_F(F^0)\log\left(\int \mathrm{d}F e^{-\frac{N\hat{m}_F}{2}F^2+\sqrt{N\hat{m}_F}\xi F}P_F(F)\right)\right)\\ &\quad \left.+\pi\left(-\frac{\hat{m}_xm_x}{2}+\int\mathcal{D}\xi\,\mathrm{d}x^0 e^{-\frac{\hat{m}_x}{2}(x^0)^2+\sqrt{\hat{m}_x}\xi x^0}P_X(x^0)\log\left(\int \mathrm{d}x e^{-\frac{\hat{m}_x}{2}x^2+\sqrt{\hat{m}_x}\xi x}P_X(x)\right)\right)\right\},\end{aligned} \tag{213}$$

where we changed integration variables as $\sqrt{N\hat{m}_F}\xi + N\hat{m}_F F^0 \to \sqrt{N\hat{m}_F}\xi$ and $\sqrt{\hat{m}_x}\xi + \hat{m}_x x^0 \to \sqrt{\hat{m}_x}\xi$ together with $\mathcal{D}\xi \to \mathcal{D}\xi e^{-\frac{N\hat{m}_F}{2}(F^0)^2+\sqrt{N\hat{m}_F}\xi F^0}$ and $\mathcal{D}\xi \to \mathcal{D}\xi e^{-\frac{\hat{m}_x}{2}(x^0)^2+\sqrt{\hat{m}_x}\xi x^0}$, respectively. The saddle point conditions can be summarized via equation that we obtained in the state evolution, notably eqs. (178-180) with $\hat{m}_F = \pi m_x\hat{m}$, $\hat{m}_x = \alpha m_F\hat{m}$.

V. ASYMPTOTIC PHASE DIAGRAMS

In this section we use the state evolution derived in section IV A 1 to analyze the asymptotic MMSE of the Bayes-optimal inference in matrix factorization for applications listed in section I C. We restrict our analysis to the Bayes-optimal inference, i.e. the case where we generate the data as specified in Sec. I C and assume that we know the corresponding distributions. In terms of the AMP algorithm and the state evolution this means we can use all the simplifications that arise under the Nishimori conditions. The AMP algorithms for matrix factorization and the asymptotic analysis derived in sections II and IV apply to all the examples, the only elements that are application-dependent are the "input" functions f_a and f_r, and the output function $g_{\rm out}$.

A. Easy/hard and possible/impossible phase transitions

Fixed points of equations (178-180) allow us to evaluate the MMSE of the matrix F as $E_F = N\langle(F^0)^2\rangle - m_F$ and of the matrix X as $E_X = \langle(x^0)^2\rangle - m_x$. We will be investigating two fixed points: The one that is reached from the "uninformative" initialization (181), and the fixed point that is reached by iterating (178-180) with the informative (planted) initialization $m_F^{t=0} = N\langle(F^0)^2\rangle$ and $m_x^{t=0} = \langle(x^0)^2\rangle$. If these two initializations lead to a different fixed point it is the one with larger value of the Bethe free entropy (213) that corresponds to the Bayes-optimal MMSE. Further in cases where the uninformative initialization does not lead to this Bayes-optimal MMSE we conjecture that the optimal solution is hard to reach for a large class of algorithms.

Depending on the application in question and the value of parameters $(\alpha, \pi, \rho, \epsilon, \dots)$ we can sometimes identify a phase transition, i.e. a sharp change in behavior of the MMSE. As in statistical physics it is instrumental to distinguish two kinds of phase transitions

- *Second order phase transition:* In this case there are two regions of parameters. In one, the recovery performance is poor (positive MMSE); in the other one the recovery is perfect (zero MMSE). This situation can arrive only in zero noise, with positive noise there is a smooth "crossover" and the transition between the phase of good and poor recovery is not sharply defined. Interestingly, as we will see, in all examples where we observed the second order phase transition its position coincides with the simple counting bounds discussed in Section I C. The phase with poor MMSE means that there is simply not enough information in the data to recover the original signal (independently of recovery algorithm or its computational complexity). Experience from statistical physics tells us that in problems where a second order phase transition happens, and more general in cases where the state evolution (178-180) has a unique fixed point, there is no fundamental barrier that would prevent good algorithms to attain the MMSE. And in particular our analysis suggest that in the limit of large system sizes the AMP algorithm derived in this paper should be able to achieve this Bayes-optimal MMSE.

- *First order phase transition:* A more subtle phenomenology can be observed in low noise regime in dictionary learning, sparse PCA and blind matrix calibration problems. In some region of parameters, that we call the "spinodal" region, the informative (planted) and uninformative initializations do not lead to the same fixed point of the state evolution equations. The spinodal regime itself may divide into two parts - the one where the uninformative fixed point has a larger free entropy, and the "solvable but hard" phase where the informative

fixed point has a larger free entropy. The boundary between these two parts is the first order phase transition point. We conjecture that in the solvable but hard phase the Bayes-optimal MMSE is not achievable by the AMP algorithm nor by a large class of other known algorithms. The first order phase transition is associated with a discontinuity in the MMSE. The MSE reached from the uninformative initializations will be denoted AMP-MSE and is also discontinuous.

In the case of the first order phase transition it will hence be useful to distinguish in our notations between the minimal achievable MSE that we denote MMSE, and the MSE achievable by the AMP-like algorithms that we denote AMP-MSE. When MMSE=AMP-MSE in the large N limit we say that AMP is asymptotically optimal. The region where in the large size limit MMSE<AMP-MSE is called the *spinodal* regime/region.

B. Dictionary learning, blind matrix calibration, sparse PCA and blind source separation

In terms of our asymptotic analysis the equations for dictionary learning, sparse PCA, and blind source separation are very close, see definitions in section I C, these problems basically differ by the region of parameters α, π that is of interest. Moreover the dictionary learning can be seen as the $\eta \to \infty$ limit of the blind calibration problem (which is trivially taken in the equations). We hence group the discussion of these problems in the present section, they all present the first order phase transition as low measurement noise.

1. Input and output functions

The matrices F and X in our setting of the dictionary learning and sparse PCA problems are generated according to eqs. (18-19). Using the definitions of the input function f_a in eq. (54), and f_r in eq. (55) we obtain explicitly:

$$f_r(Z,W) = \frac{W}{1+Z}\,, \qquad f_a(\Sigma,T) = \frac{\rho\, e^{-\frac{(T-\overline{x})^2}{2(\Sigma+\sigma)}} \frac{\sqrt{\Sigma}}{(\Sigma+\sigma)^{\frac{3}{2}}} (\overline{x}\Sigma + T\sigma)}{(1-\rho)e^{-\frac{T^2}{2\Sigma}} + \rho \frac{\sqrt{\Sigma}}{\sqrt{\Sigma+\sigma}} e^{-\frac{(T-\overline{x})^2}{2(\Sigma+\sigma)}}}\,, \tag{214}$$

The functions $f_c(\Sigma,T)$ and $f_s(Z,W)$ are then obtained from eqs. (56-57).

In case of matrix calibration we have some prior knowledge on the matrix F given by eq. (23). This leads to a function f_r of the form

$$f_r(Z,W) = \frac{W + \sqrt{N} F'_{\mu i} \frac{\sqrt{1+\eta}}{\eta}}{Z + 1 + \frac{1}{\eta}}\,. \tag{215}$$

Indeed as the uncertainty in the matrix $\eta \to \infty$ this goes to the f_r for dictionary learning or sparse PCA (214).

The output function $g_{\rm out}$ defined in eq. (20) is, for the output channel (44) with additive white Gaussian noise of variance Δ:

$$g_{\rm out}(\omega, y, V) = \frac{y-\omega}{\Delta + V}\,. \tag{216}$$

For such a simple output function the eqs. (160-165) in the density evolution simplify greatly into

$$\hat{\chi}^t = \hat{m}^t = \frac{1}{\Delta + Q_F^t Q_x^t - q_F^t q_x^t}\,, \qquad \hat{q}^t = \frac{\Delta_0 + \rho_0(\overline{x}_0^2 + \sigma_0) + q_F q_x - 2 m_F m_x}{(\Delta + Q_F^t Q_x^t - q_F^t q_x^t)^2}\,, \tag{217}$$

which under the simplification of the Nishimori line gives

$$\hat{\chi}^t = \hat{q}^t = \hat{m}^t = \frac{1}{\Delta + \rho(\overline{x}^2 + \sigma) - m_F^t m_x^t}\,. \tag{218}$$

This is the only equation in the state evolution that explicitly depends on the variance of the measurement noise Δ. Also eqs. (168-173) simplify for distributions (18) and (23) and under the Nishimori conditions they reduce to a pair of equations

$$m_F^{t+1} = \frac{\frac{1}{\eta} + \pi m_x^t \hat{m}^t}{(1+\frac{1}{\eta}) + \pi m_x^t \hat{m}^t}\,, \tag{219}$$

$$m_x^{t+1} = (1-\rho)\int \mathcal{D}z\, f_a^2\left(\frac{1}{\alpha m_F^t \hat{m}^t}, z\frac{1}{\sqrt{\alpha m_F^t \hat{m}^t}}\right) + \rho \int \mathcal{D}z\, f_a^2\left(\frac{1}{\alpha m_F^t \hat{m}^t}, z\frac{\sqrt{\alpha m_F^t \hat{m}^t + 1}}{\sqrt{\alpha m_F^t \hat{m}^t}}\right)\,. \tag{220}$$

Note that indeed only eq. (219) depends on the matrix uncertainty parameter η, and the dictionary learning limit $\eta \to \infty$ is straightforward. The MMSE of the matrix F, defined in (8), predicted by this state evolution is then E_F. The MMSE of the matrix X, defined in (6), is found equal to E_X

$$E_F = 1 - m_F \,, \qquad E_X = \rho(\overline{x}^2 + \sigma) - m_x \,. \tag{221}$$

The two sets of initial conditions that we will analyze to investigate the MMSE and the associated phase transitions are

- Random (uninformative) initialization: $m_X^{t=0} = 0$, and $m_F^{t=0} = 1/(1+\eta)$. Note that in the limit of dictionary learning $\eta \to \infty$ this initialization corresponds to a fixed point of the state evolution equations. This fixed point reflects the N permutational symmetry in the dictionary learning problem, and its instability corresponds to a spontaneous breaking of this symmetry. In the limit of dictionary learning we will hence initialize the state evolution with $m_F^{t=0}$ being a very small positive constant, and we will see the behavior will not depend on its precise value.
- Planted (informative) initialization: $m_X^{t=0} = \rho(\overline{x}^2 + \sigma) - \delta_x$, and $m_F^{t=0} = 1 - \delta_F$, where δ_x and δ_F are small positive constants to test the "stability" of the zero MMSE point.

The free entropy density from which we compute the limiting performance of the Bayes-optimal inference in case of a first order phase transition is expressed from (213) as follows. For simplicity from now on (till the end of this section V B) we analyze only the case where the mean of the elements of X^0 was zero, $\overline{x} = 0$, and the variance of the nonzero ones was one, $\sigma = 1$)

$$\begin{aligned}
\phi(m_x, m_F) &= -\frac{\alpha\pi}{2} + \frac{\alpha\pi}{2}\log\left(\Delta + \rho - m_x m_F\right) + \alpha\pi\frac{\Delta+\rho}{\Delta+\rho-m_x m_F} \\
&- \pi(1-\rho)\int \mathcal{D}z\, \log\left[1-\rho + \frac{\rho}{\sqrt{\alpha m_F \hat{m}+1}} e^{\frac{z^2 \alpha m_F \hat{m}}{2(\alpha m_F \hat{m}+1)}}\right] - \pi\rho\int \mathcal{D}z\, \log\left[1-\rho + \frac{\rho}{\sqrt{\alpha m_F \hat{m}+1}} e^{\frac{z^2 \alpha m_F \hat{m}}{2}}\right] \\
&- \frac{\alpha}{2}\left[\pi m_x \hat{m} - \log\left(1 + \frac{\pi m_x \hat{m}}{1+\frac{1}{\eta}}\right)\right] \,,
\end{aligned} \tag{222}$$

where $\hat{m}$ is given by eq. (218). The dependence on the matrix uncertainty is only in the last term and the limit of the completely unknown matrix F is easily taken by $\eta \to \infty$.

The simple Eqs. (218)-(222) is all we need at the end to analyze the MMSEs E_X and E_F of dictionary learning, blind matrix calibration, sparse PCA and blind source separation problems when the signal X^0 and F^0 are generated according to eqs. (18-19) with $\overline{x} = 0$ and $\sigma = 1$, ρ is the fraction of nonzero elements in X^0, and η is the matrix uncertainty from (22). The parameter $\alpha = M/N$ is the ratio between number of lines and the number of columns of F^0, $\pi = P/N$ is the ration between number of columns and the number of lines of X^0. The output channel has additive white Gaussian noise of variance Δ, and this information about distributions and their parameters is used in the posterior likelihood in the optimal Bayes inference.

2. *Phase diagram for blind matrix calibration and dictionary learning*

a. Indentifiability threshold in zero measurement noise: For the noiseless case $\Delta = 0$, and all positive $\eta > 0$, the linear stability analysis of eqs. (218-220) around the informative fixed point $m_x = \rho$, $m_F = 1$ (i.e. the MMSEs $E_F = E_X = 0$) leads to an update for the perturbations $\delta_F^{t+1} = (\delta_F^t \rho + \delta_x)/(\alpha\pi)$ and $\delta_x^{t+1} = \rho(\delta_F^t \rho + \delta_x)/\alpha$. By computations of the largest eigenvalue of the corresponding 2×2 matrix we obtain that this informative fixed point is stable if and only if

$$\pi > \pi^*(\alpha, \rho) \equiv \alpha/(\alpha - \rho) \,. \tag{223}$$

In other words the zero MSE fixed point is locally stable above the counting lower bound (21).

Further we notice that in the low noise limit $\Delta \to 0$, for all positive and finite η, and for $m_x = \rho - \delta_x$, $m_F = 1 - \delta_F$ with δ_x and δ_F being small positive constants of the same order as Δ the free entropy (222) becomes in the leading order $(\pi\alpha - \pi\rho - \alpha)\log(\Delta + \delta_x + \rho\delta_F)/2$. For $\pi > \pi^*(\alpha, \rho)$ this is a large positive value, and a large negative value for $\pi < \pi^*(\alpha, \rho)$.

Hence for the noiseless measurements $\Delta = 0$ the asymptotic Bayes-optimal MMSE is $E_X = 0$ and $E_F = 0$ for $\pi > \pi^*(\alpha, \rho)$ for all $\eta > 0$. This is a remarkable result as it implies that in the Bayes-optimal setting the dictionary is identifiable (or an exact calibration of the matrix possible) as soon as the number of samples P per signal element is larger than the value $\pi^*(\alpha, \rho)$ given by the trivial counting bound (21).

b. Indentifiability versus achievability gap: The next question is whether this MMSE is achievable in a tractable way. To answer this we study the state evolution starting in the uninformative initialization. First we analyze the behavior of the state evolution when $m_x = \delta_x$, and $m_F = \delta_F$ where both δ_x, and δ_F are positive and small, while we also consider η being very large. The linear expansion of state evolution update then leads to $\delta_x^{t+1} = \rho^2\alpha\delta_F/(\Delta+\rho)$ and $\delta_F = 1/\eta + \pi\delta_x/(\Delta+\rho)$. Hence for $\eta \to \infty$ the uninformative initialization is in fact a stable fixed point of the state evolution equations as long as

$$\pi \leq \pi_F \equiv (\Delta+\rho)^2/(\alpha\rho^2)\,. \tag{224}$$

This means that for $\pi^*(\rho,\alpha) < \pi < \pi_F(\Delta,\rho,\alpha)$ the MMSE is not achievable in the dictionary learning (e.g. when $\alpha < 1$ $\pi^*(0,\alpha) < \pi_F(0,0,\alpha)$) with the approximate message passing presented in this paper. This simple analysis leads us to the conclusion that a first order phase transition is in play in the dictionary learning problem, and as we will see also in the blind calibration ($\eta < \infty$) and sparse PCA ($\alpha > 1$).

As a side remark let us remind that the limit $\eta \to 0$ should lead to results known from Bayesian compressed sensing. In particular in compressed sensing for low noise the matrix X is identifiable if and only if $\alpha > \rho$. To reconcile this with the previous results notice that indeed for $\eta = c\Delta \to 0$ with $c = O(1)$ the leading term of the free entropy becomes $\pi(\alpha-\rho)\log(\Delta)$. Hence compressed sensing result is recovered. Whereas for $1 \gg \eta \gg \Delta$ it is the dictionary learning static phase transition $\pi^* = \alpha/(\alpha-\rho)$ that is the relevant one.

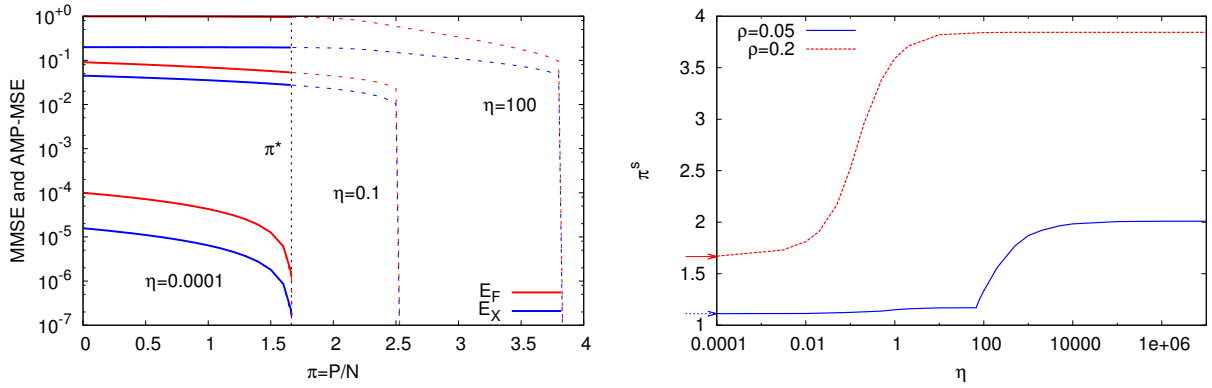

FIG. 2. Left: Blind matrix calibration: The predicted MSE E_F (for the matrix estimation) and E_X (for the signal estimation) corresponding to $\rho = 0.2$, $\alpha = 0.5$, $\Delta = 0$, for three values of η. MMSE is in full lines, AMP-MSE is in dashed lines. The MMSE jumps abruptly from a finite value to zero at the phase transition point π^* (223). However, the AMP-MSE matches the MMSE only when the sample complexity is larger than the spinodal transition, that takes place at a larger value $\pi^s(\eta) > \pi^*$. The AMP-MSE is zero for $\pi > \pi^s(\eta)$. Right: We plot the value of the spinodal transition at which the AMP-MSE has a discontinuity as function of η for $\alpha = 0.5$, $\Delta = 0$ and two different values of the sparsity ρ. Arrows on the left mark the static transition π^* and we see that $\lim_{\eta\to 0}\pi^s(\eta) = \pi^*$. In the limit of dictionary learning, $\eta \to \infty$, the spinodal transition converges to a finite value. Interestingly, for small values of the density, e.g. $\rho = 0.05$, we see a sharp phase transition in the threshold $\pi^s(\eta)$, in this case at $\eta \approx 68$.

c. Phase diagrams for blind calibration and dictionary learning Due to the close link between the two problems, we shall described the results for dictionary learning together with the case of blind matrix calibration. In both these cases we are typically trying to learn (calibrate) an overcomplete dictionary $\alpha < 1$ and a sparse signal X from as few samples P as possible. We hence first plot in Fig. 2 (left) the MSE for F (in red) and for X (in blue) as a function of $\pi = P/N$ and fixed (representative) value of undersampling ratio $\alpha = 0.5$, and density $\rho = 0.2$ in zero (or negligible) measurement noise, $\Delta = 0$. We consider several values of matrix uncertainty η (the larger the value the less we know about the matrix). The AMP-MSE achieved from the uninformative initialization is depicted in dashed lines, the MMSE achieved in this zero noise case from the planted (informative) initialization is in full lines.

Fig. 2 (right) shows the value of the spinodal transition π^s as a function of the matrix uncertainty η. We see that, as expected, $\lim_{\eta\to 0}\pi^s(\eta) = \pi^*$. A result that is less intuitive is that the large η limit is also well defined and finite $\lim_{\eta\to\infty}\pi^s(\eta) = \pi^s_{\rm DL} < 0$. This means that even in the dictionary learning where no prior information about the matrix elements is available the dictionary is identifiable with AMP for large system sizes above the spinodal transition $\pi^s_{\rm DL}$.

In Fig. 2 (right) we also see an interesting behavior in the function $\pi^s(\eta)$ for low values of ρ - there is a sharp phase transition from low η regime, where the AMP-MSE at the transition has a weak discontinuity towards a relatively low value of MSE, and a high η regime where the discontinuity is very abrupt towards a value of MSE that is close

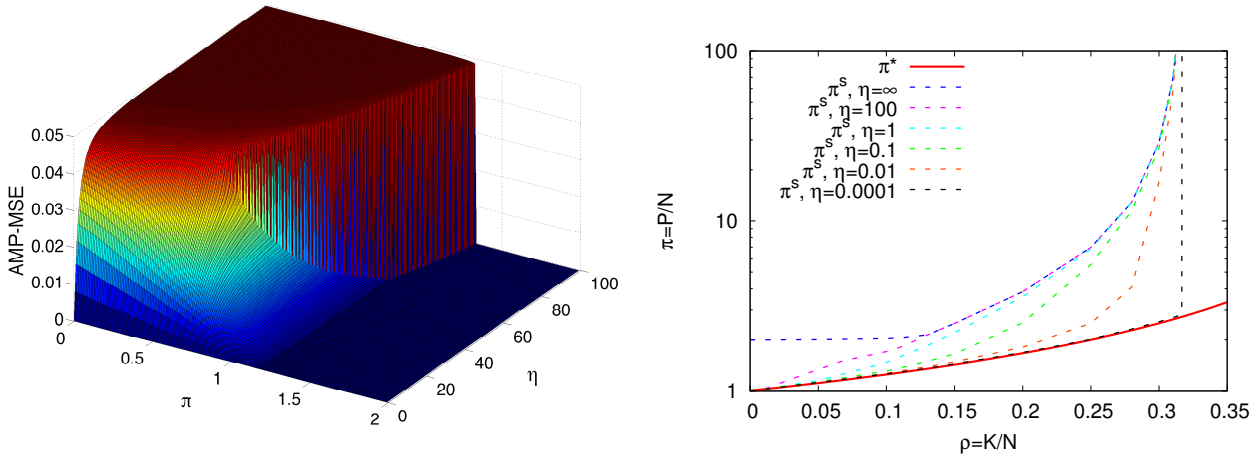

FIG. 3. Left: The AMP-MSE of the signal matrix X is plotted against π and η for the blind matrix calibration at $\rho = 0.05$, $\alpha = 0.5$ and $\Delta = 0$. The transition as a function of the number of samples π for fixed matrix uncertainty η is always discontinuous in this figure, for larger values of η this discontinuity is, however, much more pronounced. Right: The phase diagram of dictionary learning and blind matrix calibration for $\alpha = 0.5$ and $\Delta = 0$. In both cases, the matrix F is identifiable by the Bayes-optimal inference above the full red line $\pi^* = \frac{\alpha}{\alpha-\rho}$. However, the system undergoes the spinodal transition $\pi^s(\eta)$ shown here with a dashed line for dictionary learning ($\eta = \infty$) and for blind calibration (η finite). Below the spinodal transition, the MMSE is not achieved with AMP. Notice that for $\eta \to 0$ the spinodal line converges to π^* for $\rho < \rho_{\rm BP}^{\rm CS}$. For all values of η the spinodal line diverges as $\rho \to \rho_{\rm BP}^{\rm CS}$. The threshold value $\rho_{\rm BP}^{\rm CS} = 0.317$ for $\alpha = 0.5$ is the (spinodal) phase transition of pure compressed sensing (when the matrix F is fully known), from [23].

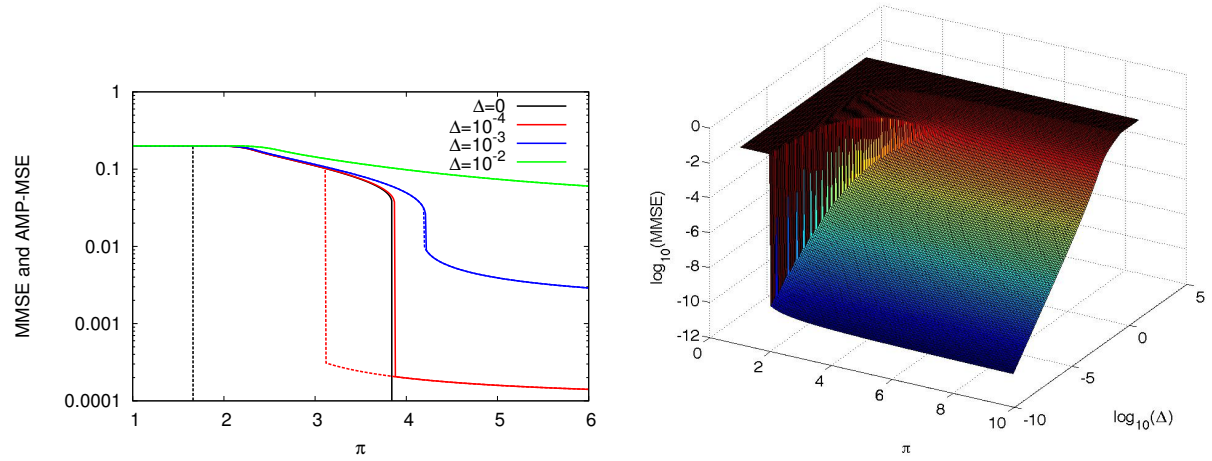

FIG. 4. Dictionary learning with noisy measurements. Left: The MMSE (dashed lines) and the AMP-MSE (full lines) for the signal matrix X as a function of the number of samples $P = \pi N$ for $\alpha = 0.5$, $\rho = 0.2$ for various values of the measurement noise Δ. The behavior is qualitatively the same as for the noiseless case, the only difference is that the MMSE at $\pi > \pi^*(\Delta)$, and the AMP-MSE at $\pi > \pi^s(\Delta)$ are no longer strictly zero but rather $O(\Delta)$. If the additive noise Δ is large enough, however, the phase transition disappears and the MMSE=AMP-MSE is continuous (this is the case in this plot for $\Delta = 10^{-2}$). Right: A surface plot of the MMSE of the signal E_X in the Δ, π-plane in the case $\alpha = 0.5$ and $\rho = 0.2$. Notice how the sharp transition disappears at large noise where it is replaced by a smooth evolution of the MMSE.

to the completely uninformative value. This behavior is also illustrated in Fig. 3 (left) where we plot the AMP-MSE as a function of π and η (for fixed $\rho = 0.05$, $\alpha = 0.5$, $\Delta = 0$).

Fig. 3 (right) depicts the phase diagram of dictionary learning ($\eta \to \infty$) and blind matrix calibration (finite η) we plot the static threshold π^* (η-independent) above which the matrix F is identifiable in full red, and the spinodal threshold π^s (for various values of η) above which the AMP identifies asymptotically the original matrix F is dashed line. Notice that $\pi^s(\rho)$ diverges as $\rho \to \rho_{\rm BP}^{\rm CS}$, where $\rho_{\rm BP}^{\rm CS}$ is the AMP phase transition in compressed sensing, see e.g.

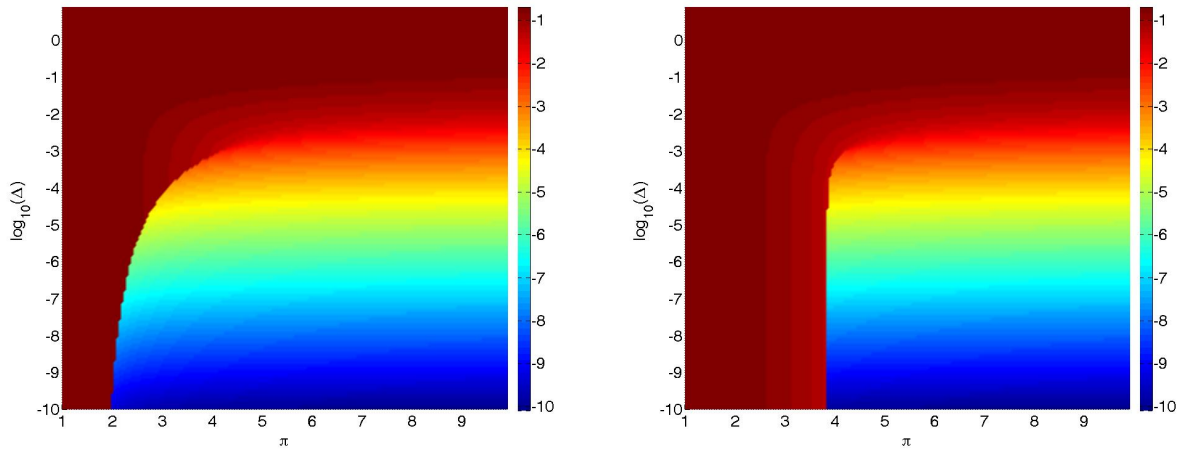

FIG. 5. Comparison of the MMSE (left) to the AMP-MSE (right) for noisy dictionary learning with $\alpha = 0.5$ and $\rho = 0.2$, the MSE of the signal matrix X is plotted on the Δ, π-plane. The color-scale is in decadic logarithm of the MSEs. We clearly see the region where MMSE<AMP-MSE and the one where the two are equal.

[23]. This is expected as for $\rho > \rho_{\rm BP}^{\rm CS}$ the sparse signal cannot be recovered with AMP even if the matrix F is fully known.

From now on (also in following subsections) we will discuss only the case $\eta \to \infty$ when no prior information about the dictionary F is available. In Fig. 4 (left) we illustrate the results for dictionary learning with non-zero measurement noise. The situation is qualitatively similar to what happens in noisy compressed sensing [23]. The first order phase transition is becoming weaker as the noise grows, until some value Δ^* above which there is no phase transition anymore and the AMP-MSE=MMSE for the whole range of π. In Fig. 4 (right) we then plot the AMP-MSE and the MMSE of the signal X as a function of the noise variance Δ and π for $\alpha = 0.5$ and $\rho = 0.2$. This surface plot demonstrates how the sharp transition disappears and is replaced by a continuous evolution of the MSE when the noise is large. This MMSE is compared to the AMP-MSE for the same case in Fig. 5.

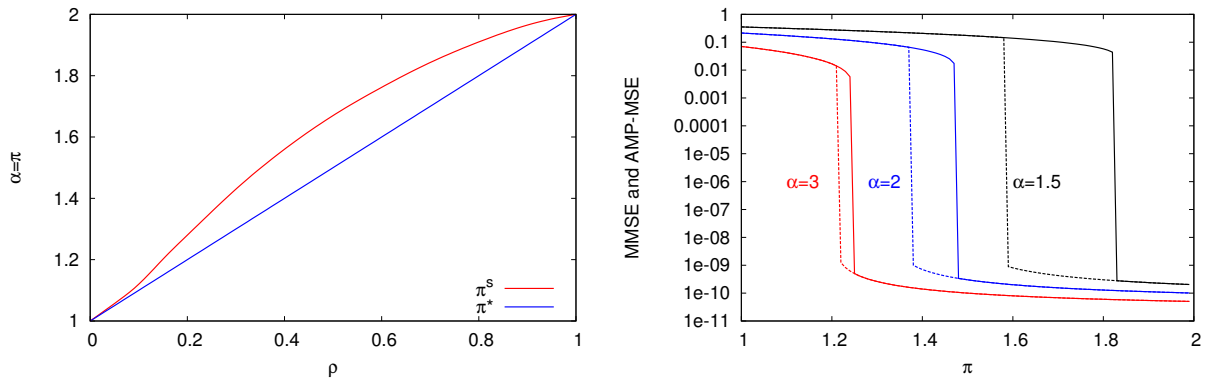

FIG. 6. Phase diagram for sparse PCA. Left: Spinodal transition π^s (upper red, where AMP-MSE goes to zero) and the optimal Bayes inference transition (lower blue, $\pi^* = \alpha^* = 1 + \rho$, where MMSE goes to zero) lines in zero measurement noise and when $\alpha = \rho$. Right: AMP-MSE and MMSE of the sparse matrix X in sparse PCA for $\rho = 0.5$, $\Delta = 10^{-10}$ and several values of ratio α as a function of π.

d. Phase diagram for sparse PCA Sparse PCA as we set it in Section I C is closely related to dictionary learning. Except that from the three sizes of matrices M, N and P the smallest one is the N – corresponding to the matrix Y to be of relatively low rank. Hence for sparse PCA we should only really consider $\alpha > 1$, $\pi > 1$.

Behavior of the state evolution is for this range of parameters qualitatively very similar to the one we just observed in the previous section. In Fig. 6 left we treat the case of $M = P$, i.e. $\pi = \alpha$, with zero measurement noise, and compute the smallest value for which the matrices F and ρ-sparse X are recoverable for a given ρ. Just as in the dictionary

learning we obtain that the Bayes optimal MMSE is zero everywhere above the counting bound $\pi = \alpha > 1 + \rho$, blue line in Fig. 6 left. The AMP-MSE is, however, zero only above the spinodal line, depicted in red in the figure. The gap between the two lines is not very large in this case.

The right part of Fig. 6 shows the AMP-MSE and the MMSE of the signal matrix X for measurement noise variance $\Delta = 10^{-10}$ and density $\rho = 0.5$.

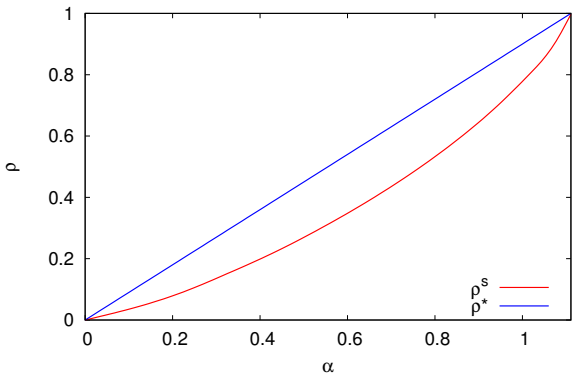

FIG. 7. Blind source separation. At the measurement noise $\Delta = 0$, $\pi = 10$ we plot the spinodal ρ^s as a function of α. Above this line AMP-MSE> 0, whereas below this line AMP-MSE$= 0$. The static transition line is $\rho^* = 0.9\alpha$.

e. Phase diagram for blind source separation In blind source separation, P corresponds to the length of the signal, N is the number of sources, and M the number of sensors. Typically the signal is very long, i.e. one has $P \gg N$ and $P \gg M$. The particularly interesting case is when there is more sources than sensors $\alpha < 1$ in that case the signal can be reconstructed only if the density of the signal ρ is smaller than a certain value. The counting bound gives us $\rho < \alpha(\pi - 1)/\pi$ and this also corresponds to the value under which the MMSE drops to zero under zero measurement noise. As in the previous case also here we observe a first order phase transition and with AMP we can reach zero error (in the noiseless case) only below ρ^s that we depict for $\pi = 10$ as a function of α in Fig. 7.

C. Low rank matrix completion

In the remaining examples we treat cases in which neither F nor X are sparse, $\rho = 1$, we start with low rank matrix completion.

In matrix completion the output function $g_{\rm out}$ is eq. (216) for the known matrix elements μl (there is ϵMP of them), and $g_{\rm out}(\omega, y, V) = 0$ for the unknown elements μl (there is $(1-\epsilon)MP$ of them). The input function f_a (214) for non-sparse matrix X, $\rho = 1$, becomes

$$f_a(\Sigma, T) = \frac{\overline{x}\Sigma + T\sigma}{\Sigma + \sigma}\,. \tag{225}$$

In the state evolution, under the Nishimori conditions, we then have

$$\hat{m}^t = \frac{\epsilon}{\Delta + \overline{x}^2 + \sigma - m_F^t m_x^t}\,, \tag{226}$$

where ϵ is the fraction of known elements of Y. Moreover for the input function (225) the state evolution equation for m_x becomes

$$m_x^{t+1} = \frac{\overline{x}^2 + (\overline{x}^2 + \sigma)\sigma\alpha m_F^t \hat{m}^t}{1 + \sigma\alpha m_F^t \hat{m}^t}\,. \tag{227}$$

The equation for m_F is the one of (219).

It is instrumental to analyze the local stability of the informative and the uninformative initialization for low rank matrix completion. For the informative initialization we consider $m_x^t = 1 - \delta_x^t$, and $m_F^t = 1 - \delta_F^t$, where δ_x^t, and δ_F^t are small positive numbers. The state evolution update equations at zero noise $\Delta = 0$ lead to $\delta_x^{t+1} = (\delta_x^t + \delta_F^t)/(\epsilon\alpha)$, and $\delta_F^{t+1} = (\delta_x^t + \delta_F^t)/(\epsilon\pi)$. The largest eigenvalue of this 2×2 system is $(\alpha + \pi)/(\epsilon\alpha\pi)$ and hence the informative

fixed point is stable for $\epsilon > \epsilon^* = (\alpha + \pi)/(\alpha\pi)$, which coincides with the counting bound eq. (25). This means that for $\epsilon > \epsilon^*$ the noiseless matrix completion, the matrices F and X can be recovered without error (asymptotically).

For the uninformative initialization we consider $m_x^t = \delta_x^t$, and $m_F^t = \delta_F^t$, where δ_x^t, and δ_F^t are again small positive numbers. This time the state evolution equations give $\delta_F^{t+1} = \pi\epsilon\delta_x^t/(1+\Delta)$ and $\delta_x^{t+1} = \alpha\epsilon\delta_F^t/(1+\Delta)$. Hence the uninformative fixed point is stable for $\epsilon < (\Delta+1)/\sqrt{\pi\alpha}$. This can indeed be verified in Fig. 8.

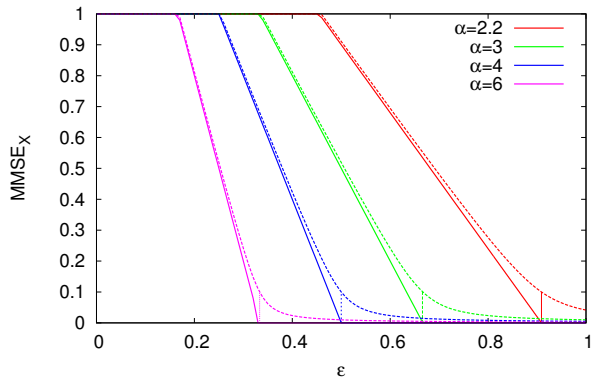

FIG. 8. Low rank matrix completion. The MMSE of the signal matrix X is plotted for $\alpha = \pi$ (for four different values of this parameter) in the noiseless (full line) and noisy (dashed lines), with $\Delta = 10^{-2}$, cases. In zero noise there is a second order phase transition which coincides with the counting bound, eq. (25), marked by short vertical lines. With non-zero noise there is no phase transition.

In matrix completion we treat matrices Y of low rank, hence N is much smaller than both P and M. The main questions concerns the fraction ϵ of elements that need to be known in order for the recovery of X and F to be possible. In this case we did not identify first order phase transition, as a result we have MMSE=AMP-MSE. Moreover we observed a phase transition in the MMSE only under zero measurement noise, and its position coincides with the counting bound (25), $\epsilon^* = (\alpha+\pi)/\alpha\pi$. With non-zero measurement noise the behavior of MMSE as a function of the other parameters is smooth and derivable (no phase transition). In Fig. 8 we plot an example of the MMSE as a function of the fraction of known elements ϵ for squared matrix Y, i.e $\alpha = \pi$. We generated the signal elements with zero mean and unit variance, $\overline{x} = 0$, $\sigma = 1$.

Our analysis suggest that compared to the cases with non-zero sparsity the low-rank matrix completion is a much easier problem, at least in the random setting considered in the present paper. The fact that the "counting" threshold can be saturated or close to saturated in the noiseless case by several algorithms can be seen e.g. in the data presented in [15].

D. Robust PCA

The input functions are the eq. (214) for matrix F and eq. (225) for matrix X. In robust PCA as defined by (27) we get for the output function

$$g_{\rm out}(\omega, y, V) = \frac{y-\omega}{V}\left[1 - \epsilon\frac{\Delta_s}{\Delta_s + V} - (1-\epsilon)\frac{\Delta_l}{\Delta_l + V}\right] . \tag{228}$$

The state evolution under the Nishimori condition becomes

$$\hat{m}^t = \frac{1}{\overline{x}^2 + \sigma - m_F^t m_x^t}\left[1 - \epsilon\frac{\Delta_s}{\Delta_s + \overline{x}^2 + \sigma - m_F^t m_x^t} - (1-\epsilon)\frac{\Delta_l}{\Delta_l + \overline{x}^2 + \sigma - m_F^t m_x^t}\right] . \tag{229}$$

Equation for m_F^t is (219), and for m_x^t is (227).

In robust PCA the informative initialization is again $m_x^t = 1 - \delta_x^t$, and $m_F^t = 1 - \delta_F^t$, where δ_x^t, and δ_F^t are small positive numbers. For small noise Δ_s and $\Delta_l = O(1)$ the corresponding fixed point is stable under the same conditions as for low rank matrix completion, i.e. for $\epsilon > \epsilon^* = (\alpha+\pi)/(\alpha\pi)$.

The uninformative fixed point is $m_x^t = \delta_x^t$, and $m_F^t = \delta_F^t$, where δ_x^t, and δ_F^t are again small positive numbers. This evolves as $\delta_x^{t+1} = \alpha\delta_F^t\hat{m}$, $\delta_F^{t+1} = \pi\delta_x^t\hat{m}$, with $\hat{m} = (1 + \delta_s = \epsilon\Delta_l - \epsilon\Delta_s)/[(1+\Delta_s)(1+\Delta_l)]$ in this limit. Hence for instance for $\Delta_s \to 0$, $\Delta_l = 1$ and $\pi = \alpha$ we have that the uninformative fixed point is stable for $\epsilon < 2/\alpha - 1$, which again corresponds to what we observe in Fig. 9.

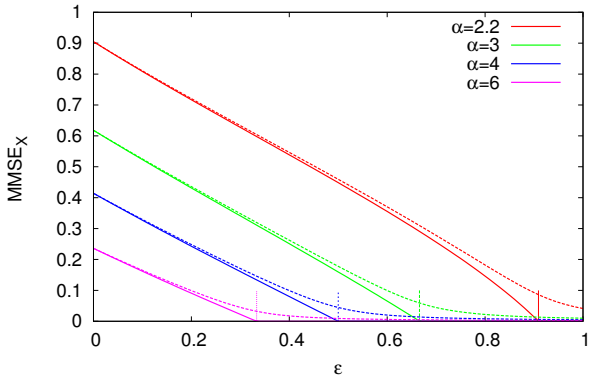

FIG. 9. MMSE of the matrix X in robust PCA with $\alpha = \pi$ in the noiseless (full line, $\Delta_s = 0$, $\Delta_l = 1$), and the noisy (dashed lines, with $\Delta_s = 10^{-2}$, $\Delta_l = 1$) cases. The situation is not very different from low rank matrix completion. Indeed in the noiseless case there is a second order phase transition and the MMSE is zero beyond the counting bound, marked by short vertical lines. In presence of noise there is no phase transition.

In the example of Fig. 9 we plot the MMSE as a function of the fraction of undistorted elements ϵ in the case of squared matrix Y, $\alpha = \pi$, the variance of the large distortions $\Delta_l = 1$ and two different values of the small measurement noise Δ_s. We see a second order phase transition at the counting bound for $\Delta_s = 0$ and a smooth decay of the MMSE for $\Delta_x > 0$.

It is interesting to compare how well robust PCA can be solved with respect to the matrix completion. In both cases ϵ is the fraction of known elements. The difference is that in matrix completion their position is known, whereas in robust PCA it is not. Intuitively the R-PCA should thus be a much harder problem. This is not confirmed in our analysis that instead suggest that robust PCA is as easy as matrix completion, since the zero noise phase transitions in the two coincide. Moreover, whereas at $\epsilon \to 0$ there is no information left in matrix completion (that is why the MMSE$= 1$), in robust PCA the largely distorted elements can still be explored and the MMSE< 1. Note, however, that algorithmically it seems less easy to saturate this theoretical asymptotic performance in R-PCA, see e.g. Figure 8 in [15].

E. Factor analysis

In factor analysis, the input functions f_r and f_a are the same as the dictionary learning (214) at $\rho = 1$. The output function (44) for factor analysis (28) is given by

$$g_{\rm out}(\omega_{\mu l}, y_{\mu l}, V_\mu) = \frac{y_{\mu l} - \omega_{\mu l}}{\psi_\mu + V_{\mu l}}, \tag{230}$$

where ψ_μ is the variance of the μ-th component of the unique factor. The variance of unique factor ψ_μ depends here on the index μ and does not on the index l, which leads to a slight modification in the derivation of the state evolution from section IV A. For simplicity, we assume that ψ_μ's are *known*; in practice, these should be estimated by the expectation and maximization scheme in conjunction with GAMP. Then, we obtain, on the Nishimori line

$$\hat{m}^t_\mu = \frac{1}{\psi_\mu + Q^0_{F,\mu}(\overline{x}^2 + \sigma) - m^t_{F,\mu} m^t_x}. \tag{231}$$

$$m^{t+1}_{F,\mu} = \frac{\pi Q^0_{F,\mu} m^t_x \hat{m}^t_\mu}{1 + \pi Q^0_{F,\mu} m^t_x \hat{m}^t_\mu}, \tag{232}$$

$$m^{t+1}_x = \frac{\alpha\sigma^2 \langle m^t_{F,\mu} \hat{m}^t_\mu \rangle_\mu}{1 + \alpha\sigma \langle m^t_{F,\mu} \hat{m}^t_\mu \rangle_\mu}, \tag{233}$$

where $\langle\cdot\rangle_\mu$ means the average over ψ_μ, the variance of the elements of the matrix F is denoted $Q^0_{F,\mu}$.

To analytically solve (233), one has to specify the distributions of ψ_μ and $Q^0_{F,\mu}$. We set $Q^0_{F,\mu} = Q^0_F = 1$ for all μ, and suppose a two-peak distribution for ψ as

$$P_\psi(\psi) = \epsilon\delta(\psi - \psi_1) + (1 - \epsilon)\delta(\psi - \psi_2). \tag{234}$$

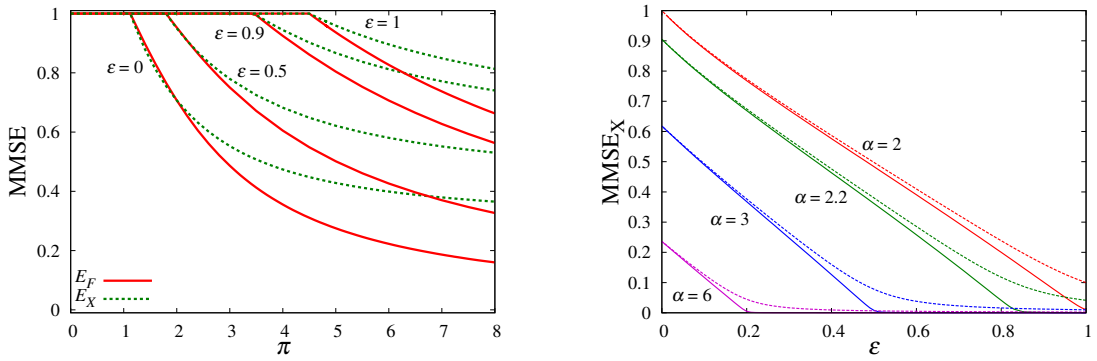

FIG. 10. Left: MMSE in factor analysis with $\psi_1 = 0.5$ and $\psi_2 = 2$, $\alpha = 2$. Right: MMSE of the signal matrix X in factor analysis at four different values of $\alpha = \pi$ at $\psi_1 = 10^{-2}$ and $\psi_2 = 1$ (dashed lines), and $\psi_1 = 10^{-5}$ and $\psi_2 = 1$ (solid lines).

Let us also assume $\overline{x} = 0$ and $\sigma = 1$. In this case the state evolution can be summarized as

$$m_{F1} = \frac{\pi m_x \hat{m}_1^t}{1 + \pi m_x \hat{m}_1^t}, \quad m_{F2} = \frac{\pi m_x \hat{m}_2^t}{1 + \pi m_x \hat{m}_2^t} \tag{235}$$

$$m_x = \frac{\alpha\{\epsilon m_{F1}\hat{m}_1^t + (1-\epsilon) m_{F2}\hat{m}_2^t\}}{1 + \alpha\{\epsilon m_{F1}\hat{m}_1^t + (1-\epsilon) m_{F2}\hat{m}_2^t\}}, \tag{236}$$

where

$$\hat{m}_1^t = \frac{1}{\psi_1 + 1 - m_{F1} m_x}, \quad \hat{m}_2^t = \frac{1}{\psi_2 + 1 - m_{F2} m_x}. \tag{237}$$

The total MMSE is given by $E_F = 1 - (\epsilon m_{F1} + (1-\epsilon) m_{F2})$ and $E_X = 1 - m_x$. Fig. 10 (left) shows the π-dependence of the MMSE at $\alpha = 2$, $\psi_1 = 0.5$, and $\psi_2 = 2$ for $\epsilon = 0$, 0.5, 0.9 and 1.

We analyze again the stability of the uninformative fixed point, $(m_x, m_{F_1}, m_{F_2}) = (0, 0, 0)$, of the state evolution. Small positive numbers δ_X, δ_{F_1}, and δ_{F_2} that give the uninformative initialization $m_x = \delta_X$, $m_{F_1} = \delta_{F_1}$, and $m_{F_2} = \delta_{F_2}$ evolve under the state evolution as

$$\delta_{F_1}^{t+1} = \frac{\pi \delta_X^t}{\psi_1 + 1}, \quad \delta_{F_2}^{t+1} = \frac{\pi \delta_X^t}{\psi_2 + 1}, \tag{238}$$

$$\delta_X^{t+1} = \alpha\pi \Big[\frac{\epsilon}{(\psi_1+1)^2} + \frac{1-\epsilon}{(\psi_2+1)^2} \Big] \delta_X^t. \tag{239}$$

These expressions indicate that the uninformative fixed point becomes unstable when $\alpha\pi \Big[\frac{\epsilon}{(\psi_1+1)^2} + \frac{1-\epsilon}{(\psi_2+1)^2} \Big] > 1$. The critical values of π given by this condition coincides with the transition point where the MMSE departs from 1 shown in Fig. 10 (left). As an example of ϵ-dependence, we show the MMSE of X at $\psi_1 = 10^{-2}$, $\psi_1 = 10^{-5}$ and $\psi_2 = 1$ for four different values of $\alpha = \pi$ in Fig. 10 (right). Consistently with our analysis, in these cases the uninformative initialization is always unstable,

The transition associated with the stability of the informative fixed point occurs only when at least one of ψs tends to be zero. For instance when $\psi_1 = 0$, the informative initialization corresponds to $\delta_{F_1} = 1 - m_{F_1}$ and $\delta_X = 1 - m_X$ that are given by

$$\delta_{F_1}^{t+1} = \frac{\delta_X^t + \delta_{F_1}^t}{\pi}, \quad \delta_X^{t+1} = \frac{\delta_{F_1}^t + \delta_X^t}{\alpha\epsilon} \tag{240}$$

without depending on ψ_2. These expressions mean that when $\epsilon > \epsilon^* = \pi/[\alpha(\pi - 1)]$, the informative fixed point is stable, consistently with Fig. 10 (right).

The state evolution of factor analysis for the two-peak case is qualitatively similar to that of robust PCA and low rank matrix completion, but the values of the phase transition points differ.

VI. DISCUSSION AND CONCLUSIONS

In this paper we have analyzed various examples of the matrix factorization problem. We obtain a matrix Y that is an element-wise noisy measurement of an unknown matrix $Z = FX$, where both Y, and Z are $M \times P$ matrices, F is a $M \times N$ matrix, and X is a $N \times P$ matrix. We have considered the tractability of this problem in the large size limit $N \to \infty$ while $\pi = P/N = O(1)$, and $\alpha = M/N = O(1)$. Our analysis concerns the teacher-student scenario where X and F are generated with random independent elements of some known probability distributions and we employ the Bayes-optimal inference scheme to recover F and X from Y.

Let us summarize our contribution: We derived the approximate message passing algorithm for matrix factorization. One version of the algorithm —for calibration and dictionary learning— was reported in [13], and a very related algorithm called Big-AMP was discussed by [15]. This algorithm is derived from belief propagation. We have presented the AMP for matrix factorization in several forms in Sections II A, and II B. We have also discussed simplifications that arise under the Nishimori conditions (Section II C), or when the matrix is large and one uses self-averaging of some of the quantities appearing in GAMP (Section II D).

Next to the AMP algorithm we have also derived the corresponding Bethe free entropy in Section III. The Bethe free entropy evaluated at a fixed point of the GAMP equations approximates the log-likelihood of the corresponding problem. We mainly use it in situations when we have more than one fixed point of GAMP, it is then the one with the largest values of the Bethe entropy that asymptotically given the MMSE of the Bayes-optimal inference. We also derived a variational Bethe free entropy in Section III B. This is a useful quantity that can serve in controlling the convergence of the AMP approach. Alternatively, a direct maximization of this expression is a promising algorithm itself (see [53] for an investigation of this idea for compressed sensing).

The AMP algorithm for matrix factorization is amenable to asymptotic analysis via the state evolution technique that was carried out rigorously for approximate message passing in compressed sensing [48]. We derive the state evolution analysis for matrix factorization using tools of statistical mechanics. In particular we use two approaches leading to equivalent results the cavity method (Section IV A) and the replica method (Section IV B). Our derivation of the state evolution is not rigorous, but we conjecture that it is nevertheless asymptotically exact as is the case in many other systems of this type including the compressed sensing. The main result of this state evolution are simple iterative equations that provide a way to compute the MMSE of the Bayes-optimal inference as well as the MSE reached theoretically in the large size limit by the AMP algorithm. The rigorous proof of the formulas derived in this paper is obviously an important topic for future work.

The main results of this paper concern analysis of MMSE and AMP-MSE for various interesting examples of the matrix factorization problem. We analyze the asymptotic phase diagrams for dictionary learning, blind matrix calibration, sparse PCA, blind source separation, matrix completion, robust PCA and factor analysis. Earlier results on this analysis appeared in [12, 13]. We find that when one of the matrices F or X is sparse the problems undergo a first order phase transition which is related to an interesting algorithmic barrier known for instance from compressed sensing [23].

It is a generic observation that for most of the problems we analyzed the theoretically achievable performance is much better than the one achievable by existing algorithms. The AMP algorithm should be able to match this performance for very large systems which is the most exciting perspective for further development of this work. If successful it could lead to an algorithmic revolution in various application of the matrix factorization.

In this paper we concentrate on the theoretical analysis and not on the performance of the algorithm itself. Some studies of the performance of some versions of the algorithm can be find in [13, 15]. We, however, observed that the performance depends strongly on the implementation details and we did not yet found a way to match the theoretically predicted performance for systems of treatable (practical) size in all cases.

It is worth discussing some of these algorithmic issues in this conclusion. One of the main problems it that the GAMP algorithm with parallel update presents instabilities that drive its evolution away from the so called Nishimori line; see a recent study of this issue in the compressed sensing problem [54]. This can be seen even in the state evolution when we do not assume explicitly that the result of the Bayes-optimal inference corresponds to a fixed point that belongs to the Nishimori line. There are ways how to avoid these issues, e.g. we observed that the difficulties basically disappear when the sequential update of the message passing algorithm from Section II A is used instead of the parallel one. This, however, does not scale very well with the systems size and our results were hence spoiled by very strong finite size effects. When learning the $M \times N$ matrix F, and the $N \times P$ matrix X we also often observed that a (very small) number of the P signals X_l were not correctly reconstructed, and these "rogue" vectors in X_l were polluting the reconstruction of F. An exemple of these finite size effects can be observed in [13].

In the work of [15] a part of the problems with convergence of the corresponding algorithm wes mitigated by adaptive damping (though maybe with not the most suitable cost function, see Sec. III) and expectation maximization learning. Why this is helpful is theoretically explained in the recent work [54] for compressed sensing. However, the implementation of [15] does not match the theoretical performance predicted in this paper either (we have explicitly

tried for the dictionary learning and robust PCA examples). This shows that more work is needed in order to reach a practical algorithm able to achieve the prediction at moderate sizes. A more proper understanding of these problems, and further developments of the algorithm is therefore the main direction of our future work.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Research Council under the European Union's 7^{th} Framework Programme (FP/2007-2013)/ERC Grant Agreement 307087-SPARCS. Financial support by the JSPS Core-to-Core Program "Non-equilibrium dynamics of soft matter and information" and JSPS/MEXT KAKENHI Nos. 23-4665 and 25120013 is also acknowledged.

[1] Olshausen B. A. *et al.* Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
[2] Olshausen B. A. & Field D. J. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1. *Vision Research* **37**, 3311 (1997).
[3] Kreutz-Delgado K. *et al.* Dictionary learning algorithms for sparse representation. *Neural computation* **15**, 349–396 (2003).
[4] Zou H., Hastie T. & Tibshirani R. Sparse principal component analysis. *Journal of computational and graphical statistics* **15**, 265–286 (2006).
[5] Belouchrani A., Abed-Meraim K., Cardoso J.-F. & Moulines E. A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on* **45**, 434–444 (1997).
[6] Candès E. J. & Recht B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* **9**, 717–772 (2009).
[7] Candès E. J. & Tao T. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on* **56**, 2053–2080 (2010).
[8] Candès E. J., Li X., Ma Y. & Wright J. Robust principal component analysis? *Journal of the ACM (JACM)* **58**, 11 (2011).
[9] Donoho D. L., Maleki A. & Montanari A. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.* **106**, 18914–18919 (2009).
[10] Mézard M., Parisi G. & Virasoro M. A. *Spin-Glass Theory and Beyond*, vol. 9 (World Scientific, Singapore, 1987).
[11] Mézard M. & Montanari A. *Information, Physics, and Computation* (Oxford Press, Oxford, 2009).
[12] Sakata A. & Kabashima Y. Sample complexity of Bayesian optimal dictionary learning. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 669–673 (2013).
[13] Krzakala F., Mézard M. & Zdeborová L. Phase diagram and approximate message passing for blind calibration and dictionary learning. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 659–663 (2013). Preprint arXiv:1301.5898.
[14] Schniter P., Parker J. & Cevher V. Bilinear Generalized Approximate Message Passing (BiG-AMP) for Matrix Recovery Problem. In *Workshop on Information Theory and Applications (ITA), San Diego CA* (2012).
[15] Parker J. T., Schniter P. & Cevher V. Bilinear generalized approximate message passing. *arXiv preprint arXiv:1310.2632* (2013).
[16] Elad M. Sparse and Redundant Representation Modeling—What Next? *Signal Processing Letters* **19** (2012).
[17] Opper M. & Haussler D. Calculation of the Learning Curve of Bayes Optimal Classification Algorithm for Learning a Perceptron With Noise. In *In Computational Learning Theory: Proceedings of the Fourth Annual Workshop*, 75–87 (Morgan Kaufmann, 1991).
[18] Iba Y. The Nishimori line and Bayesian statistics. *Journal of Physics A: Mathematical and General* **32**, 3875 (1999).
[19] Nishimori H. *Statistical Physics of Spin Glasses and Information Processing* (Oxford University Press, Oxford, 2001).
[20] Mézard M., Parisi G. & Zecchina R. Analytic and Algorithmic Solution of Random Satisfiability Problems. *Science* **297**, 812–815 (2002).
[21] Montanari A. Estimating random variables from random sparse observations. *European Transactions on Telecommunications* **19**, 385–403 (2008).
[22] Dempster A., Laird N. & Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society* **38**, 1 (1977).
[23] Krzakala F., Mézard M., Sausset F., Sun Y. & Zdeborová L. Probabilistic Reconstruction in Compressed Sensing: Algorithms, Phase Diagrams, and Threshold Achieving Matrices. *J. Stat. Mech.* (2012).
[24] Vila J. P. & Schniter P. Expectation-Maximization Bernoulli-Gaussian Approximate Message Passing. In *Proc. Asilomar Conf. on Signals, Systems, and Computers (Pacific Grove, CA)* (2011).
[25] Jöreskog K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**, 183–202 (1969).
[26] Lewicki M. S. & Sejnowski T. J. Learning overcomplete representations. *Neural computation* **12**, 337–365 (2000).

[27] Aharon M., Elad M. & Bruckstein A. M. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing* **54**, 4311 (2006).

[28] Michal Aharon, Michael Elad A. M. B. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and its Applications* **416**, 48–67 (2006).

[29] Vainsencher D., Mannor S. & Bruckstein A. M. The Sample Complexity of Dictionary Learning. *Journal of Machine Learning Research* **12**, 3259–3281 (2011).

[30] Jenatton R., Gribonval R. & Bach F. Local stability and robustness of sparse dictionary learning in the presence of noise. *arXiv:1210.0685* (2012).

[31] Spielman D. A., Wang H. & Wright J. Exact recovery of sparsely-used dictionaries. *arXiv preprint arXiv:1206.5882* (2012).

[32] Arora S., Ge R. & Moitra A. New algorithms for learning incoherent and overcomplete dictionaries. *arXiv preprint arXiv:1308.6273* (2013).

[33] Agarwal A., Anandkumar A. & Netrapalli P. Exact Recovery of Sparsely Used Overcomplete Dictionaries. *arXiv preprint arXiv:1309.1952* (2013).

[34] Gribonval R., Jenatton R., Bach F., Kleinsteuber M. & Seibert M. Sample complexity of dictionary learning and other matrix factorizations. *arXiv preprint arXiv:1312.3790* (2013).

[35] Lee T.-W., Lewicki M. S., Girolami M. & Sejnowski T. J. Blind source separation of more sources than mixtures using overcomplete representations. *Signal Processing Letters, IEEE* **6**, 87–90 (1999).

[36] Zibulevsky M. & Pearlmutter B. A. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation* **13**, 863–882 (2001).

[37] Bofill P. & Zibulevsky M. Underdetermined blind source separation using sparse representations. *Signal processing* **81**, 2353–2362 (2001).

[38] Georgiev P., Theis F. & Cichocki A. Sparse component analysis and blind source separation of underdetermined mixtures. *Neural Networks, IEEE Transactions on* **16**, 992–996 (2005).

[39] d'Aspremont A., El Ghaoui L., Jordan M. I. & Lanckriet G. R. A direct formulation for sparse PCA using semidefinite programming. *SIAM review* **49**, 434–448 (2007).

[40] Gribonval R., Lesage S. *et al.* A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges. In *ESANN'06 proceedings-14th European Symposium on Artificial Neural Networks*, 323–330 (2006).

[41] Candes E. J. & Plan Y. Matrix completion with noise. *Proceedings of the IEEE* **98**, 925–936 (2010).

[42] Keshavan R. H., Montanari A. & Oh S. Matrix completion from a few entries. *Information Theory, IEEE Transactions on* **56**, 2980–2998 (2010).

[43] Chandrasekaran V., Sanghavi S., Parrilo P. A. & Willsky A. S. Sparse and low-rank matrix decompositions. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, 962–967 (IEEE, 2009).

[44] Yuan X. & Yang J. Sparse and low-rank matrix decomposition via alternating direction methods. *preprint* (2009).

[45] Wright J., Ganesh A., Rao S., Peng Y. & Ma Y. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, 2080–2088 (2009).

[46] Donoho D., Maleki A. & Montanari A. Message passing algorithms for compressed sensing: I. motivation and construction. In *IEEE Information Theory Workshop (ITW)*, 1 –5 (2010).

[47] Rangan S. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2168 –2172 (2011).

[48] Bayati M. & Montanari A. The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing. *IEEE Transactions on Information Theory* **57**, 764 –785 (2011).

[49] Thouless D. J., Anderson P. W. & Palmer R. G. Solution of 'Solvable Model of a Spin-Glass'. *Phil. Mag.* **35**, 593–601 (1977).

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

[51] Yedidia J., Freeman W. & Weiss Y. Understanding Belief Propagation and Its Generalizations. In *Exploring Artificial Intelligence in the New Millennium*, 239–236 (Morgan Kaufmann, San Francisco, CA, USA, 2003).

[52] Rangan S., Schniter P., Riegler E., Fletcher A. & Cevher V. Fixed Points of Generalized Approximate Message Passing with Arbitrary Matrices. *arXiv preprint arXiv:1301.6295* (2013).

[53] Krzakala F., Manoel A., Tramel E. W. & Zdeborová L. Variational Free Energies for Compressed Sensing (2014). Submitted.

[54] Caltagirone F., Krzakala F. & Zdeborová L. On Convergence of Approximate Message Passing (2014). ArXiv:1401.6384 [cs.IT].