Statistical Physics of Inference and Bayesian Estimation

Florent Krzakala, Lenka Zdeborova, Maria Chiara Angelini and Francesco Caltagirone

Contents

1 Bayesian Inference and Estimators

Inference and data estimation is a fundamental interdisciplinary topic with many practical application. The problem of inference is the following: we have a set of observations $\mathbf{y}$, produced in some way (possibly noisy) by an unknown signal $\mathbf{s}$. From them we want to estimate the signal $\vec{s}$. To be concrete, we have

$$\vec{y} = f(\vec{s}; \text{noise}) \,, \tag{1}$$

and the objective is to produce an estimation $\hat{\mathbf{s}} = \hat{\mathbf{s}}(y)$ that is (hopefully) accurate under some metric.

Inference is a huge field and different approaches are possible. It can be regarded as a subfield of statistics, and lies at the merging of a number of areas of science and engineering, including data mining, machine learning, signal processing, and inverse problems. Each of these disciplines provides some information on how to model data acquisition, computation, and how best to exploit the hidden structure of the problem of interest.

Numerous techniques and algorithms have been developed over a long period of time, and they often differ in the assumptions and the objectives that they try to achieve. As an example, a few major distinctions to keep in mind are the following.

Parametric versus non-parametric In parametric estimation, stringent assumptions are made about the unknown object, hence reducing $\mathbf{s}$ to be determined by a small set of parameters. In contrast, non-parametric estimation strives to make minimal modeling assumptions, resulting in θ being an high-dimensional or infinite-dimensional object (for instance, a function).

Bayesian versus frequentist The Bayesian approach assumes $\mathbf{s}$ to be a random variable as well, whose 'prior' distribution plays an obviously important role. From a frequentist point of view, $\mathbf{s}$ is instead an arbitrary point in a set of possibilities. In these lectures we shall mainly follow the Bayesian point of view, as it fit more naturally the statistical physics approach, but the two are in fact closely related.

Statistical efficiency versus computational efficiency Within classical estimation theory, a specific estimator $\hat{\mathbf{s}}$ is mainly evaluated in terms of its accuracy: How close (or far) is $\hat{\mathbf{s}}(y)$ to $\mathbf{s}$ for typical realizations of the noise? We can broadly refer to this figure of merit as to 'statistical efficiency.' Within modern applications, computational efficiency has arisen as a second central concern. Indeed $\mathbf{s}$ is often high-dimensional: it is not uncommon to fit models with millions of parameters. The amounts of observations has grown in parallel. It becomes therefore crucial to devise estimators whose complexity scales gently with the dimensions, and with the amount of data.

These lectures will focus on a Bayesian-parametric approach and will talk mainly about performance analysis (existence and study of phase transitions), and a bit about the analysis of some algorithms.

1.1 The Bayes formula

The Bayesian inference makes use of the Bayes formula, written for the first time by Rev. Thomas Bayes (1702 - 1762). Indicating with $P(A|B)$ the probability of having an event A conditioned to the event B, the Bayes formula states that we can extract $P(A|B)$ from the knowledge of $P(B|A)$ simply as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

Translating this formula in the contest of statistical inference, if we know:

- $P(\vec{y}|\vec{s})$, often called *likelihood*, that is the probability of having a certain observation $\vec{y}$ given a signal $\vec{s}$
- $P(\vec{s})$, called *prior probability*, that is the probability of the signal

we extract the probability of the signal given the observation, called *posterior probability* as

$$P(\vec{s}|\vec{y}) = \frac{P(\vec{y}|\vec{s})P(\vec{s})}{Z} \tag{3}$$

where Z is just the renormalization of the probability. In the prior we should insert the knowledge that we have about the signal. If we don't have any information, we can simply take a uniform distribution. Once we have $P(\vec{s}|\vec{y})$, the last thing to do is to extract $\hat{\vec{x}}$ that is the estimate of the signal $\vec{s}$. Many kinds of estimators exist. We will analyze some of them in the following.
However firstly we apply what we have said to a real example.

Example 1: Inferring a decay constant (from Ref. [1])

Unstable particles are emitted from a source and decay at a distance y, a real number that has an exponential probability distribution with characteristic length λ. Decay events can only be observed if they occur in a window extending from $y = 1cm$ to $x = 20cm$. M decays are observed at locations $\{y_1, ..., y_M\}$. What is λ?

Writing things in the inference language, $\vec{y}$ are our observations and λ is our signal $\vec{s}$ of dimension $N = 1$. We know

$$P(y|\lambda) = \begin{cases} \frac{1}{\lambda} e^{-y/\lambda}/Z(\lambda) & \text{if } 1 < y < 20 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

with $Z(\lambda) = \int_1^{20} dx \frac{1}{\lambda} e^{-x/\lambda} = e^{-1/\lambda} - e^{-20/\lambda}$. From eq. (4) we can extract $P(\lambda|\vec{y})$ using the Bayes formula:

$$P(\lambda|\vec{y}) \propto \frac{1}{(\lambda Z(\lambda))^M} e^{-\sum_{i=1}^{M} y_i/\lambda} P(\lambda) \tag{5}$$

If we do not have any prior information on λ, we assume that $P(\lambda)$ is a constant that just enters in the normalization. $P(\lambda|\vec{y})$ is the final answer from Bayesian statistics. It contains all the information that we have on λ in this approach.

For a dataset consisting of several points, e.g., the six points $x = 1.5, 2, 3, 4, 5, 12$, the likelihood function is shown in the following plot, and it has a maximum in 3.7. This estimation of λ is called the maximum

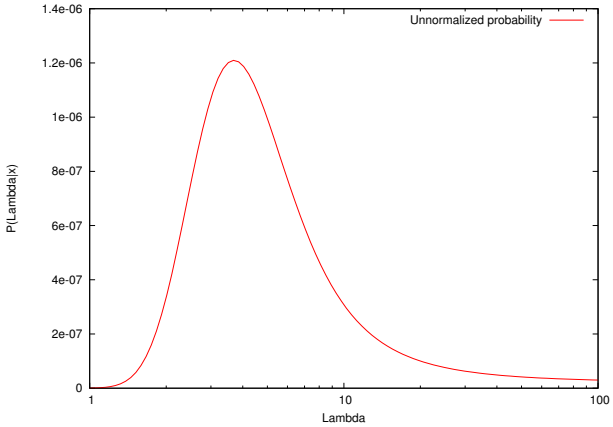

Figure 1

likelihood (ML) estimator, since it maximizes the so-called likelihood $P(x|\lambda)$. However, the way we derived it was instead to maximize the posterior probability $P(\lambda|x)$, and it turned out equal to maximum likelihood just because we used a uniform distribution for the prior $P(\lambda)$.

Probabilities are used here to quantify degrees of belief. To avoid possible confusions, it must be emphasized that λ is not a stochastic variable, and the fact that the Bayesian approachuses a probability distribution does not mean that we think of the world as stochastically changing its nature between the states described by the different hypotheses. The notation of probabilities is used here to represent the beliefs about the mutually exclusive hypotheses (here, values of λ), of which only one is actually true. That probabilities can denote degrees of belief, given assumptions, is at the hearth of Bayesian inference.

1.2 Probabilty reminder

1.2.1 A bit of probabilities

$P(A|B)P(B) = P(A, B)$
$\sum_A P(A|B) = 1.$
$\sum_B P(A|B) = something???.$
$\sum_B P(A, B) = P(A)$ (marginalisation)
$< A >= \sum_A AP(A)$ (mean)
$< A^2 > - < A >^2$ (variance)

1.2.2 Probability distribution

$\int dx P(x) = 1$
e.g. Gaussian distribution (or normal)
$\sum_X P(x) = 1$
e.g. Poisson distribution
$P(k) = frac\lambda^k e^{-\lambda} k!$
mean and variance are given by lambda!
Exponentielle
$P(x) = frac e^{-x/\lambda}\lambda$
mean is $lambda$ and variance is λ^2

1.2.3 A bit of random variables

Markov inequality : if x is a random positive variable $P(x \geq a) \leq \frac{\mathbb{E}(x)}{a}$.
Chebyshevs: $P(|X - m| \geq k\sigma) \leq \frac{1}{k^2}$

1.3 Estimators

Now that probabilty is no longer a problem for us, let us come back to the problem. Is is clear that we should have took the maximum probability? We could want to go further and infer a different value for λ. How so?

Maximum a posteriori estimator (MAP). Maybe the simplest answer to the previous question is to estimate λ as the value that maximizes the posterior probability, in formula:

$$\hat{x}_{\mathrm{MAP}} = \mathrm{argmax}_s P(\vec{s}|\vec{y}) \tag{6}$$

This is reasonable in some cases, however what about a situation in which the posterior distribution is the one in Fig. 2? The MAP estimator $\hat{\lambda}_{\mathrm{MAP}}$ chooses the value of λ that corresponds to the peak, but indeed, extracting randomly λ from that distribution, the number of times in which we will extract λ_{MAP} is really small. Thus at least in this case the MAP estimator doesn't seem a good one.

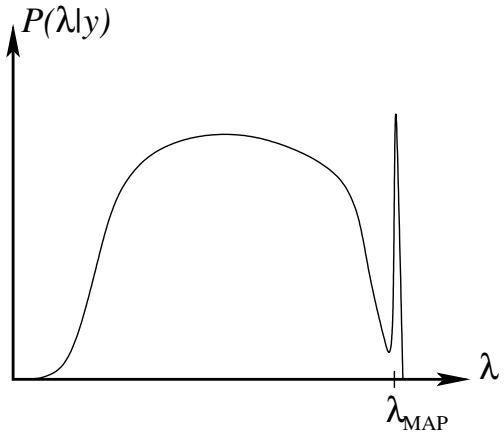

Figure 2: Example of a strange a posteriori probability and maximum a posteriori estimator

Minimal mean square error estimator (MMSE). A good estimator should have a low mean square error (MSE) between the real signal and the estimation, defined as:

$$\text{MSE} = \sum_{i=1}^{N} \frac{(x_i - s_i)^2}{N} \tag{7}$$

Unfortunately we can not calculate the MSE because we do not know $\vec{s}$. However we know $P(\vec{s}|\vec{y})$ and we can estimate the risk for the MSE, defined as:

$$\mathbb{E}\,(\text{MSE}) = \int d\vec{s} P(\vec{s}|\vec{y}) \sum_{i=1}^{N} \frac{(x_i - s_i)^2}{N} \tag{8}$$

If we try to minimize the risk for MSE imposing $\frac{d}{dx_i}\mathbb{E}(\text{MSE}) = 0$, we end up with the MMSE estimator:

$$\hat{x_i}_{\text{MMSE}} = \int ds_i P(s_i|\vec{y}) s_i \tag{9}$$

where $P(s_i|\vec{y}) = \int \prod_{j\neq i} P(\vec{s}|\vec{y})$ is the marginal probability of the variable s_i. Thus the MMSE estimator is just the mean of the posterior probability component-wise. We can see that in the case of Fig. 2, this estimator is more reasonable than the MAP one.

We can also choose to minimize other quantities than the MSE. For example if we choose to minimize the risk of the observable $\sum_{i=1}^{N} \frac{|x_i - s_i|}{N}$ we end up with an estimator that is component-wise the median of the marginals.

Minimal error assignments estimator (MARG). If we have a problem in which the signal can take only discrete values, we can define the number of errors as:

$$\text{num. of errors} = \sum_{i=1}^{N} \frac{1 - \delta(s_i, x_i)}{N} \tag{10}$$

If we minimize the risk on the number of errors:

$$\mathbb{E}\,(\text{errors}) = \int d\vec{s} P(\vec{s}|\vec{y}) \sum_{i=1}^{N} \frac{1 - \delta(s_i, x_i)}{N} \tag{11}$$

we obtain the so called minimal error assignment estimator:

$$\hat{x}_{i\mathrm{MARG}} = \mathrm{argmax}_{s_i} P(s_i|\vec{y}) \tag{12}$$

that corresponds to the optimal Bayesian decision.

This estimator is deeply different from the MAP one. In fact for the MARG we take for each component the value that maximizes the marginal of the posterior probability on that component, the operation is thus done component-wise, while for the MAP one we chose the point $\vec{s}$ that maximizes the whole probability. In a statistical physics approach we can say that the difference between the MAP estimator and the MMSE or MARG ones is the same difference that occurs between the minimization of the energy or of the free-energy. The MAP approach maximizes the total probability, but does not take into account the "entropic effects". We can make the connection with statistical physics more clear rewriting eq. (3) as:

$$P(\vec{s}|\vec{y}) = \frac{e^{\log(P(\vec{y}|\vec{s}))+\log(P(\vec{s}))}}{Z} \tag{13}$$

where $P(\vec{s}|\vec{y})$ is interpreted as the Boltzmann weight, the argument of the exponential is the Hamiltonian (assuming $\beta = -1$) and Z is the partition function. In this language the marginal probability $P(s_i|\vec{y})$ has the role of the magnetization.

1.4 A toy example in denoising

A typical example of inference is the denoising problem. A signal is transmitted trough a channel that is noisy. We want to estimate the signal from the data measured at the end of the channel, namely from the signal plus the noise.

1.4.1 Phase transition in an easy example

As first example we restrict the problem to an easy one: the original signal $\vec{s}$, of length N, is sparse and in particular only the i-th element is different from zero:

$$s_j = \begin{cases} x^* & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases} \tag{14}$$

In the transmission, Gaussian noise is added. Thus we will observe a random signal $\vec{y}$ with probability

$$P(y_j = E) = \begin{cases} \delta(E - x^*) & \text{if } j = i \\ \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{E^2}{2\Delta}} & \text{if } j \neq i \end{cases} \tag{15}$$

For simplicity we did not add noise on the i-th component of the signal. The average number of events that lie in a small interval δE is:

$$\mathcal{N}(E, E+\delta E) = \frac{N}{\sqrt{2\pi\Delta}} e^{-\frac{E^2}{2\Delta}} \delta E \propto e^{\log(N)\left(1-\frac{E^2}{2\Delta\log N}\right)}. \tag{16}$$

Defining $E_c = \sqrt{2\Delta \log N}$, the exponent is negative for $|E| > |E_c|$ and positive otherwise. Thus in the large N limit, the average number of events (and consequently the probability of having at least one event) is zero for $|E| > |E_c|$. One can define an entropy that in the large N limit takes values:

$$s(E) = \log(\mathcal{N}) = \begin{cases} \log(N)\left(1 - \frac{E^2}{2\Delta\log N}\right) & \text{if } |E| < |E_c| \\ 0 & \text{if } |E| > |E_c| \end{cases} \tag{17}$$

and has the form shown in Fig. 3. Thus if the only non trivial component of the initial signal is larger

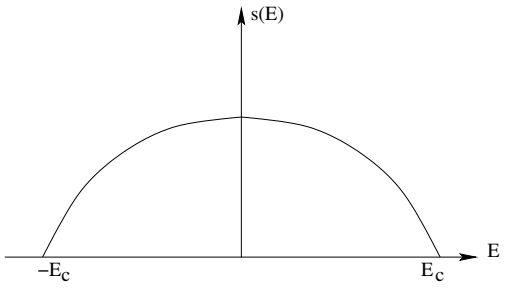

Figure 3: Entropy for the denoising toy problem as defined in eq. (17)

than the threshold, namely $|x^*| > E_c$, we are able to recognize it. For this reason in Ref. [2] Donoho and Jhonstone proposed the following *universal thresholding*: if $y_i = E$ with $|E| < E_c$, the estimated signal is $\hat{s}_i = 0$, because there is no way to understand if the observation is signal or noise; if $|y_i| > E_c$, the estimated signal is $\hat{s}_i = y_i$.

We have seen that even in this simple example we can clearly identify two phases: the first one, when $|x^*| > E_c$, is the *easy* phase for the reconstruction of the signal, the second one, when $|x^*| < E_c$, is the *hard* or impossible phase and the passage from one to another is sharp. Indeed it is a real phase transition.

1.4.2 The connection with Random Energy Model

The expert reader will have already recognized the connection with the Random Energy Model (REM), introduced many years before Donoho and Johnstone by Derrida in Ref. [3]. In the REM we have n Ising spins, thus the total number of configurations of the whole system is $N = 2^n$. The energies associated to these configurations are independent random variables extracted from a distribution $P(E) = \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{E^2}{2\Delta}}$, with $\Delta = n/2$. The Boltzmann weight of a certain configuration at temperature T (and inverse temperature β) is $P(E) = \frac{e^{-\beta E}}{Z}$. The entropy of this model is exactly the one in eq. (17). Now, using the thermodynamical relation $\frac{ds}{dE} = \frac{1}{T}$ we can obtain the equilibrium energy at a given temperature: $E(T) = -\frac{\Delta}{T}$. This is valid only when $E > -E_c$, in fact we know that in the thermodynamic limit there are no states with $E < -E_c$. Thus there is a critical temperature $T_c = \frac{1}{2\sqrt{\log(2)}}$, obtained imposing $E(T) = -E_c$, below which the equilibrium energy is always E_c. Thus the Gibbs measure for $T < T_c$ condensates only on a sub-extensive number of states (because $s(E_c) = 0$). For $T < T_c$ we are in the so called *glassy phase*.

We have seen that the critical energy that we found for the denoising problem is the same critical energy that is linked to the glass transition in the REM. This is not an isolated case. For each inference problem the critical threshold between the hard and the easy phases is linked to the critical point between paramagnetic and spin-glass phase of the associated statistical-physics disordered model. We will study deeply this link in the next section.

2 Taking averages: quenched, annealed and planted ensembles

In statistical physics of disordered systems, we very often face —by definition— the following situation: we have a Hamiltonian with spin variables $S = \pm 1$ that contains some disordered quantities whose distribution we know. For instance, one can think about the seminal case of spin glasses where the Edwards-Anderson (EA) [4] Hamiltonian reads:

$$\mathcal{H} = -\sum_{\langle i,j \rangle} J_{ij} S_i S_j \tag{18}$$

where the sum is over all pair of spins on a given graph. In general, we know that the J_{ij} are taken random from a given distribution, say for instance $P(J) = \frac{1}{2}\delta(J-1) + \frac{1}{2}\delta(J+1)$. It is then difficult to compute the partition sum, since we do not know the Hamiltonian explicitly but rather the probability of a given one! A first solution is to consider a given instance of the problem, and to study this particular problem. In fact, this is a strategy which, strangely enough, was only followed recently, however it brings statistical physics to another level and allows deep connections with many problems of computer science. A second solution, which we shall follow in this section, is averaging over many realizations of the disorder.

2.1 The Quenched ensemble

How to average is a problem that was solved a long time ago by Edwards himself in his work on spin glasses and vulcanization (see the wonderful book *Stealing the gold* [5]). If one takes a large enough system, he suggested, then the system becomes *self-averaging*: all extensive thermodynamic quantities have the same values (in densities) for almost all realizations of the Hamiltonian. Therefore, one can average them and compute the average free energy

$$f_{\text{quenched}} = \left[\frac{F}{N}\right] = \lim_{N\to\infty} -\frac{1}{\beta N}\left[\log Z\right], \tag{19}$$

where $[\,.\,]$ denotes the average over the disorder. Edwards did not stop here and also suggested (and gave credit to Mark Kac for the original idea) a way to compute the average of the (very tricky) logarithm of Z, known today as the *replica trick*, using the identity:

$$\log Z = \lim_{n\to 0} \frac{Z^n - 1}{n}. \tag{20}$$

The idea here is that, if averaging the logarithm of Z turns out to be difficult, the average of Z^n is maybe doable for any integer value of n, and performing a (risky!!) analytic continuation to $n = 0$, one might compute the averaged free energy over the disorder as

$$f_{\text{quenched}} = -\frac{1}{Nn\beta} \lim_{n\to 0} \left([Z^n] - 1\right). \tag{21}$$

This is called the quenched average, and we shall from now on refer to such computation as the quenched computation. In fact, the self-averaging hypothesis for the free energy has been proven now rigorously in many cases (in particular for all lattices in finite dimension [6] and for mean-field models [7]) for most situations with discrete variables, so this is in fact the correct computation that one should do in physics. It is, however, very difficult to solve this problem in general, and this is at the core of the statistical physics of disordered systems. We shall indeed often encounter this situation in the following.

2.2 The Annealed ensemble

It is *much easier* to consider the so called annealed ensemble. This is a very different computation, and of course, it has no reason to be equal to the quenched computation. In the annealed ensemble, one simply

averages the partition sum and only *then* takes the logarithm:

$$f_{\text{annealed}} = -\frac{1}{N\beta} \log [Z]. \tag{22}$$

It is important to see that this is *wrong* if one wants to do physics. The point is that the free energy is an extensive quantity, so that the free energy per variable should be a quantity of $O(1)$ with fluctuations going in most situation $\propto (1/\sqrt{N})$ (the exponent can be more complex, but the idea is that fluctuation are going to zero as $N \to \infty$.) The partition sum Z, however, is exponentially large in N, and so its fluctuation can be quite large: averaging over them is then far from safe, as the average could be dominated by rare, but large, fluctuations.

Consider for instance the situation where Z is $\exp(-\beta N)$ with probability $1/N$ and $\exp(-2\beta N)$ with probability $1 - 1/N$. With high probability, if one picks up a large system, its free energy should be $f = 2$, however in the quenched computation one finds

$$[Z] = \frac{1}{N} \exp(-\beta N) + \left(1 - \frac{1}{N}\right) \exp(-2\beta N) \tag{23}$$

and to leading order, the annealed free energy turns out to be $f_{\text{annealed}} = 1$.

One should not throw away the annealed computation right away, as it might be a good approximation in some cases. Moreover, it turns out to be very convenient to prove theorems! Indeed, since the logarithm is a concave function, the average of the logarithm is always smaller or equal to the logarithm of the average, so that

$$f_{\text{annealed}} \leq f_{\text{quenched}}. \tag{24}$$

This is in fact a crucial property in the demonstrations of many results in the physics of disordered systems, and in computer science as well (the celebrated "first moment method" [8]).

Furthermore, there is a reason why, in physics, one should sometimes consider the annealed ensemble instead of the quenched one: when the disorder is changing quickly in time, on timescales similar to those of configuration changes, then we indeed need to average both over configurations and disorder and the annealed average is the correct physical one: this is actually the origin of the name "annealed" and "quenched" averages.

2.3 The Planted ensemble

In fact, a third ensemble of disorder can be defined, that seems odd at first sight, but which turns out to be very interesting as well: the planted ensemble. Following André Gide, we shall first consider a solution and then create the problem: The idea of planting is precisely to *first* generate a configuration of spins that we want to be an equilibrium one, and *then* to create the disorder in the Hamiltonian such that this is precisely an equilibrium configuration. This is very practical, as we are generating at the same time an equilibrium configuration and a realization of the disorder (while doing the opposite is the very difficult task of Monte-Carlo (MC) simulation)! However, like in the annealed case, problems created this way have no reason to be typical ones —that is, the realization of the disorder will have different statistical properties than the quenched one— and indeed in general they are not: Hamiltonian created by this procedure defines a new ensemble.

The planted ensemble has many fascinating properties, and can be used in many ways, as we shall see. It sometimes allows to prove results on the quenched ensemble and to simulate it at zero computational cost. It is also the hidden link between the (complex) theory of the glass transition and the (much more understood) theory of first order transition. Maybe more importantly, it is standing at the roots of the link between statistical physics and Bayesian inference problems that we mentioned in the previous section.

In this chapter, we shall discuss briefly the properties of the planted ensemble, and its relation with the quenched and annealed one (and to another one, called the Nishimori ensemble which will turn out to be the planted one in disguise). We shall then show how to use it and discuss the wonderful applications we have just mentioned.

2.4 The fundamental properties of planted problems

2.4.1 The two golden rules

Consider for the sake of the discussion the Edwards-Anderson spin glass with N spins and Hamiltonian in eq. (18). We first generate a configuration of spins $\mathcal{C}$ totally at random (each of the 2^N configurations has a probability 2^{-N} to appear). We now create the disorder average, by choosing each link $J_{ij} = \pm 1$ such that the probability of a given realization of the disorder $\{J\}$ is

$$P(\{J\}|\mathcal{C}) \propto e^{-\beta \mathcal{H}_J(\mathcal{C})}. \tag{25}$$

This can be done easily by taking each link with the probability $P_{link}(J) = e^{-\beta J S_i^{\mathcal{C}} S_j^{\mathcal{C}}}/(2\cosh\beta)$ where $S_i^{\mathcal{C}}$ and $S_j^{\mathcal{C}}$ are the values of spins i and j in the planted configuration $\mathcal{C}$. We have now created a planted problem. Let us see what is the relation of the planted configuration with the planted problem. In order to do this, we use the Bayes theorem:

$$P(\{J\}|\mathcal{C}) = P(\mathcal{C}|\{J\}) \frac{P(\{J\})}{P(\mathcal{C})} \propto P(\mathcal{C}|\{J\}) P(\{J\}) \tag{26}$$

since the distribution $P(\mathcal{C})$ is uniform. At this point, we thus have

$$P(\mathcal{C}|\{J\}) \propto \frac{e^{-\beta \mathcal{H}_J(\mathcal{C})}}{P(\{J\})}, \tag{27}$$

and by normalization, we thus obtain

$$P(\mathcal{C}|\{J\}) = \frac{e^{-\beta \mathcal{H}_J(\mathcal{C})}}{Z_{\{J\}}}, \tag{28}$$

$$P(\{J\}) \propto Z_{\{J\}} = \sum_{\mathcal{C}'} e^{-\beta \mathcal{H}_J(\mathcal{C}')}. \tag{29}$$

We now see the two fundamental properties of the planted ensemble:

- The planted configuration is an equilibrium one (its probability is precisely given by the Boltzmann factor).
- The realization of the disorder of the planted problem is not chosen uniformly, as in the quenched ensemble, but instead each planted problem appears with a probability proportional to its partition sum: $P(\{J\}) = A(\{J\}) \cdot Z_{\{J\}}$. To ensure normalization of the probability, we require that $\sum_{\{J\}} A(\{J\}) \cdot Z_{\{J\}} = 1$. In full generality we take $A(\{J\}) = \frac{P_{\text{quenched}}(\{J\})}{\sum_{\{J\}} P_{\text{quenched}}(\{J\}) Z_{\{J\}}}$. Thus at the end one has

$$P_{\text{planted}}(\{J\}) = \frac{Z_{\{J\}}}{Z_{\text{annealed}}} P_{\text{quenched}}(\{J\}), \tag{30}$$

  where $Z_{\text{annealed}} = \sum_{\{J\}} Z_{\{J\}} \cdot P_{\text{quenched}}(\{J\})$.

We shall precise in the next section the relation between the planted, annealed and quenched ensembles, and with a fourth one, the Nishimori one.

2.5 The planted ensemble is the annealed ensemble, but the planted free energy is not the annealed free energy

From now on, we shall always precise which ensemble (quenched, planted or annealed) is used in order to compute the averages.

The average energy in the planted ensemble can be easily computed by averaging over all realizations:

$$
\begin{aligned}
[\langle E\rangle]_{\text{planted}} &= \sum_{\{J\}} P_{\text{planted}}(\{J\}) \sum_{\mathcal{C}} \frac{e^{-\beta\mathcal{H}_J(\mathcal{C})}}{Z_{\{J\}}} \mathcal{H}_J(\mathcal{C}) && (31)\\
&= \sum_{\{J\}} \frac{Z_{\{J\}}}{Z_{\text{annealed}}} P_{\text{quenched}}(\{J\}) \sum_{\mathcal{C}} \frac{e^{-\beta\mathcal{H}_J(\mathcal{C})}}{Z_{\{J\}}} \mathcal{H}_J(\mathcal{C}) && (32)\\
&= \sum_{\{J\}} P_{\text{quenched}}(\{J\}) \sum_{\mathcal{C}} \frac{e^{-\beta\mathcal{H}_J(\mathcal{C})}}{Z_{\text{annealed}}} \mathcal{H}_J(\mathcal{C}) && (33)\\
&= [\langle E\rangle]_{\text{annealed}} && (34)
\end{aligned}
$$

The equilibrium energy is simply the same as in the annealed ensemble. In fact, this computation could be repeated for many quantities, and this create a deep link between the annealed and planted ensemble. The reason why it is so is that, in the Edwards-Anderson spin glass, choosing a configuration at random and a set of couplings from the planted ensemble is totally equivalent to sampling a configuration and a set of couplings in the annealed ensemble. Indeed, the joint distribution in the annealed ensemble is by definition

$$P_{\text{annealed}}(\{J\},\mathcal{C}) \propto e^{\beta\mathcal{H}_J(\mathcal{C})}, \tag{35}$$

and since we have generated the configuration with a uniform measure in the planted ensemble, we have

$$P_{\text{planted}}(\{J\},\mathcal{C}) = P_{\text{planted}}(\{J\}|\mathcal{C})\, P_{\text{planted}}(\mathcal{C}) \propto e^{\beta\mathcal{H}_J(\mathcal{C})}. \tag{36}$$

Sampling from the planted ensemble is totally equivalent to sampling from the annealed ensemble! One could thus say that these are the same ensembles, but there is quite an important difference: the planted free energy may be totally different from the annealed free energy (notice for instance that one cannot repeat the former trick for the free energy[1]). Indeed, in the planted ensemble, we generate a problem which is typical for this ensemble, and we compute the free energy of this problem. In the annealed ensemble, we do something very different in the computation of the free energy since we average over both configurations and disorder, so we are not looking at the free energy of a typical instance at all! In fact, the annealed computation is not something one should interpret as the free energy of a single, representative, instance. The best proof of this is that the annealed computation display sometime negative entropies (see for instance the case of the Random Energy Model [3]!). This is impossible for a given problem, and this demonstrates that the annealed free energy should not, in general, be regarded as the typical free energy of an instance chosen in the annealed ensemble.

This is the source of the subtle, yet important, distinction between the annealed and planted ensembles.

2.6 Equivalence between the planted and the Nishimori ensemble

We will now show that the planted ensemble can be mapped to a quenched ensemble but with a slightly modified distribution of couplings, which satisfies the so-called Nishimori condition [9]. Consider again the planted ensemble we have just described. We have generated a configuration at random, and then created a distribution of disorder correlated with this configuration. Notice now that the original Hamiltonian of Edwards-Anderson in eq. (18) satisfies the following gauge invariance:

$$
\begin{aligned}
S_i &\rightarrow S_i\tau_i, && (37)\\
J_{ij} &\rightarrow J_{ij}\tau_i\tau_j, && (38)
\end{aligned}
$$

[1]This might sound surprising, given the energy is the same, but one needs to keep in mind that in changing the temperature, we also change the problem, so that we cannot simply perform thermodynamic integration to compute the free energy in the planted problem.

for any set of $\{\tau\} = \pm 1$. Consider now the very particular set of variables $\{\tau\}$ such that $\tau_i = S_i$. If we apply this gauge transformation, then the planted configuration is transformed into a uniform one where all spin are up, with value $S'_i = 1$ for all i. But what has happened to the couplings? Since the energy of each link has not been changed by the gauge transform, the situation is simple: if the link was frustrated, then it is still a frustrated link with respect to the uniform configuration and thus $J = -1$. If it was not frustrated, then $J = 1$. In other words, we have now a problem where each link has been assigned independently, with probability $P_{\text{link}}(J) = e^{-\beta J}/(2\cosh\beta)$. Since $\beta \geq 0$, most of these links are in fact positive (all of them are positive at $\beta = \infty$ and half of them at $\beta = 0$). So we see that, after the gauge transformation has been done, the disorder has been changed so that:

$$P(J) = \frac{e^{-\beta}}{2\cosh\beta}\delta(J-1) + \frac{e^{\beta}}{2\cosh\beta}\delta(J+1). \tag{39}$$

This is in fact a well known ensemble, called the Nishimori ensemble, and the line in the plane temperature/fraction of positive J defined by eq. (39) is called the Nishimori line [9]. It has been the subject of many studies, because of its particular properties; and we now see that it is simply the planted ensemble in disguise. In fact, almost (if not all) results on the Nishimori line can be understood right away once one notices that the planted configuration (that is, the uniform $S = 1$ one after the gauge transform) is an equilibrium one.

Following eq. (30), one can obtain new interesting properties by simply averaging the partition sum to the power n. We find

$$[Z^n]_{\text{planted}} = \frac{[Z^{n+1}]_{\text{quenched}}}{[Z]_{\text{quenched}}}. \tag{40}$$

This identity was known on the Nishimori line (which, as we just saw, is nothing but the planted ensemble) already from the work of [10]. We now understood how it appears more generally in the planted ensemble.

3 Inference on spin-glasses

3.1 Spin-glasses solution

We already introduced the spin-glass model with Hamiltonian as in Eq. (18). In the mean-field version of this model, the Sherrington-Kirkpatrick (SK) model [14], every spin is connected to all the others. In this case the solution of the model in the quenched ensemble is known and we briefly explain it in this Section. Above a certain critical temperature T_c there is a paramagnetic (PM) phase, caracterized by all the local magnetizations $m_i =< s_i >$ that are null. In this region indeed the quenched and the annealed computation for the free energy lead to the same result. Under the critical temperature T_c the phase space is divided in many different equilibrium pure states, separated by barriers of divergent height in the thermodynamic limit. If we extract a certain realization of the disorder, and then we create two replicas of the system with the same disorder that evolve independently with a certain equilibrium dynamics, than the two replicas will fall in different states. Each pure state α has a certain weight w_α in the partition function and it is characterized by certain magnetizations $\{m_i^\alpha\}$. The spins of the system will thus freeze in their position in a given state, but there will be no preferential orientation, and the total magnetization will be $M = 0$. The *overlap* parameter $q_{\alpha\beta}$ measures how much two different states α and β are similar:

$$q_{\alpha\beta} = \frac{1}{N}\sum_i m_i^\alpha m_i^\beta, \quad -1 \leq q_{\alpha\beta} \leq 1.$$

To characterize a phase of a system, one can use the *overlap distribution*, that measures if the system has different states and how much they are similar:

$$P(q) = \overline{\sum_{\alpha\beta} w_\alpha w_\beta \delta(q_{\alpha\beta} - q)}.$$

In the ferromagnet, there are only two states of equilibrium, with opposite magnetization. Thus the $P(q)$ is trivial; it has two peaks, at m^2 and $-m^2$, related by the obvious Z_2 symmetry. For the spin glass the order parameter is no more the magnetization but the overlap distribution that is not trivial [15, 16]. The $P(q)$ has a continuous support $q \in [0, q_{EA}]$, where $q_{EA} = \frac{1}{N}\sum_i m_i^\alpha m_i^\alpha$ is the self-overlap inside a state. Looking at this function we can learn that in the SK model below T_c there are many states, with different distances.

This mean field theory is called the Replica Symmetry Breaking (RSB) theory, and it predicts a thermodynamic transition also in magnetic field h at a finite temperature. In this framework, a transition line, called *deAlmeida-Thouless* (AT) [17] line, can be identified in the $T - h$ plane between PM and SG phase. It starts at $(T, h) = (T_c, 0)$ and it ends at $(T, h) = (0, h_c)$. $h_c = \infty$ for the SK model. The SK transition is a *continuous* one, in fact q_{EA} grows in a continuous way from 0 at T_c.

But there exists also some particular disordered models that have a *discontinuous* transition. The paradigmatic example is the so called *p-spin* model [18, 19, 20], that has Hamiltonian:

$$\mathcal{H} = -\sum_{\langle i_1,\dots i_p\rangle} J_{i_1\dots i_p} S_{i_1}\dots S_{i_p} \tag{41}$$

with $J_{i_1\dots i_p}$ taken random from a given distribution. In the mean-field, fully-connected case, for high enough temperature, the stable state is the paramagnet, while it undergoes a static transition towards a low temperature spin-glass phase at T_s. The structure of $P(q)$ is quite different from the case of the SK model. In fact it has a single peak at $q = 0$ in the PM phase, and discontinuously develops a second peak at $q_1 \neq 0$ at T_s. However the height of this new peak in $P(q)$ is zero at the transition and grows continuously lowering the temperature. This form of $P(q)$ is due to the fact that below T_s there are different states that dominate the partition function. The self-overlap inside a state is q_1 while the overlap between two different states is zero. The fact that q_1 is well distinct from 0 at T_s means that at the static transition the states are already well formed. There is another temperature that is important for this kind of systems: the dynamical one,

$T_d > T_s$. If we perform a dynamical simulation of our system, we will see that as we approach T_d from above, there will be a slowing down, and the system will take longer and longer time to relax to equilibrium (or to decorrelate from its initial configuration). At T_d the system is no more ergodic, and it will not relax anymore to equilibrium. What happens between T_s and T_d is that an exponential number of metastable states exists: $\mathcal{N} \sim e^{N\Sigma}$, where Σ is called *complexity* or *configurational entropy* These states are the ones who are trapping the system, because each of them is surrounded by an infinite energy barrier. But they are just metastable, so not detected by a thermodynamic calculation that will find a transition only at T_s. At T_s in fact $\Sigma = 0$, the partition function is dominated by a subexponential number of states,

3.2 Phase diagrams of the planted spin-glass

We now want to make the connection between the planted ensemble and inference problems more explicit. Thus we will ask the following question: If we choose a configuration $\mathcal{C}$ and we extract a special configuration of the couplings from the planted ensemble with probability $P(\{J\}|\mathcal{C})$ as in eq. (25) such that the configuration $\mathcal{C}$ is an equilibrium one for example for the EA model at inverse temperature β, is it possible to identify the planted $\mathcal{C}$ from the only knowledge of the couplings $\{J\}$?

This is exactly a problem of inference. From the knowledge of the observables $\{J\}$ and of the likelihood $P(\{J\}|\mathcal{C})$, we can extract the posterior probability $P(\mathcal{C}|\{J\}) = \frac{e^{-\beta \mathcal{H}_J(\mathcal{C})}}{Z_{\{J\}}}$, as explained in Sec. 2.4.1. Being the spins discrete variables, the best choice to infer $\mathcal{C}$ is the use of the MARG estimator, introduced in Sec. 1.3:

$$\hat{s}_i = \text{argmax}_{s_i} P(s_i|\{J\}). \tag{42}$$

We can use for example a MC simulation to sample $P(s_i|\{J\})$ and extract the estimate $\hat{s}_i$. We expect two different situations:

- If all the local magnetizations are zero (that is what happens for example in the PM phase of a spin glass in the quenched ensemble, that is equivalent to the planted one in the PM phase because the annealed and the quenched free energies are the same, we do not have enough information to extract the planted configuration and $P(s_i) = \frac{1}{2}\delta(s_i + 1) + \frac{1}{2}\delta(s_i - 1)$. This means that the extracted couplings were not enough correlated to $\mathcal{C}$. The limit case for example is when $\beta = 0$. In that case $P(J) = \frac{1}{2}\delta(J + 1) + \frac{1}{2}\delta(J - 1)$, the extracted couplings are not at all correlated with the planted configuration, thus there is no hope to find it.

- If the local magnetizations are not zero we can hope that our estimators will give a result that is correlated to the planted solution.

We can define the *overlap* between the estimated and the planted configuration as:

$$q = \frac{\sum_{i=1}^{N} \left(\hat{s}_i - s_i^{\mathcal{C}}\right)}{N} - \frac{1}{2} \tag{43}$$

The subtraction of $\frac{1}{2}$ assures that the random choice in the thermodynamic limit has zero overlap with the planted configuration. There are two possible situations:

- **A second order phase transition** between an *impossible* phase at high temperature, $T > T_c$, where $q = 0$ and a phase in which the reconstruction is *easy*, for $T < T_c$, and we have a positive overlap.

  This is for example the situation of the planted EA model and illustrated in the left part of Fig. 4.

- **A first order phase transition**, that has a more complex scenario. At high temperature, for $T > T_{PT}$ the free energy associated to the inference problem has a unique solution at $q = 0$ (or two solutions with the one at $q = 0$ with the lowest free-energy). This is an *impossible* phase for the reconstruction. For $T_{sp} < T < T_{PT}$ the free energy has a lower minimum at $q \neq 0$ and a highest one at $q = 0$. This is a possible but *hard* phase. In fact in principle one could find a configuration correlated to the planted

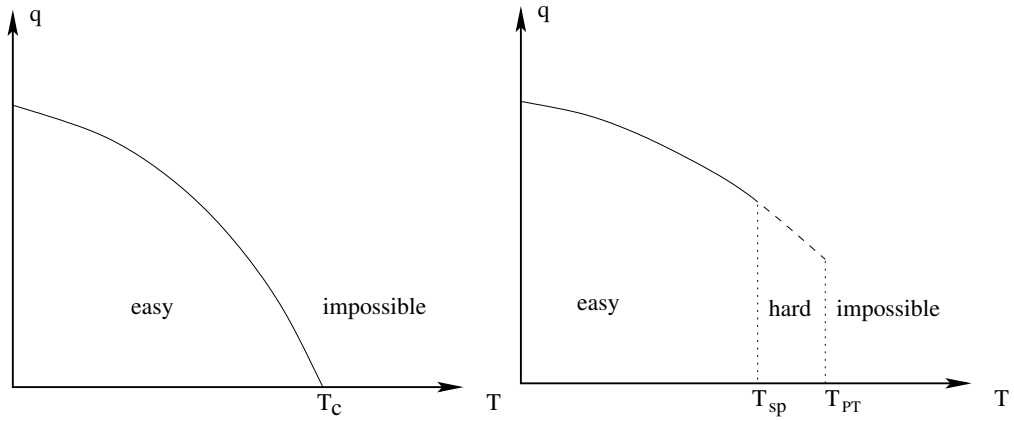

Figure 4: II order phase transition (left) and I order phase transition (right) for the overlap between the estimator and the planted configuration as a function of the temperature in different inference problems

one, but usually it will take a very large time, because one is attracted by the $q = 0$ solution. For $T < T_{sp}$ the free energy has a unique solution at $q \neq 0$ (we are below the *spinodal* point). This is the *easy* phase.

This is for example the situation of spin-glass models whit p-spin interactions, $p > 2$, and it is illustrated in the right panel of Fig. 4.

The value T_c at which the planted EA model undergoes the phase transition between possible and impossible phases is exactly the same temperature at which the quenched EA model has a phase transition between a paramagnetic and a spin-glass phase. We have already view such phenomenon in Sec. 1.4, where the threshold for the denoising problem was exactly the same as the threshold for the REM to enter in the spin-glass phase.

However there is no spin glass phase in the planted EA problem (and neither in the denoising one). There is an intuitive reason for that: we know that for $T < T_c$ we enter in a phase in which the magnetizations are correlated to the planted configuration. Thus there can not exist many different states, as in the spin-glass phase: Indeed the transition will be towards a ferromagnetic phase, polarized in the direction of the planted solution. We have previously seen that the planted ensemble is the analogous of the Nishimori ensemble up to a gauge transformation. On the Nishimori line it has been proved that no RSB can exist, that implies that the low temperature phase of the planted model is not a spin-glass phase.

There are also other things that have been proved (with a lot of work) on the Nishimori line and are quite intuitive if viewed in the planted ensemble. For example one can define an overlap between two different configurations a and b of the problem at equilibrium: $q^{ab} = \langle s_i^a s_i^b \rangle$. It can be demonstrated that on the Nishimori line the equality: $q^{ab} = m = \langle s_i \rangle$ holds, where m is the local magnetization. In the planted ensemble one can interpret this result easily: One can chose an equilibrium configuration as the planted one, and we have seen that the magnetization of the system below T_c is just the overlap with the planted configuration. [2]

3.3 Belief Propagation

We have said that in general in an inference problem we need to compute the posterior marginal probabilities, as in eq. (42). We can do it with a MC simulation. However we know that MC is often slow and in

[2] Another analogous relation that holds on the Nishimori line is the equivalence between the spin-glass and the ferromagnetic susceptibilities: $\chi_{SG} = \chi_F$

inference problems time is important. In this section we will analyze another method to compute marginal probabilities, that is exact on random graphs, and is much faster than MC. Indeed often the spatial structure in inference problems is similar to that of a random graph (for example when one want to analyze real networks like friends on Facebook).

Let us consider an Erdos-Renyi random graph $\mathcal{G}(N, E)$ with N vertices and E edges. To each vertex is associated a variable σ_i and to each edge an interaction $\psi_{ij}(\sigma_i, \sigma_j)$. To be concrete, for an Ising spins system we have $\psi_{ij}(\sigma_i, \sigma_j) = \exp(-\beta J_{ij}\sigma_i\sigma_j)$. Such a random graph can be considered locally as a tree since it can be proven that the typical loops have a length of order $\log N$. In the following computations we will therefore pretend to be in the very bulk of an actual tree. As we will see, these computations are correct as long as the system does not "feel" the presence of loops through long range correlations. We define the quantity $Z_{i\to j}(\sigma_i)$, for two adjacent sites i and j, as the partial partition function for the sub-tree rooted at i, excluding the branch directed towards j, with a fixed value σ_i of the spin variable on the site i. We also introduce $Z_i(\sigma_i)$, the partition function of the whole tree with a fixed value of σ_i. These quantities can be computed according to the following recursion rules:

$$Z_{i\to j}(\sigma_i) = \prod_{k\in\partial i\setminus j}\left(\sum_{\sigma_k} Z_{k\to i}(\sigma_k)\psi_{ik}(\sigma_i,\sigma_k)\right) , \qquad Z_i(\sigma_i) = \prod_{j\in\partial i}\left(\sum_{\sigma_j} Z_{j\to i}(\sigma_j)\psi_{ij}(\sigma_i,\sigma_j)\right) , \tag{44}$$

where $\partial i \setminus j$ indicates all the neighbors of i except spin j. We can rewrite these equations in terms of normalized quantities which can be interpreted as probability laws for the random variable σ_i, namely $\eta_{i\to j}(\sigma_i) = Z_{i\to j}(\sigma_i)/\sum_{\sigma'} Z_{i\to j}(\sigma')$ and $\eta_i(\sigma_i) = Z_i(\sigma_i)/\sum_{\sigma'} Z_i(\sigma')$. The quantity $\eta_{i\to j}(\sigma_i)$ is the marginal probability law of variable σ_i in a modified system where the link $\langle i, j\rangle$ has been removed. The recursion equations read

$$\eta_{i\to j}(\sigma_i) = \frac{1}{z_{i\to j}}\prod_{k\in\partial i\setminus j}\left(\sum_{\sigma_k} \eta_{k\to i}(\sigma_k)\psi_{ik}(\sigma_i,\sigma_k)\right) , \qquad \eta_i(\sigma_i) = \frac{1}{z_i}\prod_{j\in\partial i}\left(\sum_{\sigma_j} \eta_{j\to i}(\sigma_j)\psi_{ij}(\sigma_i,\sigma_j)\right) , \tag{45}$$

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

$$z_{i\to j} = \sum_{\sigma_i}\prod_{k\in\partial i\setminus j}\left(\sum_{\sigma_k} \eta_{k\to i}(\sigma_k)\psi_{ik}(\sigma_i,\sigma_k)\right) , \qquad z_i = \sum_{\sigma_i}\prod_{j\in\partial i}\left(\sum_{\sigma_j} \eta_{j\to i}(\sigma_j)\psi_{ij}(\sigma_i,\sigma_j)\right) . \tag{46}$$

The quantity $\eta_i(\sigma_i)$ is exactly the marginal probability law of the Gibbs-Boltzmann distribution, hence the local magnetizations can be computed as $m_i = \langle\sigma_i\rangle = \sum_\sigma \eta_i(\sigma)\sigma$. Finally, it is useful to define the object

$$z_{ij} = \sum_{\sigma_i,\sigma_j} \eta_{j\to i}(\sigma_j)\eta_{i\to j}(\sigma_i)\psi_{ij}(\sigma_i,\sigma_j) = \frac{z_j}{z_{j\to i}} = \frac{z_i}{z_{i\to j}} , \tag{47}$$

where the last two equalities are easily derived using Eqs. (45).

We can now write the free energy of the system. Clearly, for any spin σ_i the total partition function is $Z = \sum_{\sigma_i} Z_i(\sigma_i)$. Note that using Eqs. (45) and (46), we obtain

$$z_i = \sum_{\sigma_i}\prod_{j\in\partial i}\left(\sum_{\sigma_j} \eta_{j\to i}(\sigma_j)\psi_{ij}(\sigma_i,\sigma_j)\right) = \sum_{\sigma_i}\prod_{j\in\partial i}\left(\sum_{\sigma_j} \frac{Z_{j\to i}(\sigma_j)}{\sum_{\sigma'} Z_{j\to i}(\sigma')}\psi_{ij}(\sigma_i,\sigma_j)\right) = \frac{\sum_{\sigma_i} Z_i(\sigma_i)}{\prod_{j\in\partial i}\sum_{\sigma_j} Z_{j\to i}(\sigma_j)} , \tag{48}$$

and along the same steps

$$z_{j\to i} = \frac{\sum_{\sigma_j} Z_{j\to i}(\sigma_j)}{\prod_{k\in\partial j\setminus i}\sum_{\sigma_k} Z_{k\to j}(\sigma_k)} . \tag{49}$$

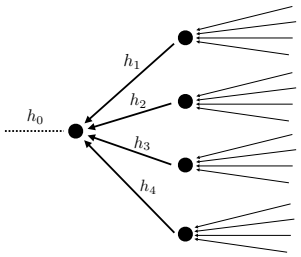

Figure 5

So we can start from an arbitrary spin i and

$$Z = \sum_{\sigma_i} Z_i(\sigma_i) = z_i \prod_{j\in\partial i} \left(\sum_{\sigma_j} Z_{j\to i}(\sigma_j) \right) = z_i \prod_{j\in\partial i} \left(z_{j\to i} \prod_{k\in\partial j\setminus i} \sum_{\sigma_k} Z_{k\to j}(\sigma_k) \right) , \tag{50}$$

and we can continue to iterate this relation until we reach the leaves of the tree. Using Eq. (47), we finally obtain

$$Z = z_i \prod_{j\in\partial i} \left(z_{j\to i} \prod_{k\in\partial j\setminus i} z_{k\to j} \cdots \right) = z_i \prod_{j\in\partial i} \left(\frac{z_j}{z_{ij}} \prod_{k\in\partial j\setminus i} \frac{z_k}{z_{jk}} \cdots \right) = \frac{\prod_i z_i}{\prod_{\langle i,j\rangle} z_{ij}} \tag{51}$$

and the free energy is

$$\begin{aligned} F &= -T \log Z = \sum_i f_i - \sum_{\langle i,j\rangle} f_{ij} \ , \\ f_i &= -T \log z_i \ , \\ f_{ij} &= -T \log z_{ij} \ . \end{aligned} \tag{52}$$

The advantage of this expression of F is that it does not depend on the arbitrary choice of the initial site i we made above.

3.3.1 Stability of the paramagnetic solution

Let us consider a spin glass on a Bethe Lattice. Through belief propagation we can write the effective "cavity" field acting on a central bulk spin 0 as (see Fig. 5)

$$\beta h_0 = \sum_{i=1}^{k} \operatorname{atanh}[\tanh(\beta J_{ij}) \tanh(\beta h_i)] + H. \tag{53}$$

A first, simple, way to check the appearance of a spin glass phase is to compute the temperature where the spin-glass susceptibility diverges, or equivalently where the solution found with the method presented in the preceding section is *locally unstable*. When this happens, a continuous phase transition towards a spin glass phase arises. This is due to the appearance of long range correlations that make the presence of loops relevant to the problem. A line is separating the paramagnetic and the spin glass phase, which is called the de Almeida-Thouless line (AT) [21]. In order to compute the AT line, one should consider the onset of the divergence of the spin-glass susceptibility [22, 23]. For a given spin S_0, due to the homogeneous local tree-like structure of the graph, the spin glass susceptibility can be written as

$$\chi_{\mathrm{SG}} = \beta \sum_i \langle S_0 S_i \rangle_{\mathrm{c}}^2 \approx \beta \sum_{r=0}^{\infty} k^r \overline{\langle S_0 S_r \rangle_{\mathrm{c}}^2}. \tag{54}$$

In eq. (54) $\overline{\cdots}$ represents a spatial average over the whole graph, whereas $\langle\cdots\rangle$ represents a thermal average and $k = c - 1$. Note that $X_{\rm c}$ is the connected version of a correlation function X. Using the Fluctuation-Dissipation relation, one obtains

$$\beta\overline{\langle S_0 S_i\rangle_{\rm c}^2} = \overline{\left(\frac{\partial\langle S_i\rangle}{\partial\langle h_0\rangle}\right)^2}. \tag{55}$$

Since the physical field is a function of the cavity fields, we can monitor the propagation of the response using the chain rule

$$\frac{\partial\langle S_r\rangle}{\partial\langle h_0^{\rm c}\rangle} = \frac{\partial\langle S_r\rangle}{\partial\langle h_r\rangle}\frac{\partial\langle h_r\rangle}{\partial\langle h_{r-1}\rangle}\frac{\partial\langle h_{r-1}\rangle}{\partial\langle h_{r-2}\rangle}\cdots\frac{\partial\langle h_1\rangle}{\partial\langle h_0\rangle}. \tag{56}$$

To check whether or not the susceptibility defined in eq. (54) diverges, it is thus convenient to consider the large-distance behavior of the following stability parameter

$$\lambda(r) = k^r\overline{\left(\frac{\partial\langle h_r\rangle}{\partial\langle h_0\rangle}\right)^2} \tag{57}$$

with

$$\frac{\partial h_0}{\partial h_i} = \frac{\tanh(\beta J_{i0})[1-\tanh^2(\beta h_i)]}{1-\tanh^2(\beta J_{i0})\tanh^2(\beta h_i)}. \tag{58}$$

If $\lambda(r)$ vanishes for large r, then the system is paramagnetic, otherwise the spin-glass susceptibility diverges. Repeating this computation for different values of H and T the complete AT line can be determined point by point with an excellent precision.

In the case of zero external field eq. (58) can be simplified since in the paramagnetic phase $h = 0$ all over the sample, and therefore the stability condition factorizes and reads

$$\overline{\left(\frac{\partial\langle h_i\rangle}{\partial\langle h_{\rm j}\rangle}\right)^2} < 1. \tag{59}$$

Using eq. (53), the critical point is thus given by

$$k\overline{\tanh(\beta_c J)^2} = 1, \tag{60}$$

as was first found by Thouless in Ref. [21]. For a discrete spin glass with e.g., bimodal interactions, we obtain

$$T_{\rm c}^{\pm J}(k) = \left[\mathrm{atanh}\frac{1}{\sqrt{k}}\right]^{-1}, \tag{61}$$

while for Gaussian disorder one has to use numerical tools to obtain $T_{\rm c}$.

It is also interesting to consider the large-connectivity limit of these computations, where k is large and where we rescale the couplings so that $J_{ij} = \frac{\epsilon_{ij}}{\sqrt{k}}$ where $\epsilon_{ij} \in \{\pm 1\}$ with equal probability in order to keep the free energy extensive [22, 23]. We obtain:

$$\beta h_0 = \sum_{i=1}^{k}\mathrm{atanh}\left(\frac{\beta}{\sqrt{k}}\epsilon_{ij}\tanh(\beta h_i)\right) + H \tag{62}$$

$$\approx \sum_{i=1}^{k}\frac{\beta}{\sqrt{k}}\epsilon_{ij}\tanh(\beta h_i) + H. \tag{63}$$

In order to make the connection with the well known result of de Almeida and Thouless, it is convenient to use the cavity magnetization of a spin $m_i = \tanh\beta h_i$ for which the recursion reads

$$m_0 = \tanh\left(\sum_{i=1}^{k}\frac{\beta}{\sqrt{k}}\epsilon_{ij}m_j + H\right). \tag{64}$$

For large k, the argument of the tanh is a sum of uncorrelated variables and thus follows a Gaussian distribution. Denoting m and q the mean and the variance of the distribution of m_i, two self-consistent equations can be derived

$$m = \frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} \tanh\Big(\beta q^{1/2} z + \beta H\Big) dz \tag{65}$$

$$q = \frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} \tanh^2\Big(\beta q^{1/2} z + \beta H\Big) dz \tag{66}$$

In the previous equations the replica symmetric relations for the magnetization and the spin overlap in the Sherrington-Kirkpatrick model[14, 24] are apparent. In addition, the condition for the stability of the solution, applied to the magnetization after one iteration, now reads

$$k\overline{\left(\frac{\partial m_0}{\partial m_i}\right)^2} < 1 \tag{67}$$

from which we immediately obtain

$$\frac{\beta^2}{\sqrt{2\pi}} \int e^{-z^2/2} \mathrm{sech}^4\Big(\beta q^{1/2} z + \beta H\Big) < 1. \tag{68}$$

This is precisely the original result obtained by de Almeida and Thouless,[25] i.e., the cavity approach reproduces these results in the large-connectivity limit exactly: a proof, if needed of the coherence of the method.

4 Community detection

A useful application of what we said in the previous sections is the problem of *community detection* (for a review see e.g. [26]). In the study of complex networks, a network is said to have community structure if it divides naturally into groups of nodes with denser connections within groups and sparser connections between groups. This type of structure, where nodes are more likely to connect to others of the same type as in a ferromagnet, is called *assortative*. The goal is to detect the communities. In other cases the network can be *disassortative*, with denser connections between different groups than within groups. For instance, a set of predators might form a functional group in a food web, not because they eat each other, but because they eat similar prey.

The problem of community detection consists in finding the labeling for each node (the group they belong to) given the graph. This has a widely applications: on-line communities, biological or metabolic networks, financial market...

The first naive idea to solve the problem in an assortative case is the so called *graph partitioning problem*: we try to make a partition minimizing the number of links between the two groups. However this is a NP complete problem! Our naive idea is not so simple to implement!

The problem can be stated also in another way: given the adjacency matrix of the graph, we want to find its hidden structure reshuffling the elements in a proper way.

In the literature there are two classes of solving methods:

- **Spectral clustering methods**, based on the computation of the top eigenvalues of different matrices associated to the graph (adjacency matrix, random-walk matrix, ...) that should be related to the groups.
- **Modularity maximization**, that consists in finding the labeling that minimizes a given cost function, for example

$$Q = \sum_{ij} \left[A_{ij} - \frac{d_i d_j}{2N} \right] \delta_{s_i} \delta_{s_j} \tag{69}$$

  where A_{ij} is the adjacency matrix, d_i is the degree of node i and s_i is the label that we assign to node i. This method is equivalent to minimize the free energy of a given Potts model.

There are two main problems for these methods:

- Extract the number of real groups present in the graphs
- Understand when there is no information in the graph avoiding overfitting. These methods will find communities even when the graph is simply an Erdos-Renyi random graph with no communities.

In the next section we will introduce a new method based on Bayesian inference, following ref. [27].

4.1 The stochastic block model

To apply Bayesian inference we first need a model. The simplest one is the stochastic block model, defined as follows. It has parameters q (the number of groups), $\{n_a\}$ (the expected fraction of nodes in each group a, for $1 \leq a \leq q$), and a $q \times q$ *affinity matrix* p_{ab} (the probability of an edge between group a and group b). We generate a random graph G on N nodes, with adjacency matrix $A_{ij} = 1$ if there is an edge from i to j and 0 otherwise, as follows. Each node i has a label $t_i \in \{1, \ldots, q\}$, indicating which group it belongs to. These labels are chosen independently, where for each node i the probability that $t_i = a$ is n_a. Between each pair of nodes i, j, we then include an edge from i to j with probability p_{t_i,t_j}, setting $A_{ij} = 1$, and set $A_{ij} = 0$ with probability $1 - p_{t_i,t_j}$. We forbid self-loops, so $A_{ii} = 0$.

We let N_a denote the number of nodes in each group a. Since N_a is binomially distributed, in the limit of large N we have $N_a/N = n_a$ with high probability. The average number of edges from group a to group

b is then $M_{ab} = p_{ab}N_aN_b$, or $M_{aa} = p_{aa}N_a(N_a - 1)$ if $a = b$. Since we are interested in sparse graphs where $p_{ab} = O(1/N)$, we will often work with a rescaled affinity matrix $c_{ab} = Np_{ab}$. In the limit of large N, the average degree of the network is then

$$c = \sum_{a,b} c_{ab} n_a n_b \,. \tag{70}$$

In the undirected case A_{ij}, p_{ab}, and c_{ab} are symmetric. A special case is when

$$c_{ab} = \begin{cases} c_{\rm in} & a = b \\ c_{\rm out} & a \neq b \end{cases} \tag{71}$$

4.2 Inferring the group assignment

The first problem that we can analyze is the following: given the graph, that was generated from the stochastic block model with known parameters $\theta = \{q, \{n_a\}, \{p_{ab}\}\}$ we want to infer the group assignment $\{q_i\}$ for each node. The prior probability is just $P(\{q_i\}) = \prod_{i=1}^{N} n_{q_i}$. The likelihood is

$$P(G|\{q_i\}) = \prod_{i \neq j} \left[p_{q_i,q_j}^{A_{ij}} (1 - p_{q_i,q_j})^{1-A_{ij}} \right] \tag{72}$$

from which we can extract the posterior probability

$$P(\{q_i\}|G) = \frac{P(G|\{q_i\})P(\{q_i\})}{Z} = \frac{e^{-H(\{q_i\}|G)}}{Z}, \tag{73}$$

where in the last equality we are emphasizing that in the language of statistical physics, this distribution is the Boltzmann distribution of a generalized Potts model with Hamiltonian

$$H(\{q_i\}\,|\,G) = -\sum_i \log n_{q_i} - \sum_{i \neq j} \left[A_{ij} \log c_{q_i,q_j} + (1 - A_{ij}) \log \left(1 - \frac{c_{q_i,q_j}}{N} \right) \right]. \tag{74}$$

The labels q_i are Potts spins taking one of the q possible values, and the logarithms of the group sizes n_{q_i} become local magnetic fields. In the sparse case $c_{ab} = O(1)$, there are strong $O(1)$ interactions between connected nodes, $A_{ij} = 1$, and weak $O(1/N)$ interactions between nodes that are not connected, $A_{ij} = 0$. Thus the resulting Potts model is fully connected even if the original graph was sparse.

At this point we can infer the labels using the MARG estimator:

$$\hat{q}_i = \text{argmax}_{q_i} \nu_i(q_i|G) \tag{75}$$

where $\nu_i(q_i|G)$ is the marginal probability on node i, the local magnetization of the Potts variable.

The solution of the inference problem seems simple in this setting. There is however a strong assumption, as usual in Bayesian inference, that is that the graph was created from the stochastic block model. This is a quite strong assumption. However, if it was actually created from this model, this is the best estimator we can obtain. In practice we can simulate the model with Hamiltonian in eq. (74) using a MC simulation to extract $\mu(q_i|G)$. However the model is now fully connected. This means that each MC step takes $O(N^2)$ time. However we can rewrite eq. (74) as:

$$H(\{q_i\}\,|\,G) = -\sum_i \log n_{q_i} - \sum_{i \neq j : A_{ij} \neq 0} \left(\log c_{q_i,q_j} + \log \left(1 - \frac{c_{q_i,q_j}}{N} \right) \right) + \sum_{i \neq j} \log \left(1 - \frac{c_{q_i,q_j}}{N} \right). \tag{76}$$

The first two terms are local and the last term depends only on the total number of spins that are in a certain color, thus can be updated easily at each step. In this way the time for each MC step is $O(N)$.

4.3 Learning the parameters of the model

The second problem that we want to solve is to assign the labels if we do not know the parameters of the block model θ. We make use once more of the Bayes formula. In general we do not have any information on the parameters, thus the prior probability is constant and the posterior probability is just:

$$P(\theta|G) \propto P(G|\theta) = \sum_{\{q_i\}} P(G, \{q_i\}|\theta) \tag{77}$$

where the sum runs over all possible group assignments. The last term is just the partition function Z in eq. (73), for the case in which we know the parameters θ.

Thus maximizing $P(\theta \mid G)$ over θ is equivalent to maximizing the partition function over θ, or equivalently minimizing the free energy density f of the Potts model (74) as a function of θ. If the function $f(\theta)$ has a non-degenerate minimum, then in the thermodynamic limit this minimum is achieved with high probability at precisely the values of the parameters that were used to generate the network. Extracting the free energy in a MC simulation is usually a long and difficult task, because it requires to calculate the averaged energy as a function of the temperature and then to compute f from the relation $\frac{\partial(\beta f)}{\partial \beta} = E$. For this reason, rather than minimizing $f(\theta)$ directly, it is useful to write explicit conditions for the stationarity of $f(\theta)$. Taking the derivative of $f(\theta)$ with respect to n_a for $1 \leq a \leq q$, subject to the condition $\sum_a n_a = 1$, and setting these derivatives equal to zero gives

$$\frac{1}{N} \sum_i \langle \delta_{q_i,a} \rangle = \frac{\langle N_a \rangle}{N} = n_a \quad \forall a = 1, \ldots, q\,, \tag{78}$$

where by $\langle f(\{q_i\}) \rangle = \sum_{\{q_i\}} f(\{q_i\}) \mu(\{q_i\}|G, \theta)$ we denote the thermodynamic average. Thus for each group a, the most likely value of n_a is the average group size; an intuitive result, but one that deserves to be stated. Analogously, taking the derivative of $f(\theta)$ by the affinities c_{ab} gives

$$\frac{1}{N n_a n_b} \sum_{(i,j)\in E} \langle \delta_{q_i,a} \delta_{q_j,b} \rangle = \frac{\langle M_{ab} \rangle}{N n_a n_b} = c_{ab} \quad \forall a, b\,. \tag{79}$$

Meaning that the most likely value of c_{ab} is proportional to the average number of edges from group a to group b. More to the point, the most likely value of $p_{ab} = c_{ab}/N$ is the average fraction of the $N_a N_b$ potential edges from group a to group b that in fact exist. In the undirected case, for $a = b$ we have

$$\frac{1}{N n_a^2/2} \sum_{(i,j)\in E} \langle \delta_{q_i,a} \delta_{q_j,a} \rangle = \frac{\langle M_{aa} \rangle}{N n_a^2/2} = c_{aa} \quad \forall a\,. \tag{80}$$

The stationarity conditions (78–80) naturally suggest an iterative way to search for the parameters θ that minimize the free energy. We start with arbitrary estimates of θ (actually not completely arbitrary, for a more precise statement see subsequent sections), measure the mean values $\langle N_a \rangle$ and $\langle M_{ab} \rangle$ in the Boltzmann distribution with parameters θ, and update θ according to (78–80) . We then use the resulting θ to define a new Boltzmann distribution, again measure $\langle N_a \rangle$ and $\langle M_{ab} \rangle$, and so on until a fixed point is reached.

In statistical physics, the stationarity conditions (78–80) can be interpreted as the equality of the quenched and annealed magnetization and correlations. In models of spin glasses (e.g. [28]) they are referred to as the *Nishimori conditions* (see Sec. 2.6). This iterative way of looking for a maximum of the free energy is equivalent to the well-known *expectation-maximization* (EM) method in statistics [29].

4.4 Belief propagation equations

In Sec. 3.3 we have introduced Belief Propagation equations, that is an alternative and more effective method with respect to MC to compute marginal probabilities on random graphs and we want to apply them in this

case. Note that in our case the "network of interactions" is fully connected, since in the Hamiltonian (74) there are weak interactions even along the non-edges, i.e., between pairs of nodes that are not connected. However, as we will see these weak interactions can be replaced with a "mean field", limiting the interactions to the sparse network.

We define conditional marginals, or *messages*, denoted $\psi_{q_i}^{i\to j}$ as the marginal probability that the node i belongs to group q_i in the absence of node j. The cavity method 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 to j recursively in terms of the messages that i receives from its other neighbors k:

$$\psi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}}\, n_{t_i} \prod_{k\neq i,j} \left[\sum_{t_k} c_{t_i t_k}^{A_{ik}} \left(1 - \frac{c_{t_i t_k}}{N}\right)^{1-A_{ik}} \psi_{t_k}^{k\to i} \right], \tag{81}$$

where $Z^{i\to j}$ is a normalization constant ensuring $\sum_{t_i} \psi_{t_i}^{i\to j} = 1$. We apply (81) iteratively until we reach a fixed point $\{\psi_{q_i}^{i\to j}\}$. Then the marginal probability is estimated to be $\nu_i(t_i) = \psi_{t_i}^{i}$, where

$$\psi_{t_i}^{i} = \frac{1}{Z^{i}}\, n_{t_i} \prod_{k\neq i} \left[\sum_{t_k} c_{t_i t_k}^{A_{ik}} \left(1 - \frac{c_{t_i t_k}}{N}\right)^{1-A_{ik}} \psi_{t_k}^{k\to i} \right]. \tag{82}$$

Since we have nonzero interactions between every pair of nodes, we have potentially $N(N-1)$ messages. 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

$$\psi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}}\, n_{t_i} \prod_{k\notin \partial i\setminus j} \left[1 - \frac{1}{N} \sum_{t_k} c_{t_k t_i} \psi_{t_k}^{k\to i} \right] \prod_{k\in\partial i} \left[\sum_{t_k} c_{t_k t_i} \psi_{t_k}^{k\to i} \right] = \psi_{t_i}^{i} + O\left(\frac{1}{N}\right). \tag{83}$$

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

$$\psi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}}\, n_{t_i} \prod_{k\notin \partial i} \left[1 - \frac{1}{N} \sum_{t_k} c_{t_k t_i} \psi_{t_k}^{k\to i} \right] \prod_{k\in\partial i\setminus j} \left[\sum_{t_k} c_{t_k t_i} \psi_{t_k}^{k\to i} \right]. \tag{84}$$

The belief propagation equations can hence be rewritten as

$$\psi_{t_i}^{i\to j} = \frac{1}{Z^{i\to j}}\, n_{t_i} e^{-h_{t_i}} \prod_{k\in\partial i\setminus j} \left[\sum_{t_k} c_{t_k t_i} \psi_{t_k}^{k\to i} \right], \tag{85}$$

where we neglected terms that contribute $O(1/N)$ to $\psi^{i\to j}$, and defined an auxiliary external field

$$h_{t_i} = \frac{1}{N} \sum_{k} \sum_{t_k} c_{t_k t_i} \psi_{t_k}^{k}. \tag{86}$$

This simplification is analogous to the one that leads to eq. (76) from eq. (74). In order to find a fixed point of Eq. (85) in linear time we update the messages $\psi^{i\to j}$, recompute ψ^{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 $\nu_i(t_i)$ is then

$$\psi^i_{t_i} = \frac{1}{Z^i}\, n_{t_i} e^{-h_{t_i}} \prod_{j\in\partial i} \left[\sum_{t_j} c_{t_j t_i} \psi^{j\to i}_{t_j} \right] . \tag{87}$$

When the cavity approach is asymptotically exact then the true marginal probabilities obey $\nu_i(t_i) = \psi^i_{t_i}$. We can also calculate the free energy using eq. (52). Now that we have the marginals we chose the labeling $\{\hat{q}_i\}$ according to eq. (75). The overlap with the original group assignment $\{q_i\}$ can be defined as:

$$Q(\{\hat{q}_i\},\{q_i\}) = \max_\pi \frac{\frac{1}{N}\sum_i \delta_{\hat{q}_i,\pi(q_i)} - \max_a n_a}{1-\max_a n_a}, \tag{88}$$

where π ranges over the permutations on q elements. The overlap is defined so that if $\hat{q}_i = q_i$ for all i, i.e., if we find the exact labeling, 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 $\{\hat{q}_i\}$ is correlated with the original one $\{q_i\}$ if in the thermodynamic limit $N \to \infty$ the overlap is strictly positive. The marginals $\nu_i(t_i)$ can also be used to distinguish nodes that have a very strong group preference from those that are uncertain about their membership (this is usually not possible with other clustering methods).

4.5 Phase transitions in group assignment

In this section we will analyze the different phase transitions that exist in group assignment with known parameters θ.

The factorized solution and its stability. The first observation to make about the belief propagation equations (85) is that

$$\psi^{i\to j}_{t_i} = n_{t_i} \tag{89}$$

is always a fixed point, as can be verified by plugging (89) into (85). In the literature, a fixed point where messages do not depend on the indexes i, j is called a *factorized fixed point*, hence our name for this case of the block model. The free energy density at this fixed point is

$$f_{\text{factorized}} = \frac{c}{2}\,(1-\log c)\ . \tag{90}$$

For the factorized fixed point we have $\psi^i_{t_i} = n_{t_i}$, in which case the overlap (88) is $Q = 0$. This fixed point does not provide any information about the original assignment—it is no better than a random guess. It is what we called *paramagnetic* fixed point in Sec. 3.2. If this fixed point gives the correct marginal probabilities and the correct free energy, we have no hope of recovering the original group assignment. For which values of q and c_{ab} is this the case? We can study its stability under random perturbations of the messages with the method of Sec. 3.3.1.

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 $O(\log N)$. Consider such a tree with d levels, in the limit $d \to \infty$. Assume that on the leaves the factorized fixed point is perturbed as

$$\psi^k_t = n_t + \epsilon^k_t\,, \tag{91}$$

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

$$T^{ab}_i \equiv \frac{\partial \psi^{k_i}_a}{\partial \psi^{k_{i+1}}_b}\Big|_{\psi_t = n_t} = \left[\frac{\psi^{k_i}_a c_{ab}}{\sum_r c_{ar}\psi^{k_{i+1}}_r} - \psi^{k_i}_a \sum_s \frac{\psi^{k_i}_s c_{sb}}{\sum_r c_{sr}\psi^{k_{i+1}}_r} \right]\Bigg|_{\psi_t = n_t} = n_a \left(\frac{c_{ab}}{c} - 1 \right) . \tag{92}$$

where this expression was derived from (85) 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

$$\epsilon_{t_0}^{k_0} = \sum_{\{t_i\}_{i=1,\dots,d}} \left[\prod_{i=0}^{d-1} T_i^{t_i,t_{i+1}} \right] \epsilon_{t_d}^{k_d} \tag{93}$$

We observe in (92) that the matrix T_i^{ab} does not depend on the index i. Hence (93) 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

$$\left\langle \left(\epsilon_{t_0}^{k_0}\right)^2 \right\rangle \approx \left\langle \left(\sum_{k=1}^{c^d} \lambda^d \epsilon_t^k \right)^2 \right\rangle \approx c^d \lambda^{2d} \left\langle \left(\epsilon_t^k\right)^2 \right\rangle . \tag{94}$$

This gives the following stability criterion,

$$c\lambda^2 = 1 . \tag{95}$$

For $c\lambda^2 < 1$ the perturbation on leaves vanishes as we move up the tree and the factorized fixed point is stable. On the other hand, if $c\lambda^2 > 1$ the perturbation is amplified exponentially, the factorized 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 $(1, 1, \dots, 1)$, and $\lambda_2 = (c_{\rm in} - c_{\rm out})/(qc)$ with eigenvectors of the form $(0, \dots, 0, 1, -1, 0, \dots, 0)$ and degeneracy $q-1$. The factorized fixed point is then unstable, and communities are easily detectable, if

$$|c_{\rm in} - c_{\rm out}| > q\sqrt{c} . \tag{96}$$

For the case when $q = 2$ it was proved rigorously in Ref. [30] that it is indeed impossible to cluster if $|c_{\rm in} - c_{\rm out}| < q\sqrt{c}$ and it is impossible even to estimate the model parameters from the graph. The other part, that indeed it is possible to have reconstruction if $|c_{\rm in} - c_{\rm out}| > q\sqrt{c}$ is proved rigorously in Ref. [31].

Continuous transition Fig. 6 (from ref. [27] 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, with $c_{aa} = c_{\rm in}$ and $c_{ab} = c_{\rm out}$ for all $a \neq b$, varying the ratio $\epsilon = c_{\rm out}/c_{\rm in}$. If $\epsilon = 1$ the probability of connection inside and outside a group is the same, thus we expect that we can not distinguish the communities, the graph is an Erdős-Rényi random graph. $\epsilon = 0$ gives completely separated groups. 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 $\psi_{t_i}^{i\to j}$ are random normalized distributions on t_i). The points in Fig. 6 are results obtained from Gibbs sampling with MC.

We can 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, $Q = 0$. BP and MC converge to the factorized solution, with free energy (90). To understand how is it possible to have no information even if $\epsilon \neq 1$, note that from the expressions for the free energy, it follows that the network generated with the block model is thermodynamically *indistinguishable* from an Erdős-Rényi random graph of the same average degree, in the sense that typical thermodynamic properties of the two ensembles are the same.

- If $|c_{\rm in} - c_{\rm out}| > q\sqrt{c}$, the graph contains significant information about the original group assignment, and using BP or MC 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. (75).

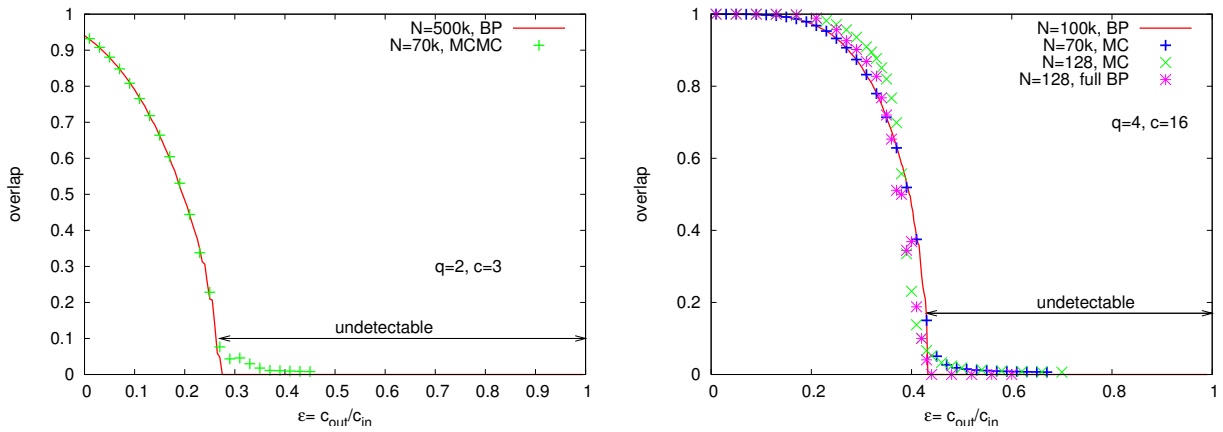

Figure 6: The overlap (88) between the original assignment and its best estimate given the structure of the graph, computed by the marginalization (75) in Ref. [27]. Graphs were generated using N nodes, q groups of the same size, average degree c, and different ratios $\epsilon = c_{\rm out}/c_{\rm in}$. Results from belief propagation (85) for large graphs (red line) are compared to Monte Carlo simulations (data points). The agreement is good, with differences in the low-overlap regime attributed to finite size fluctuations. On the right we also compare to results from the full BP (81) and MC 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 (MC) 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.

Fig. 6 hence illustrates a continuous phase transition in the detectability of communities that is the analogous of what happens in the EA model, described in the left part of Fig. 4.

Discontinuous transition The situation illustrated in Fig. 6 is, however, not the most general one. Fig. 7 (from ref. [27]) illustrates the case of planted coloring with $q = 5$, $c_{\rm in} = 0$, and $c_{\rm out} = qc/(q-1)$. In this case the condition for stability (96) leads to a threshold value $c_\ell = (q-1)^2$. The overlap obtained with BP is plotted, using two different initializations: the random one, and the planted one corresponding to the original assignment. In the latter case, the initial messages are

$$\psi_{q_i}^{i\to j} = \delta_{q_i t_i}\,, \tag{97}$$

where t_i is the original assignment. The corresponding BP free energies are also plotted. As the average degree c increases, we see four different phases in Fig. 7:

I. For $c < c_d$, both initializations converge to the factorized fixed point, so the graph does not contain any significant information about the original group assignment. The ensemble of assignments that have the proper number of edges between each pair of groups is thermodynamically indistinguishable from the uniform ensemble. The original assignment is one of these configurations, and there is no possible way to tell which one it is without additional knowledge.

II. For $c_d < c < c_c$, the planted initialization converges to a fixed point with positive overlap, and its free energy is larger than the annealed free energy. 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

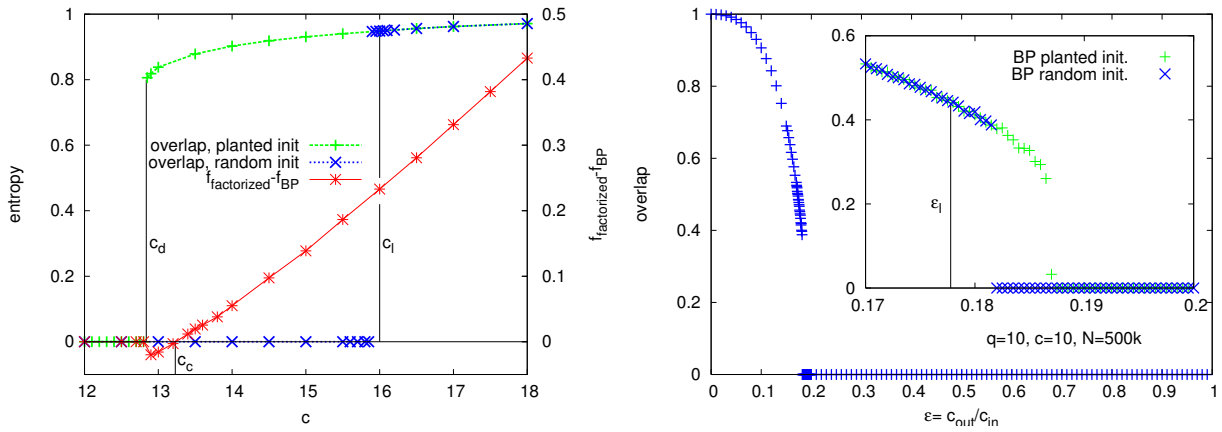

Figure 7: Left: graphs generated with $q = 5$, $c_{\rm in} = 0$, and $N = 10^5$. We compute the overlap (88) and the free energy with BP for different values of the average degree c. The green crosses show the overlap of the BP fixed point resulting from using the original group assignment as the initial condition, and the blue crosses show the overlap resulting from random initial messages. The red stars show the difference between the factorized free energy (90) and the free energy resulting from the planted 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. Right: 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 planted 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.

assignment. The annealed free energy is still the correct total free energy, 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 energy is smaller than the annealed free energy. 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 entropy and the lowest free energy. 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 factorized fixed point. In this phase, inference is possible, but exponentially hard; the state containing the original assignment is, in a sense, hidden below a glass transition. Based on the physics of glassy systems, we predict that no polynomial-time algorithm can achieve a positive overlap with the original group assignment.

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 easy, and BP achieves it in linear time. Indeed, in this easy phase, many efficient algorithms will be able to find a group assignment strongly correlated with the original one.

The case $q = 5$, $c_{\rm in} = 0$, illustrated in Fig. 7, is also investigated with MC. For the planted initialization, its performance is generally similar to BP. For the random initialization, MC agrees with BP only in phases

(I) and (IV). It follows from results on glassy systems [32] that in phases (II) and (III), the equilibration time of MC is exponentially large as a function of N, and that its performance in linear time, i.e., CN for any constant C, does not yield any information about the original group assignment.

The boundaries between different phases correspond to well-known phase transitions in the statistical physics of spin glasses as mentioned in Sec. 3.2 and the discontinuous phase transition of this case is the analogous of the right part of Fig. 4 for the p-spin glass. Specifically, c_d is the dynamical transition or reconstruction threshold, see e.g. [33, 34]. The detectability threshold c_c corresponds to the condensation transition or the Kauzmann temperature. Finally, c_ℓ is the easy/hard transition in planted models introduced in [12].

The problem of clustering is thus an example in which statistical inference and the tools developed in the contest of statistical mechanics of disordered systems are really useful to characterize the problem and identify phase transitions. This is not important only at the theoretical level, but can really help to reach a deep understanding of the problem and to develop useful algorithms if it is the case. In fact previous methods like spectral methods are not so good as the algorithm proposed in this section.

5 Compressed sensing

Images taken from a common camera are usually compressible. In fact in a proper basis (the *wavelet* one), the signal is sparse. This is for example, how *data compression* works in JPEG 2000: if an image of $N = 10^6$ pixels is acquired, once put in the wavelet basis only a limited number of components, say $M = 10^4$ components, will be sensibly different from zero. Thus only these components are maintained, with a great gain in memory. However, this brings a question: why we need to record initially N components and then throw away many of them? The answer is simple: Because we do not know in advance which are the good ones. Data compression is the standard way to compress signals. However there is an alternative way, *compressed sensing*. The idea is to record directly a compressed signal, of $M < N$ components. This will allow to save time (recording is a slow process, so it is better to record a smaller amount of data, and plus we do not need to compress the signal a-posteriori, because it is already compressed) and memory (we do not have to store the original signal). However we will need an algorithm that allows us to recover the original signal a posteriori. This alternative way of acquisition will have many applications: speeding up magnetic resonance imaging without the loss of resolution, fast data compressions, gain of weight using smaller hard disks to store data in telescopes and so on.

5.1 The problem

The idea of compressed sensing (CS) is the following: The signal (the real image) is a vector $\vec{s}$ of N components. The measurements are grouped into a M-component vector $\vec{y}$, which is obtained from $\vec{s}$ by a linear transformation $\vec{y} = \mathbf{G}\vec{s}$. $\mathbf{G}$ is thus a $M \times N$ matrix. Of course $M < N$ if we want to compress the signal. The observer knows the matrix $\mathbf{G}$ and the measurements $\vec{y}$. His aim is to reconstruct $\vec{s}$. The inverse relation $\vec{s} = \mathbf{G}^{-1}\vec{y}$ can not be used because $\mathbf{G}$ is not invertible, being $M < N$. The system is underdetermined and usually has infinitely many solutions.

But we can use another information: we know that in a proper basis the signal is sparse. Let us call $\mathbf{A}$ the matrix for the changing of basis. Thus $\vec{s} = \mathbf{A}\vec{x}$, where $\vec{x}$ is sparse, and $\vec{y} = \mathbf{F}\vec{x}$, with $\mathbf{F} = \mathbf{G} \cdot \mathbf{A}$ a new $M \times N$ matrix. At this point we can think to choose among the infinite solutions of the problem the one that gives the sparsest $\vec{x}$. Unfortunately this is a NP-hard problem. In the following we will mainly talk about asymptotic results, thus we will be interested in the case of large signals $N \to \infty$, keeping signal density ρ and measurement rate α of order one. We also want to keep the components of the signal and of the measurements of order one, hence we consider the elements of the measurement matrix to have mean and variance of order $O(1/N)$.

5.2 Exhaustive algorithm and possible-impossible reconstruction

Which is the theoretical limit for the reconstruction of $\vec{x}$ given $\vec{y}$ and $\mathbf{F}$? If K is the number of non-zero components of $\vec{x}$, we will demonstrate the following claim: We can identify $\vec{x}$ as long as $M \geq K$. To demonstrate it we construct the following algorithm: We want to find the K components of $\vec{x}$ that are different from zero. For this scope we analyze exhaustively all the $\binom{N}{K}$ possible K-component vectors $\vec{r}$ constructed choosing randomly K components from $\vec{x}$. Calling $\mathbf{R}$ the $M \times K$ matrix constructed from $\mathbf{F}$ picking only the columns corresponding to the chosen components of $\vec{r}$, we have to solve the new system $\vec{y} = \mathbf{R}\vec{r}$, where we know $\vec{y}$ and $\mathbf{R}$ as before. If $M = K$, if our choice for $\vec{r}$ is the correct one, that means, if $\vec{r}$ does not have zero elements, the solution of the system is only one and we have solved our problem. If $M > K$, the system is overdetermined and in principle there are no solutions. However, if our choice for $\vec{r}$ is the correct one, there is one solution of the problem, because we know that in the problem there is the *planted* solution, the original one, we constructed the problem (multiplying $\vec{x}$ by $\mathbf{F}$ and obtaining $\vec{y}$) starting from the planted solution.

Thus if $M \geq K$ the solution of the system $\vec{y} = \mathbf{R}\vec{r}$ is unique if and only if $\vec{r}$ is the correct one, namely it contains all the non-zero elements of $\vec{x}$.

This algorithm is able to reconstruct the signal as long as $M \geq K$. However it has a problem: it takes an exponential time in N, it is not useful for practical scopes when N is large. However it shows us that the possible-impossible threshold for the correct reconstruction of the signal is $\alpha_c = \rho$, where $\alpha = \frac{M}{N}$ and $\rho = \frac{K}{N}$.

5.3 The ℓ_1 minimization

Summarizing, the problem of compressed sensing is to infer $\hat{x}$ that solves the equation $|\vec{y} - \mathbf{F}\vec{x}| = 0$, minimizing the ℓ_0 norm: $|\vec{x}|_0 =$ number of components of $\vec{x}$ different from 0. However this is a NP-hard problem. Thus we can think to minimize some other norms, for example the ℓ_1 or ℓ_2 norms, where the ℓ_r norm of a vector $\vec{x}$ is defined as $|\vec{x}|_r = \sum_{i=1}^{N} |x_i|^r$. This is a much easier problem. In fact one should minimize the cost function: $C = |\vec{y} - \mathbf{F}\vec{x}| + \Gamma |\vec{x}|_{1,2}$ (Γ is a constant that should be taken small). The function C is a convex one, thus it can be minimized in a polynomial time. In practice the ℓ_1 norm is used, because it tends to find sparse solutions, this is not the case for the ℓ_2 norm. If one tries to reconstruct the signal using the ℓ_1 minimization, one actually find that this is possible up to $\alpha_{\ell_1} > \rho$. The theoretical threshold for reconstruction is not reached. For $\alpha_{\ell_1} < \rho$, ℓ_1 and ℓ_0 minimization actually give two different results.

5.4 Bayesian reconstruction

In this Section we will apply Bayesian inference introduced in the previous Sections to the CS problem. We can add a Gaussian white noise ξ_μ on the measurement with variance Δ_μ. Thus the problem becomes:

$$y_\mu = \sum_{i=1}^{N} F_{\mu i} x_i + \xi_\mu \quad \mu = 1, \ldots, M \,, \tag{98}$$

Using as usual the Bayes formula (eq. 3), we can write the posterior probability to have a signal $\mathbf{x}$ if we observe the measurement $\mathbf{y}$ and we know the matrix $\mathbf{F}$ as:

$$P(\mathbf{x}|\mathbf{F}, \mathbf{y}) = \frac{1}{Z} \prod_{i=1}^{N} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta_\mu}} e^{-\frac{1}{2\Delta_\mu}\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)^2} \,, \tag{99}$$

where Z, the partition function, is a normalization constant and

$$P(x) = (1-\rho)\delta(x) + \rho\phi(x) \tag{100}$$

is the sparse prior on the signal. We model the signal as stochastic with iid entries, the fraction of non-zero entries being $\rho > 0$ and their distribution being ϕ. In general the signal properties are not known. In the following we will assume that we know the exact parameters (ρ, ϕ, Δ) of the model. If this is not the case, we can infer them by the Expectation Maximization procedure described in Sec. 4.3.

Eq. (99) can be seen as the Boltzmann measure on the disordered system with Hamiltonian

$$H(\mathbf{x}) = -\sum_{i=1}^{N} \log\left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] + \sum_{\mu=1}^{M} \frac{\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)^2}{2\Delta_\mu} \,, \tag{101}$$

where the "disorder" comes from the randomness of the measurement matrix $F_{\mu i}$ and the results y_μ. Stated this way, once again, the problem is similar to a spin glass with N particles interacting with a long-range disordered potential. The real signal is a very special configuration of these particles, the "planted" one, which was used to generate the problem (i.e. the value of the vector $\mathbf{y}$).

5.5 Variational Approaches to Reconstruction in Compressed Sensing: Mean-field varionational Bayes

Consider an inference problem where we need to estimate the posterior

$$P(\vec{x}|\vec{y}) = \frac{P(\vec{y}|\vec{x})P(\vec{x})}{Z} \tag{102}$$

In full generality, the expression of the Gibbs free energy, using a "variational" distributon P_{var} is

$$\mathcal{L} = \left[\log\left(P(\vec{y}|\vec{x})P(\vec{x})\right)\right]_{P_{var}} - \int d\vec{x} P_{var}(\vec{x}) \log P_{var}(\vec{x}) = \left[-E(\vec{y},\vec{x})\right]_{P_{var}} - \int d\vec{x} P_{var}(\vec{x}) \log P_{var}(\vec{x}) \tag{103}$$

with $E = -\log\left(P(\vec{y}|\vec{x})P(\vec{x})\right)$. The mean field approach amounts in using a factorized form:

$$P_{var}(\vec{x}) = \prod_i Q_i(x_i) \tag{104}$$

for which $\mathcal{L}$ can be written

$$\mathcal{L}_{MF} = \left[-E(\vec{y},\vec{x})\right]_{P_{var}} - \sum_i \int dx_i Q_i(x_i) \log Q_i(x_i) \tag{105}$$

Consider one of these variable, say x_i. What is the maximization condition for $Q_i(x_i)$? Clearly it must maximize the following (we denote $\bar{\vec{x}}_i$ all variables x that are *not* x_i) expression:

$$\begin{aligned}
& \left[-E(\vec{y},\vec{x})\right]_{P_{var}} - \int dx_i Q_i(x_i) \log Q_i(x_i) && (106)\\
= & \int dx_i Q_i(x_i) \int \bar{\vec{x}}_i Q(\bar{\vec{x}}_i)(-E(\vec{y},\vec{x})) - \int dx_i Q_i(x_i) \log Q_i(x_i) && (107)\\
= & \int dx_i Q_i(x_i) \left[-E(\vec{y},\vec{x})\right]_{Q(\bar{x})} - \int dx_i Q_i(x_i) \log Q_i(x_i) && (108)\\
= & \int dx_i Q_i(x_i) \log \exp\left[-E(\vec{y},\vec{x})\right]_{Q(\bar{x})} - \int dx_i Q_i(x_i) \log Q_i(x_i) && (109)\\
= & \int dx_i Q_i(x_i) \left(\log \exp\left[-E(\vec{y},\vec{x})\right]_{Q(\bar{x})} - \log Q_i(x_i)\right) && (110)\\
= & \int dx_i Q_i(x_i) \log \frac{\exp\left[-E(\vec{y},\vec{x})\right]_{Q(\bar{x})}}{Q_i(x_i)} && (111)
\end{aligned}$$

One recognize that this is (minus) the KL divergence. The KL divergence is minimal (and thus its negative maximal) for

$$Q_i(x_i) = \frac{1}{Z_i} \exp\left[-E(\vec{y},\vec{x})\right]_{Q(\bar{x})} \tag{112}$$

which is indeed the expected mean-field equation.

Mean-Field Equations for compressed sensing

Let us now move to compressed sensing. Here we have $\log P(\vec{y}|\vec{x}) = -\frac{1}{2\Delta}\sum_\mu (y_\mu - \sum_i F_{\mu i} x_i)^2$. In this case, the energy for the variable i reads

$$-E_i(\vec{y},\vec{x}) = -\frac{1}{2\Delta}\sum_\mu (y_\mu - \sum_{j\neq i} F_{\mu j} x_j - F_{\mu i} x_i)^2 + \sum_{j\neq i} \log P_j^{prior}(x_j) + \log P_i^{prior}(x_i) \tag{113}$$

Its average under $Q(\bar{x})$ reads

$$[-E_i(\vec{y},\vec{x})]_{Q(\bar{x})} = -\frac{1}{2\Delta}\left[\sum_\mu (y_\mu - \sum_{j\neq i} F_{\mu j}x_j - F_{\mu i}x_i)^2 + \sum_{j\neq i}\log P_j^{prior}(x_j) + \log P_i^{prior}(x_i)\right]_{Q(\bar{x})} \quad (114)$$

$$(115)$$

so that (keeping only the relevant terms)

$$\exp\left([-E_i(\vec{y},\vec{x})]_{Q(\bar{x})}\right) \propto P^{prior}(x_i)\exp -\frac{1}{2\Delta}\left[\sum_\mu (y_\mu - \sum_{j\neq i} F_{\mu j}x_j - F_{\mu i}x_i)^2\right]_{Q(\bar{x})} \quad (116)$$

We recognise at this point that the variational distribution is a product of the prior times a Gaussian form:

$$Q_i(x_i) = \frac{1}{Z_i}P^{prior}(x_i)e^{-\frac{(x_i-R_i)^2}{2\Sigma_i^2}}\sqrt{2\pi\Sigma^2} \quad (117)$$

Let us denote by a_i and v_i the mean and variance of the variable x_i under this measure and concentrate now of the remaning term in the exponential, that is $-\frac{1}{2\Delta}\left[\sum_\mu (y_\mu - \sum_{j\neq i} F_{\mu j}x_j - F_{\mu i}x_i)^2\right]_{Q(\bar{x})}$. We have

$$-\frac{1}{2\Delta}\left[\sum_\mu (y_\mu - \sum_{j\neq i} F_{\mu j}x_j - F_{\mu i}x_i)^2\right]_{Q(\bar{x})} \quad (118)$$

$$= \text{cst} - \frac{1}{2\Delta}\sum_\mu\left[(F_{\mu i}x_i)^2 - 2(F_{\mu i}x_i)(y_\mu - \sum_{j\neq i}F_{\mu j}x_j)\right]_{Q(\bar{x})} \quad (119)$$

$$= \text{cst} - \frac{1}{2\Delta}\sum_\mu\left((F_{\mu i}x_i)^2 - 2(F_{\mu i}x_i)(y_\mu - \sum_{j\neq i}F_{\mu j}a_j)\right) \quad (120)$$

$$= \text{cst} - \frac{x_i^2}{2\Delta}\left(\sum_\mu F_{\mu i}^2\right) + \frac{x_i}{2\Delta}\left(2F_{\mu i}(y_\mu - \sum_{j\neq i}F_{\mu j}a_j)\right) = \text{cst} - \frac{x_i^2}{2\Delta}A + \frac{2x_i}{2\Delta}B \quad (121)$$

$$(122)$$

with $A = \left(\sum_\mu F_{\mu i}^2\right)$ and $B = \sum_\mu F_{\mu i}(y_\mu - \sum_{j\neq i}F_{\mu j}a_j)$. It is practical to put this equation under the form of eq(117) using $-\frac{1}{2\Sigma^2}(x-R)^2 = \text{cst} - \frac{x^2}{2\Sigma^2} + \frac{2Rx}{2\Sigma^2}$. We thus identify

$$(\Sigma^2)^{-1} = \frac{\sum_\mu F_{\mu i}^2}{\Delta} \quad (123)$$

$$R = \frac{\sum_\mu F_{\mu i}(y_\mu - \sum_{j\neq i}F_{\mu j}a_j)}{\sum_\mu F_{\mu i}^2} \quad (124)$$

The final algortimh thus reads:

$$a_i = f_a(R_i,\Sigma_i) \quad (125)$$

$$(\Sigma_i^2)^{-1} = \frac{\sum_\mu F_{\mu i}^2}{\Delta} \quad (126)$$

$$R_i = \frac{\sum_\mu F_{\mu i}(y_\mu - \sum_{j\neq i}F_{\mu j}a_j)}{\sum_\mu F_{\mu i}^2} = a_i + \frac{\sum_\mu F_{\mu i}(y_\mu - \sum_j F_{\mu j}a_j)}{\sum_\mu F_{\mu i}^2} \quad (127)$$

The interesting point is that, if using using $\sum_\mu F_{\mu i}^2 = 1$, these are EXACTLY the itereative equations used by Maleki-Donoho. Indeed, using x to denote a and f_a, we have the so-called Iterative Thresolding algortimh, which is therefore just a Bayesian variational one:

$$x^{t+1} = f(A^* z^t + x^t, \Delta) \quad (128)$$

$$z^t = y - Ax^t \quad (129)$$

One can check that the equation for f is indeed the soft and hard thresholding one (when doing the proper, subtle, limit $\Delta \to 0$) for ℓ_1 adn ℓ_0 respectivly.

Expression for the Mean-Field free energy

Let us rewrite slghtly the free energy as

$$\mathcal{L} = [\log(P(\vec{y}|\vec{x}))]_{P_{var}} - \int dx P_{var}(\vec{x}) (\log P_{var}(\vec{x}) - \log P(\vec{x})) \quad (130)$$

$$= [\log(P(\vec{y}|\vec{x}))]_{P_{var}} - \int dx P_{var}(\vec{x}) \left(\log \frac{P_{var}(\vec{x})}{P(\vec{x})} \right) \quad (131)$$

$$= [\log(P(\vec{y}|\vec{x}))]_{P_{var}} - D_{KL}(P_{var}(\vec{x})||P(\vec{x})) \quad (132)$$

$$(133)$$

the first term is easy to compute. The divergence one also; Given a prior $P(x_i)$, our algoritmh estimates the distribution as eq.117 and assume the x_i are decorrelated. We thus compute the KL divergence separatly for each terms. It reads

$$D_{KL}^i = \int dx \frac{1}{\tilde{Z}_i} P(x_i) e^{-\frac{(x_i - R_i)^2}{2\Sigma_i^2}} \log \left(\frac{1}{\tilde{Z}_i} e^{-\frac{(x_i - R_i)^2}{2\Sigma_i^2}} \right) \quad (134)$$

$$= -\log \tilde{Z}_i - \int dx_i P_{var}(x_i) \frac{(x_i - R_i)^2}{2\Sigma_i^2} = -\log \tilde{Z}_i - \frac{1}{2\Sigma_i^2} \left(< x_i^2 > + R_i^2 - 2R_i < x_i > \right) \quad (135)$$

$$= -\log \tilde{Z}_i - \frac{c_i + a_i^2 + R_i^2 - 2R_i a_i}{2\Sigma_i^2} = -\log \tilde{Z}_i - \frac{c_i + (a_i - R_i)^2}{2\Sigma_i^2} \quad (136)$$

This expression is very convenient, as it involves only the norm (which we always need to compute anyway when we perform the integral) and the average and variance of the distribution.

The full cost function thus yields

$$\mathcal{L}_{MF} = -\sum_\mu \frac{(y_\mu - \sum_i F_{\mu i} a_i)^2 + \sum_i F_{\mu i}^2 c_i}{2\Delta} - \sum_\mu \frac{1}{2} \log \Delta_\mu + \sum_i \log \tilde{Z}_i + \sum_i \frac{c_i + (a_i - R_i)^2}{2\Sigma_i^2} \quad (137)$$

Mean-Field Reloaded

The nice thing with eq.(137) is that if one forget that a_i and c_i are actually a fonction of R_i and Σ_i, then one can recover the mean field algorithm very easyly by just deriving the fonction.

A steepest descent appraoch to MF

Let us first derive a number of identity. We need first to know the properties of the derivative of the $log\tilde{Z}_i$. We have

$$\frac{d\log \tilde{Z}_i}{dR_i} = \frac{f_a(R_i,\Sigma_i)-R_i}{\Sigma_i^2} \tag{138}$$

$$\frac{d\log \tilde{Z}_i}{d\Sigma_i^2} = \frac{(f_a(R_i,\Sigma_i)-R_i)^2+f_c(R_i,\Sigma_i)}{2\Sigma_i^4} \tag{139}$$

$$\tag{140}$$

Then one needs to estimate the derivative of a and c with respect to R and Z.

$$\frac{da_i(R_i,\Sigma_i)}{dR_i} = \frac{c_i}{\Sigma_i^2} \tag{141}$$

$$\frac{da_i(R_i,\Sigma_i)}{d\Sigma_i^2} = \frac{c_i}{\Sigma_i^4}(a_i-R) \tag{142}$$

$$\frac{dc_i(R_i,\Sigma_i)}{dR_i} = f_d(R_i,,\Sigma_i) \tag{143}$$

$$\frac{dc_i(R_i,\Sigma_i)}{d\Sigma_i^2} = \frac{1}{\Sigma_2}\left[f_d(a_i-R)+\frac{c_i^2}{\Sigma_2}\right] \tag{144}$$

$$\tag{145}$$

So we just need a new fonction $f_d(R_i,,\Sigma_i)=\frac{dc_i(R_i,\Sigma_i)}{dR_i}$.

At this point, one can compute the derivative of the free energy:

$$\begin{aligned}\frac{d\mathcal{L}_{MF}}{dR_i} &= \sum_\mu \frac{2(y_\mu-\sum_i F_{\mu i}a_i)F_{\mu i}\frac{da_i(R_i,\Sigma_i)}{dR_i}-F_{\mu i}^2\frac{dc_i(R_i,\Sigma_i)}{dR_i}}{2\Delta}+\frac{a_i-R_i}{\Sigma_2^2} \\ &+ \frac{1}{2\Sigma_i^2}\left[\frac{dc_i(R_i,\Sigma_i)}{dR_i}+2(a_i-R_i)(\frac{da_i(R_i,\Sigma_i)}{dR_i}-1)\right]\end{aligned} \tag{146}$$

$$\begin{aligned}\frac{d\mathcal{L}_{MF}}{d\Sigma_i^2} &= \sum_\mu \frac{2(y_\mu-\sum_i F_{\mu i}a_i)F_{\mu i}\frac{da_i(R_i,\Sigma_i)}{d\Sigma_i^2}-F_{\mu i}^2\frac{dc_i(R_i,\Sigma_i)}{d\Sigma_i^2}}{2\Delta}+\frac{(a_i-R_i)^2}{2\Sigma_i^4} \\ &+ \frac{1}{2\Sigma_i^2}\left[\frac{dc_i(R_i,\Sigma_i)}{d\Sigma_i^2}+2(a_i-R_i)\frac{da_i(R_i,\Sigma_i)}{d\Sigma_i^2}\right]-\frac{1}{2\Sigma_i^4}\left[c_i+(a_i-R_i)^2\right]\end{aligned} \tag{147}$$

so that the algoritmh reads

$$R_i^{t+1} = R_i^t+\gamma\frac{d\mathcal{L}_{MF}}{dR_i} \tag{148}$$

$$(\Sigma_i^2)^{t+1} = (\Sigma_i^2)^{t+1}+\gamma\frac{d\mathcal{L}_{MF}}{d\Sigma_i^2} \tag{149}$$

$$\tag{150}$$

with γ a properly choosen dampening factor.

EM for the noise

Now, imagine we would learn the value of the noise. In this case EM leads to

$$\Delta = \frac{1}{M}\sum_\mu\left(\left[y_\mu-\sum_i F_{\mu i}a_i\right]^2+\sum_i F_{\mu i}v_i\right) \tag{151}$$

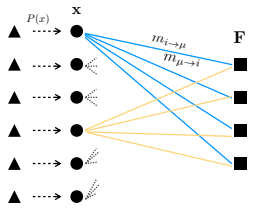

Figure 8

Which is what we would expect naivly.

5.6 The belief propagation reconstruction algorithm for compressed sensing

Exact computation of the averages requires exponential time and is thus intractable. To approximate the expectations we will use a variant of the belief propagation (BP) algorithm [35, 36]. Indeed, message passing has been shown very efficient in terms of both precision and speed for the CS problem.

5.6.1 Belief Propagation recursion

The canonical BP equations for the probability measure $P(\mathbf{x}|\mathbf{F},\mathbf{y})$, Eq. (99), are expressed in terms of $2MN$ "messages", $m_{j\to\mu}(x_j)$ and $m_{j\to\mu}(x_j)$, which are probability distribution functions (see fig. 8 for the underlying graphical model). They read:

$$m_{\mu\to i}(x_i) = \frac{1}{Z^{\mu\to i}} \int \prod_{j\neq i} \mathrm{d}x_j e^{-\frac{1}{2\Delta_\mu}(\sum_{j\neq i} F_{\mu j}x_j + F_{\mu i}x_i - y_\mu)^2} \prod_{j\neq i} m_{j\to\mu}(x_j)\,, \tag{152}$$

$$m_{i\to\mu}(x_i) = \frac{1}{Z^{i\to\mu}} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] \prod_{\gamma\neq\mu} m_{\gamma\to i}(x_i)\,, \tag{153}$$

where $Z^{\mu\to i}$ and $Z^{i\to\mu}$ are normalization factors ensuring that $\int \mathrm{d}x_i\, m_{\mu\to i}(x_i) = \int \mathrm{d}x_i\, m_{i\to\mu}(x_i) = 1$. These coupled integral equations for the messages are too complicated to be of any practical use. However, in the large N limit, when the matrix elements $F_{\mu i}$ scale like $1/\sqrt{N}$, one can simplify these canonical equations.

Using the Hubbard-Stratonovich transformation

$$e^{-\frac{\omega^2}{2\Delta}} = \frac{1}{\sqrt{2\pi\Delta}} \int \mathrm{d}\lambda\; e^{-\frac{\lambda^2}{2\Delta} + \frac{i\lambda\omega}{\Delta}}\,, \tag{154}$$

for $\omega = (\sum_{j\neq i} F_{\mu j}x_j)$ we can simplify Eq. (152) as

$$m_{\mu\to i}(x_i) = \frac{1}{Z^{\mu\to i}\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta_\mu}(F_{\mu i}x_i - y_\mu)^2} \int \mathrm{d}\lambda e^{-\frac{\lambda^2}{2\Delta_\mu}} \prod_{j\neq i} \left[\int \mathrm{d}x_j m_{j\to\mu}(x_j) e^{\frac{F_{\mu j}x_j}{\Delta_\mu}(y_\mu - F_{\mu i}x_i + i\lambda)}\right]. \tag{155}$$

Now we expand the last exponential around zero because the term $F_{\mu j}$ is small in N, we keep all terms that are of $O(1/N)$. Introducing means and variances as new "messages"

$$a_{i\to\mu} \equiv \int \mathrm{d}x_i\, x_i\, m_{i\to\mu}(x_i)\,, \tag{156}$$

$$v_{i\to\mu} \equiv \int \mathrm{d}x_i\, x_i^2\, m_{i\to\mu}(x_i) - a_{i\to\mu}^2\,, \tag{157}$$

we obtain

$$m_{\mu\to i}(x_i) = \frac{1}{Z^{\mu\to i}\sqrt{2\pi\Delta_\mu}} e^{-\frac{1}{2\Delta_\mu}(F_{\mu i}x_i - y_\mu)^2} \int \mathrm{d}\lambda e^{-\frac{\lambda^2}{2\Delta_\mu}} \prod_{j\neq i} \left[e^{\frac{F_{\mu j}a_{j\to\mu}}{\Delta_\mu}(y_\mu - F_{\mu i}x_i + i\lambda) + \frac{F_{\mu j}^2 v_{j\to\mu}}{2\Delta_\mu^2}(y_\mu - F_{\mu i}x_i + i\lambda)^2} \right] . \tag{158}$$

Performing the Gaussian integral over λ, we obtain

$$m_{\mu\to i}(x_i) = \frac{1}{\tilde{Z}^{\mu\to i}} e^{-\frac{x_i^2}{2}A_{\mu\to i} + B_{\mu\to i}x_i} , \qquad \tilde{Z}^{\mu\to i} = \sqrt{\frac{2\pi}{A_{\mu\to i}}} e^{\frac{B_{\mu\to i}^2}{2A_{\mu\to i}}} , \tag{159}$$

where the normalization $\tilde{Z}^{\mu\to i}$ contains all the x_i-independent factors, and we have introduced the scalar messages:

$$A_{\mu\to i} = \frac{F_{\mu i}^2}{\Delta_\mu + \sum_{j\neq i} F_{\mu j}^2 v_{j\to\mu}} , \tag{160}$$

$$B_{\mu\to i} = \frac{F_{\mu i}(y_\mu - \sum_{j\neq i} F_{\mu j}a_{j\to\mu})}{\Delta_\mu + \sum_{j\neq i} F_{\mu j}^2 v_{j\to\mu}} , \tag{161}$$

The noiseless case corresponds to $\Delta_\mu = 0$.

To close the equations on messages $a_{i\to\mu}$ and $v_{i\to\mu}$ we notice that

$$m_{i\to\mu}(x_i) = \frac{1}{\tilde{Z}^{i\to\mu}} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] e^{-\frac{x_i^2}{2}\sum_{\gamma\neq\mu} A_{\gamma\to i} + x_i \sum_{\gamma\neq\mu} B_{\gamma\to i}} . \tag{162}$$

Messages $a_{i\to\mu}$ and $v_{i\to\mu}$ are respectively the mean and variance of the probability distribution $m_{i\to\mu}(x_i)$. It is also useful to define the local beliefs a_i and v_i as

$$a_i \equiv \int \mathrm{d}x_i\, x_i\, m_i(x_i) , \tag{163}$$

$$v_i \equiv \int \mathrm{d}x_i\, x_i^2\, m_i(x_i) - a_i^2 , \tag{164}$$

where

$$m_i(x_i) = \frac{1}{\tilde{Z}^i} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] e^{-\frac{x_i^2}{2}\sum_{\gamma} A_{\gamma\to i} + x_i \sum_{\gamma} B_{\gamma\to i}} . \tag{165}$$

For a general function $\phi(x_i)$ let us define the probability distribution

$$\mathcal{M}_\phi(\Sigma^2, R, x) = \frac{1}{\hat{Z}(\Sigma^2, R)} \left[(1-\rho)\delta(x) + \rho\phi(x)\right] \frac{1}{\sqrt{2\pi}\Sigma} e^{-\frac{(x-R)^2}{2\Sigma^2}} , \tag{166}$$

where $\hat{Z}(\Sigma^2, R)$ is a normalization. We define the average and variance of $\mathcal{M}_\phi$ as

$$f_a(\Sigma^2, R) \equiv \int \mathrm{d}x\, x\, \mathcal{M}(\Sigma^2, R, x) , \tag{167}$$

$$f_c(\Sigma^2, R) \equiv \int \mathrm{d}x\, x^2\, \mathcal{M}(\Sigma^2, R, x) - f_a^2(\Sigma^2, R) , \tag{168}$$

(where we do not write explicitly the dependence on ϕ). Notice that:

$$f_a(\Sigma^2, R) = R + \Sigma^2 \frac{\mathrm{d}}{\mathrm{d}R} \log \hat{Z}(\Sigma^2, R) , \tag{169}$$

$$f_c(\Sigma^2, R) = \Sigma^2 \frac{\mathrm{d}}{\mathrm{d}R} f_a(\Sigma^2, R) . \tag{170}$$

The closed form of the BP update is

$$a_{i\to\mu} = f_a\left(\frac{1}{\sum_{\gamma\neq\mu}A_{\gamma\to i}},\frac{\sum_{\gamma\neq\mu}B_{\gamma\to i}}{\sum_{\gamma\neq\mu}A_{\gamma\to i}}\right), \qquad a_i = f_a\left(\frac{1}{\sum_{\gamma}A_{\gamma\to i}},\frac{\sum_{\gamma}B_{\gamma\to i}}{\sum_{\gamma}A_{\gamma\to i}}\right), \tag{171}$$

$$v_{i\to\mu} = f_c\left(\frac{1}{\sum_{\gamma\neq\mu}A_{\gamma\to i}},\frac{\sum_{\gamma\neq\mu}B_{\gamma\to i}}{\sum_{\gamma\neq\mu}A_{\gamma\to i}}\right), \qquad v_i = f_c\left(\frac{1}{\sum_{\gamma}A_{\gamma\to i}},\frac{\sum_{\gamma}B_{\gamma\to i}}{\sum_{\gamma}A_{\gamma\to i}}\right). \tag{172}$$

For a general signal model $\phi(x_i)$ the functions f_a and f_c can be computed using a numerical integration over x_i. In special cases, like the case of Gaussian ϕ these functions are easily computed analytically. Eqs. (156-157) together with (160-161) and (162) then lead to closed iterative message passing equations, which can be solved by iterations. There equations can be used for any signal $\mathbf{s}$, and any matrix $\mathbf{F}$. When a fixed point of the BP equations is reached, the elements of the original signal are estimated as $x_i^* = a_i$, and the corresponding variance v_i can be used to quantify the correctness of this estimate. Perfect reconstruction is found when the messages converge to a fixed point such that $a_i = s_i$ and $v_i = 0$.

The use of mean and variances instead of the canonical BP messages is exact in the large N limit, thanks to the fact that the matrix is not sparse (a sum like $\sum_i F_{\mu_i} x_i$ contains of order N non-zero terms), and each element of the matrix F scales as $O(1/\sqrt{N})$.

5.6.2 The TAP form of the message passing algorithm

In the message-passing form of BP described above, $2M \times N$ messages are sent, one between each variable component i and each measurement, in each iteration. In fact, it is possible to rewrite the BP equations in terms of $N + M$ messages instead of 2 $M \times N$, always within the assumption that the F matrix is not sparse, and that all its elements scale as $O(1/\sqrt{N})$. In statistical physics terms, this corresponds to the Thouless-Anderson-Palmer equations (TAP) [37] used in the study of spin glasses. In the large N limit, these are asymptotically equivalent (only $o(1)$ terms are neglected) to the BP equations. Going from BP to TAP is, in the compressed sensing literature, the step to go from the rBP [38] to the AMP [39] algorithm. Let us now show how to take this step.

In the large N limit, it is clear from (171-172) that the messages $a_{i\to\mu}$ and $v_{i\to\mu}$ are nearly independent of μ. However, one must be careful to keep the correcting "Onsager reaction terms". Let us define

$$\omega_\mu = \sum_i F_{\mu i} a_{i\to\mu}, \qquad V_\mu = \sum_i F_{\mu i}^2 v_{i\to\mu}, \tag{173}$$

$$\Sigma_i^2 = \frac{1}{\sum_\mu A_{\mu\to i}}, \qquad R_i = \frac{\sum_\mu B_{\mu\to i}}{\sum_\mu A_{\mu\to i}}. \tag{174}$$

Then we have

$$\Sigma_i^2 = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu - F_{\mu i}^2 v_{i\to\mu}}\right]^{-1} = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu}\right]^{-1}, \tag{175}$$

$$R_i = \left[\sum_\mu \frac{F_{\mu i}(y_\mu - \omega_\mu + F_{\mu i} a_{i\to\mu})}{\Delta_\mu + V_\mu - F_{\mu i}^2 v_{i\to\mu}}\right]\left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu - F_{\mu i}^2 v_{i\to\mu}}\right]^{-1} = a_i + \frac{\sum_\mu F_{\mu i}\frac{(y_\mu-\omega_\mu)}{\Delta_\mu+V_\mu}}{\sum_\mu F_{\mu i}^2 \frac{1}{\Delta_\mu+V_\mu}}. \tag{176}$$

In order to compute $\omega_\mu = \sum_i F_{\mu i} a_{i\to\mu}$, we see that when expressing $a_{i\to\mu}$ in terms of a_i we need to keep all corrections that are linear in the matrix element $F_{\mu i}$

$$a_{i\to\mu} = f_a\left(\frac{1}{\sum_\nu A_{\nu\to i} - A_{\mu\to i}},\frac{\sum_\nu B_{\nu\to i} - B_{\mu\to i}}{\sum_\nu A_{\nu\to i} - A_{\mu\to i}}\right) = a_i - B_{\mu\to i}\Sigma^2 \frac{\partial f_a}{\partial R}\left(\Sigma_i^2, R_i\right). \tag{177}$$

Therefore

$$\omega_\mu = \sum_i F_{\mu i} a_i - \frac{(y_\mu - \omega_\mu)}{\Delta_\mu + V_\mu} \sum_i F_{\mu i}^2 v_i \,. \tag{178}$$

The computation of V_μ is similar, this time all the corrections are negligible in the limit $N \to \infty$.

Finally, we get the following closed system of iterative TAP equations that involve only matrix multiplication:

$$V_\mu^{t+1} = \sum_i F_{\mu i}^2 \, v_i^t \,, \tag{179}$$

$$\omega_\mu^{t+1} = \sum_i F_{\mu i} \, a_i^t - \frac{(y_\mu - \omega_\mu^t)}{\Delta_\mu + V_\mu^t} \sum_i F_{\mu i}^2 \, v_i^t \,, \tag{180}$$

$$(\Sigma_i^{t+1})^2 = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu^{t+1}} \right]^{-1} , \tag{181}$$

$$R_i^{t+1} = a_i^t + \frac{\sum_\mu F_{\mu i} \frac{(y_\mu - \omega_\mu^{t+1})}{\Delta_\mu + V_\mu^{t+1}}}{\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu^{t+1}}} \,, \tag{182}$$

$$a_i^{t+1} = f_a \left((\Sigma_i^{t+1})^2, R_i^{t+1} \right) , \tag{183}$$

$$v_i^{t+1} = f_c \left((\Sigma_i^{t+1})^2, R_i^{t+1} \right) . \tag{184}$$

We see that the signal model $P(x_i) = (1-\rho)\delta(x_i) + \rho\phi(x_i)$ assumed in the probabilistic approach appears only through the definitions (167-168) of the two functions f_a and f_c . In the case where the signal model is chosen as Gauss-Bernoulli. Equations (179-184) are equivalent to the (generalized) approximate message passing of [39, 40].

A reasonable initialization of these equations is

$$a_i^{t=0} = \rho \int \mathrm{d}x \, x \, \phi(x) \,, \tag{185}$$

$$v_i^{t=0} = \rho \int \mathrm{d}x \, x^2 \, \phi(x) - \left(a_i^{t=0} \right)^2 , \tag{186}$$

$$\omega_\mu^{t=0} = y_\mu \,. \tag{187}$$

5.6.3 Further simplification for measurement matrices with random entries

For some special classes of random measurement matrices $\mathbf{F}$, the TAP equations (179-182) can be simplified further. Let us start with the case of a homogeneous matrix $\mathbf{F}$ with iid random entries of zero mean and variance $1/N$ (the distribution can be anything as long as the mean and variance are fixed). The simplification can be understood as follows. Consider for instance the quantity V_μ. Let us define $\overline{V}$ as the average of V_μ with respect to different realizations of the measurement matrix F.

$$\overline{V} = \sum_{i=1}^N \overline{F_{\mu i}^2} v_i = \frac{1}{N} \sum_{i=1}^N v_i \,. \tag{188}$$

The variance is

$$\begin{aligned} \mathrm{var}\, V &\equiv \overline{(V_\mu - \overline{V})^2} = \sum_{i \neq j} \overline{\left(F_{\mu i}^2 - \frac{1}{N} \right) \left(F_{\mu j}^2 - \frac{1}{N} \right)} \, v_i v_j + \sum_{i=1}^N \overline{\left(F_{\mu i}^2 - \frac{1}{N} \right)^2} \, v_i^2 \\ &= 0 + \frac{2}{N} \left(\frac{1}{N} \sum_{i=1}^N v_i^2 \right) = O\left(\frac{1}{N} \right) . \end{aligned} \tag{189}$$

Since the average is of order one and the variance of order $1/N$, in the limit of large N we can hence neglect the dependence on the index μ and consider all V_μ equal to their average. The same argument can be repeated for all the terms that contain $F_{\mu i}^2$. Hence for the homogeneous matrix $\mathbf{F}$ with iid random entries of zero mean and variance $1/N$, one can effectively "replace" every $F_{\mu i}^2$ by $1/N$ in Eqs. (181-182) and (179-180). The iteration equations then take the simpler form (assuming for simplicity that $\Delta_\mu = \Delta$)

$$V = \frac{1}{N}\sum_i v_i \,, \tag{190}$$

$$\omega_\mu = \sum_i F_{\mu i} a_i - \frac{(y_\mu - \omega_\mu)}{\Delta + V}\left[\frac{1}{N}\sum_i v_i\right] \,, \tag{191}$$

$$\Sigma^2 = \frac{\Delta + V}{\alpha} \,, \tag{192}$$

$$R_i = a_i + \sum_\mu F_{\mu i}\frac{(y_\mu - \omega_\mu)}{\alpha} \,. \tag{193}$$

$$a_i = f_a\left(\Sigma^2, R_i\right) \,, \tag{194}$$

$$v_i = f_c\left(\Sigma^2, R_i\right) \,. \tag{195}$$

These equations can again be solved by iteration. They only involve $2(M+N+1)$ variables. For a general matrix $\mathbf{F}$ one iteration of the above algorithm takes $O(NM)$ steps (and in practice we observed that the number of iterations needed for convergence is basically independent of N). For matrices that can be computed recursively (i.e. without storing all their NM elements) a speed up of this algorithm is possible, as the message passing loop takes only $O(M+N)$ steps.

5.6.4 The phase diagram for noiseless measurements and the optimal Bayes case

In this section we turn the equations from the previous section into phase diagrams to display the performance of belief propagation in CS reconstruction. We mainly discuss the noiseless case, with random homogeneous measurement matrices, this is a benchmark case that has been widely used to demonstrate the power of the ℓ_1 reconstruction. We use measurement matrices with iid entries with zero mean and variance $1/N$ (we remind that our approach is independent of the distribution of the iid matrix elements and depends only on their mean and variance). We assume to know exactly the distribution of the signal (the prior is exact).

In Fig. 9 we show the free entropy density $\Phi(D)$ at fixed squared distance D between the inferred signal and the real one. ϕ is Gaussian with zero mean and unit variance. The free entropy $\Phi(D)$ is computed using the replica method. The dynamics of the message passing algorithm (without learning) is a gradient dynamics leading to a maximum of the free-entropy $\Phi(D)$ starting from high distance D. As expected, we see in Fig. 9 that $\Phi(D)$ has a global maximum at $D = 0$ if and only if $\alpha > \rho$, which confirms that the Bayesian optimal inference is in principle able to reach the theoretical limit $\alpha = \rho$ for exact reconstruction. The figure shows the existence of a critical measurement rate $\alpha_{\rm BP}(\rho) > \rho$, below which a secondary local maximum of $\Phi(D)$ appears at $D > 0$. When this secondary maximum exists, the BP algorithm converges instead to it, and does not reach exact reconstruction. The threshold $\alpha_{\rm BP}(\rho)$ is obtained analytically as the smallest value of α such that $\Phi(D)$ is monotonic. The behavior of $\Phi(D)$ is typical of a first order transition. The equilibrium transition appears at a number of measurement per unknown $\alpha = \rho$, which is the point where the global maximum of $\Phi(D)$ switches discontinuously from being at $D = 0$ (when $\alpha > \rho_0$) to a value $D > 0$. In this sense the value $\alpha = \alpha_{\rm BP}(\rho_0)$ appears like a spinodal point: it is the point below which the global maximum of $\Phi(D)$ is no longer reached by the dynamics. Instead, in the regime below the spinodal ($\alpha < \alpha_{\rm BP}(\rho)$, the dynamical evolution is attracted to a metastable non-optimal state with $D > 0$.

The spinodal transition is the physical reason that limits the performance of the BP algorithm. The convergence time of BP diverges around the spinodal transition $\alpha_{\rm BP}$.

Notice that the ℓ_1 transition at α_{ℓ_1} is continuous (second order), whereas the spinodal transition is discontinuous (first order). The transition at $\alpha_{\rm BP}$ is called a spinodal transition in the mean field theory of

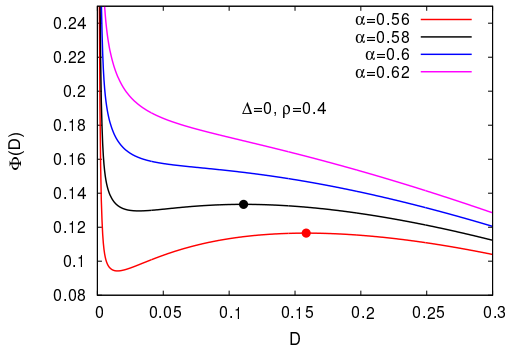

Figure 9: Left: The free entropy, $\Phi(D)$, is plotted as a function of $D = \left\langle \sum_i (\hat{x}_i - x_i)^2/N \right\rangle$ for $\rho = 0.4$ and several measurement rates α in the Bayesian approach (when both the signal and the signal model are described by a Gauss-Bernoulli distribution). The evolution of the BP algorithm is basically a steepest ascent in $\Phi(D)$ starting from a large value of D. Such ascent goes to the global maximum at $D = 0$ for large value of α but is blocked in the local maximum that appears for $\alpha < \alpha_{\rm BP}(\rho = 0.4) \approx 0.59$. For $\alpha < \rho$, the global maximum is not at $D = 0$ and exact inference is impossible.

first order phase transitions. It is similar to the one found in the cooling of liquids which go into a super-cooled glassy state instead of crystallizing, and appears in the decoding of error correcting codes [41, 42] as well. This difference might seem formal, but it is absolutely essential for what concerns the possibility of achieving the theoretically optimal reconstruction with the use of seeding measurement matrices (as discussed in the next section).

In Fig. 10 we show how the critical value $\alpha_{\rm BP}$ depends on the signal density ρ and on the type of the signal, for several Gauss-Bernoulli signals. In this figure we still assume that the signal distribution is known. We compare to the Donoho-Tanner phase transition α_{ℓ_1} that gives the limit for exact reconstruction with the ℓ_1 minimization [43, 39, 44], and to the information-theoretical limit for exact reconstruction $\alpha = \rho$.

Note that for some signals, e.g. the mixture of Gaussians $\Phi(x) = [\mathcal{N}(-1, 0.1) + \mathcal{N}(1, 0.1)]/2$, there is a region of signal densities (here $\rho \simeq 0.8$) for which the BP reconstruction is possible down to the optimal subsampling rates $\alpha = \rho$.

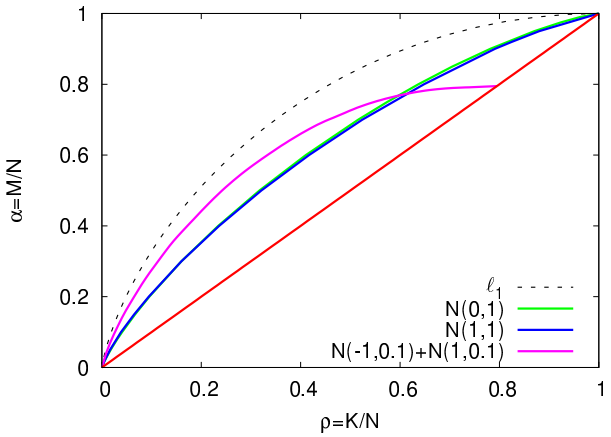

Figure 10: Phase diagram for the BP reconstruction in the optimal Bayesian case when the signal model is matching the empirical distribution of signal elements. The elements of the $M \times N$ measurement matrix $\mathbf{F}$ are iid variables with zero mean and variance $1/N$. The spinodal transition $\alpha_{\rm BP}(\rho)$ is computed with the asymptotic replica analysis and plotted for the following signal distributions: $\phi(x) = \mathcal{N}(0,1)$ (green), $\phi(x) = \mathcal{N}(1,1)$ (blue) $\phi(x) = [\mathcal{N}(-1,0.1) + \mathcal{N}(1,0.1)]/2$. The data are compared to the Donoho-Tanner phase transition $\alpha_{\ell_1}(\rho)$ (dashed) for ℓ_1 reconstruction that does not depend on the signal distribution, and to the theoretical limit for exact reconstruction $\alpha = \rho$ (red). The left hand side represents the undersampling rate α as a function of the signal density ρ_0.

References

[1] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[2] J. M. Johnstone D. L. Donoho. *Biometrika*, 81(3):425, 1994.

[3] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, 24:2613–2626, 1981.

[4] S. F. Edwards and P. W. Anderson. Theory of spin glasses. *J. Phys. F: Met. Phys.*, 5:965, 1975.

[5] P. M. Goldbart, Goldenfeld N., and Sherrington D. *Stealing the gold: a celebration of the pioneering physics of Sam Edwards*. Oxford Science Publication, Oxford, 2004.

[6] J. Wehr and M. Aizenman. Fluctuations of extensive functions of quenched random couplings . *J. Stat. Phys.*, 60:287, 1990.

[7] Francesco Guerra and Fabio Lucio Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230:71–79, 2002. 10.1007/s00220-002-0699-y.

[8] C. Moore and S. Mertens. *The Nature of Computation*. Oxford press, Oxford, 2011.

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

[10] A. Georges, D. Hansel, P. Le Doussal, and J.M. Maillard. The replica momenta of a spin-glass and the phase diagram of n-colour ashkin-teller models. *J Phys. France*, 48:1–9, 1987.

[11] Gavin E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60:2721–2726, Sep 1999.

[12] Florent Krzakala and Lenka Zdeborová. Hiding quiet solutions in random constraint satisfaction problems. *Phys. Rev. Lett.*, 102:238701, 2009.

[13] Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. *Proc. FOCS*, 2008.

[14] D. Sherrington and S. Kirkpatrick. Solvable model of a spin glass. *Phys. Rev. Lett.*, 35:1792, 1975.

[15] G. Parisi. Infinite number of order parameters for spin-glasses. *Phys. Rev. Lett.*, 43:1754, 1979.

[16] G. Parisi. The order parameter for spin glasses: a function on the interval 0–1. *J. Phys. A*, 13:1101, 1980.

[17] J. R. L. de Almeida and D. J. Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin-glass model. *J. Phys. A*, 11:983–990, 1978.

[18] B. Derrida. *Phys. Rev. Lett.*, (45):79, 1980.

[19] B. Derrida. *Phys. Rev. B*, (24):2613, 1981.

[20] A. Crisanti and H.-J. Sommers. *Z. Phys. B*, (87):341.

[21] D. J. Thouless. Spin-glass on a Bethe lattice. *Phys. Rev. Lett.*, 56:1082, 1986.

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

[23] A. P. Young, editor. *Spin Glasses and Random Fields*. World Scientific, Singapore, 1998.

[24] K. Binder and A. P. Young. Spin glasses: Experimental facts, theoretical concepts and open questions. *Rev. Mod. Phys.*, 58:801, 1986.

[25] J. R. L. de Almeida and D. J. Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin glass model. *J. Phys. A*, 11:983, 1978.

[26] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[27] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, 2011.

[28] Yukito Iba. The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, 32:3875, 1999.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:138, 1977.

[30] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. (arXiv:1202.1499), 2012.

[31] L. Massoulie. Community detection thresholds and the weak ramanujan property. (arXiv:1311.3085), 2013.

[32] Lenka Zdeborová and Florent Krzakala. Generalization of the cavity method for adiabatic evolution of gibbs states. *Phys. Rev. B*, 81:224205, 2010.

[33] Marc Mézard and Andrea Montanari. Reconstruction on trees and spin glass transition. *J. Stat. Phys.*, 124:1317–1350, september 2006.

[34] L. Zdeborová and F. Krzakala. Phase transitions in the coloring of random graphs. *Phys. Rev. E*, 76:031131, 2007.

[35] F. R. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47(2):498–519, 2001.

[36] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, pages 239–236. Morgan Kaufmann, San Francisco, CA, USA, 2003.

[37] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin-glass'. *Phil. Mag.*, 35:593–601, 1977.

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

[39] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914–18919, 2009.

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

[41] Richardson T. and Urbanke R. *Modern Coding Theory*. Cambridge University Press, 2008.

[42] Mézard M. nad Montanari A. *Information, Physics, and Computation*. Oxford Press, 2009.

[43] Donoho D. and Tanner J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Phil. Trans. R. Soc. A*, 367:4273–4293, 2009.

[44] Kabashima Y., Wadayama T., and Tanaka T. A typical reconstruction limit of compressed sensing based on lp-norm minimization. *J. Stat. Mech.*, page L09003, 2009.