Matricial and tensorial factorisation using tools coming from statistical physics

Thibault Lesieur

NNT : 2017SACLS345

Thèse de doctorat

Factorisation matricielle et tensorielle par une approche issue de la physique statistique

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'université Paris-Sud

École doctorale n°564
École Doctorale *Physique en Île-de-France (EDPIF)*
Spécialité de doctorat: Physique

Thèse présentée et soutenue à Gif-sur-Yvette le 9 octobre 2017 par

Thibault Lesieur

Composition du Jury :

M, Silvio, Franz Professeur, Paris Sud	Président
M, Manfred Opper Professeur, TU Berlin	Rapporteur
M, Coja-Oghlan, Amin Professeur, Goethe Universität	Rapporteur
M, Marc, LeLarge Chargé de recherche, INRIA	Examinateur
M, Krzakala, Florent Professeur, ENS Paris LPS	Examinateur
Mme, Lenka, Zdeborová Chargé de recherche, IPHT Cea Saclay	Directrice de thèse

Abstract

In this thesis I present the result of my work on the problem of low-rank matrix factorization and low-rank tensor matrix factorization. Because one often tries to analyze data that are encoded in a matrix form, matrices ends up being an ubiquitous object in Machine Learning. Approximating a matrix by some form of low-rank approximation (factorizing) is therefore one of the basic tasks that one could end up doing when treating such problems. Basic spectral techniques to solve this problem such as Principal Component Analysis (PCA) find their usefulness in that they are quite model agnostic and make little assumption about the structure of the underlying data. In this thesis we present a way to use such prior knowledge of the data in the setting of Bayesian inference. We will treat these inference problem using tools coming from statistical physics. We will give algorithmic solutions to solve these problem in an efficient way. We will analyze theoretically these systems and uncover the "zoology of phase diagrams" that they can exhibit.

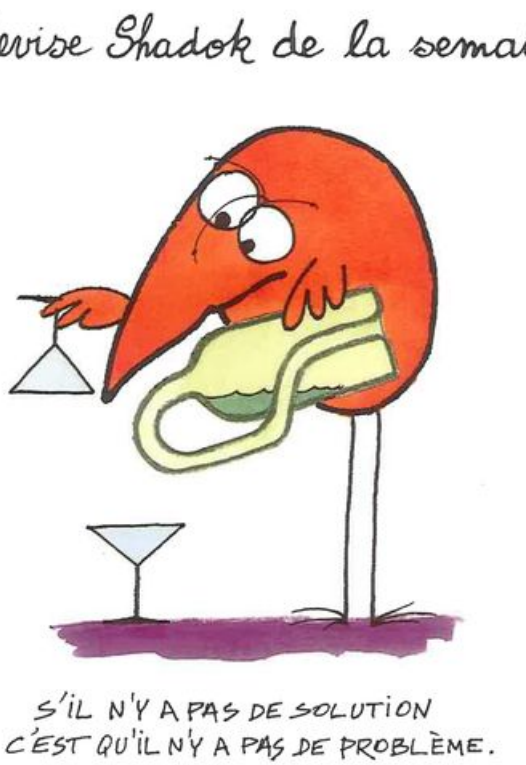
Acknowledgements

I would like to express my gratitude to my supervisor Lenka Zdeborová for giving me the opportunity to study these interesting problems as well as for all the support she provided me during these three and a half year. I would also like to thank the members of the sphinx team and members of IPHT lab for all these 3 years : Florent Krzakala, Christophe Schülke, Lais Sarem-Schunk, Soumya Sasmal, Christian Schmidt, Eric Tramel, Marylou Gabrié, Andre Manoel, Alaa Saade, Laura Fioni and Alejandro Lage Castellanos.

Special thanks go to Laure Sauboy and Sylvie Zaffanella for helping me navigating the administrative maze of the CEA and Paris-Sud during these 3 years.

I would like to thank my friend and roommate Félix Rose for his friendship and all the scientific discussions that we had.

My thanks also go to my family for their support during these three years.



'If there is no solution it's because there is no problem.'
Shadok proverb

Table des matières

Chapitre 1

Introduction

Chapitre 2

Introduction

The last years have seen an explosion of the amount of data collected in different industries.

At the time of writing of this thesis Facebook counts 1.98 billions active users that log in at least once a month [FB]. With a current estimate of the world population sitting at 7.5 billion people this means that 26% of the world population is an active user of this social network. The analysis alone of the graph (of size $N = 1.98 \times 10^9$) formed by the "friendship" connection between people contains an immense amount of data that Facebook seems to have found a usefulness for [UKBM11].

The availability of big training set combined with advance in Machine Learning (ML) techniques has allowed people to train deep neural networks with success.

Be they collected through the usage of social network, "smart" object or even collected by hand, this sudden availability of data has created a world in which data are now seen as valuable resources. The field that deals with the techniques that allow one to transform raw data into a human face classifier or uncover the existence of clusters in a data set or even program a computer to play Go [SHM+16] is called Machine Learning.

Machine learning problem tends to be separated into 3 different types.

— **Supervised learning** is the part of Machine learning that deals with learning a function $f(x)$ from a set of examples of N examples (x_i, y_i). One looks for a function f such that the $f(x_i) \approx y_i$. What does one mean by the sign "$\approx$" here, is a problem dependent question. In practice the supervised learning often translates into an optimization problem of the form

$$f^* = \operatorname{argmin} R(f) = \operatorname{argmin} \left\{ \frac{1}{N} \sum_{1 \leq i \leq N} L(f(x_i), y_i) + \Lambda(f) \right\} \tag{2.1}$$

$L(.,.)$ can be seen as a distance that ensures that $f(x_i)$ will remain "close" to y_i. The space in which we look for a minima has to be restrained. There can be multiple ways to do that. One way is to restrain the search to functions f that can be expressed in a certain way. For instance one could restrain the search to functions that are linear in x, $f(x) = \beta^\top x + c$ with $x, \beta \in \mathbb{R}^r \, c \in \mathbb{R}$. The Λ term is also here to restrict the space of solutions by penalizing some values of f. The field of ML possesses a large array of

techniques to express a function f (to name but a few, nearest neighbor approximation, perceptron, neural network, decision trees, etc.). There are also ways to combine the result of different function $f_1(x), f_2(x), \cdots$ in order to create a new function f with greater expressivity. These techniques are called boosting techniques.

— **Unsupervised learning** is the part of ML that deals with the analysis of data and the search for hidden structure in a data set. The difference between supervised learning and unsupervised learning is illustrated in fig 2.1. In unsupervised learning one does not try to predict or learn a function but tries to understand the structure of available data. A good approximation would be to say that unsupervised learning deals with the learning of the density probability (or generative model) from which we assume the observed data was created. Examples of techniques coming from unsupervised learning include k-means clustering, latent, restricted Boltzmann machine, Principal Component Analysis, etc.

— **Reinforcement learning** is a part of Machine Learning that overlaps with control theory. In that setting one tries to teach an agent how to behave in external environment in order to maximize some reward all the while having only partial information of the environment.

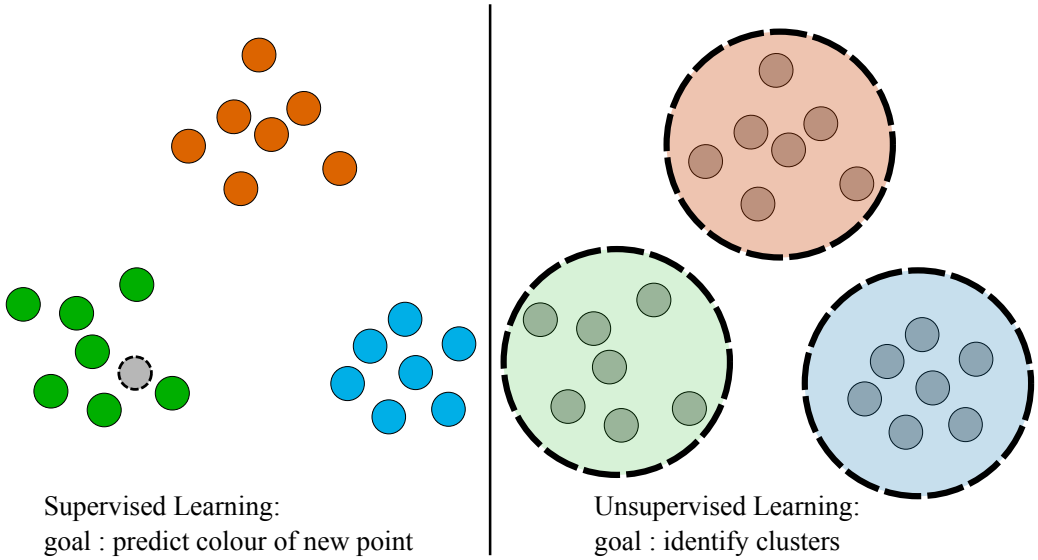


FIGURE 2.1 – **Left pannel :** Here we illustrate the concept of supervised learning. We have been given a collection of points each of a different colour. We are then provided with a new point of unknown colour (the grey one with dashed contour). Supervised learning is about answering the question "what is the colour of the new point ?" **Right pannel :** Here we illustrate the concept of unsupervised learning. We have been given a collection of points. Using techniques such as k-means one could show that this distribution of points form clusters (materialised by transparent coloured circles) and could therefore be well explained by a mixture of Gaussians model.

It is worth noting that the distinction into 3 separate "types" of Machine Learning problem is a distinction that has more to do with what one wants to do with the data and less with the type of techniques and insights one might use in the resolution (essentially because nearly all ML problem can be translated into an optimization problem). For example some problems of optimal control can be translated into problem of inference on graphical models [KGO12].

Machine Learning sits at the intersection of many fields. Some might see it as only a renaming of statistics, whereas others might see it only as a collection of techniques to solve engineering

problems. Some might even have a more "free form" vision of the field [A free form approach].

Despite its success there is little theoretical understanding of why Machine Learning techniques work. In the last years new works coming from Statistical physics have tried to fill that theoretical understanding gap. By analyzing certain systems they hope to shed some light on the inner workings of Machine Learning techniques. The reason why this might be a fruitful approach is that one can often find a translation between concept in Statistical physics and concept in ML problems. For instance solving an optimization problem is the same thing as looking for the ground state of an Hamiltonian. We will see in the next chapter that this link between the two fields goes beyond that.

This thesis belongs to this "collection" of work that makes overlap statistical physics, Machine learning and Information theory. The idea is to "import" intuitions, tools and techniques from statistical physics in order to solve and analyze problems in Machine Learning. Here are a few examples of such work [DMM09, KMM$^+$13]. A reader interested in studying these question in more details might find these books to be most precious resources [Nis01, MPV87, MM09].

In this thesis I present the result of my work on the problem of low-rank matrix/tensor factorization. In order to explain the motivation for this work, let me start from a typical unsupervised learning problem. An unsupervised learning problem is a problem where one is given data about a problem and where one attempts to find some underlying structure in the provided data. Let us give some example of such problems and of "typical" structure we would be interested in.

— **Community detection** : Suppose one is given some undirected network of size N. This network is coming from a social network such as Facebook. The nodes of the network represent people and a link between two nodes mean that these two people are friend on Facebook. We note Y the adjacency matrix $Y \in \mathbb{R}^{N\times N}$ of the network. An analysis of the network might uncover for instance an associative structure where it is possible to assign to each of the N node a color. Nodes of the same color will form "communities" and will on average have more links to nodes of the same color compared to nodes of different color. In such a case maybe one would just have uncovered some feature of real life and these two communities might map to the fact that people speak different languages or belong to different social classes.

— **Clustering** : An example of an unsupervised learning problem is the following. One is given some M data points representing some information of people entering an hospital. Information about each patient (tension, hearth rate etc.) is encoded in a vector $x_i \in \mathbb{R}^N$. We could collect all these data points in a matrix $Y \in \mathbb{R}^{N\times M}$. An analysis of these data points might uncover some cluster like structure. This would tell us that one could model incoming patient i as to belonging to a certain class depending on the cluster to which x_i belongs.

In both these cases we encode the data we want to analyze either in a square symmetric matrix $Y \in \mathbb{R}^{N\times N}$ or in a rectangular matrix $Y \in \mathbb{R}^{N\times N}$. Because one often ends up trying to analyze data that is encoded in a matrix form, matrices end up being an ubiquitous object in Machine Learning. We hope that in the matrix Y some form of structure that we want to discover is hidden. By "structure" we mean a low-rank structure this means that Y is the combination of an underlying signal and of random noise. In practice this means that we expect Y to have the following structure.

$$\mathbf{XX^\top}\ \textbf{Symmetric case :}\ Y = XX^\top + \mathcal{N}_{\mathrm{Sym}}(0,\Delta)^{N\times N} \tag{2.2}$$

$$\mathbf{UV^\top}\ \textbf{case :}\ Y = UV^\top + \mathcal{N}(0,\Delta)^{N\times M} \tag{2.3}$$

Where $X, U \in \mathbb{R}^{N\times r}$ and $V \in \mathbb{R}^{M\times r}$. $\mathcal{N}(0,\Delta)^{N\times M}$ is a random Gaussian matrix where each entry is a Gaussian variable of mean 0 and variance Δ. $\mathcal{N}_{\text{Sym}}(0,\Delta)^{N\times N}$ is a symmetric Gaussian random matrix with elements outside the diagonal of variance Δ. Knowing Y we want to find back X or U and V. Y can be seen as the perturbation of a rank r matrix by random noise.

One of the most simple way to solve this problem is to use a method called PCA where one looks for the matrix $\tilde{Y}$ which minimize the Frobenius norm to Y among all matrices of rank r. This is equivalent to using the rescaled eigenvectors of Y as estimates for X, U and V.

A technique such as PCA has the benefit to be relatively model agnostic meaning the we make little or no assumption about the structure of matrices X, U and V. It can often prove useful to incorporate some prior knowledge on X, U and V in our model. To do this we will look at this problem from the angle of Bayesian inference. We will suppose that Y was created according to the following process.

$$\begin{aligned}
&\forall i \in [1;N], \text{Sample}\, x_i \in \mathbb{R}^{r\times 1} \sim P_{X_0}(x_i) \Rightarrow w_{ij} = \frac{1}{\sqrt{N}} x_i^\top x_j \\
&\Rightarrow P\left(Y|\{w_{ij}\}\right) = \prod_{1\le i<j\le N} P_{\text{out}}(Y_{ij}|w_{ij}) \qquad (2.4)\\
&\begin{cases} \forall i\in[1;N], \text{Sample}\, u_j \in \mathbb{R}^{r\times 1} \sim P_{U_0}(u_i) \\ \forall j\in[1;M], \text{Sample}\, v_j \in \mathbb{R}^{r\times 1} \sim P_{V_0}(v_j) \end{cases} \Rightarrow w_{ij} = \frac{1}{\sqrt{N}} u_i^\top v_j \Rightarrow P\left(Y|\{w_{ij}\}\right) = \prod_{\substack{1\le i\le N\\ 1\le j\le M}} P_{\text{out}}(Y_{ij}|w_{ij}) \qquad (2.5)
\end{aligned}$$

The main difference with equations (2.2, 2.3) is that we assume $P(X)$ to have been sampled from some density distribution. We also assume a more general form for the noise model with the conditional probability function $P_{\text{out}}(Y|w)$. We will treat this problem using Bayesian inference. This yields us with

$$P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\le i\le N} P_X(x_i) \prod_{1\le i<j\le N} \exp\left(g\left(Y_{ij}\middle|\frac{x_i^\top x_j}{\sqrt{N}}\right)\right), \qquad (2.6)$$

$$P(U,V|Y) = \frac{1}{Z_{UV}(Y)} \prod_{1\le i\le N} P_U(u_i) \prod_{1\le j\le M} P_V(v_j) \prod_{1\le i\le N, 1\le j\le M} \exp\left(g\left(Y_{ij}\middle|\frac{u_i^\top v_j}{\sqrt{N}}\right)\right). \qquad (2.7)$$

Here we have replaced P_{X_0} by P_X, P_{U_0} by P_U, P_{V_0} by P_V and $P_{\text{out}}(Y|w)$ by $\exp(g(Y,w))$. The Bayes optimal case correspond to the case where

$$P_X = P_{X_0},\ P_U = P_{U_0},\ P_V = P_{V_0},\ P_{\text{out}}(Y|w) = e^{g(Y,w)} \qquad (2.8)$$

$P_{X_0}, P_{U_0}, P_{V_0}$ and P_{out} are the density probability that were used to create Y. In most of this thesis we will assume to be in the Bayes optimal setting (2.8). The term "optimal" here means that since we know from what model was sampled Y we can compute the exact posterior probability $P(X|Y)$ but more importantly we will be able to compute the optimal estimator of $\hat{X}(Y)$ that minimizes some average error. For instance the $\hat{X}_{\text{MSE}}(Y) = \mathbb{E}_{X\sim P(X|Y)}[X]$ is the estimator of X that minimizes the average squared norm between X_0 and $\hat{X}_{\text{MSE}}(Y)$. Though analyzing the problem in a more general setting will prove useful.

The factor $1/\sqrt{N}$ in the second argument of the function g ensures that the behavior of the above models is non-trivial and that there is an interesting competition between the number $O(N)$ of local magnetic fields P_X, P_U, P_V and the number of $O(N^2)$ interactions. To physics

readership familiar with the Sherrington-Kirkpatrick (SK) model this $1/\sqrt{N}$ factor will be familiar because in the SK model the interaction between the Ising spins that lead to extensive free energy are also of this order (with mean that is of order $1/N$). This is compared to the ferromagnetic Ising model on a fully connected lattice for which the interactions leading to extensive free energy scale as $1/N$.

For readers interested in inference problems, i.e. the planted setting, the $1/\sqrt{N}$ factor is the scaling of the signal-to-noise ratio for which inference of $O(N)$ unknown from $O(N^2)$ measurements is neither trivially easy nor trivially impossible. In the planted setting Y can be viewed as a random matrix with a rank-r perturbation. The regime where the eigenvalues of dense random matrices with low-rank perturbations split from the bulk of the spectral density is precisely when the strength of the perturbation is $O(1/\sqrt{N})$, see e.g. [BBAP05].

We will also treat the problem of low-rank tensor factorization, in which we now observe a symmetric tensor Y of order p (symmetric here means that Y is invariant with respect to all permutation of the p indices $i_1, \cdots, i_p$). We suppose that Y exhibit a low-rank structure defined by the following generative process.

$$\begin{aligned}
&\forall i \in [1; N], \text{Sample}\, x_i \in \mathbb{R}^{r\times 1} \sim P_X(x_i) \Rightarrow \\
&\Rightarrow w_{i_1\cdots i_p} = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{k=1\cdots r} (x_{i_1} \circ \cdots \circ x_{i_p})_k \Rightarrow P\left(Y|\{w_{ij}\}\right) = \prod_{1\le i<j\le N} P_{\text{out}}(Y_{ij}|w_{ij}) \qquad (2.9)\\
&P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\le i\le N} P_{X_0}(x_i) \prod_{1\le i<j\le N} P_{\text{out}}\left(Y_{ij}\middle|\frac{x_i^\top x_j}{\sqrt{N}}\right). \qquad (2.10)
\end{aligned}$$

The $\circ$ symbol is the Hadamard product. The $\sqrt{(p-1)!}$ is just here so that we get more convenient equations.

We will often talk in this thesis of the $XX^\top$ $UV^\top$ and X^p case to designate systems of types 2.6 2.7 2.10 respectively.

2.1 Organization of the manuscript

This thesis is organized in 2 main parts.

- **In the General Theory chapter 3** I will introduce the theoretical tools that we will use all throughout the thesis. This chapter can be thought of as having two main parts.
  - In 3.1 we will introduce the concept of Inference, both frequentist and Bayesian inference. we will talk about estimators and about what it means for an estimator to be optimal.
  - In the second part 3.2 I present the tools, model and techniques coming from statistical physics that will prove useful in order to describe analyze and treat the different statistical physics problem that we will encounter. This will include sections on Mean-field and Plefka method 3.3, A section on factor graphs 3.4, a section on Belief Propagation (BP) 3.5, a section on the Sherrington-Kirkpatrick model and spin glasses 3.6 and a section on the replica method 3.7.

— **The Results chapter 4** of this thesis will mainly contain the results of this thesis on the problem of low-rank matrix/tensor factorization. This chapter will be organized into three sub-parts.
  — In the introduction part 4.1 we introduce in more detail the setting in which we will work. We will also talk about the problem of low-rank matrix factorization and the techniques one can use to solve it 4.1.3 so that the reader may gain a simple understanding of the basic issue and difficulties at play behind the problem of inference in general and the problem of matrix/tensor factorization in general.
  — We will then move to the analysis of the problem both from an algorithmic 4.2 and theoretical point of view 4.3,4.4.
  — We will provide a number of examples of systems 4.5.1. These examples will allow us both to illustrate the malleability of matrix factorization to describe multiple situation in learning in estimation (sparse PCA, clustering of point, community detection in network, synchronization). This will allow us to establish a zoology of systems and phase transitions that one can expect in these matrix and tensor factorization problem. It will hopefully help the reader forge an intuition as to what can one expect in different setting of matrix and tensor factorization.
  — In the last section 4.6. We will talk in more detail of the problem of the problem of sparse PCA. for rank $r > 1$. The reader might think of this problem as a low rank version of the dictionary learning problem. We will see how even in the setting $1 \ll r \ll N$ this problem can be hard to solve even using AMP. We will diagnostic the problem in order to understand from where comes the difficulty and we will propose a partial fix to it. This section did not really fit anywhere else but I still wanted to put these results somewhere in my thesis.

2.2 Contributions

This thesis focuses on the study of the general type of models described by probability measures (2.6) and (2.7) (2.10). On the one hand, these probability distributions represent Boltzmann measures of vectorial-spin systems on fully connected symmetric or bipartite graphs or even hyper-graph. Examples of previously studied physical models that are special cases of the setting considered here would be the Sherrington-Kirkpatrick model [SK75], the Hopfield model [Hop82, Méz16], the p-spin model [MPV87], the inference (not learning) in the restricted Boltzmann machine [GTK15, Méz16, TM16]. Our work hence provides a unified replica symmetric solution and TAP equations for generic class of prior distributions and Hamiltonians.

On the other hand, these probability distributions (2.6), (2.7) and (2.10) combined with (2.8) represent the posterior probability distribution of a low-rank matrix/tensor estimation problem that finds a wide range of applications in high-dimensional statistics and data analysis. The thermodynamic limit $M/N = \alpha$ with $\alpha = O(1)$ whereas $N, M \to \infty$ was widely considered in the studies of spin glasses, in the context of low-rank matrix estimation this limit correspond to the challenging high-dimensional regime, whereas traditional statistics considers the case where $M/N \gg 1$. We focus on the analysis of the phase diagrams and phase transitions of low-rank matrix estimation corresponding to the Bayes optimal matrix/tensor estimation. We note that because we assume the data Y were generated from random factors U, V, X we obtain much tighter bounds, including the constant factors, control of the high-dimensional behavior $N, M \to \infty$, $\alpha = M/N = O(1)$, than some traditional bounds in statistics that aim

not to assume any generative model but instead craft proofs under verifiable conditions of the observed data Y.

We note at this point that methodologically closely related series of work on matrix factorization is concerned with the case of high rank, i.e. when $r/N = O(1)$ [KMZ13, PSC14, KKM$^+$16]. While that case also has a set of important application (among then the learning of over-complete dictionaries) it is different from the low-rank case considered here. The theory developed for the high rank case requires the prior distribution to be separable component-wise. The high rank case also does not present any known output channel universality, the details of the channel enter explicitly the resulting state evolution. Whereas the low-rank case can be viewed as a generalization of a spin model with pairwise interaction, in the graphical model for the high-rank case the interactions involve $O(N)$ variables.

No attempt is made at mathematical rigor in the present thesis. It is, however, worth mentioning that for the case of Bayes-optimal inference, a large part of the results of this paper was proven rigorously in a recent series of works [RF12, JM13, DM14a, KXZ16, DAM16, BDM$^+$16, ML16, Mio17b]. These proofs include the mutual information (related to the replica free energy) in the Bayes-optimal setting and the corresponding minimum mean-squared-error (MMSE), and the rigorous establishment that the state evolution is indeed describing asymptotic evolution of the Low-RAMP algorithm. The study out of the Bayes-optimal conditions (without the Nishimori conditions) are move involved.

It has become a tradition in related literature [ZK16] to conjecture that the performance of the Low-RAMP algorithm cannot be improved by other polynomial algorithms. We do analyze here in detail the cases where Low-RAMP does not achieve the MMSE, and we remark that since effects of replica symmetry breaking need to be taken into account when evaluating the performance of the best polynomial algorithms, the conjecture of the Low-RAMP optimality among the polynomial algorithms deserves further detailed investigation.

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

— **Approximate Message Passing** for Low-Rank matrix estimation (**Low-RAMP**) : In section 4.2 we derive and detail the approximate message passing algorithm to estimate marginal probabilities of the probability measures (2.6) and (2.7) for general prior distribution, rank and Hamiltonian (output channel). We describe various special case of these equations that arise due to the Nishimori conditions or due to self-averaging. In the physics literature this would be the TAP equations [TAP77] generalized to vectorial spins with general local magnetic fields and generic type of pairwise interactions. The Low-RAMP equations encompass as a special case the original TAP equations for the Sherrington-Kirkpatrick model, TAP equations for the Hopfield model [MPV87, Méz16], or the restricted Boltzmann machine [GTK15, TMC$^+$16, Méz16, TM16]. Within the context of low-rank matrix estimation, the AMP equations were discussed in [RF12, MT13, DM14a, LKZ15b, LKZ15a, DAM16, LDBB$^+$16]. Recently the Low-RAMP algorithm was even generalized to spin-variables that are not real vectors but live on compact groups [PWBM16a].

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

We distribute two open-source Julia and Matlab versions of LowRAMP at

`http://krzakala.github.io/LowRAMP/`. We strongly encourage the reader to download, modify, and improve on it.

— In physics, message passing equations are always closely linked with the **Bethe free energy** whose stationary points are the message passing fixed point equations. In the presence of multiple fixed points it is the value of the Bethe free energy that decides which of the fixed points is the correct one. In section 3.3 we derive the Bethe free energy on a single instance of the low-rank matrix estimation problem. The form of free energy that we derive has the convenience to be variational in the sense that in order to find the fixed point we are looking for a maximum of the free energy, not for a saddle. Corresponding free energy for the compressed sensing problem was derived in [KMTZ14, RSR$^+$13] and can be used to guide the iterations of the Low-RAMP algorithm [VSR$^+$15].

— In section 4.3 we derive the general form of the **state evolution** (under the replica symmetric assumption) of the Low-RAMP algorithm, generalizing previous works, e.g. [RF12, DM14a, LKZ15b, LKZ15a]. We present simplifications for the Bayes-optimal inference and for the conventional form of the Hamiltonian. We also give the corresponding expression for the free energy. We derive the state evolution and the free energy using both the cavity and the replica method.
For the Bayes-optimal setting the replica Bethe free energy is up to a simple term related to the mutual information from which, one can deduce the value of the **minimum information-theoretically achievable mean squared error**. Specifically, the MMSE correspond to the global maximum of the replica free energy (defined here with the opposite sign than in physics), the performance of Low-RAMP correspond to the maximum of the replica free energy that has the highest MSE.
We stress here that Low-RAMP algorithm belongs to the same class of approximate Bayesian inference algorithms as generic type of Monte Carlo Markov chains of variational mean-field methods. Yet Low-RAMP is very particular compared to these other two because of the fact that on a class of random models considered here its performance can be analyzed exactly via the state evolution and (out of the hard region) Low-RAMP asymptotically matches the performance of the Bayes-optimal estimator. Study of AMP-type of algorithms hence opens a way to put the variational mean field algorithms into more theoretical framework.

— We discuss the **output channel universality** as known in special cases in statistical physics (replica solution of the SK model depends only on the mean and variance of the quenched disorder not on other details of the distribution) and statistics [DAM16] (for the two group stochastic block model). The general form of this universality was first put into light for the Bayes-optimal estimation in [LKZ15a], proven in [KXZ16], in this thesis we discuss this universality out of the Bayes-optimal setting.

— In section 4.4.1 we show that the state evolution with a Gaussian prior can be used to analyze the asymptotic **performance of spectral algorithms** such as PCA (symmetric case) or SVD (bipartite case) and derive the corresponding spectral mean-squared errors and phase transitions as studied in the random matrix literature [BBAP05]. For a recent closely related discussion see, [PWBM16b].

— In section 4.4.2 and 4.4.4 we discuss the **typology of phase transition and phases** that arise in Bayes-optimal low-rank matrix estimation. We provide sufficient criteria for existence of phases where estimation better than random guesses from the prior distribution is not information-theoretically possible. We analyze linear stability of the fixed point of the Low-RAMP algorithm related to this phase of undetectability, to conclude that the threshold where Low-RAMP algorithm starts to have better performance than

randomly guessing from the prior agrees with the spectral threshold known in the literature on low-rank perturbations of random matrices. We also provide sufficient criteria for the existence of first order phase transitions related to existence of phases where information-theoretically optimal estimation is not achievable by existing algorithms, and analyze the three thresholds $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ related to the first order phase transition.

— In section 4.5.1 we give a number of examples of **phase diagrams** for the following models : rank-one $r = 1$ symmetric $XX^\top$ case with prior distribution being Bernoulli, Rademacher-Bernoulli, Gauss-Bernoulli, corresponding to balanced 2-groups. For generic rank $r \geq 1$ we give the phase diagram for the symmetric $XX^\top$ case for the jointly-sparse PCA, and for the symmetric community detection.

— Section 4.5.3 and appendix 4.6.2 is devoted to **small ρ analysis** of the above models. This is motivated by the fact that in most existing literature with sparsity constraints the number of non-zeros is usually considered to be a vanishing fraction of the system size. In our analysis the number of non-zeros is a finite fraction ρ of the system size, we call the the regime of *linear sparsity*. We investigate whether the $\rho \to 0$ limit corresponds to previously studied cases. What concerns the information-theoretically optimal performance and related threshold $\Delta_{\rm IT}$ our small ρ limit agrees with the results known for sub-linear sparsity. Concerning the performance of efficient algorithms, from our analysis we conclude that for linear sparsity in the leading order the existing algorithms do not beat the threshold of the naive spectral method. This corresponds to the known results for the planted dense subgraph problem (even when the size of the planted subgraph is sub-linear). However, for the sparse PCA problem with sub-linear sparsity algorithms such as covariance thresholding are known to beat the naive spectral threshold [DM14b]. In the regime of linear sparsity we do not recover such behavior, suggesting that for linear sparsity, $\rho = O(1)$, efficient algorithms that take advantage of the sparsity do not exist.

— In section 4.5.4 and appendix 4.6.2 we discuss analytical results that we can obtain for the community detection, and joint-sparse PCA models in the **limit of large rank** r. These large rank results are matching the rigorous bounds derived for these problems in [BMVX16].

2.3 Publications

The work done during this thesis has given rise to a number of publications in various conference and journal. They are presented here in chronological order.

— **Phase Transitions in Sparse PCA :** [LKZ15b] : In that first paper we analyzed a spiked jointed-sparse model. We uncovered a first order phenomena and derive the asymptotic behavior in the large rank limit of this system. This article was presented at the conference ISIT2015

— **MMSE of probabilistic low-rank matrix estimation : Universality with respect to the output channel :** [LKZ15a] : This article deals with the problem of community detection. The low-rank matrix factorization formalism allows us to treat the dense case of the stochastic block model. We uncover a first order phase transition when the number of hidden communities is larger than 4. We also talk about channel universality.

— **Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering :** [LDBB$^+$16] : In this article we treat the problem of clustering of points. Being given M points arranged in clusters in a N dimensional space. We study this question in one specific Bayesian setting and show that these system can exhibit non trivial behavior and phase diagrams.

— **Mutual information for symmetric rank-one matrix estimation : A proof of the replica formula :** [BDM$^+$16] In this paper we prove the correctness of the free energy formula provided by the replica computation for the problem of rank 1 matrix factorization. This allow us to express the Minimum Squared Error done when reconstructing the signal. We also show that an iterative algorithm called Approximate Message Passing (AMP) can be Bayes-optimal depending on the parameter of the system. My contribution to this paper was minor.

— **Constrained Low-rank Matrix Estimation : Phase Transitions, Approximate Message Passing and Applications :** [LKZ17] : In this long paper we present most of the result of past paper on matrix factorization. This paper was published in Journal of Statistical Mechanics (JSTAT).

— **Statistical and computational phase transitions in spiked tensor estimation :** [LML$^+$17] :In that paper we decide to move away from matrices and try and tackle the problem of tensor factorization once again in the Bayesian framework. We compute rigorously the mutual information and the Minimal Mean Square Error (MMSE) of the system. We study the performance of AMP algorithm and uncover a Hard phase which provides a physical interpretation to the Hardness of computing the eigenvectors of tensors [HL09].

Chapitre 3

General Theory

3.1 Frequentist and Bayesian inference

That whole thesis deals with the problem of inference. Inference can be seen as the problem of inferring a signal $X \in \mathcal{X}$ from some measurement $Y \in \mathcal{Y}$. The process that created Y from X is described by a conditional density probability $P_{\text{out}}(Y|X)$. Essentially we want to find an estimator $\hat{X} \in \mathcal{X}$, a function that takes in input $Y \in \mathcal{Y}$ and outputs $\hat{X}(Y) \in \mathcal{X}$. We want $\hat{X}(Y)$ to be a "good" reconstruction of X from Y. This is illustrated in figure 3.1.

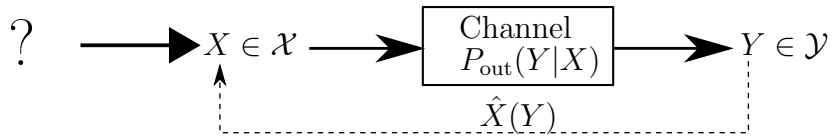


FIGURE 3.1 – We describe the typical setting of inference. We want to infer X from Y and find an estimator $\hat{X}(Y)$ (dashed line) that allows us to "come back" to X from Y. In the frequentist framework we make no other hypothesis other than $X \in \mathcal{X}$

The frequentist framework and Bayesian framework are two ways to analyze this inference problem. We will quickly describe the frequentist framework and then move on to the Bayesian framework that we use in that thesis.

The quantification of what it means for an estimator to be good is a non trivial question that is at the origin of the "schism" between frequentist and Bayesian inference. Essentially frequentist inference is interested with finding "worst-case" estimators while Bayesian inference deals with finding estimators that have good performance on average.

To quantify the quality of an estimator we introduce a distance function $d(.,.)$ between elements of $\mathcal{X}$.

frequentist inference (or worst case approach) : In frequentist inference to quantify the quality of an estimator we introduce the average error made when reconstructing the signal $\hat{I}(,\hat{X},X)$.

$$\hat{I}(\hat{X}, X) = \mathbb{E}_{Y \sim P_{\text{out}}(Y|X)} \left[d(X, \hat{X}(Y)) \right] \tag{3.1}$$

$\hat{I}(\hat{X}, X)$ is the average error made when reconstructing the hidden signal X with estimator $\hat{X}$. If the hidden is signal X. The average here is only taken over the measurement Y.

Our measure of how well estimator $\hat{X}$ is not a single number but a function $\hat{I}(\hat{X}, X_0)$ over $\mathcal{X}$ that outputs a positive number. Some estimator might reconstruct well some value of X and worst others but when designing an estimator one might have to make a choice on which value one might prefer to reconstruct (For example when designing an alarm do you prefer to have false positives or false negative). One of the difficulty of studying frequentist estimator is that except if X is exactly recoverable there does not exist an optimal estimator $\hat{X}_{\text{opt}}$ such that

$$\forall \hat{X}, \forall X \in \mathcal{X} : \hat{I}(\hat{X}_{\text{opt}}, X) \leq \hat{I}(\hat{X}, X) \tag{3.2}$$

It is easy to prove that such an estimator can not exist. To do so let us suppose that $\hat{X}_{\text{opt}}$ exists and let us create estimator $\hat{X}_1$ such that $\forall Y \in \mathcal{Y}, \hat{X}_1(Y) = X_0$. From we 3.2 we get

$$0 \leq \hat{I}(\hat{X}_{\text{opt}}, X_0) \leq \hat{I}(\hat{X}, X_0) \leq 0 \tag{3.3}$$

And therefore $\forall X \in \mathcal{X}$, $\hat{I}(\hat{X}_{\text{opt}}, X) = 0$. Which means that if $\hat{X}_{\text{opt}}$ exists it recovers the signal exactly.

Proofs on frequentist estimators are often upper bounds on the average error $\hat{I}$ conditioned on X belonging to some set $\mathcal{X}$.

$$\forall X \in \mathcal{X}, \hat{I}(\hat{X}, X_0) \leq \beta \tag{3.4}$$

Where $\mathcal{X}$ is large enough and β is small enough so that this is a non-trivial result. One could try to find the estimator $\hat{X}(Y)$ that minimizes β in (3.4) (if such an estimator exists), this could provide a notion of an optimal estimator.

Frequentist estimator are nice in the sense that they are "worst case" estimator. However they can often provide overly pessimistic bounds that are worst than the typical case, this leads us to the framework of Bayesian inference.

Bayesian inference In Bayesian inference we assume that the hidden data X does not have a fixed value but that it also has a probabilistic description and that it was sampled from some density probability $P_X(X)$.

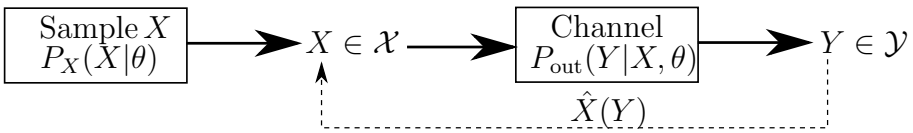


FIGURE 3.2 – We describe the typical setting of Bayesian inference. We once again want to infer back X from Y. The difference with the 3.1 is that here we know how X was generated from $P_X(X)$. θ is a variable that encode the details of the model.

Using Bayes formula we can access the posterior probability of X knowing Y.

$$P(X|Y,\theta) = \frac{P_X(X|\theta)P_{\text{out}}(Y|X,\theta)}{P(Y|\theta)} \tag{3.5}$$

This in turn makes the creation of optimal estimators possible. If one tries to find an estimator that minimizes the average error.

$$J(\hat{X}) = \mathbb{E}_{(X,Y)\sim P(X,Y|\theta)}\left[d(\hat{X}(Y),X)\right] = \mathbb{E}_{Y\sim P(Y|\theta)}\left[\mathbb{E}_{X\sim P(X|Y,\theta)}\left[d(\hat{X}(Y),X)\right]\right] \tag{3.6}$$

The difference with (3.1) is that the average is also taken over X. Taking the average with respect to X make it possible to find optimal estimators. Two example that will be useful in this thesis

— **Mean Estimator : $\mathbf{d(x,y) = \|x-y\|_2^2}$** : If we try to minimize the squared l_2 norm to the hidden solution then it is easy to prove (using (3.6)) that the optimal estimator is the posterior mean.

$$\hat{X}_{\text{Mean}}(Y) = \int XP(X|Y,\theta)\mathrm{d}X \tag{3.7}$$

In practice most of the time there is no analytical expression for (3.7). This comes from the fact that $\mathcal{X}$ might be a high dimensional space on which integration is hard. For instance if $\mathcal{X} = \{-1,1\}^N$ computing (3.7) requires the sum over $O(2^N)$ term for $N > 30$ this starts to be too long for most computers. There are ways to approximate (3.7) even in the high dimensional regime. For example on could run a Markov Chain Monte Carlo to compute (3.7). We will see in section 3.2 other ways inspired from statistical physics to approximate this integral.

— **Maximum A Posteriori estimator : $\mathbf{d(x,y)} = \mathbb{1}(\mathbf{x}=\mathbf{y})$** : If one deals with discrete variable X then one can ask for an estimator that maximizes the probability that the signal was reconstructed perfectly. This gives rise to the Maximum A Posteriori estimators or MAP estimator.

$$\hat{X}_{\text{MAP}}(Y) = \text{argmax}_X P(X|Y) = \text{argmax}_X P(Y|X)P(X) \tag{3.8}$$

Finding the maximum of $P(X|Y)$ sounds like a simpler problem to solve. If one deals with continuous variables then one can use a gradient descent algorithm, if one deals with discrete variables then one could use a zero temperature Markov Chain Monte Carlo to try and reach the global maxima of the problem. These methods might get stuck in local minimas. There are ways around that problem, for example, one might try some simulated annealing or parallel tempering scheme to try and deal with local minima.

Learning of model-parameters When trying to use a Bayesian inference one is left with the problem that we do not know in general with what model was Y created, such a model might not even exist. Nevertheless we need to be able to learn/fit a model from a finite set of data Y. The problem of learning a model is not an easy one. We will treat this problem as the problem of learning the parameter θ. We will suppose here that there is a value of θ that we need to find that corresponds to the real model used to generate Y. We could treat θ as just

another parameter on which we need to perform inference on. Therefore one has

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}\,. \tag{3.9}$$

Where here

$$P(Y|\theta) = \int \mathrm{d}X P(Y|X,\theta)P(X|\theta) \tag{3.10}$$

$\log(P(Y|\theta))$ is called the log-likelihood of the model. We still are stuck with the problem that we don't know with what probability distribution was sampled θ. To do that we could go even further down the Bayesian rabbit-hole and write that the prior probability $P(\theta) = P(\theta|\theta_1)$ Where θ_1 would be another set of parameter we would need to infer. We see that this is just displacing the problem.

Hopefully In this thesis we will be in a setting where the question of the value of $P(\theta)$ will be unimportant because we will always have

$$\log(P(\theta)) = O(1) \ll \log(P(Y|\theta)) = O(N)\,. \tag{3.11}$$

Where N will be the "size" of the system Y. $N \to +\infty$ the setting we will look at. $\log(P(\theta)) = O(1)$ means that we will need $O(1)$ parameters to describe our model. While the log-likelihood $\log(P(Y|\theta))$ coming from the observation of the data will be much larger. We can therefore write

$$P(\theta|Y) = P(Y|\theta)P(\theta)\frac{1}{P(Y)} = \frac{1}{P(Y)}\exp\left(\log(P(\theta)) + Nf(Y,\theta)\right) \tag{3.12}$$

$$f(Y,\theta) = O(1) \tag{3.13}$$

As $N \to +\infty$ we can neglect $\log(P(\theta))$ and just estimate θ with the argmax of $P(Y|\theta)$.

$$\hat{\theta}(Y) = \mathrm{argmax}_\theta \log(P(Y|\theta)) \tag{3.14}$$

The reason why this is a good estimator is that when the size of the system Y goes to infinity the function $f(Y,\theta)$ will become piqued around the true value of θ (or the best one to fit the model).

Bayes optimality - Nishimori conditions Suppose now that we know with what model parameters θ_0 was created Y. This setting is known as the Bayes-optimal setting. Let us call X_0 the hidden solution we are trying to infer. Y will be the observed data. The fact that we are in the Bayes-optimal setting has consequences when taking certain mean with respect to the posterior distribution. Essentially it makes it so that X_0 is a typical draw of the posterior mean. We mean here if one takes three configuration X, X_1, X_2 drawn from $P(Y|X)$ then for any function $f(.,.)$ one has

$$\mathbb{E}\left[f(X_0, X)\right] = \mathbb{E}\left[f(X_1, X_2)\right]\,. \tag{3.15}$$

This is easy to prove let X, X_1, X_2 be three independent samples from the posterior probability distribution $P(X|Y)$, eq. (3.5). We then consider some function $f(A,B)$ of two configurations

of the variables A, B. Consider the following two expectations

$$\mathbb{E}\left[f(X_1, X_2)\right] = \int f(X_1, X_2) P(Y) P(X_1|Y) P(X_2|Y) \mathrm{d}X_1 \, \mathrm{d}X_2 \, \mathrm{d}Y \,, \tag{3.16}$$

$$\begin{aligned} \mathbb{E}\left[f(X_0, X)\right] &= \int f(X_0, X) P(X_0, X) \mathrm{d}X \, \mathrm{d}X_0 = \int f(X_0, X) P(X_0, X, Y) \mathrm{d}X \, \mathrm{d}X_0 \mathrm{d}Y \\ &= \int f(X_0, X) P(X|Y, X_0) P_{\text{out}}(Y|X_0) P_0(X_0) \mathrm{d}X \, \mathrm{d}X_0 \mathrm{d}Y \,. \end{aligned} \tag{3.17}$$

where we used the Bayes formula. We further observe that $P(X|Y, X_0) = P(X|Y)$ because X is independent of X_0 when conditioned on Y. In the Bayes optimal case, we then obtain

$$\text{Bayes optimal}: \qquad \mathbb{E}\left[f(X_1, X_2)\right] = \mathbb{E}\left[f(X_0, X)\right] \,, \tag{3.18}$$

meaning that under expectations there is no statistical difference between the ground truth assignment of variables X_0 and an assignment sampled uniformly at random from the posterior probability distribution (3.5). This will be a source of simplification our analysis.

3.2 Statistical Physics

Statistical physics is a branch of physics that aim to describe the behavior of large systems that contain a large number of degree of liberty. A single water molecule can be quite accurately described by the classical equation of motions and solving these equations exactly or numerically is a realistic task to accomplish. Problem arise when one tries to find out what happens when one puts $N \approx 10^{23}$ water particle in a closed volume. Solving the equations of motion of this system for so many particles is not a realistic strategy to try and understand the behavior of such a system. Most of the time we are less interested in the solution of these equations as in the average property of the system.

Statistical physics aim to substitute to the deterministic description of the system a probabilistic one. Our description of the system is not a given configuration but a density probability over all configurations. For large system that can be described by an Hamiltonian dynamics statistical physics prescribe that the density probability of the system is given by the Gibbs measure. Let X be a configuration of the system then the Gibbs measure is given by

$$P_{\text{Gibbs}}(X) = \frac{1}{\mathcal{Z}(\beta)} \exp\left(\beta H(X)\right) \,. \tag{3.19}$$

Most of the time the Gibbs measure is taken with a minus in front of β out of convenience we will work in another convention where the Gibbs distribution is given by (3.19). This has no consequences on the result and will just make it so that the energy and free-energy is something that we will want to maximize. Here $\mathcal{Z}(\beta)$ is called the partition function and is here so that $P(X)$ is normalized. $\beta = 1/T$ is the inverse temperature of the system. The "miracle" of statistical physics is that macroscopic properties of the system (Energy, Entropy, Volume, Magnetization, etc.) are all encoded in a function called the thermodynamic potential. Here the thermodynamic potential is the free-energy ϕ given by

$$\phi(\beta) = \log \mathcal{Z}(\beta) \tag{3.20}$$

One can then access average of observable (with respect to the Gibbs distribution) by taking successive derivative of the free energy with respect to parameters for example the average energy can be computed from the free energy

$$\langle H \rangle_{\text{Gibbs}} = \frac{\partial \phi(\beta)}{\partial \beta} \tag{3.21}$$

Since most properties physicist are interested in can be encoded into a thermodynamic potential, studying a statistical physics problem often translates into computing a thermodynamic potential. This is where problems begin since most of the time there is no analytical formula for the partition function $\mathcal{Z}$ or the free energy of a problem ϕ. Over the year physicists have developed tools to try and tackle this problem of computing free energies and describing the statistical properties of large system (Monte Carlo Markov Chain, Mean field, Diagram Expansion, Belief Propagation techniques, etc.). In that Chapter we will describe some of these techniques that are used in that thesis.

But before jumping right into the specific section that will deal with these method let us first try and analyze a very simple system. Let us consider N ± 1 spins interacting through the following Hamiltonian

$$H(\{s_i\}) = \frac{1}{2} \sum_{1 \leq i,j \leq N} \frac{J_0}{N} s_i s_j \,. \tag{3.22}$$

The study of this simple system will help us illustrate the idea behind the techniques used in that thesis. Let us compute the free energy of this system (we set $\beta = 1$).

$$\mathcal{Z} = \sum_{\{s_i\}} \exp\left(\frac{1}{2} \sum_{1 \leq i,j \leq N} \frac{J_0}{N} s_i s_j \right) \tag{3.23}$$

$$\mathcal{Z} = \sum_{\{s_i\}} \exp\left(N \frac{J_0}{2} \left[\frac{1}{N} \sum_{1 \leq i \leq N} s_i \right]^2 \right) \tag{3.24}$$

By defining the average magnetization m

$$m = \frac{1}{N} \sum_{1 \leq i \leq N} s_i \tag{3.25}$$

We can rewrite the partition function in the following way

$$\mathcal{Z} = \sum_{m=1-2k/N, k \in \{0,\cdots,N\}} \binom{N}{N(m+1)/2} \exp\left(N \frac{J_0 m^2}{2} \right) \tag{3.26}$$

The binomial term is here to count how many way there are to have a magnetization of value m. Using the Sterling Formula the binomial can be well approximated by

$$\binom{N}{N(m+1)/2} \sim \frac{1}{\sqrt{N(1-m^2)}} \exp\left(-N \left[\frac{m+1}{2} \log\left(\frac{m+1}{2} \right) + \frac{1-m}{2} \log\left(\frac{1-m}{2} \right) \right] \right) \tag{3.27}$$

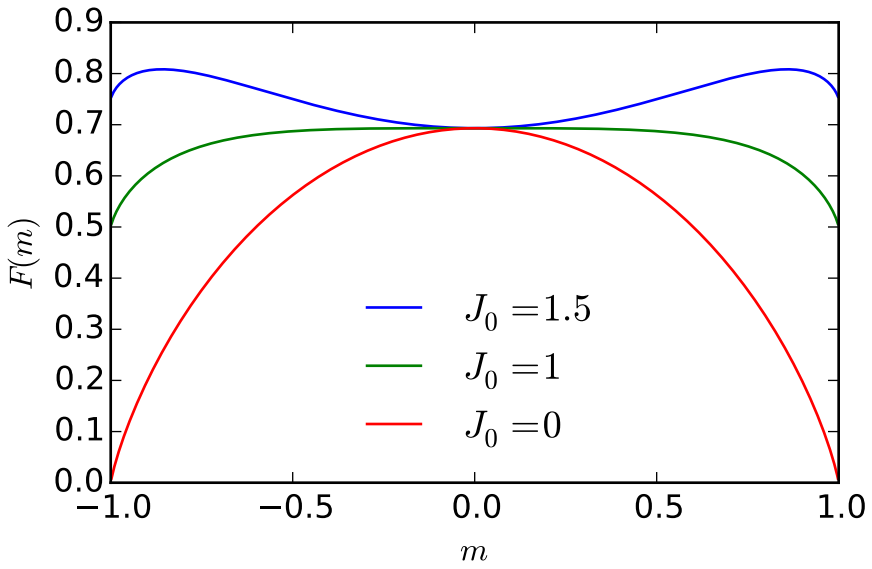


FIGURE 3.3 – We plot the profile of $F(m)$ (3.28) for different value of J_0. We plot $F(m)$ for values of $J_0 \in \{0, 1, 1.5\}$ For $J_0 \leq 1$ the maximum of $F(m)$ is located at $m = 0$, this means that the system if paramagnetic. For $J_0 > 1$ the maximum of $F(m)$ are located at $m \neq 0$ this means that the Gibbs distribution of the system is dominated by states that have a non zero magnetization the system is then said to be ferromagnetic.

This approximation holds as soon as $N(1 - m^2) \gg 1$. Since we are summing over ever closer value of magnetization $m \in [-1, 1]$. As N grows large we can approximate $\mathcal{Z}$ in (3.26) by a continuous integral. We obtain

$$\mathcal{Z} \approx \int_{-1}^{1} \mathrm{d}m \exp\left(NF(m)\right) \tag{3.28}$$

where

$$F(m) = -\frac{m+1}{2}\log\left(\frac{m+1}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right) + \frac{J_0 m^2}{2}\,. \tag{3.29}$$

We can estimate (3.28) using Laplace method. By estimating the maximum of $F(m)$ This is illustrated in Fig 3.3. All of this yields us that

$$\log \mathcal{Z}(J_0) \approx N \max_{-1 \leq m \leq 1} F(m) \tag{3.30}$$

Looking for the maximum of we look for m such that $\frac{\partial F}{\partial m} = 0 = -\tanh^{-1}(m) + J_0 m$ or

$$m = \tanh(J_0 m) \tag{3.31}$$

We can interpret this as a fixed point equation

$$m^{t+1} = \tanh(J_0 m^t) \tag{3.32}$$

The trivial fixed point $m = 0$ can then be stable or unstable depending on the value of J_0.

Let us take a few steps back and see what we just did.

— We started with the problem of computing an integral that at first looked intractable.
— We then transformed this problem of integral computation into an optimization problem over a function $F(m)$.
— This optimization problem also yields us with fixed point equations. The solution of these equation give us a description of the system (here the magnetization m).

Essentially the way we will analyze the problem encountered in this thesis will be by transforming integration problem into optimization problem. Integration problem are hard and complicated to solve. In the more general setting Monte Carlo or Monte Carlo inspired method are the only technique that remain to try and compute large dimensional integral. In comparison, optimization problem seem easier to solve since we have access to some techniques to try and tackle them (Gradient descent, Newton descent, Stochastic gradient descent, Simulated Annealing, Parallel tempering etc.). Looking for the stationary point of an optimization problem often yields us with update equations for which we look for a fixed point. Of course this optimisation might require itself exponential time to solve but one can hope that a "typical" problem can be solved in polynomial time, examples of such problems are given in [CO07].

This theory chapter will be therefore organized in the following way.

— **Naive Mean field and Plefka Expansion** can be seen as techniques to transform the computation of Free energy into optimization problem. It works by trying to compute approximation of the Free-energy as a function of the marginal density probability of the system one tries to study.
— **Factor Graphs** are a mathematical/theoretical tool to represent density probabilities in term of a graph of interactions. This graph is made out of factors nodes and variables nodes. For physicists factors can be thought of as interaction term (they might be 2 or p points interactions) while variables can be mapped to particles.
— **Belief propagation (BP)** is an approximation method that aim given a factor graph to compute marginals of variables and compute an estimate of the free energy of the system. The belief propagation relies on a set of update equations for which we look for a fixed point.
— **Spin glasses and the Sherrington-Kirkpatrick model :** In that section we introduce the concept of spin glass. Spin glass refer to a class of material that were discovered at the end of the 50s. They exhibited uncommon magnetic behavior. Theoretical efforts to understand them gave rise to a number of theoretical models such as for instance the Sherrington-Kirkpatrick model. This model (and other such as the Edward-Anderson model) exhibit quite unintuitive behavior which required new theoretical method and techniques to be analyzed. The cavity method and the replica method belong to such techniques.
— **The replica method** can be thought of as a non rigorous mathematical trick to compute the free energy of disordered systems. It was introduced to study spin glasses and relies on the following equality

$$\forall Z \in \mathbb{R},\ \log Z = \lim_{n \to 0} \frac{Z^n - 1}{n} \tag{3.33}$$

We will use the replica method to compute the free energy of all the system encountered in this thesis.

3.3 Naive Mean field and the Plefka Expansion

In that section we present different ways one could use to try and transform the computation of Partition functions and free energy which can be a hard feat into an optimization problem for which one might have tools to try and tackle it. We will show just two techniques Naive Mean-Field and the Plefka expansion technique. We will then show how these techniques could be combined. A good introduction to mean fields method can be found in [OS01, MM09].

3.3.1 Naive-Mean Field

Suppose we are given an Hamiltonian $H(X)$ (where X is a big vector that encodes the whole state of the system) and an inverse temperature β. This defines us a Gibbs distribution and a free energy.

$$P_{\text{Gibbs}}(x_1, \cdots, x_N) = \frac{1}{Z} \exp\left(\beta H(x_1, \cdots, x_N)\right) \tag{3.34}$$

$$\log Z = \log \left[\int \mathrm{d}x_1 \cdots \mathrm{d}x_N \exp\left(\beta H(x_1, \cdots, x_N)\right) \right] \tag{3.35}$$

We want to compute the free energy and different observable about the system. As always, in most case there is no analytical way to compute Z or marginals of individual variables exactly. We therefore have to settle for an approximation of all these quantity. This is where the Kullback-Leibler divergence comes in handy. The Kullback-Leibler divergence can be thought of as a distance function on density probability (even though it is not symmetric and does not satisfy the triangular inequality). Being given two density probability $p(x)$ and $q(x)$ defined on a set $\mathcal{X}$ the Kullback-Leibler $D_{\text{KL}}(p\|q)$ is defined as

$$D_{\text{KL}}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \tag{3.36}$$

The Kullback-Leibler has the following properties

— $D_{\text{KL}}(p\|q) \geq 0$ for all p and q (even when taken to the continuous limit).
— $D_{\text{KL}}(p\|q) = 0 \iff p = q$.
— $D_{\text{KL}}(p\|q)$ is a convex function of (p, q).
— $D_{\text{KL}}(p\|q)$ is non symmetric meaning that in general $D_{\text{KL}}(p\|q) \neq D_{\text{KL}}(q\|p)$.

We see that in a way $D_{\text{KL}}(p\|q)$ can be thought of as a "distance" on density probabilities (even thought it is not symmetric and does not obey the triangular inequality.)

The idea of classical mean field method is to set $q(x) = P_{\text{Gibbs}}(x)$ and to look for density probabilities $p(x)$ that minimizes the Kullback-Leibler divergence to $q = P_{\text{Gibbs}}$. We get

$$D_{\text{KL}}(p\|P_{\text{Gibbs}}) = -S_p - \sum_{X \in \mathcal{X}} p(X) \log\left(\frac{1}{Z} \exp\left(\beta H\right)\right) \tag{3.37}$$

$$D_{\text{KL}}(p\|P_{\text{Gibbs}}) = \log Z - \left(S_p + \beta \left\langle H \right\rangle_p\right) \tag{3.38}$$

Where S_p is the entropy of p and $\langle H \rangle_p$ is the average of $H(X)$ where X is sampled from $p(X)$. Since $D_{\mathrm{KL}}(p \| P_{\mathrm{Gibbs}}) \geq 0$ we see that

$$\forall p, \quad S_p + \beta \langle H \rangle_p \leq \log Z \tag{3.39}$$

Where the inequality is saturated only when $p = P_{\mathrm{Gibbs}}$. This expression is reminiscent of formula for the free energy of a system $F = U - TS$. This defines the mean field free energy estimate (Here the free energy is something that we want to maximize)

$$\Phi_{\mathrm{MF}}(p) = S_p + \beta \langle H \rangle_p \ . \tag{3.40}$$

We can try and get better estimates of the free-energy maximizing this quantity over p. These density probabilities p need to be density probabilities for which computing the entropy S_p and the mean energy $\langle H \rangle_p$ is an easy task.

The term Naive mean field in the context of classical statistical physics means that we are going to look to a maximize $\Phi_{\mathrm{MF}}(p)$ over density $p(X)$ probabilities that factorizes over components of x_i.

$$p(X) = \prod_{1 \leq i \leq N} \mu_i(x_i) \tag{3.41}$$

Where the x_i are component of X. $\Phi_{\mathrm{MF}}(p)$ becomes

$$\Phi_{\mathrm{MF}}(p) = \sum_{1 \leq i \leq N} S_{\mu_i} + \beta \langle H \rangle_p \ . \tag{3.42}$$

Optimality with respect to one parameter μ_i give us the Mean Field fixed point equations

$$\mu_i(x_i) = \frac{1}{Z_i} \exp\left(\beta \langle H \rangle_{p \setminus i}(x_i)\right) \tag{3.43}$$

Where $\langle H \rangle_{p \setminus i}$ means taking the average of H with respect to all variables x_j except x_i. For a Ising spin system this yields us with the usual mean field fixed point equations

$$\langle s_i \rangle_{\mathrm{MF}} = \tanh\left(h_i \langle s_i \rangle_{\mathrm{MF}} + \sum_{1 \leq i \leq N} J_{ij} \langle s_j \rangle_{\mathrm{MF}}\right) \tag{3.44}$$

it is worth noting that there can be multiple way to perform mean field. Over what set of "component" of X one factorizes p can have great difference on the performance of the Mean field approach.

The Gaussian variable approach The naive mean field approach aims to compute the free energy by neglecting correlations between variables and therefore looks for a minimizer $p(x)$ among probabilities that factorizes over the x_i. The reason for that is, that correlation makes thing hard to compute (free-energy, average of observable etc.). However there exists one class of density probabilities where correlation are not such a big issue, namely Gaussian variables.

One could just look for a minimizer of $p(X)$ under the form.

$$p(X) = \mathcal{N}(\mu, \Sigma, X) \tag{3.45}$$

The entropy of S_p then becomes.

$$S_p = \frac{1}{2} \log \det (2\pi e \Sigma) \tag{3.46}$$

Depending on the circumstances this might be a good maximizer or not. Of course the main difficulty with this approach is that one has now to optimize over $O(N^2)$ term in order to optimize the free-energy. This makes the mean field method harder to use. There are cases where this minimization is easier than in the general case and allow for the use of this method [OA09].

3.3.2 The Plefka expansion

As it turns out the mean-field free energy from (3.42) is just the first two order term in a high temperature expansion of the free-energy. It was derived in [Ple82, GY91] first by Plefka and then by Georges and Yedidia. What we call the TAP free energy is just then the 2nd order expansion of the free energy in β. Suppose one is given a general system with N classical variables x_i. The Hamiltonian βH defines the free energy

$$\Phi = \log \left(\mathrm{Tr} \left[\exp(\beta H(x_1, \cdots, x_N)) \right] \right) . \tag{3.47}$$

We define a new Hamiltonian that fixes the marginal probabilities of the variables x_i by introducing fields $\lambda_i(x_i)$.

$$\beta H_{\text{Field}}(\{\lambda_i\}) = \beta H + \sum_{1 \leq i \leq N} \lambda_i(x_i, \beta) , \tag{3.48}$$

$$\Phi_{\text{Field}}(\beta, \{\lambda_i\}) = \log \left(\mathrm{Tr} \left[\exp(\beta H_{\text{Field}}(\{\lambda_i\})) \right] \right) . \tag{3.49}$$

By taking the Legendre transform one gets

$$\Phi_{\text{Legendre}}(\beta, \{\mu_i(x_i)\}) = \min_{\{\lambda_i(x_i)\}} \left\{ \Phi_{\text{Field}}(\beta, \{\lambda_i\}) - \sum_{1 \leq i \leq n} \int \mathrm{d}x \mu_i(x) \lambda_i(x) \right\} , \tag{3.50}$$

$$\{\Lambda_i(x_i)\} = \mathrm{argmin}_{\{\lambda_i(x_i)\}} \left\{ \Phi_{\text{Field}}(\beta, \{\lambda_i\}) - \sum_{1 \leq i \leq n} \int \mathrm{d}x \mu_i(x) \lambda_i(x) \right\} . \tag{3.51}$$

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

$$\int \mathrm{d}x \mu_i(x) \Lambda_i(x) = 0 . \tag{3.52}$$

According to the definition of the Legendre transformation (and because Φ_{field} is convex in $\{\lambda_i(x_i)\}$ one has

$$\Phi = \max\left\{\Phi_{\text{Legendre}}(\beta, \{\mu_i(x_i)\})\right\} . \tag{3.53}$$

To compute $\Phi_{\text{Legendre}}(\beta, \{\mu_i(x_i)\})$ we resort to a high temperature expansion. This procedure is called the Plefka expansion [Ple82], it relies on the following high temperature expansion (we stop here at order 2).

$$\Phi_{\text{TAP}} = \Phi_{\text{Frustrated}}(\beta = 0) + \beta \left(\frac{\partial \Phi_{\text{Legendre}}}{\partial \beta}\right)(\beta = 0) + \frac{\beta^2}{2}\left(\frac{\partial^2 \Phi_{\text{Legendre}}}{\partial \beta^2}\right)(\beta = 0) . \tag{3.54}$$

The expansion was explained in detail in [GY91], here we remind the main steps. Let us introduce the following operator

$$U = H - \langle H \rangle - \sum_{1 \leq i \leq N} \frac{\partial \Lambda_i(x_i)}{\partial \beta} , \tag{3.55}$$

where the average is taken with respect to probability distribution induced by (3.48). One can show that for all observables O one has

$$\forall \beta, \frac{\partial \langle O \rangle}{\partial \beta} = \left\langle \frac{\partial O}{\partial \beta} \right\rangle - \langle OU \rangle . \tag{3.56}$$

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

$$\forall \beta, \frac{\partial \Phi_{\text{Legendre}}}{\partial \beta} = \langle H \rangle . \tag{3.57}$$

Using (3.57) and (3.56) we get

$$\forall \beta, \frac{\partial^2 \Phi_{\text{Legendre}}}{\partial \beta^2} = \frac{\langle H \rangle}{\partial \beta} = -\langle HU \rangle . \tag{3.58}$$

Therefore

$$\Phi_{\text{TAP}} = \Phi_{\text{Legendre}}(\beta = 0) + \beta \langle H \rangle (\beta = 0) - \frac{\beta^2}{2} \langle HU \rangle (\beta = 0) . \tag{3.59}$$

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

$$\begin{aligned} \left(\frac{\partial \langle \delta(x_i - \hat{x}_i) \rangle}{\partial \beta}\right) = \left(\frac{\partial \mu_i(\hat{x}_i)}{\partial \beta}\right) = 0 = \langle U \delta(x_i - \hat{x}_i) \rangle \\ = \left\langle \left(H - < H > - \sum_{1 \leq i \leq N} \frac{\partial \Lambda_i(x_i)}{\partial \beta} + \int \mathrm{d}\hat{x}_i \mu_i(\hat{x}_i) \frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \right) \delta(x_i - \hat{x}_i) \right\rangle . \end{aligned} \tag{3.60}$$

From this one deduces

$$\left\langle \frac{\partial \Lambda_i(x_i)}{\partial \beta} \delta(x_i - \hat{x}_i) + C(i, \beta) \delta(x_i - \hat{x}_i) s \right\rangle = \langle H \delta(x_i - \hat{x}_i) - \langle H \rangle \delta(x_i - \hat{x}_i) \rangle , \tag{3.61}$$

where $C(i,\beta)$ is

$$C(i,\beta) = \int \mathrm{d}\hat{x}_i \mu_i(\hat{x}_i) \frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \,. \tag{3.62}$$

By definition

$$\langle \delta(x_i - \hat{x}_i) \rangle = \mu_i(\hat{x}_i) \,, \tag{3.63}$$
$$\langle H\delta(x_i - \hat{x}_i) \rangle = \mu_i(x_i) \langle H \rangle_{x_i=\hat{x}_i} \,, \tag{3.64}$$

where $\langle H \rangle_{x_i=\hat{x}_i}$ is the average energy conditioned on the fact that $x_i = \hat{x}_i$. Using (3.61) we get

$$\mu_i(\hat{x}_i) \left(\frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \right) = \mu_i(\hat{x}_i) \langle H \rangle_{x_i=\hat{x}_i} - \mu_i(\hat{x}_i) \langle H \rangle + \mu_i(\hat{x}_i) C(i,\beta) \,, \tag{3.65}$$

$$\left(\frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \right) = \langle H \rangle_{x_i=\hat{x}_i} - \langle H \rangle + C(i,\beta) \,. \tag{3.66}$$

We can see from (3.48) that $\Lambda_i(\hat{x}_i, \beta)$ is defined up to a constant we fix that constant by having

$$\int \mathrm{d}\hat{x}_i \mu_i(\hat{x}_i) \Lambda_i(\hat{x}_i) = 0 \,. \tag{3.67}$$

Therefore we get

$$\forall \beta, \left(\frac{\partial \Lambda_i(\hat{x}_i)}{\partial \beta} \right) = \langle H \rangle_{x_i=\hat{x}_i} - \langle H \rangle \ = \frac{1}{\mu_i(\hat{x}_i)} \langle H\delta(x_i - \hat{x}_i) \rangle - \langle H \rangle \,, \tag{3.68}$$

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

$$P_{\text{Factorised}} = \prod_{i=1\cdots N} \mu_i(x_i) \,. \tag{3.69}$$

By using (3.54) and (3.68) around $\beta = 0$ one gets

$$\Phi_{\text{TAP}} = \sum_{1 \le i \le N} S_{\mu_i} + \beta \langle H \rangle + \frac{\beta^2}{2} \left[\langle H^2 \rangle - \langle H \rangle^2 - \sum_{1 \le i \le N} \int \mathrm{d}x_i \mu_i(\hat{x}_i) (\langle H \rangle_{x_i=\hat{x}_i} - \langle H \rangle)^2 \right] + O\left(\beta^3\right) \,, \tag{3.70}$$

One could do further expansions of this formula in β to get better estimate of the free energy. For all of the case that we will look at in this thesis the second order expansion will be enough and all further expansion will contribute sub-extensively to the free energy in the large N limit of the system.

It is also worth noting that there might be multiple way to do a Plefka expansion depending on the set of variables x_i that one takes to be independent at $\beta = 0$.

3.4 Factor Graph

Factor graphs are a way to represent an ensemble of density probabilities. In such a description one will write the density probability in terms of variables and interactions.

A factor graph is formed of an undirected bipartite graph between N variables (or nodes) and M factors which are interaction terms. The variables are numerated according to $1 \leq i \leq N$ and the factors according to $1 \leq \alpha \leq M$. A variable can only have links with factors and a factor can only have links with variables.

The density probability is then given by

$$P(x_1, \cdots, x_i, \cdots, x_N) = \frac{1}{Z} \prod_{1 \leq a \leq M} \psi_a(x_{\partial a}) \tag{3.71}$$

Where the ψ_a are interaction functions that take as argument all the variables that have a link with factor a. The variable Z is a normalization constant of the density probability. Factor graph are useful tools to represent density probabilities.

Let us give an example with N variables $x_i = \pm 1$

$$P(x_1, \cdots, x_N) = \frac{1}{Z} \exp\left(\sum_{1 \leq i < j \leq N} J_{ij} x_i x_j + \sum_{1 \leq i \leq N} h_i x_i \right) = \frac{1}{Z} \prod_{1 \leq i < j \leq N} \exp(J_{ij} x_i x_j) \prod_{1 \leq i \leq N} \exp(h_i x_i x_j) \tag{3.72}$$

This density probability can be described using the following factor graph.

3.5 Belief propagation

So far the factor graph formalism is just a reformulation of the initial density probability. If the factor graph one deals with is a tree (and therefore has no loop) then everything can be computed in polynomial time (free energy, marginal probability of variable etc.). The method that allows us to compute this is called the belief propagation. When used on tree factor graph the BP method is just a clever trick to compute the free energy exactly. When used on a general factor graph this formalism will allow us to introduce the Bethe approximation that one can use to approximate the free energy and marginals of the system. This method has many names Belief propagation method (BP), cavity method (when it's then averaged over the disorder) or even message passing algorithm [YFW03]. The Belief propagation (BP) is explained with much more details than we can hope to achieve in that thesis in [YFW03], nevertheless we will sketch in this section the main idea behind this theoretical and algorithmical tool that BP is. Here are some important properties of BP.

— Belief propagation is a method to compute marginals and the free energy of systems described by factor graphs. The main idea behind it is that variables and factors send each other beliefs (or message) which are estimates of marginal probabilities in a modified factor graph where one has removed or "turned off" some interactions or variables (hence created a "cavity" which is where the term cavity method comes from). These beliefs are then updated according to some update equations till convergence.

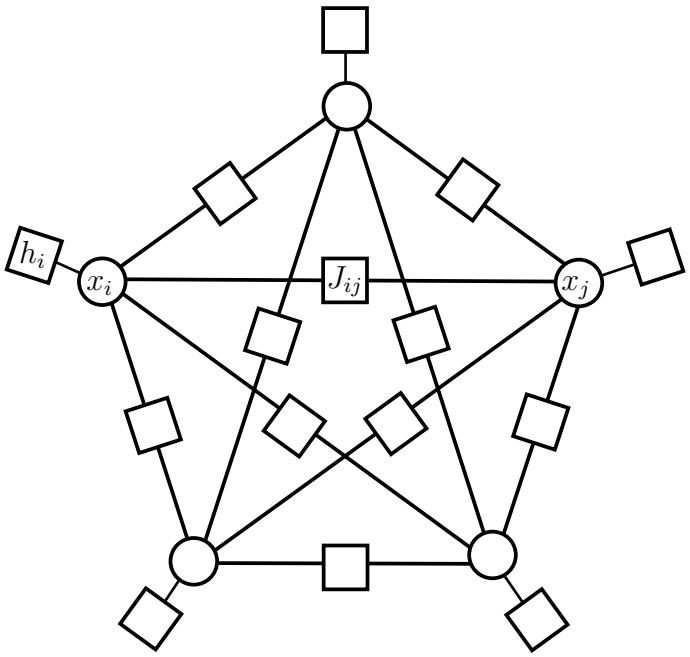


FIGURE 3.4 – This is the factor graph describing model (3.72). The variables x_i are represented by the circles. The factors are represented by squares. The factor annotated with a J_{ij} are interaction term between variables x_i and x_j. The factor annotated with an h_i are the field factors.

— Belief propagation is exact on factor graph that are trees. When working on trees Belief propagation can be seen as just a clever "trick" to be able to make the computation of the partition function with a number of operation that is not exponential in the number of variables in the system.
— Belief propagation is not exact on factor graphs that have loops. Even convergence property of the algorithm are not guaranteed. This come from the fact that when a factor graph has loops some assumption about the nature of correlation in the graph graph can not be done anymore. Nevertheless the belief propagation update equations can be applied and iterated till hopefully they converge. There exist some case where Belief Propagation approach succeeds despite the existence of loops.
— In some cases a BP approach has the added benefit that it is possible to analyze the dynamics of these equations in the large system size limit. Such an approach is called Density Evolution (or State Evolution or single letter characterization in some cases).

3.5.1 Tree Factor graph

Let us first look at the case where the factor graph one is looking at is a tree and therefore has no loops. In that case one can easily compute the partition function of the system. To do so let us introduce the following density probabilities.

$$n_{i\to\alpha}(x_i) \tag{3.73}$$

$n_{i\to\alpha}(x_i)$ is marginal probability of variables x_i in the modified system where factor α is missing.

We will illustrate the derivation of the BP equations using Fig 3.5. It is straightforward to see that

$$n_{1\to\alpha}(x_1) = \frac{1}{Z_{1\to\alpha}} \int P_{\partial_1}(x_2, x_5)\psi_\beta(x_1, x_2)\psi_\gamma(x_1, x_2) \tag{3.74}$$

Where $P_{\partial_1}(x_2, x_5)$ is the marginal probability of variables x_2 and x_5 in the system where one has removed variables x_1 and every factor it was connected to. And where $Z_{1\to\alpha}$ is a normalization constant. So far all of this is exact on any factor graph.

Because of the tree nature of the factor graph $P_{\partial_1}(x_2, x_5)$ is going to factorizes because as soon as x_1 has been removed there is no interaction left between variables x_2 and x_5 and one has.

$$P_{\partial_1}(x_2, x_5) = n_{2\to\beta}(x_2)n_{5\to\gamma}(x_5) \tag{3.75}$$

Therefore one has

$$n_{x_1\to\alpha}(x_1) = \frac{1}{Z_{1\to\alpha}} \left[\int n_{2\to\beta}(x_2)\psi_\beta(x_1, x_2)\right] \left[\int n_{5\to\gamma}(x_5)\psi_\gamma(x_1, x_2)\right] \tag{3.76}$$

One could write similar equations for all the message $n_{x_i\to a_j}$ for any tree factor graph by introducing the variables $\hat{n}_{a_j\to x_i}$. The update equations then become

$$n_{i\to\alpha}(x_i) = \frac{1}{Z_{i\to\alpha}} \prod_{\beta\in\partial i,\, \beta\neq\alpha} \tilde{n}_{\beta\to i}(x_i) \tag{3.77}$$

$$\tilde{n}_{\alpha\to i}(x_i) = \frac{1}{Z_{\alpha\to i}} \int \psi_\alpha(x_{\partial\alpha}) \prod_{k\in\partial\alpha,\, k\neq i} [\mathrm{d}x_k n_{k\to\alpha}(x_k)] \tag{3.78}$$

Here we have used the fact that the underlying factor graph is a tree to be able to write (3.75).

The marginal probability of the system can then be computed using this formula.

$$P(x_i) = \frac{1}{Z_i} \prod_{\alpha\in\partial i} \tilde{n}_{\alpha\to i}(x_i) \tag{3.79}$$

For different reason one is often interested in computing the free energy of system described by this tree factor graph. When the factor graph one deals with are trees then this can be computed exactly.

The free energy in our notation is written as

$$\log(Z) = \phi = U + S \tag{3.80}$$

where Z is defined in (3.71),S is the entropy of (3.71) and U is

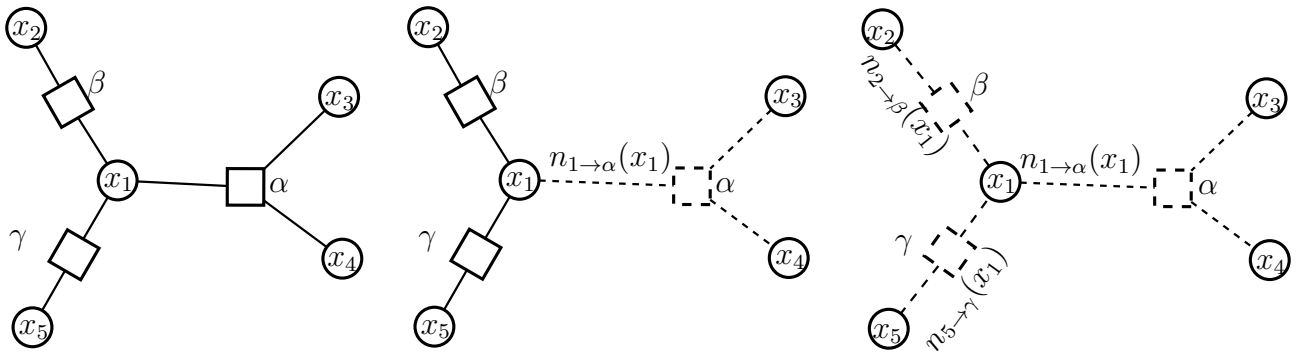


FIGURE 3.5 – On this figure we present the example of a tree Factor graph that we used to present the derivation of the cavity equations. First we present the full factor graph **(left)**. We then introduce the message $n_{1\to\alpha}(x_1)$ which is the marginal probability of variable x_1 in the modified factor graph where the factor α has been "turned off" and put in dashed line **(middle)**. Computing the marginal probability of the jointed variable x_2 and x_5 $P_{\partial 1}(x_2, x_5)$ can be done in the modified system on the **(right)** here one sees that because in that setting x_2 and x_5 belong to separated component of this modified factor graph and are therefore independent.

$$U = \left\langle \sum_\alpha \log \psi_\alpha(x_{\partial\alpha}) \right\rangle_{P(x_1,\cdots,x_N)} = \sum_\alpha \langle \log \psi_\alpha(x_{\partial\alpha}) \rangle_{\prod\limits_{i\in\partial\alpha} n_{i\to\alpha}(x_i)} \tag{3.81}$$

This decomposition in a term of energy and entropy is reminiscent of the formula $F = U - TS$.

The energy term can be easily computed using messages.

The problem is how to estimate the entropy S. Hopefully one can prove that the full density probability $P(x_1, \cdots, x_N)$ can be written as a function of local marginals and local jointed probabilities around factors.

$$P(x_1, \cdots, x_N) = \prod_\alpha P(x_{\partial\alpha}) \prod_i P(x_i)^{1-|\partial i|} \tag{3.82}$$

This is a property that comes from the fact that the factor graph is a tree. This allow us to compute the Free Energy of the system.

$$\log(Z) = \phi = \sum_\alpha \log \left[\int \psi(x_{\partial\alpha}) \prod_{i\in\partial\alpha} \mathrm{d}x_i n_{i\to\alpha}(x_i) \right] + \sum_i \log \left[\int \mathrm{d}x_i \prod_{\alpha\in\partial i} \tilde{n}_{\alpha\to i}(x_i) \right] - \sum_{\substack{i,\alpha \\ i\in\partial\alpha}} \log \left[\int \mathrm{d}x_i n_{i\to\alpha}(x_i) \tilde{n}_{\alpha\to i}(x_i) \right] \tag{3.83}$$

This is called the Bethe free energy and is exact on trees.

3.5.2 Loopy BP

We have seen that BP is exact on trees. However most situation that one encounters are ones where the factor graph is not a tree and contains loops.

The critical property that was needed in order for Belief propagation equations to allow access to the real marginals was.

$$P_{\partial i}(x_{\partial\partial i}) = \prod_{\alpha\in\partial i}\prod_{j\in\partial\beta, j\neq\alpha} n_{j\to\alpha}(x_j)\,. \tag{3.84}$$

Where $x_{\partial\partial i}$ are all the variable with which variables x_i interact. This formula is illustrated on figure 3.6. The main idea behind it is that when one hollows a cavity by removing a variable the variables at the border of the cavity become independent and that any correlation these variable might have between one another was mediated through variable x_i. When dealing with tree factor graph, this is exactly true, but when dealing with factor graph with loops this might be wrong or just an approximation

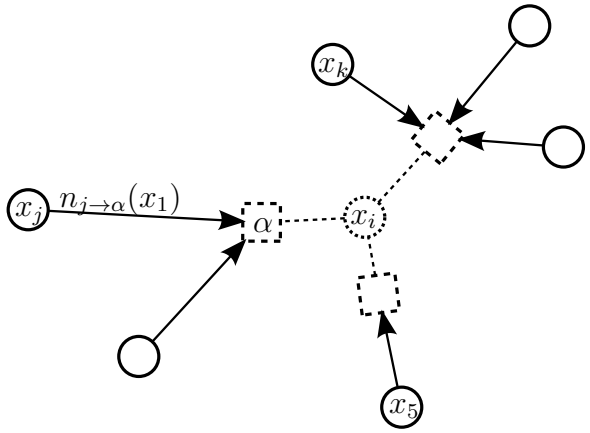


FIGURE 3.6 – On this figure we present the main assumption of Belief propagation. When removing a variable x_i and creating therefore a "cavity" in that modified system, we expect marginal probability of variables at the border of this cavity to factorize. Which is to say that any correlation between variable x_j and x_k comes from the interactions with x_i. If one removes this variable then one is left with variables x_j on the border of the cavity that are independent.

We will present two other cases in which this approximation might be justified.

— Spare random graphs : sparse Erdős-Rényi random graph are a model of random graphs in which one creates a network of size n where one decides for every pair $\langle i, j\rangle$ with a probability c/n whether this link belongs to the network or not. Given such a random graph described by and adjacency matrix $A \in \mathbb{R}^{n\times n}$ let us create the following spin system.

$$P(x_1, \cdots, x_n) = \frac{1}{Z}\exp\left(\sum_{1\leq i<j\leq n} x_i J_{ij} x_j\right),\ x_i = \pm 1 \tag{3.85}$$

$$J_{ij} = \mathcal{N}(0,1)A_{ij} \tag{3.86}$$

The interaction are non zero only when there is a link between x_i and x_j in the graph described by A. These random sparse graphs have the property of being locally tree-like, this means here that the shortest loop going through a typical node i will have a typical length of order $O(\log(N))$. On figure 3.6 this would mean that any correlation between variables x_j and x_k would have to come through a lengthy loop of order $\log(N)$ along which the correlation would decay which warrants (3.84). For instance the correctness of the BP approach was proved for two families of random sparse graph in [COP16]. Correctness here means that the true messages are asymptotic BP fixed point BP and that the Bethe Free energy is equal to the real free energy in the asymptotic regime.

— Another case in which the approximation can be warranted is in the dense case. Consider once again an Ising spin system in which the J_{ij} are taken at random according to.

$$J_{ij} = \frac{J\mathcal{N}(0,1)}{\sqrt{N}} + \frac{J_0}{N}, J_{ij} = J_{ji} \tag{3.87}$$

This model is known as the Sherrington-Kirkpatrick model. In that setting all the variables x_i interact, however these interactions are extremely weak, this makes it possible for the BP equation to be exact in the large N limit for some values of J and J_0. This model is central to the study of spin glasses and disordered systems. We will present it with more details in the next section.

3.6 Spin glasses and the Sherrington-Kirkpatrick model

Few fields have been as fertile in term of cross-field contribution as the study of spin glass system. The first encounter of physicist with spin glass system can be traced to the study of AgMn and CuMn composites by Nobel and Chatenier in 1959 [DCDNB66] (Here the Manganese is introduced as an impurity). And by Zimmerman and Hoare in 1960 [ZH60] in CuMn composite.

These researcher observed a linear capacity $C(T) \sim T$ around $T = 0$. This linear thermal capacity could not be explained by conduction electron in the material. The phenomena observed suggested that in the low temperature regime spins in the material would freeze in random direction. Further investigation would show that this freezing opens abruptly suggesting a phase transition phenomena in the system. In statistical physics phase transitions are often linked with a specific order parameter as is the average magnetization for Ferro/Para magnetic materials. However Neutron scattering experiments indicated that there was no periodicity in the way spins were ordered, this meant that this phase was neither a ferromagnetic or an antiferromagnetic materials. The question of how to describe theoretically these materials and what was that "mystery" order parameter associated with that phase transition remain opened until 1975. In their famous paper "Theory of spin glasses" [EA75] Edward and Anderson introduced the Edward-Anderson model (EA). The EA model is an Ising spin model given by the following Hamiltonian

$$P(\{s_i\} = \exp\left(\beta \sum_{\langle i,j \rangle} J_{ij} s_i s_j\right) . \tag{3.88}$$

Here the spins are placed on some lattice of dimension d Where the interaction J_{ij} have been sampled independently from a normal distribution $\mathcal{N}\left(0, \frac{J^2}{2d}\right)$. As βJ increases the system enters

a phase called the Spin- Glass phase. Edward and Anderson in their paper argue that the correct order parameter associated with this phase transition is the self overlap q_{EA} defined by

$$q_{\text{EA}} = \frac{1}{N}\sum_i \langle S_i \rangle^2 \,. \tag{3.89}$$

This parameter measures how "polarized" the spins are even though no particular magnetization direction is preferred by the system. As βJ increases it passes a threshold where q_{EA} goes from a zero value to a non zero value. We still lack a theoretical description of the low temperature phase of the Edward Anderson model (3.89). Some of the difficulties of the Edward-Anderson model come from the fact that this is not a Mean-Field model and there exist a spatial order because the spin live on a lattice.

A comparatively easier system to solve is the Sherrington-Kirkpatrick model (SK) introduced in 1975 [SK75] by Sherrington and Kirkpatrick in their famous paper "Solvable model of a Spin-Glass" [SK75]. The SK model is an infinite range spin system where the coupling J_{ij} between every spins have been sampled at random from some distribution.

$$P(s_1, \cdots, s_N) = \frac{1}{Z}\exp\left(\beta \sum_{1\leq i<j\leq N} J_{ij} s_i s_j\right) \tag{3.90}$$

with

$$J_{ij} = \frac{J_0}{N} + \frac{J}{\sqrt{N}}\mathcal{N}(0,1) \tag{3.91}$$

There are $N \gg 1$ spins in the system and all the coupling J_{ij} are of order $O(1/\sqrt{N})$ and have been sampled independently from a normal distribution. Therefore all the spins interact weakly with one another.

- The $\mathbf{J_0/N}$ term is a ferromagnetic term that can if strong enough can induce a ferromagnetic order in the system.
- The $\mathbf{J\mathcal{N}(0,1)/\sqrt{N}}$ term is the term from which comes the disorder. The bigger this term is the more disorder will play a role in the physics of the system.
- The scaling $1/N$ of the ferromagnetic term and $1/\sqrt{N}$ of the disorder term are here to ensure that the energy free-energy and capacity of the system will be extensive in the size of the system N.

Since this model has both infinite range and weak coupling J_{ij} one could first think that some form of Mean-field like approach could succeed in describing this system. Unfortunately naive mean-field method are not exact in that system. For naive mean field to work (meaning being exact in the large system limit) the coupling would need to be $O\left(1/N\right)$ rather than $O\left(1/\sqrt{N}\right)$. However for some value of βJ and J_0/J some modified version of mean-field called the Thouless-Anderson-Palmer (TAP) equations [TAP77] can succeed as long as one remains outside of the spin-glass phase.

Depending on the value of βJ and J_0/J the system might belong to different phases. This is illustrated in Fig 3.7. These phases come from different symmetries of the problem that can be potentially broken. There are 2 symmetries in that system.

- **The $\mathbb{Z}_2$ symmetry** or ± 1 symmetry is just the usual $s_i \to -s_i$ symmetry that one encounters in any Ising spin model with 0 fields. Breaking or not this symmetry marks the border between a Ferromagnetic phase and a paramagnetic phase.

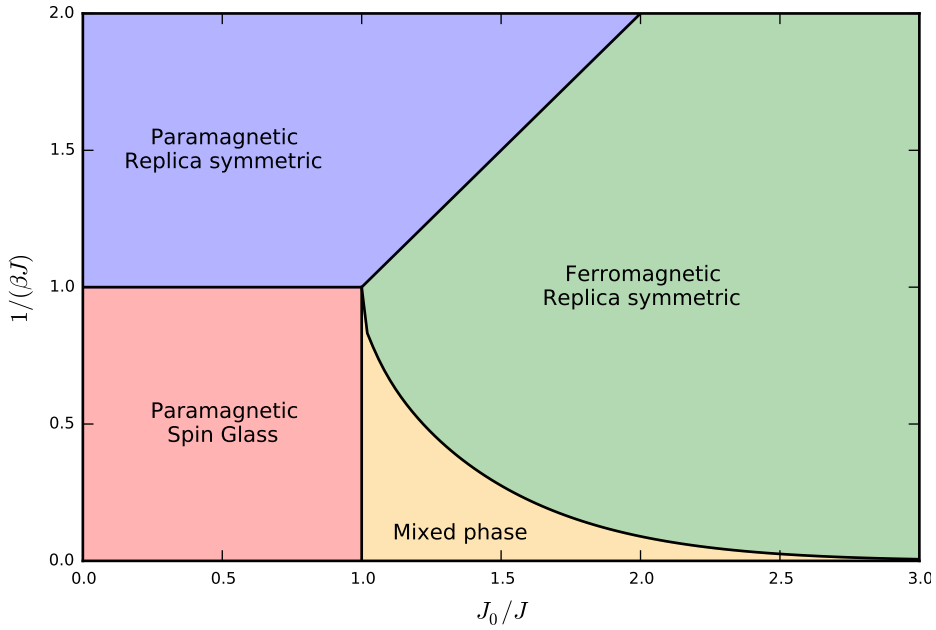


FIGURE 3.7 – Here we plot the phase diagram of the Sherrington-Kirkpatrick as a function of $1/(\beta J)$ and J_0/J. We denote 4 phases a paramagnetic phase a ferromagnetic phase a spin glass phase and a "mixed" phase. These $4 = 2 \times 2$ phases corresponds to the two symmetries that can be broken the $\mathbb{Z}_2$ symmetry and the replica symmetry. The paramagnetic phase is the replica symmetric phase where the system has a zero magnetization. The ferromagnetic phase breaks the $\mathbb{Z}_2$ and has non-zero magnetization but remains replica symmetric. The spin glass phase breaks the replica symmetry but has on average zero magnetization. The mixed phase has at the same time a breaking of the replica symmetry and a non zero magnetization.

— **The replica symmetry** is a symmetry that appears when trying to analyze spin-glass system. When doing replica computation (introduced in the following section) one introduces replicas or copies of the system (These replicas do not interact with one another). Doing this computation one would expect to see a permutation symmetry between all these replicas of the system. It is this permutation symmetry that can end up being broken. What it means exactly for the replica symmetry to be broken will be further explained in the following paragraphs.

To interpret what it means for the replica symmetry to be broken let us go back to the definition of the partition function and introduce the notion of Gibbs state. The partition function of a system is given by

$$Z = \sum_{x \in \mathcal{X}} \exp\left(\beta H(x)\right) . \tag{3.92}$$

Being given a partition $\mathcal{X}_1, \cdots, \mathcal{X}_m$ of $\mathcal{X}$ the configuration space in which x lives, we can write

$$Z = \sum_{k=1\cdots m} Z_k \,, \tag{3.93}$$

Where

$$Z_k = \sum_{x \in \mathcal{X}_k} \exp\left(\beta H(x)\right) . \tag{3.94}$$

Now the "Gibbs state representation" Amounts to choosing the $\mathcal{X}_1, \cdots, \mathcal{X}_m$ so that they satisfy three "properties".

— **Probability Weight conditions** This conditions amount to saying that most of the weight of the Gibbs distribution is held by the sets $\mathcal{X}_1, \cdots, \mathcal{X}_m$. Quantitatively this means

$$Z \approx \sum_{k=1\cdots m} Z_k \,, \tag{3.95}$$

In our setting this will mean

$$\log(Z) - \log\left(\sum_{k=1\cdots m} Z_k\right) \ll N \,, \tag{3.96}$$

Of course we ask for the intersection of the sets to be empty $\mathcal{X}_k \cap \mathcal{X}_l = \varnothing$.

— **Dynamical separation of the State :** The second conditions asks that the conditional Gibbs density $p_{\mathcal{X}_k}(x)$ are also Gibbs State. Where $p_{\mathcal{X}_k}(x)$ are

$$p_{\mathcal{X}_k}(x) = \frac{\mathbb{1}\left(x \in \mathcal{X}_k\right)}{Z_k} \exp\left(\beta H(x_k)\right) \,. \tag{3.97}$$

A Gibbs state is a density probability that is invariant by a Markov Chain which satisfy global balance of the Gibbs distribution of the system. Of course most of the time there is only one real Gibbs State namely the Gibbs distribution since as soon as a Markov Chain has a path of non zero probability to go anywhere in the configuration space then it means that a Markov Chain will equilibrate toward the Gibbs distribution. The subtility here lies in the time it takes for this equilibration to occur. If it takes typically $O\left(\exp(cN)\right)$ step for a Markov Chain initialized according to $p_{\mathcal{X}_k}(x)$ to jump from $\mathcal{X}_k$ to any $\mathcal{X}_l$ then saying that $p_{\mathcal{X}_k}(x)$ is a Gibbs state is a good approximation. To retake the example of model (3.23) for $J_0 > 1$ there exist two Gibbs state each with average magnetization $m = \pm m_{\text{eq}}$ but any Markov Chain that is only allowed to flip a finite number of spins at each step will take exponential time to go from $m = +m_{\text{eq}}$ to $m = -m_{\text{eq}}$. Of course to make this idea of "nearly" a Gibbs state work one has to reduce the scope of Markov Chain one is allowed to consider maybe for instance by putting a born on the maximum number of spins that one is allowed to flip at once during a MC step.

— **Quick Equilibration inside a given $\mathcal{X}_k$:** This condition means that no matter where one initializes a Markov Chain in $\mathcal{X}_k$ the time to equilibrate toward $p_{\mathcal{X}_k}(x)$ will be short (sub exponential in the size of the system). This is equivalent to saying that $\log(1 - \lambda_2) = o(N)$ where λ_2 is the second eigenvalue of the transition matrix of the Markov Chain limited to living inside $\mathcal{X}_k$.

This Gibbs State representation will allow us to try and understand the structure of the Gibbs distribution when the replica symmetry breaks down. Essentially the difference between different replica symmetric solution and replica breaked description lies in the number of Gibbs state that is required to accurately describe the system.

— **Replica Symmetric solution :** It is possible to describe the system using a finite number of Gibbs State (or at least a sub exponential number of them).

$$\log m = o(N) \tag{3.98}$$

For this ferromagnetic system there are two Gibbs State one with positive magnetization and one with negative. A nice mental representation of this case would be that the system can be well described by a high dimensional particle stuck inside a quadratic potential. There might be a finite number of quadratic potential this particle can be stuck in, but not so many that this would have major consequences on the physics of the system.

— **Replica Symmetric breaking :** Here in order to describe the Gibbs distribution in term of a sum of Gibbs state one requires an exponential number of Gibbs state.

$$\log(m) = O(N) \tag{3.99}$$

To understand why we say that this corresponds to a replica symmetry breaking of the system. Let us introduce $\{s_i^a\}$ and $\{s_j^b\}$ two configuration of spins sampled from the Gibbs distribution (or coming from two copies or replicas of the system). If one where to compute the overlap between these two configurations

$$q_{ab} = \frac{1}{N} \sum_{1 \leq i \leq N} s_i^a s_i^b \tag{3.100}$$

We would see that this parameter will take different value ($\pm O(1/\sqrt{N})$) depending on whether $\{s_i^a\}$ and $\{s_i^b\}$ belongs to the same $\mathcal{X}_k$ or not.

— **Refinement of the replica symmetry breaking :** There is a refinement around this idea of a replica symmetry breaking. There are multiple way for the replica symmetry to be broken to keep track of these different ways physicists talk about one step, two step, three step,... and infinite step replica symmetry breaking. This number of "step" describe the type of structure that the all the Gibbs state form. At least in the SK model there might be a structure to be investigated in how these Gibbs state are organized. The parameter to keep track of to understand this structure is the overlap function between configuration q_{ab} (3.100). This will be explained with more detail in the replica section.

The replica symmetry breaking and the breaking of the configuration space into an exponential number of Gibbs state has consequences on both BP equations and on replica computations solution.

— **For the Cavity equation** which for the SK model can be thought of as a refinement of the Naive Mean-Field equation this means that the Gibbs distribution can not be described by a single fixed point. And there now exist an exponential number of fixed point to these equations whose contribution to the free-energy have to be taken into account for all of them. The "fix" to this problem is to introduce the 1-RSB cavity equations where the parameters of the cavity equation are themselves treated as fluctuating variables. This is well explained in [MM09].

— **For the replica computation** this structure of the Gibbs state can be treated using the Parisi solution of the replica equation. This is explained with more detail in the replica section 3.7

3.7 Replica method

A lot of properties of the systems we will analyze are encoded in the free energy of the system. Like often in statistical physics the problem lies in computing the free energy.

The replica trick is a theoretical method to compute the free energy of systems exhibiting disorder. Suppose that one wants to compute the free energy of the following system.

$$P(x_1, \cdots, x_n) = \frac{1}{Z} \exp\left(\sum_{1 \leq i < j \leq N} x_i J_{ij} x_j + h \sum_{1 \leq i \leq N} x_i \right), \; x_i = \pm 1 \tag{3.101}$$

$$J_{ij} = \frac{J\mathcal{N}(0,1)}{\sqrt{N}} + \frac{J_0}{N} \tag{3.102}$$

Such a system is called the planted Sherrington-Kirkpatrick model. In order to analyze the property in the large N limit of this system we want to access the the free energy per spin ϕ of this systems.

$$\Phi(J_{ij}, h) = \log(Z) = O(N) \tag{3.103}$$

$$\Phi(J_{ij}, h)/N = O(1) \tag{3.104}$$

Any observable we can think of can be computed by taking a derivative with respect to some variable of ϕ. For example the average magnetization of the system can be computed using

$$\left\langle \sum_{1 \leq i \leq N} x_i \right\rangle = \frac{\partial \Phi}{\partial h} \,. \tag{3.105}$$

We expect two things to happen.

— Self averaging of the free energy per spin. This mathematical property states that

$$\log Z(J_{ij}) - \mathbb{E}_{J_{ij}} \left[\log Z(J_{ij}) \right] = O\left(\frac{1}{\sqrt{N}} \right) \tag{3.106}$$

Where we have averaged the free energy over the all the possible J_{ij}.
This means that in the large system size limit the free-energy per spin of a typical random draws of the disorder give rise to fluctuations of the free energy per spin of order $1/\sqrt{N}$.

— Existence of a thermodynamic limit : We also make the assumption that there is a thermodynamic limit to this problem, this means that the average free energy per spin has a limit in the large N limit.

$$\phi(J, J_0, h) = \lim_{N \to +\infty} \frac{1}{N} \mathbb{E}_{J_{ij}} \left[\log Z(J_{ij}, h) \right] \tag{3.107}$$

The reason why one wants to average the free energy and not the partition function is that the free energy is a self-averaging quantity meaning that computing it's mean will tell us something about a typical draw of the disorder. The partition function in the general case is not a self

averaging quantity. This means that in the general case one has

$$\phi(h) - \lim_{N\to+\infty} \frac{1}{N} \log \mathbb{E}_{J_{ij}} \left[Z(J_{ij}, h) \right] = O(1) \tag{3.108}$$

In the general case computing the average of the partition function does not allow us to compute ϕ. The problem stem from the fact that when computing the average of the partition function some rare configuration of the disorder that give rise to huge values of the partition function are going to dominate the average. These rare configurations would not dominate the average were they tempered by a logarithm.

We want to compute the average $\log Z$, this is hard to do. Most of the time it is much easier to compute the average of $Z^n, n \in \mathbb{N}$. The main idea behind the replica method is this

— Compute $\mathbb{E}\left[Z^n\right]$ pour $n \in \mathbb{N}$. Obtain a formula in n. The average of Z^n is computed by introducing replicas or copies of the system.
— Use the fact that $\log(x)$ can be written as

$$\log(x) = \lim_{n\to 0} \frac{x^n - 1}{n} \,. \tag{3.109}$$

Then "abuse" the analytical formula obtained for $n \in \mathbb{N}$ to allow $n \in \mathbb{R}$ to take the limit $n \to 0$ in (3.109) and obtain the formula for $\mathbb{E}\left[\log Z\right]$.

Let us present this computation in the case of the Sherrington-Kirkpatrick model.

$$Z^n = \mathrm{Tr}_X \left[\exp\left(\sum_{1\le i<j\le N} x_i J_{ij} x_j + h \sum_{1\le i\le N} x_i \right) \right]^n \tag{3.110}$$

$$Z^n = \prod_{1\le a\le n} \mathrm{Tr}_X \left[\exp\left(\sum_{1\le i<j\le N} x_i^a J_{ij} x_j^a + h \sum_{1\le i\le N} x_i^a \right) \right] \tag{3.111}$$

$$Z^n = \mathrm{Tr}_{X^a} \left[\exp\left(\sum_{1\le i<j\le N} J_{ij} \left[\sum_{1\le a\le n} x_i^a x_j^a \right] + h \sum_{1\le i\le N} \sum_{1\le a\le n} x_i^a \right) \right] \tag{3.112}$$

Here we compute the average Z^n by introducing copies or replicas of the system. The x_i^a for different a belong to copies of the system that share the same Hamiltonian. At this stage no interaction exist between different replicas of the system.

Here one can take the average with respect to the disorder J_{ij}.

$$\mathbb{E}\left[Z^n\right] = \mathrm{Tr}_{X^a} \left[\exp\left(\frac{1}{N} \sum_{1\le i<j\le N} J^2 \left[\sum_{1\le a\le n} x_i^a x_j^a \right]^2 + \frac{J_0}{N} \sum_{1\le i<j\le N} \sum_{1\le a\le n} x_i^a x_j^a + h \sum_{1\le i\le N} \sum_{1\le a\le n} x_i^a \right) \right] \tag{3.113}$$

$$\mathbb{E}\left[Z^n\right] = \mathrm{Tr}_{X^a} \left[\exp\left(\frac{J^2}{4N} \sum_{1\le a,b\le n} \sum_{1\le i,j\le N} x_i^a x_j^a x_i^b x_j^b + \frac{J_0}{2N} \sum_{1\le i,j\le N} \sum_{1\le a\le n} x_i^a x_j^a + h \sum_{1\le i\le N} \sum_{1\le a\le n} x_i^a \right) \right] \tag{3.114}$$

$$\mathbb{E}\left[Z^n\right] = \mathrm{Tr}_{X^a}\left[\exp\left(\frac{nNJ^2}{4} + \frac{NJ^2}{2}\sum_{1\leq a<b\leq n}\left(\frac{1}{N}\sum x_i^a x_i^b\right)^2 + \frac{NJ_0}{2}\sum_{1\leq a\leq n}\left(\frac{1}{N}\sum x_i^a\right)^2 + Nh\sum_{1\leq a\leq n}\left(\frac{1}{N}\sum x_i^a\right)\right)\right] \tag{3.115}$$

By using the Hubbard-Stratonovich identity one gets

Here all the spins x_i^a, x_j^b are decoupled for $i \neq j$.

$$\mathbb{E}\left[Z^n\right] = \int\prod_{a<b}\mathrm{d}q_{ab}\prod_a \mathrm{d}m_a \mathrm{Tr}_{X^a}\left[\exp\left(\frac{nNJ^2}{4} - \sum_{1\leq a<b\leq n}\frac{NJ^2q_{ab}^2}{2} - \sum_{1\leq a\leq n}\frac{NJ_0m_a^2}{2} + \sum_{1\leq i\leq N}\left(\sum_{1\leq a<b\leq n} x_i^a x_i^b q_{ab}J^2 + \sum_{1\leq a\leq n} NJ_0 x_i^a m_a + h\sum_{1\leq a\leq n} x_i^a\right)\right)\right] \tag{3.116}$$

$$\mathbb{E}\left[Z^n\right] = \int\prod_{a<b}\mathrm{d}q_{ab}\prod_a \mathrm{d}m_a \exp\left(\frac{nNJ^2}{4} - \sum_{1\leq a<b\leq n}\frac{NJ^2q_{ab}^2}{2} - \sum_{1\leq a\leq n}\frac{NJ_0m_a^2}{2} + N\log\hat{I}(q_{ab}, m_a, h)\right) \tag{3.117}$$

Where

$$\hat{I}(q_{ab}, m_a, h) = \sum_{\{x_a\}}\exp\left(\sum_{1\leq a<b\leq n} x^a x^b q_{ab}J^2 + \sum_{1\leq a\leq n} J_0 x^a m_a + h\sum_{1\leq a\leq n} x^a\right) \tag{3.118}$$

Because of the factor N that appears in (3.117) this integral can be computed using a saddle-point approximation. Looking for the maximum value of the integrand yields us with.

$$q_{ab} = \frac{\partial\log\hat{I}}{\partial q_{ab}} = \langle x_a x_b\rangle_{\hat{I}} \tag{3.119}$$

$$m_a = \frac{\partial\log\hat{I}}{\partial m_a} = \langle x_a\rangle_{\hat{I}} \tag{3.120}$$

Where the average of the spins $x_a, a \in \{1; n\}$ are taken with respect to this normalized density probability.

$$\frac{1}{\hat{I}(q_{ab}, m_a)}\exp\left(\sum_{1\leq a<b\leq n} x^a x^b q_{ab}J^2 + \sum_{1\leq a\leq n} J_0 x^a m_a + h\sum_{1\leq a\leq n} x^a\right) \tag{3.121}$$

The interpretation of the m_a is easy it is just the average magnetization of replicas a.

The q_{ab} are a bit trickier to interpret. Given that different replicas share the same Hamiltonian one can expect that the typical sample drawn from the different density probability are going to look similar. The matrix of overlap q_{ab} is a measure of how much replicas a and replicas b look alike.

Everything we have done so far for $n \in \mathbb{N}$ is exact. Such a computation is exact as long as $n \in \mathbb{N}$. For example one could use it to compute the average of the partition function, in physics this is called an annealed computation. An annealed computation can have it's usefulness it can give us a lower bound on the free energy. There are instances where such a bound are tight and where the computing the average value of $\mathbb{E}[Z]$ is enough to capture the main properties of the system.

3.7.1 Replica symmetric ansatz.

In most situation that we will encounter in this thesis the annealed computation is not enough and one has to really compute the average value of $\log(Z)$.

We want take $n \to 0$ in (3.117). In order to be able to do that we are going to look for an extrema of (3.117) for a certain form of matrix q_{ab} that allows one to take such a limit.

q_{ab} can be interpreted as an overlap between different replicas. There is a permutation symmetry between the different replicas. If the permutation symmetry between replicas is not broken then on expect to look for an extrema of (3.117) of the form.

$$\{q_{ab}\} = \begin{pmatrix} 1 & q & q & q & q \\ q & 1 & q & q & q \\ q & q & 1 & q & q \\ q & q & q & 1 & q \\ q & q & q & q & 1 \end{pmatrix} \tag{3.122}$$

$$q_{ab} = \begin{cases} 1 \,, \text{if} \,, a = b \\ q \,, \text{if} \,, a \neq b \end{cases} \tag{3.123}$$

$$m_a = m \tag{3.124}$$

By assuming this form of the matrix q_{ab} one gets starting from the integrand of (3.117).

$$\exp\left(\frac{nNJ^2}{4} - n(n-1)\frac{NJ^2q^2}{4} - \frac{nNJ_0m_a^2}{2} + N\log\hat{I}(q,m,h)\right) \tag{3.125}$$

And where

$$\hat{I}(q,m,h) = \sum_{\{x_a\}} \exp\left(\sum_{1\leq a<b\leq n} x^a x^b q J^2 + \sum_{1\leq a\leq n} J_0 x^a m_a + h \sum_{1\leq a\leq n} x^a \right) \tag{3.126}$$

$$\hat{I}(q,m,h) = \sum_{\{x_a\}} \exp\left(\frac{qJ^2}{2}\left[\sum_{1\leq a} x^a\right]^2 - \frac{qJ^2 x^{a2}}{2} + \sum_{1\leq a\leq n} J_0 x^a m_a + h \sum_{1\leq a\leq n} x^a \right) \tag{3.127}$$

$$\hat{I}(q,m,h) = \int \mathrm{d}W \exp\left(\frac{-W^2}{2}\right) \sum_{\{x_a\}} \exp\left(J\sqrt{q}W \left[\sum_{1\leq a} x^a\right] - \frac{qJ^2 x^{a2}}{2} \right. \tag{3.128}$$

$$\left. + \sum_{1\leq a\leq n} J_0 x^a m_a + h \sum_{1\leq a\leq n} x^a \right) \tag{3.129}$$

$$\hat{I}(q,m,h) = \int \mathrm{d}W \exp\left(\frac{-W^2}{2}\right) \left[2\cosh(J_0 m_a + h + J\sqrt{q}W) \exp\left(-qJ^2/2\right)\right]^n \tag{3.130}$$

Therefore when $n \to 0$ one has to first order in n.

$$\frac{\log \hat{I}(q,m,h)}{n} = \int \mathrm{d}W \exp\left(\frac{-W^2}{2}\right) \log\left(2\cosh(J_0 m + h + J\sqrt{q}W)\right) - qJ^2/2 \tag{3.131}$$

Therefore one has

$$\lim_{n\to 0} \frac{\mathbb{E}\left[Z^n\right] - 1}{nN} = \mathrm{Extrema}_{m,q} \left[\frac{J^2(1-q)^2}{4} - \frac{J_0 m^2}{2} + \log(2) \right.$$
$$\left. + \mathbb{E}_W \left[\log\cosh\left(J_0 m + h + J\sqrt{q}W\right)\right] \right] \tag{3.132}$$

The variable W on which the average is a Gaussian variable of mean 0 and variance 1. Here m and q are chosen so that the expression in (3.126) is an extrema in m and q. Extremising with respect to m and q gives us the Replica Symmetric update equations.

$$q = \mathbb{E}_W \left[\tanh\left(J_0 m + h + J\sqrt{q}W\right)^2\right] \tag{3.133}$$

$$m = \mathbb{E}_W \left[\tanh\left(J_0 m + h + J\sqrt{q}W\right)\right] \tag{3.134}$$

Equation (3.132) is called the replica symmetric free energy of the Sherrington-Kirkpatrick model.

3.7.2 Breaking of the replica symmetric

In the previous sub-section we assumed that the overlap matrix that would maximize the free energy would be preserve the symmetry between replicas. This assumption turns our to be right only for coupling constant βJ that is small enough. For high enough coupling constant

J the permutation symmetry between replicas breaks. This in turn make it so that one can find other configuration of that overlap matrix q_{ab} that give rise to a higher free energy. The solution introduced by Giorgio Parisi in [Par79] in one where this matrix q_{ab} has block structure as shown in Fig (3.8).

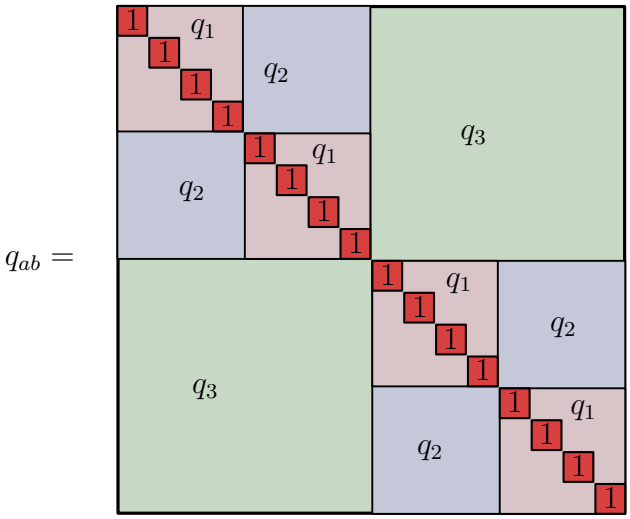


FIGURE 3.8 – In this figure we present the block structure of the q_{ab} matrix that corresponds to the Parisi solution of the SK model. This structure here corresponds to a 2 step replica symmetry breaking. The corresponding structure in term of Gibbs state is illustrated in Fig 3.9

This structure describes a Breaking of the configuration space into a structure given in Fig 3.9

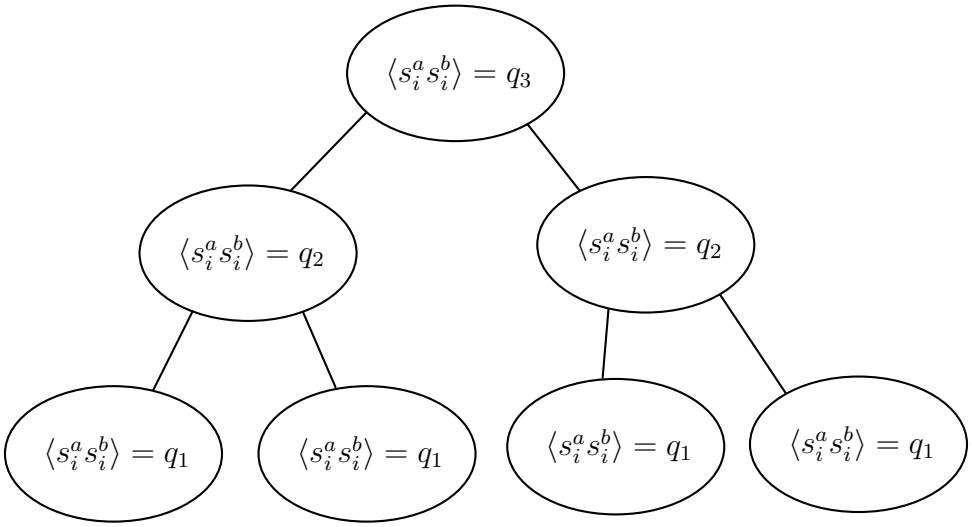


FIGURE 3.9 – In this figure we present a tree representation of the breaking of the configuration space into Gibbs state of different overlap. Two configurations taken according to the Gibbs distribution will have an average overlap between them of q_3. The Gibbs distribution can then be separated into an exponential number of Gibbs states where the average overlap between configuration is q_2 these Gibbs-State can be separated into Gibbs state where the overlap is q_1. It is a visual representation of the structure of overlap matrix given in Fig 3.8 under which one looks for a maxima of the free energy in the Parisi solution framework.

Chapitre 4

Results

4.1 Introduction

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

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

$$P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\leq i\leq N} P_X(x_i) \prod_{1\leq i<j\leq N} e^{g(Y_{ij}, x_i^\top x_j/\sqrt{N})}\,. \tag{4.1}$$

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

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

$$P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\leq i\leq N} P_X(x_i) \prod_{1\leq i<j\leq N} e^{g(Y_{ij}, f(x_i,x_j)/\sqrt{N})}\,. \tag{4.2}$$

Here $Y_{ij} \in \mathbb{R}^{N\times N}$. In this case Y_{ij} is a symmetric matrix. x_i can live in any space. The signal w is given by a general function $w = f(x_i, x_j)$. This is known as the graphon case.

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

$$P(U,V|Y) = \frac{1}{Z_{UV}(Y)} \prod_{1\leq i\leq N} P_U(u_i) \prod_{1\leq j\leq M} P_V(v_j) \prod_{1\leq i\leq N, 1\leq j\leq M} e^{g(Y_{ij}, u_i^\top v_j/\sqrt{N})}\,. \tag{4.3}$$

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

— **Symmetric p vector-spins glass model** :

$$P(X|Y) = \frac{1}{Z_{X^p}(Y)} \prod_{1\leq i\leq N} P_X(x_i) \prod_{1\leq i_1<\cdots<i_p\leq N} e^{g\left(Y_{i_1\cdots i_p}, \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{k=1\cdots r} x_{i_1 k} x_{i_2 k}\cdots x_{i_p k}\right)}. \quad (4.4)$$

Defined as above, this time $Y_{i_1 i_2\cdots i_p} \in \otimes^p \mathbb{R}^N$ is a symmetric tensor of order p. $X \in \mathbb{R}^{N\times r}$. Again we denote by x_i the *vector-spins* of dimension r that collect rows of matrices X. In this case the graph of interactions between spins is an hypergraph.

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

The general nature of the present work stems from the fact that the probability distributions P_X, P_U, P_V and the function g are very generic (assumptions are summarize in section 4.1.2). These functions can even depend on the node i or edge ij. For simplicity we will treat site-independent functions P_X, P_U, P_V and g, but the theory developed here generalizes very straightforwardly to the site or edge dependent case. From a statistical physics point of view the terms P_X, P_U, P_V play a role of generic local magnetic fields acting on the individual spins. Distributions P_X, P_U, P_V describe the nature of the vector-spin variables and the fields that act on them. The simplest example is the widely studied Ising spins for which $r = 1$ and $P_X(x) = \rho\delta(x-1)+(1-\rho)\delta(x+1)$, where ρ here would be related to the usual magnetic field h and inverse temperature β as $\rho = e^{\beta h}/(2\cosh\beta h)$. In this paper we treat a range of other examples with $r \geq 1$ and elements of x being both discrete or continuous. This involves for instance spherical spin models with P_X being Gaussian, or Heisenberg spins where $r = 3$ and each x is confined to the surface of a sphere.

Denoting

$$w_{ij} = x_i^\top x_j/\sqrt{N}\,, \quad \text{or} \quad w_{ij} = u_i^\top v_j/\sqrt{N}\,, \quad \text{or} \quad w_{i_1\cdots i_p} = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{k=1\cdots r} \left(x_{i_1}\circ\cdots\circ x_{i_p}\right)_k . \quad (4.5)$$

Where $\circ$ denotes the Hadamard product. According to symmetric, bipartite context or tensor context, the terms $g(Y, w)$ are then interactions between pairs or p tuples of spins that depend only on the scalar/Hadamard product between the corresponding vectors. We will also consider the case where the w_{ij} are given by a general function

$$w_{ij} = f(x_i, x_j)/\sqrt{N}\,. \quad (4.6)$$

The BP approach is flexible enough that we can treat this case naturally. Matrix multiplications in our equations will be replaced by integrations against a multiplicative Kernel. We introduce this case in order to analyze the behavior of objects called graphons. Graphons can be seen as a generative model for random graphs, this generative model can be seen as the continuous limit of the Stochastic Block Model.

The most commonly considered form of interaction in statistical physics is simply

$$g(Y, w) = \beta Y w \tag{4.7}$$

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

In order to complete the setting of the problem we need to specify how is the quenched disorder Y chosen. We will consider two main cases of the quenched disorder defined below. We note that even for problems where the matrix Y is not generated by either of the below our approach might still be relevant, e.g. for the restricted Boltzmann machine that is equivalent to the above bipartite model with Y that were learned from data (see for instance [GTK15, TMC+16] and the discussion below).

— **Randomly quenched disorder :** In this case the matrix elements of Y are chosen independently at random from some probability distribution $P(Y_{ij})$. In this thesis we will consider this distribution to be independent of N and in later parts for simplicity we will restrict its mean to be zero. This case of randomly quenched disorder will encompass many well known and studied spin glass models such as the Ising spin glass, Heisenberg spin glass or the spherical spin glass or the Hopfield model.

— **Planted models :** Concerning applications in data science this is the more interesting case and most of this paper will focus on it. In this case we consider that there is some *ground truth* value of $X_0 \in \mathbb{R}^{N\times r_0}$ (or $U_0 \in \mathbb{R}^{N\times r_0}$, $V_0 \in \mathbb{R}^{M\times r_0}$) with rows that are generated independently at random from some probability distribution P_{X_0} (or P_{U_0}, P_{V_0}). Then the disorder Y is generated element-wise as a noisy observation of the product $w^0_{ij} = x^{0,\top}_i x^0_j/\sqrt{N}$ (or $w^0_{ij} = u^{0,\top}_i v^0_j/\sqrt{N}$) via an output *channel* characterized by the output probability distribution $P_{\text{out}}(Y_{ij}|w^0_{ij})$.

4.1.1 Preliminaries on the planted setting

Bayes optimal inference

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

$$P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\le i\le N} P_{X_0}(x_i) \prod_{1\le i<j\le N} P_{\text{out}}\left(Y_{ij}\middle|\frac{x_i^\top x_j}{\sqrt{N}}\right) . \tag{4.8}$$

For the bipartite case it is

$$P(U, V|Y) = \frac{1}{Z_{UV}(Y)} \prod_{1\le i\le N} P_{U_0}(u_i) \prod_{1\le j\le M} P_{V_0}(v_j) \prod_{1\le i\le N, 1\le j\le M} P_{\text{out}}\left(Y_{ij}\middle|\frac{u_i^\top v_j}{\sqrt{N}}\right) . \tag{4.9}$$

Making link with the Boltzmann probability measures (2.6) and (2.7) we see that the *Bayes optimal inference* of the planted configuration is equivalent to the statistical physics of the above vector-spin models with

$$P_{X_0} = P_X\,, \quad P_{U_0} = P_U\,, \quad P_{V_0} = P_V\,, \quad P_{\rm out}(Y|w) = e^{g(Y,w)}\,. \tag{4.10}$$

This approach is optimal in the sense that the statistical estimator $\hat{X}$ computed from the data Y that minimizes the expected mean-squared error between the estimator $\hat{X}$ and the ground truth X_0 is given by the mean of the marginal of variable x_i in the probability distribution (4.8)

$$\hat{x}_i(Y) = \int \mathrm{d}x\, x\, \mu_i(x)\,, \quad \text{where} \quad \mu_i(x) = \int P(X|Y) \prod_{\{x_j\}_{j\neq i}} \mathrm{d}x_j\,. \tag{4.11}$$

Analogously for the bipartite case.

In the Bayes-optimal setting defined by conditions (2.8) the statistical physics analysis of the problem presents important simplifications known as the Nishimori conditions [Nis01, ZK16] (3.18), which will be largely used in the present paper. These conditions can be proven and stated without the usage of the methodology developed below, they are a direct consequence of the Bayesian formula for conditional probability and basic properties of probability distributions.

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

$$\forall w, \int \mathrm{d}Y\, P_{\rm out}(Y|w) = 1\,. \tag{4.12}$$

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

$$\forall w, \int \mathrm{d}Y\, P_{\rm out}(Y|w) \frac{\partial g(Y,w)}{\partial w} = \mathbb{E}_{P_{\rm out}(Y|w)}\left[\frac{\partial g(Y,w)}{\partial w}\right] = 0\,, \tag{4.13}$$

$$\forall w, \int \mathrm{d}Y\, P_{\rm out}(Y|w) \left[\left(\frac{\partial g(Y,w)}{\partial w}\right)^2 + \frac{\partial^2 g(Y,w)}{\partial w^2}\right] = \mathbb{E}_{P_{\rm out}(Y,w)}\left[\left(\frac{\partial g(Y,w)}{\partial w}\right)^2 + \frac{\partial^2 g(Y,w)}{\partial w^2}\right] = 0\,. \tag{4.14}$$

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

$$\frac{1}{\Delta} = \mathbb{E}_{P_{\rm out}(Y|w=0)}\left[\left(\frac{\partial g}{\partial w}\right)^2_{Y,w=0}\right]\,, \tag{4.15}$$

We also remind the reader of a property of Bayes optimal inference that were already discussed in section 3.1 namely equation (3.18). This is a simple yet important property that will lead to numerous simplifications in the Bayes optimal case and it will be used in several places of this paper, under the name *Nishimori condition*.

From the point of view of statistical physics of disordered systems the most striking property of systems that verify the Nishimori conditions is that there cannot be any *replica symmetry breaking* in the equilibrium solution of these systems [NS01, Nis01, ZK16]. This simplifies considerably the analysis of the Bayes-optimal inference. Note, however, that metastable

(out-of-equilibrium) properties of Bayes-optimal inference do not have to satisfy the Nishimori conditions and replica symmetry breaking might be needed for their correct description (this will be relevant in the cases of first order phase transition described in section 4.4.4).

4.1.2 The large size limit, assumptions and channel universality

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

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

For readers interested in inference problems, i.e. the planted setting, the $1/\sqrt{N}$ factor is the scaling of the signal-to-noise ratio for which inference of $O(N)$ unknown from $O(N^2)$ measurements is neither trivially easy nor trivially impossible. In the planted setting Y can be viewed as a random matrix with a rank-r perturbation. The regime where the eigenvalues of dense random matrices with low-rank perturbations split from the bulk of the spectral density is precisely when the strength of the perturbation is $O(1/\sqrt{N})$, see e.g. [BBAP05].

We are therefore looking at statistical physics models with $O(N^2)$ pairwise interactions where each of the interactions depend only weakly on the configuration of the vector-spins. As a consequence, properties of the system in the thermodynamic limit $N \to \infty$ depend only weakly on the details of the interaction function $g(Y_{ij}, w_{ij})$ with w_{ij} given by (4.5). The results of this paper hold for every function g for which the following Taylor expansion is well defined

$$e^{g(Y_{ij},w_{ij})} = e^{g(Y_{ij},0)} \left\{ 1 + \frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} w_{ij} \right. \tag{4.16}$$

$$\left. + \left[\left(\frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} \right)^2 + \frac{\partial^2 g(Y_{ij},w)}{\partial w^2}\Big|_{w=0} \right] \frac{w_{ij}^2}{2} + O(w_{ij}^3) \right\} . \tag{4.17}$$

In order to simplify the notation in the following we denote

$$S_{ij} \equiv \frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} , \tag{4.18}$$

$$R_{ij} \equiv \left(\frac{\partial g(Y_{ij},w)}{\partial w}\Big|_{w=0} \right)^2 + \frac{\partial^2 g(Y_{ij},w)}{\partial w^2}\Big|_{w=0} . \tag{4.19}$$

We will refer to the matrix S as the Fisher score matrix. The above expansion can now be

written in a more compact way

$$e^{g(Y_{ij},w_{ij})} = e^{g(Y_{ij},0)}\left[1 + S_{ij}w_{ij} + \frac{R_{ij}w_{ij}^2}{2} + O(w_{ij}^3)\right] = e^{g(Y_{ij},0)+S_{ij}w_{ij}+\frac{1}{2}(R_{ij}-S_{ij}^2)w_{ij}^2+O(w_{ij}^3)}\,. \quad (4.20)$$

Let us now analyze the orders in this expansion. In the Boltzmann measure (2.6) and (2.7) the terms $e^{g(Y_{ij},w_{ij})}$ appears in a product over $O(N^2)$ terms and $w = O(1/\sqrt{N})$. At the same time only terms of order $O(N)$ in the exponent of the Boltzmann measure influence the leading order (in N) of the marginals (local magnetizations), therefore all the terms that depend on 3rd and higher order of w are negligible. This means that the leading order of the marginals depend on the function $g(Y, w)$ only trough the matrices of its first and second derivatives at $w = 0$, denoted S and R (4.18-4.19). This means in particular that in order to understand the phase diagram of a model with general $g(Y, w)$ we only need to consider one more order than in the conventional Hamiltonian considered in statistical physics (4.7).

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

Gaussian Noise Channel :

$$g(Y,w) = \frac{-(Y-w)^2}{2\Delta} - \frac{1}{2}\log 2\pi\Delta\,, \qquad S_{ij} = \frac{Y_{ij}}{\Delta}\,, \qquad R_{ij} = \frac{Y_{ij}^2}{\Delta^2} - \frac{1}{\Delta}\,, \quad (4.21)$$

where Δ is the variance of the noise, for the specific case of the Gaussian output channel Δ is also the Fisher information as defined in eq. (4.15). The second example is the one most often considered in physics given by eq. (4.7)

Conventional Hamiltonian :

$$g(Y,w) = \beta Y w\,, \qquad S_{ij} = \beta Y_{ij}\,, \qquad R_{ij} = \beta^2 Y_{ij}^2\,. \quad (4.22)$$

with β being a constant called inverse temperature. Another quite different example of the output channel will be given to model community detection in networks in section 4.1.4.

4.1.3 Principal component analysis and spectral method

In that section we will present the spectral techniques that one might use to solve the matrix factorization problem. The reason it is useful to talk about these methods is that quite often the performance of Bayes inference method are deeply linked to the performance of spectral methods, for instance if your signal has mean 0 then there is phase transition of the posterior probability at the exact same place as the spectral transition appears. A good rule of thumb that one can apply to get a feel for what can one expect in matrix factorization problems is the following "As long as the constraints that one imposes on the prior $P_X(x)$ are not too harsh then the performance of spectral techniques and Bayesian inference will not be too far apart". Harsh constraint could mean very low sparsity, large rank r with correlations between the component of the x_i, essentially anything that make $P_X(x)$ not well approximated by a

Gaussian distribution. This "rule of thumb" is nothing but an empirical rule gained by analyzing a lot of phase diagram of matrix factorization system. Therefore having some understanding of the property of spectral methods is useful to ease the rest of the explanation.

We will focus mostly on matrices and not on tensor. Matrices are comparatively to tensors relatively simple objects to understand and analyze. Tensors have all sort of properties that make them hard to analyze, a lot of things that are easy to compute for matrices end up being NP-hard to compute for matrices this is well explained in [HL09].

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

$$\text{argmin}_{\tilde{Y}} \left[\sum_{1 \leq i \leq N, 1 \leq j \leq M} (Y_{ij} - \tilde{Y}_{ij})^2 \right] = \sum_{s=1}^{r} u_s \lambda_s v_s^\top \,, \tag{4.23}$$

where λ_s is the sth largest singular value of Y, and $u_s \in \mathbb{R}^N$, $v_s \in \mathbb{R}^M$ are the corresponding left-singular and right-singular vectors of the matrix Y. The above property is know as the Eckart-Young-Mirsky theorem [EY36]. In the symmetric case $Y_{ij} = Y_{ji}$ we simply replace the singular values by eigenvalues and the singular vectors by the eigenvectors.

$$\text{argmin}_{\tilde{Y}} \left[\sum_{1 \leq i \leq N, 1 \leq j \leq N} (Y_{ij} - \tilde{Y}_{ij})^2 \right] = \sum_{s=1}^{r} x_s \lambda_s x_s^\top \,, \tag{4.24}$$

The above unconstrained low-rank approximations of the matrix Y, eq. (4.23, 4.24), are also often referred to as *principal component analysis* (PCA), because indeed when the matrix Y is interpreted as N samples of M-dimensional data then the right-singular vectors v_s are directions of greatest variability in the data.

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

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

One might wonder how well do spectral methods perform in recovering the signal. This is a question for frequentist inference. In the case of rank 1 system with additive Gaussian noise

this question has been answered in [BGN11, BBAP05].

$$Y = \frac{XX^\top}{\sqrt{N}} + \sqrt{\Delta}\mathcal{N}(0,1)^{N\times N}_{\text{Sym}} \tag{4.25}$$

$$Y = \frac{UV^\top}{\sqrt{N}} + \sqrt{\Delta}\mathcal{N}(0,1)^{N\times M} \tag{4.26}$$

The take away message is that in each case there is a competition between the noise $\mathcal{N}(0,1)^{N\times N}$ noise the signal $\frac{XX^\top}{\sqrt{N}}$ a visual representation of this competition would be given by plotting the spectrum of Y for different value of Δ. Let us describe what happens for the symmetric case Y as one increases Δ the noise starting from zero. Our measure of how well the signal is reconstructed will be

$$\alpha(X_0, \hat{X}_{\text{PCA}}) = \frac{\left|X_0^\top \hat{X}_{\text{PCA}}\right|}{\|X_0^\top\|_2 \|\hat{X}_{\text{PCA}}\|_2} \tag{4.27}$$

$0 \leq \alpha(X_0, \hat{X}_{\text{PCA}}) \leq 1$ and is a measure of how well the signal was reconstructed. $\alpha(X_0, \hat{X}_{\text{PCA}}) = 1$ means perfect reconstruction. And $\alpha(X_0, X_{\text{Normal}}) \approx \pm\frac{1}{\sqrt{N}}$ is expected if one chooses a vector X_{Normal} from an isotropic normal distribution.

— $\boldsymbol{\Delta = 0}$: Y only has one non zero eigenvalue and PCA achieves perfect reconstruction $\alpha(X_0, \hat{X}_{\text{PCA}}) = 1$.

— $\boldsymbol{0 < \Delta < \Delta_c = \frac{\|X_0\|_2^4}{N^2}}$: The Noise created a bulk of $N-1$ eigenvalues of width $2\sqrt{N\Delta}$ centered around zero and distributed according to a semi-circle law, the eigenvectors of these eigenvalues have $O(1/\sqrt{N})$ correlation with the hidden solution X_0. One informative eigenvalue stays well separated from the bulk of eigenvalues the corresponding eigenvector is well correlated with the signal

$$\alpha(X_0, \hat{X}_{\text{PCA}}) = \sqrt{1 - \frac{\Delta}{\Delta_c}} \pm O\left(\frac{1}{\sqrt{N}}\right) \tag{4.28}$$

Where $\hat{X}_{\text{PCA}}(Y)$ is the estimator that outputs the top eigenvectors of Y. As Δ increases the separation between the bulk and the informative eigenvalue decreases.

— $\boldsymbol{\Delta > \Delta_c}$: The informative eigenvalue has penetrated the Bulk, all eigenvectors of Y have a vanishing correlation of order $O(1/\sqrt{N})$ with the solution.

$$\alpha(X_0, \hat{X}_{\text{PCA}}) = \pm O\left(\frac{1}{\sqrt{N}}\right) \tag{4.29}$$

We plot this spectrum in fig 4.1. Similar thing happen in the $UV^\top$ case where there is once again a critical value of Δ below which a singular value decomposition of Y yields us with eigenvectors that are positively correlated with U_0 and V_0. The result of the performance of naive PCA are summarized here.

$$XX^\top \textbf{ case : } \Delta_c = \frac{\|X_0\|_2^4}{N^2} \tag{4.30}$$

$$UV^\top \textbf{ case : } \Delta_c = \frac{\|U_0\|_2^2 \|V_0\|_2^2}{N^2} \tag{4.31}$$

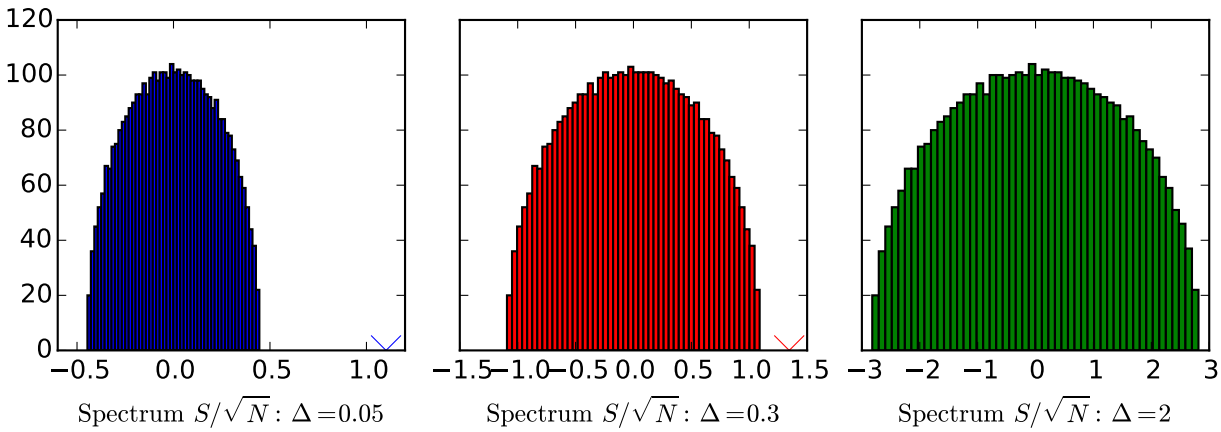


FIGURE 4.1 – We plot the spectrum of matrix $Y/\sqrt{N}$ in the $XX^\top$ case for $N = 2000$ and $\|X_0\|_2^2 = N$ and for 3 values of $\Delta = 0.05, 0.3, 2$. We see the bulk of noisy eigenvalue growing around zero until it hides the informative eigenvalues coming from the signal.

The overlap of the eigenvectors with the hidden solution for $\Delta < \Delta_c$ are then given by

$$XX^\top \textbf{ case :}$$

$$\alpha(X_0, \hat{X}_{\text{PCA}}) = \sqrt{1 - \frac{\Delta}{\Delta_c}} + O\left(\frac{1}{\sqrt{N}}\right) \tag{4.32}$$

$$UV^\top \textbf{ case :}$$

$$\alpha(U_0, \hat{U}_{\text{PCA}}) = \sqrt{\frac{1 - \frac{\Delta^2}{\Delta_c^2}}{1 + \frac{\Delta}{\Delta_c\sqrt{\alpha}}}} + O\left(\frac{1}{\sqrt{N}}\right) \tag{4.33}$$

$$\alpha(V_0, \hat{V}_{\text{PCA}}) = \sqrt{\frac{1 - \frac{\Delta^2}{\Delta_c^2}}{1 + \frac{\Delta\sqrt{\alpha}}{\Delta_c}}} + O\left(\frac{1}{\sqrt{N}}\right) \tag{4.34}$$

where

$$\alpha = \frac{M}{N} \tag{4.35}$$

The performance of spectral methods will be analyzed with more detail in section 4.4.1.

Why are spectral method a reasonable thing to do ? One might be perplexed in front of the ubiquity of spectral methods. After all if the task at hand is the analysis of data why would one expect the model

$$Y = UV^\top + \mathcal{N}(0, \Delta)^{N \times M} \tag{4.36}$$

to be a reasonable model to fit ? PCA method amount to solving this optimization problem

$$\sum_{\substack{1 \le i \le N \\ 1 \le j \le M}} \left(Y_{ij} - \frac{u_i^\top v_j}{\sqrt{N}}\right)^2 \tag{4.37}$$

Why a squared penalty ? Why a multiplication between u_i and v_j ? Let us imagine a more general model given by

$$w_{ij} = \frac{1}{\sqrt{N}} f(u_i, v_j) \Rightarrow P\left(Y|\{w_{ij}\}\right) = \prod_{\substack{1\leq i\leq N\\1\leq j\leq M}} P_{\text{out}}(Y_{ij}|w_{ij}) \tag{4.38}$$

Where we take the u_i and v_j to belong to the interval $[0;1]$. Now the signal w is given by a general function of u_i and v_j and the noise is not an additive Gaussian noise but a general Noise channel given by the conditional density probability $P_{\text{out}}(Y|w)$. Doing a MAP approach to this problem means finding u_i and v_j that maximizes this cost function

$$\text{argmax}_{\{u_i\},\{v_j\}} \prod_{\substack{1\leq i\leq N\\1\leq j\leq M}} P_{\text{out}}\left(Y_{ij}\middle|\frac{1}{\sqrt{N}} f(u_i, v_j)\right) . \tag{4.39}$$

Using the channel universality property described in section 4.1.2 this optimization problem is the same in the large N limit as this problem

$$\text{argmax}_{\{u_i\},\{v_j\}} \sum_{\substack{1\leq i\leq N\\1\leq j\leq M}} \left[S_{ij}\frac{1}{\sqrt{N}} f(u_i, v_j) + \frac{1}{2N}\left(R_{ij} - S_{ij}^2\right) f(u_i, v_j)^2\right] \tag{4.40}$$

Using the self-averaging property of the sum we will replace the term $(R_{ij} - S_{ij}^2)$ by it's averaged value $-1/\Delta'$. The function to optimize now becomes

$$\text{argmax}_{\{u_i\},\{v_j\}} \sum_{\substack{1\leq i\leq N\\1\leq j\leq M}} \left(S_{ij}\sqrt{\Delta'} - \frac{1}{\sqrt{N\Delta'}} f(u_i, v_j)\right)^2 \tag{4.41}$$

The problem still remain that $f(u_i, v_j) \neq u_i v_j$ and that the optimization problem we have written still is not a PCA optimization problem over the matrix S. To see how all of this relate to the spectrum of S let us represent the function $f(u, v)$ in a different way. Suppose that we now limit u_i and v_j to be real numbers $\in [0;1]$. Let us discretize this interval into r value $\{(k-1)/r, k \in \{1, \cdots, r\}\}$. We will now do a mapping from one representation of the problem to another. The variables u_i and v_j that before lived in $[0;1]$ will now be represented by r dimensional vectors $\vec{u}_i \in \{\vec{e}_k\} \subset \mathbb{R}^{r\times 1}$ and $\vec{v}_j \in \{\vec{e}_k\} \subset \mathbb{R}^{r\times 1}$. Where the vectors $\vec{e}_k$ form the canonical basis of $\mathbb{R}^{r\times 1}$ where, $(\vec{e}_k)_i = \delta_k^i$. We then map from one representation to the other according to these rules

$$u_i \in \left[\frac{k-1}{r}; \frac{k}{r}\right[\Leftrightarrow \vec{u}_i = \vec{e}_k \tag{4.42}$$

$$v_j \in \left[\frac{k-1}{r}; \frac{k}{r}\right[\Leftrightarrow \vec{v}_j = \vec{e}_k \tag{4.43}$$

$$f(u_i, v_j) \leftrightarrow \vec{u}_i^{\top} F \vec{v}_j \tag{4.44}$$

where

$$F_{kk'} = \vec{e}_k^{\top} F \vec{e}_{k'} = f\left(\frac{k-1}{r}, \frac{k'-1}{r}\right) \tag{4.45}$$

Of course this mapping is only exact in the $r \to +\infty$ limit. But if one takes r large enough then this might be a good enough approximation to describe accurately the function f. We now have once again transformed our optimization problem.

$$\text{argmax}_{\{\vec{u}_i\}\in\{\vec{e}_k\}^N,\{\vec{v}_j\}\in\{\vec{e}_k\}^M} \sum_{\substack{1\leq i\leq N\\1\leq j\leq M}} \left(S_{ij}\sqrt{\Delta'} - \frac{\vec{u}_i^\top F \vec{v}_j}{\sqrt{N\Delta'}}\right)^2 \tag{4.46}$$

Under this form this almost look like a Singular Value Decomposition (SVD) problem the only thing different is that we have this constraint that $\vec{u}_i, \vec{v}_j \in \{\vec{e}_k\}$. But if we just relax this constraint and allow the variables $\vec{u}_i$ and $\vec{v}_j$ to live in the whole space $\mathbb{R}^{r\times 1}$ then the problem become

$$\text{argmax}_{\{\vec{u}_i\},\{\vec{v}_j\}} \sum_{\substack{1\leq i\leq N\\1\leq j\leq M}} \left(S_{ij}\sqrt{\Delta'} - \frac{\vec{u}_i^\top F \vec{v}_j}{\sqrt{N\Delta'}}\right)^2 \tag{4.47}$$

This is an easy optimization problem to solve. Let us call $\hat{u}_i \in \mathbb{R}^{r\times 1}$,$\hat{v}_j \in \mathbb{R}^{r\times 1}$ a solution to the SVD optimization problem.

$$\text{argmax}_{\{\hat{u}_i^{\text{SVD}}\},\{\hat{v}_j^{\text{SVD}}\}} \sum_{\substack{1\leq i\leq N\\1\leq j\leq M}} \left(S_{ij} - \frac{\hat{u}_i^{\text{SVD}\top}\hat{v}_j^{\text{SVD}}}{\sqrt{N}}\right)^2 \tag{4.48}$$

It is easy to prove that a solution to (4.47) can be given by

$$\vec{u}_i = A^\top \sqrt{\frac{\Delta}{C}}\hat{u}_i^{\text{SVD}} \tag{4.49}$$

$$\vec{v}_j = B^\top \sqrt{\frac{\Delta}{C}}\hat{v}_j^{\text{SVD}} \tag{4.50}$$

where

$$A, B \in O(r), C \text{ is a positive diagonal matrix} \tag{4.51}$$

$$F = ACB \in \mathbb{R}^{r\times r} \tag{4.52}$$

We therefore see that even though we started from a much more "general" model for our data we were able to link it to the singular value decomposition of some matrix S_{ij}. This is a simple explanation to justify the usage of spectral methods. There remains a subtility though which is that we did not perform in the end a SVD on the matrix Y_{ij} directly (assuming that the Y_{ij} observed are numbers) but on the matrix S_{ij}. To go from one to the other we had to assume a shape of the channel and a corresponding function $g(Y, w)$. Some functions $g(Y, w)$ might be better that other as is illustrated in Fig 4.3. A similar treatment of the $XX^\top$ case can be performed.

4.1.4 Examples and applications

The Boltzmann measures (2.6) and (2.7) together with the model for the disorder Y can be used to describe a range of problems of practical and scientific interest studied previously in physics and/or in data sciences. In this section we list several examples and applications for each of the four categories – the symmetric and bipartite case, and the randomly quenched and planted disorder.

Examples with randomly quenched disorder

Sherrington-Kirkpatrick (SK) model. The SK model [SK75] stands at the roots of the theory of spin glasses. It can be described by the symmetric Boltzmann measure (2.6) with the conventional Hamiltonian $g(Y, w) = \beta Y w$.

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

$$P_X(x_i) = \rho\delta(x_i - 1) + (1 - \rho)\delta(x_i + 1). \tag{4.53}$$

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

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

The algorithm developed in this thesis for the general case corresponds to the Thouless-Anderson-Palmer [TAP77] equations for the SK model. The theory developed here correspond to the replica symmetric solution of [SK75]. Famously this solution is wrong below certain temperature where effects of replica symmetry breaking (RSB) have to be taken into account. In this thesis we focus on the replica symmetric solution, that leads to exact and novel phase diagrams for the planted models. The RSB solution in the present generic setting will be presented elsewhere. We present the form of the TAP equations in the general case encompassing a range of existing works.

Spherical spin glass. Next to the SK model, the spherical spin glass model [KTJ76] stands behind large fraction of our understanding about spin glass. Mathematically much simpler than the SK model this model stands as a prominent case in the development in mathematical physics. The spherical spin glass is formulated via the symmetric Boltzmann measure (2.6) with the conventional Hamiltonian $g(Y, w) = \beta Y w$. The function $P_X(x_i) = e^{-x_i^2/2}$ with $x_i \in \mathbb{R}$ enforces (canonically) the spherical constraint $\sum_i x_i^2 = N$. External magnetic field can also be included in $P_X(x_i)$.

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

Heisenberg spin glass. In Heisenberg spin glass [Som81] the Hamiltonian is again the conventional symmetric one with randomly quenched disorder. The spins are 3-dimensional vectors, $x_i \in \mathbb{R}^3$, of unit length, $x_i^\top x_i = 1$. Magnetic field influences the direction of the spin so that

$$P_X(x_i) = e^{\beta h^\top x_i}\,, \tag{4.54}$$

where $h \in \mathbb{R}^3$. The more general r-component model was also studied extensively in the spin glass literature [GT81].

Restricted Boltzmann Machine. Restricted Boltzmann machines (RBMs) are one of the triggers of the recent revolution in machine learning called deep learning [Hin10, HOT06, LBH15]. The way RBMs are used in machine learning is that one considers the bipartite Boltzmann measure (2.7). In the training phase one searches a matrix Y_{ij} such that the data represented as a set of configurations of the u-variable have high likelihood (low energy). The v-variable are called the hidden units and columns of the matrix Y_{ij} (each corresponding to one hidden unit) are often interpreted as features that are useful to explain the structure of the data.

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

We note that the disorder Y_{ij} that was obtained for an RBM trained on real datasets does not belong to the classes for which the theory developed in this thesis is valid (training introduces involved correlations). However, it was shown recently that the Low-RAMP equations as studied in the present paper can be used efficiently for training of the RBM [KR98, GTK15].

The RBM with Gaussian hidden variables is related to the well known Hopfield model of associative memory [Hop82]. Therefore the properties of the bipartite Boltzmann measure (2.7) with a randomly quenched disorder Y_{ij} are in one-to-one correspondence with the properties of the Hopfield model. This relation in the view of the TAP equations was studied recently in [Méz16].

Examples with planted disorder

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

Gaussian estimation The most elementary examples of the planted case is when the measurement channel is Gaussian as in eq. (4.21), and the distributions P_X, P_U and P_V are also

Gaussian i.e.

$$P_X(x_i) = \frac{1}{\sqrt{\mathrm{Det}(2\pi\sigma_X)}} e^{-\frac{1}{2}(x_i-\mu_X)^\top \sigma_X^{-1}(x_i-\mu_X)}\,, \tag{4.55}$$

$$P_U(u_i) = \frac{1}{\sqrt{\mathrm{Det}(2\pi\sigma_U)}} e^{-\frac{1}{2}(u_i-\mu_U)^\top \sigma_U^{-1}(u_i-\mu_U)}\,, \tag{4.56}$$

$$P_V(v_i) = \frac{1}{\sqrt{\mathrm{Det}(2\pi\sigma_V)}} e^{-\frac{1}{2}(v_i-\mu_V)^\top \sigma_V^{-1}(v_i-\mu_V)}\,, \tag{4.57}$$

where $\mu_X, \mu_U, \mu_V \in \mathbb{R}^r$ are the means of the distributions, $\sigma_X, \sigma_U, \sigma_V \in \mathbb{R}^{r\times r}$ are the covariance matrices, $Y_{ij}, w_{ij} \in \mathbb{R}$ with w_{ij} being given by (4.5).

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

$$P_{\mathrm{out}}(Y_{ij}|w^0_{ij}) = e^{g(Y_{ij},w^0_{ij})}\,, \tag{4.58}$$

where $g(Y,w)$ is given by eq. (4.21), and

$$w^0_{ij} = {x^0}_i^\top x^0_j/\sqrt{N}\,, \quad \text{or} \quad w^0_{ij} = {u^0_i}^\top v^0_j/\sqrt{N}\,. \tag{4.59}$$

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

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

$$P_V(v_j) = \sum_{s=1}^{r} n_s \delta(v_j - e_s)\,, \tag{4.60}$$

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

The columns of the matrix U then represent centroids of each of the r clusters in the N-dimensional space. The distribution P_U can as an example take the Gaussian form (4.56) with the covariance σ_V being an identity and the mean μ_V being zero. The output channel is Gaussian as in (4.21). All together this means the Y_{ij} collects positions of M points in N dimensional space that are organized in r Gaussian clusters. The goal is to estimate the centers of the clusters and the cluster membership from Y.

Standard algorithms for data clustering include those based on the spectral decomposition of the matrix Y such as principal component analysis [HTFF05, Was13], or Loyd's k-means [Llo82]. Works on Gaussian mixtures that are methodologically closely related to the present paper include application of the replica method to the case of two clusters $r = 2$ in [WN94, BS94, BM94] or the AMP algorithm of [MT13]. Note that for two clusters with the two centers being symmetric around the origin, the resulting Boltzmann measure of the case with randomly quenched disorder is equivalent to the Hopfield model as treated e.g. in [Méz16].

Note also that there are interesting variants of the Gaussian mixture clustering such as subspace clustering [PHL04] where only some of the M directions are relevant for the clustering. This can be modeled by a prior on the vectors v_i that have a non-zero weight of v_i being the null vector.

The approach described in the present paper on the Bayes-optimal inference in the Gaussian mixture clustering problem has been used in a work of the authors with other collaborators in the work presented in [LDBB$^+$16].

Sparse PCA Sparse Principal Component Analysis (PCA) [JL04, ZHT06] is a dimensionality reduction technique where one seeks a low-rank representation of a data matrix with additional sparsity constraints on the obtained representation. The motivation is that the sparsity helps to interpret the resulting representation. Formulated within the above setting sparse PCA corresponds to the bipartite case (2.7). The variables U is considered unconstrained, as an example one often considers a Gaussian prior on u_i (4.56). The variables V are such that many of the matrix-elements are zero.

In the literature the sparse PCA problem was mainly considered in the rank-one case $r = 1$ for which a series of intriguing results was derived. The authors of [JL04] suggested an efficient algorithm called diagonal thresholding that solves the sparse PCA problem (i.e. estimates correctly the position and the value of the non-zero elements of U) whenever the number of data samples is $N > CK^2 \log M$ [AW08], where K is the number of non-zeros and C is some constant. More recent works show existence of efficient algorithm that only need $N > \hat{C}K^2$ samples [DM14b]. For very sparse systems, i.e. small K, this is a striking improvement over the conventional PCA that would need $O(M)$ samples. This is why sparsity linked together with PCA brought excitement into data processing methods. At the same time, this result is not as positive as it may seem, because by searching exhaustively over all positions of the non-zeros the correct support can be discovered with high probability with number of samples $N > \tilde{C}K \log M$.

Naively one might think that polynomial algorithms that need less thank $O(K^2)$ samples might exist and a considerable amount of work was devoted to their search without success. However, some works suggest that perhaps polynomial algorithm that solve the sparse PCA problems for number of samples $N < O(K^2)$ do not exist. Among the most remarkable one is the work [KNV$^+$15] showing that the SDP algorithm, that is otherwise considered rather powerful, fails in this regime. The work of [BR13] goes even further showing that if sparse PCA can be solved for $N < O(K^2)$ then also a problem known as the planted clique problem can be solved in a regime that is considered as algorithmically hard for already several decades.

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

The above works consider the scaling when $N \to \infty$ and K is fixed (or growing slower than $O(N)$). A regime natural to many applications is when $K = \rho N$ where $\rho = O(1)$. This regime was considered in [DM14a] where it was shown that for $\rho > \rho_0 \approx 0.04139$ an efficient algorithm that achieves the information theoretical performance exists. This immediately bring a question of what exactly happens for $\rho < \rho_0$ and how does the barrier described above appear for

$K \ll N$? This question was illuminated in a work by the present authors [LKZ15b] and will be developed further in this thesis.

We consider sparse PCA in the bipartite case, with Gaussian U (4.57) and sparse V

$$P_V(v_i) = \rho\delta(v_i - 1) + (1-\rho)\delta(v_i)\,, \tag{4.61}$$

as corresponds to the formulation of [JL04, ZHT06, AW08, BR13] and others. This probabilistic setting of sparse PCA was referred to as *spiked Wishart model* in [DM14a], notation that we will adopt in the present paper. This model is also equivalent to the one studied recently in [MV15] where the authors simply integrate over the Gaussian variables.

In [DM14a] the authors also considered a symmetric variant of the sparse PCA, and refer to it as the *spiked Wigner model*. The spiked Wigner model is closer to the planted clique problem, that can be formulated using (2.6) with X having many zero elements. In the present work we will consider several models for the prior distribution P_X. The *Bernoulli model* as in [DM14a] where

$$\textbf{Bernoulli model}: \qquad P_X(x_i) = \rho\delta(x_i - 1) + (1-\rho)\delta(x_i)\,. \tag{4.62}$$

The spiked Bernoulli model can also be interpreted as a problem of submatrix localization where a submatrix of size $\rho N \times \rho N$ of the matrix Y has a larger mean than a randomly chosen submatrix. The submatrix localization is also relevant in the bipartite case, where it has many potential applications. The most striking ones being in gene expression where large-mean submatrices of the matrix of gene expressions of different patients may correspond to groups of patients having the same type of disease [MO04, CC00].

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

$$\textbf{Rademacher}-\textbf{Bernoulli model}: \qquad P_X(x_i) = \frac{\rho}{2}\left[\delta(x_i - 1) + \delta(x_i + 1)\right] + (1-\rho)\delta(x_i)\,, \tag{4.63}$$

as well as the *spiked Gauss-Bernoulli*

$$\textbf{Gauss}-\textbf{Bernoulli model}: \qquad P_X(x_i) = \rho\mathcal{N}(x_i, 0, 1) + (1-\rho)\delta(x_i)\,, \tag{4.64}$$

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

So far we discussed the sparse PCA problem in the case of rank one, $r = 1$, but the case with larger rank is also interesting, especially in the view of the question of how does the algorithmic barrier depend on the rank. To investigate this question in [LKZ15b] we also considered the jointly-sparse PCA, where the whole r-dimensional lines of X are zero at once, the non-zeros are Gaussians of mean $\vec{0}$ and covariance being the identity. Mathematically, $x_i \in \mathbb{R}^r$ with

$$P_X(x_i) = \frac{\rho}{(2\pi)^{r/2}} e^{\frac{-x^\top x}{2}} + (1-\rho)\delta(x_i)\,. \tag{4.65}$$

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

$$P_X(x_i) = \prod_{1\le k\le r}\left[\frac{\rho}{\sqrt{2\pi}} e^{\frac{-x_{ik}^2}{2}} + (1-\rho)\delta(x_{ik})\right]\,. \tag{4.66}$$

Community detection Detection of communities in networks is often modeled by the stochastic block model (SBM) where pairs of nodes get connected with probability that depends on the indices of groups to which the two nodes depend. Community detection is a widely studied problem, see e.g. the review [For10]. Studies of statistical and computationally barriers in the SBM recently became very active in mathematics and statistics starting perhaps with a series of statistical physics conjectures about the existence of phase transitions in the problem [DKMZ11b, DKMZ11a]. The particular interest of those works is that they are set in the sparse regime where every node has $O(1)$ neighbors.

Also the dense regime where every node has $O(N)$ neighbors is theoretically interesting when the difference between probabilities to connect depends only weakly on the group membership. The relevant scaling is the same as in the Potts glass model studied in [GKS85]. In fact the dense community detection is exactly the planted version of this Potts glass model. In the setting of the present model (2.6) the community detection was already considered in [DAM16] for two symmetric groups, in [LKZ15a], and in [BDM$^+$16]. In the present paper we detail the results reported briefly in [LKZ15a] and in [BDM$^+$16]. We consider the case with a general number of equal sized groups, the symmetric case. And also a case with two groups of different sizes, but such that the average degree in each of the groups is the same.

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

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

$$C = \begin{pmatrix} p_{\text{in}} & p_{\text{out}} & \cdots & p_{\text{out}} \\ p_{\text{out}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & p_{\text{out}} \\ p_{\text{out}} & \cdots & p_{\text{out}} & p_{\text{in}} \end{pmatrix} = p_{\text{in}} I_r + p_{\text{out}}(J_r - I_r)\,, \tag{4.67}$$

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

$$p_{\text{out}} = O(1)\,, \quad p_{\text{in}} = O(1)\,, \tag{4.68}$$

$$|p_{\text{in}} - p_{\text{out}}| = \frac{\mu}{\sqrt{N}}\,, \quad \mu = O(1)\,, \tag{4.69}$$

so that the average degree in the graph is extensive. Note, however, that by rather generic correspondence between diluted and dense models, that has been made rigorous recently [DAM16, CLM16], the results derived in this case hold even for average degrees that diverge only very mildly with n. The goal is to infer the group index to which each node belongs purely from the adjacency matrix of the network (up to a permutation of the indices). This problem is transformed into the low-rank matrix factorization problem through the use of the following prior probability distribution

$$P_X(x_i) = \frac{1}{r} \sum_{s=1}^{r} \delta(x_i - e_s)\,. \tag{4.70}$$

where $e_s \in \mathbb{R}^r$ is the vector with 0 everywhere except a 1 at position s. Eq. (4.70) is just a special case of (4.60). The output channel that describes the process of creation of the graph is

$$P_{\text{out}}(Y_{ij} = 1|x_i^\top x_j/\sqrt{N}) = p_{\text{out}} + \frac{\mu x_i^\top x_j}{\sqrt{N}}\,, \tag{4.71}$$

$$P_{\text{out}}(Y_{ij} = 0|x_i^\top x_j/\sqrt{N}) = 1 - p_{\text{out}} - \frac{\mu x_i^\top x_j}{\sqrt{N}}\,. \tag{4.72}$$

Next to the conventional Hamiltonian (4.22) and the Gaussian noise (4.21), the SBM output (4.71-4.72) is a third example of an output channel that we consider in this thesis. It will be used to illustrate the simplicity that arises due to the channel universality, as also considered in [DAM16] and [LKZ15a]. Here, we obtain for the output matrices

$$S_{ij}(Y_{ij} = 1) = \frac{\mu}{p_{\text{out}}}, \quad S_{ij}(Y_{ij} = 0) = \frac{-\mu}{1 - p_{\text{out}}}\,, \tag{4.73}$$

$$R_{ij}(Y_{ij} = 1) = 0, \quad R_{ij}(Y_{ij} = 0) = 0\,. \tag{4.74}$$

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

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

$$C = \begin{pmatrix} p_{\text{out}} & p_{\text{out}} \\ p_{\text{out}} & p_{\text{out}} \end{pmatrix} + \frac{\mu}{\sqrt{N}} \begin{pmatrix} \frac{1-\rho}{\rho} & -1 \\ -1 & \frac{\rho}{1-\rho} \end{pmatrix}\,. \tag{4.75}$$

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

$$P_X(x) = \rho\delta\left(x - \sqrt{\frac{1-\rho}{\rho}}\right) + (1-\rho)\delta\left(x + \sqrt{\frac{\rho}{1-\rho}}\right)\,. \tag{4.76}$$

The values in C are chosen so that each community has an average degree of $p_{\text{out}}N$. The fact that in both of these cases each community has the same average degree means that one can not hope to just use the histogram of degrees to make the distinction between the communities. The output channel here is identical to the one given in (4.71-4.72)

A third example of community detection is locating one denser community in a dense network, as considered in [Mon15] (specifically the large degree limit considered in that paper). We note that thanks to the output channel universality (Sec. 4.1.2) this case is equivalent to the spiked Bernoulli model of symmetric sparse PCA.

As a side remark we note that the community detection setting as considered here is also relevant in the bipartite matrix factorization setting where it becomes the problem of biclustering [MO04, CC00]. The analysis developed in this thesis can be straightforwardly extended to the bipartite case.

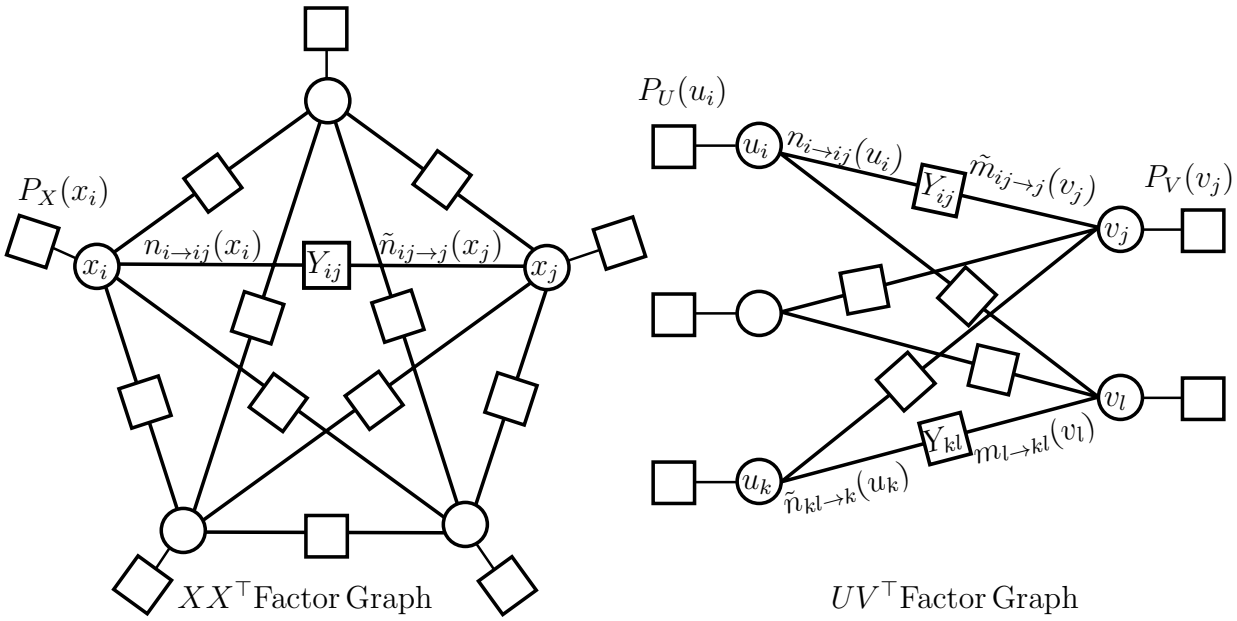


FIGURE 4.2 – This is the factor graph in the symmetric, $XX^\top$, and bipartite, $UV^\top$, matrix factorization. The squares are factors (or interaction terms), the circles represent variables. This factor graph allows us to introduce messages $n_{i\to ij}(x_i)$ and $\hat{n}_{ij\to j}(x_j)$ for the $XX^\top$ case. These are messages from variables to factors and from factors to variables. For the $UV^\top$ we introduce the four kinds of messages. $\hat{n}_{ij\to i}(u_i)$, $\hat{m}_{ij\to j}(v_j)$, $\hat{n}_{kl\to k}(u_k)$ and $m_{l\to kl}(v_l)$.

4.2 Low-rank approximate message passing

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

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

$$\tilde{n}_{ki\to i}(x_i) = \frac{1}{Z_{ki\to i}} \int \mathrm{d}x_k n_{k\to ki}(x_k) e^{g\left(Y_{ki},\frac{x_k^\top x_i}{\sqrt{N}}\right)}, \tag{4.77}$$

$$n_{i\to ij}(x_i) = \frac{P_X(x_i)}{Z_{i\to ij}} \prod_{1\le k\le N, k\neq i,j} \tilde{n}_{ki\to i}(x_i)\,. \tag{4.78}$$

The factor graph from which these messages are derived is given in Fig. 4.2. The most important assumption made in the BP equations is that the messages $\tilde{n}_{ki\to i}(x_i)$ are conditionally on the value x_i independent of each other thus allowing to write the product in eq. (4.78).

The message in (4.77) can be expanded as in (4.20) around $w = 0$ thanks to the $1/\sqrt{N}$ term. One gets

$$\tilde{n}_{ki\to i}(x_i) = \frac{e^{g(Y_{ki},0)}}{Z_{ki\to i}} \int \mathrm{d}x_k n_{k\to ki}(x_k) \left[1 + S_{ki}\frac{x_k^\top x_i}{\sqrt{N}} + \frac{x_i^\top x_k x_k^\top x_i}{2N} R_{ki} + O\left(\frac{1}{N^{3/2}}\right)\right], \tag{4.79}$$

where matrices S_{ij} and R_{ij} were defined in (4.18-4.19). One then defines the mean (r dimensional vector) and covariances matrix (of size $r \times r$) of the message $n_{k\to ki}$ as

$$\hat{x}_{k\to ki} = \int \mathrm{d}x_k n_{k\to ki}(x_k) x_k^\top \,, \tag{4.80}$$

$$\sigma_{x,k\to ki} = \int \mathrm{d}x_k n_{k\to ki}(x_k) x_k x_k^\top - \hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top \,. \tag{4.81}$$

The mean with respect to $n_{k\to ki}(x_k)$ is then taken in (4.79) and one gets

$$\tilde{n}_{ki\to i}(x_i) = \frac{1}{Z_{ki\to i}} \exp\left[g(Y_{ki},0) + S_{ki}\frac{\hat{x}_{k\to ki}^\top x_i}{\sqrt{N}} - \frac{x_i^\top \hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top x_i}{2N} S_{ki}^2 + \right.$$
$$\left. \frac{x_i^\top(\hat{x}_{k\to ik}\hat{x}_{k\to ik}^\top + \sigma_{x,k\to ik})x_i}{2N} R_{ki} + O\left(\frac{1}{N^{3/2}}\right)\right] \,. \tag{4.82}$$

Eqs. (4.78) and (4.82) are combined to get

$$n_{i\to ij}(x_i) = \frac{P_X(x_i)}{Z_{i\to ij}} \exp\left(B_{X,i\to ij}^\top x_i - \frac{x_i^\top A_{X,i\to ij} x_i}{2} \right) \,, \tag{4.83}$$

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

$$B_{X,i\to ij} = \frac{1}{\sqrt{N}} \sum_{1\le k\le N, k\neq j} S_{ki}\hat{x}_{k\to ki} \,, \tag{4.84}$$

$$A_{X,i\to ij} = \frac{1}{N} \sum_{1\le k\le N, k\neq j} \left[S_{ki}^2 \hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top - R_{ki}\left(\hat{x}_{k\to ki}\hat{x}_{k\to ki}^\top + \sigma_{x,k\to ki}\right)\right] \,. \tag{4.85}$$

The new mean and variance of the message (4.83) then needs to be computed. For this we define the function f_{in}^x

$$f_{\text{in}}^x(A,B) \equiv \frac{\partial}{\partial B} \log\left(\int \mathrm{d}x P_X(x) \exp\left(B^\top x - \frac{x^\top A x}{2} \right) \right) = \frac{\int \mathrm{d}x P_X(x) \exp\left(B^\top x - \frac{x^\top A x}{2} \right) x}{\int \mathrm{d}x P_X(x) \exp\left(B^\top x - \frac{x^\top A x}{2} \right)} \tag{4.86}$$

as the mean of the normalized density probability

$$\mathcal{W}(x,A,B) = \frac{1}{Z_x(A,B)} P_X(x) \exp\left(B^\top x - \frac{x^\top A x}{2} \right) \,. \tag{4.87}$$

The variance of the message (4.83) can be computed by writing the derivative of $f_{\text{in}}^x(A,B)$ with respect to B and getting

$$\frac{\partial f_{\text{in}}^x(A,B)}{\partial B} = \mathbb{E}_{\mathcal{W}(\mathcal{A},\mathcal{B})}(xx^\top) - f_{\text{in}}^x(A,B) f_{\text{in}}^{x\,\top}(A,B) \,. \tag{4.88}$$

This expression is the covariance matrix of distribution (4.87). Also it is worth nothing that

$$\frac{\partial \log(Z_x(A,B))}{\partial B} = f_{\text{in}}^x(A,B) \,. \tag{4.89}$$

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

$$B^t_{X,i\to ij} = \frac{1}{\sqrt{N}} \sum_{1\le k\le N, k\neq j} S_{ki}\hat{x}^t_{k\to ki}\,, \tag{4.90}$$

$$A^t_{X,i\to ij} = \frac{1}{N} \sum_{1\le k\le N, k\neq j} \left[S^2_{ki}\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} - R_{ki}(\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} + \sigma^t_{x,k\to ki})\right]\,, \tag{4.91}$$

$$\hat{x}^{t+1}_{i\to ij} = f^x_{\text{in}}(A^t_{i\to ij}, B^t_{i\to ij})\,, \tag{4.92}$$

$$\sigma^{t+1}_{x,i\to ij} = \frac{\partial f^x_{\text{in}}}{\partial B}(A^t_{i\to ij}, B^t_{i\to ij})\,. \tag{4.93}$$

4.2.1 Low-RAMP : TAPyfication and Onsager terms

The above relaxed BP algorithm uses $O(N^2)$ messages which can be memory demanding. But all the messages depend only weakly on the target node, and hence the algorithm can be reformulated using only $O(N)$ messages and collecting correcting terms called the Onsager terms in order to get estimators of the marginals that in the large size limit are equivalent to the previous ones. We call this formulation the TAPyfication, because of its analogy to the work of [TAP77]. We present the derivation in the case of symmetric low-rank matrix estimation. We notice that the variables $B_{i\to ij}$ and $A_{i\to ij}$ depend only weakly on the target node j. One can use this fact to close the equations on the marginals of the system. In order to do that we introduce the variables $B_{X,i}$ and $A_{X,i}$ as

$$B^t_{X,i} = \frac{1}{\sqrt{N}} \sum_{1\le k\le N} S_{ki}\hat{x}^t_{k\to ki}\,, \tag{4.94}$$

$$A^t_{X,i} = \frac{1}{N} \sum_{1\le k\le N} \left[S^2_{ki}\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} - R_{ki}\left(\hat{x}^t_{k\to ki}\hat{x}^{t,\top}_{k\to ki} + \sigma^t_{x,k\to ki}\right)\right]\,. \tag{4.95}$$

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

$$\hat{x}^{t+1}_i = f^x_{\text{in}}(A^t_{X,i}, B^t_{X,i})\,, \tag{4.96}$$

$$\sigma^{t+1}_{x,i} = \frac{\partial f^x_{\text{in}}}{\partial B}(A^t_{X,i}, B^t_{X,i})\,. \tag{4.97}$$

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

From the definition of the parameters A and B, we have that $\forall j$, $B^t_{X,i} - B^t_{X,i\to ij} = \frac{S_{ij}}{\sqrt{N}}\hat{x}^t_{j\to ij} = O\left(\frac{1}{\sqrt{N}}\right)$. $A_{X,i} - A_{X,i\to ij} = O\left(\frac{1}{N}\right)$. One deduces from (4.92) and (4.93), (4.96) and (4.97), and the Taylor expansion, that in the leading order the difference between the messages and the

node estimators is

$$\begin{aligned}\hat{x}^t_{k\to ki} - \hat{x}^t_k &= f(A^{t-1}_{X,k\to ki}, B^{t-1}_{X,k\to ki}) - f(A^{t-1}_{X,k}, B^{t-1}_{X,k}) \\ &= -\frac{S_{ki}}{\sqrt{N}}\sigma^t_{x,k}\hat{x}^{t-1}_{i\to ki} + O\left(\frac{1}{N}\right) = -\frac{S_{ki}}{\sqrt{N}}\sigma^t_{x,k}\hat{x}^{t-1}_i + O\left(\frac{1}{N}\right) . \end{aligned} \tag{4.98}$$

By plugging (4.98) into (4.94) one gets in the leading order

$$B^t_{X,i} = \frac{1}{\sqrt{N}}\sum_{k=1}^N S_{ki}\hat{x}^t_k - \hat{x}^{t-1}_i \frac{1}{N}\sum_{k=1}^N S^2_{ki}\sigma^t_{x,k} , \tag{4.99}$$

$$A^t_{X,i} = \frac{1}{N}\sum_{k=1}^N \left[S^2_{ki}\hat{x}^t_k\hat{x}^{t,\top}_k - R_{ki}\left(\hat{x}^t_k\hat{x}^{t,\top}_k + \sigma^t_{x,k}\right)\right] . \tag{4.100}$$

These two equations, together with eqs. (4.96-4.97) give us the low-rank approximate message passing algorithm (Low-RAMP) with $O(N)$ variables. The second term in the equation (4.99) is called the Onsager reaction term. Notice the iteration index $t-1$ which is non-intuitive on the first sight and was often misplaced until recently, see e.g. discussion in [ZK16]. Note also that there is no Onsager reaction term in the expression for the covariance $A_{X,i}$, that is because the individual terms in a sum in A are of order $O(1/N)$ and not $O(1/\sqrt{N})$. This is a common pattern in AMP-type algorithm, the Onsager terms appear only in the terms that estimate means, not in the variances. The Low-RAMP algorithm is related in spirit to the AMP algorithm for linear sparse estimation [DMM09], for instance the function f_{in} is the same thresholding function as in the linear-estimation AMP. However, the linear estimation AMP is more involved for a generic output channel and the structure of the two algorithms are quite different, stemming from the fact that in the present case all interactions are pairwise whereas for the linear estimation each interaction involves all the variables, giving rise to non-trivial terms that do not appear in Low-RAMP. The following pseudocode summarizes our implementation of the Low-RAMP algorithm :

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

Initialize each $\hat{x}_i \in \mathbb{R}^{r\times 1}$ vector using $\hat{x}^{\text{init}}$: $\forall i$, $\hat{x}_i \leftarrow \hat{x}^{\text{init}}_i$.
Initialize each $\hat{x}^{\text{old}}_i \in \mathbb{R}^r$ vector to zero : $\forall i$, $\hat{x}^{\text{old}}_i \leftarrow 0$.
Initialize each vector $B_{X,i} \in \mathbb{R}^{r\times 1}$ to zero, $B_{X,i} \leftarrow 0$.
Initialize each N $r\times 1$ vector $B^{\text{old}}_{X,i}$ to zero, $\forall i$, $B^{\text{old}}_{X,i} \leftarrow 0$.
Initialize to zero each N matrix, $r\times r$ matrix $A_{X,i}$ with, $A_{X,i} \leftarrow 0$.
Initialize to zero each N matrix, $r\times r$ matrix $A^{\text{old}}_{X,i}$ with, $A^{\text{old}}_{X,i} \leftarrow 0$.
Initialize to zero each N matrix, $r\times r$ matrix $\sigma_{X,i}$ with, $\sigma_{X,i} \leftarrow 0$.
while conv $*\,\lambda > \epsilon_{\text{criterium}}$ and $t < t_{\max}$:
    do $t \leftarrow t+1$;
        $\forall i$, $B^{\text{new}}_{X,i} \leftarrow$ Update with equation (4.99) or (4.173).
        $\forall i$, $A^{\text{new}}_{X,i} \leftarrow$ Update with equation (4.100) or (4.174).
        $\forall i$, $B_{X,i} \leftarrow \lambda B^{\text{new}}_{X,i} + (1-\lambda)B^{\text{old}}_{X,i}$,
        $\forall i$, $A_{X,i} \leftarrow \lambda A^{\text{new}}_{X,i} + (1-\lambda)A^{\text{old}}_{X,i}$,
        $\forall i$, $\hat{x}^{\text{old}}_i \leftarrow \hat{x}_i$, $\hat{x}_i \leftarrow f^x_{\text{in}}(A_X, B_{X,i})$,
        $\forall i$, $\sigma_{X,i} \leftarrow \frac{\partial f^x_{\text{in}}}{\partial B}(A_X, B_{X,i})$,
conv $\leftarrow \frac{1}{N}\sum \|\hat{x}_i - \hat{x}^{\text{old}}_i\|$.
return signal components $\mathbf{x}$.

The canonical initialization we use is

$$\forall i \in [1; N],\ \ \hat{x}_i^{\text{init}} \leftarrow 10^{-3}\mathcal{N}(0, I_r)\,. \tag{4.101}$$

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

$$\forall i \in [1; N],\ \ \hat{x}_i^{\text{init}} \leftarrow x_i^0\,, \tag{4.102}$$

where x_i^0 is the planted (ground truth) configuration. In the above pseudocode the damping factor λ is chosen constant for the whole duration of the algorithm. It is possible to choose λ dynamically in order to improve the convergence. Using the fact that the Low-RAMP algorithm finds a stationary fixed point of the Bethe free energy given in (3.3) one can choose the damping factor λ so that at each step so that the Bethe-free-energy increases, this is described in [VSR+15]. Another way to choose λ is by ensuring that $\sum \|\hat{x}^t - \hat{x}^{t+1}\|$ does not oscillate too much. If one sees too much oscillations one increases the damping and decreases it otherwise. Further way to improve convergence is related to randomization of the update scheme as argued for the related compressed sensing problem in [CZK14].

4.2.2 Low-RAMP and TAP equations for the graphon case

One might be interested in case where the signal w might not be given by $w_{ij} \neq x_i x_j$ but could be given by a general function $f(x_i, x_j)$. The motivation to consider such a case might come from the study of graphons. Graphons are The main difference with the $w = x_i^\top x_j$ case lies in the fact the variables $B \in \mathbb{R}^{r\times 1}$ and $A^{r\times r}$ are now replaced with just one functions $B(x_i)$. Out of convenience of notation we will define the notations

$$(C * \lambda)(x) = \int \mathrm{d}\hat{x} C(x, \hat{x})\lambda(\hat{x}) \tag{4.103}$$

$$(\lambda * C)(x) = \int \mathrm{d}\hat{x} \lambda(\hat{x}) C(\hat{x}, x) \tag{4.104}$$

$$(C_1 * C_2)(x_1, x_2) = \int \mathrm{d}y C_1(x, y) C_2(y, x_2)\lambda(\hat{x}) \tag{4.105}$$

$$(C^\top)(x_1, x_2) = C(x_2, x_1) \tag{4.106}$$

the act of multiplying the multiplicative kernel $C(x_1, x_2)$ by a function $\lambda(x)$ on the left/right and the act of multiplying two multiplicative kernel. This can be thought of as a generalization of matrix multiplication where the indices x of the matrix rather than living in $\mathbb{N}$ can live in any set.

The beginning of the derivation is the same as in the multiplicative case.

$$\tilde{n}_{ki\to i}(x_i) = \frac{1}{Z_{ki\to i}} \int \mathrm{d}x_k \hat{P}_{X,k\to ki}(x_k) e^{g\left(Y_{ki}, \frac{f(x_i,x_j)}{\sqrt{N}}\right)}\,, \tag{4.107}$$

$$\hat{P}_{X,i\to ij}(x_i) = \frac{P_X(x_i)}{Z_{i\to ij}} \prod_{1\leq k\leq N, k\neq i,j} \tilde{n}_{ki\to i}(x_i)\,. \tag{4.108}$$

Here we have replaced the notation of $n_{i\to ij}(x_i)$ with the notation $\hat{P}_{X,i\to ij}(x_i)$ the reason being that in that version of the AMP algorithm one has to keep track of the whole posterior probability of the message. And $\hat{P}_{X,i\to ij}$ is going to "replace" the role of the variable $\hat{x}_{i\to ij}$ in the AMP equations.

The message in (4.77) can be expanded as in (4.20) around $w = 0$ thanks to the $1/\sqrt{N}$ term. One gets

$$\tilde{n}_{ki\to i}(x_i) = \frac{e^{g(Y_{ki},0)}}{Z_{ki\to i}} \int \mathrm{d}x_k \hat{P}_{X,k\to ki}(x_k) \left[1 + S_{ki}\frac{f(x_i,x_k)}{\sqrt{N}} + \frac{f(x_i,x_k)^2}{2N}R_{ki} + O\left(\frac{1}{N^{3/2}}\right)\right], \tag{4.109}$$

The mean with respect to $\hat{P}_{X,k\to ki}(x_k)$ is then taken in (4.109) and one gets

$$\tilde{n}_{ki\to i}(x_i) = \frac{1}{Z_{ki\to i}} \exp\left[g(Y_{ki},0) + D_{ki\to i}(x_i)\right]. \tag{4.110}$$

Where

$$D_{ki\to i}(x_i) = \frac{1}{\sqrt{N}} S_{ki} \int \mathrm{d}x_k \hat{P}_{X,k\to ki}(x_k) f(x_i,x_k) + \frac{R_{ki}}{2N} \int \mathrm{d}x_k \hat{P}_{X,k\to ki}(x_k) f(x_i,x_k)^2 \tag{4.111}$$

$$- \frac{1}{2N} S_{ki}^2 \left[\int \mathrm{d}x_k \hat{P}_{X,k\to ki}(x_k) f(x_i,x_k)\right]^2 \tag{4.112}$$

$$D_{ki\to i}(x_i) = \frac{1}{\sqrt{N}} S_{ki} (f * \hat{P}_{X,k\to ki})(x_i) + \frac{R_{ki}}{2N} ([f^2] * \hat{P}_{X,k\to ki})(x_i) - \frac{1}{2N} S_{ki}^2 \left[(f * \hat{P}_{X,k\to ki})(x_i)\right]^2 \tag{4.113}$$

Eqs. (4.108) and (4.110) are combined to get

$$n_{i\to ij}(x_i) = \frac{P_X(x_i)}{Z_{i\to ij}} \exp\left(B_{X,i\to ij}(x_j)\right), \tag{4.114}$$

Where

$$B_{X,i\to ij}(x_i) = \sum_{1\le k\le N, k\neq j} D_{ki\to i}(x_i). \tag{4.115}$$

The new posterior probability $\hat{P}_{X,i\to ij}(x_i)$ can then be computed as

$$\hat{P}_{X,i\to ij}(x_i) = \frac{1}{Z_x(B_{X,i\to ij})} P_X(x) e^{B_{X,i\to ij}(x)}. \tag{4.116}$$

Where $\mathcal{Z}(B)$ here is defined as

$$\mathcal{Z}_x(B) = \int \mathrm{d}x P_X(x) e^{B(x)} \tag{4.117}$$

TAPyfication

To compute the TAP form of these equations we will need two things. First we need to decide of an update scheme to reach a fixed point of the equations (4.115,4.116) We will decide on the

following Parallel update

$$B^t_{X,i\to ij}(x_i) = \sum_{1\le k\le N, k\neq j} D^t_{ki\to i}(x_i)\,. \tag{4.118}$$

$$\hat{P}^{t+1}_{X,i\to ij}(x_i) = \frac{1}{Z_x(B^t_{X,i\to ij})} P_X(x) e^{B^t_{X,i\to ij}(x)}\,. \tag{4.119}$$

We will also need to compute how $\hat{P}_X(x)$ (4.116) evolve when we perturb the function B a little. To do this let us introduce a perturbation field $\lambda(x)$ and compute to first order in ϵ the following quantity

$$= \frac{1}{Z_x(B+\epsilon\lambda)} P_X(x) e^{B(x)+\epsilon\lambda(x_i)} - \frac{1}{Z_x(B)} P_X(x) e^{B(x)} \tag{4.120}$$

$$= \frac{1}{Z_x(B)(1+\epsilon\overline{\lambda(x)})} P_X(x) e^{B(x)}(1+\epsilon\lambda(x)) - \frac{1}{Z_x(B)} P_X(x) e^{B(x)} + o(\epsilon) \tag{4.121}$$

$$= \frac{1}{Z_x(B)(1+\epsilon\overline{\lambda(x)})} P_X(x) e^{B(x)}(1+\epsilon\lambda(x)) - \frac{1}{Z_x(B)} P_X(x) e^{B(x)} + o(\epsilon) \tag{4.122}$$

$$= \epsilon \hat{P}_X(x)(\lambda(x) - \overline{\lambda}) + o(\epsilon) \tag{4.123}$$

Where $\overline{\lambda}$ is the mean value of $\lambda(x)$ with x taken from density probability (4.116). The $\overline{\lambda}$ term is just here to ensure that the perturbed density probability sums to 1. Therefore when adding a small field lambda to the field $B(x)$. The marginal probability is modified to first order by

$$= \hat{P}_X(x)(\lambda(x) - \overline{\lambda}) \tag{4.124}$$

$$= \int \mathrm{d}\hat{x} \left[\hat{P}_X \delta(x-\hat{x}) - \hat{P}_X(x)\hat{P}_X(\hat{x})\right] \lambda(\hat{x}) \tag{4.125}$$

$$= \int \mathrm{d}\hat{x} C(x,\hat{x})\lambda(\hat{x}) \tag{4.126}$$

$$= (C * \lambda)(x) \tag{4.127}$$

where

$$C(x_1,x_2) = \left[\hat{P}_X(x_1)\delta(x_1-x_2) - \hat{P}_X(x_1)\hat{P}_X(x_2)\right]\,. \tag{4.128}$$

Here $C(x_1,x_2)$ is an integral kernel that is the sum of two terms, a dirac term and a continuous term outside of the diagonal. $C(x_1,x_2)$ can be thought of as a covariance matrix since

$$C(x_1,x_2) = \int \mathrm{d}x \delta(x-x_1)\delta(x-x_2)\hat{P}_X(x) - \left(\int \mathrm{d}x \delta(x-x_1)\hat{P}_X(x)\right)\left(\int \mathrm{d}x \delta(x-x_2)\hat{P}_X(x)\right) \tag{4.129}$$

It is an linear operator that measures the linear response $\hat{P}_X(x)$ with respect to the field B. Therefore we will note $C^t_{X,i}$ the linear response Kernel of the posterior probability of variables i at time t. With this result in our pocket we are now ready to compute the TAP form the AMP

equations for this problem. Let us define the B_i^t

$$B_i^t(x_i) = \sum_{1\leq k\leq N} D_{ki\to i}^t(x_i)\,, \tag{4.130}$$

$$\hat{P}_{X,i}^{t+1}(x_i) = \frac{1}{Z_x(B_i)} e^{B_i^t(x_i)} \tag{4.131}$$

Now as before let us notice that to leading order the difference between $P_{X,i\to ij}$ and $P_{X,i}$ is given by

$$\hat{P}_{X,i\to ij}^{t+1}(x_i) - \hat{P}_{X,i}^{t+1}(x_i) = -\frac{S_{ij}}{\sqrt{N}}\left[C_{X,i}^{t+1} * f * \hat{P}_{X,i}^t\right](x_i) + O\left(\frac{1}{N}\right) \tag{4.132}$$

Therefore by combining (4.132) and (4.131) one obtain the TAP equations

$$\begin{aligned} B_i^t(x_i) = \sum_{1\leq i\leq N}\Bigg[&\frac{1}{\sqrt{N}} S_{ki}(f * \hat{P}_{X,k}^t)(x_i) - \frac{1}{N} S_{ki}^2 (f * C_{X,i}^t * f * \hat{P}_{X,k}^{t-1})(x_i) \\ &+ \frac{R_{ki}}{2N}([f^2] * \hat{P}_{X,k}^t)(x_i) - \frac{1}{2N} S_{ki}^2 \left[(f * \hat{P}_{X,k}^t)(x_i)\right]^2\Bigg] \end{aligned} \tag{4.133}$$

$$\hat{P}_{X,i}^{t+1}(x_i) = \frac{1}{Z_x(B_{X,i})} e^{B_{X,i}^t(x_i)} \tag{4.134}$$

$$C_{X,i}^t(x_1, x_2) = \delta(x_1 - x_2)\hat{P}_{X,i}^t(x_1) - \hat{P}_{X,i}^t(x_1)\hat{P}_{X,i}^t(x_2) \tag{4.135}$$

Low rank approximation of $\mathbf{f(x_1, x_2)}$

These TAP equations we have written are equations that require us to store density probabilities and to compute integrals therefore, when implementing this algorithm we will unavoidably run into the problem of approximating/representing the functions $P_{X,i}^t$ as well as estimating the multiplication against a multiplicative kernel. Suppose that the x_i variables live in $[0;1]^{10}$ a naive approach to store $\hat{P}_{X,i}$ would be to cut $[0;1]^{10}$ into little hypercubes of size ϵ. Already for $\epsilon = 0.1$ this means that we will need to store 10^{10} numbers to represent one density probability $\hat{P}_{X,i}$, which is way to much if we want to actually use this algorithm.

In this subsection we will focus our attention on an approximation method that will help us understand what is the rank of the problem. In the previous case where $f(x_i, x_j) = x_i^\top x_j$ a great number of simplifications could be made so that at each step of the algorithm at worst $O(r^2 N)$ numbers needed to be stored to represent the state of the algorithm. We could make such a simplification because $x_i^\top x_j$ could be written as

$$f(x_1, x_2) = x_i^\top x_j = \sum_{k=1\cdots r} x_{i,k} x_{j,k} \tag{4.136}$$

If the function $f(x_1, x_2)$ could be written in this form

$$f(x_1, x_2) = \sum_{k=1\cdots r} f_k(x_1)\lambda_k f_k(x_2)\,, \tag{4.137}$$

Where

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_r| > 0 \tag{4.138}$$

$$\forall k, \int \mathrm{d}x f_k^2(x) = 1 \tag{4.139}$$

then we could do the same simplification and rewrite the AMP algorithm by just storing for each variables x_i the following mean at each time t.

$$\overline{f_k(x_i)} = \int \mathrm{d}x_i \hat{P}_{X,i}^t(x_i) f_k(x_i) \tag{4.140}$$

$$\overline{f_k(x_i) f_{k'}(x_i)} = \int \mathrm{d}x_i \hat{P}_{X,i}^t(x_i) f_k(x_i) f_{k'}(x_i) \tag{4.141}$$

By introducing (4.137) into (4.133,4.134). It is easy to see that one can close the equations on a set of variables. Of course the function $f(x_1, x_2)$ might not be low-rank and we might need an infinity of term in (4.137) to accurately describe $f(x_1, x_2)$. The rank of the problem will then be the number of terms in (4.137). Cutting the sum to a finite number of term would provide an approximation of $f(x_1, x_2)$.

4.2.3 Summary of Low-RAMP for the bipartite low-rank estimation

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

$$B_{U,i\to ij}^t = \frac{1}{\sqrt{N}} \sum_{1\leq l\leq M, l\neq j} S_{il} \hat{v}_{l\to il}^t \,, \tag{4.142}$$

$$A_{U,i\to ij}^t = \frac{1}{N} \sum_{1\leq l\leq M, l\neq j} \left[S_{il}^2 \hat{v}_{l\to il}^t \hat{v}_{l\to il}^{t,\top} - R_{il} \left(\hat{v}_{l\to il}^t \hat{v}_{l\to il}^{t,\top} + \sigma_{v,l\to il}^t \right) \right] , \tag{4.143}$$

$$\hat{u}_{i\to ij}^t = f_{\text{in}}^u(A_{U,i\to ij}^t, B_{U,i\to ij}^t) \,, \tag{4.144}$$

$$\sigma_{u,i\to ij}^t = \frac{\partial f_{\text{in}}^u}{\partial B}(A_{U,i\to ij}^t, B_{U,i\to ij}^t) \,, \tag{4.145}$$

$$B_{V,j\to ij}^t = \frac{1}{\sqrt{N}} \sum_{1\leq k\leq N, k\neq j} S_{kj} \hat{u}_{k\to kj}^t \,, \tag{4.146}$$

$$A_{V,j\to ij}^t = \frac{1}{N} \sum_{1\leq k\leq N, k\neq i} \left[S_{kj}^2 \hat{u}_{k\to kj}^t \hat{u}_{k\to kj}^{t,\top} - R_{kj} \left(\hat{u}_{k\to kj}^t \hat{u}_{k\to kj}^{t,\top} + \sigma_{u,k\to kj}^t \right) \right] , \tag{4.147}$$

$$\hat{v}_{j\to ij}^{t+1} = f_{\text{in}}^v(A_{V,j\to ij}^t, B_{V,j\to ij}^t) \,, \tag{4.148}$$

$$\sigma_{v,j\to ij}^{t+1} = \frac{\partial f_{\text{in}}^v}{\partial B}(A_{V,j\to ij}^t, B_{V,j\to ij}^t) \,. \tag{4.149}$$

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

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

tion read

$$B_{U,i}^t = \frac{1}{\sqrt{N}} \sum_{l=1}^{M} S_{il} \hat{v}_l^t - \left(\frac{1}{N} \sum_{l=1}^{M} S_{il}^2 \sigma_{v,l}^t \right) \hat{u}_i^{t-1} , \tag{4.150}$$

$$A_{U,i}^t = \frac{1}{N} \sum_{l=1}^{M} \left[S_{il}^2 \hat{v}_l^t \hat{v}_l^{t,\top} - R_{il} \left(\hat{v}_l^t \hat{v}_l^{t,\top} + \sigma_{v,l}^t \right) \right] , \tag{4.151}$$

$$\hat{u}_i^t = f_{\text{in}}^u (A_{U,i}^t, B_{U,i}^t) , \tag{4.152}$$

$$\sigma_{u,i}^t = \left(\frac{\partial f_{\text{in}}^u}{\partial B} \right) (A_{U,i}^t, B_{U,i}^t) , \tag{4.153}$$

$$B_{V,j}^t = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} S_{kj} \hat{u}_k^t - \left(\frac{1}{N} \sum_{k=1}^{N} S_{kj}^2 \sigma_{u,k}^t \right) \hat{v}_j^t , \tag{4.154}$$

$$A_{V,j}^t = \frac{1}{N} \sum_{k=1}^{N} \left[S_{kj}^2 \hat{u}_k^t \hat{u}_k^{t,\top} - R_{kj} \left(\hat{u}_k^t \hat{u}_k^{t,\top} + \sigma_{u,k}^t \right) \right] , \tag{4.155}$$

$$\hat{v}_j^{t+1} = f_{\text{in}}^v (A_{V,j}^t, B_{V,j}^t) , \tag{4.156}$$

$$\sigma_{v,j}^{t+1} = \left(\frac{\partial f_{\text{in}}^v}{\partial B} \right) (A_{V,j}^t, B_{V,j}^t) . \tag{4.157}$$

Due to independence between messages assumed in the BP algorithm we can simplify the Low-RAMP algorithm introducing (4.175) and (4.176) also for the bipartite case. As well as take advantage of the Bayes optimality and the Nishimori conditions or of the particular form of the conventional Hamiltonian.

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

Initialize each N, $r \times 1$ vector $\hat{u}_i$ using $\hat{u}^{\text{init}}$, $\forall i$, $\hat{u}_i \leftarrow \hat{u}^{\text{init}}_i$.
Initialize each M, $r \times 1$ vector $\hat{v}_j$ using $\hat{v}^{\text{init}}$, $\forall j$, $\hat{v}_j \leftarrow \hat{v}^{\text{init}}_j$.
Initialize each M, $r \times 1$ vector $\hat{u}^{\text{old}}_i$ to 0, $\forall i$, $\hat{u}^{\text{old}}_i \leftarrow 0$.
Initialize each M, $r \times 1$ vector $\hat{v}^{\text{old}}_j$ to 0, $\forall j$, $\hat{v}^{\text{old}}_j \leftarrow 0$.
Initialize to zero each N, $r \times 1$ vector $B_{U,i}$ to zero, $\forall i$, $B_{U,i} \leftarrow 0$.
Initialize to zero each M, $r \times 1$ vector $B_{V,j}$ to zero, $\forall j$, $B_{V,j} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A_{U,i}$ with, $\forall j$, $A_{U,i} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A^{\text{old}}_{U,j}$ with, $\forall i$, $A^{\text{old}}_{U,i} \leftarrow 0$.
Initialize to zero each M matrix, $r \times r$ matrix $A_{V,i}$ with, $\forall j$, $A_{V,j} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $A^{\text{old}}_{V,j}$ with, $\forall j$, $A^{\text{old}}_{V,j} \leftarrow 0$.
Initialize to zero each N matrix, $r \times r$ matrix $\sigma_{U,i}$ with, $\sigma_{U,i} \leftarrow 0$.
Initialize to zero each M matrix, $r \times r$ matrix $\sigma_{V,j}$ with, $\sigma_{V,j} \leftarrow 0$.
while conv $* \lambda > \epsilon_{\text{criterium}}$ and $t < t_{\max}$:
    do $t \leftarrow t + 1$;
        Update variables U
        $\forall i,\ B^{\text{new}}_{U,i} \leftarrow$ Update with equation (4.150).
        $\forall i,\ A^{\text{new}}_{U,i} \leftarrow$ Update with equation (4.151).
        $\forall i,\ B_{U,i} \leftarrow \lambda B^{\text{new}}_{U,i} + (1-\lambda) B^{\text{old}}_{U,i}$,
        $\forall i,\ A_{U,i} \leftarrow \lambda A^{\text{new}}_{U,i} + (1-\lambda) A^{\text{old}}_{U,i}$,
        $\forall i,\ \hat{u}^{\text{old}}_i \leftarrow \hat{u}_i,\ \hat{u}_i \leftarrow f^u_{\text{in}}(A_{U,i}, B_{U,i})$,
        $\forall i,\ \sigma_{U,i} \leftarrow \frac{\partial f^u_{\text{in}}}{\partial B}(A_{U,i}, B_{U,i})$.
        Update variables V
        $\forall j,\ sB^{\text{new}}_{V,j} \leftarrow$ Update with equation (4.154).
        $\forall j,\ A^{\text{new}}_{V,j} \leftarrow$ Update with equation (4.155).
        $\forall j,\ B_{V,j} \leftarrow \lambda B^{\text{new}}_{V,j} + (1-\lambda) B^{\text{old}}_{V,j}$,
        $\forall j,\ A_{V,j} \leftarrow \lambda A^{\text{new}}_{V,j} + (1-\lambda) A^{\text{old}}_{V,j}$,
        $\forall j,\ \hat{v}^{\text{old}}_j \leftarrow \hat{u}_i,\ \hat{u}_i \leftarrow f^v_{\text{in}}(A_{V,j}, B_{V,j})$,
        $\forall j,\ \sigma_{V,j} \leftarrow \frac{\partial f^v_{\text{in}}}{\partial B}(A_{V,j}, B_{V,j})$.
        Compute distance with previous iteration
        conv $\leftarrow \frac{1}{N}\sum \|\hat{u}_i - \hat{u}^{\text{old}}_i\| + \frac{1}{M}\sum \|\hat{v}_j - \hat{v}^{\text{old}}_j\|$.
return signal components **x**

The initialization and damping factor λ are chosen similarly as for the symmetric case as discussed in section 4.2.1.

4.2.4 Low-RAMP : Tensor factorization

This AMP algorithm can also be computed in the tensor case for the density probability (2.10).

Once again the variables will be the vectors x_i. The factor will designated by their indices $i_1 \cdots i_p$ and will be the

$$\exp\left[g\left(Y_{i_1\cdots i_p}, \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{k=1\cdots r} x_{i_1 k} x_{i_2 k} \cdots x_{i_p k}\right)\right] \tag{4.158}$$

The factor $Y_{i_1\cdots i_p}$ are defined up to a permutation of the p indices.

We once again define the messages from factors to variables and from factors to variables.

$$\tilde{n}_{ik_2\cdots k_p\to i}(x_i) = \frac{1}{Z_{ik_2\cdots k_p\to i}} \int e^{g\left(Y_{ik_2\cdots k_p}, \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum\limits_{l=1\cdots r} \left(x_i \circ x_{k_2} \circ \cdots \circ x_{k_p}\right)_l\right)} \prod_{l=2\cdots p} \mathrm{d}x_l n_{l\to ik_2\cdots k_p}(x_l)\,, \tag{4.159}$$

$$n_{i\to ii_2\cdots i_p}(x_i) = \frac{P_X(x_i)}{Z_{i\to ii_2\cdots i_p}} \prod_{\substack{1\le k_2<k_3<\cdots<k_p\le N\\ \forall l, i_l\neq i\\ (k_2,\cdots,k_p)\neq(i_2,\cdots,i_p)}} \tilde{n}_{ik_2\cdots k_p\to i}(x_i)\,. \tag{4.160}$$

We can once again expand equation 4.159 to order 2 in order to express update equations as a function of only the moments of the message. One gets the AMP equations for this system.

$$B^t_{i\to ii_2\cdots i_p} = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{\substack{1\le k_2<k_3<\cdots<k_p\le N\\ \forall l, k_l\neq i\\ (k_2,\cdots,k_p)\neq(i_2,\cdots,i_p)}} S_{i,k_2,k_3\cdots k_p} \hat{x}^t_{k_2\to ik_2\cdots k_p} \circ \hat{x}^t_{k_3\to ik_2\cdots k_p} \circ \cdots \circ \hat{x}^t_{k_p\to ik_2\cdots k_p} \tag{4.161}$$

$$\begin{aligned} A^t_{i\to ii_2\cdots i_p} &= \frac{1}{N^{p-1}} \sum_{\substack{1\le k_2<k_3<\cdots<k_p\le N\\ \forall l, k_l\neq i\\ (k_2,\cdots,k_p)\neq(i_2,\cdots,i_p)}} S^2_{i,k_2,k_3\cdots k_p} \left(\hat{x}^t_{i_2\to ik_2\cdots k_p} \hat{x}^{t\top}_{i_2\to ik_2\cdots k_p}\right) \circ \cdots \circ \left(\hat{x}^t_{i_p\to ik_2\cdots k_p} \hat{x}^{t\top}_{i_p\to ik_2\cdots k_p}\right) \\ &\quad - \frac{1}{N^{p-1}} \sum_{\substack{1\le k_2<k_3<\cdots<k_p\le N\\ \forall l, k_l\neq i\\ (k_2,\cdots,k_p)\neq(i_2,\cdots,i_p)}} R_{i,k_2,k_3\cdots k_p} \left(\begin{smallmatrix}\hat{x}_{i_2\to ik_2\cdots k_p} \hat{x}^\top_{i_2\to ik_2\cdots k_p}\\ +\sigma_{i_2\to ik_2\cdots k_p}\end{smallmatrix}\right) \circ \cdots \circ \left(\begin{smallmatrix}\hat{x}_{i_p\to ik_2\cdots k_p} \hat{x}^\top_{i_p\to ik_2\cdots k_p}\\ +\sigma_{i_p\to ik_2\cdots k_p}\end{smallmatrix}\right)\,, \end{aligned} \tag{4.162}$$

$$\hat{x}^{t+1}_{i\to ii_2\cdots i_p} = f_{\text{in}}(A^t_{i\to ii_2\cdots i_p}, B^t_{i\to ii_2\cdots i_p})\,, \tag{4.163}$$

$$\sigma^{t+1}_{i\to ii_2\cdots i_p} = \partial_B f_{\text{in}}(A^t_{i\to ii_2\cdots i_p}, B^t_{i\to ii_2\cdots i_p})\,, \tag{4.164}$$

The $\circ$ notation is the Hadamard product which multiplies vector and matrices elements by elements. Once again these equations can be expressed and closed on the marginals of the equations.

One gets

$$B_i^t = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{\substack{1\le k_2<k_3<\cdots<k_p\le N \\ \forall l,k_l\neq i}} S_{i,k_2,k_3\cdots k_p} \hat{x}^t_{k_2} \circ \hat{x}^t_{k_3} \circ \cdots \circ \hat{x}^t_{k_p}$$
$$- \frac{(p-1)!}{N^{p-1}} \sum_{\substack{1\le k_2\le N \\ 1\le k_3<\cdots<k_p\le N \\ \forall l,k_l\neq i}} S^2_{i,k_2,k_3\cdots k_p} \left[\sigma^t_{k_2} \circ \left(\hat{x}^t_{k_3}\hat{x}^{t-1^\top}_{k_3}\right) \circ \cdots \circ \left(\hat{x}^t_{k_p}\hat{x}^{t-1^\top}_{k_p}\right)\right] \hat{x}^{t-1}_i \tag{4.165}$$

$$A_i^t = \frac{(p-1)!}{N^{p-1}} \sum_{\substack{1\le k_2<k_3<\cdots<k_p\le N \\ \forall l,i_l\neq i}} S^2_{i,k_2,k_3\cdots k_p} \left(\hat{x}^t_{k_2}\hat{x}^{t^\top}_{k_2}\right) \circ \cdots \circ \left(\hat{x}^t_{k_p}\hat{x}^{t^\top}_{k_p}\right) \tag{4.166}$$
$$- \frac{(p-1)!}{N^{p-1}} \sum_{\substack{1\le k_2<k_3<\cdots<k_p\le N \\ \forall l,k_l\neq i}} R_{i,k_2,k_3\cdots k_p} \left(\hat{x}^t_{k_2}\hat{x}^{t^\top}_{k_2}+\sigma^t_{k_2}\right) \circ \cdots \circ \left(\hat{x}^t_{k_p}\hat{x}^{t^\top}_{k_p}+\sigma^t_{k_p}\right) ,$$

$$\hat{x}^{t+1}_i = f_{\text{in}}(A^t_i, B^t_i) , \tag{4.167}$$
$$\sigma^{t+1}_i = \partial_B f_{\text{in}}(A^t_i, B^t_i) , \tag{4.168}$$

One can use the self averaging in the large N limit to further simplify these equations.

$$B_i^t = \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{i_2<k_3<\cdots<k_p} S_{i,k_2,k_3\cdots k_p} \hat{x}^t_{k_2} \circ \hat{x}^t_{k_3} \circ \cdots \circ \hat{x}^t_{k_p}$$
$$\frac{(p-1)}{\widetilde{\Delta}} \left[\left(\frac{1}{N}\sum_{1\le k\le N} \sigma^t_k\right) \circ \left(\frac{1}{N}\sum_{1\le k\le N} \hat{x}^t_k \hat{x}^{t-1^\top}_k\right)^{\circ(p-2)} \right] \hat{x}^{t-1}_i, \tag{4.169}$$

$$A^t = \frac{1}{\widetilde{\Delta}} \left(\frac{1}{N}\sum_{1\le k\le N} \hat{x}^t_k \hat{x}^t_k\right)^{\circ(p-1)} - \overline{R} \left(\frac{1}{N}\sum_{1\le k\le N} \hat{x}^t_k \hat{x}^{t^\top}_k + \sigma^t_k\right)^{\circ(p-1)} , \tag{4.170}$$

$$\hat{x}^{t+1}_i = f_{\text{in}}(A^t, B^t_i) , \tag{4.171}$$
$$\sigma^{t+1}_i = \partial_B f_{\text{in}}(A^t, B^t_i) , \tag{4.172}$$

Advantage of self-averaging

We can further simplify these equations by noticing that in all the expressions where S^2_{ij} appears we can replace it by its mean without changing the leading order of the quantities. This follows from the assumption made in the BP equations, that states that the messages incoming to a node are independent conditionally on the value of the node. Consequently the sums in eqs. (4.90-4.91) are sum of $O(N)$ independent variables and can hence in the leading order be replaced by their means.

This allows us to write the Low-RAMP equations (4.99-4.100) in an even simpler form

$$B^t_{X,i} = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} S_{ki} \hat{x}^t_k - \left(\frac{1}{N\widetilde{\Delta}} \sum_{k=1}^{N} \sigma^t_{x,k} \right) \hat{x}^{t-1}_i \,, \tag{4.173}$$

$$A^t_X = \frac{1}{N\widetilde{\Delta}} \sum_{k=1}^{N} \hat{x}^t_k \hat{x}^{t,\top}_k - \overline{R} \frac{1}{N} \sum_{k=1}^{N} \left(\hat{x}^t_k \hat{x}^{t,\top}_k + \sigma^t_{x,k} \right) \,, \tag{4.174}$$

where we defined

$$\frac{1}{\widetilde{\Delta}} \equiv \frac{2}{N^2} \sum_{1 \le i < j \le N} S^2_{ij} \,, \tag{4.175}$$

$$\overline{R} \equiv \frac{2}{N^2} \sum_{1 \le i < j \le N} R_{ij} \,. \tag{4.176}$$

Whereas $\widetilde{\Delta}$ is always positive, $\overline{R}$ can be positive or negative. Together with eqs. (4.96-4.97) the above two expressions are a closed set of equations. Note in particular that in (4.174) the dependence on index i disappeared in the leading order.

Bayes-optimal case

In the Bayes-optimal inference case we derived expression (4.14). Putting it together with the definition of the matrix R_{ij} in eq. (4.19) and realizing that the average over sites i, j and the average over $P(Y|w)$ act the same way we get that for $\overline{R}$ as defined in (4.176) we have $\overline{R} = 0$. This property belongs to the class of properties called the Nishimori conditions. In the Bayes optimal case the expression for A^t_X simplified further into

$$A^t_X = \frac{1}{N\Delta} \sum_{k=1}^{N} \hat{x}^t_k \hat{x}^{t,\top}_k \,, \tag{4.177}$$

where we used the Bayes-optimality once more to realize that $\widetilde{\Delta} = \Delta$ as defined in eq. (4.15). The convenient property of the Bayes-optimality is that the quantity A^t_X now has to be non-negative.

4.2.5 Bethe Free Energy

We define the free energy of a given probability measure as the logarithm of its normalization (in physics it is usually the negative logarithm, but in this thesis we adopt the definition without the minus sign). Notably for the symmetric vector-spin glass model (2.6) we define

$$\Phi_{XX^\top}(Y) = \log\left(Z_X(Y)\right) - \sum_{1 \le i < j \le N} g(Y_{ij}, 0) \,, \tag{4.178}$$

where $Z_X(Y)$ is the normalization, i.e. the partition function, in (2.6). We subtract the constant term on the right hand side for convenience. For the bipartite vector-spin glass model (2.7)

analogously

$$\Phi_{UV^\top}(Y) = \log\left(Z_{UV}(Y)\right) - \sum_{1\le i<N, 1\le j\le M} g(Y_{ij}, 0)\,. \tag{4.179}$$

We remove the constant term $g(Y_{ij}, 0)$ so that the Free energy $\Phi_{XX^\top}$ and $\Phi_{UV^\top}$ will be $O(N)$ and self averaging in the large N limit. The exact free energies are intractable to compute, therefore we use approximations equivalent to those used for derivation of the Low-RAMP algorithm to derive the so-called *Bethe free energy*. Under the assumption of replica symmetry this free energy is exact in the leading order in N and with high probability over the ensemble of instances. To derive the Bethe free energy we use the Plefka expansion, later extended by Georges and Yedidia [Ple82, GY91]. The derivation is presented in the section 3.3.

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

$$\Phi_{\text{Bethe},XX^\top} = \max_{\{A_{X,i}\},\{B_{X,i}\}} \Phi_{\text{Bethe},XX^\top}(\{A_{X,i}\},\{B_{X,i}\}) \tag{4.180}$$

$$\begin{aligned}\Phi_{\text{Bethe},XX^\top}(\{A_{X,i}\},\{B_{X,i}\}) = &\sum_{1\le i\le N} \log(Z_x(A_{X,i}, B_{X,i})) - B_{X,i}^\top \hat{x}_i + \frac{1}{2}\mathrm{Tr}\left[A_{X,i}(\hat{x}_i\hat{x}_i^\top + \sigma_{x,i})\right] \\ &+ \frac{1}{2}\sum_{1\le i,j\le N}\left[\frac{1}{\sqrt{N}}S_{ij}\hat{x}_i^\top\hat{x}_j + \frac{R_{ij}}{2N}\mathrm{Tr}\left[(\hat{x}_i\hat{x}_i^\top + \sigma_{x,i})(\hat{x}_j\hat{x}_j^\top + \sigma_{x,j})\right]\right. \\ &\left. - \frac{S_{ij}^2}{2N}\mathrm{Tr}\left[\hat{x}_i\hat{x}_i^\top\hat{x}_j\hat{x}_j^\top\right] - \frac{1}{N}S_{ij}^2\mathrm{Tr}\left[\sigma_{x,i}\sigma_{x,j}\right]\right]\,,\end{aligned} \tag{4.181}$$

where $\hat{x}_i = f_{\text{in}}^x(A_{X,i}, B_{X,i})$ and $\sigma_{x,i} = \partial_B f_{\text{in}}^x(A_{X,i}, B_{X,i})$ are considered as explicit functions of $A_{X,i}$ and $B_{X,i}$, where the function f_{in}^x depends on the prior probability distribution P_X via eq. (4.86).

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

$$\Phi_{\text{Bethe},UV^\top} = \max_{\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}} \Phi_{\text{Bethe},UV^\top}(\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\})\,, \tag{4.182}$$

$$\begin{aligned}&\Phi_{\text{Bethe},UV^\top}(\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}) = \\ &\qquad\sum_{1\le i\le N} \log(Z_u(A_{U,i}, B_{U,i})) - {B_{U,i}}^\top \hat{u}_i + \frac{1}{2}\mathrm{Tr}\left[A_{U,i}(\hat{u}_i\hat{u}_i^\top + \sigma_{u,i})\right] \\ &\qquad+ \sum_{1\le j\le M} \log(Z_v(A_{V,j}, B_{V,j})) - {B_{V,j}}^\top \hat{v}_j + \frac{1}{2}\mathrm{Tr}\left[A_{V,j}(\hat{v}_j\hat{v}_j^\top + \sigma_{v,j})\right] \\ &\qquad+ \sum_{1\le i\le N, 1\le j\le M}\left[\frac{1}{\sqrt{N}}S_{ij}\hat{u}_i^\top\hat{v}_j + \frac{1}{2N}R_{ij}\mathrm{Tr}\left[(\hat{u}_i\hat{u}_i^\top + \sigma_{u,i})(\hat{v}_j\hat{v}_j^\top + \sigma_{v,j})\right]\right. \\ &\qquad\left. - \frac{S_{ij}^2\mathrm{Tr}\left(\hat{u}_i\hat{u}_i^\top\hat{v}_j\hat{v}_j^\top\right)}{2N} - \frac{1}{N}S_{ij}^2\mathrm{Tr}\left[\sigma_{u,i}\sigma_{v,j}\right]\right]\,,\end{aligned} \tag{4.183}$$

where the $\hat{u}_i = f_{\text{in}}^u(A_{U,i}, B_{U,i})$, $\hat{v}_j = f_{\text{in}}^v(A_{V,j}, B_{V,j})$ and the $\sigma_{u,i} = \partial_B f_{\text{in}}^u(A_{U,i}, B_{U,i})$, $\sigma_{v,j} = \partial_B f_{\text{in}}^v(A_{V,j}, B_{V,j})$, are again seen as a function of variables A, and B. Note that fixed points

of the Low-RAMP algorithm are stationary points of the Bethe free energy as can be checked explicitly by taking the derivatives of the formulas.

The main usage of the free energy is when there exist multiple fixed points of the Low-RAMP equations then the one that corresponds to the best achievable mean squared error is the one for which the free energy is the largest. Another way to use the free energy is in order to help the convergence of the Low-RAMP equations, the adaptive damping is used [RSR+13, VSR+15] and relies on the knowledge of the above expression for the Bethe free energy.

To conclude, we recall that Low-RAMP is distributed in Matlab and Julia at `http://krzakala.github.io/LowRAMP/`, in a version that include the use of the Bethe free energy as a guide to increase convergence, as in [VSR+15].

Conventional Hamiltonian : SK model

Another case that is worth specifying is the conventional Hamiltonian where g is given by (4.22). One gets

$$A_X^t = -\overline{R}\frac{1}{N}\sum_{k=1}^{N}\sigma_{x,k}^t\,. \tag{4.184}$$

It is slightly counter-intuitive that this variance-like term is negative, but it is always used only in the function f_{in}^x defined in (4.86) where it gets multiplied by the $P_X(x_i)$ therefore if P_X is decaying fast enough or has a bounded support, the corresponding integrals exist and are finite.

This is a convenient point where we can make the link between the Low-RAMP algorithm and the TAP equations for the SK model. For the Ising spins (4.53) the function f_{in}^x becomes

$$f_{\text{in}}^x(A,B) = \tanh\left(\beta h + B\right), \quad \frac{\partial f_{\text{in}}^x(A,B)}{\partial B} = 1 - \tanh^2\left(\beta h + B\right). \tag{4.185}$$

Notice the independence on the parameter A. The conventional Hamiltonian of the SK model corresponds to $g(Y,w) = \beta Y w$ so that $S = \beta Y$. Which gives us for the update of the parameter B eq. (4.99)

$$B_{X,i}^t = \frac{\beta}{\sqrt{N}}\sum_{k=1}^{N} Y_{ki}\hat{x}_k^t - \hat{x}_i^{t-1}\frac{\beta^2}{N}\sum_{k=1}^{N} Y_{ki}^2(1-\hat{x}_k^{2,t})\,. \tag{4.186}$$

Together with (4.185) we get the well known TAP equations [TAP77]

$$\hat{x}_i^{t+1} = \tanh\left[\beta h + \frac{\beta}{\sqrt{N}}\sum_{k=1}^{N} Y_{ki}\hat{x}_k^t - \hat{x}_i^{t-1}\frac{\beta^2}{N}\sum_{k=1}^{N} Y_{ki}^2(1-\hat{x}_k^{2,t})\right]. \tag{4.187}$$

4.3 State Evolution

An appealing property of the Low-RAMP algorithm is that its large-system-size behavior can be analyzed via the so called state evolution (or single letter characterization in information theory). In the statistical physics context the state evolution is the cavity method [MPV87] thanks to which one can derive the replica symmetric solution from the TAP equations, taking

properly into account the distribution of the disorder (random or quenched). Mathematically, at least in the Bayes optimal setting, the state evolution for the present systems is a rigorous statement about the asymptotic behavior of the Low-RAMP algorithm [JM13].

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

4.3.1 Derivation for the symmetric low-rank estimation

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

$$M_x^t = \frac{1}{N} \sum_{1 \leq i \leq N} \hat{x}_i^t x_i^{0,\top} \in \mathbb{R}^{r \times r_0}, \tag{4.188}$$

$$Q_x^t = \frac{1}{N} \sum_{1 \leq i \leq N} \hat{x}_i^t \hat{x}_i^{t,\top} \in \mathbb{R}^{r \times r}, \tag{4.189}$$

$$\Sigma_x^t = \frac{1}{N} \sum_{1 \leq i \leq N} \sigma_{x,i}^t \in \mathbb{R}^{r \times r}\,, \tag{4.190}$$

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

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

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

We do assume, however, that (4.13) holds for our choice of $g(Y, w)$ and $P_{\text{out}}(Y|w)$ even when $g(Y, w) \neq \log P_{\text{out}}(Y|w)$. This will indeed hold in all examples presented in this thesis. Using self-averaging arguments such as in Sec. 4.2.4 the averages over the quenched randomness $P_{\text{out}}(Y|w)$ and over elements (ij) are interchangeable. Eq. (4.13) then in practice means that, in order for the state evolution as derived in this section to be valid, the Fisher score matrix S should have an empirical mean of elements of order $o\left(1/\sqrt{N}\right)$. If this assumption is not met, i.e. we have $\mathbb{E}(S) = a/\sqrt{N}$ with $a \gg 1$, it means that the matrix $S/\sqrt{N}$ will have an eigenvalue of order a, while the eigenvalues corresponding to the planted signal will be $O(1)$. This means that for $a \gg 1$ the eigenvalues corresponding to the signal will be subdominant and this would require additional terms in the state evolution.

We know that $\hat{x}_i^{t+1} = f_{\text{in}}(A_{X,i}^t, B_{X,i}^t)$ and $\sigma_{x,i}^{t+1} = \frac{\partial f_{\text{in}}}{\partial B}(A_{X,i}^t, B_{X,i}^t)$, eqs. (4.96-4.97). Therefore to compute the updated order parameters in the large N limit one needs to compute the

probability distribution of $B^t_{X,i}$ and $A^t_{X,i}$

$$P(B^t_{X,i}|x^0_i, Q^t_x, M^t_x, \Sigma^t_x)\,, \tag{4.191}$$
$$P(A^t_{X,i}|x^0_i, Q^t_x, M^t_x, \Sigma^t_x)\,. \tag{4.192}$$

Quantities $B^t_{X,i}$ and $A^t_{X,i}$ are defined by eq. (4.94) and (4.95). Notably, by the assumptions of belief propagation the terms in the sums on the right hand side of eqs. (4.94) and (4.95) are independent. By the central limit theorem $B^t_{X,i}$ and $A^t_{X,i}$ then behave as Gaussian random variables.

Therefore all one needs to compute is their mean and variance with respect to the output channel. Using eq. (4.94) we get

$$\mathbb{E}(B^t_{X,i}) = \frac{1}{\sqrt{N}} \sum_{1\leq k\leq N} \int \mathrm{d}Y_{ki} P_{\mathrm{out}} \left(Y_{ki} \middle| \frac{x^{0,\top}_k x^0_i}{\sqrt{N}} \right) \left(\frac{\partial g}{\partial w} \right)_{Y_{ki},0} \hat{x}^t_{k\to ki}\,. \tag{4.193}$$

Let us now expand P_{out} around 0

$$\mathbb{E}(B^t_{X,i}) = \frac{1}{\sqrt{N}} \sum_{1\leq k\leq N} \int \mathrm{d}Y_{ki} P_{\mathrm{out}}\left(Y_{ki}|0\right) \left[1 + \frac{x^{0,\top}_k x^0_i}{\sqrt{N}} \left(\frac{\partial \log P_{\mathrm{out}}(Y_{ki}|w)}{\partial w} \right)_{Y_{ki},0} + O\left(\frac{1}{N}\right) \right] \left(\frac{\partial g}{\partial w} \right)_{Y_{ki},0} \hat{x}^t_{k\to ki}\,. \tag{4.194}$$

Using the above stated assumption of validity of eq. (4.13) we can simplify into

$$\mathbb{E}(B^t_{X,i}) = \frac{1}{N\widehat{\Delta}} \sum_{1\leq k\leq N} \hat{x}^t_k x^{0,\top}_k x^0_i = \frac{M^t_x}{\widehat{\Delta}} x^0_i\,, \tag{4.195}$$

where we used the definition (4.188) of the order parameter M^t and where we defined $\widehat{\Delta}$ via

$$\frac{1}{\widehat{\Delta}} \equiv \mathbb{E}_{P_{\mathrm{out}}} \left[\left(\frac{\partial g(Y,w)}{\partial w} \right)_{Y,0} \left(\frac{\partial \log(P_{\mathrm{out}}(Y|w))}{\partial w} \right)_{Y,0} \right]. \tag{4.196}$$

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

$$\mathrm{Cov}(B^t_{X,i}) = \frac{1}{N} \sum_{1\leq i\leq N} \mathrm{Cov}(S_{ki}\hat{x}^t_{k\to ki})\,. \tag{4.197}$$

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

$$\mathrm{Cov}(B^t_{X,i}) = \frac{1}{N\widetilde{\Delta}} \sum_{1\leq i\leq N} \hat{x}^t_{k\to ki} \hat{x}^{t,\top}_{k\to ki} = \frac{Q^t_x}{\widetilde{\Delta}}\,, \tag{4.198}$$

where $\widetilde{\Delta}$ was introduced in (4.175) and thanks to self-averaging it also equals

$$\frac{1}{\widetilde{\Delta}} = \mathbb{E}_{P_{\rm out}(Y|w)} \left[\left(\frac{\partial g}{\partial w} \right)^2_{Y,0} \right] . \tag{4.199}$$

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

The distribution of the $A^t_{X,i}$ now needs to be computed. Using the definition of $A^t_{X,i}$ eq. (4.95) and the self-averaging of section 4.2.4 we obtain directly that

$$\mathbb{E}(A^t_{X,i}) = \frac{Q^t_x}{\widetilde{\Delta}} - \overline{R} \left(Q^t_x + \Sigma^t_x \right) , \tag{4.200}$$

where $\overline{R}$ is defined in (4.176) and also equals

$$\overline{R} = \mathbb{E}_{P_{\rm out}(Y|w)} \left[\left(\frac{\partial g}{\partial w} \right)^2_{Y,0} + \left(\frac{\partial^2 g}{\partial w^2} \right)_{Y,0} \right] . \tag{4.201}$$

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

Overall, using (4.96) and (4.97) one gets for the state evolution equations

$$M^{t+1}_x = \mathbb{E}_{x_0,W} \left[f^x_{\rm in} \left(A^t, \hat{M}^t_x x_0 + \sqrt{\hat{Q}^t_x} W \right) x_0^\top \right] , \tag{4.202}$$

$$Q^{t+1}_x = \mathbb{E}_{x_0,W} \left[f^x_{\rm in} \left(A^t, \hat{M}^t_x x_0 + \sqrt{\hat{Q}^t_x} W \right) f^x_{\rm in}(\cdots, \cdots)^\top \right] , \tag{4.203}$$

$$\Sigma^{t+1}_x = \mathbb{E}_{x_0,W} \left[\frac{\partial f^x_{\rm in}}{\partial B} \left(A^t, \hat{M}^t_x x_0 + \sqrt{\hat{Q}^t_x} W \right) \right] , \tag{4.204}$$

Where ,

$$A^t = \frac{Q^t_x}{\widetilde{\Delta}} - \overline{R}(Q^t_x + \Sigma^t_x) \tag{4.205}$$

$$\hat{M}^t_x = \frac{M^t_x}{\widehat{\Delta}} \tag{4.206}$$

$$\hat{Q}^t_x = \frac{M^t_x}{\widehat{\Delta}} \tag{4.207}$$

where W and x_0 are two independent random variables. W is a Gaussian noise of mean 0 and covariance matrix I_r and x_0 is a random variable of probability distribution P_{X_0}. The thresholding function $f^x_{\rm in}$ is defined in eq. (4.86).

The interpretation of this formula is simple. Because of the dense nature of the interactions the messages received by a variable are Gaussians messages parameterized by two variables A_i and B_i (The A in AMP means that the messages are Gaussian). These variables B_i and A_i are going to be the sum of a large number of random variables. A sum of a large number of random

variables tends to be distributed according to a Gaussian. The idea of State Evolution equations is to characterize how these messages are distributed using only a few order parameters of the system.

For every system that will be treated using a replica symmetric ansatz in that thesis the formula (4.202,4.203,4.204) will apply (except for the bipartite $UV^\top$ case that will be slightly modified). The tensor case will just provide use with different update equations for the matrices A^t,$\hat{M}_x^t$ and $\hat{Q}_x^t$.

This also allows us to know that at a fixed point the marginals x_i will be distributed according to

$$\hat{x_i}^{t=+\infty} = f_{\text{in}}^x\left(A^t, \hat{M}_x^t x_0 + \sqrt{\hat{Q}_x^t} W\right) \tag{4.208}$$

or (4.209)

$$\hat{x_i}^{t=+\infty} = f_{\text{in}}^x\left(\frac{Q_x}{\widetilde{\Delta}} - \overline{R}(Q_x + \Sigma_x), \frac{M_x}{\widehat{\Delta}} x_0 + \sqrt{\frac{Q_x}{\widetilde{\Delta}}} W\right) , \tag{4.210}$$

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

$$\text{MSE}_X = \text{Tr}\left[\langle x_0 x_0^\top\rangle - 2M_x + Q_x\right] , \tag{4.211}$$

where $\langle x_0 x_0^\top\rangle = \mathbb{E}_{x_0}(x_0 x_0^\top)$ is the average with respect to the distribution P_{X_0}.

Note that the state evolution equations only depend on the assumed and truth noise channels through three variables $\widetilde{\Delta}$, $\widehat{\Delta}$ and $\overline{R}$. In the Bayes-optimal case these equations will simplify even further and the noise channel will be described through one parameter Δ, the Fisher information, this is derived in section 4.3.6. This universality with respect to the output channel has been observed elsewhere in a special case of the present problem [DM15] (see e.g. their remark 2.5) in the study of detection of a small hidden clique with approximate message passing.

Finally one additional assumption made in this whole section is that no Replica Symmetry Breaking (RSB) appears. It is known that RSB does appear for some regimes of parameters out of the equilibrium Bayes-optimal case. We let the investigation of RSB in the context of low-rank matrix estimation for future work, in the examples analyzed in this thesis we will restrict ourselves to the Bayes-optimal case where RSB at equilibrium cannot happen [ZK16].

4.3.2 Summary for the graphon case

State evolution can also be written similarly for the $w_{ij} = f(x_i, x_j)$ case. The order parameters one has to keep track of are now 3 multiplicative kernels $M_x^t(x_1, x_2), Q_x^t(x_1, x_2)$ and $\Sigma_x^t(x_1, x_2)$

defined by

$$M_x^t(x_1,x_2)=\frac{1}{N}\sum_{1\leq i\leq N}\hat{P}_{X,i}^t(x_1)\delta(x_2-x_i^0),\quad Q_x^t(x_1,x_2)=\frac{1}{N}\sum_{1\leq i\leq N}\hat{P}_{X,i}^t(x_1)\hat{P}_{X,i}^t(x_2), \tag{4.212}$$

$$\Sigma_x^t(x_1,x_2)+Q_x^t(x_1,x_2)=\frac{1}{N}\sum_{1\leq i\leq N}\hat{P}_{X,i}^t(x_1)\delta(x_1-x_2)\,, \tag{4.213}$$

These order parameters are then updated according to the following equations

$$M_x^{t+1}(x_1,x_2)=\mathbb{E}_{x_0,W}\left[\frac{1}{Z_x(B_{x_0,W})}P_X(x_1)\exp(B_{x_0,W}(x_1))\delta(x_2-x_0)\right], \tag{4.214}$$

$$Q_x^{t+1}(x_1,x_2)=\mathbb{E}_{x_0,W}\left[\frac{P_X(x_1)P_X(x_2)\exp(B_{x_0,W}(x_1)+B_{x_0,W}(x_2))}{Z_x(B_{x_0,W})^2}\right], \tag{4.215}$$

$$\Sigma_x^{t+1}(x_1,x_2)+Q_x^{t+1}(x_1,x_2)=\mathbb{E}_{x_0,W}\left[\frac{1}{Z_x(B_{x_0,W})}P_X(x_1)\exp(B_{x_0,W}(x_1))\delta(x_1-x_2)\right], \tag{4.216}$$

where

$$B_{x_0,W}(x)=\frac{\left[f*M*f^0\right](x,x_0)}{\widehat{\Delta}}+f*W-\frac{1}{2\widetilde{\Delta}}\left[f*Q_x*f\right](x,x)+\frac{\overline{R}}{2}\left[f*(\Sigma_x^t+Q_x^t)*f\right](x,x) \tag{4.217}$$

And where $W(x)$ is a Gaussian process of mean 0 and with a covariance given by

$$\overline{W(x_1)W(x_2)}=\frac{Q_x^t(x_1,x_2)}{\widetilde{\Delta}} \tag{4.218}$$

x_0 is sampled from P_X^0. f^0 is the function $w=f^0(x_1,x_2)$ with which the data were created. Estimating (4.214,4.215,4.216) is a hard task in itself. Integrating the variable x_0 and variable in x are the easy part of the computation, since these can be approximated using numerical approximations. Integrating against the Gaussian process $W(x)$ is the hard part of the computation since $W(x)$ can be an infinite dimension variable. The only way I see to reliably compute this integral is through a Monte-Carlo simulation. To do this one first need to compute a low-rank approximation of Q_x^t.

$$Q_x^t(x_1,x_2)\approx\sum_{k=1\cdots h}W_k(x_1)\lambda_k W_k(x_2) \tag{4.219}$$

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_h>0 \tag{4.220}$$

$$\int \mathrm{d}x W_k^2(x)=1 \tag{4.221}$$

A sample of W is then given by

$$W(x) = \sum_{k=1\cdots h} \epsilon_k \sqrt{\lambda_k} W_k(x) \tag{4.222}$$

where

$$\epsilon_k \approx \mathcal{N}(0,1) \tag{4.223}$$

Of course we would actually need a infinite number of terms h to describe Q_x^t and sample W. In practice we will have to stop at a high enough value of number of term to describe (4.214,4.215,4.216) accurately.

4.3.3 Summary for the bipartite low-rank matrix factorization

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

$$M_u^t = \frac{1}{N}\sum_{1\le i\le N} u_i^t u_i^{0,\top}, \quad Q_u^t = \frac{1}{N}\sum_{1\le i\le N} u_i^t u_i^{t,\top}, \quad \Sigma_u^t = \frac{1}{N}\sum_{1\le i\le N} \sigma_{u,i}^t\,, \tag{4.224}$$

$$M_v^t = \frac{1}{M}\sum_{1\le j\le M} v_j^t v_j^{0,\top}, \quad Q_v^t = \frac{1}{M}\sum_{1\le j\le M} v_j^t v_j^{t,\top}, \quad \Sigma_v^t = \frac{1}{M}\sum_{1\le j\le M} \sigma_{v,j}^t\,. \tag{4.225}$$

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

$$M_u^t = \mathbb{E}_{u_0,W}\left[f_{\text{in}}^u\left(\frac{\alpha Q_v^t}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v^t+\Sigma_v^t), \alpha\frac{M_v^t}{\widetilde{\Delta}}u_0 + \sqrt{\frac{\alpha Q_v^t}{\widetilde{\Delta}}}W\right) u_0^\top\right], \tag{4.226}$$

$$Q_u^t = \mathbb{E}_{u_0,W}\left[f_{\text{in}}^u\left(\alpha\frac{Q_v^t}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v), \alpha\frac{M_v^t}{\widetilde{\Delta}}u_0 + \sqrt{\frac{\alpha Q_v^t}{\widetilde{\Delta}}}W_v\right) f_{\text{in}}^u(\cdots,\cdots)^\top\right], \tag{4.227}$$

$$\Sigma_u^t = \mathbb{E}_{u_0,W}\left[\frac{\partial f_{\text{in}}^u}{\partial B}\left(\alpha\frac{Q_v^t}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v^t+\Sigma_v^t), \alpha\frac{M_v^t}{\widetilde{\Delta}}u_0 + \sqrt{\frac{\alpha Q_v^t}{\widetilde{\Delta}}}W\right)\right], \tag{4.228}$$

$$M_v^{t+1} = \mathbb{E}_{v_0,W}\left[f_{\text{in}}^v\left(\frac{Q_u^t}{\widetilde{\Delta}} - \overline{R}(Q_u^t+\Sigma_u^t), \frac{M_u^t}{\widetilde{\Delta}}v_0 + \sqrt{\frac{Q_u^t}{\widetilde{\Delta}}}W\right) v_0^\top\right], \tag{4.229}$$

$$Q_v^{t+1} = \mathbb{E}_{v_0,W}\left[f_{\text{in}}^v\left(\frac{Q_u^t}{\widetilde{\Delta}} - \overline{R}(Q_u^t+\Sigma_u^t), \frac{M_u^t}{\widetilde{\Delta}}v_0 + \sqrt{\frac{Q_u^t}{\widetilde{\Delta}}}W\right) f_{\text{in}}^v(\cdots,\cdots)^\top\right], \tag{4.230}$$

$$\Sigma_v^{t+1} = \mathbb{E}_{v_0,W}\left[\frac{\partial f_{\text{in}}^v}{\partial B}\left(\frac{Q_u^t}{\widetilde{\Delta}} - \overline{R}(Q_u^t+\Sigma_u^t), \frac{M_u^t}{\widetilde{\Delta}}v_0 + \sqrt{\frac{Q_u^t}{\widetilde{\Delta}}}W\right)\right]. \tag{4.231}$$

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

The large size limit of the $\text{MSE}_U = \sum_{i=1\ldots N} ||\hat{u}_i - u_i^0||_2^2/N$ and $\text{MSE}_V = \sum_{j=1\ldots M} ||\hat{v}_j - v_j^0||_2^2/M$ can

be computed from the order parameters as

$$\mathrm{MSE}_U = \mathrm{Tr}\left[\langle u_0 u_0^\top\rangle - 2M_u + Q_u\right] , \tag{4.232}$$

$$\mathrm{MSE}_V = \mathrm{Tr}\left[\langle v_0 v_0^\top\rangle - 2M_v + Q_v\right] , \tag{4.233}$$

with $\langle u_0 u_0^\top\rangle = \mathbb{E}_{u_0}(u_0 u_0^\top)$ and $\langle v_0 v_0^\top\rangle = \mathbb{E}_{v_0}(v_0 v_0^\top)$.

4.3.4 Summary for the tensor case

State evolution can also be written similarly for the tensor case. In that case there are six order parameters are given by (4.1884.1894.190)

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

$$M_x^{t+1} = \mathbb{E}_{x_0,W}\left[f_{\mathrm{in}}^x\left(\hat{Q}_x^t - \hat{\Sigma}_x^t, \hat{M}_x^t x_0 + \sqrt{\hat{Q}_x^t}W\right) x_0^\top\right] , \tag{4.234}$$

$$Q_x^{t+1} = \mathbb{E}_{x_0,W}\left[f_{\mathrm{in}}^x\left(\hat{Q}_x^t - \hat{\Sigma}_x^t, \hat{M}_x^t x_0 + \sqrt{\hat{Q}_x^t}W\right) f_{\mathrm{in}}^x(\cdots,\cdots)^\top\right] , \tag{4.235}$$

$$\Sigma_x^{t+1} = \mathbb{E}_{x_0,W}\left[\frac{\partial f_{\mathrm{in}}^x}{\partial B}\left(\hat{Q}_x^t - \hat{\Sigma}_x^t, \hat{M}_x^t x_0 + \sqrt{\hat{Q}_x^t}W\right)\right] , \tag{4.236}$$

Where

$$\hat{M}_x^t = \frac{{M_x^t}^{\circ^{p-1}}}{\widehat{\Delta}} \tag{4.237}$$

$$\hat{Q}_x^t = \frac{{Q_x^t}^{\circ^{p-1}}}{\widetilde{\Delta}} \tag{4.238}$$

$$\hat{\Sigma}_x^t = \overline{R}(Q_x^t + \Sigma_x^t)^{\circ^{p-1}} \tag{4.239}$$

In these equations W and x_0 are independent random variables, W is r dimensional Gaussian variable of mean $\vec{0}$ and covariance matrix I_r, X_0 is sampled from density probability P_{X_0}.

The large size limit of the $\mathrm{MSE}_X = \sum_{i=1\ldots N} ||\hat{x}_i - x_i^0||_2^2/N$ is once again given by (4.211).

4.3.5 Replica Free Energy

State evolution can also be used to derive large size limit of the Bethe free energy (4.181) and (4.183) defined as

$$\Phi_{\mathrm{RS},XX^\top} \equiv \lim_{N\to+\infty} \frac{1}{N} \left\langle \log(Z_X(Y)) - \sum_{1\leq i<j\leq N} g(Y_{ij},0) \right\rangle , \tag{4.240}$$

$$\Phi_{\mathrm{RS},UV^\top} \equiv \lim_{N\to+\infty} \frac{1}{N} \left\langle \log(Z_{UV}(Y)) - \sum_{1\leq i\leq N, 1\leq j\leq M} g(Y_{ij},0) \right\rangle , \tag{4.241}$$

$$\Phi_{\mathrm{RS},X^p} \equiv \lim_{N\to+\infty} \frac{1}{N} \left\langle \log(Z_{X^p}(Y)) - \sum_{1\leq i_1<i_2<\cdots<i_p\leq N} g(Y_{i_1 i_2\cdots i_p},0) \right\rangle , \tag{4.242}$$

where the average is taken with respect to density probability (2.6), (2.7) or (2.10) the $Z_X(Y)$, $Z_{UV}(Y)$ and $Z_{X^p}(Y)$ are the corresponding partition functions. We subtract the constant $g(Y_{ij},0)$ for convenience in order to get a quantity that is self averaging in the large N limit.

Alternatively to the state evolution, the average free energy can be derived using the replica method as we summarize in Appendix 4.6.2. The resulting replica free energy for the symmetric $XX^\top$ case is (assuming the replica symmetric ansatz to hold)

$\mathbf{XX}^\top$ symmetric case

$$\Phi_{\mathrm{RS},XX^\top} = \max \left(\phi_{\mathrm{RS},XX^\top}(M_x,Q_x,\Sigma_x), \frac{\partial\phi_{\mathrm{RS}}}{\partial M_x} = \frac{\partial\phi_{\mathrm{RS}}}{\partial Q_x} = \frac{\partial\phi_{\mathrm{RS}}}{\partial\Sigma_x} = 0 \right\} , \tag{4.243}$$

where

$$\begin{aligned}\phi_{\mathrm{RS},XX^\top}(M_x,Q_x,\Sigma_x) = {} & \frac{\mathrm{Tr}(Q_x Q_x^\top)}{4\widetilde{\Delta}} - \frac{\mathrm{Tr}(M_x M_x^\top)}{2\widehat{\Delta}} - \frac{\overline{R}}{2}\mathrm{Tr}((Q_x+\Sigma_x)(Q_x+\Sigma_x)^\top) \\ & + \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(\frac{Q_x}{\widetilde{\Delta}} - \overline{R}(Q_x+\Sigma_x), \frac{M_x}{\widehat{\Delta}}x_0 + \sqrt{\frac{Q_x}{\widetilde{\Delta}}}W \right)\right)\right] ,\end{aligned} \tag{4.244}$$

where the function $Z_x(A,B)$ is defined as the normalization in eq. (4.87).

$\mathbf{XX}^\top$ graphon symmetric case $w_{ij} = f(x_i,x_j)$

$$\Phi_{\mathrm{RS,Graphon}} = \max \left(\phi_{\mathrm{RS,Graphon}}(M_x,Q_x,\Sigma_x), \frac{\partial\phi_{\mathrm{RS}}}{\partial M_x} = \frac{\partial\phi_{\mathrm{RS}}}{\partial Q_x} = \frac{\partial\phi_{\mathrm{RS}}}{\partial\Sigma_x} = 0 \right\} , \tag{4.245}$$

where

$$\phi_{\text{RS,Graphon}}(M_x, Q_x, \Sigma_x) = \frac{\text{Tr}(f * Q_x * Q_x^\top)}{4\widetilde{\Delta}} - \frac{\text{Tr}(f * M_x * f_0 * M_x^\top)}{2\widehat{\Delta}}$$
$$s - \frac{\overline{R}}{2}\text{Tr}(f * (Q_x + \Sigma_x) * f * (Q_x + \Sigma_x)^\top) + \mathbb{E}_{W,x_0}\left[\log Z_x\left(B_{x_0,W}\right)\right] , \quad (4.246)$$

Where $B_{x_0,W}$ is defined at (4.217). The function $Z_x(B)$ is defined as the normalization in eq. (4.116).

The trace of an multiplicative kernel is here defined as

$$\text{Tr}\left(C\right) = \int \text{d}x C(x,x) \quad (4.247)$$

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

$$\Phi_{\text{RS},UV^\top} = \max\Bigg\{\phi_{\text{RS},UV^\top}(M_u, Q_u, \Sigma_u, M_v, Q_v, \Sigma_v)$$
$$, \frac{\partial\phi_{\text{RS}}}{\partial M_u} = \frac{\partial\phi_{\text{RS}}}{\partial Q_u} = \frac{\partial\phi_{\text{RS}}}{\partial\Sigma_u} = \frac{\partial\phi_{\text{RS}}}{\partial M_v} = \frac{\partial\phi_{\text{RS}}}{\partial Q_v} = \frac{\partial\phi_{\text{RS}}}{\partial\Sigma_v} = 0\Bigg\} , \quad (4.248)$$

where

$$\phi_{\text{RS},UV^\top}(M_u, Q_u, \Sigma_u, M_v, Q_v, \Sigma_v) = \frac{\alpha\text{Tr}(Q_v Q_u^\top)}{2\widetilde{\Delta}} - \frac{\alpha\text{Tr}(M_v M_u^\top)}{\widehat{\Delta}} - \alpha\overline{R}\text{Tr}((Q_v + \Sigma_v)(Q_u + \Sigma_u)^\top)$$
$$+ \mathbb{E}_{W,u_0}\left[\log\left(Z_u\left(\frac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v + \Sigma_v), \frac{\alpha M_v}{\widehat{\Delta}}u_0 + W\sqrt{\frac{\alpha Q_v}{\widetilde{\Delta}}}\right)\right)\right]$$
$$+ \alpha\mathbb{E}_{W,v_0}\left[\log\left(Z_v\left(\frac{Q_u}{\widetilde{\Delta}} - \overline{R}(Q_u + \Sigma_u), \frac{M_u}{\widehat{\Delta}}v_0 + \sqrt{\frac{Q_u}{\widetilde{\Delta}}}W\right)\right)\right] , \quad (4.249)$$

with the function $Z_u(A, B)$ and $Z_v(A, B)$ also defined as the normalization in eq. (4.87).

$\mathbf{X}^{\circ\mathbf{P}}$ tensor case

$$\Phi_{\text{RS},X^p} = \max\left(\phi_{\text{RS},X^p}(M_x, Q_x, \Sigma_x), \frac{\partial\phi_{\text{RS}}}{\partial M_x} = \frac{\partial\phi_{\text{RS}}}{\partial Q_x} = \frac{\partial\phi_{\text{RS}}}{\partial\Sigma_x} = 0\right\} , \quad (4.250)$$

$$\phi_{\text{RS},X^p}(M_x, Q_x, \Sigma_x) = (p-1)\frac{\text{Sum}(Q_x^{\circ p})}{2p\widetilde{\Delta}} - (p-1)\frac{\text{Sum}(M_x^{\circ p})}{p\widehat{\Delta}} - (p-1)\frac{\overline{R}}{2p}\text{Sum}((Q_x + \Sigma_x)^{\circ p})$$
$$+ \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(\frac{Q_x^{\circ p-1}}{\widetilde{\Delta}} - \overline{R}(Q_x + \Sigma_x)^{\circ p-1}, \frac{M_x^{\circ p-1}}{\widehat{\Delta}}x_0 + \sqrt{\frac{Q_x^{\circ p-1}}{\widetilde{\Delta}}}W\right)\right)\right] , \quad (4.251)$$

Where the Sum means doing the sum over all the coordinate of a vector or a matrix.

It is worth noting that there is a close link between the expression of the replica free energy and the state evolution equations. Namely fixed points of the state evolution equations are stationary points of the replica free energy and vice versa. Therefore, by looking for a stationary point of these equations one finds back the state evolution equations (4.202-4.204,4.226-4.231).

4.3.6 Simplification of the SE equations

Simplification in the Bayes optimal setting

The state evolution equations simplify considerably when we restrict ourselves to the Bayes-optimal setting defined in Sec. 4.1.1 by eq. (2.8).

From the definitions of $\widehat{\Delta}$ in eq. (4.196) and $\widetilde{\Delta}$ in eq. (4.175) and using the identity (2.8) defining the Bayes optimal setting we obtain

$$\frac{1}{\widehat{\Delta}} = \frac{1}{\widetilde{\Delta}} = \frac{1}{\Delta} = \mathbb{E}_{P_{\rm out}(Y,w=0)} \left[\left(\frac{\partial g}{\partial w} \right)^2_{Y,w=0} \right] , \tag{4.252}$$

where Δ is the Fisher information of the output channel defined in eq. (4.15). Note for instance that for the Gaussian input channel (4.21), Δ is simply the variance of the Gaussian noise. The bigger the Δ the harder the inference problem becomes. The smaller the Δ the easier the inference is.

Further consequence of having Bayes optimality (2.8) is that $\overline{R} = \mathbb{E}(R_{ij}) = 0$ as proven in equation (4.14). This simplifies greatly the state evolution equations into

$$M_x^{t+1} = \mathbb{E}_{x_0,W} \left[f^x_{\rm in} \left(\frac{Q_x^t}{\Delta}, \frac{M_x^t}{\Delta} x_0 + \sqrt{\frac{Q_x^t}{\Delta}} W \right) x_0^\top \right] , \tag{4.253}$$

$$Q_x^{t+1} = \mathbb{E}_{x_0,W} \left[f^x_{\rm in} \left(\frac{Q_x^t}{\Delta}, \frac{M_x^t}{\Delta} x_0 + \sqrt{\frac{Q_x^t}{\Delta}} W \right) f^x_{\rm in}(\cdots, \cdots)^\top \right] , \tag{4.254}$$

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

Another property that arises in the Bayes optimal setting (2.8) and follows from the Nishimori condition (3.18) and the definition of the order parameters M_x, Q_x and Σ_x^t in (4.188-4.190) is that

$$Q_x^t = M_x^t = M_x^\top , \qquad Q_x^t + \Sigma_x^t = Q_x + \Sigma_x = \langle x_0 x_0^\top \rangle . \tag{4.255}$$

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

$$M_x^{t+1} = \mathbb{E}_{x_0,W} \left[f^x_{\rm in} \left(\frac{M_x^t}{\Delta}, \frac{M_x^t}{\Delta} x_0 + \sqrt{\frac{M_x^t}{\Delta}} W \right) x_0^\top \right] , \tag{4.256}$$

where x_0 and W are independent random variables distributed as above. For the rest of the

article we define the Bayes-optimal state evolution function $f_{P_X}^{\mathrm{SE}}$ for prior P_X

$$M_x^{t+1} = f_{P_X}^{\mathrm{SE}}\left(\frac{M_x^t}{\Delta}\right) , \tag{4.257}$$

Let us comment on the output channel universality as discussed in Sec. 4.1.2. In the Bayes-optimal setting the channel universality becomes particularly simple and striking. For an arbitrary output channel $P_{\mathrm{out}}(Y|w)$ (for which the expansion done in section 4.1.2 is meaningful) we have the following

— The Low-RAMP algorithm in the Bayes optimal setting depends on the noise channel only through the Fisher score matrix S as defined in eq. (4.18). This is specified in section (4.2.4).
— The state evolution in the Bayes-optimal setting depends on the output channel through the Fisher information of the channel Δ (4.15) as described in section 4.3.6. As a consequence the minimal achievable error, the minimal efficiently achievable error and all other quantities that can be obtained from the state evolution depend on the output channel only trough the Fisher information Δ.

The replica free energy (4.244) in the Bayes-optimal case becomes

$$\phi_{\mathrm{RS},XX^\top}(M_x) = \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(\frac{M_x}{\Delta}, \frac{M_x}{\Delta}x_0 + \sqrt{\frac{M_x}{\Delta}}W\right)\right)\right] - \frac{\mathrm{Tr}(M_x M_x^\top)}{4\Delta} . \tag{4.258}$$

This was first derived in [LKZ15a], and proven for a special case in [DAM16], and later in full generality in [KXZ16, ML16, Mio17b].

For numerical reasons it will often prove useful to compute (4.258) by integrating the derivative with respect to M_x therefore the following equation will prove useful.

$$\frac{\partial \phi_{\mathrm{RS},XX^\top}(M_x)}{\partial M_x} = \frac{1}{2\Delta}\left(M_x' - M_x\right) \tag{4.259}$$

Where the term M_x' here means the update of M_x using equation (4.256). To define the gradient with respect to matrices we use the canonical scalar product between matrices $\mathrm{Tr}[AB^\top]$. We also remind that the order parameter M_x used in the state evolution is related to the mean-squared error as

$$\mathrm{MSE}_X = \mathrm{Tr}\left[\langle x_0 x_0^\top\rangle - M_x\right] , \tag{4.260}$$

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

$$M_u^t = \mathbb{E}_{u_0,W}\left[f_{\mathrm{in}}^u\left(\frac{\alpha M_v^t}{\Delta}, \alpha\frac{M_v^t}{\Delta}u_0 + \sqrt{\frac{\alpha M_v^t}{\Delta}}W\right)u_0^\top\right] , \tag{4.261}$$

$$M_v^{t+1} = \mathbb{E}_{v_0,W}\left[f_{\mathrm{in}}^v\left(\frac{M_u^t}{\Delta}, \frac{M_u^t}{\Delta}v_0 + \sqrt{\frac{M_u^t}{\Delta}}W\right)v_0^\top\right] . \tag{4.262}$$

The replica free energy (4.249) in the Bayes optimal setting becomes

$$\phi_{\mathrm{RS,UV^\top}}(M_u, M_v) = \mathbb{E}_{W,u_0}\left[\log\left(Z_u\left(\frac{\alpha M_v}{\Delta}, \frac{\alpha M_v}{\Delta}u_0 + W\sqrt{\frac{\alpha M_v}{\Delta}}\right)\right)\right] + \alpha\mathbb{E}_{W,v_0}\left[\log\left(Z_v\left(\frac{M_u}{\Delta}, \frac{M_u}{\Delta}v_0 + \sqrt{\frac{M_u}{\Delta}}W\right)\right)\right] - \frac{\alpha\mathrm{Tr}(M_v M_u^\top)}{2\Delta}, \quad (4.263)$$

where once again $M_u = M_u^\top$ and $M_v = M_v^\top$. The derivative with respect to M_u and M_v are given by

$$\frac{\partial\phi_{\mathrm{RS,UV^\top}}(M_u, M_v)}{\partial M_u} = \frac{\alpha}{2\Delta}(M_v' - M_v) \quad (4.264)$$

$$\frac{\partial\phi_{\mathrm{RS,UV^\top}}(M_u, M_v)}{\partial M_v} = \frac{\alpha}{2\Delta}(M_u' - M_u) \quad (4.265)$$

Where once again M_u' and M_v' refer to the updated value of M_u and M_v according to (4.261 and (4.262 respectively. The global maximum of the free energy is asymptotically the equilibrium free energy, the value of M_u and M_v at this maximum is related to the MMSE via

$$\mathrm{MSE}_U = \mathrm{Tr}\left[\langle u_0 u_0^\top\rangle - M_u\right], \quad (4.266)$$

$$\mathrm{MSE}_V = \mathrm{Tr}\left[\langle v_0 v_0^\top\rangle - M_v\right]. \quad (4.267)$$

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

For the tensor of order p case the State evolution in the Bayes optimal case is.

$$M_x^{t+1} = \mathbb{E}_{x_0,W}\left[f_{\mathrm{in}}^x\left(\frac{M_x^{t^{\circ p-1}}}{\Delta}, \frac{M_x^{t^{\circ p-1}}}{\Delta}x_0 + \sqrt{\frac{M_x^{t^{\circ p-1}}}{\Delta}}W\right)x_0^\top\right], \quad (4.268)$$

The Free Energy is then given by

$$\phi_{\mathrm{RS,X^P}}(M_x) = \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(\frac{M_x^{\circ p-1}}{\Delta}, \frac{M_x^{\circ p-1}}{\widehat{\Delta}}x_0 + \sqrt{\frac{M_x^{\circ p-1}}{\Delta}}W\right)\right)\right] - (p-1)\frac{\mathrm{Sum}(M_x^{\circ p})}{2p\Delta}, \quad (4.269)$$

The derivative with respect to M_x of (4.269) yields us with

$$\frac{\partial\phi_{\mathrm{RS,UV^\top}}(M_u, M_v)}{\partial M_u} = \frac{p-1}{2\Delta}(M_x' - M_x)\circ M_x^{\circ p-2} \quad (4.270)$$

Where M_x' is given by (4.268). The MSE is given by (4.260).

Simplification for the conventional Hamiltonian and randomly quenched disorder

Another illustrative example of the state evolution we give in this section is for the conventional Hamiltonian (4.7) with randomly quenched disorder, as this is the case most commonly consi-

dered in the existing physics literature. In that case the model (2.6) corresponds to a generic vectorial spin glass model. To take into account that the disorder is not planted, but random, we plug into the generic state evolution

$$P_{\text{out}}(Y, w) = \frac{1}{\sqrt{2\pi J^2}} \exp\left(-\frac{Y^2}{2J^2}\right) \tag{4.271}$$

such that $P_{\text{out}}(Y, w)$ does not depend on w, meaning that the disorder Y is chosen independently, there is no planting. For the conventional Hamiltonian (4.7) and output channel (4.271) we obtain

$$\overline{R} = \frac{1}{\widetilde{\Delta}} = \mathbb{E}\left[Y^2\right] = J^2\,, \tag{4.272}$$

$$\frac{1}{\widehat{\Delta}} = 0\,. \tag{4.273}$$

The state evolution (4.202) and (4.203) then becomes

$$M_x^{t+1} = 0\,, \tag{4.274}$$

$$Q_x^{t+1} = \mathbb{E}_W\left[f_{\text{in}}^x\left(-J^2\Sigma_x^t, J\sqrt{Q_x^t}W\right) f_{\text{in}}^x(\cdots,\cdots)^\top\right]\,, \tag{4.275}$$

$$\Sigma_x^{t+1} = \mathbb{E}_W\left[\frac{\partial f_{\text{in}}^x}{\partial B}\left(-J^2\Sigma_x^t, J\sqrt{Q_x^t}W\right)\right]\,. \tag{4.276}$$

The free energy (4.244) is given by

$$\begin{aligned}\phi_{\text{RS}}(M_x, Q_x, \Sigma_x) = \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(-J^2\Sigma_x, J\sqrt{Q_x}W\right)\right)\right] \\ + \frac{J^2\text{Tr}(Q_xQ_x^\top)}{4} - \frac{J^2}{2}\text{Tr}((Q_x+\Sigma_x)(Q_x+\Sigma_x)^\top)\,.\end{aligned} \tag{4.277}$$

Specifically, for the Sherrington-Kirkpatrick model [SK75], where the rank is one and the spins are Ising eq. (4.53) with $\rho = 1/2$ the $f_{\text{in}}^x(A, B)$ is given by $\tanh(B)$, the state evolution becomes

$$M_x^{t+1} = 0\,, \tag{4.278}$$

$$Q_x^{t+1} = \mathbb{E}_W\left[\tanh\left(J\sqrt{Q_x^t}W\right)^2\right]\,, \tag{4.279}$$

$$\Sigma_x^{t+1} = 1 - Q_x^t\,. \tag{4.280}$$

where W is a Gaussian random variables of zero mean and unit variance. With a free energy (4.244) given by

$$\phi_{\text{RS}}(Q_x) = \frac{J^2(1-Q_x)^2 - J^2}{4} + \mathbb{E}_W\left[\log\left(\cosh\left(J\sqrt{Q_x}W\right)\right)\right]\,. \tag{4.281}$$

The reader will notice that these are just the replica symmetric equations of the Sherrington-Kirkpatrick solution [SK75].

4.4 General results about low-rank matrix estimation

4.4.1 Analysis of the performance of PCA : (Matrix case)

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

PCA method for $XX^\top$ case

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

$$P(X|Y) = \frac{1}{Z(Y)} \prod_{1\le i\le N} \frac{\exp(-\|x_i\|_2^2/2)}{\sqrt{2\pi}^r} \prod_{1\le i<j\le N} \exp\left(\beta g\left(Y, \frac{x_i^\top x_j}{\sqrt{N}}\right)\right) . \tag{4.282}$$

The function $g(Y,w)$ defines the parameters $\widehat{\Delta}$ (4.196), $\widetilde{\Delta}$ (4.175), and $\overline{R}$ (4.176). We want to analyze the overlap between $\hat{X}$ and X_0 in the limit $N \to \infty$, $\beta \to \infty$ since then the posterior will be dominated by the likelihood terms $g(Y,w)$. We put here a prior P_X Gaussian to ensure that $Z(Y) < +\infty$. We could have chosen any β-independent prior $P_X(x)$ as long as the support of $P_X(x)$ is the whole $\mathbb{R}^r$. As $\beta, N \to \infty$ the details of $P_X(x)$ will be washed away. One can write the state evolution equations (4.202-4.204)

$$M_x^{t+1} = \beta\Sigma_x^{t+1} \frac{M_x^t \Sigma_0}{\widehat{\Delta}} , \tag{4.283}$$

$$Q_x^{t+1} = \beta\Sigma_x^{t+1} \left[\frac{M_x^t \Sigma_0 {M_x^t}^\top}{\widehat{\Delta}^2} + \frac{Q_x^t}{\widetilde{\Delta}}\right] \beta\Sigma_x^{t+1} , \tag{4.284}$$

$$\beta\Sigma_x^{t+1} = \left[\frac{1}{\beta} + Q_x^t\left(\frac{1}{\widetilde{\Delta}} - \overline{R}\right) - \Sigma_x^t\left(\frac{(\beta-1)}{\widetilde{\Delta}} + \overline{R}\right)\right]^{-1} , \tag{4.285}$$

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

$$M_x^{t+1} = \Sigma'^{t+1} \frac{M_x^t \Sigma_0}{\widehat{\Delta}} , \tag{4.286}$$

$$Q_x^{t+1} = \Sigma_x'^{t+1} \left[\frac{M_x^t \Sigma_0 {M_x^t}^\top}{\widehat{\Delta}^2} + \frac{Q_x^t}{\widetilde{\Delta}}\right] \Sigma_x'^{t+1} , \tag{4.287}$$

$$\Sigma_x'^{t+1} = \left[Q_x^t\left(\frac{1}{\widetilde{\Delta}} - \overline{R}\right) - \frac{\Sigma_x'^t}{\widetilde{\Delta}}\right]^{-1} , \tag{4.288}$$

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

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

Following the derivation of the state evolution, eq. (4.210), we get that at the fixed point of the state evolution the spectral estimator $\hat{x}_i$ is distributed according to

$$\hat{x_i} = \Sigma'_x \left[\frac{M_x}{\widehat{\Delta}} x_i^0 + \sqrt{\frac{Q_x}{\widetilde{\Delta}}} W_i \right] , \tag{4.289}$$

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

MSE achieved by the spectral methods

When one uses spectral methods to solve a low rank matrix estimation problem one computes the r leading eigenvalues of the corresponding matrix and then one is left with the problem of what to do with the eigenvectors. Depending on what problem one tries to solve one can for instance cluster the eigenvectors using the k-means algorithm. A more systematic way is the following : We know by (4.289) that the elements $\hat{x}_i$ of the eigenvectors can be written as a random variable distributed as

$$\hat{x}_i = \hat{M} x_i^0 + \sqrt{\hat{Q}} W_i \,, \tag{4.290}$$

with $\hat{M} = \Sigma'_x M_x/\widehat{\Delta}$, and $\hat{Q} = Q_x(\Sigma'_x)^2/\widetilde{\Delta}$, where M_x, Q_x and Σ'_x are fixed points of the state evolution equations (4.286-4.288). This formula allows us to approach the problem as a low-dimensional Bayesian denoising problem. Writing

$$P(\hat{x}_i, x_i^0) = P(\hat{x}_i|x_i^0)P_{X_0}(x_i^0) = \frac{P_{X_0}(x_i^0)}{\sqrt{\mathrm{Det}\left(2\pi\hat{Q}\right)}} \exp\left(-\left(\hat{x}_i^\top - x_i^{0,\top}\hat{M}^\top\right)\hat{Q}^{-1}\left(\hat{x}_i - \hat{M}x_i^0\right)/2\right) , \tag{4.291}$$

one gets

$$P(x_i^0|\hat{x}_i) = \frac{P(\hat{x}_i|x_i^0)P_X(x_i^0)}{P(\hat{x}_i)} = \frac{P_{X_0}(x_i^0)}{P(\hat{x}_i)\sqrt{\mathrm{Det}\left(2\pi\hat{Q}\right)}} \exp\left(-\frac{\left(\hat{x}_i^\top - x_i^{0,\top}\hat{M}^\top\right)\hat{Q}^{-1}\left(\hat{x}_i - \hat{M}x_i^0\right)}{2}\right) . \tag{4.292}$$

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

$$\mathbb{E}_{P(x_i^0|\hat{x}_i)}\left[x_i^0\right] = f_{\mathrm{in}}^x\left(\hat{M}^\top\hat{Q}^{-1}\hat{M}, \hat{M}^\top\hat{Q}^{-1}\hat{x}_i\right) . \tag{4.293}$$

By combining (4.293) with (4.290) one gets that the mean-squared error achieved by the spectral method is given by

$$\mathrm{MSE}_{\mathrm{PCA}} = \mathbb{E}_{x_0,W}\left\{\left[x_0 - f_{\mathrm{in}}^x\left(\hat{M}^\top\hat{Q}^{-1}\hat{M}, \hat{M}^\top\hat{Q}^{-1}\hat{M}x_0 + \sqrt{\hat{M}^\top\hat{Q}^{-1}\hat{M}}W\right)\right]^2\right\}, \qquad (4.294)$$

where the W are once again Gaussian variables of zero mean and unit covariance, and the variable x^0 is distributed according to P_{X_0}. In the figures presented in subsequent sections we evaluate the performance of PCA via (4.294).

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

This section summarizes properties of problems for which the prior distribution P_x has zero mean. We will see that in those cases a particularly simple fixed point of both the Low-RAMP and its state evolution exists. We analyze the stability of this fixed points, and note that linearization around it leads to a spectral algorithm on the Fisher score matrix. As a result we observe equivalence between the corresponding spectral phase transition and a phase transition beyond which Low-RAMP performs better than a random guess based on the prior.

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

Linearization around the uniform fixed point- Matrix case ($XX^\top$, $UV^\top$ case)

From the definition of the thresholding function (4.86), it follows that that $\hat{x}_i = 0$, $\forall 1 \leq i \leq n$ is a fixed point of the self-averaged low-RAMP equations (4.96,4.97) or (4.169 4.170) whenever

$$\int \mathrm{d}x P_X(x)x = \langle x\rangle = 0, \quad \text{and} \quad \overline{R} = 0\,. \qquad (4.295)$$

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

In previous work on inference and message passing algorithms [KMM$^+$13, ZK16] we learned that when a uniform fixed point of the message passing update exists it is instrumental to expand around it and investigate the spectral algorithm to which such a procedure leads. We follow this strategy here and expand the Low-RAMP equations around the uniform fixed point. In the linear order in $\hat{x}$, the term A_X^t is negligible, from the definition of the thresholding

function (4.86) one gets

$$\hat{X}^{t+1} = \left(\frac{S}{\sqrt{N}} \hat{X}^t - \hat{X}^{t-1} \frac{\Sigma_x}{\widetilde{\Delta}} \right) \langle x x^\top \rangle + o\left(\|X^t\|_2 \right) \tag{4.296}$$

$$\hat{x}_i^{t+1} = \Sigma_x \frac{\sqrt{(p-1)!}}{N^{\frac{p-1}{2}}} \sum_{1 \le k_2 < \cdots < k_p \le N} S_{k_2 \cdots k_p} \hat{x}_{k_2}^t \circ \cdots \circ \hat{x}_{k_p}^t + o\left(\|X^t\|_2^{p-1} \right) . \tag{4.297}$$

This is an expansion of the AMP equations for the $XX^\top$ AMP equations and symmetric tensor ($p > 2$) Where Σ_x is the average value of the variance of the estimators, as defined in (4.190), at the uniform fixed point. In the Bayes optimal setting we remind from equation (4.255) that we moreover have $\Sigma_x = \langle x x^\top \rangle$.

$XX^\top$, $UV^\top$ case (p=2) : If we consider (4.296) as a fixed point equation for $\hat{X}$ we see that columns of $\hat{X}$ are related to the eigenvectors of the Fisher score matrix S. Expanding around the uniform fixed point the Low-RAMP equations thus yields a spectral algorithm that is essentially PCA applied to the matrix S (not the original dataset Y).

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

X^p tensor case ($p > 2$) : From equation (4.297)we see that the trivial fixed point is always stable whatever the value of the matrix S might be.

Example of the spectral decomposition of the Fisher score matrix ($XX^\top$ case)

Spectral method always come to mind when thinking about estimation of low-rank matrices. Analysis of the linearized Low-RAMP suggests that in cases where we have some guess about the form of the output channel $P_{\text{out}}(Y|w)$ then the optimal spectral algorithm should not be ran on the data matrix Y_{ij} but instead on the Fisher score matrix S_{ij} defined by (4.18). This was derived in [LKZ15a] and further studied in [PWBM16b]. In this section we give an example of a case where the spectrum of Y_{ij} does not carry any information for some region of parameter, but the one of S_{ij} does.

Consider as an example the output channel to be

$$P_{\text{out}}(Y|w) = \frac{1}{2} \exp\left(-|Y - w| \right) . \tag{4.298}$$

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

$$S_{ij} = \text{Sign}(Y_{ij}) . \tag{4.299}$$

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

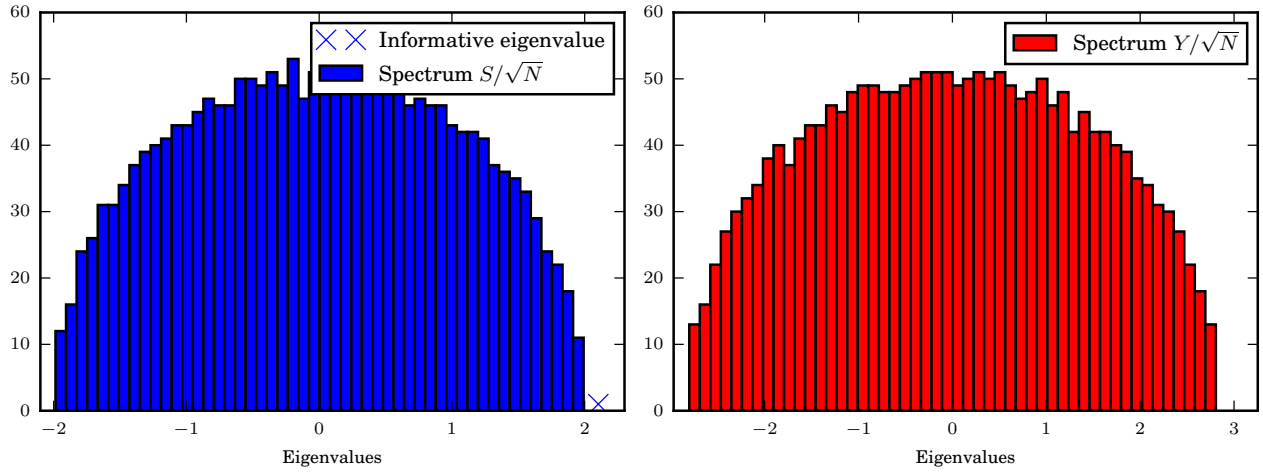


FIGURE 4.3 – Spectrum of the Fisher score matrix $S/\sqrt{N}$ and of the data matrix $Y/\sqrt{N}$ for the same instance of a problem with exponential output noise in the rank-one symmetric $XX^\top$ case. The planted configuration is generated from a Gaussian of zero mean and variance 1.4. We see that an eigenvalue is out of the bulk for S but not for Y. The data were generated on a system of size $N = 2000$.

Now let us look at the spectrum of both Y and S in Fig. 4.3. We plot the spectrum of S and Y for $\sigma = 1.4$. For this value of variance we see that an eigenvalue associated with an eigenvector that carries information about the planted configuration gets out of the bulk of S but not of Y. Even though some information on the signal was encoded into Y in that specific case one had to take the absolute value of Y to be able to recover an informative eigenvalue.

This situation can be quantified using the spectral analysis of section 4.4.1 applied to two different noise channels $g_1(Y, w) = -\beta(\,\mathrm{sign}(Y) - w)^2/2$ and $g_2(Y, w) = -\beta(Y - w)^2/4$ and taking the limit $\beta \to \infty$ as in (4.282). For these noise channels a theoretical analysis of the top eigenvectors of S and Y can be performed using theory presented in section 4.4.1. Taking square of (4.286) and dividing by (4.287) one can show that the state evolution equation describing the overlap of the top positive eigenvectors of Y and S is given by the only stable fixed point of the following update equation

$$\frac{(m_x^{t+1})^2}{q_x^{t+1}} = \frac{\frac{{m_x^t}^2}{q_x^t}\frac{\sigma^2\widetilde{\Delta}}{\widehat{\Delta}^2}}{1 + \frac{{m_x^t}^2}{q_x^t}\frac{\sigma\widetilde{\Delta}}{\widehat{\Delta}^2}}\,, \tag{4.300}$$

$$\widehat{\Delta} = \widetilde{\Delta} = 1 : \text{for } g_1(Y, w)\,, \tag{4.301}$$

$$\widehat{\Delta} = \widetilde{\Delta} = 2 : \text{for } g_2(Y, w)\,, \tag{4.302}$$

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

$$\sigma^2 \geq \frac{\widehat{\Delta}^2}{\widetilde{\Delta}}\,. \tag{4.303}$$

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

planted solution.

Stability of the uniform fixed point in Bayes-optimal setting

In this section we restrict for simplicity to the Bayes optimal setting defined by eq. (2.8). As we derived in section 4.3 the evolution of the Low-RAMP algorithm can be tracked using the state evolution equations. In the Bayes optimal case we have $\overline{R} = 0$ therefore the (sufficient) condition for the existence of the uniform fixed point is to have prior P_X (or both P_U and P_V) of zero mean. The existence of a uniform fixed point of the Low-RAMP algorithm translates into the existence of a fixed point of the state evolution with $M_x^t = 0$ for the symmetric matrix and tensor case (or $M_u^t = M_v^t = 0$ for the bipartite case).

The stability of this fixed point is analyzed by expanding in linear order the state evolution equation (4.256) around the uniform fixed point, taking into account the definition of the thresholding function (4.86). For the $XX^\top$ case this gives

$$M_x^{t+1} = \frac{\Sigma_x M_x^t \Sigma_x}{\Delta} + O(\|M_x^t\|_2^2)\,, \tag{4.304}$$

where Σ_x in the Bayes-optimal case is the covariance matrix of the signal (and prior) distribution as given by eq. (4.255). Calling $\lambda^x_{\max}$ the largest eigenvalue of the covariance of the distribution of the signal-elements, Σ_x, we obtain a simple criterion for the stability of the uniform fixed point

$$\begin{cases} \Delta_c = (\lambda^x_{\max})^2 < \Delta \Rightarrow \text{stable} \\ \Delta < \Delta_c = (\lambda^x_{\max})^2 \Rightarrow \text{unstable} \end{cases}. \tag{4.305}$$

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

$$\begin{cases} \Delta_c = \langle x_0^2 \rangle^2 < \Delta \Rightarrow \text{stable} \\ \Delta < \Delta_c = \langle x_0^2 \rangle^2 \Rightarrow \text{unstable} \end{cases}. \tag{4.306}$$

Interestingly, the criteria (4.305) is the same as the criteria for the spectral phase transition of the Fisher score matrix S. When the uniform fixed point is not stable the Fisher score matrix has an eigenvalue going out of the bulk [BBAP05, HR04]. We stress here that this analysis is particular to signals of zero mean. If the mean is non-zero the spectral threshold does not change, but the Bayes optimal performance gets better and hence superior to PCA.

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

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

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

leads to

$$M_u^t = \alpha \frac{\Sigma_u M_v^t \Sigma_u}{\Delta} + O(\|M_v^t\|_2^2)\,, \tag{4.307}$$

$$M_v^{t+1} = \frac{\Sigma_v M_u^t \Sigma_v}{\Delta} + O(\|M_u^t\|_2^2)\,, \tag{4.308}$$

where in the Bayes-optimal case the Σ_u and Σ_v are simply the covariances of the prior distribution P_U and P_V, i.e. $\Sigma_u = \langle uu^\top \rangle$ and $\Sigma_v = \langle vv^\top \rangle$. By replacing (4.308) in (4.307) one gets

$$M_u^{t+1} = \left(\frac{\sqrt{\alpha}\Sigma_u\Sigma_v}{\Delta}\right) M_u^t \left(\frac{\sqrt{\alpha}\Sigma_u\Sigma_v}{\Delta}\right)^\top . \tag{4.309}$$

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

$$\begin{cases} \Delta_c = \sqrt{\alpha}\lambda_{\max}^{uv} < \Delta \Rightarrow \text{Stable} \\ \Delta < \Delta_c = \sqrt{\alpha}\lambda_{\max}^{uv} \Rightarrow \text{Unstable} \end{cases} . \tag{4.310}$$

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

For the tensor case the expansion around the fixed point gives.

$$M_x^{t+1} = \frac{\Sigma_x (M_x^t)^{\circ^{p-1}} \Sigma_x}{\Delta} + O(\|M_x^t\|_2^{2p-2})\,, \tag{4.311}$$

As soon as $p > 2$ the trivial fixed point will be stable whatever the value of Δ since the expansion will be zero to first order. In term of algorithm this means that the fixed point positively correlated with the hidden solution will be hidden behind a Free-Energy barrier. Which will make AMP fail in reconstructing the hidden signal even when it is Information theoretically possible.

4.4.3 Symmetry of the system : difference between matrix and tensor factorization.

It often happens (at least in the system studied in that thesis) that the system exhibit a symmetry. For instance a classical spin systems with zero fields has a ± 1 or $\mathbb{Z}_2$ symmetry meaning that.

$$P(x_1, x_2, \cdots, x_N) = \frac{1}{Z} \exp\left(\sum_{1 \le i < j \le N} x_i J_{ij} x_j\right) = P(-x_1, -x_2, \cdots, -x_N) \tag{4.312}$$

When dealing with vectorial spins symmetries we define the symmetry group of the system as

the set of matrices $R \in \mathbb{R}^{r\times r}$ as

$$R \in \text{Sym} \iff R \in \mathbb{R}^{r\times r} \text{ and } \forall x_1, x_2, \cdots, x_N \in \mathbb{R}^r, P(x_1, \cdots, x_N) = P(Rx_1, \cdots, Rx_N) \tag{4.313}$$

and in the $UV^\top$ case.

$$(A, B) \in \text{Sym} \iff A, B \in \mathbb{R}^{r\times r} \text{ and } \forall u_1, u_2, \cdots, u_N \in \mathbb{R}^r, \forall v_1, v_2, \cdots, v_M \in \mathbb{R}^r \\ P(u_1, \cdots, u_N, v_1, \cdots, v_M) = P(Au_1, \cdots, Au_N, Bv_1, \cdots, Bv_M) \tag{4.314}$$

If there is no symmetry in the system then $\text{Sym} = \{I_r\}$.

In term of state evolution this means if M_x, Q_x, Σ_x is a fixed point of the SE equations (4.202-4.204) then $\forall R \in \text{Sym}$ such that

$$M_x \longrightarrow RM_x,\ Q_x \longrightarrow RQ_xR^\top,\ \Sigma_x \longrightarrow R\Sigma_xR^\top \tag{4.315}$$

is also a fixed point of the SE equations. And in the $UV^\top$ case a similar transformation of the SE equations (4.226- 4.231) is given by, $\forall (A, B) \in \text{Sym}$

$$M_u \longrightarrow AM_u,\ Q_u \longrightarrow AQ_uA^\top,\ \Sigma_u \longrightarrow A\Sigma_uA^\top \tag{4.316}$$
$$M_v \longrightarrow BM_v,\ Q_v \longrightarrow BQ_vB^\top,\ \Sigma_v \longrightarrow B\Sigma_vB^\top \tag{4.317}$$

The Matrix Case Because of the structure of the problem of matrix factorization it is easy for these problem to exhibit a rotational symmetry. It is straightforward to see that

$$\forall R \in O(r) \tag{4.318}$$
$$XX^\top = XR(XR)^\top, \tag{4.319}$$

$$\forall A \in GL_r(\mathbb{R}), B = {A^{-1}}^\top \tag{4.320}$$
$$UV^\top = (UA)(VB)^\top, \tag{4.321}$$

All the matrices that satisfy (4.318) and (4.321) form a continuum (respectively $O(N)$ and $GL_n(\mathbb{R})$). This means that the matrix factorization problem can exhibit a rotation symmetry if the prior distribution $P(X)$ or $P_U(u), P_V(v)$ does not break the rotational symmetry.

This is an usual problem encountered when performing PCA. The minima of the PCA problem (4.24) are defined up to a rotation. This means that if $X \in \mathbb{R}^{N\times r}$ is such that $\left\|Y - \frac{1}{\sqrt{N}}XX^\top\right\|_2^2$ is minimal. Then for any rotation matrix R then XR will also be a minima.

The Tensor Case In the case of tensor factorization ($p \geq 3$) no such continuous symmetry exists. Let us suppose a matrix $R \in \mathbb{R}^{r\times r}$ such that $\forall X \in \mathbb{R}^{N\times r}$

$$\forall X \in \mathbb{R}^{N\times r}, \overbrace{X \otimes X \otimes \cdots \otimes X}^{p} = (XR) \otimes (XR) \otimes \cdots \otimes (XR) \tag{4.322}$$

it is easy to prove that this can only be the case if

$$\forall 1 \leq i_1, \cdots, i_p \leq r : \sum_{l=1\cdots r} R_{i_1 l} R_{i_2 l} \cdots R_{i_p l} = \delta^{i_2}_{i_1} \delta^{i_3}_{i_2} \cdots \delta^{i_p}_{i_{p-1}} \tag{4.323}$$

Only equations for ordered $i_1, \cdots, i_p$ are independent.

$$\forall 1 \leq i_1 \leq \cdots \leq i_p \leq r : \sum_{l=1\cdots r} R_{i_1 l} R_{i_2 l} \cdots R_{i_p l} = \delta^{i_2}_{i_1} \delta^{i_3}_{i_2} \cdots \delta^{i_p}_{i_{p-1}} \tag{4.324}$$

This specify $\binom{p+r-1}{r} \sim \frac{r^p}{p!}$ equations while there are only r^2 unknowns variables. As soon as $p \geq 3$ and $r \geq 2$ one ends up with more equations than unknown which makes the problem of continuous symmetry group over-constrained. This means that for $p \geq 3$ and $r \geq 2$ there can be no continuous symmetry group. And any solution of equations (4.324) will be isolated. We conjecture that for $p \geq 3$ and $r \geq 2$ The only solution to (4.324) are permutations matrix if p is odd and permutation matrices where every 1 might be replaced by a -1 if p is even.

4.4.4 Multiple stable fixed points : First order phase transitions

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

We encounter two types of situations with multiple stable fixed points

— **Equivalent fixed points due to symmetry :** The less interesting type of multiple fixed points arises when there is an underlying symmetry in the definition of the problem, then both the state evolution and and Low-RAMP equations have multiple fixed points equivalent under the symmetry. All these fixed points have the same Bethe and replica free energy. One example of such a symmetry is a Gaussian prior of zero mean and isotropic covariance matrix. Then there is a global rotational symmetry present. Another example of a symmetry is given by the community detection problem defined in section 4.1.4. In the case of r symmetric equally sized groups with connectivity matrix (4.67) there is a permutation symmetry between communities so that any fixed of the state evolution or Low-RAMP equations exists in $r!$ versions.

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

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

Typical first order phase transition : Algorithmic interpretation

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

In Fig. 4.4 we plot all the fixed points of the state evolution equation and the corresponding value of the replica free energy (4.258) as a function of Δ/ρ^2. The equations for the state evolution specific to the spiked Rademacher-Bernoulli model are given by (4.344). For this model the uniform fixed point exists and is stable down to $\Delta_c = \rho^2$, eq. (4.305). The numerically stable fixed points are drawn in blue, the unstable ones in red. We focus on the interval of Δ where more than one stable fixed point of the state evolution (4.256) exists. We use the example of the spiked Rademacher-Bernoulli model for the purpose of being specific in figure 4.4. The definitions and properties defined in the rest of this section are generic and apply to all the settings considered in this thesis, not only to the spiked Rademacher-Bernoulli model.

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

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

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

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

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

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

— Δ_{IT} called the *information theoretic phase transition*, is the value of Δ at which the two fixed points $M_{\text{Uninformative}} \neq M_{\text{Informative}}$ exist and have the same replica free energy (4.258).

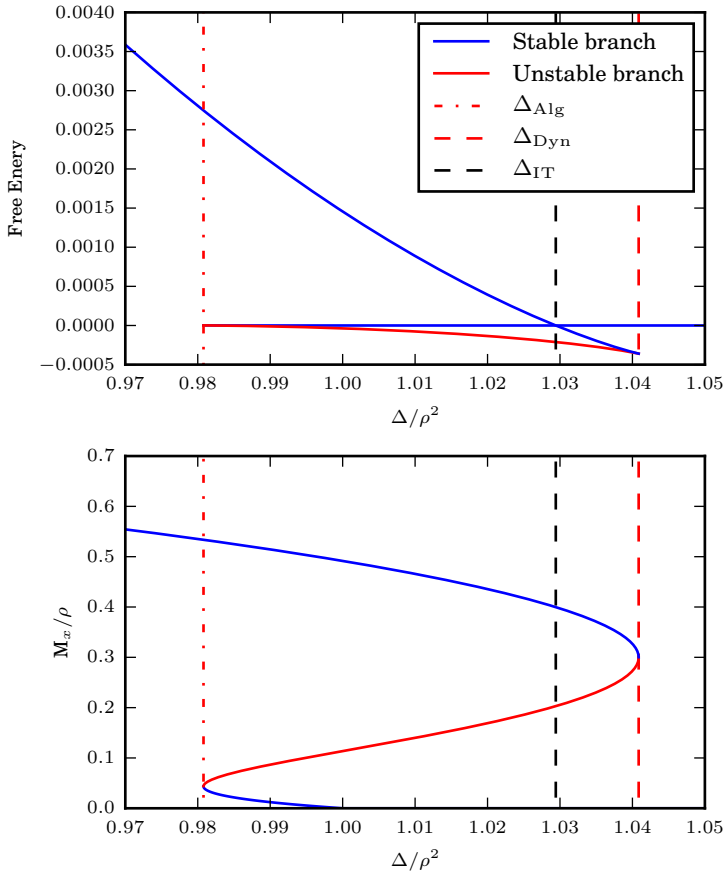


FIGURE 4.4 – In the lower panel, we plot as a function of Δ/ρ^2 all the fixed points M_x/ρ of state evolution equations (4.256) for the spiked Rademacher-Bernoulli model with fraction of non-zero being $\rho = 0.08$. Numerically stable fixed points are in blue, unstable in red. We remind that the order parameter M_x is related to the mean-squared error as $\text{MSE}_X = \text{Tr}\left[\langle x_0 x_0^\top \rangle - M_x\right]$. In the upper panel, we plot the replica free energy (4.258) corresponding to these fixed points, again as a function of Δ/ρ^2. When there are multiple stable fixed points to the SE equations the one that corresponds to performance of the Bayes optimal estimation is the one with the largest free energy (we remind that with respect to the most common physics notation we defined the free energy as the positive logarithm of the partition function). The Δ for which these two branches cross in free-energy is called the information theoretic phase transition, $\Delta_{\rm IT}$. The two spinodal transitions $\Delta_{\rm Alg}$ and $\Delta_{\rm Dyn}$ are where the lower MSE an higher MSE stable fixed point disappears. The $\Delta_c/\rho^2 = 1$ corresponds to the spectral transition at which the uniform fixed point become unstable.

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

In Fig. 4.4 these three transition are marked by vertical dashed lines, the information theoretic transition in black and the two spinodal transition in red.

We recall that for the priors of zero mean, where the uniform fixed points discussed in section 4.4.2 exists, the stability point of the uniform fixed point Δ_c (4.305) is in general unrelated to the Δ_{Alg}, Δ_{IT} and Δ_{Dyn}. In the example presented in Fig. 4.4 of spiked Rademacher-Bernoulli model at $\rho = 0.08$ we have $\Delta_{\text{Alg}} < \Delta_c < \Delta_{\text{IT}}$. In general the position of Δ_c with respect to Δ_{Alg}, Δ_{IT} and Δ_{Dyn} can be arbitrary, in the following sections we will observe several examples. A notable situation is when $\Delta_c = \Delta_{\text{Alg}}$, cases where this happen are discussed in section 4.4.4. It should be noted that this is the case in the community detection problem, and since this is a well known and studied example it is sometimes presented in the literature as the generic case. But from the numerous examples presented in this thesis we see that cases where $\Delta_c \neq \Delta_{\text{Alg}}$ are also very common.

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

— **Phase where Low-RAMP is asymptotically Bayes-optimal :** For $\Delta \geq \Delta_{\text{IT}}$ and for $\Delta \leq \Delta_{\text{Alg}}$ the Low-RAMP algorithm in the limit of large system sizes gives the information theoretically optimal performance. This is either because the fixed point it reaches is unique (up to symmetries) or because it is the one with larger free energy.

— **The hard phase :** For $\Delta_{\text{Alg}} < \Delta < \Delta_{\text{IT}}$ the estimation error achieved by the Low-RAMP algorithm is strictly larger that the lowest information-theoretically achievable error. On the other hand, and in line with other statistical physics works on inference problems in the Bayes optimal setting, we conjecture that in this region no other polynomial algorithm that would achieve better error than Low-RAMP exists. This conjecture could be slightly modified by the fact that the branch of stable fixed points that do not correspond to the MMSE could present aspects of replica symmetry breaking which could modify its position. Investigation of this is left for future work.

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

— (a) Those that are rigorous, known from existing literature that is not related to statistical physics considerations. An example is given by the performance of the spectral methods that is better that random guessing for $\Delta < \Delta_c$ [BBAP05].

— (b) A second part regroups the results directly following from the analysis of the Bayes-optimal Low-RAMP and the MMSE that are not presented rigorously in this thesis, but were made rigorous in a series of recent works [KXZ16, BDM$^+$16, ML16, Mio17b]. Most of these results are proven and although some cases are still missing a rigorous proof, it is safe to assume that it is a question of time that researchers will fill the corresponding gaps and weaken the corresponding assumptions. For instance the state evolution of Low-RAMP [RF12, JM13, DM14a], its Bayes-Optimality in the easy phase (at least for problem where the paramagnetic fixed point is not symmetric) and the value of the information theoretically optimal MMSE [KXZ16, DAM16, BDM$^+$16, ML16, Mio17b] are all proven rigorously.

— (c) The third kind of claims are purely conjectures. For instance, the claim that among all polynomial algorithms the performance of Low-RAMP cannot be improved (so that the hard phase is indeed hard). Of course proving this in full generality would imply that

P≠NP and so we cannot expect that such a proof would be easy to find. At the same time from a broad perspective of understanding average computational complexity this is the most intriguing claim and is worth detailed investigation and constant aim to find a counter-example.

Note about computation of the first order phase transitions

In this section we discuss how to solve efficiently the SE equations in the $XX^\top$ Bayes optimal case for rank one. We notice that for a given prior distribution P_X the only way the symmetric Bayes optimal state evolution, eq. (4.256), depends on the noise parameter Δ is via the ratio m^t/Δ. One can write the SE equations in the form (4.257)

$$m^{t+1} = f^{\rm SE}_{P_X}\left(\frac{m^t}{\Delta}\right) . \tag{4.325}$$

Let us further define fixed points of (4.325) in a parametric way

$$\Delta = \frac{f^{\rm SE}_{P_X}(x)}{x} , \tag{4.326}$$

$$m = f^{\rm SE}_{P_X}(x) . \tag{4.327}$$

To get a fixed point (m, Δ) of (4.325) we choose a value of x and compute $(f(x), f(x)/x)$. We observe from the form of the state evolution (rank-one symmetric Bayes optimal case) that $f(x)$ is a non-decreasing function of x. We further observe that (m, Δ) is a stable fixed points if and only if $\partial_x \Delta(x) < 0$. The two spinodal thresholds $\Delta_{\rm Alg}$ and $\Delta_{\rm Dyn}$ are defined by loss of existence of corresponding stable fixed points and they can hence be computed as

$$\Delta_{\rm Alg}, \Delta_{\rm Dyn} \in \left\{ \Delta(x), x \in \mathbb{R}^+, \frac{\partial \Delta(x)}{\partial x} = 0 \right\} . \tag{4.328}$$

The information theoretic transition $\Delta_{\rm IT}$ relies on computation of the replica free energy (4.258). Using (4.259) one gets

$$\frac{\partial \phi(m, \Delta)}{\partial m} = \frac{1}{2\Delta}\left(f^{\rm SE}_{P_X}\left(\frac{m}{\Delta}\right) - m \right) \tag{4.329}$$

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

$$\phi(m(x), \Delta(x)) - \phi(0, \Delta(x)) = \frac{1}{2}\left[\int_0^x {\rm d}u \, f^{\rm SE}_{P_X}(u) - \frac{x f^{\rm SE}(x)}{2} \right] . \tag{4.330}$$

The $x_{\rm IT}$ for which (4.330) is zero then gives the information theoretic phase transition $\Delta_{\rm IT} = f(x_{\rm IT})/x_{\rm IT}$.

Computation in the tensor X^p case. The same computation can be made in the tensor of case. Similarly as in the $XX^\top$ case. The SE equations in the Bayes optimal setting can be written as.

$$m^{t+1} = f_{P_X}^{\mathrm{SE}}\left(\frac{m^{tp-1}}{\Delta}\right) \tag{4.331}$$

The following parameter are a fixed point of this equation.

$$\Delta = \frac{f_{P_X}^{\mathrm{SE}}(x)^{p-1}}{x}\,, \tag{4.332}$$

$$m = f_{P_X}^{\mathrm{SE}}(x)\,. \tag{4.333}$$

Δ_{Alg} and Δ_{Dyn} are once again defined by equation (4.328). The Δ_{IT} transition can be found using the following equation.

$$\phi(m(x), \Delta(x)) - \phi(0, \Delta(x)) = \frac{1}{2}\left[\int_0^x \mathrm{d}u\, f_{P_X}^{\mathrm{SE}}(u) - (p-1)\frac{x f^{\mathrm{SE}}(x)}{p}\right]. \tag{4.334}$$

The x_{IT} for which (4.334) is zero then gives the information theoretic phase transition $\Delta_{\mathrm{IT}} = f(x_{\mathrm{IT}})^{p-1}/x_{\mathrm{IT}}$.

Sufficient criteria for existence of the hard phase : $XX^\top$ case

This section is specific to the Bayes-optimal cases when the prior P_X has a zero mean and hence the uniform fixed point of the Low-RAMP and the state evolution exists. In section 4.4.2 we derived that the uniform fixed point is stable at $\Delta > \Delta_c$ and unstable for $\Delta < \Delta_c$. It follows from the theory of bifurcations that the critical point where a fixed-point changes stability must be associated with an onset of another close-by fixed point. In general there are two possibilities

- **2nd order bifurcation.** If the fixed point close to the uniform fixed point departs in the direction of smaller $\Delta < \Delta_c$, where the uniform fixed point is unstable, then this close-by fixed point is stable. This case corresponds to Fig. 4.4. Behavior in the vicinity of the uniform fixed point then does not let us distinguish between (a) existence of a first order phase transition at lower Δ (as in Fig. 4.4), or (b) continuity on the MMSE down to $\Delta = 0$ with no algorithmically hard phase existing in that case.
- **1st order bifurcation.** If the fixed point close to the uniform fixed point departs in the direction of larger $\Delta > \Delta_c$, where the uniform fixed point is stable, then this close-by fixed point is unstable. In that case, the fixed point that is stable from $\Delta < \Delta_c$ is not close to the uniform fixed point and this case forces existence of a first order phase transition with $\Delta_c = \Delta_{\mathrm{Alg}}$.

Expansion of the state evolution (4.256) around the uniform fixed point gives us a closed form criteria to distinguish whether the phase transition happening at Δ_c is a 1st or 2nd order bifurcation. In case of a 2nd order bifurcation, this expansion allows us to compute the mean squared error obtained with Low-RAMP close to Δ_c.

For specificity we consider the rank-one, $r = 1$ case, and expand eq. (4.256) to 2nd order to get

$$m^{t+1} = m^t \frac{\langle x_0^2\rangle^2}{\Delta} - \frac{(m^t)^2}{\Delta^2}\left(\langle x_0^2\rangle^3 - \frac{\langle x_0^3\rangle^2}{2}\right) + O\left((m^t)^3\right), \tag{4.335}$$

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

$$\forall i, j \quad \frac{\partial^{i+j} f^x_{\text{in}}}{\partial A^i \partial B^j}(A = 0, B = 0) \tag{4.336}$$

are linked to moments of the density probability P_X.

The stability criteria $\Delta_c = \langle x_0^2 \rangle^2$ appears once again as in section 4.4.2. Below $\Delta < \Delta_c$ the uniform fixed point $m = 0$ is unstable and m^t will converge towards another fixed point different from $m = 0$. For $\Delta > \Delta_c = \langle x_0^2 \rangle^2$ the uniform fixed point $m = 0$ is stable. Using expansion (4.335) near $\Delta \simeq \Delta_c = \langle x_0^2 \rangle^2$ we can write what is the other fixed point next to $m_{\text{uniform}} = 0$, we get

$$m_{\text{close-by}} = \frac{\Delta_c(\Delta_c - \Delta)}{\Delta_c^{3/2} - \frac{\langle x_0^3 \rangle^2}{2}} + O\left((\Delta - \Delta_c)^2\right) . \tag{4.337}$$

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

— If $\langle x_0^3 \rangle^2 < 2\langle x_0^2 \rangle^3$, eq. (4.337) is a stable fixed point in the region $\Delta < \Delta_c$. Eq. (4.337) then gives the expansion of this fixed point. This situation corresponds to Fig. 4.4 where the Rademacher-Bernoulli prior has zero 3rd moment. This is the 2nd order bifurcation at Δ_c.

— $\langle x_0^3 \rangle^2 > 2\langle x_0^2 \rangle^3$ eq. (4.337) is an unstable (and hence irrelevant) fixed point in the region $\Delta > \Delta_c$. But also in this case there must be a stable fixed point for $\Delta < \Delta_c$, but this fixed point cannot have a small values of m. The only way a stable fixed point can appear in this case is by a discontinuous (1st order) transition at Δ_c. This is the 1st order bifurcation at Δ_c.

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

$$\begin{cases} \langle x_0 \rangle = 0 \\ \langle x_0^3 \rangle^2 > 2\langle x_0^2 \rangle^3 \end{cases} . \tag{4.338}$$

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

On the other hand when the criteria (4.338) is not met, then for $\Delta < \Delta_c$ the MSE achieved by the Low-RAMP algorithm is in first approximation equal to

$$\text{MSE}(\Delta) = \langle x_0^2 \rangle^2 - \frac{\Delta_c(\Delta_c - \Delta)}{\Delta_c^{3/2} - \frac{\langle x_0^3 \rangle^2}{2}} + O\left((\Delta - \Delta_c)^2\right) . \tag{4.339}$$

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

Interestingly spectral method also give an MSE linear in $\Delta - \Delta_c$. As derived in section 4.4.1 the MSE one achieves using the eigenvectors of matrix S is

$$\text{MSE}_{\text{Spectral}}(\Delta) = \langle x_0^2 \rangle - \frac{\Delta_c - \Delta}{\sqrt{\Delta_c}} . \tag{4.340}$$

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

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

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

4.5 Phase diagrams for Bayes-optimal low-rank matrix/tensor estimation

From now on we restrict our analysis to the Bayes optimal setting as defined in section 4.1.1, eq. (2.8). The motivation is to investigate performance of the Bayes-optimal and the Low-RAMP estimators for a set of benchmark problems. We investigate phase diagrams stemming from the state evolution equations and from the corresponding replica free energies summarized in section 4.3.6.

4.5.1 Examples of phase diagram

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

Spiked Bernoulli model

The spiked Bernoulli model is defined by prior (4.62) with density ρ of ones, and $1-\rho$ of zeros. This prior has a positive mean and consequently the state evolution does not have the uniform fixed point. This is a problem where one tries to recover a submatrix of size $\rho N \times \rho N$ with mean of elements equal to 1, in a $N \times N$ matrix of lower mean. Using the Bayes-optimal state evolution (4.202) one gets,

$$m^t = f^{\rm SE}_{\rm Bernoulli}\left(\frac{m^t}{\Delta}\right) , \tag{4.341}$$

$$f^{\rm SE}_{\rm Bernoulli}(x) = \rho \mathbb{E}_W \left[\frac{\rho}{\rho + (1-\rho)\exp\left(\frac{-x}{2} + W\sqrt{x}\right)}\right] , \tag{4.342}$$

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

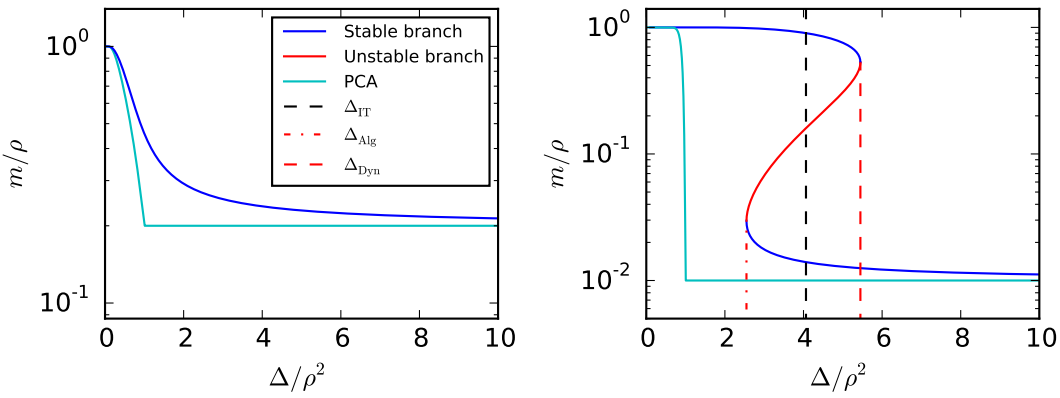


FIGURE 4.5 – Plots of the fixed points of the state evolution for the spiked Bernoulli model, MSE being given by MSE $= \rho - m$ (4.260). The performance of PCA as analyzed in section 4.4.1 is plotted for comparison. These two plots are made for $\rho = 0.2$ (left) and $\rho = 0.01$ (right).

Depending on the value of ρ there are two kinds of behavior of the fixed points of these equations as a function of the effective noise Δ. We plot the two cases in Fig. 4.5. For larger values of ρ there is a unique fixed point corresponding to the MMSE that is asymptotically achieved by the Low-RAMP algorithm, this is the regime in which the proof of [DM14a] applies. For small enough values of ρ we do observe a region of $\Delta_{\rm Alg} < \Delta < \Delta_{\rm Dyn}$ where there are 3 fixed points, two stable and one unstable. The replica free energy associated to the these fixed points crosses at $\Delta_{\rm IT}$ so that the higher fixed point in the relevant one at $\Delta < \Delta_{\rm IT}$, and the lower fixed point at $\Delta > \Delta_{\rm IT}$. These phase transitions were defined in section 4.4.4. In Fig. 4.6 we plot the phase transitions $\Delta_{\rm Alg}$, $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ as a function of ρ. The y-axes in the left panel is simply the effective noise parameter Δ, on the right panel the same data are plotted with Δ/ρ^2 on the y-axes. We observe that the $\Delta_{\rm Alg} =_{\rho\to 0} e\rho^2$, with e being the Euler number, this is the same asymptotic behavior as obtained previously in [Mon15] (Fig. 5).

Rademacher-Bernoulli and Gauss-Bernoulli

Next we analyze the spiked Rademacher-Bernoulli and Gauss-Bernoulli priors defined by (4.63) and (4.66). The first thing we notice is that both these distribution have zero mean and variance ρ. This means according to (4.4.2) that there is a uniform fixed point of the SE equations that is stable for $\Delta > \rho^2$. The skewness of both these distribution is 0, which means that at Δ_c there is no discontinuity of the MSE (4.4.4). The SE equations for these models are

$$m^{t+1} = f^{\rm SE}_{\rm Rademacher-Bernoulli}\left(\frac{m^t}{\Delta}\right), \tag{4.343}$$

$$f^{\rm SE}_{\rm Rademacher-Bernoulli}(x) = \rho\mathbb{E}_W\left[\tanh\left(x + W\sqrt{x}\right)\frac{\rho}{(1-\rho)\frac{\exp(x/2)}{\cosh\left(x+W\sqrt{x}\right)} + \rho}\right], \tag{4.344}$$

$$m^{t+1} = f^{\rm SE}_{\rm Gaussian-Bernoulli}\left(\frac{m^t}{\Delta}\right), \tag{4.345}$$

$$f^{\rm SE}_{\rm Gauss-Bernoulli}(x)/\rho = \frac{x}{1+x}\mathbb{E}_W\left[W^2\hat{\rho}(x, W\sqrt{x^2+x})\right], \tag{4.346}$$

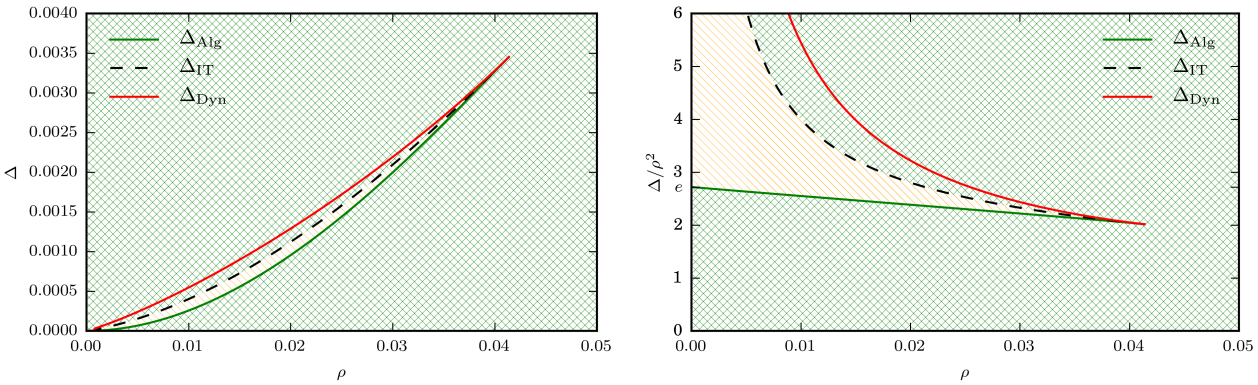


FIGURE 4.6 – The phase diagram of the spiked Bernoulli model (4.62) as a function of the density ρ and effective noise Δ (left) or Δ/ρ^2 (right). There is no phase transition in the system for $\rho > 0.04139$ and a 1st order phase transition for $\rho < 0.04139$. The lower green curve is the algorithmic spinodal $\Delta_{\rm Alg}$ curve, that converges to $\Delta_{\rm Alg} =_{\rho\to 0} e\rho^2$. The dashed black line is the information theoretic threshold $\Delta_{\rm IT}$. The upper red curve is the dynamical spinodal $\Delta_{\rm Dyn}$. The orange hashed zone is the hard region in which Low-RAMP does not reach the Bayes-optimal error. In the rest of the phase diagram (green hashed) the Low-RAMP provides in the large size limit the Bayes-optimal error. Note that this is exactly the same phase diagram as presented in [Mon15] (Fig. 5) for the problem of finding one dense subgraph, this is thanks to output channel universality and the fact that large degree sparse graphs have upon rescaling the same phase diagram as dense graphs.

where $\hat{\rho}$ is

$$\hat{\rho}(a, b^2) = \frac{\rho}{(1-\rho)\exp\left(\frac{-b^2}{2(1+a)}\right)(1+a)^{\frac{r}{2}} + \rho}\,. \tag{4.347}$$

Both these models have similar phase diagram.

We first illustrate in Fig. 4.7 the different types of phase transition that we observe for the spiked Rademacher-Bernoulli model as the density ρ is varied. We plot all the fixed points of equation (4.344) for several values of ρ as a function of the effective noise Δ/ρ^2. The four observed case are the following

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

Finally Fig. 4.9 presents the four thresholds $\Delta_{\rm Dyn}$, $\Delta_{\rm IT}$, $\Delta_{\rm Alg}$ and Δ_c as a function of the density ρ.

In Fig. 4.10 we present for completeness the comparison between the fixed points on the state evolution and the fixed points of the Low-RAMP algorithm for the Gauss-Bernoulli model,

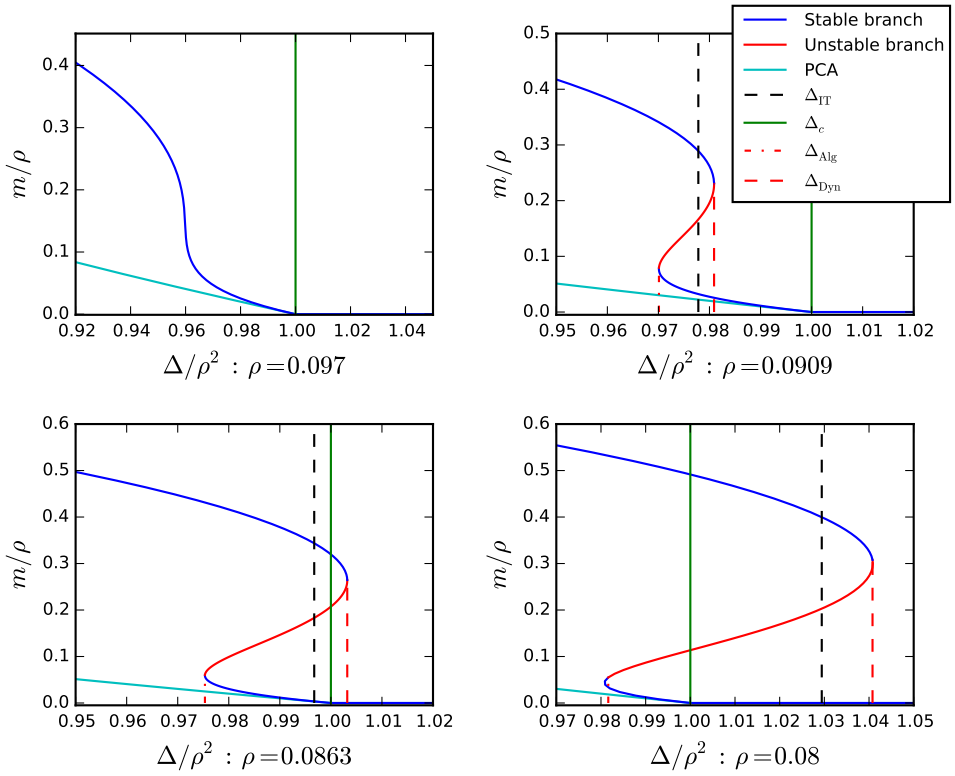


FIGURE 4.7 – We plot all the fixed points of the state evolution equations for the spiked Rademacher-Bernoulli model in the Bayes optimal setting as a function of Δ/ρ^2 for four representative values of ρ. The blue curves are stable fixed points, red are the unstable ones. MSE is given by (4.260). The vertical lines mark the two spinodal transition (dynamical $\Delta_{\rm Dyn}$ and algorithmic $\Delta_{\rm Alg}$, red dashed), the information theoretic transition $\Delta_{\rm IT}$. The vertical green full line mark the stability point of the uniform fixed point $\Delta_c = \rho^2$. We remind that the error MSE $= \rho - m$ achieved by the Bayes optimal estimator corresponds to the upper branch for $\Delta < \Delta_{\rm IT}$, and to the lower branch for $\Delta > \Delta_{\rm IT}$. Error achieved by the Low-RAMP algorithm always correspond to the lower branch (larger error). Note that in the three panels where multiple fixed points exists, the only element that changes is the position of the (spectral) stability threshold Δ_c with respect to the other thresholds.

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

We also compare to the MSE reached by the PCA spectral algorithm and from its analysis eq. (4.294). We see that whereas both Low-RAMP and PCA work better than random guesses below Δ_c, the MSE reached by Low-RAMP is considerably smaller.

Two balanced groups

The next example of phase diagram we present is for community detection with two balanced (i.e. one group is smaller ρN, but both have the same average degree) groups as defined in

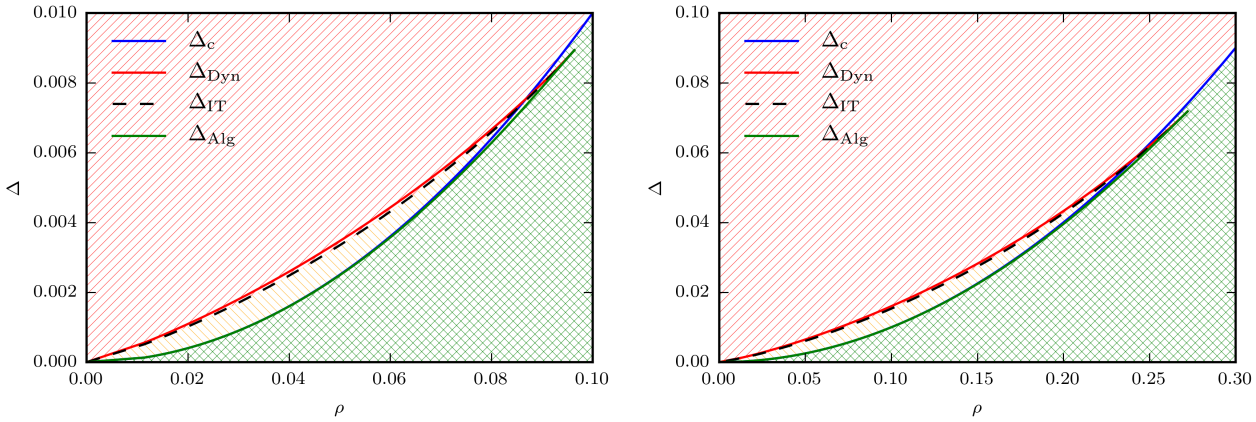


FIGURE 4.8 – Phase diagram of the spiked Rademacher-Bernoulli (left hand side panel) and spiked Gauss-Bernoulli (right hand side panel) models. We plot Δ as a function of ρ. The local stability threshold of the uniform fixed point $\Delta_c/\rho^2 = 1$ is in blue. The algorithmic spinodal $\Delta_{\rm Alg}$ (green), the dynamical spinodal $\Delta_{\rm Dyn}$ (red) and the information theoretic transition $\Delta_{\rm IT}$ (black dashed) all join into a tri-critical point located at $(\Delta_{\rm tri} = 0.008935, \Delta_{\rm tri}/\rho^2_{\rm tri} = 0.9612, \rho_{\rm tri} = 0.09641)$ for the Rademacher-Bernoulli model (left panel), and at $(\Delta_{\rm tri} = 0.07182, \Delta_{\rm tri}/\rho^2_{\rm tri} = 0.9693, \rho_{\rm tri} = 0.2722)$ for the Gauss-Bernoulli model (right panel). The hash materializes the different phases. The easy phase where the Low-RAMP algorithm is Bayes-optimal and achieves better error than random guessing is hashed in green crossed lines, the hard phase where Low-RAMP is sub-optimal is hashed in yellow \\, and the impossible phase where even the best achievable error is as bad as random guessing is hashed in red //.

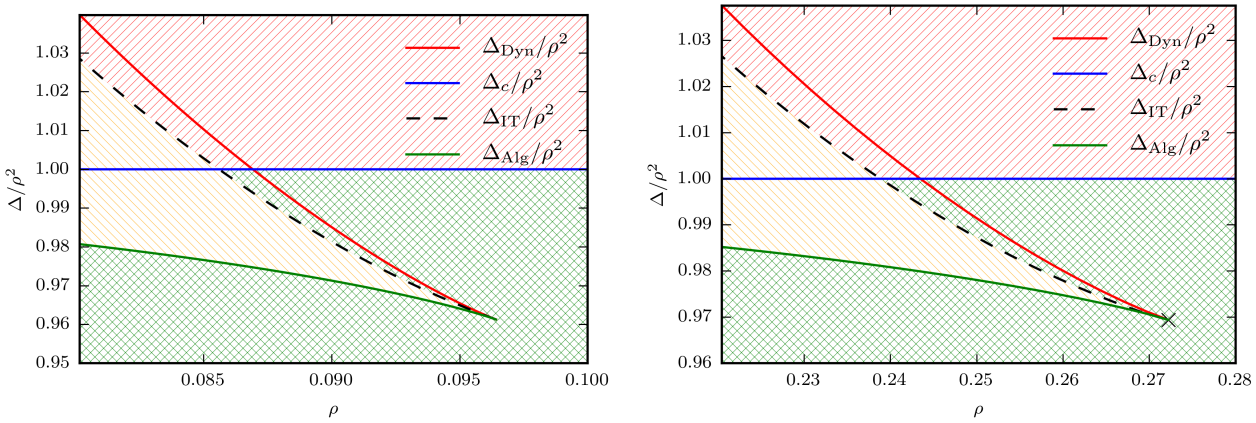


FIGURE 4.9 – The same plot as in Fig. 4.8 zoomed into the region of the tri-critical point with y-axes rescaled by ρ^2. The spiked Rademacher-Bernoulli mode on left hand side panel, the spiked Gauss-Bernoulli model on the right hand side.

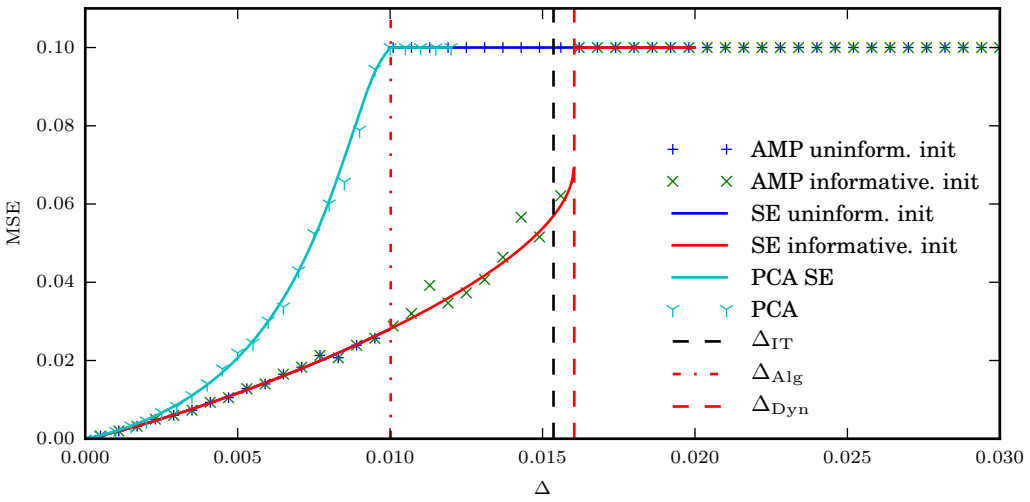


FIGURE 4.10 – Comparison between the state evolution and the fixed point of the Low-RAMP algorithm, for the spiked Gauss-Bernoulli model of sparse PCA with rank one and density $\rho = 0.1$. The phase transitions stemming from state evolution are $\Delta_{\rm Alg} \approx \Delta_{\rm c} = 0.01, \Delta_{\rm IT} = 0.0153, \Delta_{\rm Dyn} = 0.0161$. The points are the fixed points of the Low-RAMP algorithm run on one typical instance of the problem of size $N = 20000$. Blue pluses is the MSE reached from an uninformative initialization of the algorithm. Green crosses is the MSE reached from the informative initialization of the algorithm.

eqs. (4.75-4.76). This is an example of a system where the bifurcation at Δ_c is of a first order, with $\Delta_c = \Delta_{\rm Alg}$. In this problem the prior given by eq. (4.76), with $\langle x_0 \rangle = 0$, $\langle (x_0)^2 \rangle = 1$. The output channel is of the stochastic block model type eq. (4.71-4.72), leading to effective noise parameter

$$\Delta = \frac{p_{\rm out}(1-p_{\rm out})}{\mu^2}\,, \tag{4.348}$$

where μ and $p_{\rm out}$ are the parameters from (4.71-4.72). Therefore the uniform fixed point becomes unstable when $\Delta < 1$.

Using eqs. (4.202) and (4.76) we get for the state evolution for community detection with two balanced groups

$$m^{t+1} = f^{\rm SE}_{\rm TwoBalanced}\left(\frac{m^t}{\Delta}\right), \forall t, m^t \in [0;1]\,, \tag{4.349}$$

where

$$f^{\rm SE}_{\rm TwoBalanced}(x) = \int_{-\infty}^{+\infty} \frac{e^{\frac{-u^2}{2}}}{\sqrt{2\pi}} \frac{2\rho(1-\rho)\sinh\left(\frac{x}{2\rho(1-\rho)} + u\sqrt{\frac{x}{\rho(1-\rho)}}\right)}{1 + 2\rho(1-\rho)\left(\cosh\left(\frac{x}{2\rho(1-\rho)} + u\sqrt{\frac{x}{\rho(1-\rho)}}\right) - 1\right)}\,. \tag{4.350}$$

To investigate whether $\Delta_c = 1$ is a 1st or 2nd order bifurcation we compute the second order expansion of the state evolution equations as we have done in (4.335). We find that the expansion

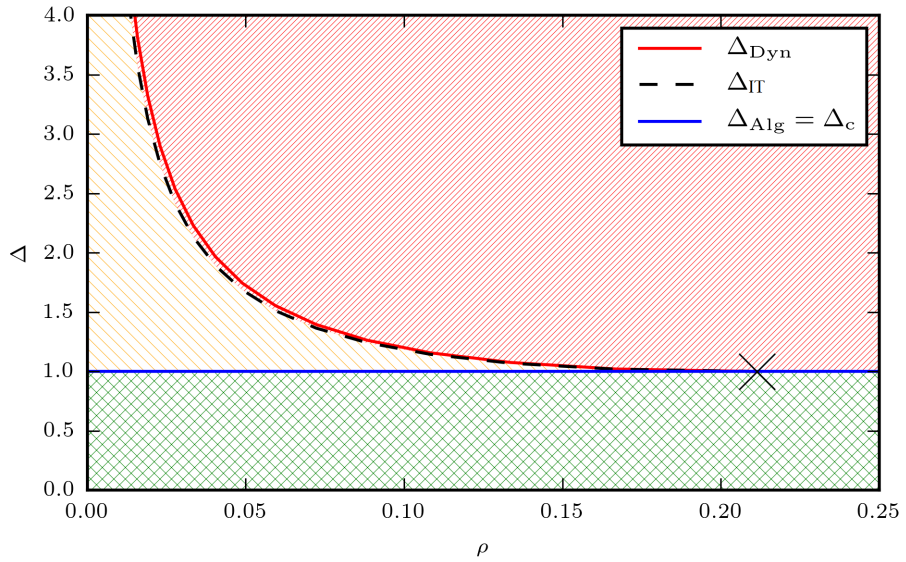


FIGURE 4.11 – We plot here $\Delta_{\text{Alg}} = \Delta_c$, Δ_{IT} and Δ_{Dyn} as a function of ρ for the community detection with two balanced groups. All the curves merge at $\rho = \frac{1}{2} - \frac{1}{\sqrt{12}}$. The hashed regions have the same meaning as in previous figures, red is the impossible inference phase, green is easy and yellow is hard inference.

of the state evolution up to second order for the two balanced communities is

$$m^{t+1} = f_\rho\left(\frac{m^t}{\Delta}\right) = \frac{m^t}{\Delta} + \left(\frac{m^t}{\Delta}\right)^2 \frac{1-6\rho(1-\rho)}{2\rho(1-\rho)}\,. \tag{4.351}$$

In a similar fashion as in section (4.4.4) it is the sign of the second order terms that decides between 1st or 2nd order bifurcation at $\Delta_c = 1$. We find that if $\rho(1-\rho) < 1/6$ then the second order derivative of (4.350) is positive leading to a jump in MSE when Δ crosses the value $\Delta = 1$ which means that there will be first order phase transition for all

$$\rho \in \left[0; \frac{1}{2} - \frac{1}{\sqrt{12}} \approx 0.21\right] \cup \left[\frac{1}{2} + \frac{1}{\sqrt{12}} \approx 0.79; 1\right]\,. \tag{4.352}$$

It turns out that for the two balanced groups this criteria is both sufficient and necessary. Out of the interval (4.352) the phase transition at Δ_c is of second order, with no discontinuities. Defining the phase transitions Δ_{Alg}, Δ_{IT}, Δ_{Dyn} as before in section 4.4.4, we have $\Delta_{\text{Alg}} = \Delta_c$ and we plot the three phase transitions for community detection for two balanced groups in the phase diagram Fig. 4.11.

Jointly-sparse PCA generic rank

In this section we discuss analysis of the Gauss-Bernoulli jointly sparse PCA as defined by the prior distribution (4.65) for a generic rank r. This just means that each vector $x_i \in \mathbb{R}^r$ is either $\vec{0}$ with probability $1-\rho$ or is taken from Gaussian density probability of mean zero and covariance matrix I_r with probability ρ. The prior distribution (4.65) has a zero mean, therefore the uniform fixed point exist and according to (4.305) is stable down to $\Delta_c = \rho^2$.

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

$$M^t = m^t I_r, \tag{4.353}$$

with m^t being a scalar parameter. The problem is hence greatly simplified, one can then treat the r dimensional integral in (4.256). After integration by parts and integration on the sphere one gets

$$m^{t+1} = f^{\text{SE}}_{\text{Joint}-\text{GB}}\left(\frac{m^t}{\Delta}\right), \tag{4.354}$$

$$f^{\text{SE}}_{\text{Joint}-\text{GB}}(x) = \frac{\rho x}{1+x}\int \mathrm{d}u \frac{1}{(2\pi)^{\frac{r}{2}}} \exp\left(\frac{-u^2}{2}\right) S_r u^{r-1} \left\{1 + \frac{xu^2\left[1-\hat{\rho}(x,(x^2+x)u^2)\right]}{r}\right\} \hat{\rho}(x,(x^2+x)u^2)\,, \tag{4.355}$$

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

$$\hat{\rho}(a,b^2) = \frac{\rho}{(1-\rho)\exp\left(\frac{-b^2}{2(1+a)}\right)(1+a)^{\frac{r}{2}} + \rho}\,. \tag{4.356}$$

An expansion of (4.354) around $m^t = 0$ yields

$$m^{t+1} = \frac{\rho^2 m^t}{\Delta} - \rho^3\left(\frac{m^t}{\Delta}\right)^2 + O\left((m^t)^3\right)\,. \tag{4.357}$$

Analogously to the conclusions we reached when studying expansion (4.335), we conclude that since the second term is negative there is always a 2nd order bifurcation at $\Delta_c = \rho^2$ with a stable fixed point for $\Delta < \Delta_c$ that stays close to the uniform fixed point. At the same time this close-by fixed point typically exists only in a very small interval of $(\Delta_{\text{Alg}}, \Delta_c)$, similarly as in Fig. 4.10.

Community detection with symmetric groups

In this section we discuss the phase diagram of the symmetric communities detection model as defined in section 4.1.4. The corresponding prior distribution is (4.70) which leads to the function f^x_{in}

$$\forall k \in [1:r], f^x_{\text{in}}(A,B)_k = \frac{\exp\left(B_k - A_{kk}/2\right)}{\sum\limits_{k'=1\cdots r}\exp\left(B_{k'} - A_{k'k'}/2\right)}\,. \tag{4.358}$$

The corresponding output is given by (4.71-4.72), corresponding to the effective noise

$$\Delta = \frac{p_{\text{out}}(1-p_{\text{out}})}{\mu^2} = \frac{p_{\text{out}}(1-p_{\text{out}})}{N(p_{\text{in}}-p_{\text{out}})^2}\,. \tag{4.359}$$

Once again we study the phase diagram by analyzing the state evolution (4.256). We can verify that the following form of the order parameter in invariant under iterations of the Bayes-optimal

state evolution

$$M^t = b^t \frac{I_r}{r} + \frac{(1-b^t)J}{r^2}\,, \tag{4.360}$$

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

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

— Having $b^t = 1$ means that the communities have been perfectly reconstructed. This corresponds to initializing the algorithm in the planted solution.

Using (4.256) the state evolution equations for b^t can be written.

$$b^{t+1} = \mathcal{M}_r\left(\frac{b^t}{\Delta}\right)\,, \tag{4.361}$$

where

$$\mathcal{M}_r\left(x\right) = \frac{1}{r-1}\left[r\int \frac{\exp\left(\frac{x}{r} + u_1\sqrt{\frac{x}{r}}\right)}{\exp\left(\frac{x}{r} + u_1\sqrt{\frac{x}{r}}\right) + \sum_{i=2}^{r}\exp\left(u_i\sqrt{\frac{x}{r}}\right)}\prod_{i=1}^{r}\mathrm{d}u_i\frac{\exp\left(\frac{-u_i^2}{2}\right)}{\sqrt{2\pi}} - 1\right]\,. \tag{4.362}$$

This can be proven by computing M_{11}^{t+1} using (4.256) and (4.358) which yields

$$\begin{aligned} &b^{t+1}\left(\frac{1}{r} - \frac{1}{r^2}\right) + \frac{1}{r^2} = \\ &= \frac{1}{r}\left[\int \frac{\exp\left(\frac{x}{r} + u_1\sqrt{\frac{x}{r}} + u_0\sqrt{\frac{1-b^t}{r^2\Delta}}\right)}{\exp\left(\frac{x}{r} + u_1\sqrt{\frac{x}{r}} + u_0\sqrt{\frac{1-b^t}{r^2\Delta}}\right) + \sum_{i=2}^{r}\exp\left(u_i\sqrt{\frac{x}{r}} + u_0\sqrt{\frac{1-b^t}{r^2\Delta}}\right)}\prod_{i=0}^{r}\mathrm{d}u_i\frac{\exp\left(\frac{-u_i^2}{2}\right)}{\sqrt{2\pi}}\right]\,. \end{aligned} \tag{4.363}$$

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

One observes that eq. (4.361) has always the uniform fixed point $b^t = 0$. This is an example of a non-zero mean prior for which nevertheless there is a uniform fixed point because other kind of symmetry is present in the model. Let us expand (4.361) around $b^t = 0$ to determine the stability of this fixed point, one gets

$$b^{t+1} = \frac{b^t}{\Delta r^2} + \frac{r-4}{2\Delta^2 r^4}b^{t2} + O\left(b^{t3}\right)\,. \tag{4.364}$$

The uniform fixed point hence becomes unstable for $\Delta > \Delta_c = \frac{1}{r^2}$. Translated back into the parameters of the stochastic block model this gives the easy/hard phase transition at

$$|p_{\text{in}} - p_{\text{out}}| = r\sqrt{p_{\text{out}}(1-p_{\text{out}})/N} \tag{4.365}$$

well known in the sparse case where $p = \text{const}/N$ from [DKMZ11a]. In terms of the type of

phase transitions there are two cases

— 2nd order for $r \leq 3$. The second term in (4.361) is negative, this means that there will be a fixed point close-by to the uniform one for $\Delta < \Delta_c$. The transition is of second order.

— 1st order for $r \geq 5$. The second term in (4.361) is positive, this means that there will be a jump in the order parameter at the transition $\Delta_c = \Delta_{\text{Alg}}$. This is the signature a first order phase transition.

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

In Fig. 4.12 left panel, we illustrate the first order phase transition in the state evolution and in the behavior of the Low-RAMP algorithms for $r = 15$ groups.

To compute the values of Δ_{IT} and Δ_{Dyn}, we write $b^{t+1} = \mathcal{M}_r(b^t/\Delta)$ and carry a similar analysis as in section 4.4.4, as detailed in appendix 4.6.2. Fig. 4.12 (right panel) summarizes the values in a scaling that anticipated the large rank expansion done is section 4.5.4.

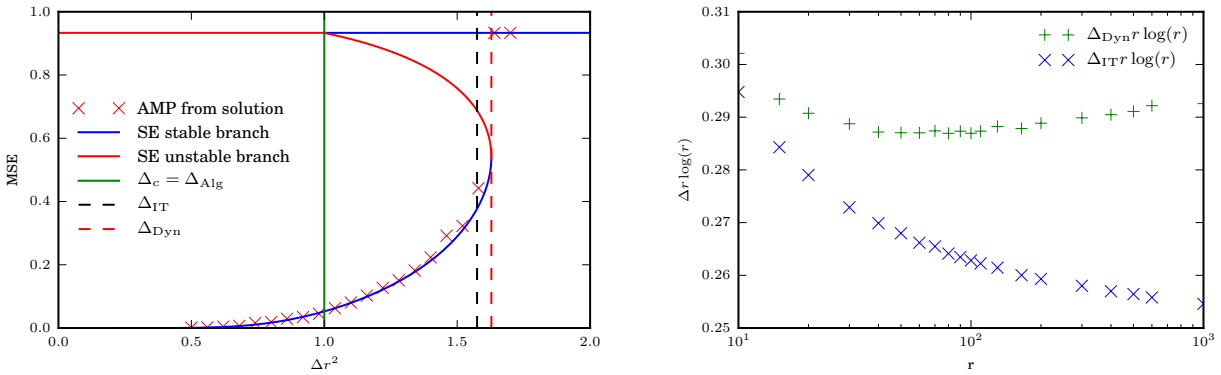


FIGURE 4.12 – **Left :** We plot MSE deduced from state evolution (lines) and from Low-RAMP algorithm (marks) for $r = 15$ groups and $N = 20000$ as a function of Δr^2. The vertical full green line is $\Delta_c r^2 = 1$. The vertical dashed black line is Δ_{IT} and the full lines correspond to the MSE obtained from the informative initialization and have discontinuities at Δ_{Dyn}. Note that the MSE from does not go to zero at finite positive Δ instead at small noise one has MSE $\sim \exp(-\text{const.}/\Delta)$. **Right :** We plot $\Delta r \log r$ for the information theoretic Δ_{IT} and dynamical spinodal Δ_{Dyn} phase transitions obtained from the state evolution using the protocol described in Appendix 4.6.2. We rescale the Δ in this way to compare with the large rank expansion in (4.408) and (4.409).

Gaussian Mixture Clustering

In this section we discuss the phase diagram of the Gaussian mixture clustering model defined in 4.1.4. We suppose that the centroids were sampled from isotropic Gaussian. This means that $P_U(u)$ is given by

$$P_U(u) = \mathcal{N}(0, \rho I_r, u) \tag{4.366}$$

ρ will be a measure of how well are separated the centroids (In that setting this is a more natural variable than Δ). While $P_V(v)$ is given by (4.60). Δ will be set to 1 and ρ will be the variable that will change. f_{in}^v will be given by (4.358) while f_{in}^u will be given by

$$f_{\text{in}}^u(A,B) = \left(\frac{I_r}{\rho} + A\right)^{-1} B\,. \tag{4.367}$$

Once again we analyze the phase diagram by analyzing the state evolution and the ansatz that allows us to solve the SE equation easily is the following

$$M_u^t = Q_u^t = b_u^t I_r + b_{u,J}^t \frac{J_r}{r}, \tag{4.368}$$

$$M_v^t = Q_v^t = \frac{I_r b_v^t}{r} + (1 - b_v^t)\frac{J_r}{r^2} \tag{4.369}$$

where $(b_v^t, b_u^t, b_{u,J}^t) \in [0;1] \times [0;\rho]^2$, and I_r and J_r are respectively the identity matrix and the $r \times r$ matrix filled with 1. Having $(b_v^t, b_u^t, b_{u,J}^t) = (1, \rho, 0)$ would mean that we have achieved perfect reconstruction of the ground truth, while $(b_v^t, b_u^t) = (0,0)$ means that we are not able to extract any information from the matrix Y except for the average of the k clusters U_k. Using (4.261) and (4.262) we get the following SE equations :

$$b_u^t = \frac{b_v^t \rho^2}{\frac{r}{\alpha} + b_v^t \rho}\,, \qquad b_v^{t+1} = \mathcal{M}_r\left(b_u^t r\right)\,, \tag{4.370}$$

where $\mathcal{M}_r$ is given by (4.362). We can combine these to obtain a single update equation for the variables b_v^t :

$$b_v^{t+1} = \mathcal{M}_r\left(b_v^t \frac{\rho^2}{\frac{1}{\alpha} + \frac{\rho b_v}{r}}\right)\,. \tag{4.371}$$

Notice that this equations is closed on b_u^t, which got eliminated. $b_v = 0$ is a trivial fixed point of (4.371) by expanding $\mathcal{M}_r(x)$ (using 4.364) around the trivial fixed point we get

$$b_v^{t+1} = \frac{\alpha b_v^t \rho^2}{r^2} + \frac{\alpha^2 {b_v^t}^2}{2}\left[r - 4 - \frac{2r}{\rho}\right]\frac{\rho^4}{r^4}\,. \tag{4.372}$$

Notice that $(b_u = 0, b_v = 0)$ is always a fixed point. We are interested in when this 'uninformative' fixed point becomes unstable. From the expansion of (4.372) one deduces that this occurs when :

$$\rho > \rho_c = \frac{r}{\sqrt{\alpha}}\,. \tag{4.373}$$

Looking at the second derivative of (4.372), we deduce that when the uninformative fixed point becomes unstable, the second derivative is proportional to $r - 4 - 2\sqrt{\alpha}$; if this is negative then this means that another fixed point appears close to 0. If on the other hand the second derivative is positive when ρ increases and crosses ρ_c, we see a jump in MSE as $b_v^{t=+\infty}$ jumps non-continuously to another value (when $b_v^{t=0} \approx 0$) and that means that there is a first order transition phenomena in the system. This means that for some value of ρ, r and α there can be multiple stable fixed point to the SE equations. If one fixes the number r and α, it is possible to have first order transition in the system if and only if

$$r > 4 + 2\sqrt{\alpha}\,. \tag{4.374}$$

It turns out that this sufficient criteria is also necessary (this comes from SE numerical experiments). We plot the overlap b_v as a function of ρ in Fig 4.13.

To compute the values of ρ_{IT} and ρ_{Dyn}, we write $b_v^{t+1} = \mathcal{M}_r\left(b_v^t \frac{\rho^2}{\frac{1}{\alpha}+\frac{\rho b_v}{r}}\right)$ and carry a similar analysis as in section 4.4.4, as detailed in appendix 4.6.2. Fig. 4.13 (right panel) summarizes the values in a scaling that anticipated the large rank expansion done is section 4.5.4.

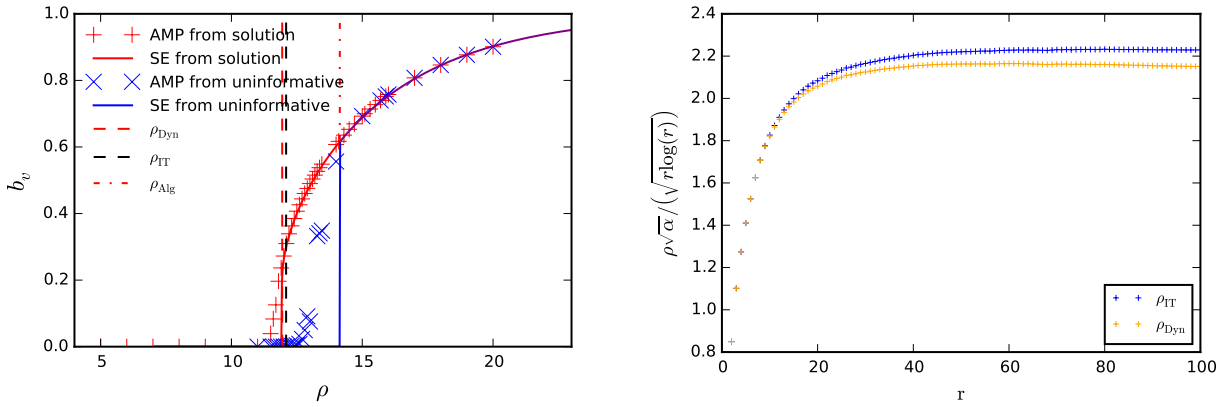


FIGURE 4.13 – **Left :** AMP simulations vs. theory for $r = 20$ clusters. Points are results of numerical simulations using the AMP algorithm; lines are theoretical results using State Evolution. All results are for $\alpha = 2$, $N = 10000$ and $M = 20000$. We observe a phase transition at $\rho_c = \frac{r}{\sqrt{\alpha}} \approx 14.14$ and a discontinuity in overlap at ρ_c. The blue and red curve follow the b_v that can be achieved by initializing at $b_u^{t=0} = 0$ and $b_v^{t=0} = 1$ respectively. The vertical dashed lines indicate positions of the three transitions deduced from the SE equations (Here higher ρ means an easier problem). Because of the discontinuous nature of the transition the finite size effect are sizable. For instance, the red points with non-zero overlap below ρ_c correspond to the algorithm being able to reconstruct a fraction of the $r = 20$ clusters which leads to these intermediate value of the overlap. Increasing n decreases these finite size effects. **Right :** We plot the $\rho_{\text{Spinodal}}(r)\sqrt{\alpha}/(\sqrt{r\log r})$ and $\rho_{\text{Static}}(r)\sqrt{\alpha}/(\sqrt{r\log r})$ for $\alpha = 2$. In the large r limit these curves should go to $\sqrt{2}$ and 2 respectively as derived in 4.5.4. The asymptotic behavior of the static transition can not yet be easily observed : We conjecture that one would need to go much larger values of r to see the limit behavior of the spinodal.

4.5.2 Tensor factorization.

In this subsection we present the result of the analysis of the Bayes optimal tensor factorization. Tensor are mathematical objects that are far less convenient to manipulate than matrix. A lot of basic operation that can be achieved through polynomial algorithm on matrices are NP hard for tensors ([HL09]).

To begin to understand tensor factorization a good first way to do it is to analyze the tensor factorization in the Bayes optimal case where the data is taken to come from a rank 1 Gaussian.

Rank 1 Gaussian Bayes optimal Tensor factorization

The prior is given by

$$P_X(x) = \mathcal{N}(\mu, 1, x) \tag{4.375}$$

First let us analyze the SE equations in the case where the mean of the distribution μ is 0.

The SE equations sums up to

$$m_x^{t+1} = \frac{{m_x^t}^{p-1}}{\Delta + {m_x^t}^{p-1}} \tag{4.376}$$

it is easy to see that the trivial fixed point of this equation will always be stable. This means that at $\mu = 0$ there will be no easy phase. And AMP starting from an uninformative fixed point will fail to reconstruct the signal even for $\Delta \ll 1$.

As soon as $\mu \neq 0$ the trivial fixed point will stop to exist. One can use equations (4.332 4.3334.334) to compute the $\Delta_{\text{Dyn}}(\mu), \Delta_{\text{IT}}(\mu)$ and $\Delta_{\text{Alg}}(\mu)$.

Having a non non zero μ makes it easier to solve the problem. For $0 < \mu < \mu_{\text{Tri}} = \frac{p-2}{2\sqrt{p-1}}$. The 3 phase

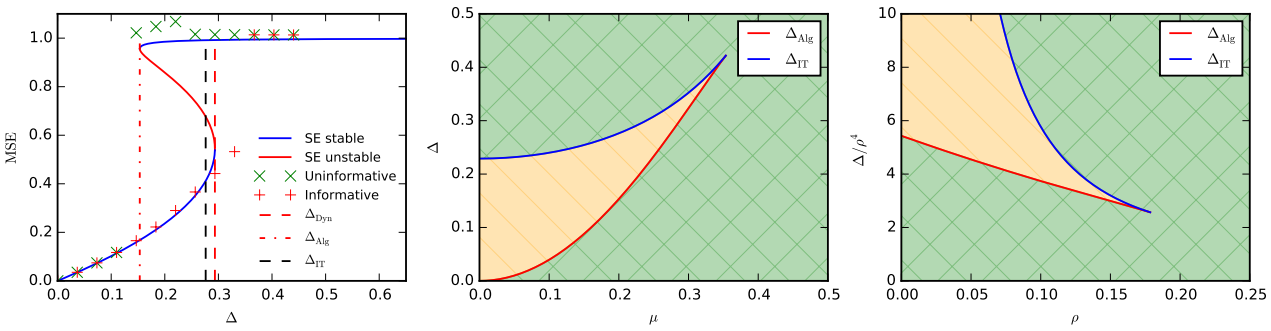


FIGURE 4.14 – **Left panel :** Comparison between the AMP fixed point reached from uninformative (marked with crosses) or informative (i.e. strongly correlated with the ground truth, marked with pluses) initialization and the fixed point of the SE equations (stable fixed point in blue, unstable in red). The data are for the Gaussian prior with mean $\mu = 0.2$, unit variance, $p = 3$, $r = 1$. The AMP runs are done on a system of size $N = 1000$. **Central panel :** Phase diagram for the order $p = 3$ tensor factorization, rank $r = 1$, Gaussian prior of mean μ (x-axes) and unit variance. In the green-shaded zone AMP matches the information-theoretically optimal performance, MMSE = MSE_{AMP}. In the orange-shaded zone MMSE < MSE_{AMP}. The tri-critical point is located at $\mu_{\text{Tri}} = (p-2)/(2\sqrt{p-1})$ and $\Delta_{\text{Tri}} = x_{\text{Tri}}^{p-2}/(1+x_{\text{Tri}})^{p-1}$ where $x_{\text{Tri}} = (p-2)(3p-4)/p^2$. **Right panel :** Phase diagram for the order $p = 3$ tensor factorization, rank $r = 1$, the Bernoulli prior as a function of ρ and Δ/ρ^4. The tri-critical point is located at $\rho_{\text{Tri}} = 0.178$ and $\Delta_{\text{Tri}}/\rho^4 = 2.60$. As $\rho \to 0$ we observed $\Delta_{\text{Alg}}/\rho^4 \to 2e$. Compare to Fig. 5 in [LKZ17] where the same phase diagram is presented for the matrix factorization $p = 2$ case.

4.5.3 Large sparsity (small ρ) expansions

In existing literature the sparse PCA was mostly studied for sparsity levels that are much smaller than a finite fraction of the system size (as considered in this thesis). In order to compare with

existing results we hence devote this section to the study of small density ρ expansions of the results obtained from the state evolution for some of the models studied.

Spiked Bernoulli, Rademacher-Bernoulli, and Gauss-Bernoulli models

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

The small ρ limit behavior of the information theoretic Δ_{IT} threshold and the dynamical spinodal threshold Δ_{Dyn} are given by

Bernoulli and Rademacher-Bernoulli

$$\Delta_{\text{Dyn}}(\rho) \sim_{\rho\to 0} \frac{-\rho}{2\log(\rho)}\,, \tag{4.377}$$

$$\Delta_{\text{IT}}(\rho) \sim_{\rho\to 0} \frac{-\rho}{4\log(\rho)}\,, \tag{4.378}$$

Gaussian-Bernoulli

$$\Delta_{\text{Dyn}}(\rho) \sim_{\rho\to 0} \frac{-\rho}{\log(\rho)} \max\left\{ \frac{\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)}{\beta}, \beta \in \mathbb{R}^+ \right\} \sim 0.595 \frac{-\rho}{\log(\rho)}\,, \tag{4.379}$$

$$\Delta_{\text{IT}}(\rho) \sim_{\rho\to 0} \frac{-\rho}{\log(\rho)} \max\left\{ \frac{\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)}{\beta}, \right.$$
$$\left. \int_0^\beta \mathrm{d}u\, \frac{2\exp\left(\frac{-1}{u}\right)}{\sqrt{\pi u}} + \left(\frac{1}{\sqrt{u}}\right) = \frac{1}{2}\beta \left[\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right) \right]; \beta \in \mathbb{R}^+ \right\} \sim 0.528 \frac{-\rho}{\log(\rho)}\,. \tag{4.380}$$

The information theoretic transitions for all these 3 models scale like $O\left(\frac{-\rho}{\log(\rho)}\right)$ while the algorithmic transition scales like $O(\rho^2)$. This means that for small ρ there is a large gap between what is information theoretically and algorithmically achievable. Note at this point that the bounds derived in [BMVX16] for sparse PCA have the same leading order behavior when ρ is small as (4.377) and (4.378).

To derive the above small ρ expressions we combine (4.328) and (4.330) and the following small

ρ limit of the state-evolution functions $f^{\mathrm{SE}}(x)$

$$\forall \beta \in \mathbb{R}^+, \lim_{\rho\to 0} \frac{f^{\mathrm{SE}}(-\beta \log(\rho))}{\rho} = 1(\beta > 2), \textbf{ Bernoulli} \tag{4.381}$$

$$\forall \beta \in \mathbb{R}^+, \lim_{\rho\to 0} \frac{f^{\mathrm{SE}}(-\beta \log(\rho))}{\rho} = 1(\beta > 2), \textbf{ Rademacher-Bernoulli} \tag{4.382}$$

$$\forall \beta \in \mathbb{R}^+, \lim_{\rho\to 0} \frac{f^{\mathrm{SE}}(-\beta \log(\rho))}{\rho} = \frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\beta\pi}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right), \textbf{ Gaussian-Bernoulli} \tag{4.383}$$

Here the functions f^{SE} are the state-evolution update functions stated in (4.342), (4.344) and (4.346) for the different models (when the model is clear from context we will omit the lower index specifying the model). The above is proven by deriving the state evolution equations for each of these models. The computation is done in appendix 4.6.2.

Two balanced groups, limit of small planted subgraph

In this section we analyze the small ρ limit for the two balanced groups of section 4.5.1. From the definition of the function f_ρ (4.351), and a computation done in appendix 4.6.2 we get

$$\lim_{\rho\to 0} f_\rho(-\beta\rho(1-\rho)\log(\rho(1-\rho))) = 1(\beta > 2)\,. \tag{4.384}$$

By combining this with (4.328) and (4.330) one gets

$$\Delta_{\mathrm{Dyn}}(\rho) \sim_{\rho\to 0} \frac{1}{-2\rho(1-\rho)\log(\rho(1-\rho))}\,, \tag{4.385}$$

$$\Delta_{\mathrm{IT}}(\rho) \sim_{\rho\to 0} \frac{1}{-4\rho(1-\rho)\log(\rho(1-\rho))}\,. \tag{4.386}$$

The derivations of these limits is done in the appendix 4.6.2.

Note that the limit of small ρ in the two balanced groups model is closely related to the problem of planted clique. However, in the planted clique problem the size of the clique to recover is much smaller that the size of the graph, typically $O(\sqrt{N})$ for efficient recovery and $O(\log N)$ for information theoretically possible recovery. Whereas our equation were derives when $\rho = O(1)$, we can try to see what would the above scaling imply for $k = \rho N = o(N)$. In the planted clique problem the first (smaller) group is fully connected, this means that the entry C_{11} of the connectivity matrix C (4.67) is equal to one. Therefore for $\rho \ll 1$ one has

$$\mu = (1 - p_{\mathrm{out}})\rho\sqrt{N}\,. \tag{4.387}$$

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

Recalling the definition of Δ (4.348) for the community detection output channel, and using $\Delta_c = 1$ for the spectral threshold, and $\Delta_{\mathrm{IT}} = -1/(4\rho\log\rho)$ for the information theoretic

threshold at small ρ we get

$$k_c = \sqrt{N}\sqrt{\frac{p_{\text{out}}}{1-p_{\text{out}}}}\,, \tag{4.388}$$

$$k_{\text{IT}} = \log(N)\frac{4p_{\text{out}}}{1-p_{\text{out}}}\,. \tag{4.389}$$

We indeed recover the scaling known from the planted clique problem, see e.g. [DM15]. The p_{out}-dependent constant are indeed the tight constants (for efficient and information theoretically optimal recovery) for the balanced version of the planted clique where the expected degree of every node is the same independently of the fact in the node is in the clique or not.

Sparse PCA at small density ρ

We investigate here the small ρ limit of the bipartite $UV^\top$ spiked Gaussian-Bernoulli, and Rademacher-Bernoulli model. We remind that $U \in \mathbb{R}^N$, $V \in \mathbb{R}^M$, while $\alpha = M/N$. In the model we consider that elements of U are Gaussian of zero mean and unit variance, while P_V is given by (4.63) for the Rademacher-Bernoulli model, and by (4.66) for the Gauss-Bernoulli model. The state evolution equations then read

$$m_u^t = \frac{\alpha m_v^t}{\Delta + \alpha m_v^t}\,, \tag{4.390}$$

$$m_v^{t+1} = f^{\text{SE}}_{\text{Gauss}-\text{Bernoulli}}\left(\frac{m_u^t}{\Delta}\right)\,, \tag{4.391}$$

$$m_v^{t+1} = f^{\text{SE}}_{\text{Rademacher}-\text{Bernoulli}}\left(\frac{m_u^t}{\Delta}\right)\,, \tag{4.392}$$

where $f^{\text{SE}}_{\text{Gauss}-\text{Bernoulli}}$ and $f^{\text{SE}}_{\text{Rademacher}-\text{Bernoulli}}$ are defined in (4.344) and (4.346) respectively. By combining these equations one gets

$$m_v^{t+1} = f^{\text{SE}}_{\text{Gauss}-\text{Bernoulli}}\left(\frac{\alpha m_v^t}{\Delta^2 + \alpha\Delta m_v^t}\right)\,, \tag{4.393}$$

$$m_v^{t+1} = f^{\text{SE}}_{\text{Rademacher}-\text{Bernoulli}}\left(\frac{\alpha m_v^t}{\Delta^2 + \alpha\Delta m_v^t}\right)\,. \tag{4.394}$$

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

$$\frac{\rho^2\alpha}{\Delta^2} > 1, \text{or}, \Delta < \Delta_c = \rho\sqrt{\alpha}\,. \tag{4.395}$$

Also in this case the stability transition Δ_c corresponds to the spectral transition where one sees informative eigenvalues get out of the bulk of the matrix S, as is known in the theory of low-rank perturbations of random matrices [BBAP05] (these methods are known not to take advantage of the sparsity). For ρ small enough there will be again a first order phase transitions with Δ_{Alg}, Δ_{IT}, Δ_{Dyn} defined as in section 4.4.4. The asymptotic behavior of these thresholds

is

Rademacher-Bernoulli

$$\Delta_{\mathrm{Dyn}}(\rho) \sim_{\rho\to 0} \sqrt{\frac{-\rho\alpha}{2\log(\rho)}}\,, \tag{4.396}$$

$$\Delta_{\mathrm{IT}}(\rho) \sim_{\rho\to 0} \sqrt{\frac{-\rho\alpha}{4\alpha\log(\rho)}}\,, \tag{4.397}$$

Gaussian-Bernoulli

$$\Delta_{\mathrm{Dyn}}(\rho) \sim_{\rho\to 0} \sqrt{\frac{-\rho\alpha}{\log(\rho)}\max\left\{\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right), \beta\in\mathbb{R}^+\right\}} \sim 0.771\sqrt{\frac{-\rho\alpha}{\log(\rho)}}\,, \tag{4.398}$$

$$\Delta_{\mathrm{IT}}(\rho) \sim_{\rho\to 0} \sqrt{\frac{-\rho\alpha}{\log(\rho)}}$$

$$\sqrt{\max\left\{\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right), \int_0^\beta \mathrm{d}u\,\frac{2\exp\left(\frac{-1}{u}\right)}{\sqrt{\pi u}} + \operatorname{erfc}\left(\frac{1}{\sqrt{u}}\right) = \frac{1}{2}\beta\left[\frac{2\exp\left(\frac{-1}{\beta}\right)}{\sqrt{\pi\beta}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right)\right]; \beta\in\mathbb{R}^+\right\}}$$

$$\sim 0.726\sqrt{\frac{-\rho\alpha}{\log(\rho)}}\,. \tag{4.399}$$

Reminding that the value of Δ_{Alg} scales in the same way as Δ_c (4.395), we see that a large hard phase opens as $\rho \to 0$.

To put the above results in relation to existing literature on sparse PCA [AW08, DM14b]. The main difference is that the regime considered in existing literature is that the number of non-zero element in the matrix V, $\rho N = o(1)$. The information theoretic threshold found in eq. (4.397) correspond (up to a constant) to information theoretic bounds found in [AW08]. However, the algorithmic performance that is found in the case of very small sparsity, by e.g. the covariance thresholding of [DM14b], is not reproduced in our analysis of linear sparsity $\rho = O(1)$. This suggest that in case when the sparsity is small but linear in N, efficient algorithms that take advantage of the sparsity might not exist. This regime should be investigated further.

To derive the small ρ limit we follow a similar strategy as we did in the symmetric $XX^\top$ case. The main idea is to find the value of Δ for which $m_v = f^{\mathrm{SE}}_{P_X}(x)$ is a fixed point of the SE equations (4.394) or (4.393).

$$\Delta(x) = \frac{\alpha f^{\mathrm{SE}}(x) + \sqrt{\alpha^2 f^{\mathrm{SE}2}(x) + 4\alpha\frac{f^{\mathrm{SE}}(x)}{x}}}{2}\,, \tag{4.400}$$

This means that

$$(m_u, m_v) = (x\Delta(x), f^{\mathrm{SE}}(x)) \tag{4.401}$$

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

$$\begin{aligned}\phi(m_u = x\Delta(x), m_v = f(x), \Delta = \Delta(x)) - \phi(m_u = 0, m_v = 0, \Delta = \Delta(x)) = \\ = \alpha \int_0^x \mathrm{d}u f^{\mathrm{SE}}(u) + \int_0^{\alpha f^{\mathrm{SE}}(x)/\Delta(x)} \mathrm{d}u \frac{u}{1+u} - x f^{\mathrm{SE}}(x)\,. \end{aligned} \tag{4.402}$$

Combining this with (4.383) and (4.382) one gets the above asymptotic behavior.

4.5.4 Large rank expansions

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

Large-rank limit for jointly-sparse PCA

We analyze the large r limit of jointly-sparse PCA for which the state evolution equation is given by (4.354). We notice that u^2 will have mean r and standard deviation $\sqrt{2r}$. This essentially means that to get the large r limit of the density evolution equations one need to replace u^2 by r everywhere it appears. Expanding in large r then gives

$$m^{t+1} = \frac{\rho m^t}{m^t + \Delta} \left(1 + \frac{m^t}{\Delta}(1 - \hat{\rho})\right) \hat{\rho}\,, \tag{4.403}$$

where

$$\hat{\rho} = \lim_{r \to +\infty} \frac{\rho}{(1-\rho)\exp\left(\frac{r}{2}\left[\frac{m^t}{\Delta} + \log(1 + \frac{m^t}{\Delta})\right]\right) + \rho} = \begin{cases} 1 \text{ if } r m^{t2} \to 0 \\ \rho \text{ if } r m^{t2} \to +\infty \end{cases}\,. \tag{4.404}$$

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

$$m^{t+1} = \frac{m^t \rho}{m^t + \Delta} + o(1)\,. \tag{4.405}$$

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

$$\max(\rho - \Delta, 0)\,. \tag{4.406}$$

Analogous expansion of the replicated free energy leads to the result that in the large rank limit we have $\Delta_{\mathrm{Dyn}} = \Delta_{\mathrm{IT}} = \rho$ whereas $\Delta_{\mathrm{Alg}} = \Delta_c = \rho^2$. This is plotted in the large rank phase diagram (4.15).

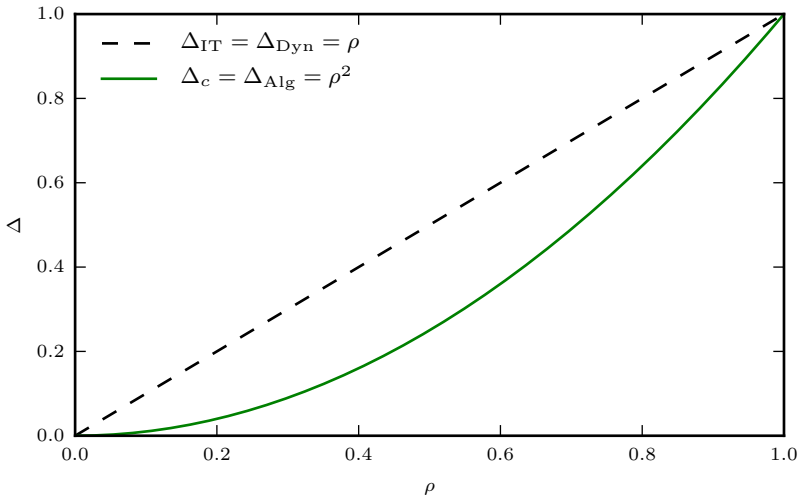


FIGURE 4.15 – Phase diagram of the jointly-sparse PCA model at large rank r. In the large r limit the algorithmic spinodal merges with Δ_c, the dynamical spinodal and the information theoretic one converge toward the line $\Delta = \rho$.

Community detection

The large rank limit is analyzed for the problem of symmetric community detection in appendix 4.6.2. The asymptotic behavior of $\Delta_{\rm IT}(r)$ and $\Delta_{\rm Dyn}(r)$ as $r \to +\infty$ are

$$\Delta_c = \frac{1}{r^2}\,, \tag{4.407}$$

$$\Delta_{\rm Dyn} = \frac{1}{2r\ln(r)}[1 + o_r(1)]\,, \tag{4.408}$$

$$\Delta_{\rm IT} = \frac{1}{4r\ln(r)}[1 + o_r(1)]\,. \tag{4.409}$$

We see that a large gap opens between Δ_c and $\Delta_{\rm IT}$ as r grows. The behavior $\Delta_{\rm IT}$ and $\Delta_{\rm Dyn}$ for moderately large r is illustrated in figure 4.12 and we see that the above limit is reached very slowly. Using eq. (4.359) this translates into the large r limit phase transition in terms of parameters of the stochastic block model as discussed in [LKZ15a], and proven in [BMVX16].

Gaussian mixture model

The large rank limit (large number of clusters) is analyzed for the problem of Gaussian mixture clustering in appendix 4.6.2. The asymptotic behavior of $\rho_{\rm IT}(r,\alpha)$ and $\rho_{\rm Dyn}(r,\alpha)$ as $r \to +\infty$

are

$$\rho_c(r,\alpha) = \rho_{\text{Alg}}(r,\alpha) = \frac{r}{\sqrt{\alpha}} \tag{4.410}$$

$$\rho_{\text{IT}}(r,\alpha) = 2\sqrt{\frac{r\log r}{\alpha}}(1+o_r(1))\,, \tag{4.411}$$

$$\rho_{\text{Dyn}}(r,\alpha) = \sqrt{\frac{2r\log r}{\alpha}}(1+o_r(1))\,. \tag{4.412}$$

We see that a large gap opens between ρ_c and ρ_{IT} as r grows. The behavior ρ_{IT} and ρ_{Dyn} for moderately large r is illustrated in figure 4.13 and we see that the above limit is reached very slowly.

4.6 Sparse PCA and the rotational symmetry

In that section I will present a result/idea that I thought were interesting but I did not feel belong in any paper in particular but that I still think are worth mentioning. During the few months I spent working on the problem of sparse PCA and the Gauss-Bernoulli prior I realized that the behavior of the AMP algorithm did not match the replica computation prediction. This was a problem and I think/hope I was able to diagnostic and understand why the problem of matrix factorization with a Gauss-Bernoulli prior (4.66) is a harder task that one might think, especially when r grows large ($r > 20$).

4.6.1 Sparse PCA and rotational symmetry

By sparse PCA I mean the inference problem when one of the priors be it P_X or P_U is given by (4.66) (and P_V is just a Gaussian prior).

$$P_X(x_i) = \prod_{1\leq k\leq r}\left[\frac{\rho}{\sqrt{2\pi}}e^{\frac{-x_{ik}^2}{2}} + (1-\rho)\delta(x_{ik})\right]\,. \tag{4.413}$$

Problem arises when the rank r grows large and the signal ρ is not small enough. Even though we are completely able to derive the free energy for this problem and compute easily the order parameter $M = Q$ that minimizes the replica free energy matching these performances with an algorithm can be quite a challenge. These difficulties arise from 2 main reason.

— **Computing the $\mathbf{f}_{\text{in}}(\mathbf{A},\mathbf{B})$** function associated with (4.66) can be a difficult task. Because of matrix A there can be correlations between components of the x_k for different k. As a result this means that the computation of $f_{\text{in}}(A,B)$ requires the sum of 2^r term (one for every term when ones develops expression (4.66)). For $r = 30$ this means that most of the computation time will be spent computing the update functions $f_{\text{in}}(A,B)$. There are ways around this problem and I have found that a simple field approximation was a reasonable approach to approximating this function $f_{\text{in}}(A,B)$.

— **Weak breaking of the rotational symmetry :** The prior P_X is supposed to break the rotational symmetry of the problem (remember that the interaction term $g(Y_{ij},w)$ has a rotation symmetry). The problem we observe when running AMP is that if one starts

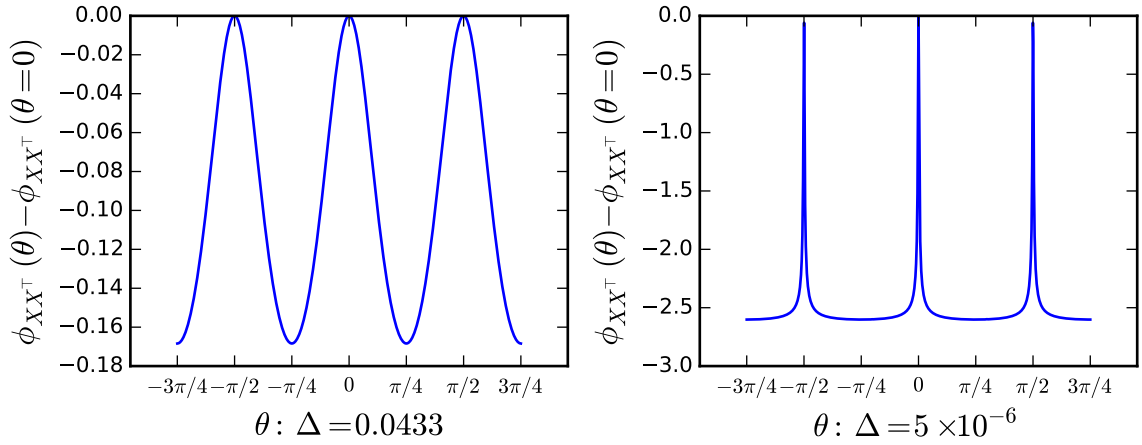


FIGURE 4.16 – In this figure we plot the free energy $\phi_{\text{RS}XX^\top}(M, Q)$ (4.244) for the rank 2 Gauss-Bernoulli (4.66) prior for $r = 2$, $\rho = 0.5$, $Q = mI_2$ and $M^t = mR(\theta)$, where R is a 2×2 rotation matrix of angle θ. We plot the free energy as a function of θ. m is chosen such that the free $\phi(m, \theta = 0)$ is minimal. Essentially this means that we are plotting the Free-energy cost to recovering the signal up to the wrong rotation. The free-energy is maximal when $\theta = k\pi/2$ (right rotation) and minimal when $\theta = k\pi/2 + \pi/4$. The profile of free energy will have different consequence for the AMP algorithm **Left :** Here $\Delta = 0.0433$ even if one end up with a wrong rotation angle θ there is a gradient of free energy in θ that one could follow to get the right value. **Right :** Here the profile in free-energy is much more piqued around $\theta = k\pi/2$ outside of these valleys the free-energy profile gets very flat and the free energy gradient very small. In practice for the algorithm (and therefore a finite N) this means that the AMP equation find a fixed point with the wrong angle θ even though this is not the configuration that actually minimizes the free-energy. The problem gets worse as the rank r increases.

with a high rank $r > 7$ and low Δ then AMP will tend to converge to a configuration where

$$M = m^t R,\ Q = q^t I_r \tag{4.414}$$

Where R will be some rotation matrix that is not a simple permutation matrix (if R was a permutation matrix then one could consider the problem solved). This is strange after all the Gauss-Bernoulli prior P_X is supposed to break the rotational symmetry. And even if one ended up in a configuration of the AMP algorithm given by (4.414) then we would expect that the AMP algorithm would be able to follow some free energy gradient and reach a state where R is just a permutation matrix. Maybe the problem lies with the mean field approximation we did to compute $f_{\text{in}}(A, B)$? It turns out that $f_{\text{in}}(A, B)$ is not the problem. The problem lies in the fact that gradient one is supposed to follow to reach the correct solution can vanish. We illustrate that in fig 4.16 We plot the Replica symmetric free energy for the Gauss-Bernoulli $XX^\top$ case for $r = 2$ for two values of Δ as a function of θ the angle of the rotation matrix $R(\theta)$.
It is clear from this figure that the free-energy is maximized when $\theta = k\pi/2, k \in \mathbb{N}$. If θ is in-between these values then the free energy is lower. If Δ is be enough then there is non vanishing gradient to follow to reach $\theta = 0$. But as Δ gets smaller the profile of free energy gets flatter and the algorithm gets stuck on these plateau. The problem gets worse as we increases the rank r since there are no more ways to get the wrong symmetry. Essentially the algorithm gets stuck in a configuration where it thinks that by

chance every component of x_i is non zero and it behaves as is P_X was a Gaussian prior. Another way to interpret this is to say that if one gets rotation R a little modification of R will not change the free-energy in a significant way. This creates the bizarre situation where it looks as if solving the problem for low Δ is harder than solving it at large Δ, which is what I observe with the AMP experiments. Therefore to solve this problem we need an optimization cost function that will not be constant when one is away from the solution.

4.6.2 Breaking the rotation symmetry using l_1 regularization

One way to deal with sparsity is to introduce a l_1 regularization to the problem. This l_1 regularization will "punish" configuration of x_i that are far away from being sparse. The new optimization problem we want to solve is now.

$$\text{argmin}_{X\in\mathbb{R}^{N\times r}}\frac{1}{2}\sum_{1\leq i,\leq N}\left(Y_{ij}-\frac{x_i x_j^\top}{\sqrt{N}}\right)^2+\lambda\sum_{1\leq i\leq N}\|x_i\|_{l_1} \tag{4.415}$$

Where λ is a parameter that controls how sparse we want our solution to be. Setting $\lambda\to+\infty$ would make it so that the minima of this cost function would be given by $\forall i,\ x_i=0$. Setting $\lambda=0$ make it so that the solution to this problem is the usual PCA solution given by the top r eigenvectors of Y. This can be a tricky problem to solve since it is non convex. But if we only want to use the l_1 regularization to break the rotational symmetry even a very small value of λ would suffice. When λ is very small the l_1 regularization problem translates into the problem

"among the minimizer of $\frac{1}{2}\sum_{1\leq i,\leq N}\left(Y_{ij}-\frac{x_i x_j^\top}{\sqrt{N}}\right)^2$ find the one that minimizes $\sum_{1\leq i\leq N}\|x_i\|_{l_1}$".

This is way simpler problem to solve since all the minimizer of $\frac{1}{2}\sum_{1\leq i,\leq N}\left(Y_{ij}-\frac{x_i x_j^\top}{\sqrt{N}}\right)^2$ are of the form $X_{\text{PCA}}R$ where $X_{\text{PCA}}\in\mathbb{R}^{N\times r}$ is the matrix that contain all the r top eigenvectors of Y and $R\in\mathbb{R}^{r\times r}$ is just a rotational matrix. The cost function we want to minimize is now

$$U(R)=\sum_{1\leq i\leq N}\|Rx_i^{\text{PCA}}\|_{l_1} \tag{4.416}$$

This looks like a much simpler problem to solve since we have gone from a problem over Nr variables to a problem over r^2 variables. In practice it is often convenient to replace the l_1 cost function by the following function

$$U(R,\epsilon)=\sum_{1\leq i\leq N}g(Rx_i^{\text{PCA}}) \tag{4.417}$$

$$g(x,\epsilon)=\sum_{1\leq k\leq r}\sqrt{x_k^2+\epsilon^2} \tag{4.418}$$

$$\text{Where}\,\epsilon\ll 1 \tag{4.419}$$

One has $g(x,\epsilon=0)=\|x\|_{l_1}$. Replacing $\|\|_{l_1}$ with $g(x,\epsilon)$ just yields us with better convergence in the gradient descent algorithm that we then use.

We will treat treat this problem using a gradient descent approach on R.The only difficulty lies with the fact that we need R to be a rotational matrix $RR^\top = I_r$. This constraint will be treated by writing the gradient descent algorithm in a specific way. Let R^t be our estimate of R at time t, we write

$$R^{t+1} = \exp\left(H^t\right) R^t \,. \tag{4.420}$$

Where the exponential is here a matrix exponential and H^t will be an anti symmetric matrix. Looking for R^{t+1} under this form ensures that R^{t+1} will be a rotational matrix. All that remain is then to compute the gradient of $U(\exp\left(H^t\right) R^t, \epsilon)$ in H^t around $H^t = 0$. Using the fact that to first order matrix exponentiation is given by

$$\exp(H^t) = I_r + H^t + O\left(\|H^t\|_2^2\right) \,. \tag{4.421}$$

One can compute that to first order in H^t and get

$$U(\exp\left(H^t\right) R^t, \epsilon) - U(R^t, \epsilon) = \frac{1}{2}\mathrm{Tr}\left[C^t H^t\right] + O\left(\|H^t\|_2^2\right) \,, \tag{4.422}$$

Where

$$C^t = \mathrm{Sign}(X_{\mathrm{PCA}} R^t, \epsilon)^\top X_{\mathrm{PCA}} R^t - {R^t}^\top X_{\mathrm{PCA}}^\top \mathrm{Sign}(X_{\mathrm{PCA}} R^t, \epsilon) X_{\mathrm{PCA}} R^t \,. \tag{4.423}$$

Where $\mathrm{Sign}(\cdots, \epsilon)$ is the function that computes the expression $\frac{x}{\sqrt{x^2+\epsilon^2}}$ of a matrix component by component. C^t is just the gradient of $U(\exp(H)R^t, \epsilon)$ expressed in matrix form. Therefore in our gradient descent algorithm we will set $H^t = \gamma C^t$ and do a line search in γ to minimize the cost function. This yields use with a gradient descent algorithm to solve this problem. In practice this work quite well and if Δ is small enough we indeed get a solution that is well aligned with the hidden solution meaning the $R^\top X_{\mathrm{PCA}}^\top X_0$ is indeed a permutation matrix. One could then use the output of the gradient descent algorithm to initialize an AMP algorithm and try and reach optimal reconstruction performance. In practice it works well and solves the issue for small Δ but there still remain a domain in which I was not able to match the SE performance with the AMP algorithm.

This lead me to think that the difficult issues with sparse PCA have little to do with low-sparsity but more to do with large rank r and the breaking of the rotational symmetry. (It is worth noting that this issue does not appear in tensor factorization problem as the interaction term break the rotational symmetry "for free".)

Conclusion

In this thesis we analyzed the problem of low rank matrix/tensor factorization in the Bayes-optimal setting using the theoretical tools and techniques coming from statistical physics and initially developed to understand spin glasses. These methods gave us algorithms to solve instances of factorization problem and theoretical tools to analyze the behavior of these algorithms. These theoretical tools allowed us to explore different systems and establish what we like to call a "zoology" of phase diagrams. I hope this zoology will help the reader reach a qualitative understanding of the factorization problem. Nevertheless a few questions remain open. Here they are presented in a separate format.

Is the AMP algorithm useful for practical inference problems ? : At least for the matrix factorization problem one might be tempted to ask the question "*Why should anyone care about Bayes-optimal factorization of matrices ? After all it appears clearly from your analysis that spectral method often give performances that are comparable with Bayes-optimal performances.*" This is a valid remark and it is true that for some type of signal (defined by P_{X_0} or P_{U_0}, P_{V_0}) there is often little or no difference between the theoretical performance of AMP and PCA. Trying to use AMP on real data creates the need to learn the parameters of the system. The system for which the performance of AMP and PCA differed greatly always turned up to be systems where the signal was highly structured, it was very sparse or it could take only a few well separated value. Essentially, AMP shined whenever there was a structure in the data that could be used to decrease greatly the reconstruction error. AMP for inference problem should not be compared with off-the-shelf unsupervised learning method, but should be thought of as a lever one can use to transform prior information about the data into much lower reconstruction error in cases where this is possible. It is my hope that the zoology of phase diagrams presented here should help the reader get a sense of when and where AMP techniques are worth the trouble.

In the tensor factorization setting things are a bit more clear, computing the spectrum of tensor for $p \geq 3$ is an NP-hard problem [HL09]. There are methods that try to go around this NP-hardness, such as unfolding a tensor or doing a gradient descent on the likelihood but they all require a noise Δ level that is order of magnitude lower ($\Delta = O(n^{(2-p)/2})$) in practice than the information theoretical threshold $\Delta_{\text{IT}} = O(1)$ [RM14]. The analysis of the SE equations gives us a physical explanation to this Hardness but it also tells us when the AMP equations will reach the optimal reconstruction. In that case (at least for synthetic data) AMP method by taking advantage of prior information can potentially vastly outperform other "spectral" methods.

Is the analysis rigorous ? The short answer is "yes". Recently, proofs of the replica formula appeared as long as one remains in the Bayes-optimal setting (on the Nishimori line). Bayes-optimality is an important property required by the main results of the papers presented in the next paragraph. The proof of the replica formula did not come in one go but was proven step by step.

The first proofs dealt with showing that the behavior of the AMP algorithm was indeed described by SE equations (also sometime called single letter characterization). This was established for the rank 1 $UV^\top$ case in [RF12] and for additive Gaussian noise. In [JM13] Javanmard and Montanari proved that the behavior of a class of AMP algorithm with interaction matrices that are Gaussian could indeed be described by SE equations.

Being able to prove the validity of the SE equations just gives an upper bound on the reconstruction error one can hope to achieve. To compute rigorously the theoretical optimal reconstruction one has essentially to prove that the replica symmetric free energy formula is correct. Deshpande and Montanari proved in 2014 that the optimal reconstruction was reached by an AMP algorithm in the Bernoulli $XX^\top$ case for $\rho > 0.05$ [DM14a]. Their proof did not work for all priors P_X or even for all ρ. The reason why, is that first order phenomena that can appear in those systems do not only create difficulties for algorithms, but they also make it hard for proofs techniques to work as-well. Finally the proof of the replica formula (even where $\phi_{XX^\top}$ is not a convex function) was established in [BDM$^+$16] for the rank one $XX^\top$ case. One of the important ingredient of the proof was the spatial coupling method, whose role is to destroy the first order phenomena all while hardly changing the free energy and therefore making the free-energy function convex. The $XX^\top$, $UV^\top$ and tensor case for higher rank were proven in [ML16, Mio17a, LML$^+$17].

The property of universality of the noise channel were proven in [DAM16, KXZ16].

What to do with these results ? What are the interesting open problems. As of today I see three axes along which this work could provide us with interesting questions and problems.

— **The dense case as a tool to understand sparse problems :** The fact that the systems we analyzed here were dense rather than sparse was the source of a great number of simplifications. State Evolution (SE) equations with their Gaussian distributed messages are order of magnitude easier to analyze and to interpret than Density Evolution (DE) equations (where one has to resort to population dynamics). For instance in the community detection problem we were able to summarize the state of the whole system into two scalars b^t, Δ and a function $\mathcal{M}_r$, which allowed us not only to predict what the AMP equations would give us but also to understand it in a way that population dynamics would not allow us to. Of course sparse and dense regime do not have the same properties and the same phase diagrams. However I have found that when presented with a new sparse estimation problem translating it into its dense version and then getting a grasp of the properties of the dense version was often an easy (and labor cheap) way to get a first understanding of the scaling laws and phase transition that one can expect in the sparse regime. An example of a property translating well from the sparse to the dense regime be given by 4.365 criteria can be also used in the sparse limit where $p_{\text{out}} = O(1/N)$ [DKMZ11a]. Of course this parallel between the sparse and dense regime might not hold or may require for the average degree to grow to infinity

sufficiently fast with the size of the system. For example in the community detection problem where $p_{\text{in}} = O(1/N)$, $p_{\text{out}} = O(1/N)$, and $r = 5$ by translating directly the properties of the dense regime to the sparse regime one might predict to see a discontinuity of the reconstruction error when $n|p_{\text{out}} - p_{\text{out}}| = r\sqrt{n/r(p_{\text{in}} + p_{\text{out}}(r-1))}$, but in the disassortative case where $p_{\text{out}} > p_{\text{in}}$ no such jump occurs if the average degree is small enough $n/r(p_{\text{in}} + p_{\text{out}}(r-1))$, in contradiction with the properties of the dense regime. Often taking the average degree to grow with the size of the system (but not grow too quickly) is enough to smooth the differences between the properties of the dense and the sparse regime, for example the balanced communities problem analyzed in 4.5.1 has very similar properties when analyzed in the regime where the average degree remains small but still grows to infinity $p_{\text{out}} = O(\log(n)/n)$[CLM16].

— **Large rank matrix factorization $\mathbf{r} = \mathbf{O}(\mathbf{N})$:** What happens when the rank r grows with N, $r = O(N)$? This problem is related to the dictionary learning and sparse coding problem [KDMR$^+$03, OF96]. This, at least from an algorithmic point of view, is a much harder problem to treat. AMP equations that can be derived for this problem do not always converge to a fixed point described by the replica computation [KKM$^+$14]. After having explored this problem for several months during my PhD (and failing to solve it). I think that the difficulties in that problem arise from us not being able to treat the rotation symmetry in this problem both from an algorithmic point of view and a theoretical point of view. The analysis of the r is large but no too large limit with the AMP equations could be useful in understanding this problem. A breakthrough in that problem would, I'm sure, yield a great deal of interesting results because of the link between large rank matrix factorization and the training of neural networks.

— **Other matrix ensembles :** In this whole thesis we always assumed that the interaction matrices Y_{ij} had its components sampled independently. This assumption allowed the Bethe approximation we made to be exact in the large N limit and made it so that the typical error we made on the estimation of the marginals was of order $O(1/\sqrt{N})$. But if these AMP techniques are to be used in order to do inference they will need to be robust and adapt to the situation. For instance in the Hopfield model the correlation between the J_{ij} changes the shape of the correct TAP equations [Méz17]. Techniques such as adaptative TAP try to tackle this problem by learning the correct Onsager reaction term on the fly [OW01] in order to compute the marginals of the system. Recently a method to treat problem of compressed sensing for all sensing matrices that are right rotationally invariant [SRF16] was derived. Maybe an algorithm inspired from this techniques could provide us with a path to a reliable and robust AMP algorithm.

Résumé en français de la thèse : thesis summary in french

Introduction

Depuis quelques années la combinaison d'avancées techniques dans le domaine des statistiques et la production de données en grande quantité dans différents domaines de l'industrie notamment informatiques (Google ,Facebook et autre), a permis au domaine du machine learning de prospérer. Le machine learning peut être vu soit comme un domaine de l'ingénierie soit comme un autre terme pour désigner le domaine des statistiques. On divise en général les problèmes de machine learning en 3 classes.

— **L'apprentissage supervisé** a pour but d'apprendre une fonction $f(x)$ à partir de M exemples de couples (x_i, y_i) où $\forall i, f(x_i) \approx y_i$. En pratique résoudre un tel problème revient souvent à résoudre un problème d'optimisation sur la fonction f.

$$f^* = \operatorname{argmin} R(f) = \operatorname{argmin} \left\{ \frac{1}{N} \sum_{1 \leq i \leq N} L(f(x_i), y_i) + \Lambda(f) \right\} \tag{4.424}$$

$L(.,.)$ est une distance qui assure que $f(x_i)$ reste près de y_i. Pour que le problème ait un intérêt, l'espace dans lequel on recherche f doit être réduit. Cela peut être accompli de 2 façons. On peut rechercher une fonction f qui s'écrit d'une certaine façon par exemple une fonction linéaire en x , $f(x) = \beta^\top x + c$ où $x, \beta \in \mathbb{R}^r \, c \in \mathbb{R}$. Le terme Λ a aussi pour but de restreindre ou biaiser la recherche de f en pénalisant certaines valeurs de f.

— **L'apprentissage non supervisé** a pour but d'analyser des données dans des situations dans lesquelles on n'a pas accès à une valeur de fonction y_i. Dans le cas d'apprentissage non supervisé on a simplement accès aux données $\{x_i\}$ et on cherche à savoir si ces données exhibent une structure particulière. Pour simplifier là ou l'apprentissage supervisé cherche à apprendre des fonctions, l'apprentissage non supervisé cherche à apprendre des densités de probabilités.

— **L'apprentissage par renforcement** est un type d'apprentissage automatique où le but est de faire apprendre une "stratégie" à un agent interagissant avec un environnement.

Les distinctions que l'on pourrait faire entre les différents types de problèmes en apprentissage automatique sont plus affaires de conventions que de vraies différences dans les outils théoriques et algorithmiques que l'on devrait déployer pour les résoudre. En effet, la grande majorité des problèmes d'apprentissages automatiques peuvent être transcris en un problème d'optimisation mathématique à résoudre.

Étonnamment les méthodes d'apprentissages automatiques ont été couronnées de succès dans les dernières années. On arrive aujourd'hui en résolvant certains problèmes d'optimisations par descente de gradient à apprendre des fonctions qui déterminent étant donnée une image si cette image contient un chien ou un chat.

Si les méthodes d'apprentissages automatiques ont réussi à résoudre des problèmes d'une grande complexité, on ne sait pas vraiment pourquoi elles fonctionnent. Les résultats théoriques sur les méthodes d'apprentissages automatiques restent insuffisantes pour comprendre par exemple pourquoi il est possible d'entrainer des réseaux de neurones profonds. Dans ce manuscrit je m'efforce d'apporter un début de réponse à cette question. Le problème que je cherche à résoudre est le problème de factorisation de matrice de petit rang. Essentiellemnent cela revient

à résoudre le problème d'inférence suivant.

$$Y = XX^\top + \text{Bruit} \tag{4.425}$$

On veut retrouver (inférer) X dans une situation où l'on observe Y. La raison pour laquelle c'est un problème intéressant est que de nombreux problèmes d'apprentissages automatiques (pour la plupart d'apprentissages non supervisés) peuvent être traduit en problème de factorisation de matrice. Nous chercherons à répondre à des questions telles que

— Étant donné ce problème de factorisation de matrice. Comment le résoudre ?
— Quelle est la meilleure reconstruction qu'il est théoriquement possible d'atteindre ?
— Comment atteindre ces performances optimales ?

Pour pouvoir répondre à ces questions avec un modèle théorique nous aurons besoin d'un modèle génératif pour décrire la matrice Y. Nous utiliserons l'inférence Bayésienne pour répondre aux questions plus haut.

Cette thèse se trouvant à l'intersection du domaine des statistiques et de la physique des verres de spin Nous analyserons des modèles d'inférences statistiques à l'aide d'outils théoriques développés pour analyser des systèmes de verres de spin.

Le problème que l'on veut résoudre.

Le problème que l'on veut résoudre est le suivant

$$\begin{aligned} Y &= XX^\top + \text{Bruit} \in \mathbb{R}^{N\times N} \\ Y &= UV^\top + \text{Bruit} \in \mathbb{R}^{N\times M} \\ Y &= X \circ X \circ \cdots \circ X + \text{Bruit} \in \mathbb{R}^{N\times M} \end{aligned}$$

Y est une matrice provenant de la multiplication de 2 matrices $U \in \mathbb{R}^{N\times r}$ et $V \in \mathbb{R}^{M\times r}$ et auxquelles on a rajouté du bruit (le bruit n'a d'ailleurs pas besoin d'être Gaussien). On observe Y et l'on veut reconstruire X ou U, V. Nous nous limiterons dans ce résumé au cas $XX^\top$.

Le modèle génératif sera choisi tel que chaque ligne $x_i \in \mathbb{R}^{r\times 1}$ de X a été tirée au hasard d'une densité de probabilité $P_X(x_i)$. Puis l'on fait passer chaque coordonnée de la matrice $XX^\top$ à travers un canal décrit par une densité de probabilité $P_{\text{out}}(Y|w)$. Cela nous donne accès à la probabilité postérieure de X connaissant Y.

$$P(Y) = \prod_{1\le i\le N} P_{X_0}(x_i) \prod_{1\le i<j\le N} P_{\text{out}}\left(Y_{ij}\middle|\frac{x_i^\top x_j}{\sqrt{N}}\right). \tag{4.426}$$

$$P(X|Y) = \frac{1}{Z_X(Y)} \prod_{1\le i\le N} P_{X_0}(x_i) \prod_{1\le i<j\le N} P_{\text{out}}\left(Y_{ij}\middle|\frac{x_i^\top x_j}{\sqrt{N}}\right). \tag{4.427}$$

Nous travaillerons dans le cadre de l'inférence bayésienne où nous supposerons que Y a été créé à partir du même modèle que l'on essaye d'ajuster aux données Y. Nous appellerons X_0 la solution plantée, c'est à dire la vraie valeur de X que l'on essaie de reconstruire. L'observable $\hat{X}(Y)$ (fonction de Y) qui minimise la distance à la solution plantée est la fonction qui calcule la moyenne de X selon la probabilité postérieur $P(Y|X)$.

Le problème est que calculer ces valeurs moyennes est un calcul qui nécessite un nombre exponentiel d'opérations en N. Si l'on veut résoudre des instances données de ce problème l'on doit avoir des méthodes d'estimations qui ne soient pas exponentielles en la taille du système.

Algorithme de reconstruction : Low-RAMP

Un algorithme pour estimer les marginales des variables x_i est donné par l'algorithme Low-RAMP. Donné par les equations suivantes

$$B_{X,i}^t = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} S_{ki} \hat{x}_k^t - \frac{\hat{x}_i^{t-1}}{\Delta N} \sum_{k=1}^{N} \sigma_{x,k}^t \,, \tag{4.428}$$

$$A_X^t = \frac{1}{N\Delta} \sum_{k=1}^{N} \hat{x}_k^t \hat{x}_k^{t,\top} \,, \tag{4.429}$$

$$\hat{x}_i^{t+1} = f_{\text{in}}^x(A_{X,i}^t, B_{X,i}^t) \,, \tag{4.430}$$

$$\sigma_{x,i}^{t+1} = \frac{\partial f_{\text{in}}^x}{\partial B}(A_{X,i}^t, B_{X,i}^t) \,. \tag{4.431}$$

où

$$S_{ij} = \frac{\partial \log P_{\text{out}}(Y|w)}{\partial w} \quad , \frac{1}{\Delta} = \mathbb{E}_{Y \partial P_{\text{out}}(Y|w=0)} \left[\left(\frac{\partial \log P_{\text{out}}(Y|w)}{\partial w} \right)^2 \right] \tag{4.432}$$

$$f_{\text{in}}^x(A,B) \equiv \frac{\partial}{\partial B} \log \left(\int \mathrm{d}x P_X(x) \exp\left(B^\top x - \frac{x^\top A x}{2} \right) \right) = \frac{\int \mathrm{d}x P_X(x) \exp\left(B^\top x - \frac{x^\top A x}{2} \right) x}{\int \mathrm{d}x P_X(x) \exp\left(B^\top x - \frac{x^\top A x}{2} \right)} \tag{4.433}$$

A chaque itération de l'algorithme, on calcule une estimation de la moyenne $(\hat{x}_i^t)$ et de la matrice de covariance $(\sigma_{x,i}^t)$ des variables x_i à l'aide la fonction f_{in}^x (si les x_i étaient des spins ± 1 alors $f_{\text{in}}^x(A,B) = \tanh(B)$). Cette algorithme est essentiellement le même algorithme que l'on obtiendrait en écrivant les équations mean-field naïve. On peut résumer les interactions de l'ensemble du systèmes sur une variable par un champs distribué aléatoirement selon une Gaussienne.

La seule différence avec les équations mean-field se trouve dans les formules équations (4.4284.429)

Les équations de State Evolution

L'algorithme Low-RAMP nous a permis de réaliser des expériences numériques. Toutefois l'on aimerait analyser les performances de cet algorithme quand la taille du système N tend vers $+\infty$. L'idée derrière les équations de State-Evolution est de décrire l'état de l'algorithme à l'aide d'un nombre fini de paramètres d'ordre.

$$M_x^t = \frac{1}{N} \sum_{1 \le i \le N} \hat{x}_i^t x_i^{0,\top} \in \mathbb{R}^{r \times r}, \quad Q_x^t = \frac{1}{N} \sum_{1 \le i \le N} \hat{x}_i^t \hat{x}_i^{t,\top} \in \mathbb{R}^{r \times r} \,. \tag{4.434}$$

On peut ensuite calculer comment ces paramètres d'ordre évoluent après une itération de l'algorithme.

$$M_x^{t+1} = \mathbb{E}_{x_0,W}\left[f_{\text{in}}^x\left(\frac{Q_x^t}{\Delta}, \frac{M_x^t}{\Delta}x_0 + \sqrt{\frac{Q_x^t}{\Delta}}W\right)x_0^\top\right], \tag{4.435}$$

$$Q_x^{t+1} = \mathbb{E}_{x_0,W}\left[f_{\text{in}}^x\left(\frac{Q_x^t}{\Delta}, \frac{M_x^t}{\Delta}x_0 + \sqrt{\frac{Q_x^t}{\Delta}}W\right)f_{\text{in}}^x(\cdots,\cdots)^\top\right], \tag{4.436}$$

où

$$W \sim \mathcal{N}(0, I_r) \quad , x_0 \sim P_X(x) \tag{4.437}$$

En utilisant la méthode des répliques on peut aussi calculer l'énergie libre associée à chacun de ces paramètres d'ordres.

$$\phi_{\text{RS,XX}^\top}(M_x) = \mathbb{E}_{W,x_0}\left[\log\left(Z_x\left(\frac{M_x}{\Delta}, \frac{M_x}{\Delta}x_0 + \sqrt{\frac{M_x}{\Delta}}W\right)\right)\right] - \frac{\text{Tr}(M_x M_x^\top)}{4\Delta}. \tag{4.438}$$

où

$$Z_x(A,B) = \int \mathrm{d}x P_X(x)\exp\left(B^\top x - \frac{x^\top A x}{2}\right) \tag{4.439}$$

Une zoologie de diagramme de phase

En fonction de Δ et $P_x(x)$ les systèmes étudiés dans cette thèse présentent une variété de comportements. J'ai étudié les propriétés de nombreux systèmes définis par leur prior P_X.

Une des propriétés les plus importantes des ces diagrammes de phase est la présence où non d'une phase dite difficile (HARD en anglais). Dans tous les diagrammes de phase suivant on fera la distinction entre

- La phase dite facile (EASY en anglais). Dans cette phase l'algorithme Low-RAMP est capable de reconstruire le signal X_0 avec des performances qui sont optimales du point de vue de la théorie de l'information. De plus ces performances sont non triviales.
- Dans la phase impossible (IMPOSSIBLE en anglais) l'algorithme Low-RAMP est capable de reconstruire le signal de façon optimale. Malheureusement car le ratio signal sur bruit est trop important le signal X_0 est impossible à reconstruire.
- Dans la phase dite dure (HARD en anglais). Il y a une grande différence entre les performances de l'algorithme Low-RAMP et les performances optimales que l'on pourrait obtenir en calculant les moyennes exactes sur la probabilité postérieur $P(X|Y)$.

Je trace un exemple d'un tel diagramme de phase dans la fig 4.17 pour le cas $XX^\top$ et où $P_X(x)$ est donné par

$$P_X(x) = \rho\delta\left(x - \sqrt{\frac{1-\rho}{\rho}}\right) + (1-\rho)\delta\left(x + \sqrt{\frac{\rho}{1-\rho}}\right). \tag{4.440}$$

Cette densité de probabilité peut prendre 2 valeurs chacune avec une probabilité ρ et $1-\rho$ les 2 valeurs prises sont telles que P_x a une moyenne 0 et une variance de 1.

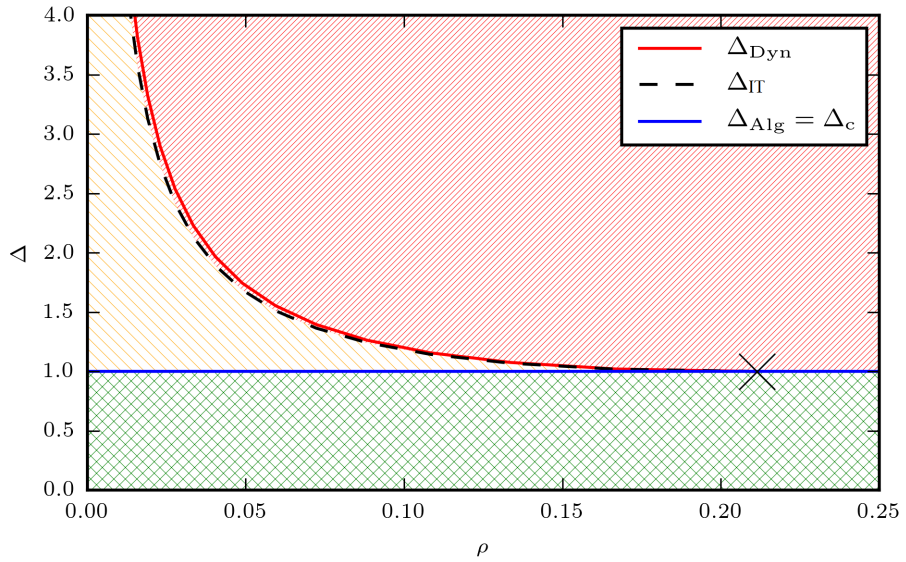


FIGURE 4.17 – Je trace ici le diagramme de phase pour 2 communautés en fonction de ρ et Δ pour un prior P_x donné par (4.440). On y voit 3 phases, la phase facile en vert où la reconstruction du signal est possible et l'on peut atteindre la reconstruction optimale, la phase impossible en rouge où l'on ne peut pas reconstruire le signal car il est impossible de le reconstruire et la phase difficile en orange où il est théoriquement possible de reconstruire le signal mais où cette reconstruction est impossible avec l'algorithme Low-RAMP.

Conclusion

Dans cette thèse je me suis intéressé au problème de factorisation de matrice et tenseur de petit rang. J'ai pu analyser ce problème en utilisant des outils issus de la physique des verres de spins. Bien que dans toute cette analyse je me sois restreint à rester dans un cadre Bayésien optimal ce qui assure de rester dans une phase où la symétrie des répliques n'est pas brisée. On voit que ces systèmes peuvent avoir une grand nombre de comportements. L'analyse des équations de State Evolution nous a permis de mettre en évidence des phénomènes de transition du premier ordre dans ces systèmes. Cette analyse peut aussi nous renseigner sur l'arbitrage à faire sur le rapport coût de calcul sur performance que proposent les méthodes d'inférence Bayésienne.

Appendices

Mean Field equations

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

$$B^t_{X,i} = \sum_{k=1}^{N} \frac{1}{\sqrt{N}} S_{ki} \hat{x}^t_k \,, \tag{441}$$

$$A^t_{X,i} = \frac{1}{N} \sum_{k=1}^{N} (S^2_{ki} - R_{ki}) \left(\hat{x}^t_k \hat{x}^{t,\top}_k + \sigma^t_{x,k} \right) \,, \tag{442}$$

$$\hat{x}^{t+1}_i = f^x_{\rm in}(A^t_{X,i}, B^t_{X,i}) \,, \tag{443}$$

$$\sigma^{t+1}_{x,i} = \frac{\partial f^x_{\rm in}}{\partial B}(A^t_{X,i}, B^t_{X,i}) \,. \tag{444}$$

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

$$F^{\rm MF}_{XX^\top}(\{A_{X,i}\}, \{B_{X,i}\}) = \sum_{1 \le i \le N} \log(\mathcal{Z}_x(A_{X,i}, B_{X,i})) - B^\top_{X,i} \hat{x}_i + \frac{1}{2} {\rm Tr} \left[A_{X,i} (\hat{x}_i \hat{x}_i^\top + \sigma_{x,i}) \right]$$
$$+ \frac{1}{2} \sum_{1 \le i,j \le N} \left[\frac{1}{\sqrt{N}} S_{ij} \hat{x}_i^\top \hat{x}_j + \frac{(R_{ij} - S^2_{ij})}{2N} {\rm Tr} \left[(\hat{x}_i \hat{x}_i^\top + \sigma_{x,i})(\hat{x}_j \hat{x}_j^\top + \sigma_{x,j}) \right] \right] . \tag{445}$$

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

$$B^t_{U,i} = \frac{1}{\sqrt{N}} \sum_{l=1}^{M} S_{il} \hat{v}^t_l \,, \tag{446}$$

$$A^t_U = \frac{1}{N} \sum_{l=1}^{M} \left(S^2_{il} - R_{il} \right) \left(\hat{v}^t_l \hat{v}^{t,\top}_l + \sigma^t_{u,l} \right) \,, \tag{447}$$

$$\hat{u}^t_i = f^u_{\rm in}(A^t_U, B^t_{U,i}) \,, \tag{448}$$

$$\sigma^t_{u,i} = \left(\frac{\partial f^u_{\rm in}}{\partial B} \right) (A^t_U, B^t_{U,i}) \,, \tag{449}$$

$$B^t_{V,j} = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} S_{kj} \hat{u}^t_k \,, \tag{450}$$

$$A^t_{V,j} = \frac{1}{N} \sum_{k=1}^{N} (S^2_{kj} - R_{kj}) \left(\hat{u}^t_k \hat{u}^{t,\top}_k + \sigma^t_{u,k} \right) \,, \tag{451}$$

$$\hat{v}^{t+1}_j = f^v_{\rm in}(A^t_{V,j}, B^t_{V,j}) \,, \tag{452}$$

$$\sigma^{t+1}_{v,j} = \left(\frac{\partial f^v_{\rm in}}{\partial B} \right) (A^t_{V,j}, B^t_{V,j}) \,. \tag{453}$$

The mean field free energy for the bipartite case reads

$$\begin{aligned} F^{\text{MF}}_{UV^\top}(\{A_{U,i}\},\{B_{U,i}\},\{A_{V,j}\},\{B_{V,j}\}) = \\ \sum_{1\le i\le N} \log(\mathcal{Z}_u(A_{U,i},B_{U,i})) - {B_{U,i}}^\top \hat{u}_i + \frac{1}{2}\text{Tr}\left[A_{U,i}(\hat{u}_i\hat{u}_i^\top + \sigma_{u,i})\right] \\ + \sum_{1\le j\le M} \log(\mathcal{Z}_v(A_{V,j},B_{V,j})) - {B_{V,j}}^\top \hat{v}_j + \frac{1}{2}\text{Tr}\left[A_{V,j}(\hat{v}_j\hat{v}_j^\top + \sigma_{v,j})\right] \\ + \sum_{1\le i\le N, 1\le j\le M} \left[\frac{1}{\sqrt{N}} S_{ij}\hat{u}_i^\top\hat{v}_j + \frac{1}{2N}(R_{ij}-S_{ij}^2)\text{Tr}\left[(\hat{u}_i\hat{u}_i^\top+\sigma_{u,i})(\hat{v}_j\hat{v}_j^\top+\sigma_{v,j})\right]\right]. \end{aligned} \tag{454}$$

The difference between the mean field equations and the Low-RAMP equations from section 4.2.1 and 4.2.3 can be seen for both variables A and B.

Replica computation $UV^\top$ case.

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

$$\mathcal{Z}(\{Y_{ij}\}) = \int \prod_i \mathrm{d}u_i P_U(u_i) \prod_j \mathrm{d}v_j P_U(v_j) \prod_{ij} \exp\left(g\left(Y_{ij}, \frac{u_i^\top v_j}{\sqrt{N}}\right) - g\left(Y_{ij},0\right)\right). \tag{455}$$

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

$$\langle\log(\mathcal{Z}(\{Y_{ij}\}))\rangle_Y = \lim_{n\to 0}\frac{\langle\mathcal{Z}^n\rangle - 1}{n} = \lim_{n\to 0}\frac{\log(\langle\mathcal{Z}^n\rangle)}{n}. \tag{456}$$

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

$$\mathcal{Z}^n(\{Y_{ij}\}) = \int \prod_{a=1\cdots n}\prod_{i=1\cdots N} \mathrm{d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} \mathrm{d}v_j^a P_V^a(v_j^a) \exp \prod_{\substack{i=1\cdots N\\ j=1\cdots M}} \exp\left(g\left(Y_{ij}, \frac{u_i v_j}{\sqrt{N}}\right) - g\left(Y_{ij},0\right)\right). \tag{457}$$

Therefore one has

$$\begin{aligned} \mathbb{E}(\mathcal{Z}^n) = \int \prod_{i=1\cdots N, j=1\cdots M} \mathrm{d}Y_{ij} P_{\text{out}}\left(Y_{ij}, w=0\right) \prod_{a=0\cdots n}\left(\left[\prod_{i=1\cdots N}\mathrm{d}u_i^a P_U^a(u_i^a)\prod_{j=1\cdots M}\mathrm{d}v_j^a P_V^a(v_j^a)\right]\right. \\ \left.\left[\prod_{i=1\cdots N, j=1\cdots M}\exp\left(\sum_{a=0\cdots n} g^a\left(Y_{ij}, \frac{u_i^a v_j^a}{\sqrt{N}}\right) - g^a\left(Y_{ij},0\right)\right)\right]\right), \end{aligned} \tag{458}$$

where

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

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

$$\mathbb{E}(\mathcal{Z}^n) = \int \prod_{i=1\cdots N, j=1\cdots M} \mathrm{d}Y_{ij} P_{\text{out}}\left(Y_{ij}, w=0\right) \prod_{a=0\cdots n} \left(\left[\prod_{i=1\cdots N} \mathrm{d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} \mathrm{d}v_j^a P_V^a(v_j^a) \right] \left[\prod_{i=1\cdots N, j=1\cdots M} \exp\left(\sum_{a=0\cdots n} \left(\frac{\partial g^a}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a}{\sqrt{N}} + \left(\frac{\partial^2 g^a}{\partial w^2}\right)_{Y_{ij},0} \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^{1.5}}\right) \right) \right] \right) . \quad (459)$$

By expanding the exponential to order two one gets

$$\mathbb{E}(\mathcal{Z}^n) = \int \prod_{i=1\cdots N, j=1\cdots M} \mathrm{d}Y_{ij} P_{\text{out}}\left(Y_{ij}, w=0\right) \prod_{a=0\cdots n} \left[\prod_{i=1\cdots N} \mathrm{d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} \mathrm{d}v_j^a P_V^a(v_j^a) \right] \prod_{i=1\cdots N, j=1\cdots M} \left[1 + \sum_{a=0\cdots n} \left(\frac{\partial g^a}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a}{\sqrt{N}} + \sum_{a=1\cdots n} \left(\frac{\partial g}{\partial w}\right)_{Y_{ij},0} \left(\frac{\partial g^0}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a u_i^0 v_j^0}{N} + \sum_{1\le a<b\le n} \left(\frac{\partial g}{\partial w}\right)_{Y_{ij},0} \left(\frac{\partial g}{\partial w}\right)_{Y_{ij},0} \frac{u_i^a v_j^a u_i^b v_j^b}{N} + \sum_{a=0\cdots n} \left[\left(\frac{\partial^2 g^a}{\partial w^2}\right)_{Y_{ij},0} + \left(\frac{\partial^2 g^a}{\partial w^2}\right)^2_{Y_{ij},0} \right] \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^{1.5}}\right) \right] . \quad (460)$$

By averaging with respect to the Y_{ij} one gets

$$\begin{aligned} \mathbb{E}(\mathcal{Z}^n) &= \int \prod_{a=0\cdots n} \left[\prod_{i=1\cdots N} \mathrm{d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} \mathrm{d}v_j^a P_V^a(v_j^a) \right] \prod_{i=1\cdots N, j=1\cdots M} \left[1 + \sum_{a=1\cdots n} \frac{u_i^a v_j^a u_i^0 v_j^0}{N\widehat{\Delta}} + \sum_{1\le a<b\le n} \frac{u_i^a v_j^a u_i^b v_j^b}{N\widetilde{\Delta}} + \sum_{a=1\cdots n} \overline{R} \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^{1.5}}\right) \right] . \\ &= \int \left[\prod_{a=0\cdots n} \prod_{i=1\cdots N} \mathrm{d}u_i^a P_U^a(u_i^a) \prod_{j=1\cdots M} \mathrm{d}v_j^a P_V^a(v_j^a) \right] \\ &\exp\left(\sum_{i=1\cdots N, j=1\cdots M} \sum_{a=1\cdots n} \frac{u_i^a v_j^a u_i^0 v_j^0}{N\widehat{\Delta}} + \sum_{1\le a<b\le n} \frac{u_i^a v_j^a u_i^b v_j^b}{N\widetilde{\Delta}} + \sum_{a=1\cdots n} \overline{R} \frac{(u_i^a v_j^a)^2}{2N} + O\left(\frac{1}{N^{1.5}}\right) \right) . \end{aligned} \quad (461)$$

We now introduce the order parameters

$$q_{ab}^u = \frac{1}{N} \sum_{i=1\cdots N} u_i^a u_i^b , \quad (462)$$

$$q_{ab}^v = \frac{1}{M} \sum_{j=1\cdots M} v_j^a v_j^b . \quad (463)$$

This leads to

$$\mathbb{E}(\mathcal{Z}^n) = NM \int \prod_{0\le a\le b\le n} \mathrm{d}q^u_{ab}\mathrm{d}q^v_{ab} \exp\left(\frac{N\alpha}{\widehat{\Delta}}\sum_{a=1\cdots n} q^u_{a0}q^v_{a0} + \frac{N\alpha}{\widetilde{\Delta}}\sum_{1\le a<b\le n} q^u_{ab}q^v_{ab} + \frac{N\alpha}{2}\sum_{a=\cdots n}\overline{R}q^u_{aa}q^v_{aa}\right)\hat{I}_u(\{q^u_{ab}\})\hat{I}_v(\{q^v_{ab}\})\,, \quad (464)$$

where

$$\hat{I}_u(\{q^u_{ab}\}) = \int\left(\prod_{a=1\cdots n}\prod_{i=1\cdots N}\mathrm{d}u^a_i P^a_V(u^a_i)\right)\prod_{0\le a\le b\le n}\delta\left(\sum_{i=1\cdots N}u^a_i u^b_i - Nq^u_{ab}\right), \quad (465)$$

and

$$\hat{I}_v(\{q^v_{ab}\}) = \int\left(\prod_{a=1\cdots n}\prod_{j=1\cdots M}\mathrm{d}v^a_j P^a_V(v^a_j)\right)\prod_{0\le a\le b\le n}\delta\left(\sum_{i=1\cdots N}v^a_j v^b_j - Mq^v_{ab}\right). \quad (466)$$

Here $\hat{I}_u(\{q^u_{ab}\})$ and $\hat{I}_v(\{q^v_{ab}\})$ are the entropy costs one pays in order for the order parameters to take one specific value. We can treat this constraint by going to Fourier space and then rotating the path of integration

$$I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\}) = \int\prod_{a=0\cdots n}P^a_U(u_a)\mathrm{d}u^a\exp\left(\sum_{0\le a\le b\le n}\hat{q}^u_{ab}(u^a u^b - q^u_{ab})\right). \quad (467)$$

$$I_v(\{q^v_{ab}\},\{\hat{q}^v_{ab}\}) = \int\prod_{a=0\cdots n}P^a_V(v_a)\mathrm{d}v^a\exp\left(\sum_{0\le a\le b\le n}\hat{q}^v_{ab}(v^a v^b - q^v_{ab})\right). \quad (468)$$

Therefore

$$\mathbb{E}(\mathcal{Z}^n) = NM\int\prod_{0\le a\le b\le n}\mathrm{d}q^u_{ab}\mathrm{d}\hat{q}^u_{ab}\mathrm{d}q^v_{ab}\mathrm{d}\hat{q}^v_{ab} \exp\left(\frac{N\alpha}{\widehat{\Delta}}\sum_{a=1\cdots n}q^u_{a0}q^v_{a0} + \frac{N\alpha}{\widetilde{\Delta}}\sum_{1\le a<b\le n}q^u_{ab}q^v_{ab} + \frac{N\alpha}{2}\sum_{a=\cdots n}\overline{R}q^u_{aa}q^v_{aa}\right) I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\}))^N I_v(\{q^v_{ab}\},\{\hat{q}^v_{ab}\}))^M \quad (469)$$

We need to extremize this function with respect to all variables. By taking the derivative equal to 0 with respect to variables q^u_{ab} and q^v_{ab} we get

$$\hat{q}^u_{a0} = \frac{\alpha\hat{q}^v_{a0}}{\widehat{\Delta}}, \qquad \hat{q}^v_{a0} = \frac{\hat{q}^u_{a0}}{\widehat{\Delta}}\,, \quad (470)$$

$$\forall 1\le a<b\le n, \qquad \hat{q}^u_{ab} = \frac{\alpha\hat{q}^v_{ab}}{\widetilde{\Delta}}, \qquad \hat{q}^v_{ab} = \frac{\hat{q}^u_{ab}}{\widetilde{\Delta}}\,, \quad (471)$$

$$\hat{q}^u_{aa} = \overline{R}\alpha\hat{q}^v_{ab}, \qquad \hat{q}^v_{ab} = \overline{R}\hat{q}^u_{ab}\,. \quad (472)$$

We now assume the replica symmetric ansatz

$$\forall a,b>1, \quad q^u_{ab} = \delta^a_b(\Sigma_u + Q_u) + (1-\delta^a_b)Q_u\,, \tag{473}$$
$$\forall a>1, \quad q^u_{a0} = M_u\,, \tag{474}$$
$$\forall a,b>1, \quad q^v_{ab} = \delta^a_b(\Sigma_v + Q_v) + (1-\delta^a_b)Q_v\,, \tag{475}$$
$$\forall a>1, \quad q^v_{a0} = M_v\,. \tag{476}$$

We can then express the free energy. Let us compute I_u and I_v in that RS ansatz

$$I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\})\exp\left(\sum_{0\le a\le b\le n}\hat{q}^u_{ab}q^u_{ab}\right) = \int\prod_{a=0\cdots n}P^a_U(u_a)\mathrm{d}u^a\exp\left(\sum_{0\le a\le b\le n}\hat{q}^u_{ab}u^a u^b\right)\,. \tag{477}$$

By using Hubbard-Stratonovich identity one gets

$$I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\})\exp\left(\sum_{0\le a\le b\le n}\hat{q}^u_{ab}q^u_{ab}\right) = \tag{478}$$
$$= \int\mathcal{D}W P_{U_0}(u_0)\mathrm{d}u^0\left[\int\mathrm{d}u P_U(u)\exp\left(\frac{\alpha M_v}{\widehat{\Delta}}uu_0 + W\sqrt{\frac{\alpha Q_v}{\widetilde{\Delta}}}u - \left(\frac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)\right)\frac{u^2}{2}\right)\right]^n\,. \tag{479}$$

We can now compute the limit as $n\to 0$.

$$\lim_{n\to0}\frac{\log(I_u(\{q^u_{ab}\},\{\hat{q}^u_{ab}\}))}{n} = \frac{\alpha Q_v Q_u}{2\widetilde{\Delta}} - \frac{\alpha M_v M_u}{\widehat{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)(Q_u+\Sigma_u) +$$
$$\mathbb{E}_{W,u_0}\left[\log\left[\int\mathrm{d}u P_U(u)\exp\left(\frac{\alpha M_v}{\widehat{\Delta}}uu_0 + W\sqrt{\frac{\alpha Q_v}{\widetilde{\Delta}}}u - \left(\frac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)\right)\frac{u^2}{2}\right)\right]\right]$$
$$\lim_{n\to0}\frac{I_v(\{q^v_{ab}\},\{\hat{q}^v_{ab}\})}{N} = \frac{Q_v Q_u}{2\widetilde{\Delta}} - \frac{M_v M_u}{\widehat{\Delta}} - \overline{R}(Q_v+\Sigma_v)(Q_u+\Sigma_u) +$$
$$\mathbb{E}_{W,v_0}\left[\log\left[\int\mathrm{d}v P_V(v)\exp\left(\frac{M_u}{\widehat{\Delta}}vv_0 + W\sqrt{\frac{Q_u}{\widetilde{\Delta}}}u - \left(\frac{Q_u}{\widetilde{\Delta}} - \overline{R}(Q_u+\Sigma_u)\right)\frac{u^2}{2}\right)\right]\right]\,. \tag{480}$$

This finally gives us the replica free energy.

$$\Phi_{\mathrm{RS}}(M_u,Q_u,\Sigma_u,M_v,Q_v,\Sigma_v) = \frac{\alpha Q_v Q_u}{2\widetilde{\Delta}} - \frac{\alpha M_v M_u}{\widehat{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v)(Q_u+\Sigma_u)$$
$$+\mathbb{E}_{W,u_0}\left[\mathcal{Z}_u\left(\frac{\alpha Q_v}{\widetilde{\Delta}} - \alpha\overline{R}(Q_v+\Sigma_v), \frac{\alpha M_v}{\widehat{\Delta}}u_0 + W\sqrt{\frac{\alpha Q_v}{\widetilde{\Delta}}}\right)\right]$$
$$+\alpha\mathbb{E}_{W,v_0}\left[\mathcal{Z}_v\left(\frac{Q_u}{\widetilde{\Delta}} - \overline{R}(Q_u+\Sigma_u), \frac{M_u}{\widehat{\Delta}}v_0 + Ws\sqrt{\frac{Q_u}{\widetilde{\Delta}}}\right)\right]\,. \tag{481}$$

This can also be computed with vectorial notations in both the $XX^\top$ and $UV^\top$ setting leading to eqs. (4.249) and (4.244).

Small ρ expansion

In this appendix we give the small-ρ limits of the state evolution update functions for the Rademacher-Bernoulli, Gauss-Bernoulli, 2 balanced groups and Bernoulli models.

Rademacher-Bernoulli model. We want to compute $\forall \beta > 0$ the limit of $f_{\text{Rademacher}-\text{Bernoulli}}(-\beta \log(\rho))/\rho$ (4.344) as $\rho \to 0$,

$$\frac{f^{\text{SE}}_{\text{Rademacher}-\text{Bernoulli}}(-\beta \log(\rho))}{\rho} = \mathbb{E}_W \left[\frac{\rho \tanh\left(-\beta \log(\rho) + W\sqrt{-\beta \log(\rho)}\right)}{(1-\rho)\frac{\exp(-\beta\log(\rho)/2)}{\cosh\left(-\beta \log(\rho) + W\sqrt{-\beta \log(\rho)}\right)} + \rho} \right]. \tag{482}$$

Taking the small ρ limit here we get

$$\lim_{\rho=0} \frac{f^{\text{SE}}_{\text{Rademacher}-\text{Bernoulli}}(-\beta \log(\rho))}{\rho} = \lim_{\rho=0} \frac{1}{\rho^{\beta/2-1}+1} = 1(\beta > 2)\,, \tag{483}$$

where we used the fact that the noise $W\sqrt{-\beta \log(\rho)}$ is negligible compared to $\log(\rho)$ when $\rho \to 0$.

Gauss-Bernoulli model. We want to compute $\forall \beta > 0$ the limit $f_{\text{Gauss}-\text{Bernoulli}}(-\beta \log(\rho))/\rho$ (4.346) as $\rho \to 0$ and $r = 1$.

$$f^{\text{SE}}_{\text{Gauss}-\text{Bernoulli}}(x)/\rho = \mathbb{E}_{W,x_0}\left[f^{\text{Gauss}-\text{Bernoulli}}_{\text{in}}\left(x, xx_0 + \sqrt{x}W\right) x_0 \right], \tag{484}$$

$$= \mathbb{E}_W \left[\int \frac{\exp\left(-x_0^2/2\right)}{\sqrt{2\pi}} x_0 f^{\text{Gauss}-\text{Bernoulli}}_{\text{in}}\left(x, xx_0 + \sqrt{x}W\right) \right], \tag{485}$$

By intregrating by part one gets

$$= x\mathbb{E}_W \left[\int \frac{\exp\left(-x_0^2/2\right)}{\sqrt{2\pi}} \frac{\partial f^{\text{Gauss}-\text{Bernoulli}}_{\text{in}}}{\partial B}\left(x, xx_0 + \sqrt{x}W\right) \right], \tag{486}$$

Here $xx_0 + \sqrt{x}W$ is a Gaussian random variable of mean 0 and variance $x + x^2$. Therefore one has

$$f^{\text{SE}}_{\text{Gauss}-\text{Bernoulli}}(x)/\rho = x\mathbb{E}_W \left[\frac{\partial f^{\text{Gauss}-\text{Bernoulli}}_{\text{in}}}{\partial B}\left(x, \sqrt{x^2+x}W\right) \right], \tag{487}$$

By making another integration by part one gets

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

$$= \frac{x}{1+x} \mathbb{E}_W \left[W^2 \hat{\rho}(x, W\sqrt{x^2+xs} \right], \tag{488}$$

where

$$\hat{\rho}(x, W^2(x^2+x)) = \frac{\rho}{(1-\rho)\exp\left(\frac{-xW^2}{2}\right)\sqrt{1+x} + \rho}\,. \tag{489}$$

By writing $x = -\beta \log(\rho)$ Depending on the value of W, $\hat{\rho}$ will either go to zero or one as $\rho \to 0$. This will appear in the limit of $\hat{\rho}$. By taking $x = -\beta \log(\rho)$ one gets

$$\lim_{\rho \to 0} \hat{\rho}(x, W^2(x^2+x)) = \lim_{\rho \to 0} \frac{1}{(1-\rho)\rho^{\frac{\beta W^2}{2}-1}\sqrt{1-\beta\log(\rho)}+1} = 1(\beta W^2 > 2). \tag{490}$$

From this we deduce that

$$\begin{aligned}\lim_{\rho \to 0} \frac{f^{\mathrm{SE}}_{\mathrm{Gauss-Bernoulli}}(-\beta\log(\rho))}{\rho} &= \lim_{\rho \to 0} \frac{-\beta\log(\rho)}{1-\beta\log(\rho)} \mathbb{E}_W\left[W^2 1\left(|W| > \sqrt{\frac{2}{\beta}}\right)\right] \\ &= \frac{2\exp(-1/\beta)}{\sqrt{\beta\pi}} + \operatorname{erfc}\left(\frac{1}{\sqrt{\beta}}\right). \end{aligned} \tag{491}$$

Here we used again the fact that the noise $W\sqrt{-\beta\log(\rho)}$ is negligible compared to $\log(\rho)$ when $\rho \to 0$.

2 balanced groups. We want to compute $\forall \beta > 0$ the limit $f_{\mathrm{Balanced}}(-\beta\rho(1-\rho)\log(\rho(1-\rho)))$ (4.350) as $\rho \to 0$.

$$f^{\mathrm{SE}}_{\mathrm{Balanced}}(-\beta\rho(1-\rho)\log(\rho(1-\rho))) = \tag{492}$$

$$\mathbb{E}_W\left[\frac{2\rho(1-\rho)\sinh\left(\frac{-\beta\log(\rho(1-\rho))}{2} + W\sqrt{-\beta\log(\rho(1-\rho))}\right)}{1+2\rho(1-\rho)\left(\cosh\left(\frac{-\beta\log(\rho(1-\rho))}{2} + W\sqrt{-\beta\log(\rho(1-\rho))}\right)-1\right)}\right], \tag{493}$$

$$\lim_{\rho \to 0} f^{\mathrm{SE}}_{\mathrm{Balanced}}(-\beta\rho(1-\rho)\log(\rho(1-\rho))) = \lim_{\rho \to 0} \frac{2\left[\rho(1-\rho)\right]^{1-\beta/2}}{1+2\left[\rho(1-\rho)\right]^{1-\beta/2}} = 1(\beta > 2). \tag{494}$$

Here we used the fact that the noise $W\sqrt{-\beta\log(\rho(1-\rho))}$ is negligible compared to $\log(\rho(1-\rho))$ when $\rho \to 0$.

Spiked Bernoulli model The state evolution update function in this model is given by (4.342). Once again we set $x = -\beta\log(\rho)$ to get

$$\frac{f^{\mathrm{SE}}_{\mathrm{Bernoulli}}(-\beta\log(\rho))}{\rho} = \mathbb{E}_W\left[\frac{1}{1+(1-\rho)\exp\left((\beta/2-1)\log(\rho) + W\sqrt{-\beta\log(\rho)}\right)}\right], \tag{495}$$

$$\lim_{\rho \to 0} \frac{f^{\mathrm{SE}}_{\mathrm{Bernoulli}}(-\beta\log(\rho))}{\rho} = \lim_{\rho \to 0} \frac{1}{1+(1-\rho)\rho^{\beta/2-1}} = 1(\beta > 2). \tag{496}$$

Here we used the fact that the noise $W\sqrt{-\beta\log(\rho)}$ is negligible compared to $\log(\rho)$ when $\rho \to 0$.

To compute the limiting behavior of the phase transitions we analyzed the functions $f_{\mathrm{Rademacher-Bernoulli}}$, $f_{\mathrm{Gauss-Bernoulli}}$ or $f_{\mathrm{Bernoulli}}$ (that we will call f_ρ) as follows. We remind

$$\Delta_{\mathrm{Dyn}}(\rho) = \max_{x \in \mathbb{R}^+} \frac{f_\rho(x)}{x} = \frac{\rho}{-\log(\rho)} \max_{\beta \in \mathbb{R}^+} \frac{f_\rho(-\beta\log(\rho))}{\rho\beta}. \tag{497}$$

In the small ρ limit

$$\Delta_{\mathrm{Dyn}}(\rho) \sim_{\rho\to 0} \frac{\rho}{-\log(\rho)} \max_{\beta\in\mathbb{R}^+} \frac{1(\beta>2)}{\beta} = \frac{\rho}{-2\log(\rho)}\,. \tag{498}$$

The information-theoretic phase transition in the small ρ limit is computed as follows

$$\Delta_{\mathrm{IT}}(\rho) = \max_{x\in\mathbb{R}^+} \left\{ \Delta(x), \int_0^x \mathrm{d}m f_\rho(u) = \frac{x f_{\rho(x)}}{2} \right\}, \tag{499}$$

$$\Delta_{\mathrm{IT}}(\rho) \sim_{\rho\to 0} \max_{\beta\in\mathbb{R}^+} \left\{ \Delta(-\beta\log(\rho)), \int_0^{-\beta\log(\rho)} \mathrm{d}u f_\rho(u) = \frac{-\beta\log(\rho) f_\rho(-\beta\log(\rho))}{2} \right\}, \tag{500}$$

$$\Delta_{\mathrm{IT}}(\rho) \sim_{\rho\to 0} \frac{\rho}{-\log(\rho)} \max_{\beta\in\mathbb{R}^+} \left\{ \frac{1(\beta>2)}{\beta}, \int_0^\beta \mathrm{d}u 1(u>2) = \frac{\beta 1(\beta>2)}{2} \right\}, \tag{501}$$

$$\Delta_{\mathrm{IT}}(\rho) \sim_{\rho\to 0} \frac{\rho}{-4\log(\rho)}\,. \tag{502}$$

Large rank behavior for the symmetric community detection

To locate the phase transitions Δ_{IT} and Δ_{Dyn} in the symmetric community detection case we make a couple of remarks about the state evolution. First we remark that for $\forall x \in \mathbb{R}^+$ $b = \mathcal{M}_r(x)$ is a fixed point of (4.361) for $\Delta = \mathcal{M}_r(x)/x$. By definition Δ_{Dyn} is the greatest Δ for which a fixed point exists

$$\Delta_{\mathrm{Dyn}}(r) = \max_{x\in\mathbb{R}^+} \left\{ \frac{\mathcal{M}_r(x)}{x} \right\}. \tag{503}$$

To find the Δ_{IT} we notice that by taking the derivative with respect to Q and M of (4.244) one finds a combination of (4.202) and (4.203). Therefore we have

$$\phi(b_2,\Delta) - \phi(b_1,\Delta) = \frac{r-1}{2r^2\Delta} \int_{b_1}^{b_2} \mathrm{d}u \mathcal{M}_r\left(\frac{u}{\Delta}\right) - u\,. \tag{504}$$

We deduce a way to compute Δ_{IT} as

$$\Delta_{\mathrm{IT}}(r) = \max_{x\in\mathbb{R}^+} \left\{ \frac{\mathcal{M}_r(x)}{x}, \text{s.t.} \int_0^x \mathrm{d}u \mathcal{M}_r(u) = \frac{x\mathcal{M}_r(x)}{2} \right\}. \tag{505}$$

To compute Δ_{IT} we evaluate $\mathcal{M}_r$ on a whole interval, then for each x draw a line between point $(0,0)$ and $(x, H(x))$. We then compute the area between $\mathcal{M}_r$ and this line. When this area is zero then $\mathcal{M}_r(x)/x$ gives us Δ_{IT}.

In order to compute $\mathcal{M}_r$ we study the function $\mathcal{M}_r(x)$ where we take $x = \beta r\log(r)$, with

$\beta = \Omega(1)$. The important part of $\mathcal{M}_r$ is the integral

$$\int \frac{\exp\left(\frac{x}{r} + \sqrt{\frac{x}{r}}u_1\right)}{\exp\left(\frac{x}{r} + \sqrt{\frac{x}{r}}u_1\right) + \sum_{i=2}^{r} \exp\left(\sqrt{\frac{x}{r}}u_i\right)} \prod_{i=1}^{r} \mathcal{D}u_i \,. \tag{506}$$

The important variables to look at are (when taking $x = \beta r \log(r)$)

$$F_1 = \exp\left(\frac{x}{r} + \sqrt{\frac{x}{r}}u_1\right) = \exp\left(\beta \log(r) + \sqrt{\beta \log(r)}u_1\right) \,, \tag{507}$$

$$F_2 = \sum_{i=2}^{r} \exp\left(\sqrt{\frac{x}{r}}u_i\right) = \sum_{i=2}^{r} \exp\left(\sqrt{\beta \log(r)}u_i\right) . \tag{508}$$

If the typical value of F_1 dominates F_2 as $r \to +\infty$ then $\mathcal{M}_r = 1$, otherwise if F_2 dominates F_1 then $\mathcal{M}_r = 0$. To estimate F_1, F_2 let us notice that with high probability the maximum value of the u_i will be of order $\sqrt{2\log(r)}$. This is a general property of Gaussian variables that the maximum of r independent Gaussian variables of variance σ^2 is of order $\sigma\sqrt{2\log(r)}$. We can therefore compute the mean of F_2 while conditioning on the fact that all of the u_i are smaller than $\sqrt{2\log(r)}$. This allows us to compute the typical value of F_1 as $F_1 \sim r^{\beta}$. For F_2 we obtain : when $\beta < 2$ then $F_2 \sim r^{\frac{\beta}{2}+1}$, and when if $\beta > 2$ then $F_2 \sim r^{\sqrt{2\beta}}$. We have to look at which of the F_1 or F_2 has a higher exponent. $\beta = 2$ is the value at which these two exponent cross. Therefore we have

$$\lim_{r \to +\infty} \mathcal{M}_r(\beta r \log(r)) = 1 \, (\beta > 2) \,. \tag{509}$$

Now let us remind (503) while keeping $x = \beta r \log(r)$ to get

$$\Delta_{\mathrm{Dyn}}(r) \sim_{r\to\infty} \max\left\{ \frac{1\,(\beta > 2)}{\beta r \log(r)}, \beta \in \mathbb{R}^+ \right\} = \frac{1}{2r\log(r)} \,. \tag{510}$$

To get the information-theoretic transition we use (505). Let us find the β that satisfies equation (505) we get

$$\beta r \log(r) \left[1 - \frac{2}{\beta}\right] = \frac{\beta r \log(r)}{2} \,. \tag{511}$$

We deduce $\beta = 4$ and therefore

$$\Delta_{\mathrm{IT}}(r) \sim_{r\to\infty} \frac{1}{4r\log(r)} \,. \tag{512}$$

Large rank behavior for the mixture of Gaussian clustering

To compute these transitions numerically ρ_{IT} and ρ_{alg} we consider one value of $b_v = \mathcal{M}_r(x)$ and ask what is the value of ρ such that $b_s = \mathcal{M}_r(x)$ is a fixed point. Using (4.371) the answer is :

$$\rho(x, r) = \frac{x}{2r} + \sqrt{\frac{x^2}{4r} + \frac{x}{\alpha \mathcal{M}_r(x)}} \,. \tag{513}$$

The spinodal transition being the minimum value of ρ for which a fixed point other than 0 exists we can get the spinodal by minimizing (513). The static transition is obtained by expressing the difference in free energy between $b_v = 0$ and $b_v = \mathcal{M}_r(x)$ at a given ρ and requiring this quantity to be 0. It is possible to express the RS free energy using $\mathcal{M}_r$. If one integrates the gradient of $\phi_{UV^\top}$ along the path $g(u) = (b_u(u), b_v(u))$ (in the space of order parameter b_v, b_u)defined by :

$$\forall u \in [0, \mathcal{M}_r(x)], g(u) = \left(u \frac{\rho(x)^2/r}{\frac{1}{\alpha} + \frac{\rho(x)u}{r}}, u \right) , \tag{514}$$

after using (4.264,4.265) and integrating by parts we get

$$\begin{aligned} \phi_{UV^\top}(\mathcal{M}_r(x), \rho(x,\alpha,r), \alpha, r) - \phi_{UV^\top}(0, \rho(x,\alpha,r), \alpha; r) = \\ - \alpha \frac{r-1}{2r^2} \left[\int_0^x \mathrm{d}u \mathcal{M}_r(u) + \int_0^{\mathcal{M}_r(x)} \mathrm{d}u \frac{u\rho^2}{\frac{1}{\alpha} + \frac{u\rho(x)}{r}} - x\mathcal{M}_r(x) \right] . \end{aligned} \tag{515}$$

The static transition is found where $b_v = \mathcal{M}_r(x)$ and $b_v = 0$ both have the same free energy and therefore (515) is equal to 0.

Formulas (513,515) also allow us to explore the large r limit of these systems. From 4.6.2 we know that

$$\forall \beta > 0, \lim_{r \to +\infty} \mathcal{M}_r(\beta r \log r) = 1_{\beta > 2} \, . \tag{516}$$

The fixed point to our equation is of the SE form $b_v = \mathcal{M}_r(x)$ where $x = \beta r \log r$ and $\beta > 2$ (since we are looking for a fixed point that solves the problem) as r grows large it is easy to prove that $\rho(x, \alpha) \sim \sqrt{\frac{x}{\alpha \mathcal{M}_r(x)}} \sim \sqrt{\frac{\beta r \log r}{\alpha}} = O(r \log(r))$. For $\beta < 2$ we get $\rho(\beta r \log(r), r) = O(r)$. Therefore minimizing $\rho(\beta r \log r, \alpha)$ gives us

$$\rho_{\mathrm{Dyn}}(r, \alpha) = \sqrt{\frac{2r \log r}{\alpha}} (1 + o_r(1)) \, . \tag{517}$$

For $\beta > 2$ we have that $u\rho(x)/r \ll 1$ in (515) which allows us to rewrite the zero condition on (515) zero as

$$\phi_{UV^\top}(\mathcal{M}_r(x), \rho(x,\alpha,r), \alpha, r) - \phi_{UV^\top}(0, \rho(x,\alpha,r), \alpha; r) = \tag{518}$$

$$-\alpha \frac{r-1}{2r^2} \left[\int_0^x \mathrm{d}u \mathcal{M}_r(u) - \frac{x\mathcal{M}_r(x)}{2} \right] = 0 \, . \tag{519}$$

Which as in 4.6.2 give us as r goes to $+\infty$

$$\beta r \log(r) \left[1 - \frac{2}{\beta} \right] = \frac{\beta r \log(r)}{2} \, . \tag{520}$$

We deduce $\beta = 4$ and therefore

$$\rho(r,\alpha)_{\mathrm{IT}} = 2\sqrt{\frac{r\log r}{\alpha}}(1+o_r(1))\,, \tag{521}$$

$$\rho(r,\alpha)_{\mathrm{Spinodal}} = \sqrt{\frac{2r\log r}{\alpha}}(1+o_r(1))\,. \tag{522}$$

Bibliographie

[AW08] Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *IEEE International Symposium on Information Theory*, pages 2454–2458. IEEE, 2008.

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005.

[BDM+16] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborov'a. Mutual information for symmetric rank-one matrix estimation : A proof of the replica formula. In *Advances In Neural Information Processing Systems*, pages 424–432, 2016.

[BGN11] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1) :494 – 521, 2011.

[BM94] Michael Biehl and Andreas Mietzner. Statistical mechanics of unsupervised structure recognition. *Journal of Physics A : Mathematical and General*, 27(6) :1885, 1994.

[BMVX16] Jess Banks, Cristopher Moore, Roman Vershynin, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *arXiv preprint arXiv :1607.05222*, 2016.

[BR13] Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse PCA. *arXiv preprint arXiv :1304.0828*, 2013.

[BS94] N Barkai and Haim Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. *Physical Review E*, 50(3) :1766, 1994.

[CC00] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[CLM16] Francesco Caltagirone, Marc Lelarge, and Léo Miolane. Recovering asymmetric communities in the stochastic block model. *arXiv preprint arXiv :1610.03680*, 2016.

[CO07] Amin Coja-Oghlan. Solving np-hard semirandom graph problems in polynomial expected time. *J. Algorithms*, 62(1) :19–46, January 2007.

[COP16] Amin Coja-Oghlan and Will Perkins. Belief propagation on replica symmetric random factor graph models. *arXiv preprint arXiv :1603.08191*, 2016.

[CZK14] Francesco Caltagirone, Lenka Zdeborová, and Florent Krzakala. On convergence of approximate message passing. In *IEEE International Symposium on Information Theory*, pages 1812–1816. IEEE, 2014.

[DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the binary stochastic block model. In *IEEE International Symposium on Information Theory (ISIT)*, pages 185–189. IEEE, 2016.

[DCDNB66] FJ Du Chatenier, J De Nobel, and BM Boerstoel. Specific heats of gold and dilute alloys of manganese, chromium, iron and vanadium in gold at low temperatures. *Physica*, 32(3) :561–570, 1966.

[DKMZ11a] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6) :066106, 2011.

[DKMZ11b] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6) :065701, 2011.

[DM14a] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2197–2201. IEEE, 2014.

[DM14b] Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–342, 2014.

[DM15] Yash Deshpande and Andrea Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, pages 1–60, 2015.

[DMM09] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45) :18914–18919, 2009.

[EA75] Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F : Metal Physics*, 5(5) :965, 1975.

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3) :211–218, 1936.

[For10] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3) :75–174, 2010.

[GKS85] D. J. Gross, I. Kanter, and H. Sompolinsky. Mean-field theory of the potts glass. *Phys. Rev. Lett.*, 55(3) :304–307, Jul 1985.

[GT81] Marc Gabay and Gérard Toulouse. Coexistence of spin-glass and ferromagnetic orderings. *Physical Review Letters*, 47(3) :201, 1981.

[GTK15] Marylou Gabrié, Eric W Tramel, and Florent Krzakala. Training restricted Boltzmann machine via the Thouless-Anderson-Palmer free energy. In *Advances in Neural Information Processing Systems*, pages 640–648, 2015.

[GY91] A Georges and J S Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A : Mathematical and General*, 24(9) :2173, 1991.

[Hin10] Geoffrey Hinton. A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1) :926, 2010.

[HL09] Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP hard. *CoRR*, abs/0911.1393, 2009.

[Hop82] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8) :2554–2558, 1982.

[HOT06] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7) :1527–1554, 2006.

[HR04] David C Hoyle and Magnus Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E*, 69(2) :026124, 2004.

[HTFF05] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning : data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2) :83–85, 2005.

[JL04] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *Unpublished manuscript*, 7, 2004.

[JM13] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference*, page iat004, 2013.

[KDMR+03] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2) :349–396, 2003.

[KGO12] Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2) :159–182, 2012.

[KKM+14] Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in bayes-optimal matrix factorization. *CoRR*, abs/1402.1298, 2014.

[KKM+16] Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7) :4228–4265, 2016.

[KMM+13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52) :20935–20940, 2013.

[KMTZ14] Florent Krzakala, Andre Manoel, Eric W Tramel, and Lenka Zdeborová. Variational free energies for compressed sensing. In *IEEE International Symposium on Information Theory*, pages 1499–1503. IEEE, 2014.

[KMZ13] Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Phase diagram and approximate message passing for blind calibration and dictionary learning. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 659–663. IEEE, 2013.

[KNV+15] Robert Krauthgamer, Boaz Nadler, Dan Vilenchik, et al. Do semidefinite relaxations solve sparse PCA up to the information limit ? *The Annals of Statistics*, 43(3) :1300–1322, 2015.

[KR98] Hilbert J Kappen and FB Rodriguez. Boltzmann machine learning using mean field theory and linear response correction. *Advances in neural information processing systems*, pages 280–286, 1998.

[KTJ76] JM Kosterlitz, DJ Thouless, and Raymund C Jones. Spherical model of a spin-glass. *Physical Review Letters*, 36(20) :1217, 1976.

[KXZ16] Florent Krzakala, Jiaming Xu, and Lenka Zdeborová. Mutual information in rank-one matrix estimation. *to appear in ITW 2016, preprint arXiv :1603.08447*, 2016.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553) :436–444, 2015.

[LDBB+16] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering. *to appear in Allerton 2016, preprint arXiv :1610.02918*, 2016.

[LKZ15a] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. MMSE of probabilistic low-rank matrix estimation : Universality with respect to the output channel. In *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 680–687. IEEE, 2015.

[LKZ15b] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse PCA. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1635–1639, 2015.

[LKZ17] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation : Phase transitions, approximate message passing and applications. *preprint arXiv :1701.00858 [math.ST]*, 2017.

[Llo82] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2) :129–137, 1982.

[LML+17] Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. *arXiv preprint arXiv :1701.08010*, 2017.

[Méz16] Marc Mézard. Mean-field message-passing equations in the Hopfield model and its generalizations. *arXiv preprint arXiv :1608.01558*, 2016.

[Méz17] Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2) :022117, 2017.

[Mio17a] L. Miolane. Fundamental limits of low-rank matrix estimation : the non-symmetric case. *ArXiv e-prints*, February 2017.

[Mio17b] Léo Miolane. Fundamental limits of low-rank matrix estimation. *arXiv preprint arXiv :1702.00473*, 2017.

[ML16] Léo Miolane Marc Lelarge. Fundamental limits of symmetric low-rank matrix estimation. *arXiv :1611.03888 [math.PR]*, 2016.

[MM09] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[MO04] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis : a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1) :24–45, 2004.

[Mon15] Andrea Montanari. Finding one community in a sparse graph. *Journal of Statistical Physics*, 161(2) :273–299, 2015.

[MPV87] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin-Glass Theory and Beyond*, volume 9 of *Lecture Notes in Physics*. World Scientific, Singapore, 1987.

[MT13] Ryosuke Matsushita and Toshiyuki Tanaka. Low-rank matrix reconstruction and clustering via approximate message passing. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 917–925. Curran Associates, Inc., 2013.

[MV15] Rémi Monasson and Dario Villamaina. Estimating the principal components of correlation matrices from all their empirical eigenvectors. *EPL (Europhysics Letters)*, 112(5) :50001, 2015.

[Nis01] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing : An Introduction*. Oxford University Press, Oxford, UK, 2001.

[NS01] Hidetoshi Nishimori and David Sherrington. Absence of replica symmetry breaking in a region of the phase diagram of the ising spin glass. In *American Institute of Physics Conference Series*, volume 553, pages 67–72, 2001.

[OA09] Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural computation*, 21(3) :786–792, 2009.

[OF96] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583) :607, 1996.

[OS01] Manfred Opper and David Saad. *Advanced mean field methods : Theory and practice*. MIT press, 2001.

[OW01] Manfred Opper and Ole Winther. Adaptive and self-averaging thouless-anderson-palmer mean-field theory for probabilistic modeling. *Physical Review E*, 64(5) :056131, 2001.

[Par79] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23) :1754, 1979.

[PHL04] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data : a review. *ACM SIGKDD Explorations Newsletter*, 6(1) :90–105, 2004.

[Ple82] T Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A : Mathematical and General*, 15(6) :1971, 1982.

[PSC14] Jason T Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing âĂŤ part I : Derivation. *IEEE Transactions on Signal Processing*, 62(22) :5839–5853, 2014.

[PWBM16a] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Message-passing algorithms for synchronization problems over compact groups. *arXiv preprint arXiv :1610.04583*, 2016.

[PWBM16b] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of PCA for spiked random matrices and synchronization. *arXiv preprint arXiv :1609.05573*, 2016.

[RF12] Sundeep Rangan and Alyson K Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1246–1250. IEEE, 2012.

[RM14] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[RSR+13] Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson Fletcher, and Volkan Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 664–668. IEEE, 2013.

[SHM+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587) :484–489, 2016.

[SK75] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35 :1792–1796, 1975.

[Som81] Hans-Juergen Sommers. Theory of a Heisenberg spin glass. *Journal of Magnetism and Magnetic Materials*, 22(3) :267–270, 1981.

[SRF16] Philip Schniter, Sundeep Rangan, and Alyson K Fletcher. Vector approximate message passing for the generalized linear model. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 1525–1529. IEEE, 2016.

[TAP77] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3) :593–601, 1977.

[TM16] Jérôme Tubiana and Rémi Monasson. Emergence of compositional representations in restricted Boltzmann machines. *arXiv preprint arXiv :1611.06759*, 2016.

[TMC+16] Eric W Tramel, Andre Manoel, Francesco Caltagirone, Marylou Gabrié, and Florent Krzakala. Inferring sparsity : Compressed sensing using generalized restricted Boltzmann machines. In *IEEE Information Theory Workshop (ITW)*, pages 265–269. IEEE, 2016.

[UKBM11] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv :1111.4503*, 2011.

[VSR+15] Jeremy Vila, Philip Schniter, Sundeep Rangan, Florent Krzakala, and Lenka Zdeborová. Adaptive damping and mean removal for the generalized approximate message passing algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2021–2025. IEEE, 2015.

[Was13] Larry Wasserman. *All of statistics : a concise course in statistical inference*. Springer Science & Business Media, 2013.

[WN94] TLH Watkin and J-P Nadal. Optimal unsupervised learning. *Journal of Physics A : Mathematical and General*, 27(6) :1899, 1994.

[YFW03] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Exploring artificial intelligence in the new millennium. In Gerhard Lakemeyer and Bernhard Nebel, editors, *Exploring artificial intelligence in the new millennium*, chapter Understanding Belief Propagation and Its Generalizations, pages 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[ZH60] Jo E Zimmerman and FE Hoare. Low-temperature specific heat of dilute cu-mn alloys. *Journal of Physics and Chemistry of Solids*, 17(1-2) :52–56, 1960.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2) :265–286, 2006.

[ZK16] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference : Thresholds and algorithms. *Advances in Physics*, 65 :453–552, 2016.

ÉCOLE DOCTORALE
Physique en île-de-France (EDPIF)

Titre : Factorisation matricielle et tensorielle par une approche issue de la physique statistique.
.....

Mots clés : Physique Statistique, Apprentissage Machine, Informatique

Résumé : Dans cette thèse je présente des résultats sur la factorisation de matrice et de tenseur. Les matrices étant un objet omniprésent en mathématique un grand nombre de problème d'apprentissage machine peuvent être transcrit en un problème de factozisation de matrice de petit rang. C'est une des méthode les plus basiques utilisé dans les méthodes d'apprentissage non supervisé et les problèmes de réductions dimensionnelle.
Les résultats présentés dans cette thèse ont pour la plupart déjà été inclus dans des publications antérieures [LKZ 2015].
Le problème de la factorisation de matrice de petit rang devient de plus en plus en difficile quand on rajoute des contraintes additionnelles, comme par exemple la positivité d'un des facteurs.
Nous présentons içi un cadre dans lequel analyser ce problème sous un angle Bayesien où le priors sur les facteurs peuvent être générique et où l'output channel à travers duquel la matrice est observé peut être générique aussi.
Nous tracerons un parallèle entre le problème de factorisation matriciel et les problème de verre de spin vectoriel. Ce cadre permet d'abborder d'une façon unifié des problèmes qui étaient abordé de façon séparés dans des publications précédentes. Nous dérivons en détail la forme générale des equations de Low-rank Approximate Message Passing (Low-RAMP) ce qui donnera un algorithm de factorisation. Ces équations sont connues dans en physique statistique sous le nom des equations TAP. Nous dérivons ces equations dans différents cas, pour le modèle de Sherrington-Kirkpatrick, les restricted Boltzmann machine, le modèle de Hopfield ou encore le modèle xy. La dynamique des équations Low-RAMP peuvent être analysé en utilisant les equation de State Evolution ces equations sont équivalentes à un calcul des répliques symmétrique. Dans la section dévolue aux résultats nous étudierons de nombreux diagramme de phase et transition de phase dans le cas Bayes-optimale. Nous présentons different différents typologies de diagramme de phase et leurs interprétations en terme de performances algorithmiques.

Title : Matricial and tensorial factorisation using tools coming from statistical physics.....
.....

Keywords : Statistical Physics, Machine Learning, Computer Science

Abstract :
In this thesis we present the result on low rank matrix and tensor factorization. Matrices being such an ubiquitous mathematical object a lot of machine learning can be mapped to a low-rank matrix factorization problem. It is for example one of the basic methods used in data analysis for unsupervised learning of relevant features and other types of dimensionality reduction. The result presented in this thesis have been included in previous work [LKZ 201].
The problem of low rank matrix becomes harder once one adds constraint to the problem like for instance the positivity of one of the factor of the factorization. We present a framework to study the constrained low-rank matrix estimation for a general prior on the factors, and a general output channel through which the matrix is observed. We draw a parallel with the study of vector-spin glass models -- presenting a unifying way to study a number of problems considered previously in separate statistical physics works. We present a number of applications for the problem in data analysis.
We derive in detail a general form of the low-rank approximate message passing (Low-RAMP) algorithm that is known in statistical physics as the TAP equations. We thus unify the derivation of the TAP equations for models as different as the Sherrington-Kirkpatrick model, the restricted Boltzmann machine, the Hopfield model or vector (xy, Heisenberg and other) spin glasses. The state evolution of the Low-RAMP algorithm is also derived, and is equivalent to the replica symmetric solution for the large class of vector-spin glass models. In the section devoted to result we study in detail phase diagrams and phase transitions for the Bayes-optimal inference in low-rank matrix estimation. We present a typology of phase transitions and their relation to performance of algorithms such as the Low-RAMP or commonly used spectral methods. excessit e vita aetatis nono anno atque vicensimo cum quadriennio imperasset. natus apud Tuscos in Massa Veternensi, patre Constantio Constantini fratre imperatoris, matreque Galla.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France