Statistical Physics Methods in Optimization and Machine Learning

An Introduction to Replica, Cavity & Message-Passing techniques

Florent Krzakala and Lenka Zdeborová

October 12, 2023

Introduction

> I thought fit to [...] explain in detail in the same book the peculiarity of a certain method, by which it will be possible [...] to investigate some of the problems in mathematics by means of mechanics. This procedure is [...] no less useful even for the proof of the theorems themselves; for certain things first became clear to me by a mechanical method, although they had to be demonstrated by geometry afterwards [...] But it is of course easier, when we have previously acquired, by the method, some knowledge of the questions, to supply the proof than it is to find it without any previous knowledge. I am persuaded that it [...] will be of no little service to mathematics; for I apprehend that some, either of my contemporaries or of my successors, will, by means of the method when once established, be able to discover other theorems in addition, which have not yet occurred to me.
>
> *The method of mechanical theorems*
> Letter from Archimedes to Eratosthenes – circa 220BC

This set of lectures will be discussing probabilistic models, and focus on problems coming from Statistics, Machine Learning and Constrained Optimization, while using tools and techniques from Statistical Physics. The focus will be more theoretical than practical, so you have been warned! Our goal is to show how some methods from Statistical Physics allow us to derive precise answers to many mathematical questions. As pointed out by Archimedes, once these answers are given, even if they are obtained through heuristic methods, it is a simpler task (but still non-trivial) to prove them rigorously. Over the last few decades, there has been an increasing convergence of interest and methods between Theoretical Physics and Applied Mathematics, and many theoretical and applied works in Statistical Physics and Computer Science have relied on a connection with the Statistical Physics of Spin Glasses. The aim of this course is to present the background necessary for entering this fast-developing field.

At first glance, it may seem surprising that Physics has any connection with minimization and probabilistic inference problems. The connection lies in the Gibbs (or Boltzmann) distribution, the fundamental object in Statistical Mechanics. From the point of view of Statistics and Optimization we will be interested in two types of problems: a) minimizing a cost function and b) sampling from a distribution. In both cases, the Statistical Physics approach, or more precisely the Boltzmann measure, turns out to be convenient.

Say that you have a "cost" function $E(\boldsymbol{x})$ of $\boldsymbol{x} \in \mathbb{R}^d$. In statistical mechanics, one associates a temperature-dependent "Boltzmann" probability measure to each possible value of $\boldsymbol{x}$ as follows:

$$\mathbb{P}^{\text{Boltzmann}}(\boldsymbol{x}) = \frac{1}{Z_N(\beta)} e^{-\beta E(\boldsymbol{x})}$$

where $\beta = 1/T$ is called the inverse temperature, and

$$Z_N(\beta) = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{x}\; e^{-\beta E(\boldsymbol{x})}$$

is called the partition function, or the partition sum. The introduction of the temperature is very practical. For instance one can check that the limit $\beta \to \infty$ allows us to study minimization

problems, since

$$\lim_{\beta\to\infty} -\partial_\beta \log Z(\beta) = \min_{\boldsymbol{x}} E(\boldsymbol{x})$$

One can also obtain the number of minima by computing $\lim_{\beta\to\infty} Z(\beta)$.

The formalism is equally interesting for sampling problems. A typical problem that arises in Statistical Inference, Information Theory, Signal Processing or Machine Learning is the following: let X be an unknown quantity that we would like to infer, but which we don't have access to. Instead, we are given access to a quantity Y, related to X (usually a noisy version of X). For concreteness, assume that X is simply a scalar Gaussian variable with zero mean and unit variance, and that $Y = X + Z$ for another standard Gaussian variable Z. Given Y, what can we say about X? In other words: what is the probability of X given our measurement Y? The quantity $\mathbb{P}_{X|Y}(x, y)$ is called the *posterior distribution* of X given Y. Bayes' formula famously states that

$$\mathbb{P}_{X|Y}(x, y) = \frac{\mathbb{P}_{Y|X}(x, y)P_X(x)}{P_Y(y)},$$

so that the posterior $\mathbb{P}_{X|Y}(x, y)$ is given by the product of the prior probability on X, $P_X(x)$, times the likelihood $\mathbb{P}_{Y|X}(x, y)$, divided by the *evidence* $P_Y(y)$, which is just a normalization constant (in the sense that it is not a function of X, the random variable of interest).

This, of course, can be trivially rewritten as

$$\mathbb{P}_{X|Y}(x, y) = \frac{e^{\log\left[\mathbb{P}_{Y|X}(x,y)P_X(x)\right]}}{P_Y(y)},$$

and thus can be represented as a Boltzmann measure provided we identify

$$\beta = 1, \qquad Z = P_Y(y), \qquad E = -\log \mathbb{P}_{Y|X}(x, y)P_X(x).$$

Thus, the evidence P_Y is simply the partition sum. This simple rewriting is behind the popularity of Statistical Mechanics language in Bayesian inference. Indeed, many words in the vocabulary of the Machine Learning community are borrowed directly from Physics (such as "energy-based model", "free energy", "mean-field", "Boltzmann machine"...).

In what follows, we shall be interested in the accuracy of the resulting estimates, for instance the mean squared error over a given set of models. We will thus need to apply the methods from Statistical Physics, not only to derive the posterior distribution and the partition sum, but also to take averages over many models or many realizations, in order to determine the typical behavior. For example, we would like to access information-theoretical quantities such as the entropy. Computing the partition sum Z is already difficult, but computing such averages is notoriously even harder.

Conveniently, there is a part of Statistical Physics that focuses exactly on this task: the field of disordered systems and spin glasses. Spin glasses are magnets in which the interaction strength between each pair of particles is random. Starting in the late 70s with the groundbreaking work of Sir Sam Edwards and Nobel prize winner Philip W. Anderson, the Statistical Physics of disordered systems and spin glasses has grown into a versatile theory, with powerful heuristic tools such as the replica and the cavity methods. In itself, the idea of using Statistical Physics methods to study some problems in computer science is not a new. It was the inspiration, for instance, behind the creation of Simulated Annealing. Anderson, on the one hand, and Parisi

and Mézard on the other, used this connection back in 1986 to study optimisation problems, and since then many applied it with great success to study a large variety of problems in optimization (random satisfiability & coloring), error correcting codes, inference and machine learning.

In the lectures, we wish to address these questions with an interdisciplinary approach, leveraging tools from mathematical physics and statistical mechanics, but also from information theory and optimization. Our modelling strategy and the analysis originate in studies of phase transitions for models of Condensed Matter Physics. Yet, most of our objectives and applications belong to the fields of Machine Learning, Computer Science, and Statistical Data Processing.

Notations

We shall use probabilistic notations: random variables are uppercase, and a particular value is lowercase. For instance, we will speak of the probability $\mathbb{P}(X = x)$ that the random variable X takes the value x.

$=:$	definition of a new quantity
$\asymp$	asymptotically equal, used for large deviation
$\mathbb{P}(X = x)$	probability that the random variable X takes value x

Contents

Part I

Techniques and methods

Chapter 1

The mean-field ferromagnet of Curie & Weiss

> The world was so recent that many things lacked names, and in order to indicate them it was necessary to point. Every year during the month of March a family of ragged gypsies would set up their tents near the village, and with a great uproar of pipes and kettledrums they would display new inventions. First they brought the magnet.
>
> *100 años de soledad*
> Gabriel Garcia Márquez – 1967

The *Curie-Weiss model* is a simple model for ferromagnetism. The essential phenomenon associated with a ferromagnet is that below a certain critical temperature, a magnetization will spontaneously appear in the absence of an external magnetic field. The Curie-Weiss model is simple enough that all the thermodynamic functions characterising its macroscopic properties can be computed *exactly*. And yet, it is rich enough to capture the basic phenomenology of phase transitions — namely the transition between a *disordered* paramagnetic phase (not magnetized) and an *ordered* ferromagnetic phase (magnetized). Because of its simplicity and because of the correctness of at least some of its predictions, the Curie-Weiss model occupies an important place in the Statistical Mechanics literature.

In this model, the magnetic moments are encoded by N microscopic *spin variables* $S_i \in \{-1, +1\}$ for $i = 1, \dots, N$. Every magnetic moment S_i interacts with *every other* magnetic moment S_j for $j \neq i$. Ferromagnetism can be modelled by a collective alignment of the magnetic moments in the same direction. Therefore, to encourage a ferromagnetic phase, we add an energy cost for spins which are not aligned. In its simplest flavour, this cost takes the form of a two-body interaction $-S_i S_j$. The total cost function associated with a given configuration of spins $\boldsymbol{S} \in \{-1, +1\}^N$, also known in Physics as the *Hamiltonian*, is given by:

$$\mathcal{H}_N^0(\boldsymbol{S}) = -\frac{1}{2N} \sum_{ij} S_i S_j \, .$$

It is actually convenient to add a constant external field $h \in \mathbb{R}$ enforcing alignment in $S_i = +1$

for $h > 0$ and $S_i = -1$ for $h < 0$, so we shall work with the Hamiltonian:

$$\mathcal{H}_N(\boldsymbol{S}) = \mathcal{H}_N^0(\boldsymbol{S}) - h\sum_i S_i = -\frac{1}{2N}\sum_{ij} S_i S_j - h\sum_i S_i\,. \tag{1.1}$$

The probability of finding the system at the configuration $\boldsymbol{s} \in \{-1,+1\}^N$ is given by the Boltzmann measure[1]:

$$\mathbb{P}_{N,\beta,h}\left(\boldsymbol{S}=\boldsymbol{s}\right) = \frac{e^{-\beta\mathcal{H}(\boldsymbol{s})}}{Z_N(\beta,h)}.$$

where $\beta = T^{-1} \geq 0$ is the inverse temperature. Note that for $\beta > 0$ the Boltzmann measure gives more weight to configurations with lower cost or energy. In particular, when $\beta \to \infty$ (or equivalently $T = 0$) it concentrates around configurations that minimise $\mathcal{H}$. In the opposite limit, when $\beta = 0$ (or equivalently $T \to \infty$) it assigns equal weight to every configuration, yielding the uniform measure. The normalization of the Boltzmann measure plays a very important role in Statistical Physics, and is known as the *partition sum* or *partition function*:

$$Z_N(\beta,h) = \sum_{\boldsymbol{S}\in\{-1,+1\}^N} e^{\frac{\beta}{2N}\sum_{ij} S_i S_j + \beta h \sum S_i}. \tag{1.2}$$

As we will show next, the partition sum is closely related to the *thermodynamic functions* that characterize the macroscopic properties of the model.

1.1 Rigorous solution

As mentioned in the introduction, ferromagnetism is modelled in this context by the alignment of the magnetic moments or spins. Therefore, a good probe for a ferromagnetic state is given by the *magnetization per spin*:

$$\bar{S} =: \frac{1}{N}\sum_{i=1}^{N} S_i.$$

Note that $\bar{S}$ is the empirical average of the spins, and therefore is itself a random variable. It is interesting to note that the Hamiltonian of the Curie-Weiss model is actually only a function of the magnetization:

$$\mathcal{H}(\bar{S}) = -N\left(\frac{1}{2}\bar{S}^2 + h\bar{S}\right),$$

making it explicit that it is an extensive quantity $\mathcal{H}(\bar{S}) \propto N$. And each time we flip a single spin from -1 to $+1$, the magnetization per spin increases by $2/N$. Therefore, what is the probability that the magnetization per spin takes a particular value in the set $\mathcal{S}_N = \{-1, -1 + 2/N, -1 + 4/N, \ldots, 1\}$? According to the Boltzmann measure, this is given by:

$$\mathbb{P}(\bar{S} = m) = \frac{\Omega(m,N)}{Z_N(\beta,h)} e^{\beta N(\frac{1}{2}m^2 + hm)}$$

[1] Also known as Gibbs measure or Gibbs-Boltzmann measure.

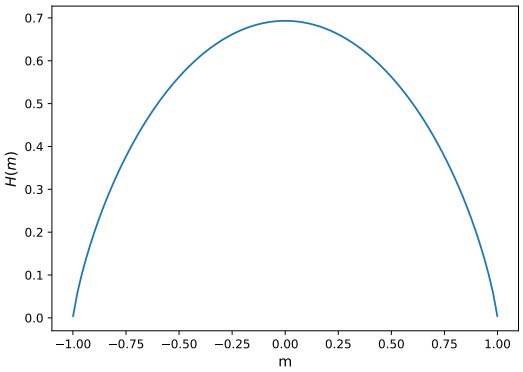


Figure 1.1: Binary entropy $H(m)$ defined in equation 1.3 as a function of the magnetization m.

where $\Omega(m, N)$ is the number of configurations with magnetization $\bar{S} = m$. This can be computed explicitly, and is simply given by a standard binomial:

$$\Omega(m, N) = \frac{N!}{\left(\frac{N-Nm}{2}\right)!\left(\frac{N+Nm}{2}\right)!}$$

The expression above is different from the usual binomial distribution because the S_i take their values in $\{-1, 1\}$ instead of $\{0, 1\}$. At first glance it is not very friendly, but with some work refining Stirling's approximation (see exercise 1.1), one can show that

$$\frac{e^{NH(m)}}{N+1} \leq \Omega(m, N) \leq e^{NH(m)},$$

with $H(m)$, often called the *binary entropy* , given by

$$H(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right). \tag{1.3}$$

Note that the binary entropy is usually defined with $\log_2$ in Information Theory. We shall stick to the natural logarithm here.

We thus reach our first result:

Lemma 1. *Let $\phi(m) = H(m) + \frac{1}{2}\beta m^2 + \beta h m$, then*

$$\frac{1}{N+1}\frac{e^{N\phi(m)}}{Z_N(\beta, h)} \leq \mathbb{P}(\bar{S} = m) \leq \frac{e^{N\phi(m)}}{Z_N(\beta, h)} \tag{1.4}$$

One can also compute, and bound, the value of $Z_N(\beta, h)$. Indeed, summing over m on both sides of the right part of equation 1.4 one reaches

$$1 \leq \sum_m \frac{e^{N\phi(m)}}{Z_N(\beta, h)} \leq (N+1)\frac{e^{N\phi(m^*)}}{Z_N(\beta, h)}, \tag{1.5}$$

where we have defined the value $m^* \in [-1, 1]$ that maximizes $\phi(m)$ (note that m^* depends on β and h). Therefore, taking the logarithm on both sides:

$$\frac{\log Z_N(\beta, h)}{N} \leq \phi(m^*) + \frac{\log(N+1)}{N}.$$

Additionally, the left part of equation 1.4 teaches us that

$$\frac{1}{N+1}\frac{e^{N\phi(m)}}{Z_N(\beta,h)} \leq \mathbb{P}(\bar{S}=m) \leq 1$$
$$-\frac{\log(N+1)}{N} + \phi(m) \leq \frac{\log Z_N(\beta,h)}{N}\,.$$

This is true for all the discrete values $m \in \mathcal{S}_N$, and in particular for the value $m^{\max}$ that maximizes $\phi(m)$ over this set. It is easy to see that maximizing over $[-1,1]$ instead of $\mathcal{S}_N$ does not change the result substantially since $\phi(m^{\max}) > \phi(m^*) - \log N/N$ [2]. Therefore, for N large enough we finally obtain

Lemma 2. *Let* $\Phi_N(\beta,h) = \log Z_N(\beta,h)/N$, *then*

$$\phi(m^*) - \frac{\log(N(N+1))}{N} \leq \Phi_N(\beta,h) \leq \phi(m^*) + \frac{\log(N+1)}{N} \tag{1.6}$$

We shall call $\log Z_N$ the *free entropy*, while $\Phi_N =: \frac{1}{N}\log Z_N$ is the *free entropy density*. Asymptotically, it reads:

Theorem 1. *Let* $\Phi_N(\beta,h) = \log Z_N(\beta,h)/N$, *then*

$$\Phi(\beta,h) =: \lim_{N\to\infty} \Phi_N(\beta,h) = \max_{m\in[-1,1]} \phi(m) = \phi(m^*) \tag{1.7}$$

Additionally,

$$\lim_{N\to\infty} \frac{\log \mathbb{P}(\bar{S}=m)}{N} = \phi(m) - \phi(m^*) \tag{1.8}$$

Note that this result is quite remarkable: we have turned the seemingly impossible sum over 2^N states of equation 1.2 into a simple maximization of a one-dimensional *potential function* $\phi(m)$. In particular, equation 1.8 shows that in the *thermodynamic limit* $N \to \infty$ the potential $\phi(m)$ fully characterizes the probability of finding the system at a given macroscopic state $\bar{S} = m$. More importantly, we have done this without any approximation. All the steps are rigorous! This is, of course, thanks to the simplicity of the Curie-Weiss model.

1.1.1 Phase transition in the Curie-Weiss model

From this exact solution, we can now analyze the phenomenology of the Curie-Weiss model. The extremization $\phi'(m) = 0$ leads to the condition

$$\frac{1}{2}\log\left(\frac{1+m}{1-m}\right) = \beta(m+h)$$

which, using $\tanh^{-1}(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$ gives the so-called Curie-Weiss *mean-field* or *saddle point* equation:

$$m = \tanh\left(\beta(h+m)\right). \tag{1.9}$$

[2] The mean value theorem applied between m^* and $m^{\max}$ yields a $c \in [m^*, m^{\max}]$ which satisfies $\phi(m^*) = \phi(m^{\max}) + \phi'(c)(m^* - m^{\max})$, and such that $|m^* - m^{\max}| \leq 2/N$. Since the difference is K/N for some constant K, it will be eventually smaller than $\log N/N$.

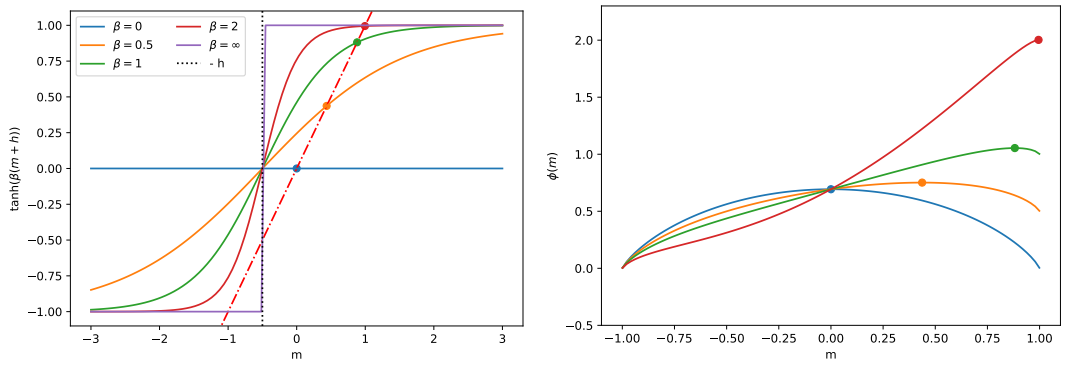


Figure 1.2: (Left) Right-hand side of equation 1.9 as a function of m, for fixed $h = 0.5$ and different values of the inverse temperature (β solid lines). Solutions of equation 1.9 (dots) are given by the intersection of $f(m) = \tanh(\beta(h + m))$ with the line $f(m) = m$ (red dashed). (Right) Same picture in terms of the potential $\phi(m)$, where the solutions of equation 1.9 correspond to the global maximum of $\phi(m)$. Note that for $\beta \gg 1$ an unstable solution corresponding to a minimum of ϕ appear.

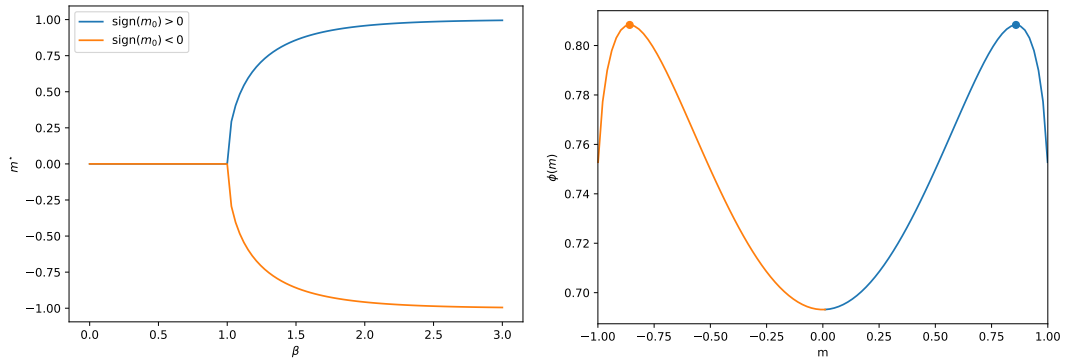


Figure 1.3: (Left)Fixed point $m^\star$ of the mean-field equation $m = \tanh(\beta m)$ above the critical temperature ($h = 0$, $\beta = 1.5$) as a function of the inverse temperature β. Depending on the sign of the initialisation m_0, we reach one of the two global maxima of $\phi(m)$ (right).

Figure 1.2 shows the right-hand side of the mean-field equation equation 1.9 for a fixed external field h and different values of the inverse temperature β. The solution m^* of these self-consistent equations is given by the intersection of $f(m) = \tanh(\beta(m + h))$ with the line $f(m) = m$. Depending on the value of the parameters, there can be up to three solutions.

The property $\frac{1}{N} \log \mathbb{P}(\bar{S} = m) \to \phi(m) - \phi(m^*)$ is usually called a *Large Deviation Principle* in mathematics. It basically tell us that the probability that $\bar{S}$ takes any other value than m^* is exponentially small, i.e. a very rare event. In a nutshell, we can write that the probability to find the system in a given value of m is approximately:

$$\mathbb{P}(\bar{S} = m) \underset{N\to\infty}{\asymp} e^{N(\phi(m)-\phi(m^*))},$$

where we have used the symbol $\underset{N\to\infty}{\asymp}$ to denote an equality valid asymptotically in N. If the maximum is unique, then the magnetization is found to be equal to m^* with probability one. This convergence in probability is called a *concentration phenomena* in probability theory. For

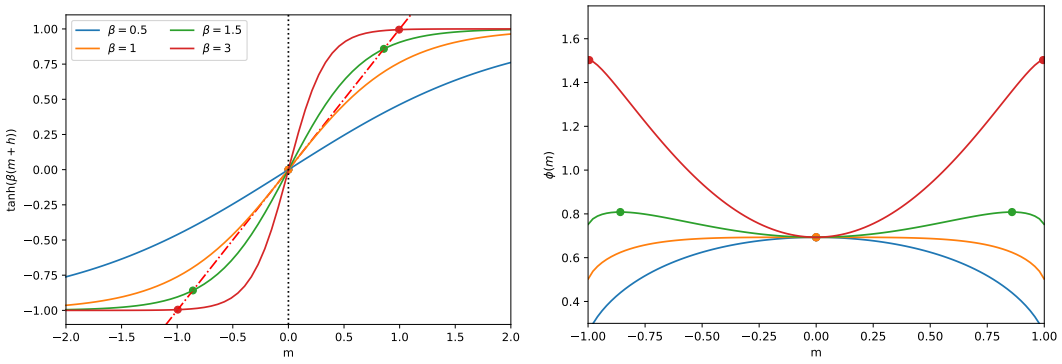


Figure 1.4: Same setting as Fig. 1.2, but for zero external field $h = 0$. For $\beta < 1$, the potential $\phi(m)$ has only one maximum corresponding to a disordered phase $m^\star = 0$. For $\beta > 1$, the system has two ordered ferromagnetic phases corresponding to the emergence of two symmetric global maxima $\pm m^\star$.

the physicist, it means that "macroscopic" quantities such as the magnetization are entirely deterministic, as their random fluctuations around the mean are negligible: this concentration of the measure is at the root of the success of Statistical Mechanics.

In the Curie-Weiss model, however, if $h = 0$ and $\beta > 1$, then the global maximum is *not* unique: there are two degenerate maxima at $\pm m^*$. This means that if one samples a configuration from the Boltzmann measure, then with a probability $1/2$ the configuration will have a magnetization $\pm m^*$, see Figs. 1.3 and 1.4. This situation is called *phase coexistence* in Physics, and is a fundamental property of liquids, solids and gases in nature. Phase coexistence arises only for $h = 0$ and $\beta > 1$ in the Curie-Weiss model. Indeed, as shown in Fig. 1.2, for $h \neq 0$ there might be more than one solution to the mean-field equations, but there is only one global maximum corresponding to a single phase, also known as a single *Gibbs state*.

1.2 The free energy/entropy is all you need

In more involved models one cannot hope to directly access the statistical distribution of relevant quantities with a direct computation, as we did in Theorem 1. However, we should expect to find similar phenomenology: concentration of macroscopic quantities in the thermodynamic limit, a large deviation principle for the Boltzmann measure, single phase vs phase coexistence, and of course, phase transitions. It turns out that almost all these phenomena can be understood from the computation of the asymptotic free entropy density $\Phi(\beta, h)$. This is why computing this quantity is the single most important analytical task in Statistical Physics.

First, a piece of trivia: most physicists do not use the free entropy or the free entropy density — for reasons rooted in the history of thermodynamics, dating back to Carnot and Clausius — but rather the *free energy* and the corresponding *free energy density*, which are the same as the

free entropy up to a $-\beta^{-1}$ term. In the Curie-Weiss model, this gives:

$$\begin{aligned} f_N(\beta,h) &=: -\frac{1}{\beta N}\log Z_N(\beta,h) \qquad \left(= -\frac{1}{\beta}\Phi_N(\beta,h)\right) \\ f(\beta,h) &=: \lim_{N\to\infty} f_N(\beta,h) = \min_m \; f(m,\beta,h) \\ f(m,\beta,h) &=: -\frac{1}{2}m^2 - hm - \frac{1}{\beta}H(m) \qquad \left(= -\frac{1}{\beta}\phi(m)\right) \end{aligned}$$

The fact that physicists (since Clausius) use the free energy with a factor $-\beta^{-1}$ in front of the log seems to be a notational problem for many mathematicians, who just cannot understand why they should bother with a trivial minus sign. Thus, many of them simply refer to $\Phi(\beta,h)$ as the free energy density (or worse, sometime using the terminology "pressure" from the theory of gases) which should make Clausius turn in his grave. It is also common for mathematicians to define the Hamiltonian with a global minus sign with respect to the one used by Physicists. Since this monograph is not concerned with actual applications in Physics, we might forgive these bad habits. Nevertheless, we will attempt to use the correct terminology, so that we shall not, for instance, "maximize" the energy and "minimize" the entropy!

1.2.1 Derivatives of the free entropy

We now discuss how knowing the free entropy actually allows one to rediscover all the phenomena we have discussed. First, we notice that for any finite value of N, the free entropy is a generating functional for the (connected) moments of the magnetization $\bar{S}$. Denoting by $\langle\cdot\rangle_N$ the average with respect to the Boltzman measure, and recalling that

$$Z_N(\beta,h) = \sum_{\boldsymbol{S}\in\{-1,1\}^N} e^{-\beta\mathcal{H}_N^0 + \beta N h\bar{S}}$$

we have:

$$\frac{1}{\beta}\frac{\partial}{\partial h}\Phi_N(\beta,h) = \frac{\partial}{\partial h}\frac{1}{\beta N}\log Z_N(\beta,h) = \sum_{\boldsymbol{S}\in\{-1,1\}^N}\frac{\bar{S}e^{-\beta\mathcal{H}_N}}{Z_N(\beta,h)} = \langle\bar{S}\rangle_N = m_N, \tag{1.10}$$

This also shows why it is useful to introduce an external magnetic field h at the beginning of our derivations: we can take obtain moments the moments of the magnetization by taking derivatives with respect to h. While the second derivative yield the variance, etc... However, it is far from trivial that this relation holds when the limit $N\to\infty$ is taken: the mathematical conditions for switching the limit and the derivative are non-trivial. Indeed, this can only be done away from the phase transition, in which case it follows from the convexity of the free entropy. Consider the second derivative with respect to h, we find:

$$\frac{1}{\beta}\frac{\partial^2}{\partial h^2}\Phi_N(\beta,h) = \frac{\partial}{\partial h}\sum_{\boldsymbol{S}\in\{-1,1\}^N}\frac{\bar{S}e^{-\beta\mathcal{H}_N}}{Z_N(\beta,h)} = N\beta(\langle\bar{S}^2\rangle_N - \langle\bar{S}\rangle_N^2) \geq 0\,. \tag{1.11}$$

Therefore Φ_N is convex. This result is also known in the statistical physics context as the *fluctuation-dissipation theorem*. A fundamental theorem on the limit of a sequence of convex functions f_n as $n\to\infty$ tells us that if $f_n(x)\to f(x)$ for all x, and if $f_n(x)$ is convex, then $f_n'(x)\to f'(x)$ for all x where $f(x)$ is differentiable. Out of the phase transition points, where

the free entropy is singular, the derivative of the asymptotic free entropy thus yields the asymptotic magnetization. Let us check that this is true. We know that $\Phi(\beta, h) = \phi(m^*(h, \beta))$, therefore we write:

$$\frac{1}{\beta}\frac{\partial}{\partial h}\Phi(\beta, h) = m^*(h, \beta) + \left.\frac{\partial \phi}{\partial m}\right|_{m^*} \frac{\partial m^*}{\partial h} = m^*(h, \beta)\,. \tag{1.12}$$

given that the derivative of $\phi(m)$ is zero when evaluated at m^*. The derivative of the free entropy has thus given us the *equilibrium* magnetization m^*, as it should.

1.2.2 Legendre transforms

Another instructive way to look at the problem arises using two important mathematical facts coming from prominent French mathematicians: Laplace and Legendre. First, let's state the very useful Laplace method for computing integrals.

Theorem 2 (Laplace method)**.** *Suppose $f(x)$ is a twice continuously differentiable function on $[a, b]$ and there exists a unique point $x_0 \in (a, b)$ such that:*

$$f(x_0) = \max_{x\in[a,b]} f(x) \quad \textit{and} \quad f''(x_0) < 0\,,$$

then:

$$\lim_{n\to\infty} \frac{\int_a^b e^{nf(x)}\,\mathrm{d}x}{e^{nf(x_0)}\sqrt{\frac{2\pi}{n(-f''(x_0))}}} = 1\,,$$

and in particular

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\log\int_a^b e^{nf(x)}\,\mathrm{d}x &= f(x_0) \\ \lim_{n\to\infty} \frac{\int_a^b g(x)e^{nf(x)}dx}{\int_a^b e^{nf(x)}\,\mathrm{d}x} &= g(x_0) \end{aligned} \tag{1.13}$$

These formulas, proven by Laplace in his fondamental text "Mémoire sur la probabilité des causes par les évènements" in 1774, have profound implications when combined with the large deviation principle.

Consider the Curie-Weiss model with zero field ($h = 0$). By definition, we have

$$\begin{aligned} \mathbb{P}(\bar{S} = m; h = 0) &= \frac{\sum_{\boldsymbol{S}\in\{-1,1\}^N} e^{-\beta\mathcal{H}_N^0}\mathbf{1}(\bar{S} = m)}{Z_N(\beta, h = 0)} \\ &= \frac{\sum_{\boldsymbol{S}\in\{-1,1\}^N} e^{-\beta\mathcal{H}_N^0}\mathbf{1}(\bar{S} = m)}{e^{N\log\Phi_N(\beta,0)}} \underset{N\to\infty}{\asymp} e^{-NI_0^*(m)} \end{aligned}$$

where we have denoted $I_0^*(m)$ the true large deviation rate at zero external field. Simply by assuming this large deviation principle, we can reach deep conclusions even if we *do not know* the actual expression of $I_0^*(m)$. Indeed, we can write the total partition sum of the

system *in presence of an external field*, as a **Laplace integral** over the possible values of m. Using $\mathcal{H}_N = \mathcal{H}_N^0 - Nh\bar{S}$ we write:

$$Z_N(\beta, h) = \sum_{m \in \mathcal{S}_N} \left(\sum_{\boldsymbol{S} \in \{-1,1\}^N} e^{-\beta \mathcal{H}_N^0} \mathbf{1}(\bar{S} = m) \right) e^{N\beta h m}$$
$$\underset{N\to\infty}{\asymp} \int_{-1}^{1} \mathrm{d}m \; e^{N\left(\Phi(\beta,0) - I_0^*(m,\beta) + \beta h m\right)} . \tag{1.14}$$

At this point the Laplace method applied to the limit of $Z_N(\beta, h)$ gives us automatically that

$$\Phi(\beta, h) - \Phi(\beta, 0) = \max_m \left[\beta h m - I_0^*(m)\right] = \phi(m^*)$$

We thus obtain a very generic relation between the free entropy of the system in a field and the large deviation rate (without field) $I_0^*(m)$: they are related through a **Legendre transform**:

$$\Phi(\beta, h) - \Phi(\beta, 0) = \max_m \left[\beta h m - I_0^*(m)\right]$$

In fact, the theory of Legendre transforms tells us slightly more: if we further take the Legendre transform of $\Phi(\beta, h)$ we can (almost) recover the true rate $I_0^*(m)$. Let us define

$$I_0(m) = \max_h \left(\beta h m - \Phi(\beta, h)\right) + \Phi(\beta, 0)$$

then a fundamental property of the Legendre transform reveals that $I_0(m)$ is the *convex* envelope of $I_0^*(m)$. We thus can recover the large deviation rate even simply by "Legendre-transforming" the free entropy. Again, we see that everything can be computed through the knowledge of the free entropy $\Phi(\beta, h)$. Truly, the free entropy is all you need.

Note however that there is a fundamental limitation to the ability to compute large deviation rates with this technique. Given we can only recover the convex envelope of the true rate, if the true rate is not convex there is a part of the curve that we cannot not obtain! This is illustrated in Figure 1.5: we only get the part of $I_0^*(m)$ that coincides with the convex envelope $I_0(m)$ (this set is called the "exposed points" of $I_0^*(m)$), while the other points are just given an upper bound. These considerations are classical in statistical mechanics, and at the basis of the "equivalence of ensembles" as well as to the derivation of thermodynamics (which is really nothing but Legendre transforms).

1.2.3 Gartner-Ellis Theorem

What we have said above can also be written rigorously using the language of modern large deviation theory. In particular, the Gartner-Ellis Theorem connects the large deviation rates with the Legendre transform of the partition sum in a very generic way:

Theorem 3 (Gartner-Ellis, informal)**.** *If*

$$\lambda(k) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}(\exp(NkA_N))$$

exists and is differentiable for all $k \in \mathbb{R}$, then defining

$$I(a) = \sup_k (ka - \lambda(k)) ,$$

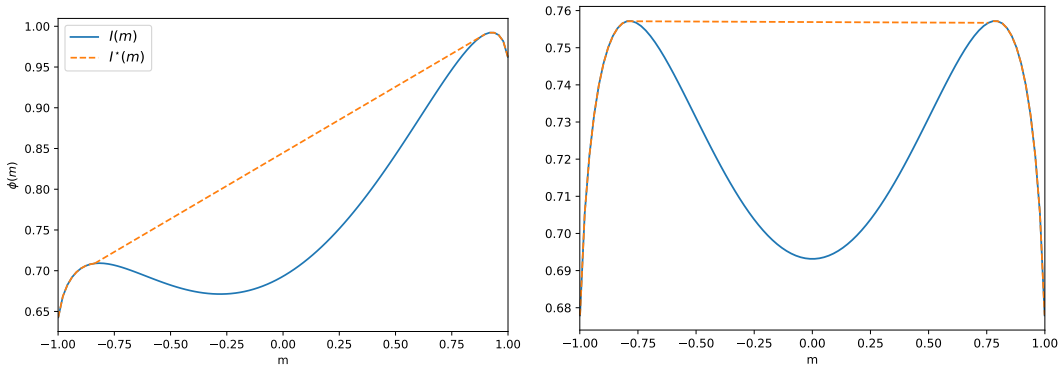


Figure 1.5: The true large deviation rate $I(m)$, and the function $I^*(m)$ obtained after taking two consecutive Legendre transforms. $I^*(m)$ is the convex envelope of $I(m)$. The two functions coincide for all the "exposed points", where the tangent is different from the curve, but they differ in the dashed region, which is not "exposed". Legendre transforms allows only to compute sharp large deviation rates for the exposed points, and upper bounds otherwise.

we have the large deviation principle

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(A_N = a) \leq -I(a)$$

with equality for the exposed points

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(A_N = a) = -I(a) \; \forall a \in \{\text{exposed points}\}\,.$$

In the Curie-Weiss model, the connection with our former derivation is immediate: consider the model without a field (i.e. $h = 0$) and let $A_N = \bar{S}$ be the averaged magnetization in the system. The theorem (which in this particular case is called Cramer's theorem) thus tells us that we need to compute:

$$\lambda(k) = \lim_{N\to\infty} \frac{1}{N} \log \sum_{\boldsymbol{S}} \frac{1}{Z_N} e^{-\beta \mathcal{H}_N^0 + Nk\bar{S}} = \lim_{N\to\infty} (\Phi_N(\beta, h = k/\beta) - \Phi_N(\beta, h = 0))$$

and we thus recognise

$$\lambda(k) = \max_m \phi(m, h = k/\beta) = \max_m I(m) + \beta h m$$

while the rate function for the magnetization (at zero field) is given by

$$I(m) = \max_k km - \lambda(k)$$

Given the property of the Legendre transform, $I(m)$ is indeed given by the convex envelope of $\phi(m, 0)$, as expected.

1.3 Toolbox: Gibbs free entropy and the variational approach

One cannot hope that the computation of the partition sum will always be so easy, so it is worth learning a few tricks. A fundamental result in statistical physics, that is at the roots

of many analytical and practical approaches, and that has been used extensively in machine learning as well, is the following one:

Theorem 4 (Gibbs variational approach)**.** *Consider a Hamiltonian $\mathcal{H}_N(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^N$, and the associated Boltzmann-Gibbs measure $\mathcal{P}_{\text{Gibbs}}(\boldsymbol{x}) = e^{-\beta \mathcal{H}_N(\boldsymbol{x})}/Z_N(\beta)$. Given an arbitrary probability distribution $\mathcal{Q}(\boldsymbol{x})$ over $\mathbb{R}^N$ and its entropy $\mathcal{S}[\mathcal{Q}] = -\langle \log \mathcal{Q} \rangle_{\mathcal{Q}}$, let the Gibbs functional be*

$$N\phi^{\text{Gibbs}}(\mathcal{Q}) =: \mathcal{S}[\mathcal{Q}] - \beta \langle \mathcal{H}_N \rangle_{\mathcal{Q}}$$

where $\langle . \rangle_{\mathcal{Q}}$ denotes the expectation with respect the distribution $\mathcal{Q}$. Then

$$\forall \mathcal{Q}, \;\; \Phi_N(\beta) = \frac{1}{N} \log Z_N(\beta) \geq \phi^{\text{Gibbs}}(\mathcal{Q})$$

with equality when $\mathcal{Q} = \mathcal{P}_{\text{Gibbs}}$.

Let us first prove a simpler result. We shall introduce the following quantity, known as the Kullback-Leibler divergence (or "relative entropy"), which measures how much two distributions $\mathcal{P}(\boldsymbol{x})$ and $\mathcal{Q}(\boldsymbol{x})$ differ:

$$D_{\text{KL}}(\mathcal{P}||\mathcal{Q}) =: \int d\boldsymbol{x} \mathcal{P}(\boldsymbol{x}) \log \frac{\mathcal{P}(\boldsymbol{x})}{\mathcal{Q}(\boldsymbol{x})}$$

We can prove the following lemma:

Lemma 3 (Gibbs inequality)**.**

$$D_{\text{KL}}(\mathcal{P}||\mathcal{Q}) \geq 0$$

with equality if and only if $\mathcal{P} = \mathcal{Q}$ almost everywhere.

Proof. For any scalar $u \in \mathbb{R}^+$ we have $\log(u) \leq u - 1$, with equality at $u = 1$. Therefore

$$\begin{aligned} -D_{\text{KL}}(\mathcal{P}||\mathcal{Q}) &= \int d\boldsymbol{x} \mathcal{P}(\boldsymbol{x}) \log \frac{\mathcal{Q}(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{x})} \\ &\leq \int d\boldsymbol{x} \mathcal{P}(\boldsymbol{x}) \left(\frac{\mathcal{Q}(\boldsymbol{x})}{\mathcal{P}(\boldsymbol{x})} - 1 \right) = \int d\boldsymbol{x} \mathcal{Q}(\boldsymbol{x}) - \int d\boldsymbol{x} \mathcal{P}(\boldsymbol{x}) = 1 - 1 = 0 \end{aligned}$$

Given the $\log(u) < u - 1$ unless $u = 1$, the inequality is strict unless $\mathcal{P} = \mathcal{Q}$ almost everywhere. □

We then write the difference between the Gibbs free entropy and the actual entropy using the Kullback-Leibler divergence:

Lemma 4. *Denoting the Boltzmann-Gibbs probability distribution as $\mathcal{P}_{\text{Gibbs}}(\boldsymbol{x}) = e^{-\beta \mathcal{H}(\boldsymbol{x})}/Z_N$, and an arbitrary distribution as $\mathcal{Q}$, we have*

$$N\Phi_N = N\phi^{\text{Gibbs}}(\mathcal{Q}) + D_{\text{KL}}(\mathcal{Q}||\mathcal{P}_{\text{Gibbs}})$$

Proof. The proof is a trivial application of definition of the divergence:

$$\begin{aligned} \langle \log \mathcal{P}_{\text{Gibbs}} \rangle_{\mathcal{Q}} &= -\beta \langle \mathcal{H}_N \rangle_{\mathcal{Q}} - \log Z_N \\ \langle \log \mathcal{P}_{\text{Gibbs}} \rangle_{\mathcal{Q}} - \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} &= -\beta \langle \mathcal{H}_N \rangle_{\mathcal{Q}} - N\Phi_N - \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} \\ -D_{\text{KL}}(\mathcal{Q}||\mathcal{P}_{\text{Gibbs}}) &= N\phi^{\text{Gibbs}}(\mathcal{Q}) - N\Phi_N \\ N\Phi_N &= N\phi^{\text{Gibbs}}(\mathcal{Q}) + D_{\text{KL}}(\mathcal{Q}||\mathcal{P}_{\text{Gibbs}}) \end{aligned}$$

□

Together, lemma 3 and lemma 4 imply theorem 4. Why is this interesting? Basically, it gives us a way to approximate the partition sum (and the true distribution) by using many "trial" distributions, and piking up the one with the largest free entropy. This has been used in countless many ways since the birth of statistical and quantum physics, under many names (for instance "Gibbs-Bogoliubov-Feynman"), and it is at the root of many applications of Bayesian Statistics and machine learning as well, in which case the Gibbs free entropy is called Evidence Lower BOund, or ELBO in short.

Let us see how it can be used for the Curie-Weiss model. The simplest thing we could try is a factorized distribution, identical for all spins:

$$\mathcal{Q}(\boldsymbol{S}) = \prod_i \mathcal{Q}_i(S_i) = \prod_i \left(\frac{1+m}{2}\delta(S_i - 1) + \frac{1-m}{2}\delta(S_i + 1) \right) \tag{1.15}$$

Then, we can write the Gibbs free entropy density as

$$\phi^{\text{Gibbs}}(\mathcal{Q}) = \beta \langle (\bar{S})^2/2 + h\bar{S} \rangle_{\mathcal{Q}} + \frac{1}{N}\sum_i H(\mathcal{Q}_i) \tag{1.16}$$

$$= \beta m^2/2 + \beta h m + H(m) \tag{1.17}$$

with $H(m)$ the binary entropy, and we find back the correct free entropy — at any fixed m — through this simple variational ansatz.

1.4 Toolbox: The cavity method

Here, we are going to see another important method: the cavity trick. It will be at the root of many of our computations in the course. The whole idea is based on the following question: what happens when you add one variable to the system? Physicists like imagery: one could visualize a system of N variables, making a little "hole" or "cavity" in it, and delicately adding one variable, hence the name "cavity method".

Let us see how it works by comparing the two Hamiltonians with N and $N+1$ variables. Denoting the new spin as the number "0" we have, for a system at inverse temperature β' and field h':

$$\begin{aligned} -\beta' \mathcal{H}_{N+1} &= \beta' \frac{1}{2}(N+1) \left(\frac{S_0 + \sum_i S_i}{N+1} \right)^2 + \beta' h' (S_0 + \sum_i S_i) \\ &= \beta' \frac{1}{2(N+1)} + \frac{\beta'}{2}\frac{N^2}{N+1}\left(\frac{\sum_i S_i}{N} \right)^2 + \beta' S_0 \frac{N}{N+1}\left(\frac{\sum_i S_i}{N} \right) + \beta' h' \sum_i S_i + \beta' h' S_0 \end{aligned} \tag{1.18}$$

If we define $\beta' = \beta(N+1)/N$, and our new field h' as $h' = hN/(N+1)$, we get

$$\begin{aligned} -\beta' \mathcal{H}_{N+1}(h') &= \text{cst} + \frac{\beta}{2} N \left(\frac{\sum_i S_i}{N} \right)^2 + \beta S_0 \left(\frac{\sum_i S_i}{N} \right) + \beta h \sum_i S_i + \beta h S_0 \\ &= \text{cst} - \beta \mathcal{H}_N + \beta S_0 \left(\frac{\sum_i S_i}{N} \right) + \beta h S_0 \end{aligned}$$

We thus have two systems: one with $N+1$ spins at temperature and fields (β', h') and one with N spins at temperature and fields (β, h). The relation we just derived makes the expectation over the $N+1$ variable easy to compute in the new system, as a function of the sum in the old system. In fact, one can directly write the expectation of the new variable as follows:

$$\langle S_0 \rangle_{N+1,\beta'} = \frac{\sum\limits_{S_0,\boldsymbol{S}} S_0 e^{-\beta' \mathcal{H}'_{N+1}}}{\sum\limits_{S_0,\boldsymbol{S}} e^{-\beta' \mathcal{H}'_{N+1}}} = \frac{\sum\limits_{\boldsymbol{S}} \sum\limits_{S_0} S_0 e^{-\beta \mathcal{H}_N + \beta S_0 \bar{S} + \beta h S_0}}{\sum\limits_{\boldsymbol{S}} \sum\limits_{S_0} e^{-\beta \mathcal{H}_N + \beta S_0 \bar{S} + \beta h S_0}} = \frac{\langle \sinh\left(\beta(\bar{S}+h)\right) \rangle_{N,\beta}}{\langle \cosh\left(\beta(\bar{S}+h)\right) \rangle_{N,\beta}}$$

Assuming, as a physicist would do, that $\bar{S}$ concentrates on a deterministic value m^* (at least out of the phase co-existence line), and assuming that m^* should be the same for N and $N+1$ systems when N is large enough (as it should), we have immediately:

$$\langle S_0 \rangle_{N+1,\beta'} = \frac{\langle \sinh\left(\beta(\bar{S}+h)\right) \rangle_{N,\beta}}{\langle \cosh\left(\beta(\bar{S}+h)\right) \rangle_{N,\beta}} \approx \frac{\sinh\left(\beta(m^*+h)\right)}{\cosh\left(\beta(m^*+h)\right)} = \tanh\left(\beta(m^*+h)\right)$$

As $N \to \infty$, the difference between β and β' vanished, and we recover the mean field equation:

$$m^* = \tanh(\beta(m^*+h))$$

We can also recover the free energy from a similar cavity argument. First we note that

$$\frac{1}{N} \log Z_N = \frac{1}{N} \log \frac{Z_N}{Z_{N-1}} Z_{N-1} = \frac{1}{N} \log \frac{Z_N}{Z_{N-1}} \frac{Z_{N-1}}{Z_{N-2}} \dots \frac{Z_1}{1} = \frac{1}{N} \sum_{n=0}^{N} \log \frac{Z_{n+1}}{Z_n} \tag{1.19}$$

so that (a rigorous argument can be made using Cesaro sums, see appendix 1.C):

$$\Phi(\beta, h) = \lim_{N \to \infty} \log \frac{Z_{N+1}(\beta, h)}{Z_N(\beta, h)}$$

Note, however, that the β should be the same at N and $N+1$ systems in this computation. Starting from equation 1.18, we thus write, making sure we keep all terms that are not $o(1)$:

$$\begin{aligned} -\beta \mathcal{H}_{N+1} &= o(1) + \frac{\beta}{2}(N - 1 + o(1)) \left(\frac{\sum_i S_i}{N} \right)^2 + \beta S_0 (1 + o(1)) \left(\frac{\sum_i S_i}{N} \right) + \beta h \sum_i S_i + \beta h S_0 \\ &= o(1) - \beta \mathcal{H}_N - \frac{\beta}{2} \left(\frac{\sum_i S_i}{N} \right)^2 + \beta S_0 \left(\frac{\sum_i S_i}{N} \right) + \beta h S_0 \end{aligned} \tag{1.20}$$

Notice the presence of the term $-\beta(\bar{S})^2/2$, which we could have overlooked, have we been less cautious. We did not keep this term in the previous computation, because we could include it in the β', and it was not making any difference in the large N limit and could be absorbed in the normalisation at the price of a minimal $o(1)$ change in β. Here, however, we need to compute Z_{N+1}/Z_N, which is $O(1)$, so we need to pay attention to any constant correction. We can now finally compute[3]

$$\frac{Z_{N+1}}{Z_N} = \langle e^{-\frac{\beta}{2}\bar{S}^2} 2 \cosh\left(\beta(\bar{S}+h)\right) \rangle_{N,\beta} \tag{1.21}$$

[3]Let us make a remark that shall be useful later on, when we shall discuss the cavity method on sparse graphs: the new Hamiltonian has two terms in addition to the old one: a "site" term that is a function of S_0, and a "link" term, that appears because, on top of adding one spin, we added N links to the systems. This will turn out to be a generic property.

Using now the concentration of $\bar{S}$ to m^*, we get, taking the log

$$\Phi(\beta, h) = -\beta \frac{m^{*2}}{2} + \log\left[2\cosh(\beta(m^* + h))\right] \tag{1.22}$$

This might seem a little bit odd, since it does not look the same as the former expression! Indeed we now find:

$$\Phi(\beta, h) = \max_m \tilde{\phi}(m) \qquad \text{with} \qquad \tilde{\phi}(m) =: -\beta \frac{m^2}{2} + \log 2\cosh(\beta(m + h)) \tag{1.23}$$

And indeed, we are taking the maximum since the derivative of $\tilde{\phi}(m)$ with respect to m yields $m^* = \tanh(\beta(m^* + h))$. Did we make a mistake? No, we did not! While $\tilde{\phi}(m)$ and $\phi(m)$ are different, both have the same value at any of their extrema, as can be checked by a simple plot. Indeed, we can show analytically that their fixed points are identical: Remember that $m^* = \tanh(\beta(m^* + h))$ so that $\beta(m^* + h) = \operatorname{atanh}(m^*)$, and therefore, using the identity

$$\log\left[2\cosh\left(\operatorname{atanh}(x)\right)\right] - x\,\operatorname{atanh}(x) = H(x) = -\left(\frac{1+x}{2}\log\frac{1+x}{2} + \frac{1-x}{2}\log\frac{1-x}{2}\right)$$

we obtain:

$$\begin{aligned}
\tilde{\phi}(m^*) &= -\beta\frac{m^{*2}}{2} + \log\left[2\cosh(\beta(m^* + h))\right] \\
&= -\beta\frac{m^{*2}}{2} + \log\left[2\cosh(\operatorname{atanh}(m^*))\right] \\
&= -\beta\frac{m^{*2}}{2} + m^*\operatorname{atanh}(\mathrm{m}^*) + H(m^*) \\
&= -\beta\frac{m^{*2}}{2} + \beta m^{*2} + \beta m^* h + H(m^*) \\
&= \phi(m^*)
\end{aligned}$$

One should not, however, assume that the last expression $\tilde{\phi}(m)$ is the correct large deviation quantity: it is not. The reader is invited to check that the Legendre transform of $\Phi(\beta, h)$ is giving back the correct large deviation function ϕ, as it should.

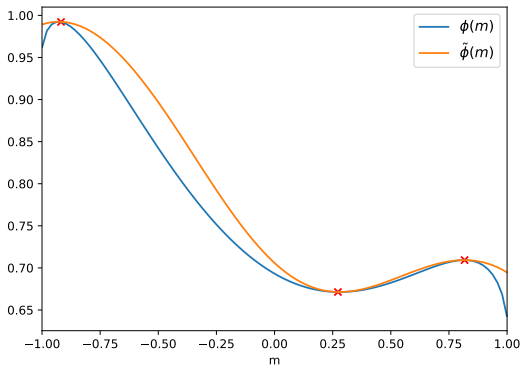


Figure 1.6: The functions $\phi(m)$ (1.4) and $\tilde{\phi}(m)$ (1.23) are different but coindide for all their extremums.

All these considerations can be made entirely rigorous, as we show in Appendix 1.C. This is one of the very important tools that mathematicians may use to prove some of the results discussed in this monograph.

1.5 Toolbox: The "field-theoretic" computation

To conclude this lecture, it will be also worth learning a trick that physicists use a lot, and that will be central to all replica computations in the next chapters. It is probably a good idea to learn it in the context of the simple Curie-Weiss model. We write again the Hamiltonian

$$\mathcal{H}_N^0 = -\frac{1}{N}\sum_{i<j} S_i S_j \approx -\frac{1}{2N}\sum_{i,j} S_i S_j - \sum_i S_i$$

as well as the partition sum

$$Z_N = \sum_{\boldsymbol{S}} e^{\frac{N\beta}{2}(\sum_j \frac{S_j}{N})^2 + N\beta h \sum_i \frac{S_i}{N}}$$

Now, let us pretend we do not know how to compute the binomial coefficient $\Omega(m)$. Instead, we are going to use the so-called "Dirac-Fourier" method, which starts with the following identity for the delta "function":

$$\int \mathrm{d}u\, f(u)\delta(u-x) = f(x)$$

In this case, physicists would actually use the following version of the identity:

$$\begin{aligned}\int \mathrm{d}m\, f(m)\delta(Nm-x) &= \int \mathrm{d}m\, \delta\left(N\left(m-\frac{x}{N}\right)\right) f(m) = \frac{1}{N}\int \mathrm{d}m\, f(m)\delta\left(m-\frac{x}{N}\right)\\ &= \frac{1}{N} f\left(\frac{x}{N}\right)\end{aligned}$$

but notice that the additional $1/N$ term does not matter since we take the log and divide by N: it thus makes no difference asymptotically. We then write:

$$\begin{aligned}Z_N &= \sum_{\boldsymbol{S}} e^{-\beta\mathcal{H}_N} = N\sum_{\boldsymbol{S}} \int \mathrm{d}m\, \delta\left(Nm - \sum_j S_j\right) e^{\frac{N\beta}{2}m^2 + N\beta h m}\\ &= N\int \mathrm{d}m\, e^{\frac{N\beta}{2}m^2+\beta h N m} \sum_{\boldsymbol{S}} \delta\left(Nm - \sum_j S_j\right)\end{aligned}$$

We recognize that one would need to compute the entropy at fixed m. Again, let us pretend we cannot compute it (the idea is to do so without any combinatorics). Instead, using a Fourier transform of the delta "function" (which is really a distribution), one finds that

$$Z_N = N\int \mathrm{d}m \int \mathrm{d}\lambda\, e^{\frac{N\beta}{2}m^2+\beta h m N} \sum_{\boldsymbol{S}} e^{i2\pi\lambda N\left(m - \sum_j \frac{S_j}{N}\right)}$$

We do not like to keep the i explicitly, so instead, we write $\hat{m} = i2\pi\lambda$ and integrate in the complex plane

$$\begin{aligned}Z_N &= 2iN\pi \int_{-1}^{1} \mathrm{d}m \int_{-i2\pi\infty}^{i2\pi\infty} \mathrm{d}\hat{m}\, e^{\frac{N\beta}{2}m^2+\beta h N m + N\hat{m}m} \sum_{\boldsymbol{S}} e^{-\hat{m}\sum_j S_j}\\ &= 2iN\pi \int_{-1}^{1} dm \int_{-i2\pi\infty}^{i2\pi\infty} d\hat{m}\, e^{N\frac{\beta}{2}m^2 + N\beta h m + N\hat{m}m} \left(2\cosh\hat{m}\right)^N\end{aligned}$$

We are interested only in the density of the log, and thus we write:

$$\frac{\log Z_N}{N} = \frac{1}{N} \log \left\{ \int_{-1}^{1} dm \int_{-i2\pi\infty}^{i2\pi\infty} d\hat{m}\, e^{\frac{N\beta}{2}m^2 + N\beta h m + N\hat{m}m + N\log 2 + N\log\cosh\hat{m}} \right\} + o(1)$$

We have gotten read of the combinatoric sums, at the price of introducing integrals. It turns out, however, that these are very easy integrals, at least when N is large. The trick is to use Cauchy's theorem to deform the contour integral in $\hat{m}$ and put the path *right into* a saddle in the complex plane. This is called the saddle point method, and it is a generalization of Laplace's method to the complex plane. More information is provided in Exercice 1.3, but what the saddle point methods tell us is that for integrals of this type, we have

$$\Phi(\beta, h) = \lim_{N\to\infty} \frac{\log Z}{N} = \mathrm{extr}_{m,\hat{m}}\, g(m, \hat{m})$$

with

$$g(m, \hat{m}) = \frac{\beta}{2}m^2 + \beta h m + \hat{m}m + \log 2 + \log\cosh\hat{m}$$

In other words, the complex two-dimensional integral has been replaced by a much simpler differentiation where we just need to find the extremum of $g(m, \hat{m})$!

How consistent is this expression with the previous one we obtained? Let us do the differentiation over $\hat{m}$: the extrema condition imposes $m = -\tanh\hat{m}$, or $\hat{m} = -\tanh^{-1} m$. If we plug this into the expression, we finally find

$$g(m) = \frac{\beta}{2}m^2 + \beta h m - m\,(\tanh^{-1} m) + \log\left(2\cosh\tanh^{-1} m\right)$$

This does not look directly like our good old $\phi(m)$, but by using the following identity:

$$\log(2\cosh(\mathrm{atanh}(m)) - m\mathrm{atanh}(m) = H(m),$$

we get back

$$g(m) = \phi(\beta, m) + \beta h m = \frac{\beta}{2}m^2 + \beta h m + H(m) = \phi(m, \beta)$$

and $\Phi(\beta, h) = \mathrm{extr}_m\, \phi(\beta, m)$ as it should.

It is worth noting, however, that most physicists choose to write things in a slightly different, but equivalent, way. Indeed, starting again from

$$g(m, \hat{m}) = \frac{\beta}{2}m^2 + \beta h m + \hat{m}m + \log 2 + \log\cosh\hat{m}$$

the typical physicist imposes $\hat{m} = -\beta(m + h)$ instead of $m = -\tanh\hat{m}$. They would then get rid of $\hat{m}$ instead of m, since it is most convenient, and write:

$$\Phi(\beta, h) = \mathrm{extr}_m \left(-\frac{\beta}{2}m^2 + \log(2\cosh\beta(m + h)) \right) = \max \tilde{\phi}(m, h, \beta)$$

As we have seen in the previous section in the cavity computation, this is not a problem, since this formula is correct as well. In fact, it is reassuring that both formulations can be found using this method.

Bibliography

The Curie-Weiss model is inspired from the pioneering works of the Frenchmen Curie (1895) and Weiss (1907). The history of mean-field models and variational approaches in physics is well described in Kadanoff (2009). The presentation of the rigorous solution of the Curie-Weiss model follows Dembo et al. (2010a). The Laplace method was introduced by Pierre Simon de Laplace in its revolutionary work (Laplace, 1774), where he founded the field of statistics. The saddle point method was first published by Debye (1009), who in turns credited it to an unpublished note by Riemann (1863). A classical reference on large deviation is Dembo et al. (1996). The nice review by Touchette (2009) covers the large deviation approach to statistical mechanics, and is a recommended read for physicists. Variational approaches in statistics and machine learning are discussed in great detail in Wainwright and Jordan (2008).

1.6 Exercises

Exercise 1.1: Bounds on the binomial

We wish to prove the bounds on the Binomial coefficient used in the derivation at the beginning of the chapter:

$$\frac{e^{nH(k/m)}}{n+1} < \binom{n}{k} = \frac{n!}{k!(n-k)!} < e^{nH(k/n)}$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ is the Binomial entropy (we used $p = (1+m)/2$ in the chapter).

(a) Use the Binomial to prove that, for any $0 < p < 1$

$$1 = \sum_{i=0}^{n} \binom{n}{i} (1-p)^i p^{n-i}$$

(b) Using a particular value of p, and keeping only one term of the sum, show that

$$\binom{n}{k} < e^{nH(k/n)}$$

(c) If one makes n draws from a Bernoulli variable with probability of positive outcome $p = k/n$, what is the most probable value for the number of positive outcomes? Deduce that

$$(n+1)\binom{n}{k} > e^{nH(k/n)}$$

Exercise 1.2: Laplace method

In this exercise, we will be interested in the asymptotic behaviour ($\lambda \to \infty$) of the following class of real integrals:

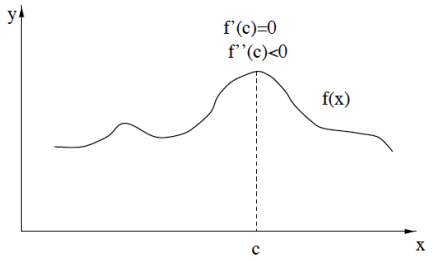
$$I(\lambda) = \int_a^b \mathrm{d}t\; h(t) e^{\lambda f(t)}$$

(a) Intuitively, what are the regions in the interval $[a, b]$ which will contribute more to the value of $I(\lambda)$?

(b) Suppose the function $f(t)$ has a single global maximum at a point $c \in [a, b]$ such that $f''(c) < 0$, and assume $h(c) \neq 0$. Using a Taylor expansion for f, show that for $\lambda \gg 1$ we expect the integral to behave as follows:

$$I(\lambda) \underset{\lambda \gg 1}{=} \int_{c-\epsilon}^{c+\epsilon} \mathrm{d}t\; h(c) e^{\lambda\left[f(c) + \frac{1}{2}f''(c)(t-c)^2\right]}$$

where $\epsilon > 0$ is a positive but small real number.



(c) Using your result above, conclude that:

$$I(\lambda) \asymp \frac{h(c)e^{\lambda f(c)}}{\sqrt{-\lambda f''(c)}} \int_{\mathbb{R}} e^{-t^2}\,\mathrm{d}t.$$

(d) (Gaussian integral) Show that:

$$\int_{\mathbb{R}} e^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$$

(e) Deduce Laplace's formula:

$$\int_a^b \mathrm{d}t\; h(t)e^{\lambda f(t)} \asymp \sqrt{\frac{2\pi}{-\lambda f''(c)}} e^{\lambda f(c)} h(c)$$

Exercise 1.3: Saddle point method

The saddle point method is a generalisation of Laplace's method to the complex plane. As before, we search for an asymptotic formula for integrals of the type:

$$I(\lambda) = \int_\gamma \mathrm{d}z\; h(z)e^{\lambda f(z)}$$

where $\gamma : [a, b] \to \mathbb{C}$ is a curve in the complex plane $\mathbb{C}$ and $\lambda > 0$ is a real positive number which we will take to be large. If the complex function f is holomorphic on a connected open set $\Omega \subset \mathbb{C}$, the integral $I(\lambda)$ is independent of the curve γ. The goal is therefore to choose γ wisely.

Part I: Geometrical properties of holomorphic functions
Let $f : \mathbb{C} \to \mathbb{C}$ be a holomorphic function, and let $z = x + iy \in \mathbb{C}$ for real $x, y \in \mathbb{R}$. Without loss of generality, we can write $f(z) = u(x, y) + iv(x, y)$ for $u, v : \mathbb{R}^2 \to \mathbb{R}$ real-valued functions. The goal of this exercise is to study the properties of f around a critical point $f'(z_0) = 0$ for $z_0 \in \mathbb{C}$.

(a) Show that at a critical point, the gradients of u and v are zero.

(b) Using the Cauchy integral formula, show that for all $z_0 = x_0 + iy_0$ in an open

convex set Ω where f is holomorphic we have:

$$u(x_0, y_0) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\theta \; u(x_0 + r\cos\theta, y_0 + r\sin\theta)$$
$$v(x_0, y_0) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\theta \; v(x_0 + r\cos\theta, y_0 + r\sin\theta)$$

for all circles of radius $r > 0$ centred at z_0 contained inside Ω. This result is known as the Mean Value Theorem in complex analysis.

(c) Conclude that neither u or v can have a local extremum (maximum or minimum) inside Ω. Therefore, all critical points $z_0 \in \Omega$ are necessarily saddle points of u and v.

(d) Let z_0 be a critical point of f such that $f''(z_0) \neq 0$. Using the Taylor series of f around z_0 and using the polar decompositions $f''(z_0) = \rho e^{i\alpha}$, $z - z_0 = re^{i\theta}$, find the values of $\theta \in [0, 2\pi)$ corresponding to the two directions of steepest-descent of u as a function of α in the complex plane.

Part II: Choosing the good curve γ

(a) What are the regions of γ which dominate the integral $I(\lambda)$?

(b) Let z_0 be a critical point $f'(z_0) = 0$. Explain why should we choose γ to pass through z_0 following the steepest-descent directions of the real part $\mathrm{Re[f]}$?

(c) Show that such a γ, we can rewrite the integral as:

$$I(\lambda) = e^{i\lambda \mathrm{Im[f(z_0)]}} \int_\gamma \mathrm{d}z \; h(z) e^{\lambda \mathrm{Re[f(z)]}}$$

(d) Let $\gamma(t) = x(t) + iy(t)$ for $t \in [a, b]$ be a parametrisation of the curve passing through $z_0 = \gamma(t_0)$ through the steepest-descent direction of $\mathrm{Re[f]}$. Letting $f(t) = f(\gamma(t))$, $h(t) = h(\gamma(t))$, $u(t) = \mathrm{Re[f(t)]}$ and $v(t) = \mathrm{Im[f(t)]}$, show that the problem boils down to the evaluation of the following integral:

$$\int_a^b \mathrm{d}t \; \gamma'(t) h(t) e^{\lambda u(t)}$$

Part III: Back to Laplace's method

(a) Suppose $h(t_0) \neq 0$, and note we can choose a parametrisation of γ such that $\gamma'(t_0) \neq 0$. Use Laplace's method to show that $I(\lambda)$ admits the following asymptotic expansion for $\lambda \gg 1$:

$$I(\lambda) \asymp h(t_0)\gamma'(t_0) \sqrt{\frac{2\pi}{-\lambda u''(t_0)}} e^{\lambda f(t_0)}$$

(b) Write the second derivative of $f(t)$ with respect to t and show that at the critical point z_0 we have:

$$\frac{\mathrm{d}^2 f(t_0)}{\mathrm{d}t^2} = \gamma'(t_0)^2 \frac{\mathrm{d}^2 f(z_0)}{\mathrm{d}z^2}$$

(c) Show that the second derivative $f''(t_0)$ is necessarily real and negative. Conclude that:

$$u''(t_0) = -|f''(z_0)||\gamma'(t_0)|^2$$

(d) Let θ be the angle between the curve γ and the real axis at the critical point z_0, see figure below. Show that:

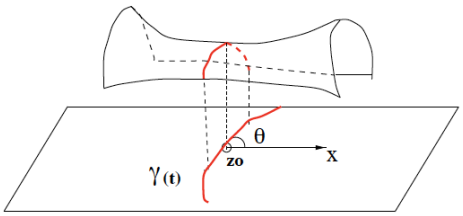
$$\gamma'(t_0) = |\gamma'(t_0)|e^{i\theta}$$

(e) Letting $f''(z_0) = |f''(z_0)|e^{i\alpha}$, show that $\theta = \frac{1}{2}(\pi - \alpha)$ or $\theta = \frac{1}{2}(\pi - \alpha) + \pi$ depending on the orientation of the curve γ.

(f) Conclude that:

$$I(\lambda) \asymp \pm h(z_0)e^{\lambda f(z_0)}e^{i\frac{\pi-\alpha}{2}}\sqrt{\frac{2\pi}{\lambda|f''(z_0)|}} = h(z_0)e^{\lambda f(z_0)}\sqrt{\frac{2\pi}{-\lambda f''(z_0)}}$$

where the $\pm$ is given by the orientation of the steepest-descent curve.



Exercise 1.4: Metropolis-Hastings algorithm

Consider again the Hamiltonian of the Curie-Weiss model. A very practical way to sample configurations of N spins from the Gibbs probability distribution

$$\mathbb{P}(\boldsymbol{S} = \boldsymbol{s}; \beta, h) = \frac{\exp\left(-\beta \mathcal{H}_N(\boldsymbol{s}; h)\right)}{Z_N(\beta, h)} \tag{1.24}$$

is the Monte-Carlo-Markov-Chain (MCMC) method, and in particular the Metropolis-Hastings algorithm. It works as follows:

1. Choose a starting configuration for the N spins values $s_i = \pm 1$ for $i = 1, \ldots, N$.
2. Choose a spin i at random. Compute the current value of the energy E_{now} and the value of the energy E_{flip} if the spins i is flipped (that is if $S_i^{\text{new}} = -S_i^{\text{old}}$).
3. Sample a number r uniformly in $[0, 1]$ and, if $r < e^{\beta(E_{\text{now}} - E_{\text{flip}})}$ perform the flip (i.e. $S_i^{\text{new}} = -S_i^{\text{old}}$) otherwise leave it as it is.
4. Go back to step 2.

If one is performing this program long enough, it is guarantied that the final configuration $\boldsymbol{S}$ will have been chosen with the correct probability.

(a) Write a code to perform the MCMC dynamics, and start by a configuration where all spins are equal to $S_i = 1$. Take $h = 0$, $\beta = 1.2$ and try your dynamics for a long enough time (say, with $t_{\max} = 100N$ attempts to flip spins) and monitor the value of the magnetization per spin $m = \sum_i S_i/N$ as a function of time. Make a plot for $N = 10, 50, 100, 200, 1000$ spins. Compare with the exact solution at $N = \infty$. Remarks? Conclusions?

(b) Start by a configuration where all spins are equal to 1 and take $h = -0.1$, $\beta = 1.2$. Monitor again the value of the magnetization per spin $m = \sum_i s_i/N$ as a function of time. Make a plot for $N = 10, 50, 100, 200, 1000$ spins. Compare with the exact solution at $N = \infty$. Remarks? Conclusions?

Exercise 1.5: Glauber Algorithm

An alternative local algorithm to sample from the measure eq. 1.24 is known as the Glauber or heat bath algorithm. Instead of flipping a spin at random, the idea is to thermalise this spin with its local environment.

Part I: The algorithm

(a) Let $\bar{S} = \frac{1}{N}\sum_{i=1}^{N} s_i$ be the total magnetisation of a system of N spins. Show that for all $i = 1, \cdots, N$, the probability of having a spin at $S_i = \pm 1$ given that all other spins are fixed is given by:

$$\mathbb{P}\left(S_i = \pm 1 | \{S_j\}_{j\neq i}\right) \equiv P_{\pm} = \frac{1 \pm \tanh(\beta(\bar{S} + h))}{2}$$

(b) The Glauber algorithm is defined as follows:

1. Choose a starting configuration for the N spins. Compute the magnetisation m_t and the energy E_t corresponding to the configuration.
2. Choose a spin S_i at random. Sample a random number uniformly $r \in [0, 1]$. If $r < P_+$, set $S_i = +1$, otherwise set $S_i = -1$. Update the energy and magnetisation.
3. Repeat step 2 until convergence.

Write a code implementing the Glauber dynamics. Repeat items (a) and (b) of exercise 1.4 using the same parameters. Compare the dynamics. Comment on the observed differences.

Part II: Mean-field equations from Glauber

Let's now derive the mean-field equations for the Curie-Weiss model from the Glauber algorithm.

(a) Let m_t denote the total magnetisation at time t, and define $P_{t,m} = \mathbb{P}(m_t = m)$. For simplicity, consider $\beta = 1$ and $h = 0$. Show that for $\delta \ll 1$ we can write:

$$\begin{aligned} P_{t+\delta t,m} = & P_{t,m+\frac{2}{N}} \times \left\{\frac{1}{2}\left(1 + m + \frac{2}{N}\right)\right\} \times \frac{1 - \tanh(m + 2/N)}{2} \\ & + P_{t,m-\frac{2}{N}} \times \left\{\frac{1}{2}\left(1 - m + \frac{2}{N}\right)\right\} \frac{1}{2}\left(1 + \tanh(m - 2/N)\right) \\ & + P_{t,m}\left\{\frac{1}{2}(1+m)\frac{1+\tanh(m)}{2} + \frac{1}{2}(1-m)\frac{1-\tanh(m)}{2}\right\}. \end{aligned}$$

This is known as the **master equation**.

(b) Defining the mean magnetisation with respect to $P_{t,m}$

$$\langle m(t)\rangle = \int \mathrm{d}m\; m\; P_{t,m}$$

and using the master equation above, show we can get an equations for the expected magnetisation:

$$\begin{aligned}\langle m(t+\delta t)\rangle = &\int P_{t,m+2/N} \times \left\{\frac{1}{2}\left(1+m+2/N\right)\right\} \times \frac{1-\tanh(m+2/N)}{2} \times m\,\mathrm{d}m \\ &+ \int P_{t,m-2/N} \left\{\frac{1}{2}(1-m+2/N)\right\} \times \frac{1+\tanh(m-2/N)}{2} \times m\,\mathrm{d}m \\ &+ \int P_{t,m} \times \left\{\frac{1+m}{2} \times \frac{1+\tanh(m)}{2} + \frac{1-m}{2} \times \frac{1-\tanh(m)}{2}\right\} \times m\,\mathrm{d}m\end{aligned}$$

(c) Making the change of variables $m \to m+2/N$ in the first integral and $m \to m-2/N$ in the second and choosing $\delta = \frac{1}{N}$, conclude that for $N \to \infty$ we can write the following continuous dynamics for the mean magnetisation:

$$\frac{d}{dt}\langle m(t)\rangle = -\langle m(t)\rangle + \tanh\langle m(t)\rangle$$

(d) Conclude that the stationary expected magnetisation satisfies the Curie-Weiss mean-field equation. Generalise to arbitrary β and h.

(e) We can now repeat the experiment of the previous exercise, but using the theoretical ordinary differential equation: start by a configuration where all spins are equal to 1 and take different values of h and β. For which values will the Monte-Carlo chain reach the equilibrium value? When will it be trapped in a spurious maximum of the free entropy $\phi(m)$? Compare your theoretical prediction with numerical simulations.

Exercise 1.6: Potts model

The Potts model is a variant of the Ising model where there could be more that two states: here the Potts spins can take up to q values. It is one of the most fundamental model of statistical physics. In its fully connected version, the Hamiltonian reads:

$$\mathcal{H} = \sum_{i,j} \frac{1}{N}(1-\delta_{\sigma_i,\sigma_j})$$

with $\sigma_i = 1 \dots q$, for all $i = 1 \dots N$.

(a) Using the variational approach of Section 1.3, write the Gibbs free-energy of the Potts model, as a function of the fractions $\{\rho_\tau\}_{i=1,\dots N,\tau=1\dots q}$ of spins in state τ.

(b) Write the mean-field self-consistent equation governing the $\{\rho_\tau\}$ by extremizing the Gibbs free-energy and solve it numerically.

(c) Using the cavity approach of section 1.4, show that one can recover the mean-field self-consistent equation using this method.

Appendix

1.A The jargon of Statistical Physics

In these notes we will often adopt a terminology from Statistical Physics. While this is standard for someone who already took a course on the subject, it can often be confusing for newcomers from other fields.

As we have discussed in the introduction, a *system* is defined by a set of *degrees of freedom* (d.o.f.) $\{x_i\}_{i\in\mathcal{I}}$ and a *Hamiltonian* function $\mathcal{H}\left(\{x_i\}_{i\in\mathcal{I}}\right)$ assigning an energy or cost to each *configuration* of d.o.f. For example, in the Curie-Weiss model the d.o.f. are *spins* $x_i \in \{-1, 1\}$ indexed by $i = 1, \cdots, N$, and the Hamiltonian is given by equation 1.1. Note however that the *index set* $\mathcal{I}$ and the *state space* $\mathcal{X}$ can be more general. For instance, in Chapter 6 we will discuss the Graph Colouring problem, an example of a system where the d.o.f. are defined in the nodes of a graph G, and they can take q different values, such that $\mathcal{X} = \{1, \cdots, q\}$. Instead, in Chapter 16 we will study inference problems in which the state space can be uncountable, such as the weights of a single-layer neural network. Note that we can even study systems indexed by an uncountable set — e.g. $\mathcal{I} \subset \mathbb{R}$ — in which case it is common to refer to the d.o.f. as *fields*. However, in these notes we will be only dealing with discrete and finite (therefore countable) indices. Therefore, it will be convenient to adopt a vector notation for a configuration $\boldsymbol{x} \in \mathcal{X}^N$, where $N \equiv |\mathcal{I}|$ is refereed to as the *system size*. Quantities that scale with the system size N are often refereed to as *extensive* (in mathematical notation we write $O_N(N)$), while quantities that do not scale with the system size are often called *intensive* (and we write $O_N(1)$). Typical extensive quantities are the energy, entropy and volume, while typical intensive quantities are the temperature and pressure. Note that we can always produce an intensive quantity by considering the density of an extensive quantity (dividing it by N). The limit of infinite system size $N \to \infty$ is called the *thermodynamic limit*, and a major part of Statistical Mechanics is devoted to studying the behavior of intensive quantities in the thermodynamic limit.

In Classical Mechanics, the goal is to study the *microscopic* properties of a system, for instance the trajectory of each molecule of a gas or the dynamics of each neuron of a neural network during training. Instead, in Statistical Mechanics the goal is to study *macroscopic* properties of the system, which are collective properties of the d.o.f. In our previous examples, a macroscopic property of the gas is its mean energy, its temperature or its pressure, while a macroscopic property of the neural network is the generalization error. To make these notions more precise, in Statistical Physics we define an *ensemble* over the *configurations*, which is simply a probability measure over the space of all possible configurations. Different ensembles can be defined for the same system, but in these notes we will mostly focus on the *canonical ensemble*, which is

defined by the *Boltzmann-Gibbs distribution*:

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z_N(\beta)} e^{-\beta \mathcal{H}(\boldsymbol{x})} \tag{1.25}$$

where the normalization constant $Z_N(\beta)$ is known as the *partition function*. Note that the partition function is closely related to the moment generating function for the energy. From this probabilistic perspective, a configuration $\boldsymbol{x} \in \mathcal{X}^N$ is simply a random sample from the Boltzmann-Gibbs distribution, and a macroscopic quantity can be seen as a statistic from the ensemble. Physicists often denote the average with respect to the Boltzmann-Gibbs distribution with brackets $\langle \cdot \rangle_\beta$ and refer to it as a *thermal average*. We now define the most important quantity in these notes, the *free energy density*:

$$-\beta f_N(\beta) = \frac{1}{N} \log Z_N(\beta). \tag{1.26}$$

Note that since the Hamiltonian is typically extensive $\mathcal{H} = O(N)$, the free entropy (the logarithm of the partition sum) is also extensive, and therefore its density is an intensive quantity. It is closely related to the cumulant generating function for the energy. For this reason, and as discussed in the introduction, from the free entropy we can access many of the important macroscopic properties of our system. Two macroscopic quantities physicists are often interested in are the *energy* and *entropy* densities:

$$e_N(\beta) = \left\langle \frac{1}{N} \mathcal{H}(\boldsymbol{x}) \right\rangle_\beta = \partial_\beta (\beta f_N(\beta)), \qquad s_N(\beta) = \beta^2 \partial_\beta f_N(\beta) \tag{1.27}$$

which satisfy the useful relation:

$$f_N(\beta) = e_N(\beta) - \frac{1}{\beta} s_N(\beta) \tag{1.28}$$

1.B The ABC of phase transitions

One of the main goals of the Statistical Physicist is to characterize the different *phases* of a system. A phase can be loosely defined as a region of parameters defining the model that share common macroscopic properties. For example, the Curie-Weiss model studied in Chapter 1 is defined by the parameters $(\beta, h) \in \mathbb{R}_+ \times \mathbb{R}$, and we have identified two phases in the thermodynamic limit: a paramagnetic phase characterised by no net system magnetization and a ferromagnetic phase characterized by net system magnetization. The macroscopic quantities characterising the phase of the system (in this case the net magnetization) are known as the *order parameters* for the system. In Chapter 16 we will study a model for Compressive Sensing, which is the problem of reconstructing a compressed sparse signal corrupted by noise. The parameters of this system are the sparsity level (density of non-zero elements) $\rho \in [0, 1]$ and the noise level $\Delta \in \mathbb{R}_+$, and we will identify the existence of three phases: one in which reconstruction is easy, one in which it is hard and a one in which it is impossible. The order parameter in this case will be the correlation between the estimator and the signal. Although all examples studied here have clear order parameters, it is not always easy to identify one, and it might not always be unique. For instance, in the Compressive sensing example we could also have chosen the mean squared error as an order parameter.

When a system changes phase by varying a parameter (say the temperature), we say the system undergoes a *phase transition*, and we refer to the boundary (in parameter space) separating the two phases as the *phase boundary*. In Physics, we typically summarize the information about the phases of a system with a *phase diagram*, which is just a plot of the phase boundaries in parameter space. See Figure 1.B.1 for two examples of well-known phase diagrams in Physics.

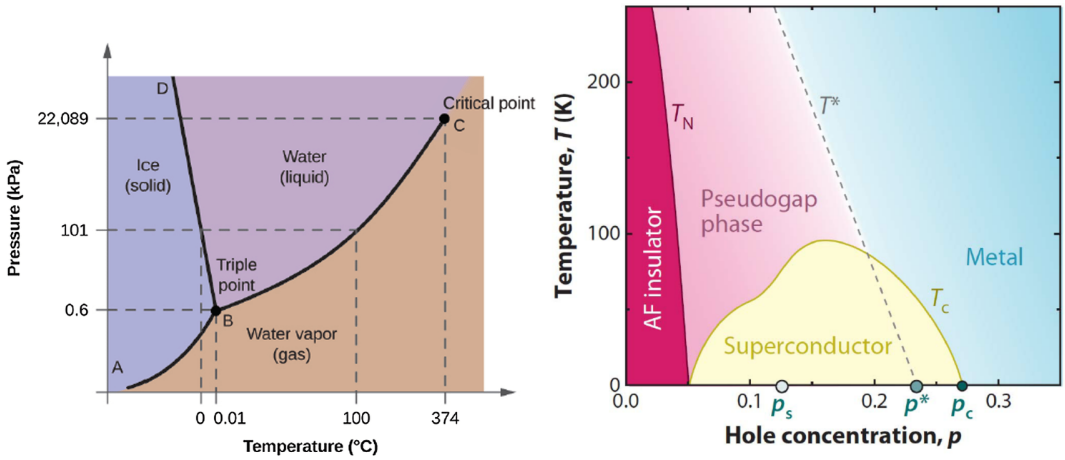


Figure 1.B.1: Phase diagrams of water (left) and of the cuprate (right), from Taillefer (2010) and Schwarz et al. (2020) respectively.

Phase transitions manifest physically by a macroscopic change in the behavior of the system (think about what happens to the water when it starts boiling). Therefore, the reader who followed our discussion from Chapter 1 and Appendix 1.A should not be surprised by the fact that phase transitions can be characterised and classified from the free energy. Indeed, the classification of phase transitions in terms of the analytical properties of the free energy dates back to the work of Paul Ehrenfest in 1933 (see Jaeger (1998) for a historical account). At this point, the mathematically inclined reader might object: the free energy from equation 1.26 is an analytic function of the model parameters, so how can it change behavior across phases? Indeed, for finite N the free energy is an analytic function of the model parameters. However, the limit of a sequence of analytic functions need not be analytic, and in the thermodynamic limit the free energy can develop singularities. Studying the singular behavior of the limiting free energy is the key to Ehrenfest's classification of phase transitions. The two most common types of phase transitions are:

First order phase transition: A first order phase transition is characterised by the discontinuity in the first derivative of the limiting free energy with respect to a model parameter. The most common example is the transition of water from a liquid to a gas as we change the temperature at fixed pressure (what you do when you cook pasta), see Fig. 1.B.1 (left). In this example, the derivative of the free energy with respect to the temperature, also known as the *entropy*, discontinuously jumps across the phase boundary.

Second order phase transition: A second order phase transition is characterised by a discontinuity in the second derivative of the limiting free energy with respect to a model parameter. Therefore, in a second order transition the free energy itself and its first derivative are continuous. Note that second order derivatives of the free energy are typically associated with response functions such as the susceptibility. Perhaps the most

famous example of a second order transition is the spontaneous magnetization of certain metals as a function of temperature, known as ferromagnetic transition.

Although less commonly used, we can define an n-th order phase transition in terms of the discontinuity of the n-th derivative of the limiting free energy. The order of a phase transition is associated to a rich phenomenology, which we now discuss, for the sake of concreteness, in our favourite model: the Curie-Weiss model for ferromagnetism.

Phase transitions in the Curie-Weiss model Recall that in Chapter 1 we have computed the thermodynamic limit of the free energy density for the Curie-Weiss model:

$$-\beta f_\beta = \lim_{N\to\infty} \frac{1}{N} \log Z_N = \max_{m\in[-1,1]} \phi(m)$$

where:

$$\phi(m) = \frac{\beta}{2} m^2 + \beta h m + H(m)$$
$$H(m) = -\frac{1+m}{2} \log \frac{1+m}{2} - \frac{1-m}{2} \log \frac{1-m}{2}.$$

As we have shown, the parameter $m^\star$ solving the minimization problem above gives the order parameter of the system, the net magnetization at equilibrium:

$$m^\star = \operatorname*{argmax}_{m\in[-1,1]} \phi(m) = -\partial_h f(\beta, h) = \langle \bar{S} \rangle_\beta \tag{1.29}$$

In particular, the limiting average energy and entropy densities are given by:

$$e(\beta, h) = \partial_\beta(\beta f(\beta, h)) = -m^\star \left(\frac{m^\star}{2} + h \right) \qquad s(\beta, h) = \beta^2 \partial_\beta f(\beta, h) = H(m^\star) \tag{1.30}$$

In particular, note that the entropy density depends on the model parameters (β, h) only indirectly through the magnetization $m^\star = m^\star(\beta, h)$. The potential $\phi(m)$ is an analytic function of m and the parameters (β, h). However, due to the optimization over m, the free energy density can develop a non-analytic behavior as a function of (β, h), signaling the presence of phase transitions, which we now recap.

Zero external field and the second order transition: Note that the decomposition $f = e - \beta s$ (see equation 1.27) makes it explicitly that the free energy is a competition between two parabolas: the energy (convex) and the entropy (concave). At zero external field $h = 0$, we note that the potential is a symmetric function of the magnetization $\phi(-m) = \phi(m)$. At high temperatures $\beta \to 0^+$, the dominant term is given by the entropy, which has a single global minimum at $m^\star = 0$ (see Fig. 1.1): this is the *paramagnetic phase* in which the system has no net magnetization. At the critical temperature $\beta_c = 1$, $m = 0$ becomes a maximum of the system, with two global minima (having the same free energy) continuously appearing, see Fig. 1.4. This signals a phase transition towards a *ferromagnetic phase* defined by a net system magnetization $|m^\star| > 0$. Note that the first derivative of the free energy with respect to β (proportional to the entropy) remains a continuous function across the transition. However, we notice that the second derivative

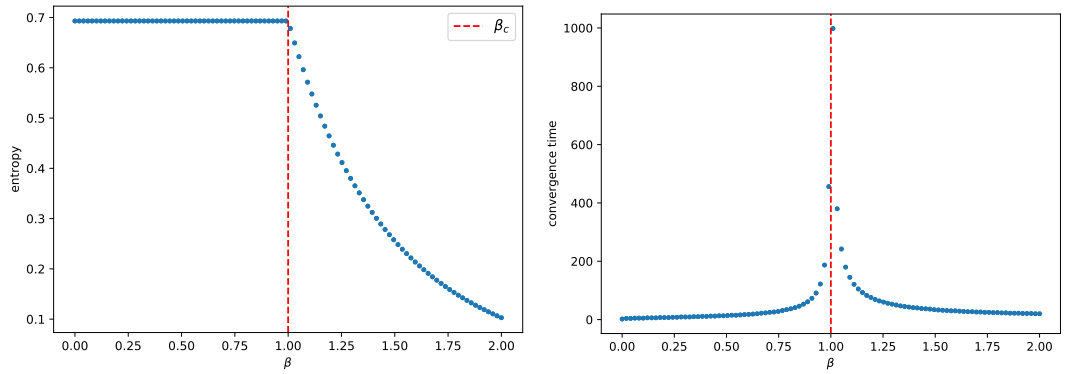


Figure 1.B.2: (Left) Entropy as a function of the inverse temperature β at zero external field $h = 0$. Note that the entropy is a continuous function of the temperature, with a cusp at the critical point $\beta_c = 1$, indicating that its derivative (proportional to the second derivative of the free energy) has a discontinuity. (Right) Convergence time of the saddle-point equation as a function of the inverse temperature β at zero external field $h = 0$. Note the critical slowing down close to the second order critical point $\beta_c = 1$.

of the free energy is discontinuous, indicating this is a *second order phase transition*. This transition corresponds to a significant change in the statistical behavior of the system at macroscopic scales: while for $\beta < 1$ a typical configuration from the Boltzmann-Gibbs distribution has no net magnetization $m^\star = \langle \bar{S} \rangle_\beta \approx 0$ (*disordered phase*), for $\beta > 1$ a typical configuration has a net magnetization $|m^\star| = |\langle \bar{S} \rangle_\beta| > 0$ (*ordered phase*). This is an example of an important concept in Physics known as *spontaneous symmetry breaking*: while the Hamiltonian of the system is invariant under the $\mathbb{Z}_2$ symmetry $\bar{S} \to -\bar{S}$, for $\beta > 1$ a typical draw of the Gibbs-Boltzmann distribution $\boldsymbol{S} \sim \mathbb{P}_{N,\beta}$ breaks this symmetry at the macroscopic level. Second order transitions carry a rich phenomenology. Since the transition is second order (i.e. continuous first derivative), the critical temperature can be obtained by studying the expansion of the free energy potential around $m = 0$:

$$\phi(m) \underset{m \to 0}{=} \log 2 + \frac{m^2}{2}(\beta - 1) + O(m^3)$$

which give us the critical $\beta_c = 1$ as the point in which the second derivative changes sign ($m = 0$ goes from a minimum to a maximum). It is also useful to have the picture in terms of the saddle-point equation:

$$m = \tanh(\beta m).$$

The fact that $m = 0$ is always a fixed point of this equation signals it is always an extremizer of the free energy potential. From this perspective, the critical temperature $\beta_c = 1$ corresponds to a change of stability of this fixed point. Seeing the saddle-point equations as a discrete dynamical system $m^{t+1} = f(m^t)$, the stability of a fixed point can be determined by looking at the Jacobian of the update function $f : [-1, 1] \to [-1, 1]$ around the fixed point $m = 0$:

$$f(x) = \tanh(\beta x) \underset{m \to 0}{=} \beta x + O(m^3) \tag{1.31}$$

For $\beta < 1$, the fixed point is *stable* (attractor/sink of the dynamics), while for $\beta > 1$ it becomes an *unstable* (repeller/source of the dynamics). Note that this implies that

close to the transition $\beta \approx 1^+$, iterating the saddle point equations starting close to zero $m^{t=0} = \epsilon \ll 1$ (but not exactly at zero) takes long to converge to a non-zero magnetization $m > 0$, with the time diverging as we get closer to the transition. This phenomenon is known is Physics as the *critical slowing down*, and together with the expansion of the free energy and the stability analysis of the equations give yet another way to characterise a second order critical point. See Figure 1.B.2 (right) for an illustration.

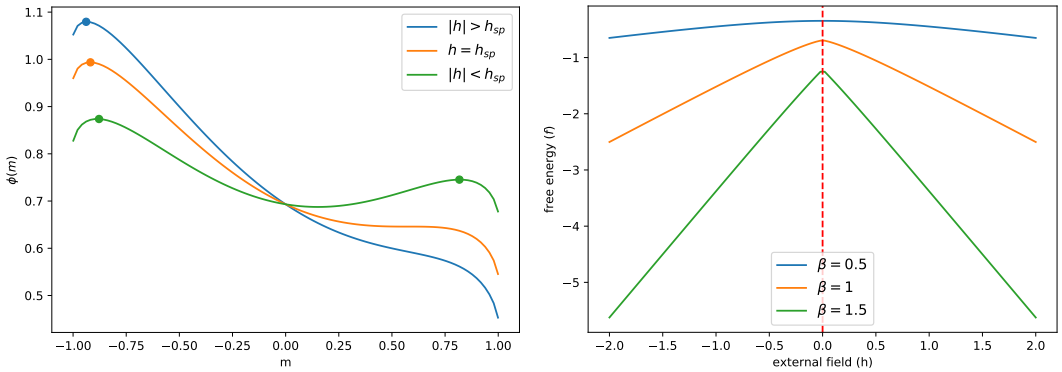


Figure 1.B.3: (Left) Free energy potential $\phi(m)$ as a function of m for fixed inverse temperature $\beta = 1.5$ and varying external field $h < 0$. Note that the free energy potential has a local maximum for $|h| > h_{sp}$ that disappears at the spinodal transition $h = h_{sp}$. (Right) Free energy as a function of the external field h at different temperatures. Note the non-analytical cusp at $h = 0$.

Finite external field and the first order transition: Turning on the external magnetic field $h \neq 0$ can dramatically change the discussion above. First, note that the Hamiltonian loses the $\mathbb{Z}_2$ symmetry: this is known in Physics as *explicit symmetry breaking*. At high temperatures $\beta \to 0^+$, the free energy potential is convex, with a single minimum at $m = h$ aligned with the field. As temperature is lowered and we enter what previously was the ferromagnetic phase ($\beta > 1$), two behaviors are possible. For small h, the field simply has the effect of breaking the symmetry between the previous two global minima and making the with opposite sign a local minimum, see Fig. 1.B.3 (left). In this situation, even though the equilibrium free energy is given by the now unique global minimum of the potential, the presence of a local minimum has an important effect in the dynamics. Indeed, if we initialize the saddle-point equations close to the magnetization corresponding to the local minimum, it will converge to this local minimum, since it is also a stable fixed point of the corresponding dynamical system, see Fig. 1.B.4 (left). This phenomenon is known as *metastability* in Physics. Note that metastability can be a misleading name, since in the thermodynamic limit $N \to \infty$ metastable states are stable fixed points of the free energy potential. However, at finite system size N, the system will dynamically reach equilibrium in a time of order $t = O(e^N)$. Metastability will play a major role in the Statistical Physics analysis of inference problems, since it is closely related to algorithmic hardness.

As the external field h is increased, the difference in the free energy potential between the two minima increases, and eventually at a critical field $h_{\rm sp}$, known as the *spinodal point*, the local minimum disappears, making the potential convex again, see Fig. 1.B.3 (left). The spinodal points can be derived from the expression of the free energy potential, and

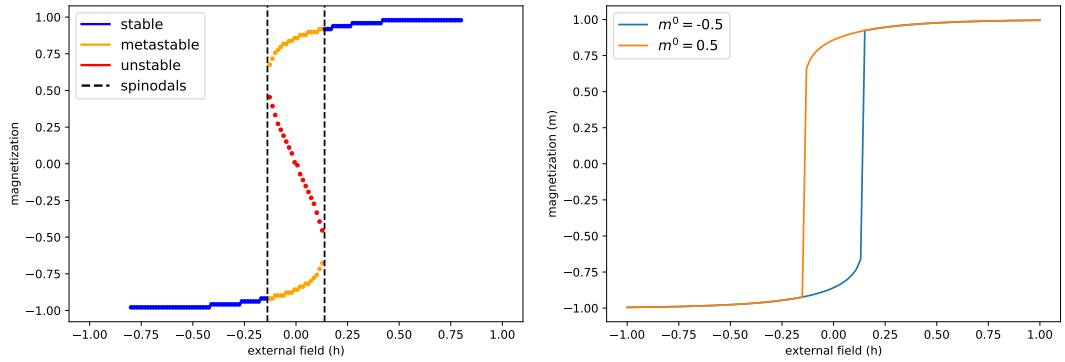


Figure 1.B.4: (Left) Stable, metastable and unstable branches of the magnetization as a function of the external field at fixed inverse temperature $\beta = 1.5$. (Right) Magnetization obtained by iterating the saddle-point equations from different initial conditions $m^{t=0}$ as a function of the external field h and fixed inverse temperature $\beta = 1.5$. Note the hysteresis loop: point at which the magnetization discontinuously jumps from negative to positive depends on the initial state of the system.

is given by:

$$h_{\rm sp}(\beta) = \pm\sqrt{\frac{1}{\beta}\left(1-\frac{1}{\beta}\right)} \mp \frac{1}{\beta}\tanh^{-1}\left(\sqrt{1-\frac{1}{\beta}}\right), \qquad \beta > 1$$

From this discussion, it is clear that for $\beta > 1$ the magnetization (which is the derivative of the free entropy with respect to h) has a discontinuity at $h = 0$, since for $h \neq 0$ we have a non-zero magnetization and for $h = 0$ we are in the paramagnetic phase $h = 0$. This is a first order phase transition of the system with respect to the external field h, see Fig. 1.B.3 (right). Note that as a consequence of metastability, in the region $|h| < |h_{\rm sp}|$ the system magnetization will depend of the state in which it was initially prepared. This memory of the initial state is known as *hysteresis* in Physics, see Fig. 1.B.4 (right).

1.C A rigorous version of the cavity method

The cavity method presented in the main chapter can also be done entirely rigorously, as we now show. This appendix can be skipped for non mathematically-minded readers, although it is always interesting to see how things can be done precisely. In fact, this section is a good training for the later Chapters where the rigorous proofs are more involved, despite using the very same techniques.

First, we want to show that the magnetization, as well as many other observables, does indeed converge to a fixed value as N increases. This is done through the following lemma that tells us that, if indeed $\bar{S}$ concentrates, then any observable will concentrate as well.

Lemma 5. *For any bounded observable $O(\{S\}_N)$ there exists a constant $C = \|O(.)\|_\infty$ such that:*

$$|\langle O(\boldsymbol{S})\rangle_{N+1,\beta'} - \langle O(\boldsymbol{S})\rangle_{N,\beta}| \leq C\beta\sinh(\beta h+\beta)\sqrt{\mathrm{Var}_{N,\beta}(\bar{S})} \tag{1.32}$$

Proof. The proof proceeds as follow. First we compute directly

$$\langle O(\boldsymbol{S})\rangle_{N+1,\beta'} = \frac{\langle O(\boldsymbol{S})\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta}}{\langle\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta}}$$

From this, it follows that

$$\begin{aligned}
&\langle O(\boldsymbol{S})\rangle_{N+1,\beta'} - \langle O(\boldsymbol{S})\rangle_{N,\beta} \\
&= \frac{\langle O(\boldsymbol{S})\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta} - \langle O(\boldsymbol{S})\rangle_{N,\beta}\langle\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta}}{\langle\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta}} \\
&\leq \langle O(\boldsymbol{S})(\cosh\left(\beta(\bar{S}+\beta h)\right)\rangle_{N,\beta} - \langle O(\boldsymbol{S})\rangle_{N,\beta}\langle\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta} \\
&= \mathrm{COV}_{N,\beta}\left[O(\boldsymbol{S}),\cosh\left(\beta(\bar{S}+\beta h)\right)\right]
\end{aligned}$$

where we have used the fact that $\cosh(x) \geq 1$. When one sees a covariance, one should always try to use the Cauchy-Schwartz formula that states hat $\mathrm{Cov}^2(X,Y) < \mathrm{Var}(X)\mathrm{Var}(Y)$. In our case, this leads to

$$(\langle O(\boldsymbol{S})\rangle_{N+1,\beta'} - \langle O(\boldsymbol{S})\rangle_{N,\beta})^2 \leq \mathrm{Var}(O(\boldsymbol{S}))\mathrm{Var}(\cosh\left(\beta(\bar{S}+h)\right)) \tag{1.33}$$

Finally we note that $\cosh(\beta(\bar{S}+h))$ is a convex Lipschitz function for the variable $\bar{S}$ with constant $\beta\sinh(\beta(1+h))$ (since $\bar{S}$ is bounded by one), and therefore by applying Jensen's inequality, and then the Lipschitz inequality, we have

$$\begin{aligned}
\cosh\left(\beta(\bar{S}+h)\right) - \langle\cosh\left(\beta(\bar{S}+h)\right)\rangle &\leq \cosh\left(\beta(\bar{S}+h)\right) - \cosh\left(\beta(\langle\bar{S}\rangle+h)\right) \\
&\leq \beta\sinh(\beta(1+h))((\bar{S}-\langle\bar{S}\rangle))
\end{aligned}$$

so that

$$\langle(\cosh\left(\beta(\bar{S}+h)\right) - \langle\cosh\left(\beta(\bar{S}+h)\right)\rangle)\rangle^2 \leq \beta^2\sinh(\beta(1+h))^2\mathrm{Var}(\langle\bar{S}\rangle)$$

Plugging this relation in (1.33) finishes the proof. □

We can now prove the main thesis and obtain the mean field equation:

Theorem 5. *There exists a constant $C(\beta,h)$ such that*

$$|\langle S_i\rangle_{N,\beta} - \tanh\beta(h+\langle\bar{S}\rangle_{N,\beta})| \leq C(\beta,h)\sqrt{\mathrm{Var}(\bar{S})}$$

Proof. By direct computation, we get that

$$\langle S_i\rangle_{N+1,\beta'} = \langle S_0\rangle_{N+1,\beta'} = \frac{\langle\sinh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta}}{\langle\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta}}$$

Using again Jensen and the Lipschitz property of the $\cosh$ and $\sinh$, we get concentration of both terms in the fraction as

$$\begin{aligned}
|\langle\sinh\left(\beta(\bar{S}+h)\right)\rangle - \sinh\left(\beta(\langle\bar{S}\rangle+h)\right)| &\leq \beta\cosh\left(\beta(1+h)\right)\sqrt{\mathrm{Var}(\bar{S})} \\
|\langle\cosh\left(\beta(\bar{S}+h)\right)\rangle - \cosh\left(\beta(\langle\bar{S}\rangle+h)\right)| &\leq \beta\sinh\left(\beta(1+h)\right)\sqrt{\mathrm{Var}(\bar{S})}
\end{aligned}$$

Together with the inequality $|a_1/b_1 - a_2/b_2| \leq (a_1-a_2)/b_1 + a_2|b_1-b_2|/b_1b_2$, we get using $a_i \geq 0$ and $b_i \geq \max(1,a_i)$ that

$$|\langle S_i\rangle_{N+1,\beta'} - \tanh\left(\beta(h+\langle\bar{S}\rangle_{N,\beta})\right)| \leq C(\beta,h)\sqrt{\mathrm{Var}(\bar{S})}$$

Which, using lemma 5, finishes the proof. □

All is left to do to get the mean field equation is showing that the variance of the magnetization is going to zero outside of the phase transition/coexistence line. This is not entirely easy to do, but we can easily show that this true almost everywhere in the plane (β, h) using the so-called fluctuation-dissipation approach:

Lemma 6 (Bound on the variance)**.** *For any β, and values h_1, h_2 of the magnetic field, one has*

$$\int_{h_1}^{h_2} \mathrm{Var}(\bar{S})_{\beta,N,h} dh \leq \frac{2}{\beta N}$$

so that $\mathrm{Var}(\bar{S})_{\beta,N,h} \leq \frac{2}{N}$ for almost every h.

Proof. The proof start by noticing that $\langle \bar{S} \rangle_{\beta,N} = \frac{\partial}{\partial \beta h} \Phi_N(\beta, h)$ and by direct computation

$$N\mathrm{Var}(\bar{S})_{\beta,N,h} = \frac{\partial^2}{\partial(\beta h)^2} \Phi_N(\beta, h) = \frac{\partial}{\partial(\beta h)} \langle \bar{S} \rangle_{\beta,N}$$

Therefore

$$N \int_{\beta h_1}^{\beta h_2} \mathrm{Var}(\bar{S})_{\beta,N,h} d\beta h = \langle \bar{S} \rangle_{\beta,N,h_2} - \langle \bar{S} \rangle_{\beta,N,h_1} \leq 2$$

□

We can also prove the free entropy, using a technique that shall be very useful for more complex problems:

Theorem 6 (Free entropy by the cavity method)**.**

$$\Phi(\beta, h) = \max_m \tilde{\phi}(m; \beta, h)$$

Proof. Writing the seemingly trivial equality:

$$\begin{aligned} \phi_N(\beta, h) &= \frac{1}{N} \log Z_N = \frac{1}{N} \log \frac{Z_N}{Z_{N-1}} Z_{N-1} = \frac{1}{N} \log \frac{Z_N}{Z_{N-1}} \frac{Z_{N-1}}{Z_{N-2}} \cdots \frac{Z_1}{1} \\ &= \frac{1}{N} \sum_{n=0}^{N-1} A_n(\beta, h); \qquad A_N(\beta, h) =: \log \frac{Z_{N+1}}{Z_N} \end{aligned}$$

we have, by the Stolz–Cesàro theorem, that if A_n converges to A, then $\Phi_N(\beta, h)$ also converges to A. Using equation 1.21 we thus write

$$A_N(\beta, h) =: \log Z_{N+1}/Z_N = \log \langle e^{-\frac{\beta}{2}\bar{S}^2} 2 \cosh\left(\beta(\bar{S} + h)\right) \rangle_{N,\beta} + o(1)\,.$$

By lemma 5 and 6 we thus obtain for any $\beta > 0$ and almost everywhere in h, that

$$\lim_{N \to \infty} A_N(\beta, h) = -\beta \frac{m^{*2}}{2} + \log 2 \cosh(\beta(m^* + h))$$

with $m^* = \langle \bar{S} \rangle$. Therefore almost everywhere in field we have, following eq.(1.21) that the free entropy is given by one of the extremum m^* of eq.(1.21):

$$\Phi(\beta, h) = \tilde{\phi}(\langle \bar{S} \rangle)\,.$$

It just remains to show that the correct extremum, if they are many of them, is the maximum one. This can be done by noting that it is necessary the maximum since $\tilde{\phi}(m^*) = \phi(m^*)$, and that by the Gibbs variational approach, theorem 4, we already proved that $\Phi \geq \phi(m) \forall m$. Given the entropy is Lipschitz continuous in h for all N (its derivative is the magnetization, which is bounded), its limit is continuous, so that the if the free entropy is given by the maximum of $\phi(m)$ almost everywhere, it is true everywhere. □

Chapter 2

A simple example: The Random Field Ising Model

> Let's start at the very beginning. A very good place to start.
>
> *The sound of music*
> 1965

Let us move to a more challenging example. We continue to consider a system of N spins $s_i \in \{\pm 1\}$ with Hamiltonian

$$\mathcal{H}_N(\boldsymbol{s}, \boldsymbol{h}) = -\frac{N}{2}\left(\sum_i \frac{s_i}{N}\right)^2 - \sum_i h_i s_i,$$

but now, the additional fields $\boldsymbol{h}$ are fixed, once and for all. We choose $h_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Delta)$.

This is a simple variation on the Ising model, but now we have these new random fields $\boldsymbol{h}$. This makes the problem a bit more complicated. For this reason, it is called the Random Field Ising Model, or RFIM. We may ask many questions. For instance: what is the assignment of the spins that minimizes the energy? This is a non-trivial question since there is a competition between aligning all variables together in the same direction, and aligning them to the direction of the local random fields. What will be the energy of this assignment? Will the energy be very different when one picks up another random value for $\boldsymbol{h}$? How do we find such an assignment in practice? With which algorithm?

2.1 Self-averaging and Concentration

We shall be interested in the behavior of this system in the large N limit. How are we going to deal with the random variables? The entire field of statistical mechanics of disordered systems, and its application to optimization and statistics, is based on the idea of self-averaging: that is, the idea that the particular realization of $\boldsymbol{h}$ does not matter in the large size limit $N \to \infty$ (the

asymptotic limit for mathematicians, or the "thermodynamic" limit for physicists). This is a powerful idea which allows us to average over all $\boldsymbol{h}$.

Let us see how this can be proven rigorously. First, a word of caution, the partition sum Z_N is exponentially large in N, so we expect its fluctuation around its mean to be very large as well. Fortunately, as we have seen, our focus will be on $\log Z_N(\boldsymbol{h})/N$, which we hope converges to a constant $O(1)$ value as $N \to \infty$. The quantity $\log Z_N(\boldsymbol{h})/N$ is random, it depends on the value of $\boldsymbol{h}$, but we may expect that the typical value of $\log Z_N(\boldsymbol{h})/N$ is close to its mean. This notion, that a large enough system is close to its mean, is called "self-averaging" in statistical physics. In probability theory, this is a concentration of measure phenomenon.

Indeed for the Random Field Ising model, we can show that the free entropy concentrates around its mean value for large N:

Theorem 7 (Self-averaging)**.** *Let $\Phi_N(\boldsymbol{h},\beta) =: \log Z_N(\beta,\boldsymbol{h})/N$ be the free entropy density of the RFIM, then:*

$$\mathrm{Var}[\Phi_N(\boldsymbol{h},\beta)] \leq \frac{\Delta\beta^2}{N}$$

Proof. The proof is a consequence of the very useful Gaussian Poincaré inequality (see the theorem in exercise 9): Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a smooth function and X has multivariate Gaussian distribution $X \sim \mathcal{N}(0,\Gamma)$, where $\Gamma \in \mathbb{R}^{n\times n}$. Then

$$\mathrm{Var}[f(X)] \leq \mathbb{E}\left[\Gamma\nabla f(X)\cdot\nabla f(X)\right].$$

For the RFIM, given that

$$\partial_{h_i}\Phi_N(\boldsymbol{h},\beta) = \frac{\beta}{N}\langle S_i\rangle$$

we find that

$$\nabla\Phi_N(\boldsymbol{h},\beta)\cdot\nabla\Phi_N(\boldsymbol{h},\beta) = \frac{\beta^2}{N}\sum_i \frac{\langle S_i\rangle^2}{N} \leq \frac{\beta^2}{N},$$

and the Gaussian Poincaré inequality with $\Gamma = \Delta I$ (covariance matrix of $\boldsymbol{h}$) yields the final result. □

Hence, instead of computing the free entropy density for each realization of $\boldsymbol{h}$, we turn to computing the expectation over all possible $\boldsymbol{h}$, i.e. we define

$$\Phi(\beta,\Delta) \triangleq \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{h}}\Phi_N(\beta,\delta,\boldsymbol{h}) = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\boldsymbol{h}}\left[\log Z_N(\beta,\boldsymbol{h})\right]$$

and the self-averaging property guarantees that $\mathbb{E}_{\boldsymbol{h}}\left[\log Z_N(\beta,\boldsymbol{h})\right]/N$ is close to $\left[\log Z_N(\beta,\boldsymbol{h})\right]/N$ when N is large. In probability theory, when the variance goes to zero one says that the random variable converges in probability. Here, we have thus showed that $\Phi_N(\beta,\boldsymbol{h},\Delta)$ converges in probability to $\Phi(\beta,\Delta)$. In fact, we could work a bit more and show that the probability to have a deviation larger than the variance is exponentially low, so we are definitely safe, even at moderate values of N. This leaves us with the question of how to compute $\Phi(\beta,\Delta)$, as it involves the expectation of a logarithm.

2.2 Replica Method

In order to compute the average of the logarithm, a powerful heuristic method has been used widely in statistical physics: the replica method, proposed in the 70's by Sir Sam Edwards (who credits it to Marc Kac). Here is the argument: Suppose n is close to zero, then

$$Z^n = e^{n \log Z} = 1 + n \log(Z) + \mathrm{o}(n) \qquad \Longrightarrow \qquad \log Z = \lim_{n \to 0} \frac{Z^n - 1}{n}$$

If Z is a random variable and we suppose that swapping limit and expectation is valid (which is by no means evident), then we find the following identity, often referred to as "the replica trick:"

$$\mathbb{E}\left[\log Z\right] = \mathbb{E}\left[\lim_{n \to 0} \frac{Z^n - 1}{n}\right] = \lim_{n \to 0} \frac{\mathbb{E}[Z^n] - 1}{n}$$

This is at the root of the replica method: we replace the average of the logarithm of Z, which is hard to compute, by the average of powers of Z. If n is integer, we may hope that we shall indeed be able to compute averages of Z^n. We could then "pretend" that our computation with n finite is valid for $n \in \mathbb{R}$ (which is again not a trivial step), and perform an analytic continuation from $\mathbb{N}$ to $\in \mathbb{R}$, and send $n \to 0$. While this sounds acrobatic and certainly not like rigorous mathematics, it seems, amazingly, to work quite well when we stick to the guidelines that physicists —following the trail of Giorgio Parisi and Marc Mézard— have proposed over the last few decades. Indeed, when it can be applied, this method appears to seemingly always lead to the correct result, at least when we can compare it to rigorous computation. Nowadays, there is a deep level of trust in the results given by the replica method.

We shall come back on what we can say with rigorous mathematics later on, but first, let us see how it works in detail for the computation of $\Phi(\beta, \Delta)$ in the Random Field Ising Model, using our "field-theoretic" toolbox of the previous chapter.

2.2.1 Computing the replicated partition sum

Let n be the number of replicas and $\alpha = 1, \ldots, n$ be the index of replicas (these are traditional notation in the replica literature) we have

$$Z^n = \sum_{\boldsymbol{s}^{(1)}} \sum_{\boldsymbol{s}^{(2)}} \cdots \sum_{\boldsymbol{s}^{(n)}} \exp\left\{ \sum_{\alpha=1}^{n} \beta \left[\frac{N}{2} \left(\sum_i \frac{s_i^{(\alpha)}}{N} \right)^2 + \sum_i h_i s_i^{(\alpha)} \right] \right\} .$$

We now write the average over the random fields, and proceed to fix the magnetization, as we did for the Curie-Weiss model, this time for each of the n "replicas" indexed from $\alpha = 1, ..., n$:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[Z^n\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\sum_{\left\{\boldsymbol{s}^{(\alpha)}\right\}_{\alpha=1}^{n}} e^{\beta\frac{N}{2}\sum_{\alpha=1}^{n}\left(\sum_i \frac{s_i^{(\alpha)}}{N}\right)^2+\beta\sum_{\alpha=1}^{n}\sum_i h_i s_i^{(\alpha)}}\right] \\
&\overset{(a)}{=} N^n\mathbb{E}_{\boldsymbol{h}}\left[\sum_{\left\{\boldsymbol{s}^{(a)}\right\}_\alpha}\int\prod_\alpha \mathrm{d}m_\alpha\,\delta\left(\sum_i s_i^{(\alpha)} - Nm_\alpha\right)e^{\beta\frac{N}{2}\sum_\alpha m_\alpha^2+\beta\sum_\alpha\sum_i h_i s_i^{(\alpha)}}\right] \\
&\overset{(b)}{=} (2\mathbb{i}\pi N)^n\mathbb{E}_{\boldsymbol{h}}\left[\sum_{\left\{\boldsymbol{s}^{(a)}\right\}_\alpha}\int\prod_\alpha \mathrm{d}m_\alpha\,\mathrm{d}\hat{m}_\alpha\, e^{\sum_\alpha \hat{m}_\alpha\left[\sum_i s_i^{(\alpha)}-Nm_\alpha\right]}e^{\beta\frac{N}{2}\sum_\alpha m_\alpha^2+\beta\sum_\alpha\sum_i h_i s_i^{(\alpha)}}\right] \\
&\propto \int\prod_\alpha \mathrm{d}m_\alpha\,\mathrm{d}\hat{m}_\alpha\, e^{\left(\beta\frac{N}{2}\sum_\alpha m_\alpha^2 - N\sum_\alpha \hat{m}_\alpha m_\alpha\right)}\mathbb{E}_{\boldsymbol{h}}\left[\sum_{\left\{\boldsymbol{s}^{(a)}\right\}_\alpha} e^{\sum_\alpha \hat{m}_\alpha\sum_i s_i^{(\alpha)}+\beta\sum_\alpha\sum_i h_i s_i^{(\alpha)}}\right]
\end{aligned}$$

where (a) is obtained by splitting the sum by magnetization and (b) is obtained by taking the Fourier transform of the Dirac delta function, followed by a change of variables $\hat{m}_\alpha = 2\pi\mathbb{i}\lambda_\alpha$. We then chose to ignore the irrelevant prefactors in front of the integral. Then:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[Z^n\right] &\propto \int\prod_\alpha \mathrm{d}m_\alpha\,\mathrm{d}\hat{m}_\alpha\, e^{\left(\beta\frac{N}{2}\sum_\alpha m_\alpha^2 - N\sum_\alpha \hat{m}_\alpha m_\alpha\right)}\mathbb{E}_{\boldsymbol{h}}\left[\sum_{\left\{\boldsymbol{s}^{(a)}\right\}_\alpha}\prod_\alpha\prod_i e^{\hat{m}_\alpha s_i^{(\alpha)}+\beta h_i s_i^{(\alpha)}}\right] \\
&\overset{(c)}{\propto} \int\prod_\alpha \mathrm{d}m_\alpha\,\mathrm{d}\hat{m}_\alpha\, e^{\left(\beta\frac{N}{2}\sum_\alpha m_\alpha^2 - N\sum_\alpha \hat{m}_\alpha m_\alpha\right)}\mathbb{E}_{\boldsymbol{h}}\left[\prod_i\prod_\alpha\sum_{s_i^{(\alpha)}=\pm 1} e^{\hat{m}_\alpha s_i^{(\alpha)}+\beta h_i s_i^{(\alpha)}}\right] \\
&\overset{(d)}{\propto} \int\prod_\alpha \mathrm{d}m_\alpha\,\mathrm{d}\hat{m}_\alpha\, e^{\left(\beta\frac{N}{2}\sum_\alpha m_\alpha^2 - N\sum_\alpha \hat{m}_\alpha m_\alpha\right)}\left\{\mathbb{E}_h\left[\prod_\alpha 2\cosh\left(\beta h+\hat{m}_\alpha\right)\right]\right\}^N \\
&= \int\prod_\alpha \mathrm{d}m_\alpha\,\mathrm{d}\hat{m}_\alpha\, e^{N\left[\frac{\beta}{2}\sum_\alpha m_\alpha^2 - \sum_\alpha \hat{m}_\alpha m_\alpha + \log\left(\mathbb{E}_h\left[\prod_\alpha 2\cosh(\beta h+\hat{m}_\alpha)\right]\right)\right]}
\end{aligned}$$

where (c) is obtained by writing the sum of products into a product of sums and (d) comes from the fact that all h_i's are i.i.d. so that the integral over the vectors $\mathbf{h}$ has been replaced by a single intergal over a scalar h. We have again got read entirely of the combinatoric sums but at the price of introducing integrals.

2.2.2 Replica Symmetry Ansatz

At this point, we seem to have reached a quite complicated expression: we now have to somehow manage to integrate over all the $m_\alpha, \hat{m}_\alpha$, and magically take the $n \to 0$ limit. These integrals can be see seen as an integral over the two n-dimensional vectors matrices $\boldsymbol{m}$ and $\hat{\boldsymbol{m}}$, i.e:

$$\int\prod_\alpha \mathrm{d}m_\alpha\,\mathrm{d}\hat{m}_\alpha =: \int \mathrm{d}\boldsymbol{m}\,\mathrm{d}\hat{\boldsymbol{m}}$$

From the structure of the integral, we should expect that a saddle point method will hold, so that we should extremize the expression in the exponential over these two vectors. While this looks like a formidable challenge (maximization over all possible vectors!) we may *guess* how these vectors will look like at the extremum. A very reasonable assumption, called the *replica symmetry* (RS) ansatz, is that at the extremum all the replicas are equivalent, so that

$$m_\alpha \equiv m, \qquad \hat{m}_\alpha \equiv \hat{m}, \qquad \forall\, \alpha$$

Physicists, who have been trained to follow the steps of the giant german scientists of the late XIX$^{\text{th}}$ and early XX$^{\text{th}}$ century, call such a guess an *ansatz*. Following then the *replica symmetric ansatz*, the seemingly huge monster $\mathbb{E}_{\boldsymbol{h}}\left[Z^n\right]$ is now reduced to the more gentle:

$$\mathbb{E}_{\boldsymbol{h}}\left[Z^n\right] = \propto \int \mathrm{d}m\, \mathrm{d}\hat{m}\, \exp\left\{N\left[\frac{\beta}{2}nm^2 - n\hat{m}m + \log\left(\mathbb{E}_h\left[2^n \cosh^n\left(\beta h + \hat{m}\right)\right]\right)\right]\right\}$$

We have thus quite simplified the problem and the averaged free entropy we are looking for reads

$$\begin{aligned}
\Phi(\beta, \Delta) &= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\boldsymbol{h}}\left[\log\left(Z(\beta, \boldsymbol{h})\right)\right] \\
&\overset{(a)}{=} \lim_{N\to\infty} \frac{1}{N} \lim_{n\to 0} \frac{\mathbb{E}_{\boldsymbol{h}}\left[\left(Z(\beta, \boldsymbol{h})\right)^n\right] - 1}{n} \\
&\overset{(b)}{=} \lim_{n\to 0} \frac{1}{n} \lim_{N\to\infty} \frac{\mathbb{E}_{\boldsymbol{h}}\left[\left(Z(\beta, \boldsymbol{h})\right)^n\right] - 1}{N} \\
&\overset{(c)}{=} \lim_{n\to 0} \frac{1}{n} \mathsf{Extr}_{m,\hat{m}} \left[\frac{\beta}{2}nm^2 - n\hat{m}m + \log\left(\mathbb{E}_h\left[2^n \cosh^n\left(\beta h + \hat{m}\right)\right]\right)\right]
\end{aligned}$$

where (a) is simply applying the replica trick; (b) is a non-rigorous swap of two limits which is assumed to be correct in the replica method; and (c) is the saddle point method. We have almost finished the replica computation, the last step is to get rid of the remaining n. This can be done by using the replica trick once more:

$$\begin{aligned}
\Phi(\beta, \Delta) &\overset{(d)}{=} \lim_{n\to 0} \frac{1}{n} \mathsf{Extr}_{m,\hat{m}} \left[\frac{\beta}{2}nm^2 - n\hat{m}m + n\mathbb{E}_h\left[\log\left(2\cosh\left(\beta h + \hat{m}\right)\right)\right]\right] \\
&= \mathsf{Extr}_{m,\hat{m}} \left[\frac{\beta}{2}m^2 - \hat{m}m + \mathbb{E}_h\left[\log\left(2\cosh\left(\beta h + \hat{m}\right)\right)\right]\right]
\end{aligned}$$

where (d) comes from the trick that as $n \to 0^+$:

$$\mathbb{E}[X^n] = \mathbb{E}\left[e^{n\log(X)}\right] \approx \mathbb{E}\left[1 + n\log(X)\right] = 1 + n\mathbb{E}\left[\log(X)\right] \approx e^{n\mathbb{E}[\log(X)]}$$

which implies

$$\log\left(\mathbb{E}[X^n]\right) \approx \log\left(1 + n\mathbb{E}\left[\log(X)\right]\right) \approx n\mathbb{E}\left[\log(X)\right]$$

This kind of voodoo replica magic should be astonishing: essentially, we see that we can push the expectation within a function when n is going to 0! This already hints at the fact that, for this to be really valid, we shall require some concentration property for the random variables! In any case, we have finished the replica part of our computation, and have managed to bring back the computation to an extremization of a two-dimensional function, just like we did for the Curie-Weiss model:

$$\Phi(\beta, \Delta) = \mathsf{extr}_{m,\hat{m}} \left[\frac{\beta}{2}m^2 - \hat{m}m + \mathbb{E}_h\left[\log\left(2\cosh\left(\beta h + \hat{m}\right)\right)\right]\right]$$

2.2.3 Computing the Saddle points: mean-field equation

We now need to compute the saddle points. We have:

$$\frac{\partial}{\partial m}\left[\frac{\beta}{2}m^2 - \hat{m}m + \mathbb{E}_h\left[\log\left(2\cosh\left(\beta h + \hat{m}\right)\right)\right]\right] = \beta m - \hat{m} \quad \Rightarrow \quad \hat{m} = \beta m$$

Plugging this back, we reach a formula very similar to the one obtained for Curie-Weiss. Defining

$$\Phi_{\rm RS}(m,\beta,\Delta) \triangleq -\frac{\beta}{2}m^2 + \mathbb{E}_h\left[\log\left(2\cosh\left(\beta(h+m)\right)\right)\right]$$

we find:

$$\Phi(\beta,\Delta) = \mathsf{extr}_m\, \Phi_{\rm RS}(m) = \Phi_{\rm RS}(m^*) \tag{2.1}$$

where m^* will satisfy the self-consistent mean-field equation

$$m = \mathbb{E}_h\left[\tanh\left(\beta(h+m)\right)\right] = \int \mathrm{d}h\, \frac{e^{-\frac{h^2}{2\Delta}}}{\sqrt{2\pi\Delta}} \tanh(\beta(h+m))$$

As we did in the Curie-Weiss model, we can also compute the large deviation function that gives us the free entropy for a fixed value of m. In fact, this is self-averaging as well, since we could repeat all the steps of theorem 7 with an indicator function. The free entropy is obtained by doing the saddle point in the correct order and first differentiating wrt $\hat{m}$, leading to an implicit equation on $\hat{m}^*$:

$$m = \mathbb{E}_h\left[\tanh\left(\beta h + \hat{m}^*\right)\right]$$

so that the replica symmetric approach predicts

$$\mathbb{P}(\bar{S} = m) \asymp e^{N\Phi^{\rm LD}(\beta,\Delta,m)}$$

where

$$\Phi^{\rm LD}(\beta,\Delta,m) = \mathsf{extr}_{\hat{m}}\left[\frac{\beta}{2}m^2 - \hat{m}m + \mathbb{E}_h\left[\log\left(2\cosh\left(\beta h + \hat{m}\right)\right)\right]\right]$$

where the extremization over $\hat{m}$ imposes that it solves eq. (2.2.3). Indeed it is easy to check that this recover the large deviation function of the previous chapter.

We illustrate the result of the replica predictions are shown in Figure.2.2.1 where we plot the minimial cost of the assignement versus the variance of the random field. This is obtained by solving the self constant equation 2.1, and computing the equilibrium energy by deriving the free entropy with respect to the (inverse) temperature :

$$\mathbb{E}\langle e\rangle = \partial_\beta \Phi(\beta,\Delta) = -\frac{(m^*)^2}{2} + \mathbb{E}_h\left[(h+m^*)\tanh\left(\beta(h+m^*)\right)\right] \tag{2.2}$$

and taking the zero temperature limit so that

$$\mathbb{E}[e_{\rm min}] = -\frac{(m^*)^2}{2} + \mathbb{E}_h\left[(h+m^*)\,\mathrm{sign}\left(h+m^*\right)\right] \tag{2.3}$$

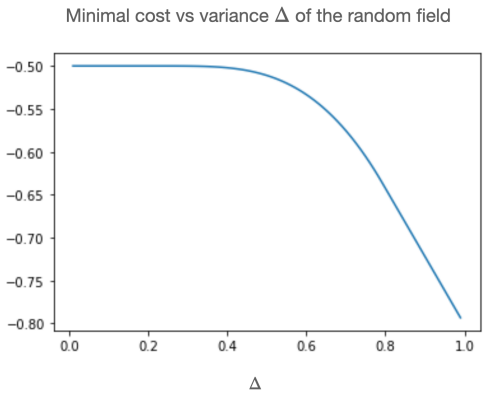


Figure 2.2.1: Minimum energy in the Random Field Ising Model depending on the variance Δ.

2.3 A rigorous computation with the interpolation technique

Given this computation was somehow acrobatic, it would be only natural to seek rigorous re-assurance that the result we reached is exact. In order to do so, we shall use the interpolation method introduced by Francesco Guerra to prove results from the replica method.

2.3.1 A Simple Problem

First we start by a different, but simpler, problem. Consider a system with the Hamiltonian:

$$\mathcal{H}_0\left(\boldsymbol{s},\boldsymbol{h};m\right)=-\sum_i s_i(h_i+m)$$

The corresponding partition function and free entropy per spin read

$$Z_0(\beta,\boldsymbol{h};m)=\sum_{\boldsymbol{s}} e^{\beta\sum_i s_i(h_i+m)}=\prod_i\sum_{s_i=\pm1} e^{\beta s_i(h_i+m)}=\prod_i 2\cosh\left(\beta(h_i+m)\right)$$
$$\Phi_0(\beta,\Delta;m)=\mathbb{E}_{\boldsymbol{h}}\left[\frac{\log\left(Z_0(\beta,\boldsymbol{h};m)\right)}{N}\right]=\mathbb{E}_h\left[\log\left(2\cosh\left(\beta(h+m)\right)\right)\right]$$

In fact, we can even define this partition sum at fixed value of $\bar{S}=s$, to access large deviations:

$$Z_0(\beta,\boldsymbol{h};m,s)=\sum_{\boldsymbol{s}}\mathbf{1}(\bar{S}=s)e^{\beta\sum_i s_i(h_i+m)}$$

We do not know how to do this computation directly but we can, however, do it using the Gartner-Ellis theorem, or Legendre transform, since as $N\to\infty$ this is equivalent to computing the rate. We thus write

$$\tilde{Z}_0(\beta,\boldsymbol{h};m,k)=\sum_{\boldsymbol{s}} e^{\beta\sum_i s_i(h_i+m)+k\sum_i s_i}$$
$$\frac{1}{N}\tilde{Z}_0(\beta,\boldsymbol{h};m,k)\to\mathbb{E}_h\left[\log\left(2\cosh\left(\beta(h+m)+k\right)\right)\right]$$

and get, from Gartner-Ellis, imposing $s = m$:

$$\Phi_0(\beta, m, \Delta) = \lim_{N\to\infty} \frac{1}{N} \log Z_0(\beta, \boldsymbol{h}; m, \bar{S} = m) = \mathbb{E}_h \left[\log\left(2\cosh\left(\beta(h+m)+k^*\right)\right)\right] - k^* m$$
$$m = \mathbb{E}_h \left[\tanh\left(\beta(h+m)+k^*\right)\right]$$

we can make the trivial change of variable $k^* = \hat{m} - \beta m$ and we reach

$$\Phi_0(\beta, m, \Delta) = \mathrm{extr}_{\hat{m}}\, \mathbb{E}_h \left[\log\left(2\cosh\left(\beta h + \hat{m}\right)\right)\right] - m\hat{m} + \beta m^2$$

and we now have something that looks already very close to the replica prediction!

2.3.2 Guerra's Interpolation

Guerra's method consists in modifying the Hamiltonian $\mathcal{H}_0$ to transform it progressively into the actual problem. This is done as follows: we define a family of Hamiltonians and their associated partition functions at a different "times" t as:

$$\mathcal{H}_t\left(\boldsymbol{s}, \boldsymbol{h}; m\right) = -\sum_i s_i \left[h_i + m(1-t)\right] - t\frac{N}{2}\left(\sum_i \frac{s_i}{N}\right)^2,$$
$$Z_t\left(\beta, \boldsymbol{h}; m\right) = \sum_{\boldsymbol{s}} \mathbf{1}(\bar{S} = m) e^{-\beta \mathcal{H}_t(\boldsymbol{s}, \boldsymbol{h}; m)}, \quad \forall\, t \in (0, 1]$$

It is easy to verify that that, for $t = 0$, we recover the model of the previous paragraph, while for $t = 1$ we have:

$$\mathcal{H}_1\left(\boldsymbol{s}, \boldsymbol{h}; m\right) \equiv \mathcal{H}_{\mathrm{RFIM}}\left(\boldsymbol{s}, \boldsymbol{h}\right), \qquad Z_1\left(\beta, \boldsymbol{h}; m\right) \equiv Z_{\mathrm{RFIM}}\left(\beta, \boldsymbol{h}, m\right), \qquad \forall\, m \in [-1:1]$$

We are now going to interpolate the model at time 1 from the one at time 0, and write, using the fundamental theorem of calculus:

$$\begin{aligned}
\Phi(\beta, m, \Delta) &= \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{h}} \left[\frac{\log\left(Z_{\mathrm{RFIM}}\left(\beta, \boldsymbol{h}\right)\right)}{N}\right] \\
&= \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{h}} \left[\frac{\log\left(Z_1\left(\beta, \boldsymbol{h}; m\right)\right)}{N}\right] \\
&= \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{h}} \left[\frac{\log\left(Z_0\left(\beta, \boldsymbol{h}; m\right)\right)}{N} + \int_0^1 \mathrm{d}\tau\ \frac{\partial}{\partial t} \frac{\log\left(Z_t\left(\beta, \boldsymbol{h}; m\right)\right)}{N}\bigg|_{t=\tau}\right] \\
&= \Phi_0(m, \Delta, \beta) + \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{h}} \left[\int_0^1 \mathrm{d}\tau\ \frac{\partial}{\partial t} \frac{\log\left(Z_t\left(\beta, \boldsymbol{h}; m\right)\right)}{N}\bigg|_{t=\tau}\right]
\end{aligned}$$

What is left to do is to compute the additional integral term. We find:

$$\begin{aligned}
\frac{\partial}{\partial t}\frac{\log\left(Z_t\left(\beta,\boldsymbol{h};m\right)\right)}{N} &= \frac{1}{N}\frac{1}{Z_t\left(\beta,\boldsymbol{h};m\right)}\frac{\partial}{\partial t}\sum_{\boldsymbol{s}}\mathbf{1}(\bar{S}=m)\mathrm{e}^{-\beta\mathcal{H}_t(\boldsymbol{s},\boldsymbol{h};m)} \\
&= \frac{1}{N}\frac{1}{Z_t\left(\beta,\boldsymbol{h};m\right)}\sum_{\boldsymbol{s}}\mathbf{1}(\bar{S}=m)\mathrm{e}^{-\beta\mathcal{H}_t(\boldsymbol{s},\boldsymbol{h};m)}(-\beta)\frac{\partial}{\partial t}\mathcal{H}_t\left(\boldsymbol{s},\boldsymbol{h};m\right) \\
&= \frac{\beta}{NZ_t\left(\beta,\boldsymbol{h};m\right)}\sum_{\boldsymbol{s}}\mathbf{1}(\bar{S}=m)\mathrm{e}^{-\beta\mathcal{H}_t(\boldsymbol{s},\boldsymbol{h};m)}\left[-m\sum_i s_i+\frac{N}{2}\left(\sum_i\frac{s_i}{N}\right)^2\right] \\
&= \beta\sum_{\boldsymbol{s}}\underbrace{\frac{\mathbf{1}(\bar{S}=m)\mathrm{e}^{-\beta\mathcal{H}_t(\boldsymbol{s},\boldsymbol{h};m)}}{Z_t\left(\beta,\boldsymbol{h};m\right)}}_{=P_{\beta,t,\boldsymbol{h},m}(\boldsymbol{s})}\left[-m\sum_i\frac{s_i}{N}+\frac{1}{2}\left(\sum_i\frac{s_i}{N}\right)^2\right] \\
&= \beta\left\{-m\left\langle\sum_i\frac{s_i}{N}\right\rangle_{\beta,t,\boldsymbol{h},m}+\frac{1}{2}\left\langle\left(\sum_i\frac{s_i}{N}\right)^2\right\rangle_{\beta,t,\boldsymbol{h},m}\right\} \\
&= \beta\left\{\frac{1}{2}\left\langle\left(m-\sum_i\frac{s_i}{N}\right)^2\right\rangle_{\beta,t,\boldsymbol{h},m}-\frac{m^2}{2}\right\}
\end{aligned}$$

Substituting this back we obtain

$$\begin{aligned}
\Phi(\beta,m,\Delta) &= \Phi_0(\beta,m,\Delta)+\lim_{N\to\infty}\mathbb{E}_{\boldsymbol{h}}\left[\int_0^1\mathrm{d}\tau\,\beta\left\{\frac{1}{2}\left\langle\left(m-\sum_i\frac{s_i}{N}\right)^2\right\rangle_{\beta,\tau,\boldsymbol{h},m}-\frac{m^2}{2}\right\}\right] \\
&= \underbrace{\Phi_0(\beta,m,\Delta)-\frac{\beta m^2}{2}}_{=\Phi_{\mathrm{RS}}(m)}+\underbrace{\frac{\beta}{2}\lim_{N\to\infty}\mathbb{E}_{\boldsymbol{h}}\left[\int_0^1\mathrm{d}\tau\left\langle\left(m-\sum_i\frac{s_i}{N}\right)^2\right\rangle_{\beta,\tau,\boldsymbol{h},m}\right]}_{=0} \\
&= \mathrm{extr}_{\hat{m}}\Phi_{\mathrm{RS}}(m,\hat{m}), \qquad \forall\, m\in[-1:1]
\end{aligned}$$

where the last equality to 0 arises because we have restricted the magnetization to be precisely equal to m. We have thus succeed in proving the replica symmetric equation for the free entropy at all values of m. More importantly, we saw that the replica method was trustworthy!

Bibliography

A nice review on the random field Ising model in physics is Nattermann (1998). It played a fundamental role in the development of disordered systems. The replica method was introduced by Sam Edwards, who credited it to Marc Kac (Edwards et al. (2005)). It has been turned into a powerful and versatile tool by the work of a generation of physicists led by Parisi, Mézard and Virasoro (Mézard et al. (1987b)). The interpolation trick we discussed to prove the replica formula was famously introduced by Guerra (2003). The peculiar technique we used here fixing the magnetization is inspired from El Alaoui and Krzakala (2018). Probabilistic inequalities such as Gaussian Poincaré are fundamental to modern probability and statistics theories. A good reference is Boucheron et al. (2013). These concentration inequalities are the cornerstone of all approaches to rigorous mathematical treatments of statistical physics models.

2.4 Exercises

Exercise 2.1: Gaussian Poincaré inequality and Efron-Stein

In order to prove the Gaussian Poincaré inequality, we first need to prove the very generic Efron-Stein inequality, which is at the root of many important result in probability theory:

Theorem 8 (Efron-Stein)**.** *Suppose that $X_1, \dots, X_n$ and $X'_1, \dots, X'_n$ are independant random variable, with X_i and X'_i having the same law for all i. Let $X = (X_1, \dots, X_i, \dots X_n)$ and $X^{(i)} = (X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots X_n)$. Then for any function $f : \mathbb{R}^n \to \mathbb{R}$ we have:*

$$\mathrm{var}(f(X)) \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}[(f(X) - f(X^{(i)}))^2].$$

We are going to prove Efron-Stein using the so-called Lindeberg trick, by considering averages over mixed ensembles of the X_i and X'_i. First we define the set $X_{(i)}$ as the random vector equal to X' to i, and equal to X for all larger indices, i.e. $X_{(i)} = (X'_1, \dots, X'_{i-1}, X'_i, X_{i+1}, \dots X_n)$. In particular $X_{(0)} = X$ and $X_{(n)} = X'$.

1. Show that (this is called the Lindeberg replacement trick):

$$\mathrm{Var}\left[f(X)\right] = \mathbb{E}[f(X)(f(X) - f(X'))] = \sum_{i=1}^{n} \mathbb{E}[f(X)(f(X_{(i-1)}) - f(X_{(i)}))]$$

2. Show that for all i:

$$\begin{aligned}\mathbb{E}[f(X)(f(X_{(i-1)}) - f(X_{(i)}))] &= \mathbb{E}[f(X^{(i)})(f(X_{(i)}) - f(X_{(i-1)}))] \\ &= \frac{1}{2}\mathbb{E}\left[(f(X) - f(X^{(i)})(f(X_{(i-1)}) - f(X_{(i)})\right]\end{aligned}$$

3. Show that by Cauchy-Schwartz:

$$|\mathbb{E}[f(X)(f(X_{(i-1)}) - f(X_{(i)}))]| \leq \frac{1}{2}\mathbb{E}\left[(f(X) - f(X^{(i)})^2\right]$$

   and proove the Efron-Stein theorem.

Now that we have Efron-Stein, we can prove Poincare's inequality for Gaussian random variables. We shall do it for a single variable, and let the reader generalize the proof to the multi-valued case.
With X_i a ± 1 random variable that takes each value with probability $1/2$ (this is called a Rademacher variable), define:

$$S_n = X_1 + X_2 + \ldots + X_n\,.$$

4. Using Efron-Stein, show that

$$\mathrm{Var}[f(\frac{S_n}{\sqrt{n}})] \leq \frac{n}{4}\mathbb{E}\left[\left(f\left(\frac{S_{n-1}}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) - f\left(\frac{S_{n-1}}{\sqrt{n}} - \frac{1}{\sqrt{n}}\right)\right)^2\right]$$

5. Using the central limit theorem, show that this leads, as $n \to \infty$, to the following theorem:

   Theorem 9 (Gaussian-Poincaré)**.** *Suppose f: $\mathbb{R} \mapsto \mathbb{R}$ is a smooth function and X is Gaussian $X \sim \mathcal{N}(0,1)$, then*

   $$Var[f(X)] \leq \mathbb{E}\left[(f'(X))^2\right].$$

EXERCISE 2.2: RANDOM FIELD ISING MODEL BY THE CAVITY METHOD

The goal of this exercise is to provide an alternative derivation for the free entropy of the random field Ising model, using a technique close to the cavity method.

1. Show that, by adding one spin to a system of N spins, one has:

   $$A_N(\beta,\Delta) =: \mathbb{E}_{h,\boldsymbol{h}} \log \frac{Z_{N+1}}{Z_N} = \mathbb{E}_h \log\langle e^{-\frac{\beta}{2}\bar{S}^2} 2\cosh\left(\beta(\bar{S}+h)\right)\rangle_{N,\beta,\boldsymbol{h}} + o(1)$$

2. Show that, by adding an external magnetic field B to the Hamiltonian (i.e. a term $B\sum_i S_i$, one can get a concentration of the magnetization for almost all B so that, for any $\boldsymbol{h}$, we have:

   $$\int_{B1}^{B2} \langle \bar{S}^2\rangle_{N,\beta,\boldsymbol{h}} - \langle \bar{S}\rangle^2_{N,\beta,\boldsymbol{h}} dB \leq 2/\beta N$$

   Note that this gives the concentration over the Boltzmann averages, but not over the disorder (the fields $\boldsymbol{h}$). This means we showed that the magnetization converges to a value $m(\boldsymbol{h})$ that could — a priori — depend on the given realization of the disorder $\boldsymbol{h}$.

3. Explain why this implies, almost everywhere in B, that at large N:

   $$A_N(\beta,\Delta,B) = \mathbb{E}_{h,\boldsymbol{h}}\left[-\frac{\beta}{2}\langle \bar{S}\rangle^2_{N,\beta,\boldsymbol{h}} + \log 2\cosh\left(\beta(\langle \bar{S}\rangle_{N,\beta,\boldsymbol{h}} + h + B)\right)\right] + o(1)$$

4. Given that we proved that $\langle S\rangle_{N,\beta,\boldsymbol{h}}$ concentrates as N grows to a value $m(\boldsymbol{h})$, show that this implies, as $N \to \infty$, the bound:

   $$\Phi(\beta,\Delta,B) \leq \sup_m \left[-\frac{\beta}{2}m^2 + \mathbb{E}_h \log 2\cosh\left(\beta(m+h+B)\right)\right]$$

5. Use the variational approach of lecture 1 to obtain the converse bound and finally show:

   $$\Phi(\beta,\Delta) = \sup_m -\frac{\beta}{2}m^2 + \mathbb{E}_h \log 2\cosh\left(\beta(m+h)\right)$$

Exercise 2.3: Mean-field algorithm and state evolution for the RFIM

Our aim in this exercise is to provide an algorithm for finding the lowest energy configuration, and to analyse its property.

1. Using the variational approach of section 1.3 or the cavity method of section 1.4, explain why the following iterative algorithm might be a good one for finding the lowest energy in practice (tip: this is the zero temperature limit of a finite temperature iteration):

$$S_i^{t+1} = \text{sign}\left(h_i + \sum_i S_i^t/N\right)$$

2. Implement the code of this algorithm, check that it indeed finds configurations with minimum values that match the replica predictions for the minimum energy when the system is large enough.

3. Show that in the large N limit, the dynamics of this algorithm obey a "state evolution" equation, that is, that at each time the average magnetization $m^t = \sum_i S_i^t/N$ is given by the deterministic equation:

$$m^{t+1} = \mathbb{E}_h \text{sign}(h + m^t)$$

and conclude that the algorithm is performing a fixed point iteration of the replica symmetric free energy equation 2.2.3.

Chapter 3

A first application: The spectrum of random matrices

> Unfortunately, no one can be told what the Matrix is. You have to see it for yourself
>
> *The Matrix* (Morpheus)
> 1999

We shall move now to a first non-trivial application of the replica method to compute the spectrum of random matrices. Random matrices were introduced by Eugene Wigner to model the nuclei of heavy atoms. He postulated that the spacings between the lines in the spectrum of a heavy atom nucleus should resemble the spacings between the eigenvalues of a random matrix, and should depend only on the symmetry class of the underlying evolution. Since then, the study of random matrices has become a field in itself, with numerous applications ranging from solid-state physics and quantum chaos to machine learning and number theory. The simplest of all random matrix is the Wigner one:

$$\mathrm{A}_N = \frac{1}{\sqrt{2N}} \left(\mathrm{G} + \mathrm{G}^\top \right)$$

where G is a random matrix where each element $G_{ij} \sim \mathcal{N}(0,1)$ is i.i.d. chosen from $\mathcal{N}(0,1)$. The central question is: what does the distribution of eigenvalues $\nu_{A_N}(\lambda)$ of such random matrices look like as $N \to \infty$? It turns out that $\nu_{A_N}(\lambda)$ converges towards a well-defined deterministic function $\nu(\lambda)$. We shall see how one can use the replica and the cavity method to compute it, as a first real application of our tools.

3.1 The Stieltjes transform

We will use a technique called the Stieltjes transform. It starts with a very useful identity, defined in the sense of distributions, called Sokhotsky's Formula. It is often used in field theory in physics, where people refer to it as the "Feynman trick". It can also be useful to compute the Kramers-Kronig relations in optics, or to define the Hilbert transform of causal

functions[1]

$$\delta(x-x_0) = -\lim_{\epsilon\to 0}\frac{1}{\pi}\Im\frac{1}{x-x_0+i\epsilon}$$

This formula is at the roots of the theoretical approach to random matrix theory. Indeed, given a probability distribution $\nu(x)$ that can take N values with uniform probability, we have:

$$\nu(x) = \frac{1}{N}\sum_i \delta(x-x_i) = -\lim_{\epsilon\to 0}\frac{1}{N\pi}\sum_i \Im\frac{1}{x-x_0+i\epsilon}$$

Now we use the fact that $1/x$ is the derivative of the logarithm to write

$$\frac{1}{N}\sum_i \frac{1}{x-x_i} = \frac{1}{N}\partial_x\sum_i \log\left(x-x_i\right) = \frac{1}{N}\partial_x \log\prod_i (x-x_i)$$

So that, given an $N\times N$ matrix A with eigenvalues $\lambda_1,\dots,\lambda_N$, and using the fact that the determinant is the product of eigenvalues, we have:

$$\nu(\lambda) = \frac{1}{N}\sum_i \delta(\lambda-\lambda_i) = -\frac{1}{N\pi}\Im\lim_{\epsilon\to 0}\partial_\lambda \log\det(A-(\lambda+i\epsilon)\mathbf{1})$$

This is the basis of the computation techniques used in random matrix theory. For a given matrix A, we introduce the Stieltjes transform, as

$$S_A(\lambda) = -\frac{1}{N}\sum_i \frac{1}{\lambda-\lambda_i} = -\frac{1}{N}\partial_\lambda \log\det(A-\lambda\mathbf{1})\,,$$

and once we have the Stieltjes transform, we can access the probably density via its imaginary part:

$$\nu_A(\lambda) = \frac{1}{\pi}\lim_{\epsilon\to 0}\Im S_A(\lambda+i\epsilon)\,.$$

The entire field of random matrix theory is thus reduced to the computation of the Stieltjes transform associated with the probability distribution of eigenvalues.

3.2 The replica method

Our aim is now to compute the Stieltjes transform of the Wigner matrix using the replica method. We shall not aim for mathematical rigor here, as our goal is merely the demonstration of the power of our tools.

[1]This formula is easily proven using Cauchy integration in the complex plane. Alternatively, one can simply use $x^2-\varepsilon^2 = (x+i\varepsilon)(x-i\varepsilon)$. Indeed,

$$\lim_{\varepsilon\to 0^+}\int_{\mathbb{R}}\frac{f(x)}{x\pm i\varepsilon}\,dx = \mp i\pi\lim_{\varepsilon\to 0^+}\int_{\mathbb{R}}\frac{\varepsilon}{\pi(x^2+\varepsilon^2)}f(x)\,dx + \lim_{\varepsilon\to 0^+}\int_{\mathbb{R}}\frac{x^2}{x^2+\varepsilon^2}\,\frac{f(x)}{x}\,dx.$$

The first integral approaches a Dirac delta function as $\varepsilon\to 0^+$ (it is a nascent delta function) and therefore, the first term equals $\mp i\pi f(0)$. The second term converges to a (real) Cauchy principal-value integral, so that

$$\Im\lim_{\varepsilon\to 0^+}\int_{\mathbb{R}}\frac{f(x)}{x\pm i\varepsilon}\,dx = \mp i\pi f(0) = \mp i\pi\int_{\mathbb{R}}\delta(x)f(x)dx\,. \tag{3.1}$$

3.2.1 Averaging replicas

Assuming that both the Stieltjes transform and the density of eigenvalues are self-averaging, we need to compute their expectation in the large size limit.

$$\lim_{N\to\infty} \mathbb{E} S_{\mathrm{A}_N}(\lambda) = -\partial_\lambda \lim_{N\to\infty} \frac{1}{N} \mathbb{E} \log\det\left(\mathrm{A}_N - \lambda \mathrm{I}_N\right). \tag{3.2}$$

Instead of directly using the replica trick on the log-det, we shall instead use the following approach, which will turn out to be more practical:

$$\mathbb{E}\log\det\left(\mathrm{A}_N - \lambda \mathrm{I}_N\right) = -2\mathbb{E}\log\left[\det\left(\mathrm{A}_N - \lambda \mathrm{I}_N\right)^{-1/2}\right]$$

so that, following the replica strategy of computing $\log X$ by computing instead $(X^n - 1)/n$, we shall need to compute average of the n-th power of $\det\left(\mathrm{A}_N - \lambda \mathrm{I}_N\right)^{-1/2}$. We now use the Gaussian integral [2] to express the square root of determinant as an integral and write:

$$\begin{aligned}\mathbb{E}\det\left(\mathrm{A}_N - \lambda \mathrm{I}_N\right)^{-n/2} &= \mathbb{E}\left[\prod_{a=1}^{n} \int_{\mathbb{R}^N} \frac{\mathrm{d}\boldsymbol{x}}{(2\pi)^{N/2}} e^{-\frac{1}{2}\boldsymbol{x}^\top(\mathrm{A}_N - \lambda \mathrm{I}_N)\boldsymbol{x}}\right] \\ &= \int_{\mathbb{R}^N} \prod_{a=1}^{n} \frac{\mathrm{d}\boldsymbol{x}^a}{(2\pi)^{N/2}} e^{\frac{\lambda}{2}\sum_{a=1}^{n}\|\boldsymbol{x}^a\|_2^2} \mathbb{E}\left[e^{-\frac{1}{2}\sum_{a=1}^{n}\boldsymbol{x}^{a\top}\mathrm{A}_N\boldsymbol{x}^a}\right]\end{aligned} \tag{3.3}$$

For our Wigner matrix (the so-called GOE(N) ensemble) we can write $\mathrm{A} = \frac{1}{\sqrt{2N}}\left(\mathrm{G} + \mathrm{G}^\top\right)$ for $\mathrm{G} \sim \mathcal{N}(0,1)$ i.i.d. We can thus write the average in eq. equation 3.3 as:

$$\mathbb{E}_{\mathrm{A}}\left[e^{-\frac{1}{2}\sum_{a=1}^{n}\boldsymbol{x}^{a\top}\mathrm{A}_N\boldsymbol{x}^a}\right] = \mathbb{E}_{\mathrm{G}}\left[e^{-\frac{1}{\sqrt{2N}}\sum_{a=1}^{n}\boldsymbol{x}^a\mathrm{G}\boldsymbol{x}^a}\right] = \prod_{ij}\mathbb{E}_{G_{ij}}\left[e^{-\frac{G_{ij}}{\sqrt{2N}}\sum_{a=1}^{n}x_i^a x_j^a}\right].$$

At this point, it is a good idea to refresh our knowledge of Gaussian integrals, as we shall use them often. In particular, we have

$$\int e^{-ax^2+bx}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}}, \text{ or } \mathbb{E}_x\left[e^{bx}\right] = e^{\frac{b^2}{2a}}$$

so that

$$\mathbb{E}_{\mathrm{A}}\left[e^{-\frac{1}{2}\sum_{a=1}^{n}\boldsymbol{x}^{a\top}\mathrm{A}_N\boldsymbol{x}^a}\right] = \prod_{ij} e^{\frac{1}{4N}\left(\sum_{a=1}^{n}x_i^a x_j^a\right)^2} = \prod_{ij} e^{\frac{N}{4}\sum_{a,b=1}^{n}\left(\frac{x_i^a x_j^a x_i^b x_j^b}{N^2}\right)} = e^{\frac{N}{4}\sum_{a,b=1}^{n}\left(\frac{\boldsymbol{x}^a\cdot\boldsymbol{x}^b}{N}\right)^2}.$$

[2] Remember that Gaussian distributions are normalized so that

$$\frac{1}{\sqrt{\det\left(\mathrm{A}_N - \lambda \mathrm{I}_N\right)}} \int_{\mathbb{R}^N} \frac{\mathrm{d}\boldsymbol{x}}{(2\pi)^{N/2}} e^{-\frac{1}{2}\boldsymbol{x}^\top(\mathrm{A}_N - \lambda \mathrm{I}_N)\boldsymbol{x}} = 1\,.$$

A more generic formula, that turns out to be used in 90% of replica computations, is that if A is a symmetric positive-definite matrix, then

$$\int e^{-\frac{1}{2}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x} + \boldsymbol{B}^T\boldsymbol{x}}\,\mathrm{d}^n x = \sqrt{\frac{(2\pi)^n}{\det A}} e^{\frac{1}{2}\boldsymbol{B}^T\mathbf{A}^{-1}\boldsymbol{B}}.$$

This is a very typical step of a replica computation! We have now performed the integration over the disorder (randomness) and reached:

$$\mathbb{E}\det\left(\mathrm{A}_N-\lambda\mathrm{I}_N\right)^{-n/2}=\int_{\mathbb{R}^N}\prod_{a=1}^{n}\frac{\mathrm{d}\boldsymbol{x}^a}{(2\pi)^{N/2}}e^{N\frac{\lambda}{2}\sum\limits_{a=1}^{n}(\frac{1}{N}\boldsymbol{x}^a\cdot\boldsymbol{x}^a)+\frac{N}{4}\sum\limits_{a,b=1}^{n}(\frac{1}{N}\boldsymbol{x}^a\cdot\boldsymbol{x}^b)^2}.$$

A very important phenomenon has occurred: we see that the integration over the disorder has "coupled" the previously independent replicas. This is indeed what always happens when integrating over the disorder. This new term, that coupled the replicas, is of fundamental importance, and has a name: it is called the *overlap* between replicas. We shall thus define:

$$q^{ab}=:\frac{1}{N}\boldsymbol{x}^a\cdot\boldsymbol{x}^b\,.$$

We now introduce a "delta function" to free the overlap order parameter, just like we did previously for the magnetization in the random field Ising model. For any function f, we have:

$$f\left(\frac{1}{N}\boldsymbol{x}^a\cdot\boldsymbol{x}^b\right)=N^n\int\prod_{1\leq a\leq b\leq n}\mathrm{d}q^{ab}\,\delta\left(Nq^{ab}-\boldsymbol{x}^a\cdot\boldsymbol{x}^b\right)f(\boldsymbol{x}^a\cdot\boldsymbol{x}^b)$$

As before, we shall drop the N^n prefactor, that does not count as we shall eventually take the normalized logarithm and send $N\to\infty$, and write

$$\mathbb{E}\det\left(\mathrm{A}_N-\lambda\mathrm{I}_N\right)^{-n/2}\approx\int\prod_{1\leq a\leq b\leq n}\delta\left(q^{ab}-\frac{\boldsymbol{x}^a\cdot\boldsymbol{x}^b}{N}\right)\mathrm{d}q^{ab}e^{N\frac{\lambda}{2}\sum\limits_{a=1}^{n}q_{aa}+\frac{N}{4}\sum\limits_{a,b=1}^{n}q_{ab}^2}.\tag{3.4}$$

Performing the exact same steps we used in the previous chapters, we now take the Fourier representation of the delta functions (and change variables so that they appear "real" instead of "complex"):

$$\prod_{1\leq a\leq b\leq n}\delta\left(Nq^{ab}-\boldsymbol{x}^a\cdot\boldsymbol{x}^b\right)=\int\prod_{1\leq a\leq b\leq n}\frac{\mathrm{d}\hat{q}^{ab}}{2\pi}\,e^{-\sum\limits_{1\leq a\leq b\leq n}\hat{q}^{ab}\left(Nq^{ab}-\boldsymbol{x}^a\cdot\boldsymbol{x}^b\right)},$$

and inserting it in the above thus allow us to write:

$$\mathbb{E}\det\left(\mathrm{A}_N-\lambda\mathrm{I}_N\right)^{-n/2}\approx\int\prod_{1\leq a\leq b\leq n}\frac{\mathrm{d}q^{ab}\,\mathrm{d}\hat{q}^{ab}}{2\pi}\,e^{N\Phi(q^{ab},\hat{q}^{ab})}\tag{3.5}$$

where:

$$\Phi(q^{ab},\hat{q}^{ab})=-\sum_{1\leq a\leq b\leq n}\hat{q}^{ab}q^{ab}+\frac{\lambda}{2}\sum_{a=1}^{n}q^{aa}+\frac{1}{4}\sum_{a,b=1}^{n}\left(q^{ab}\right)^2+\Psi_x(\hat{q}^{ab})$$

with (given now the sites are decoupled):

$$\begin{aligned}\Psi_x(\hat{q}^{ab})&=\frac{1}{N}\log\int\prod_{a=1}^{n}\frac{\mathrm{d}\boldsymbol{x}^a}{(2\pi)^{N/2}}e^{\sum\limits_{1\leq a\leq b\leq n}\hat{q}^{ab}\boldsymbol{x}^a\cdot\boldsymbol{x}^b}=\frac{1}{N}\log\left(\int\prod_{a=1}^{n}\frac{\mathrm{d}x^a}{\sqrt{2\pi}}e^{\sum\limits_{1\leq a\leq b\leq n}x^a\hat{q}^{ab}x^b}\right)^N\\&=\log\int\prod_{a=1}^{n}\frac{\mathrm{d}x^a}{\sqrt{2\pi}}e^{\sum\limits_{1\leq a\leq b\leq n}x^a\hat{q}^{ab}x^b}\end{aligned}$$

At $N \to \infty$, the integral in equation 3.5 can be evaluated with the saddle point method, and therefore:

$$\mathbb{E} \det \left(\mathrm{A}_N - \lambda \mathrm{I}_N\right)^{-n/2} \approx \exp\left(N \underset{q^{ab},\hat{q}^{ab}}{\operatorname{extr}} \Phi(q^{ab}, \hat{q}^{ab})\right)$$

This concludes the replica computation.

3.2.2 Replica symmetric ansatz

Again, we are trapped with the extremization of a function, but this time it should be over a $n \times n$ matrix, which seems like a complicated space. To make progress in the extremization problem, we need can restrict our search for particular solutions, and we are going, again, to assume replica symmetry:

$$q^{ab} = \delta^{ab} q, \qquad \hat{q}^{ab} = -\frac{1}{2}\delta^{ab}\hat{q}$$

With this ansatz, we have:

$$\sum_{1 \le a \le b \le n} \hat{q}^{ab} q^{ab} = -\frac{n}{2}\hat{q}q, \qquad \sum_{a=1}^{n} q^{aa} = nq, \qquad \sum_{a,b=1}^{n} \left(q^{ab}\right)^2 = nq^2$$

Finally,

$$\Psi_x(\hat{q}) = \log \int \prod_{a=1}^{n} \frac{\mathrm{d}x^a}{\sqrt{2\pi}} e^{\hat{q}\sum_{a=1}^{n}(x^a)^2} = n \log \int \frac{\mathrm{d}x}{\sqrt{2\pi}} e^{-\frac{1}{2}\hat{q}x^2} = -\frac{n}{2}\log(\hat{q})$$

Putting together and applying the saddle point method, we thus reach

$$\begin{aligned} -\frac{2}{N}\mathbb{E} \log\det\left(\mathrm{A}_n - \lambda \mathrm{I}_n\right)^{-1/2} &\approx -2 \lim_{n\to 0} \frac{e^{-\frac{nN}{2}\operatorname{extr}_{q,\hat{q}}\left\{\log\hat{q} - q\hat{q} - \lambda q - \frac{1}{2}q^2\right\}} - 1}{n} \\ &\approx \operatorname{extr}\left\{\log\hat{q} - q\hat{q} - \lambda q - \frac{1}{2}q^2\right\} \end{aligned}$$

To solve this problem, we look at the saddle point equations obtained by taking the derivatives with respect to the parameters $(q, \hat{q})$:

$$\hat{q} = \frac{1}{q}, \qquad q = -\hat{q} - \lambda \qquad \Leftrightarrow \qquad \frac{1}{q} + q + \lambda = 0$$

This has two solutions:

$$q^\star_\pm = \frac{-\lambda \pm \sqrt{\lambda^2 - 4}}{2}$$

Finally, to get the Stieltjes transform, we use the relation in eq. equation 3.2:

$$\lim_{N\to\infty} \mathbb{E} S_{\mathrm{A}_N}(\lambda) = -\partial_\lambda \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\log\det\left(\mathrm{A} - \lambda \mathrm{I}_N\right) = q^\star_\pm = \frac{-\lambda \pm \sqrt{\lambda^2 - 4}}{2}$$

we have thus found the Stieltjes transform, and we have two (!) solutions for $\lambda > 0$ and $\lambda < 0$. Only one of them will be the correct one when we shall inverse the Stieltjes transform, but it will be easy to check which one, since probabilities needs to be positive.

3.2.3 From Stieltjes to the spectrum

We can now compute the spectrum using the relation

$$\nu_A(\lambda) = \frac{1}{\pi} \lim_{\epsilon \to 0} \Im S_A(\lambda + i\epsilon) \, .$$

The first term $-\lambda$ will not give us any non trivial computation, we concentrate on the second one. When $|\lambda| > 2$, we have

$$\sqrt{(\lambda + i\epsilon)^2 - 4} = \sqrt{\lambda^2 - \epsilon^2 - 4 + 2i\epsilon\lambda} \approx \sqrt{\lambda^2 - \epsilon^2 - 4} \left(1 - \frac{i\epsilon\lambda}{\lambda^2 - \epsilon^2 - 4}\right)$$

which, again, has no imaginary part as $\epsilon \to 0$. We are thus forced to conclude that $\nu(\lambda) = 0$ when $|\lambda| > 2$. If $|\lambda| < 2$, on the other hand, we get that

$$\sqrt{(\lambda + i\epsilon)^2 - 4} \approx \sqrt{\lambda^2 - \epsilon^2 - 4} \sqrt{1 - \frac{i\epsilon\lambda}{\lambda^2 - \epsilon^2 - 4}} = i\sqrt{4 - \lambda^2} + O(\epsilon)$$

so that we find (finally choosing the correct replica solution to be the "+" to have a positive distribution):

$$\nu(\lambda) = \frac{1}{\pi}\sqrt{4 - \lambda^2} \text{ if } 2 > \lambda > -2$$
$$\nu(\lambda) = 0 \text{ otherwise}$$

which is indeed the correct Wigner solution, aka the famous "semi-circle" law.

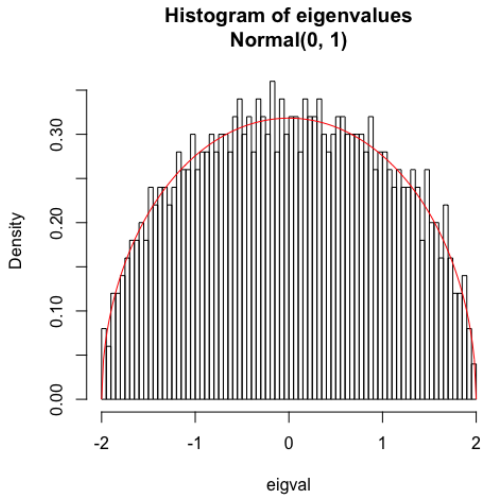


Figure 3.2.1: Simulation of the semicircle law using a 1000 by 1000 Wigner matrix.

3.3 The Cavity method

It is a useful exercise to use the cavity method instead of the replica one to compute the Stieltjes transform. We start from the very definition

$$S_{A_N}(\lambda) = -\frac{1}{N} \sum_i \frac{1}{\lambda - \lambda_i}$$

The idea behind the cavity computation consists in finding a recursion equation between the transform for an $N \times N$ matrix and the one for an $(N+1) \times (N+1)$ matrix. Let us define:

$$M_N = \lambda \mathbf{1} - A_N \qquad R_N = [\lambda \mathbf{1} - A_N]^{-1} = M_N^{-1}$$

where R_N is called the resolvant matrix. Using the formula for computing inverse of matrices via their matrix of cofactors, we find

$$(R_{N+1})_{N+1,N+1} = \frac{\det M_N}{\det M_{N+1}}$$

Additionally, we can compute the determinant of the $N+1$ matrix with the Laplace expansion along the last row, and then on the last column, so that:

$$\det M_{N+1} = (M_{N+1})_{N+1,N+1} \det M_N - \sum_{k,l=1}^{N} (M_{N+1})_{N+1,k} (M_{N+1})_{l,N+1} C_{l,k}^N$$

with $C_{l,k}^N$ the matrix of co-factors of M_N. Dividing the previous expression by $\det M_N$, we thus find

$$\begin{aligned}
\frac{\det M_{N+1}}{\det M_n} &= \frac{1}{(R_{N+1})_{N+1,N+1}} \\
&= (M_{N+1})_{N+1,N+1} - \frac{1}{\det M_N} \sum_{k,l=1}^{N} (M_{N+1})_{N+1,k} (M_{N+1})_{l,N+1} C_{l,k}^N \\
&= \lambda - (A_{N+1})_{N+1,N+1} - \sum_{k,l=1}^{N} (A_{N+1})_{N+1,k} (A_{N+1})_{l,N+1} (R_N)_{k,l}
\end{aligned}$$

At this point, we make the additional assumption that the off-diagonal elements of the resolvent R_N are of order $O(N^{-1/2})$. This can be checked, for instance by expanding in powers of λ with pertubation theory, and observing that, at each order in this expansion, the off-diagonal elements are indeed of order $N^{-1/2}$. In which case the equation further simplifies to

$$(R_{N+1})_{N+1,N+1} = \frac{1}{\lambda - N^{-1} \sum_{l=1}^{N} (R_N)_{l,l}}$$

where we have used that the A_N have i.i.d. elements. Given the matrix elements of the diagonal of R_{N+1} are identically distributed, we find

$$\frac{1}{N} \mathrm{Tr} R_{N+1} = \frac{1}{\lambda - \frac{1}{N} \mathrm{Tr} R_N}$$

so that the Stieltjes tranform verifies

$$S_{A_{N+1}}(\lambda) = -\frac{1}{\lambda + S_{A_N}(\lambda)}$$

and we thus expect, as N increases, that the Stieltjes transform will converge to the fixed point. This is indeed the same (correct) equation that we found with the replica method. We have thus checked that both methods give the correct solution in this case.

Bibliography

The Wigner random matrix was famously introduced in Wigner (1958). The use of the replica method for random matrices iniated with the seminal work of Edwards and Jones Edwards and Jones (1976). It has now grown into a field in iteself, with hundred of deep non-trivial results. A good set of lecture notes on the subject can be found in Livan et al. (2018); Potters and Bouchaud (2020). A classical mathematical reference is Bai and Silverstein (2010).

3.4 Exercises

Exercise 3.1: Wishart, or Marcenko-Pastur law

The goal of this exercise is to repeat the replica computation for Wishart-Matrices, and to derive their distribution of Eigenvalues, also called the Marcenko-Pastur law. Wishart matrices are defined as follows: Consider a $M \times N$ random matrix X with i.i.d. coefficients distributed from a standard normalized Gaussian $\mathcal{N}(0,1)$. The Wishart matrix $\hat{\Sigma}$ is:

$$\hat{\Sigma} =: \frac{1}{N} X X^T \, .$$

In other words, these are the correlation matrices between random data points. As such they are used in many concrete situations in data science and machine learning, in particular to filter out the signal from the noise in data.

1. Denoting $\alpha = M/N$, repeat the replica computation of the Stieltjes Transform, but now for the Wishart matrix in the limit where $M, N \to \infty$, with α fixed. Show that it leads to
$$S_{\hat{\Sigma}}(\lambda) = \frac{1 - \alpha - \lambda \pm \sqrt{(\lambda - \alpha - 1)^2 - 4\alpha}}{2\alpha\lambda}$$

2. For $\alpha < 1$, show that this implies the Marcenko-Pastur law for the distribution of eigenvalues:
$$\nu_{\hat{\Sigma}}(\lambda) = \frac{1}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\alpha\lambda} \text{ with } \lambda \in [\lambda_-, \lambda_+]$$
$$\lambda_\pm = (1 \pm \sqrt{\alpha})^2$$

3. Perform simulation of such random matrices, for large values of M and N and check your predictions for the distribution of eigenvalues.

4. Repeat the simulation for a value $\alpha > 1$, and compare again with the distribution $\nu_{\hat{\Sigma}}(\lambda)$. Are we missing something? Hint: the Stieltjes transform of a delta function $\delta(\lambda)$ is $-1/\lambda$.

Chapter 4

The Random Field Ising Model on Random Graphs

I think that I shall never see
A poem lovely as a tree.

Joyce Kilmer – 1913

We shall continue our exploration of the Boltzmann-Gibbs distribution

$$\mathbb{P}_{N,\beta}\left(\boldsymbol{S}=\boldsymbol{s}\right)=\frac{e^{-\beta\mathcal{H}(\boldsymbol{s})}}{Z_N(\beta)}.$$

for the random field Ising model, but we shall now look at a more complex, and more interesting, topology. We assume the existence of a graph $\mathcal{G}_{ij}$ that connects some of the nodes, such that $\mathcal{G}_{ij}=1$ if i and j are connected, and 0 otherwise. So far, the last chapters dealt only with the case of fully connected graphs $\mathcal{G}_{ij}=1$ for all i,j. Now our Hamiltonian can be written as:

$$\mathcal{H}_{N,J,\{h\},\mathcal{G}}(\boldsymbol{s})=-J\sum_{(i,j)\in\mathcal{G}}S_iS_j-\sum_{i=1}^{N}h_iS_i$$

where J is a scalar coupling constant.

4.1 The root of all cavity arguments

Let us assume we have N spins $i=1,\ldots,N$, that are all isolated. In this case, their probability distribution is simple enough, we have $m_i=\tanh(\beta h_i)$. Imagine now we are connecting these N spins to a new spin S_0, in the spirit of the cavity method. Of course, the m_i are now changed! Let us thus refer to the old values of m_i as the "cavity ones" and write:

$$m_i^c\equiv\tanh\beta h_i\,.$$

With this definition, note that $(1+S_i m_i^c)/2 = e^{\beta h_i S_i}/\cosh(\beta h_i)$. Our main question of interest is: can we write the magnetization of the new spin m_0 as a function of the $\{m_i^c\}$? Let us try! Clearly

$$\begin{aligned} m_0 &= \frac{\sum_{S_0,\{s\}} S_0 e^{\beta S_0 h_0 + \sum_i \beta J S_i S_0 + \beta S_i h_i}}{\sum_{S_0,\{s\}} e^{\beta S_0 h_0 + \sum_i \beta J S_i S_0 + \beta S_i h_i}} = \frac{\sum_{S_0} S_0 e^{\beta S_0 h_0} \prod_i \sum_{S_i} e^{\beta J S_i S_0} \frac{1+S_i m_i^c}{2}}{\sum_{S_0} e^{\beta S_0 h_0} \prod_i \sum_{S_i} e^{\beta J S_i S_0} \frac{1+S_i m_i^c}{2}} \\ &= \frac{X^+ - X^-}{X^+ X^-} \text{ with } X^s = e^{\beta s h_0} \prod_i \sum_{S_i} e^{\beta J S_i s} \frac{1+S_i m_i^c}{2} \end{aligned}$$

Using the identity $\mathrm{atanh} x = \frac{1}{2}\log\frac{1+x}{1-x}$, and applying the atanh on both side, we reach,

$$\begin{aligned} \mathrm{atanh}(m_0) &= \frac{1}{2}\log\frac{1+\frac{X^+-X^-}{X^++X^-}}{1-\frac{X^+-X^-}{X^++X^-}} = \frac{1}{2}\log\frac{X^+}{X^-} \\ &= \beta h_0 + \sum_i \frac{1}{2}\log\frac{e^{\beta J}(1+m_i^c)+e^{-\beta J}(1-m_i^c)}{e^{-\beta J}(1+m_i^c)+e^{\beta J}(1-m_i^c)} \\ &= \beta h_0 + \sum_i \frac{1}{2}\log\frac{\cosh(\beta J)+m_i^c\sinh(\beta J)}{\cosh(\beta J)-m_i^c\sinh(\beta J)} \\ &= \beta h_0 + \sum_i \frac{1}{2}\log\frac{1+m_i^c\tanh(\beta J)}{1-m_i^c\tanh(\beta J)} \\ &= \beta h_0 + \sum_i \mathrm{atanh}\left(m_i^c\tanh(\beta J)\right) . \end{aligned}$$

We can thus finally write the magnetization of the new spins as the function of the spins in the "old" system in a relatively simple form:

$$m_0 = \tanh\left(\beta h_0 + \sum_i \mathrm{atanh}\left(m_i^c\tanh(\beta J)\right)\right) \tag{4.1}$$

4.2 Exact recursion on a tree

It is quite simple to repeat the same argument iteratively on a tree. If we start with initial conditions then we can write, at any layer of the tree,

$$m_{i\to j} = \tanh\left(\beta h_i + \sum_{k\in\partial_i\neq j} \mathrm{atanh}\left(m_{k\to i}\tanh(\beta J)\right)\right) \tag{4.2}$$

What if we want to know the true marginals? This is easy, we just write

$$m_i = \tanh\left(\beta h_i + \sum_{k\in\partial_i} \mathrm{atanh}\left(m_{k\to i}\tanh(\beta J)\right)\right) \tag{4.3}$$

This is the root of the so-called Belief propagation approach. We solve the problem for the cavity marginals $m_{i\to j}$, which has a convenient interpretation as a message passing problem. Once we know them all, we can compute the true marginal!

This is a method that first appeared in statistical physics, when Bethe and Peierls used it as an approximation of the regular lattice. Indeed, we could iterate this method on a large infinite tree of connectivity, say $c = 2d$ to approximate a hypercubic lattice of dimension d (where each node has $2d$ neighbors). Consider for instance the situation with zero field. In this case, the magnetization at distance ℓ from the leaves follows

$$m^{\ell+1} = \tanh\left((c-1)\text{atanh}\left(m^{\ell}\tanh(\beta J)\right)\right)$$

We can look for a fixed point of this equation, and check for which values of the (inverse) temperature a non-zero value for the magnetization is possible. This is the same phenomenon as in the Ising model, just with a slightly more complicated fixed point. By ploting this equation, one realizes that, assuming $d = 2c$, this happens at $\beta^{\text{BP}} = \infty$ for $d = 1$, $\beta^{\text{BP}} = 0.346$ for $d = 2$, $\beta^{\text{BP}} = 0.203$ for $d = 3$, $\beta^{\text{BP}} = 0.144$ for $d = 4$, and $\beta^{\text{BP}} = 0.112$ for $d = 5$. If we compare these numbers to the actual transition on a real hypercubic lattice, we find $\beta^{\text{lattice}} = \infty$ for $d = 1$, $\beta^{\text{lattice}} = 0.44$ for $d = 2$, $\beta^{\text{lattice}} = 0.221$ for $d = 3$, $\beta^{\text{lattice}} = 0.149$ for $d = 4$ and $\beta^{\text{lattice}} = 0.114$ for $d = 5$. Not so bad, and in fact we see that the predictions become exact as d grows! This approach is quite a good one to estimate a critical temperature. In fact, one can show that it gives a rigorous upper bound on the ferromagnetic transition, in any topology!

4.2.1 Belief propagation on trees for pairwise models

The iterative approach we just discussed can be made completely generic on a tree graph! The example that we have been considering so far reads, in full generality

$$\mathcal{H} = -\sum_{(ij)\in\mathcal{G}} J_{ij}S_iS_j - \sum_i h_iS_i\,.$$

This is an instance of a very generic type of model: those with pairwise interactions, where the probability of each configuration is given by

$$\begin{aligned} P\left((S)\right) &= \frac{1}{Z}\left[\prod_i \psi_i(S_i)\prod_{(ij)\in\mathcal{G}}\psi_{ij}(S_i,S_j)\right] \\ Z &= \sum_{\{S_{i=1,\ldots,N}\}}\left[\prod_i \psi_i(S_i)\prod_{(ij)\in\mathcal{G}}\psi_{ij}(S_i,S_j)\right] \end{aligned}$$

and the connection is clear once we define $\psi_{ij}(S_i,S_j) = \exp(\beta J_{ij}S_iS_j)$ and $\psi_i(S_i) = \exp(\beta h_iS_i)$.

We want to compute Z on a tree, like we did for the RFIM. For two adjacent sites i and j, the trick is to consider the variable $Z_{i\to j}(S_i)$, defined as the *partial* partition function for the sub-tree rooted at i, when excluding the branch directed towards j, with a fixed value S_i of the i spin variable. We also need to introduce $Z_i(S_i)$, the partition function of the entire complete tree when, again, the variable i is fixed to a value S_i. On a tree, these intermediate variables can be computed exactly according to the following recursions

$$Z_{i\to j}(S_i) = \psi_i(S_i)\prod_{k\in\partial i\setminus j}\left(\sum_{S_k} Z_{k\to i}(S_k)\psi_{ik}(S_i,S_k)\right) \tag{4.4}$$

$$Z_i(S_i) = \psi_i(S_i)\prod_{j\in\partial i}\left(\sum_{S_j} Z_{j\to i}(S_j)\psi_{ij}(S_i,S_j)\right) \tag{4.5}$$

where ∂i denotes the set of all neighbors of i. In order to write these equations, the only assumption that has been made was that, for all $k \neq k' \in \partial i \setminus j$, the messages $Z_{k\to i}(S_k)$ and $Z_{k'\to i}(S_{k'})$ are independent. On a tree, this is obviously true: since there are no loops, the sites k and k' are connected only through i and we have "cut" this interaction when considering the partial quantities. This recursion is very similar, in spirit, to the standard transfer matrix method for a one-dimensional chain.

In practice, however, it turns out that working with partition functions (that is, numbers that can be exponentially large in the system size) is somehow impractical. We can thus normalize equation 4.4 and rewrite these recursions in terms of probabilities. Denoting $\eta_{i\to j}(S_i)$ as the marginal probability distribution of the variable S_i when the edge (ij) has been removed, we have

$$\eta_{i\to j}(S_i) = \frac{Z_{i\to j}(S_i)}{\sum_{S_i'} Z_{i\to j}(S_i')}, \qquad \eta_i(S_i) = \frac{Z_i(S_i)}{\sum_{S_i'} Z_i(S_i')}.$$

So that the recursions equation 4.4 and equation 4.5 now read

$$\eta_{i\to j}(S_i) = \frac{\psi_i(S_i)}{z_{i\to j}} \prod_{k\in\partial i\setminus j} \left(\sum_{S_k} \eta_{k\to i}(S_k)\psi_{ik}(S_i,S_k) \right), \tag{4.6}$$

$$\eta_i(S_i) = \frac{\psi_i(S_i)}{z_i} \prod_{j\in\partial i} \left(\sum_{S_j} \eta_{j\to i}(S_j)\psi_{ij}(S_i,S_j) \right), \tag{4.7}$$

where the $z_{i\to j}$ and z_i are normalization constants defined by:

$$z_{i\to j} = \sum_{S_i} \psi_i(S_i) \prod_{k\in\partial i\setminus j} \left(\sum_{S_k} \eta_{k\to i}(S_k)\psi_{ik}(S_i,S_k) \right), \tag{4.8}$$

$$z_i = \sum_{S_i} \psi_i(S_i) \prod_{j\in\partial i} \left(\sum_{S_j} \eta_{j\to i}(S_j)\psi_{ij}(S_i,S_j) \right). \tag{4.9}$$

The iterative equations equation 4.6 and equation 4.7), along with their normalization equation 4.8 and equation 4.9, are called the belief propagation equations. Indeed, since $\eta_{i\to j}(S_i)$ is the distribution of the variable S_i when the edge to variable j is absent, it is convenient to interpret it as the "belief" of the probability of S_i in absence of j. It is also called a "cavity" probability since it is derived by removing one node from the graph. The belief propagation equations are used to define the belief propagation algorithm

1. Initialize the cavity messages (or beliefs) $\eta_{i\to j}(S_i)$ randomly or following a prior information $\psi_i(S_i)$ if we have one.

2. Update the messages in a random order following the belief propagation recursion equation 4.6 and equation 4.7 until their convergence to their fixed point.

3. After convergence, use the beliefs to compute the complete marginal probability distribution $\eta_i(S_i)$ for each variable. This is the belief propagation estimate on the marginal probability distribution for variable i.

Using the resulting marginal distributions, one can compute, for instance, the equilibrium local magnetization via $m_i = \langle S_i \rangle = \sum_{S_i} \eta_i(S_i) S_i$, or basically any other local quantity of interest.

At this point, since we have switched from partial partition sums to partial marginals, the astute reader could complain that we have lost sight of our prime objective: the computation of the partition function. Fortunately, one can compute it from the knowledge of the marginal distributions. To do so, it is first useful to define the following quantity for every edge (ij):

$$z_{ij} = \sum_{S_i, S_j} \eta_{j\to i}(S_j)\eta_{i\to j}(S_i)\psi_{ij}(S_i, S_j) = \frac{z_j}{z_{j\to i}} = \frac{z_i}{z_{i\to j}} \,,$$

where the last two equalities are obtained by plugging equation 4.6 into the first equality and realizing that it almost gives equation 4.9. Using again equation 4.6 and equation 4.9, we obtain

$$\begin{aligned} z_i &= \sum_{S_i} \psi_i(S_i) \prod_{j\in\partial i} \left(\sum_{S_j} \eta_{j\to i}(S_j)\psi_{ij}(S_i, S_j) \right) \\ &= \sum_{S_i} \psi_i(S_i) \prod_{j\in\partial i} \left(\sum_{S_j} \frac{Z_{j\to i}(S_j)}{\sum_{S'} Z_{j\to i}(S')} \psi_{ij}(S_i, S_j) \right) = \frac{\sum_{S_i} Z_i(S_i)}{\prod_{j\in\partial i} \sum_{S_j} Z_{j\to i}(S_j)} \,, \end{aligned}$$

and along the same steps

$$z_{j\to i} = \frac{\sum_{S_j} Z_{j\to i}(S_j)}{\prod_{k\in\partial j\setminus i} \sum_{S_k} Z_{k\to j}(S_k)} \,.$$

For any spin S_i, the total partition function can be obtained using $Z = \sum_{S_i} Z_i(S_i)$. We can thus start from an arbitrary spin i

$$Z = \sum_{S_i} Z_i(S_i) = z_i \prod_{j\in\partial i} \left(\sum_{S_j} Z_{j\to i}(S_j) \right) = z_i \prod_{j\in\partial i} \left(z_{j\to i} \prod_{k\in\partial j\setminus i} \sum_{S_k} Z_{k\to j}(S_k) \right) \,,$$

and we continue to iterate this relation until we reach the leaves of the tree. Using eq4.2.1, we obtain

$$Z = z_i \prod_{j\in\partial i} \left(z_{j\to i} \prod_{k\in\partial j\setminus i} z_{k\to j} \cdots \right) = z_i \prod_{j\in\partial i} \left(\frac{z_j}{z_{ij}} \prod_{k\in\partial j\setminus i} \frac{z_k}{z_{jk}} \cdots \right) = \frac{\prod_i z_i}{\prod_{(ij)} z_{ij}} \,.$$

We thus obtain the expression of the free energy in a convenient form, that can be computed directly from the knowledge of the cavity messages, often called the *Bethe free energy* on a tree:

$$\begin{aligned} f_{\text{Tree}} N &= -T \log Z = \sum_i f_i - \sum_{(ij)} f_{ij} \,, \\ f_i &= -T \log z_i \,, \quad f_{ij} = -T \log z_{ij} \,, \end{aligned}$$

where f_i is a "site term" coming from the normalization of the marginal distribution of site i, and is related to the change in Z when the site i (and the corresponding edges) is added to the system. Meanwhile, f_{ij} is an "edge" term that can be interpreted as the change in Z when

the edge (ij) is added. This provides a convenient interpretation of the Bethe free energy equation 4.2.1: it is the sum of the free energy f_i for all sites but, since we have counted each edge twice we correct this by subtracting f_{ij}.

We have now entirely solved the problem on a tree. There is, however, nothing that prevents us from applying the same strategy on any graph. Indeed the algorithm we have described is well defined on any graph, but we are not assured that it gives exact results nor that it will converge. Using these equations on graphs with loops is sometimes referred to as *loopy belief propagation* in Bayesian inference literature.

One may wonder if there is a connection between the BP approach and the variational one. We may even wonder if this could be simply the same as using our variational bound with a better approach than the naive mean field one! Sadly, the answer is no! We cannot prove in general that the BP free entropy is a lower bound on any graph: indeed there are examples where it is larger than $\log Z$, and some where it is lower.

Two important remarks:

- There *is* however a connection between the variational approach and the BP one. If one writes the variational approach and uses the following parametrization for the guess (where $c_i = |\partial i|$):
$$\mathcal{Q}(S) = \frac{\prod_{ij} b_{ij}(S_i, S_j)}{\prod_i b_i(S)^{c_i - 1}}$$
then it is possible to show that optimizing on the function b_{ij} and b_i one finds the BP free entropy. The sad news, however, is that $\mathcal{Q}(S)$ does not really correspond to a true probability density, it is not always normalizable, so one cannot apply the variational bound.

- In the case of ferromagnetic models, or more exactly, on attractive potential Ψ_{ij}, and only in this case, it can be shown rigorously that BP *does* give a lower bound on the free entropy, on any graph. For such models (thus including the RFIM!) it is thus effectively equivalent to a variational approach. In fact, it can be further shown that in the limit of zero temperature, BP finds the ground state of the RFIM on any graph (through a mapping to linear programming).

4.3 Cavity on random graphs

4.3.1 Random graphs

We shall now discuss the basic properties of sparse *Erdős-Rényi* (ER) random graphs.

An ER random graph is taken uniformly at random from the ensemble, denoted $\mathcal{G}(N, M)$, of graphs that have N vertices and M edges. To create such a graph, one has simply to add M random edges to an empty graph. Alternatively, one can also define the so called $\mathcal{G}(N, p)$ ensemble, where an edge exists independently for each pair of nodes with a given probability $0 < c/N < 1$. The two ensembles are asymptotically equivalent in the large N limit, when $M = c(N-1)/2$. The constant c is called the average degree. We denote by c_i the degree

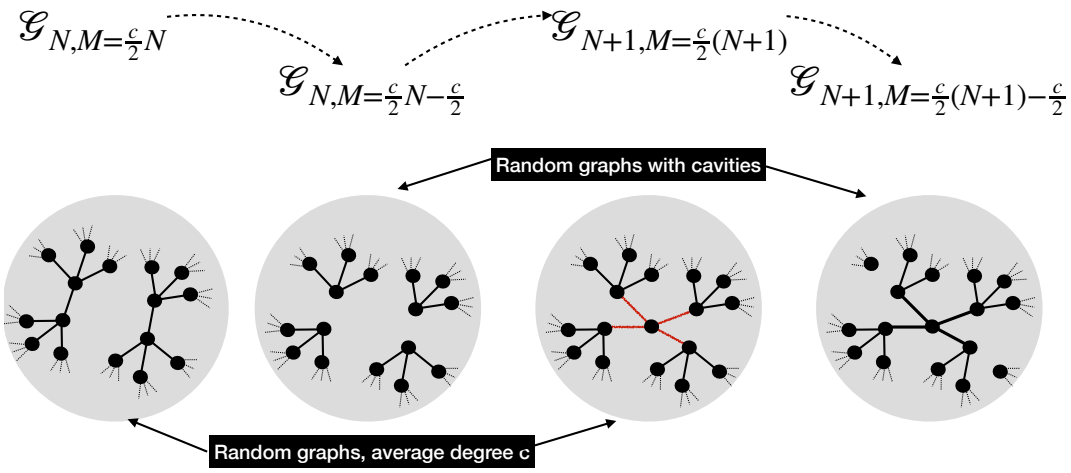


Figure 4.3.1: Iterative construction of a random graph for the Cavity method. The average degree of the graph is c.

of a node i, i.e. the number of nodes to which i is connected. The degrees are distributed according to Poisson distribution, with average c.

Alternatively, one can also construct the so-called *regular* random graphs from the ensemble $\mathcal{R}(N, c)$ with N vertices but where the degree of each vertex is fixed to be exactly c. This means that the number of edges is also fixed to $M = cN/2$.

At the core of the cavity method is the fact that such random graphs locally look like trees, i.e. there are no short cycles going trough a typical node. The key point is thus that, in this limit, such random graphs can be considered locally as trees. The intuitive argument for this result is the following one: starting from a random site, and moving following the edges, in ℓ steps c^ℓ sites will be reached. In order to have a loop, we thus need $c^\ell \sim N$ to be able to come back on the initial site, and this gives $\ell \sim \log(N)$.

4.3.2 Cavity method

Let us now apply our beloved cavity method on an ER random graph with N links and on average $M = cN/2$ links. We will see how this method leads to free entropies using the same telescopic sum trick as before. However, this should be done with caution. If we apply our usual Cesaro trick naively and write

$$Z_{N,M} = \frac{Z_{N,M}}{Z_{N-1,M-m_1}} \frac{Z_{N-1,M-m_1}}{Z_{N-2,M-m_1-m_2}} \dots Z_{1,0}$$

then we encounter a problem when choosing the value of m at each step. We want the new spin to have on average c neighbors. That means we must add a Poissonian variable m with mean c at each step, so that the new spin has the correct numbers of neighbors. But this also adds a link to c spins in the previous graph, so that, on average, we went from $M = cN/2$ to $M' = cN/2 + c$ while we actually wanted to get $M' = c(N+1)/2$. The difference is $\Delta M = c/2$, so we need to construct our sequence of graphs such that we remove $m/2$ links on average

every time we add one spin connected to m previous spins. Therefore, we write instead

$$Z_{N,M} = \frac{Z_{N,M}}{Z_{N-1,M-m_1}} \frac{Z_{N-1,M-m_1}}{Z_{N-1,M-m_1+m_2/2}} \frac{Z_{N-1,M-m_1+m_2/2}}{Z_{N-2,M-m_1+m_1'/2-m_2}} \ldots\ldots Z_{1,0}$$

Concretely, this means that we must apply the Cesaro theorem to the following term to compute the free entropy

$$A_N = \log \frac{Z_{N,M=N\frac{c}{2}}}{Z_{N-1,M-c}} \frac{Z_{N-1,M-c}}{Z_{N-1,M-c+\frac{c}{2}}} = \log \frac{Z_{N,M=N\frac{c}{2}}}{Z_{N-1,M-c}} - \log \frac{Z_{N-1,M-c+\frac{c}{2}}}{Z_{N-1,M-c}} .$$

And we thus obtain

$$\lim_{N,M\to\infty} \mathbb{E}\frac{1}{N} \log Z_N = \lim_{N\to\infty} \mathbb{E} A_N = \lim_{N\to\infty} \mathbb{E}\Phi_N^{(\text{site})} - \lim_{N\to\infty} \mathbb{E}\Phi_N^{(\text{link})}$$

We thus have two terms to compute to compute the free entropy, that corresponds to the 2-step iteration on the graph depicted in Fig.4.3.1

1. The first one corresponds to the change in free entropy when *one adds one spin to a graph, connecting it to c spins with c cavities*.
2. The second one corresponds to the change in free entropy when *one adds $c/2$ links to a graph, connecting c spins with cavities pairwise*.

This corresponds exactly to what we found on the tree! Indeed, the in 4.2.1, we see that we have N sites terms, minus $M = cN/2$ links terms! It is reassuring to find that this construction gives us the same answer as on the tree.

Mathematically, these two free entropy shift can be expressed as

$$\begin{aligned}
\Phi_N^{(\text{site})} &= \mathbb{E}\log\langle \sum_{S_0} e^{\beta h_0 S_0 + \beta J \sum_{i\in\partial 0} S_0 S_i} \rangle_{N-1,M-c} \\
&= \mathbb{E}\log\langle 2\cosh\left(\beta h_0 + \beta J \sum_{i\in\partial 0} S_i\right)\rangle_{N-1,M-c} \\
\Phi_N^{(\text{link})} &= \mathbb{E}\log\langle \prod_{(ij)} e^{\beta J J S_i S_j} \rangle_{N-1,M-c} = \frac{c}{2}\mathbb{E}\log\langle e^{\beta J J S_i S_j}\rangle_{N-1,M-1} .
\end{aligned}$$

Interestingly, both these equations depend on the distribution of the *joint cavity magnetizations* in the graph so we can write

$$\begin{aligned}
\Phi^{(\text{site})} &= \mathbb{E}\int \mathrm{d}P_{\mathrm{e}}(d) \int \prod_{i=1}^{d} \mathrm{d}m_i Q^c(\{m_i\}) \log\langle 2\cosh\left(\beta h_0 + \beta J \sum_{i\in\partial 0} S_i\right)\rangle_{\{m_i\}} \\
\Phi^{(\text{link})} &= \mathbb{E}\frac{c}{2}\int \mathrm{d}m_1 Q(m_1,m_2) \log\langle e^{\beta J J S_i S_j}\rangle_{\{m_1,m_2\}}
\end{aligned}$$

where $P_{\mathrm{e}}(d)$ is the excess degree probability. For a regular graph with fixed connectivity c, $P_{\mathrm{e}}(d) = \delta(d-(c-1))$ while for a Erdos-renyi random graph, interestingly $P_{\mathrm{e}}(d) = P(d)$ again! See exercises section.

At the level of rigor used in physics, these formulas can be further simplified! First, we make the assumption that the distribution of cavity fields converges to a limit distribution, independently of the disorder. Secondly, the crucial point is now to *assume the distribution of these cavity marginals factorizes*! This makes sense: the different cavities $\{m_i\}$ are all far from each other in the graph. If we are not *exactly* at a critical point (a phase transition) then the correlations are not infinite range. This is case, our formula depends only on the single point distribution $Q^{\mathrm{c}}(m)$ and we thus write:

$$\begin{aligned}\Phi^{(\mathrm{site})} &= \int \mathrm{d}P_{\mathrm{e}}(d) \int \prod_{i=1}^{d} \int \mathrm{d}m_i Q^{c}(m_i) \log\langle 2\cosh\left(\beta h_0 + \beta J \sum_{i\in\partial 0} S_i\right)\rangle_{\{m_i\}} \\ \Phi^{(\mathrm{link})} &= \frac{c}{2} \int \mathrm{d}m_1 Q(m_1)\, \mathrm{d}m_2 Q(m_2) \log\langle e^{\beta J J S_i S_j}\rangle_{\{m_1,m_2\}}\end{aligned}$$

Our task is thus to find the asymptotic distribution $Q(m)$. At the level of rigor used in physics, this is easily done. We obviously assume that $Q(m)$ is unique, and does not depend on the realization of the disorder (this is not so trivial). Then we realize that the distribution of cavity fields must satisfy a recursion such as

$$Q^{\mathrm{c}}_{N+1}(m) = \int \mathrm{d}h_0 P(h_0) \int \mathrm{d}P_{\mathrm{e}}(d) \prod_{i=1}^{d} \int \mathrm{d}m_i Q^{\mathrm{c}}_{N+1}(m_i) \delta\left(m - f_{\mathrm{BP}}(\{m_i\}, h_0)\right)$$

with

$$f_{\mathrm{BP}}(\{m_i\}, h_0) = \tanh\left(\beta h_0 + \sum_i \operatorname{atanh}\left(m_i \tanh(\beta J)\right)\right)$$

Obviously, once we find the fixed point, we can compute the distribution to total magnetization, which reads almost exactly the same, except now we have to use the actual distribution of neighbors:

$$Q(m) = \int \mathrm{d}h_0 P(h_0) \int \mathrm{d}P_{\mathrm{e}}(d) \prod_{i=1}^{d} \int \mathrm{d}m_i Q^{\mathrm{c}}(m_i) \delta\left(m - f_{\mathrm{BP}}(\{m_i\}, h_0)\right)$$

1. Explain the population dynamic
2. Note that the free energy is the same as the one on graphs!!! Dictionary GRAPH to POPULATION
3. Add discussion regular vs random, and excess degree

4.3.3 The relation between Loopy Belief Propagation and the Cavity method

- tree-like etc....
- Single graph vs POPULATION !

4.3.4 Can we prove it?

Well, we can try !!! First $Q^c(m)$ is not clearly self-averaging, but for sure

$$\Phi \leq \max_{Q^c(m)} \int \mathrm{d}P_{\mathrm{e}}(d) \prod_{i=1}^{d} \int \mathrm{d}m_i Q^c(\{m_i\}) \log\langle 2\cosh\left(\beta h_0 + \beta J \sum_{i\in\partial 0} S_i\right)\rangle_{\{m\}} - \frac{c}{2} \int \mathrm{d}m_1 Q^c(\{m_1\})\, \mathrm{d}m_2 Q^c(\{m_2\}) \log\langle e^{\beta J J S_i S_j}\rangle_{m_1,m_2}$$

It is possible to show that the extremization leads to $Q(m)$ being a solution of the cavity recursion. This means that we obtain a bound! If we found a distribution $Q(m)$ (or actually, all of them) that satisfies eq., then we have a bound.

Can we get the converse bound easily? Sadly, no. The point is that BP is *not* a variational method on a given instance, so we cannot use the mean-field technics! Fortunately, it can be shown rigorously, that, for any ferromagnetic model, BP *does* give a lower bound on the free entropy.

It is also instructive to compare to what we had in the fully connected model. Indeed, if $Q(m)$ become a delta (which we expect as c grows) we obtain, using $J = 1/N$

$$\Phi \leq \max_m \mathbb{E}_h \log 2\cosh\left(\beta h + \beta m\right) - \frac{N}{2}\log e^{\beta m^2/N} = \mathbb{E}_h \log 2\cosh\left(\beta h + \beta m\right) - \beta\frac{m^2}{2}$$

which is indeed the result we had in the fully connected limit!

4.3.5 When do we expect this to work?

When do we expect belief propagation to be correct? As we have have discussed, random graphs are locally tree-like: they are trees up to any finite distance. Further assuming that we are in a pure thermodynamic state, we expect that we have short range correlations, so that the large $O(\log N)$ loops should not matter in a large enough system, and these equations should provide a correct description of the model.

Clearly, this must be a good approach for describing a system in a paramagnetic phase, or even a system with a ferromagnetic transition (where we should expect to have two different fixed points of the iterations). It could be, however, that there exists a huge number of fixed points for these equations: how to deal with this situation? Should they all correspond to a given pure state? Fortunately, we do not have such worries, as the situation we just described is the one arising when there is a glass transition. In this case, one needs to use the cavity method in conjunction with the so-called "replica symmetry breaking" approach as was done by Mézard, Parisi, and Virasoro.

Bibliography

The cavity method is motivated by the original ideas from Bethe (1935) and Peierls (1936) and used in detail to study ferromagnetism (Weiss, 1948). Belief propagation first appeared in

computer science in the context of Shanon error correction (Gallager, 1962) and was rediscovered in many different contexts. The name "Belief Propagation" comes in particular from Pearl (1982). The deep relation between loopy belief propagation and the Bethe "cavity" approach was discussed in the early 2000s, for instance in Opper and Saad (2001) and Wainwright and Jordan (2008). The work by Yedidia et al. (2003) was particularly influential. The construction of the cavity method on random graphs presented in this chapter follows the classical papers Mézard and Parisi (2001, 2003). That loopy belief propagation gives a lower bound on the true partition *on any graph* in the case of ferromagnetic (and in general attractive models) is a deep non-trivial result proven (partially) by Willsky et al. (2007) using the loop calculus of Chertkov and Chernyak (2006), and (fully) by Ruozzi (2012). Chertkov (2008) showed how beleif propagation finds the ground state of the RFIM at zero temperature. Finally, that the cavity method gives rigorous upper bounds on the critical temperature was shown in Saade et al. (2017).

4.4 Exercises

Exercise 4.1: Excess degree in Erdos-Renyi graphs

Consider a Erdos-Renyi random graph with N nodes and M links in the asymptotic regime where $N \to \infty$, with $c = 2M/N$ the average degree.

1. Consider one given node, what is the probability p that it is connected with another given node, say j? Since it has $N-1$ potential neighbors, show that the probability distribution of the number of neighbors for each node follows

$$\mathbb{P}(d = k) = \frac{c^k e^{-c}}{k!} \tag{4.10}$$

In the cavity method, we are often interested as well in the excess degree distribution, that is, given one site i that has a neighbor j, what is its distribution of *additional* neighbors d?

2 Argue that finding first a link (ij) and then looking to i is equivalent to sampling all nodes with a probability that is proportional to their number of neighbors:

$$\mathbb{P}_i = \frac{d_i}{c} \tag{4.11}$$

2 Finally show that the probability distribution of having $k+1$ neighbors when one chose each sites with probability $\frac{d_i}{c}$ is

$$\mathbb{P}(d = k+1) = \frac{c^{k+1} e^{-c}}{k+1!} \frac{k+1}{c} \tag{4.12}$$

so that the probability distribution of excess degree reads

$$\mathbb{P}^{\mathrm{e}}(d = k) = \frac{c^k e^{-c}}{k!} \tag{4.13}$$

Exercise 4.2: The random field ising model on a regular random graph

We have seen that the BP update equation for the RFIM is given by

$$f_{\rm BP}(\{m_i\}, h_0) = \tanh\left(\beta h_0 + \sum_i \operatorname{atanh}\left(m_i \tanh(\beta J)\right)\right) \tag{4.14}$$

and that the distribution of cavity fields follows (for a random graph with fixed connectivity $c-1$):

$$Q^{\text{cav.}\cdot}(m) = \int dh_0 \mathcal{N}(h_0; 0, \Delta) \prod_{i=1}^{c-1} \int dm_i Q^{\rm cav}(m_i) \delta\left(m - f_{\rm BP}(\{m_i\}, h_0)\right) \tag{4.15}$$

This can be solved in practice using the population dynamics approach where we represent $Q^{\text{cav.}\cdot}(m)$ by a population of $\mathcal{N}_{\rm pop}$ elements. In this case, formally we iterate a collection, or a pool, of elements. Starting from $Q^{\rm cav.}_{t=0}(m) = \left[m_1, m_2, m_3, \ldots, m_{\mathcal{N}_{\rm pop}}\right]$ with, for instance all $m = 1$ or random initial conditions, we iterate as follows:

- For T steps:
- For all $i = 1 \to \mathcal{N}_{\rm pop}$:
- Draw a random $h_0 \sim \mathcal{N}(0, \Delta)$, and $c-1$ random $\{m_i\}$ from $Q^{\text{cav.}\cdot}_t$.
- Compute $m^{\rm new} = f_{\rm BP}(\{m_i\}, h_0)$.
- Assign m to the i elements of the new population $Q_{t+1_i} = m^{\rm new}$.

If $\mathcal{N}_{\rm pop}$ is large enough (say 10^5) then this is a good approximation of the true population density, and if T is large enough, then we should have converged to the fixed point. Once this is done, we can compute the average magnetization by computing the true marginal as follows:

- Set $\overline{m} = 0$
- For $\mathcal{N}$ steps:
- Draw c random $\{m_i\}$ from $Q^{\rm cav.}_{\rm eq}$.
- Compute $m^{\rm new} = f_{\rm BP}(\{m_i\}, h_0)$ using this time the c values.
- $\overline{m} = \overline{m} + m^{\rm new}/\mathcal{N}_{\rm pop}$

1. Consider the RFIM on regular random graphs with connectivity $c = 4$. Compute analytically the phase transition point in $beta$, denoted β_c at zero disordered fields ($\Delta = 0$).
2. Implement the population dynamics and find numerically the phase transition point when $\overline{m}(\beta, \Delta)$ become non zero. Draw the phase diagram in (β, Δ) separating the phase where $\overline{m} = 0$ with the one where $\overline{m} \neq 0$.

3. Now, let us specialize to the low (and eventually zero) temperature limit. Using the change of variable $m_i = \tanh(\beta \tilde{h}_i)$, show that the iteration has the following limit when $\beta \to \infty$

$$\tilde{h}^{\text{new}} = \lim_{\beta \to \infty} \frac{1}{\beta} \text{atanh} f_{\text{BP}}(\{m_i\}, h_0) = h_0 + \sum_i \phi(\tilde{h}_i) \tag{4.16}$$

with

$$\phi(x) = \begin{cases} x, \text{ if } |\text{x}| < 1 \\ \text{sign(x) if } |\text{x}| > 1 \end{cases} \tag{4.17}$$

4. Using this equation, perform the population dynamics at zero temperature and compute the critical value of Δ where a non zero magnetization appears.

Chapter 5

Sparse Graphs & Locally Tree-like Graphical Models

Auprès de mon arbre je vivais heureux
J'aurais jamais dû m'éloigner de mon arbre

Auprès de mon arbre
Georges Brassens – 1955

A common theme in the previous chapters has been the study of the Boltzmann-Gibbs distribution:

$$\mathbb{P}_{N,\beta}\left(\boldsymbol{S}=\boldsymbol{s}\right)=\frac{e^{-\beta\mathcal{H}(\boldsymbol{s})}}{Z_N(\beta)}.$$

This is a joint probability distribution over the random variables $S_1,\dots,S_N$ defined through the energy function $\mathcal{H}$. In the two examples we have seen so far, the energy function is composed of two pieces: an *interaction term* that couples different random variables and a *potential term* which acts on each random variable separately. For instance, for the Curie-Weiss model:

$$\mathcal{H}_{N,h}(\boldsymbol{s})=\underbrace{-\frac{1}{2N}\sum_{i,j=1}^{N}s_is_j}_{\text{interaction}}\underbrace{-h\sum_{i=1}^{N}s_i}_{\text{potential}}$$

Notice that it is the interaction term that correlates the random variables: if it was zero, the Gibbs-Boltzmann distribution would factorize and we would be able to fully characterize the system by studying each variable independently. It is the interaction term that makes the problem truly multi-dimensional.

In the Curie-Weiss model and RFIM, the interaction term is quadratic: it couples the random variables pairwise. In the Chapters that follow, we will study many other examples of Gibbs-Boltzmann distributions, each defined by different flavours of variables, potentials and interactions terms. Therefore, it will be useful to introduce a very general way to think about and represent multi-dimensional probability distributions. This is the subject of this Chapter.

5.1 Graphical Models

To proceed with the study of probabilistic models, we introduce a tool called *Graphical Models* that will give us a neat and very generic way to think about a broad range of probability distributions. A Graphical Model is a way to represent relations or correlations between variables.

In this section we give basic definitions and introduce a couple of examples that will be studied in more detail later in the class.

5.1.1 Graphs

An undirected graph $G(V, E)$ is defined by:

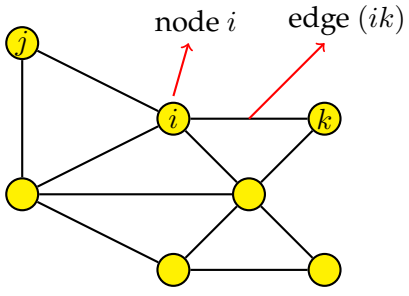
- A set of nodes V, which we will index by $i \in V$.
- A set of edges E, which we will index by a pair of nodes $(ij) \in E$.

Let $|S|$ denote the size of set S. For the purpose of this section we denote the total number of nodes by $|V| = N$ and the total number of edges by $|E| = M$. The *adjacency matrix* of the graph $G(V, E)$ is a symmetric $N \times N$ binary matrix $A \in \{0, 1\}^{N \times N}$ with entries:

$$A_{ij} =: \begin{cases} 1 & \text{if } (ij) \in E \\ 0 & \text{if } (ij) \notin E \end{cases}$$

We will denote as ∂ the neighborhood operator, i.e. $\partial i := \{j \in V \,|\, (ij) \in E\}$ is the set of neighbors of node i. For a fixed node $i \in V$, define the *degree of node* i as $d_i = |\partial i|$, i.e. the total number of neighbors of i. Note that d_i can be easily obtained from the adjacency matrix A,

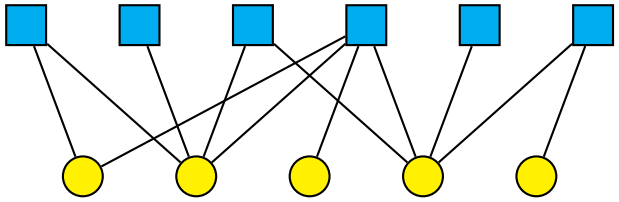
$$d_i = \sum_{j=1}^{N} A_{ij} = \sum_{j=1}^{N} \mathbb{1}_{(ij)\in E} \,.$$



5.1.2 Factor Graphs

- A factor graph is a graph with nodes of type 'circle' and of type 'square'

  - a 'circle' node is a *variable node*, indexed by $i, j, k, \ldots$
  - a 'square' node is *factor node*, indexed by $a, b, c, \ldots$
- A factor graph is bipartite: only edges of type 'circle'—'square' exist, no 'circle'—'circle' edges and no 'square'—'square' edges.



$$\begin{aligned} \partial i &:= \{a \,|\, (ia) \in E\}\,, \qquad |\partial i| = d_i, \qquad \forall\, i \in \{1, \ldots, N\} \\ \partial a &:= \{i \,|\, (ia) \in E\}\,, \qquad |\partial a| = d_a, \qquad \forall\, a \in \{1, \ldots, M\} \end{aligned}$$

- Every variable node i represents a *random variable* s_i taking values in Λ.
- Every factor node a represents a *non-negative function* $f_a\left(\{s_i\}_{i\in\partial a}\right)$.

A graphical model represents a joint probability distribution over the variables $\{s_i\}_{i=1}^N$:

$$P\left(\{s_i\}_{i=1}^N\right) =: \frac{1}{Z_N} \prod_{a=1}^{M} f_a\left(\{s_j\}_{j\in\partial a}\right),$$

where Z_N is the normalization constant

$$Z_N =: \sum_{\{s_i\}_{i=1}^N} \prod_{a=1}^{M} f_a\left(\{s_j\}_{j\in\partial a}\right).$$

In this lecture we will use graphical models extensively as a language to represent probability distributions arising in optimization, inference and learning problems. We will study a variety of f_a, Λ and graphical models. Let us start by giving several examples.

Example 1 (Physics, spin glass)

Consider a graph $G(V, E)$ as a graph of interactions, e.g. 3D cubic lattice with $N = 10^{23}$ nodes. The nodes may represent N Ising spins $s_i \in \Lambda = \{-1, +1\}$. In statistical physics, systems are often defined by their energy function, which we call Hamiltonian. One can think of the Hamiltonian as a simple cost function where lower values are better. The Hamiltonian of a spin glass then reads

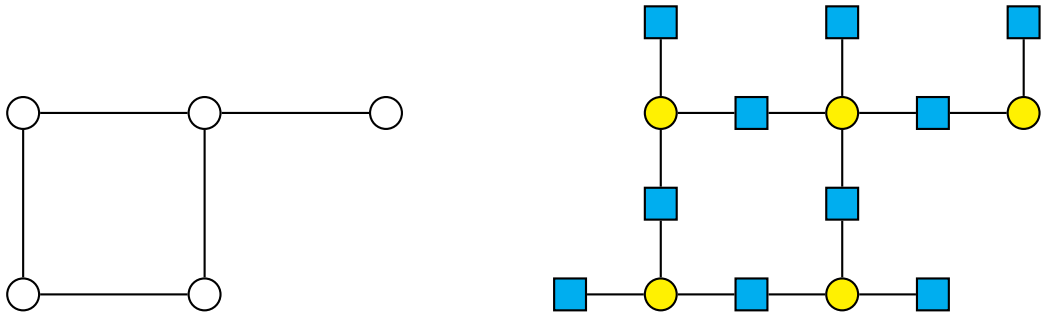
$$\mathcal{H}\left(\{s_i\}_{i=1}^N\right) = -\sum_{(ij)\in E} J_{ij} s_i s_j - \sum_i h_i s_i$$

where $J_{ij} \in \mathbb{R}$ are the interactions and $h_i \in \mathbb{R}$ are magnetic fields. Associated with the Hamiltonian, we consider the Boltzmann probability distribution defined as:

$$P\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z_N} e^{-\beta \mathcal{H}\left(\{s_i\}_{i=1}^N\right)} = \frac{1}{Z_N} \prod_{i=1}^N e^{\beta h_i s_i} \prod_{(ij)\in E} e^{\beta J_{ij} s_i s_j},$$

where the normalization constant $Z_N = \sum_{\{s_i\}_{i=1}^N} e^{-\beta \mathcal{H}\left(\{s_i\}_{i=1}^N\right)}$ is called the partition function.

The graphical model associated with a spin glass defined on a graph $G(V, E)$ (left) is drawn on the right.



The graphical model has:

- For each node i one factor node $f_i(s_i) = e^{\beta h_i s_i}$ corresponding to the magnetic field h_i.
- For each edge $(ij) \in E$ one factor node $f_{(ij)}(s_i, s_j) = e^{\beta J_{ij} s_i s_j}$ corresponding to the interaction J_{ij}.

Example 2 (Combinatorial optimization, graph coloring)

Consider a graph $G(V, E)$, and a set of q colors $s_i \in \{\text{red}, \text{blue}, \text{green}, \text{yellow}, \ldots, \text{black}\} = \{1, 2, \ldots, q\}$. In graph coloring we seek to assign a color to each node so that neighbors do not have the same color.

In the figure we show a proper 4-coloring of the corresponding (planar) graph. Note that the same graph can be also colored using only 3 colors, but not 2 colors.

Figure 5.1.1: Illustration on how coloring of maps corresponds to coloring of graphs.

To set up graph coloring in the language of graphical models, we write the number of proper colorings of the graph $G(V, E)$ as

$$Z_N = \sum_{\{s_i\}_{i=1}^N} \prod_{(ij)\in E} \left(1 - \delta_{s_i,s_j}\right)$$

The graph is colorable if and only if $Z_N \geq 1$, in that case we can also define a probability measure uniform over all proper colorings as

$$P\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z_N} \prod_{(ij)\in E} \left(1 - \delta_{s_i,s_j}\right)$$

We can also soften the constraint on colors and introduce a more general probability distribution:

$$P\left(\{s_i\}_{i=1}^N, \beta\right) = \frac{1}{Z_N(\beta)} \prod_{(ij)\in E} e^{-\beta\delta_{s_i,s_j}}$$

As $\beta \nearrow \infty$ we recover the case with strict constraints. The graphical model for graph coloring then corresponds to factor nodes on each edge (ij) of the form $f_{(ij)}(s_i, s_k) = e^{-\beta\delta_{s_i,s_j}}$.

Example 3 (Statistical inference, unsupervised learning)

Probability distributions that are readily represented via graphical models also naturally arise in statistical inference. We can give the example of the *Stochastic Block Model*, which is a commonly considered model for community detection in networks. In the SBM, N nodes are divided in q groups, the group of node i being written $s_i^* \in \{1, 2, \ldots, q\}$ for $i = 1, \ldots, N$. Node i is assigned in group $s_i^* = a$ with probability (fraction of expected group size) $n_a \geq 0$, where $\sum_{a=1}^q n_a = 1$. Pairs of nodes are then connected with probability that corresponds to their group memberships. Specifically:

$$\begin{cases} P\left((ij) \in E \,\middle|\, s_i^*, s_j^*\right) = p_{s_i^* s_j^*} \\ P\left((ij) \notin E \,\middle|\, s_i^*, s_j^*\right) = 1 - p_{s_i^* s_j^*} \end{cases} \quad \Rightarrow \quad G(V, E) \quad \& \quad A_{ij}$$

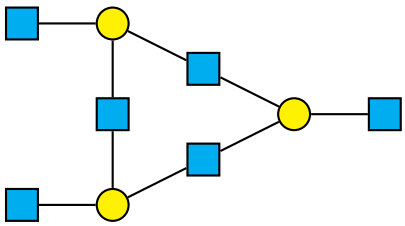
where p_{ab} forms a symmetric $q \times q$ matrix.

The question of community detection is whether, given the edges A_{ij} and the parameters $\boldsymbol{\theta} = (n_a, p_{ab}, q)$, one can retrieve s_i^* for all $i = 1, \ldots, N$. A less demanding challenge is to find an estimator $\hat{s}_i$ such that $\hat{s}_i = s_i^*$ for as many nodes i as possible.

We adopt the framework of Bayesian inference. All the information we have about s_i^* is included in the posterior probability distribution[1]:

$$\begin{aligned} P\left(\{s_i\}_{i=1}^N \,\middle|\, A, \boldsymbol{\theta}\right) &= \frac{1}{Z_N(A, \boldsymbol{\theta})} P\left(A \,\middle|\, \{s_i\}_{i=1}^N, \boldsymbol{\theta}\right) P\left(\{s_i\}_{i=1}^N \,\middle|\, \boldsymbol{\theta}\right) \\ &= \frac{1}{Z_N(A, \boldsymbol{\theta})} \prod_{i<j} \left[\left(1 - p_{s_i,s_j}^{1-A_{ij}}\right) p_{s_i,s_j}^{A_{ij}}\right] \prod_{i=1}^N n_{s_i} \end{aligned}$$

[1]Note that s_i in the posterior distribution is just a dummy variable, the argument of a function.



The corresponding graphical model has one factor node per variable (field), and one factor node per pair of nodes (interaction). Indeed, we notice that even pairs without edge (where $A_{ij} = 0$) have a factor node with $f_{ij}(s_i, s_j) = 1 - p_{s_i,s_j}$. In the class we will study this posterior quite extensively.

Example 4 (Generalized linear model)

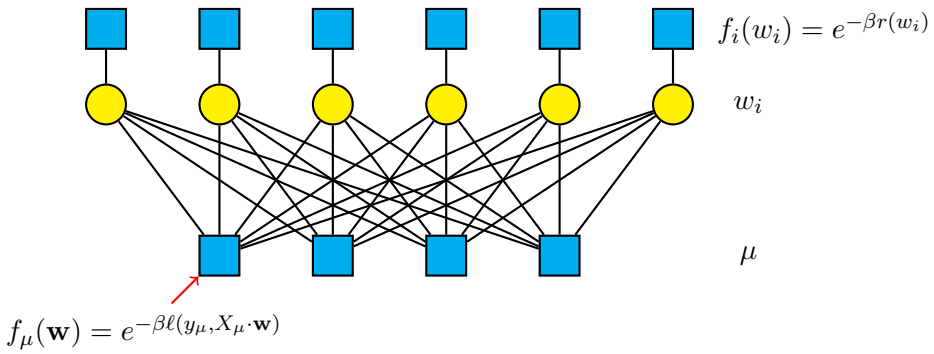
Let n be the number of data samples and d be the dimension of data. Samples are denoted $X_\mu \in \mathbb{R}^d$ and labels $y_\mu \in \{-1, +1\}$ (cats/dogs), for $\mu = 1, \dots, n$. Generalized linear regression is then formulated as the minimization of a loss function of the form

$$\mathcal{L}(\mathbf{w}) = \sum_{\mu=1}^{n} \ell(y_\mu, X_\mu \cdot \mathbf{w}) + \sum_{i=1}^{d} r(w_i)\,.$$

The loss minimization problem can then be regarded as the $\beta \to \infty$ limit of the following probability measure

$$P(\mathbf{w}) = \frac{1}{Z_N(\mathbf{X}, \mathbf{y}, \beta)} e^{-\beta \mathcal{L}(\mathbf{w})} = \frac{1}{Z_N(\mathbf{X}, \mathbf{y}, \beta)} \prod_{i=1}^{d} e^{-\beta r(w_i)} \prod_{\mu=1}^{n} e^{-\beta \ell(y_\mu, X_\mu \cdot \mathbf{w})}$$

The corresponding graphical model then looks as follows



5.1.3 Some properties and usage of factor graphs

As we saw in the above examples the factors are often of two types: (i) factor nodes that are related to only one variable — such as the magnetic field in the spin glass, the regularization

in the generalized linear regression, or the prior in the stochastic block model — and (ii) factor nodes that are related to interactions between the variables. For convenience, we will thus treat those two types separately and denote the type (i) as $g_i(s_i)$, and the type (ii) as $f_a(\{s_i\}_{i\in\partial a})$. Factor nodes of type (i) will be denoted with indices $i, j, k, l, \ldots$, factors of type (ii) will be denoted with indices $a, b, c, d, \ldots$. In this notation the probability distribution of interest becomes:

$$P\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z}\prod_{i=1}^N g_i(s_i)\prod_{a=1}^M f_a\left(\{s_i\}_{i\in\partial a}\right)$$

As we saw already and will see throughout the course, quantities of interest can be extracted from the value of the normalization constant, or partition function in physics jargon, that reads

$$Z = \sum_{\{s_i\}_{i=1}^N}\prod_{i=1}^N g_i(s_i)\prod_{a=1}^M f_a\left(\{s_i\}_{i\in\partial a}\right) .$$

We will in particular be interested in the value of the free entropy $N\Phi = \log Z$. Another quantity of interest is the marginal distribution for each variable, related to the local magnetization in physics, defined as

$$\mu_i(s_i) =: \sum_{\substack{\{s_j\}_{j=1}^N \\ j\neq i}} P\left(\{s_j\}_{j=1}^N\right)$$

The hurdle with computing the partition function and the marginals for large system sizes is that it entails evaluating sums over a number of terms that is exponential in N. From the computational complexity point of view, we do not know of exact polynomial algorithms able to compute these sums for a general graphical model. In the rest of the lecture, we will cover cases where the marginals and the free entropy can be computed exactly up to the leading order in the system size N.

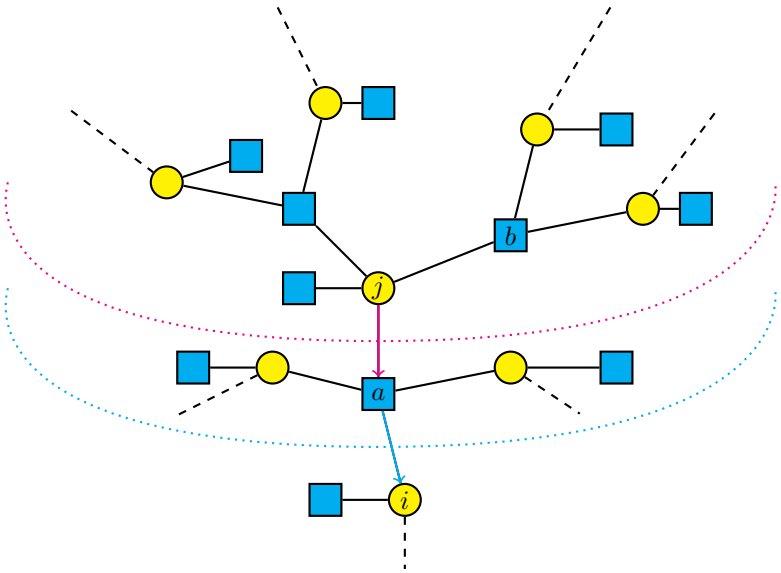
A special case of graphical models where the marginals and the free entropy can be computed exactly is that of tree graphical models, i.e. graphs that do not contain loops. How to approach this case is explained in the next section. Then, we will show that problems on graphs that locally look like trees, i.e. the shortest loop going trough a typical node is long, can be solved by carefully employing a tree-like approximation. This type of approximation holds for graphical models corresponding to random sparse graphs, i.e. those where the average degree is constant as the size N grows. We will see that also a range of problems defined on densely connected factor graphs can be solved exactly in the large size limit.

5.2 Belief Propagation and the Bethe free energy

5.2.1 Derivation of Belief Propagation on a tree graphical model

We will start by computing the partition function Z and marginals $\mu_i(s_i)$ for tree graphical models. We recall that the probability distribution under consideration is

$$P\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z}\prod_{i=1}^N g_i(s_i)\prod_{a=1}^M f_a\left(\{s_i\}_{i\in\partial a}\right)$$



In order to express the marginals and the partition function we define *auxiliary partition functions* for every $(ia) \in E$.

$$R^{j\to a}_{s_j} = g_j(s_j) \sum_{\{s_k\}_{\text{all } k \text{ above } j}} \prod_{\text{all } k \text{ above } j} g_k(s_k) \prod_{\text{all } b \text{ above } j} f_b\left(\{s_l\}_{l\in\partial b}\right)$$

$$V^{a\to i}_{s_i} = \sum_{\{s_j\}_{\text{all } j \text{ above } a}} f_a\left(\{s_k\}_{k\in\partial a}\right) \prod_{\text{all } j \text{ above } a} g_j(s_j) \prod_{\text{all } b \text{ above } a} f_b\left(\{s_k\}_{k\in\partial b}\right)$$

The meaning of these quantities can be understood from the figure above. $R^{j\to a}_{s_j}$ represents the partition function of the part of the system above the red dotted line, with variable node j restricted to taking value s_j. Analogously $V^{a\to i}_{s_i}$ is the partition function of the subsystem above the blue dotted line, with variable node i restricted to taking value s_i [2].

Since the graphical model is a tree, i.e. it has no loops, the restriction of the variable j to s_j makes the branches above j independent. We can thus split the sum over the variables, according to the branch to which they belong, to obtain

$$\begin{aligned} R^{j\to a}_{s_j} &= g_j(s_j) \prod_{b\in\partial j\setminus a} \underbrace{\left[\sum_{\{s_k\}_{\text{all } k \text{ above } b}} f_b\left(\{s_k\}_{k\in\partial b}\right) \prod_{\text{all } k \text{ above } b} g_k(s_k) \prod_{\text{all } c \text{ above } b} f_c\left(\{s_l\}_{l\in\partial c}\right)\right]}_{=V^{b\to j}_{s_j}} \\ &= g_j(s_j) \prod_{b\in\partial j\setminus a} V^{b\to j}_{s_j} \end{aligned}$$

where we recognized the definition of $V^{b\to j}_{s_j}$ and used it. Analogously, for the $V^{a\to i}_{s_i}$ we can

[2] Note s_i appears in the argument of f_a since $i \in \partial a$

split the sum over the different branches of the tree above factor node a, to get

$$V_{s_i}^{a\to i} = \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right) \prod_{j\in\partial a\setminus i} \underbrace{\left[\sum_{\{s_k\}_{\text{all } k \text{ above } j}} g_j(s_j) \prod_{\text{all } k \text{ above } j} g_k(s_k) \prod_{\text{all } b \text{ above } j} f_b\left(\{s_l\}_{l\in\partial b}\right)\right]}_{=R_{s_j}^{j\to a}}$$

$$= \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right) \prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a} \qquad (*)$$

If we start on the leaves of the tree, i.e. variable nodes that only belong to one factor a, we have from the definition $R_{s_j}^{j\to a} = g_j(s_j)$, for j being a leaf. This, together with the relations above, would allow us to collect recursively the contribution from all branches and compute the total partition function of a tree graphical model rooted in node j as

$$Z = \sum_{s_j} g_j(s_j) \prod_{b\in\partial j} V_{s_j}^{b\to j}\,.$$

This value will not depend on the node in which we rooted the tree, as all the contributions to the partition function are accounted for regardless of the root.

The partition function usually scales like $\exp(cN)$, exponentially in the system size (simply because it is a sum over exponentially many terms), which is a huge number. A more convenient way to deal with the above restricted partition functions R and V is to define messages $\chi_{s_j}^{j\to a}$ and $\psi_{s_i}^{a\to i}$ as follows:

$$\chi_{s_j}^{j\to a} =: \frac{R_{s_j}^{j\to a}}{\sum_s R_s^{j\to a}} \quad \text{so that} \quad \sum_s \chi_s^{j\to a} = 1, \qquad \forall\,(ja)\in E$$

$$\psi_{s_i}^{a\to i} =: \frac{V_{s_i}^{a\to i}}{\sum_s V_s^{a\to i}} \quad \text{so that} \quad \sum_s \psi_s^{a\to i} = 1, \qquad \forall\,(ia)\in E\,. \qquad (**)$$

We will call $\chi_{s_j}^{j\to a}$ a message from variable j to factor a and interpret it as the probability that variable j takes value s_j in the restricted system where only the parts above the red dotted line are considered. Analogously, $\psi_{s_i}^{a\to i}$ will be called the message from factor a to variable i and is interpreted as the probability that variable i takes value s_i in the system where only the parts above the blue dotted line are considered. Using these definitions we rewrite the above recursive relations as:

$$\chi_{s_j}^{j\to a} \overset{\text{by } (*)}{=} \frac{g_j(s_j)\prod_{b\in\partial j\setminus a} V_{s_j}^{b\to j}}{\sum_s g_j(s)\prod_{b\in\partial j\setminus a} V_s^{b\to j}} \times \underbrace{\frac{\prod_{b\in\partial j\setminus a}\sum_{s'} V_{s'}^{b\to j}}{\prod_{b\in\partial j\setminus a}\sum_{s''} V_{s''}^{b\to j}}}_{=1}$$

$$= \frac{g_j(s_j)\prod_{b\in\partial j\setminus a} \frac{V_{s_j}^{b\to j}}{\sum_{s''} V_{s''}^{b\to j}}}{\sum_s g_j(s)\prod_{b\in\partial j\setminus a} \frac{V_s^{b\to j}}{\sum_{s'} V_{s'}^{b\to j}}}$$

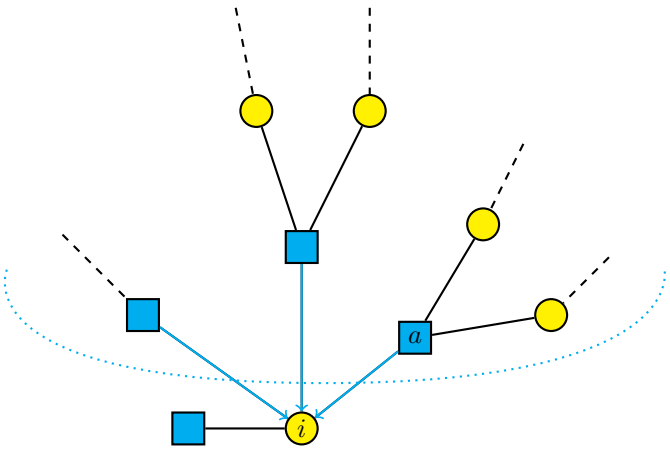
$$\overset{\text{by } (**)}{=} \frac{g_j(s_j)\prod_{b\in\partial j\setminus a}\psi_{s_j}^{b\to j}}{\sum_s g_j(s)\prod_{b\in\partial j\setminus a}\psi_s^{b\to j}} = \frac{1}{Z^{j\to a}} g_j(s_j) \prod_{b\in\partial j\setminus a} \psi_{s_j}^{b\to j}$$

$$Z^{j\to a} =: \sum_s g_j(s) \prod_{b\in\partial j\setminus a} \psi_s^{b\to j}$$

and similarly

$$
\begin{aligned}
\psi_{s_i}^{a\to i} &\overset{\text{by } (*)}{=} \frac{\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a}}{\sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} R_{s_j}^{j\to a}} \times \underbrace{\frac{\prod_{j\in\partial a\setminus i}\sum_{s'_j} R_{s'_j}^{j\to a}}{\prod_{j\in\partial a\setminus i}\sum_{s''_j} R_{s''_j}^{j\to a}}}_{=1} \\
&= \frac{\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} \frac{R_{s_j}^{j\to a}}{\sum_{s''_j} R_{s''_j}^{j\to a}}}{\sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} \frac{R_{s_j}^{j\to a}}{\sum_{s'_j} R_{s'_j}^{j\to a}}} \\
&\overset{\text{by } (**)}{=} \frac{\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a}}{\sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a}} \\
&= \frac{1}{Z^{a\to i}} \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a} \\
Z^{a\to i} &=: \sum_{s_i}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a} = \sum_{\{s_j\}_{j\in\partial a}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a}
\end{aligned}
$$

Above, we just obtained the *self-consistent equations* for the messages χ's and ψ's that are called the Belief Propagation equations.



Now, we want the marginals $\mu_i(s_i)$ and the partition function Z expressed in a way that can be generalized to factor graphs that are not trees. With this in mind, we first write the marginals

(still in a tree factor graph assumption) as

$$\mu_i(s_i) = \frac{1}{Z} g_i(s_i) \prod_{a\in\partial i} V_{s_i}^{a\to i} = \frac{g_i(s_i)\prod_{a\in\partial i} V_{s_i}^{a\to i}}{\sum_s g_i(s)\prod_{a\in\partial i} V_s^{a\to i}} \times \underbrace{\frac{\prod_{a\in\partial i}\sum_{s'} V_{s'}^{a\to i}}{\prod_{a\in\partial i}\sum_{s''} V_{s''}^{a\to i}}}_{=1}$$

$$= \frac{g_i(s_i)\prod_{a\in\partial i}\frac{V_{s_i}^{a\to i}}{\sum_{s''} V_{s''}^{a\to i}}}{\sum_s g_i(s)\prod_{a\in\partial i}\frac{V_s^{a\to i}}{\sum_{s'} V_{s'}^{a\to i}}} = \frac{g_i(s_i)\prod_{a\in\partial i}\psi_{s_i}^{a\to i}}{\sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i}} = \underbrace{\frac{1}{Z^i} g_i(s_i)\prod_{a\in\partial i}\psi_{s_i}^{a\to i}}_{=:\chi_{s_i}^i}$$

$$Z^i =: \sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i}$$

Here we see that each marginal $\mu_i(s_i)$ is a very simple function of the incoming messages $\psi_{s_i}^{a\to i}$, $a\in\partial i$.

The partition function Z, instead, can be compute quite directly by rooting the tree in node i and noticing (independently of which node i we chose as the root)

$$Z = \sum_{s_i} g_i(s_i)\prod_{a\in\partial i} V_{s_i}^{a\to i}\,. \tag{$\dagger$}$$

We want an expression for Z that only involves the messages χ and ψ, in a way that the result does not depend explicitly on the rooting of the tree. To do this, we first define

$$Z^i =: \sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i} = \frac{\sum_s g_i(s)\prod_{a\in\partial i} V_s^{a\to i}}{\prod_{a\in\partial i}\sum_{s'} V_{s'}^{a\to i}}$$

$$Z^a =: \sum_{\{s_i\}_{i\in\partial a}} f_a\left(\{s_i\}_{i\in\partial a}\right)\prod_{i\in\partial a}\chi_{s_i}^{i\to a} = \frac{\sum_{\{s_i\}_{i\in\partial a}} f_a\left(\{s_i\}_{i\in\partial a}\right)\prod_{i\in\partial a} R_{s_i}^{i\to a}}{\prod_{i\in\partial a}\sum_{s'} R_{s'}^{i\to a}}$$

$$Z^{ia} =: \sum_s \chi_s^{i\to a}\psi_s^{a\to i} = \frac{\sum_s V_s^{a\to i} R_s^{i\to a}}{\sum_{s'} V_{s'}^{a\to i}\sum_{s''} R_{s''}^{i\to a}}$$

We will now prove that

$$Z = \frac{\prod_{i=1}^N Z^i \prod_{a=1}^M Z^a}{\prod_{(ia)\in E} Z^{ia}}$$

Let's go:

$$\frac{\prod_{i=1}^N Z^i\prod_{a=1}^M Z^a}{\prod_{(ia)\in E} Z^{ia}} = \frac{\prod_{i=1}^N \frac{\sum_s g_i(s)\prod_{a\in\partial i} V_s^{a\to i}}{\prod_{a\in\partial i}\sum_{s'} V_{s'}^{a\to i}} \cdot \prod_{a=1}^M \frac{\sum_{\{s_i\}_{i\in\partial a}} f_a\left(\{s_i\}_{i\in\partial a}\right)\prod_{i\in\partial a} R_{s_i}^{i\to a}}{\prod_{i\in\partial a}\sum_{s'} R_{s'}^{i\to a}}}{\prod_{(ia)\in E}\frac{\sum_s V_s^{a\to i} R_s^{i\to a}}{\sum_{s'} V_{s'}^{a\to i}\sum_{s''} R_{s''}^{i\to a}}}$$

$$= \frac{\cancel{\frac{\prod_{i=1}^N\sum_s g_i(s)\prod_{a\in\partial i} V_s^{a\to i}}{\prod_{(ia)\in E}\sum_{s'} V_{s'}^{a\to i}}}\cdot\cancel{\frac{\prod_{a=1}^M\sum_{\{s_i\}_{i\in\partial a}} f_a\left(\{s_i\}_{i\in\partial a}\right)\prod_{i\in\partial a} R_{s_i}^{i\to a}}{\prod_{(ia)\in E}\sum_{s'} R_{s'}^{i\to a}}}}{\frac{\prod_{(ia)\in E}\sum_s V_s^{a\to i} R_s^{i\to a}}{\cancel{\prod_{(ia)\in E}\sum_{s'} V_{s'}^{a\to i}}\cdot\cancel{\prod_{(ia)\in E}\sum_{s''} R_{s''}^{i\to a}}}}$$

We root the tree at node j. By (*), we have

$$\frac{\prod_{i=1}^{N} Z^i \prod_{a=1}^{M} Z^a}{\prod_{(ia)\in E} Z^{ia}} = \left[\sum_s g_j(s) \prod_{a\in\partial j} V_s^{a\to j}\right] \times \left[\prod_{\substack{i=1\\ i\neq j}}^{N} \sum_s V_s^{b\to i} \overbrace{g_i(s) \prod_{a\in\partial i\setminus b} V_s^{a\to i}}^{=R_s^{i\to b}}\right]$$
$$\times \left[\prod_{a=1}^{M} \sum_{s_i} R_{s_i}^{i\to a} \overbrace{\sum_{\{s_k\}_{k\in\partial a\setminus i}} f_a\left(\{s_k\}_{k\in\partial a}\right) \prod_{k\in\partial a\setminus i} R_{s_k}^{k\to a}}^{=V_s^{a\to i}}\right]$$
$$\times \left[\prod_{(ia)\in E} \sum_s V_s^{a\to i} R_s^{i\to a}\right]^{-1}$$

And by (†) we conclude

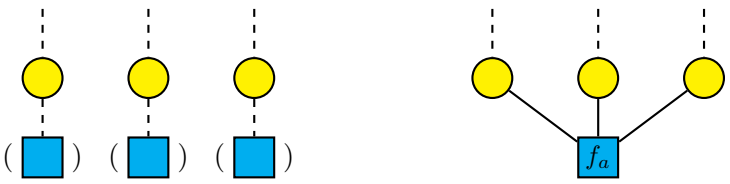
$$\frac{\prod_{i=1}^{N} Z^i \prod_{a=1}^{M} Z^a}{\prod_{(ia)\in E} Z^{ia}} = Z \cdot \frac{\prod_{\substack{i=1\\ i\neq \text{root } j}}^{N} \overbrace{\sum_s R_s^{i\to b} V_s^{b\to i}}^{b \text{ towards root from } i} \cdot \prod_{a=1}^{M} \overbrace{\sum_s V_s^{a\to i} R_s^{i\to a}}^{i \text{ towards root from a}}}{\prod_{(ia)\in E} \sum_s V_s^{a\to i} R_s^{i\to a}} = Z$$

The last step comes from the fact that, in a tree, if we take for all variables node the edges towards the root, and for all factor nodes the edge towards the root we accounted for exactly all the edges.

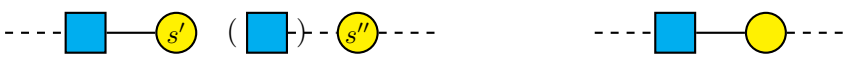
It is quite instructive to keep in mind the interpretation of free energy terms (deduced from their definitions)

- Z^i: change of the partition function Z when variable node i is added to the factor graph

- Z^a: change of the partition function Z when factor node a is added to the factor graph



- Z^{ia}: change of Z when variable node i and factor node a are connected



The formula for the partition function is hence quite intuitive: first, we are adding up all the contributions from nodes and factors. Since with every new addition we account for all the connected edges, we end up counting each edge exactly twice (since it is connected both to a node and a factor). Thus, we have to subtract the edge contributions once, in order to correct for the double-counting.

5.2.2 Belief propagation equations summary

After a bulky derivation, let us summarize the Belief Propagation equations, and the formulas for the marginals and for the free entropy.

We consider a graphical model for the following probability distribution

$$P\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z}\prod_{i=1}^N g_i(s_i)\prod_{a=1}^M f_a\left(\{s_i\}_{i\in\partial a}\right)$$

The Belief Propagation equations read

$$\chi_{s_j}^{j\to a} = \frac{1}{Z^{j\to a}} g_j(s_j)\prod_{b\in\partial j\setminus a}\psi_{s_j}^{b\to j}$$
$$\psi_{s_i}^{a\to i} = \frac{1}{Z^{a\to i}}\sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right)\prod_{j\in\partial a\setminus i}\chi_{s_j}^{j\to a}$$

where $Z^{j\to a}$ and $Z^{a\to i}$ are normalization factors set so that $\sum_s \chi_s^{j\to a} = 1$ and $\sum_s \psi_s^{a\to i} = 1$.

The free entropy density Φ, which is exact on trees and is called the Bethe free entropy on more general graphs, reads

$$N\Phi = \log Z = \sum_{i=1}^N \log Z^i + \sum_{a=1}^M \log Z^a - \sum_{(ia)} \log Z^{ia} \tag{5.1}$$
$$Z^i =: \sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i}$$
$$Z^a =: \sum_{\{s_i\}_{i\in\partial a}} f_a\left(\{s_i\}_{i\in\partial a}\right)\prod_{i\in\partial a}\chi_{s_i}^{i\to a}$$
$$Z^{ia} =: \sum_s \chi_s^{i\to a}\psi_s^{a\to i}$$

The marginals of the variable i are given as

$$\mu_i(s_i) = \frac{1}{Z^i} g_i(s_i)\prod_{a\in\partial i}\psi_{s_i}^{a\to i}$$

A landmark property of Belief Propagation and the Bethe entropy is that the BP equations can be obtained from the stationary point condition of the Bethe free entropy. Show this for homework. This is a crucial property in the task of finding the correct fixed point of BP when several of them exist.

As it often happens for successful methods/algorithms, Belief Propagation has been independently discovered in several fields. Notable works introducing BP in various forms are:

- Hans Bethe & Rudolf Peierls 1935, in magnetism to approximate the regular lattice by a tree of the same degree.
- Robert G. Gallager 1962 Gallager (1962), in information theory for decoding sparse error correcting codes.
- Judea Pearl 1982 Pearl (1982), in Bayesian inference.

5.2.3 How do we use Belief Propagation?

Above we derived the BP equations and the free entropy on tree factor graphs. To use BP as an algorithm on trees, we initialize the messages on the leaves according to definition $\chi^{j\to a}_{s_j} = g_j(s_j)$, and then spread towards the root. A single iteration is sufficient. The trouble is that basically no problems of interest are defined on tree graphical models. Thus its usage in this context is very limited.

On graphical models with loops BP can always be used as a heuristic iterative algorithm

$$\chi^{j\to a}_{s_j}(t+1) = \frac{1}{Z^{j\to a}(t)} g_j(s_j) \prod_{b\in\partial j\setminus a} \psi^{b\to j}_{s_j}(t)$$

$$\psi^{a\to i}_{s_i}(t) = \frac{1}{Z^{a\to i}(t)} \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right) \prod_{j\in\partial a\setminus i} \chi^{j\to a}_{s_j}(t)$$

Initialization has various options

- $\chi^{j\to a}_{s_j}(t=0) = g_j(s_j)/\sum_s g_j(s)$ in the "prior".
- $\chi^{j\to a}_{s_j}(t=0) = g_j(s_j) + \varepsilon^{j\to a}_{s_j}$ + normalize, $\varepsilon^{j\to a}_{s_j}$ are small perturbation of the "prior".
- $\chi^{j\to a}_{s_j}(t=0) = \varepsilon^{j\to a}_{s_j}$ + normalize, random initialization.
- $\chi^{j\to a}_{s_j}(t=0) = \delta_{s_j,s^*_j}$, planted initialization

We will discuss in the follow-up lectures the usage of these different initializations and their properties.

Then, we iterate the Belief Propagation equations until convergence or for a given number of steps. The order of iterations can be parallel, or random sequential. Again, we will discuss in what follows the advantages and disadvantages of the different update schemes.

It is useful to remark that, if the graph is not a tree, then the independence of branches when conditioning on a value of a node is in general invalid. In general, the products in the BP equations should involve joint probability distributions. In this case, the simple recursive structure of the algorithm would get replaced by joint probability distributions over neighborhoods

(up to a large distance), which would be just as intractable as the original problem. We will see in this lecture that there are many circumstances in which the independence between the incoming messages $\psi_{s_j}^{b\to i}$ for $b \in \partial j \setminus a$ and between $\chi_{s_j}^{j\to a}$ for $j \in \partial a \setminus i$ is approximately true. We will focus on cases when it leads to results that are exact up to the leading order in N. One important class of graphical models on which BP and its variants lead to asymptotically exact results is that of sparse random factor graphs. This is the case because such graphs look like trees up to a distance that grows (logarithmically) with the system size.

Sparse random graphs are locally tree-like

Informally, a graph is *locally tree-like* if, for almost all nodes, the neighborhood up to distance d is a tree, and $d \to \infty$ as $N \to \infty$.

Importantly, sparse random factor graphs are locally tree-like. Sparse here refers to the fact that the average degree of both the variable node and the factor nodes are constants while the size of the graph $N, M \to \infty$.

We will illustrate this claim in the case of sparse random graphs (for sparse random factor graphs the argument is analogous). In a random graph, each edge is present with probability $\frac{c}{N-1}$, where $c = \mathrm{O}(1)$ and $N \to \infty$, the average degree of nodes is c.

In order to compute the length of the shortest loop that goes trough a typical node i, consider a non-backtracking spreading process starting in node i. The probability of the spreading process returning to i in d steps (through a loop) is

$$1 - \Pr(\text{does not return to } i) \approx 1 - \left(1 - \frac{1}{N}\right)^{c^d}$$

where c^d is the expected number of explored nodes after d steps. In the limit $N \to \infty, c = O(1)$ so this probability is exponentially small for small distances d, and exponentially close to one for large distances d. The distance d at which this probability becomes $O(1)$ marks the order of the length of the shorted loop going trough node i:

$$c^d \log\left(1 - \frac{1}{N}\right) = c^d \left(-\frac{1}{N} - \frac{1}{2N^2} + \cdots\right) \approx \mathrm{O}(1)$$

This happens when $c^d \approx \mathrm{O}(N) \quad \Rightarrow \quad d \approx \log(N)/\log(c)$. We conclude that a random graph with average degree $c = \mathrm{O}(1)$ is such that the length of the shortest loop going through a random node i is $\mathrm{O}(\log(N))$ with high probability. We thus see that up to distance $\mathrm{O}(\log(N))$ the neighborhood of a typical node is a tree.

We will hence attempt to use belief propagation and the Bethe free entropy on locally tree-like graphs. The key assumption BP makes is the independence of the various branches of the tree. If the branches are connected through loops of length $\mathrm{O}(\log(N)) \to \infty$ and the correlation between the root and the leaves of the tree decays fast enough (we will make this condition much more precise), the independence of branches gets asymptotically restored and BP and the Bethe free entropy lead to asymptotically exact results. We will investigate cases when the correlation decay is fast enough, and also those when it is not, and show how to still use BP-based approach to obtain adjusted asymptotically exact results (this will lead us to the notion of replica symmetry breaking).

5.3 Exercises

Exercise 5.1: Representing problems by graphical models and Belief Propagation

Write the following problems (i) in terms of a probability distribution and (ii) in terms of a graphical model by drawing a (small) example of the corresponding factor graph. Finally (iii) write the Belief Propagation equations for these problems (without coding or solving them) and the expression for the Bethe free energy that would be computed from the BP fixed points.

(1) **p-spin model**

One model that is commonly studied in physics is the so-called Ising 3-spin model. The Hamiltonian of this model is written as

$$\mathcal{H}(\{S_i\}_{i=1}^N) = -\sum_{(ijk)\in E} J_{ijk} S_i S_j S_k - \sum_{i=1}^{N} h_i S_i \tag{5.2}$$

where E is a given set of (unordered) triplets $i \neq j \neq k$, J_{ijk} is the interaction strength for the triplet $(ijk) \in E$, and h_i is a magnetic field on spin i. The spins are Ising, which in physics means $S_i \in \{+1, -1\}$.

(2) **Independent set problem**

Independent set is a problem defined and studied in combinatorics and graph theory. Given a (unweighted, undirected) graph $G(V, E)$, an independent set $S \subseteq V$ is defined as a subset of nodes such that if $i \in S$ then for all $j \in \partial i$ we have $j \notin S$. In other words in for all $(ij) \in E$ only i or j can belong to the independent set.

(a) Write a probability distribution that is uniform over all independent sets on a given graph.
(b) Write a probability distribution that gives a larger weight to larger independent sets, where the size of an independent set is simply $|S|$.

(3) **Matching problem** Matching is another classical problem of graph theory. It is related to a dimer problem in statistical physics. Given a (unweighted, undirected) graph $G(V, E)$ a matching $M \subseteq E$ is defined as a subset of edges such that if $(ij) \in M$ then no other edge that contains node i or j can be in M. In other words a matching is a subset of edges such that no two edges of the set share a node.

(a) Write a probability distribution that is uniform over all matchings on a given graph.
(b) Write a probability distribution that gives a larger weight to larger matchings, where the size of a matching is simply $|M|$.

Exercise 5.2: Bethe free entropy

A key connection between Belief Propagation and the Bethe free entropy:

Show that the BP equations we derived in the lecture

$$\chi_{s_j}^{j\to a} = \frac{1}{Z^{j\to a}} g_j(s_j) \prod_{b\in\partial j\setminus a} \psi_{s_j}^{b\to j}$$
$$\psi_{s_i}^{a\to i} = \frac{1}{Z^{a\to i}} \sum_{\{s_j\}_{j\in\partial a\setminus i}} f_a\left(\{s_j\}_{j\in\partial a}\right) \prod_{j\in\partial a\setminus i} \chi_{s_j}^{j\to a}$$

are stationarity conditions of the Bethe free entropy equation 5.1 under the constraint that both $\sum_s \psi_s^{a\to i} = 1$ and $\sum_s \chi_s^{i\to a} = 1$ for all $(ia) \in E$.

Appendix

5.A BP for pair-wise models, node by node

In Section 5.2, we have derived the Belief propagation (BP) equations for a general tree-like graphical model. To first sight, the BP equations might look daunting. The goal of this Appendix is to provide additional intuition behind the BP equations by constructing them from scratch, node by node, in a concrete setting.

Let $G = (V, E)$ be a graph with $N = |V|$ nodes, and let's consider for concreteness the case of a pair-wise interacting spin model on G:

$$\mathbb{P}(\boldsymbol{S} = \boldsymbol{s}) = \frac{1}{Z} \prod_{i=1}^{N} g_i(s_i) \prod_{(ij)\in E} f_{(ij)}(s_i, s_j)$$

This encompasses many cases of interest, for instance the RFIM we studied in Chapter 2, for which:

$$g_i(s_i) = e^{\beta h_i s_i}, \qquad f_{(ij)}(s_i, s_j) = e^{\beta s_i s_j} \qquad \text{(RFIM)}$$

and the $q = 2$ graph coloring problem studied in this Chapter:

$$g_i(s_i) = 1, \qquad f_{(ij)}(s_i, s_j) = e^{\beta \delta_{s_i s_j}}. \qquad \text{(Graph coloring)}$$

The reader can keep any of these two problems in mind in what follows. To lighten notation, we will write $\mu(\boldsymbol{s}) =: \mathbb{P}(\boldsymbol{S} = \boldsymbol{s})$ for the probability distribution, with μ understood as a function $\mu : \{-1, 1\}^N \to [0, 1]$. We will also use $\propto$ to denote "equal up to a multiplicative factor", which here it will always denote the constant which normalizes the probability distributions or messages. Recall that the BP equations are self-consistent equations for the messages or beliefs from variables to factors $\chi^{i\to a}_{s_i}$ and from factors to variables $\psi^{a\to i}_{s_i}$. For pair-wise models, we have one factor per edge, and the BP equations read:

$$\chi^{j\to(ij)}_{s_j} = \frac{1}{Z^{j\to(ij)}} g_j(s_j) \prod_{(kj)\in\partial j\setminus(ij)} \psi^{(kj)\to j}_{s_j}$$

$$\psi^{(ij)\to i}_{s_i} = \frac{1}{Z^{(ij)\to i}} \sum_{\{s_j\}_{j\in\partial(ij)\setminus i}} f_{(ij)}(s_i, s_j) \prod_{j\in\partial(ij)\setminus i} \chi^{j\to(ij)}_{s_j}$$

Since in pair-wise models there is a one-to-one correspondence between edges and factor nodes, we can simply rewrite the BP equations directly on the graph G. For each node $i \in V$

of the graph, define the outgoing messages $\chi_{s_i}^{i\to j} =: \chi_{s_i}^{i\to(ij)}$ and the incoming messages $\psi_{s_i}^{j\to i} = \psi_{s_i}^{(ij)\to i}$. The BP equations in terms of these "new" messages read:

$$\chi_{s_i}^{i\to j} = \frac{1}{Z^{i\to j}} g_i(s_i) \prod_{k\in\partial i\setminus j} \psi_{s_i}^{k\to i}, \qquad \psi_{s_i}^{k\to i} = \frac{1}{Z^{k\to i}} \sum_{s_k\in\{-1,1\}} f_{(ik)}(s_i, s_k)\, \chi_{s_k}^{k\to i}$$

Recall that the marginal for variable s_i is given in terms of the messages as:

$$\mu_i(s_i) =: \mathbb{P}(S_i = s_i) \propto g_i(s_i) \prod_{j\in\partial i} \psi_{s_i}^{j\to i} \tag{5.3}$$

Which basically tell us that the probability that $S_i = s_i$ is simply given by the "local belief" (or prior) g_i times the incoming beliefs $\psi_{s_i}^{j\to i}$ from all neighbors of i. In other words, BP factorizes the marginal distribution at every node in terms of independent beliefs.

Note that in this case it is also easy to solve for one of the messages to obtain a self-consistent equation for only one of them. For instance, solving for the incoming messages $\psi_{s_k}^{k\to i}$ gives:

$$\chi_{s_i}^{i\to j} = \frac{g_i(s_i)}{Z^{i\to j}} \prod_{k\in\partial i\setminus j} \sum_{s_k\in\{-1,+1\}} f_{(ik)}(s_i, s_k)\chi_{s_k}^{k\to i}.$$

5.A.1 One node

When the graph has a single node $N = 1$, we have only a single spin $S_1 \in \{-1, 1\}$ and therefore $E = \varnothing$. The corresponding factor graph has a single variable node and a single factor node for the "local field" g_i. In this case, the marginal distribution over the spin S_1 is simply given by the "local belief":

$$\mu(\pm 1) =: \mathbb{P}(S_1 = \pm 1) = \frac{1}{Z} g_i(\pm 1) = \frac{g_1(\pm 1)}{g_1(+1) + g_1(-1)}$$

Note that in the absence of a prior $g_1(s_1) = 1$, the marginal is simply the uniform distribution $\mu_1(s_1) = \frac{1}{2}$.

5.A.2 Two nodes

For two nodes $N = 2$, we have two spin variables $S_1, S_2 \in \{-1, +1\}$ and two possible graphs: either the spins are decoupled and $E = \varnothing$, or they interact through an edge $E = \{(12)\}$.

No edges: In the first case, the joint distribution is given by:

$$\mu(s_1, s_2) = \frac{1}{Z} g_1(s_1) g_2(s_2)$$
$$= \frac{g_1(s_1) g_2(s_2)}{g_1(+1)g_2(+1) + g_1(-1)g_2(+1) + g_1(+1)g_2(-1) + g_1(-1)g_2(-1)}$$

and the marginal distribution of S_1 is given by:

$$\begin{aligned}\mu_1(s_1) &= \sum_{s_2\in\{-1,+1\}} \mu(s_1,s_2) = \mu(s_1,-1)+\mu(s_1,+1)\\ &= \frac{g_2(-1)+g_2(+1)}{g_1(+1)g_2(+1)+g_1(-1)g_2(+1)+g_1(+1)g_2(-1)+g_1(-1)g_2(-1)} g_1(s_1)\end{aligned}$$

The marginal distribution of S_2 is obtained by interchanging $1 \leftrightarrow 2$. Notice that in this case, the joint distribution factorizes into the product of the marginals:

$$\mu(s_1,s_2) = \mu_1(s_1)\mu_2(s_2)$$

This is a direct consequence of the absence of an interaction term coupling the two spins. Note that we cannot write BP equations for this case since there are no factors.

One edge: The case in which $E = \{(12)\}$ is more interesting. Now the joint distribution is given by:

$$\mu(s_1,s_2) = \frac{1}{Z} g_1(s_1)g_2(s_2)f_{(12)}(s_1,s_2)$$

Notice that very quickly it becomes cumbersome to write the exact expression for the normalization Z, which we keep implicit from now on. The marginals are now given by:

$$\begin{aligned}\mu_1(s_1) &= \frac{1}{Z}\left[g_2(+1)f_{(12)}(s_1,+1)+g_2(-1)f_{(12)}(s_1,-1)\right] g_1(s_1)\\ \mu_2(s_2) &= \frac{1}{Z}\left[g_1(+1)f_{(12)}(+1,s_2)+g_1(-1)f_{(12)}(-1,s_2)\right] g_2(s_2)\end{aligned}$$

Crucially, it is easy to check that the joint distribution doesn't factorise anymore:

$$\mu(s_1,s_2) \neq \mu_1(s_1)\mu_2(s_2)$$

Let's now look at what the BP equations are telling us. For instance, the outgoing messages are given by:

$$\chi^{1\to 2}_{s_1} \propto g_1(s_1), \qquad\qquad \chi^{2\to 1}_{s_2} \propto g_2(s_2)$$

while the incoming messages are given by:

$$\begin{aligned}\psi^{2\to 1}_{s_1} &\propto \sum_{s_2\in\{-1,+1\}} f_{(12)}(s_1,s_2)\chi^{2\to 1}_{s_2} \propto \sum_{s_2\in\{-1,+1\}} f_{(12)}(s_1,s_2)f_{(12)}(s_1,s_2)g_2(s_2)\\ &\propto f_{(12)}(s_1,-1)g_2(-1)+f_{(12)}(s_1,+1)g_2(+1)\\ \psi^{1\to 2}_{s_2} &\propto \sum_{s_1\in\{-1,+1\}} f_{(12)}(s_1,s_2)\chi^{1\to 2}_{s_1} \propto \sum_{s_1\in\{-1,+1\}} f_{(12)}(s_1,s_2)f_{(12)}(s_1,s_2)g_1(s_1)\\ &\propto f_{(12)}(-1,s_2)g_1(-1)+f_{(12)}(+1,s_1)g_1(+1)\end{aligned}$$

From that, it is pretty clear that the marginals factorize in terms of the local beliefs times the incoming beliefs eq. equation 5.3.

5.A.3 Three nodes

Finally, let's consider the more involved case of three nodes $N = 3$. The case in which there is no edge $E = \varnothing$ or only one edge $|E| = 1$ reduces to one of the cases we have seen before. Therefore, the interesting cases are when we have either two or three edges.

Two edges: For two edges, there are two nodes which have degree 1 and one node with degree two. Without loss of generality, we can choose node 2 to have degree 2, and the joint distribution read:

$$\mu(s_1, s_2, s_2) = \frac{1}{Z} f_{(12)}(s_1, s_2) f_{(23)}(s_2, s_3) g_1(s_1) g_2(s_2) g_3(s_3)$$

The marginal probability of $S_1 = s_1$ is then given by:

$$\begin{aligned}\mu_1(s_1) &= \sum_{s_2 \in \{-1,1\}} \sum_{s_3 \in \{-1,1\}} \mu(s_1, s_2, s_2) \\ &\propto \big[\big(f_{(23)}(-1,-1)g_3(-1) + f_{(23)}(-1,+1)g_2(-1)g_3(+1)\big) f_{(12)}(s_1,-1)g_2(-1) + \\ &\quad \big(f_{(23)}(+1,-1)g_3(-1) + f_{(23)}(+1,+1)g_3(+1)\big) f_{(12)}(s_1,+1)g_2(+1)\big] g_1(s_1)\end{aligned}$$

Note that each two of the four terms share a common term. The outgoing BP messages now read:

$$\chi^{1\to 2}_{s_1} \propto g_1(s_1), \qquad \chi^{2\to 1}_{s_2} \propto g_2(s_2) \sum_{s_3 \in \{-1,+1\}} f_{(23)}(s_2, s_3) \chi^{3\to 2}_{s_3}, \qquad \chi^{3\to 2}_{s_3} \propto g_3(s_3)$$

Inserting the first and third into the second:

$$\begin{aligned}\chi^{2\to 1}_{s_2} &\propto g_2(s_2) \sum_{s_3 \in \{-1,+1\}} f_{(23)}(s_2, s_3) g_3(s_3) \\ &\propto \big[f_{(23)}(s_2,-1)g_3(-1) + f_{(23)}(s_2,+1)g_3(+1)\big] g_2(s_2)\end{aligned}$$

It is easy to check that this allow us to reconstruct the marginal of $S_1 = s_1$ from eq. equation 5.3:

$$\begin{aligned}\mu_1(s_1) &\propto g_1(s_1) \sum_{s_2 \in \{-1,+1\}} f_{(12)}(s_1, s_2) \chi^{2\to 1}_{s_2} \\ &\propto g_1(s_1) \sum_{s_2 \in \{-1,+1\}} f_{(12)}(s_1, s_2) \big[f_{(23)}(s_2,-1)g_3(-1) + f_{(23)}(s_2,+1)g_3(+1)\big] g_2(s_2)\end{aligned}$$

Three edges: For three edges, all nodes are connected and have degree two. The joint distribution read:

$$\mu(s_1, s_2, s_2) = \frac{1}{Z} f_{(12)}(s_1, s_2) f_{(13)}(s_1, s_3) f_{(23)}(s_2, s_3) g_1(s_1) g_2(s_2) g_3(s_3)$$

The marginal probability of $S_1 = s_1$ now reads:

$$\begin{aligned}\mu_1(s_1) &= \sum_{s_2 \in \{-1,1\}} \sum_{s_3 \in \{-1,1\}} \mu(s_1, s_2, s_2) \\ &\propto \big[f_{(12)}(s_1,-1)f_{(13)}(s_1,-1)g_2(-1)g_3(-1) + f_{(12)}(s_1,+1)f_{(13)}(s_1,-1)g_2(+1)g_3(-1) \\ &\quad + f_{(12)}(s_1,-1)f_{(13)}(s_1,+1)g_2(-1)g_3(+1) + f_{(12)}(s_1,+1)f_{(13)}(s_1,+1)g_2(+1)g_3(+1)\big] g_1(s_1)\end{aligned}$$

Note that, different from the 2 nodes case the four terms above don't share any common factor apart from $g_1(s_1)$. The outgoing BP messages now read:

$$\chi^{1\to 2}_{s_1} \propto g_1(s_1) \sum_{s_3\in\{-1,1\}} f_{(13)}(s_1,s_3)\chi^{3\to 1}_{s_3}, \qquad \chi^{1\to 3}_{s_1} \propto g_1(s_1) \sum_{s_2\in\{-1,1\}} f_{(12)}(s_1,s_2)\chi^{2\to 1}_{s_2}$$
$$\chi^{2\to 1}_{s_2} \propto g_2(s_2) \sum_{s_3\in\{-1,1\}} f_{(23)}(s_2,s_3)\chi^{3\to 2}_{s_3}, \qquad \chi^{2\to 3}_{s_2} \propto g_2(s_2) \sum_{s_1\in\{-1,1\}} f_{(21)}(s_2,s_1)\chi^{1\to 2}_{s_1}$$
$$\chi^{3\to 1}_{s_3} \propto g_3(s_3) \sum_{s_2\in\{-1,1\}} f_{(32)}(s_3,s_2)\chi^{2\to 3}_{s_2}, \qquad \chi^{3\to 2}_{s_3} \propto g_3(s_3) \sum_{s_1\in\{-1,1\}} f_{(31)}(s_3,s_1)\chi^{1\to 3}_{s_1}$$

Note that in this case it is not simple to solve for the messages. For instance, the marginal of $S_1 = s_1$ is given by:

$$\mu_1(s_1) \propto g_1(s_1) \sum_{s_2\in\{-1,+1\}} f_{(12)}(s_1,s_2)\chi^{2\to 1}_{s_2}$$

and therefore it depends on $\chi^{2\to 1}_{s_2}$. However, $\chi^{2\to 1}_{s_2}$ depends on $\chi^{3\to 2}_{s_3}$, which itself depends on $\chi^{1\to 3}_{s_1}$ which depends on $\chi^{2\to 1}_{s_2}$. In other words: we cannot factorize the marginals in terms of independent messages because removing an edge, say (12), doesn't make the variable S_1 independent from variable S_2; they are correlated through their common link to S_3. This is a direct consequence of the fact that the graph we consider has a loop.

Chapter 6

Belief propagation for graph coloring

> The picture will have charm when each colour is very unlike the one next to it.
>
> Leon Battista Alberti – 1404-1472

As discussed previously, the probability distribution that we want to investigate in graph coloring, given graph $G(V, E)$ and colors $s_i \in \{1, 2, \dots, q\}$ reads

$$P\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z_G(\beta)} \prod_{(ij)\in E} \mathrm{e}^{-\beta\delta_{s_i,s_j}} \,. \tag{6.1}$$

In physics that variables of the type are called Potts spins, and the model is called correspondingly the Potts model. In what follows we will consider both the repulsive (anti-ferromagnetic) case of $\beta > 0$ and the attractive (ferromagnetic) case of $\beta < 0$. In our notation, the node-factors will simply be $g_i(s_i) = 1$ for all i, and the interaction factors $f_{(ij)}(s_i, s_j) = \mathrm{e}^{-\beta\delta_{s_i,s_j}}$ for all $(ij) \in E$.

6.1 Deriving quantities of interest

Let us now illustrate how to compute quantities of interest such as the number of configurations having a given energy. We define the *energy* $e = \sum_{(ij)\in E} \delta_{s_i,s_j}/N$ as the number of monochromatic edges per node. We denote the number of colorings of a given energy as $\mathcal{N}(e)$ and define the entropy $s(e)$ via

$$\mathcal{N}(e) = \mathrm{e}^{Ns(e)} \tag{6.2}$$

Letting Φ be the free entropy density, we can then rewrite the partition function as

$$\mathrm{e}^{N\Phi(\beta)} = Z_G = \sum_{\{s_i\}_{i=1}^N} \mathrm{e}^{-\beta\sum_{(ij)\in E}\delta_{s_i,s_j}} = \sum_e \sum_{\text{all coloring of energy } e} \mathrm{e}^{-N\beta e} = \int \mathrm{d}e\, \mathrm{e}^{Ns(e)-N\beta e}$$

where in the third step we split the sum into the sum of all the configurations that will be at energy e and the sum over all the energies and in the last step we replaced the sum over the

discrete values of e by an integral over e and we used the definition introduced in equation 6.2. This is well justified since we consider the limit $N \to \infty$ and we are interested in the leading order (in N) of Φ, e and s. The saddle-point method then gives us

$$\left.\frac{\partial s(e)}{\partial e}\right|_{e=e^*} = \beta, \qquad \Phi(\beta) = s(e^*) - \beta e^* \tag{6.3}$$

A random configuration (sampled from the Boltzmann distribution at temperature β) will have energy concentrating at e^* w.h.p. $\Rightarrow \langle e \rangle_{\text{Boltz}} = e^*$. We also remind that the Boltzmann distribution is convenient because

$$\frac{\mathrm{d}\Phi(\beta)}{\mathrm{d}\beta} = -\langle e \rangle_{\text{Boltz}} \,. \tag{6.4}$$

Thus, if we compute the free entropy density Φ as a function of β, we can compute also its derivative and consequently access the number of configurations of a given energy $s(e)$. Doing these calculations exactly is in general a computationally intractable task. We will use BP and the Bethe free entropy to obtain approximate — and in some cases asymptotically exact — results.

How can we use Belief Propagation to evaluate the above quantities? The BP fixed point gives us a set of messages $\chi^{i\to j}$ and the Bethe free entropy Φ_{Bethe} as a function of the messages. In general, the messages give us an approximation of the true marginals of the variables, and the Bethe free entropy Φ_{Bethe} give us an approximation for the true free entropy Φ. Remembering that a BP fixed point is a stationary point of the Bethe free entropy we get:

$$\frac{\mathrm{d}\Phi_{\text{Bethe}}(\beta)}{\mathrm{d}\beta} = \underbrace{\frac{\partial \Phi_{\text{Bethe}}}{\partial \chi}}_{=0 \text{ at BP fixed point}} \cdot \frac{\partial \chi}{\partial \beta} + \frac{\partial \Phi_{\text{Bethe}}(\beta)}{\partial \beta} \tag{6.5}$$

Thus, given a BP fixed point we can approximate both the free entropy and the average energy. Once we evaluated Φ_{Bethe} and e^* we can readily obtain the entropy $s(e)$, which we defined as the logarithm of the number of colorings at energy e:

$$s(e^*) = \Phi(\beta) + \beta e^*. \tag{6.6}$$

Remember, however, that the correctness of the resulting $s(e)$ relies on the correctness of the Bethe free entropy for a given fixed point of the BP equations.

6.2 Specifying generic BP to coloring

We now apply the BP equations, as derived in the previous section, even though in general the graph $G(V, E)$ is not a tree. In what follows, we will see under what circumstances this can lead to asymptotically exact results for coloring of random graphs. Note that, on generic graphs this approach can always be considered as a heuristic approximation of the quantities of interest (in most cases the approximation is hard to control). In this section we will manipulate the BP equations to see what can be derived from them and, when needed, we will restrict our considerations to random sparse graphs.

Applying the recipe from the previous section we obtain:

$$\chi^{j\to(ij)}_{s_j} = \frac{1}{Z^{j\to(ij)}} \prod_{(kj)\in\partial j\setminus(ij)} \psi^{(kj)\to j}_{s_j}$$
$$\psi^{(ij)\to i}_{s_i} = \frac{1}{Z^{(ij)\to i}} \sum_{s_j} \mathbb{e}^{-\beta\delta_{s_i,s_j}} \chi^{j\to(ij)}_{s_j} = \frac{1}{Z^{(ij)\to i}} \left[1 - \left(1-\mathbb{e}^{-\beta}\right)\chi^{j\to(ij)}_{s_i}\right]$$

We see that since every factor node has two neighbors, the second BP equations has a simple form (as the product is only over one term). We can thus eliminate the messages ψ from the equations while still keeping a simple form. We get

$$\chi^{j\to(ij)}_{s_j} = \frac{1}{Z^{j\to(ij)}\prod_{(kj)\in\partial j\setminus(ij)} Z^{(kj)\to j}} \prod_{(kj)\in\partial j\setminus(ij)} \left[1 - \left(1-\mathbb{e}^{-\beta}\right)\chi^{k\to(kj)}_{s_j}\right]$$
$$\Rightarrow \quad \chi^{j\to i}_{s_j} = \frac{1}{Z^{j\to i}} \prod_{k\in\partial j\setminus i} \left[1 - \left(1-\mathbb{e}^{-\beta}\right)\chi^{k\to j}_{s_j}\right]$$

In the last equation we defined an overall normalization term $Z^{j\to i}$ and went back to a graph notation, where ∂i denotes the set of neighbouring nodes of i (instead, in a factor graph notation ∂i would denote the set of neighbouring factors of i). On a first sight, this change of notation can be confusing, since we use the same letter χ to denote the messages on the original graph and on the factor graph. However, note it presents no ambiguity: messages between two variable nodes $i \to j$ can only refer to the original graph, since in a factor graph we cannot connect two variable nodes directly.

The resulting belief propagation equations have quite an intuitive meaning. Recall that $\chi^{j\to i}_{s_j}$ represents the probability that node j takes color s_j if the connection between j and i was temporarily removed. Keeping this in mind, the terms in the above equations have the following meaning

- $1 - \left(1-\mathbb{e}^{-\beta}\right)\chi^{k\to i}_{s_j} = \sum_{s_k\neq s_j}\chi^{k\to i}_{s_k} + \mathbb{e}^{-\beta}\chi^{k\to i}_{s_j}$, is the probability that neighbor k allows node j to take color s_j.
- $\prod_{k\in\partial j\setminus i}[\cdots]$ is the the probability that all the neighbors let node j take color s_j (i was excluded, as the edge (ij) is removed). The product is used because of the implicit assumption of BP, about the independence of the neighbors when conditioning on the value s_j.

In computer science belief propagation is most commonly discussed as an algorithm. We note that in the context of graph coloring the messages provide the marginal probabilities, not directly a proper coloring. In some regimes a proper graph coloring can be deduced from BP with decimation of variables, i.e. variables are set one by one to values deduced from the values of the message at convergence or after a given number of iterations. We will discuss how well this algorithm performs in follow up chapters. In this chapter we will be using BP rather as an analysis tool.

6.3 Bethe free energy for coloring

In a homework problem you will show that using similar simplifications as above we can rewrite the generic Bethe free entropy for graph coloring as

$$N\Phi_{\rm Bethe}(\beta) = \sum_{i=1}^{N} \log Z^{(i)} - \sum_{(ij)\in E} \log Z^{(ij)} \tag{6.7}$$

where

$$Z^{(i)} = \sum_{s} \prod_{k\in\partial i} \left[1 - \left(1 - \mathrm{e}^{-\beta}\right) \chi_s^{k\to i}\right]$$
$$Z^{(ij)} = \sum_{s_i,s_j} \mathrm{e}^{-\beta\delta_{s_i,s_j}} \chi_{s_i}^{i\to j}\chi_{s_j}^{j\to i} = 1 - \left(1 - \mathrm{e}^{-\beta}\right) \sum_{s} \chi_s^{i\to j}\chi_s^{j\to i}$$

We keep in mind that the Bethe free entropy is evaluated at a fixed point of the BP equations. Sometimes we will think of the Bethe free entropy as a function of all the messages $\chi^{i\to j}$ or of a parametrization of the messages.

We will always denote the free entropy with the index "Bethe" when we are evaluating it using the BP approximation. While on a tree we showed that $\Phi = \Phi_{\rm Bethe}$, in the case of a generic (even locally tree-like) graph we will discuss the relation between Φ and $\Phi_{\rm Bethe}$ (more precisely its global maximizers) in more detail in the next lectures.

We note that, so far, all we wrote depends explicitly on the graph $G(V, E)$ through the list of nodes V and edges E. No average over the graph (i.e. the disorder) was taken! This is rather different from the replica method, where the first step in the computation is in fact taking the average over the disorder.

6.4 Paramagnetic fixed point for graph colorings

We notice that for graph coloring $\chi_{s_j}^{j\to i} = q^{-1}$ for all (ij) and s_j is a fixed point of the BP equations on any graph $G(V, E)$. We will call this the paramagnetic fixed point. In general there might be, and often are, other fixed points, as we will discuss in next sections. To check that q^{-1} is indeed a fixed point of BP, call d_i the degree of node i (i.e. the number of neighbors of node i) and obtain the BP recursion

$$\frac{\left[1 - \left(1 - \mathrm{e}^{-\beta}\right)\frac{1}{q}\right]^{d_i-1}}{q\left[1 - (1 - \mathrm{e}^{-\beta})\frac{1}{q}\right]^{d_i-1}} = \frac{1}{q} \tag{6.8}$$

which is indeed always true. For the Bethe free entropy this paramagnetic fixed point gives us:

$$N\Phi_{\rm Bethe}(\beta) = \sum_{i=1}^{N} \log\left\{q\left[1 - \left(1 - \mathrm{e}^{-\beta}\right)\frac{1}{q}\right]^{d_i}\right\} - \sum_{(ij)\in E} \log\left\{q\cdot\frac{1}{q^2}\mathrm{e}^{-\beta} + (q^2 - q)\cdot\frac{1}{q}\right\}$$
$$\Phi_{\rm Bethe}(\beta) = c\log\left[1 - \left(1 - \mathrm{e}^{-\beta}\right)\frac{1}{q}\right] + \log(q) - \frac{c}{2}\log\left[1 - \left(1 - \mathrm{e}^{-\beta}\right)\frac{1}{q}\right]$$
$$= \log(q) + \frac{c}{2}\log\left[1 - \left(1 - \mathrm{e}^{-\beta}\right)\frac{1}{q}\right]$$

where we called $c = \sum_i d_i/N = 2M/N$ the average degree of the graph. For the average energy and entropy at inverse temperature β, we get

$$\text{energy} \quad e^* = -\frac{\partial \Phi_{\text{Bethe}}(\beta)}{\partial \beta} = \frac{c}{2} \frac{\mathrm{e}^{-\beta} \frac{1}{q}}{1-(1-\mathrm{e}^{-\beta})\frac{1}{q}} = \frac{c}{2} \underbrace{\frac{\mathrm{e}^{-\beta}}{(q-1)+\mathrm{e}^{-\beta}}}_{\text{prob. an edge is monochromatic}}$$

$$\text{entropy} \quad s(e^*) = \Phi_{\text{Bethe}}(\beta) + \beta e^* = \log(q) + \frac{c}{2}\log\left[1-\left(1-\mathrm{e}^{-\beta}\right)\frac{1}{q}\right] + \frac{\beta c}{2}\frac{\mathrm{e}^{-\beta}}{(q-1)+\mathrm{e}^{-\beta}}$$

Notice that these expressions give us a parametric form for $s(e)$. Sometimes we can even exclude β and write $s(e)$ in a closed form, but generically we simply plot parametrically $e(\beta), s(\beta)$. In the figure we take $q = 4$ colors and several values of the average degrees c.

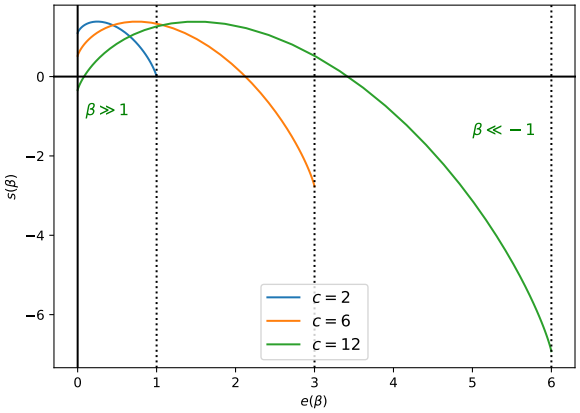


Figure 6.4.1: Entropy as a function on the energy cost for graph coloring with $q = 4$ colors corresponding to the paramagnetic fixed point of belief propagation. Average degree of the graph is c.

For a given pair of parameters c, q, the curve for $s(e)$ ranges from 0 to $c/2$ because at most all the edges can be violated. The curve achieves a maximum equal to $\log(q)$ at $e = c/2q$, because that is the typical cost of a random coloring (each edge is violated with probability $1/q$). The slope of the curve $s(e)$ corresponds to the inverse temperature β since $\frac{\partial s(e)}{\partial e} = \beta$.

Recalling that the entropy s is a logarithm of an integer (number of colorings with given energy), it is clear it should not take negative values. The fact that we see negative values for large energies and also for energy close to zero for $c = 12$ indicates that either coloring with such energy do not exist with high probability or that there was a flaw in what we did. Note also that so far we did not specify anything about the graph, except its average degree. Clearly we could have graphs with average degree $c = 6$ that contain one or several (even linearly many in N) 5-cliques and thus will not have any valid coloring. Thus the result we obtained cannot be valid for those graphs (indicating exponentially many valid 4-colorings $e = 0$ for $c = 6$). At this point, we thus restrict to random sparse graphs and we investigate whether the results we obtained are plausible at least in that case.

We notice, from the results we obtained for the paramagnetic fixed point, that $\chi^i = \frac{1}{q}$ for all i implies that values of energies larger than certain values strictly smaller than $c/2$ are not accessible, as they have negative entropy. At the same time, if all nodes had the same color this

would automatically achieve the energy $e = c/2$. To reconcile this paradox we must realize that the paramagnetic fixed point $\chi^i = \frac{1}{q}$ for all i assumes that every color is represented the same number of times, which is clearly not the case if all nodes have the very same color. With BP, it is often the case that there are several fixed points and we need to select the correct one. In general we need to find the fixed point with larger free entropy (from the saddle point method we know this is the one that dominates the probability measure). For the graphs coloring problems at large energy, i.e. $\beta < 0$ this motivates the investigation of a different fixed point that is able to break the equal representation of every color. We will call it the ferromagnetic fixed point.

6.5 Ferromagnetic Fixed Point

We now investigate whether the BP equations for graph coloring have fixed points of the following form for all $(ij) \in E$:

$$\chi_1^{i\to j} = a, \quad \chi_s^{i\to j} = b = \frac{1-a}{q-1}, \quad \forall\, s \neq 1 \tag{6.9}$$

Then, the graph coloring BP equations would read

$$\begin{aligned}
\chi_{s_j}^{j\to i} &= \frac{1}{Z^{j\to i}} \prod_{k\in\partial j\setminus i} \left[1 - \left(1 - \mathrm{e}^{-\beta}\right)\chi_{s_j}^{k\to j}\right] \\
a &= \frac{1}{Z^{j\to i}} \left[1 - \left(1 - \mathrm{e}^{-\beta}\right) a\right]^{d_j-1} =: \frac{1}{Z^{j\to i}} A^{d_j-1} \\
b &= \frac{1}{Z^{j\to i}} \left[1 - \left(1 - \mathrm{e}^{-\beta}\right) b\right]^{d_j-1} =: \frac{1}{Z^{j\to i}} B^{d_j-1}
\end{aligned}$$

with the normalization

$$Z^{j\to i} = (q-1)B^{d_j-1} + A^{d_j-1}, \quad a = \frac{A^{d_j-1}}{(q-1)B^{d_j-1} + A^{d_j-1}}, \quad b = \frac{B^{d_j-1}}{(q-1)B^{d_j-1} + A^{d_j-1}} \tag{6.10}$$

For such a ansatz to be a fixed point for every $(ij) \in E$ we need $d_j \equiv d$ for all j. This is the case with *random d-regular graphs* where every variable node has degree d and satisfies this condition. We could, of course, have a ferromagnetic fixed point where the a depends on (ij) and solve the corresponding distributional equations, but in this section we look for a simpler solution to illustrate the basic concepts and we will thus restrict to random d-regular graphs, $d \geq 3$. For d-regular random graphs we obtain the following self-consistent equation for the parameter a, given the degree d, inverse temperature β and number of colors q

$$a = \frac{\left[1 - \left(1 - \mathrm{e}^{-\beta}\right) a\right]^{d-1}}{\left[1 - \left(1 - \mathrm{e}^{-\beta}\right) a\right]^{d-1} + (q-1)\left[1 - \left(1 - \mathrm{e}^{-\beta}\right)\frac{1-a}{q-1}\right]^{d-1}} =: \mathsf{RHS}(a;\beta,d,q) \tag{6.11}$$

To express the Bethe free entropy corresponding to this ferromagnetic fixed point we use eq. (6.7) and plug (6.9) in it to get:

$$\begin{aligned}
\Phi_{\mathrm{Bethe}}(\beta) &= \log\left\{(q-1)\left[1 - (1-\mathrm{e}^{-\beta})\frac{1-a}{q-1}\right]^d + \left[1 - (1-\mathrm{e}^{-\beta})a\right]^d\right\} \\
&- \frac{d}{2}\log\left\{1 - (1-\mathrm{e}^{-\beta})\left[\frac{(1-a)^2}{q-1} + a^2\right]\right\}
\end{aligned} \tag{6.12}$$

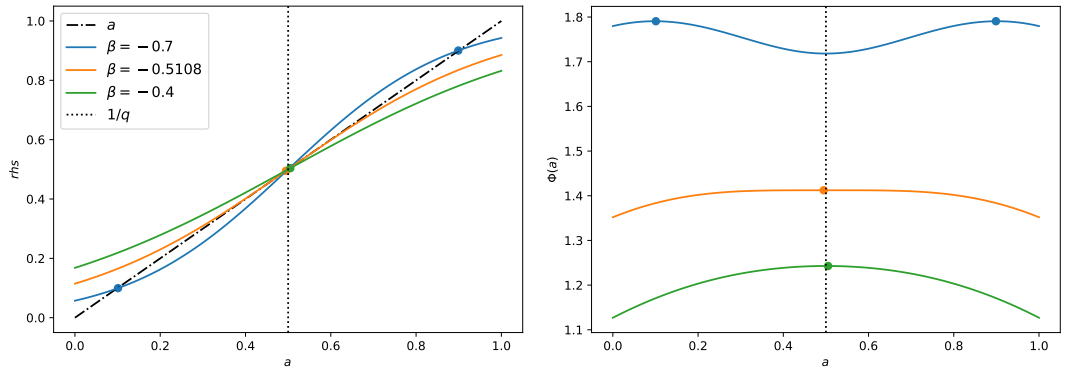


Figure 6.5.1: Illustration of a second order (continuous) phase transition for 2-coloring on 5-regular graph and several values of the inverse temperature β. In the left panel, the right hand side of eq. (6.11) is plotted against the parameter a. In the right panel, the Bethe free entropy is plotted for the same inverse temperatures. The stable fixed points are marked and correspond to the local maxima of the Bethe free entropy.

In the last expressions we can think of the Bethe free entropy as a function of the parameter a, keeping in mind that we are seeking to evaluate it at the global maximizer.

6.5.1 Ising ferromagnet, $q = 2$

In Fig. 6.5.1 we plot the left hand side and the right hand size of eq. (6.11) as a function of a, for a given value of degree d and number of colors q and several values of the inverse temperature β. We also plot the Bethe free entropy as a function of a.

We observe that for $\beta = -0.4$ (green curve) the only fixed point of (6.11) and maximum of (6.12) is reached at $a = 1/q$. This is the paramagnetic fixed point we investigated previously. For $\beta = -0.7$ (blue curve) we see, however, a different picture. The fixed point $a = 1/q$ is unstable under iterations of (6.11) and corresponds to a local minimum of the Bethe free entropy. There are two new stable fixed points that appear and correspond to the maxima of the Bethe entropy.

At what value of the inverse temperature β do the additional fixed points appear? For this we need to evaluate the stability of the paramagnetic fixed point, i.e. when the derivative at the paramagnetic fixed point $\frac{\partial \mathsf{RHS}}{\partial a}\big|_{a=1/q} = 1$:

$$1 = \frac{\partial \mathsf{RHS}}{\partial a}\bigg|_{a=\frac{1}{q}} = -\frac{(d-1)(1-\mathrm{e}^{-\beta})}{\mathrm{e}^{-\beta}+(q-1)} \quad \Rightarrow \quad \beta_{\mathrm{stab}} = -\log\left(1+\frac{q}{d-2}\right) \tag{6.13}$$

We can investigate other values of the degree d and still two colors $q = 2$ and we will observe the same picture for all d. We recognize the second order phase transition from a paramagnet to a ferromagnet that we saw already in the Curie-Weiss model. Indeed the only difference between the present model and the Curie-Weiss model is that the graph of interactions was fully connected in the Curie-Weiss model while it is a d-regular random graph in the present case.

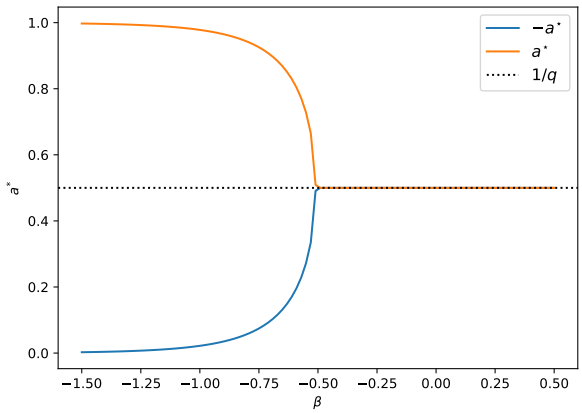


Figure 6.5.2: Magnetization for 2-coloring on 5-regular graph as a function of the inverse temperature β.

This is the Bethe approximation to the solution of the Ising model on regular cubic lattices. Of course lattices are not trees, but if we match the degree of the random graph to the coordination number of a cubic lattice in D dimension we obtain $d = 2D$ and we can observe that for $D = 1$ ($d = 2$) the $\beta_{\rm stab} = -\infty$. This is actually an exact solution because the 1-dimensional cubic lattice is just a chain which is a tree-graph. Thus the BP solution is exact. For a 2-dimensional Ising model that has been famously solved by Onsager we have $\beta_{\rm Onsager} = -\log(1+\sqrt{2}) = -0.881$ which is relatively close to the Bethe approximation $\beta_{\rm stab}(d=4) = -0.693$. Note that the sign is opposite from what can be found in the literature because here we defined positive temperature for the coloring (the anti-ferromagnet), and also that there is a multiplicative factor two as here the energy cost for a variable change is 1 (whereas in the usual Ising model it is 2). For the 3-dimensional Ising model no closed form solution exists yet, the critical temperature has been evaluated numerically to very high precision and reads $\beta_{3D} = -0.4433$, again to be compared with its Bethe approximation $\beta_{\rm stab}(d=6) = -0.4055$ which is remarkably close. As the degree grows the Bethe approximation actually gets closer and closer to the finite dimensional values. And eventually as $d \to \infty$ we recover exactly the Curie-Weiss solution that we studied in the first lecture with a proper rescaling on the interaction strength. You will show this for a homework.

6.5.2 Potts ferromagnet $q \geq 3$

For more than two colors, $q \geq 3$, we find a somewhat different behaviour leading to a 1st order phase transition. Let us again start by plotting the fixed point equations and the Bethe free entropy in Fig. 6.5.3. We see that, as before, $\beta_{\rm stab}$ marks the inverse temperature at which the paramagnetic fixed point $1/q$ becomes unstable and the corresponding Bethe entropy maximum becomes a minimum. But there is another stable fixed point, corresponding to a local maximum of the Bethe entropy appearing at $\beta_{\rm s} > \beta_{\rm stab}$. This inverse temperature where a new stable fixed point appears discontinuously is called the spinodal temperature in physics.

When there are more than one stable fixed points, more than one local maximas in Bethe free entropy, eq. (6.12), we must compare their free entropies, the larger one dominates the corresponding saddle point and hence is the correct solution.

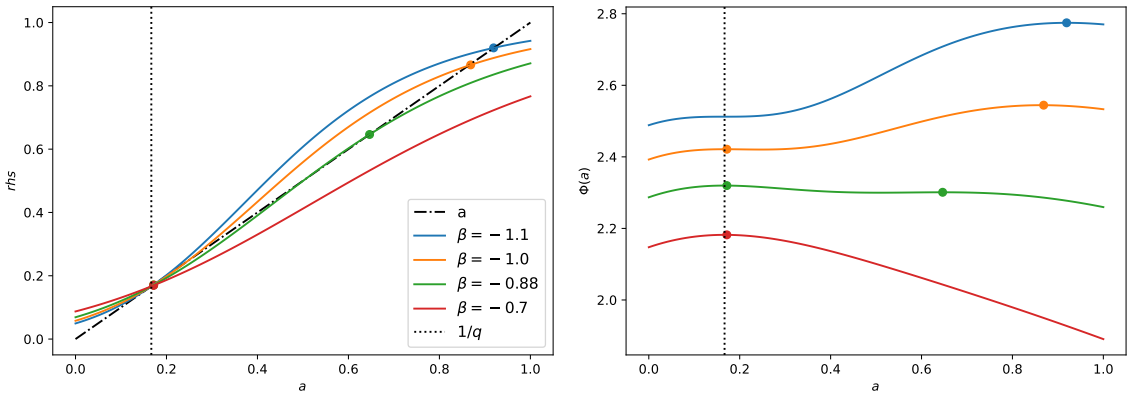


Figure 6.5.3: Illustration of a first order (discontinuous) phase transition for 6-coloring on 5-regular graph and several values of the inverse temperature β. In the left panel, the right hand side of eq. (6.11) is plotted against the parameter a. In the right panel, the Bethe free entropy is plotted for the same inverse temperatures. The stable fixed points are marked and correspond to the local maxima of the Bethe free entropy.

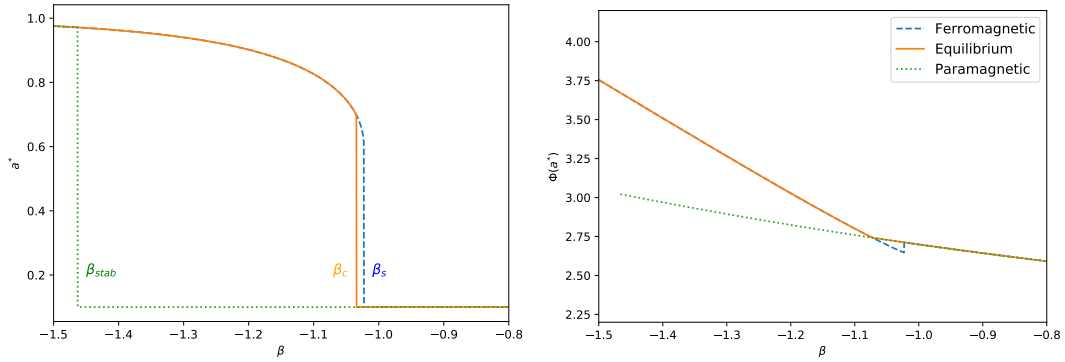


Figure 6.5.4: The magnetization (left panel) and Bethe free entropy as a function of the inverse temperature β for $q = 10$ and $d = 5$. (Note a different values of q from previous figure.)

The inverse temperature at which the ferromagnetic fixed point $a > 1/q$ becomes the global maximum of the Bethe entropy, instead of the paramegnetic fixed point, will be denoted β_c and corresponds to a first order phase transition. We have $\beta_{\rm stab} < \beta_c < \beta_s$. The order parameter a changes discontinuously at β_c in a first order phase transition. In the case of $q = 2$, instead, we saw a continuous 2nd order phase transition, where the ferromagnetic fixed points appear at the same temperature at which the corresponding free entropy becomes a global maximum, together with the instability of the paramagnetic fixed point. In other words, in 2nd order phase transitions $\beta_s = \beta_c = \beta_{\rm stab}$.

With the knowledge of the ferromagnetic point we can hence correct the result for the number of colorings at a given cost, in Fig. 6.5.5.

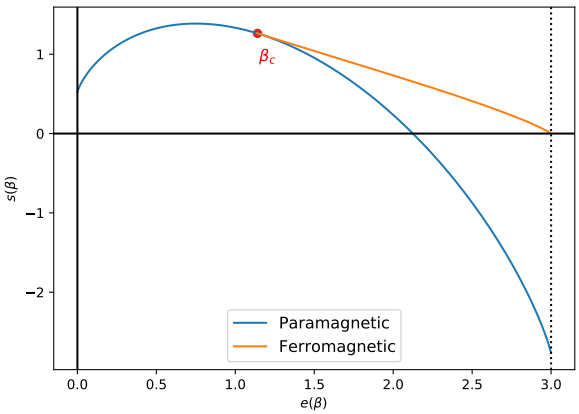


Figure 6.5.5: Entropy as a function of the energy for graph coloring with $q = 4$ colors and average degree $c = d = 6$ (same as Fig. 6.4.1).

Two important comments are in place:

- The results obtained with BP on random graphs are exact for $\beta < 0$ in the sense that

$$\forall\, \varepsilon > 0, \qquad Pr\left(|\Phi(\beta) - \Phi_{\text{Bethe}}(\beta)| < \varepsilon\right) \overset{N\to\infty}{\longrightarrow} 1 \quad \text{w.h.p} \tag{6.14}$$

  - For $q = 2$ this has been proven by Dembo & Montanari 2010 Dembo et al. (2010b) (on sparse random graphs even beyond d-regular).
  - For $q \geq 3$ this has been proven by Dembo, Montanari, Sly, Sun, 2014 Dembo et al. (2014) on d-regular graphs.

  The situation for $\beta > 0$ is much more involved and will be treated in the following.

- We note that the paramagnetic and ferromagnetic fixed points are BP fixed points even for finite size regular graphs, not only in the limit $N \to \infty$. But of course not every regular graph has the same number of colorings of a given energy, especially not the non-random ones. Also, the discussed phase transitions exist only in the limit $N \to \infty$, but BP fixed point behave the same way we discussed even at finite size N. This means that BP ignores part of the finite size effects and gives us an interesting proxy of a phase transition even in finite size, where formally no phase transition can exists since the free entropy is an analytic function for every finite N.

6.6 Back to the anti-ferromagnetic ($\beta > 0$) case

Let us now go back to the anti-ferromagnetic case, $\beta > 0$, with $\beta \to \infty$ corresponding to the original coloring problem on random sparse graph. We aim to obtain a condition on the average degree for which the graphs are with high probability colorable, or not. We also want to know, in the phase where colorings exist with high probability, how many of them are there, i.e. the value of $s(e = 0)$.

Previously, we evaluated the entropy $s(e)$ of the paramagnetic fixed point $\chi_s^{i\to j} \equiv \frac{1}{q}$. We can observe that the corresponding Bethe free entropy is equal to the so-called *annealed* free entropy.

To define this term, let us remind the general probability distribution we are considering:

$$P\left(\{s_i\}_{i=1}^N\right) = \frac{1}{Z_G}\prod_{i=1}^N g_i(s_i)\prod_{a=1}^M f_a\left(\{s_i\}_{i\in\partial a}\right). \tag{6.15}$$

The free entropy $\Phi_G = \frac{1}{N}\log(Z_G)$ then depends explicitly on the graph G. We expect the free entropy to be self-averaging, i.e. concentrating around its mean as

$$\forall\,\varepsilon > 0, \qquad \Pr\left(|\Phi_G - \mathbb{E}_G[\Phi_G]| > \varepsilon\right) \overset{N\to\infty}{\longrightarrow} 0 \tag{6.16}$$

When $N \to \infty$, computing Φ_G and $\mathbb{E}_G[\Phi_G]$ should thus lead to the same result.

We define:

$$\begin{aligned}\text{quenched free entropy} \quad & \Phi_{\text{quench}} \equiv \mathbb{E}_G\left[\frac{1}{N}\log(Z_G)\right]\\ \text{annealed free entropy} \quad & \Phi_{\text{anneal}} \equiv \frac{1}{N}\log\left(\mathbb{E}_G[Z_G]\right)\end{aligned}$$

Naively, we could expect to have $\left(\mathbb{E}[\log(Z)] - \log(\mathbb{E}[Z])\right)/N \to 0$, but this is often not the case. The partition function is of order $Z = O\left(\exp(N)\right)$ as $N \to \infty$, and concentration holds only for $\frac{1}{N}\log(Z)$. The annealed free entropy can get dominated by rare instances of the graph G. Let us give a simple example.

Consider an artificial problems where we have

$$Z_G = \begin{cases} \mathbb{e}^N, & \text{w.p. } 1-\mathbb{e}^{-N} \\ \mathbb{e}^{3N}, & \text{w.p. } \mathbb{e}^{-N}\end{cases}$$

The quenched and annealed averages are then given by:

$$\begin{aligned}\mathbb{E}_G\left[\frac{1}{N}\log(Z_G)\right] &= \left(1-\mathbb{e}^{-N}\right) + \mathbb{e}^{-N}3 = 1 + \mathbb{e}^{-N}(2) \to 1\\ \frac{1}{N}\log\left(\mathbb{E}_G[Z_G]\right) &= \frac{1}{N}\log\left\{\mathbb{e}^N - 1 + \mathbb{e}^{2N}\right\} = 2 + \frac{1}{N}\log\left(1 + \mathbb{e}^{-N} - \mathbb{e}^{-2N}\right) \to 2\end{aligned}$$

We see that while the quenched entropy represents the typical values, the annealed entropy got influenced by exponentially rare values and could completely mislead us about the properties of the typical instance. In general, since $\log(\cdot)$ is a concave function, by Jensen's inequality we have $\Phi_{\text{anneal}} \geq \Phi_{\text{quench}}$, therefore the annealed free entropy will at least provide us with an upper bound. Of course, Φ_{anneal} is usually much easier to compute than Φ_{quench}.

For the coloring problem, let $G(N, M)$ represent a random graph with N nodes, and M edges chosen at random among all possible edges. The annealed free entropy then follows from

$$\mathbb{E}_{G(N,M)}[Z_G(\beta)] = q^N \mathbb{E}_{\{s_i\}_{i=1}^N}\mathbb{E}_{G(N,M)}\left[\mathbb{e}^{-\beta\sum_{(ij)\in E}\delta_{s_i,s_j}}\right] = q^N\left[\underbrace{\mathbb{e}^{-\beta}\frac{1}{q} + \left(1-\frac{1}{q}\right)}_{\text{free entropy of one edge}}\right]^M$$

Here we use the fact that edges in the random graph are independent and that, for $\beta > 0$, the contributions to the average are dominated by colorings where each color is represented roughly equally. The annealed entropy is thus positive and vanishes at average degree

$$c_{\rm ann}(\beta) = -\frac{2\log q}{\log\left[1-(1-\mathrm{e}^{-\beta})\frac{1}{q}\right]} \tag{6.17}$$

Notice that the annealed free entropy and the Bethe one corresponding to the paramagnetic fixed point are the same $\Phi_{\rm anneal} = \Phi_{\rm Bethe}|_{\chi\equiv\frac{1}{q}}$.

The question is whether the annealed/paramagnetic free entropy is correct for all average degrees c and inverse temperatures $\beta > 0$. Unfortunately, the answer is negative. For instance for $\beta \to 0$ in Coja-Oghlan (2013) the authors show that all proper colorings disappear with high probability for values of average degree strictly smaller than the average degree $c_{\rm ann}(\beta\to\infty)$ at which $\Phi_{\rm anneal}(\beta\to\infty)$ becomes negative. This means that $\Phi_{\rm anneal}$ cannot be equal to the quenched free entropy all the way to $c_{\rm ann}$.

What could possibly go wrong with the paramagnetic fixed point $\chi_s^{i\to j} \equiv \frac{1}{q}$? One immediate thing we should ask is whether belief propagation converges to this fixed point on large random graphs. Is the paramagnetic fixed point even a stable one, i.e. if we initialize as $\chi_s^{i\to j} = \frac{1}{q} + \varepsilon_s^{i\to j}$ (with $\sum_s \varepsilon_s^{i\to j} = 0$ due to normalization), does BP converge back to $\chi_s^{i\to j} \equiv \frac{1}{q}$? To investigate this question one could implement the iterations on a large single graph and simply try out.

A computationally more precise way to find the answer to these questions is to perform the linear stability analysis. Let $\theta = 1-\mathrm{e}^{-\beta}$, and consider the first-order Taylor expansion around the fixed point $\chi \equiv \frac{1}{q}$ (or equivalently $\varepsilon_{s_k}^{k\to j} \equiv 0$)

$$\chi_{s_j}^{j\to i}(t+1) = \frac{1}{q} + \sum_{k\in\partial j\setminus i}\sum_{s_k} \left.\frac{\partial\chi_{s_j}^{j\to i}(t+1)}{\partial\chi_{s_k^{k\to i}}(t)}\right|_{\chi\equiv\frac{1}{q}} \varepsilon_{s_k}^{k\to j}$$

Note that

$$\frac{1}{q} = \frac{1}{Z^{j\to i}|_{\chi\equiv\frac{1}{q}}}\prod_{k\in\partial j\setminus i}\left(1-\theta\frac{1}{q}\right) = \frac{1}{Z^{j\to i}|_{\chi\equiv\frac{1}{q}}}\left(1-\frac{\theta}{q}\right)^{d_j-1}, \quad Z^{j\to i}\big|_{\chi\equiv\frac{1}{q}} = q\left(1-\frac{\theta}{q}\right)^{d_j-1} \tag{6.18}$$

$$\begin{aligned}
\left.\frac{\partial\chi_{s_j}^{j\to i}(t+1)}{\partial\chi_{s_k}^{k\to j}(t)}\right|_{\chi\equiv\frac{1}{q}} &= \frac{1}{Z^{j\to i}}\frac{\partial}{\partial\chi_{s_k}^{k\to j}(t)}\left[\prod_{\ell\in\partial j\setminus i}\left(1-\theta\chi_{s_j}^{\ell\to j}(t)\right)\right]\Bigg|_{\chi\equiv\frac{1}{q}} \\
&\quad - \frac{1}{(Z^{j\to i})^2}\prod_{\ell\in\partial j\setminus i}\left(1-\theta\chi_{s_j}^{\ell\to j}(t)\right)\left.\frac{\partial Z^{j\to i}}{\partial\chi_{s_k}^{k\to j}(t)}\right|_{\chi\equiv\frac{1}{q}} \\
&= \delta_{s_j,s_k}\frac{-\theta\left(1-\frac{\theta}{q}\right)^{d_j-2}}{Z^{j\to i}|_{\chi\equiv\frac{1}{q}}} - \frac{\left(1-\frac{\theta}{q}\right)^{d_j-1}}{\left(Z^{j\to i}_{\chi\equiv\frac{1}{q}}\right)^2}\left[-\theta\left(1-\frac{\theta}{q}\right)^{d_j-2}\right] \\
&= -\delta_{s_j,s_k}\frac{\theta}{q-\theta} + \frac{\theta}{q(q-\theta)}
\end{aligned}$$

In order to proceed, we define a $q \times q$ matrix $\mathbf{T}$ such that:

$$T_{jk} = \begin{cases} a := \frac{\theta(1-q)}{q(q-\theta)}, & \text{if } j = k \\ b := \frac{\theta}{q(q-\theta)}, & \text{if } j \neq k \end{cases} \quad \Rightarrow \quad \mathbf{T} = \begin{bmatrix} a & b & \cdots & b \\ b & a & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ b & \cdots & b & a \end{bmatrix} \tag{6.19}$$

Notice that the matrix T has $q-1$ degenerate eigenvalues

$$\lambda_{\max} = a - b = \frac{-\theta}{q-\theta} \Rightarrow \begin{cases} < 0, & \text{for } \beta > 0 \\ > 0, & \text{for } \beta < 0 \end{cases}, \quad \text{with eigenvector } \begin{bmatrix} +1 & -1 & 0 & \cdots & 0 \end{bmatrix}^{\mathrm{T}}$$

and a single zero eigenvalue.

Keeping this in mind, we can see that the linear expansion of the BP equations for ε's reads:

$$\varepsilon_{s_j}^{j\to i}(t+1) = \sum_{k\in\partial j\setminus i} \sum_{s_k} T_{s_j,s_k} \varepsilon_{s_k}^{k\to j}(t)\,. \tag{6.20}$$

We now define the *excess degree distribution*, $\tilde{p}_k = (k+1)p_{k+1}/c$, for a random graph ensemble with degree distribution p_k and average degree c. $\tilde{p}_k$ represents the probability that a randomly chosen edge is incident to a node has k other edges except the chosen edge, i.e. has degree $k+1$. And similarly, let $\tilde{c} = \sum_k k\tilde{p}_k$ be the *average excess degree*. With it we obtain:

$$\langle \varepsilon(t+1) \rangle = \tilde{c}\,\lambda_{\max} \langle \varepsilon(t) \rangle\ , \tag{6.21}$$

where the $\langle\cdot\rangle$ is the average over edges. A phase transition occurs at $\tilde{c}\,\lambda_{\max} = 1$ that determines whether $\lim_{t\to\infty} \langle \varepsilon(t+1) \rangle$ blows up to infinity or converges to zero.

$$\tilde{c}\,\lambda_{\max} = \frac{-\tilde{c}\left(1 - \mathrm{e}^{-\beta_{\mathrm{stab}}}\right)}{q - \left(1 - \mathrm{e}^{-\beta_{\mathrm{stab}}}\right)} = 1 \quad \Rightarrow \quad \beta_{\mathrm{stab}} = -\log\left(1 + \frac{q}{\tilde{c}-1}\right) \tag{6.22}$$

This is the stability transition that we have already computed for the ferromagnetic solution for $\beta < 0$. For $\beta > 0$ we notice that $\lambda_{\max} < 0$ and thus the corresponding instability corresponds to an oscillation from one color to another at each parallel iteration. Such an oscillatory behaviour would be possible on a bipartite graph, where indeed one side of the graph could have one color and the other side another color. But this two-color solution is not compatible with the existence of many loops of even and odd length in random graphs. The temperature β_{stab} will thus not have any significant bearing on the anti-ferromagnetic case $\beta > 0$.

Is it possible that the mean of the perturbation $\langle\varepsilon\rangle \to 0$ but the variance $\langle\varepsilon^2\rangle \nearrow \infty$? Let us investigate:

$$\left\langle \left(\varepsilon_{s_j}^{j\to i}(t+1)\right)^2 \right\rangle = \left\langle \sum_{k\in\partial j\setminus i} \left[\sum_{s_k} T_{s_j,s_k} \varepsilon_{s_k}^{k\to j}(t)\right]^2 + \underbrace{\sum_{\substack{k,\ell\in\partial j\setminus i \\ k\neq\ell}} \left[\sum_{s_k} T_{s_j,s_k} \varepsilon_{s_k}^{k\to j}(t)\right]\left[\sum_{s_\ell} T_{s_j,s_\ell} \varepsilon_{s_\ell}^{\ell\to j}(t)\right]}_{=0 \text{ since neighbors are independent } \left\langle \varepsilon_{s_k}^{k\to j} \varepsilon_{s_\ell}^{\ell\to j} \right\rangle = 0} \right\rangle$$

$$= \tilde{c} \left\langle \left[\sum_{s_k} T_{s_j,s_k} \varepsilon_{s_k}^{k\to j}(t)\right]^2 \right\rangle .$$

Therefore, the variance will be determined by the maximum eigenvalue $\lambda_{\max}$ of $\mathbf{T}$, such that

$$\mathsf{Var}(t+1) = \tilde{c}\lambda_{\max}^2 \mathsf{Var}(t) \tag{6.23}$$

We can thus distinguish two cases for the graph coloring problem:

- For $\tilde{c} < \frac{(q-\theta)^2}{\theta^2}$ BP converge back to $\chi \equiv \frac{1}{q}$. Specifically for $\beta \to \infty$ we get

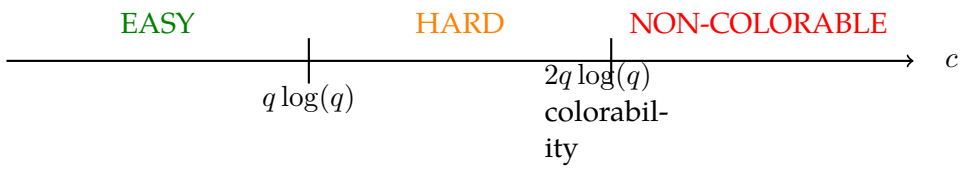
$$\tilde{c}_{\rm KS} = \frac{(q-\theta)^2}{\theta^2} \to (q-1)^2 \tag{6.24}$$

- For $\tilde{c} > \frac{(q-\theta)^2}{\theta^2}$ BP goes away from $\frac{1}{q}$ and actually does not converege.

Note that the abbreviation KS comes from the works on Kesten and Stigum (1967), and in physics this transition is related to the works of de Almeida and Thouless (1978).

For Erdös-Rényi random graphs, where the average degree and the excess degree are equal $\tilde{c} = c$, in the setting $q = 3$ and $\beta \to \infty$ we have that $c_{\rm KS}(q=3) = 4 < c_{\rm ann}(q=3) = 5.4$. $c_{\rm KS}$ in this case is also smaller that the upper bound on colorability threshold from Coja-Oghlan (2013). At average degree $c > 4$ the BP equations do not converge anymore and another approach, based on replica symmetry breaking, will be needed to understand what is going on. It is interesting to note that algorithms that are provably able to find proper 3-coloring exist for all $c < 4.03$ Achlioptas and Moore (2003), so even slightly above the threshold $c_{\rm KS}(q=3) = 4$.

For $q \geq 4$ and $\beta \to \infty$ we have that $c_{\rm KS} > c_{\rm ann}$ and the investigated instability cannot be used to explain what goes wrong in the colorable regime. An important motivation to understand what is happening comes from the algorithmic picture that appears for large values of the number of colors q. In that case the annealed upper bound on colorability scales as $2q \log q$, and probabilistic lower bounds (see e.g. Coja-Oghlan and Vilenchik (2013)) imply that this is indeed the right scaling for the colorability threshold. Yet, for what concerns polynomial algorithms that provably find proper colorings, we only know of algorithms that work up to degree $q \log q$ Achlioptas and Coja-Oghlan (2008), i.e. up to half of the colorable region. Design of tractable algorithms able to find proper coloring for average degrees $(1+\epsilon)q \log q$ is a long-standing open problem. What is happening in the second half of the colorable regime will be clarified in the follow up lectures.



6.7 Exercises

Exercise 6.1: Bethe free entropy

(a) Show from the generic formula for Bethe free entropy density that we derived in the lecture

$$\Phi_{\text{Bethe}}^{\text{general}} = \frac{1}{N}\sum_{i=1}^{N}\log Z^i + \frac{1}{N}\sum_{a=1}^{M}\log Z^a - \frac{1}{N}\sum_{ia}\log Z^{ia} \qquad (\dagger)$$

where

$$\begin{aligned} Z^i &= \sum_s g_i(s)\prod_{a\in\partial i}\psi_s^{a\to i} \\ Z^a &= \sum_{\{s_i\}_{i\in\partial a}} f_a\left(\{s_i\}_{i\in\partial a}\right)\prod_{i\in\partial a}\chi_{s_i}^{i\to a} \\ Z^{ia} &= \sum_s \psi_s^{a\to i}\chi_s^{i\to a} \end{aligned}$$

that the Bethe free entropy density for graph coloring can be written as

$$\Phi_{\text{Bethe}}^{\text{coloring}} = \frac{1}{N}\sum_{i=1}^{N}\log Z^{(i)} - \frac{1}{N}\sum_{(ij)\in E}\log Z^{(ij)}$$

where

$$\begin{aligned} Z^{(i)} &= \sum_s\prod_{k\in\partial i}\left[1-\left(1-\mathrm{e}^{-\beta}\right)\chi_s^{k\to i}\right] \\ Z^{(ij)} &= 1-\left(1-\mathrm{e}^{-\beta}\right)\sum_s\chi_s^{i\to(ij)}\chi_s^{j\to(ij)} \end{aligned}$$

Exercise 6.2: Fully connected limit from Belief Propagation

The goal of this exercise is to show that, when the average degree is large, the Belief Propagation solution gives back the fully connected limit, and in particular the one obtained with the "fully-connected" cavity method shown in eq. (1.22). We shall do this starting from the Potts ferromagnetic model, using $q = 2$ to make the connection with the ferromagnetic solution of section 1.4.

(a) First, we check that we can find back the mean field self-consistent equation. To do so, first rewrite the recursion relation derived in section 6.5 using $q = 2$, and denoting $\chi_\pm = (1 \pm m)/2$, show that the belief propagation recursion fixed point can be written as

$$m = -\frac{\left(\cosh\frac{\beta}{2} + m\sinh\frac{\beta}{2}\right)^{d-1} - \left(\cosh\beta - m\sinh\frac{\beta}{2}\right)^{d-1}}{\left(\cosh\frac{\beta}{2} + m\sinh\frac{\beta}{2}\right)^{d-1} + \left(\cosh\frac{\beta}{2} - m\sinh\frac{\beta}{2}\right)^{d-1}}\,. \qquad (6.25)$$

Using the relation $\operatorname{atanh} x = \frac{1}{2}\log\frac{1+x}{1-x}$, show that this yields

$$m = -\tanh\left[(d-1)\operatorname{atanh}\left(m\tanh\left(\frac{\beta}{2}\right)\right)\right]. \qquad (6.26)$$

(b) We now move the large d limit, and use $\beta = -\frac{2\beta_{\rm MF}}{d}$. Why is this necessary in order to recover the fully connected model? Show that it leads indeed to the mean field equation

$$m = \tanh\left(\beta_{\rm MF} m\right). \tag{6.27}$$

Another approach is to look directly to the free entropy, and to recover directly the free entropy expression of the fully connected model by taking the large connectivity limit of the Bethe free entropy:

(a) Show that the Bethe free entropy reads, according to Belief Propagation:

$$\Phi(\beta, m) = \log\left(\left(1-(1-e^{-\beta})\frac{1+m}{2}\right)^d + \left(1-(1-e^{-\beta})\frac{1-m}{2}\right)^d\right) - \frac{d}{2}\log\left(1-(1-e^{-\beta})\frac{1}{2}\left[1+m^2\right]\right)$$

(b) We use again $\beta = \frac{2\beta_{\rm MF}}{d}$. Show that, to leading order in d, we recover, as $d \to \infty$ the expression (1.22). Hint: First rewrite the expressions inside the log of first line as

$$\left(1-(1-e^{-\beta})\frac{1\pm m}{2}\right)^d = e^{d\log\left[1-(1-e^{-\frac{2}{d}\beta_{\rm MF}})\frac{1\pm m}{2}\right]}. \tag{6.28}$$

Notice however that there is an additional trivial term $\beta_{\rm MF}/2$ and explain why this additional term is here.

Exercise 6.3: Belief propagation for the matching problem on random graphs

Consider now the matching problem on sparse random graphs. Use the graphical model representation from the previous homework.

(a) Write belief propagation equations able to estimate the marginals of the probability distribution

$$P\left(\left\{S_{(ij)}\right\}_{(ij)\in E}\right) = \frac{1}{Z(\beta)}\prod_{(ij)\in E} \mathrm{e}^{\beta S_{(ij)}} \prod_{i=1}^{N} \mathbb{I}\left(\sum_{j\in\partial i} S_{(ij)} \leq 1\right)$$

Be careful that in the matching problem the nodes of the graph play the role of factor nodes in the graphical model and edges in the graph carry the variable nodes in the graphical model.

(b) Write the corresponding Bethe free entropy in order to estimate $\log(Z(\beta))$. Use results of the previous homework to suggest how to estimate the number of matchings of a given size on a given randomly generated large graph G.

(c) Consider now d-regular random graphs and draw the number of matchings as a function of their size for several values of d. Comment on what you obtained, does it correspond to your expectation? If not, explain the differences.

Part II

Probabilistic Inference

Chapter 7

Denoising, Estimation and Bayes Optimal Inference

> Si un événement peut être produit par un nombre n de causes différentes, les probabilités de l'existence de ces causes prises de l'événement sont entre elles comme les probabilités de l'événement prises de ces causes, et la probabilité de l'existence de chacune d'elles est égale à la probabilité de l'événement prise de cette cause, divisée par la somme de toutes des probabilités de l'événement prises de chacune de ces causes.
>
> Pierre Simon de Laplace – 1774

7.1 Bayes-Laplace inverse problem

In this chapter, we shall discuss the estimation, or the learning, of a quantity that we do not know directly, but only through some indirect, noisy measurements. They are actually many different ways to think of the problem depending on whether we are in the context of signal processing, Bayesian statistics, or information theory, but it boils down to separating the signal from the noise in some data.

To be concrete, consider the following situation: assume an unknown signal $\mathbf{x}$ (a vector, or a scalar, or a matrix...) is generated from a known distribution $P_X(\mathbf{x})$. We would like to know this signal, to "estimate" its value. We are not given $\mathbf{x}$, however, but instead a measurement $\mathbf{y}$ obtained through some noisy process, whose characteristic are also known. In other words we have: $X \to Y$, with

$$X \sim P_X(\mathbf{x}), \qquad Y \sim P_{Y|X}(\mathbf{y}|\mathbf{x})\,, \tag{7.1}$$

and we aim at finding back, in the best way we could, $\mathbf{x}$.

This is the setting of Bayesian estimation. P_X is called the "prior" distribution, as it tells us what we know on the variable X *a priori*, before any measurement is done. $P_{Y|X}$ is telling us the probability to obtain a given result $\mathbf{y}$, given the value of $\mathbf{x}$. Seen as a function of $\mathbf{x}$ for a

given value of $\mathbf{y}$, $\mathcal{L} = (\mathbf{x}; \mathbf{y}) = \mathbb{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}, \mathbf{x})$ is called the Likelihood of $\mathbf{x}$. What we are really interested in, however, is the value of $\mathbf{x}$ (the signal) if we measure $\mathbf{y}$ (the data). This is called the posterior probability of X given Y: $\mathbb{P}_{X|Y}(x, y)$. To obtain the latter with the former, we follow the direction given by Laplace and Bayes in the late XVIII century, and we just write the celebrated "Bayes" formula:

$$\mathbb{P}_{X|Y}(x, y) = \frac{\mathbb{P}_{Y|X}(x, y)\mathbb{P}_X(x)}{\mathbb{P}_Y(y)} \tag{7.2}$$

so that the posterior $\mathbb{P}_{X|Y}(x, y)$ is given by the product of the prior probability on X, $\mathbb{P}_X(x)$, times the likelihood $\mathbb{P}_{Y|X}(x, y)$, divided by the "evidence" $\mathbb{P}_Y(y)$ (which is just a normalization constant). Of course, if we deal with continuous variable, we can write the same formula with probability density instead:

$$P_{X|Y}(x, y) = \frac{P_{Y|X}(x, y)P_X(x)}{P_Y(y)} \tag{7.3}$$

Note that this Bayesian setting is not entirely general! Unfortunately, we often do not know what P_X is in many estimation problems (and sometime, we do not know $P_{Y|X}$ either) which makes the use of this formalism tricky (and has generated a long standing dispute between so-called frequentist and Bayesian statisticians), but in this chapter, we shall forget about these problems, and restricted ourselves to the situation where we *do* know these distributions, so that one can safely use Bayesian statistics. There are many concrete problems where this is the case (central to fields such as information theory and error correction, signal processing, denoising, ...) and so this will be enough to keep us busy for some time. We will move to more complicated situations later.

7.2 Scalar estimation

7.2.1 Posterior distribution

Let us look at three concrete problems where the "true value", a scalar that we shall denote x^*, is generated by:

1. A Rademacher random variable : $X = \pm 1$ with probability $1/2$.
2. A Gaussian random variable with mean 0 and variance 1: $X \sim \mathcal{N}(x; 0, 1)$.
3. A Gauss-Bernoulli random variable that is 0 with probability $1/2$, and a Gaussian with mean 0 and variance 1 otherwise: $X \sim \mathcal{N}(x; 0, 1)/2 + \delta(x)/2$.

We shall concentrate on noisy measurements with Gaussian noise. In this case, we are given n measurements

$$y_i = x^* + \sqrt{\Delta} z_i, \qquad i = 1, \cdots, N \tag{7.4}$$

with $z_i \sim \mathcal{N}(0,1)$ a standard Gaussian noise, with zero mean and unit variance. Following Bayes formula, we can now compute the posterior probability for our estimate of x^* as:

$$P_{X|Y}(x,\mathbf{y}) = \frac{1}{P_Y(\mathbf{y})}\frac{1}{(2\pi\Delta)^{N/2}} e^{-\sum_i \frac{(y_i-x)^2}{2\Delta}} P_X(x) \tag{7.5}$$

The posterior tells us all there is to know about the inference of the unknown x^*.

For instance, in the Rademacher example, if we are given the five measurement numbers:

$$1.04431591, 2.55352006, 1.43665582, 1.37069702, 0.77697312\,. \tag{7.6}$$

It is very likely that the $x^* = 1$ rather than $x^* = -1$. How likely? We can compute explicitly the posterior and find

$$P^{\text{Rademacher}}_{X|Y}(x|\mathbf{y}) = \frac{1}{1+e^{-2x\sum\limits_{i=1}^{N}\frac{y_i}{\Delta}}} = \sigma\left(2x\sum_{i=1}^{N}\frac{y_i}{\Delta}\right) \tag{7.7}$$

where $\sigma(x)$ is the sigmoid function. We thus find that we can estimate the probability that $x^* = 1$ to be larger that 0.999. So we are indeed pretty sure of our estimation. What this number really mean is "if we repeat many time such experiments: when we measure such outcomes for the y, then less than 1 in 1000 times it would have been with $x^* = -1$".

Let us move to the more difficult Gaussian example. In this case, x^* is chosen randomly from a Gaussian distribution, and we are given 10 measurements:

$$\begin{aligned} &0.04724576, 1.26855971, -0.19887457, 1.09534511, -1.46442807 \\ &0.44767123, 2.6244575, 1.94488421, 0.58953688, 0.572018\,. \end{aligned} \tag{7.8}$$

We compute explicitly the posterior and find that it is itself also a Gaussian, given by:

$$P^{\text{Gaussian}}_{X|Y}(x|\mathbf{y}) = \mathcal{N}\left(x; \frac{\sum\limits_{i=1}^{N} y_i}{N+\Delta}, \frac{\Delta}{N+\Delta}\right)\,. \tag{7.9}$$

The posterior distribution for this particular data set is shown in figure 7.2.1: it is Gaussian with mean $0.630.$ and variance $0.091.$. Actually, the true value of x^* is this case was 0.67.

Let us consider finally the third example. In this case the posterior is slightly more complicated and reads

$$P^{\text{Gauss}-\text{Bernoulli}}_{X|Y}(x|\mathbf{y}) = \frac{\delta(x)}{1+\sqrt{\frac{\Delta}{N+\Delta}}e^{\frac{\left(\sum\limits_{i=1}^{N} y_i\right)^2}{2\Delta(N+\Delta)}}} + \frac{\mathcal{N}\left(x; \frac{\sum\limits_{i=1}^{N} y_i}{N+\Delta}, \frac{\Delta}{N+\Delta}\right)}{1+\sqrt{\frac{N+\Delta}{\Delta}}e^{\frac{-\left(\sum\limits_{i=1}^{N} y_i\right)^2}{2\Delta(N+\Delta)}}} \tag{7.10}$$

We performed the experiment,where x^* is chosen randomly from a Gauss-Bernoulli, distribution, and we are given 10 measurements:

$$\begin{aligned} &0.99978688, 1.7956116, -0.43158072, 3.07234211, 1.11920946] \\ &0.53248943, 0.80011329, -0.52783428, 0.40378413, -0.00223177\,. \end{aligned} \tag{7.11}$$

The posterior is shown in Figure 7.2.2 (here $x^* = 0.8$, and p non zero is 0.8233787142909471).

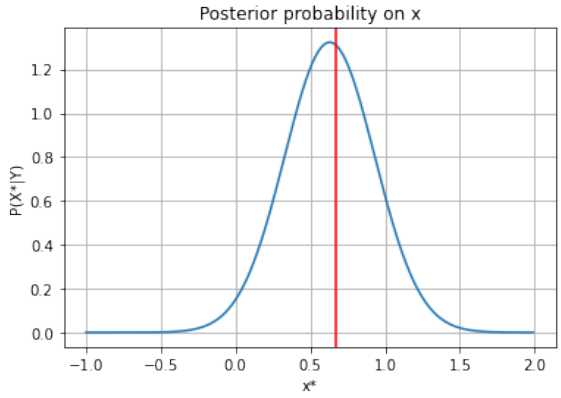


Figure 7.2.1: Posterior distribution for the Gaussian prior example and the data (7.8). The true value x^* is marked by a red vertical line.

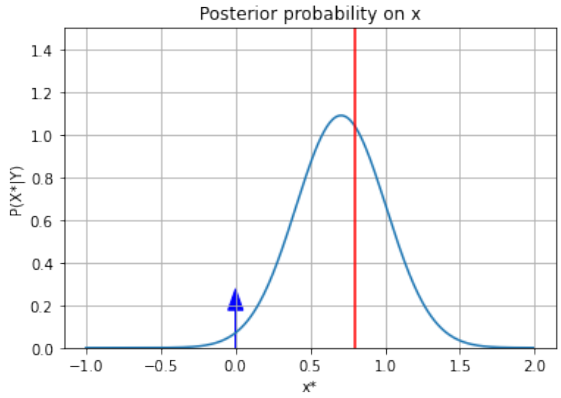


Figure 7.2.2: Posterior distribution for the Gaussian prior example and the data (7.11). Note the point mass component represented by the blue arrow and the Gaussian component. The true value x^* is marked by the red line.

7.2.2 Point-estimate inference: MAP, MMSE and all that.

Often, we are not really interested by the posterior distribution, but rather by a given estimate of the unknown x^*. We would really like to give a number and make our best guess! Such estimates are denoted as $\hat{x}(\mathbf{y})$. $\hat{x}(\mathbf{y})$ should be a function of the data $\mathbf{y}$ that gives us a number that is, ideally, as close as possible to the true value x^*. The first idea that comes to mind is to use the most probable value, that is the "mode" of the posterior distribution. This is called maximum a posteriori estimation:

$$\hat{x}_{\rm MAP}(\mathbf{y}) =: \underset{x}{\rm argmax}\, P_{X|Y}(x, \mathbf{y}) \tag{7.12}$$

The MAP estimate is the default-estimator. This is the one of choice in many situations, in particular because it is very often simple to compute.

However, it is not (at least for finite amount of data) always the best estimator. For instance what if the posterior has bad looking as below? Is $\hat{x}_{\rm MAP}$ still reasonable for this case? We need to think of a way to define a "best" estimator. The particular choice of the estimator depends

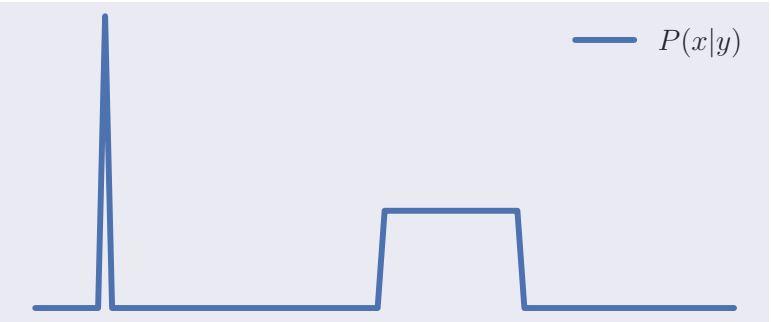


Figure 7.2.3: A nasty looking posterior

on our definition of "error". For instance one could decide to minimize the squared error $(\hat{x}(\mathbf{y}) - x^*)^2$, or the absolute error $|\hat{x}(\mathbf{y}) - x^*|$. If x^* is discrete, however, we might instead be interested by minimizing the probability of having a wrong value and use $1 - \delta_{\hat{x}(\mathbf{y}),x^*}$. Depending on our objective, we shall see that we should use a different estimator.

Let us consider the expected error one can get using a given estimator. Formally, we define a "risk" as the average of the loss function $\mathcal{L}(\hat{x}(\mathbf{y}), x^*)$ over the jointed distribution of signal and measurements. This is called the averaged posterior risk, as it is —indeed— the average of the posterior risk:

$$\mathcal{R}^{\text{averaged}}(\hat{x}) = \mathbb{E}_{x^*,\mathbf{y}}\left[\mathcal{L}(\hat{x}, x^*)\right] = \int P_{X,Y}(x^*, \mathbf{y}) dx^* d\mathbf{y} \mathcal{L}(\hat{x}(\mathbf{y}), x^*) \tag{7.13}$$

$$= \int d\mathbf{y} P_Y(\mathbf{y}) \int P_{X|Y}(x^*, \mathbf{y}) dx^* \mathcal{L}(\hat{x}(\mathbf{y}), x^*) \tag{7.14}$$

$$= \mathbb{E}_{\text{Y}}\left[\mathcal{R}^{\text{posterior}}(\hat{x}, \mathbf{y})\right] \tag{7.15}$$

Our goal, of course, is to find a way to minimize this risk. This minimal value, that is the "best" possible error one can possibly get (on average) is called the Bayes risk:

$$\mathcal{R}^{\text{Bayes}} = \min_{\hat{x}} \mathcal{R}^{\text{averaged}}(\hat{x}) \tag{7.16}$$

Our goal is two-fold: we want to know what is the best possible error, the Bayes error, as well as how to get it: we want to know what is the Bayes-optimal estimator that gives us the Bayes risk.

This line of reasoning leads, for the square loss, to the following theorem:

Theorem 10 (MMSE Estimator)**.** *The optimal estimator for the square error, called the Minimal Mean Square Error (MMSE) estimator, is given by the posterior mean:*

$$\hat{x}_{\text{MMSE}}(\mathbf{y}) = \mathbb{E}_{X|Y}(x, \mathbf{y}) \tag{7.17}$$

and the minimal mean square error is given by the variance of the estimator with respect to the posterior distribution:

$$\text{MMSE} =: \min_{\hat{x}} \mathcal{R}^{\text{Bayes}}(\mathbf{y}, \hat{x}(.)) = \int \mathrm{d}x^* \, \mathrm{d}\mathbf{y} P_{X|Y}(x^*, \mathbf{y})(\mathbb{E}_{X|Y}(x, \mathbf{y}) - x^*)^2 = \text{Var}_{P_{X|Y}}[X] \tag{7.18}$$

Proof. Consider the posterior risk for the square loss:

$$\mathcal{R}^{\text{Posterior}}(\mathbf{y}, \hat{x}(.)) = \int P_{X|Y}(x^*, \mathbf{y})(\hat{x}(\mathbf{y}) - x^*)^2 \, \mathrm{d}x^* \,. \tag{7.19}$$

Conditioned on $\mathbf{y}$ $\hat{x}$ is simple a scalar variable, it is a just number, so we can differentiate this expression with respect to $\hat{x}$ to obtain its minimum:

$$\frac{\partial}{\partial \hat{x}} \mathcal{R}^{\text{Posterior}}(\mathbf{y}, \hat{x}(.)) = 2 \int P_{X|Y}(x^*, \mathbf{y})(\hat{x}(\mathbf{y}) - x^*) \, \mathrm{d}x^\star = 0 \tag{7.20}$$

$$\hat{x}(\mathbf{y}) = \int P_{X|Y}(x^*, \mathbf{y}) x^* \, \mathrm{d}x^* = \int P_{X|Y}(x, \mathbf{y}) x \, \mathrm{d}x = \mathbb{E}_{X|Y}(x, \mathbf{y}) \tag{7.21}$$

Conditioned on $\mathbf{y}$, the minimum is thus obtained using the posterior mean. □

In what follows we shall study many such problems with Gaussian noise, so it is rather convenient to define the optimal denoising function as the MMSE estimator for a given problem:

$$\eta(R, \Sigma) =: \mathbb{E}_{P(X|X+R+\Sigma Z)} = \frac{\int \mathrm{d}x x e^{-\frac{(x-R)^2}{2\Sigma^2}} P(x)}{\int \mathrm{d}x e^{-\frac{(x-R)^2}{2\Sigma^2}} P(x)} \tag{7.22}$$

It is interesting to see, however, that for other errors, the optimal function can be different. If one choose the absolute value as the cost, then we find instead that one should use the Median:

Theorem 11 (MMAE Estimator)**.** *The optimal estimator for the absolute error, called the Minimal Mean absolute error (MMAE) estimator, is given by the posterior median:*

$$\hat{x}_{\text{MMAE}}(\mathbf{y}) = \text{Median}_{X|Y}(x|\mathbf{y}) \,. \tag{7.23}$$

Proof. Here we have

$$\mathcal{R}^{\text{Posterior}}(\mathbf{y}, \hat{x}(.)) = \int dx^* P_{X|Y}(x^*, \mathbf{y}) |\hat{x}(\mathbf{y}) - x^*| \tag{7.24}$$

$$= \int_{-\infty}^{\hat{x}} dx^* P_{X|Y}(x^*, \mathbf{y})(-(\hat{x}(\mathbf{y}) - x^*)) + \int_{\hat{x}}^{\infty} dx^* P_{X|Y}(x^*, \mathbf{y})(\hat{x}(\mathbf{y}) - x^*) \tag{7.25}$$

Performing the derivative with Leibniz integral rule, we find that we require:

$$\frac{\partial}{\partial \hat{x}} \mathcal{R}^{\text{Posterior}}(\mathbf{y}, \hat{x}(.)) = -\int_{-\infty}^{\hat{x}} dx^* P_{X|Y}(x^*, \mathbf{y}) + \int_{\hat{x}}^{\infty} dx^* P_{X|Y}(x^*, \mathbf{y}) = 0 \tag{7.26}$$

and this is achieved for

$$\hat{x}_{\text{MMAE}}(\mathbf{y}) = \text{Median}_{X|Y}(x|\mathbf{y}) \tag{7.27}$$

□

Finally, if we are interested to choose between a finite number of hypothesis, like in the case ± 1, or if we want to know if the number was exactly zero in the Gauss-Bernoulli case, a good measure of error is to look to the optimal decision version and to minimize the number of mistakes:

Theorem 12 (Optimal Decision)**.** *The Optimal Bayesian decision estimator is the one that maximizes the* (*marginal*) *probability for each class:*

$$\hat{x}_{\mathrm{OBD}}(\mathbf{y}) = \underset{x}{\operatorname{argmax}}\, P_{X|Y}(x, \mathbf{y}) \tag{7.28}$$

7.2.3 Back to free entropies

Let us now discuss how to think about these problems with a statistical physics formalism. We can write down the posterior distribution as

$$P\left(x \,|\, y\right) = \frac{\exp\left(\log\left(P\left(\mathbf{y} \,|\, x\right) P(x)\right)\right)}{P(\mathbf{y})} \equiv P_{\mathrm{Gibbs},\mathbf{y}}(x) = \frac{\mathbb{e}^{-\mathcal{H}(x;\mathbf{y})}}{Z(\mathbf{y})} \tag{7.29}$$

A way to define our Boltzmann measure would be to use $\beta = 1$, $\mathcal{H}(x;\mathbf{y}) = -\log\left(P\left(\mathbf{y} \,|\, x\right)\right) - \log\left(P(x)\right)$, and $Z(\mathbf{y}) = P(\mathbf{y})$. In practice, for such problems with a Gaussian noise, we shall employ a slightly different convention that is more practical, and use instead

$$P\left(x \,|\, \mathbf{y}\right) = \frac{\exp\left(\log\left(P\left(\mathbf{y} \,|\, x\right) P(x)\right)\right)}{P(\mathbf{y})} = \frac{1}{(2\pi\Delta)^{N/2}} \frac{e^{-\sum_i \frac{(y_i - x)^2}{2\Delta}} P(x)}{P(\mathbf{y})} \tag{7.30}$$

$$= \frac{e^{-\sum_i \frac{y_i^2}{2\Delta}}}{(2\pi\Delta)^{N/2} P(\mathbf{y})} e^{\sum_i \left(\frac{y_i x}{\Delta} - \frac{x^2}{2\Delta}\right)} P(x) \tag{7.31}$$

$$=: \frac{e^{\sum_i \left(-\frac{x^2}{2\Delta} + \frac{y_i x}{\Delta}\right)} P(x)}{Z(\mathbf{y})} \tag{7.32}$$

with

$$Z(\mathbf{y}) = \int \mathrm{d}x\, e^{\sum_i \left(-\frac{x^2}{2\Delta} + \frac{y_i x}{\Delta}\right)} P(x) = \left[\frac{e^{-\sum_i \frac{y_i^2}{2\Delta}}}{(2\pi\Delta)^{N/2} P(\mathbf{y})}\right]^{-1}$$

Interestingly, with this definition, the partition sum is also equal to the ratio between the probability that y is a pure random noise (a Gaussian with variance Δ), and that y has been actually generated by a noisy process from x:

$$Z(\mathbf{y}) = \frac{P_Y^{\mathrm{model}}(\mathbf{y})}{P_Y^{\mathrm{random}}(\mathbf{y})}\,.$$

This is called the likelihood ratio in hypothesis testing. Obviously, if the two distributions are the same, then $Z = 1$ for all values of y. With this definition, we define the expected free entropy, as before, as

$$F_N = \mathbb{E}_Y \log Z(\mathbf{y})\,. \tag{7.33}$$

In fact, the free entropy turns out to be nothing more than the Kullback-Liebler divergence between $P_Y^{\mathrm{model}}(\mathbf{y})$ and $P_Y^{\mathrm{random}}(\mathbf{y})$:

$$F_N = \mathbb{E}_Y^{\mathrm{model}} \log \frac{P_Y^{\mathrm{model}}(\mathbf{y})}{P_Y^{\mathrm{random}}(\mathbf{y})} = D_{KL}(P_Y^{\mathrm{model}}(\mathbf{y}) | P_Y^{\mathrm{random}}(\mathbf{y})) \tag{7.34}$$

Many other information quantities would have been equally interesting, but they are all equivalent. We could have used for instance the entropy of the variable y, which is related trivially to our free entropy.

$$H(Y) = -\mathbb{E}_Y \log P(\mathbf{y}) = N\frac{\mathbb{E}_y[y^2]}{2\Delta} + \frac{N}{2}\log 2\pi\Delta - F_N \tag{7.35}$$

Information theory practitioners would, typically, use the mutual information between X and Y, that is the Kullback-Leibler distance between the jointed and factorized distribution of Y and X.

$$I(X,Y) = D_{\rm KL}(P_{X,Y}||P_X P_Y)\,. \tag{7.36}$$

Again, this can be expressed directly as a function of the free entropy, using (see exercise section for basic properties of the mutual information and conditional entropies):

$$I(X,Y) = H(Y) - H(Y|X) = H(Y) - \frac{N}{2}\log(2\pi e\Delta) \tag{7.37}$$

$$= -F_N - \frac{N}{2} + N\frac{\mathbb{E}_y[y^2]}{2\Delta} = F - \frac{N}{2} + N\frac{\mathbb{E}_x[x^2]+\Delta}{2\Delta} \tag{7.38}$$

$$= -F_N + N\frac{\mathbb{E}_x[x^2]}{2\Delta} \tag{7.39}$$

Given these equivalences, we shall thus focus on the free entropy.

7.2.4 Some useful identities: Nishimori and Stein

Before going further, we need to note some important mathematical identities that we shall use all the time, especially in the context of Bayesian inference.

The first one is a generic property of the Gaussian integrals, a simple consequence of integration by part, called Stein's lemma:

Lemma 7 (Stein's Lemma)**.** *Let $X \sim \mathcal{N}(\mu,\sigma^2)$. Let g be a differentiable function such that the expectation $\mathbb{E}\left[(X-\mu)g(X)\right]$ and $\mathbb{E}\left[g'(X)\right]$ exists, then we have*

$$\mathbb{E}\left[g(X)(X-\mu)\right] = \sigma^2\mathbb{E}\left[g'(X)\right]$$

Particularly, when $X \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}\left[Xg(X)\right] = \mathbb{E}\left[g'(X)\right]$$

Proof. The proof is a trivial application of the integration by part formula $\int f'g = [fg] - \int fg'$ applied on $g\ f = -e^{-x^2/2}/\sqrt{2\pi}$. □

Additionally, there is a set of identities that are extremly useful, that are usually called " Nishimori symmetry" in the context of physics and error correcting codes. In its more general form, it reads

Theorem 13 (Nishimori Identity)**.** *Let $X^{(1)}, \dots, X^{(k)}$ be k i.i.d. samples (given Y) from the distribution $P(X = \cdot \,|\, Y)$. Denoting $\langle \cdot \rangle$ the "Boltzmann" expectation, that is the average with respect to the $P(X = \cdot \,|\, Y)$, and $\mathbb{E}[\cdot]$ the "Disorder" expectation, that is with respect to (X^*, Y). Then for all continuous bounded function f we can switch one of the copies for X^*:*

$$\mathbb{E}\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{(k)}\right)\right\rangle_k\right] = \mathbb{E}\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{*}\right)\right\rangle_{k-1}\right] \tag{7.40}$$

Proof. The proof is a consequence of Bayes theorem and of the fact that both x^* and any of the copy $X^{(k)}$ are distributed from the posterior distribution. Denoting more explicitly the Boltzmann average over k copies for any function g as

$$\left\langle g(X^{(1)}, \dots, X^{(k)})\right\rangle_k =: \int \prod_{i=1}^{k} \mathrm{d}x_i P(x_i|Y) g(X^{(1)}, \dots, X^{(k)}) \tag{7.41}$$

we have, starting from the right hand side

$$\begin{aligned}
&\mathbb{E}_{Y,X^*}\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{*}\right)\right\rangle_{k-1}\right] \\
&= \int dx^* dy P(x^*|Y)P(Y)\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{*}\right)\right\rangle_{k-1} \\
&= \mathbb{E}_Y \int \mathrm{d}x^k P(x^k|y)\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{k}\right)\right\rangle_{k-1} \\
&= \mathbb{E}_Y\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{(k)}\right)\right\rangle_k\right]
\end{aligned}$$

□

We shall drop the subset "k" from Boltzmann averages from now on. The Nishimori property has many useful consequences that we can now discuss. First let us look at the expression of the MMSE. It has a nice expression in terms of overlaps:

$$\mathsf{MMSE}(\lambda) = \mathbb{E}_{y,x^*}\left[\left(\langle x\rangle_y - x^*\right)^2\right] = \mathbb{E}_{y,x^*}\left[\langle x\rangle_y^2 + (x^*)^2 - 2x^*\langle x\rangle_y\right] = q + q_0 - 2m$$

where

- $q \triangleq \mathbb{E}_y\left[\langle x\rangle_y^2\right] = \mathbb{E}_y\left[\left\langle x^{(1)}x^{(2)}\right\rangle_y\right]$ is overlap between two copies
- $q_0 \triangleq \mathbb{E}_{x^*}\left[(x^*)^2\right]$ is the self overlap
- $m \triangleq \mathbb{E}_{y,x^*}\left[x^*\langle x\rangle_y\right]$ is the overlap with the truth.

Using now the Nishimori Identity this can be simplified as

$$\mathbb{E}_{y,x^*}\left[\langle f(x, x^*)\rangle_y\right] = \mathbb{E}_y\left[\left\langle f(x^{(1)}, x^{(2)})\right\rangle_y\right]$$

Using this result we have $q \equiv m$ and thus $\mathsf{MMSE}(\lambda) = q_0 - m$

7.2.5 I-MMSE theorem

Theorem 14 (I-MMSE Theorem). *For a single measurement and Gaussian noise, if* $Y = \sqrt{\Delta}Z + X^*$, *The derivative of the mutual information, or the free entropy, with respect to the inverse noise gives the MMSE*

$$\frac{\partial}{\partial \Delta^{-1}} \mathcal{I}(\Delta) = \frac{1}{2}\mathsf{MMSE}(\lambda) = \frac{1}{2}(q_0 - m) \tag{7.42}$$

$$\frac{\partial}{\partial \Delta^{-1}} F(\Delta) = \frac{1}{2} m \tag{7.43}$$

Proof. Writing F explicitly as a function of the Gaussian noise, we have

$$F = \mathbb{E}_{x^*,z} \log \int \mathrm{d}x \; P(x) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{zx}{\sqrt{\Delta}}} \tag{7.44}$$

Performing the derivative, we find

$$\partial_{\Delta^{-1}} F = \mathbb{E}_{x^*,z} \frac{\int \mathrm{d}x \; P(x) \left(-\frac{x^2}{2} + xx^* + \frac{zx}{2}\sqrt{\Delta}\right) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{zx}{\sqrt{\Delta}}}}{\int \mathrm{d}x \; P(x) e^{-\frac{x^2}{2\Delta} + \frac{xx^*}{\Delta} + \frac{zx}{\sqrt{\Delta}}}} \tag{7.45}$$

$$= -\frac{1}{2}\mathbb{E}_{z,x_*}\left[\langle x^2 \rangle\right] + \mathbb{E}_{z,x_*}\left[\langle x \rangle x^*\right] + \mathbb{E}_{z,x_*}\left[\frac{\sqrt{\Delta}}{2}\langle x \rangle z\right] \tag{7.46}$$

Using Stein's lemma on the variable z, the third term can be written as

$$\mathbb{E}_{z,x_*}\left[\frac{\sqrt{\Delta}}{2}\langle x \rangle z\right] = \mathbb{E}_{z,x_*}\left[\frac{\sqrt{\Delta}}{2}\partial_z \langle x \rangle\right] = \frac{1}{2}\mathbb{E}_{z,x_*}\left[\langle x^2 \rangle - \langle x \rangle^2\right] \tag{7.47}$$

so that

$$\partial_{\Delta^{-1}} F = -\frac{1}{2}\mathbb{E}_{z,x_*}\left[\langle x^2 \rangle\right] + \mathbb{E}_{z,x_*}\left[\langle x \rangle x^*\right] + \frac{1}{2}\mathbb{E}_{z,x_*}\left[\langle x^2 \rangle - \langle x \rangle^2\right] \tag{7.48}$$

$$= \mathbb{E}_{z,x_*}\left[\langle x \rangle x^*\right] - \frac{1}{2}\mathbb{E}_{z,x_*}\left[\langle x \rangle^2\right] \tag{7.49}$$

$$= m - \frac{q}{2} = \frac{m}{2} \tag{7.50}$$

Where the last step follows from Nishimori. □

7.3 Application: Denoising a sparse vector

It is obvious to check that all the theorems that we have discussed applied equally to d-dimensional vectors. We can thus apply our newfound knowledge to a more interesting problem: denoising a sparse vector.

Consider a vector $\mathbf{x}^*$ of dimension d. Often, in computers, d is a power of 2, so we take $d = 2^N$, with only ONE single non zero component, exactly equal to 1. In other words, x is one of

the d vectors $x = [10000\ldots]$, $x = [01000\ldots]$, etc. Instead of this vector, you are given a noisy d-dimensional vector $\mathbf{y}$ which has been polluted by a *very* small Gaussian noise

$$y = x^* + \sqrt{\frac{\Delta}{N}} z \tag{7.51}$$

Can we recover x^*? We proceed in the Bayesian way, and write that

$$P(\mathbf{x}|\mathbf{y}) = \frac{1}{P(\mathbf{y})} P(\mathbf{x}) \prod_i \frac{e^{-\frac{(y_i - x_i)^2}{2\Delta/N}}}{\sqrt{2\pi\Delta/N}} = \frac{\prod_i \mathcal{N}(y_i, 0, \Delta/N)}{P(\mathbf{y})} \prod_i e^{-\frac{x_i^2 - 2x_i y_i}{2\Delta/N}} P(\mathbf{x}) \tag{7.52}$$

We recognize that $\prod_i \mathcal{N}(y_i, 0, \Delta/N)$ is just the probability of the null model (if the vector $\mathbf{y}$ was simply a random one). Additionally, using the prior on $\mathbf{x}$ tell us that it has to be on the corner of the hypercube (one of the vector $\mathbf{x}_i$ which are zero everywhere but $x_i = 1$), we write

$$P(\mathbf{x}_i|\mathbf{y}) = \frac{P^{\text{random}}(y)}{P^{\text{model}}(y)} \frac{1}{2^N} e^{\frac{N}{\Delta} y_i - \frac{N}{2\Delta}} =: \frac{1}{Z} \frac{1}{2^N} e^{\frac{N}{\Delta} y_i - \frac{N}{2\Delta}} . \tag{7.53}$$

As before, the partition sum stands for the ratio of probability between the model and the null, and reads

$$Z = \frac{1}{2^N} \sum_{i=1}^{d} e^{-\frac{N}{2\Delta} + \frac{N y_i}{\Delta}} = \frac{1}{2^N} \sum_{i=1}^{d} e^{-\frac{N}{2\Delta} + \frac{N\delta_{i,i^*}}{\Delta} + \sqrt{\frac{N}{\Delta}} z_i} \tag{7.54}$$

Our goal is to compute the free entropy as a function of Δ

$$\Phi(\Delta) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E} \log Z \tag{7.55}$$

The application of the I-MMSE theorem tells us in particular that

$$q =: \lim_{N\to\infty} \mathbb{E}\left[\sum_i \frac{\langle \mathbf{x} \rangle \mathbf{x}^*}{N}\right] = 2\partial_{\Delta^{-1}} \Phi(\Delta) \tag{7.56}$$

This can be done rigorously, as we shall now see. In fact we can prove the following expression for the free energy:

Theorem 15. *Let $f(\Delta) : \mathbb{R} \to \mathbb{R}$ be*

$$f(\Delta) = \frac{1}{2\Delta} - \log 2 \,, \text{if} \,, \Delta \leq 1/2 \log 2 \tag{7.57}$$

$$f(\Delta) = 0 \,, \text{if} \, \Delta \geq 1/2 \log 2 \tag{7.58}$$

Then the limit of the free entropy is $f(\Delta)$

We shall prove this theorem by proving an upper and lower bound. Let us start by

Lemma 8 (Upper bound)**.**

$$\Phi_N(\Delta) \geq f(\Delta) \tag{7.59}$$

Proof. The bound comes from using only one term in the sum, the one corresponding to the correct position i^*:

$$\Phi_N(\Delta) = \frac{1}{N}\mathbb{E}\left[\log Z\right] \geq \frac{1}{N}\mathbb{E}\left[\log\left(\frac{1}{2^N}e^{-\frac{N}{2\Delta}+\frac{N}{\Delta}+z\sqrt{\frac{N}{\Delta}}}\right)\right] \tag{7.60}$$

$$\geq \frac{1}{2\Delta} - \log 2 \tag{7.61}$$

Additionally, since the two distributions become indistinguishable for infinite noise and that Z is just the likelihood ratio, we have $\Phi_N(\Delta=\infty)=0$. Since $\partial_\Delta \Phi_N(\Delta) = (\partial_{\Delta^{-1}}\Phi_N(\Delta))(\partial_\Delta(1\Delta)) = -\frac{q}{2\Delta^2} \leq 0$, we have $\Phi_N(\Delta)\geq 0$. □

Lemma 9 (Lower bound)**.**

$$\Phi_N(\Delta) \leq f(\Delta) + o(1) \tag{7.62}$$

Proof. The bound comes from the Jensen inequality (the annealed bound):

$$\begin{aligned}\Phi_N(\Delta) = \frac{1}{N}\mathbb{E}\left[\log Z\right] &\leq \frac{1}{N}\mathbb{E}_{z_i^*}\left[\log\left(e^{\frac{N}{2\Delta}+z_i^*\sqrt{\frac{N}{\Delta}}-N\log 2} + \sum_{i\neq i^*}\mathbb{E}_{z_i}\left[\frac{1}{2^N}e^{-\frac{N}{2\Delta}+z_i\sqrt{\frac{N}{\Delta}}}\right]\right)\right] \\ &\leq \frac{1}{N}\mathbb{E}_{z_i^*}\left[\log\left(e^{\frac{N}{2\Delta}+z_i^*\sqrt{\frac{N}{\Delta}}-N\log 2} + \left(1-\frac{1}{2^N}\right)e^{\frac{N}{2\Delta}-\frac{N}{2\Delta}}\right)\right] \\ &\leq \frac{1}{N}\mathbb{E}_{z_i^*}\left[\log\left(e^{N\left(\frac{1}{2\Delta}-\log 2\right)+z_i^*\sqrt{\frac{N}{\Delta}}}+1\right)\right]\end{aligned} \tag{7.63}$$

It is intuitively clear that, depending on where or not the term $1/2\Delta - \log 2$ in the exponential is positive or negative, then we should expect to either completely dominate the expression, or to disappear exponentially.

We can show this with rigor, for instance by defining the monotonic growing function

$$g(z_{i^*}) =: e^{N\left(\frac{1}{2\Delta}-\log 2\right)+z_i^*\sqrt{\frac{N}{\Delta}}}+1\,. \tag{7.64}$$

We have $g(z_{i^*}) \leq g(|z_{i^*}|)$, and

$$\log g(|z_{i^*}|) \leq \log g(0) + |z_{i^*}|\max\frac{g'}{g} = \log g(0) + |z_{i^*}|\sqrt{\frac{N}{\Delta}} \tag{7.65}$$

so that

$$\frac{1}{N}\mathbb{E}\log g(z_{i^*}) \leq \frac{1}{N}\mathbb{E}\log g(|z_{i^*}|) \leq \frac{1}{N}\log\left(1+e^{N\left(\frac{1}{2\Delta}-\log 2\right)}\right) + \mathbb{E}|z|\frac{1}{\sqrt{N\Delta}} \tag{7.66}$$

$$\leq \frac{1}{N}\log\left(1+e^{N\left(\frac{1}{2\Delta}-\log 2\right)}\right) + o(1) \tag{7.67}$$

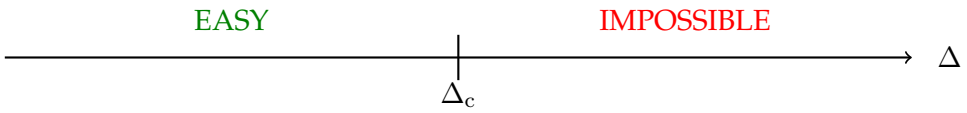
We conclude by noting that the bounds tends to $f(\Delta)$ as $N\to\infty$ (taking the exponential of $1/2\Delta-\log 2$ out of the log) and using $\log(1+x)\leq x$. □

Now that we know the free entropy, we can apply the I-MMSE theorem. A phase transition occurs depending on Δ being lower or larger than Δ_c

$$\Delta_c = \frac{1}{2\log 2}$$

If $\Delta > \Delta_c$, the MMSE is 1, and we cannot find the signal. Even the best guess is not better than a random one. If, on the other end, $\Delta < \Delta_C$, then we should be able to solve the problem, and finds a perfect MMSE (that is, a zero error).

Can we do it in practice? What should be our algorithm? It is a well-known result that the maximum of d i.i.d. Gaussian random variables is asymptotically $\sqrt{2\log d}$ with fluctuations of order $1/\sqrt{\log d}$. In our case, this means that in absence of a signal and with a variance $\sigma^2 = \Delta/N = \Delta \log 2/\log d$ the largest number will be $\sqrt{\Delta 2 \log 2}$, which is indeed smaller than one when $\Delta < \Delta_c$. This means that the simplest algorithm (keep only the largest component), will be optimal with high probab Therefore, we have the interesting following phase diagram:



For signal of dimension d, and a noise σ^2 this yields $\sigma_c^2 = 1/(2\log(d))$.

Actually, one can prove an even stronger result in the regime $\Delta \geq \Delta_c$. As we show in appendix 7.A, not only the free entropy divided by N goes to zero, but the total free entropy as well. We recall that it is nothing but the KL divergence between the distribution of the model and the random one:

$$F_N = \mathbb{E}_Y^{\text{model}} \log \frac{P_Y^{\text{model}}(\mathbf{y})}{P_Y^{\text{random}}(\mathbf{y})} = D_{KL}(P_Y^{\text{model}}(\mathbf{y}) | P_Y^{\text{random}}(\mathbf{y})) \to_{N\to\infty} 0 \tag{7.68}$$

This actually means that the two distributions are eventually just becoming just the same one, and are thus indistinguishable: not only we cannot find the signal but there is just no way to know that a signal has been hidden for $\Delta > \Delta_c$, as the data looks perfectly like Gaussian noise.

Bibliography

The legacy of the Bayes theorem, and the fundamental role of Laplace in the invention of "inverse probabilities" is well discussed in McGrayne (2011). Bayesian estimation is a fundamental field at the frontier between information theory and statistics, and is discussed in many references such as Cover and Thomas (1991). The I-MMSE theorem was introduced by Guo et al. (2005). Nishimori symmetries were introduced in physics is Nishimori (1980) and soon realized to have deep connection to information theory Nishimori (1993) and Bayesian inference Iba (1999). The model of denoising a sparse vector was discussed in Donoho et al. (1998). This problem has deep relation to Shannon's Random codes Shannon (1948) and the Random energy model in statistical physics Derrida (1981).

7.4 Exercises

Exercise 7.1: Few useful equalities on Entropies

In what follow, we shall denote the entropy of a random variable X with a distribution $p_X(x)$ as

$$\mathrm{H}(X) = -\int dx\; p(x) \log p(x)$$

- **Entropy of a Gaussian variable:**

  Show that the entropy of a Gaussian variable sampled from $\mathcal{N}(m, \Delta)$ is given by

  $$\mathrm{H}(X) = \frac{1}{2} \log 2\pi e \Delta$$

- **Mutual information:**

  The mutual information between two (potentially) correlated variable X and Y is defined as the Kullback-Leibler divergence between their joint distribution and the factorized one. In other words, it reads

  $$\mathrm{I}(X;Y) = D_{\mathrm{KL}}\left(P_{X,Y} || P_X P_Y\right) = \int dxdy\; p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)\; p_Y(y)}$$

  Show that the mutual information satisfies the following chain rules:

  $$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$$

  where the so-called conditional entropy $\mathrm{H}(X|Y)$ is defined as

  $$\mathrm{H}(X|Y) = -\int dy\; p_Y(y) \int dx\; p_{X|Y}(x|y) \log p_{X|Y}(x|y)$$

- **Conditional entropy of a Gaussian**:

  Given a random variable X, whose distribution is $p_X(x)$ and a random variable Z whose distribution is Gaussian with variance Δ, we define a new random variable Y given by

  $$Y = X + Z$$

  Shows that the conditional entropy $\mathrm{H}(X|Y)$ is then given by

  $$\mathrm{H}(X|Y) = \frac{1}{2} \log 2\pi e \Delta$$

Exercise 7.2: Numerical tests

Perform simulation of the 3 models discussed in section 6.2.1 with the MMSE, MAP, and MMAE estimators discussed in section 6.2.2
Using different values for the number of observation n (from 10 to 1000, or even more) and averaging your finding on many instances, plots how, for each problems the error

evolves with N for different Risk and different estimators.

Exercise 7.3: Second derivative

We saw in section 6.2.5 that the first derivative of the free entropy (with Gaussian noise) with respect to Δ^{-1} is (one-half) the overlap m.
Compute the second derivative (use again Stein and Nishimori) and relate it to a variance of a quantity. Show that it implies the convexity of the free entropy with respect to Δ^{-1}.

Appendix

7.A A tighter computation of the likelihood ratio

We shall here prove that the average partition sum is not only going to zero when $\Delta > \Delta_c$, but that the corrections to the free entropy are actually exponentially small. This require a tighter analysis of the likelihood ratio and of the upper bound in 9.

Our starting point is eq.7.63 that states

$$\Phi_N(\Delta) \leq \frac{1}{N}\mathbb{E}_{z_i^*}\left[\log\left(e^{N(\frac{1}{2\Delta}-\log 2)+z_i^*\sqrt{\frac{N}{\Delta}}}+1\right)\right]$$

Our goal is to shat that for $\Delta > \Delta_c$, this is exponentially small. Let us simplify notation a bit and denote $f = -\frac{1}{2\Delta} + \log 2 > 0$, and write, spitting the integral in two:

$$\Phi_N(\Delta) \leq \frac{1}{N}\mathbb{E}_z\left[\log\left(e^{-fN+z\sqrt{\frac{N}{\Delta}}}+1\right)\right] = \int_{-\infty}^{+\infty} dz \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \log\left(e^{-fN+z\sqrt{\frac{N}{\Delta}}}+1\right) \tag{7.69}$$

$$\leq \int_{-\infty}^{\sqrt{N\Delta}f} dz \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \log\left(e^{-fN+z\sqrt{\frac{N}{\Delta}}}+1\right) + \int_{\sqrt{N\Delta}f}^{\infty} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \log\left(e^{-fN+z\sqrt{\frac{N}{\Delta}}}+1\right) \tag{7.70}$$

$$\leq I_1 + I_2 \tag{7.71}$$

We first deal with I_1. Since the exponential term is positive, we can write, using the "worst" possible value of z:

$$\begin{aligned} I_1 = \int_{-\infty}^{\sqrt{N\Delta}f(1-\epsilon)} dz \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \log\left(e^{-fN+z\sqrt{\frac{N}{\Delta}}}+1\right) &\leq \int_{-\infty}^{\infty} dz \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \log\left(e^{-fN+Nf(1-\epsilon)}+1\right) \\ &\leq \int_{-\infty}^{\infty} dz \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \log\left(e^{-\epsilon fN}+1\right) \leq e^{-\epsilon fN} \end{aligned}$$

where we have used $\log(1+x) \leq x$. So indeed I_1 is exponentially small.

What about I_2? Again uwe use $\log(1+x) \leq x$:

$$I_2 = \int_{\sqrt{N\Delta}f(1-\epsilon)}^{\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \log\left(e^{-fN+z\sqrt{\frac{N}{\Delta}}}+1\right) \leq e^{-fN} \int_{\sqrt{N\Delta}f(1-\epsilon)}^{\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} e^{z\sqrt{\frac{N}{\Delta}}} \tag{7.72}$$

With a bit or rewriting we can write

$$I_2 \leq e^{-fN+\frac{N}{2\Delta}} \int_{\sqrt{N\Delta}f(1-\epsilon)}^{\infty} dz \frac{e^{-\frac{(z-\sqrt{\frac{N}{\Delta}})^2}{2}}}{\sqrt{2\pi}} = e^{-fN+\frac{N}{2\Delta}} \int_{\sqrt{N\Delta}f(1-\epsilon)-\sqrt{\frac{N}{\Delta}}}^{\infty} dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \quad (7.73)$$

$$\leq e^{-fN+\frac{N}{2\Delta}} \mathbb{P}\left(Z > \sqrt{N\Delta}f(1-\epsilon) - \sqrt{\frac{N}{\Delta}}\right) \quad (7.74)$$

For a Gaussian random variable, we have that $\mathbb{P}\left(Z > b\right) \leq e^{-b^2/2}$ thus

$$I_2 \leq e^{-fN+\frac{N}{2\Delta}} e^{-\frac{1}{2}\left(\sqrt{N\Delta}f(1-\epsilon)-\sqrt{\frac{N}{\Delta}}\right)^2} \quad (7.75)$$

$$= e^{-fN+\frac{N}{2\Delta}} e^{-\frac{1}{2}N\Delta f(1-\epsilon)^2} e^{-\frac{1}{2}\frac{N}{\Delta}} e^{Nf(1-\epsilon)} \quad (7.76)$$

$$= e^{-\epsilon N} e^{-\frac{1}{2}N\Delta f(1-\epsilon)^2} \quad (7.77)$$

This is again decaying exponentially fast. We thus obtain the following result:

Lemma 10 (Exponential decay of the Kullback-Leibler divergence)**.**

$$NF_n(\Delta) = D_{KL}(P_Y^{\text{model}}(\mathbf{y})|P_Y^{\text{random}}(\mathbf{y})) =\leq Ne^{-KN} \text{ for } \Delta > \Delta_{\text{c}} \quad (7.78)$$

Note that for the divergence to go to zero, one needs the *total* free entropy to go to zero, not just the one divided by N (that is, the density).

7.B A replica computation for vector denoising

It is instructive, and a good exercise, to redo the computation of the free entropy in theorem:15 using the replica method. The computation is very reminiscent of the one for the random energy model in chap 15, which we encourage the reader to consult. In this, the present is nothing but a "Bayesian" version of the random energy model.

Let us see how the replica computation goes. We first remind that the partition sum reads

$$Z = \frac{1}{2^N} \sum_{i=1}^{d} e^{-\frac{N}{2\Delta}+\frac{N\delta_{i,i^*}}{\Delta}+\sqrt{\frac{N}{\Delta}}z_i} = e^{-N\left(\log 2+\frac{1}{2\Delta}\right)} \sum_{i=1}^{d} e^{\frac{N\delta_{i,i^*}}{\Delta}+\sqrt{\frac{N}{\Delta}}z_i} \quad (7.79)$$

We now move to the computation of the averaged free entropy by the replica method, starting zith the replicated partition sum:

$$Z^n = e^{-nN\left(\log 2+\frac{1}{2\Delta}\right)} \prod_{a=1}^{n} \left(\sum_{i=1}^{d} e^{\frac{N}{\Delta}\delta_{i,i^*}+\sqrt{\frac{N}{\Delta}}z_i} \right) . \quad (7.80)$$

First, let us start using seemingly trivial rewriting, using $i^* = 1$ without loss of generality:

$$Z^n e^{nN\left(\log 2 + \frac{1}{2\Delta}\right)} = \sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\left(\sqrt{\frac{N}{\Delta}} z_{i_a} + \frac{N}{\Delta}\delta_{i_a,1}\right)} \tag{7.81}$$

$$= \sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\frac{N}{\Delta}\delta_{i_a,1}} e^{\sum_{a=1}^{n}\sqrt{\frac{N}{\Delta}} z_{i_a}} \tag{7.82}$$

$$= \sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\frac{N}{\Delta}\delta_{i_a,1}} e^{\sum_{a=1}^{n}\sum_{j=1}^{d}\sqrt{\frac{N}{\Delta}} z_j \delta_{j,i_a}} \tag{7.83}$$

$$= \sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\frac{N}{\Delta}\delta_{i_a,1}} \prod_{j=1}^{d} e^{z_j \sum_{a=1}^{n}\sqrt{\frac{N}{\Delta}}\delta_{j,i_a}} \tag{7.84}$$

Now we perform the expectation over disorder, using the fact that we have now a product of independent Gaussians:

$$\mathbb{E}\left[Z^n\right] = e^{-nN\left(\log 2 + \frac{1}{2\Delta}\right)} \mathbb{E}\left[\sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\frac{N}{\Delta}\delta_{i_a,1}} \prod_{j=1}^{d} e^{z_j \sum_{a=1}^{n}\sqrt{\frac{N}{\Delta}}\delta_{j,i_a}}\right] \tag{7.85}$$

$$= e^{-nN\left(\log 2 + \frac{1}{2\Delta}\right)} \sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\frac{N}{\Delta}\delta_{i_a,1}} \prod_{j=1}^{d} \mathbb{E}\left[e^{z_j \sum_{a=1}^{n}\sqrt{\frac{N}{\Delta}}\delta_{j,i_a}}\right] \tag{7.86}$$

Using $\mathbb{E}\left[e^{bz}\right] = e^{b^2/2}$ for Gaussian variables, we thus find

$$\mathbb{E}\left[Z^n\right] = e^{-nN\left(\log 2 + \frac{1}{2\Delta}\right)} \sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\frac{N}{\Delta}\delta_{i_a,1}} \prod_{j=1}^{d} e^{\frac{N}{2\Delta}\left(\sum_{a=1}^{n}\delta_{j,i_a}\right)^2} \tag{7.87}$$

$$= e^{-nN\left(\log 2 + \frac{1}{2\Delta}\right)} \sum_{i_1,\dots,i_n=1}^{d} e^{\sum_{a=1}^{n}\frac{N}{\Delta}\delta_{i_a,1}} e^{\frac{N}{2\Delta}\sum_{j=1}^{d}\sum_{a,b=1}^{n}\delta_{j,i_a}\delta_{j,i_b}} \tag{7.88}$$

$$= e^{-nN\left(\log 2 + \frac{1}{2\Delta}\right)} \sum_{i_1,\dots,i_n=1}^{d} e^{\frac{N}{\Delta}\left(\sum_{a=1}^{n}\delta_{i_a,1} + \frac{1}{2}\sum_{a,b=1}^{n}\delta_{i_a,i_b}\right)} \tag{7.89}$$

Given the replicas configurations $(i_1 ... i_n)$, that can take values in $1, \dots, d$, we now denote the so-called $n \times n$ overlap matrix $Q_{ab} = \delta_{i_a,i_b}$, that takes elements in $0, 1$, respectively if the two replicas (row and column) have different or equal configuration. We also write the $n \times n$ magnetization vector $M_a = \delta_{i_a,1}$. With this notation, we can write the replicated sum as

$$\mathbb{E}[Z^n] = e^{-nN\left(\log 2 + \frac{1}{2\Delta}\right)} \sum_{i_1,\dots,i_n=1}^{d} e^{\frac{N}{\Delta}\left(\sum_{a=1}^{n} M_a + \frac{1}{2}\sum_{a,b=1}^{n} Q_{a,b}\right)} \tag{7.90}$$

$$= e^{-nN\left(\log 2 + \frac{1}{2\Delta}\right)} \sum_{\{Q\},\{M\}} \#(Q, M) e^{\frac{N}{\Delta}\left(\sum_{a=1}^{n} M_a + \frac{1}{2}\sum_{a,b=1}^{n} Q_{a,b}\right)} \tag{7.91}$$

where $\sum_{\{Q\},\{M\}}$ is the sum over all possible such matrices and vectors, while $\#(Q, M)$ is the numbers of configurations that leads to the overlap matrix Q and magnetization vector M.

In this form, it is not yet possible to perform the analytic continuation when $n \to 0$. Keeping for a moment n integer, it is however natural to expect that the number of such configurations

(for a given overlap matrix and magnetization vector), to be exponentially large. Denoting $\#(Q,M) = e^{Ns(Q,M)}$ we thus write

$$e^{nN\left(\log 2+\frac{1}{2\Delta}\right)}\mathbb{E}Z^n \approx \int \mathrm{d}Q\,\mathrm{d}M e^{Ns(Q,M)+\frac{N}{\Delta}\left(\sum_{a=1}^n M_a+\frac{1}{2}\sum_{a,b=1}^n Q_{a,b}\right)} =: \int \mathrm{d}Q,M e^{Ng(\Delta,Q,M)}$$

As N is large, we thus expect to be able to perform a Laplace (or Saddle) approximation by choosing the "right" structure of the matrix Q and vector, that will "dominate" the sum. A quite natural guess is the replica symmetric ansatz, where we assume that all replicas are identical, and therefore the system should be invariant under the relabelling of the replicas (permutation symmetry).

In this case, we have only three natural choices for the entries of Q and M:

1. All the replicas are in the same, identical configuration, that is $i_a = i \forall a$. Let us further assume that $i \neq 1$. In this case $Q_{ab} = 1$ for all a, b, and $M_a = 0$ for all a. There are $d - 1 \approx 2^N$ possibility for this so $s(Q, M) = \log 2$ and we find $g(\beta, Q) = \log 2 + n^2/\Delta$. This does not look right: this expression does *not* have a limit with a linear part in n, so we cannot use this solution in the replica method. Clearly, this is a wrong analytical continuation.

2. All the replica are in the same, identical configuration, which is the "correct" one $i_a = i = 1$. Then $Q_{ab} = 1$ for all a, b, and $M_a = 1$ for all a. There is only one possibility, so that $s(Q, M) = 0$ and $g(\beta, Q) = n/\Delta + n^2/2\Delta$.

3. If instead all replicas are in a different, random, configurations then $Q_{aa} = 1$, $Q_{ab} = 0$ for all $a \neq b$ and $M_a = 0$. In this case $\#(Q) = 2^N(2^N - 1)\dots(2^N - n + 1)$, so that $s(Q) \approx n \log 2$ if $n \ll N$. Therefore $g(\beta, Q) = n/2\Delta + n \log 2$.

At the replica symmetric level, we thus find that the free entropy is given by two possible solutions as $n \to 0$. In the first one all replicas are in the correct, hidden solution:

$$\mathbb{E}Z^n = e^{-nN\left(\log 2+\frac{1}{2\Delta}\right)+nN\frac{1}{\Delta}} = e^{-nN\left(\log 2-\frac{1}{2\Delta}\right)} \tag{7.92}$$

while in the second case, all replicas are distributed randomly over all states and

$$\mathbb{E}Z^n = e^{-nN\left(\log 2+\frac{1}{2\Delta}\right)+nN\frac{1}{2\Delta}+n\log 2} = 0\,. \tag{7.93}$$

We thus have recovered exactly the rigorous solution from the replica method. Indeed, choosing the right solution is easy: the free entropy is continuous and convex in Δ, non-negative, and goes from ∞ to 0 as Δ grows, so that we the free energy *must* be $\log 2 - 1/2\Delta$ for $\Delta < \Delta_c = 1/2 \log 2$ and 0 for $\Delta > \Delta_c$.

Chapter 8

Low-Rank Matrix Factorization: the Spike model

> The signal is the truth. The noise is what distracts us from the truth [...] Distinguishing the signal from the noise requires both scientific knowledge and self-knowledge: the serenity to accept the things we cannot predict, the courage to predict the things we can, and the wisdom to know the difference.
>
> Nate Silver, The Signal and the Noise —- 2012

Now that we presented Bayesian estimation problems, we can apply our techniques to a non-trivial problem. This is a perfect example for testing our newfound knowledge.

8.1 Problem Setting

The Spike-Wigner Model

Suppose we are given as data a $n \times n$ symmetric matrix Y created as follows:

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}} \underbrace{\mathbf{x}^*\mathbf{x}^{*\intercal}}_{N\times N \text{ rank-one matrix}} + \underbrace{\boldsymbol{\xi}}_{\text{symmetric iid noise}}$$

where $\mathbf{x}^* \in \mathbb{R}^N$ with $x_i^* \overset{\text{i.i.d.}}{\sim} P_X(x)$, $\xi_{ij} = \xi_{ji} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i \leq j$.

This is called the Wigner spike model in statistics. The name "Wigner" refer to the fact that Y is a Wigner matrix (a symmetric random matrix with components sampled randomly from a Gaussian distribution) plus a "spike", that is a rank one matrix $\mathbf{x}^*\mathbf{x}^{*\intercal}$.

Our task shall be to recover the vector $\mathbf{x}$ from the knowledge of Y. As we just learned, this

can be achieved using the posterior estimation:

$$P\left(\mathbf{x} \,|\, \mathbf{Y}\right) = \frac{1}{Z(\mathbf{Y})} \left[\prod_{i=1}^{N} P_X(x_i)\right] \left[\prod_{i\leq j} \frac{e^{-\frac{1}{2}\left(y_{ij} - \sqrt{\frac{\lambda}{N}} x_i^* x_j^*\right)^2}}{\sqrt{2\pi}}\right]$$

Spike-Wishart Model (Bipartite Vector-Spin Glass Model)

For completeness, we present here an alternative model, which is also extremely interesting, called the Wishart-spike model. In this case

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}} \underbrace{\mathbf{u}^* \mathbf{v}^{*\intercal}}_{M\times N \text{ rank-one matrix}} + \underbrace{\boldsymbol{\xi}}_{\text{iid noise}}$$

where $\mathbf{u}^* \in \mathbb{R}^M$ with $u_i^* \overset{\text{i.i.d.}}{\sim} P_U(u)$, $\mathbf{v}^* \in \mathbb{R}^N$ with $v_i^* \overset{\text{i.i.d.}}{\sim} P_V(v)$, $\xi_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Strictly speaking, the name "Wishart" might sounds strange here. This is coming from the fact that this model, for Gaussian vectors $\mathbf{u}$, is exactly the same as another model involving a Wishart matrix, a model also called the Spiked Covariance Model. Indeed, when the factors are independent, the model can be viewed as a linear model with additive noise and scalar random design:

$$\mathbf{y}_i = \sqrt{\frac{\lambda}{N}} v_j \mathbf{u} + \xi_j \,, \tag{8.1}$$

Assuming the v_j have zero mean and unit variance, this indeed is a model of spiked covariance: the mean of the empirical covariance matrix $\Sigma = \frac{YY^Y}{N}$ is a rank one perturbation of the identity $\mathbf{1} + \mathbf{u}\mathbf{u}^T$ Random covariance matrices are called Wishart matrices, so this is a model with a rank one perturbation of a Wishart matrix.

Regardless of its name, given the matrix Y, the posterior distribution over X reads

$$P\left(\mathbf{u}, \mathbf{v} \,|\, \mathbf{Y}\right) = \frac{1}{Z(\mathbf{Y})} \left[\prod_{i=1}^{M} P_U(u_i)\right] \left[\prod_{j=1}^{N} P_V(v_i)\right] \left[\prod_{i,j} \frac{e^{-\frac{1}{2}\left(y_{ij} - \sqrt{\frac{\lambda}{N}} u_i^* v_j^*\right)^2}}{\sqrt{2\pi}}\right]$$

8.2 From the Posterior Distribution to the partition sum

We shall now make a mapping to a Statistical Physics formulation. Consider the spike-Wigner model, using Bayes rule we write:

$$
\begin{aligned}
P\left(\mathbf{x}\,|\,\mathbf{Y}\right) = \frac{P\left(\mathbf{Y}\,|\,\mathbf{x}\right)P(\mathbf{x})}{P(\mathbf{Y})} &\propto \left[\prod_i P_X(x_i)\right]\left[\prod_{i\leq j}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(y_{ij}-\sqrt{\frac{\lambda}{N}}x_ix_j\right)^2}\right] \\
&\propto \left[\prod_i P_X(x_i)\right]\exp\left(\sum_{i\leq j}\left[-\frac{\lambda}{2N}x_i^2x_j^2+\sqrt{\frac{\lambda}{N}}y_{ij}x_ix_j\right]\right) \\
\Rightarrow \quad P\left(\mathbf{x}\,|\,\mathbf{Y}\right) &= \frac{1}{Z(\mathbf{Y})}\left[\prod_i P_X(x_i)\right]\exp\left(\sum_{i\leq j}\left[-\frac{\lambda}{2N}x_i^2x_j^2+\sqrt{\frac{\lambda}{N}}y_{ij}x_ix_j\right]\right) \\
\Rightarrow \quad \hat{\mathbf{x}}_{\mathrm{MSE}}(\mathbf{Y}) &= \begin{bmatrix}\hat{x}_{\mathrm{MSE},1}(\mathbf{Y})\\ \vdots \\ \hat{x}_{\mathrm{MSE},N}(\mathbf{Y})\end{bmatrix}, \qquad \hat{x}_{\mathrm{MSE},i}(\mathbf{Y}) = \langle x_i\rangle_{\mathbf{Y}} = \int \mathrm{d}\mathbf{x}\, P\left(\mathbf{x}\,|\,\mathbf{Y}\right)\, x_i
\end{aligned}
$$

8.3 Replica Method

$$
\begin{aligned}
P\left(\mathbf{x}\,|\,\mathbf{Y}\right) &= \frac{1}{Z(\mathbf{Y})}\left[\prod_i P_X(x_i)\right]\exp\left(\sum_{i\leq j}\left[-\frac{\lambda}{2N}x_i^2x_j^2+\sqrt{\frac{\lambda}{N}}y_{ij}x_ix_j\right]\right) \\
&= \frac{1}{Z(\mathbf{Y})}\left[\prod_i P_X(x_i)\right]\exp\left(\sum_{i\leq j}\left[-\frac{\lambda}{2N}x_i^2x_j^2+\frac{\lambda}{N}x_ix_jx_i^*x_j^*+\sqrt{\frac{\lambda}{N}}\xi_{ij}x_ix_j\right]\right)
\end{aligned}
$$

We are interested in

$$
\lim_{N\to\infty}\mathbb{E}_{\mathbf{Y}}\left[\frac{1}{N}\log\left(Z(\mathbf{Y})\right)\right] = \lim_{N\to\infty}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[\frac{1}{N}\log\left(Z(\mathbf{Y})\right)\right]
$$

Using replica method, we have

$$\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[Z^n\right] = \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[\left(\int \mathrm{d}\mathbf{x}\,\left[\prod_i P_X(x_i)\right]\exp\left(\sum_{i\le j}\left[-\frac{\lambda}{2N}x_i^2x_j^2+\frac{\lambda}{N}x_ix_jx_i^*x_j^*+\sqrt{\frac{\lambda}{N}}\xi_{ij}x_ix_j\right]\right)\right)^n\right]$$

$$= \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[\prod_{\alpha=1}^n\int \mathrm{d}\mathbf{x}^{(\alpha)}\prod_i P_X\left(x_i^{(\alpha)}\right)\exp\left(\sum_{i\le j}\left[-\frac{\lambda}{2N}\left(x_i^{(\alpha)}\right)^2\left(x_j^{(\alpha)}\right)^2+\frac{\lambda}{N}x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^*+\sqrt{\frac{\lambda}{N}}\xi_{ij}x_i^{(\alpha)}x_j^{(\alpha)}\right]\right)\right]$$

$$= \mathbb{E}_{\mathbf{x}^*}\left[\int\prod_{\alpha,i}P_X\left(x_i^{(\alpha)}\right)\mathrm{d}x_i^{(\alpha)}\,\exp\left(\sum_{i\le j}\left[-\frac{\lambda}{2N}\sum_\alpha\left(x_i^{(\alpha)}\right)^2\left(x_j^{(\alpha)}\right)^2+\frac{\lambda}{N}\sum_\alpha x_i^{(\alpha)}x_j^{(\alpha)}x_i^*x_j^*\right]\right)\right.$$
$$\left.\underbrace{\prod_{i\le j}\mathbb{E}_{\xi_{ij}}\left[e^{\xi_{ij}\left(\sqrt{\frac{\lambda}{N}}\sum_\alpha x_i^{(\alpha)}x_j^{(\alpha)}\right)}\right]}_{\overset{(a)}{=}\exp\left(\frac{\lambda}{2N}\sum_{i\le j}\sum_{\alpha,\beta}x_i^{(\alpha)}x_j^{(\alpha)}x_i^{(\beta)}x_j^{(\beta)}\right)}\right]$$

$$\overset{(b)}{=} \mathbb{E}_{\mathbf{x}^*}\left[\int\prod_{\alpha,i}P_X\left(x_i^{(\alpha)}\right)\mathrm{d}x_i^{(\alpha)}\,\exp\left(-\frac{\lambda N}{4}\sum_\alpha\left(\sum_i\frac{\left(x_i^{(\alpha)}\right)^2}{N}\right)^2+\frac{\lambda N}{2}\sum_\alpha\left(\sum_i\frac{x_i^*x_i^{(\alpha)}}{N}\right)^2+\frac{\lambda N}{4}\sum_{\alpha,\beta}\left(\sum_i\frac{x_i^{(\alpha)}x_i^{(\beta)}}{N}\right)^2\right)\right.$$

$$= \mathbb{E}_{\mathbf{x}^*}\left[\int\prod_{\alpha,i}P_X\left(x_i^{(\alpha)}\right)\mathrm{d}x_i^{(\alpha)}\,\exp\left(\frac{\lambda N}{2}\sum_\alpha\left(\sum_i\frac{x_i^*x_i^{(\alpha)}}{N}\right)^2+\frac{\lambda N}{2}\sum_{\alpha<\beta}\left(\sum_i\frac{x_i^{(\alpha)}x_i^{(\beta)}}{N}\right)^2\right)\right]$$

$$\overset{(c)}{=} \mathbb{E}_{\mathbf{x}^*}\left[\int\prod_{\alpha,i}P_X\left(x_i^{(\alpha)}\right)\mathrm{d}x_i^{(\alpha)}\int\prod_\alpha\delta\left(m_\alpha-\frac{1}{N}\sum_i x_i^{(\alpha)}x_i^*\right)\mathrm{d}m_\alpha\int\prod_{\alpha<\beta}\delta\left(q_{\alpha\beta}-\frac{1}{N}\sum_i x_i^{(\alpha)}x_i^{(\beta)}\right)\mathrm{d}q_{\alpha\beta}\right.$$
$$\left.\exp\left(\frac{\lambda N}{2}\left[\sum_\alpha m_\alpha^2+\sum_{\alpha<\beta}q_{\alpha\beta}^2\right]\right)\right]$$

$$\overset{(d)}{=} \mathbb{E}_{\mathbf{x}^*}\left[\int\prod_{\alpha,i}P_X\left(x_i^{(\alpha)}\right)\mathrm{d}x_i^{(\alpha)}\int\prod_\alpha e^{\hat{m}_\alpha\left[Nm_\alpha-\sum_i x_i^{(\alpha)}x_i^*\right]}\mathrm{d}\hat{m}_\alpha\,\mathrm{d}m_\alpha\int\prod_{\alpha<\beta}e^{\hat{q}_{\alpha\beta}\left[Nq_{\alpha\beta}-\sum_i x_i^{(\alpha)}x_i^{(\beta)}\right]}\mathrm{d}\hat{q}_{\alpha\beta}\,\mathrm{d}q_{\alpha\beta}\right.$$
$$\left.\exp\left(\frac{\lambda N}{2}\left[\sum_\alpha m_\alpha^2+\sum_{\alpha<\beta}q_{\alpha\beta}^2\right]\right)\right]$$

$$\overset{(e)}{=}\int\prod_\alpha\mathrm{d}\hat{m}_\alpha\,\mathrm{d}m_\alpha\int\prod_{\alpha<\beta}\mathrm{d}\hat{q}_{\alpha\beta}\,\mathrm{d}q_{\alpha\beta}\,\exp\left(\frac{\lambda N}{2}\left[\sum_\alpha m_\alpha^2+\sum_{\alpha<\beta}q_{\alpha\beta}^2\right]+N\left[\sum_\alpha m_\alpha\hat{m}_\alpha+\sum_{\alpha<\beta}q_{\alpha\beta}\hat{q}_{\alpha\beta}\right]\right)$$
$$\left\{\mathbb{E}_{x_*}\left[\int\prod_\alpha P_X(x_\alpha)\,\mathrm{d}x_\alpha\,\exp\left(-\sum_\alpha\hat{m}_\alpha x_*x_\alpha-\sum_{\alpha<\beta}\hat{q}_{\alpha\beta}x_\alpha x_\beta\right)\right]\right\}^N$$

where

(a) uses the fact that $\int\mathcal{D}z\,e^{az}=e^{a^2/2}$

(b) uses the fact that

$$\sum_{i\le j}\frac{a_i}{N}\frac{a_j}{N}=\frac{1}{2}\left(\sum_i\frac{a_i}{N}\right)^2+\frac{1}{2}\sum_i\frac{a_i^2}{N^2}$$

and neglect the second term since it scales like $\mathrm{O}(N^{-1})$.

(c) partitions the huge integral according to overlap with true signal m_α and overlap between two distinct replicas $q_{\alpha\beta}$ with definitions

$$m_\alpha=\frac{1}{N}\sum_i x_i^{(\alpha)}x_i^*,\quad\forall\,\alpha,\qquad\qquad q_{\alpha\beta}=\frac{1}{N}\sum_i x_i^{(\alpha)}x_i^{(\beta)},\quad\forall\,\alpha<\beta$$

(d) applies Fourier transformation and change of variable

(e) change the order of integral and expectation and x_i^* are iid, for each i, the tuple $\left(x_i^*, x_i^{(1)}, \cdots, x_i^{(n)}\right)$ are identical distributed, so we switch to subscript notation $\left(x_*, x_{(1)}, \cdots, x_{(n)}\right)$ to get rid of component index i but keep the replica index α.

Replica Symmetry Ansatz

Under replica symmetry Ansatz, we have $m_\alpha \equiv m$, $\hat{m}_\alpha \equiv \hat{m}$, $q_{\alpha\beta} \equiv q$, $\hat{q}_{\alpha\beta} \equiv \hat{q}$.

$$
\begin{aligned}
\mathbb{E}[Z^n] =& \int \mathrm{d}\hat{m}\,\mathrm{d}m \int \mathrm{d}\hat{q}\,\mathrm{d}q\; e^{\frac{\lambda N}{2}\left[nm^2+\frac{n^2-n}{2}q^2\right]+N\left[nm\hat{m}+\frac{n^2-n}{2}q\hat{q}\right]}\times \\
&\times\left\{\mathbb{E}_{x_*}\left[\int\prod_\alpha P_X\left(x_\alpha\right)\mathrm{d}x_\alpha\; e^{-\hat{m}\sum_\alpha x_* x_\alpha-\hat{q}\sum_{\alpha<\beta}x_\alpha x_\beta}\right]\right\}^N \\
\overset{(a)}{=}& \int \mathrm{d}\hat{m}\,\mathrm{d}m\,\mathrm{d}\hat{q}\,\mathrm{d}q\; e^{nN\left[\frac{\lambda}{2}m^2+\frac{\lambda}{4}(n-1)q^2+m\hat{m}+\frac{n-1}{2}q\hat{q}\right]}\times \\
&\times\left\{\mathbb{E}_{x_*}\left[\int\prod_\alpha P_X\left(x_\alpha\right)\mathrm{d}x_\alpha\; e^{\frac{\hat{q}}{2}\sum_\alpha x_\alpha^2-\hat{m}\sum_\alpha x_* x_\alpha}\int\mathcal{D}z\; e^{-\mathbb{i}z\sqrt{\hat{q}}\sum_\alpha x_\alpha}\right]\right\}^N \\
\overset{(b)}{=}& \int \mathrm{d}\hat{m}\,\mathrm{d}m\,\mathrm{d}\hat{q}\,\mathrm{d}q\; e^{nN\left[\frac{\lambda}{2}m^2+\frac{\lambda}{4}(n-1)q^2+m\hat{m}+\frac{n-1}{2}q\hat{q}\right]}\times \\
&\times\left\{\mathbb{E}_{x_*}\left[\int\mathcal{D}z\prod_\alpha\left\{\int P_X\left(x_\alpha\right)\mathrm{d}x_\alpha\; e^{\frac{\hat{q}}{2}x_\alpha^2-\hat{m}x_* x_\alpha-\mathbb{i}z\sqrt{\hat{q}}x_\alpha}\right\}\right]\right\}^N \\
\overset{(c)}{=}& \int \mathrm{d}\hat{m}\,\mathrm{d}m\,\mathrm{d}\hat{q}\,\mathrm{d}q\; e^{nN\left[\frac{\lambda}{2}m^2+\frac{\lambda}{4}(n-1)q^2+m\hat{m}+\frac{n-1}{2}q\hat{q}\right]}\times \\
&\times\left\{\mathbb{E}_{x_*}\left[\int\mathcal{D}z\;\left\{\mathbb{E}_x\left[e^{\frac{\hat{q}}{2}x^2-\hat{m}x_* x-\mathbb{i}z\sqrt{\hat{q}}x}\right]\right\}^n\right]\right\}^N \\
\overset{(d)}{=}& \int \mathrm{d}\hat{m}\,\mathrm{d}m\,\mathrm{d}\hat{q}\,\mathrm{d}q\; e^{nN\left[\frac{\lambda}{2}m^2-\frac{\lambda}{4}q^2+m\hat{m}-\frac{1}{2}q\hat{q}\right]}\times \\
&\times e^{nN\,\mathbb{E}_{x_*}\left[\int\mathcal{D}z\;\log\left(\int P_X(x)\,\mathrm{d}x\; e^{\frac{\hat{q}}{2}x^2-\hat{m}x_* x-\mathbb{i}z\sqrt{\hat{q}}x}\right)\right]} \\
=& \int \mathrm{d}\hat{m}\,\mathrm{d}m\,\mathrm{d}\hat{q}\,\mathrm{d}q\; e^{nN\Phi(m,q,\hat{m},\hat{q})}
\end{aligned}
$$

where

$$
\Phi(m,q,\hat{m},\hat{q}) = \frac{\lambda}{4}(2m^2-q^2)+m\hat{m}-\frac{1}{2}q\hat{q}+\mathbb{E}_{x_*}\left[\int\mathcal{D}z\;\log\left(\int P_X(x)\,\mathrm{d}x\; e^{\frac{\hat{q}}{2}x^2+(\sqrt{\hat{q}}z-\hat{m}x_*)x}\right)\right]
$$

and

(a) uses the Hubbard–Stratonovich transformation

$$
\exp\left(-\frac{a}{2}x^2\right) = \frac{1}{\sqrt{2\pi a}}\int_{-\infty}^{\infty}\exp\left(-\frac{z^2}{2a}-\mathbb{i}xz\right)\mathrm{d}z, \qquad \forall\, a>0
$$

Then we have

$$\exp\left(-\hat{q}\sum_{\alpha<\beta}x_\alpha x_\beta\right)=\exp\left(-\frac{\hat{q}}{2}\sum_{\alpha\neq\beta}x^\alpha x^\beta\right)=\exp\left(\frac{\hat{q}}{2}\sum_\alpha x_\alpha^2-\frac{\hat{q}}{2}\left(\sum_\alpha x_\alpha\right)^2\right)$$
$$=\exp\left(\frac{\hat{q}}{2}\sum_\alpha x_\alpha^2\right)\int\frac{1}{\sqrt{2\pi}}\mathrm{d}z\ \exp\left(-\frac{z^2}{2}-\mathbb{i}z\sqrt{\hat{q}}\sum_\alpha x_a\right)$$
$$=\exp\left(\frac{\hat{q}}{2}\sum_\alpha x_\alpha^2\right)\int\mathcal{D}z\ \exp\left(-\mathbb{i}z\sqrt{\hat{q}}\sum_\alpha x_a\right)$$

(b) exchange order of integrals, and split x_α's

(c) use the fact that x_α's are i.i.d.

(d) take limit $n\to 0$ and use the fact that when n is small we have

$$\mathbb{E}[X^n]=\mathbb{E}\left[e^{n\log(X)}\right]\simeq\mathbb{E}\left[1+n\log(X)\right]=1+n\mathbb{E}\left[\log(X)\right]$$
$$=\exp\left(\log\left(1+n\mathbb{E}\left[\log(X)\right]\right)\right)\simeq\exp\left(n\mathbb{E}\left[\log(X)\right]\right)$$

Recall that according to the Nishimori identity we have $q=m$ and $\hat{q}=\hat{m}$, it simplifies to

$$\Phi_{\text{Nishi}}(m,\hat{m})\triangleq\Phi(m,q,\hat{m},\hat{q})|_{q=m,\hat{q}=\hat{m}}=\frac{\lambda}{4}m^2+\frac{1}{2}m\hat{m}$$
$$+\mathbb{E}_{x_*}\left[\int\mathcal{D}z\ \log\left(\int P_X(x)\,\mathrm{d}x\ e^{\frac{\hat{m}}{2}x^2-(\mathbb{i}\sqrt{\hat{m}}z+\hat{m}x_*)x}\right)\right]$$

We can further reduce the problem by taking partial derivative w.r.t. m and set it to zero

$$\frac{\partial}{\partial m}\Phi_{\text{Nishi}}(m,\hat{m})=\frac{\lambda}{2}m+\frac{1}{2}\hat{m}=0\quad\Rightarrow\quad\hat{m}=-\lambda m$$

Plug this back to $\Phi_{\text{Nishi}}(m,\hat{m})$ we will obtain the final free entropy function under replica symmetry Ansatz

$$\Phi_{\text{RS}}(m)\triangleq\Phi_{\text{Nishi}}(m,\hat{m})|_{\hat{m}=-\lambda m}$$
$$=-\frac{\lambda}{4}m^2+\mathbb{E}_{x_*}\left[\int\mathcal{D}z\ \log\left(\int P_X(x)\,\mathrm{d}x\ e^{\frac{-\lambda m}{2}x^2-(\mathbb{i}\sqrt{-\lambda m}z-\lambda m x_*)x}\right)\right]$$
$$=-\frac{\lambda}{4}m^2+\mathbb{E}_{x_*}\left[\int\mathcal{D}z\ \log\left(\int P_X(x)\,\mathrm{d}x\ e^{\frac{-\lambda m}{2}x^2+(\lambda m x_*-\sqrt{\lambda m}z)x}\right)\right]$$
$$=-\frac{\lambda}{4}m^2+\mathbb{E}_{x_*,z}\left[\log\left(\int P_X(x)\,\mathrm{d}x\ e^{\frac{-\lambda m}{2}x^2+(\lambda m x_*+\sqrt{\lambda m}z)x}\right)\right]$$

where the last step is because the standard normal distribution is symmetric around zero and a change of variable $z\leftarrow -z$.

The self-consistent equation on m thus reads

$$m=\mathbb{E}_{x_*,z}\frac{\int\mathrm{d}x\ P_X(x)\ xx^*e^{\frac{-\lambda m}{2}x^2+(\lambda m x_*-\sqrt{\lambda m}z)x}}{\int\mathrm{d}x\ P_X(x)e^{\frac{-\lambda m}{2}x^2+(\lambda m x_*-\sqrt{\lambda m}z)x}}\tag{8.2}$$

which involves, at worst, 3 integrals, and is therefore tractable numerically.

8.4 A rigorous proof via Interpolation

Preliminaries

Lemma 11 (Stein's Lemma). *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Let g be a differentiable function such that the expectation $\mathbb{E}\left[(X-\mu)g(X)\right]$ and $\mathbb{E}\left[g'(X)\right]$ exists, then we have*

$$\mathbb{E}\left[g(X)(X-\mu)\right] = \sigma^2 \mathbb{E}\left[g'(X)\right]$$

Particularly, when $X \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}\left[Xg(X)\right] = \mathbb{E}\left[g'(X)\right]$$

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

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

Corollary 1 (Nishimori Identity for Two Replicas). *Consider model $y = g(x^*) + w$, where g is a continuous bounded function and w is the additive noise. Let us denote $\langle \cdot \rangle_{x^*,w}$ the expectation w.r.t. $P(X = \cdot \,|\, Y = g(x^*) + w)$. Then we have*

$$\mathbb{E}_{X^*,W}\left[\langle f(X, X^*)\rangle_{X^*,W}\right] = \mathbb{E}_{X^*,W}\left[\left\langle f\left(X^{(1)}, X^{(2)}\right)\right\rangle_{X^*,W}\right]$$

where $X^{(1)}, X^{(2)}$ are two independent replicas distributed as $P(X = \cdot \,|\, Y = g(x^*) + w)$

Two Problems

Problem A: From previous lecture we studied the scalar denoising problem that $y = \sqrt{\lambda}x^* + \omega$

$$P(x \,|\, y) \propto \exp\left(\log\left(P_X(x)\right) - \frac{\lambda}{2}x^2 + \lambda x^* x + \sqrt{\lambda}\omega x\right) \tag{8.3}$$

$$\Phi_{\text{denoising}}(\lambda) = \mathbb{E}_{x^*,\omega}\left[\log\left\{\int P_X(x)\,\mathrm{d}x\ \exp\left(-\frac{\lambda}{2}x^* + \lambda x^* x + \sqrt{\lambda}xz\right)\right\}\right] \tag{8.4}$$

Suppose we solve N such problem parallelly such that $\mathbf{y} = \sqrt{\lambda m}\mathbf{x}^* + \boldsymbol{\omega}$ as problem A and define the Hamiltonian $\mathcal{H}_A(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}; m)$

$$P(\mathbf{x} \,|\, \mathbf{y}) \propto \exp\left(\sum_i \log\left(P_X(x_i)\right) + \sum_i \left[-\frac{\lambda m}{2}x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\right]\right) \tag{8.5}$$

$$\mathcal{H}_A(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}; m) \triangleq -\sum_i \log\left(P_X(x_i)\right) - \sum_i \left[-\frac{\lambda m}{2}x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\right] \tag{8.6}$$

with partition function

$$
\begin{aligned}
Z_A\left(\lambda, \mathbf{x}^*, \boldsymbol{\omega}; m\right) &= \int \prod_i \mathrm{d}x_i \ \exp\left(\sum_i \left[\log\left(P_X(x_i)\right) - \frac{\lambda m}{2} x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\right]\right) \\
&= \prod_i \int P_X(x_i)\,\mathrm{d}x_i \ \exp\left(-\frac{\lambda m}{2} x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\right) \\
\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}}\left[\frac{\log\left(Z_A\left(\lambda, \mathbf{x}^*, \boldsymbol{\omega}; m\right)\right)}{N}\right] &= \frac{1}{N}\sum_i \mathbb{E}_{x_i^*,\omega_i}\left[\log\left\{\int P_X(x_i)\,\mathrm{d}x_i \ \exp\left(-\frac{\lambda m}{2} x_i^2 + \lambda m x_i^* x_i + \sqrt{\lambda m}\omega_i x_i\right.\right.\right. \\
&= \mathbb{E}_{x_1^*,\omega_1}\left[\log\left\{\int P_X(x_1)\,\mathrm{d}x_1 \ \exp\left(-\frac{\lambda m}{2} x_1^2 + \lambda m x_1^* x_1 + \sqrt{\lambda m}\omega_1 x_1\right)\right\}\right] \\
&= \Phi_{\text{denoising}}(\lambda m)
\end{aligned}
$$

Problem B: Our target rank-one matrix factorization problem $\mathbf{Y} = \sqrt{\frac{\lambda}{N}}\mathbf{x}^*\mathbf{x}^{*\intercal} + \boldsymbol{\xi}$, define the Hamiltonian $\mathcal{H}_B\left(\mathbf{x}, \mathbf{x}^*, \lambda\right)$

$$
P\left(\mathbf{x} \,|\, \mathbf{Y}\right) \propto \exp\left(\sum_i \log\left(P_X(x_i)\right) + \sum_{i\le j}\left[-\frac{\lambda}{2N} x_i^2 x_j^2 + \frac{\lambda}{N} x_i^* x_j^* x_i x_j + \sqrt{\frac{\lambda}{N}}\xi_{ij} x_i x_j\right]\right) \tag{8.9}
$$

$$
\mathcal{H}_B\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\xi}\right) \triangleq -\sum_i \log\left(P_X(x_i)\right) - \sum_{i\le j}\left[-\frac{\lambda}{2N} x_i^2 x_j^2 + \frac{\lambda}{N} x_i^* x_j^* x_i x_j + \sqrt{\frac{\lambda}{N}}\xi_{ij} x_i x_j\right] \tag{8.10}
$$

with partition function $Z_B\left(\lambda, \mathbf{x}^*, \boldsymbol{\xi}\right)$ is the quantity that we are interested in.

Lower Bound by Guerra's Interpolation

Define

$$
\begin{aligned}
\tilde{\mathcal{H}}_A^{(t)}\left(\mathbf{x}\lambda, \mathbf{x}^*, \boldsymbol{\omega}; m\right) &\triangleq \mathcal{H}_A\left(\mathbf{x}, t\lambda, \mathbf{x}^*, \boldsymbol{\omega}; m\right) \\
&= -\sum_i \log\left(P_X(x_i)\right) - \sum_i\left[-\frac{t\lambda m}{2} x_i^2 + t\lambda m x_i^* x_i + \sqrt{t\lambda m}\omega_i x_i\right] \\
\tilde{\mathcal{H}}_B^{(t)}\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\xi}\right) &\triangleq \mathcal{H}_B\left(\mathbf{x}, t\lambda, \mathbf{x}^*, \boldsymbol{\xi}\right) \\
&= -\sum_i \log\left(P_X(x_i)\right) - \sum_{i\le j}\left[-\frac{t\lambda}{2N} x_i^2 x_j^2 + \frac{t\lambda}{N} x_i^* x_j^* x_i x_j + \sqrt{\frac{t\lambda}{N}}\xi_{ij} x_i x_j\right]
\end{aligned}
$$

And the interpolated Hamiltonian as

$$
\begin{aligned}
\mathcal{H}_t\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right) &\triangleq \tilde{\mathcal{H}}_A^{(1-t)}\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}; m\right) + \tilde{\mathcal{H}}_B^{(t)}\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\xi}\right) \\
&= -\sum_i\left[-\frac{(1-t)\lambda m}{2} x_i^2 + (1-t)\lambda m x_i^* x_i + \sqrt{(1-t)\lambda m}\omega_i x_i\right] \\
&\quad -\sum_{i\le j}\left[-\frac{t\lambda}{2N} x_i^2 x_j^2 + \frac{t\lambda}{N} x_i^* x_j^* x_i x_j + \sqrt{\frac{t\lambda}{N}}\xi_{ij} x_i x_j\right]
\end{aligned} \tag{8.11}
$$

with partition function $Z_t\left(\lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right)$.

This is a problem where we have ground truth $\mathbf{x}^*$ and observations

$$\begin{cases} Y_{ij} = \sqrt{\frac{t\lambda}{N}} x_i^* x_j^* + \xi_{ij}, & \forall\, 1 \le i \le j \le N \\ y_i = \sqrt{(1-t)\lambda m} x_i^* + \omega_i, & \forall\, 1 \le i \le N \end{cases}$$

Besides, it is easy to verify that

$$\mathcal{H}_0\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right) \equiv \mathcal{H}_A\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}; m\right), \qquad Z_0\left(\lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right) \equiv Z_A\left(\lambda, \mathbf{x}^*, \boldsymbol{\omega}; m\right), \quad (8.12)$$

$$\mathcal{H}_1\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right) \equiv \mathcal{H}_B\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\xi}\right), \qquad Z_1\left(\lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right) \equiv Z_B\left(\lambda, \mathbf{x}^*, \boldsymbol{\xi}\right), \qquad \forall\, m \quad (8.13)$$

Notice that

$$\begin{aligned} &\frac{\partial}{\partial t}\mathcal{H}_t\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right) \\ &\quad = -\sum_i \left[\frac{\lambda m}{2}\sum_i x_i^2 - \lambda m x_i^* x_i - \frac{\sqrt{\lambda m}}{2\sqrt{1-t}}\omega_i x_i\right] - \sum_{i\le j}\left[-\frac{\lambda}{2N}x_i^2 x_j^2 + \frac{\lambda}{N}x_i^* x_j^* x_i x_j + \frac{\sqrt{\lambda/N}}{2\sqrt{t}}\xi_{ij} x_i x_j\right] \end{aligned} \quad (8.14)$$

Therefore this derivative can be into several Boltzmann average terms associated to $\mathcal{H}_t\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m\right)$. For short we denote $\boldsymbol{\theta} = \{\lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}\}$

$$\begin{aligned} \frac{\partial}{\partial t}\frac{\log\left(Z_t\left(\boldsymbol{\theta}; m\right)\right)}{N} &= \frac{1}{N}\frac{1}{Z_t\left(\boldsymbol{\theta}; m\right)}\frac{\partial}{\partial t}\int \mathrm{d}\mathbf{x}\, e^{-\mathcal{H}_t(\mathbf{x}, \boldsymbol{\theta}; m)} \\ &= -\frac{1}{N}\int \mathrm{d}\mathbf{x}\, \underbrace{\frac{e^{-\mathcal{H}_t(\mathbf{x}, \boldsymbol{\theta}; m)}}{Z_t\left(\boldsymbol{\theta}; m\right)}}_{=P_{t,\boldsymbol{\theta}; m}(\mathbf{x})}\frac{\partial}{\partial t}\mathcal{H}_t\left(\mathbf{x}, \boldsymbol{\theta}; m\right) = -\frac{1}{N}\left\langle \frac{\partial}{\partial t}\mathcal{H}_t\left(\mathbf{x}, \boldsymbol{\theta}; m\right)\right\rangle_{t,\boldsymbol{\theta},m} \\ &= -\frac{1}{N}\left\{\frac{\lambda}{N}\sum_{i\le j}\left[\frac{\left\langle x_i^2 x_j^2\right\rangle_{t,\boldsymbol{\theta},m}}{2} - \left\langle x_i^* x_j^* x_i x_j\right\rangle_{t,\boldsymbol{\theta},m}\right] - \lambda m\sum_i\left[\frac{\left\langle x_i^2\right\rangle_{t,\boldsymbol{\theta},m}}{2} - \left\langle x_i^* x_i\right\rangle_{t,\boldsymbol{\theta},m}\right]\right. \\ &\quad \left. -\frac{\sqrt{\lambda/N}}{2\sqrt{t}}\sum_{i\le j}\xi_{ij}\left\langle x_i x_j\right\rangle_{t,\boldsymbol{\theta},m} + \frac{\sqrt{\lambda m}}{2\sqrt{1-t}}\sum_i \omega_i\left\langle x_i\right\rangle_{t,\boldsymbol{\theta},m}\right\} \end{aligned} \quad (8.15)$$

Let $\Phi_{\mathrm{MF}}(\lambda)$ be the free entropy density of the rank-one matrix factorization problem under

SNR λ (problem B), from fundamental theorem of calculus we have

$$
\begin{aligned}
\Phi_{\mathrm{MF}}(\lambda) &= \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}} \left[\frac{\log\left(Z_B\left(\lambda,\mathbf{x}^*,\boldsymbol{\xi}\right)\right)}{N}\right] \\
&= \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\frac{\log\left(Z_1\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m\right)\right)}{N}\right] \\
&= \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\frac{\log\left(Z_0\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m\right)\right)}{N} + \int_0^1 \mathrm{d}\tau \ \frac{\partial}{\partial t}\frac{\log\left(Z_t\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m\right)\right)}{N}\bigg|_{t=\tau}\right] \\
&\overset{(a)}{=} \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}} \left[\frac{\log\left(Z_A\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m\right)\right)}{N}\right] + \int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\frac{\partial}{\partial t}\frac{\log\left(Z_t\left(\boldsymbol{\theta};m\right)\right)}{N}\bigg|_{t=\tau}\right] \\
&\overset{(b)}{=} \lim_{N\to\infty} \Bigg\{ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}} \left[\frac{\log\left(Z_A\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m\right)\right)}{N}\right] \\
&\qquad -\frac{\lambda}{N^2}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i\le j}\left\{\frac{\left\langle x_i^2 x_j^2\right\rangle_{\tau,\boldsymbol{\theta},m}}{2} - \left\langle x_i^* x_j^* x_i x_j\right\rangle_{\tau,\boldsymbol{\theta},m}\right\}\right] \\
&\qquad +\frac{\lambda m}{N}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i}\left\{\frac{\left\langle x_i^2\right\rangle_{\tau,\boldsymbol{\theta},m}}{2} - \left\langle x_i^* x_i\right\rangle_{\tau,\boldsymbol{\theta},m}\right\}\right] \\
&\qquad +\frac{\lambda}{2N^2}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i\le j}\left\{\left\langle x_i^2 x_j^2\right\rangle_{\tau,\boldsymbol{\theta},m} - \left\langle x_i x_j\right\rangle^2_{\tau,\boldsymbol{\theta},m}\right\}\right] \\
&\qquad -\frac{\lambda m}{2N}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i}\left\{\left\langle x_i^2\right\rangle_{\tau,\boldsymbol{\theta},m} - \left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m}\right\}\right]\Bigg\} \\
&= \lim_{N\to\infty} \Bigg\{ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}} \left[\frac{\log\left(Z_A\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m\right)\right)}{N}\right] \\
&\qquad -\frac{\lambda}{2N^2}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i\le j}\left\{\left\langle x_i x_j\right\rangle^2_{\tau,\boldsymbol{\theta},m} - 2\left\langle x_i^* x_j^* x_i x_j\right\rangle_{\tau,\boldsymbol{\theta},m}\right\}\right] \\
&\qquad +\frac{\lambda m}{2N}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i}\left\{\left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m} - 2\left\langle x_i^* x_i\right\rangle_{\tau,\boldsymbol{\theta},m}\right\}\right]\Bigg\} \\
&\overset{(c)}{=} \lim_{N\to\infty} \Bigg\{ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}} \left[\frac{\log\left(Z_A\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m\right)\right)}{N}\right] \\
&\qquad +\frac{\lambda m}{2N}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i}\left\{\left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m} - 2\left\langle x_i^* x_i\right\rangle_{\tau,\boldsymbol{\theta},m}\right\}\right] \\
&\qquad -\frac{\lambda}{4N^2}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i,j}\left\{\left\langle x_i x_j\right\rangle^2_{\tau,\boldsymbol{\theta},m} - 2\left\langle x_i^* x_j^* x_i x_j\right\rangle_{\tau,\boldsymbol{\theta},m}\right\}\right]\Bigg\} \\
&= \lim_{N\to\infty} \Bigg\{ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}} \left[\frac{\log\left(Z_A\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m\right)\right)}{N}\right] \\
&\qquad -\frac{\lambda}{4N^2}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i,j}\left\langle x_i x_j\right\rangle^2_{\tau,\boldsymbol{\theta},m} - 2Nm\sum_i \left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m}\right] \\
&\qquad +\frac{\lambda}{2N^2}\int_0^1 \mathrm{d}\tau\, \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}} \left[\sum_{i,j}\left\langle x_i^* x_j^* x_i x_j\right\rangle^2_{\tau,\boldsymbol{\theta},m} - 2Nm\sum_i \left\langle x_i^* x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m}\right]\Bigg\} \\
&\overset{(d)}{=} \lim_{N\to\infty} \Bigg\{ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}} \left[\frac{\log\left(Z_A\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m\right)\right)}{N}\right]
\end{aligned}
$$

where

(a) Uses Eqn 8.12, and the short hand notation $\boldsymbol{\theta} = \{\lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}\}$.

(b) Plug in Eqn 8.15 and uses the Stein's Lemma to deal with terms containing ξ_{ij} and ω_i

$$
\begin{aligned}
\frac{\partial}{\partial \xi_{ij}} \mathcal{H}_t\left(\mathbf{x}, \boldsymbol{\theta}; m\right) &= -\sqrt{\frac{t\lambda}{N}} x_i x_j \\
\frac{\partial}{\partial \xi_{ij}} Z_t\left(\boldsymbol{\theta}; m\right) &= -\int \mathrm{d}\mathbf{x}' \, e^{-\mathcal{H}_t(\mathbf{x}', \boldsymbol{\theta}, m)} \frac{\partial}{\partial \xi_{ij}} \mathcal{H}_t\left(\mathbf{x}', \boldsymbol{\theta}; m\right) \\
&= Z_t\left(\boldsymbol{\theta}; m\right) \cdot \sqrt{\frac{t\lambda}{N}} \int \mathrm{d}\mathbf{x}' \frac{e^{-\mathcal{H}_t(\mathbf{x}', \boldsymbol{\theta}; m)}}{Z_t\left(\boldsymbol{\theta}; m\right)} x_i' x_j' = Z_t\left(\boldsymbol{\theta}; m\right) \cdot \sqrt{\frac{t\lambda}{N}} \left\langle x_i' x_j' \right\rangle_{t, \boldsymbol{\theta}, m} \\
\mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\xi_{ij} \left\langle x_i x_j \right\rangle_{t, \boldsymbol{\theta}, m} \right] &= \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\frac{\partial}{\partial \xi_{ij}} \left\langle x_i x_j \right\rangle_{t, \boldsymbol{\theta}, m} \right] \\
&= \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\int \mathrm{d}\mathbf{x} \, x_i x_j e^{-\mathcal{H}_t(\boldsymbol{\theta}; m)} \left\{ \frac{-\frac{\partial}{\partial \xi_{ij}} \mathcal{H}_t\left(\mathbf{x}, \boldsymbol{\theta}; m\right)}{Z_t\left(\boldsymbol{\theta}; m\right)} - \frac{\frac{\partial}{\partial \xi_{ij}} Z_t\left(\boldsymbol{\theta}; m\right)}{\left[Z_t\left(\boldsymbol{\theta}; m\right)\right]^2} \right\} \right] \\
&= \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\int \mathrm{d}\mathbf{x} \, x_i x_j \frac{e^{-\mathcal{H}_t(\boldsymbol{\theta}; m)}}{Z_t\left(\boldsymbol{\theta}; m\right)} \sqrt{\frac{t\lambda}{N}} \left\{ x_i x_j + \left\langle x_i' x_j' \right\rangle_{t, \boldsymbol{\theta}, m} \right\} \right] \\
&= \sqrt{\frac{t\lambda}{N}} \, \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\left\langle x_i^2 x_j^2 \right\rangle_{t, \boldsymbol{\theta}, m} - \left\langle x_i x_j \right\rangle_{t, \boldsymbol{\theta}, m}^2 \right]
\end{aligned}
$$

Similarly, for the term containing ω_i, we have

$$
\begin{aligned}
\frac{\partial}{\partial \omega_i} \mathcal{H}_t\left(\mathbf{x}, \boldsymbol{\theta}; m\right) &= -\sqrt{(1-t)\lambda m} x_i \\
\frac{\partial}{\partial \omega_i} Z_t\left(\boldsymbol{\theta}; m\right) &= -\int \mathrm{d}\mathbf{x}' \, e^{-\mathcal{H}_t(\mathbf{x}', \boldsymbol{\theta}, m)} \frac{\partial}{\partial \omega_i} \mathcal{H}_t\left(\mathbf{x}', \boldsymbol{\theta}; m\right) \\
&= Z_t\left(\boldsymbol{\theta}; m\right) \cdot \sqrt{(1-t)\lambda m} \int \mathrm{d}\mathbf{x}' \frac{e^{-\mathcal{H}_t(\mathbf{x}', \boldsymbol{\theta}; m)}}{Z_t\left(\boldsymbol{\theta}; m\right)} x_i' x_j' \\
&= Z_t\left(\boldsymbol{\theta}; m\right) \cdot \sqrt{(1-t)\lambda m} \left\langle x_i' \right\rangle_{t, \boldsymbol{\theta}, m} \\
\mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\omega_i \left\langle x_i \right\rangle_{t, \boldsymbol{\theta}, m} \right] &= \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\frac{\partial}{\partial \omega_i} \left\langle x_i \right\rangle_{t, \boldsymbol{\theta}, m} \right] \\
&= \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\int \mathrm{d}\mathbf{x} \, x_i e^{-\mathcal{H}_t(\boldsymbol{\theta}; m)} \left\{ \frac{-\frac{\partial}{\partial \omega_i} \mathcal{H}_t\left(\mathbf{x}, \boldsymbol{\theta}; m\right)}{Z_t\left(\boldsymbol{\theta}; m\right)} - \frac{\frac{\partial}{\partial \omega_i} Z_t\left(\boldsymbol{\theta}; m\right)}{\left[Z_t\left(\boldsymbol{\theta}; m\right)\right]^2} \right\} \right] \\
&= \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\int \mathrm{d}\mathbf{x} \, x_i \frac{e^{-\mathcal{H}_t(\boldsymbol{\theta}; m)}}{Z_t\left(\boldsymbol{\theta}; m\right)} \sqrt{(1-t)\lambda m} \left\{ x_i + \left\langle x_i' \right\rangle_{t, \boldsymbol{\theta}, m} \right\} \right] \\
&= \sqrt{(1-t)\lambda m} \, \mathbb{E}_{\mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}} \left[\left\langle x_i^2 \right\rangle_{t, \boldsymbol{\theta}, m} - \left\langle x_i \right\rangle_{t, \boldsymbol{\theta}, m}^2 \right]
\end{aligned}
$$

(c) Uses the fact that when a_i's are $\mathrm{O}(1)$

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N^2}\sum_{i\le j}a_ia_j &= \lim_{N\to\infty}\frac{1}{2N^2}\left[\sum_{i,j}a_ia_j+\sum_i a_i^2\right]\\ &= \lim_{N\to\infty}\frac{1}{2N^2}\sum_{i,j}a_ia_j+\mathrm{O}\left(N^{-1}\right)\\ &= \lim_{N\to\infty}\frac{1}{2N^2}\sum_{i,j}a_ia_j\end{aligned}$$

(d) Write as different replicas for the second term:

$$\begin{aligned}&\frac{\lambda}{4N^2}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i,j}\langle x_ix_j\rangle^2_{\tau,\boldsymbol{\theta},m}-2Nm\sum_i\langle x_i\rangle^2_{\tau,\boldsymbol{\theta},m}\right]\\ &\quad=\frac{\lambda}{4N^2}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\sum_{i,j}x_i^{(1)}x_j^{(1)}x_i^{(2)}x_j^{(2)}-2Nm\sum_i x_i^{(1)}x_i^{(2)}\right\rangle_{\tau,\boldsymbol{\theta},m}\right]\\ &\quad=\frac{\lambda}{4N^2}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\left(\sum_i x_i^{(1)}x_i^{(2)}\right)^2-2Nm\left(\sum_i x_i^{(1)}x_i^{(2)}\right)+N^2m^2\right\rangle_{\tau,\boldsymbol{\theta},m}-N^2m^2\right]\\ &\quad=\frac{\lambda}{4}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\left(\frac{1}{N}\sum_i x_i^{(1)}x_i^{(2)}-m\right)^2\right\rangle_{\tau,\boldsymbol{\theta},m}\right]-\frac{\lambda m^2}{4}\end{aligned}$$

Analogously, we have

$$\begin{aligned}&\frac{\lambda}{2N^2}\int_0^1\mathrm{d}\tau\,\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i,j}\left\langle x_i^*x_j^*x_ix_j\right\rangle^2_{\tau,\boldsymbol{\theta},m}-2Nm\sum_i\langle x_i^*x_i\rangle^2_{\tau,\boldsymbol{\theta},m}\right]\\ &\quad=\frac{\lambda}{2N^2}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\sum_{i,j}x_i^*x_j^*x_ix_j-2Nm\sum_i x_i^*x_i\right\rangle_{\tau,\boldsymbol{\theta},m}\right]\\ &\quad=\frac{\lambda}{2N^2}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\left(\sum_i x_i^*x_i\right)^2-2Nm\left(\sum_i x_i^*x_i\right)+N^2m^2\right\rangle_{\tau,\boldsymbol{\theta},\boldsymbol{\xi}}-N^2m^2\right]\\ &\quad=\frac{\lambda}{2}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\left(\sum_i x_i^*x_i-m\right)^2\right\rangle_{\tau,\boldsymbol{\theta},m}\right]-\frac{\lambda m^2}{2}\end{aligned}$$

(e) Utilize Eqn 8.8 and the Nishimori identity so that

$$\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\left(\sum_i x_i^*x_i-m\right)^2\right\rangle_{\tau,\boldsymbol{\theta},m}\right]=\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\left\langle\left(\sum_i x_i^{(1)}x_i^{(2)}-m\right)^2\right\rangle_{\tau,\boldsymbol{\theta},m}\right]$$

Hence, we found a lower bound that

$$\Phi_{\mathrm{MF}}(\lambda)\ge\Phi_{\mathrm{denoising}}(\lambda m)-\frac{\lambda m^2}{4},\qquad\forall\, m\quad\Rightarrow\quad\Phi_{\mathrm{MF}}(\lambda)=\max_m\left[\Phi_{\mathrm{denoising}}(\lambda m)-\frac{\lambda m^2}{4}\right]$$

Upper Bound by Fixed-Magnetization Models

To get start, we first redefine the Hamiltonian of problem A by replacing some m into θ:

$$\hat{\mathcal{H}}_A\left(\mathbf{x},\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q\right) \triangleq -\sum_i \log\left(P_X(x_i)\right) - \sum_i \left[-\frac{\lambda q}{2}x_i^2 + \lambda m x^* x + \sqrt{\lambda q}\omega_i x_i\right] \tag{8.16}$$

And as before, we define the interpolated Hamiltonian

$$\begin{aligned}
\tilde{\hat{\mathcal{H}}}_A^{(t)} &\triangleq \hat{\mathcal{H}}_A\left(\mathbf{x},t\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q\right) \\
&= -\sum_i \log\left(P_X(x_i)\right) - \sum_i \left[-\frac{t\lambda q}{2}x_i^2 + t\lambda m x^* x + \sqrt{t\lambda q}\omega_i x_i\right] \\
\hat{\mathcal{H}}_t\left(\mathbf{x},\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m,q\right) &\triangleq \tilde{\hat{\mathcal{H}}}_A^{(1-t)}\left(\mathbf{x},\lambda,\mathbf{x}^*,\boldsymbol{\omega};m\right) + \tilde{\mathcal{H}}_B^{(t)}\left(\mathbf{x},\lambda,\mathbf{x}^*,\boldsymbol{\xi}\right) \\
&= -\sum_i \left[-\frac{(1-t)\lambda q}{2}x_i^2 + (1-t)\lambda m x_i^* x_i + \sqrt{(1-t)\lambda q}\omega_i x_i\right] \\
&\quad -\sum_{i\le j}\left[-\frac{t\lambda}{2N}x_i^2 x_j^2 + \frac{t\lambda}{N}x_i^* x_j^* x_i x_j + \sqrt{\frac{t\lambda}{N}}\xi_{ij}x_i x_j\right]
\end{aligned} \tag{8.17}$$

Consider models at fixed value of magnetization $M = \frac{1}{N}\sum_i x_i^* x_i$, i.e. we use the same family of Hamiltonians as above, but the system only contains configurations with the given magnetization M

$$Z_A^{\text{fixed}}\left(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q,M\right) \triangleq \int \mathrm{d}\mathbf{x}\, e^{-\hat{\mathcal{H}}_A(\mathbf{x},\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q)}\delta\left(M - \frac{1}{N}\sum_i x_i^* x_i\right) \tag{8.18}$$

$$Z_B^{\text{fixed}}\left(\lambda,\mathbf{x}^*,\boldsymbol{\xi};M\right) \triangleq \int \mathrm{d}\mathbf{x}\, e^{-\mathcal{H}_B(\mathbf{x},\lambda,\mathbf{x}^*,\boldsymbol{\xi})}\delta\left(M - \frac{1}{N}\sum_i x_i^* x_i\right) \tag{8.19}$$

$$Z_t^{\text{fixed}}\left(\boldsymbol{\theta};m,q,M\right) \triangleq \int \mathrm{d}\mathbf{x}\, e^{-\hat{\mathcal{H}}_t(\mathbf{x},\boldsymbol{\theta};m,q)}\delta\left(M - \frac{1}{N}\sum_i x_i^* x_i\right) \tag{8.20}$$

It is easy to verify that

$$Z_1^{\text{fixed}}\left(\boldsymbol{\theta};m,q,M\right) \equiv Z_B^{\text{fixed}}\left(\lambda,\mathbf{x}^*,\boldsymbol{\xi};M\right), \qquad \forall\, m,q$$

Again, we are interested in the quantity

$$\Phi_{\text{MF}}^{\text{fixed}}\left(\lambda;M\right) \triangleq \lim_{N\to\infty}\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\frac{\log\left(Z_B^{\text{fixed}}\left(\lambda,\mathbf{x}^*,\boldsymbol{\xi};M\right)\right)}{N}\right]$$

Notice that

$$\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{H}}_t\left(\mathbf{x},\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m,q\right) &= -\sum_i\left[\frac{\lambda q}{2}\sum_i x_i^2 - \lambda m x_i^* x_i - \frac{\sqrt{\lambda q}}{2\sqrt{1-t}}\omega_i x_i\right] \\
&\quad -\sum_{i\le j}\left[-\frac{\lambda}{2N}x_i^2 x_j^2 + \frac{\lambda}{N}x_i^* x_j^* x_i x_j + \frac{\sqrt{\lambda/N}}{2\sqrt{t}}\xi_{ij}x_i x_j\right]
\end{aligned} \tag{8.21}$$

Therefore this derivative can be into several Boltzmann average terms associated to $\hat{\mathcal{H}}_t\left(\mathbf{x}, \lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}; m, q\right)$ with fixed magnetization M. For short we denote $\boldsymbol{\theta} = \{\lambda, \mathbf{x}^*, \boldsymbol{\omega}, \boldsymbol{\xi}\}$

$$
\begin{aligned}
\frac{\partial}{\partial t} \frac{\log Z_t^{\text{fixed}}\left(\boldsymbol{\theta}; m, q, M\right)}{N} &= \frac{1}{N} \frac{1}{Z_t^{\text{fixed}}\left(\boldsymbol{\theta}; m, q, M\right)} \frac{\partial}{\partial t} \int \mathrm{d}\mathbf{x}\, e^{-\hat{\mathcal{H}}_t(\mathbf{x}, \boldsymbol{\theta}; m, q)} \delta\left(M - \frac{1}{N} \sum_i x_i^* x_i\right) \\
&= -\frac{1}{N} \int \mathrm{d}\mathbf{x} \underbrace{\frac{e^{-\hat{\mathcal{H}}_t(\mathbf{x}, \boldsymbol{\theta}; m, q)} \delta\left(M - \frac{1}{N} \sum_i x_i^* x_i\right)}{Z_t\left(\boldsymbol{\theta}; m\right)}}_{= P_{t, \boldsymbol{\theta}, m, q, M}(\mathbf{x})} \frac{\partial}{\partial t} \hat{\mathcal{H}}_t\left(\mathbf{x}, \boldsymbol{\theta}; m, q\right) \\
&= -\frac{1}{N} \left\langle \frac{\partial}{\partial t} \hat{\mathcal{H}}_t\left(\mathbf{x}, \boldsymbol{\theta}; m, q\right) \right\rangle_{t, \boldsymbol{\theta}, m, q, M} \\
&= -\frac{1}{N} \Bigg\{ \frac{\lambda}{N} \sum_{i \leq j} \left[\frac{\left\langle x_i^2 x_j^2 \right\rangle_{t, \boldsymbol{\theta}, m, q, M}}{2} - \left\langle x_i^* x_j^* x_i x_j \right\rangle_{t, \boldsymbol{\theta}, m, q, M} \right] \\
&\quad - \lambda \sum_i \left[\frac{q}{2} \left\langle x_i^2 \right\rangle_{t, \boldsymbol{\theta}, m, q, M} - m \left\langle x_i^* x_i \right\rangle_{t, \boldsymbol{\theta}, m, q, M} \right] \\
&\quad - \frac{\sqrt{\lambda / N}}{2\sqrt{t}} \sum_{i \leq j} \xi_{ij} \left\langle x_i x_j \right\rangle_{t, \boldsymbol{\theta}, m, q, M} + \frac{\sqrt{\lambda q}}{2\sqrt{1-t}} \sum_i \omega_i \left\langle x_i \right\rangle_{t, \boldsymbol{\theta}, m, q, M} \Bigg\}
\end{aligned}
\tag{8.22}
$$

Redo everything using same trick above gives

$$
\begin{aligned}
\Phi_{\mathrm{MF}}^{\mathrm{fixed}}(\lambda;M) &= \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\xi}}\left[\frac{\log\left(Z_B^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\xi};M)\right)}{N}\right] \\
&= \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\frac{\log\left(Z_1^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m,q,M)\right)}{N}\right] \\
&= \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\frac{\log\left(Z_0^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m,q,M)\right)}{N} + \int_0^1 \mathrm{d}\tau\ \frac{\partial}{\partial t}\frac{\log\left(Z_t^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi};m,q,M)\right)}{N}\bigg|_{t=\tau}\right] \\
&= \lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}}\left[\frac{\log\left(Z_A^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q,M)\right)}{N}\right] + \int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\frac{\partial}{\partial t}\frac{\log\left(Z_t^{\mathrm{fixed}}(\boldsymbol{\theta};m,q,M)\right)}{N}\bigg|_{t=\tau}\right] \\
&= \lim_{N\to\infty}\Bigg\{\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}}\left[\frac{\log\left(Z_A^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q,M)\right)}{N}\right] \\
&\qquad -\frac{\lambda}{N^2}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i\le j}\left\{\frac{\left\langle x_i^2x_j^2\right\rangle_{\tau,\boldsymbol{\theta},m,q,M}}{2} - \left\langle x_i^*x_j^*x_ix_j\right\rangle_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right] \\
&\qquad +\frac{\lambda}{N}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_i\left\{\frac{q}{2}\left\langle x_i^2\right\rangle_{\tau,\boldsymbol{\theta},m,q,M} - m\left\langle x_i^*x_i\right\rangle_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right] \\
&\qquad +\frac{\lambda}{2N^2}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i\le j}\left\{\left\langle x_i^2x_j^2\right\rangle_{\tau,\boldsymbol{\theta},m,q,M} - \left\langle x_ix_j\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right] \\
&\qquad -\frac{\lambda q}{2N}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_i\left\{\left\langle x_i^2\right\rangle_{\tau,\boldsymbol{\theta},m,q,M} - \left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right]\Bigg\} \\
&= \lim_{N\to\infty}\Bigg\{\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}}\left[\frac{\log\left(Z_A^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q,M)\right)}{N}\right] \\
&\qquad -\frac{\lambda}{2N^2}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i\le j}\left\{\left\langle x_ix_j\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M} - 2\left\langle x_i^*x_j^*x_ix_j\right\rangle_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right] \\
&\qquad +\frac{\lambda}{2N}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_i\left\{q\left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M} - 2m\left\langle x_i^*x_i\right\rangle_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right]\Bigg\} \\
&= \lim_{N\to\infty}\Bigg\{\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}}\left[\frac{\log\left(Z_A^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q,M)\right)}{N}\right] \\
&\qquad -\frac{\lambda}{4N^2}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i,j}\left\{\left\langle x_ix_j\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M} - 2\left\langle x_i^*x_j^*x_ix_j\right\rangle_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right] \\
&\qquad +\frac{\lambda}{2N}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_i\left\{q\left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M} - 2m\left\langle x_i^*x_i\right\rangle_{\tau,\boldsymbol{\theta},m,q,M}\right\}\right]\Bigg\} \\
&= \lim_{N\to\infty}\Bigg\{\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}}\left[\frac{\log\left(Z_A^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q,M)\right)}{N}\right] \\
&\qquad -\frac{\lambda}{4N^2}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i,j}\left\langle x_ix_j\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M} - 2Nq\sum_i\left\langle x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M}\right] \\
&\qquad +\frac{\lambda}{2N^2}\int_0^1 \mathrm{d}\tau\ \mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega},\boldsymbol{\xi}}\left[\sum_{i,j}\left\langle x_i^*x_j^*x_ix_j\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M} - 2Nm\sum_i\left\langle x_i^*x_i\right\rangle^2_{\tau,\boldsymbol{\theta},m,q,M}\right]\Bigg\} \\
&= \lim_{N\to\infty}\Bigg\{\mathbb{E}_{\mathbf{x}^*,\boldsymbol{\omega}}\left[\frac{\log\left(Z_A^{\mathrm{fixed}}(\lambda,\mathbf{x}^*,\boldsymbol{\omega};m,q,M)\right)}{N}\right]
\end{aligned}
$$

Hence, we found a upper bound that

$$\begin{aligned} & \Phi_{\mathrm{MF}}^{\mathrm{Fixed}}(\lambda; M) \leq \Phi_{\mathrm{denoising}}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(M-m)^2 - \frac{\lambda m^2}{2}, \qquad \forall\, m, q \\ \Rightarrow \quad & \Phi_{\mathrm{MF}}^{\mathrm{fixed}}(\lambda; M) \leq \min_{m,q} \left[\Phi_{\mathrm{denoising}}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(M-m)^2 - \frac{\lambda m^2}{2} \right] \end{aligned}$$

Sandwich the $\Phi_{\mathrm{MF}}(\lambda)$

In the large N limit, the Boltzmann distribution will be dominated by the configurations with specific magnetization, so by Laplace method (we are a bit sloppy here) we have

$$\begin{aligned} \Phi_{\mathrm{MF}}(\lambda) &= \max_M \Phi_{\mathrm{MF}}^{\mathrm{fixed}}(\lambda, M) \\ &\leq \max_M \min_{m,q} \left[\Phi_{\mathrm{denoising}}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(M-m)^2 - \frac{\lambda m^2}{2} \right] \\ &\leq \max_M \left[\Phi_{\mathrm{denoising}}(\lambda m) + \frac{\lambda q^2}{4} + \frac{\lambda}{2}(M-m)^2 - \frac{\lambda m^2}{2} \right] \Big|_{m=M, q=M} \\ &= \max_M \left[\Phi_{\mathrm{denoising}}(\lambda M) - \frac{\lambda M^2}{4} \right] \end{aligned}$$

Finally, combine the bound from both sides and note that $\Phi_{\mathrm{MF}}(\lambda)$ does not depend on m and M

$$\begin{cases} \Phi_{\mathrm{MF}}(\lambda) \geq \max_m \left[\Phi_{\mathrm{denoising}}(\lambda m) - \frac{\lambda m^2}{4} \right] \\ \Phi_{\mathrm{MF}}(\lambda) \leq \max_M \left[\Phi_{\mathrm{denoising}}(\lambda M) - \frac{\lambda M^2}{4} \right] \end{cases} \quad \Rightarrow \quad \Phi_{\mathrm{MF}}(\lambda) = \max_m \left[\Phi_{\mathrm{denoising}}(\lambda m) - \frac{\lambda m^2}{4} \right]$$

Bibliography

Nishimori demonstrated that his symmetry implied replica-symmetry in Nishimori (1980). The modern approach in terms of perturbations is discussed in Korada and Macris (2009); Abbe and Montanari (2013); Coja-Oghlan et al. (2018). The spiked Wigner model has been studied in great detail over the last decades, and has been the topics of many fundamental papers Johnstone (2001); Baik et al. (2005). The replica approach to this problem is reviewed in details in Lesieur et al. (2017). The results was first proved in Barbier et al. (2016). We presented here the alternative later proof of El Alaoui and Krzakala (2018). Thanks to an universality theorem Krzakala et al. (2016), many problems can be reduced to variants of such low-rank factorization problems and generic formula have been proven for these Lelarge and Miolane (2019); Miolane (2017); Barbier and Macris (2019). Extensions can also be made for almost arbitrary priors, including neural networks generating models Aubin et al. (2020).

8.5 Exercises

Exercise 8.1: The BBP transition

Consider the rank-one factorization model with vectors x where each component is sampled uniformly from $\mathcal{N}(0, 1)$.

a) Using the replica expression for the free entropy, show that the overlap m between the posterior estimate $\langle X \rangle$ and the real value x^* obeys a self consistent equation.

b) Solve this equation numerically, and show that m is non zero only for SNR $\lambda > 1$.

c) Once this is done, perform simulations of the model by creating matrices

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}} \underbrace{\mathbf{x}^*\mathbf{x}^{*\intercal}}_{N\times N \text{ rank-one matrix}} + \underbrace{\boldsymbol{\xi}}_{\text{symmetric iid noise}}$$

and compare the MMSE obtained with this approach with the one of any algorithm you may invent so solve the problem. A classical algorithm for instance, is to use as an estimator the eigenvector of Y corresponding to its largest eigenvalue.

Exercise 8.2: More phase transitions

Consider the rank-one factorization model with vectors X where each components is

model 1: sampled uniformly from ± 1

model 2: sampled uniformly from ± 1 (with probability ρ), otherwise 0 (with probability $1 - \rho$)

a) Using the replica expression for the free entropy, show that the overlap m between the posterior means estimate $\langle X \rangle$ and the real value X^* obeys a self consistent equation.

b) Solve this equation numerically, and show that m is non zero only for SNR $\lambda > 1$, for models 1 and for a non-trivial critical value for model 2. Check also that, for ρ small enough, the transition is a first order one for model 2.

Chapter 9

Cavity method and Approximate Message Passing

9.1 Self-consistent equation

N variables

$$-\mathcal{H}_N = \sum_{1\leq i\leq j} -\frac{x_i^2 x_j^2 \lambda}{2N} + \frac{x_i x_j x_i^* x_j^* \lambda}{N} + x_i x_j \xi_{ij}\sqrt{\frac{\lambda}{N}} \tag{9.1}$$

$N+1$ variables

$$-\mathcal{H}_{N+1} = \sum_{0\leq i\leq j} -\frac{x_i^2 x_j^2 \lambda}{2(N+1)} + \frac{x_i x_j x_i^* x_j^* \lambda}{N+1} + x_i x_j \xi_{ij}\sqrt{\frac{\lambda}{N+1}} \tag{9.2}$$

$$= \sum_{1\leq i\leq j}\left(-\frac{x_i^2 x_j^2 \lambda}{2(N+1)} + \frac{x_i x_j x_i^* x_j^* \lambda}{N+1} + x_i x_j \xi_{ij}\sqrt{\frac{\lambda}{N+1}}\right) - \frac{x_0^2\lambda}{2}\sum_i \frac{x_i^2}{N+1}$$
$$+ x_0 x_0^* \lambda \sum_i \frac{x_i x_i^*}{N+1} + x_0 \sum_i x_i \xi_{0i}\sqrt{\frac{\lambda}{N+1}} \tag{9.3}$$

$$= -\mathcal{H}_N(\lambda\frac{N}{N+1}) - x_0^2\frac{\lambda}{2}\sum_i \frac{x_i^2}{N} + x_0 x_0^* \sum_i \frac{x_i x_i^*}{N} + x_0 \sum_i x_i \xi_{0i}\sqrt{\frac{\lambda}{N}} + o(1) \tag{9.4}$$

Let us now look at the average magnetization of the new spin. It must satisfies:

$$m = \mathbb{E}_{\xi,x^*,x_0^*,\xi_0} \frac{\int dx_0 P_X(x_0)\, x_0^* x_0 \,\langle e^{-x_0^2\frac{\lambda}{2}\sum_i \frac{x_i^2}{N} + x_0 x_0^* \sum_i \frac{x_0 x_0^*}{N} + x_0 \sum_i x_i \xi_{0i}\sqrt{\frac{\lambda}{N}}}\rangle_N}{\int dx_0 P_X(x_0)\langle e^{-x_0^2\frac{\lambda}{2}\sum_i \frac{x_i^2}{N} + x_0 x_0^* \sum_i \frac{x_i x_i^*}{N} + x_0 \sum_i x_i \xi_{0i}\sqrt{\frac{\lambda}{N}}}\rangle_N} \tag{9.5}$$

Let us therefore evaluate this term in brakets. Using concentration of measure for the overlaps, we find

$$\langle e^{-x_0^2\frac{\lambda}{2}\sum_i \frac{x_i^2}{N} + x_0 x_0^* \sum_i \frac{x_i x_i^*}{N} + x_0 \sum_i x_i \xi_{0i}\sqrt{\frac{\lambda}{N}}}\rangle_N \approx \langle e^{-x_0^2\frac{\lambda}{2}\rho + x_0 x_0^* m + x_0 \sum_i x_i \xi_{0i}\sqrt{\frac{\lambda}{N}}}\rangle_N \tag{9.6}$$

$$= e^{-x_0^2\frac{\lambda}{2}\rho + x_0 x_0^* m}\langle e^{x_0 \sum_i x_i \xi_{0i}\sqrt{\frac{\lambda}{N}}}\rangle_N \tag{9.7}$$

How to deal with the last term? We could expand in power of m! Indeed, concentration suggest that the x_i are x_j are only weekly correlated so that we could hope that:

$$\begin{aligned}
\langle e^{x_0 \sum_i x_i \xi_{0i} \sqrt{\frac{\lambda}{N}}} \rangle_N &= \langle \prod_i e^{x_0 x_i \xi_{0i} \sqrt{\frac{\lambda}{N}}} \rangle_N \approx \langle \prod_i (1 + x_0 x_i \xi_{0i} \sqrt{\frac{\lambda}{N}} + \frac{(x_0 x_i)^2}{2} \xi_{0i}^2 \frac{\lambda}{N}) \rangle_N \\
&\approx \prod_i (1 + x_0 \langle x_i \rangle \xi_{0i} \sqrt{\frac{\lambda}{N}} + \frac{1}{2} x_0^2 \langle x_i^2 \rangle \xi_{0i}^2 \frac{\lambda}{N}) \approx \prod_i (1 + x_0 \langle x_i \rangle \xi_{0i} \sqrt{\frac{\lambda}{N}} + \frac{1}{2} x_0^2 \langle x_i^2 \rangle \frac{\lambda}{N}) \\
&\approx e^{x_0 \sum_i \langle x_i \rangle \xi_{0i} \sqrt{\frac{\lambda}{N}} - \frac{x_0^2}{2} \sum_i \langle x_i \rangle^2 \xi_{0i}^2 \frac{\lambda}{N} + \frac{x_0^2}{2} \sum_i \langle x_i^2 \rangle \xi_{0i}^2 \frac{\lambda}{N}} \approx e^{x_0 \sum_i \langle x_i \rangle \xi_{0i} \sqrt{\frac{\lambda}{N}} - \frac{x_0^2}{2} q \lambda + \frac{x_0^2}{2} \rho \lambda}
\end{aligned} \tag{9.8}$$

This can be actually proved rigorously. Indeed consider the two following expressions:

$$\mathcal{A} = -\frac{x_0^2}{2} \lambda \rho + x_0 \sum_i \xi_{i0} x_i \sqrt{\frac{\lambda}{N}} \tag{9.9}$$

$$\mathcal{B} = -\frac{x_0^2}{2} \lambda q + x_0 \sum_i \xi_{i0} \langle x_i \rangle \sqrt{\frac{\lambda}{N}} \tag{9.10}$$

It is easy to check via Gaussian integration, and using concentration, that

$$\mathbb{E} \langle (e^{\mathcal{A}} - e^{\mathcal{B}})^2 \rangle \to 0 \tag{9.11}$$

We thus obtain

$$\langle e^{-x_0^2 \frac{\lambda}{2} \sum_i \frac{x_i^2}{N} + x_0 x_0^* \sum_i \frac{x_0 x_0^*}{N} + x_0 \sum_i x_i \xi_{0i} \sqrt{\frac{\lambda}{N}}} \rangle_N \approx e^{-x_0^2 \frac{\lambda}{2} q + x_0 x_0^* m + x_0 \sum_i \langle x_i \rangle \xi_{0i} \sqrt{\frac{\lambda}{N}}} \tag{9.12}$$

Recognizing that the last term is actually a random Gaussian variable thanks to the CLT, we have

$$m = \mathbb{E}_{x_0^*, z} \frac{\int dx_0 P_X(x_0) \, x_0^* x_0 \, e^{-x_0^2 \frac{q\lambda}{2} + x_0 x_0^* m + x_0 \sqrt{\lambda q} z}}{\int dx_0 P_X(x_0) e^{-x_0^2 \frac{q\lambda}{2} + x_0 x_0^* m + x_0 \sqrt{\lambda q} z}} \tag{9.13}$$

and we have recovered our self-consistent equation from a rigorous computation. This is the power of the cavity method!

9.2 Rank-one by the cavity method

We can derive the free energy as well. To do this, we need to be keep track of all order 1 constant, so we write:

$$-\mathcal{H}_{N+1} = \sum_{0 \le i \le j} -\frac{x_i^2 x_j^2 \lambda}{2(N+1)} + \frac{x_i x_j x_i^* x_j^* \lambda}{N+1} + x_i x_j \xi_{ij} \sqrt{\frac{\lambda}{N+1}} \tag{9.14}$$

$$\begin{aligned}
&= \sum_{1 \le i \le j} \left(-\frac{x_i^2 x_j^2 \lambda}{2(N+1)} + \frac{x_i x_j x_i^* x_j^* \lambda}{N+1} + x_i x_j \xi_{ij} \sqrt{\frac{\lambda}{N+1}} \right) - \frac{x_0^2 \lambda}{2} \sum_i \frac{x_i^2}{N+1} \\
&\quad + x_0 x_0^* \lambda \sum_i \frac{x_i x_i^*}{N+1} + x_0 \sum_i x_i \xi_{0i} \sqrt{\frac{\lambda}{N+1}}
\end{aligned} \tag{9.15}$$

In the parenthesis, we recognise, expanding in N and keeping only $O(1)$ variables

$$(.) = -\mathcal{H}_N + \frac{\lambda}{2}\sum_{i\leq j}\frac{x_i^2x_j^2}{N^2} - \lambda\sum_{i\leq j}\frac{x_ix_i^*x_jx_j^*}{N^2} - \frac{1}{2N}\sqrt{\frac{\lambda}{N}}\sum_{i\leq j}x_ix_i\xi_{ij} + o(1) \tag{9.16}$$

$$= -\mathcal{H}_N + \frac{\lambda}{4}\sum_{i,j}\frac{x_i^2x_j^2}{N^2} - \frac{\lambda}{2}\sum_{i,j}\frac{x_ix_i^*x_jx_j^*}{N^2} - \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j}x_ix_i\xi_{ij} + o(1) \tag{9.17}$$

$$= -\mathcal{H}_N + \frac{\lambda}{4}\left(\sum_i\frac{x_i^2}{N}\right)^2 - \frac{\lambda}{2}\left(\sum_i\frac{x_ix_i^*}{N}\right)^2 - \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j}x_ix_i\xi_{ij} + o(1) \tag{9.18}$$

We thus write

$$\begin{aligned}-\mathcal{H}_{N+1} =& -\mathcal{H}_N + \frac{\lambda}{4}\left(\sum_i\frac{x_i^2}{N}\right)^2 - \frac{\lambda}{2}\left(\sum_i\frac{x_ix_i^*}{N}\right)^2 - \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j}x_ix_i\xi_{ij} + o(1)\\ &- x_0^2\frac{\lambda}{2}\sum_i\frac{x_i^2}{N} + x_0x_0^*\sum_i\frac{x_ix_i^*}{N} + x_0\sum_i x_i\xi_{0i}\sqrt{\frac{\lambda}{N}}\end{aligned}$$

We can now compute the free energy using the Cavity method and write

$$\begin{aligned}F \approx \mathbb{E}\log\frac{Z_{N+1}}{Z_N} =& \mathbb{E}_{\xi,x^*}\log\langle e^{\frac{\lambda}{4}\left(\sum_i\frac{x_i^2}{N}\right)^2 - \frac{\lambda}{2}\left(\sum_i\frac{x_ix_i^*}{N}\right)^2 + \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j}x_ix_i\xi_{ij}}\rangle_N\\ &+ \mathbb{E}_{\xi,x^*,\xi_0}\log\int dx_0P(x_0)\langle e^{-x_0^2\frac{\lambda}{2}\sum_i\frac{x_i^2}{N} + x_0x_0^*\sum_i\frac{x_ix_i^*}{N} + x_0\sum_i x_i\xi_{0i}\sqrt{\frac{\lambda}{N}}}\rangle_N\end{aligned} \tag{9.19}$$

All these terms are somehow simple, and more importantly, look like they depends on our order parameters, excepts for the Gaussian random variables ξ_{ij} and ξ_{0i}. In fact, we already took care of the second line, so we can further simply the expression as

$$\begin{aligned}F \approx \mathbb{E}\log\frac{Z_{N+1}}{Z_N} =& \mathbb{E}_{\xi,x^*}\log\langle e^{\frac{\lambda}{4}\left(\sum_i\frac{x_i^2}{N}\right)^2 - \frac{\lambda}{2}\left(\sum_i\frac{x_ix_i^*}{N}\right)^2 + \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j}x_ix_i\xi_{ij}}\rangle_N\\ &+ \mathbb{E}_{\xi,x^*,\xi_0}\log\int dx_0P_X(x_0)e^{-x_0^2\frac{\lambda}{2}q + x_0x_0^*m + x_0\sum_i\langle x_i\rangle\xi_{0i}\sqrt{\frac{\lambda}{N}}}\end{aligned} \tag{9.20}$$

Let us this deal with the first line. First notice that, because of concentration, we have:

$$\sum_i\frac{x_ix_i^*}{N} \to m\,, \sum_i\frac{x_i^2}{N} \to \rho\,. \tag{9.21}$$

We thus can write, via this concentration property:

$$\langle e^{\frac{\lambda}{4}\left(\sum_i\frac{x_i^2}{N}\right)^2 - \frac{\lambda}{2}\left(\sum_i\frac{x_ix_i^*}{N}\right)^2 + \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j}x_ix_j\xi_{ij}}\rangle = o(1) + \langle e^{\frac{\rho^2\lambda}{4} - \frac{m\lambda}{2} + \frac{1}{4N}\sqrt{\frac{\lambda}{N}}x_ix_i\xi_{ij}}\rangle \tag{9.22}$$

$$= o(1) + e^{\frac{\rho^2\lambda}{4} - \frac{m\lambda}{2}}\langle\prod_{ij}e^{\frac{1}{4N}\sqrt{\frac{\lambda}{N}}x_ix_j\xi_{ij}}\rangle \tag{9.23}$$

This last terms look complicated. However, it also follows a concentration property. First, let us notice that, using use Stein lemma, we have

$$\mathbb{E}\langle \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j} x_i x_j \xi_{ij}\rangle = \mathbb{E}\langle \sum \left(\frac{\lambda}{4N^2}\langle x_i^2 x^2\rangle - \langle x_i x_j\rangle^2\right)\rangle \tag{9.24}$$

We can also compute the variance of this quantities, and check that it concentrates. This is, actually, nothinig but the Matrix-MMSE. We can thus write

$$\langle e^{\frac{\lambda}{4}\left(\sum_i \frac{x_i^2}{N}\right)^2 - \frac{\lambda}{2}\left(\sum_i \frac{x_i x_i^*}{N}\right)^2 + \frac{1}{4N}\sqrt{\frac{\lambda}{N}}\sum_{i,j} x_i x_j \xi_{ij}}\rangle \tag{9.25}$$

$$\approx e^{\frac{\rho^2\lambda}{4} - \frac{m\lambda}{2}} e^{\frac{\lambda}{4N^2}\sum_{ij}\langle x_i^2 x^2\rangle - \langle x_i x_j\rangle^2} \tag{9.26}$$

$$\approx e^{\frac{\lambda q}{4} - \frac{\lambda m}{2}} \tag{9.27}$$

We thus find (using also Nishimori so that $m = q$):

$$F \approx -\frac{\lambda m}{4} + \mathbb{E}_{x^*,z} \log \int dx_0 P_X(x_0) e^{-x_0^2 \frac{\lambda m}{2} + x_0 x_0^* m + x_0 x_0^* \sqrt{\lambda m} z} \tag{9.28}$$

9.3 AMP

We come back to the cavity equation. We have the new spins x_0 that "sees" an effective model as

$$x_0 \sim \frac{1}{Z} P_X(x_0) e^{-x_0^2 \frac{\lambda}{2}\sum_i \frac{\langle x_i\rangle^2}{N} + x_0 x_0^* \frac{\lambda}{N}\sum_i \frac{\langle x_i\rangle x_i^*}{N} + x_0 \sum_i \langle x_i\rangle \xi_{0i}\sqrt{\frac{\lambda}{N}}} \tag{9.29}$$

Can we turn this into an algorithm? If we remember that $y_{ij} = x_i^* x_j^* \sqrt{\lambda/n} + \xi_{ij}$, this can be written as

$$x_0 \sim \frac{1}{Z} P_X(x_0) e^{-x_0^2 \frac{\lambda}{2}\sum_i \frac{\langle x_i\rangle^2}{N} + \sqrt{\frac{\lambda}{N}} x_0 \sum_i y_{0i}\langle x_i\rangle} \tag{9.30}$$

Defining the denoising function $\eta(A, B)$ as

$$\eta(A, B) =: \frac{\int dx_0 P_X(x_0) x_0 e^{-x_0^2 A/2 + x_0 B}}{\int dx P_X(x_0) e^{-x_0^2 A/2 + x_0 B}} \tag{9.31}$$

We have the mean and second moments of x_0 expressed as

$$\langle x_0\rangle = \eta\left(\lambda\sum_i \langle x_i\rangle_c^2/N, \sqrt{\frac{\lambda}{N}}\sum_i \langle x_i\rangle_c y_{0i}\right) \tag{9.32}$$

$$\langle x_0^2\rangle - \langle x_0\rangle^2 = \eta'\left(\sum_i \langle x_i\rangle_c^2/N, \sqrt{\frac{\lambda}{N}}\sum_i \langle x_i\rangle_c y_{0i}\right) \tag{9.33}$$

where η' refer to the derivative with respect to B.

It is VERY tempting to turn this into an iterative algorithm, and write, denoting $\hat{x}$ the estimator of the marginal of x at time t

$$\mathbf{h}^t = \sqrt{\frac{\lambda}{N}}\hat{\mathbf{x}}^t\mathbf{Y} \tag{9.34}$$

$$\hat{\mathbf{x}}^{t+1} = \eta\left(\lambda\frac{\hat{\mathbf{x}}^t\cdot\hat{\mathbf{x}}^{t+1}}{N}, \mathbf{h}^t\right) \tag{9.35}$$

However, there is a problem! In these, we have indeed the mean of X_0, but it is not expressed as a function of the mean of x_i, but the mean of x_i in a system WITHOUT X_0 (the cavity system, where x_0 has been removed). What we need would be to express the mean of X_0 as a function of the mean of xi instead, not its cavity mean!

This problem was solved by Thouless, Anderson and Palmer, using Onsager's retraction term. To first order, we can express the cavity mean of x_i (in absence of x_0) as a function of the actual mean x_i (in presence of x_0) as

$$\langle x_i\rangle_c \approx \eta\left(\lambda\sum_{j\neq 0}\langle x_j\rangle_c^2/N, \sqrt{\frac{\lambda}{N}}\sum_{j\neq 0}\langle x_j\rangle_c y_{ij}\right) \approx \eta\left(\lambda\sum_{j\neq 0}\langle x_j\rangle^2/N, \sqrt{\frac{\lambda}{N}}\sum_{j\neq 0}\langle x_j\rangle y_{ij}\right) \tag{9.36}$$

$$\approx \eta\left(\lambda\sum_{j}\langle x_j\rangle^2/N - \lambda\langle x_0\rangle^2/N, \sqrt{\frac{\lambda}{N}}\sum_{j}\langle x_j\rangle y_{ij} - \sqrt{\frac{\lambda}{N}}\langle x_0\rangle y_{i0}\right) \tag{9.37}$$

$$\approx \eta\left(\lambda\sum_{j}\langle x_j\rangle^2/N, \sqrt{\frac{\lambda}{N}}\sum_{j}\langle x_j\rangle y_{ij}\right) - (\partial_B\eta)\sqrt{\lambda/N}\langle x_0\rangle y_{i0} + O(1/N) \tag{9.38}$$

$$\approx \langle x_i\rangle - \eta'\sqrt{\lambda/N}\langle x_0\rangle y_{i0} + O(1/N) \tag{9.39}$$

So the algorithms need to be slightly modified! The local field acting on on the spin 0 is now

$$\sqrt{\frac{\lambda}{N}}\sum_i y_{0i}\langle x_i\rangle_c \to \sqrt{\frac{\lambda}{N}}\sum_i y_{0i}\left(\langle x_i\rangle - \eta_i'\sqrt{\frac{\lambda}{N}}\langle x_0\rangle y_{i0}\right)$$

$$= \sqrt{\frac{\lambda}{N}}\sum_i y_{0i}\langle x_i\rangle - \lambda\left(\frac{1}{N}\sum_i \eta_i' y_{01}^2\right)\langle x_0\rangle \tag{9.40}$$

$$\approx \sqrt{\frac{\lambda}{N}}\sum_i y_{0i}\langle x_i\rangle - \lambda\left(\frac{1}{N}\sum_i \eta_i'\right)\langle x_0\rangle \tag{9.41}$$

so that our iterative algorithm can be written

$$\mathbf{h}^t = \sqrt{\frac{\lambda}{N}}\hat{\mathbf{x}}^t\mathbf{Y} - \lambda\left(\frac{1}{N}\sum_i \eta_i'\right)\hat{\mathbf{x}}^{t-1} \tag{9.42}$$

$$\hat{\mathbf{x}}^{t+1} = \eta\left(\lambda\frac{\hat{\mathbf{x}}^t\cdot\hat{\mathbf{x}}^t}{N}, \mathbf{h}^t\right) \tag{9.43}$$

This algorithm is often called Approximate Message Passing (AMP in short). The correction in $\mathbf{z}$ is called the *Onsager* retro-action term.

The really power-full things about this algorithm is that z is really the cavity field, at each, time, and we know the distribution of cavity fields! In fact, from eqs.(9.30) and (9.13), we expect that $\mathbf{h}$ is distributed as Gaussian so that

$$\mathbf{h}^t = \lambda m^t \mathbf{x}^* + \sqrt{\lambda m^t}\mathbf{z} \tag{9.44}$$

with

$$m^{t+1} = \mathbb{E}_{x_0^*,z} \frac{\int dx_0 P_X(x_0)\, x_0^* x_0\, e^{-x_0^2 \frac{m^t \lambda}{2} + \lambda x_0 x_0^* m^t - x_0\sqrt{\lambda m^t} z}}{\int dx_0 P_X(x_0) e^{-x_0^2 \frac{m^t \lambda}{2} + \lambda x_0 x_0^* m^t - x_0\sqrt{\lambda m^t} z}} \tag{9.45}$$

In other words, we can track the performance of the algorithm step by step. This last equation is often called the "state evolution" of the algorithm.

This can actually be made rigorous (Bolthausen, Bayati-Montanari, Fletcher-Rangan):

Theorem 16 (AMP and State Evolution, informal)**.** *Given the AMP algorithm* (*eqs.9.42*)*, the algorithm behaves as if* $\mathbf{h}^t = \mathbf{x}^* m^t + \sqrt{\lambda q^t}\mathbf{z}$*, with* m^t *given by eq.9.13 with high probability as* $n \to \infty$

9.4 Exercises

EXERCISE 9.1: AMP

Implement AMP for the problems discussed in the Exercises in chap 7, and compare its performance with the optimal ones

Bibliography

The cavity method as described in this section was developed by Parisi, Mézard, & Virasoro Mézard et al. (1987a,b). It is often used in mathematical physics as well, as initiated in Aizenman et al. (2003), and has been applied to the problem discussed in this chapter in Lelarge and Miolane (2019). The vision of this cavity method as an algorithm is initially due to Thouless-Anderson & Palmer Thouless et al. (1977), an article that extraordinary influential in the statistical physics community. Its study as an iterative algorithm with a rigorous state evolution is due to Bolthausen (2014); Bayati and Montanari (2011). For the present low-rank problem, it was initially studied by Rangan and Fletcher (2012) and later discussed in great details in Lesieur et al. (2017), where it was derived starting from Belief-Propagation.

Chapter 10

Stochastic Block Model & Community Detection

10.1 Definition of the model

In this lecture we will discuss clustering of sparse networks also known as community detection problem. To have a concrete example in mind you can picture part of the Facebook network corresponding to students in a high-school where edges are between user who are friends. Knowing the graph of connections the aim is to recover from such a network a division of students corresponding to the classes in the high-school. The signal comes from the notion that students in the same class will more likely be friends and hence connected than students from two different classes. A widely studied simple model for such a situation is called the the *Stochastic Block Model* that we will now introduce a study. Each of N nodes $i = 1, \dots, N$ belong of one among q classes/groups. The variable denoting to which class the node i belongs will be denoted as $s_i^* \in \{1, 2, \dots, q\}$. A node is in group a with probability (fraction of expected group size) $n_a \geq 0$, where $\sum_{a=1}^{q} n_a = 1$.

The edges in the graph are generated as follows: For each pair of node i, j we decide independently whether the edge (ij) is present or not with probability

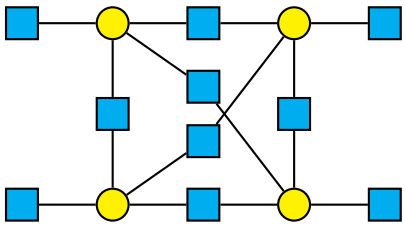
$$\begin{cases} P\left((ij) \in E, A_{ij} = 1 \,\middle|\, s_i^*, s_j^*\right) = p_{s_i^* s_j^*} \\ P\left((ij) \notin E, A_{ij} = 0 \,\middle|\, s_i^*, s_j^*\right) = 1 - p_{s_i^* s_j^*} \end{cases} \quad \Rightarrow \quad G(V, E) \quad \& \quad A_{ij}$$

Here p_{ab} is a symmetric $q \times q$ matrix.

The goal of community detection is given the adjacency matrix A_{ij} to find $\hat{s}_i$ so that $\hat{s}_i$ is as close as s_i^* as possible according to some natural measure of distance, such as the number of misclassification of node into wrong classes. In what follows we will discuss the case where the parameters of the model $\overbrace{n_a, p_{ab}, q}^{\theta}$ are known to the community detection algorithms, but also when they are not and need to be learned. Following previous lecture on inference a natural approach is to consider Bayesian inference where for known values of parameters θ, all information about s_i^* we can extract from the knowledge of the graph is included in the

posterior distribution:

$$P\left(\{s_i\}_{i=1}^N \middle| G,\theta\right) = \frac{1}{Z_G} P\left(G \middle| \{s_i\}_{i=1}^N,\theta\right) P\left(\{s_i\}_{i=1}^N \middle| \theta\right)$$
$$= \frac{1}{Z_G} \prod_{i<j} \left[\left(1-p_{s_i,s_j}\right)^{1-A_{ij}} p_{s_i,s_j}^{A_{ij}}\right] \prod_{i=1}^N n_{s_i}$$



Leading to the fully connected graphical model with one factor node per every variable and per every (non-ordered) pair of variables.

We will study this posterior and investigate what is the algorithm that gives the highest performance. Whether this algorithm achieves information-theoretically optimal performance and whether there are interesting phase transition in this problem. In order to obtain self-averaging results, we study the thermodynamic limit where $N \to \infty$ and $p_{ab} = c_{ab}/N$ with $c_{ab}, n_a, q = O(1)$. The intuition behind this limit is that in this way every node has on average $O(1)$ neighbors while the side of the graph goes large, just as in real world where one has one average the same number of friends independently of the size of the world. This limit also ends up challenging and presents intriguing behaviour.

The average degree of a node in group a, i.e. $s_i^* = a$ is

$$c_a = \sum_{b=1}^q p_{ab} N n_b = \sum_{b=1}^q \frac{c_{ab}}{\cancel{N}} \cancel{N} n_b \sum_{b=1}^q c_{ab} n_b \tag{10.1}$$

Thus we can see that the average degree of a node in group a does not depends on N, i.e. even if the number of nodes is increasing we have always the same average degree.
The overall average degree is then

$$c = \sum_{a,b} c_{ab} n_a n_b \,. \tag{10.2}$$

10.2 Bayesian Inference and Parameter Learning

We start by defining a natural measure of performance the *agreement* between the original assignment $\{s_i^*\}$ and its estimate $\{t_i\}$ as

$$A(\{s_i^*\},\{t_i\}) = \max_\pi \frac{1}{N} \sum_i \delta_{s_i^*,\pi(t_i)} \,, \tag{10.3}$$

where π ranges over the permutations on q elements. We also define a normalized agreement that we call the *overlap*,

$$Q(\{s_i^*\},\{t_i\}) = \max_{\pi} \frac{\frac{1}{N}\sum_i \delta_{s_i^*,\pi(t_i)} - \max_a n_a}{1-\max_a n_a}\,. \tag{10.4}$$

The overlap is defined so that if $s_i^* = t_i$ for all i, i.e., if we find the original labeling without error, then $Q = 1$. If on the other hand the only information we have are the group sizes n_a, and we assign each node to the largest group to maximize the probability of the correct assignment of each node, then $Q = 0$. We will say that a labeling $\{t_i\}$ is correlated with the original one $\{s_i^*\}$ if in the thermodynamic limit $N \to \infty$ the overlap is strictly positive, with $Q > 0$ bounded above some constant.

10.2.1 Community detection with known parameters

The probability that the stochastic block model generates a graph G, with adjacency matrix A, along with a given group assignment $\{s_i\}$, conditioned on the parameters $\theta = \{q, \{n_a\}, \{c_{ab}\}\}$ is

$$P(G,\{s_i\}\,|\,\theta) = \prod_{i<j}\left[\left(\frac{c_{s_i,s_j}}{N}\right)^{A_{ij}}\left(1-\frac{c_{s_i,s_j}}{N}\right)^{1-A_{ij}}\right]\prod_i n_{s_i}\,. \tag{10.5}$$

Note that the above probability is normalized, i.e. $\sum_{G,\{s_i\}} P(G,\{s_i\}\,|\,\theta) = 1$. Assume now that we know the graph G and the parameters θ, and we are interested in the probability distribution over the group assignments. Using Bayes' rule we have

$$P(\{s_i\}\,|\,G,\theta) = \frac{P(G,\{s_i\}\,|\,\theta)}{\sum_{t_i} P(G,\{t_i\}\,|\,\theta)} = \frac{1}{Z_G}\prod_{i<j}\left[\left(c_{s_i,s_j}\right)^{A_{ij}}\left(1-\frac{c_{s_i,s_j}}{N}\right)^{1-A_{ij}}\right]\prod_i n_{s_i}\,. \tag{10.6}$$

Theorem 12 from previous lectures tells us that in order to maximize the overlap $Q(\{s_i^*\},\{\hat{s}_i\})$ between the ground truth assignment and an estimator $\hat{s}_i$ we need to compute the

$$\hat{s}_i = \mathrm{argmax}_{s_i}\mu_i(s_i)\,, \tag{10.7}$$

where $\mu_i(t_i)$ is the marginal probability of the posterior probability distribution. We remind that $\mu_i(s_i) = \sum_{\{s_j\}_{j\neq i}} P(\{s_i\}\,|\,G,\theta)$.

Note the key difference between this optimal decision estimator and the maximum likelihood estimator that is evaluating the configuration at which the posterior distribution has the largest value. In high-dimensional, $N \to \infty$, noisy setting as we consider in the SBM the maximum likelihood estimator is sub-optimal with respect to the overlap with the ground truth configuration. Note also that the posterior distribution is symmetric with respect to permutations of the group labels. Thus the marginals over the entire distribution are uniform. However, we will see that when communities are detectable this permutation symmetry is broken, and we obtain that marginalization is the optimal estimator for the overlap defined in equation 10.4, where we maximize over all permutations.

When the graph G is generated from the SBM using indeed parameters of value θ then Nishimori identities derived in Section 7.2.4 hold and their consequence in the SBM is that in

the thermodynamic limit we can evaluate the overlap $Q(\{s_i\},\{s_i^*\})$ even without the explicit knowledge of the original assignment $\{s_i^*\}$. Due to the Nishimori identity it holds

$$\lim_{N\to\infty} \frac{\frac{1}{N}\sum_i \mu_i(\hat{s}_i) - \max_a n_a}{1-\max_a n_a} = \lim_{N\to\infty} Q(\{\hat{s}_i\},\{s_i^*\})\,. \tag{10.8}$$

The marginals $\mu_i(q_i)$ can also be used to distinguish nodes that have a very strong group preference from those that are uncertain about their membership.

Another consequence of the Nishimori identities is that two configurations taken at random from the posterior distribution have the same agreement with each other as one such configuration with the original assignment s_i^*, i.e.

$$\lim_{N\to\infty} \frac{1}{N} \max_{\pi} \sum_i \mu_i\left(\pi(s_i^*)\right) = \lim_{N\to\infty} \frac{1}{N} \sum_i \sum_a \mu_i(a)^2\,, \tag{10.9}$$

where π again ranges over the permutations on q elements. Overall we are seeing that in the spirit of the Nishimori identities the ground truth configuration s^* has exactly the same properties as any other configuration drawn uniformly from the posterior measure. This property lets us use the ground truth configuration to probe the equilibrium properties of the posterior, a fact that we will use heavily when coming back to the graph coloring problem.

10.2.2 Learning the parameters

Now assume that the only knowledge we have about the system is the graph G, and not the parameters θ. The general goal in Bayesian inference is to learn the most probable values of the parameters θ of an underlying model based on the data known to us. In this case, the parameters are $\theta = \{q, \{n_a\}, \{c_{ab}\}\}$ and the data is the graph G, or rather the adjacency matrix A_{ij}. According to Bayes' rule, the probability $P(\theta\,|\,G)$ that the parameters take a certain value, conditioned on G, is proportional to the probability $P(G\,|\,\theta)$ that the model with parameters θ would generate G. This in turn is the sum of $P(G,\{s_i\}\,|\,\theta)$ over all group assignments $\{s_i\}$:

$$P(\theta\,|\,G) = \frac{P(\theta)}{P(G)}\,P(G\,|\,\theta) = \frac{P(\theta)}{P(G)} \sum_{\{s_i\}} P(G,\{s_i\}\,|\,\theta)\,. \tag{10.10}$$

The *prior distribution* $P(\theta)$ includes any graph-independent information we might have about the values of the parameters. In our setting, we wish to remain perfectly agnostic about these parameters; for instance, we do not want to bias our inference process towards assortative structures. Thus we assume a uniform prior, i.e., $P(\theta) = 1$ up to normalization. Note, however, that since the sum in (10.10) typically grows exponentially with N, we could take any smooth prior $P(\theta)$ as long as it is independent of N; for large N, the data would cause the prior to "wash out," leaving us with the same distribution we would have if the prior were uniform. Thus maximizing $P(\theta \mid G)$ over θ is equivalent to maximizing the partition function over θ, or equivalently the free energy entropy over θ.

10.3 Belief propagation for SBM

We now write Belief Propagation as we derived it in previous lectures for the probability distribution (10.6). We note that this distribution corresponds to a fully-connected factor

graph, while we argued Belief Propagation is designed for trees and works well on tree-like graphs. At the same time the main reason we needed tree-like graphs was the assumption of independence of incoming messages, if the incoming messages are changing only very weekly the outcoming one then the needed independence can be weaker. Indeed results for other models that we have studied so far, such as the Curie-Weiss model or the low-rank matrix estimation can be derived from belief propagation. In the case of the Curie-Weiss model the interactions were very weak, every neighbor was influencing the magnetization by interaction of strength inversely proportional to N. We note that the probability distribution (10.6) is a mixture of the terms corresponding to edges that are $O(1)$ and organized on a tree-like graph, and non-edged that are dense but each of them contributing by a interaction of strength inversely proportional to N. It is thus an interesting case where a sparse and dense graphical model combine.

The canonical Belief Propagation equations for the graphical model (10.6) read

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}} n_{t_i} \prod_{k\neq i,j} \left[\sum_{t_k} c_{t_i t_k}^{A_{ik}} \left(1 - \frac{c_{t_i t_k}}{N}\right)^{1-A_{ik}} \chi_{t_k}^{k\to i} \right], \tag{10.11}$$

where $Z^{i\to j}$ is a normalization constant ensuring $\sum_{t_i} \chi_{t_i}^{i\to j} = 1$. We remind that we interpret the *messages* $\chi_{q_i}^{i\to j}$ as a marginal probability that node i is in group t_i conditional to the absence of factor ij. The BP assumes that the only correlations between i's neighbors are mediated through i, so that if i were missing—or if its label were fixed—the distribution of its neighbors' states would be a product distribution. In that case, we can compute the message that i sends j recursively in terms of the messages that i receives from its other neighbors k.

Then the marginal probability is then estimated from a fixed point of BP to be $\mu_i(t_i) \approx \chi_{t_i}^i$, where

$$\chi_{t_i}^{i} = \frac{1}{Z^{i}} n_{t_i} \prod_{k\neq i} \left[\sum_{t_k} c_{t_i t_k}^{A_{ik}} \left(1 - \frac{c_{t_i t_k}}{N}\right)^{1-A_{ik}} \chi_{t_k}^{k\to i} \right]. \tag{10.12}$$

Since we have nonzero interactions between every pair of nodes, we have potentially $N(N-1)$ messages, and indeed (10.11) tells us how to update all of these for finite N. However, this gives an algorithm where even a single update takes $O(N^2)$ time, making it suitable only for networks of up to a few thousand nodes. Happily, for large sparse networks, i.e., when N is large and $c_{ab} = O(1)$, we can neglect terms of sub-leading order in N. In that case we can assume that i sends the same message to all its non-neighbors j, and treat these messages as an external field, so that we only need to keep track of $2M$ messages where M is the number of edges. In that case, each update step takes just $O(M) = O(N)$ time. To see this, suppose that $(i,j) \notin E$. We have

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}} n_{t_i} \prod_{k\notin \partial i\setminus j} \left[1 - \frac{1}{N} \sum_{t_k} c_{t_k t_i} \chi_{t_k}^{k\to i} \right] \prod_{k\in\partial i} \left[\sum_{t_k} c_{t_k t_i} \chi_{t_k}^{k\to i} \right] = \chi_{t_i}^{i} + O\left(\frac{1}{N}\right). \tag{10.13}$$

Hence the messages on non-edges do not depend to leading order on the target node j. On the other hand, if $(i,j) \in E$ we have

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}} n_{t_i} \prod_{k\notin \partial i} \left[1 - \frac{1}{N} \sum_{t_k} c_{t_k t_i} \chi_{t_k}^{k\to i} \right] \prod_{k\in\partial i\setminus j} \left[\sum_{t_k} c_{t_k t_i} \chi_{t_k}^{k\to i} \right]. \tag{10.14}$$

The belief propagation equations can hence be rewritten as

$$\chi^{i\to j}_{t_i} = \frac{1}{Z^{i\to j}}\, n_{t_i} e^{-h_{t_i}} \prod_{k\in\partial i\setminus j} \left[\sum_{t_k} c_{t_k t_i} \chi^{k\to i}_{t_k}\right] , \tag{10.15}$$

where we neglected terms that contribute $O(1/N)$ to $\chi^{i\to j}$, and defined an auxiliary external field that summarizes the contribution and overall influence of the non-edges

$$h_{t_i} = \frac{1}{N}\sum_k \sum_{t_k} c_{t_k t_i} \chi^k_{t_k} \,. \tag{10.16}$$

In order to find a fixed point of Eq. (10.15) in linear time we update the messages $\chi^{i\to j}$, recompute χ^j, update the field h_{t_i} by adding the new contribution and subtracting the old one, and repeat. The estimate of the marginal probability $\mu_i(t_i)$ is then

$$\chi^{i}_{t_i} = \frac{1}{Z^{i}}\, n_{t_i} e^{-h_{t_i}} \prod_{j\in\partial i} \left[\sum_{t_j} c_{t_j t_i} \chi^{j\to i}_{t_j}\right] . \tag{10.17}$$

The role of the magnetic field h_{t_i} in the belief propagation update is similar as the one of the prior on group sizes n_{t_i}. Contrary to the fixed n_{t_i} the field h_{t_i} adapts to the current estimation of the groups sizes. For the assortative communities this term is crucial as if one group a was becoming more represented then the corresponding field h_a would be larger and would weaken all the messages χ_a. The field is hence adaptively keeping the group sizes of the correct size preventing all node to fall into the same group. For the disassortative case the field plays a similar role of adjusting the sizes of the groups. A particular case if the disassortative structure with groups of the same size and the same average degree of every group in which case the term $e^{-h_{t_i}}$ does not change the behaviour of the BP equations. This will stand on the basis of the mapping to planted coloring problem.

When the Belief Propagation is asymptotically exact then the true marginal probabilities are given as $\mu_i(t_i) = \chi^i_{t_i}$. The estimator maximizing the overlap Q is then

$$\hat{s}_i = \mathrm{argmax}_{t_i} \chi^i_{t_i} \tag{10.18}$$

The overlap with the original group assignment is then computed from (10.8) under the assumption that the assumed parameters θ were indeed the ones used to generate the graph, leading to

$$Q = \lim_{N\to\infty} \frac{\frac{1}{N}\sum_i \chi^i_{\hat{s}_i} - \max_a n_a}{1-\max_a n_a} \tag{10.19}$$

In order to write the Bethe free entropy we use similar simplification and neglect subleading terms. To write the resulting formula we introduce

$$Z^{ij} = \sum_{a<b} c_{ab}(\chi^{i\to j}_a \chi^{j\to i}_b + \chi^{i\to j}_b \chi^{j\to i}_a) + \sum_a c_{aa}\chi^{i\to j}_a \chi^{j\to i}_a \quad \text{for} \quad (i,j)\in E \tag{10.20}$$

$$\tilde{Z}^{ij} = \sum_{a,b} \left(1 - \frac{c_{ab}}{N}\right) \chi^i_a \chi^j_b \quad \text{for} \quad (i,j)\notin E \,, \,, \tag{10.21}$$

$$Z^{i} = \sum_{t_i} n_{t_i} e^{-h_{t_i}} \prod_{j\in\partial i} \sum_{t_j} c_{t_j t_i} \chi^{j\to i}_{t_j} \tag{10.22}$$

we can then write the *Bethe free entropy*, in the thermodynamic limit as

$$\Phi_{\rm BP}(q,\{n_a\},\{c_{ab}\}) = \frac{1}{N}\sum_i \log Z^i - \frac{1}{N}\sum_{(i,j)\in E}\log Z^{ij} + \frac{c}{2}\,, \tag{10.23}$$

where c is the average degree given by (10.2) and the third term comes from the edge-contribution of non-edges (i.e. $\sum_{i,j}\log(\tilde{Z}^{ij})$). Derivation of this expression will be your homework.

Under the assumption that the model parameters θ are known, hence Nishomori conditions hold, we can now state a conjecture about exactness of the belief propagation in the asymptotic limit of $N\to\infty$. The fixed point of belief propagation corresponding to the largest Bethe free entropy provides the Bayes-optimal estimator for the SBM in the following sense:

- The Bethe free entropy $\Phi_{\rm BP}$ is exact, i.e. $\forall\varepsilon>0$
$$|\Phi_{BP}(\theta) - \Phi_{\rm exact}(\theta)| < \varepsilon \tag{10.24}$$
with high probability over the randomness in the model and as $N\to\infty$.
- The corresponding BP overlap is equal to the one obtained by the Bayes-optimal estimator $\forall\varepsilon>0$
$$|Q_{\rm BP} - Q_{\rm BO}| < \varepsilon \tag{10.25}$$
with high probability over the randomness in the model and as $N\to\infty$.
- Finally for all up to a vanishing fraction of node the BP marginals are correct $\forall\varepsilon>0$
$$|\chi^i_{s_i} - \mu_i(s_i)| < \varepsilon \tag{10.26}$$
for almost all $i=1,\dots,N$ with high probability as $N\to\infty$

In case the parameters θ are not known and need to be learned the Bethe free energy is greatly useful. In previous sections we concluded that the most likely parameters are those maximizing the Bethe free entropy. One can thus simply run gradient descent on the parameters to maximize the Bethe entropy. Even more conveniently, we can write the stationarity conditions of the Bethe free entropy with respect to the parameters n_a and c_{ab} and use them as an iterative procedure to estimate the parameters. Keeping in mind that a BP fixed point is a stationary point of the Bethe free entropy, and that for n_a we need to impose the normalization $\sum_a n_a = 1$ and thus we are looking for a constraint optimizer we obtain

$$n_a = \frac{1}{N}\sum_i \chi^i_a\,, \tag{10.27}$$

$$c_{ab} = \frac{1}{N}\frac{1}{n_b n_a}\sum_{(i,j)\in E}\frac{c_{ab}(\chi^{i\to j}_a\chi^{j\to i}_b + \chi^{i\to j}_b\chi^{j\to i}_a)}{Z^{ij}}\,, \tag{10.28}$$

where Z^{ij} is defined in (10.20). Derivation of this expression will be your homework. The interpretation of these expressions are very intuitive and again stems from Nishmori identities for Bayes-optimal estimation. The first equations states that the fraction of nodes in group a should be the expected number of nodes in group a according to the BP prediction. Similarly for c_{ab} the meaning of the second equations is that the expected number of edges between group a and group b is the same as when computed from the BP marginals. Therefore, BP can also be used readily to learn the optimal parameters.

10.4 The phase diagram of community detection

We will now consider the case when the parameters $q, \{n_a\}, \{c_{ab}\}$ used to generate the network are known. We will further limit ourselves to a particularly algorithmically difficult case of the block model, where every group a has the same average degree c and hence there is no information about the group assignment simply in the degree distribution. The condition reads:

$$\sum_{d=1}^{q} c_{ad} n_d = \sum_{d=1}^{q} c_{bd} n_d = c\,, \quad \text{for all } a, b\,. \tag{10.29}$$

If this is not the case, we can achieve a positive overlap with the original group assignment simply by labeling nodes based on their degrees. The first observation to make about the belief propagation equations (10.15) in this case is that

$$\chi_{t_i}^{i\to j} = n_{t_i} \tag{10.30}$$

is always a fixed point, as can be verified by plugging (10.30) into (10.15). The free entropy at this fixed point is

$$\Phi_{\text{para}} = -\frac{c}{2}\left(1 - \log c\right)\,. \tag{10.31}$$

For the marginals we have $\chi_{t_i}^{i} = n_{t_i}$, in which case the overlap (10.8) is $Q = 0$. This fixed point does not provide any information about the original assignment—it is no better than a random guess. If this fixed point gives the correct marginal probabilities and the correct free entropy, we have no hope of recovering the original group assignment. For which values of q, n_a and c_{ab} is this the case?

10.4.1 Second order phase transition

Fig. 10.4.1 represents two examples where the overlap Q is computed on a randomly generated graph with q groups of the same size and an average degree c. We set $c_{aa} = c_{\text{in}}$ and $c_{ab} = c_{\text{out}}$ for all $a \neq b$ and vary the ratio $\epsilon = c_{\text{out}}/c_{\text{in}}$. The continuous line is the overlap resulting from the BP fixed point obtained by converging from a random initial condition (i.e., where for each i, j the initial messages $\chi_{t_i}^{i\to j}$ are random normalized distributions on t_i). The convergence time is plotted in Fig. 10.4.2. The points in Fig. 10.4.1 are results obtained from Gibbs sampling, using the Metropolis rule and obeying detailed balance with respect to the posterior distribution, starting with a random initial group assignment $\{q_i\}$. We see that $Q = 0$ for $c_{\text{out}}/c_{\text{in}} > \epsilon_c$. In other words, in this region both BP and MCMC converge to the paramagnetic state, where the marginals contain no information about the original assignment. For $c_{\text{out}}/c_{\text{in}} < \epsilon_c$, however, the overlap is positive and the paramagnetic fixed point is not the one to which BP or MCMC converge.

Fig. 10.4.1(b) shows the case of $q = 4$ groups with average degree $c = 16$. We show the large N results and also the overlap computed with MCMC for a rather small size $N = 128$. Again, up to symmetry breaking, marginalization achieves the best possible overlap that can be inferred from the graph by any algorithm.

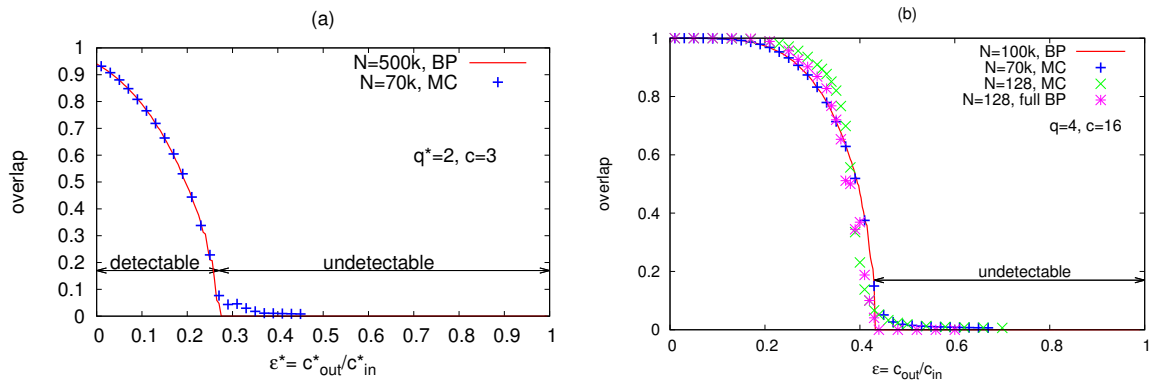


Figure 10.4.1: (color online): The overlap (10.4) between the original assignment and its best estimate given the structure of the graph, computed by the marginalization (10.7). Graphs were generated using N nodes, q groups of the same size, average degree c, and different ratios $\epsilon = c_{\rm out}/c_{\rm in}$. Thus $\epsilon = 1$ gives an Erdős-Rényi random graph, and $\epsilon = 0$ gives completely separated groups. Results from belief propagation (10.15) for large graphs (red line) are compared to Gibbs sampling, i.e., Monte Carlo Markov chain (MCMC) simulations (data points). The agreement is good, with differences in the low-overlap regime that we attribute to finite size fluctuations. In the part (b) we also compare to results from the full BP (10.11) and MCMC for smaller graphs with $N = 128$, averaged over 400 samples. The finite size effects are not very strong in this case, and BP is reasonably close to the exact (MCMC) result even on small graphs that contain many short loops. For $N \to \infty$ and $\epsilon > \epsilon_c = (c - \sqrt{c})/[c + \sqrt{c}(q-1)]$ it is impossible to find an assignment correlated with the original one based purely on the structure of the graph. For two groups and average degree $c = 3$ this means that the density of connections must be $\epsilon_c^{-1}(q = 2, c = 3) = 3.73$ greater within groups than between groups to obtain a positive overlap.

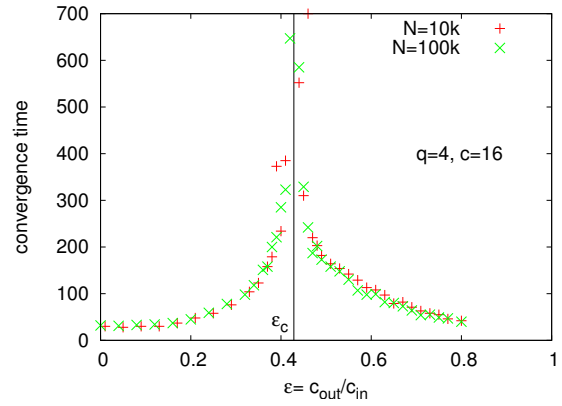


Figure 10.4.2: (color online): The number of iterations needed for convergence of the BP algorithm for two different sizes. The convergence time diverges at the critical point ϵ_c. The equilibration time of Gibbs sampling (MCMC) has qualitatively the same behavior, but BP obtains the marginals much more quickly.

10.4.2 Stability of the paramagnetic fixed point

Let us now investigate the stability of the paramagnetic fixed point under random perturbations to the messages when we iterate the BP equations. In the sparse case where $c_{ab} = O(1)$, graphs generated by the block model are locally treelike in the sense that almost all nodes have a neighborhood which is a tree up to distance $O(\log N)$, where the constant hidden in the O depends on the matrix c_{ab}. Equivalently, for almost all nodes i, the shortest loop that i belongs to has length $O(\log N)$. Consider such a tree with d levels, in the limit $d \to \infty$. Assume that on the leaves the paramagnetic fixed point is perturbed as

$$\chi_t^k = n_t + \epsilon_t^k \,, \tag{10.32}$$

and let us investigate the influence of this perturbation on the message on the root of the tree, which we denote k_0. There are, on average, c^d leaves in the tree where c is the average degree. The influence of each leaf is independent, so let us first investigate the influence of the perturbation of a single leaf k_d, which is connected to k_0 by a path $k_d, k_{d-1}, \ldots, k_1, k_0$. We define a kind of transfer matrix

$$T_{ab}^i \equiv \frac{\partial \chi_a^{k_i}}{\partial \chi_b^{k_{i+1}}}\Big|_{\chi_t = n_t} = \left[\frac{\chi_a^{k_i} c_{ab}}{\sum_r c_{ar} \chi_r^{k_{i+1}}} - \chi_a^{k_i} \sum_s \frac{\chi_s^{k_i} c_{sb}}{\sum_r c_{sr} \chi_r^{k_{i+1}}} \right] \Big|_{\chi_t = n_t} = n_a \left(\frac{c_{ab}}{c} - 1 \right) . \tag{10.33}$$

where this expression was derived from (10.15) to leading order in N. The perturbation $\epsilon_{t_0}^{k_0}$ on the root due to the perturbation $\epsilon_{t_d}^{k_d}$ on the leaf k_d can then be written as

$$\epsilon_{t_0}^{k_0} = \sum_{\{t_i\}_{i=1,\ldots,d}} \left[\prod_{i=0}^{d-1} T_{t_i, t_{i+1}}^i \right] \epsilon_{t_d}^{k_d} \tag{10.34}$$

We observe in (10.33) that the matrix T_{ab}^i does not depend on the index i. Hence (10.34) can be written as $\epsilon^{k_0} = T^d \epsilon^{k_d}$. When $d \to \infty$, T^d will be dominated by T's largest eigenvalue λ, so $\epsilon^{k_0} \approx \lambda^d \epsilon^{k_d}$.

Now let us consider the influence from all c^d of the leaves. The mean value of the perturbation on the leaves is zero, so the mean value of the influence on the root is zero. For the variance, however, we have

$$\left\langle \left(\epsilon_{t_0}^{k_0} \right)^2 \right\rangle \approx \left\langle \left(\sum_{k=1}^{c^d} \lambda^d \epsilon_t^k \right)^2 \right\rangle \approx c^d \lambda^{2d} \left\langle \left(\epsilon_t^k \right)^2 \right\rangle . \tag{10.35}$$

This gives the following stability criterion,

$$c\lambda^2 = 1 \,. \tag{10.36}$$

For $c\lambda^2 < 1$ the perturbation on leaves vanishes as we move up the tree and the paramagnetic fixed point is stable. On the other hand, if $c\lambda^2 > 1$ the perturbation is amplified exponentially, the paramagnetic fixed point is unstable, and the communities are easily detectable.

Consider the case with q groups of equal size, where $c_{aa} = c_{\rm in}$ for all a and $c_{ab} = c_{\rm out}$ for all $a \neq b$. If there are q groups, then $c_{\rm in} + (q-1)c_{\rm out} = qc$. The transfer matrix T_{ab} has only two distinct eigenvalues, $\lambda_1 = 0$ with eigenvector $(1, 1, \ldots, 1)$, and $\lambda_2 = (c_{\rm in} - c_{\rm out})/(qc)$ with

eigenvectors of the form $(0, \ldots, 0, 1, -1, 0, \ldots, 0)$ and degeneracy $q - 1$. The paramagnetic fixed point is then unstable, and communities are easily detectable, if

$$|c_{\rm in} - c_{\rm out}| > q\sqrt{c}\,. \tag{10.37}$$

The stability condition (10.36) is known in the literature on spin glasses as the de Almeida-Thouless local stability condition de Almeida and Thouless (1978), in information science as the Kesten-Stigum bound on reconstruction on trees Kesten and Stigum (1967).

We observed empirically that for random initial conditions both the belief propagation converges to the paramagnetic fixed point when $c\lambda^2 < 1$. On the other hand when $c\lambda^2 > 1$ then BP converges to a fixed point with a positive overlap, so that it is possible to find a group assignment that is correlated (often strongly) to the original assignment. We thus conclude that if the parameters $q, \{n_a\}, \{c_{ab}\}$ are known and if $c\lambda^2 > 1$, it is possible to reconstruct the original group assignment.

For the cases presented in Fig. 10.4.1 we can thus distinguish two phases:

- If $|c_{\rm in} - c_{\rm out}| < q\sqrt{c}$, the graph does not contain any significant information about the original group assignment, and community detection is impossible. Moreover, the network generated with the block model is *indistinguishable* from an Erdős-Rényi random graph of the same average degree.
- If $|c_{\rm in} - c_{\rm out}| > q\sqrt{c}$, the graph contains significant information about the original group assignment, and using BP or MCMC yields an assignment that is strongly correlated with the original one. There is some intrinsic uncertainty about the group assignment due to the entropy, but if the graph was generated from the block model there is no better method for inference than the marginalization introduced by Eq. (10.7).

Fig. 10.4.1 hence illustrates a phase transition in the detectability of communities. Unless the ratio $c_{\rm out}/c_{\rm in}$ is far enough from 1, the groups that truly existed when the network was generated are undetectable from the topology of the network. Moreover, unless the condition (10.37) is satisfied the graph generated by the block model is indistinguishable from a random graph, in the sense that typical thermodynamic properties of the two ensembles are the same.

10.4.3 First order phase transition

The situation of a continuous (2nd order) phase transition, illustrated in Fig. 10.4.1 is, however, not the most general one. Fig. 10.4.3 illustrates the case of a discontinuous (1st order) phase transition that occurs e.g. for $q = 5$, $c_{\rm in} = 0$, and $c_{\rm out} = qc/(q-1)$. In this case the condition for stability (10.37) leads to a threshold value $c_\ell = (q-1)^2$. We plot again the overlap obtained with BP, using two different initializations: the random one, and the planted/informed one corresponding to the original assignment. In the latter case, the initial messages are

$$\chi_{q_i}^{i\to j} = \delta_{q_i s_i^*}\,, \tag{10.38}$$

where s_i^* is the original assignment. We also plot the corresponding BP free energies. As the average degree c increases, we see four different phases in Fig. 10.4.3:

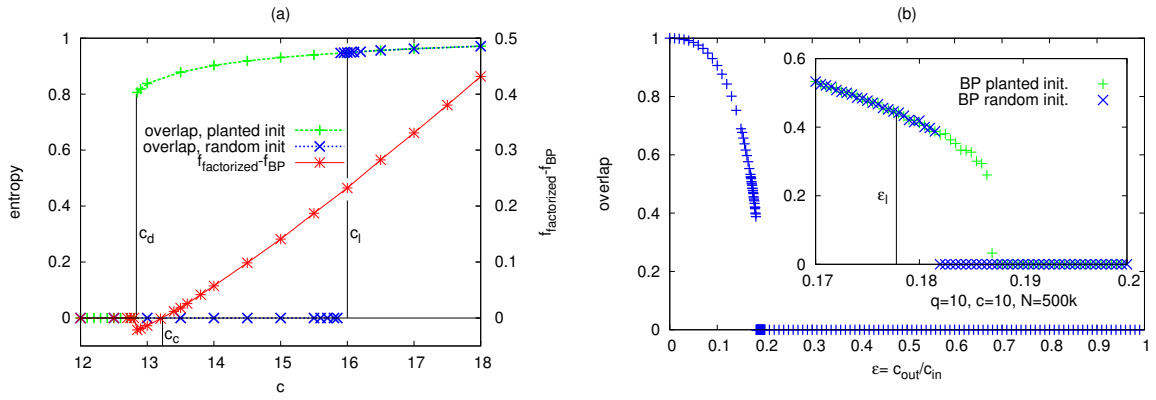


Figure 10.4.3: (color online): (a) Graphs generated with $q = 5$, $c_{\rm in} = 0$, and $N = 10^5$. We compute the overlap (10.4) and the free energy with BP for different values of the average degree c. The green crosses show the overlap of the BP fixed point resulting from using the original group assignment as the initial condition, and the blue crosses show the overlap resulting from random initial messages. The red stars show the difference between the paramagnetic free energy (10.31) and the free energy resulting from the informed initialization. We observe three important points where the behavior changes qualitatively: $c_d = 12.84$, $c_c = 13.23$, and $c_\ell = 16$. We discuss the corresponding phase transitions in the text. (b) The case $q = 10$ and $c = 10$. We plot the overlap as a function of ϵ; it drops down abruptly from about $Q = 0.35$. The inset zooms in on the critical region. We mark the stability transition ϵ_ℓ, and data points for $N = 5 \cdot 10^5$ for both the random and informed initialization of BP. In this case the data are not so clear. The overlap from random initialization becomes positive a little before the asymptotic transition. We think this is due to strong finite size effects. From our data for the free energy it also seems that the transitions ϵ_c and ϵ_d are very close to each other (or maybe even equal, even though this would be surprising). These subtle effects are, however, relevant only in a very narrow region of ϵ and are, in our opinion, not likely to appear for real-world networks.

I. For $c < c_d$, both initializations converge to the paramagnetic fixed point, so the graph does not contain any significant information about the original group assignment.

II. For $c_d < c < c_c$, the planted initialization converges to a fixed point with positive overlap, and its free entropy is smaller than the paramagnetic free entropy. In this phase there are exponentially many basins of attraction (states) in the space of assignments that have the proper number of edges between each pair of groups. These basins of attraction have zero overlap with each other, so none of them yield any information about any of the others, and there is no way to tell which one of them contains the original assignment. The paramagnetic free entropy is still the correct total free entropy, the graphs generated by the block model are thermodynamically indistinguishable from Erdős-Rényi random graphs, and there is no way to find a group assignment correlated with the original one.

III. For $c_c < c < c_\ell$, the planted initialization converges to a fixed point with positive overlap, and its free entropy is larger than the paramagnetic free entropy. There might still be exponentially many basins of attraction in the state space with the proper number of edges between groups, but the one corresponding to the original assignment is the one with the largest free entropy. Therefore, if we can perform an exhaustive search of the state space, we can infer the original group assignment. However, this would take

exponential time, and initializing BP randomly almost always leads to the paramagnetic fixed point. In this phase, inference is possible, but conjectured to be exponentially hard computationally.

IV. For $c > c_\ell$, both initializations converge to a fixed point with positive overlap, strongly correlated with the original assignment. Thus inference is both possible and tractable, and BP achieves it in linear time.

We saw in our experiments that for assortative communities where $c_{\rm in} > c_{\rm out}$, phases (II) and (III) are extremely narrow or nonexistent. For $q \leq 4$, these phases do not exist at all, and the overlap grows continuously from zero in phase (IV), giving a continuous phase transition as illustrated in Fig. 10.4.1. We can think of the continuous (2nd order) phase transition as a degenerate case of the 1st order phase transition where the three discussed thresholds $c_d = c_c = c_\ell$ are the same. For $q \geq 5$, phases (II) and (III) exist but occur in an extremely narrow region, as shown in Fig. 10.4.3(b). The overlap jumps discontinuously from zero to a relatively large value, giving a discontinuous phase transition. In the disassortative (antiferromagnetic) case where $c_{\rm in} < c_{\rm out}$, phases (II) and (III) are more important. For instance, when $c_{\rm in} = 0$ and the number of groups is large, the thresholds scale as $c_d \approx q \log q$, $c_c \approx 2q \log q$ and $c_\ell = (q-1)^2$. However, in phase (III) the problem of inferring the original assignment is hard and conjectured to be insurmountable to all polynomial algorithms in this region.

10.4.4 The non-backtracking matrix and spectral method

Having seen that BP and also MCMC are able to attain the algorithmic threshold c_l a natural question is what other classes of algorithms are able to do so. Very natural spectral methods such as principal component analysis, based on leading eigenvalues of the associated matrices, are commonly used for clustering. For community detection in sparse graphs, however, most of the commonly used spectral methods do not attain the threshold because in sparse graphs their leading eigenvectors become localized on subgraphs that have nothing to do with the communities (e.g. the adjacency matrix spectrum localizes on high degree nodes in the limit of large sparse graphs).

A very generic and powerful idea to design spectral methods achieving optimal thresholds is to realize that the way we computed the threshold c_l in the first place was by linearizing the belief propagation around it paramagnetic fixed point by introducing its perturbations $\chi_t^{i\to j} = n_t + \varepsilon_t^{i\to j}$ we obtained

$$\varepsilon_t^{i\to j} = \sum_{k\in\partial i\setminus j} \sum_q \frac{\partial \chi_t^{i\to j}}{\partial \chi_q^{k\to i}}\Big|_{\chi_q^{k\to i}=n_q} \varepsilon_q^{k\to i} = \sum_{k\in\partial i\setminus j} \sum_q T_{tq}\, \varepsilon_q^{k\to i} \tag{10.39}$$

where the matrix T was computed in (10.33). Introducing the so-called non-backtracking matrix as a $2M$ but $2M$ matrix (M being the number of edges) with coordinates corresponding to oriented edges

$$B_{i\to j,k\to l} = \delta_{i,l}(1-\delta_{j,k}) \tag{10.40}$$

we can write the linearized BP as

$$\varepsilon = (T\otimes B)\varepsilon\,. \tag{10.41}$$

We thus see that the linearized belief propagation corresponds to a power-iteration of a tensor product of a small $q \times q$ matrix T and the non-backtracking matrix B. The spectrum of B indeed provides information about the communities that is more accurate than other commonly used spectral methods and it attains the detectability threshold in sparse SBM.

10.5 Exercises

Exercise 10.1: SBM free energy

Show that the Bethe free entropy for the stochastic block model can be written using

$$Z^{ij} = \sum_{a<b} c_{ab}(\chi_a^{i\to j}\chi_b^{j\to i} + \chi_b^{i\to j}\chi_a^{j\to i}) + \sum_a c_{aa}\chi_a^{i\to j}\chi_a^{j\to i} \quad \text{for} \quad (i,j) \in E \tag{10.42}$$

$$Z^{i} = \sum_{t_i} n_{t_i} e^{-h_{t_i}} \prod_{j\in\partial i} \sum_{t_j} c_{t_j t_i} \chi_{t_j}^{j\to i} \tag{10.43}$$

as

$$\Phi_{\rm BP}(q, \{n_a\}, \{c_{ab}\}) = \frac{1}{N}\sum_i \log Z^i - \frac{1}{N}\sum_{(i,j)\in E} \log Z^{ij} + \frac{c}{2}\,, \tag{10.44}$$

where c is the average degree given by (10.2).

Exercise 10.2: Parameter learning with BP

Show that in the stochastic block model maximization of the Bethe free entropy with respect to the parameters n_a and c_{ab} at a BP fixed point leads to the following conditions for stationarity that can be then used for iterative learning of the parameters n_a and c_{ab}.

$$n_a = \frac{1}{N}\sum_i \chi_a^i\,, \tag{10.45}$$

$$c_{ab} = \frac{1}{N}\frac{1}{n_b n_a}\sum_{(i,j)\in E} \frac{c_{ab}(\chi_a^{i\to j}\chi_b^{j\to i} + \chi_b^{i\to j}\chi_a^{j\to i})}{Z^{ij}}\,. \tag{10.46}$$

Chapter 11

The spin glass game from sparse to dense graphs

He deals the cards as a meditation
And those he plays, never suspect
He doesn't play for the money he wins
He don't play for respect
He deals the cards to find the answer
The sacred geometry of chance
The hidden law of a probable outcome
The numbers lead a dance

Sting – Shape of my heart, 1993

11.1 The spin glass game

Consider the following problem: N people are given a red or black card, or, equivalently, a value $S_i^* = \pm 1$. You are allowed to ask M pairs of people, randomly, to tell you if they had the same card (without telling you which one). Can we figure out the two groups?

This is simple enough to answer formally with Bayesian statistics. Given the M answers $J_{ij} = \pm 1$ (1 for the same card, -1 for different ones) the posterior probability assignment is given by

$$P_{\rm post}(\boldsymbol{S}|J) = \frac{1}{\mathcal{N}} P_{\rm post}(J|\boldsymbol{S}) P_{\rm prior}(\boldsymbol{S}) = \frac{1}{\mathcal{N} 2^N} \prod_{ij \in \mathcal{G}} P(J_{ij} = S_i S_j) \tag{11.1}$$

where $\mathcal{G}$ is the graph where two sites are connected if you asked the question to the pairs. Given some people are not trustworthy, let us denote the probability of lies as p. Then

$$P_{\rm post}(\boldsymbol{S}|J) \quad = \tfrac{1}{\mathcal{N} 2^N} \textstyle\prod_{ij \in \mathcal{G}} \left[(1-p)\delta(J_{ij} = S_i S_j) + p\delta(J_{ij} = -S_i S_j) \right] \tag{11.2}$$

and using the change of variable $p = e^{-\beta}/(e^{-\beta} + e^{\beta}) = 1/(1 + e^{2\beta})$ we find

$$P_{\text{post}}(\boldsymbol{S}|J) = \frac{1}{Z} e^{\beta \sum_{ij\in\mathcal{G}} J_{ij} S_i S_j} \tag{11.3}$$

This is the spin glass problem with Hamiltonian $\mathcal{H} = -\sum J_{ij} S_i S_j$, thus the name the "Spin Glass Game" for this particular inference problem.

11.2 Sparse graph

11.2.1 Belief Propagation

First, we discuss the problem on sparse graphs. In this case, as $N, M \to \infty$, we have random tree-like regular graph, and we can thus write the BP equations. We can use the result from appendix 5.A, For such pair-wise models, we have one factor per edge, and the BP equations read:

$$\chi_{s_j}^{j\to(ij)} = \frac{1}{Z^{j\to(ij)}} \prod_{(kj)\in\partial j\setminus(ij)} \psi_{s_j}^{(kj)\to j} \tag{11.4}$$

$$\psi_{s_i}^{(ij)\to i} = \frac{1}{Z^{(ij)\to i}} \sum_{s_j} e^{\beta J_{ij} s_i s_j} \chi_{s_j}^{j\to(ij)} \tag{11.5}$$

It is convenient to write the iteration using the parametrization

$$\chi_{s_j}^{j\to(ij)} = \frac{1 + s_j m^{j\to(ij)}}{2} \tag{11.6}$$

$$\psi_{s_i}^{(ij)\to i} = \frac{e^{\beta s_i h^{(ij)\to i}}}{2\cosh(\beta h^{(ij)\to i})} \tag{11.7}$$

so that

$$m^{j\to(ij)} = \tanh\left(\beta \sum_{(kj)\in\partial j\setminus(ij)} h^{(kj)\to j}\right) \tag{11.8}$$

$$\frac{e^{\beta s_i h^{(ij)\to i}}}{2\cosh(\beta h^{(ij)\to i})} = \frac{1}{Z^{(ij)\to i}} \sum_{s_j} e^{\beta J_{ij} s_i s_j} \frac{1 + s_j m^{j\to(ij)}}{2} \tag{11.9}$$

The second equation can equivalently be written as

$$\tanh \beta h^{(ij)\to i} = \frac{X_+ - X_-}{X_+ + X_-} \tag{11.10}$$

$$X_s = \sum_{s_j} e^{\beta J_{ij} s s_j} \frac{1 + s_j m^{j\to(ij)}}{2} \tag{11.11}$$

$$= \frac{1}{2} e^{\beta s J_{ij}} (1 + m^{j\to(ij)}) + \frac{1}{2} e^{-\beta s J_{ij}} (1 - m^{j\to(ij)}) \tag{11.12}$$

Using now the relation $\text{atanh}y = \frac{1}{2}\log\frac{1+x}{1-x}$, and applying the atanh on both side, we reach

$$\beta h^{(ij)\to i} = \frac{1}{2}\log\frac{X_+}{X_-} \tag{11.13}$$

$$= \frac{1}{2}\log\frac{e^{\beta J_{ij}}(1+m^{j\to(ij)}) + e^{-\beta J_{ij}}(1-m^{j\to(ij)})}{e^{-\beta J_{ij}}(1+m^{j\to(ij)}) + e^{\beta J_{ij}}(1-m^{j\to(ij)})} \tag{11.14}$$

$$= \frac{1}{2}\log\frac{\cosh(\beta J_{ij}) + m^{j\to(ij)})\sinh(\beta J_{ij})}{\cosh(\beta J_{ij}) - m^{j\to(ij)})\sinh(\beta J_{ij})} \tag{11.15}$$

$$= \frac{1}{2}\log\frac{1+m^{j\to(ij)})\tanh(\beta J_{ij})}{1-m^{j\to(ij)})\tanh(\beta J_{ij})} \tag{11.16}$$

$$= \text{atanh}\left(m^{j\to(ij)})\tanh(\beta J_{ij})\right) \tag{11.17}$$

We thus finally write BP more conveniently as

$$m^{j\to(ij)} = \tanh\left(\beta \sum_{(kj)\in\partial j\setminus(ij)} h^{(kj)\to j}\right) \tag{11.18}$$

$$h^{(ij)\to i} = \frac{1}{\beta}\text{atanh}\left(m^{j\to(ij)})\tanh(\beta J_{ij})\right) \tag{11.19}$$

Or even better (realizing that the notation $m^{j\to(ij)}$ can be written without ambiguity as $m^{j\to i}$)

$$m^{j\to i} = \tanh\left(\sum_{k\in\partial j\setminus i} \text{atanh}\left(m^{k\to j}\tanh(\beta J_{kj})\right)\right) \tag{11.20}$$

With this choice of notation, the iteration is really practical.

11.2.2 Population dynamics

One can prove this algorithm perfectly solve the spin glass game. We can also analyze the Bayes performances, using population dynamics. Defining

$$\mathcal{F}_{\text{BP}}\left(\{m^{k\to j}\}\right) = \tanh\left(\sum_{k\in\partial j\setminus i} \text{atanh}\left(m^{k\to j}\tanh(\beta J_{kj})\right)\right) \tag{11.21}$$

The probability density distribution of messages is given by the fixed point of

$$\mathcal{P}(m^{\text{cav}}) = \int dc P^{\text{excess}}(c)\prod_{i=1}^{c} dm_i \mathcal{P}(m_i)\delta(m - \mathcal{F}_{\text{BP}}\{m\}) \tag{11.22}$$

11.2.3 Phase transition

It is easy to locate the phase transition, as it turns out to be a second-order one, using the local perturbation approach. Writing $m^{j\to i} = \epsilon^{j\to i}S_*^j$ we find to linear order than

$$\epsilon^{j\to i}S_*^j \approx \tanh\left(\sum_{k\in\partial j\setminus i}\left(\epsilon^{k\to j}S_*^k\tanh(\beta J_{kj})\right)\right) \approx \left(\sum_{k\in\partial j\setminus i}\left(\epsilon^{k\to j}S_*^k\tanh(\beta J_{kj})\right)\right) \tag{11.23}$$

or equivalently for an homogeneous growth

$$\epsilon \approx \left(\sum_{k \in \partial j \setminus i} \left(\epsilon S_*^k S_*^j \tanh(\beta J_{kj}) \right) \right) \tag{11.24}$$

so that on average, the magnetization will increase if

$$1 = c\mathbb{E}\left[S_*^i S_*^j \tanh(\beta J_{ij}) \right] = c(1-2p)\tanh(\beta) = c\tanh(\beta)^2 \tag{11.25}$$

so that the transition arise at

$$\beta = \sqrt{\frac{1}{\operatorname{atanh}(2M/N)}} = \sqrt{\frac{1}{\operatorname{atanh}(c)}} \tag{11.26}$$

11.3 Dense graph limit: TAP/AMP

Now we look to the problem with DENSE graph, in fact we may assume we observe ALL pairs, but to make the problem interesting, we take the probability of lies close to $1/2$. We write

$$p_{\text{Dense}} = e^{-\beta/\sqrt{N}}/(e^{-\beta/\sqrt{N}} + e^{\beta/\sqrt{N}}) = 1/(1+e^{2\beta/\sqrt{N}}) \tag{11.27}$$

In this limit, the problem becomes

$$P_{\text{post}}(\boldsymbol{S}|J) = \frac{1}{Z} e^{\frac{\beta}{\sqrt{N}} \sum_{i<j} J_{ij} S_i S_j} \tag{11.28}$$

Of course, we can write the same BP algorithm

$$m^{j\to i} = \tanh\left(\sum_{k \in \partial j \setminus i} \operatorname{atanh}\left(m^{k\to j} \tanh(\beta/\sqrt{N} J_{kj}) \right) \right) \tag{11.29}$$

and the transition arise at $\beta = 1$ since the criterion becomes

$$1 = \frac{2N^2/2}{N} \tanh(\beta/\sqrt{(N)})^2 \approx \beta^2 \tag{11.30}$$

11.4 Approximate Message Passing

This is a very nice limit to study the problem, however there is a little annoying fact: we have to update N^2 messages! This is way too much!!!! The trick is now to Taylor expand BP. We start by the BP iteration, which reads at first order:

$$m_{t+1}^{j\to i} = \tanh\left(\sum_{k \in \partial j \setminus i} \operatorname{atanh}\left(m_t^{k\to j} \tanh(\beta/\sqrt{N} J_{kj}) \right) \right) \approx \tanh\left(\frac{\beta}{\sqrt{N}} \sum_{k \in \partial j \setminus i} \left(m_t^{k\to j} J_{kj} \right) \right) \tag{11.31}$$

At this point, we realize that we can close the equation on the full marginal defined as

$$m_{t+1}^j = \tanh\left(\frac{\beta}{\sqrt{N}}\sum_k \left(m_t^{k\to j} J_{kj}\right)\right) \tag{11.32}$$

Indeed

$$m_{t+1}^{j\to i} = \tanh\left(\frac{\beta}{\sqrt{N}}\sum_k \left(m_t^{k\to j} J_{kj}\right) - \frac{\beta}{\sqrt{N}}\left(m_t^{i\to j} J_{ij}\right)\right) \tag{11.33}$$

$$\approx \tanh\left(\frac{\beta}{\sqrt{N}}\sum_k \left(m_t^{k\to j} J_{kj}\right)\right) - \frac{\beta}{\sqrt{N}}\left(m_t^{i\to j} J_{ij}\right)\tanh'\left(\frac{\beta}{\sqrt{N}}\sum_k \left(m_t^{k\to j} J_{kj}\right)\right) \tag{11.34}$$

$$\approx m_{t+1}^j - \frac{\beta}{\sqrt{N}}\left(m_t^{i\to j} J_{ij}\right)\left(1 - {m_{t+1}^j}^2\right) \tag{11.35}$$

$$\approx m_{t+1}^j - \frac{\beta}{\sqrt{N}}\left(m_t^i J_{ij}\right)\left(1 - {m_{t+1}^j}^2\right) \tag{11.36}$$

$$\tag{11.37}$$

where we keep the first correction in N. Finally, combining the two following equations:

$$m_t^{k\to j} = m_t^k - \frac{\beta}{\sqrt{N}}\left(m_{t-1}^j J_{jk}\right)\left(1 - m_t^{k\,2}\right) \tag{11.38}$$

$$m_{t+1}^j = \tanh\left(\frac{\beta}{\sqrt{N}}\sum_k \left(m_t^{k\to j} J_{kj}\right)\right) \tag{11.39}$$

we reach

$$m_{t+1}^j = \tanh\left(\frac{\beta}{\sqrt{N}}\sum_k J_{kj}\left(m_t^k - \frac{\beta}{\sqrt{N}}\left(m_{t-1}^j J_{jk}\right)\left(1 - m_t^{k\,2}\right)\right)\right) \tag{11.40}$$

$$m_{t+1}^j = \tanh\left(\frac{\beta}{\sqrt{N}}\sum_k \left(J_{kj} m_t^k - \frac{\beta}{\sqrt{N}}\left(m_{t-1}^j\right)\left(1 - m_t^{k\,2}\right)\right)\right) \tag{11.41}$$

$$m_{t+1}^j = \tanh\left(\frac{\beta}{\sqrt{N}}\sum_k J_{jk} m_t^k - \beta^2 m_{t-1}^j\left(1 - \sum_k \frac{m_t^{k\,2}}{N}\right)\right) \tag{11.42}$$

This can be conveniently written as

$$\mathbf{h}_t = \frac{1}{\sqrt{N}} J\mathbf{m}_t - \beta\mathbf{m}_{t-1}(1 - \overline{\mathbf{m_t}^2}) \tag{11.43}$$

$$\mathbf{m}_{t+1} = \tanh\beta\mathbf{h}^t \tag{11.44}$$

This is the TAP, or AMP algorithm. The second term in the first equation is called the Onsager term, and it makes a subtle difference with the naive mean field approx!

In full generality we can write

$$\mathbf{h}_t = \frac{1}{\sqrt{N}} J\mathbf{m}_t - \mathbf{m}_{t-1}\overline{\partial_h \eta(\beta \mathbf{h}^{t-1})} \tag{11.45}$$

$$\mathbf{m}_{t+1} = \eta(\beta \mathbf{h}^t) \tag{11.46}$$

Note how convenient and easy is it to write this algorithm! In fact in this form, this is known as the AMP algorithm!

11.4.1 State Evolution

Let us come back on the iteration

$$m_{t+1}^{j\to i} = \tanh\left(\sum_{k\in\partial j\setminus i} \operatorname{atanh}\left(m_t^{k\to j}\tanh(\beta/\sqrt{N}J_{kj})\right)\right) \tag{11.47}$$

Instead of focusing on the population dynamics on m, it seems like a good idea to decompose the iteration as

$$m_{t+1}^{j\to i} = \tanh h_t^{j\to i} \tag{11.48}$$

$$h_t^{j\to i} = \left(\sum_{k\in\partial j\setminus i} \operatorname{atanh}\left(m_t^{k\to j}\tanh(\beta/\sqrt{N}J_{kj})\right)\right) \tag{11.49}$$

With this notation, the distribution of $h_t^{j\to i}$ is a sum of uncorrelated terms on a tree, so we expect that for large tree, in the limit of large connectivities, we could simply things a bit. Let us first rewrite in the large connectivity limits:

$$m_{t+1}^{j\to i} = \tanh h_t^{j\to i} \tag{11.50}$$

$$h_t^{j\to i} = \frac{\beta}{\sqrt{N}}\sum_k \left(m_t^{k\to j}J_{kj})\right) \tag{11.51}$$

we have

$$h_t^{j\to i} \sim \mathcal{N}(\overline{h}^t, \Delta h^t) \tag{11.52}$$

With this, we can close the equations as follows: the mean and variance of the distribution of h is m and q, then

$$h_t^{j\to i}S_i^* = S_i^*\frac{\beta}{\sqrt{N}}\sum_k \delta(J=OK)S_*^iS_k^*m_t^{k\to j} - \delta(J=lies)S_*^iS_k^*m_t^{k\to j}) \tag{11.53}$$

$$= \frac{\beta}{\sqrt{N}}\sum_k \delta(J=OK)S_k^*m_t^{k\to j} - \delta(J=lies)S_k^*m_t^{k\to j}) \tag{11.54}$$

Let us denote the mean and second moments of the distribution of the m in the direction of the hidden states as

$$q^t = \mathbb{E}[(m_{i\to j}S_i^*)^2] \tag{11.55}$$
$$m^t = \mathbb{E}[m_{i\to j}S_i^*] \tag{11.56}$$

then we have

$$\overline{h}^t = N\frac{\beta}{\sqrt{N}}(1-2p)m_t = N\frac{\beta}{\sqrt{N}}\tanh(\beta/\sqrt{N})m_t = \beta^2 m_t \tag{11.57}$$

and

$$\Delta_t = \beta^2 q_t \tag{11.58}$$

so that we can write

$$q^t = \mathbb{E}_z \tanh\left(\beta^2 m_t + z\beta\sqrt{q_t}\right)^2 \tag{11.59}$$
$$m^t = \mathbb{E}_z \tanh\left(\beta^2 m_t + z\beta\sqrt{q_t}\right) \tag{11.60}$$

Which is the same as what we could have had with the replica method.

11.5 From the spin glass game to rank-one factorization

It turns out what we wrote here is entirely general in the large connectivity limit. Assuming the following problem:

$$Y = \sqrt{\frac{1}{N}}x^*(x^*)^T + \sqrt{\Delta}W \tag{11.61}$$

where W is a random matrix, and x^* is sampled from a distribution P_0, it turns out that if we wanted to solve the problem, we could simply write the AMP algorithm as

$$\mathbf{h}_t = \frac{1}{\sqrt{N}}J\mathbf{m}_t - \mathbf{m}_{t-1}\beta\overline{\eta'(\beta\mathbf{h}^{t-1})} \tag{11.62}$$
$$\mathbf{m}_{t+1} = \eta(\beta\mathbf{h}^t) \tag{11.63}$$

where this time we should actually use a different denoiser:

$$\eta_t(h) = \eta(A = \beta\overline{\mathbf{m}_t^2}, B = \beta h) =: \frac{\int dx P_X(x)xe^{-x^2A/2+xB}}{\int dx P_X(x)e^{-x^2A/2+xB}} \tag{11.64}$$

Using this algorithm, we, again, can get the state evolution as

$$m^t = \mathbb{E}_{z,x^*}[x^*\eta\left(\beta m^t, \beta m_t x^* + z\right)] \tag{11.65}$$

Chapter 12

Approximate Message Passing and State Evolution

I hope that someone gets my message in a bottle.

Sting, the Police – 1979

12.1 AMP And SE

Let us follow Bolthausen Bolthausen (2009) and Bolthausen (2014) and consider the iteration using symmetric matrices.

12.1.1 Symmetric AMP

Consider the symmetric AMP iteration with the same prescriptions as in i.e. assumptions on the matrices, functions and inputs.

$$\mathbf{x}^{t+1} = \mathbf{A}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.1}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{12.2}$$

with initialization at $\mathbf{x}^0$ and

$$b_t = \mathbb{E}\left[\mathrm{div} f_t(\mathbf{Z}^t)\right] \tag{12.3}$$

where $\mathbf{Z}^t \sim \mathbf{N}(0, \kappa_{t,t}\mathbf{I}_n)$. We claim that in the iterations, all xs are behaving as Gaussian variables, with $\mathbf{x}^t \sim \mathcal{N}(0, \kappa_t)$, and $\mathbf{X} \sim \mathcal{N}(0, \kappa_{t,s})$. Additionally, we also have

$$\kappa_{s,t} = \frac{1}{N}\mathbf{m}^{s-1}\mathbf{m}^{t-1} =: q_{s-1,t-1} \tag{12.4}$$

This is precisely what would happen WITHOUT the Onsager term IF the matrix A would be a new one at each iteration time. The Onsager correction makes it work in such a way that it remains true when A remains the same over iteration.

12.1.2 Conditioning on linear observations

Define the σ-algebra $\mathfrak{S}_t = \sigma(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^t)$. We then have :

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \mathbf{A}|_{\mathfrak{S}_t}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.5}$$

because $\mathbf{m}^t, \mathbf{m}^{t-1}$ are $\mathfrak{S}_t$-measurable. Ok.
A simple recursion shows that conditioning on $\mathfrak{S}_t$ is equivalent to conditioning on the gaussian space generated by $\mathbf{A}\mathbf{m}^0, \mathbf{A}\mathbf{m}^1, ..., \mathbf{A}\mathbf{m}^{t-1}$. This is a subspace of the Gaussian space generated by the entries of $\mathbf{A}$. Conditioning a Gaussian space on its subspace amounts to doing orthogonal projections, which gives, for GOE(n) :

$$\mathbf{A}|_{\mathfrak{S}_t} = \mathbb{E}\left[\mathbf{A}|\mathfrak{S}_t\right] + \mathcal{P}_t(\mathbf{A}) \tag{12.6}$$

$$= \mathbf{A} - \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}} + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}} \tag{12.7}$$

where $\mathbf{M}_{t-1} = \left[m^0|...|m^{t-1}\right]$ and $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$. Using this on symmetric AMP, we get :

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \left(\mathbf{A} - \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}} + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\right)\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.8}$$

$$= \left(\mathbf{A} - \left(Id - \mathbf{P}_{\mathbf{M}_{t-1}}\right)\mathbf{A}\left(Id - \mathbf{P}_{\mathbf{M}_{t-1}}\right)\right)\mathbf{m}^t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.9}$$

$$= \left(\mathbf{A}\mathbf{P}_{\mathbf{M}_{t-1}} + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{P}^T_{\mathbf{M}_{t-1}}\right)\mathbf{m}^t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.10}$$

$$= \mathbf{A}\mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1} \tag{12.11}$$

assuming $\mathbf{M}_{t-1}$ has full rank, defining (unique way) $\boldsymbol{\alpha}_t$ as the coefficients of the projection of $\mathbf{m}^t$ onto the columns of $\mathcal{M}_{t-1}$, we have :

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \mathbf{A}\mathbf{M}_{t-1}\boldsymbol{\alpha}_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1} \tag{12.12}$$

using the definition of the symmetric AMP iteration, we have $\mathbf{A}\mathbf{M}_{t-1} = \mathbf{X}_{t-1} + [0|\mathbf{M}_{t-2}]\,\mathbf{B}_t$ where $\mathbf{X}_{t-1} = \left[\mathbf{x}^1|...|\mathbf{x}^t\right]$ and $\mathbf{B}_t$ is the diagonal matrix of Onsager terms. Then:

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \left(\mathbf{X}_{t-1} + [0|\mathbf{M}_{t-2}]\,\mathbf{B}_t\right)\boldsymbol{\alpha}_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1} \tag{12.13}$$

$$= \underbrace{\mathbf{X}_{t-1}\boldsymbol{\alpha}_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t}_{Part\ 1} + \underbrace{[0|\mathbf{M}_{t-2}]\,\mathbf{B}_t\boldsymbol{\alpha}_t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1}}_{Part\ 2} \tag{12.14}$$

It is clear that Part 1 in the above expression is a combination of previous terms with an additional Gaussian one (product of independent GOE with frozen Ms). Part 2 cancels out in high-dimensional limit, isometry+Stein's lemma, AKA Onsager magic. Then recursion intuitively gives Gaussian equivalent model at each and across iterations.

Let us show how to cancel part 2! We shall focus on the term

$$\mathcal{A} = \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} \tag{12.15}$$

$$= \mathbf{M}_{t-1}(\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}\mathbf{M}^T_{t-1}\mathbf{A}\mathbf{m}^t_{\perp} \tag{12.16}$$

$$= \mathbf{M}_{t-1}(\mathbf{M}^T_{t-1}\mathbf{M}_{t-1})^{-1}(\mathbf{A}\mathbf{M}_{t-1})^T\mathbf{m}^t_{\perp} \tag{12.17}$$

But $(\mathbf{A}\mathbf{M}_{t-1})^T = (\mathbf{X}_{t-1} - [0|\mathbf{M}_{t-2}]\,\mathbf{B}_t)^T$ so that

$$(\mathbf{A}\mathbf{M}_{t-1})^T\mathbf{m}_\perp^t = (\mathbf{X}_{t-1}^T - \mathbf{B}_t^T\,[0|\mathbf{M}_{t-2}]^T)\mathbf{m}_\perp^t = \mathbf{X}_{t-1}^T\mathbf{m}_\perp^t \tag{12.18}$$

so that

$$\begin{aligned}
\mathcal{A} = & \quad \mathbf{M}_{t-1}(\mathbf{M}_{t-1}^T\mathbf{M}_{t-1})^{-1}\mathbf{X}_{t-1}^T\mathbf{m}_\perp^t & (12.19)\\
= & \quad \mathbf{M}_{t-1}(\mathbf{M}_{t-1}^T\mathbf{M}_{t-1})^{-1}\mathbf{X}_{t-1}^T(\mathbf{m}^t - \mathbf{m}_\parallel^t) & (12.20)\\
= & \quad \mathbf{M}_{t-1}(\mathbf{M}_{t-1}^T\mathbf{M}_{t-1})^{-1}\mathbf{X}_{t-1}^T(f^t(\mathbf{x}^t) - \mathbf{M}^{t-1}\alpha^t) & (12.21)
\end{aligned}$$

Here we need to be a bit carefful with the scaling of these terms. Remember than this should be a vector with order one compotent. To underline this, we can write:

$$\mathcal{A} = \quad \left(\mathbf{M}_{t-1}(\mathbf{M}_{t-1}^T\mathbf{M}_{t-1})^{-1}N\right)\left(\frac{1}{N}\mathbf{X}_{t-1}^T(f^t(\mathbf{x}^t) - \mathbf{M}^{t-1}\alpha^t)\right) \tag{12.22}$$

Now we see that we have a matrix with $O(1)$ element that multiply vectors with $O(1)$ elements, and when we multiply them we have a again an $O(1)$ quantities (since we sum over t values, and t is finite). Let us focus on the two terms in the second parenthesis:

$$\mathcal{B} = \quad \frac{1}{N}\mathbf{X}_{t-1}^T(f^t(\mathbf{x}^t) - \mathbf{M}^{t-1}\alpha^t) = \mathcal{C} + \mathcal{D} \tag{12.23}$$

These terms can be simplifies using Stein lemma. Indeed

$$\begin{aligned}
\mathcal{C} = & \quad \frac{1}{N}\mathbf{X}_{t-1}^T f^t(\mathbf{x}^t) & (12.24)\\
= & \quad \begin{bmatrix} \frac{1}{N}\sum_i x_i^{(1)} f(x_i^{(t)}) \\ \frac{1}{N}\sum_i x_i^{(2)} f(x_i^{(t)}) \\ \cdots \\ \frac{1}{N}\sum_i x_i^{(t)} f(x_i^{(t)}) \end{bmatrix} & (12.25)\\
=_{N\to\infty} & \quad \begin{bmatrix} \mathbb{E}[z^1 f(z^{(t)})] \\ \mathbb{E}[z^2 f(z^{(t)})] \\ \cdots \\ \mathbb{E}[z^t f(z^{(t)})] \end{bmatrix} & (12.26)\\
& & (12.27)
\end{aligned}$$

where we have used the recurence hypothesis. Now we can use Stein lemma[1] and we find

$$\mathcal{C} = \quad \begin{bmatrix} \kappa_{1,t}\mathbb{E}[f'(z^{(t)})] \\ \kappa_{2,t}\mathbb{E}[f'(z^{(t)})] \\ \cdots \\ \kappa_{t,t}\mathbb{E}[f'(z^{(t)})] \end{bmatrix} = \quad b_t \begin{bmatrix} \kappa_{1,t} \\ \kappa_{2,t} \\ \cdots \\ \kappa_{t,t} \end{bmatrix} \tag{12.28}$$

But we also have that

$$\kappa_{s,t} = \frac{1}{N}\mathbf{m}^{s-1}\mathbf{m}^{t-1} \tag{12.29}$$

[1] Stein's lemma: Suppose X is a normally distributed random variable with and variance κ. Then $E\big(g(X)(X)\big) = \kappa E\big(g'(X)\big)$.. In general, suppose X and Y are jointly normally distributed. Then $\mathrm{Cov}(g(X), Y) = \mathrm{Cov}(X, Y)E(g'(X))$.

and therefore

$$\mathcal{C} = \frac{1}{N} b_t \begin{bmatrix} (\mathbf{m}^0)^T \mathbf{m}^{t-1} \\ (\mathbf{m}^1)^T \mathbf{m}^{t-1} \\ \dots \\ (\mathbf{m}^{t-1})^T \mathbf{m}^{t-1} \end{bmatrix} = \frac{1}{N} b_t \mathbf{M}_{t-1}^T \mathbf{m}^{t-1} \tag{12.30}$$

We can deal in a similar way with $\mathcal{D}$ and we find

$$\mathcal{D} = \frac{1}{N} \mathbf{M}_{t-1}^T \left[0 | \mathbf{M}_{t-2}\right] \mathbf{B}_t \boldsymbol{\alpha}_t \tag{12.31}$$

so that finally

$$\mathcal{A} = \left(\mathbf{M}_{t-1} (\mathbf{M}_{t-1}^T \mathbf{M}_{t-1})^{-1} N\right) \left(\frac{1}{N} \mathbf{X}_{t-1}^T (f^t(\mathbf{x}^t) - \mathbf{M}^{t-1} \alpha^t)\right) \tag{12.32}$$

$$= \mathbf{M}_{t-1} (\mathbf{M}_{t-1}^T \mathbf{M}_{t-1})^{-1} \mathbf{M}_{t-1}^T \left(b_t \mathbf{m}^{t-1} - \left[0 | \mathbf{M}_{t-2}\right] \mathbf{B}_t \boldsymbol{\alpha}_t\right) \tag{12.33}$$

$$= \mathbf{P}_{\mathbf{M}_{t-1}} \left(b_t \mathbf{m}^{t-1} - \left[0 | \mathbf{M}_{t-2}\right] \mathbf{B}_t \boldsymbol{\alpha}_t\right) \tag{12.34}$$

This is precisely the part needed to cancel part 2! At this point, we have proven the claim.

Additionally, we can now write the evolution of the κ_t, or equivalently, of the q_t. This is called state evolution!

$$q_t = \frac{1}{N} \mathbf{m}^t \mathbf{m}^t = \mathbb{E}\left[\left(f_t(\sqrt{\kappa_t} Z)\right)^2\right] = \mathbb{E}\left[\left(f_t(\sqrt{q_{t-1}} Z)\right)^2\right] \tag{12.35}$$

What is interesting here is the generality of the statement. In fact, f not even need to be separable or even well defined for it to be true Berthier et al. (2020). This allows to use such iteration with neural networks **?**Gerbelot and Berthier (2021) and black-box functions in signal processing **?**.

12.1.3 Convergence of AMP

A crucial question at this point can be asked about whether or not x^t (or m^t) does converge. This can be studied by looking at

$$\frac{1}{N} \left|\mathbf{m}^{t+1} - \mathbf{m}^t\right|_2^2 = q^{t+1} + q^t - 2q^{t,t+1} \tag{12.36}$$

If we assume that we have a convergence in q (but not necessary in x!!) such that we are on the orbit of AMP are $q^{t+1} = q^t = q^*$, then it amonts to wther or not $q^{t,t+1}$ is convergeing to q^*. We can write

$$C^{t+1} = q_{t+1,t} = \frac{1}{N} \mathbf{m}^{t+1} \mathbf{m}^t = \mathbb{E}\left[\left(f_t(X^{t+1})\right)\left(f_t(X^t)\right)\right] \tag{12.37}$$

Using our Gaussian equivalance, we can now replcae the two Xs by correlated gaussians! We have:

$$X^t = \sqrt{\kappa_t - \kappa_{t,s}} Z' + \sqrt{\kappa_{t,s}} Z \tag{12.38}$$

$$X^s = \sqrt{\kappa_s - \kappa_{t,s}} Z'' + \sqrt{\kappa_{t,s}} Z \tag{12.39}$$

so that at the fixed point when $\kappa_s = \kappa_t = q^*$ we have

$$\begin{aligned} C^{t+1} &= \mathbb{E}\left[\left(f_t(\sqrt{\kappa_{t+1} - \kappa_{t+1,t}}Z' + \sqrt{\kappa_{t+1,t}}Z)\right)\left(f_t(\sqrt{\kappa_t - \kappa_{t+1,t}}Z'' + \sqrt{\kappa_{t+1,t}}Z)\right)\right] && (12.40)\\ &= \mathbb{E}\left[\left(f_t(\sqrt{q^* - C^t}Z' + \sqrt{C^t}Z)\right)\left(f_t(\sqrt{q^* - C^t}Z'' + \sqrt{C^t}Z)\right)\right] && (12.41)\\ &= g(C^t) && (12.42)\end{aligned}$$

The question is thus is this fixed-point iteration of g, defined for 0 to q^* is converging. Accoriding to the fixed point theorem, this will be the case if the map is contrative, that is if for any x and y, we have

$$\|g(x) - g(y)\| \leq |x - y| \tag{12.43}$$

In other words, we need g to be 1-Lipschitz. One can show that the first two derivative of g are positive (takes the derivative and use Stein lemma!), so that the derivative is maximum at q^*! We thus obtain, after a quick computation, the criterion:

$$\mathbb{E}\left[\left(f_t'(\sqrt{q^*}Z)\right)^2\right] \leq 1 \tag{12.44}$$

12.1.4 A first application: the TAP equation!

Consider how the function

$$f_t(x) = \tanh(\beta x + h) \tag{12.45}$$

In this case, we see that AMP simplifies as follows:

$$b_t = \beta(1 - \mathbb{E}[\tanh(\beta \mathbf{x}^t + h)^2]) = \beta(1 - q^t) \tag{12.46}$$

and

$$\begin{aligned} \mathbf{x}^{t+1} &= \mathbf{A}\mathbf{m}^t - \beta(1 - q^t)\mathbf{m}^{t-1} && (12.47)\\ \mathbf{m}^{t+1} &= \tanh(\beta \mathbf{x}^{t+1} + \beta h) && (12.48)\end{aligned}$$

so that

$$\mathbf{m}^{t+1} = \tanh(\beta\left(\mathbf{A}\mathbf{m}^t - \beta(1 - q^t)\mathbf{m}^{t-1}\right) + \beta h) \tag{12.49}$$

There are the TAP equations Thouless et al. (1977) for the SK model, and q obbey the parisi replica symmetric equation:

$$q_t = \mathbb{E}\left[(\tanh(\beta\sqrt{q_{t-1}}Z + \beta h))^2\right] \tag{12.50}$$

Additionally, we also can have convergence when

$$\mathbb{E}\left[\left(\tanh'(\beta\sqrt{q^*}Z + \beta h)\right)^2\right] \leq 1 \tag{12.51}$$

which is precisely the Almeida-Thouless criterion **?**. This was the main results of the original paper by Bolthausen (2014).

12.1.5 Parisi equation reloaded: Approximate Survey propagation

Since the Parisi RS's equation of the SK model seems to appear, we may wonder, are the RSB equation a state evolution of something? The answer is yes, as shown in **?**. Define in full generality the function

$$f_t(x) = \frac{\mathbb{E}_V \left[\cosh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)^m \tanh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)\right]}{\mathbb{E}_V \left[\cosh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)^m\right]} \tag{12.52}$$

$$q_1^{t+1} = \mathbb{E}_U \left[\frac{\mathbb{E}_V \left[\cosh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)^m \tanh^2\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)\right]}{\mathbb{E}_V \left[\cosh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)^m\right]}\right] \tag{12.53}$$

$$q_0^{t+1} = \mathbb{E}_U \left[\left(\frac{\mathbb{E}_V \left[\cosh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)^m \tanh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)\right]}{\mathbb{E}_V \left[\cosh\left(\beta x + \beta\sqrt{q_1^t - q_0^t}V + \beta h\right)^m\right]}\right)^2\right] \tag{12.54}$$

then the AMP iteration follows has a state evolution that follows Parisi's 1RSB equations, in particular, the norm of f_t is q_0!

We thus see that the Parisi equation can be interpreted as a way to follow some iterative algorithm in time.

12.2 Applications

12.2.1 Wigner Spike Model

Consider the following recursion:

$$\mathbf{x}^{t+1} = \left(\mathbf{A} + \frac{\sqrt{\lambda}}{N}\mathbf{x}_*\mathbf{x}_*^T\right)\mathbf{m}^t - b_t\mathbf{m}^{t-1} = Y\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.55}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{12.56}$$

How do we deal with this? Let us make a change of variable! We write simply

$$\mathbf{x}^{t+1} = A\mathbf{m}^t - b_t\mathbf{m}^{t-1} + \sqrt{\lambda}\mathbf{x}_* q_0^t \tag{12.57}$$

with

$$\tilde{\mathbf{x}}^{t+1} = A\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.58}$$

$$q_0^t = \frac{1}{N}\mathbf{x}_*^T\mathbf{m}^t \tag{12.59}$$

with this change of variable, we can write, defining $\tilde{f}_t(\mathbf{x}) = f_t(\mathbf{x} + \mathbf{x}_0 q_0^t)$ and

$$\tilde{\mathbf{x}}^{t+1} = \mathbf{A}\mathbf{m}^t - b_t \mathbf{m}^{t-1} \tag{12.60}$$

$$\mathbf{m}^t = \tilde{f}_t(\tilde{\mathbf{x}}^t) \tag{12.61}$$

so that we can use state evolution for the tilde variable! This leads again to

$$q_t = \frac{1}{N}\mathbf{m}^t\mathbf{m}^t = \mathbb{E}\left[\left(\tilde{f}_t(\sqrt{q_{t-1}}Z)\right)^2\right] \tag{12.62}$$

$$= \mathbb{E}\left[\left(f_t(\sqrt{q_{t-1}}Z + \sqrt{\lambda}q_0^{t-1}X^*)\right)^2\right] \tag{12.63}$$

and from the definition of q_0^t, we have the jointed state evolution:

$$q^t = \mathbb{E}\left[\left(f_t(\sqrt{q_{t-1}}Z + \sqrt{\lambda}q_0^{t-1}X^*)\right)^2\right] \tag{12.64}$$

$$q_0^t = \mathbb{E}\left[\left(f_t(\sqrt{q_{t-1}}Z + \sqrt{\lambda}q_0^{t-1}X^*)\right)X_*\right] \tag{12.65}$$

This is very nice! Let's look a concrete example: the **Rademacher spike model!**

Say we are given $Y = \frac{1}{\sqrt{N}}\left(\mathbf{W} + \sqrt{\frac{\lambda}{N}}\mathbf{x}_*\mathbf{x}_*^T\right)$, with $x_*^i = \pm 1$. Can we recover the vector ? We can use the AMP algorithm for this,

$$\mathbf{x}^{t+1} = Y\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.66}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{12.67}$$

using the binary denoiser, that is using $f_t(x) = \tanh\sqrt{\beta}x$. This is a particular case of the TAP equation (this is often called the planted SK model). In this case we have the state evolution:

$$q^t = \mathbb{E}\left[\left(\tanh(\beta\sqrt{q_{t-1}}Z + \beta\sqrt{\lambda}q_0^{t-1}X^*)\right)^2\right] \tag{12.68}$$

$$q_0^t = \mathbb{E}\left[\left(\tanh(\beta\sqrt{q_{t-1}}Z + \beta\sqrt{\lambda}q_0^{t-1}X^*)\right)X_*\right] \tag{12.69}$$

These are the one of the SK model with a ferromagnetic bias (λ is often called J_0 in this context).

12.2.2 A primer on Bayesian denoising

Why did we use the hyperbolic tangent? This has to do with Bayesian methods in statistics. After all, we know that we are given X, and that $X = q_0^{t-1}\sqrt{\lambda}X_* + \sqrt{q^{t-1}}Z$. Given we know X is just $X0$ up to a Gaussian noise, what is the best we can do to estimates it?

It is a general rule that the "best" estimates in Bayesian method is the mean of the posterior, in terms of MMSE. This is easily proven: assume that we have an observable Y which is given by a polluted version of X_*. Then for all estimators we have:

$$\text{MMSE} = \mathbb{E}_{X_*,Y}\left[(\hat{X}(Y) - X_*)^2\right] = \mathbb{E}_Y\left[\mathbb{E}_{X^*|Y}\left[(\hat{X}(Y) - X_*)^2\right]\right] \tag{12.70}$$

Let us minimize it! COnditioned on Y, $\hat{X}$ is just a number so we can derive with respect to it and find that for each Y, we should choose

$$\hat{(X)} = \mathbb{E}_{X^*|Y} X^* = \int P(X|Y)X \tag{12.71}$$

This is the Bayesian optimal! In our problem, this is easyly done: Given we are given an observable $X = q_0^{t-1}\sqrt{\lambda}X_* + \sqrt{q^{t-1}}Z$, we see that

$$P(X^*|X) \propto \qquad P^{\text{prior}}(X)P^{\text{X|X}^*}(X) \propto P^{\text{prior}}(X)e^{-\frac{(X-\sqrt{\lambda}q_0^{t+1}X^*)^2}{2q^{t-1}}} \tag{12.72}$$

$$\mathbb{E}[X^*|X] = \qquad \frac{\int dx x P^{\text{prior}}(x)e^{-\frac{(X-\sqrt{\lambda}q_0^{t+1}x)^2}{2q^{t-1}}}}{\int dx P^{\text{prior}}(x)e^{-\frac{(X-\sqrt{\lambda}q_0^{t+1}x)^2}{2q^{t-1}}}} \tag{12.73}$$

This is called the Optimal Bayesian Gaussian Denoiser.

Another simplification can be done, using the so called Nishimori property:

Theorem 17 (Nishimori Identity)**.** *Let $X^{(1)}, \dots, X^{(k)}$ be k i.i.d. samples (given Y) from the distribution $P(X = \cdot \,|\, Y)$. Denoting $\langle\cdot\rangle$ the "Boltzmann" expectation, that is the average with respect to the $P(X = \cdot \,|\, Y)$, and $\mathbb{E}[\cdot]$ the "Disorder" expectation, that is with respect to (X^*, Y). Then for all continuous bounded function f we can switch one of the copies for X^*:*

$$\mathbb{E}\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{(k)}\right)\right\rangle_k\right] = \mathbb{E}\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{*}\right)\right\rangle_{k-1}\right] \tag{12.74}$$

Proof. The proof is a consequence of Bayes theorem and of the fact that both x^* and any of the copy $X^{(k)}$ are distributed from the posterior distribution. Denoting more explicitly the Boltzmann average over k copies for any function g as

$$\left\langle g(X^{(1)}, \dots, X^{(k)})\right\rangle_k =: \int \prod_{i=1}^{k} \mathrm{d}x_i P(x_i|Y) g(X^{(1)}, \dots, X^{(k)}) \tag{12.75}$$

we have, starting from the right hand side

$$\begin{aligned}
&\mathbb{E}_{Y,X^*}\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{*}\right)\right\rangle_{k-1}\right] \\
&= \int dx^* dy P(x^*|Y)P(Y)\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{*}\right)\right\rangle_{k-1} \\
&= \mathbb{E}_Y \int \mathrm{d}x^k P(x^k|y)\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{k}\right)\right\rangle_{k-1} \\
&= \mathbb{E}_Y\left[\left\langle f\left(Y, X^{(1)}, \dots, X^{(k-1)}, X^{(k)}\right)\right\rangle_k\right]
\end{aligned}$$

□

We shall drop the subset "k" from Boltzmann averages from now on.

Using a binary prior leads indeed to $f^(t) = \tanh \sqrt{\lambda} X$ (indeed we have only a factor $e^{x\sqrt{\lambda}q/qx}$

As far as we know, this is the best algorithm there is in polynomial times!

Sparse? : HARD PHASE: No algo known!

12.2.3 Wishart Spike Model

Given two vectors $\mathbf{u}_* \in \mathbb{R}^n$ and $\mathbf{v}_* \in \mathbb{R}^m$ we are given the $n \times m$ matrix (with $\alpha = m/n$):

$$W = \sqrt{\frac{\lambda}{n}} \mathbf{u}_* \mathbf{v}_*^T + \xi \tag{12.76}$$

The AMP algorithm **?** reads (conveniently removing the fact that variances can be pre-computed so that the functions f is defined at each time t):

$$\mathbf{u}^{t+1} = \frac{1}{\sqrt{n}} W g_v^t(\mathbf{v}^t) - \alpha < g_v^{t\prime}(\mathbf{v}) > g_u^{t-1}(\mathbf{u}^{t-1}) \tag{12.77}$$

$$\mathbf{v}^{t+1} = \frac{1}{\sqrt{n}} W g_u^t(\mathbf{u}^t) - < g_u^{t\prime}(\mathbf{u}^t) > g_v^{t-1}(\mathbf{v}^{t-1}) \tag{12.78}$$

Conveniently, one can recast these equations into a new form, that is amenable to a rigorous analysis. Define the $n + m \times 2$ matrix X

$$\mathbf{x} = \begin{pmatrix} \mathbf{u} & 0 \\ 0 & \mathbf{v} \end{pmatrix} \tag{12.79}$$

and the interaction matrix

$$Y = \sqrt{\frac{\lambda}{n}} \mathbf{x}_* \mathbf{x}_*^T + Z = \sqrt{\frac{\lambda}{n}} \begin{pmatrix} \mathbf{u}_* \mathbf{u}_*^T & \mathbf{u}_* \mathbf{v}_*^T \\ (\mathbf{u}_* \mathbf{v}_*^T)^T & \mathbf{v}_* \mathbf{v}_*^T \end{pmatrix} + \begin{pmatrix} \xi_{up} & \xi \\ \xi^T & \xi_{low} \end{pmatrix} \tag{12.80}$$

Now we define $f^t : \mathbb{R}^{m+n\times 2} \to \mathbb{R}^{m+n\times 2}$ defined as each time t by the following rule at each lines (with matrix-like notations):

$$\text{for } i = 1 \dots n \quad f_i = [0 \;\; g_u^t(x[i,1])] \tag{12.81}$$

$$\text{for } i = n+1 \dots m+n \quad f_i = [g_v^t(x[i,2]) \;\; 0] \tag{12.82}$$

$$\tag{12.83}$$

Then the following AMP recursion is equivalent to eq.(12.77,12.78):

$$x^{t+1} = \frac{Y}{\sqrt{n}} f^t(x^t) - \frac{1}{n} < f^{t\prime}(x^t) > f^{t-1}(X^{t-1}) \tag{12.84}$$

$$= \frac{Y}{\sqrt{n}} f^t(x^t) - b_t f^{t-1}(x^{t-1}) \tag{12.85}$$

where we have put the Onsager terms into the moniker b_t. We thus obtain the state evolution equation for eq.(12.77,12.78) directly from the Wigner ones:

$$q_u^t = \mathbb{E}\left[\left(g_u^t(\sqrt{\alpha q_v^{t-1}}Z + \alpha\sqrt{\lambda}q_v^{0^{t-1}}V^*)\right)^2\right] \tag{12.86}$$

$$q_u^{0^t} = \mathbb{E}\left[\left(g_u^t(\sqrt{\alpha q_v^{t-1}}Z + \alpha\sqrt{\lambda}q_v^{0^{t-1}})V^*\right)U_*\right] \tag{12.87}$$

$$q_v^t = \mathbb{E}\left[\left(g_v^t(\sqrt{q_u^{t-1}}Z + \sqrt{\lambda}q_u^{0^{t-1}}U^*)\right)^2\right] \tag{12.88}$$

$$q_v^{0^t} = \mathbb{E}\left[\left(g_v^t(\sqrt{{q_u}^{t-1}}Z + \sqrt{\lambda}q_u^{0^{t-1}}U^*)\right)V_*\right] \tag{12.89}$$

Again, they are many application (inclyding denoising spike models, analysing Mixture of Gaussians, Hopfield models, etc....)

if $r > 4 + 2\sqrt{\alpha}$, there is a gap between the threshold for informationtheoretically optimal performance and the threshold at which known algorithms succeed

12.3 From AMP to proofs of replica prediction: The Bayes Case

Let us start from the Bayes case! Suppose we are given as data a $n \times n$ symmetric matrix Y created as follows:

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}} \underbrace{\mathbf{x}^*\mathbf{x}^{*\intercal}}_{N\times N \text{ rank-one matrix}} + \underbrace{\xi}_{\text{symmetric iid noise}}$$

where $\mathbf{x}^* \in \mathbb{R}^N$ with $x_i^* \overset{\text{i.i.d.}}{\sim} P_X(x)$, $\xi_{ij} = \xi_{ji} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i \leq j$.

This is called the Wigner spike model in statistics. The name "Wigner" refer to the fact that Y is a Wigner matrix (a symmetric random matrix with component sampled randomly from a Gaussian distribution) plus a "spike", that is a rank one matrix $\mathbf{x}^*\mathbf{x}^{*\intercal}$.

We shall now make a mapping to a Statistical Physics formulation. Consider the spike-Wigner model, using Bayes rule we write:

$$\begin{aligned}
P(\mathbf{x}\,|\,\mathbf{Y}) = \frac{P(\mathbf{Y}\,|\,\mathbf{x})\,P(\mathbf{x})}{P(\mathbf{Y})} &\propto \left[\prod_i P_X(x_i)\right]\left[\prod_{i\leq j}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(y_{ij}-\sqrt{\frac{\lambda}{N}}x_ix_j\right)^2}\right] \\
&\propto \left[\prod_i P_X(x_i)\right]\exp\left(\sum_{i\leq j}\left[-\frac{\lambda}{2N}x_i^2x_j^2 + \sqrt{\frac{\lambda}{N}}y_{ij}x_ix_j\right]\right) \\
\Rightarrow \quad P(\mathbf{x}\,|\,\mathbf{Y}) &= \frac{1}{Z(\mathbf{Y})}\left[\prod_i P_X(x_i)\right]\exp\left(\sum_{i\leq j}\left[-\frac{\lambda}{2N}x_i^2x_j^2 + \sqrt{\frac{\lambda}{N}}y_{ij}x_ix_j\right]\right) \\
\Rightarrow \quad \hat{\mathbf{x}}_{\text{MSE}}(\mathbf{Y}) &= \begin{bmatrix}\hat{x}_{\text{MSE},1}(\mathbf{Y}) \\ \vdots \\ \hat{x}_{\text{MSE},N}(\mathbf{Y})\end{bmatrix}, \qquad \hat{x}_{\text{MSE},i}(\mathbf{Y}) = \langle x_i\rangle_{\mathbf{Y}} = \int \mathrm{d}\mathbf{x}\, P(\mathbf{x}\,|\,\mathbf{Y})\, x_i
\end{aligned}$$

Interestingly, we see that the partition function $Z(Y)$ is given by the ratio between

$$Z(Y) = \frac{P(Y)}{e^{-\sum_{ij} y_{ij}^2/2}/\sqrt{2\pi}^{N^2}} \tag{12.90}$$

This is the so-called likelihood ratio betweem the probablity that our model has been generated randomly, and the one that it has been generated by accident from a pure random problem.

It is convenient to use the notation of information theory, and to compute instead the mutual information. It is defined as

$$I(X,Y) = \mathbb{E}_{X,Y} \log \frac{P(X,Y)}{P(X),P(Y)} = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{12.91}$$

In our model, the relation between the free energy and the mutual information is trivial to compute, one simply finds:

$$\frac{I(X,Y)}{N} = f + \frac{\lambda(\mathbb{E}[X^2])^2}{4} \tag{12.92}$$

Why work with mutual information since it is just the free energy ? It has actually nice properties that we now list:

- The mutual information is monotonic in λ from 0 to $H(X)$ (the first is trivial from the definition, since then the joint distribution factorize, while the second comes from the fact that if we recover X perfectly from Y, then $H(Y|X) = 0$).
- Call the matrix $M = X^*X$, it is easy to show using Bayesian technics that the best possible error in reconstructing M from Y is given by the derivative of the mutual information. This is the I-MMSE theorem:

$$\mathrm{M-MMSE} = \frac{1}{4}\frac{\partial I(\lambda)}{\partial \lambda} \tag{12.93}$$

Additionally, we see that the model looks like a SK problem, where the role of the inverse temperature is played by $\sqrt{\lambda}$ This is often called the Nishimori condition in the litterature. It is a very peculiar condition, in fact we can show that $q = m$ if we impose this!

Using the replica method, one can work hard and derive the free energy! It reads (here $q = m$)

$$f(m) \approx \frac{\lambda m^2}{4} - \mathbb{E}_{x^*,z} \log \int dx_0 P(x_0) e^{-x_0^2 \frac{\lambda m}{2} + \lambda x_0 x_0^* m + x_0 x_0^* \sqrt{\lambda m} z} \tag{12.94}$$

Notice that we also find an interesting identity : $df_{\mathrm{rs}}/d\lambda = 4m^2/4$, so that the replica prediction for the mutual information is just

$$f_{rs} + \lambda\rho^2/4 \tag{12.95}$$

This is interesting, can we proove it ?????

Let us notice that AMP cannot be better than the MMSE, so let us use the following estimator:

$$\mathrm{AMP-M} = \mathrm{m_i m_j} \tag{12.96}$$

What is the error of this estimator? This is easily computed

$$\mathrm{AMP-MSE} = \frac{1}{\mathrm{N}^2}\sum_{\mathrm{i}}(\mathrm{f_t(X_i)f_t(X_j)} - \mathrm{X^i_* X^j_*}) \tag{12.97}$$

In expectation, this is simply

$$\mathrm{AMP-MSE} = \rho^2 - \mathrm{m}^2 \tag{12.98}$$

For sure, this error is larger, or equal to the Bayes one, so

$$\mathrm{AMP-MSE}(\lambda) > \mathrm{MMSE}(\lambda) \tag{12.99}$$

so

$$\int d\lambda \frac{1}{4}\mathrm{AMP-MSE}(\lambda) > \int \frac{1}{4}\mathrm{MMSE}(\lambda) = \mathrm{I}(\infty) - \mathrm{I}(0) = \mathrm{H(X)} \tag{12.100}$$

Now, it turns out that the derivative of the replica symmetric free energy is just AMP-MSE! Indeed the derivative of the replica mutual information is just

$$-\frac{1}{4}m^2 + \frac{1}{4}\rho^2 + \partial_\lambda T \tag{12.101}$$

but the derivative of the free energy gives the correct term and we reach indeed the MSE at the end! Imagine we can integrate without discontinuity (that is, m is continuous) then:

$$I_{\mathrm{replica}}(\infty) - I_{\mathrm{replica}}(0) = \int d\lambda \frac{1}{4}\mathrm{AMP-MSE}(\lambda) > \int \frac{1}{4}\mathrm{MMSE}(\lambda) = \mathrm{I}(\infty) - \mathrm{I}(0) = \mathrm{H(X)} \tag{12.102}$$

but (defining $E = \rho - m$) and $\Sigma = 1/\lambda m$ we find

$$I_{\mathrm{replica}}(\lambda) = \frac{\lambda(\rho^2+m^2)}{4} - \mathbb{E}_{x^*,z}\log\int dx_0 P(x_0) e^{-x_0^2\frac{\lambda m}{2} + \lambda x_0 x_0^* m + x_0 x_0^*\sqrt{\lambda m}z} = I(X; X+\Sigma Z) + \frac{\lambda E^2}{4} \tag{12.103}$$

But clearly, we see tghat $I(0) = 0$ and $I(\infty) = H(x)$ so we reach

$$H(X) = \int d\lambda \frac{1}{4}\mathrm{AMP-MSE}(\lambda) > \int \frac{1}{4}\mathrm{MMSE}(\lambda) = \mathrm{H(X)} \tag{12.104}$$

So we must have AMP reaching the MMSE almost everywhere and the free energy to be correct!

12.4 From AMP to proofs of replica prediction: The MAP case

Consider again our favorite problem, aka the spike model!

Say we are given $Y = \frac{1}{\sqrt{N}}\left(\mathbf{W} + \sqrt{\frac{\lambda}{N}}\mathbf{x}_*\mathbf{x}_*^T\right)$, and we do not know anything about X_*, except that it is well normalized (on the sphere). A typical thing to do would be to use maximum

likelihood. We want to minimize $\|Y - \sqrt{\lambda}xx^T\|_2^2$ constainted to X being on the sphere. This can be done using the following cost fucntion

$$\mathcal{L} = -\sum_{i,j} x_i Y_{ij} x_j + \mu(\sum_i x_i^2 - N)^2 \tag{12.105}$$

where we use μ as a Lagrange parameter to fix the norm. This problem is solved when the following condition is satisfied:

$$\sum_j Y_{ij} x_j = \mu x_i \tag{12.106}$$

or equivalently when

$$Y\hat{\mathbf{x}} = \mu\hat{\mathbf{x}} \tag{12.107}$$

that is, of course, when $\mathbf{x}$ is an eigenvector. Them, the loss is minimized using the largest possible eigevnvalue.

Let us see how we can use AMP for this! We simply use the function that will preserve the norm and write

$$\mathbf{x}^{t+1} = \left(\mathbf{A} + \frac{\sqrt{\lambda}}{N}\mathbf{x}_*\mathbf{x}_*^T\right)\mathbf{m}^t - b_t\mathbf{m}^{t-1} = Y\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{12.108}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{12.109}$$

with a non separable function written as

$$f_t(\mathbf{x}^t) = \mathbf{x}^t\frac{\sqrt{N}}{\|\mathbf{x}^t\|_2} \tag{12.110}$$

clearly, we can compute the Onsager term

$$b_t = \frac{\sqrt{N}}{\|\mathbf{x}^t\|_2} \tag{12.111}$$

Imagine that we iterate AMP and find a fixed point. Then what is this fixed point? Well, we have

$$\mathbf{x} = Y\,\mathbf{x}\frac{\sqrt{N}}{\|\mathbf{x}^t\|_2} - \frac{\sqrt{N}}{\|\mathbf{x}\|_2}\mathbf{x}\frac{\sqrt{N}}{\|\mathbf{x}\|_2} \tag{12.112}$$

$$\mathbf{x}\left(1 + \frac{N}{\|\mathbf{x}\|_2^2}\right) = Y\,\mathbf{x}\frac{\sqrt{N}}{\|\mathbf{x}\|_2} \tag{12.113}$$

$$\mathbf{m}\left(\frac{\|\mathbf{x}\|_2}{\sqrt{N}} + \frac{\sqrt{N}}{\|\mathbf{x}\|_2}\right) = Y\mathbf{m} \tag{12.114}$$

Clearly, this is the same fixed point as the one we are looking for! In fact, we even are given the value of the Lagrange parameter! So it seems that, if it converges, AMP can solved the problem for us! Now we can analyse the solution that AMP gives using state evolution! We have $\mathbf{x} = Z + q_0\sqrt{\lambda}\mathbf{x}_*$, so clearly $|\mathbf{x}|_2^2 = N(1 + q_0^2\lambda)$. Additionally we have

$$q_0 = \mathbb{E}\frac{1}{N}\left(\frac{Z + q_0\sqrt{\lambda}\mathbf{x_0}}{\sqrt{1 + q_0^2\lambda}}\mathbf{x_0}\right) = q_0\frac{\sqrt{\lambda}}{\sqrt{1 + q_0^2\lambda}} \tag{12.115}$$

From this, we can deduce that

$$q_0 = 0 \text{ when } \lambda < 1 \tag{12.116}$$

$$q_0 = \sqrt{1 - \frac{1}{\lambda}} \text{ when } \lambda > 1 \tag{12.117}$$

This is the BBP transition! Additionally, we can even compute the value of the eigenvalue! INdeed $|\mathbf{x}|_2^2 = N(1 + q_0^2\lambda)$ become $|\mathbf{x}|_2^2 = N\lambda$ so that

$$\lambda_{\max} = \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right) \tag{12.118}$$

This is given a generic strategy for prooving replica prediction: find an AMP that matches the fixed point of the minimum, prove that it converges, then simply study its state evolution prediction!

12.5 Rectangular AMP and GAMP

This is from Bayati and Montanari (2011); Rangan (2011), but the present discussion uses Berthier et al. (2020); Gerbelot and Berthier (2021).

There are many problem defined for rectangular non symmetric matrices $A \in \mathbb{R}^{m\times n}$, and for such, it will be interesting to look at the following iteration:

$$\mathbf{u}^{t+1} = A^T g_t(\mathbf{v}^t) - d_t e_f(\mathbf{u}^t) \tag{12.119}$$

$$\mathbf{v}^t = A e_t(\mathbf{u}^t) - b_t g_{t-1}(\mathbf{v}^{t-1}) \tag{12.120}$$

with, assuming $\mathbb{E}[A_{ij}^2] = 1/n$ and $\alpha = m/n$,

$$d_t = \frac{\alpha}{m}\text{div} g_t(\mathbf{v}^t), b_t = \frac{1}{n}\text{div} g_t(\mathbf{u}^t) \tag{12.121}$$

It would be great to have a state evolution for those equations, so that $\mathbf{u}$ and $\mathbf{v}$ can be treated as Gaussian at all times! This can be done by mapping (reducing) this recursion to the original symmetric one! Define:

$$A_s = \sqrt{\frac{1}{1+\alpha}} \begin{pmatrix} B & A \\ A^T & C \end{pmatrix} \text{ and } \mathbf{x^0} = \begin{pmatrix} 0 \\ \mathbf{u^0} \end{pmatrix} \tag{12.122}$$

and

$$f_{2t+1} = \sqrt{1+\alpha} \begin{pmatrix} g_t(x_1, \dots, x_m) \\ 0 \end{pmatrix}, \; f_{2t} = \sqrt{\frac{1+\alpha}{\alpha}} \begin{pmatrix} 0 \\ e_t(x_{m+1}, \dots, x_{m+n}) \end{pmatrix}. \tag{12.123}$$

and then we see, again, that $\mathbf{x^t}$ follows the regular AMP, equation. Making the change of variable, we see that both $\mathbf{u}$ and $\mathbf{v}$ are Gaussian variables, and we can compute their variances etc using state evolution,

A very important application of this rectangular AMP is the Approximate Message Passing for generalized linear problem, called Generalized AMP (GAMP in short), written slightly differently, as

$$\mathbf{u}^{t+1} = A^T g_t^{\text{out}}(\omega^t) + \Sigma_t^{-1} f_t^{\text{in}}(\mathbf{u}^t) \tag{12.124}$$

$$\omega^t = A f_t^{\text{in}}(\mathbf{u}^t) - V_t g^{\text{out}}(\omega^{t-1}) \tag{12.125}$$

with V_t and $-\Sigma_t^{-1}$ playing the role of the Onsager term, and where the state evolution reads so that

$$\mathbf{u}^t = \sqrt{q_u^t} Z \, , \omega^t = \sqrt{q_\omega^t} Z' \tag{12.126}$$

$$q_u^{t+1} = \alpha \mathbb{E}[(g_t^{\text{out}}(\sqrt{q_\omega^t} Z))^2] \, , q_\omega^t = \mathbb{E}[f_t^{\text{in}}(\sqrt{q_u^t} Z))^2] \tag{12.127}$$

In practice, however, since we are interested in cases where there is a hidden $\mathbf{x_0}$ and a set of measurement given by a function of $\mathbf{z} = A\mathbf{x_0}$, we need to be slightly more generic (in the same as in the wigner spike model). In fact, we want to use function g tht depends potentially on $\mathbf{z}$ (through $\mathbf{y}$ and function f tgat correlates with $\mathbf{x_0}$. In full generality GAMP reads as

$$\mathbf{u}^{t+1} = A^T g_t^{\text{out}}(\omega^t, \mathbf{Y}) + \Sigma_t^{-1} f_t^{\text{in}}(\mathbf{u}^t) \tag{12.128}$$

$$\omega^t = A f_t^{\text{in}}(\mathbf{u}^t) - V_t g^{\text{out}}(\omega^{t-1}, \mathbf{Y}) \tag{12.129}$$

and this add a new direction for u (that can get correlated with x_0) and for ω (that can get correlated with $\mathbf{z}_0$). In this case, the state evolution reads

$$q_u^{t+1} = \alpha \mathbb{E}[(g_t^{\text{out}}(\omega, z_0, \phi(z_0))^2] \tag{12.130}$$

$$q_\omega^t = \mathbb{E}[f_t^{\text{in}}(\sqrt{q_u^t} Z + m_u^t \mathbf{x_0})^2] \tag{12.131}$$

$$m_u^{t+1} = \alpha \mathbb{E}[\partial_{z_0}(g_t^{\text{out}}(\omega, \phi(z_0))] \tag{12.132}$$

$$m_\omega^t = \mathbb{E}[f_t^{\text{in}}(\sqrt{q_u^t} Z + m_u^t \mathbf{x_0}) x_0] \tag{12.133}$$

where the equation on m_u^{t+1} comes from applying the Stein lemma to

$$m_u^{t+1} = \alpha \mathbb{E}[z_0(g_t^{\text{out}}(\omega, \phi(z_0))] = \alpha \mathbb{E}[\partial_{z_0}(g_t^{\text{out}}(\omega, \phi(z_0))] \tag{12.134}$$

and where z, ω have a joint Gaussian distribution with covariance $\begin{pmatrix} \rho & m^t \\ m^t & q^t \end{pmatrix}$

12.6 Learning problems: the LASSO

12.6.1 Gradient Descent

We start with a initial guess x^0 and we iterate to tend to the minimum of the function. At each iteration, the next approximation of x is given by:

$$x^{t+1} = x^t - \eta f'(x^t)$$

To find the next value x^{t+1}, the function is approximated around the current point using a quadratic approximation and x^{t+1} is given by the minimum of the quadratic approximation.

By the Taylor's theorem, we have:

$$f(x) \approx f(x^t) + (x - x^t)f'(x^t) + \frac{(x - x^t)^2}{2}f''(x^t)$$

By replacing $f''(x^t)$ by η^{-1}, that represents the learning rate, we have:

$$f(x) \approx f(x^t) + (x - x^t)f'(x^t) + \frac{(x - x^t)^2}{2\eta}$$

The minimum of this function is given by:

$$f'(x^t) + \frac{x^* - x^t}{\eta} = 0$$

$$\Rightarrow x^{t+1} \overset{\text{Def}}{=} x^* = x^t - \eta f'(x^t)$$

We see that we can write GD as the minimum, at each steps, of the approximate $f(x)$ cost function

12.6.2 Proximal Gradient Descent

Let $f(x)$ be a function we want to optimize, which is not derivable somewhere but can be decomposed as a sum of two functions, one derivable and the other not.

$$f(x) = g(x) + h(x)$$

$$\text{Where} \begin{cases} g(x) & \text{is derivable} \\ h(x) & \text{is not derivable, e.g. } |x| \end{cases}$$

Therefore, the gradient approximation can be done, up to the second derivative, only for $g(x)$. Since multiplying terms won't change the optimum value, the expression for the minimum x can be derived like so:

$$f(x) \simeq g(x^t) + (x - x^t)g'(x^t) + \frac{(x - x^t)^2}{2\eta} + h(x)$$

$$x^{t+1} = \text{argmin}\left[(x - x^t)g'(x^t) + \frac{(x - x^t)^2}{2\eta} + h(x)\right]$$

$$= \text{argmin}\left[h(x) + \frac{1}{2\eta}\left(x - x^t + g'(x^t)\eta\right)^2\right]$$

$$= \text{argmin} \left[h(x)\eta + \frac{1}{2} \left(x - (x^t - g'(x^t)\eta) \right)^2 \right]$$

If $h(x)$ is not very complicated, the former expression can be solved analytically (this is a fundamental object of study in Convex Optimization).

Definition: The proximal operator of a function.

$$\underset{l(.)}{Prox}(z) = \underset{x}{argmin} \left[l(x) + \frac{1}{2}(x - z)^2 \right]$$

Example: Soft Thresholding $l(x) = |x|a$

$$\underset{l(.)}{Prox}(z) = \begin{cases} z - a & \text{if} \quad z \geq a \\ 0 & \text{if} \quad z \in [-a;\, a] \\ z + a & \text{if} \quad z \leq -a \end{cases}$$

The recursion is then:

$$x^{t+1} = \underset{\eta h(.)}{Prox} \left[x^t - \eta g'(x^t) \right]$$

12.6.3 The LASSO and other linear problems

A very typical problem is the following one: find the argmin of

$$\mathcal{L}(w) = \frac{1}{2} \|y - A\mathbf{w}\|_2^2 + h(\lambda) \tag{12.135}$$

This can be solve, for convex h, using PGD, with the iteration:

The recursion is then:

$$\mathbf{w}^{t+1} = \underset{\eta h(.)}{Prox} \left[\mathbf{w}^t - \eta L'(x^t) \right] = \underset{\eta h(.)}{Prox} \left[\mathbf{w}^t + \eta A^T (y - Aw) \right]$$

12.6.4 AMP reloaded

Consider the following AMP recursion:

$$\mathbf{x}^{t+1} = f_t(\mathbf{x}^t + A^T r_t) \tag{12.136}$$

$$r_t = y - Ax + b_t r^{t-1} \tag{12.137}$$

Clearly, this is an instance of GAMP! Use $g_t(v^t) = y - v^t = r^t$ and $e(u) = Prox$, then GAMP reads

$$\mathbf{u}^{t+1} = A^T(\mathbf{y} - \mathbf{v}^t) + \text{Prox}(\mathbf{u}^t) \tag{12.138}$$

$$\mathbf{v}^t = A\text{prox}(\mathbf{u}^t) - b_t g_{t-1}(\mathbf{v}^{t-1}) \tag{12.139}$$

so using $x = prox(u)$ and $r = y - v$ we have

$$\mathbf{u}^{t+1} = A^T r_t + \mathbf{x} \tag{12.140}$$

$$r^t = y - \mathbf{v}^t = y - A\text{prox}(\mathbf{u}^t) + b_t g_{t-1}(\mathbf{v}^{t-1}) = y - A\mathbf{x} + b_t r \tag{12.141}$$

So state evolution can be written for the AMP considered! At the fixed point, if this AMP converges, we find

$$\mathbf{x} = f_t(\mathbf{x} + A^T(y - Ax)/(1 - b) \tag{12.142}$$

So this recursion solved the LASSO (or any convex optimization problem for that matter). Just takes its AMP recursion, and voila, you just solved LASSO!

Let us write the state evolution in the simplest case where $y = Ax_0$. Then we have $\mathbf{u}$ and $\mathbf{v}$ Gaussian and we need to track their variance and projection to $\mathbf{x}_0$ and $\mathbf{y}$.

$$\begin{aligned}
q_u^{t+1} &= \alpha\mathbb{E}[(g_t^{\text{out}}(\omega, \phi(z_0))^2] = \alpha\mathbb{E}[(y - w)^2] = \alpha(\rho + q_\omega - 2m_\omega) = \alpha E^t & (12.143)\\
q_\omega^t &= \mathbb{E}[f_t^{\text{in}}(\sqrt{q_u^t}Z + m_u^t\mathbf{x_0})^2] & (12.144)\\
m_u^{t+1} &= \alpha\mathbb{E}[\partial_{z_0}(g_t^{\text{out}}(\sqrt{q_\omega^t}Z' + m_\omega^t z_0, \phi(z_0))] = \alpha & (12.145)\\
m_\omega^t &= \mathbb{E}[f_t^{\text{in}}(\sqrt{q_u^t}Z + m_u^t\mathbf{x_0})x_0] & (12.146)
\end{aligned}$$

so that we can write, keeping only q and m

$$\begin{aligned}
q^{t+1} &= \mathbb{E}[f_t^{\text{in}}(\sqrt{\alpha(\rho + q^t - 2m^t)}Z + \alpha\mathbf{x_0})^2] & (12.147)\\
m^{t+1} &= \mathbb{E}[f_t^{\text{in}}(\sqrt{\alpha(\rho + q^t - 2m^t)}Z + \alpha\mathbf{x_0})\mathbf{x_0}] & (12.148)
\end{aligned}$$

or even more directly

$$E^{t+1} = \rho + \mathbb{E}[f_t^{\text{in}}(\sqrt{\alpha E^t}Z + \alpha\mathbf{x_0})^2] - 2\mathbb{E}[f_t^{\text{in}}(\sqrt{\alpha E^t}Z + \alpha\mathbf{x_0})\mathbf{x_0}]$$

12.7 Inference with GAMP

The inference problem addressed in this section is the following. We observe a M dimensional vector y_μ, $\mu = 1, \ldots, M$ that was created component-wise depending on a linear projection of an unknown N dimensional vector x_i, $i = 1, \ldots, N$ by a known matrix $F_{\mu i}$

$$y_\mu = f_{\text{out}}\left(\sum_{i=1}^{N} F_{\mu i} x_i\right), \tag{12.149}$$

where f_{out} is an output function that includes noise. Examples are the additive white Gaussian noise (AWGN) were the output is given by $f_{\text{out}}(x) = x + \xi$ with ξ being random Gaussian variables, or the linear threshold output where $f_{\text{out}}(x) = \text{sign}(x - \kappa)$, with κ being a threshold value. The goal in linear estimation is to infer $\mathbf{x}$ from the knowledge of $\mathbf{y}$, F and f_{out}. An alternative representation of the output function f_{out} is to denote $z_\mu = \sum_{i=1}^{N} F_{\mu i} x_i$ and talk about the likelihood of observing a given vector $\mathbf{y}$ given $\mathbf{z}$

$$P(\mathbf{y}|\mathbf{z}) = \prod_{\mu=1}^{M} P_{\text{out}}(y_\mu|z_\mu)\,. \tag{12.150}$$

One uses

$$g_{\rm out}(\omega, y, V) \equiv \frac{\int {\rm d}z P_{\rm out}(y|z)\,(z-\omega)\,e^{-\frac{(z-\omega)^2}{2V}}}{V\int {\rm d}z P_{\rm out}(y|z) e^{-\frac{(z-\omega)^2}{2V}}}\,. \tag{12.151}$$

and (using $R = \mathbf{u} * \Sigma$)

$$f_{\rm in}(\Sigma, R) = \frac{\int {\rm d}x\, x\, P_X(x)\, e^{-\frac{(x-R)^2}{2\Sigma}}}{\int {\rm d}x\, P_X(x)\, e^{-\frac{(x-R)^2}{2\Sigma}}}\,, \qquad f_v(\Sigma, R) = \Sigma \partial_R f_a(\Sigma, R)\,. \tag{12.152}$$

Again, these are not new equations! This was written as early as in 1989 by Mézard as the TAP equation for the so-called perceptron problem **?**. The same equation were written also by Kabashima **?** and Krzakala et al **?**. The form presented here is due to Rangan Rangan (2011).

Part III

Random Constraint Satisfaction Problems

Chapter 13

Graph Coloring II: Insights from planting

13.1 SBM and planted coloring

Consider here a special case of the stochastic block model with $n_a = 1/q$ and $c_{aa} = c_{\rm in}$, and $c_{ab} = c_{\rm out}$ for $a \neq b$. We call assortative/ferromagnetic the case with $c_{\rm in} > c_{\rm out}$, i.e. connections withing groups being more likely than between different groups. We call disassortative/anti-ferromagnetic the case with $c_{\rm in} < c_{\rm out}$, i.e. connections withing groups being less likely than between different groups.

In the last lecture we derived the belief propagation equations for the stochastic block model that read

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}}\, n_{t_i} e^{-h_{t_i}} \prod_{k\in\partial i\setminus j}\left[\sum_{t_k} c_{t_k t_i}\chi_{t_k}^{k\to i}\right] = \frac{1}{Z^{i\to j}}\, n_{t_i} e^{-h_{t_i}} \prod_{k\in\partial i\setminus j}\left[c_{\rm out} - (c_{\rm in}-c_{\rm out})\chi_{t_i}^{k\to i}\right], \tag{13.1}$$

with an auxiliary external field that summarizes the contribution and overall influence of the non-edges

$$h_{t_i} = \frac{1}{N}\sum_k \sum_{t_k} c_{t_k t_i}\chi_{t_k}^{k}\,. \tag{13.2}$$

We notice that defining

$$\mathbb{e}^{-\beta} = \frac{c_{\rm in}}{c_{\rm out}} \tag{13.3}$$

and including terms independent of the value t_i info the normalization $Z^{i\to j}$ we obtain

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}}\, n_{t_i} e^{-h_{t_i}} \prod_{k\in\partial i\setminus j}\left[c_{out} - (c_{out}-c_{in})\chi_{t_i}^{k\to i}\right] \tag{13.4}$$

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}}\, n_{t_i} e^{-h_{t_i}} \prod_{k\in\partial i\setminus j} c_{out}\left[1 - (1-\frac{c_{in}}{c_{out}})\chi_{t_i}^{k\to i}\right] \tag{13.5}$$

and including c_{out} in $\frac{1}{Z^{i\to j}}$, we finally obtain:

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}} n_{t_i} e^{-h_{t_i}} \prod_{k\in\partial i\setminus j} \left[1-(1-\mathrm{e}^{-\beta})\chi_{t_i}^{k\to i}\right] , \tag{13.6}$$

where we abused the notation as we denoted the new normalization in the same way. We notice that this is very close to the BP that we wrote for graph coloring in Section 6

$$\chi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}} \prod_{k\in\partial i\setminus j} \left[1-(1-\mathrm{e}^{-\beta})\chi_{t_i}^{k\to i}\right] , \tag{13.7}$$

the only difference is the missing term $n_{t_i}e^{-h_{t_i}}$ that corresponds to the prior fixing the proper sizes of groups. In the disassortative/anti-ferromagnetic case with $c_{\rm in} < c_{\rm out}$ or equivalently $\beta > 0$, and equal group sizes $n_a = 1/q$ for all $a = 1, \dots, q$, this term is not needed and does not asymptotically influence the behaviour of the BP equations upon iterations. This claim can be checked empirically by running the BP algorithm with and without the term, showing it formally mathematically is non-trivial.

The stochastic block model under the setting of this section ($n_a = 1/q$, $c_{aa} = c_{\rm in}$, and $c_{ab} = c_{\rm out}$ for $a \neq b$) can be seen as a planted coloring problem at finite inverse temperature $\beta > 0$. Planted coloring is defined as follows:

- Start with N nodes, q colors, each node gets randomly one color $s_i^* \in [q] \triangleq \{1, \dots, q\}$ as the true configuration.
- Put at random $M = cN/2$ edges between nodes so that fraction $c_{\rm out}(q-1)/(cq)$ (chosen so that the expected number of edges between groups agrees with the SBM) of them is between nodes with different colors and the rest between nodes with the same colors. This produces the adjacency matrix $\mathbf{A} = [A_{ij}]_{i,j=1}^N$.

The goal is to find the true configuration $\mathbf{s}^*$ (up to permutation of colors) from the knowledge of the adjacency matrix $\mathbf{A}$, and the parameters θ. We call this way of generating coloring instances planted because we picture the ground truth group assignment to be the planted configuration around which is the graph to be colored constructed. The inverse temperature β is then related to the fraction of monochromatic edges.

The threshold c_d, c_c and c_ℓ that we described to occur in the SBM for $q = 5$ colors at $\beta \to \infty$ in Fig. 10.4.3 thus exist in the planted coloring. More interestingly they bear consequences on the original non-planted graph coloring problem as we describe in what follows.

13.2 Relation between planted and random graph coloring

13.2.1 BP convergence and algorithmic threshold

For the stochastic block model we derived that for

$$|c_{\rm in} - c_{\rm out}| > q\sqrt{c} \tag{13.8}$$

the BP algorithm converges to a configuration with positive overlap with the ground truth assignment into groups. In terms of inverse temperature β and the average degree c this condition is written as

$$c > c_\ell = \left(\frac{q - 1 + \mathrm{e}^{-\beta}}{1 - \mathrm{e}^{-\beta}} \right)^2 . \tag{13.9}$$

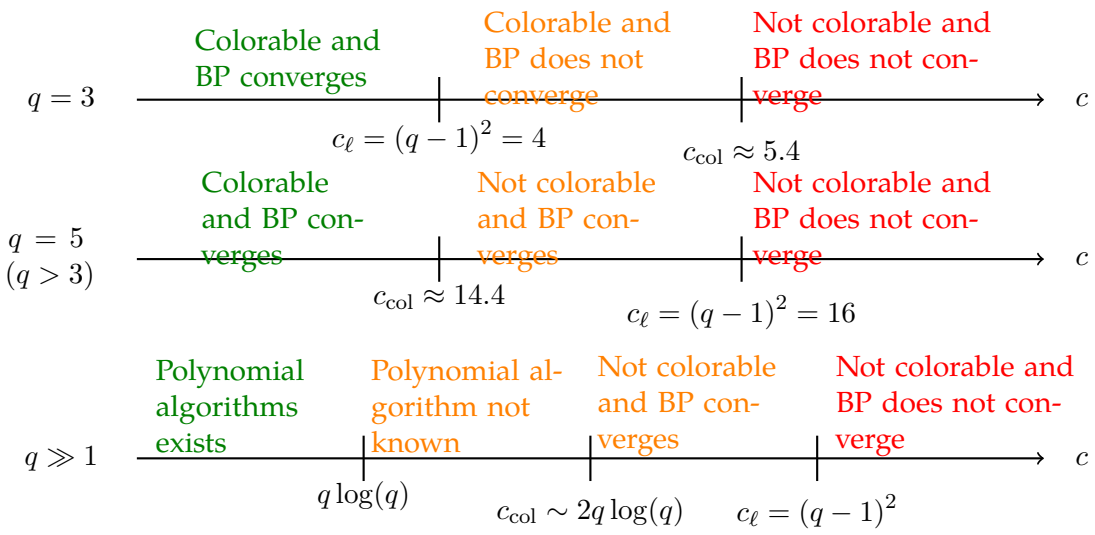
This is thus the algorithmic transition that in the SBM marks the onset of a phase where belief propagation is able to find the optimal overlap with the ground truth group assignment.

We derived this phase transition via a stability of the paramagnetic BP fixed point. Since the BP equations for the planted coloring are the same as the one for the random graph coloring the stability of the BP fixed point also applies to the random graph coloring. Indeed, at the end of Section 6 we derived a condition for when the BP for random graph coloring converges to the paramagnetic BP fixed point and when it goes away from it (6.24). This condition was exactly eq. (13.9). In fact in the normal (non-planted) coloring the BP algorithm does not converge for $c > c_\ell$.

To summarize the consequence of the threshold c_ℓ, we have 4 cases:

- $c > c_\ell$ planted coloring: randomly initialized BP converges to a configuration with positive overlap with the planted configuration.
- $c > c_\ell$ normal coloring: randomly initialized BP does not converge.
- $c < c_\ell$ planted coloring: randomly initialized BP converges to the paramagnetic fixed point.
- $c < c_\ell$ normal coloring: randomly initialized BP converges to the paramagnetic fixed point.

Let us now review in the following table how is this BP convergence threshold related to the estimation of the colorability threshold for $c_{\mathrm{in}} = 0$ or equivalently $\beta \to \infty$ and to the average degree beyond which we do not know of polynomial algorithms that would provably find proper colorings for large number of colors. We see from the table that the convergence of BP had little to do with colorability nor with hardness of finding a proper coloring. In the planted coloring this is the algorithmic threshold beyond which the inference of the planted coloring starts to the algorithmically easy, but in the non-planted coloring the c_ℓ BP convergence threshold does not have algorithmic consequences.



13.2.2 Contiguity of random and planted ensemble and clustering

We recall that for the SBM we found a threshold c_c (in continuous phase transition $c_c = c_\ell$) such that

- For $c < c_c$ the BP fixed point with the largest free entropy is the paramagnetic fixed point with $\chi_a^{i\to j} = n_a$ and $Q = 0$.
- For $c > c_c$ the BP fixed point with the largest free entropy is the feromagnetic fixed point with $Q > 0$.

Recall Fig. 10.4.3 for illustration of this threshold and also recall the fact that for $c < c_c$ the paramagnetic fixed point thus provides the exact marginals for most variables and the exact free entropy in the leading order.

A consequence of the paramagnetic fixed point being the one describing the correct marginals, overlap and free energy is that there is no information in the graph that allows us to find a configuration correlated with the planted configuration. This can only be true if all properties that are true with high-probability in the limit of large graphs are the same in the planted graph as they would be in a non-planted graph. Properties that hold with high-probability in the large size limit are called thermodynamic properties in physics. Mathematically, this kind of indistinguishability of two graph ensembles (the random and the planted one) by high-probability properties is termed *contiguity*. For $c < c_c$ the random and planted ensemble are hence contiguous, i.e. not distinguishable by thermodynamics properties. This among other things implies that for $c < c_c$ the paramagnetic BP fixed point and the corresponding Bethe free entropy are exact in the leading order not only for the planted graphs but also for the random ones. We will see in the next lecture that for $c > c_c$ this is not the case, thus answering one of the main questions we have about correctness of BP to graph coloring.

Nishimori identities for Bayes optimal inference are implying that the planted configuration has all the properties of a configuration randomly sampled from the posterior distribution. Putting this property together with the contiguity of the random and planted ensemble for

$c < c_c$ this allows us to investigate in a unique way properties of equilibrium configuration in the random ensemble.

In the section on SBM we observed empirically that the fixed point of BP and marginals reached by correspondingly initialized Monte Carlo Marlo chain in time linear in the size of the system are equivalent in the large size limit (MCMC was just slower to convergence). We thus conclude that if BP converges to the paramagnetic fixed point from the planted initialization, i.e. for $c < c_d$, then MCMC would be able to *equilibrate*, i.e. estimate in linear time correctly in the leading order all quantities that concentrate. Thus for $c < c_d$ dynamics such as MCMC is able to equilibrate in a number of steps that is $O(1)$ per node, this happens if from the physics point of view the phase is liquid and not glassy.

In the cases where we have a 1st order phase transition in the planted model, i.e. when $c_d < c_c$ the phase occurring for $c_d < c < c_c$ has interesting properties. Recall that for all $c < c_c$ the planted and random ensemble are contiguous. Yet in the planted ensemble the BP initialized in the planted configuration converges to a fixed point strongly correlated with the planted configuration $Q > 0$. Because of the contiguity this means that also in the random ensemble BP initialized in a configuration drawn uniformly at random from the Boltzmann distribution would converge to an analogous fixed point. Considering that such a BP fixed point describes the subspace that would be sampled by MCMC if initialized in the same way we conclude that dynamics initialized at equilibrium would remain stuck close to the initialisation and would not explore in linear time the whole phase space. The domain of attraction where the dynamics is stuck will be denoted as a cluster. Randomly initialized MCMC will not be able to equilibrate, i.e. sample the space of configuration almost uniformly for $c_d < c < c_c$. In this phase is thus conjectured that sampling configurations uniformly from the Boltzmann measure is algorithmically hard.

Having introduced the notion of clustering, let us state one possible definition of how to split configurations into clusters via equivalence classes of BP fixed points: Consider the random graph coloring problem at inverse temperature β, consider BP equations (at the same temperature) initialized in all possible configurations and iterate each of those BP to a fixed point. Then define a cluster of configurations as all configurations from which BP converges to the same fixed point. Cluster are then equivalence classes of configurations with respect to BP fixed points. In case BP does not reach convergence from some configurations we can think that one cluster corresponds to all such configurations. While it may be harder to relate this definition to the more physical notion such as Gibbs states it is one that is most readily translated into a method of analysis that is able to describe organization of clusters, e.g. how many there are of a given free entropy.

The size of the cluster is described by the free entropy of the corresponding BP fixed point Φ_{ferro}. We remind that for $c < c_c$ the total equilibrium free entropy is the paramagnetic one Φ_{para}, that is in the phase $c_d < c < c_c$ strictly larger than the ferromagnetic one $\Phi_{\text{para}} > \Phi_{\text{ferro}}$. If equilibrium solutions are in cluster (basins of attraction) corresponding to free entropy Φ_{ferro} and the total free entropy is Φ_{para} is must mean that there are $\mathrm{e}^{N(\Phi_{\text{para}}-\Phi_{\text{ferro}})}$ of such clusters. We will define the *complexity* of clusters as the logarithm of the number of such clusters per node, defined as:

$$e^{N\Sigma} = \text{number of clusters} \tag{13.10}$$

where:

$$\Sigma = \Phi_{\text{para}} - \Phi_{\text{ferro}} \tag{13.11}$$

for $c \in (c_{\rm d}, c_{\rm c})$. Since at c_d the ferromagnetic fixed point appears discontinuously, it means that exponentially many clusters appear discontinuously. As $c \to c_c$ the complexity Σ is then going to zero.

To summarize

- For $c < c_d$ most configurations belong to the same cluster, MCMC is able to equilibrate in linear time.
- For $c_d < c < c_c$ configurations belong to one of exponentially many clusters, MCMC is not able to equilibrate in linear time.

This relation to dynamics being able to equilibrate before the threshold and not being able to equilibrate in linear time after the threshold gives the name *dynamical* threshold.

So far the method we described to locate the clustering threshold c_d is based on running BP equations from the random and planted fixed point and comparing the free entropies of the corresponding fixed points. This is a somewhat demanding procedure and does not lead to a simple closed-form expression for the threshold as we obtained e.g. for the linear stability thresholds c_ℓ. In the next section we will give a closed-form upper bound on the clustering threshold c_d for the case of proper colorings, i.e. when $c_{\rm in} = 0$ or equivalently when $\beta \to \infty$.

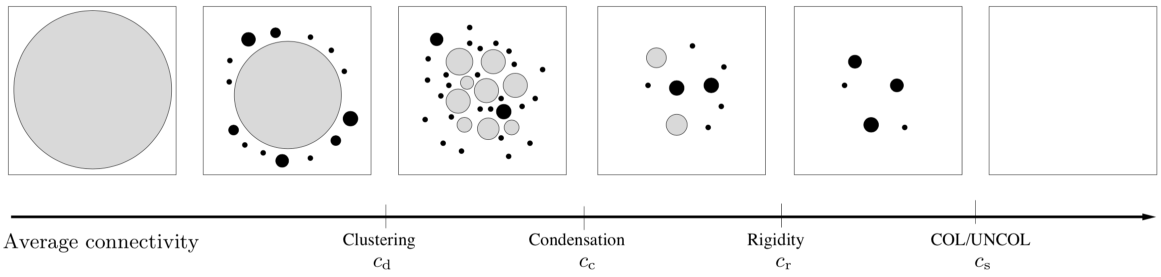


Figure 13.2.1: Cartoon of the space of proper colorings in random graph coloring problem. The grey circles correspond to unfrozen clusters, the black circles to frozen clusters. The size of the circle illustrates the size of the clusters.

13.2.3 Upper bound on clustering threshold

For graph coloring at zero temperature, $\beta \to \infty$, the BP fixed points initialized in a solution (proper coloring) may have so-called *frozen variables*, i.e. variables i that stay at their initial value up until the fixed point. This means that the frozen variables take the same value in all the solutions belonging to the cluster. *Frozen variables*: Variable i is frozen if in the fixed point of BP we still have $\chi_s^{j\to i} = \delta_{s,s_j^*}$. *Frozen cluster*: A cluster in frozen if a finite fraction of variables are frozen in the cluster.

In order to monitor whether clusters in which a randomly samples solution lives are frozen of not we recall the contiguity between the random and planted ensemble and the fact that the planted configuration has all the properties of the equilibrium configurations. This allows

us to monitor the threshold where typical clusters get frozen by tracking the updates of BP initialized in the planted configuration and monitoring how many variables are frozen.

Let us consider a simple case, for degree-d regular random graph, that is locally-treelike in the large size limit $N \to \infty$. Consider then a rooted tree around of on the nodes. The planted solution can be seen as a broadcasting process on that tree where the root took a given color and then recursively the children (i.e. the neighbors in the direction of the leaves) are given at random one of the other colors. The question of existence of frozen variables is then translated into the question whether the leaves of such a tree of large depth imply the value of the root or whether the root can take another color while staying compatible with the coloring rule and the colors on the leaves.

Denote η_ℓ the probability that a node in ℓ-th generation is implied by the leaves above it, i.e. can take only one color (the planted). A node will be implied to have its planted color if each of the other colors is implied in at least one of its children.

$$\begin{aligned}
\eta_\ell &= \Pr\left(\text{a node in } \ell\text{-th generation is implied by its leaves}\right) \\
&= \Pr\left(\text{for a node in } \ell\text{-th generation, all other } (q-1) \text{ colors are implied on children}\right) \\
&= 1 - \Pr\left(\text{for a node in } \ell\text{-th generation, at least one of the } (q-1) \text{ colors is not implied from children}\right)
\end{aligned}$$

Assume now that r of the remaining $q-1$ colors are not implied on any of its $d-1$ children. This probability can be written as:

$$\begin{aligned}
&\Pr\left(\text{for a node in } \ell\text{-th generation, given } r \text{ colors are not implied by any children}\right) \\
&= \left\{\Pr\left(\text{for a node in } \ell\text{-th generation, the given } r \text{ colors are not implied by a given child}\right)\right\}^{d-1} \\
&= \left\{1 - \Pr\left(\text{for a node in } \ell\text{-th generation, a given child is implied to take one of the } r \text{ given colors}\right)\right\}^{d-1} \\
&= \left[1 - \frac{r\,\eta_{\ell+1}}{q-1}\right]^{d-1}
\end{aligned}$$

We then proceed by inclusion-exclusion principle. For $r=1$ we over-counted the cases where in fact two colors were not implied etc. obtaining

$$\begin{aligned}
\eta_\ell &= 1 - \sum_{r=1}^{q-1} (-1)^{r-1} \binom{q-1}{r} \Pr\left(\text{for a node in } \ell\text{-th generation, the given } r \text{ colors are not implied by any children}\right) \\
&= 1 + \sum_{r=1}^{q-1} (-1)^{r} \binom{q-1}{r} \left[1 - \frac{r\,\eta_{\ell+1}}{q-1}\right]^{d-1}
\end{aligned}$$

Therefore, we end up with a one-variable recursion

$$\eta_\ell = 1 + \sum_{r=1}^{q-1} (-1)^{r} \binom{q-1}{r} \left[1 - \frac{r\,\eta_{\ell+1}}{q-1}\right]^{d-1},$$

For a tree of depth d we start from $\eta_d = 1$, since the nodes in d-th generation are a leaves themselves. The probability η_ℓ is then updates as we proceed down to the root up to a fixed point.

We now define a rigidity threshold c_{r} such that

- When $d > c_{\mathrm{r}}$, the fixed point $\eta > 0$.

- When $d < c_{\rm r}$, the fixed point $\eta = 0$.

The rigidity threshold c_r then provides an upper bound on the dynamical threshold c_d. For small number of colors the two threshold are not particularly close to each other and the rigidity threshold in even larger than c_c for $q < 9$. But for large number of colors $q \gg 1$ we get by expansion $c_{\rm r} = q \log(q) + {\rm o}\left(q \log(q)\right)$. This thus tells us that for large number of colors clusters containing typical solution are frozen with high probability starting from average degree at least c_r.

We note that while we presented this argument for d-regular trees, for the random ones with fluctuating degree distribution we can get an analogous closed-form expression that has the same larger q behaviour in the leading order.

13.2.4 Comments and bibliography

Contiguity between planted and random ensembles only holds when paramagnetic (or analogous) fixed point exists. In general problems such as random K-SAT there the planted ensemble is always different from the random one. However, the notion of clustering and clusters corresponding to basins of attraction of Monte Carlo sampling as well as their definition via BP fixed point is more general and holds beyond the models where planted and random ensembles are related.

We also note that the scaling of the rigidity threshold at large number of colors is $c_r \approx q \log(q)$ and this coincides with the scaling beyond which we do not have any known analyzable algorithms working. Moreover numerical tests in the literature show that solutions found by polynomial algorithms do not lead to a frozen BP fixed point. This leads to a conjecture that frozen solutions are hard to find. At the same time the rigidity threshold concerns freezing of the typical clusters and there might be atypical ones that remain unfrozen up to much larger average degree. This is indeed expected to be the case in analogy with related constraint satisfaction problems that have been analyzed Braunstein et al. (2016). Overall the precise threshold at which solution become hard to find remain open, even on the level of a conjecture.

Chapter 14

Graph coloring III: Colorability threshold and properties of clusters

14.1 Analyzing clustering: Generalities

In the last section we deduced that in a region of parameters the space of solutions in the random graph coloring problem is split into so-called clusters. We associated clusters to BP fixed points, and their free entropy to the Bethe free entropy $\Phi_{\rm Bethe}$ corresponding to the fixed point. This allows us to analyse how many clusters there are of a given free entropy in the same way as we analyzed how many configurations there are of a given energy/cost. We define the complexity function $\Sigma(\Phi_{\rm Bethe})$ as the logarithm of the number of BP fixed points with a given Bethe free entropy $\Phi_{\rm Bethe}$ per node. With this definition we can define the so-called replicated free entropy $\Psi(m)$ as

$$\mathbb{e}^{\Psi(m)N} = \sum_{\text{BP fixed points}} \mathbb{e}^{Nm\Phi_{\rm Bethe}} = \int_{\Phi} \mathbb{e}^{N[\Sigma(\Phi)+m\Phi]} \tag{14.1}$$

Just as before for the relation between energy, entropy and free entropy, we have from the properties of the saddle point and the Legendre transform that

$$\frac{\partial\Sigma(\Phi)}{\partial\Phi} = -m\,, \quad \frac{\partial\Psi(m)}{\partial m} = \Phi. \tag{14.2}$$

Thus if we are able to compute the replicated free entropy $\Psi(m)$ we can compute from it the number of clusters of a given free entropy $\Sigma(\Phi)$ just as we did when computing the number of configurations of a given energy.

We would now like to determine what is the free entropy Φ of clusters that contain the equilibrium configurations, i.e. configurations samples at random from the Boltzmann measure. For this we need to consider the free entropy Φ that maximizes the total free entropy $\Phi + \Sigma(\Phi)$. We also need to take into account that the BP fixed point are structures that can be counted and those that exist must be at least one, and hence $\Sigma(\Phi)$ cannot be negative, since logarithm of positive integers are non-negative. Thus the equilibrium of the system is described by

$$\max_{\Phi}[\Phi + \Sigma(\Phi)|\Sigma(\Phi) \geq 0] \tag{14.3}$$

This expression is maximized in two possible ways. Assume that for neighborhood of $m = 1$ the complexity $\Sigma(\Phi)$ has non-zero value (except maybe one point where it crosses zero). Then define Φ_1 as the value of the free entropy at which the function $\Sigma(\Phi)$ has slope -1 i.e.

$$\frac{\partial \Sigma(\Phi)}{\partial \Phi}\Big|_{\Phi_1} = -1 \tag{14.4}$$

then

- Either $\Sigma(\Phi_1) > 0$, in this case eq. (14.3) is maximized at Φ_1 and the number of clusters of that free entropy is exponentially large corresponding to $\Sigma(\Phi_1)$. We call such a clustered phase the dynamical one-step replica symmetry breaking phase.
- Or $\Sigma(\Phi_1) < 0$, in this case the value of free entropy that maximized the expression (14.3) is the largest Φ such that $\Sigma(\Phi) > 0$. We denote this the Φ_0. When this happens the dominating part of the free entropy is included in the largest clusters that are not exponentially numerous and in fact arbitrarily large fraction of the weight is carried by a finite number of largest clusters. The equilibrium of the system thus condensed in a few of the larger clusters. We call this phase the static one-step replica symmetry breaking phase.

In the last lecture we computed the complexity of clusters in random graph coloring for $c_d < c < c_c$ as the difference between the free entropy of the paramagnetic and the feromagnetic fixed point. In problems where we have contiguity between the planted and the random ensembles, this complexity from eq. (13.11) is exactly equal to $\Sigma(\Phi_1)$ and thus goes to zero at the condensation threshold c_c. Above the condensation threshold $c > c_c$ the equilibrium of the system is no longer described by cluster of free entropy Φ_1 (that is equal to the free entropy of the planted cluster) instead it is given by Φ_0 as defined above by the free entropy at which the complexity becomes zero.

As the average degree grows further in the random graph coloring problem the maximum of the complexity curve $\Sigma(\Phi)$ becomes zero at which point the last existing clusters disappear and this thus marks the colorability threshold. To compute the colorability threshold we thus need to count the total number of clusters, corresponding to the maximum of the curve $\Sigma(\Phi)$.

14.2 Analyzing BP fixed points

From the previous section we see that we can learn many details about clusters if we are able to compute the replicated free entropy $\Psi(m)$ from eq. (14.1) where we sum the exponential of the Bethe free entropy to the power m over all the BP fixed points. More explicitly this means to compute

$$\mathbb{e}^{\Psi(m)N} = \sum_{\text{BP fixed points}} \mathbb{e}^{m[\sum_i \log Z^i + \sum_a \log Z^a - \sum_{(ia)} \log Z^{ia}]} \tag{14.5}$$

where the terms Z^i, Z^a, and Z^{ia} are the contributions to the Bethe free entropy that we derived when we were deriving the BP equations. We will write BP equations schematically as

$$\chi^{i\to a} = \mathcal{F}_\chi(\{\psi^{b\to i}_{b\in\partial i\setminus a}\}) \tag{14.6}$$

$$\psi^{a\to i} = \mathcal{F}_\psi(\{\chi^{j\to a}_{j\in\partial a\setminus i}\}) \tag{14.7}$$

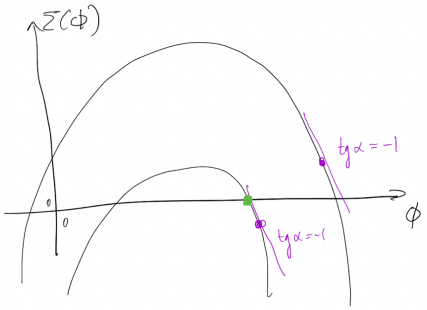


Figure 14.1.1: Illustration of the complexity curve $\Sigma(\Phi)$ that is the logarithm of the number of clusters per nodes versus the free entropy Φ. The point where the curve has slope -1 is marked in purple. Recall the properties of the Legendre transform that imply that the slope of this curve is equal to $-m$.

The replicated free entropy thus reads

$$\mathrm{e}^{\Psi(m)N} = \int \prod_{(ia)} \left[\mathrm{d}\chi^{i\to a}\mathrm{d}\psi^{a\to i}\right] \mathrm{e}^{m[\sum_i \log Z^i + \sum_a \log Z^a - \sum_{(ia)} \log Z^{ia}]} \tag{14.8}$$

$$\prod_i \prod_{a\in\partial i} \delta[\chi^{i\to a} - \mathcal{F}_\chi(\{\psi^{b\to i}_{b\in\partial i\setminus a}\})] \prod_a \prod_{i\in\partial a} \delta[\psi^{a\to i} - \mathcal{F}_\psi(\{\chi^{j\to a}_{j\in\partial a\setminus i}\})] \,. \tag{14.9}$$

What we just wrote is in fact very naturally a partition function of an auxiliary problem living on the original graph where the variables $(\chi^{i\to a}, \psi^{a\to i})$ and fields $(Z^{ia})^m$ live on the original edges and the factors live on the original variable i and factor nodes a.

We thus put the problem of computing the replicated free energy $\Psi(m)$ into a form in which we can readily apply belief propagation as we derived it in Section 5.2. The only difference now is that the variables are continuous real numbers and that the messages in the auxiliary problem are probability distributions. Moreover we realize that it is consistent to assume that the new BP message in the auxiliary problem in the direction from $i \to a$ does not depend on the original message in the opposite direction, because of the assumed independence between BP messages. This allows us to write the BP for the auxiliary problem that is a generic form of a survey propagation (instead of messages we are now passing surveys). These equations are also called the one step replica symmetry breaking cavity equations (but deriving the

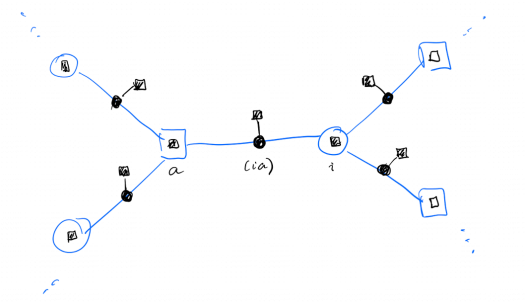


Figure 14.2.1: Illustration of the graphical model constructed from the original factor graph used to analyze BP fixed points.

same result with the replica approach is rather demanding and much less transparent). The parameter m is called the Parisi parameter. They read

$$P^{a\to i}(\psi^{a\to i}) = \frac{1}{\mathcal{Z}^{a\to i}} \int \prod_{j\in\partial a\setminus i} \mathrm{d}P^{j\to a}(\chi^{j\to a})(Z^{a\to i})^m \delta[\psi^{a\to i} - \mathcal{F}_\psi(\{\chi^{j\to a}_{j\in\partial a\setminus i}\})] \tag{14.10}$$

$$P^{i\to a}(\chi^{i\to a}) = \frac{1}{\mathcal{Z}^{i\to a}} \int \prod_{b\in\partial i\setminus a} \mathrm{d}P^{b\to j}(\psi^{b\to j})(Z^{i\to a})^m \delta[\chi^{i\to a} - \mathcal{F}_\chi(\{\psi^{b\to i}_{b\in\partial i\setminus a}\})] \tag{14.11}$$

where the messages are now probability distributions over the original BP messages. As before we need to find a fixed point of these equations which in a single graph is numerically demanding because we need to update a full probability distribution (in fact two) on every edge.

In problems on random regular graphs without any additional disordered, e.g. in the graph coloring on random regular graphs, the correct solution often has a form that is independent on the edge (ia). In that case we have the same probability distribution on every edge and the fixed point can be found naturally by so-called population dynamics where the probability distribution of edges is represented by a large set of samples from the probability distributions and the so-called reweighting factor $(Z^{i\to a})^m$ or $(Z^{a\to i})^m$ is proportional to the probability with which each element appears in the population.

The replicated free entropy can then be computed from the fixed point using the recipe for

the Bethe entropy on the auxiliary problem and reads:

$$N\Psi(m) = \sum_i \Phi^i + \sum_a \Phi^a - \sum_{ia} \Phi^{ia} \tag{14.12}$$

where

$$\mathrm{e}^{m\Phi^i} = \int \prod_{a\in\partial i} \mathrm{d}P^{a\to i}(\psi^{a\to i})(Z^i)^m \tag{14.13}$$

$$\mathrm{e}^{m\Phi^a} = \int \prod_{i\in\partial a} \mathrm{d}P^{i\to a}(\chi^{i\to a})(Z^a)^m \tag{14.14}$$

$$\mathrm{e}^{m\Phi^{ia}} = \int \mathrm{d}P^{a\to i}(\psi^{a\to i})\mathrm{d}P^{i\to a}(\chi^{i\to a})(Z^{ia})^m\,. \tag{14.15}$$

The curve $\Sigma(\Phi)$ is then computed from the Legendre transform of $\Psi(m)$.

It might happen that also the clusters cluster into super-clusters, or split into mini-clusters. This would then lead to the so-called two-step replica symmetry breaking. And one could continue to speak about K-step RSB, or full-RSB in case an infinite hierarchy of step is needed. Currently we do not know how to solve the full-RSB equations on random sparse graphs, but in dense optimization problems this can be done, but this is beyond the scope of the present lecture.

14.3 Computing the colorability threshold

The concept of frozen variables that we introduced to upper bound the clustering threshold enables a key simplification to compute the colorability threshold c_{col} that we aimed at from the very first lecture on graph coloring. We will again restrict to random d-regular graphs as the computation simplifies in this case. We learned that the space of solutions is divided in *clusters* and some clusters have *frozen variables*.

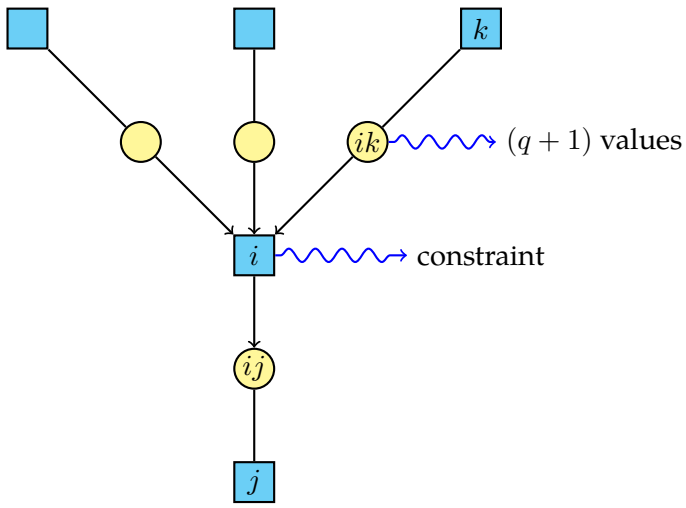
Note that frozen variables are needed in order for a cluster of solution to vanish when the average degree is infinitesimally increasing (addition of new edge will create contradiction with high probability). Further it is not possible to have a very small fraction of frozen variables because they must sustain each other in a sort of backbone structure. We will thus assume that the colorability threshold is given by the average degree at which clusters with frozen variables all disappear. Since clusters were identified with BP fixed points, counting frozen clusters to see when all disappear becomes counting BP fixed points with frozen variables.

The key idea of describing clusters is again that being a BP fixed point is just another type of constraints for another type of variables. It is an auxiliary problem that can be formulated using a tree-like graphical model and solved using again belief propagation and counting frozen clusters comes from the value of the corresponding free entropy. This "BP on frozen BP fixed points" is called the *survey propagation*.

Let us now describe the structure of a frozen BP fixed point. The frozen BP messages are $\chi^{j\to i}_{s_j} = \delta_{s_j,s}$ = "s" for q possible values of s; the not frozen ones are $\chi^{j\to i}_{s_j}$ = something else = "$*$" and will be called "joker". For each edge (ij) in each direction the message will be either "joker" or frozen in one of the actual colors. The variables will now be living on the edges, let

$\nu_{ij} \in \{*, 1, \dots, \mathsf{q}\}$ denote the value for variable node (ij), in total this variables can be in one of the $(q+1)$ possible states. In the new graphical model BP equations lead to the constraints, terms in Bethe entropy are the new factor weights, BP messages are the new variables. The new messages are probability distribution on the original BP messages.

Let us not discuss the constraints (factor nodes) that will be in place of the original nodes (as in graph matching problems)



The constraint $\mathfrak{C}_i$ on node i requires the following:

- if an incoming value $\nu_{ki} = \mathsf{s}$, the outgoing value ν_{ij} cannot be s since the neighbor k is frozen to s; if an incoming value $\nu_{ki} = *$, then it will not impose any constraint on the outgoing value ν_{ij}.

- incoming values do not forbid all colors

- if $\nu_{ij} = *$ at least two colors are not forbidden by the incoming edges, if $v_{ij} = \mathsf{s}$, only s is not forbidden by the incoming edges.

We can now write the survey propagation equation, which is just BP on this new graphical model

$$\mathsf{n}^{i\to j}_{\nu_{ij}} = \frac{1}{Z^{i\to j}} \sum_{\{\nu_{ki}\}_{k\in\partial i\setminus j}} \mathfrak{C}\left(\nu_{ij}, \{\nu_{ki}\}_{k\in\partial i\setminus j}\right) \prod_{k\in\partial i\setminus j} \mathsf{n}^{k\to j}_{\nu_{ki}}$$

which is generic form of BP on variables ν_{ij}.

Survey propagation (SP) equations can be made more explicit by using the specific form of the constraints imposed above. We proceed again according to the inclusion-exclusion principle

- When $\nu_{ij} \neq *$

$$\mathsf{n}_{\nu_{ij}}^{i\to j} = \frac{1}{Z^{i\to j}}\Bigg\{ \underbrace{\prod_{k\in\partial i\setminus j}\left(1-\mathsf{n}_{\nu_{ij}}^{k\to i}\right)}_{\text{neighbors are not forbidding } \nu_{ij}} + \text{correct for the cases that also allow } \mathsf{p}\neq\nu_{ij}\Bigg\}$$

$$= \frac{1}{Z^{i\to j}}\Bigg\{ \prod_{k\in\partial i\setminus j}\left(1-\mathsf{n}_{\nu_{ij}}^{k\to i}\right) - \sum_{\mathsf{p}\neq\nu_{ij}}\prod_{k\in\partial i\setminus j}\left(1-\mathsf{n}_{\nu_{ij}}^{k\to i}-\mathsf{n}_{\mathsf{p}}^{k\to i}\right) + \cdots - (-1)^q \underbrace{\prod_{k\in\partial i\setminus j}\left(1-\sum_{p=1}^{q}\mathsf{n}_{\mathsf{p}}^{k\to i}\right)}_{\text{no body is forbidden}}\Bigg\}$$

$$= \frac{1}{Z^{i\to j}}\sum_{l=0}^{q-1}(-1)^l \sum_{\substack{V\subseteq\{1,\ldots,\mathsf{q}\}\setminus\nu_{ij}\\ |V|=l}}\prod_{k\in\partial i\setminus j}\left(1-\mathsf{n}_{\nu_{ij}}^{k\to i}-\sum_{\mathsf{v}\in V}\mathsf{n}_{\mathsf{v}}^{k\to i}\right)$$

- When $\nu_{ij} = *$

$$\mathsf{n}_{*}^{i\to j} = \frac{1}{Z^{i\to j}}\{\text{not forbidding at least two colors}\}$$
$$= \frac{1}{Z^{i\to j}}\sum_{l=2}^{q}(-1)^l \sum_{\substack{V\subseteq\{1,\ldots,\mathsf{q}\}\setminus\nu_{ij}\\ |V|=l}}\prod_{k\in\partial i\setminus j}\left(1-\sum_{\mathsf{v}\in V}\mathsf{n}_{\mathsf{v}}^{k\to i}\right).$$

- Because of the normalization we need $\mathsf{n}_{*}^{i\to j} + \sum_{p=1}^{q}\mathsf{n}_{\mathsf{p}}^{i\to j} = 1$, thus

$$Z^{i\to j} = \Pr(\text{no contradiction (at least one color not forbidden)})$$
$$= \sum_{p=1}^{q}\prod_{k\in\partial i\setminus j}\left(1-\mathsf{n}_{\mathsf{p}}^{k\to i}\right) - \sum_{p<r}\prod_{k\in\partial i\setminus j}\left(1-\mathsf{n}_{\mathsf{p}}^{k\to i}-\mathsf{n}_{\mathsf{r}}^{k\to i}\right) + \cdots - (-1)^q\prod_{k\in\partial i\setminus j}\left(1-\sum_{p=1}^{q}\mathsf{n}_{\mathsf{p}}^{k\to i}\right)$$
$$= \sum_{l=1}^{q}(-1)^{l-1}\sum_{\substack{V\subseteq\{1,\ldots,\mathsf{q}\}\\ |V|=l}}\prod_{k\in\partial i\setminus j}\left(1-\sum_{\mathsf{v}\in V}\mathsf{n}_{\mathsf{v}}^{k\to i}\right)$$

The Bethe free entropy of this problem is the *complexity* Σ, counting all the frozen clusters. We remind that when we expressed complexity in the previous section of was the one of clusters that contained typical solutions. In the next lecture we will see the relation between the two. The complexity can be computed from the fixed point of SP using the usual recipe for Bethe free entropy

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}\log\left(\Sigma^{(i)}\right) - \frac{1}{N}\sum_{(ij)\in E}\log\left(\Sigma^{(ij)}\right)$$
$$\Sigma^{(i)} = \sum_{l=1}^{q}(-1)^{l-1}\sum_{\substack{V\subseteq\{1,\ldots,\mathsf{q}\}\\ |V|=l}}\prod_{k\in\partial i}\left(1-\sum_{\mathsf{v}\in V}\mathsf{n}_{\mathsf{v}}^{k\to i}\right)$$
$$\Sigma^{(ij)} = 1-\sum_{p=1}^{q}\mathsf{n}_{\mathsf{p}}^{i\to j}\mathsf{n}_{\mathsf{p}}^{j\to i}$$

Here the term $\Sigma^{(i)}$ is related to the normalization of the messages that does not exclude the node j, the term $\Sigma^{(ij)}$ is then counting the probability that the messages in the two directions of one edge are compatible, i.e. not imposing the same color on the two ends.

On a d-regular graph and assuming symmetry among colors ($\mathsf{n_p} = \eta, \forall\, p = 1, \ldots, q$), this leads to a very concrete conjecture for the colorability threshold of random d-regular graphs:

$$\eta = \frac{\sum_{\ell=1}^{q} (-1)^{\ell-1} \binom{q-1}{\ell-1} (1-\ell\eta)^{d-1}}{\sum_{\ell=1}^{q} (-1)^{\ell-1} \binom{q}{\ell} (1-\ell\eta)^{d-1}}$$

$$\Sigma = \log\left\{\sum_{\ell=1}^{q} (-1)^{\ell-1} \binom{q}{\ell} (1-\ell\eta)^{d}\right\} - \frac{d}{2}\log\left(1-q\eta^2\right)$$

We can first compute η to be the fixed point, and then plug the fixed point in Σ to compute the complexity. If $\Sigma > 0$, then it means a random d-regular graph is colorable w.h.p. for large N, and uncolorable vise versa. For the random graph with fluctuating variables degree the expressions are only slightly more complex, but derived in the very same spirit.

Going back to the overall picture we now described the whole regime of average degrees, except $c_c < c < c_{\text{col}}$. This is the condensed phase which has rather peculiar properties and will be discussed in the next lecture.

14.4 Exercises

Exercise 14.1: Random subcube model

The random-subcube model is defined by its solution space $S \subset \{0,1\}^N$ (not by a graphical model). We define S as the union of $\lfloor 2^{(1-\alpha)N} \rfloor$ random clusters (where $\lfloor x \rfloor$ denotes the integer value of x). A random cluster A being defined as:

$$A = \left\{ \boldsymbol{\sigma} \,\middle|\, \sigma_i \in \pi_i^A, \quad \forall\, i \in \{1, \ldots, N\} \right\} \tag{14.16}$$

where π^A is a random mapping:

$$\begin{aligned} \pi^A \colon \{1, \ldots, N\} &\to \{\{0\}, \{1\}, \{0,1\}\} \\ i &\mapsto \pi_i^A \end{aligned}$$

such that for each variable i, $\pi_i^A = \{0\}$ with probability $p/2$, $\{1\}$ with probability $p/2$, and $\{0,1\}$ with probability $1-p$. A cluster is thus a random subcube of $\{0,1\}$. If $\pi_i^A = \{0\}$ or $\{1\}$, variable i is said "frozen" in A; otherwise it is said "free" in A. One given configuration $\boldsymbol{\sigma}$ might belong to zero, one or several clusters. A "solution" belongs to at least one cluster.
We will analyze the properties of this model in the limit $N \to \infty$, the two parameters α and p being fixed and independent of N. The internal entropy s of a cluster A is defined as $\frac{1}{N}\log_2(|A|)$, i.e. the fraction of free variables in A. We also define complexity $\Sigma(s)$ as the (base 2) logarithm of the number of clusters of internal entropy s per variable (i.e. divide by N).

(a) What is the analog of the satisfiability threshold α_s in this model?

(b) Compute the $\alpha_{\rm d}$ threshold below which most configurations belong to at least one cluster.

(c) For $\alpha > \alpha_{\rm d}$ write the expression for the complexity $\Sigma(s)$ as a function of the parameters p and α. Compute the total entropy defined as $s_{\rm tot} = \max_s \left[\Sigma(s) + s \,|\, \Sigma(s) \geq 0\right]$. Observe that there are two regimes in the interval $\alpha \in (\alpha_{\rm d}, 1)$, discuss their properties and write the value of the "condensation" threshold $\alpha_{\rm c}$.

Chapter 15

Replica Symmetry Breaking: The Random Energy Model

> If you can't solve a problem, then there is an easier problem you can solve: find it.
>
> George Pòlya , How to Solve It - 1945

In this chapter, we come back to the replica method, and will in particular discuss replica symmetry breaking (in its simpler form, the one-step replica symmetry breaking, 1RSB in short), and will make contact with the form of 1RSB discussed with the cavity method.

To do so, we shall discuss a very simple model, that was introduced by Bernard Derrida in 1980 to understand spin glasses and the replica method. It turns out to be indeed a very good toy model to understand key concepts, and in fact, it is also a very important model in its own, connecting with important concepts in denoising and information theory.

The random energy model is a trivial spin models with N variables, in the sense that the energy of the possible $M = 2^N$ configurations are choosen, once and for all, randomly from a Gaussian distribution.

Formally, in the Random Energy Model (REM), we have 2^N configuration, with random fixed energies E_i sampled from

$$P(E) = \mathcal{N}\left(0, \frac{N}{2}\right) = \frac{1}{\sqrt{\pi N}} e^{-\frac{E^2}{N}} \tag{15.1}$$

so that the partition sum reads

$$Z_N = \sum_{i=1}^{2^N} e^{-\beta E_i} \tag{15.2}$$

15.1 Rigorous Solution

We start to computing the exact asymptotic solution of the model, without using the replica method. To do so, we shall show that the entropy (in the sense of Boltzmann) density as a function of the energy has a deterministic asymptotic limit, that we can compute.

To do so, we first ask: if we sample 2^N energy randomly, what is the number of energies that fall between $[Ne, N(e+de)]$? Let us call this number $\#(E)$. This is a random variable, it depends on the specific draw of the 2^N configuration, but we can compute its mean and variance. First we compute the average; when de is small enough, we have

$$\mathbb{E}(\#Ne) = 2^N \mathbb{E}[\mathbb{1}(E \in [Ne, N(e+\mathrm{d}e)])] \approx \frac{2^N}{\sqrt{\pi N}} e^{-Ne^2}\,\mathrm{d}e = \frac{1}{\sqrt{\pi N}} e^{-N(e^2-\log 2)}\,\mathrm{d}e \quad (15.3)$$

Also, we see that, the probability of the energy taken randomly from $P(E)$ to be between e and $e+de$ being small, this follows a Poisson law, so that the variance is equal to the mean. Defining the function $s_{\rm ann}(e) = \log 2 - e^2$ (physicists call this quantity the annealed entropy density) we thus have two regimes:

- If $s_{\rm ann}(e) < 0$, then there the average number of configuration with energy e is exponentially close to 0. This happens when $e < -\sqrt{\log 2}$ and $e > \sqrt{\log 2}$. In this case Markov inequality tells us that with high probability the number of configuration (and not only its average) is actually 0.
  Indeed, from Markov:
  $$\mathbb{P}(X \geq 1) \leq \mathbb{E}(X) \quad (15.4)$$
  so we have, with high probability as N is going to infinity, all configurations have energy densities between $[-\sqrt{\log 2}, \sqrt{\log 2}]$.

  Markov inequality states that if X is a non-negative random variable, then $\mathbb{P}(X \geq k) \leq \mathbb{E}[X]/k$.

- If $s_{\rm ann}(e) > 0$, we have instead an exponential number of configurations, with a variance also exponential. However the entropy density concentrates to a deterministic value, as can be seen again from Markov inequality and the fact that the variance is equal to the mean:

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\#(e)}{\mathbb{E}\#(e)} - 1\right| \geq k\right) &= \mathbb{P}\left(\left(\frac{\#(e)}{\mathbb{E}\#(e)} - 1\right)^2 \geq k^2\right) \leq \frac{\mathbb{E}\left[(\#(e) - \mathbb{E}[\#(e)])^2\right]}{k^2(\mathbb{E}[\#(e)])^2} \\ &\leq \frac{\mathrm{Var}([\#(e)]}{k^2(\mathbb{E}[\#(e)])^2} \simeq \frac{\cancel{\mathbb{E}[\#(e)]}}{k^2(\mathbb{E}[\#(e)])^{\cancel{2}}} \propto \frac{e^{-N s_{\rm ann}(e)}}{k^2}\,. \end{aligned} \quad (15.5)$$

  Where we have used the fact that in this case, since the probability of the energy taken randomly from $P(E)$ to be between e and $e+de$ being small follows a Poisson law, the variance is equal to the mean. As N grows, the probability of deviation is going exponentially to zero when $s_{\rm ann}(e) > 0$, so that with high probability, $\frac{\#(e)}{\mathbb{E}\#(e)}$ is arbitrary close to 1.

Thanks to this very tight concentration, we now can state that, with high probability, the number of configurations is either 0 —if $s(E)$ is negative— or close to $e^{N s_{\rm ann}(e)}$ otherwise. More precisely, we can write the entropy density as a function of e:

$$s(e) =: s_{\rm ann}(e) \text{ if}, s_{\rm ann}(e) \geq 0, \text{ and } s(e) = -\infty \text{ otherwise}\,. \quad (15.6)$$

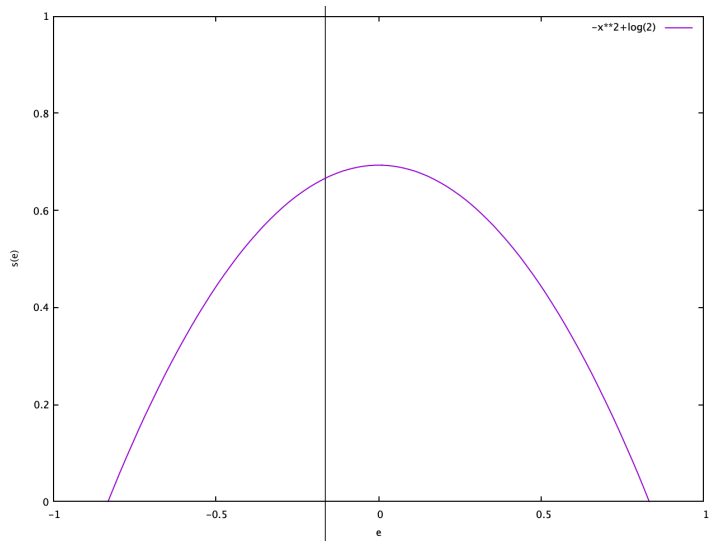


Figure 15.1.1: Entropy density in the REM

With this knowledge, we can now solve the random energy model easily, without any disorder averaging. We write that, with high probability, we have

$$Z_N = \sum_E \#(E)\mathrm{e}^{-\beta E} \approx \int_{-\sqrt{\log 2}}^{\sqrt{\log 2}} de\mathrm{e}^{-N(\beta e - s(e))} \tag{15.7}$$

and using Laplace integral, with the sum is dominated by its maximum, we have the free entropy given by the Legendre transform of the entropy:

$$\lim_{N\to\infty} \frac{1}{N} \log Z = s(e^*) - \beta e^* \tag{15.8}$$

with

$$e^* = \min_{[-\sqrt{\log 2},\sqrt{\log 2}]} [\beta e - s(e)] \tag{15.9}$$

Again, there are two situations: the minimum can be reached when the derivative is 0, that is when

$$\beta = \partial_e s(e) = -2e \tag{15.10}$$

but this can only be the case when $s(e) > 0$. However, at $e = -\sqrt{\log 2}$, $s(e)$ reaches zero, and at this point $s'(e) = 2\sqrt{\log 2}$. So this minimum is only valid when $\beta < \beta_c = 2\sqrt{\log 2}$, after which $e^* = -\sqrt{\log 2}$. In a nutshell, we have:

$$\lim_{N\to\infty} \frac{1}{N} \log Z = -(e^*)^2 + \log 2 - \beta e^* = \log 2 + \frac{\beta^2}{4}, \text{if} \beta < \beta_c = 2\sqrt{\log 2} \tag{15.11}$$

$$\lim_{N\to\infty} \frac{1}{N} \log Z = \sqrt{\log 2}\beta, \text{if} \beta \geq \beta_c = 2\sqrt{\log 2} \tag{15.12}$$

The phase transition arising a β_c is called a condensation transition. This is because, at this point, all the probability measure condensate on the lowest energy configurations of energy $-N\sqrt{\log 2}$. This phenomenon is of crucial importance in understanding the 1RSB phase transition.

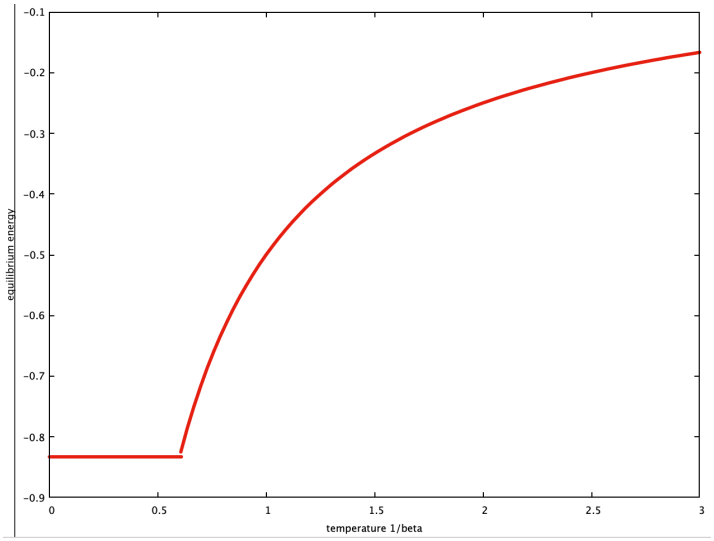


Figure 15.1.2: Equilibrium energy in the REM

15.2 Solution via the replica method

We now move to the computation by the replica method. We start by the replicated partition sum:

$$Z^n = \prod_n \left(\sum_{i=1}^{2^N} e^{-\beta E_i} \right) \tag{15.13}$$

Let us do a bit of seemingly trivial rewriting

$$Z^n = \sum_{i_1,\dots,i_n=1}^{2^N} e^{-\beta(E_{i_1}+\dots+E_{i_n})} \tag{15.14}$$

$$= \sum_{i_1,\dots,i_n=1}^{2^N} e^{-\beta \sum_{a=1}^n E_{i_a}} = \sum_{i_1,\dots,i_n=1}^{2^N} e^{-\beta \sum_{a=1}^n \left(\sum_{j=1}^{2^N} E_j \mathbb{1}(j=i_a) \right)} \tag{15.15}$$

$$= \sum_{i_1,\dots,i_n=1}^{2^N} \prod_{j=1}^{2^N} e^{-\beta E_j (\sum_{a=1}^n \mathbb{1}(j=i_a))} \tag{15.16}$$

We now compute the average over the disorder. By linearity of expectation, we can push the average in the sum. Using independence of the different 2^N energies, we can also push the average over each E_j in the product. Using the Gaussian integral , we thus find that:

Remember that if X is Gaussian with zero mean and variance Δ, then $\mathbb{E}[e^{bX}] = e^{\frac{b^2\Delta}{2}}$.

$$\mathbb{E}_{E_j} e^{-\beta E_j (\sum_{a=1}^n I(j=i_a))} = e^{\frac{N\beta^2}{4} (\sum_{a=1}^n \mathbb{1}(j=i_a))^2} = e^{\frac{N\beta^2}{4} \sum_{a,b=1}^n \mathbb{1}(j=i_a)\mathbb{1}(j=i_b)} \tag{15.17}$$

and therefore the expectation of the replicated partition sum reads:

$$\mathbb{E}[Z^n] = \sum_{i_1,\dots,i_n=1}^{2^N} e^{\frac{N\beta^2}{4} \sum_{a,b=1}^n \mathbb{1}(i_a=i_b)} \tag{15.18}$$

Following our tradition of using overlaps, we see that, given the replicas configurations $(i_1...i_n)$, it is convenient to introduce the $n \times n$ overlap matrix $Q_{ab} = \mathbb{1}(i_a = i_b)$, that takes elements in $0, 1$, respectively if the two replicas (row and column) have different or equal configuration. We can finally write the replicated sum over configurations as

$$\mathbb{E}[Z^n] = \sum_{i_\alpha=1,...,i_n}^{2^N} e^{\frac{N\beta^2}{4}\sum_{a,b=1}^n Q_{a,b}} = \sum_{\{Q\}} \#(Q) e^{N\frac{\beta^2}{4}\sum_{a,b=1}^n Q_{a,b}} \tag{15.19}$$

where $\sum_{\{Q\}}$ is the sum over all possible such matrices, and $\#(Q)$ is the numbers of configurations that leads to the overlap matrix Q.

15.2.1 An instructive bound

Considering, for a moment, the n to be integers, we can derive an instructive bound on all the moments of Z. Indeed, a possible value of the matrix Q is the one where all the n replica are in the same configuration, in which case $Q_{a,b} = 1,\ \forall a, b$. There are 2^N such matrices, so

$$\mathbb{E}Z^n > 2^N e^{N\frac{\beta^2}{4}n^2} > e^{N\frac{\beta^2}{4}n^2} \tag{15.20}$$

so that we now know, rigorously that the moment of Z grow at least as e^{n^2}. This is a bad new. Indeed, if the moments do not grow faster than exponentially, their knowledge completely determines the distribution of Z, and thus the expectation of its logarithm, according to Carleman's condition. Since, in our case, they do grow faster, this means that the moments *do not* necessarily determine the distribution of Z, and in particular the analytic continuation at $n < 1$ may not be unique. We thus will need to choose the "right" one, which is the essence of the replica method. This is precisely what Parisi's ansatz is doing for us: it provides a well defined class of analytic continuations, which turns out to be the correct one.

15.2.2 The replica symmetric solution

Let us try to perform the analytic continuation when $n \to 0$. Keeping for a moment n integer, it is natural to expect that the number of such configurations , for a given overlap matrix, to be exponential so that, denoting $\#(Q) = e^{Ns_q(Q)}$ we write

$$\mathbb{E}Z^n \approx \int \mathrm{d}Q e^{N\left(\frac{\beta^2}{4}\sum_{a,b=1}^n Q_{a,b} + s_q(Q)\right)} =: \int \mathrm{d}Q e^{Ng(\beta,Q)} \tag{15.21}$$

As N is large, we thus expect to be able to perform a Laplace (or Saddle point) approximation by choosing the "right" structure of the matrix Q, which will "dominate" the sum. A quite natural guess (Physicists like to call this an *ansatz*) is to assume that all replicas are identical, and therefore the system should be invariant under the relabelling of the replicas (permutation symmetry). In this case, we only have two choices for the entries of Q: 1) $Q_{ab} = 1$ for all a, b or 2),$Q_{aa} = 1$ for $a = 1, \cdots, n$ and $Q_{ab} = 0$ for $a \neq b$. In both case, we can easily evaluate $s_q(Q)$.

1. If $Q_{ab} = 1$ for all a, b, then all the replica are in a single configuration, as we have already seen. Then $\mathcal{N} = 2^N$, and we find $g(\beta, Q) = n^2\beta^2/4 + \log 2$. This is actually very frustrating, as this does *not* have a limit with a linear part in n, so we cannot use this solution in the replica method. Clearly, this is a wrong analytical continuation.

2. If instead $Q_{aa} = 1$ and $Q_{ab} = 0$ for all $a \neq b$ then all replicas are in a different configuration; $\#(Q) = 2^N(2^N - 1)\dots(2^N - n + 1)$, so that $s(Q) \approx n \log 2$ if $n \ll N$. Therefore $g(\beta, Q) = n\beta^2/4 + n \log 2$.

At the replica symmetric level, we thus find that the free entropy is given, at all temperature, by

$$f_{\rm RS}(\beta) = \beta^2/4 + \log 2 \tag{15.22}$$

This is not bad, and indeed we know it is the correct solution for $\beta < \beta_c$. However, this solution is obviously wrong for $\beta > \beta_c$.

15.2.3 The replica symmetry breaking solution

We now follow Parisi's prescription, and use a different ansatz for the saddle point. We assume the n replica are divided into many groups of m replica, and that the overlap is 1 within the group, and 0 between different groups. The number of group is of course n/m. In order to count the number of such matrices, we can choose one configuration by group, so that $\mathcal{N} = 2^N(2^N-1)\dots(2^N-n/m+1)$, $s_q(Q) = \frac{\log(2^N)^{\frac{n}{m}}}{N} = (n/m)\log 2$ and $\sum_{a,b} Q_{a,b} = \frac{n^2}{\frac{n}{m}} = nm$. Thus, we have

$$\mathbb{E}[Z^n] \simeq e^{N[\frac{n}{m}\log 2 + \frac{\beta^2}{4}nm]} \tag{15.23}$$

and again $1 \leq m \leq n$

$$g(q) = nm\beta^2/4 + n/m \log 2 \tag{15.24}$$

This has a nice limit as n going to zero, and we thus find:

$$f_{\rm 1RSB}(\beta, m) = m\beta^2/4 + log2/m \tag{15.25}$$

We now need to choose m. The historical reasoning to choose m proposed by Parisi is almost a joke, as it sounds crazy. First, we know that for n integer, we have obviously $n \geq m \geq 1$. Parisi tells us than, when $n \to 0$, we must instead choose $n \leq m \leq 1$. Secondly, as it was not crazy enough, we will see that it corresponds to a *minimum* rather than to a maximum. We shall see later on how to rationalize these results, thanks to the cavity method, but so far, let us follow the replica prescriptions. We extremize the replica free entropy with respect to m and find:

$$\beta^2/4 = \log 2/m^2 \tag{15.26}$$

so that $m = \beta_c/\beta$. Since, however, we want $m \leq 1$, we write instead

$$m = \min[\beta_c/\beta, 1] \tag{15.27}$$

This, amazingly, turns out to be the correct solution! Indeed, we have now:

1. If $\beta < \beta_c$ (high-temperature), we have $m = 1$ so that

$$f_{\rm 1RSB}(\beta, m) = \beta^2/4 + \log 2 = f_{\rm RS}(\beta, m) \tag{15.28}$$

2. while, if instead $\beta > \beta_c$ (low-temperature), we have $m = \beta_c/\beta$ so that

$$f_{\text{1RSB}}(\beta, m) = \beta_c\beta/4 + \beta \log 2/\beta_c = \beta\sqrt{\log 2} \tag{15.29}$$

which is indeed the correct free energy. There is definitely something magical, but it works.

15.2.4 The condensation transition and the participation ratio

The replica method can also be used to shed some light on the type of phase transition happening at β_c. Indeed, we know that for low temperature, the Boltmzann measure condensate on the lowest energy level, close to $e = -\sqrt{\log 2}$, but we may ask, how many are they, really? This can be answered by computing the participation ratio defined as

$$Y = \sum_{i=1}^{2^N} \left(\frac{e^{-\beta E_j}}{Z} \right)^2 \tag{15.30}$$

Handwavingly speaking, the participation ratio is the inverse of the number of configuration that matters in the sum. Let's compute its average by the replica method! We find

$$\begin{aligned}
\mathbb{E}Y &= \mathbb{E}\left[\sum_{i=1}^{2^N} \left(\frac{e^{-\beta E_j}}{Z} \right)^2 \right] = \mathbb{E}\left[\frac{1}{Z^2} \sum_{i=1}^{2^N} e^{-2\beta_j} \right] = \lim_{n\to 0} \mathbb{E}\left[Z^{n-2} \sum_{i=1}^{2^N} e^{-2\beta E_j} \right] && (15.31)\\
&= \lim_{n\to 0} \mathbb{E}\left[\sum_{i_1,\dots,i_{n-2}} e^{-\beta(E_{i_1}+E_{i_2}+\dots+E_{i_{n-2}})} \sum_{i=1}^{2^N} e^{-2\beta E_j} \right] && (15.32)\\
&= \lim_{n\to 0} \mathbb{E}\left[\sum_{i_1,\dots,i_n} e^{-\beta(E_{i_1}+E_{i_2}+\dots+E_{i_n})} \mathbb{1}(i_{n-1} = i_n) \right] && (15.33)
\end{aligned}$$

Now, using the invariance of all the replica, we symmetrize the last expression so that

$$\mathbb{E}Y = \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{a\neq b} \mathbb{E}\left[\sum_{i_1,\dots,i_n} e^{-\beta(E_{i_1}+E_{i_2}+\dots+E_{i_n})} \mathbb{1}(i_a = i_b) \right] \tag{15.34}$$

Finally, since the limit when $n \to 0$ of Z^n is one, we write

$$\begin{aligned}
\mathbb{E}Y &= \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{a\neq b} \frac{\mathbb{E}\left[\sum_{i_1,\dots,i_n} e^{-\beta(E_{i_1}+E_{i_2}+\dots+E_{i_n})} \mathbb{1}(i_a = i_b) \right]}{\mathbb{E}\left[\sum_{i_1,\dots,i_n} e^{-\beta(E_{i_1}+E_{i_2}+\dots+E_{i_n})} \right]} && (15.35)\\
&= \lim_{n\to 0} \frac{1}{n(n-1)} \langle Q_{a,b} \rangle && (15.36)
\end{aligned}$$

and assuming, as we saw in the 1RSB computation, that the replica computation is dominated by its saddle points, we find

$$\begin{aligned}
\mathbb{E}Y &= \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{a\neq b} Q_{a,b}^{1RSB} && (15.37)\\
&= \lim_{n\to 0} \frac{1}{n(n-1)} n/m \left[m^2 - m \right] = \lim_{n\to 0} \frac{1}{(n-1)} \left[m - 1 \right] && (15.38)\\
&= 1 - m = 1 - \beta_c/\beta && (15.39)
\end{aligned}$$

We can now plot the participation ratio and see that it foes from 1 at zero temperature and diverges at $T = T_c = 1./\beta_c$. A similar computation shows that second moment is finite, so that this number is NOT self-averaging, the number of configurations participating in the states fluctuate from instance to instance.

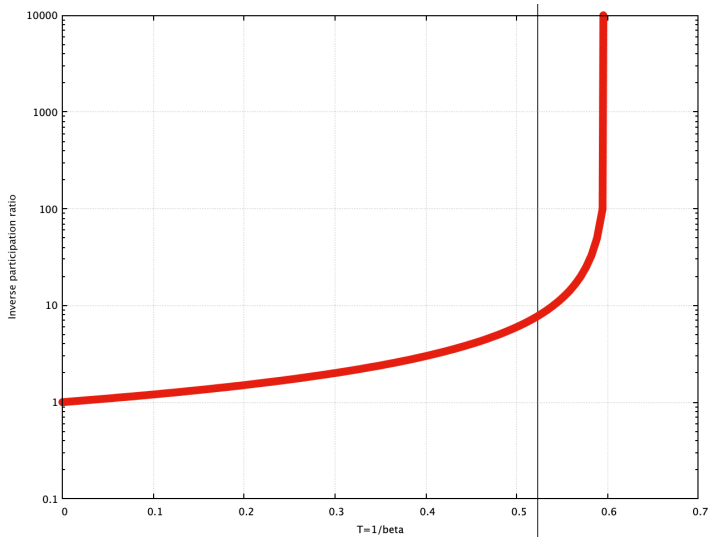


Figure 15.2.1: Inverse participation ratio in the REM

15.2.5 Distribution of overlaps

Another way to think about the replica solution is through the distribution of possible overlaps. Consider two *actual real* replica of the system. We may ask, what if we take two independent copies of the same system, what is the probability that they are in the same configuration? Or, equivalently, what is the probability that their overlap is one? This is precisely what we computed with the participation ratio! Indeed, the (average) probability that two copies are in the same configurations reads

$$\mathbb{E}(\mathbb{P}(q=1)) = \mathbb{E}\left[\sum_{i,j=1}^{2^N}\left(\frac{e^{-\beta E_i-\beta E_j}}{Z^2}\mathbb{1}(i=j)\right)\right] = \mathbb{E}[Y] = \lim_{n\to 0}\frac{\sum_{a\neq b}\mathbb{1}(Q_{a,b})}{n(n-1)}\,. \tag{15.40}$$

In the high-temperature solution, when the RS approach is correct, we see that the only possible overlap is 0. If we sample two configurations in the high-temperature region, then with high-probability, they are different ones.

In the 1RSB low temperature solution, we see instead that with probability $1-m$ we have $q=1$ (and therefore, with probability m, we have $q=0$). If we sample two configurations, they will be the same ones with probability $1-m$. This is another way to think of the condensation phenomenon. In spin glass parlance, one often says that the distribution of overlap become non trivial. We went from a simple delta-peak at high-temperature

$$\mathbb{E}P_{\rm RS}(q) = \delta(q) \tag{15.41}$$

to a distribution with two pick at low temperature

$$\mathbb{E}P_{\rm 1RSB}(q) = (1-m)\delta(q-1) + m\delta(q) \tag{15.42}$$

15.3 The connection between the replica potential, and the complexity

We have seen that the replica method is indeed very powerful, and allows to compute everything we want, pretty much. Can we understand how this version of replica symmetry breaking is connected with the one introduced in the cavity method, while at the same time shed some light on the strange replica ansatz and minimization? The answer is yes, using the construction derived in the cavity method.

The 1RSB potential — Let us discuss again the idea introduced at the previous chapter. The whole concept is that there are many "Gibbs states", and that we would like to count them. So the entire partition sum can be divided into "many" partition sums, each of them defining a Gibbs states (so that quantities of interest are well defined and concentrate within each state). We thus have

$$Z = \sum_{\alpha} Z_{\alpha} \tag{15.43}$$

The idea is then that we can count those states. To do so, we introduce this modified partition sum, that weights all partition sum by a power m:

$$Z(m) = \sum_{\alpha} Z_{\alpha}^{m} = \sum_{\alpha} e^{N m f_{\alpha}} \tag{15.44}$$

where we have denoted the free entropy of each state α as $f_{\alpha} = (\log Z_{\alpha})/N$. We assume that there are exponential many of them, so we need to replace this discrete sum, and approximate it by an integral and hope that the version we compute using the our approximation (BP, RS etc...) will be correct:

$$Z(m) = \int df \#(f) e^{N m f_{\alpha}} = \int df e^{N(m f_{\alpha} + \Sigma(f)} \tag{15.45}$$

where we have introduced the "complexity" $\Sigma(f)$, that is the logarithm of the number of state of free entropy f (divided by N). At this point, we can perform the Laplace integral and write

$$Z(m) = \int df \#(f) e^{N m f_{\alpha}} = \int df e^{N(m f_{\alpha} + \Sigma(f)} \tag{15.46}$$

so that we compute the thermodynamic potential

$$\Psi(m) = \frac{1}{N} \log Z(m) \to m f_{\alpha}^{*} + \Sigma(f^{*}) \tag{15.47}$$

$$m = -\partial_f \Sigma(f)|_{f^*} \tag{15.48}$$

This construction was at the basis of the 1RSB approach in the cavity method, and it allowed to compute the complexity function. Additionally, the Legendre transform ensure that

$$\partial_m \Psi(m) = f^{*} \tag{15.49}$$

As discussed in the previous chapter, we have to pay attention to the fact that Σ is actually an average version of the true logarithm of the number of state, and can be negative, so that we need to evaluate the integral on values of m such that $\Sigma \geq 0$.

This is reminiscent of the solution for the REM, except that for the REM, a "configuration" is the same as a "Gibbs" state (this is just because the REM is very simple model). Still, the same thing happens, with condensation on the lowest energy configuration, so that m can be different from one when Σ is negative.

With replica: the "Real replica" method — Can we compute the same potential $\Psi(m)$ with the replica method? The answer is yes, and very instructively so.

First, we notice that if we *force* m copies (or real replica) to be in the same "states", then the replicated sum

$$Z^m = (\sum_\alpha Z_\alpha)^m = \sum_{i_1,i_2,\dots i_m} Z_{i_1} Z_{i_2} \dots Z_{i_m} \,, \tag{15.50}$$

becomes

$$Z^m_{\text{constrained}} = \sum_\alpha Z^m_\alpha \,, \tag{15.51}$$

which is what we want to compute the potential.

Now, we need to compute the average of the log of this quantity and write, using the replica method

$$\Psi(m) = \mathbb{E}\frac{1}{N}\log Z^m_{\text{constrained}} = \lim_{n'\to\infty} \frac{(Z^m_{\text{constrained}})^{n'} - 1}{Nn'} \tag{15.52}$$

This would be the way to compute the potential that gives the Legendre transform of the complexity. However, we also see that

$$\begin{aligned} \frac{1}{m}\mathbb{E}\frac{1}{N}\log Z^m_{\text{constrained}} &= \lim_{n'\to 0} \frac{Z^{mn'}_{\text{constrained}} - 1}{n'm} && (15.53)\\ &= \lim_{n\to 0} \frac{Z^{n}_{\text{constrained}} - 1}{Nn} && (15.54)\\ & && (15.55) \end{aligned}$$

We thus see that when we perform the replica computation, constraining m replica to be in the same Gibbs states, and thus perform the 1RSB ansatz, what we do is nothing but the computation of the Legendre transform of the complexity (up to a trivial $1/m$ factor)! The 1RSB ansatz is nothing but a fancy way (in replica space) to construct the function $\Psi(m)$.

This is, we believe, a much clearer motivation for the 1RSB ansatz! indeed, with this motivation in mind, it is clearer to see why m should be 1, unless the complexity is negative, so that indeed $m \in [0:1]$. Additionally, it also explain why we must extremize the replica 1RSB formula with respect to m. Indeed

$$\partial_m \frac{\Psi(m)}{m} = \frac{\Psi'(m)m - \Psi(m)}{m^2} \tag{15.56}$$

which is zero when

$$\begin{aligned} \Psi'(m)m - \Psi(m) &= 0 && (15.57)\\ f^*m - f^*m + \Sigma(f^*) &= 0 && (15.58)\\ \Sigma(f^*) &= 0 && (15.59) \end{aligned}$$

so indeed, we see that extremizing over m in the replica formalism is the same as looking for the zero complexity. This is exactly what we had to do for the REM at low temperature!

Bibliography

Replica symmetry Breaking was invented by Parisi in a serie of deep thought provoking papers (Parisi, 1979, 1980) that ultimately led him to receive the Physics Nobel prize in 2021. In particular, the distribution of overlaps was introduced in (Parisi, 1983) and its fundamental role in the replica theory described in Mézard et al. (1984). The Random Energy Model was introduced by Derrida (1980, 1981) as a toy model to understand replicas and spin glasses, and it played an important role in the clarification of the replica method, especially as it leads to the study of the celebrated p-spin model by (Gross and Mézard, 1984). The construction in terms of counting Gibbs states is discussed in Monasson (1995); Mézard and Parisi (1999). It was instrumental in the creation of the modern version of the cavity method (Mézard and Parisi, 2001; Mézard et al., 2002; Mézard and Parisi, 2003) discussed in the previous chapters.

15.4 Exercises

Exercise 15.1: REM as a p-spin model

The p-spin model is one of the cornerstones of spin glass theory. It is defined as follows: there are 2^N possible configurations for the N spins variables $S_i = \pm 1$, and the Hamiltonian is given by all possible p-body (or p-upplet) interactions:

$$\mathcal{H} = -\sum_{i_1,i_2,\dots,i_p} J_{i_1,\dots,i_p} S_{i_1} \dots S_{i_p} \tag{15.60}$$

with

$$P(J) = \sqrt{\frac{\pi p!}{N^{p-1}}} e^{-\frac{N^{p-1}}{p!}J^2} \tag{15.61}$$

1. Computing the moment of the partition function using Gaussian integrals, show that

$$\mathbb{E}[Z^n] = \sum_{\{S_i^a\},\{S_i^b\},\dots,\{S_i^n\}} \exp\left[\frac{\beta^2}{4N^{p-1}} \sum_{a,b} \left(\sum_i S_i^a S_i^b\right)^p\right] \tag{15.62}$$

2. introducing delta functions to fix the overlap and taken its Fourier transform, show that

$$\mathbb{E}[Z^n] \approx \int \prod_{a<b} dq_{a,b} d\hat{q}_{a,b} e^{-NG(Q,\hat{Q})} \tag{15.63}$$

with

$$G(Q,\hat{Q}) = -n\frac{\beta^2}{4} - \frac{\beta^2}{2}\sum_{a<b} q_{a,b}^p + i\sum_{a<b} \hat{q}_{a,b} q_{a,b} - \log\left[\sum_{\{S_i^a,\dots,S_i^n\}} e^{\sum_{a<b} i\hat{q}_{a,b}\frac{1}{N}\sum_i S_i^a S_i^b}\right] \tag{15.64}$$

3. Using the replica method, with the replica symmetric solution, show that the free entropy is given by

$$f_{\text{RS}} = \frac{\beta^2}{4} + \log 2 \tag{15.65}$$

as in the REM. It is possible, of course, to break the symmetry and to obtain the 1RSB solution, which turns out to be correct at low temperature (at least for p large enough, and for a range of temperature).

4. It is interesting that the free energy is the same as the one of the REM. Show that, when $p \to \infty$, the energies of the $p-spin$ model becomes uncorrelated and that the $p-spin$ model become the REM in this limit.

Exercise 15.2: Second moment of the participation ratio in REM

We consider again the REM. Using the replica method, show that at for $\beta \geq \beta_c$, the second moment of the participation ratio is given by

$$\mathbb{E}[Y^2] = \frac{3 - 5m + 2m^2}{3} \tag{15.66}$$

Deduce that Y is not self-averaging. Actually, these results do not depends on the REM, but on the 1RSB structure, and are universal to all 1RSB models.

Exercise 15.3: Maximum of Ω Gaussians numbers and denoising

- We are given Ω Gaussian numbers $Z_i, i = 1, \ldots, \Omega$, all sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0, \Delta)$. Using the same reasoning as for the random energy model, show that when $\log \Omega$ is large, then the maximum value $\omega_{\max} = \max \frac{Z_i}{\sqrt{\log \Omega}}$ is, with high probability, $\sqrt{2\Delta}$. Check this prediction numerically, by computing the average value of the maximum of Ω Gaussian numbers for many values of Ω.

- Imagine we are interested in a sparse $\Omega-$dimensional vector $\mathbf{x}^*$. All of its components are zero, expect a few (say one or two) that are $O(1)$.

  Unfortunately, we are not given $\mathbf{x}^*$, but instead $\mathbf{y} = \mathbf{x}^* + \mathbf{z}$, a vector where all the component of $\mathbf{x}^*$ have been polluted by independent random Gaussian noises of variance Δ. Can you propose a simple algorithm to recover $\mathbf{x}^*$ from $\mathbf{y}$ that will work with high probability when Ω is large? For which value of Δ will this algorithm works? Test your algorithm in simulation.

Part IV

Statistics and machine learning

Chapter 16

Linear Estimation and Compressed Sensing

He doesn't play for the money he wins
He don't play for respect
He deals the cards to find the answer
The sacred geometry of chance
The hidden law of a probable outcome

Sting – Shape of my heart, 1993

16.1 Problem Statement

The inference problem addressed in this section is the following. We observe a M dimensional vector y_μ, $\mu = 1, \ldots, M$ that was created component-wise depending on a linear projection of an unknown N dimensional vector x_i, $i = 1, \ldots, N$ by a known matrix $F_{\mu i}$

$$y_\mu = f_{\text{out}} \left(\sum_{i=1}^{N} F_{\mu i} x_i \right) , \tag{16.1}$$

where f_{out} is an output function that includes noise. Examples are the additive white Gaussian noise (AWGN) were the output is given by $f_{\text{out}}(x) = x + \xi$ with ξ being random Gaussian variables, or the linear threshold output where $f_{\text{out}}(x) = \text{sign}(x - \kappa)$, with κ being a threshold value. The goal in linear estimation is to infer $\mathbf{x}$ from the knowledge of $\mathbf{y}$, F and f_{out}. An alternative representation of the output function f_{out} is to denote $z_\mu = \sum_{i=1}^{N} F_{\mu i} x_i$ and talk about the likelihood of observing a given vector $\mathbf{y}$ given $\mathbf{z}$

$$P(\mathbf{y}|\mathbf{z}) = \prod_{\mu=1}^{M} P_{\text{out}}(y_\mu|z_\mu) . \tag{16.2}$$

16.2 relaxed-BP

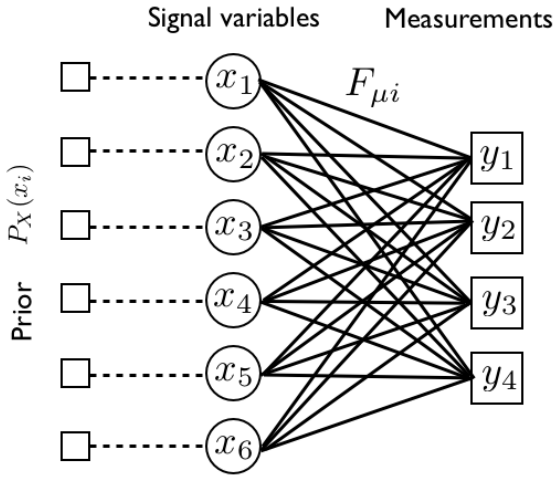


Figure 16.2.1: Factor graph of the linear estimation problem corresponding to the posterior probability for generalized linear estimation . Circles represent the unknown variables, whereas squares represent the interactions between variables.

We shall describe the main elements necessary for the reader to understand where the resulting algorithm comes from. We are given the posterior distribution

$$P(\mathbf{x}|\mathbf{y}, F) = \frac{1}{Z}\prod_{\mu=1}^{M} P_{\rm out}(y_\mu|z_\mu)\prod_{i=1}^{N} P_X(x_i)\,, \quad \text{where} \quad z_\mu = \sum_{i=1}^{N} F_{\mu i}x_i\,, \tag{16.3}$$

where the matrix $F_{\mu i}$ has independent random entries (not necessarily identically distributed) with mean and variance $O(1/N)$. This posterior probability distribution corresponds to a graph of interactions y_μ between variables (spins) x_i called the *graphical model* as depicted in Fig. 16.2.1.

A starting point in the derivation of AMP is to write the belief propagation algorithm corresponding to this graphical model. The matrix $F_{\mu i}$ plays the role of randomly-quenched disorder, the measurements y_μ are planted disorder. As long as the elements of $F_{\mu i}$ are independent and their mean and variance of of order $O(1/N)$ the corresponding system is a mean field spin glass. In the Bayes-optimal case (i.e. when the prior is matching the true empirical distribution of the signal) the fixed point of belief propagation with lowest free energy then provides the asymptotically exact marginals of the above posterior probability distribution.

For model such as (16.3) BP implements a message-passing scheme between nodes in the graphical model of Fig. 16.2.1, ultimately allowing one to compute approximations of the posterior marginals. Messages $m_{i\to\mu}$ are sent from the variable-nodes (circles) to the factor-nodes (squared) and subsequent messages $m_{\mu\to i}$ are sent from factor-nodes back to variable-nodes that corresponds to algorithm's current "beliefs" about the probabilistic distribution of

the variables x_i. It reads

$$m_{i\to\mu}(x_i) = \frac{1}{z_{i\to\mu}} P_X(x_i) \prod_{\gamma\neq\mu} m_{\gamma\to i}(x_i) \tag{16.4}$$

$$m_{\mu\to i}(x_i) = \frac{1}{z_{\mu\to i}} \int \prod_{j\neq i} [\mathrm{d}x_j m_{j\to\mu}(x_j)]\, P_{\rm out}(y_\mu|\sum_l F_{\mu l} x_l) \tag{16.5}$$

While being easily written, this BP is not computationally tractable, because every interaction involves N variables, and the resulting belief propagation equations involve $(N-1)$-uple integrals. Furthermore, here we are considering continuous variable, thus the messages would be densities and, as we have seen in Chapter 8, it would be quite hard dealing with them numerically.

Two facts enable the derivation of a tractable BP algorithm: the central limit theorem, on the one hand, and a projection of the messages to only two of their moments (also used in algorithms such as Gaussian BP or non-parametric BP.
This results in the so-called relaxed-belief-propagation (r-BP): a form of equations that is tractable and involves a pair of means and variances for every pair variable-interaction.

Let us go through the steps that allow to go from BP to r-BP and first concentrate our attention to eq. (16.5). Consider the variable $z_\mu = F_{\mu i} x_i + \sum_{j\neq i} F_{\mu j} x_j$. In the summation in eq. (16.5), all x_js are independent, and thus we expect, according to the central limit theorem, $\sum_{j\neq i} F_{\mu j} x_j$ to be a Gaussian random variable with mean $\omega_{\mu\to i} = \sum_{j\neq i} F_{\mu j} a_{j\to\mu}$ and variance $V_{\mu\to i} = \sum_{j\neq i} F_{\mu j}^2 v_{j\to\mu}$, where we have denoted

$$a_{i\to\mu} \equiv \int \mathrm{d}x_i\, x_i\, m_{i\to\mu}(x_i) \qquad v_{i\to\mu} \equiv \int \mathrm{d}x_i\, x_i^2\, m_{i\to\mu}(x_i) - a_{i\to\mu}^2\,. \tag{16.6}$$

We thus can replace the multi-dimensional integral in eq. (16.5) by a scalar Gaussian one over the random variable z:

$$m_{\mu\to i}(x_i) \propto \int dz_\mu P_{\rm out}(y_\mu|z_\mu) e^{-\frac{(z-\omega_{\mu\to i}-F_{\mu i}x_i)^2}{2V_{\mu\to i}}}\,. \tag{16.7}$$

We have replaced a complicated multi-dimensional integral involving full probability distributions by a single scalar one that involves only first two moments.

One can further simplify the equations: The next step is to rewrite $(z-\omega_{\mu\to i}-F_{\mu i}x_i)^2 = (z-\omega_{\mu\to i})^2 + F_{\mu i}^2 x_i^2 - 2(z-\omega_{\mu\to i})F_{\mu i}x_i F_{\mu i}x_i$ and to use the fact that $F_{\mu i}$ is $O(1/\sqrt{N})$ to expand the exponential

$$\begin{aligned} m_{\mu\to i}(x_i) \propto & \int dz_\mu P_{\rm out}(y_\mu|z_\mu) e^{-\frac{(z-\omega_{\mu\to i})^2}{2V_{\mu\to i}}} \\ & \times \left(1 + F_{\mu i}^2 x_i^2 - 2(z-\omega_{\mu\to i})F_{\mu i}x_i + \frac{1}{2}(z-\omega_{\mu\to i})^2 F_{\mu i}^2 x_i^2 + o(1/N)\right)\,. \end{aligned} \tag{16.8}$$

At this point, it is convenient to introduce the output function $g_{\rm out}$, defined via the output probability $P_{\rm out}$ as

$$g_{\rm out}(\omega, y, V) \equiv \frac{\int \mathrm{d}z P_{\rm out}(y|z)\,(z-\omega)\, e^{-\frac{(z-\omega)^2}{2V}}}{V\int \mathrm{d}z P_{\rm out}(y|z) e^{-\frac{(z-\omega)^2}{2V}}}\,. \tag{16.9}$$

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

$$\frac{\int \mathrm{d}z P_{\rm out}(y|z)\,(z-\omega)^2\, e^{-\frac{(z-\omega)^2}{2V}}}{V^2\int \mathrm{d}z P_{\rm out}(y|z) e^{-\frac{(z-\omega)^2}{2V}}} = \frac{1}{V} + \partial_\omega g_{\rm out}(\omega, y, V) + g^2_{\rm out}(\omega, y, V)\,. \tag{16.10}$$

Using definition (16.9), and re-exponentiating the x_i-dependent terms while keeping all the leading order terms, we obtain (after normalization) that the iterative form of equation eq. (16.5) reads:

$$m_{\mu\to i}(t, x_i) = \sqrt{\frac{A^t_{\mu\to i}}{2\pi N}} e^{-\frac{x_i^2}{2N}A^t_{\mu\to i} + B^t_{\mu\to i}\frac{x_i}{\sqrt{N}} - \frac{(B^t_{\mu\to i})^2}{2A^t_{\mu\to i}}} \tag{16.11}$$

$$\tag{16.12}$$

with

$$B^t_{\mu\to i} \;=\; g_{\rm out}(\omega^t_{\mu\to i}, y_\mu, V^t_{\mu\to i})\, F_{\mu i} \quad \text{and} \quad A^t_{\mu\to i} = -\partial_\omega g_{\rm out}(\omega^t_{\mu\to i}, y_\mu, V^t_{\mu\to i})\, F^2_{\mu i}\,.$$

Given that $m_{\mu\to i}(t, x_i)$ can be written with a quadratic form in the exponential, we just write it as a Gaussian distribution. We can now finally close the equations by writing (16.4) as a product of these Gaussians with the prior

$$m_{i\to\mu}(x_i) \;\propto\; P_X(x_i) e^{-\frac{(x_i - R_{i\to\mu})^2}{2\Sigma_{i\to\mu}}} \tag{16.13}$$

so that we can finally give the iterative form of the r-BP algorithm Algorithm 1 (below) where we denote by ∂_ω (resp. ∂_R) the partial derivative with respect to variable ω (resp. R) and we define the "input" functions as

$$f_a(\Sigma, R) = \frac{\int \mathrm{d}x\, x\, P_X(x)\, e^{-\frac{(x-R)^2}{2\Sigma}}}{\int \mathrm{d}x\, P_X(x)\, e^{-\frac{(x-R)^2}{2\Sigma}}}\,, \qquad f_v(\Sigma, R) = \Sigma \partial_R f_a(\Sigma, R)\,. \tag{16.14}$$

Examples of priors and outputs

The r-BP algorithm is written for a generic prior on the signal P_X (as long as it is factorized over the elements) and a generic element-wise output channel $P_{\rm out}$. The algorithm depends on their specific form only trough the function f_a and $g_{\rm out}$ defined by (16.14) and (16.9). It is useful to give a couple of explicit examples.

The sparse prior that is most commonly considered in probabilistic compressed sensing is the Gauss-Bernoulli prior, that is when in (**??**) we have $\phi(x) = \mathcal{N}(\overline{x}, \sigma)$ Gaussian with mean $\overline{x}$ and variance σ.

For this prior the input function f_a reads

$$f_a^{\rm Gauss-Bernoulli}(\Sigma, T) = \frac{\rho\, e^{-\frac{(T-\overline{x})^2}{2(\Sigma+\sigma)}} \frac{\sqrt{\Sigma}}{(\Sigma+\sigma)^{\frac{3}{2}}}(\overline{x}\Sigma + T\sigma)}{(1-\rho)e^{-\frac{T^2}{2\Sigma}} + \rho\frac{\sqrt{\Sigma}}{\sqrt{\Sigma+\sigma}} e^{-\frac{(T-\overline{x})^2}{2(\Sigma+\sigma)}}}\,, \tag{16.25}$$

The most commonly considered output channel is simply additive white Gaussian noise (AWGN) (**??**).

The output function then reads

$$g_{\rm out}^{AWGN}(\omega, y, V) = \frac{y-\omega}{\Delta + V}\,. \tag{16.26}$$

As we anticipated above, the example of linear estimation that was most broadly studied in statistical physics is the case of the perceptron problem discussed in detail e.g. in Watkin et al. (1993). In the perceptron problem each of the M N-dimensional patterns F_μ is multiplied by a vector of synaptic weights x_i in order to produce an output y_μ according to

$$y_\mu = 1 \qquad \text{if} \qquad \sum_{i=1}^{N} F_{\mu i} x_i > \kappa\,, \tag{16.27}$$

$$y_\mu = -1 \qquad \text{otherwise}\,, \tag{16.28}$$

where κ is a threshold value independent of the pattern. The perceptron is designed to classify patterns, i.e. one starts with a training set of patterns and their corresponding outputs y_μ and aims to learn the weights x_i in such a way that the above relation between patterns and outputs is satisfied. To relate this to the linear estimation problem above, let us consider the perceptron problem in the teacher-student scenario where the teacher perceptron generated the output y_μ using some ground-truth set of synaptic weights x_i^*. The student perceptron knows only the patterns and the outputs and aims to learn the weights. How many patterns are needed for the student to be able to learn the synaptic weights reliably? What are efficient learning algorithms?

In the simplest case where the threshold is zero, $\kappa = 0$ one can redefine the patterns $F_{\mu i} \leftarrow F_{\mu i} y_\mu$ in which case the corresponding redefined output is $y_\mu = 1$. The output function in that case reads

$$g_{\rm out}^{\rm perceptron}(\omega, V) = \frac{1}{\sqrt{2\pi V}} \frac{e^{-\frac{\omega^2}{2V}}}{H\left(-\frac{\omega}{\sqrt{V}}\right)}\,, \tag{16.29}$$

where

$$H(x) = \int_x^\infty \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,. \tag{16.30}$$

In physics a case of a perceptron that was studied in detail is that of binary synaptic weights $x_i \in \{\pm 1\}$. To take that into account in the G-AMP we consider the binary prior $P_X(x) = [\delta(x-1) + \delta(x+1)]/2$ which leads to the input function

$$f_a^{\rm binary}(\Sigma, T) = \tanh\left(\frac{T}{\Sigma}\right)\,. \tag{16.31}$$

16.3 State Evolution

Let us try to write the cavity equation here, using our simplified formulation

16.3.1 The hat variables

First, let us now see how R_i behaves (note the difference between the letters ω and w):

$$\begin{aligned}
\frac{R_i}{\Sigma_i} &= \sum_\mu B_{\mu\to i} = \sum_\mu F_{\mu i} g_{\rm out}(\omega_{\mu\to i}, y_\mu, V_{\mu\to i}) && (16.32)\\
&= \sum_\mu F_{\mu i} g_{\rm out}(\omega_{\mu\to i}, f_{\rm out}(\sum_{j\neq i} F_{\mu j} s_j + F_{\mu i} s_i, \xi_\mu), V) && (16.33)\\
&= \sum_\mu F_{\mu i} g_{\rm out}(\omega_{\mu\to i}, f_{\rm out}(\sum_{j\neq i} F_{\mu j} s_j, \xi_\mu), V) + s_i \alpha \hat{m}\,, && (16.34)
\end{aligned}$$

where we define

$$\hat{m} = \mathbb{E}_{\omega,z,\xi}\left[\partial_z g_{\rm out}(\omega, f_{\rm out}(z,\xi), V)\right]\,. \tag{16.35}$$

We further write

$$\frac{R_i}{\Sigma_i} \sim \mathcal{N}(0,1)\sqrt{\alpha\hat{q}} + s_i \alpha \hat{m} \tag{16.36}$$

with $\mathcal{N}(0,1)$ being a random Gaussian variables of zero mean and unit variance, and where

$$\hat{q} = \mathbb{E}_{\omega,z,\xi}\left[g^2_{\rm out}(\omega, f_{\rm out}(z,\xi)\, V)\right] \tag{16.37}$$

Finally, we observe that Σ should not fluctuate between sites, and we thus expect they are close to the value

$$\Sigma^{-1}(t) = \alpha\hat{\chi} \tag{16.38}$$

where we have defined

$$\hat{\chi} = -\mathbb{E}_{\omega,z,\xi}\left[\partial_\omega g_{\rm out}(\omega, f_{\rm out}(z,\xi), V)\right]\,. \tag{16.39}$$

16.3.2 Order parameters

Given these three variables, one can express the order parameters as simple integrals. Indeed, if we define:

$$\begin{aligned}
q &= \mathbb{E}_s\left[\mathbb{E}_{R,\Sigma} f_a^2(\Sigma, R)\right] && (16.40)\\
m &= \mathbb{E}_s\left[\mathbb{E}_{R,\Sigma} s f_a(\Sigma, R)\right] && (16.41)\\
V &= \mathbb{E}_s\left[\mathbb{E}_{R,\Sigma} f_c(\Sigma, R)\right] && (16.42)
\end{aligned}$$

we have immediately, denoting ξ as a standard Gaussian variable:

$$\begin{aligned}
q^t &= \mathbb{E}_s\left[\mathbb{E}_\xi f_a^2((\alpha\hat{\chi}^t)^{-1}, (\alpha\hat{\chi}^t)^{-1}(\sqrt{\hat{q}^t} + \alpha\hat{m})s\xi)\right] && (16.43)\\
m^t &= \mathbb{E}_s\left[\mathbb{E}_\xi s f_a((\alpha\hat{\chi}^t)^{-1}, \alpha\hat{\chi}^t)^{-1}(\sqrt{\hat{q}^t} + \alpha\hat{m})\right] && (16.44)\\
V^t &= \mathbb{E}_s\left[\mathbb{E}_\xi f_c((\alpha\hat{\chi}^t)^{-1}, \alpha\hat{\chi}^t)^{-1}(\sqrt{\hat{q}^t} + \alpha\hat{m})\right] && (16.45)
\end{aligned}$$

16.3.3 Back to the hats

We can now close the equation by writing the hat variables as a function of the order parameters. First, we realize that from the definition, we have z and ω jointly Gaussian with covariance

$$\begin{bmatrix} z \\ \omega \end{bmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} q_0 & m^t \\ m^t & q^t \end{pmatrix}\right) \tag{16.46}$$

So now, we can simply collect the definition, which become a 3-d integral:

$$\begin{aligned} \hat{m}^{t+1} &= \mathbb{E}^t_{\omega,z,\xi}\left[\partial_z g_{\text{out}}(\omega, f_{\text{out}}(z,\xi), V^t)\right] . && (16.47)\\ \hat{q}^{t+1} &= \mathbb{E}^t_{\omega,z,\xi}\left[g^2_{\text{out}}(\omega, f_{\text{out}}(z,\xi), V^t)\right] && (16.48)\\ \hat{\chi}^{t+1} &= -\mathbb{E}^t_{\omega,z,\xi}\left[\partial_\omega g_{\text{out}}(\omega, f_{\text{out}}(z,\xi)), V^t)\right] . && (16.49) \end{aligned}$$

16.4 From r-BP to G-AMP

While one can implement and run the r-BP, it can be simplified further without changing the leading order behavior of the marginals by realizing that the dependence of the messages on the target node is weak. This is exactly what we have done to go from the standard belief propagation to the TAP equations in sec. 9.

After the corresponding Taylor expansion the corrections add up into the so-called *Onsager reaction terms* Thouless et al. (1977). The final G-AMP iterative equations are written in terms of means a_i and their variances c_i for each of the variables x_i. The whole derivation is done in a way that the leading order of the marginal a_i are conserved. Given the BP was asymptotically exact for the computations of the marginals so is the G-AMP in the computations of the means and variances. Let us define

$$\omega_\mu^{t+1} = \sum_i F_{\mu i} a^t_{i\to\mu} , \tag{16.50}$$

$$V_\mu^{t+1} = \sum_i F^2_{\mu i} v^t_{i\to\mu} , \tag{16.51}$$

$$\Sigma_i^{t+1} = \frac{1}{\sum_\mu A^{t+1}_{\mu\to i}} , \tag{16.52}$$

$$R_i^{t+1} = \frac{\sum_\mu B^{t+1}_{\mu\to i}}{\sum_\mu A^t_{\mu\to i}} . \tag{16.53}$$

First we notice that the correction to V are small so that

$$V_\mu^{t+1} = \sum_i F^2_{\mu i} v^t_{i\to\mu} \approx \sum_i F^2_{\mu i} v^t_i . \tag{16.54}$$

From now on we will use the notation $\approx$ for equalities that are true only in the leading order.

From the definition of A and B we get

$$\begin{aligned}
\Sigma_i &= \left[-\sum_\mu F_{\mu i}^2 \partial_\omega g_{\rm out}(\omega_{\mu\to i}^{t+1}, y_\mu, V_{\mu\to i}^{t+1})\right]^{-1} \approx \left[-\sum_\mu F_{\mu i}^2 \partial_\omega g_{\rm out}(\omega_{\mu}^{t+1}, y_\mu, V_{\mu}^{t+1})\right]^{-1} \\
R_i &= \left[-\sum_\mu F_{\mu i}^2 \partial_\omega g_{\rm out}(\omega_{\mu}^{t+1}, y_\mu, V_{\mu}^{t+1})\right]^{-1} \times \left[\sum_\mu F_{\mu i} g_{\rm out}(\omega_{\mu\to i}^{t+1}, y_\mu, V_{\mu\to i}^{t+1})\right]
\end{aligned}$$

Then we write that

$$\begin{aligned}
g_{\rm out}(\omega_{\mu\to i}^{t+1}, y_\mu, V_{\mu\to i}^{t+1}) &\approx g_{\rm out}(\omega_{\mu}^{t+1}, y_\mu, V_{\mu}^{t+1}) - F_{\mu i} a_{i\to\mu}^t \partial_\omega g_{\rm out}(\omega_{\mu}^{t+1}, y_\mu, V_{\mu}^{t+1}) \quad (16.55)\\
&\approx g_{\rm out}(\omega_{\mu}^{t+1}, y_\mu, V_{\mu}^{t+1}) - F_{\mu i} a_{i}^t \partial_\omega g_{\rm out}(\omega_{\mu}^{t+1}, y_\mu, V_{\mu}^{t+1}) \quad (16.56)
\end{aligned}$$

So that finally

$$\begin{aligned}
(\Sigma_i)^{t+1} &= \left[-\sum_\mu F_{\mu i}^2 \partial_\omega g_{\rm out}(\omega_{\mu}^{t+1}, y_\mu, V_{\mu}^{t+1})\right]^{-1} \quad (16.57)\\
R_i^{t+1} &= ((\Sigma_i)^{t+1})^{-1} \times \left[\sum_\mu F_{\mu i} g_{\rm out}(\omega_\mu, y_\mu, V_\mu) - F_{\mu i}^2 a_i \partial_\omega g_{\rm out}(\omega_\mu, y_\mu, V_\mu)\right] \quad (16.58)\\
&= a_i^t + ((\Sigma_i)^{t+1})^{-1} \sum_\mu F_{\mu i} g_{\rm out}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}) \quad (16.59)
\end{aligned}$$

Now let us consider $a_{i\to\mu}$:

$$\begin{aligned}
a_{i\to\mu}^t &= f_a(R_{i\to\mu}^t, \Sigma_{i\to\mu}) \approx f_a(R_{i\to\mu}^t, \Sigma_i) \quad (16.60)\\
&\approx f_a(R_i^t, \Sigma_i) - B_{\mu\to i}^t f_v(R_i^t, \Sigma_i) \quad (16.61)\\
&\approx a_i^t - g_{\rm out}(\omega_\mu^t, y_\mu, V_\mu^t) F_{\mu i} v_i^t \quad (16.62)
\end{aligned}$$

So that

$$\omega_\mu^{t+1} = \sum_i F_{\mu i} a_i^t - \sum_i g_{\rm out}(\omega_\mu^t, y_\mu, V_\mu^t) F_{\mu i}^2 v_i^t = \sum_i F_{\mu i} a_i - V_\mu^t g_{\rm out}(\omega_\mu^t, y_\mu, V_\mu^t) \quad (16.63)$$

An important aspect of these equations to note is the index $t-1$ in the Onsager reaction term in eq. (16.65) that is crucial for convergence and appears for the same reason as in the TAP equations in sec. 9. Note that the whole algorithm is comfortably written in terms of matrix multiplications only, this is very useful for implementations where fast linear algebra can be used.

16.5 Rigorous results for Bayes and Convex Optimization

16.5.1 Bayes-Optimal

In the Bayes-optimal case, we have the Nishimori condition so that $q = m$ and $V = q_0 - m$, together with $\hat{q} = \hat{m} = \hat{\chi}$, and therefore, we can reduce everything to simpler equations.

In fact, one can prove rigorously the following theorem: the "the averaged free energy" in the statistical physics literature, in the asymptotic limit –when the number of variable is growing—of the entropy density of the variable Y reads:

$$\lim_{n\to\infty} \frac{H(Y|\mathbf{\Phi})}{n} = -\sup_{r\geq 0} \inf_{q\in[0,\rho]} \psi_{P_0}(r) + \alpha\Psi_{P_{\rm out}}(q;\rho) - \frac{rq}{2}\,. \tag{16.70}$$

where $q_0 = \rho = \mathbb{E}[x^2]$ and, denoting Gaussian integrals as $\mathcal{D}$,

$$\psi_{P_0}(r) := \int \mathcal{D}Z dP_0(X_0) \ln \int dP_0(x) e^{rxX_0+\sqrt{r}xZ-rx^2/2}\,, \tag{16.71}$$

$$\begin{aligned}\Psi_{P_{\rm out}}(q;\rho) := &\int \mathcal{D}V\mathcal{D}W d\widetilde{Y}_0 P_{\rm out}(\widetilde{Y}_0|\sqrt{q}\,V + \sqrt{\rho-q}\,W) \\ &\ln \int \mathcal{D}w P_{\rm out}(\widetilde{Y}_0|\sqrt{q}\,V + \sqrt{\rho-q}\,w)\end{aligned} \tag{16.72}$$

It is a simple check to see that q follows state evolution with the Nishimori points:

$$m^{t+1} = \psi'_{P_0}(\hat{m}^t)/2 \tag{16.73}$$

$$\hat{m}^t = \alpha\Psi'_{P_{\rm out}}(m^t;\rho)/2 \tag{16.74}$$

16.5.2 Bayes-Optimal

Another set of problem where the eisults is rigorous is when the optimization corresponds a convex problem

16.6 Application: LASSO, Sparse estimation, and Compressed sensing

16.6.1 AMP with finite Δ

In this case $f_{\rm out}^{\rm AWGN} = x + \xi$ and the output function reads

$$g_{\rm out}^{AWGN}(\omega, y, V) = \frac{y-\omega}{\Delta + V} \tag{16.75}$$

16.6.2 LASSO: infinite Δ

Let us change variable: we now denote $V = \Delta\tilde{V}$ and $\Sigma = \Delta\tilde{\Sigma}$. When Δ is large f_a and f_b can now be evaluated as Laplace integrals, and finally, all equation closes on the orginal algo where Δ is replaced by one, and where the update function are given by:

$$f_a^{\rm MAP}(\Sigma, R) = \frac{\int {\rm d}x\, x\, P_X(x)\, e^{-\frac{(x-R)^2}{2\Delta\tilde{\Sigma}}}}{\int {\rm d}x\, P_X(x)\, e^{-\frac{(x-R)^2}{2\Delta\tilde{\Sigma}}}}\,, = {\rm argmin} - \frac{(x-R)^2}{2\tilde{\Sigma}} + \lambda f_{\rm reg}(x) \tag{16.82}$$

$$f_v(\Delta\tilde{\Sigma}, R) = \Delta\tilde{\Sigma}\partial_R f_a(\tilde{\Sigma}, R)\,. \tag{16.83}$$

For instance, for a ℓ_1 regularization, we find $f_a^{\text{MAP}}(\Sigma, R) = \text{argmin}(x-R)^2/2\tilde{\Sigma} + \lambda|x|$ so that

$$\begin{aligned} f_a^{\text{ST}}(\Sigma, R) &= 0 \text{ if} |R| < \lambda\Sigma && (16.84)\\ &= R - \lambda\Sigma \text{ if} R > \lambda\Sigma && (16.85)\\ &= R + \lambda\Sigma \text{ if} R < \lambda\Sigma && (16.86)\\ & && (16.87) \end{aligned}$$

and

$$\begin{aligned} f_c^{\text{ST}}(\Sigma, R) &= 0 \text{ if} |R| < \lambda\Sigma && (16.88)\\ &= \Sigma \text{ otherwise} && (16.89) \end{aligned}$$

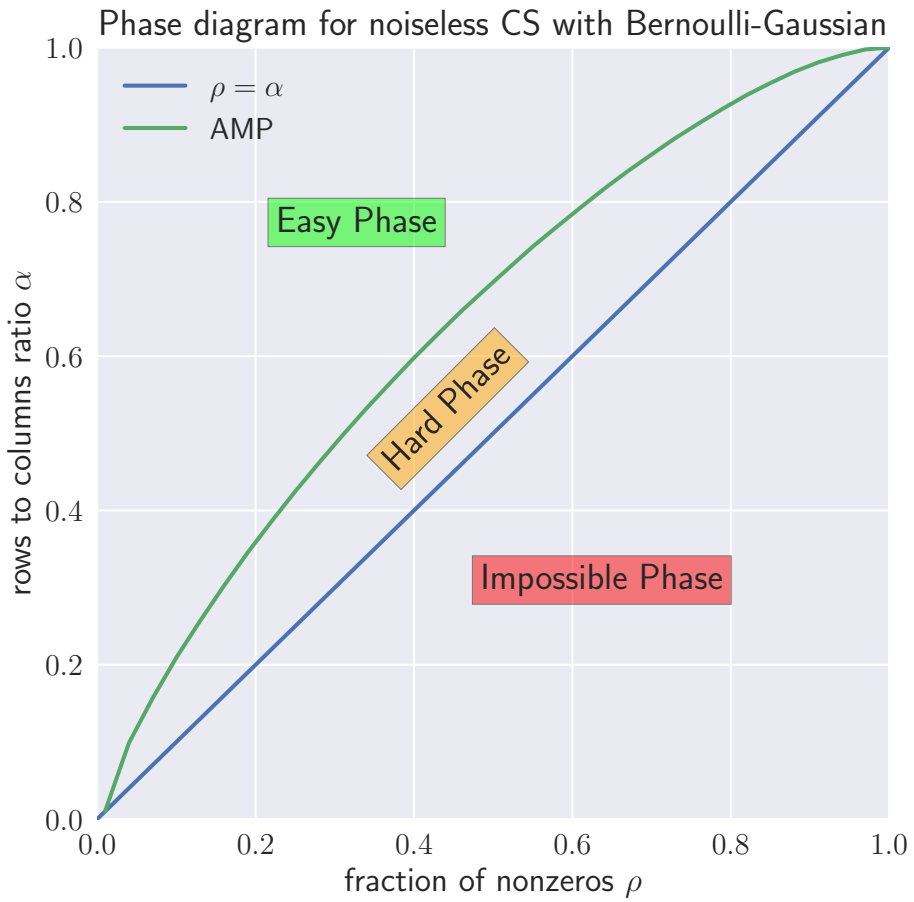


Figure 16.6.1: Phase diagram of noiseless compressed sensing for Gauss-Bernulli prior

- In easy phase, AMP can find the optimal solution.
- In hard phase, it is possible to find the optimal solution, for example by exhaustive search, but AMP gives the wrong answer because it get stuck at a local minima.
- In impossible phase, no algorithm can find the solution.

Conjecture: In the "hard" phase, no algorithm works on polynomial time.

Bibliography

Donoho et al. (2009), Krzakala et al. (2012) and Zdeborová and Krzakala (2016).

16.7 Exercises

Exercise 16.1: AMP for sparse estimation

Consider a N-dimensional sparse bernoulli-signal $\mathbf{x_0}$ where each element is sampled from

$$x_0 \sim (1-\rho)\delta(x) + \rho\mathcal{N}(0,1) \tag{16.96}$$

Say you measure $\mathbf{y} = F\mathbf{x_0}$ where F is a $M \times N$ random matrix as in the main chapter. Run both the Bayes optimal and the LASSO version of AMP for the problem of estimating $\mathbf{x_0}$ from $\mathbf{y}$ for different values of ρ and α, using moderate values of N (few hundreds) Show that both algorithm can be efficient.
Bonus point: compare with the state evolution of each algorithm.

Input: y
Initialize: $\mathbf{a}_{i\to\mu}(t=0)$,$\mathbf{v}_{i\to\mu}(t=0)$, $\mathrm{t}=1$
repeat
r-BP Update of $\{\omega_{\mu\to i}, V_{\mu\to i}\}$

$$V_{\mu\to i}(t) \leftarrow \sum_{j\neq i} F_{\mu j}^2 v_{j\to\mu}(t-1) \tag{16.15}$$

$$\omega_{\mu\to i}(t) \leftarrow \sum_{j\neq i} F_{\mu j} a_{j\to\mu}(t-1) \tag{16.16}$$

r-BP Update of $\{A_{\mu\to i}, B_{\mu\to i}\}$

$$B_{\mu\to i}(t) \leftarrow g_{\rm out}(\omega_{\mu\to i}(t), y_\mu, V_{\mu\to i}(t))\, F_{\mu i}\,, \tag{16.17}$$

$$A_{\mu\to i}(t) \leftarrow -\partial_\omega g_{\rm out}(\omega_{\mu\to i}(t), y_\mu, V_{\mu\to i}(t)) F_{\mu i}^2 \tag{16.18}$$

r-BP Update of $\{R_{\mu\to i}, \Sigma_{\mu\to i}\}$

$$\Sigma_{i\to\mu}(t) \leftarrow \frac{1}{\sum_{\nu\neq\mu} A_{\nu\to i}(t)} \tag{16.19}$$

$$R_{i\to\mu}(t) \leftarrow \Sigma_{i\to\mu}(t) \sum_{\nu\neq\mu} B_{\nu\to i}(t) \tag{16.20}$$

AMP Update of the estimated partial marginals $\mathbf{a}_{i\to\mu}(t)$,$\mathbf{v}_{i\to\mu}(t)$

$$a_{i\to\mu} \leftarrow f_a\left(\Sigma_{i\to\mu}(t), R_{i\to\mu}(t)\right)\,, \tag{16.21}$$

$$v_{i\to\mu} \leftarrow f_v\left(\Sigma_{i\to\mu}(t), R_{i\to\mu}(t)\right)\,. \tag{16.22}$$

$\mathrm{t} \leftarrow \mathrm{t}+1$
until Convergence on $\mathbf{a}_{i\to\mu}(t)$,$\mathbf{v}_{i\to\mu}(t)$
output: Estimated marginals mean and variances:

$$a_i \leftarrow f_a\left(\frac{1}{\sum_\nu A_{\nu\to i}}, \frac{\sum_\nu B_{\nu\to i}}{\sum_\nu A_{\nu\to i}}\right)\,, \tag{16.23}$$

$$v_i \leftarrow f_v\left(\frac{1}{\sum_\nu A_{\nu\to i}}, \frac{\sum_\nu B_{\nu\to i}}{\sum_\nu A_{\nu\to i}}\right)\,. \tag{16.24}$$

Algorithm 1: relaxed Belief-Propagation (r-BP)

Input: $\mathbf{y}$
Initialize: $\mathbf{a}^0$,$\mathbf{v}^0$, $\mathrm{t} = 1$ $g^0_{\mathrm{out},\mu}$
repeat
AMP Update of ω_μ, V_μ

$$V_\mu^t \leftarrow \sum_i F_{\mu i}^2 v_i^{t-1} \tag{16.64}$$

$$\omega_\mu^t \leftarrow \sum_i F_{\mu i} a_i^{t-1} - V_\mu^t g_{\mathrm{out}}^{t-1} \tag{16.65}$$

AMP Update of Σ_i, R_i and $g_{\mathrm{out},\mu}$

$$\Sigma_i^t \leftarrow \left[-\sum_\mu F_{\mu i}^2 \partial_\omega g_{\mathrm{out}}(\omega_\mu^t, y_\mu, V_\mu^t) \right]^{-1} \tag{16.66}$$

$$R_i^t \leftarrow a_i^{t-1} + (\Sigma_i^t)^{-1} \sum_\mu F_{\mu i} g_{\mathrm{out}}(\omega_\mu^t, y_\mu, V_\mu^t) \tag{16.67}$$

AMP Update of the estimated marginals a_i, v_i

$$a_i^{t+1} \leftarrow f_a(\Sigma_i, R_i^{t+1},) \tag{16.68}$$

$$v_i^{t+1} \leftarrow f_v(\Sigma_i, R_i^{t+1}) \tag{16.69}$$

$\mathrm{t} \leftarrow \mathrm{t} + 1$
until Convergence on $\mathbf{a}$,$\mathbf{v}$
output: $\mathbf{a}$,$\mathbf{v}$.

Algorithm 2: Generalized Approximate Message Passing (G-AMP)

Input: $\mathbf{y}$
Initialize: $\mathbf{a}^0$,$\mathbf{v}^0$, $\mathrm{t} = 1$ $g^0_{\mathrm{out},\mu}$
repeat
AMP Update of ω_μ, V_μ

$$V^t \leftarrow \frac{1}{N}\sum_i v_i^{t-1} \tag{16.76}$$

$$\omega_\mu^t \leftarrow \sum_i F_{\mu i} a_i^{t-1} - V_\mu^t g_{\mathrm{out}}^{t-1} \tag{16.77}$$

AMP Update of Σ_i, R_i and $g_{\mathrm{out},\mu}$

$$\Sigma^t \leftarrow \frac{\Delta + V^t}{\alpha} \tag{16.78}$$

$$R_i^t \leftarrow a_i^{t-1} + \sum_\mu F_{\mu i}\frac{y_\mu - \omega_\mu}{\alpha} \tag{16.79}$$

AMP Update of the estimated marginals a_i, v_i

$$a_i^{t+1} \leftarrow f_a(\Sigma^t, R_i^{t+1},) \tag{16.80}$$

$$v_i^{t+1} \leftarrow f_v(\Sigma^t, R_i^{t+1}) \tag{16.81}$$

$\mathrm{t} \leftarrow \mathrm{t} + 1$
until Convergence on $\mathbf{a}$,$\mathbf{v}$
output: $\mathbf{a}$,$\mathbf{v}$.

Algorithm 3: Approximate Message Passing (AMP) for Linear estimation

Input: $\mathbf{y}$
Initialize: $\mathbf{a}^0$,$\mathbf{v}^0$, $\mathrm{t} = 1$ $g^0_{\mathrm{out},\mu}$
repeat
AMP Update of ω_μ, V_μ

$$V^t \leftarrow \frac{1}{N}\sum_i v_i^{t-1} \tag{16.90}$$

$$\omega_\mu^t \leftarrow \sum_i F_{\mu i} a_i^{t-1} - V_\mu^t g_{\mathrm{out}}^{t-1} \tag{16.91}$$

AMP Update of Σ_i, R_i and $g_{\mathrm{out},\mu}$

$$\Sigma^t \leftarrow \frac{1+V^t}{\alpha} \tag{16.92}$$

$$R_i^t \leftarrow a_i^{t-1} + \sum_\mu F_{\mu i}\frac{y_\mu - \omega_\mu}{\alpha} \tag{16.93}$$

AMP Update of the estimated marginals a_i, v_i

$$a_i^{t+1} \leftarrow f_a^{\mathrm{ST}}(\Sigma^t, R_i^{t+1},) \tag{16.94}$$

$$v_i^{t+1} \leftarrow f_v^{\mathrm{ST}}(\Sigma^t, R_i^{t+1}) \tag{16.95}$$

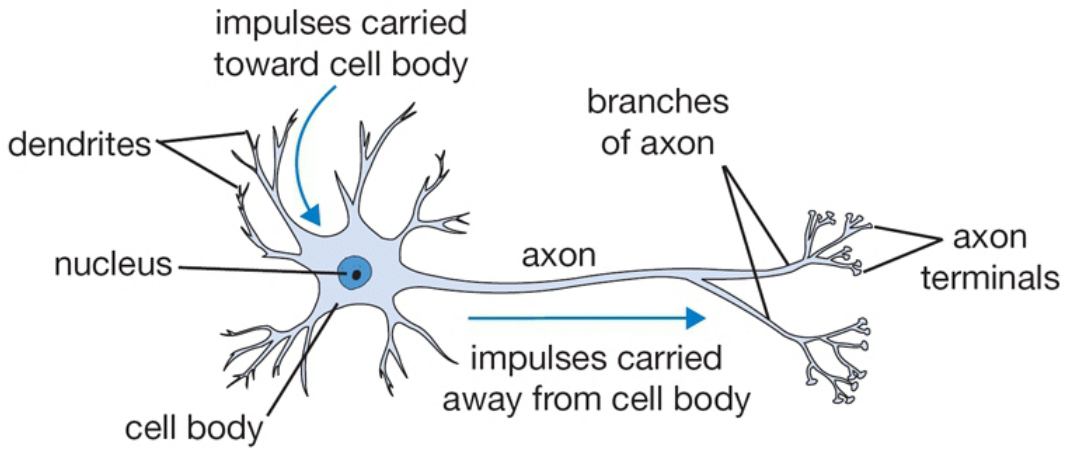
$\mathrm{t} \leftarrow \mathrm{t}+1$
until Convergence on $\mathbf{a}$,$\mathbf{v}$
output: $\mathbf{a}$,$\mathbf{v}$.

Algorithm 4: Approximate Message Passing (AMP) for LASSO

Chapter 17

Perceptron

17.1 Perceptron



1943, McCulloch-Pitts model: $y = \mathsf{Heavyside}(\boldsymbol{\xi} \cdot \mathbf{w} - \kappa)$, where $\boldsymbol{\xi}$ is the signal coming through dendrites, $\mathbf{w}$ is the synaptic weights.

1958, Rosenblatt build McCulloch-Pitts model mechanically: Given set of $\boldsymbol{\xi}$ patterns: $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$ and objectives $y_i \in \{0, 1\}$. Goal is given new pattern $\boldsymbol{\xi}_{\text{new}}$, determine its objective.

1969, Minsky and Papert introduced Perceptron: the McCulloch-Pitts model is leaning a separating hyperplane, and hence cannot learn cases that are not linearly separable, for example, the XOR problem.

But by embedding the points in a higher dimensional space, it is possible to find a separating hyperplane, this is called representation learning.

- Idea 1 – linear embedding: the points will still live in a low-dimensional space, BAD.
- Idea 2 – Nonlinear embedding, e.g. kernels, Deep neuron networks.

Statistical Physics to Perceptron

Now we change the Heavyside function to sign function, since the bijection from $\{0,1\}$ to $\{-1,1\}$ does not change the analysis but in physics it is more common to use ± 1 spins.

Capacity

1988, Derrida, Gartner.

Consider we have M length-N binary patterns $\boldsymbol{\xi}_\mu$, with $\mu = 1, \ldots, M$. The labels $y_\mu \in \{-1, 1\}$ are generated i.i.d.

Goal: learn $\mathbf{w} \in \{-1, +1\}^N$ such that we can make prediction $\hat{y}(\boldsymbol{x}) = \mathsf{sign}(\boldsymbol{\xi} \cdot \mathbf{w})$ and minimize the cost function

$$\mathcal{H}\left(\mathbf{w}; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M\right) = M - \sum_\mu \delta\left(y_\mu, \hat{y}(\boldsymbol{\xi}_\mu)\right)$$

Let $\alpha = M/N$, it is observed that there exists a threshold α_{c} such that at in the large-N limit with α fixed

- When $\alpha < \alpha_{\mathrm{c}}$, w.h.p. there exists some $\mathbf{w}$ so that the cost function $\mathcal{H}\left(\mathbf{w}; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M\right) = 0$.
- When $\alpha > \alpha_{\mathrm{c}}$, w.h.p. for all $\mathbf{w}$ the cost function $\mathcal{H}\left(\mathbf{w}; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M\right) > 0$.

And the storage capacity is defined as $M_{\mathrm{capacity}} = \alpha_{\mathrm{c}} N$.

Replica Method

1989, Knuth, Mezard, Storage capacity of memory networks with binary couplings

Let's consider the partition function of Boltzmann distribution under inverse temperature β

$$Z(\beta; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M) = \sum_{\mathbf{w}} \exp\left(-\beta \mathcal{H}\left(\mathbf{w}; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M\right)\right)$$

As $\beta \to \infty$, the partition function will converge to the number of 'good' $\mathbf{w}$, i.e.

$$\lim_{\beta\to\infty} Z(\beta; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M) = \left|\left\{\mathbf{w} \,\middle|\, \mathcal{H}\left(\mathbf{w}; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M\right) = 0\right\}\right|$$

We are interested in the average free entropy for all possible 'training set' $\{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M$

$$\Phi(\beta, \alpha) = \lim_{N\to\infty} \mathbb{E}_{\{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M} \left[\frac{\log\left(Z(\beta; \{\boldsymbol{\xi}_\mu, y_\mu\}_{\mu=1}^M)\right)}{N}\right]$$

By replica method, we can compute

$$\Phi_{\mathrm{RS}}(\beta,\alpha)=\mathsf{Ext}_{q,\hat{q}}\left\{\frac{1}{2}q\hat{q}-\frac{1}{2}\hat{q}+\int\mathcal{D}z\ \log\left(2\cosh\left(z\sqrt{\hat{q}}\right)\right)+\alpha\int\mathcal{D}z\ \log\left(\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})H\left(z\sqrt{\frac{q}{1-q}}\right)\right)\right\}$$

with

$$H(x)=\int_x^\infty \mathrm{d}u\ \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}=\frac{1}{2}\mathsf{erfc}\left(\frac{x}{\sqrt{2}}\right)$$

It is still an open problem to rigorously prove that $\Phi_{\mathrm{RS}}(\beta,\alpha)$ is correct under $\alpha_{\mathrm{c}}\approx 0.83$, we just believe it is correct from the replica formula. When $\alpha>\alpha_{\mathrm{c}}$, the Φ_{RS} is incorrect, one needs to turn to 1RSB analysis.

Teacher-Student Scenario

- Teacher: there is a $y^*=\mathsf{sign}(\boldsymbol{\xi}\cdot\mathbf{w}^*)$, where $\mathbf{w}^*$ is also binary.
  $\{\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_M\}$ is still random but their corresponding $\{y_1,\dots,y_M\}$ are generated according to the above rule.
- Student: Need to learn the rule
  - There is no capacity here because there is a true $\mathbf{w}^*$.
  - How many patterns does the student need to see to learn the rules?

Treat $\boldsymbol{\xi}_\mu$ are the μ-th row of matrix $\mathbf{F}$, treat $\mathbf{w}^*$ as the true signal $\mathbf{x}^*$ we want to recover, then

$$\mathbf{y}=\mathsf{sign}\left(\mathbf{F}\,\mathbf{x}\right)$$

Let's put this problem in a probabilistic framework. We introduce a Gaussian noise inside the sgn, so as to have a probit likelihood; moreover we will assume that the weights are binary, $\mathbf{x}\in\{+1,-1\}^N$. The full model reads

$$\begin{cases}x_i\sim\dfrac{1}{2}\,\delta(x_i-1)+\dfrac{1}{2}\,\delta(x_i+1),\\ y_\mu\sim\dfrac{1}{2}\,\mathrm{erfc}\,\big(\dfrac{1}{\sqrt{2\sigma^2}}\mathbf{F}_\mu\cdot\mathbf{x}\big).\end{cases}$$

Hence, this problem is same as the generalized linear model, and we can use AMP to solve it.

Let $P_X(x)$ be the prior of x_i and $\rho=\mathbb{E}[x^2]$. Then the replica formula reads

$$\lim_{N\to\infty}\mathbb{E}_{\mathbf{F},\mathbf{y}}\left[\frac{\log\left(Z\left(\mathbf{F},\mathbf{y}\right)\right)}{N}\right]=\mathsf{Ext}_{q,\hat{q}}\,\Phi_{\mathrm{RS}}(q,\hat{q})$$

where

$$\begin{aligned}\Phi_{\mathrm{RS}}(q,\hat{q})=&-\frac{\alpha}{2}q\hat{q}+\alpha\int\mathrm{d}x^*\,\mathrm{d}y\,\mathcal{D}z\ P_{\mathrm{out}}\left(y\,\middle|\,x^*\sqrt{\rho-q}+z\sqrt{q}\right)\times\log\left\{\int\mathrm{d}x\ P_{\mathrm{out}}\left(y\,\middle|\,x\sqrt{\rho-q}+z\sqrt{q}\right)\right\}\\&+\int\mathrm{d}x^*\,\mathcal{D}z\ P_X(x^*)\mathrm{e}^{-\frac{\alpha\hat{q}}{2}(x^*)^2+zx^*\sqrt{\hat{q}}}\times\log\left\{\int\mathrm{d}x\ P_X(x)\mathrm{e}^{-\frac{\alpha\hat{q}}{2}x^*+zx\sqrt{\hat{q}}}\right\}\end{aligned}$$

q^* is the extremizer of Φ_{RS}, and equals to $\langle\mathbf{x}^*,\hat{\mathbf{x}}\rangle_{\mathbf{F},\psi}$.

AMP State Evolution

$$\hat{q}^{(t)} = -\int \mathrm{d}p\,\mathrm{d}x\,\frac{\exp\left(-\frac{p^2}{2q^{(t)}} - \frac{(x-p)^2}{2\left(\rho-q^{(t)}\right)}\right)}{2\pi\sqrt{q^{(t)}\left(\rho-q^{(t)}\right)}}\int \mathrm{d}y\; P_{\mathrm{out}}\left(y\,|\,x\right)\partial_p g_{\mathrm{out}}\left(p, y, \rho - q^{(t)}\right)$$

$$q^{(t+1)} = \int \mathrm{d}x\; P_X(x)\int \mathcal{D}z\; f_a^2\left(\frac{1}{\alpha\hat{q}^{(t)}}, x + \frac{z}{\sqrt{\alpha\hat{q}^{(t)}}}\right)$$

- Impossible Phase: When $\alpha < 1.245$, $q = 1$ is not the global maxima of Φ_{RS}, so it is impossible to find a solution.

- Hard Phase: When $\alpha \in (1.245, 1.493)$, $q = 1$ is the global maxima, but there is another local maxima at which AMP will get stuck.

- Easy Phase: When $\alpha > 1.493$, $q = 1$ is the only local maxima, AMP will always converge to it.

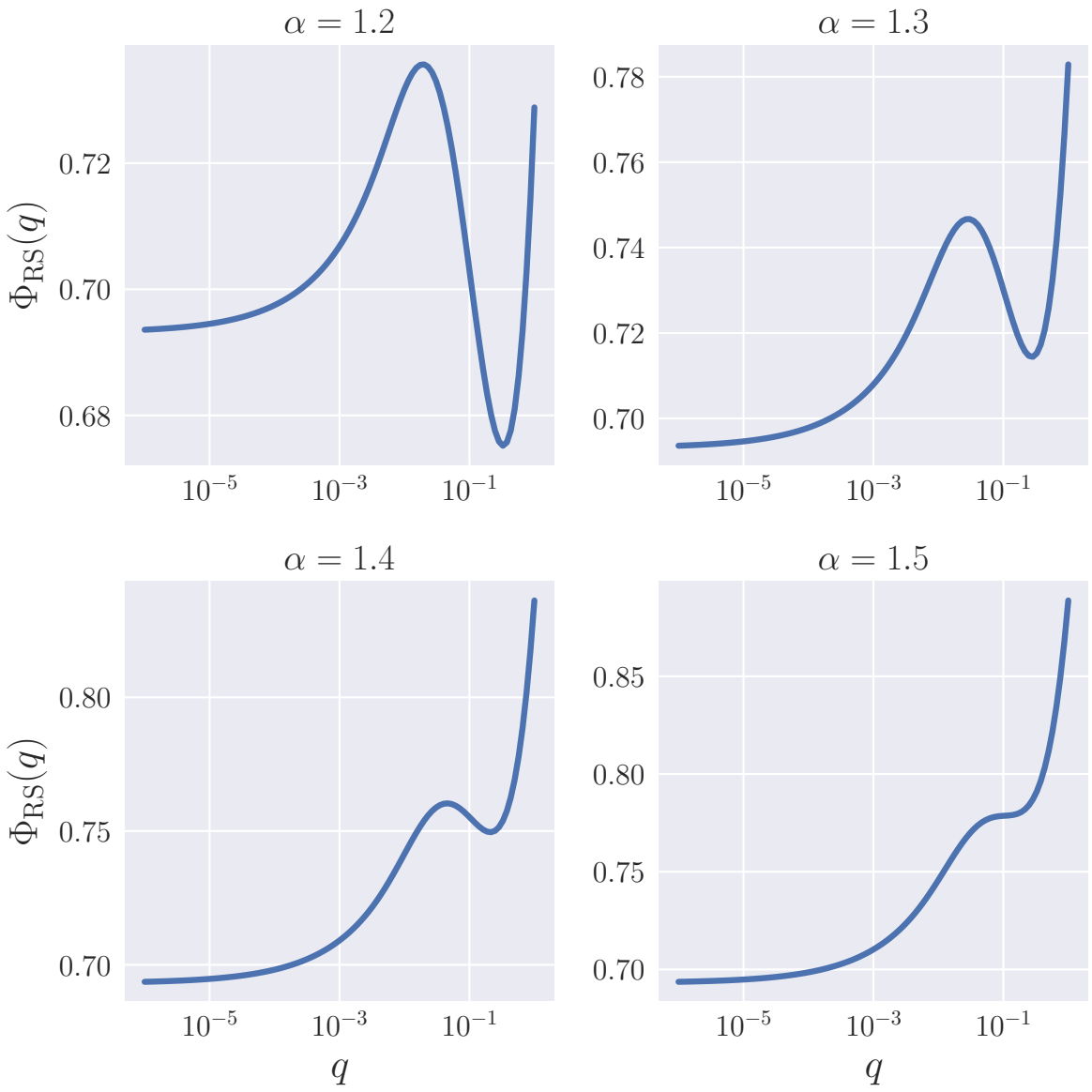


Figure 17.1.1: Free entropy of a perceptron for different values of α

Part V

Appendix

Chapter 18

A bit of probabilty theory

> But wait, if I could shake the crushing weight of expectations? Would that free some room up for joy? Or relaxation? Or simple pleasure? Instead, we measure!
>
> Lin-Manuel Miranda, Surface Pressure, Encanto – 2021

Bibliography

Abbe, E. and Montanari, A. (2013). Conditional random fields, planted constraint satisfaction and entropy concentration. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 332–346. Springer.

Achlioptas, D. and Coja-Oghlan, A. (2008). Algorithmic barriers from phase transitions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 793–802. IEEE.

Achlioptas, D. and Moore, C. (2003). Almost all graphs with average degree 4 are 3-colorable. *Journal of Computer and System Sciences*, 67(2):441–471.

Aizenman, M., Sims, R., and Starr, S. L. (2003). Extended variational principle for the sherrington-kirkpatrick spin-glass model. *Physical Review B*, 68(21):214403.

Aubin, B., Loureiro, B., Maillard, A., Krzakala, F., and Zdeborová, L. (2020). The spiked matrix model with generative priors. *IEEE Transactions on Information Theory*.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Baik, J., Arous, G. B., Péché, S., et al. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697.

Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T., and Zdeborová, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *arXiv preprint arXiv:1606.04142*.

Barbier, J. and Macris, N. (2019). The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability theory and related fields*, 174(3):1133–1185.

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

Berthier, R., Montanari, A., and Nguyen, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79.

Bethe, H. A. (1935). Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 150(871):552–575.

Bolthausen, E. (2009). On the high-temperature phase of the sherrington-kirkpatrick model. In *Seminar at EURANDOM, Eindhoven*.

Bolthausen, E. (2014). An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Braunstein, A., Dall'Asta, L., Semerjian, G., and Zdeborová, L. (2016). The large deviations of the whitening process in random constraint satisfaction problems. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(5):053401.

Chertkov, M. (2008). Exactness of belief propagation for some graphical models with loops. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10016.

Chertkov, M. and Chernyak, V. Y. (2006). Loop series for discrete statistical models on graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06):P06009.

Coja-Oghlan, A. (2013). Upper-bounding the k-colorability threshold by counting covers. *arXiv preprint arXiv:1305.0177*.

Coja-Oghlan, A., Krzakala, F., Perkins, W., and Zdeborová, L. (2018). Information-theoretic thresholds from the cavity method. *Advances in Mathematics*, 333:694–795.

Coja-Oghlan, A. and Vilenchik, D. (2013). Chasing the k-colorability threshold. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 380–389. IEEE.

Cover, T. M. and Thomas, J. A. (1991). Information theory and statistics. *Elements of Information Theory*, 1(1):279–335.

Curie, P. (1895). *Propriétés magnétiques des corps a diverses températures*. Number 4. Gauthier-Villars et fils.

de Almeida, J. R. and Thouless, D. J. (1978). Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983.

Debye, P. (1009). Näherungsformeln für die zylinderfunktionen für große werte des arguments und unbeschränkt veränderliche werte des index. *Math. Ann.*, 67:535–558.

Dembo, A., Montanari, A., et al. (2010a). Gibbs measures and phase transitions on sparse random graphs. *Brazilian Journal of Probability and Statistics*, 24(2):137–211.

Dembo, A., Montanari, A., et al. (2010b). Ising models on locally tree-like graphs. *The Annals of Applied Probability*, 20(2):565–592.

Dembo, A., Montanari, A., Sly, A., and Sun, N. (2014). The replica symmetric solution for potts models on d-regular graphs. *Communications in Mathematical Physics*, 327(2):551–575.

Dembo, A., Zeltouni, O., and Fleischmann, K. (1996). Large deviations techniques and applications. *Jahresbericht der Deutschen Mathematiker Vereinigung*, 98(3):18–18.

Derrida, B. (1980). Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, 45(2):79.

Derrida, B. (1981). Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613.

Donoho, D. L., Johnstone, I. M., et al. (1998). Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921.

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

Edwards, S. F., Goldbart, P. M., Goldenfeld, N., Sherrington, D. C., and Edwards, S. F., editors (2005). *Stealing the gold: a celebration of the pioneering physics of Sam Edwards*. Number 126 in International series of monographs on physics. Clarendon Press ; Oxford University Press, Oxford : New York.

Edwards, S. F. and Jones, R. C. (1976). The eigenvalue spectrum of a large symmetric random matrix. *Journal of Physics A: Mathematical and General*, 9(10):1595.

El Alaoui, A. and Krzakala, F. (2018). Estimation in the spiked wigner model: A short proof of the replica formula. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1874–1878. IEEE.

Gallager, R. (1962). Low-density parity-check codes. *IRE Transactions on information theory*, 8(1):21–28.

Gerbelot, C. and Berthier, R. (2021). Graph-based approximate message passing iterations. *arXiv preprint arXiv:2109.11905*.

Gross, D. J. and Mézard, M. (1984). The simplest spin glass. *Nuclear Physics B*, 240(4):431–452.

Guerra, F. (2003). Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12.

Guo, D., Shamai, S., and Verdú, S. (2005). Mutual information and minimum mean-square error in gaussian channels. *IEEE transactions on information theory*, 51(4):1261–1282.

Iba, Y. (1999). The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21):3875.

Jaeger, G. (1998). The ehrenfest classification of phase transitions: Introduction and evolution. *Arch Hist Exact Sc.*, 53:51–81.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327.

Kadanoff, L. P. (2009). More is the same; phase transitions and mean field theories. *Journal of Statistical Physics*, 137(5):777–797.

Kesten, H. and Stigum, B. P. (1967). Limit theorems for decomposable multi-dimensional galton-watson processes. *Journal of Mathematical Analysis and Applications*, 17(2):309–338.

Korada, S. B. and Macris, N. (2009). Exact solution of the gauge symmetric p-spin glass model on a complete graph. *Journal of Statistical Physics*, 136(2):205–230.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. (2012). Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009.

Krzakala, F., Xu, J., and Zdeborová, L. (2016). Mutual information in rank-one matrix estimation. In *2016 IEEE Information Theory Workshop (ITW)*, pages 71–75. IEEE.

Laplace, P. S. d. (1774). Memoire sur les probabilites des causes par lesevenements.

Lelarge, M. and Miolane, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929.

Lesieur, T., Krzakala, F., and Zdeborová, L. (2017). Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403.

Livan, G., Novaes, M., and Vivo, P. (2018). *Introduction to random matrices: theory and practice*, volume 26. Springer.

McGrayne, S. B. (2011). *The theory that would not die*. Yale University Press.

Mézard, M. and Parisi, G. (1999). Thermodynamics of glasses: A first principles computation. *Journal of Physics: Condensed Matter*, 11(10A):A157.

Mézard, M. and Parisi, G. (2001). The bethe lattice spin glass revisited. *The European Physical Journal B-Condensed Matter and Complex Systems*, 20(2):217–233.

Mézard, M. and Parisi, G. (2003). The cavity method at zero temperature. *Journal of Statistical Physics*, 111(1):1–34.

Mézard, M., Parisi, G., Sourlas, N., Toulouse, G., and Virasoro, M. (1984). Nature of the spin-glass phase. *Physical review letters*, 52(13):1156.

Mézard, M., Parisi, G., and Virasoro, M. (1987a). Sk model: The replica solution without replicas. *SPIN GLASS THEORY AND BEYOND: AN INTRODUCTION TO THE REPLICA METHOD AND ITS APPLICATIONS. Edited by MEZARD M ET AL. Published by World Scientific Press*, pages 232–237.

Mézard, M., Parisi, G., and Virasoro, M. A. (1987b). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company.

Mézard, M., Parisi, G., and Zecchina, R. (2002). Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815.

Miolane, L. (2017). Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*.

Monasson, R. (1995). Structural glass transition and the entropy of the metastable states. *Physical review letters*, 75(15):2847.

Nattermann, T. (1998). Theory of the random field ising model. In *Spin glasses and random fields*, pages 277–298. World Scientific.

Nishimori, H. (1980). Exact results and critical properties of the ising model with competing interactions. *Journal of Physics C: Solid State Physics*, 13(21):4071.

Nishimori, H. (1993). Optimum decoding temperature for error-correcting codes. *Journal of the Physical Society of Japan*, 62(9):2973–2975.

Opper, M. and Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press.

Parisi, G. (1979). Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754.

Parisi, G. (1980). A sequence of approximated solutions to the sk model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115.

Parisi, G. (1983). Order parameter for spin-glasses. *Physical Review Letters*, 50(24):1946.

Pearl, J. (1982). *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science ....

Peierls, R. (1936). Statistical theory of adsorption with interaction between the adsorbed atoms. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 32, pages 471–476. Cambridge University Press.

Potters, M. and Bouchaud, J.-P. (2020). *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge University Press.

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

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

Ruozzi, N. (2012). The bethe partition function of log-supermodular graphical models. *Advances in Neural Information Processing Systems*, 25.

Saade, A., Krzakala, F., and Zdeborová, L. (2017). Spectral bounds for the ising ferromagnet on an arbitrary given graph. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(5):053403.

Schwarz, A., Bluhm, J., and Schröder, J. (2020). Modeling of freezing processes of ice floes within the framework of the tpm. *Acta Mechanica*, 231.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Taillefer, L. (2010). Scattering and pairing in cuprate superconductors. *Annual Review of Condensed Matter Physics*, 1.

Thouless, D. J., Anderson, P. W., and Palmer, R. G. (1977). Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601.

Touchette, H. (2009). The large deviation approach to statistical mechanics. *Physics Reports*, 478(1):1–69.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Watkin, T. L. H., Rau, A., and Biehl, M. (1993). The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499–556.

Weiss, P. (1907). L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.*, 6(1):661–690.

Weiss, P. (1948). The application of the bethe-peierls method to ferromagnetism. *Physical Review*, 74(10):1493.

Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, pages 325–327.

Willsky, A., Sudderth, E., and Wainwright, M. J. (2007). Loop series and bethe variational bounds in attractive graphical models. *Advances in neural information processing systems*, 20.

Yedidia, J. S., Freeman, W. T., Weiss, Y., et al. (2003). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236-239):0018–9448.

Zdeborová, L. and Krzakala, F. (2016). Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552.

Todo list