

# Statistical physics of inference: thresholds and algorithms

Lenka Zdeborová & Florent Krzakala

**To cite this article:** Lenka Zdeborová & Florent Krzakala (2016) Statistical physics of inference: thresholds and algorithms, *Advances in Physics*, 65:5, 453-552, DOI: [10.1080/00018732.2016.1211393](https://doi.org/10.1080/00018732.2016.1211393)

**To link to this article:** <https://doi.org/10.1080/00018732.2016.1211393>



Published online: 19 Aug 2016.



Submit your article to this journal [↗](#)



Article views: 5191



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 96 View citing articles [↗](#)

## REVIEW ARTICLE

### Statistical physics of inference: thresholds and algorithms

Lenka Zdeborová<sup>a\*</sup> and Florent Krzakala<sup>b</sup>

<sup>a</sup>*Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette, France;*

<sup>b</sup>*Laboratoire de Physique Statistique, Sorbonne Universités, Université Pierre et Marie Curie Paris 6, Ecole Normale Supérieure, 75005 Paris, France*

(Received 28 September 2015; accepted 30 June 2016)

Many questions of fundamental interest in today's science can be formulated as inference problems: some partial, or noisy, observations are performed over a set of variables and the goal is to recover, or infer, the values of the variables based on the indirect information contained in the measurements. For such problems, the central scientific questions are: Under what conditions is the information contained in the measurements sufficient for a satisfactory inference to be possible? What are the most efficient algorithms for this task? A growing body of work has shown that often we can understand and locate these fundamental barriers by thinking of them as phase transitions in the sense of statistical physics. Moreover, it turned out that we can use the gained physical insight to develop new promising algorithms. The connection between inference and statistical physics is currently witnessing an impressive renaissance and we review here the current state-of-the-art, with a pedagogical focus on the Ising model which, formulated as an inference problem, we call the planted spin glass. In terms of applications we review two classes of problems: (i) inference of clusters on graphs and networks, with community detection as a special case and (ii) estimating a signal from its noisy linear measurements, with compressed sensing as a case of sparse estimation. Our goal is to provide a pedagogical review for researchers in physics and other fields interested in this fascinating topic.

**PACS:** 89.70.Eg interdisciplinary; computational complexity; 75.50.Lk magnetic properties of materials: spin glasses and other random magnets; 64.60.aq networks; 89.75.-k complex systems

**Keywords:** Bayesian inference; spin glass theory; compressed sensing; stochastic block model; belief propagation; phase transitions in computer science

	Contents	PAGE
1	Introduction	455
1.1	Organization of the manuscript	457
1.2	What is statistical inference?	458
1.2.1	Terminology and simple examples	459
1.2.2	Teacher–student scenario	462
1.2.3	Bayes-optimal versus mismatched	463
1.2.4	Estimators	464
1.2.5	The high-dimensional limit	466
1.3	Statistical physics and inference	467
1.3.1	Basic dictionary	467
1.3.2	The teacher–student scenario defines the planted ensemble	468
1.3.3	Inference via Monte Carlo and variational mean-field	470
1.3.4	Useful statistical physics concepts	471
2	Planted SG as a paradigm of statistical inference	473

---

\*Corresponding author. Email: [lenka.zdeborova@cea.fr](mailto:lenka.zdeborova@cea.fr)

2.1	Definition of the planted SG	473
2.2	Nishimori's mapping via gauge transformation	475
2.3	Bayes-optimality on the Nishimori line	476
2.4	Parameter mismatch and learning	478
2.5	Relations between planted, quenched and annealed disorders	479
2.5.1	The quenched and annealed average	480
2.5.2	Properties of the planted ensemble	481
2.5.3	Quiet planting	482
2.6	Replica symmetry breaking	482
2.7	No RSB in Bayes-optimal inference	484
2.8	Applications of planting for studies of glasses	485
2.8.1	Equilibration for free in glasses	485
2.8.2	Following states and analyzing slow annealing	485
2.8.3	Melting and glass transition	487
3	Phase diagram from the cavity method	488
3.1	Random graphs	488
3.2	BP and the cavity method	489
3.3	Properties of the randomly quenched ensemble	492
3.4	Phase diagram of the planted SG	495
3.5	Systems with first-order phase transitions	496
3.6	Comments on the phase transitions in various problems	499
4	From physics insight to new algorithmic ideas	500
4.1	Non-backtracking spectral method	500
4.2	On computational hardness of the hard phase	502
4.3	Spatial coupling	503
4.4	The spatially coupled C–W model	504
4.5	A brief history of time indices in the TAP equations	505
5	Clustering of networks and community detection	508
5.1	Biased overview of community detection	508
5.2	Asymptotically exact analysis of the SBM	509
5.3	Examples of phase transitions in the SBM	512
5.4	Spectral redemption for clustering sparse networks	514
5.5	More on the non-backtracking operator	518
5.6	The deceptiveness of the variational Bayes method	519
5.7	Real networks are not generated by the SBM	519
6	Linear estimation and compressed sensing	520
6.1	Some examples of linear estimation	520
6.2	Compressed sensing	522
6.3	Approximate message passing	523
6.3.1	Relaxed-belief propagation	523
6.3.2	From r-BP to G-AMP	525
6.3.3	The potential	528
6.3.4	Note on the convergence and parameter learning	529
6.3.5	Examples of priors and outputs	530
6.4	Replica free energy, state evolution and phase transitions	531
6.4.1	State evolution	531
6.4.2	Free energy	533
6.4.3	Phase transitions	533
6.5	Spatial coupling	535

6.6	Non-random matrices and non-separable priors	536
6.6.1	Structured priors	536
6.6.2	Orthogonal matrices	536
6.6.3	Reconstruction in discrete tomography	537
6.7	An application in physics: optics in complex media	537
6.8	Beyond linear estimation: matrix factorization	538
6.8.1	Learning random sparse dictionaries	539
6.8.2	Low-rank decomposition	540
7	Perspectives	540
	Acknowledgments	542
	Disclosure statement	542
	Funding	542
	References	542

## 1. Introduction

Our goal in this review is to describe recent developments in a rapidly evolving field at the interface between statistical inference and statistical physics. Statistical physics was developed to derive macroscopic properties of material from microscopic physical laws. Inference aims to discover structure in data. On the first look, these goals seem quite different. Yet, it was the very same man, Pierre Simon, Marquis de Laplace (1749–1827), who did one of the first sophisticated derivations of the gas laws within the caloric theory, and who also created the field of statistical inference [1]. This suggests there may be a link after all. In fact, the methods and tools of statistical physics are designed to describe large assemblies of small elements – such as atoms or molecules. These atoms can quite readily be replaced by other elementary constituents – bits, nodes, agents, neurons, data points. This simple fact, and the richness and the power of the methods invented by physicists over the last century, is at the roots of the connection, and the source of many fascinating work in the last decades (Table 1).

There is a number of books discussing this interdisciplinary connection. Without claim of exhaustivity, see for instance [2–6]. Our focus here is to present some more recent contributions that were obtained using methodology that originated in the field of spin glasses (SG) [7]. In particular the replica and cavity methods, with the related message passing algorithms, are at the roots of our toolbox. We focus on the analogy between phase transitions in thermodynamics, and the threshold phenomena that arise in the ability to infer something useful from data. The most interesting outcome of these considerations are deep consequences for the possibility of development of new algorithms, and their relations to computational feasibility. This focus is summarized in the title by the words: “threshold and algorithms”.

As we shall see, the natural setup in which the connection between inference and the statistical physics of disordered systems is particularly transparent is the so-called *teacher–student* scenario. In a nutshell, the teacher–student scenario can be described as follows: the teacher uses some ground-truth information and a probabilistic model to generate data that he then hands to the student. The student is then supposed to recover the ground truth as well as he can only from the knowledge of the data and the model. This formulation of inference problems is particularly popular in neural networks [4,8,9], where the teacher uses a rule to generate output, and the student is then trying to learn the rule to reproduce the same output. The advantage of this setting is that it is easily amenable to the Bayesian approach and strategies that recover the ground truth in an optimal way can be analyzed.

Let us make several comments: first we do not make this Bayesian choice because we think the Bayesian approach is more correct than other lines of thought common in statistics. We do it

Table 1. Glossary of abbreviations with the number of the section where the term is defined.

Abbreviation	Definition section	Meaning
MMO	1.2.4	Maximum mean overlap
MMSE	1.2.4	Minimum mean squared error
MAP	1.2.4	Maximum a posteriori estimator
MCMC	1.3.3	Monte Carlo Markov chain
BP	1.3.3	Belief propagation
RS	2.6	Replica symmetry
RSB	2.6	Replica symmetry breaking
1RSB	2.6	One-step replica symmetry breaking
d1RSB	2.6	Dynamical one-step replica symmetry breaking
FRSB	2.6	Full-step replica symmetry breaking
XOR-SAT	2.8.2	XOR-satisfiability problem
ER	3.1	Erdős–Rényi
V–B	3.3	Viana–Bray model
LDPC	3.6	Low-density parity check codes
CSP	3.6	Constraint satisfaction problem
K-SAT	3.6	K-satisfiability problem
PCA	4.1	Principal component analysis
C–W	4.3	Curie–Weiss model
TAP	4.5	Thouless–Anderson–Palmer
SBM	5.1	Stochastic block model
AWGN	6	Additive white Gaussian noise
CDMA	6.1	Code-division multiple access
LASSO	6.1	Least absolute shrinkage and selection operator
AMP	6.3	Approximate message passing
G-AMP	6.3	Generalized approximate message passing
r-BP	6.3.1	Relaxed-belief propagation
RBM	6.6	Restricted Boltzmann machine
CT	6.6.3	Computed tomography

because it is in this setting that the connection between statistical inference and statistical physics becomes the most transparent. Our main point is to explain and illustrate the insight that can be obtained from this connection in terms of phase transitions. In a sense the strategy followed in this review is closer to the approach in information theory where we know the properties of the noise that corrupted the signal than to traditional approach in statistics where we aim to assume as little as possible.

Second, we focus on the teacher–student scenario and consequently we *do not* analyze problems where one is given data or observations and has absolutely no idea of the underlying system that generated them. Nor do we discuss model selection where one needs to judge from data which model is more suitable. There is a range of textbooks on statistics where such problems are studied [10–12]. In theoretical physics it is common to study abstract models that display interesting behavior and use gained knowledge to generalize about what is happening in the real world.

Third, while the subject is certainly algorithmic in nature, we shall concern ourselves primarily, as the title shows, with phase transitions and threshold phenomena, where the connection with statistical physics is clear. This means that while we certainly are interested in analyzing and discussing algorithms, and ask if it is possible for them to reach the phase transition boundaries, we do *not* review, for any given problem, the vast amount of literature used to actually solve it in practice. Our goal, instead, is (in a typical statistical physics attitude) to reach some “universal” understanding of the problem.

The typical reader we have in mind has some knowledge of statistical physics, knows, for example, about the Ising model, phase transition and free energy. We have in mind a reader who

is curious to know how methods and concepts from statistical physics such as phase transitions translate into the context of machine learning and inference. Perhaps the most important message we want to pass in this review is that many calculations that physics uses to understand the world can be translated into algorithms that machine learning uses to understand data. We also have in mind readers from statistics and computer science who might want to learn more about the statistical physics approach. While we do explain some basic physics concepts, such readers could further read a short introduction to statistical physics, as for instance Appendix B (“some physics”) in the book of David McKay [12], or Chapter II in [5].

### 1.1. *Organization of the manuscript*

The present review is organized as follows.

In Section 1.2 we introduce the basic concepts of statistical inference and the terminology. Most of this section is standard knowledge in statistics and the main purpose of this section is to set the notations and present the concepts to readers from physics who are not particularly familiar with probability and statistics. The only part of Section 1.2 that is perhaps not standard, but certainly crucial for the present discussion, is the idealized teacher–student setting of Section 1.2.2. Traditional statistics would typically not assume knowledge of the model that generated the data and large part of statistics is precisely concerned with determining a good model. The main focus of this review is, however, on the connections to statistical physics. In the teacher–student setting, the analysis of exact phase diagrams and phase transitions is possible and this is our primary goal.

In Section 1.3 we present some basic concepts of statistical physics that non-physics readers might not be familiar with. The emphasis is given on formulation of the teacher–student setting as a planted statistical ensemble – this *planted* ensemble is not standard in physics but we do believe it is a really fundamental concept, that has become popular in the last decades, and that introducing it greatly elucidates the deep connection between properties of high-dimensional inference problems, optimization problem and disordered systems.

Having introduced the basic concepts we then choose a bottom-up approach and present in Section 2 a prototypical inference problem that we call the planted SG. We discuss how the physics of this inference problem can be deduced using the existing knowledge from the physics of SG. The main highlight of this section is that Bayes-optimal inference can be translated into thermodynamics on the Nishimori line in the underlying planted model. The physical properties of the Nishimori line are well understood for a wide range of models, this connection hence provides a unifying picture for a large class of inference problems. Consequently, concepts discussed in Section 2 bring a lot of insight into many inference problems other than the planted SG. We conclude Section 2 with a discussion on how the concept of planting turns out to be innovative in study of glasses in Section 2.8. All the results of this section are presented in a way that should be understandable without going into any calculations. The purpose of this is twofold: readers who are not familiar with the replica and the cavity calculations can still understand the reasoning and the main results. Readers who are familiar with the calculations can appreciate the reasoning, connections and consequences without having to separate them from the technicalities.

Section 3, however, is a more analytical chapter, that should be read with a pencil so that one can redo the computations. It summarizes the basic analytical technique based on belief propagation (BP) and the cavity method that we can use to analyze the properties and phase diagrams of high-dimensional inference problems. We do not go into many details and the related methods are presented in greater depths in other existing literature and textbooks, e.g. [5]. Readers familiar with these techniques can skip this section, except perhaps for its last part, Section 3.6, where we present the zoology of the main phase transitions that are encountered in inference problems.

Armed with the statistical physics' understanding of inference from Sections 2,3, we discuss related recent algorithmic ideas in Section 4. The various algorithmic ideas are discussed again on the example of the planted SG model or its simple variants. This section can be read separately from the previous ones by readers interested mainly in the algorithmic consequences stemming from physics.

The planted SG problem is a simple example. To illustrate how to generalize from it to other, more practically relevant, inference problems we present in Section 5 the problem of clustering of networks with its recent development, and in Section 6 the generalized linear estimation problem with the compressed sensing as its special case. The two examples of applications in Sections 5 and 6 are chosen with respect to author's own work. Each of these last two sections can be seen as a concise introduction into these practically relevant problems and is therefore mainly for readers that are interested in various applications of physics concepts or who want to specifically start learning about clustering or compressed sensing.

Finally, a word of caution about the subjects that we have chosen to cover (and more importantly *not to cover*). Historically, one of the most prominent example of statistical inference studied in physics are the error correcting codes. Their studies in physics were pioneered by Sourlas [13] and the results are reviewed nicely, for example, in [3,5,14]. We do not discuss error correction in this review, except for a couple of interesting and/or historical connections.

Another very prominent problem that can be formulated as statistical inference and that has been extensively studied by tools of statistical physics are neural networks. The book [3] covers very nicely the early results. A very detailed presentation, with an emphasis on teacher–student protocols, can be found in the book [4]. The reviews [8,9] are also very good references. While we do not focus on this application that is rather well documented, we will see in Section 6 that the problem of perceptron is a special case of the generalized linear estimation setting.

The two applications, network clustering and compressed sensing, that we chose to present in this review have been studied only more recently (in the last decade) and there are interesting results stemming from the connection between inference and physics that are not (yet) well covered in other reviews.

## 1.2. What is statistical inference?

Inference is the act or process of deriving logical conclusions from premises known or assumed to be true. Statistical inference is the process of deducing properties of an underlying distribution by analysis of data. A way to think about inference problems, that is particularly relevant in the context of today's data deluge (often referred to as *big data*), are procedures to extract useful information from large amounts of data. As the boundaries between fields are breaking, nowadays inference problems appear literally everywhere, in machine learning, artificial intelligence, signal processing, quantitative biology, medicine, neuroscience and many others.

A more formal way to set an inference problem is the following: one considers a set of variables  $\mathbf{x} = \{x_i\}_{i=1,\dots,N}$  on which one is able to perform some partial observations or measurements  $\mathbf{y} = \{y_\mu\}_{\mu=1,\dots,M}$ . The goal is to deduce, or infer, the values of the variables  $\mathbf{x}$  based on the perhaps indirect and noisy information contained in the observed data  $\mathbf{y}$ . To give a couple of examples, the variables may correspond for instance to image pixels, an acoustic signal or to an assignment of data points into clusters, while the corresponding observations would correspond for instance to a blurred photograph, a recording of a song performed on a platform while the train was arriving or positions of points in space.

For inference problems the two main scientific questions of interest are:

- The question of sufficient information – Under what conditions is the information contained in the observations sufficient for satisfactory recovery of the variables?



- The question of computational efficiency – Can the inference be done in an algorithmically efficient way? What are the optimal algorithms for this task?

These two questions are of very different nature. The first one is a question that belongs to the realm of statistics and information theory. The second one, instead, is in the realm of computer science and computational complexity. Before the era of big data, statistics was often not concerned with computational tractability, because the number of parameters to be estimated was usually small. Nowadays the trade-off between statistical and computational efficiency has changed, making it most reasonable to study the above two questions together.

We will see that these two questions have a very clear interpretation in terms of phase transitions in physics. More specifically, the first one is connected to a genuine phase transition while the second one is connected to the type of the transition (first or second order), and metastability associated with first-order phase transitions is related to algorithmic hardness.

### 1.2.1. Terminology and simple examples

In this review, we will follow the approach of Bayesian inference. Developed by Bayes and Laplace, it concentrates on conditional probabilities. The probability of having an event  $A$  conditioned on the event  $B$  is denoted  $P(A | B)$  and defined as

$$P(A | B) = \frac{P(A, B)}{P(B)}, \quad (1)$$

where  $P(A, B)$  is the joint probability of both events  $A$  and  $B$ , and  $P(A) = \sum_B P(A, B)$  is the probability of event  $A$  (the sum is taken over all the possible realizations of event  $B$ ). When events  $A, B$  are taken from a continuous domain the  $P(A)$  denotes probability distribution function, whereas  $P(A)dA$  denotes the probability of observing an outcome in the interval  $(A, A + dA)$ . For discrete variables the probability of one event and the probability distribution are interchangeable. Note that in this review we discuss both models with discrete variables (Section 2 or 5) and with continuous variables (Section 6), the concepts that we build hold for both these cases.

The central point of the approach is the so-called Bayes formula that follows directly from the definition of the conditional probability (1)

$$P(\mathbf{x} | \mathbf{y}) = \frac{P(\mathbf{y} | \mathbf{x})}{P(\mathbf{y})} P(\mathbf{x}). \quad (2)$$

There is a precise terminology for all the terms entering this equation, so we shall start by discussing the Bayesian inference vernacular:

- $P(\mathbf{x} | \mathbf{y})$  is called the *posterior* probability distribution for  $\mathbf{x}$  given all the data  $\mathbf{y}$ .
- $P(\mathbf{y} | \mathbf{x})$  is called the *likelihood*.
- $P(\mathbf{y})$ , also denoted as a normalization  $Z(\mathbf{y})$ , is called the *evidence*.
- $P(\mathbf{x})$  is called the *prior* probability distribution for  $\mathbf{x}$ . It reflects our belief of what  $\mathbf{x}$  should be before we received any data.

*Example of decaying particles.* Our first example is borrowed from Ref. [12]. Unstable particles are emitted from a source and decay at a distance  $y$  from the source. Unfortunately, we cannot build a perfect infinite length detector, and decay events can only be observed if they occur in a window extending from distances  $y_{\min} = 1$  cm to  $y_{\max} = 20$  cm from the source.

We run the experiment for some time, and we observe  $M$  decays at locations  $\mathbf{y} = \{y_1, \dots, y_M\}$ . We also know that we should expect an exponential probability distribution with characteristic



length  $x$  for the event. Writing things in the inference language,  $\mathbf{y}$  are the observations and  $x$  is the unknown signal (of dimension  $N = 1$ ). The question we would like to answer is: Given the observations, how to estimate  $x$ ?

In this particular example physics tells us exactly what is the probability distribution of a particle decaying at distance  $y$  given the characteristic length  $x$ . In terms of Bayesian probability this means that we know exactly the model that created the observations. This is of course not true in general, but when it is true we can set up the so-called Bayes-optimal inference: the likelihood for one observation is

$$P(y|x) = \begin{cases} \frac{1}{x\mathcal{M}(x)} e^{-y/x} & \text{if } 1 < y < 20, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

with a normalization  $\mathcal{M}(x) = \int_1^{20} dz (1/x) e^{-z/x} = e^{-1/x} - e^{-20/x}$ . The observations are independent, and therefore in this case their joint probability distribution is simply a product of the single observation distributions. From Equation (3) we can now extract  $P(x|\mathbf{y})$  using the Bayes' formula and get

$$P(x|\mathbf{y}) = \frac{P(x)}{Z(\mathbf{y})[x\mathcal{M}(x)]^M} e^{-\sum_{i=1}^M y_i/x}. \quad (4)$$

Looking at this equation, we see that we have  $P(x)$  on the right hand side. If we do not have any additional information on  $x$ , we assume that the prior distribution  $P(x)$  is a constant that just enters in the  $x$ -independent normalization, denoting then  $\tilde{Z}(\mathbf{y}) = Z(\mathbf{y})/P(x)$ , we obtain

$$P(x|\mathbf{y}) = \frac{1}{\tilde{Z}(\mathbf{y})[x\mathcal{M}(x)]^M} e^{-\sum_{i=1}^M y_i/x}. \quad (5)$$

This is the final answer from Bayesian statistics. It contains all the information that we have on  $x$  in this approach. For a dataset consisting of several points, for instance the six points  $y = 1.5, 2, 3, 4, 5, 12$  cm, one can compute the most probable value of  $x$  to be around  $\hat{x} = 3.7$  cm. Note that this is considerably different from the mere average of the points which is  $\bar{y} = 4.6$  cm. The reader can convince herself (for instance performing simulation and comparing to the true correct value) that the presented approach gives a way better result than the naive average.

*Example of linear regression.* Most physicists are used to fit a theoretical model to experimental data. Consider for instance that we measure the height of an object falling due to Newton's gravity, we assume no initial speed and no friction. We expect that the height  $h(t)$  of the object at time  $t$  will follow  $h(t, g, h_0) = h_0 - gt^2/2$ , where  $g$  is the acceleration of gravity. We can do an experiment and measure heights of the object  $y_i$  for many values of  $t_i$ , with  $i = 1, \dots, M$ . We expect, if we plot  $y_i$  against  $t_i^2/2$ , to get a straight line that will give us the value of the acceleration  $g$ .

Due to small but real measurement errors, the measured points will not lie exactly on a line. Used to such a situation, a physicist would in this case perform a linear regression to minimize the so-called sum of squares of errors. The "best" value of  $g$ , or to use the statistics language the estimator  $\hat{g}$  of  $g$ , would simply be the one that minimizes the cost function  $\sum_{i=1}^M [h(t_i, g, h_0) - y_i]^2$ .

In order to get more familiarity with the concepts of inference we discuss how this familiar procedure can be cast into the Bayesian language. In general we may not know much about the nature of the measurement noise. In such cases it is often reasonable to expect that the noise is a Gaussian random variable of zero mean and variance  $\Delta$ . Under this assumption the likelihood  $P(y_i|g, \Delta)$  of a single data point is a Gaussian with mean  $h(t_i, g, h_0)$  and variance  $\Delta$ . We assume

that the noise for different measurements is independent which leads to the total likelihood

$$P(\mathbf{y} | g, h_0, \Delta) = \frac{1}{(2\pi\Delta)^{M/2}} \prod_{i=1}^M e^{-[h(t_i, g, h_0) - y_i]^2 / 2\Delta}. \quad (6)$$

As in the previous example we assume a uniform prior distribution on the acceleration constant  $P(g)$ , initial height  $P(h_0)$  and the variance  $P(\Delta)$ , the posterior probability distribution is then given by

$$P(\Delta, g, h_0 | \mathbf{y}) = \frac{1}{Z(\mathbf{y})(2\pi\Delta)^{M/2}} \prod_{i=1}^M e^{-[h(t_i, g, h_0) - y_i]^2 / 2\Delta}. \quad (7)$$

Looking for the most probable value of  $g$  indeed amounts to minimizing the sum of squares  $\sum_{i=1}^M [h(t_i, g, h_0) - y_i]^2$ , which is the standard solution. Note, however, that the assumption of Gaussian noise independent for every measurement might not be justified in general. Also, depending on the precise setting, there might be better prior distributions for the parameters  $g$ ,  $h_0$ ,  $\Delta$ . For more advanced discussion of classical linear regression, see for instance [15]. In Section 6 of this review we will discuss a high-dimensional version of linear regression, when the number of parameters to be estimated is comparable to the number of observations  $M$ .

Let us make some important comments on the above considerations, which have been at the roots of the Bayesian approach to inference from the very start. First, we notice that probabilities are used here to quantify degrees of belief. To avoid possible confusions, it must thus be emphasized that the true  $\mathbf{x}$  in the real world *is not* a random variable, and the fact that Bayesians use a probability distribution  $P(\mathbf{x} | \mathbf{y})$  does not mean that they 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  $x$ ), of which only one is actually true. The fact that probabilities can denote degrees of belief is at the heart of Bayesian inference. The reader interested in these considerations is referred to the very influential work of De Finetti [16].

Another subtle point is how to choose the prior distribution  $P(\mathbf{x})$ . In fact, discussions around this point have been rather controversial in the beginning of the history of statistical inference, and even now there are many schools of thought. It is not the goal of this review to present these discussions and debates. In the spirit of a reductionism that is usually done in physics, we introduce in the next section the so-called teacher–student scenario in which we know what the model and the prior distribution really are, so that the use of the Bayes formula is straightforward and fully justified. Majority of this review is then set within this teacher–student framework. This can be seen as a simplified model of the generic situation. As we will see the picture of what is happening is interestingly rich already in this framework. A first step toward the generic situation of unknown models that generated the data and unknown prior probabilities is the case when the teacher hides some parameters of the model or/and of the prior distribution from the student. Analyzing this case provides insight into the generic case, but still within the realm of a solvable simplified teacher–student framework.

In Section 1.3.2 we then argue that the teacher–student scenario for a generic inference problem can be seen as a so-called *planted ensemble* of the corresponding statistical physics model. This is a known but not widely used way of presenting inference in statistical physics, but the authors feel that it is an extremely useful and insightful way. It allows to extend the intuition gathered over several decades in studies of disordered systems such as glasses and SG into many currently studied inference problems that otherwise appear highly non-trivial. One of the goals of the present review is to bring this insight forward and establish the way of thinking about inference problems as planted statistical ensembles.

### 1.2.2. Teacher–student scenario

The *teacher–student scenario* is defined as follows:

- Teacher: In a first step the teacher generates a realization of variables/parameters  $\mathbf{x}^*$  from some probability distribution  $P_{\text{tp}}(\mathbf{x})$ , where “tp” stands for teacher’s prior. In a second step he uses these *ground-truth* values  $\mathbf{x}^*$  to generate the data  $\mathbf{y}$  from some statistical model characterized by a likelihood of  $\mathbf{y}$  given  $\mathbf{x}^*$ , denoted as  $P_{\text{tm}}(\mathbf{y} | \mathbf{x}^*)$ , where “tm” stands for teacher’s model. Finally, the teacher hands the data  $\mathbf{y}$  to the student together with some information about the distributions  $P_{\text{tp}}(\mathbf{x})$  and  $P_{\text{tm}}(\mathbf{y} | \mathbf{x})$ .
- Student: The goal of the student is to infer as precisely as possible (and tractably, when computational complexity is included into considerations) the original values  $\mathbf{x}^*$  of the variables from the information provided by the teacher, i.e. from the data  $\mathbf{y}$  and the available information about the distributions  $P_{\text{tp}}(\mathbf{x})$  and  $P_{\text{tm}}(\mathbf{y} | \mathbf{x})$ .

The variables with a  $*$  (e.g.  $\mathbf{x}^*$ ) denote the ground truth that was used by the teacher, but that is not available to the student. In the present manuscript we will consider cases where both the  $\mathbf{x}$  and  $\mathbf{y}$  are high-dimensional vectors, and where the teacher’s prior  $P_{\text{tp}}(\mathbf{x})$  and model  $P_{\text{tm}}(\mathbf{y} | \mathbf{x})$  are parameterized by some scalar (or low-dimensional vector) values  $\theta^*$  and  $\Delta^*$ , respectively. This is a case relevant to many real-world problems where we assume existence of many latent variables (collected into the vector  $\mathbf{x}$ ), and we assume that the model and prior distribution is parametrized by a small number of unknown parameters.

We will in particular be interested in inference problems where the teacher’s prior distribution is separable, i.e. can be written as a product over components

$$P_{\text{tp}}(\mathbf{x}) = \prod_{i=1}^N P_{\text{tp}}(x_i). \quad (8)$$

Further, in this review, we are interested in cases where also the observations  $\{y_\mu\}_{\mu=1,\dots,M}$  are independent and each of them depends on a subset of variables denoted  $\partial\mu$ . The likelihood of the teacher’s model is then written as

$$P_{\text{tm}}(\mathbf{y} | \mathbf{x}) = \prod_{\mu=1}^M P_{\text{tm}}(y_\mu | \{x_i\}_{i \in \partial\mu}). \quad (9)$$

Given the information available to the student, she then follows the strategy of Bayesian statistics. The prior information she has from the teacher is denoted as  $P(\mathbf{x})$ , where  $\mathbf{x}$  are the variables to be inferred. The statistical model is represented by its likelihood  $P(\mathbf{y} | \mathbf{x})$ . Again we assume that both these can be written in a product form as above. She then considers Bayes’ theorem (2) and writes the posterior probability which contains all the information available to her

$$P(\mathbf{x} | \mathbf{y}) = \frac{1}{Z(\mathbf{y})} \prod_{i=1}^N P(x_i) \prod_{\mu=1}^M P(y_\mu | \{x_i\}_{i \in \partial\mu}). \quad (10)$$

A very useful quantity to define that derives directly from the posterior distribution is the marginal probability of one variable

$$\mu_i(x_i) = \int P(\mathbf{x} | \mathbf{y}) \prod_{j \neq i} dx_j. \quad (11)$$

Examples:

- Let us set the problem of simple linear regression that we used as an example in Section 1.2.1 in the teacher–student setting. Consider the teacher generated a number  $g^*$  from some probability distribution  $P_{\text{tp}}(g)$ , this  $g^*$  is the ground-truth value we aim to find back as precisely as possible from the noisy measurements. Then the teacher generated heights  $y_i$  independently from a Gaussian probability distribution of mean  $-g^*t_i^2/2$  and variance  $\Delta^*$ , where  $t_i$  are the known measurement times,  $i = 1, \dots, N$ . He then handed the heights  $\mathbf{y}$  and the distribution  $P_{\text{tp}}(g), P_{\text{tm}}(y_i | g)$  to the student. Set in this way the Bayesian solution presented in Section 1.2.1 has no ambiguity in terms of assuming the noise was Gaussian and independent, nor in terms of the choice of the prior distribution for  $g$ .
- Historically the teacher–student setting was introduced for the problem of perceptron, called model B in [17]. In perceptron the aim is to learn *weights*  $J_i$ ,  $i = 1, \dots, N$ , in such a way that the scalar product with each of  $P$  patterns  $\xi_i^\mu$ ,  $\mu = 1, \dots, P$ , is constraint to be either smaller ( $y_\mu = -1$ ) or larger ( $y_\mu = 1$ ) than some threshold  $\kappa$ . Mathematically stated one requires for all  $\mu = 1, \dots, M$

$$y_\mu = \text{sign} \left( \sum_{i=1}^N J_i \xi_i^\mu - \kappa \right). \quad (12)$$

This perceptron setting models classification of patterns into two classes. One way to set a solvable version of the model is to consider both  $\xi_i^\mu$  and  $y_\mu$  to be independent identically distributed (iid) random variables. The goal is then to find a set of weights  $J_i$  such that the constraints (12) are satisfied. Another way to set the problem that corresponds to the teacher–student scenario is to consider some teacher weights  $J_i^*$ , generated from some probability distribution, e.g.  $\pm 1$  uniformly at random. The teacher perceptron then generates the output as  $y_\mu = \text{sign}(\sum_{i=1}^N J_i^* \xi_i^\mu - \kappa)$ , where  $\xi_i^\mu$  are some known iid patterns. The goal is then to recover the teacher weights  $J_i^*$  from the knowledge of  $\xi_i^\mu$  and  $y_\mu$ . A very comprehensive presentation and solution of the perceptron that uses this teacher–student terminology appears in [3].

### 1.2.3. Bayes-optimal versus mismatched

Throughout this manuscript we will be distinguishing two main cases depending on what information about the distributions  $P_{\text{tp}}(\mathbf{x})$  and  $P_{\text{tm}}(\mathbf{y} | \mathbf{x})$  the teacher gave to the student

- Bayes-optimal case: In this case the teacher hands to the student the full and correct form of both the prior distribution  $P_{\text{tp}}(\mathbf{x})$  and the likelihood  $P_{\text{tm}}(\mathbf{y} | \mathbf{x})$ .
- Mismatching prior and/or model: In this case the teacher hands no or only partial information about the prior distribution  $P_{\text{tp}}(\mathbf{x})$  and/or the statistical model  $P_{\text{tm}}(\mathbf{y} | \mathbf{x})$ . In the present manuscript, we will assume that the teacher handed the correct functional form of the prior distribution and of the statistical model, but not the values of the parameters  $\theta$  and  $\Delta$ , respectively.

Let us now consider the following exercise. Take  $\mathbf{x}^*$  to be the ground-truth values of the variables generated by the teacher and take  $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$  to be three independent samples from the posterior probability distribution  $P(\mathbf{x} | \mathbf{y})$ . We then consider some function  $f(\mathbf{a}, \mathbf{b})$  of two

configurations of the variables  $\mathbf{a}, \mathbf{b}$ . Consider the following two expectations

$$\mathbb{E}[f(\mathbf{x}_1, \mathbf{x}_2)] = \int f(\mathbf{x}_1, \mathbf{x}_2) P(\mathbf{y}) P(\mathbf{x}_1 | \mathbf{y}) P(\mathbf{x}_2 | \mathbf{y}) d\mathbf{x}_1 d\mathbf{x}_2 d\mathbf{y}, \quad (13)$$

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}^*, \mathbf{x})] &= \int f(\mathbf{x}^*, \mathbf{x}) P(\mathbf{x}^*, \mathbf{x}) d\mathbf{x} d\mathbf{x}^* = \int f(\mathbf{x}^*, \mathbf{x}) P(\mathbf{x}^*, \mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{x}^* d\mathbf{y} \\ &= \int f(\mathbf{x}^*, \mathbf{x}) P(\mathbf{x} | \mathbf{y}, \mathbf{x}^*) P_{\text{tm}}(\mathbf{y} | \mathbf{x}^*) P_{\text{tp}}(\mathbf{x}^*) d\mathbf{x} d\mathbf{x}^* d\mathbf{y}, \end{aligned} \quad (14)$$

where we used the Bayes formula. We further observe that  $P(\mathbf{x} | \mathbf{y}, \mathbf{x}^*) = P(\mathbf{x} | \mathbf{y})$  because  $\mathbf{x}$  is independent of  $\mathbf{x}^*$  when conditioned on  $\mathbf{y}$ . Remarkably, in the Bayes-optimal case, i.e. when  $P(\mathbf{x}) = P_{\text{tp}}(\mathbf{x})$  and  $P(\mathbf{y} | \mathbf{x}) = P_{\text{tm}}(\mathbf{y} | \mathbf{x})$ , we then obtain

$$\text{Bayes optimal : } \mathbb{E}[f(\mathbf{x}_1, \mathbf{x}_2)] = \mathbb{E}[f(\mathbf{x}^*, \mathbf{x})], \quad (15)$$

meaning that under expectations there is no statistical difference between the ground-truth assignment of variables  $\mathbf{x}^*$  and an assignment sampled uniformly at random from the posterior probability distribution. This is a simple yet immensely important property that will lead to numerous simplifications in the Bayes-optimal case and it will be used in several places of this manuscript, mostly under the name *Nishimori condition*.

In the case of mismatching prior or model the equality (15) typically does not hold.

#### 1.2.4. Estimators

What is the optimal estimator  $\hat{\mathbf{x}}$  for the variables  $\mathbf{x}$ ? The answer naturally depends on what the quantity we aim to optimize is. The most commonly considered estimators are:

*Maximum a posteriori (MAP) estimator.* MAP simply maximizes the posterior distribution and is given by

$$\hat{\mathbf{x}}^{\text{MAP}} = \underset{\mathbf{x}}{\operatorname{argmax}} P(\mathbf{x} | \mathbf{y}). \quad (16)$$

The main disadvantage of the MAP estimator is the lack of confidence interval and related overfitting issues.

*Minimum mean squared error (MMSE).* Ideally we would like to minimize the squared error between  $\hat{\mathbf{x}}$  and  $\mathbf{x}^*$

$$\text{SE}(\hat{\mathbf{x}}, \mathbf{x}^*) = \frac{1}{N} \sum_{i=1}^N (\hat{x}_i - x_i^*)^2. \quad (17)$$

In general, however, we do not know the ground-truth values of the variables  $\mathbf{x}^*$ . The best we can do within Bayesian statistics is to assume that  $\mathbf{x}^*$  is distributed according to the posterior probability distribution. We then want to find an estimator  $\hat{\mathbf{x}}$  that minimizes the squared error averaged over the posterior

$$\text{MSE}(\hat{\mathbf{x}}) = \frac{1}{N} \int P(\mathbf{x} | \mathbf{y}) \sum_{i=1}^N (\hat{x}_i - x_i)^2 d\mathbf{x}. \quad (18)$$

By a simple derivative with respect to  $\hat{x}_i$  we see that the minimum of this quadratic function is achieved when

$$\hat{x}_i^{\text{MMSE}} = \int P(\mathbf{x} | \mathbf{y}) x_i d\mathbf{x} = \int \mu_i(x_i) x_i dx_i, \quad (19)$$

where the right hand side is actually the mean of the marginal on the  $i$ th variable defined by Equation (11).

The value of the MMSE is then computed as the MSE (18) evaluated at  $\hat{x}_i$  (19).

*Maximum mean overlap (MMO).* In cases where the support of the variables is discrete and the variables represent indices of types rather than continuous values, the mean-squared error is not very meaningful. In that case we would rather count how many of the  $N$  positions obtained the correct type. We define the overlap  $O$  as the fraction of variables for which the estimator  $\hat{\mathbf{x}}$  agrees with the ground truth

$$O(\hat{\mathbf{x}}, \mathbf{x}^*) = \frac{1}{N} \sum_i \delta_{x_i^*, \hat{x}_i}. \quad (20)$$

As in the case of MMSE, when we do not have the knowledge of the ground truth, the best Bayesian estimate of the overlap is its average over the posterior distribution

$$MO(\hat{\mathbf{x}}) = \frac{1}{N} \int P(\mathbf{x} | \mathbf{y}) \sum_i \delta_{x_i, \hat{x}_i} d\mathbf{x}. \quad (21)$$

The mean overlap MO maximized over the estimator  $\hat{\mathbf{x}}$  leads to the MMO estimator

$$\hat{x}_i^{\text{MMO}} = \underset{x_i}{\operatorname{argmax}} \mu_i(x_i), \quad (22)$$

where  $\mu(x_i)$  is the marginal probability of variable  $i$  having type  $x_i$ , defined by Equation (11). We should note that the MMO is not an estimator very commonly considered in statistics, but given that statistical physics often deals with discrete variables, and anticipating the close relation, the MMO will turn out instrumental.

Let us compare the MAP estimator with the ones based on marginals. In most cases the MAP estimator is easier to approach algorithmically (from the computational complexity point of view optimization is generically easier than enumeration). Moreover, in many traditional settings, where the number of samples is much higher than the number of variables, there is usually very little difference between the two since the marginal probabilities are basically peaked around the value corresponding to the MAP estimator. For instance, let us consider the teacher–student scenario with unknown parameters  $\theta$  of the teacher’s prior, or  $\Delta$  of the teacher’s model. In situations where the number of samples  $M$  is large, but the parameters  $\theta$  and  $\Delta$  are scalar (or low-dimensional), the posterior probability of these parameters is so closely concentrated around its expectation that there is no practical difference between the MAP and the marginals-based estimators for  $\theta$  and  $\Delta$ . When it comes to estimating the variables  $\mathbf{x}$  then in the setting of high-dimensional statistics (see Section 1.2.5), where the number of samples  $M$  is comparable to the number of variables  $N$  the MAP estimator for  $\mathbf{x}$  will generically provide crude overfitting and we will prefer estimators based on marginals. Another reason to favor marginalization-based estimators is because they come directly with a notion of statistical significance of the estimate. Yet another reason is that from the perspective of SG theory there might even be computational advantages (ground states might be glassy, whereas Bayes-optimal computations are not, see Section 2.7).

Let us now investigate the MMSE (and the MMO) estimator in the Bayes-optimal case in the view of equality (15). In Section 1.2.3 we concluded that in the Bayes optimal case (i.e. when the teacher handed to the student the precise form of the prior distribution and of the model likelihood) and in expectation, a random sample  $\mathbf{x}$  from the posterior distribution can be replaced by the ground-truth assignment  $\mathbf{x}^*$  without changing the values of the quantity under

consideration. It hence follows that

$$\text{Bayes optimal : } \text{MMSE} = \mathbb{E}[\text{SE}(\hat{\mathbf{x}}, \mathbf{x}^*)]. \quad (23)$$

In words, the MMSE computed solely using the posterior distribution is on average equal to the squared error between the MMSE estimator and the ground-truth value of the variables  $\mathbf{x}^*$ . This is a very nice property. An analogous identity holds between the MMO and the average of the overlap between the MMO estimator and the ground-truth assignment

$$\text{Bayes optimal : } \text{MMO} = \mathbb{E}[\text{O}(\hat{\mathbf{x}}, \mathbf{x}^*)]. \quad (24)$$

### 1.2.5. The high-dimensional limit

Let us denote by  $N$  the number of variables  $\mathbf{x} = \{x_i\}_{i=1,\dots,N}$  (referred to as dimensionality in statistics), and by  $M$  the number of observations (or samples)  $\mathbf{y} = \{y_i\}_{i=1,\dots,M}$ . Traditionally, statistical theory would consider the case of a finite number  $N$  of variables/parameters to estimate and a diverging number of samples  $M$ . Think for instance of the two examples (decaying particles and linear regression) from Section 1.2.1.

In today's data deluge more than ever, it is crucial to extract as much information as possible from available data. From the information theoretic perspective, we would expect that useful information is extractable even when the number of variables (dimensionality)  $N$  is comparably large (or even somewhat larger) than the number of observations (samples)  $M$ . In that case, separating the useful information about  $\mathbf{x}$  from noise is much more challenging. This is referred to as the *high-dimensional statistical theory*. It is, in particular, in this high-dimensional framework, where the amount of data to analyze is huge, that we need to pay attention not only to solvability but also to computational efficiency, keeping in mind both questions from Section 1.2.

Typically, we will consider the limit where  $\alpha = M/N$  is fixed, whereas  $N \rightarrow \infty$ . In this limit the following scenario, illustrated schematically in Figure 1, very often applies: for low values of  $\alpha < \alpha_c$ , successful inference of the variables is not possible for any algorithm, the corresponding observed information is simply insufficient. For high values of  $\alpha > \alpha_s$ , computationally efficient inference algorithms do exist. In an intermediate regime  $\alpha_c < \alpha < \alpha_s$ , successful inference is in principle possible but algorithmically considerably harder than above  $\alpha_s$ . In many settings, the values  $\alpha_c$  and  $\alpha_s$  can be defined in a precise mathematical way as values at which some measure of performance (or its first derivative) presents a discontinuity in the limit of large size. In such a case, we talk about thresholds or phase transitions (see Section 1.3.4 for a more precise meaning of the second). In many problems the intermediate hard phase is missing and  $\alpha_c = \alpha_s$ . On the other hand, settings and examples where the hard phase exists ( $\alpha_c < \alpha_s$ ) are theoretically very interesting and challenging many researchers to prove some formal results about this hardness. For a general inference problem, the values of  $\alpha_c$  and  $\alpha_s$  are not known, and neither are efficient

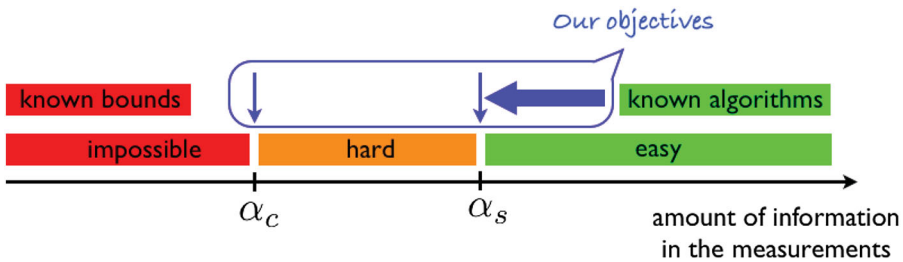


Figure 1. Schematic representation of a typical high-dimensional inference problem.



algorithms that would work down to  $\alpha_s$ . The state-of-the-art results usually only give a lower bound on  $\alpha_c$ , and an algorithmic upper bound on  $\alpha_s$ .

The objective of a large part of the statistical physicists interested in these directions and included in this manuscript could be summarized as follows (see also the schema Figure 1): determine the threshold  $\alpha_c$  between the *inference-impossible* and *inference-possible* region, and the threshold  $\alpha_s$  between the *inference-hard* and *inference-easy* region. And develop efficient algorithms that would be able to infer the variables successfully with as little information contained in the observations as corresponds to  $\alpha_s$ .

### 1.3. Statistical physics and inference

The connection between statistical physics and statistical inference has a long history. Arguably, the most influential early works are due to E.T. Jaynes, starting with his *Statistical mechanics and information theory* in 1957 [18] where he develops the maximum entropy principle, thus building upon the pioneering work of Shannon [19], who named the amount of information as entropy inspired by mathematically analogous formulas in thermodynamics. Jaynes actually suggests that statistical mechanics does not have to be regarded as a physical theory dependent on its assumptions, but simply as a special case of a more general theory of Bayesian inference. An inspiring reference is a collection of lecture notes *Statistical Physics to Statistical Inference and Back* from a meeting held in Cargèse in 1992 [2]. The goal of that meeting was precisely to revise the foundational connections between the two fields. This close connection leads to a continued cross-fertilization between the two fields. Important methodological developments in one field find numerous applications in the other, and a new applicational area in one field profits greatly from the methodology more common to the other.

The present manuscript is mainly concerned with recent applications of methods from SG and glasses [7] to inference problems and associated algorithms. This direction was perhaps most influenced by the work on simulated annealing by Kirkpatrick et al. [20]. Connections between SG and error correcting codes [13], or neural networks such as the perceptron [8,9], belong among the early very influential works. This line of research is relatively well explored and illustrated on a number of problems as seen from classical books such as the one by Nishimori [3], or more recently by Mézard and Montanari [5].

Data processing is, however, a very fast developing field with many emerging yet far-reaching applications. On top of that, there was an important development in the theory of SG on sparse random graphs in the early 2000s [21,22]. This manuscript aims at exposing some of this recent development.

#### 1.3.1. Basic dictionary

Let us first establish a dictionary between statistical inference and statistical physics. The posterior probability (10) can be written as

$$P(\mathbf{x} | \mathbf{y}) = \frac{1}{Z(\mathbf{y})} \exp \left\{ \beta \sum_{\mu=1}^M \log P(y_\mu | \{x_i\}_{i \in \partial \mu}) + \beta \sum_{i=1}^N \log P(x_i) \right\} = \frac{1}{Z(\mathbf{y})} \exp -\beta \mathcal{H}(\mathbf{x}, \mathbf{y}). \quad (25)$$

We have used the exponential form to note that it takes the form of the Boltzmann probability distribution on variables (discrete or continuous) with  $Z$  being the partition function,  $\log P(x_i)$  the local magnetic field, and  $\log P(y_\mu | \{x_i\}_{i \in \partial \mu})$  being the interaction between variables,  $\mathcal{H}(\mathbf{x}, \mathbf{y})$  being the Hamiltonian. We also introduced the auxiliary parameter  $\beta$ , that was previously  $\beta =$

1, and plays the role of the inverse temperature. The whole exponent divided by the inverse temperature is then minus the Hamiltonian  $\mathcal{H}$  of the physical system.

Note that the distribution (25) encompasses a wide range of models, the variables  $\mathbf{x}$  can be discrete (e.g. spins) or continuous, the interactions can depend on few or many variables (e.g. in the planted SG of Section 2 the interactions are pairwise, and in the linear estimation of Section 6 each of the interactions involves all of the variables).

From a physics point of view the marginal probabilities (11) are nothing else than local magnetizations obtained by summing the Boltzmann distribution over all but one variable. The MMSE (or the MMO) estimator is then computed using these local magnetizations (at  $\beta = 1$ ). On the other hand the MAP estimator can be viewed as the ground state of the physical system, i.e. the minimum of the Hamiltonian. Another way to think about the MAP estimator is as the zero-temperature limit,  $\beta \rightarrow \infty$ , of MMSE estimator.

### 1.3.2. The teacher–student scenario defines the planted ensemble

Looking at the Boltzmann distribution (25) it could represent quite a generic statistical physics model where the observations  $\mathbf{y}$  play the role of a quenched disorder. Keeping in mind that we want to study Bayesian inference in the teacher–student scenario of Section 1.2.2 we have to emphasize that the disorder  $\mathbf{y}$  is generated in a very particular way (by the teacher) related to the form of the Boltzmann distribution itself. This becomes particularly striking in the Bayes-optimal setting in which the teacher handed to the student all the information about how he generated the data  $\mathbf{y}$  except the realization  $\mathbf{x}^*$ .

In the rest of this manuscript we call models created in the teacher–student scenario of Section 1.2.2 the *planted models*. We call a disorder  $\mathbf{y}$  that results from the teacher–student scenario the *planted disorder*. We contrast it with the most commonly considered case of randomly quenched disorder where the component  $y_\mu$  are iid variables. In the planted disorder the components  $y_\mu$  are not independent, they are correlated because they all depend on the ground-truth  $\mathbf{x}^*$  that the teacher used to generate  $\mathbf{y}$ . We call  $\mathbf{x}^*$  the *planted configuration*. This is the same difference between solving a random system of linear equations  $\mathbf{y} = \mathbf{F}\mathbf{x}$  (which will have no solution if  $\mathbf{F}$  is a  $M \times N$  matrix and  $M > N$ ) and solving one where the left hand side was generated as  $\mathbf{y} = \mathbf{F}\mathbf{x}^*$ , instead of just being random. In the second case, we have planted a solution  $\mathbf{x}^*$  in the problem.

A generic strategy to define the planted statistical ensemble is to take an inference problem in the teacher–student setting. The Bayes-optimal posterior probability distribution of that inference problem then corresponds to the Boltzmann distribution of the planted model at temperature  $T = 1/\beta = 1$ . What makes the planted model particular is the form of the quenched disorder that is not independently random but instead correlated in a very particular way.

Perturbing a random ensemble by hiding a solution in it is often a natural thing to do. For instance, protein folding is often modeled, in statistical physics, by a random energy model (to take into account its glassiness) in which one special configuration (the so-called native configuration) is artificially created with a very low energy [23]. Thinking of Bayesian inference in terms of the planted ensemble of statistical physics models is implicitly done in a number of works. For instance the pioneering articles of Sourlas on finite-temperature decoding in error correcting codes [13,24,25] and the relation to the Nishimori line, summarized excellently in [3], were very influential. In mathematics, planting was introduced by Jerrum [26] as a model to study hardness of finding a large clique planted in a graph, and also by Jerrum and Sorkin [27] as a model to study graph bisection. Planting was used in many mathematical works either as a useful model or as a proof technique. It is the use of planting in proofs for constraint satisfaction problems (CSP) [28] that brought the authors' attention to this concept. The authors find that making the

connection between planting and inference very explicitly brings a lot of insight into a generic inference problem and for this reason we build the present review around this connection.

To clarify further what is exactly the planted ensemble we give several examples and counter-examples:

- *Planted coloring*. Graph coloring is a famous mathematical problem where nodes of a graph are assigned colors with a constraint that two connected nodes should not have the same color. Coloring of random graphs is a CSP, well studied in statistical physics [29,30]. In planted coloring, the teacher first generates a random assignment of colors to nodes  $\mathbf{x}^*$ , and then he generates edges of the graph at random but in a way that two nodes of the same color are not connected. The edges of the graph play the role of the planted disorder  $\mathbf{y}$ . The graph (i.e. its set of edges) is then handed to the student, but not the planted configuration  $\mathbf{x}^*$ . The role of the student is to find a configuration as correlated to the planted one as possible. Physics study of the planted coloring appeared in [31], and we will see that it is a special case of the stochastic block model (SBM) to which we devote Section 5.
- *Planted perceptron*. In Section 1.2.2 we presented the perceptron as the problem for which the teacher–student terminology was established. Formulated as a planted problem, the set of weights  $\mathbf{J}^*$  the teacher perceptron uses is the planted configuration. The patterns  $\xi_i^\mu$  are in the physics works considered as iid, they play the role of the ordinary randomly quenched disorder. The teacher generates the outputs  $y_\mu$  using the patterns and the planted weights  $\mathbf{J}^*$  (12). The output  $\mathbf{y}$  then plays the role of the planted disorder. The planted weights  $\mathbf{J}^*$  are then hidden from the student and her goal is to recover them as precisely as possible. The associated posterior probability distribution is the Boltzmann measure of the planted perceptron. On the other hand, when asking how many random patterns a perceptron can classify (the so-called capacity [17]), one is studying a non-planted problem.
- Hopfield model [32] is a well-known model studied in statistical physics. The goal in the Hopfield model is to construct an Ising spin model for which a certain pre-defined set of randomly generated patterns  $\xi_i^\mu \in \{\pm 1\}$ ,  $i = 1, \dots, N$  and  $\mu = 1, \dots, M$ , serve as configurations attracting local dynamics. It is, in fact, a model of *associative memory* where, when starting close to one of the patterns, the dynamics would converge back to it. The Ising model studied by Hopfield has Hebb-like interactions  $J_{ij} = \sum_{\mu=1}^M \xi_i^\mu \xi_j^\mu$ . We could study the inference problem for recovering the patterns  $\xi_i^\mu$  from the knowledge on the  $J_{ij}$ . For a single pattern,  $M = 1$ , the posterior distribution of this inference problem would indeed correspond to the Hopfield model, i.e. Ising spin model with interactions  $J_{ij}$ . However, for more than one pattern  $M > 1$  the posterior of the inference problem will be a function of  $NM$  variables, and will be very different from the Boltzmann distribution of the Hopfield model (which is just a function of only  $N$  spins). This shows that inferring or remembering are very different tasks. Clearly, the associative memory studied in the Hopfield model, where one start the dynamics close to one of the patterns, is different from inference of patterns from the  $J_{ij}$ . The Hopfield model is *not* a planted model in the sense to which we adhere in this review.

Since this is important we reiterate that the planted model is a statistical physics model where the Boltzmann distribution is at the same time the posterior probability distribution of the corresponding inference problem. There are many models considered in the literature where the quenched disorder is created using some auxiliary variables, but the Boltzmann distribution is not the posterior probability for inference of these auxiliary variables. Then we do not talk about a planted model, an example is given above using the Hopfield model with more than one pattern.

### 1.3.3. Inference via Monte Carlo and variational mean-field

When it comes to solving an inference problem, i.e. evaluating the corresponding MMO of MMSE estimators, see Section 1.2.4, there are many algorithmic methods studied and used in computer science, statistics and machine learning. For nice and comprehensive textbooks, see, for example [12,15]. Two of the classical inference algorithms are particularly interesting in the present context because they originate in statistical physics. These are the Gibbs sampling of Monte Carlo Markov chain (MCMC)-based methods [33–35] and the variational mean-field inference methods [36–39]. Both these methods and their applications are well covered in the existing literature and in this review we will not focus on them. We only describe in the present section their basic concepts and ideas and relation to the message-passing-based algorithms on which we focus in the manuscript.

The development of Monte Carlo goes back to the well-known paper of Metropolis et al. [40] and was widely used since in physics and other scientific fields. The Metropolis–Hastings algorithm [40,41], most commonly used in physics, is based on the comparison of energy cost in one state of a system and a small perturbation of the state (typically the value of one variable is changed). When the cost decreases the change is accepted, when the cost increases the change is accepted with a probability exponentially proportional to the cost difference,  $p = e^{-\beta\Delta E}$ , where  $\beta$  is a inverse-temperature-like parameter. Another perturbation is introduced and the procedure is repeated.

Gibbs sampling [33,34], known as heat-bath in physics, is a variant on the Metropolis–Hastings algorithm where given a fixed state of all but one variable, the probability distribution of the one variable is evaluated conditioning on the fixed state of the other variables. The value of the variables is then sampled from this probability distribution. The process is then repeated. Both Metropolis–Hastings and Gibbs sampling are designed to sample the Boltzmann probability measure. But they can also be famously turned into a generic purpose optimization algorithms called simulated annealing [20].

The advantage of Monte Carlo-based methods is that if iterated for sufficiently long time they provide the exact answer for a very generic and broad class of systems (only ergodicity and balance condition are required). The disadvantage is that the time needed to obtain the exact result can be as large as exponential in the size of the system and it is in general very hard to decide what is the running time necessary for satisfactory precision. Nevertheless, Monte Carlo-based methods are very useful in many cases, in particular when the domain of variables is discrete, and throughout this manuscript we will be comparing to them.

Tracing the precise origins of the use of variational mean-field for inference from data is a little harder. Ideas related to variational mean-field were used in statistical physics since Gibbs (together with Maxwell and Boltzmann) created the field, and are certainly present in the works of Van der Waals [42] and Weiss and Foëx [43] around the 1900s. The modern view on variational mean-field for inference and learning is very nicely summarized in [12,37].

The basic idea of variational inference is to approximate the posterior probability distribution  $P(\mathbf{x} | \mathbf{y})$  (25) by another probability distribution  $Q(\mathbf{x} | \mathbf{y})$  such that (a) for  $Q(\mathbf{x} | \mathbf{y})$  the Bayesian estimators (MMO or MMSE) can be evaluated tractably and (b) the Kullback–Leibler divergence

$$D_{\text{KL}}(Q || P) = \sum_{\mathbf{x}} Q(\mathbf{x} | \mathbf{y}) \log \frac{Q(\mathbf{x} | \mathbf{y})}{P(\mathbf{x} | \mathbf{y})} \quad (26)$$

between  $P$  and  $Q$  is the smallest possible. The variational inference methods are very fast and in many settings and practical problems they provide precise approximations. However, in some settings they provide crude approximations and sometimes misleading results (see an example in Section 5.6).

The present manuscript focuses on message passing inference algorithms, such as BP. In physics the origins of these methods trace back to the Bethe approximation [44], the work of Peierls [45] and the Thouless–Anderson–Palmer (TAP) equations [46]. Interestingly, these equations can be seen as “improved” variational mean-field equations, and this is at the roots of many perturbation approaches (such as [47,48] or [49]).

In information theory BP was invented by Gallager [50] and in Bayesian inference by Pearl [51]. In an excellent article Yedidia et al. [52] relate BP and the variational mean-field methods. However, strictly speaking BP is not a variational method, because the related Bethe free energy does not have to be larger than the true free energy. Compared to variational mean-field methods BP is as fast, and generally more precise. An important property on which we focus in this manuscript, is that on a class of models, that can be called mean-field SG, BP method gives asymptotically exact results (whereas variational mean field does not) and it is amenable to exact analysis that enables detailed understanding of many systems of interest.

#### 1.3.4. Useful statistical physics concepts

Traditional physical systems are composed of a number of spins or molecules that is comparable to the Avogadro number  $N_A \approx 10^{23}$ . It is hence almost always relevant to investigate the limit of large system size that is in statistical physics referred to as the *thermodynamic limit*.

*Self-averaging.* In the analogy between inference and statistical physics of disordered systems, the observations  $\mathbf{y}$  play the role of the so-called *quenched disorder* in the Hamiltonian in Equation (25). Generically, quantities of interest, e.g. the magnetization of variable number 10, depend on the realization of this quenched disorder.

Statistical physics, however, mostly focuses on quantities that have the so-called *self-averaging* property, i.e. in the thermodynamic limit their value does not depend on the realization of the disorder  $\mathbf{y}$  (but only on its statistical properties). More formally, a quantity  $A(\mathbf{y})$  is self-averaging if for every  $\epsilon > 0$

$$\lim_{N \rightarrow \infty} \text{Prob}\{|A(\mathbf{y}) - \mathbb{E}[A(\mathbf{y})]| > \epsilon\} = 0. \quad (27)$$

One of the most important quantities that we expect to be self-averaging is the free energy density

$$f(\beta, \mathbf{y}, N) = -\frac{1}{N\beta} \log Z(\mathbf{y}) \quad (28)$$

that is assumed to have a thermodynamic limit

$$f(\beta) = \lim_{N \rightarrow \infty} f(\beta, \mathbf{y}, N) \quad (29)$$

that does not depend on the realization of  $\mathbf{y}$ .

Among the quantities discussed previously in this manuscript, the ones for which the self-averaging property is particularly interesting are the MMSE and the MMO. Under self-averaging, Equalities (23) and (24) hold not only on average, but also when the averages on the right hand side are removed. This means that the typical squared distance between the ground truth and the MMSE (MMO) estimator is equal, in the thermodynamic limit, to the one evaluated without the knowledge of the ground truth. Clearly, focusing on the analysis of self-averaging quantities greatly simplifies the description of the behavior in the thermodynamic limit and enables a large part of the results obtained in statistical physics.

It should be noted, however, that from a rigorous point of view, proving that the above quantities indeed are self-averaging poses a non-trivial technical challenge. This is a recurring

discrepancy between theoretical physics and rigorous approaches, in physics we make assumptions that are correct and enable progress and understanding of the problems, but that are not always easy to prove.

*Phase transitions.* Another concept borrowed from physics that will turn out very influential in the analysis of high-dimensional inference is phase transitions. Intuitively, a phase transition is an abrupt change in behavior as a parameter (commonly the temperature) is tuned. Since not everything that is termed *phase transition* in computer science or engineering literature is a genuine phase transition from a physics point of view, it is useful to make some reminders.

First of all, true phase transitions do not exist at finite system size. In mathematical physics a phase transition is defined as a non-analyticity in the free energy density  $f(\beta)$  (29). Since the finite  $N$  free energy density is a logarithm of a sum of exponentials, it is always an analytical function. Only when the limit  $N \rightarrow \infty$  is taken can the non-analyticities appear. Therefore when we talk about a phase transition, we should always make sure that it indeed corresponds to a non-analyticity and not to a smooth (even though perhaps very rapid) change.

Secondly, a traditional phase transition is always associated to either a *critical slowing down* or to *metastability*. According to the Ehrenfest's classification of phase transitions, there are two main types:

- *First-order phase transitions* are associated to a discontinuity in the first derivative of the free energy. The first derivative of the free energy is usually related to some measurable quantity that consequently has a discontinuity. Since a number of quantities, including, for example, the MMSE or the magnetization, are related to this first derivative, they also display a discontinuity at a first-order phase transition. The mechanism behind a first-order phase transition is a competition between two stable phases of the system. For the sake of vocabulary let us call the two phases “paramagnet” and “ferromagnet”. Around the phase transition there is a region of phase coexistence in which the properties of the system are ruled by properties of both the phases. In systems that cannot be embedded into a (finite dimensional) Euclidean space we can define sharp boundaries of this parameter-region, called the *spinodal* points. The phase with lower free energy (i.e. higher log-likelihood) is said to be thermodynamically stable, the other one is called metastable. Metastability has profound consequence on the behavior of dynamics and algorithms. We will come back to this in detail.
- *Second-order phase transitions* are associated to a discontinuity in the second derivative of the free energy. This type of phase transition is associated to a so-called *critical slowing down*, meaning that dynamical processes become increasingly slow next to the phase transition. Diverging timescales are closely linked to diverging length scales and diverging strength of fluctuations. These are described by the theory of criticality and critical exponents that ruled statistical physics in the second half of the last century.

Genuine phase transitions are quite peculiar creatures, and this is largely why they fascinate many physicists. To make a distinction in vocabulary in physics we use *cross-over* for a more generic rapid change in behavior as a parameter is tuned. We will keep the term *phase transition* or *threshold* for what is described above. Going back to Figure 1, a physicist's mind becomes very intrigued when for a given inference problem we find that, as the amount of information in the measurements is increased, there exist genuine phase transitions. For a physicist, a phase transition is something intrinsic, so we immediately expect a profound meaning and non-trivial consequences for the system under investigation.



## 2. Planted SG as a paradigm of statistical inference

As biology has its drosophila, or computational complexity its  $K$ -SAT problems, in statistical physics the canonical example is the Ising model

$$\mathcal{H}(\mathbf{S}) = - \sum_{(ij) \in E} J_{ij} S_i S_j - \sum_{i=1}^N h_i S_i, \quad (30)$$

where  $\mathbf{S} = \{S_i\}_{i=1, \dots, N}$  and the spins are  $S_i \in \{-1, 1\}$ ,  $E$  is the set of interacting pairs,  $J_{ij}$  is the interaction strength between spins and  $h_i$  is the magnetic field. Statistical physics is the study of the Boltzmann probability measure on spin configurations

$$\mu(\mathbf{S}) = \frac{1}{Z(\beta, \mathbf{J}, \mathbf{h})} e^{-\beta \mathcal{H}(\mathbf{S})}. \quad (31)$$

In this section we study an inference problem, that we call *the planted SG*, for which, in the teacher–student setting of Section 1.2.2, the posterior probability distribution is the Boltzmann distribution of the Ising model. The purpose of this exercise is simply that we will be able to use a large part of the knowledge existing in the literature on SG to describe in detail the properties of this inference problem. Moreover, the physical picture that arises in this section generalizes to other inference problems in the teacher–student setting. Once we understood in detail the planted SG problem, we can basically deduce without calculation what will qualitatively happen in a wide range of other inference problems. Whereas the problem presented in this section may appear as a very special and relatively simple case, all the concepts that we illustrate on this example are very generic and apply to many other highly non-trivial inference problems. Sections 5 and 6 then give two recently studied examples.

The Ising model with the Hamiltonian (30) is presented in numerous textbooks. The most commonly considered geometries of interactions are finite-dimensional lattices, fully connected graphs, random graphs, or for instance social networks. Interactions  $J_{ij}$  are chosen such that  $J_{ij} \geq 0$  in the ferromagnetic case (most commonly  $J_{ij} = J$  for all  $(ij) \in E$ ) while in the antiferromagnetic case  $J_{ij} \leq 0$ . In the random field Ising model the fields  $h_i$  are random. Perhaps the most intriguing case of an Ising model is the SG, in which the interactions  $J_{ij}$  are taken independently at random from some distribution including both negative and positive numbers. In most of the situations considered in physics literature, the interaction couplings  $\mathbf{J}$  are chosen independently of any specific configuration  $\mathbf{S}$ .

### 2.1. Definition of the planted SG

To motivate the planted SG we consider the following problem: we have a (large) number  $N$  of people and a pack of cards with the number  $+1$  or  $-1$  on each of them. There is about the same number of each. We deal one card to every person. In the next step we choose pairs of people and ask them whether they have the same card or not. Their instructions are to roll a dice and according to the result answer truthfully with probability  $\rho$  and falsely with probability  $(1 - \rho)$ . We collect the answers into a matrix  $J_{ij}$ ,  $J_{ij} = 1$  if the pair  $(ij)$  answered that they had the same card,  $J_{ij} = -1$  if the answer was that they did not have the same card, and  $J_{ij} = 0$  if we did not ask this pair. We want to model this system, and one of the questions we have is whether and how we can estimate which person has which card.

Thinking back about the teacher–student scenario of Section 1.2.2 the teacher distributes the cards, i.e. generates a latent variable  $S_i^* = \pm 1$ ,  $1 \leq i \leq N$ , for each of the nodes/people independently at random. We then select the graph of observations. Let us assume that the pairs to whom



we ask questions are chosen at random, and that we ask  $M$  of them. The set of queried pairs  $E$  can then be represented by an Erdős–Rényi (ER) random graph [53,54] on  $N$  nodes having average degree  $c = 2M/N$ , as defined in Section 3.1. Graphs created in this manner are at the core of many results in this review. The teacher collects the answers into a matrix  $J_{ij} = \pm 1$ ,  $(ij) \in E$ , taken from the following distribution:

$$P(J_{ij}|S_i^*, S_j^*) = \rho \delta(J_{ij} - S_i^* S_j^*) + (1 - \rho) \delta(J_{ij} + S_i^* S_j^*), \quad (32)$$

where  $\rho$  is a bias of the dice (i.e. the probability of getting the correct answer) and  $\delta(\cdot)$  is the Dirac delta function. The latent values  $S_i^*$  are then hidden from the student and the goal is to infer from the values of  $\mathbf{J}$  as precisely as possible which node had which card. Depending on the situation, the teacher did or did not reveal the value  $\rho$  he used. For the moment we will assume  $\rho$  is known.

Let us look at the simplest cases: in the noiseless case,  $\rho = 1$ , each connected component of the graph has only two possible configurations (related by global flip symmetry). When the graph is densely connected, and hence has a giant component, one can therefore easily recover the two groups. When  $\rho = 1/2$ , it is as if we chose the couplings independently at random just as in the most commonly considered SG model. The values  $\mathbf{J}$  do not contain any information about the latent variables  $S_i^*$ , and recovery is clearly impossible. What is happening in between, for  $1/2 < \rho < 1$ ? (The behavior for  $0 < \rho < 1/2$  is then deduced by symmetry.)

The above model is simple to state, but its solution is very interesting, and rather non-trivial. This problem appeared in several different contexts in the literature, and is called *censored block model* in [55–57]. From a rigorous point of view this problem is still a subject of current research. In the physics literature a closely related model is known as the *Mattis model* [58]. Here, we will call the above problem the *planted SG model* to differentiate it from the standard SG model, where we chose all the interactions  $\mathbf{J}$  randomly and uncorrelated. It should be noted that “SG” here is used as a name for the model, it does not necessarily imply that this model must exhibit glassy behavior. The phase diagram on the planted SG is discussed in detail in Section 3.4. We stress that the main goal in inference is to retrieve the planted configuration, whereas the usual goal in physics of SG is to understand the glassy behavior of the model. These two goals are of course related in some aspects, but also complementary in others.

More formally defined, to create an instance, that is a triple  $(\mathbf{S}^*, G, \mathbf{J})$ , of the planted SG we proceed as follows:

- We consider that the hidden assignment  $S_i^*$ , called the *planted configuration* of spins, is chosen uniformly at random from all the  $2^N$  different possibilities.
- We then choose a graph of interactions  $G = (V, E)$ , but not yet the associated values of  $J_{ij}$ , independently of the planted configuration  $\mathbf{S}^*$ .
- For each edge  $(ij) \in E$ , the value of  $J_{ij}$  is chosen to be either  $+1$  or  $-1$  with a probability taken from Equation (33). The signs of the couplings then carry information about the planted configuration.

So far the relation with the Ising model may not be very clear. To see it very explicitly we write Equation (33) differently. Without losing generality, we can define a parameter  $\beta^*$  by

$$\rho = \frac{e^{\beta^*}}{2 \cosh \beta^*} \quad \text{and} \quad 1 - \rho = \frac{e^{-\beta^*}}{2 \cosh \beta^*} \quad \text{with} \quad \beta^* = \frac{1}{2} \log \left( \frac{\rho}{1 - \rho} \right), \quad (33)$$

so that Equation (33) reads

$$P(J_{ij}|S_i^*, S_j^*) = \frac{e^{\beta^* J_{ij} S_i^* S_j^*}}{2 \cosh \beta^*}. \quad (34)$$

Since we want to know the values of the spins  $\mathbf{S}$  given the knowledge of the couplings  $\mathbf{J}$ , we write the Bayes formula

$$P(\mathbf{S} | \mathbf{J}) = \frac{P(\mathbf{J} | \mathbf{S})P(\mathbf{S})}{P(\mathbf{J})} = \frac{\prod_{(ij) \in E} e^{\beta^* J_{ij} S_i S_j}}{2^N (2 \cosh \beta^*)^M P(\mathbf{J})}, \quad (35)$$

where we used Equation (34). The posterior probability distribution has to be normalized, therefore

$$P(\mathbf{J}) = \frac{Z(\beta^*, \mathbf{J})}{2^N (2 \cosh \beta^*)^M}, \quad (36)$$

where  $Z(\beta^*, \mathbf{J})$  is the partition function of the Ising model (31) at inverse temperature  $\beta^*$  (in zero magnetic field). In other words, the posterior distribution is simply the standard Boltzmann measure of the Ising model at a particular inverse temperature  $\beta^*$ . In general we shall consider this Ising model at a generic inverse temperature  $\beta \neq \beta^*$ . This is because a priori we do not know the proportion of errors. In the teacher–student setting, the teacher may not have handed to the student the value  $\beta^*$ .

The special case  $\beta = \beta^*$  corresponds to the Bayes-optimal inference as discussed in Section 1.2.3. One of the most physically important properties of the Bayes-optimal case is that the planted configuration can be exchanged (under averages, or in the thermodynamic limit) for a random configuration taken uniformly from the posterior probability measure, which is in physics simply called an *equilibrium configuration*. All equilibrium configurations with high probability share their macroscopic properties, e.g. their energy, magnetization, various overlaps, etc. This is directly implied by the Nishimori condition (15) derived in Section 1.2.3 for the Bayes-optimal inference. This property will turn out to be crucial in the present context.

As discussed in Section 1.2.4, since we are dealing with a problem with discrete variables (spins), it is desirable to minimize the number of mismatches in the estimation of the planted assignment. We hence aim to evaluate the MMO estimator (22) which in the case of an Ising model is simply

$$\hat{S}_i = \text{sign}(m_i), \quad (37)$$

where  $m_i$  is the equilibrium value of the magnetization of node  $i$  in the Ising model. We reduced the inference of the planted configuration to the problem of computing the magnetizations in a disordered Ising model as a function of temperature.

An off-the-shelf method for this that every physicist has in her handbag is a Monte Carlo simulation, for instance with the Metropolis algorithm. In Section 3 we shall see that for the mean-field setting (i.e. the underlying graph being random or fully connected) there exists an exact solution of this problem via the cavity method. However, before turning to the mean-field setting we want to discuss concepts that are more generic and we will hence keep Monte Carlo in our mind as a simple and universal method to evaluate phase diagrams of Hamiltonians with discrete degrees of freedom.

## 2.2. Nishimori's mapping via gauge transformation

For the planted SG there is an alternative way that provides understanding of its phase diagram and quite interesting additional insights [3, 59, 60]. One takes advantage of the so-called gauge transformation

$$S_i S_i^* \rightarrow \tilde{S}_i, \quad J_{ij} S_i^* S_j^* \rightarrow \tilde{J}_{ij}. \quad (38)$$

If this transformation is applied, the Ising Hamiltonian is conserved since all  $S_i = \pm 1$ . In the new variables the planted configuration becomes ferromagnetic:  $\tilde{S}_i^* = (S_i^*)^2 = 1$ . What happened to

the couplings? The energy of each edge in the planted configuration has not been changed by the gauge transform. Since the frustrated interactions were chosen independently at random in the planting, after the gauge transformation we end up with a standard SG with a fraction

$$\rho = \frac{e^{\beta^*}}{2 \cosh \beta^*} \quad (39)$$

of positive  $J_{ij} = +1$  couplings and the rest  $-1$ , where the  $-1$ 's are chosen independently at random.

After the gauge transformation, the planted assignment has been transformed to the ferromagnetic one where all spins are positive. The overlap between the planted and another randomly chosen equilibrium configuration becomes simply the magnetization. The system is now a standard SG with iid interactions, albeit with a ferromagnetic bias  $\rho$ . The question of the identification of the hidden assignment is now simply mapped to the question of finding the ferromagnetic “all spins up” configuration. This is a step that from the physics point of view simplifies the discussion. It should be stressed, however, that the theory of Bayes-optimal inference, that is the main subject of the present manuscript, applies to a class of planted models for which a similar gauge transformation does not always exist.

The phase diagram of the SG models, with a ferromagnetic bias  $\rho$ , versus a temperature  $1/\beta$ , has been determined and one has simply to look it up in the literature [61–63]. The non-zero value of magnetization in the ferromagnetically biased model then translates into the ability to be able to infer the planted configuration better than by a random guessing. The value of the local magnetization of spin  $i$  is then related to the MMO estimator via Equation (37). In a remarkable contribution, Nishimori realized that the physics on the line  $\rho = e^\beta / (2 \cosh \beta)$  is particularly simple and quite different from a generic  $\beta$ . This line is now called the *Nishimori line* and corresponds to nothing else but the condition for Bayes-optimality  $\beta = \beta^*$  in the planted model.

Consequently, a number of interesting properties, that do not hold in general, hold on the Nishimori line. These are called *Nishimori conditions*. In general, Nishimori conditions are properties that hold in the Bayes-optimal case, such as Equations (23) and (24). Another useful Nishimori identity concerns the average energy  $E = \langle \mathcal{H} \rangle / N$  (30). For more Nishimori conditions see [3,64]. Using again the property that the planted configuration can replace an equilibrium configuration we obtain immediately for the planted SG on ER random graph (as defined in Section 3.1) of average degree  $c$

$$E = -\frac{c}{2} \tanh \beta^*. \quad (40)$$

In Figure 2 we give the phase diagram of the SG at temperature  $T = 1/\beta$  with ferromagnetic bias  $\rho$ . As an example we use the SG on a random graph in which case an exact solution exists (see Section 3). However, on a generic geometry the phase diagram could be obtained using Monte Carlo simulation and would be qualitatively similar. Due to the Nishimori mapping this is also the phase diagram of the planted SG, where the ferromagnetic phase is understood as correlated to the planted configuration. The phenomenology of a rather generic Bayesian inference problem can be directly read off from this phase diagram.

### 2.3. Bayes-optimality on the Nishimori line

First, let us assume that we know (or that we learn from data, see next section) the true value of  $\beta^*$  (or equivalently  $\rho$ ) for which the couplings  $\mathbf{J}$  have been generated. Then, our consideration is restricted to the Nishimori line. If the temperature is large enough, i.e.  $\beta^*$  is low enough, we

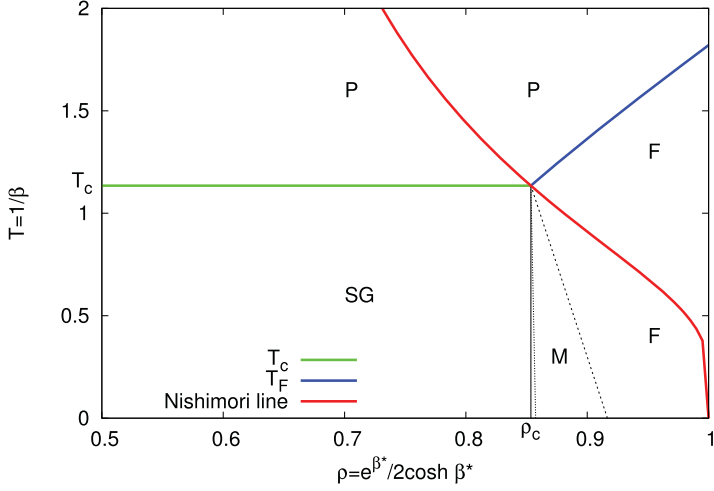


Figure 2. Phase diagram of the planted SG as a function of temperature  $T = 1/\beta$  and density  $\rho$  of edge interactions that agree with product of spins in the planted configuration. This phase diagram is for random regular graphs of degree  $c = 3$ . The Nishimori line,  $\rho = e^{\beta^*}/(2 \cosh \beta)$  (red online), corresponds to the Bayes-optimal inference. There is the SG phase where the system is in a glassy state, the ferromagnetic (F) phase, the paramagnetic (P) phase and the mixed (M) one where the system is both SG and F. The boundary of the mixed phase (black dashed) is only a guide for the eye between the multi-critical point and the zero-temperature point from [65]. The same for the boundary of the SG phase (black dotted) that is obtained from the inexact 1RSB approximation and is hardly distinguishable from the vertical line (black full).

are in the paramagnetic phase, and we find that all marginals (magnetizations) are zero. It would thus be difficult to find their signs. In fact, this simply means that, in the paramagnetic phase, inference of the hidden configuration is impossible. There is just not enough information to say anything about the hidden/planted assignment.

As the temperature is lowered, i.e.  $\beta^*$  is increased, the situation changes. If we used  $\beta^*$  corresponding to an inverse temperature above a certain critical value  $\beta_c$ , then a Monte Carlo simulation would rapidly discover the ferromagnetic phase where local magnetizations are (mostly) positive. Applying rule (37) would give at this point a decent overlap with the true assignment (which is, after the gauge transformation, all +1).

In terms of inference of the planted configuration these two phases have the following interpretation:

- *Undetectable*: For  $\beta^* < \beta_c$ ,  $\rho < \rho_c$ , the planted SG system is in the paramagnetic state. Meaning that the set of observed variables  $\mathbf{J}$  does not contain any information about the planted configuration and its inference is hence information-theoretically impossible.
- *Detectable*: For  $\beta^* > \beta_c$ ,  $\rho > \rho_c$ , the planted SG is in a ferromagnetic phase where equilibrium configurations are all correlated with the planted one and inference is possible, and algorithmically tractable, for example via Monte Carlo sampling.

For the planted SG of a random graph, the critical temperature  $T_c = 1/\beta_c$  where one goes from *impossible inference* to *easy inference* can be taken from the literature [62,63,66] (we will show how to compute it for this case in Section 3) and in the case of random ER graphs (as

defined in Section 3.1) with average degree  $c$  it reads

$$T_c^{\text{ER}}(c) = \left[ \operatorname{atanh} \frac{1}{\sqrt{c}} \right]^{-1}, \quad (41)$$

while for regular random graphs it reads

$$T_c^{\text{REG}}(c) = \left[ \operatorname{atanh} \frac{1}{\sqrt{c-1}} \right]^{-1}. \quad (42)$$

On an ER random graph (as defined in Section 3.1), this indicates that it is only possible to infer something about the planted assignment if the number of edges in the graph  $cN/2$  is such that

$$c > c_{\text{detect}} = \frac{1}{(2\rho - 1)^2}. \quad (43)$$

If this is not satisfied, the system is in the paramagnetic phase and nothing can be said about the planted configuration. If, on the other hand, Equation (43) is satisfied, then the magnetization aligns partly with the planted configuration. This is a very generic situation in Bayes-optimal inference when we encounter a second-order phase transition. We go from a phase where inference is impossible, to a phase where it is rather straightforward. Note that this is not only a statistical physics prediction, but one that can be proven rigorously. The impossibility result has been shown in [67,68] and the possibility one (if Equation (43) is satisfied) in [57].

#### 2.4. Parameter mismatch and learning

Let us now discuss what happens if one does not know the value  $\beta^*$ . One approach that seems reasonable is to use  $\beta \rightarrow \infty$ , so that we look for the most probable assignment (this would correspond to the MAP estimator) which is the ground state of the corresponding Ising model. We observe from the phase diagram in Figure 2 that for  $\rho < \rho_c$ , at  $T = 0$  the system is in the SG phase. In terms of Monte Carlo, the presence of such a phase manifests itself via a drastic growth of the equilibration time. Computing the corresponding magnetizations becomes very hard. Moreover, even in the undetectable phase,  $\rho < \rho_c$ , the zero-temperature local magnetizations will not be zero. This is a sign of overfitting that the MAP estimator suffers from.

In the detectable phase,  $\rho > \rho_c$ , the ferromagnetic phase extends only up to a temperature that for an ER random graph corresponds to a certain

$$\beta_F = \operatorname{atanh} \left( \frac{1}{c \tanh \beta^*} \right). \quad (44)$$

Hence if we chose too low value of  $\beta$  as a starting point we could miss the detectability phase. At low temperatures we might again encounter a SG phase that may or may not be correlated to the planted configuration. In either case, such a phase causes equilibration problems and should better be avoided. Fortunately, there is no such glassy phase on the Nishimori line as we shall explain in Section 2.7.

The above considerations highlight the necessity of learning parameters (in this case  $\beta^*$ ). Within the theory described above there is one particularly natural strategy for learning such parameters, closely related to the well-known algorithm of expectation maximization [69]. Let us illustrate this parameter learning on the example of the planted Ising model where we assume that the parameter  $\beta^*$  is not known.

In Bayesian inference, we always aim to write a posterior probability distribution of what we do not know conditioned to what we do know. In the present case we need  $P(\beta | \mathbf{J})$ . Using the Bayes formula we obtain

$$P(\beta | \mathbf{J}) = \sum_{\mathbf{S}} P(\mathbf{S}, \beta | \mathbf{J}) = \frac{P(\beta)}{2^N P(\mathbf{J})} \frac{Z(\beta, \mathbf{J})}{(2 \cosh \beta)^M}. \quad (45)$$

We notice that as long as the hyper-prior  $P(\beta)$  is independent of  $N$ , in the thermodynamic limit this probability distribution  $P(\beta | \mathbf{J})$  converges to a Dirac delta function on the maximum of the second fraction on the right hand side. This maximum is achieved at

$$E = - \left\langle \frac{1}{N} \sum_{(ij) \in E} J_{ij} S_i S_j \right\rangle = - \frac{cN}{2} \tanh \beta. \quad (46)$$

The corresponding iterative algorithm for learning  $\beta$  is as follows. We start with some initial choice of  $\beta$  and compute the average energy to update the value of  $\beta$  according to Equation (46) until a fixed point is reached. Depending on the initial value of  $\beta$ , and on the algorithm used to estimate the average energy, in some cases we may obtain several different fixed points of Equation (46). In those cases we choose the fixed point that is the global maximum of Equation (45).

Notice now a very useful property, the condition for stationarity of Equation (45) is the same as the Nishimori condition (40) and indeed  $\beta = \beta^*$  is a fixed point of Equation (46). This strategy is very reasonable, in the Bayes-optimal case all the parameters are the ground truth ones and a number of *Nishimori conditions* hold only in this case. Hence in order to learn the right values of parameters we iteratively impose the Nishimori conditions to hold.

Finally, let us note that in the above planted Ising model the underlying phase transition is of second order, therefore there is no particular computational issue. Inference is either impossible, or possible and tractable. The hard phase outlined in Section 1.2.5 is absent. The picture is richer in problems where in the random case there is a discontinuous phase transition. Examples of such a case are described in detail in Section 3.5.

## 2.5. Relations between planted, quenched and annealed disorders

From the early works, it sometimes seems necessary to have a gauge transformation (and consequently a special form of the Hamiltonian) to be able to speak about the Nishimori conditions or other properties of the Nishimori line [59,60,70]. Importantly this is not the case. The properties of the Nishimori line are more general and apply to every case where it makes sense to talk about Bayes-optimal inference. In fact, a more general notion of the Nishimori line can be identified with the notion of Bayes-optimality. A nice article on this connection was given by Iba [71].

In this section, we shall discuss properties of the planted disorder, and its relation with the more commonly considered – the randomly quenched and the annealed disorders. This clarifies the generic connection between statistical physics of SG and Bayesian inference. We stress again that none of this requires a gauge symmetry of the underlying Hamiltonian.

Let us consider back the Hamiltonian of the Ising system (30). For a given system, for example, the magnet on the fridge in your house, the set of couplings  $J_{ij}$  is a given fixed finite set of numbers. In the planted SG above, this set was the values of interactions  $\mathbf{J}$  we observed.

In statistical physics, where in the vast majority of cases we can take advantage of self-averaging, see Section 1.3.4, we often do not think of a given set of numbers  $J_{ij}$  but instead about

a probability distribution over such sets. For instance randomly quenched disorder is generated following

$$P(J_{ij}) = \frac{1}{2}\delta(J_{ij} - 1) + \frac{1}{2}\delta(J_{ij} + 1) \quad (47)$$

for all  $(ij) \in E$ , where  $E$  are the edges of the graph of interactions. In this way we do not have a given Hamiltonian, but rather its probability of occurrence, or equivalently an ensemble where instances appear proportionally to this probability. Properties of the system then have to be averaged over the disorder (47). This view is adopted in statistical physics because the self-averaging is assumed to hold in most physical systems of interest and the theory is then easier to develop. This notion of averaging over disorder is so well engraved into a physicist's mind that going away from this framework and starting to use the related (replica and cavity) calculations for a single realization of the disorder has lead to a major paradigm shift and to the development of revolutionary algorithms such as the survey propagation [72]. In this section we do, however, want to stay in the mindset of averaging over the disorder and explain how to fit in the notion of the *planted ensemble*, i.e. ensemble of instances where we average over the disorder that was planted.

### 2.5.1. The quenched and annealed average

How to take the average over disorder of  $J_{ij}$  meaningfully is a problem that was solved a long time ago by Edwards in his work on SG and vulcanization (see the wonderful book *Stealing the Gold* [73]). 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 needs to compute the average free energy

$$f_{\text{quenched}} = \left[ \frac{F}{N} \right] = \lim_{N \rightarrow \infty} -\frac{1}{\beta N} [\log Z], \quad (48)$$

where  $[\cdot]$  denotes the average over the disorder. This is called the quenched average, and the corresponding computations are the quenched computations. In fact, the self-averaging hypothesis for the free energy has now been proven rigorously in many cases, in particular for all lattices in finite dimension [74] and for some mean-field models [75]. 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 logarithm of  $Z$ , known today as the replica trick, using the identity

$$\log Z = \lim_{n \rightarrow 0} \frac{Z^n - 1}{n}. \quad (49)$$

The idea here is that while averaging the logarithm of  $Z$  is usually a difficult task, taking the average of  $Z^n$  might be feasible for any integer value of  $n$ . Performing a (risky) analytic continuation to  $n = 0$ , one might compute the averaged free energy over the disorder as

$$f_{\text{quenched}} = -\lim_{N \rightarrow \infty} \frac{1}{N\beta} \lim_{n \rightarrow 0} \frac{[Z^n] - 1}{n}. \quad (50)$$

The computation of the quenched free energy is in general still very difficult and remains a problem at the core of studies in statistical physics of disordered systems. It is much easier to consider the so-called annealed average, one simply averages the partition sum and only *then* takes the logarithm

$$f_{\text{annealed}} = -\frac{1}{N\beta} \log [Z]. \quad (51)$$

This is a very different computation, and of course there is no reason for it to be equal to the quenched computation.



For the example of the Ising model with binary uniformly distributed couplings (47) we obtain (independently of the dimension or geometry)

$$-\beta f_{\text{annealed}}^{\text{Ising}} = \log 2 + \frac{c}{2} \log(\cosh \beta), \quad (52)$$

where  $c$  is the average coordination number (degree) of the interaction graphs.

It is important to notice that the annealed average 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 situations as  $\propto (1/\sqrt{N})$  (the exponent can be different, but the idea is that fluctuations are going to zero as  $N \rightarrow \infty$ ). The partition sum  $Z$ , however, is exponentially large in  $N$ , and so its fluctuations can be huge. The average can be easily 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  $f_{\text{quenched}} = 2$ , however in the annealed computation one finds

$$[Z] = \frac{1}{N} \exp(-\beta N) + \left(1 - \frac{1}{N}\right) \exp(-2\beta N) \quad (53)$$

and to the 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 we shall see, 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

$$f_{\text{annealed}} \leq f_{\text{quenched}}. \quad (54)$$

This is in fact a crucial property in the demonstrations of many results in the physics of disordered systems, and in computer science this is the inequality behind the “first moment method” [76].

Furthermore, there is a reason why, in physics, one should sometimes consider the annealed average 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.

### 2.5.2. Properties of the planted ensemble

In Section 2.1 we described the planted SG, where we first considered the planted configuration  $\mathbf{S}^*$  and a graph  $G = (V, E)$ , and then we considered the probability over interactions  $P(\mathbf{J} | \mathbf{S}^*)$  (34). The ensemble of instances  $(\mathbf{S}^*, G, \mathbf{J})$  generated by the planting, i.e. by the three steps of Section 2.1, is called the planted ensemble.

From a physics point of view the two most important properties of the planted ensemble, the two golden rules to remember, are:

- The planted configuration is an equilibrium configuration of the Hamiltonian derived from the posterior distribution. By this we mean that the planted configuration has exactly the same macroscopic properties (e.g. its energy or magnetization) as any other configuration that is uniformly sampled from the posterior distribution. Indeed, when we view planting as an inference problem, we can write the corresponding posterior probability distribution as in Equation (36). In Section 1.2.3 we derived that in the Bayes-optimal inference the

planted configuration behaves exactly in the same way as a configuration sampled from the posterior.

Given that generating an equilibrium configuration from a statistical physics model is in general a difficult task (this is, after all, why the Monte Carlo method has been invented!) this is certainly a very impressive property. It actually means that, by creating a planted instance, we are generating at the same time an equilibrium configuration for the problem, something that would normally require a long Monte Carlo simulation.

- The realization of the disorder of the planted problem is not chosen uniformly, as in the randomly quenched ensemble (47), but instead each planted instance appears with a probability proportional to its partition sum. This can be seen from Equation (36) where the probability distribution over  $\mathbf{J}$  in the planted ensemble is related to the partition function of the planted system  $Z(\mathbf{J})$ . A more explicit formula can be obtained using the annealed partition function (52)

$$P_{\text{planted}}(\mathbf{J}) = \frac{Z(\mathbf{J})}{\Lambda Z_{\text{annealed}}}, \quad (55)$$

where  $\Lambda$  is the number of all possible realizations of the disorder,  $\Lambda = 2^M$  for the Ising model. Note the  $1/\Lambda$  can also be interpreted as the probability to generate a given set of couplings in the randomly quenched ensemble where the couplings are chosen uniformly at random.

The average over the planted ensemble might seem related to the annealed average because for instance the planted energy (40) is always equal to the annealed average energy, which is easily derived from the annealed free energy (52). However, the free energy of the planted ensemble is in general different from the annealed free energy (52), as illustrated, for example, in Figure 3.

### 2.5.3. Quiet planting

Let us now imagine we have found a system where the (quenched) free energy is exactly equal to the annealed free energy. Then according to Equation (55) we see that for such a system  $P_{\text{planted}}(\mathbf{J}) = 1/\Lambda = P_{\text{quenched}}(\mathbf{J})$ , meaning that generating instances from the planted ensemble is the same thing as generating from the randomly quenched ensemble. Such planting is denoted as *quiet*. In fact, we do not even need the free energies to be exactly equal, it is sufficient that the free energy densities are the same. This is because atypical instances are usually exponentially rare and hence any difference in the free energies that does not show up in the leading order will not generate atypical instances.

It so happens that in mean-field systems the (self-averaging) free energies of paramagnets are indeed equal to the annealed free energies. The equilibrium in these systems can hence be studied using the planted ensemble and this is greatly advantageous as we will see in Section 2.8.1.

The idea of quiet planting comes from a rigorous work [28] where the above notion was proven rigorously. In mathematics this property is formalized and the two ensembles termed *contiguous* [28,78]. The paper that pioneered the usage of quiet planting in physics, and that also established the name “quiet”, is *Hiding quiet solutions in random CSP* [31].

## 2.6. Replica symmetry breaking

The replica and cavity method, that are among the main assets that statistical physics offers for analysis of inference problems, were originally developed to understand behavior of glasses and SG via the so-called replica symmetry breaking (RSB). We will explain in the next section, that

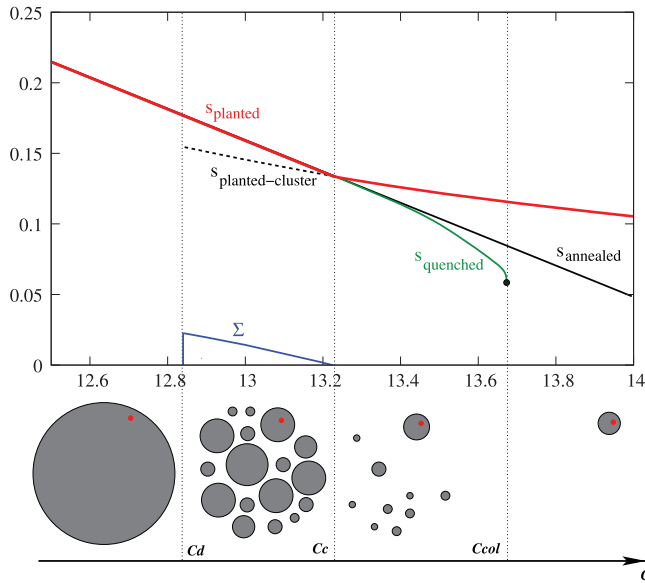


Figure 3. Entropy density (which is equal to the free energy at zero temperature) for the five-coloring problem on an ER random graph (as defined in Section 3.1) of average degree  $c$ . The planted entropy (in red) is compared to the quenched one (in green) and the annealed one (in black). For  $c < c_c$  the three curves agree. The quenched entropy goes to  $-\infty$  at the colorability threshold  $c_{\text{col}}$ . The region between  $c_d$  and  $c_c$  is relevant only for cases with a first-order phase transition as discussed in Section 3.5. In mean field theory of SG this region is called the d1RSB phase [77]. We also show the entropy of the planted cluster below  $c_c$  (black dashed), which is less than the total entropy, the difference is called the “complexity”  $\Sigma$  (in blue) and is the logarithm of the number of pure states. Bottom: sketch of the shape of the space of all valid colorings in the planted case. At an average degree  $c_d$  the space of solutions shatters into exponentially many clusters, the planted cluster being one of them. Beyond  $c_c$  the planted cluster contains more solutions than all the others together. At  $c_s$  the last non-planted cluster disappears (this is the coloring threshold in a non-planted, purely random model). Figure taken from [31].

for Bayes-optimal inference RSB is not needed. But to appreciate fully the results exposed in this manuscript it is useful to have a very basic idea of what RSB is and what it implies.

RSB is usually explained together with the related (relatively involved) calculations, see, for example [3,5]. In this section we try to convey the basic ideas behind RSB without going into the computations. We rely on the descriptions of what would happen to a Monte Carlo simulation in a system that exhibits various versions of RSB.

Let us start with the concept of *replica symmetry*. For the purpose of the present manuscript we can say that a probability measure  $P(\mathbf{x} | \mathbf{y})$  is replica symmetric if a MCMC that is initialized at equilibrium, i.e. in a configuration sampled uniformly at random from  $P(\mathbf{x} | \mathbf{y})$ , is able to sample  $P(\mathbf{x} | \mathbf{y})$  (close to) uniformly in a number of iterations that is linear in the size of the system. A probability measure  $P(\mathbf{x} | \mathbf{y})$  exhibits dynamical one-step of replica symmetry breaking (d1RSB) if MCMC initialized at equilibrium explores in linear time only configurations that belong into an exponentially small (in  $N$ ) fraction of the equilibrium configurations. In both RS and d1RSB the distance between two configurations sampled from the Boltzmann probability measure is the same number with probability going to one (as  $N \rightarrow \infty$ ).

We speak about static one-step replica symmetry breaking (1RSB) if MCMC initialized at equilibrium is able to explore configurations belonging to a finite fraction of all equilibrium

configurations. Moreover the distance between two randomly sampled configurations can take (with probability going to one) one of the two possible values.

A probability measure  $P(\mathbf{x} | \mathbf{y})$  corresponds to full-step replica symmetry breaking (FRSB) if the distance between two randomly chosen configurations is distributed according to some probability distribution with continuous non-trivial support. In terms of behavior of MCMC, in a FRSB regime the time needed to sample configurations uniformly starting from equilibrium has a more complex behavior: relaxations are critical and a power-law behavior is expected [79].

## 2.7. No RSB in Bayes-optimal inference

From the methodological point of view, the Nishimori line has one absolutely crucial property: the absence of glass phase at equilibrium. In the words of the mean-field theory of SG, there is no static RSB on the Nishimori line. Note that there might be, and indeed is in all the cases where the underlying transition is of the first order, the dynamical one-step RSB phase (d1RSB) [77]. The marginals in the d1RSB phase are, however, still exactly described by the BP algorithm (as described in Section 3), and this is what matters to us in the Bayes-optimal inference.

A RSB phase (or equivalently the static SG phase) can be defined by the non-self-averaging of the overlap  $q$  between two configurations randomly sampled from the Boltzmann distribution. Depending on the realization of the sampling, the Hamming distance between the two configurations is with non-zero probability different from its average.

In Section 1.2.3 we derived that, on the Nishimori line, the overlap  $m$  between a typical configuration and the planted one is equal to the overlap between two typical configurations  $q$ . With similar arguments one can derive equality of the distribution (over realizations of the sampling)  $P(q)$  of the overlaps between two configurations randomly sampled from the posterior, and the distribution of the magnetizations  $P(m)$ , which is the overlap between the planted configuration and a random sample from the posterior [3]. In physics, the magnetization is always argued to be self-averaging (the same is true for the free energy density, and the magnetization is its derivative with respect to the magnetic field), and hence on the Nishimori line the overlap is also self-averaging. From this, one argues that  $P(q)$  is trivial, which in mean-field SG indicates the absence of a SG phase. However, from a rigorous point of view, self-averaging of the magnetization remains an open problem.

Another way to see the problem is to remark that, given that the SG susceptibility is equal to the ferromagnetic one for planted systems (see for instance [64,80]), then it is hard to reconcile the presence of an equilibrium ferromagnet phase (where the susceptibility is finite) with an equilibrium SG phase (where the susceptibility diverges).

A complementary (rigorous) argument was presented by Montanari [81] who showed that in sparse systems (i.e. each observations depends on a bounded number of variables) in the Bayes-optimal setting, two point correlations decay. In the static SG phase these correlations could not decay.

This means that we can exclude the presence of a static equilibrium transition to a glass phase on the Nishimori line. Note, however, that the dynamics can still be complicated, and in fact as we mentioned, a dynamic SG (d1RSB) phase can appear. As soon as we deviate from the Bayes optimal setting and mismatch the prior, the model or their parameters, the stage for glassiness to arrive is open. And indeed as we see in Figure 2 there are regions with RSB when  $\beta \neq \beta^*$ , typically when we take the temperature too small. When studying inference with a mismatched prior or model we thus always have to keep this in mind before making claims about exactness.

## 2.8. Applications of planting for studies of glasses

Now that we have explored the concept of planting and quiet planting, and its connection to inference, we will show that it also has a number of consequences for the study of structural glasses in physics. While these are not directly relevant to inference, it illustrates in many ways the fascinating properties of the planted ensemble and the deep connection between inference, optimization problems and the statistical physics of glasses, and we thus summarize these properties in this section.

### 2.8.1. Equilibration for free in glasses

Let us discuss the physical consequences of the fact that in quietly planted systems the planted configuration behaves exactly the same way as any other equilibrium configuration. We can hence literally equilibrate for free. This is particularly interesting in the study of glassy models. As a matter of fact, the very concept of glassiness can be defined as super-fast growth of the equilibration time as the glass transition temperature is approached [77].

The core of the interest in glasses is concerned by finite (usually three) dimensional models. In those cases the annealed and quenched free energies are known not to coincide. A class of theories of glasses is, however, based on mean-field (i.e. systems living on random geometries of fully connected graphs) disordered systems that present a discontinuous phase transition. The analog of  $c_d$  in these systems is called the dynamical glass transition, and the analog of  $c_c$  is called the Kauzmann glass temperature [82]. Properties of the equilibration time are well studied and it turns out that the equilibration time (i.e. number of Monte Carlo steps per variable in the thermodynamic limit) diverges as the dynamical temperature is approached [83] and is exponentially large in the system size between the dynamical and Kauzmann temperature. This in practice means that numerical simulations of these models are very cumbersome.

In most of the mean-field models used for the studies of glasses, the quiet planting works. Numerical simulations studying the dynamics from equilibrium can hence be speeded up considerably by initializing in the planted configuration. A particular version from this trick was first used in [84] where it enabled for the first time precise simulation of the dynamics starting from equilibrium. It was later popularized by Krzakala and Zdeborová [31] and has now been used in a considerable number of works. Without being exhaustive let us mention Refs. [85–90].

Even if one does not care about the equivalence to the randomly quenched ensemble one can use planting to obtain equilibrium configurations even in general finite-dimensional systems. Indeed the property of planted configuration being the equilibrium one for a corresponding Boltzmann measure is general. This has been used in numerical studies of so-called pinning in glasses, see, for example [91].

### 2.8.2. Following states and analyzing slow annealing

In a very influential work, Franz and Parisi [92] considered the so-called *potential* method, that aimed to analyze the properties of a single pure state in glasses. Perhaps more interestingly, this allows to describe the adiabatic evolution of these glassy Gibbs states as an external parameter, such as the temperature, is slowly tuned. Again, there is a deep connection with the planted ensemble which we shall now discuss. Krzakala and Zdeborová [31] studied the state that is created in the planted coloring around the planted configuration. Adding to this the concept of quiet planting, this framework allows us to study the *shape* of states in mean-field glassy systems and is equivalent to the Franz–Parisi construction.

When a macroscopic system is in a given phase, and if one tunes a parameter, say the temperature, very slowly then all observables, such as the energy or the magnetization in a magnet,

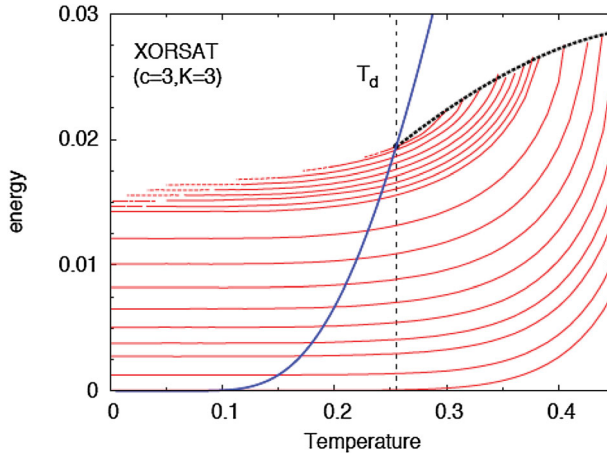


Figure 4. State following in regular 3-XOR-SAT. Figure taken from [97].

will be given by the equilibrium equation of state. In a glassy system, where there could be many such states, it is very interesting to see how they evolve upon cooling, as they tend to fall off equilibrium very fast. This is for instance what is happening in the so-called 3-XOR-SAT problem, also called 3-spins SG [93–95]. This is nothing but an Ising SG where the two-body interaction of Hamiltonian (30) is replaced by a three-spin interaction (i.e. each term is a triplet  $J_{ijk}S_iS_jS_k$ ).

A plot that illustrates this phenomena is given in Figure 4. The blue line is the equilibrium (in technical terms, the one-step RSB solution) energy of the 3-regular hypergraph (this means that every variables has exactly three triplets of interactions, chosen randomly) 3-XOR-SAT problem as a function of temperature  $T$ . For this system quiet planting works at *all* temperatures. The red lines are energies that we obtain when we plant quietly (at temperature where the red lines crosses the blue one) and *then* lower or increase the temperature and monitor the corresponding energy given by the dynamics. This procedure has been called *state following* in [96,97], because physically this is a study of the corresponding state at temperature at which the state is no longer among the equilibrium ones.

Note that such state following was often described in physics of glasses on the level of a thought-experiment. Methods for quantitative analysis existed only for the particular case of spherical  $p$ -spin model, via an exact solution of the dynamics [77,98,99]. For other systems, in particular diluted ones, state following was considered to be an open problem, see, for example, the concept suggested but not analyzed in [100]. The method for state following is mathematically very much related to the concept of Franz–Parisi potential at two different temperatures [92,101].

From an algorithmic point of view, the method of state following opens the exciting possibility of analyzing limiting energy of a very slow annealing. It is well known that in a finite size system annealing always finds the ground state, in a time that is in general exponential in the system size. A crucially interesting question is: What is the lowest achievable energy when the annealing rate scales linearly with the size of the system? In mean-field models of glasses equilibration is possible in a time that is linear in the system size down to the dynamical temperature  $T_d$  (or equivalently up to  $c_d$ ), after that the dynamics gets blocked in one of the exponentially many states. If we take an equilibrium state at  $T_d - \epsilon$  and follow it down to zero temperature the resulting energy is a very good candidate for the behavior and for the limiting energy of very



slow simulated annealing. Very slow annealing for the Ising  $p$ -spin model was compared to the theoretical prediction from state following in [102].

*Paradox in state following.* There is, however, a problem that appears when one tries to carry out the calculation for following states that are at equilibrium close to  $T_d$ . When following the state down in temperature, the replica symmetry gets broken at some point (the boundary between full and dashed lines in Figure 4). Moreover, at yet lower temperature the solution correlated to the planted configuration disappears (in a mathematically analogous way as when a spinodal line is encountered). Even using the 1RSB computation to describe the internal structure of the state we were not able to follow it down to zero temperature. A possible explanation is that the correct description of this region is the FRSB. To test this hypothesis, Sun et al. [103] analyzed the mixed spherical  $p$ -spin model where the FRSB is tractable.

The findings of Sun et al. [103], however, raised even more questions. It is found by the Nishimori mapping to the model with a ferromagnetic bias (as described in Section 2.2) that there is no FRSB solution to the problem, yet there is no magnetized 1RSB solution at very low temperature either. Yet in MCMC simulation one always confirms the expectation that as temperature is lowered the bottom of the state is correlated to the reference configuration. In other words, using the Nishimori mapping, the non-magnetized SG phase that is assumed to exist at low temperatures for  $\rho > \rho_c$  in Figure 2 is not physical in the context of state following. These observations are hence an inconsistency in the theory. Loose ends such as this one are always interesting and basically always eventually lead to new theoretical development. This state following paradox should hence not be forgotten and should be revisited as often as new possibilities open.

One such possibility might be related to the calculation involving a chain of reference (planted) configurations at lower and lower temperatures as suggested in [104].

### 2.8.3. *Melting and glass transition*

Using the planted models and extended form of the Nishimori conditions the analogy between glassy dynamics and melting was studied in [64,80]. Perhaps the theoretically most intriguing and very fundamental question in the theory of glasses is: Is there a true glass transition in a finite-dimensional system (i.e. system where the graph of interactions can be embedded into finite-dimensional Euclidean space)? Such an ideal glass transition has to come with diverging time and length scales when approaching the transition. A definitive answer to this question is difficult to obtain as both simulations and experiments are faced with the extremely slow dynamics. According to the random first-order theory of the glass transition [105,106], there exist systems with time and length scales with a genuine divergence at the ideal glass transition temperature. Moreover, this theory is able to predict in which systems such an ideal glass transition exists and in which it does not. However, the random first-order theory is not free from criticisms, see for instance [107] and references therein.

The question of existence of the ideal glass transition remains open, but in [64] it was derived that if there is a finite-dimensional model with a first-order phase transition on the Nishimori line (or equivalently in the Bayes-optimal inference) then there is an ideal glass transition as well. This result was reached by using the Nishimori mapping from a planted model to a ferromagnet, Section 2.2, and comparing their dynamical properties. In particular, if there is a first-order phase transition on the Nishimori line then it comes with two quite uncommon properties – absent latent heat, and the melting dynamics being equivalent to the stationary equilibrium dynamics. Arguably, first-order phase transitions are easier to describe or exclude than the tricky glass transition. This is hence an interesting path toward the fundamental question of existence of the ideal glass transition.



### 3. Phase diagram from the cavity method

Now we move to special cases of inference problems where the phase diagram and phase transitions can be computed analytically exactly, and for which related algorithms more efficient than Monte Carlo can be derived. We shall achieve this goal using extended mean-field methods. Slightly abusively we will call mean-field systems those where these methods lead to exact results.

Mean-field methods are often the first step in the understanding of systems in physics, and disordered systems are no exception. In SG, in particular, the mean-field model proposed in 1975 by Sherrington and Kirkpatrick [108] has been particularly important. We refer to the classical literature [7] for the replica solution, involving the so-called RSB, devised by Parisi. It is not that we are not interested in the replica method, quite the contrary, the replica method is at the roots of many results in this review. We shall, however, focus here in the alternative cavity method [21,22,109] since this leads more naturally to specific algorithms and algorithmic ideas. The development of more efficient algorithms is the goal in inference problems.

There are two main types of lattices for which the mean-field theory is exact. The first type is the fully connected lattice underlying the canonical Sherrington–Kirkpatrick model. Similar solutions have been derived for the  $p$ -spin model [93] and for the Potts glass model [110], and these played a major role in the development of the mean-field theory of the structural glass transition [111–113]. The linear estimation problem that we treat in Section 6 is an example of a problem on a fully connected lattice.

The second type of lattice for which the mean-field theory is exact is given by large random graphs with constant average degree, a case commonly referred to as “Bethe lattice” in the physics literature [114]. In 2001, Mézard and Parisi [21,22], using the cavity method, adapted the RSB scheme to solve models on such sparse lattices. There is a number of reasons why Bethe lattices are interesting, the two major ones are: (a) because of the finite degree, they provide a better approximation of the finite dimensional case and because the notions of distance and neighboring can be naturally defined, and (b) the related random graphs are fundamental in many interdisciplinary cases, for instance in computer science problems, which is our main motivation here.

The cavity method is a generalization of the Bethe and Onsager ideas [115,116] for disordered systems. It was initially developed by Mézard et al. [109] as a tool to recover the solution of the Sherrington–Kirkpatrick model without the use of replicas. It was subsequently developed by Mézard and Parisi to deal with the statistical physics of disordered systems on random graphs [21]. There is a deep global connection between the Bethe approximation and the so-called BP approach in computer science (in error correction [117] and Bayesian networks [51]) as was realized in connection to error correcting codes by Kabashima and Saad [118], and was put in a more general setting by Yedidia et al. [52]. For a more recent textbook that covers in detail the cavity method and BP see [5], for a very comprehensive explanation of BP see [52].

#### 3.1. Random graphs

We will remind basic properties of sparse ER [53,54] random graphs, that we used briefly already in Section 2.1. At the core of the cavity method is the fact that such random graphs locally look like trees, i.e. there are no short cycles going through a typical node.

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

The constant  $c$  is called the average degree. We denote by  $c_i$  the degree of a node  $i$ , i.e. the number of nodes to which  $i$  is connected. The degrees are distributed according to Poisson distribution, with average  $c$ .

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

We will mainly be interested in the sparse graph limit, where  $N \rightarrow \infty$ ,  $c \sim O(1)$ ,  $M \sim N$ . The key point is that, in this limit, such random graphs can be considered locally as trees [119,120]. The intuitive argument for this result is the following one: starting from a random site, and moving following the edges, in  $\ell$  steps  $c^\ell$  sites will be reached. In order to have a loop, we thus need  $c^\ell \sim N$  to be able to come back on the initial site, and this gives  $\ell \sim \log(N)$ .

### 3.2. BP and the cavity method

Now that we have shown that random graphs are locally tree-like, we are ready to discuss the cavity method. We will consider a system with Hamiltonian

$$\mathcal{H} = - \sum_{(ij) \in E} J_{ij} S_i S_j - \sum_i h_i S_i, \quad (56)$$

and derive the thermodynamics of this model on a tree. We have to deal with  $N$  variables  $S_i$  with  $i = 1, \dots, N$ , and  $M$  interactions, that we denote  $\psi_{ij}(S_i, S_j)$ , on each edge  $(ij)$  of the graph  $G(V, E)$ , with  $\psi_{ij}(S_i, S_j) = \exp(\beta J_{ij} S_i S_j)$ . Further denoting  $\psi_i(S_i) = \exp(\beta h_i S_i)$  the “prior” information the Hamiltonian gives about  $S_i$ , the total partition sum reads

$$Z = \sum_{\{S_i=1,\dots,N\}} \left[ \prod_i \psi_i(S_i) \prod_{(ij) \in E} \psi_{ij}(S_i, S_j) \right]. \quad (57)$$

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

$$Z_{i \rightarrow j}(S_i) = \psi_i(S_i) \prod_{k \in \partial i \setminus j} \left( \sum_{S_k} Z_{k \rightarrow i}(S_k) \psi_{ik}(S_i, S_k) \right), \quad (58)$$

$$Z_i(S_i) = \psi_i(S_i) \prod_{j \in \partial i} \left( \sum_{S_j} Z_{j \rightarrow i}(S_j) \psi_{ij}(S_i, S_j) \right), \quad (59)$$

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

In practice, however, it turns out that working with partition functions (i.e. numbers that can be exponentially large in the system size) is somehow impractical, and we can thus normalize Equation (58) and rewrite these recursions in terms of probabilities. Denoting  $\eta_{i \rightarrow j}(S_i)$  as the marginal probability distribution of the variable  $S_i$  when the edge  $(ij)$  has been removed, we have

$$\eta_{i \rightarrow j}(S_i) = \frac{Z_{i \rightarrow j}(S_i)}{\sum_{S'} Z_{i \rightarrow j}(S')}, \quad \eta_i(S_i) = \frac{Z_i(S_i)}{\sum_{S'} Z_i(S')}. \quad (60)$$

So that the recursions (58)–(59) now read

$$\eta_{i \rightarrow j}(S_i) = \frac{\psi_i(S_i)}{z_{i \rightarrow j}} \prod_{k \in \partial i \setminus j} \left( \sum_{S_k} \eta_{k \rightarrow i}(S_k) \psi_{ik}(S_i, S_k) \right), \quad (61)$$

$$\eta_i(S_i) = \frac{\psi_i(S_i)}{z_i} \prod_{j \in \partial i} \left( \sum_{S_j} \eta_{j \rightarrow i}(S_j) \psi_{ij}(S_i, S_j) \right), \quad (62)$$

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

$$z_{i \rightarrow j} = \sum_{S_i} \psi_i(S_i) \prod_{k \in \partial i \setminus j} \left( \sum_{S_k} \eta_{k \rightarrow i}(S_k) \psi_{ik}(S_i, S_k) \right), \quad (63)$$

$$z_i = \sum_{S_i} \psi_i(S_i) \prod_{j \in \partial i} \left( \sum_{S_j} \eta_{j \rightarrow i}(S_j) \psi_{ij}(S_i, S_j) \right). \quad (64)$$

The iterative equations (61) and (62), and their normalization equations (63) and (64), are called the BP equations. Indeed, since  $\eta_{i \rightarrow j}(S_i)$  is the distribution of the variable  $S_i$  when the edge to variable  $j$  is absent, it is convenient to interpret it as the “belief” of the probability of  $S_i$  in the absence of  $j$ . It is also called a “cavity” probability since it is derived by removing one node from the graph. The BP equations are used to define the BP algorithm:

- Initialize the cavity messages (or “beliefs”)  $\eta_{i \rightarrow j}(S_i)$  randomly or following a prior information  $\psi_i(S_i)$  if we have one.
- Update the messages in a random order following the BP recursion equations (61) and (63) until their convergence to their fixed point.
- After convergence, use the beliefs to compute the complete marginal probability distribution  $\eta_i(S_i)$  for each variable. This is the BP estimate on the marginal probability distribution for variable  $i$ .

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

At this point, since we have switched from partial partition sums to partial marginals, the astute reader could complain that it seems that we have lost out prime objective: the computation of the partition function. Fortunately, one can compute it from the knowledge of the marginal

distributions. To do so, it is first useful to define the following quantity for every edge  $(ij)$ :

$$z_{ij} = \sum_{S_i, S_j} \eta_{j \rightarrow i}(S_j) \eta_{i \rightarrow j}(S_i) \psi_{ij}(S_i, S_j) = \frac{z_j}{z_{j \rightarrow i}} = \frac{z_i}{z_{i \rightarrow j}}, \quad (65)$$

where the last two equalities are obtained by plugging Equation (61) into the first equality and realizing that it almost gives Equation (64). Using again Equations (61)–(64), we obtain

$$\begin{aligned} z_i &= \sum_{S_i} \psi_i(S_i) \prod_{j \in \partial i} \left( \sum_{S_j} \eta_{j \rightarrow i}(S_j) \psi_{ij}(S_i, S_j) \right) \\ &= \sum_{S_i} \psi_i(S_i) \prod_{j \in \partial i} \left( \sum_{S_j} \frac{Z_{j \rightarrow i}(S_j)}{\sum_{S'} Z_{j \rightarrow i}(S')} \psi_{ij}(S_i, S_j) \right) = \frac{\sum_{S_i} Z_i(S_i)}{\prod_{j \in \partial i} \sum_{S_j} Z_{j \rightarrow i}(S_j)}, \end{aligned} \quad (66)$$

and along the same steps

$$z_{j \rightarrow i} = \frac{\sum_{S_j} Z_{j \rightarrow i}(S_j)}{\prod_{k \in \partial j \setminus i} \sum_{S_k} Z_{k \rightarrow j}(S_k)}. \quad (67)$$

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

$$Z = \sum_{S_i} Z_i(S_i) = z_i \prod_{j \in \partial i} \left( \sum_{S_j} Z_{j \rightarrow i}(S_j) \right) = z_i \prod_{j \in \partial i} \left( z_{j \rightarrow i} \prod_{k \in \partial j \setminus i} \sum_{S_k} Z_{k \rightarrow j}(S_k) \right), \quad (68)$$

and we continue to iterate this relation until we reach the leaves of the tree. Using Equation (65), we obtain

$$Z = z_i \prod_{j \in \partial i} \left( z_{j \rightarrow i} \prod_{k \in \partial j \setminus i} z_{k \rightarrow j} \cdots \right) = z_i \prod_{j \in \partial i} \left( \frac{z_j}{z_{ij}} \prod_{k \in \partial j \setminus i} \frac{z_k}{z_{jk}} \cdots \right) = \frac{\prod_i z_i}{\prod_{(ij)} z_{ij}}. \quad (69)$$

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

$$\begin{aligned} fN &= -T \log Z = \sum_i f_i - \sum_{(ij)} f_{ij}, \\ f_i &= -T \log z_i, \quad f_{ij} = -T \log z_{ij}, \end{aligned} \quad (70)$$

where  $f_i$  is a “site term” coming from the normalization of the marginal distribution of site  $i$ , and is related to the change in  $Z$  when the site  $i$  (and the corresponding edges) is added to the system.  $f_{ij}$  is an “edge” term that can be interpreted as the change in  $Z$  when the edge  $(ij)$  is added. This provides a convenient interpretation of the Bethe free energy equation (70): it is the sum of the free energy  $f_i$  for all sites but, since we have counted each edge twice we correct this by subtracting  $f_{ij}$  (see for instance [22]).

We have now entirely solved the problem on a tree. There is, however, nothing that prevents us from applying the same strategy on any graph. Indeed the algorithm we have described is well defined on any graph, but we are not assured that it gives exact results nor that it will converge.

Using these equations on graphs with loops is sometimes referred to as “loopy BP” in Bayesian inference literature [52].

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

Clearly, this must be a good approach for describing a system in a paramagnetic phase, or even a system with a ferromagnetic transition (where we should expect to have two different fixed points of the iterations). It could be, however, that there exists a huge number of fixed points for these equations: How to deal with this situation? Should they all correspond to a given pure state? Fortunately, we do not have such worries, as the situation we just described is the one arising when there is a glass transition. In this case, one needs to use the cavity method in conjunction with the so-called *RSB* approach as was done by Mézard and Parisi [21,22,72]. In this review, however, we discuss optimal Bayesian inference and we are guaranteed from the Nishimori conditions discussed in Section 2.7 that no such RSB occurs. We can assume safely that in the Bayes-optimal setting the simple BP approach is always providing the correct description (unless there are other sources of correlations in the system, e.g. the graph is not random but a lattice or the quenched disorder is correlated in a non-local manner).

### 3.3. Properties of the randomly quenched ensemble

The first system on which we will illustrate how to use the cavity method and the related BP to compute the phase diagram is the Viana–Bray (V–B) model for the so-called diluted SG [114]. The reason we discuss first the solution for this well-known model is that, using the theory developed in Section 2, we will be able to read off the solution for the planted SG model from the phase diagram of the randomly quenched model, see Figure 2. Viana and Bray studied the Ising SG Hamiltonian (30) with the lattice being an ER random graph with average degree  $c$ . Interactions  $J_{ij}$  are chosen from  $\pm 1$  independently and uniformly at random and, for simplicity, we consider the absence of a magnetic field. We call the ensemble of instances generated this way the randomly quenched ensemble.

The description of the V–B model within this approach was originally done by Bowman and Levin [121]. Instead of working with the iteration (61), it is convenient to rewrite it in terms of an “effective” local cavity fields  $u_{i \rightarrow j}$  defined as

$$\eta_{i \rightarrow j}(S_i) = \frac{e^{u_{i \rightarrow j} S_i}}{2 \cosh e^{u_{i \rightarrow j}}}. \quad (71)$$

With a bit of algebra, the BP Equation (61) can be written as

$$u_t^{i \rightarrow j} = f(J_{ij}, \{u_{t-1}^{k \rightarrow i}\}_{k \in \partial i \setminus j}) = \frac{1}{\beta} \operatorname{atanh} \left[ \tanh(\beta J_{ij}) \tanh \left( \beta \sum_{k \in \partial i \setminus j} u_{t-1}^{k \rightarrow i} \right) \right], \quad (72)$$

where we have added an index  $t$  that stands for a “time” iteration. As stated, iterating this procedure allows to actually solve the problem on a tree, which is at the root of the cavity method. Here, however, we apply this procedure on the graph, iteratively, hence the time indices in Equation (72). The local magnetization is then computed from the BP fixed point using the

hyperbolic tangent of the local field acting on the spin, and thus reads

$$m_i = \tanh \left( \beta \sum_{j \in \partial i} u^{j \rightarrow i} \right). \quad (73)$$

Applying the same notation to the Bethe free energy per site Equation (70) leads to

$$\beta N f = \sum_i (c_i - 1) \log[2 \cosh(\beta h^i)] - \sum_{(ij) \in E} \log \left[ \sum_{S=\pm 1} 2 e^{S \beta J_{ij}} \cosh(\beta h^{i \rightarrow j} + S \beta h^{j \rightarrow i}) \right], \quad (74)$$

where  $h^i = \sum_{k \in \partial i} u^{k \rightarrow i}$  and  $h^{i \rightarrow j} = \sum_{k \in \partial i \setminus j} u^{k \rightarrow i}$ , and  $c_i$  is the degree of node  $i$ . Physical properties of the system are computed from the fixed points of the BP equations (72). At this point, there are two different strategies to solve the problem. The first one is to investigate numerically what is happening on a given realization of the problem, that is to generate a large random graph and to actually run the BP iteration on it. This is actually very instructive (as we shall see) but no matter how large our graph will be, it will not be at the thermodynamic limit.

The second strategy is to work on an infinite graph directly via a so-called *population dynamics* [21]. We attempt to describe the model by the knowledge of the probability distribution  $\mathcal{P}(u)$  of the messages  $u$  on an edge with value  $J$  over the (infinite) graph. Consider for instance the case of a random graph with fixed degree  $c_i = c$ . Clearly, within a pure Gibbs states,  $\mathcal{P}(u)$  satisfies the BP equations (72) so that

$$\mathcal{P}(u) = \frac{1}{\mathcal{Z}} \int P(J) dJ \delta[u - f(J, \{u_d\}_{d=1, \dots, c-1})] \prod_{d=1}^{c-1} \mathcal{P}(u_d) du_d, \quad (75)$$

where  $f(\cdot)$  is given by Equation (72). While this appears to be a difficult equation to solve explicitly, it can be conveniently approximated to arbitrary precision with the so-called population dynamics method [21]. In this case, one models the distribution  $\mathcal{P}(u)$  by a population of  $\mathcal{M}$  values

$$\mathcal{P}_{\text{pop}}(u) = \frac{1}{\mathcal{M}} \sum_{l=1}^{\mathcal{M}} \delta(u - u_l). \quad (76)$$

Plugging this expression in Equation (75) leads to the population dynamics algorithm:

- Initialize all values  $u_l$  randomly.
- For a given element  $l$  of the population, draw  $c - 1$  values of  $u$  from the population, and a value  $J$  of the coupling from  $P(J)$ , compute a new value  $u_{\text{new}} = f(J, \{u\})$ , and replace the element  $u_l$  by  $u_{\text{new}}$ .
- Repeat the last step until  $\mathcal{P}_{\text{pop}}(u)$  has converged to a good approximation of  $\mathcal{P}(u)$ .

This method is very simple and, given the population dynamics distribution, allows to compute the value of the free energy per site via Equation (70).

Either way, depending on the inverse temperature  $\beta$  in the VB model, one observes one of the two following possibilities: at low  $\beta$  there is a stable *paramagnetic* fixed point  $u^{i \rightarrow j} = 0$  for all  $(ij) \in E$ . The corresponding paramagnetic free energy reads

$$- \beta f(\beta, \mathbf{J}) = \log 2 + \frac{c}{2} [\log(2 \cosh \beta) - \log 2]. \quad (77)$$

We do notice that the same expression is obtained for the annealed free energy for this model.

At high  $\beta$  (low temperature) the iterations of Equation (72) do not converge on a single graph. The last statement, of course, cannot be checked simply when performing the population dynamics algorithm. A similar effect can, however, be observed by performing two-population dynamics with the same random numbers starting from infinitesimally close initial populations. At high  $\beta$  the two populations will diverge to entirely different values [122]. This is a clear sign of long-range correlations and of the appearance of a SG phase. We are, of course, interested to know the precise location of the phase transition.

Thouless [66] analyzed the threshold value of  $\beta_c$  by linearizing the update (72) around the uniform fixed point, thereby obtaining

$$u_t^{i \rightarrow j} = \tanh(\beta J_{ij}) \sum_{k \in \partial i \setminus j} u_{t-1}^{k \rightarrow i}. \quad (78)$$

Thouless then argues that averaged over the disorder  $J_{ij}$  the above equation gives zero, and when its square is averaged we get

$$\langle u_t^2 \rangle = c \tanh^2 \beta \langle u_{t-1}^2 \rangle, \quad (79)$$

where we remind that  $c$  is the average degree of the ER graph, that is also equal to the average branching factor of the tree that approximates locally the random graph. We thus expect that the paramagnetic phase becomes unstable when  $c \tanh^2 \beta \geq 1$ .

In fact, by analyzing such perturbations, one is investigating the behavior of the susceptibility. Indeed the SG susceptibility can be written as

$$\chi_{\text{SG}} \equiv \beta \sum_i \langle S_0 S_i \rangle_c^2 \approx \beta \sum_{r=0}^{\infty} c^r \overline{\langle S_0 S_r \rangle_c^2}. \quad (80)$$

In Equation (80),  $\overline{\cdots}$  represents a spatial average over the whole graph,  $\langle \cdots \rangle$  represents a thermal average,  $S_r$  are all spins at distance  $r$  from  $S_0$ , and  $X_c$  is the connected version of the correlation function  $X$ :  $\langle S_0 S_i \rangle_c = \langle S_0 S_i \rangle - \langle S_0 \rangle \langle S_i \rangle$ . The correlation can be computed using derivatives with respect to an external magnetic field on the spin  $S_0$

$$\beta \overline{\langle S_0 S_i \rangle_c^2} = \overline{\left( \frac{\partial \langle S_i \rangle}{\partial \langle h_0 \rangle} \right)^2}. \quad (81)$$

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

$$\frac{\partial \langle S_r \rangle}{\partial \langle h_0 \rangle} = \frac{\partial \langle S_r \rangle}{\partial \langle h_r \rangle} \frac{\partial \langle h_r \rangle}{\partial \langle h_{r-1} \rangle} \frac{\partial \langle h_{r-1} \rangle}{\partial \langle h_{r-2} \rangle} \cdots \frac{\partial \langle h_1 \rangle}{\partial \langle h_0 \rangle}, \quad (82)$$

where we indicate with  $h_i$  the cavity field going from spin  $i$  to spin  $i+1$  for simplicity and  $h_r, h_{r-1}, \dots, h_0$  is the sequence of cavity fields on the shortest path from  $S_0$  to  $S_r$  (we are sure that there exists just one shortest path because we are locally on a tree). To check whether or not the susceptibility defined in Equation (80) diverges, one should thus simply analyze the square of the derivative of the BP equations, as we have done in Equation (79). The conclusion is that the



critical temperature is given by the condition that a perturbation would grow, so that

$$\beta_c^{\text{ER}} = \text{atanh}\left(\frac{1}{\sqrt{c}}\right). \quad (83)$$

Given the branching ratio is  $c - 1$  for a random regular graph, we get in this case:

$$\beta_c^{\text{reg}} = \text{atanh}\left(\frac{1}{\sqrt{c-1}}\right). \quad (84)$$

It follows that for  $\beta < \beta_c$  the state of the V-B model is correctly described by the paramagnetic fixed point. For  $\beta > \beta_c$  the Bethe approximation fails and for the correct physical description the so-called RSB framework needs to be applied [21]. Since, in this review, we are concerned by the Bayes-optimal inference that effectively lives on the Nishimori line and does not need the concept of replica symmetric breaking, we refer the reader interested in this to the textbook [5].

### 3.4. Phase diagram of the planted SG

Now we compute the phase diagram of the planted SG, where the interactions  $\mathbf{J}$  are chosen conditionally on the planted configuration  $\mathbf{S}^*$ , following Equation (34). In this case the Bethe approximation and its linearization are still described by Equations (72) and (78). In the planted model the stability analysis of the paramagnetic fixed point is slightly different, since the distribution from which the interactions  $J_{ij}$  are taken is not random but depends on the planted configuration. This is where our mapping to the Nishimori ensemble is making the problem rather easy to analyze anyway. Indeed, according to our analysis in the last section, we simply need to ask whenever, on the Nishimori line with a bias  $\rho = e^{\beta^*}/2 \cosh \beta^*$ , the system acquires a spontaneous magnetization. Or more simply, we just need to compute the phase diagram when a ferromagnetic bias is present. Taking the average of Equation (78) and counting properly the probability of  $J_{ij}$  given  $S_i^*$  and  $S_j^*$  we immediately see that, while the SG transition is unchanged, the ferromagnetic one is now given by

$$\langle u_i \rangle = c(\tanh \beta^*)(\tanh \beta) \langle u_{i-1} \rangle. \quad (85)$$

According to this computation, the system thus starts to polarize toward the ferromagnetic configuration when

$$\beta > \beta_F = \text{atanh}\left(\frac{1}{c \tanh \beta^*}\right). \quad (86)$$

For the Bayes-optimal case (or equivalently, on the Nishimori line) when  $\beta^* = \beta$ , this immediately gives the transition from the no-detectability phase to the recovery phase that we gave in Equation (43). We see now how easy it was to compute it using the cavity method and the Nishimori mapping.

A couple of subtle but interesting points can be made about the phase diagram, always keeping in mind the fact that the inference problem can be seen as looking for the ferromagnetic state on the Nishimori line.

- As  $\beta^* \rightarrow \infty$  ( $\rho = 1$ ) we recover the ordinary ferromagnetic Ising model with its well-known critical temperature.
- It is interesting to observe that there is a multi-critical point at  $\beta = \text{atanh}(1/\sqrt{c})$ , where all the three conditions (83), (86) and Bayes-optimality,  $\beta = \beta^*$ , meet. This point, of course, is on the Nishimori line, and is precisely the transition point (43).

- For the Bayes-optimal case,  $\beta^* = \beta$ , for all temperature higher than the transition point  $\beta^* < \beta_c$ , the paramagnetic fixed point is the *only fixed point* of the recursion (72). Since the paramagnetic and annealed free energies are equal, this is a case where the planting is quiet, as defined in Section 2.5.3. This means that in the whole region  $\beta^* < \beta_c$  the planted ensemble behaves exactly the same as the randomly quenched ensemble. This is indeed reflected in the phase diagram in Figure 2, where everything on the left of the multi-critical point is independent of the value of  $\rho$ .
- Out of the Bayes-optimal case, i.e. when  $\beta^* \neq \beta$ , there exists at very low temperature (large  $\beta$ ) a so-called mixed phase where equilibrium configurations are correlated to the planted one (the ferromagnetic after our mapping) but where the equilibrium phase is actually a SG one.

The phase diagram of the planted SG is summarized in Figure 2 where the boundary between the paramagnet (P) and the SG is plotted in green, between the paramagnet (P) and ferromagnet (F) in blue, and the Nishimori line corresponding to the Bayes-optimal inference in red. We also sketch the position of the mixed phase (M).

### 3.5. Systems with first-order phase transitions

In the above planted Ising SG model the underlying phase transition is of second order. In that case there is no particular computational issue. Either inference is impossible, or it is possible and tractable. The hard phase outlined in Section 1.2.5 is absent.

The picture is richer in problems when a discontinuous phase transition is found in the randomly quenched ensemble. Famously, this is the case in the low-density parity check (LDPC) error correcting codes as described, for example in [3,5], and has been analyzed also in details in the pioneering work of Nishimori and Wong [123]. The examples of clustering of networks and compressed sensing (to which we dedicate Sections 5 and 6) also fall in this category.

In order to observe such a phenomenon, we have to change the Hamiltonian equation (30) and move from two-body interactions to three-body interactions. This leads to the so-called three-spin SG model, which we already mentioned in Section 2.8.2 under the name 3-XOR-SAT [93–95]. The Hamiltonian now reads

$$\mathcal{H}(\mathbf{S}) = - \sum_{(ijk) \in E} J_{ijk} S_i S_j S_k. \quad (87)$$

Again, we can create this model on a random hypergraph. For instance, we can consider a  $L$  regular hypergraph where each variable belongs to exactly  $L$  triplets, chosen randomly at random. Again we can study this model under a planted setting, or equivalently, on the Nishimori line. What is changing with respect to the preceding section? First of all, we need to rewrite the BP equation. It is an exercise (see for instance [94,95] or Appendix A in [97]) to see that they now read

$$m^{i \rightarrow a} = \tanh \left( \beta \sum_{b \in \partial i \setminus a} J_b \prod_{j \in \partial b \setminus i} m^{j \rightarrow b} \right), \quad (88)$$

where  $a, b$  denote the triplets of interactions. Repeating the analysis of the previous section, one can obtain the phase diagram which we show in Figure 5.

At first sight, this phase diagram looks really similar to the one from Figure 2. What has changed now is the fact that the transition is a first-order one. Let us first concentrate on the Nishimori line. As soon as  $T < T_{\text{spin}}$  there are two fixed points of BP, a paramagnetic and a ferromagnetic one, and the one dominating the measure is the one with lower free energy. Only

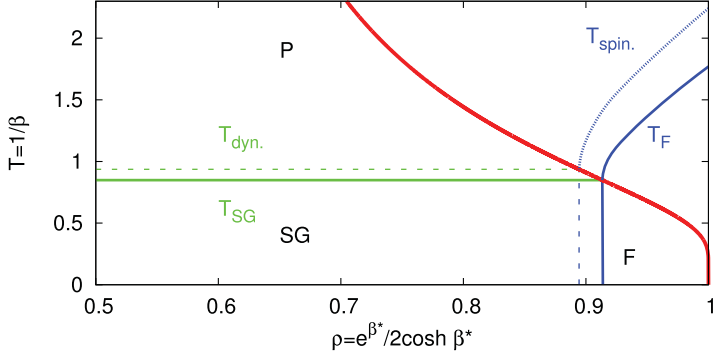


Figure 5. Phase diagram of the 3-XOR-SAT (or three-SG) problem as a function of temperature  $T = 1/\beta$  and density of ferromagnetic bonds  $\rho$ . This phase diagram is for random regular hypergraphs of degree  $L=5$ . The Nishimori line,  $\rho = e^{\beta^*}/(2 \cosh \beta)$ , corresponds to the Bayes-optimal inference of a planted configuration. As in Figure 2 there is the SG phase where the system is in a glassy state, the ferromagnetic (F) phase, and the paramagnetic (P) phase (for simplicity, the mixed phase is not shown). A major difference with respect to Figure 2, however, is the fact that the ferromagnetic transition is now a first-order one, with a spinodal line. The SG transition is also a first-order one, with the so-called dynamic transition happening *before* the static transition.

when  $T < T_F$  does the free energy of the ferromagnetic fixed point become lower than the paramagnetic one. This means that, for  $T < T_F$ , a perfect sampling would give a configuration with a positive magnetization. Using the inference-Nishimori connection, this means that inference is possible for  $T < T_F$ , and impossible for  $T > T_F$ : the equilibrium phase transition marks the information theoretic transition.

Now, let us try to actually perform inference in practice, for instance by running BP or a Monte Carlo algorithm. Unfortunately, we have no idea about what is the planted configuration so at best we can initialize randomly. The problem is that this will put us in the paramagnetic state, which exists and is locally stable at all positive temperatures. Both BP or MCMC will simply explore this paramagnetic state, and, in any sub-exponential time, will miss the existence of the ferromagnetic one. This is the classical picture of a first-order phase transition in a mean-field model. Because of the first-order transition, we are trapped into a high free energy minima that is stable at all temperatures, and therefore, we cannot perform inference. While we *should* be able to find a configuration correlated with the planted one (in the sense that it is information-theoretically possible), we cannot do it within the class of existing polynomial algorithms. Related ideas, in the context of the glass transition, were originally present in [124]. This illustrates the difference between a “possible” inference and an “easy” one in the sense of Section 1.2.5.

XOR-SAT is actually a little bit special in the sense that the easy detectable phase (see below) is missing. There is a spinodal transition ( $T_{\text{spin}}$ ) where the ferromagnetic phase becomes unstable, but not a second one where the paramagnet become unstable. In most models, however, if the temperature is lowered a second spinodal is reached, the paramagnetic phase becomes eventually unstable and inference is possible.

Let us thus discuss another example where such a transition arises. For this, we turn to the planted coloring problem [31]. In planted coloring one first chooses a random assignment of  $q$  colors out of the  $q^N$  possible ones. Then, one puts  $M = cN/2$  edges at random among all the edges that do not have the same color on their two adjacent nodes. The Bayes-optimal inference in this setting evaluates marginals over the set of all valid colorings of the created graph. Computation of these marginals builds on previous works [29,30], where the problem of coloring

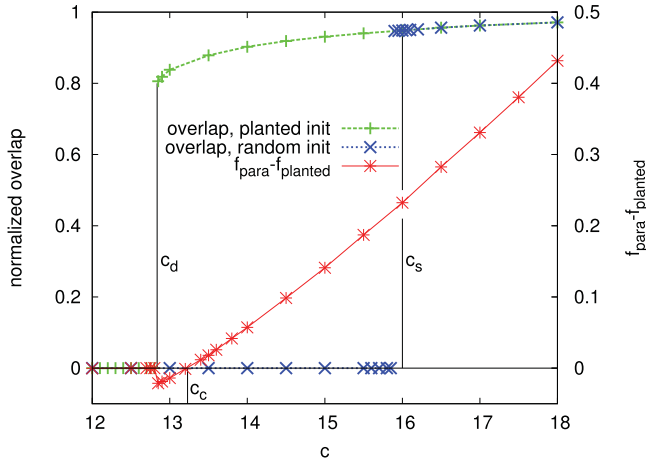


Figure 6. The phase transitions for planted five-coloring. The overlap (normalized in such a way that for a random configuration it is zero) achieved by BP initialized in the random and planted configuration. The difference between the corresponding free energies is also plotted. Figure taken from [125].

random graphs was studied, and goes as follows: one considers BP initialized in two different ways. In the random initialization the messages are chosen at random but close to the paramagnetic ones. In the planted initialization the initial values of the messages are set to point in the direction of the planted configurations. A fixed point is reached from both these initializations and its free energy is evaluated. One obtains the following picture, illustrated in Figures 3 and 6 for planted five-coloring:

- Undetectable paramagnetic phase, for average degree  $c < c_d$ . The planting is quiet and BP converges to the uniform fixed point both from random and planted initialization. We will call  $c_d$  the *dynamical* or the *clustering* phase transition [31,126].
- Undetectable clustered phase, or d1RSB phase, for average degree  $c_d < c < c_c$ . The planting is still quiet, but the set of solutions is split into exponentially many clusters that are uncorrelated and all look the same as the one containing the planted solutions. We will refer to  $c_c$  as the *detectability* phase transition. In the randomly quenched ensemble  $c_c$  is the condensation transition [31,126].
- Hard detectable phase, for average degree  $c_c < c < c_s$ . This is arguably the most peculiar phase of the four. The Bayes-optimal marginals are correlated to the planted configuration, however, there is a metastable paramagnetic phase to which BP converges from a random initialization. A ghost of the quiet planting is visible even for  $c > c_c$  in the sense that the space of configurations looks exactly like the one of the randomly quenched ensemble except for the planted cluster.
- Easy detectable phase, for average degree  $c > c_s$ . BP finds a configuration correlated to the planted coloring. The phase transition  $c_s$  can be located doing a local stability analysis of the uniform BP fixed point, along the same lines as done for the planted SG in Section 3.4. We will refer to  $c_s$  as the *spinodal* or the *hard-easy* phase transition. In the randomly quenched ensemble,  $c_s$  is the point starting from which BP stops converging, related to the local instability toward the SG phase [30,31].

A careful reader will recognize that this corresponds to the situation we have described in the introduction when we draw Figure 1. The situation in this figure is that of a problem with a first-order transition, where the possible inference is defined by the equilibrium transition, and the

easy/hard transition is defined by the presence of a spinodal marking the instability of the paramagnet toward the planted state. This is arising in learning problems as diverse as the planted perceptron, community detection with four or more communities, compressed sensing, error correction, etc. If, however, the transition is a second order one, then these two transition coincides as in the case of the planted SG (or, as we shall see, community detection with two equal sizes communities). We find it remarkable that the century old classification of phase transition in physics, introduced by Paul Ehrenfest in 1933 [127] turns out to be the right tool to discuss a deep phenomena of statistical inference and computational complexity.

### 3.6. Comments on the phase transitions in various problems

One more question we want to clarify is whether the concept of *quiet planting* works for every inference problem. Here the answer is both yes and no. First of all, for quiet planting to work, we need to be in a setting where BP has the uniform (sometimes confusingly called factorized) fixed point, where the messages in the fixed point do not depend on the indices of the edge. If such a fixed point exists then the annealed free energy density equals the quenched one [128], and consequently the argument about planting being quiet works. There are many problems where BP on the randomly quenched ensemble does not have a uniform fixed point, the most studied examples are perhaps the problem of  $K$ -satisfiability ( $K$ -SAT), or the Ising model in a random external magnetic field  $h_i$ .

Even in problems such as  $K$ -SAT, where canonical BP does not have a uniform fixed point [129], we can introduce a reweighted Boltzmann measure, as done in [130], in such a way that the corresponding BP fixed point is uniform. Planting according to this reweighted measure then leads to instances that are contiguous to the random ones. Note that this reweighting and corresponding planted instances were used in the proof of [131,132] and studied also in [133,134]. The planted solution is not an equilibrium solution with respect to the canonical Boltzmann measure but with respect to the reweighted measure.

Another comment is related to the following natural question: How to distinguish between problems where the transition is continuous or discontinuous? And when it is discontinuous, how wide is the hard phase? Is there some criteria simpler than performing the whole analysis? To distinguish between the first- and second-order transition, we do not know of any other way. We only have a couple of rules of thumb – in Potts model with sufficient number of states continuous transitions become discontinuous, and in models where more than two variables interact, e.g. the  $p$ -spin model or  $K$ -SAT, the continuous transition becomes discontinuous for sufficiently large  $K$  and  $p$  [126]. But there are binary pairwise problems such as the independent set problem where at sufficiently large average degree the continuous transition also becomes discontinuous [135].

Interestingly, one can identify a class of CSP where the hard phase is very wide. In [130] these problems were identified as those with  $c_s \rightarrow \infty$ . Previously, only XOR-SAT was known to have this property [124], but in [136] a whole hierarchy of such problems was identified and is related to  $k$ -transitivity of the constraints. A constraint is  $k$ -transitive if and only if the number of satisfying assignments compatible with any given assignment of any  $k$  variables is the same. For all CSPs that are at least two-transitive the hard phase extends over the whole region of linearly many constraints, equivalently  $c_s \rightarrow \infty$ . Calling  $r$  the smallest number such that a constraint is not  $r$ -transitive, the number of constraints predicted to render the problem tractable scales as  $N^{r/2}$  [136]. Interestingly, this classification makes the  $K$ -XOR-SAT the hardest of all CSPs (despite the fact that it can be solved by Gaussian elimination, which is not robust to any kind of noise). The notion of two-transitivity, and consequently infinitely large hard phase, extends also to continuous variables, and as it was recently pointed out, renders tensor factorization and completion very difficult [137]. Physically, the corresponding systems are such when the paramagnetic phase is

locally stable down to zero temperature [124]. Examples of such systems indeed include the Ising  $p$ -spin (related to  $K$ -XOR-SAT) or the spherical  $p$ -spin model related to tensor factorization.

Having examples where the hard phase is huge is very useful when thinking about formal reasons of why this phase is hard. In [136] the authors proved hardness for a class of the so-called *statistical algorithms* using precisely this kind of CSP. Arguably, the class of statistical algorithms is too restrictive, but we believe that the instances of CSP with such a large hard phase should be instrumental for other attempts in this direction. We shall summarize in Section 4.2 arguments for why we think there is something fundamentally hard about this phase.

In [138] we discussed the example of planted locked CSP. These are particularly interesting for a couple of reasons. First of all, the valid solutions are separated by extensive Hamming distance from each other and therefore BP either converges to a paramagnet or points exactly toward the planted configuration. Note that LDPC error correcting codes have precisely this property. Locked CSP can be thought of as non-linear generalization of LDPC. A second reason why locked CSPs are interesting is that in some cases the hard phase is very wide and moreover the planted assignment is with high probability the only one (perhaps up to a global flip symmetry). The planted locked CSP are hence attractive candidates for a one-way function that could find applications in cryptography. Having hard CSP instances that have exactly one valid assignment was used in studies of quantum annealing where having non-degenerate ground state is greatly advantageous, see, for example [139].

#### 4. From physics insight to new algorithmic ideas

In the introduction, Section 1.2, we presented two main questions about inference that we aim to answer. One question is more fundamental, about phase diagrams, phase transitions and related insights. The second question is more practical, about the development of algorithms. This section is about algorithmic contributions obtained using the insight resulting from the statistical physics analysis. For clarity, all is again illustrated on the planted model.

With the work of Mézard et al. on survey propagation [72], the statistical physics community learned that the replica and cavity calculations are as powerful when used as algorithms as they are when used as tools for analyzing the phase diagram. The connection between the Bethe approximation and BP was noticed earlier in connection to error correcting codes by Kabashima and Saad [118], and in a more general setting by Yedidia et al. [52]. But looking at the volume of related literature, it is only after survey propagation that the whole statistical physics community started to look for explicit algorithmic applications of the Bethe approximation.

Here we review some recent physics-related contributions to algorithmic development for Bayes-optimal inference.

##### 4.1. Non-backtracking spectral method

Notice, within the historical perspective, that the Bethe solution for the SG on a random graph was written by Bowman and Levin [121], i.e. about 16 years before realizing that the same equation is called BP and that it can be used as an iterative algorithm in a number of applications.

The paper by Thouless [66], who analyzed the SG critical temperature, implicitly includes another very interesting algorithm, that can be summarized by Equation (78) where BP is linearized. The fact that this equation has far-reaching implications when viewed as a basis for a spectral algorithm waited for its discovery even longer. Notably until 2013 when the methodologically same algorithm was suggested for clustering of networks [140]. This work will be discussed in Section 5.4. We now turn back to our pedagogical example, the  $\pm J$  planted SG, and related spectral algorithm as described and analyzed in [57]. Related ideas also appeared in [141].



First note that a large class of inference algorithms is based on a spectral method of some kind. Probably the most widely used is the so-called principal component analysis (PCA). Its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. In PCA the information is contained in singular vectors (or eigenvectors for symmetric matrices) corresponding to the largest singular values. Resulting spectral algorithms are used everyday to solve numerous real-world problems. The same strategy of computing leading singular (or eigen-) vectors of some matrix related to the data is applied in many settings including the planted partitioning problem closely related to the planted SG model [142,143], or in understanding financial data [144,145].

A natural question that arises is whether a spectral algorithm can find a configuration correlated to the planted assignment in the planted SG in the whole region  $\beta^* > \beta_c$  where BP can asymptotically do so. Facing this question, many experts would probably suggest to take the matrix  $J_{ij}$  (with  $J_{ij} = 0$  when  $(ij) \notin E$ ) or its associated Laplacian (perhaps normalized in some way) and compute the leading eigenvector. This is indeed a good strategy, but on sparse ER random graphs it does not work down to the threshold  $\beta_c$ . The problem is that for an Ising model living on a large sparse ER graph the tail of the spectra of the commonly considered matrices is dominated by eigenvectors localized on some small subgraphs (e.g. around high-degree nodes or on some kind of hanging sub-trees).

The strategy in designing spectral method that does not have the problem with localized eigenvectors (at least on a much larger class of graphs) is to try to mimic what BP is doing when departing from the paramagnetic fixed point. Equation (78) is precisely the linearization of BP around that fixed point and suggests a definition of a so-called *non-backtracking matrix*  $B$  defined on directed edges as

$$B_{i \rightarrow j, k \rightarrow l} = J_{ij} \delta_{il} (1 - \delta_{jk}). \quad (89)$$

Denoting  $M$  the number of edges in the non-directed graph, this is a  $2M \times 2M$  matrix. The term  $(1 - \delta_{jk})$  means that a walk that follows non-zero elements of the matrix is not going back on its steps, this motivates the name of the matrix. Building on [140,146], it was proven in [57] that the planted configuration is indeed correlated with the signs of elements of the eigenvector of  $B$  corresponding to the largest (in module) eigenvalue as soon as  $\beta^* > \beta_c$ .

Figure 7 depicts the spectrum on the matrix  $B$  (89) for a planted SG model on an ER graph of average degree  $c$  and having  $N = 2000$  nodes. The  $J_{ij} = \pm 1$  were generated following the procedure described in Section 2.1 with  $\beta^* = 0.55$ . We observe (the proof is found in [57]) that on sparse random graphs

- For  $\beta^* < \beta_c = \operatorname{atanh}(1/\sqrt{c})$ , corresponding to the paramagnetic phase of the planted SG model: the spectrum of  $B$  is the same as it would be for  $\beta^* = 0$ , bounded by a circle of radius  $\sqrt{c}$ .
- For  $\beta^* > \beta_c = \operatorname{atanh}(1/\sqrt{c})$ , corresponding to the ferromagnetic phase of the planted SG model: the spectrum of  $B$  has the same property as in the previous case, except for two eigenvalues on the real axes. One of them is out of the circle and has value  $c \tanh \beta^*$ . Its corresponding eigenvector is positively correlated with the planted assignment.

The performance of the non-backtracking spectral algorithms will be further discussed and compared to other algorithms for the example of clustering of networks in Section 5.4. In Section 5.7 we will also illustrate how this method performs on some non-random graphs. Just as all the other spectral algorithms, this method is not robust to adversarial changes in a small part of the graph (such as planting a small dense sub-graph). This can hence be a disadvantage with respect to methods that optimize some global cost function that is not affected by small local

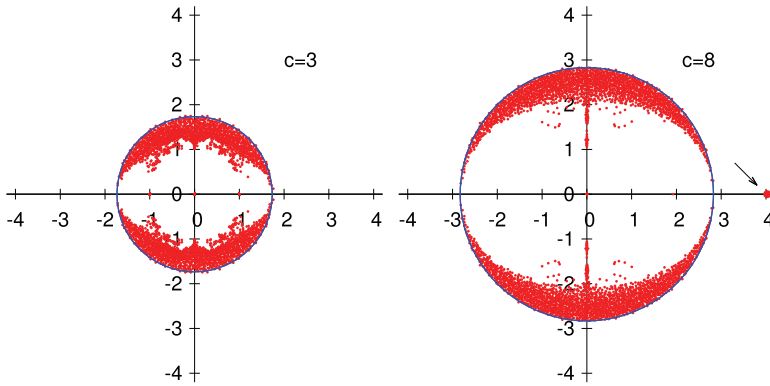


Figure 7. Spectrum of the non-backtracking operator for the  $\pm J$  planted SG model. Right: In the undetectable phase where all eigenvalues are inside a circle of radius  $\sqrt{c}$ . Left: In the detectable phase where one eigenvalue is outside the circle on the real axis, its value is  $c \tanh \beta^*$ . Figure taken from [57].

changes. We, however, illustrate in Figure 13 that spectrum of the non-backtracking matrix of many real graphs does not deviate excessively from its behavior on random graphs.

#### 4.2. On computational hardness of the hard phase

In Section 3.5 we described the four phases that we observe in the Bayes-optimal setting of inference problems. The most theoretically intriguing is the hard detectable phase that appears between the detectability transition  $c_c$ , and the spinodal transition  $c_s$ . In this phase, both Monte Carlo and BP, that aim to sample the Bayes-optimal posterior distribution, fail and get blocked in the metastable paramagnetic phase. The spinodal transition  $c_s$  is also a barrier for all known spectral algorithms. Some of the spectral methods including the one based on the non-backtracking matrix saturate this threshold.

There are many other problems where a phase of the same physical origin was discovered. Among them the best known are the LDPC error correcting codes [5] and the planted clique problem [26, 147–149]. Some recent works prove hardness in other problems assuming the hardness of planted clique (in the corresponding region) [150–153]. There is a lot of current interest in this kind of results.

When a physicist sees a phenomenon as the one above she immediately thinks about universality and tends to conjecture that the spinodal transition in inference problems will be a barrier for all polynomial algorithms. There is, however, the case of planted XOR-SAT that presents the same phenomenology as described in Section 3.5. Moreover, the hard phase in planted XOR-SAT extends to infinity (the paramagnetic state is always locally stable). The work of Feldman et al. [136] actually suggests that the right scaling for the number of clauses at which planted  $K$ -XOR-SAT becomes easy is when the number of clauses is at least  $O(N^{K/2})$ . Yet, thanks to its linearity planted  $K$ -XOR-SAT can be solved in polynomial time by Gaussian elimination. Hence the above conjecture about the hard phase is false. But the physicist’s intuition is rarely entirely wrong, and in those rare cases when it is indeed wrong it is for absolutely fundamental reasons. In our opinion, the question:

In which exact sense is the hard phase hard?

is the most fundamental question related to the interface between statistical physics, statistical inference and average computational complexity.

A line of research that should be explored, and seems quite accessible, is proving or disproving that there is no way to use BP or Monte Carlo sampling with a mismatched prior or model in order to improve the position of the spinodal transition.

Let us also note here that the spinodals of first-order phase transitions in Bayes-optimal inference problems create a much cleaner algorithmic barrier than phase transitions described in optimization or in CSP such as  $K$ -SAT. In  $K$ -SAT the idea of hard instances being close to the satisfiability threshold goes back to the works by Cheeseman et al. [154] Selman et al. [155] and Monasson et al. [156]. However, even after more than 20 years of study it is still not clear what is the limiting constraint density up to which polynomial algorithms are able to find satisfiable solutions in random  $K$ -SAT (even at very large  $K$ ). For a related discussion, see [157].

### 4.3. Spatial coupling

Spatial coupling is, from a physics point of view, a very natural strategy to beat the hard-easy spinodal transition in systems where the structure of the graphical model can be designed. It is closely related to nucleation in physics. An exponentially long living metastability is only possible in mean-field systems. Let us describe here why it is not possible in systems embedded in finite Euclidean dimension  $d$ . Metastability means that there are two competing phases. Let us call  $\Delta f$  the free energy difference between them. Let us consider a large system in the metastable state with a droplet of radius  $L$  (small compared to the size of the whole system) that by random fluctuation appears to be in the stable state. Creating such a droplet will cost some energy due to its interface where the two phases do not match. This is quantified by the surface tension  $\Gamma$ . The total energy balance can be written as

$$E(L) = \Gamma L^{d-1} - \Delta f L^d. \quad (90)$$

Physical dynamics can be roughly thought of as gradient descent in this energy. Notice that  $E(L)$  is increasing for  $L < L^* = \Gamma(d-1)/(\Delta f)$  and decreasing for  $L > L^*$ . This means that if fluctuations created a droplet larger than the critical size  $L^*$  it will continue growing until it invades the whole system. The above argument works only in geometries where the surface of a large droplet is much smaller than its volume. On random sparse graphs or fully connected geometries it does not work. This is why nucleation does not exist in mean-field systems. Notice also that nucleation is simpler for lower dimensions  $d$ , the lowest meaningful case being  $d = 1$ .

We can conclude that in order to avoid long living metastability we need to work with finite-dimensional systems. But the situation is not so simple either. In systems where the mean-field geometries are used, they are used for a good reason. In LDPC codes they assure large distance from spurious codewords, in compressed sensing random projections are known to conserve well the available information, etc. To avoid metastability we need to mix the mean-field geometry on a local scale with the finite-dimensional geometry on the global scale. This is exactly what spatial coupling is doing. Originally developed for LDPC error correcting codes [158,159] and closely related to convolutional LDPC that have longer history [160], spatial coupling has been applied to numerous other settings. Again we will discuss some of them in more details in Section 6.5. To introduce the concept here, we restrict ourselves to the spatially coupled Curie–Weiss (C–W) model as introduced in [161] and studied further in [162].

The C–W model in a magnetic field is arguably among the simplest models that present a first-order phase transition. The C–W model is a system of  $N$  Ising spins  $S_i \in \{-1, +1\}$ , that are interacting according to the fully connected Hamiltonian

$$\mathcal{H}_{\text{C-W}}(\mathbf{S}) = -\frac{J}{N} \sum_{(i,j)} S_i S_j - h \sum_{i=1}^N S_i, \quad (91)$$

where the notation  $\langle \cdot, \cdot \rangle$  is used to denote all unique pairs,  $J > 0$  is the ferromagnetic interaction strength and  $h \in \mathbb{R}$  is the external magnetic field.

Let us briefly remind what is obvious to everybody trained in statistical physics. For positive magnetic field  $h > 0$  the equilibrium magnetization  $m(J, h) > 0$  is positive, and vice versa. There exists a critical value of the interaction strength  $J_c = 1$  such that: for  $J < J_c$  the  $\lim_{h \rightarrow 0^+} m(J, h) = \lim_{h \rightarrow 0^-} m(J, h) = 0$ , and for  $J > J_c$  we have  $\lim_{h \rightarrow 0^+} m(J, h) > 0 > \lim_{h \rightarrow 0^-} m(J, h)$ . The latter is a first-order phase transition, in the “low temperature” regime  $J > J_c$  the system keeps a non-zero magnetization even at zero magnetic field  $h$ . There exists a spinodal value of the magnetic field

$$h_s(J) = \sqrt{J(J-1)} - \operatorname{atanh} \left( \sqrt{\frac{J-1}{J}} \right), \quad (92)$$

with the following properties: if the magnetizations are initialized to negative values and the magnetic field is of strength  $0 < h < h_s(J)$ , then both local physical dynamics and local inference algorithms, such as the Monte Carlo sampling, will stay at negative magnetization  $m^-(J, h) < 0$  for times exponentially large in the size of the system. The spinodal value of the magnetic field  $h_s(J)$  acts as an algorithmic barrier to equilibration, when initialized at  $S_i = -1$  for all  $i = 1, \dots, N$ , and hence to successful inference. For  $h > h_s(J)$  it is, on the other hand, easy to reach the equilibrium magnetization  $m^+(J, h)$ .

#### 4.4. The spatially coupled C–W model

We introduce here the construction from [161] that might not appear very transparent at first. But its idea is really simple. We create a one-dimensional chain of C–W models, put a nucleation seed somewhere and let the sub-systems interact in order to ensure that the nucleation proceeds to the whole chain.

As illustrated in Figure 8, we consider a one-dimensional chain of  $(2L + 1)$  C–W systems, where each of the C–W system has  $n$  spins (referred to as a “block”) and is labeled by the index  $z \in \{-L, \dots, L\}$ . The result is that a configuration  $\mathbf{S}$  of the full system is now given by the values of  $N = n(2L + 1)$  spins, each labeled by a compound index:

$$\mathbf{S} = \{S_{iz} \in \{+1, -1\} : i \in \{1, \dots, n\}, z \in \{-L, \dots, L\}\}. \quad (93)$$

In a similar way, the uniform external magnetic field  $h$  for a single system is replaced by an external field profile  $h_z$ . As far as the coupling is concerned, every spin not only connects to all spins in the same location  $z$  but also all spins within  $w$  blocks from  $z$ . The corresponding Hamiltonian is then

$$\mathcal{H}_{n,L}(\mathbf{S}) = -\frac{1}{n} \sum_{(iz, jz')} J_{zz'} S_{iz} S_{jz'} - \sum_{z=-L}^L h_z \sum_{i=1}^n S_{iz}. \quad (94)$$

The couplings between spins are  $J_{zz'} = Jg(|z - z'|/w)/w$ , where the function  $g$  satisfies the following condition  $g(|x|) = 0$ ,  $\forall |x| > 1$  and we choose its normalization to be  $(1/w) \sum_{z=-\infty}^{+\infty} g(w^{-1}|z|) = 1$ .

In order to ensure that the system *nucleates*, we must increase the field  $h_z$  at some point on the chain. We choose the magnetic field profile in such a way that  $h_z = h_{\text{seed}} > h_{\text{sp}}$  in some small region given by  $z \in \{0, \dots, w_{\text{seed}}\}$ . Everywhere else,  $h_z = h \ll h_{\text{sp}}$ , such that the average field strength is still small. To illustrate the nucleation in Figure 8 we initialize to  $S_{iz} = -1$  everywhere and let the system evolve under Monte Carlo dynamics. We indeed observe a nucleation “wave”

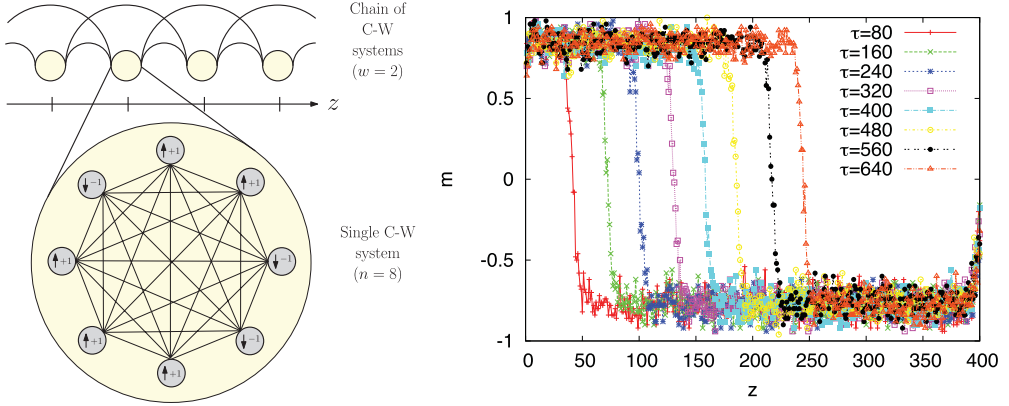


Figure 8. Left: A schematic graphical representation of the spatially coupled C-W model. A chain of C-W models interacting within a certain range  $w$  ( $w = 2$  in the figure). In the zoomed part the fully connected structure of the single C-W model is shown. A connection between C-W models along the chain indicates connections between all spins contained in both models. Right: The propagating wave obtained by Monte Carlo heat-bath simulation for the following parameters:  $J = 1.4$ ,  $L = 400$ ,  $n = 100$ ,  $h = 0.05$ ,  $w = 5$ ,  $w_{\text{seed}} = 10$  and  $h_{\text{seed}} = 0.3$ . Simulations were performed using random sequential update and therefore time must be rescaled by  $\tau = t/N$ . The figure is taken from [162].

that invades the whole system in a time proportional to the number of blocks. The system can be designed in a such a way that the spinodal regime entirely disappears [161].

#### 4.5. A brief history of time indices in the TAP equations

In physics of disordered system, the historically first form of iterative procedure that was used to compute marginals (local magnetizations) in a densely connected SG model are the TAP equations [46]. In the early statistical physics works on error correcting codes TAP equations were used for decoding. The primary goal of the paper by Kabashima and Saad [118], who pointed out the connection between BP and the Bethe approximation, was to compare the performance of TAP and BP. In dense systems the fixed points of both BP and TAP agree. However, it was reported, for example, in [163,164] that TAP equations did not converge in some cases even when BP does and there is no RSB. With the renewed interest in message passing of dense systems that followed the work of Donoho et al. on compressed sensing [165], it is definitely worth making some up-to-date remarks on the TAP equations and their convergence.

The fixed point of the TAP equations

$$m_i = \tanh \left[ \beta h + \beta \sum_{j=1}^N J_{ij} m_j - \beta^2 m_i \sum_{j=1}^N J_{ij}^2 (1 - m_j^2) \right] \quad (95)$$

describes the replica symmetric marginals of the following fully connected SG Hamiltonian

$$\mathcal{H}_{\text{SK}}(\mathbf{S}) = - \sum_{\langle i,j \rangle} J_{ij} S_i S_j - h \sum_i S_i, \quad (96)$$

where  $S_i$  are the Ising spins, and  $J_{ij}$  are independent random variables of zero mean and variance  $\tilde{J}/N$ . Hence a typical value of  $J_{ij} = O(1/\sqrt{N})$ .

Fixed points of the TAP equations are stationary points of the corresponding TAP free energy [46]. Derived as such, TAP equations do not include any time indices and were mostly iterated in the simplest form of having time  $t-1$  on all magnetizations on the right hand side, and  $t$  on the left hand side. At the same time, as we said above, a number of authors reported that the TAP equations have problems to converge even when BP does converge and replica symmetry is not broken [163,164]. No satisfactory explanation for this existed in the literature until very recently [166]. The reason behind this non-convergence was that the time indices were wrong and when done correctly, convergence is restored, and on sufficiently dense systems TAP equations behave as nicely as BP.

For the Ising SG this was first remarked by Bolthausen [166], who proved the convergence of TAP with the right time indices in a non-zero magnetic field (and out of the glassy spin glassy phase). His paper inspired the proof technique of [167] but otherwise passed quite unnoticed. In the context of linear estimation, as discussed in Section 6, the analog of the TAP equations is the approximate message passing (AMP) algorithm [168] (but see also [169]). AMP is often thought of as the large degree limit of BP and in that way of thinking the correct time indices come up naturally. Here we re-derive TAP for the Ising SG as a large degree limit of BP to clarify this issue of time indices.

Let us go back to the BP for sparse graphs equation (72). Define messages  $m_{i \rightarrow j}^t \equiv \tanh(h + \beta \sum_{k \neq j} u_{k \rightarrow i}^{t-1})$ . Recall that  $J_{ij}$ , and hence also  $u$ , is of order  $O(1/\sqrt{N})$  and expand to the leading order. BP then becomes

$$m_{i \rightarrow j}^t = \tanh \left( \beta h + \beta \sum_{k \neq j} J_{ki} m_{k \rightarrow i}^{t-1} \right). \quad (97)$$

In this dense-graph message passing there is not yet any surprise concerning time indices. The TAP equations are closed on full magnetizations (not cavity ones) defined as

$$m_i^t = \tanh \left( \beta h + \beta \sum_k J_{ki} m_{k \rightarrow i}^{t-1} \right). \quad (98)$$

In order to write the TAP equations we need to write messages in terms of the marginals up to all orders that contribute to the final equations. We get by Taylor expansion with respect to the third term

$$m_{i \rightarrow j}^t = \tanh \left( \beta h + \beta \sum_k J_{ki} m_{k \rightarrow i}^{t-1} - \beta J_{ij} m_{j \rightarrow i}^{t-1} \right) \approx m_i^t - \beta J_{ij} m_j^{t-1} [1 - (m_i^t)^2], \quad (99)$$

which leads to the iterative TAP equations

$$m_i^t = \tanh \left\{ \beta h + \beta \sum_j J_{ij} m_j^{t-1} - \beta^2 m_i^{t-2} \sum_j J_{ij}^2 [1 - (m_j^{t-1})^2] \right\}. \quad (100)$$

Observe the time index ( $t-2$ ) before the last sum, which is not intuitive on the first sight. These equations are the ones written and analyzed by Bolthausen [166].

In his proof, Bolthausen showed that the behavior of these iterative equations can be followed using what can be called the *state evolution* of the TAP algorithm, at least outside of the SG phase (the algorithm does *not* converge inside the SG phase). It works as follows: we consider the interactions  $J_{ij}$  to have zero mean and variance  $\tilde{J}/N$ . Going back to Equation (98), the argument of

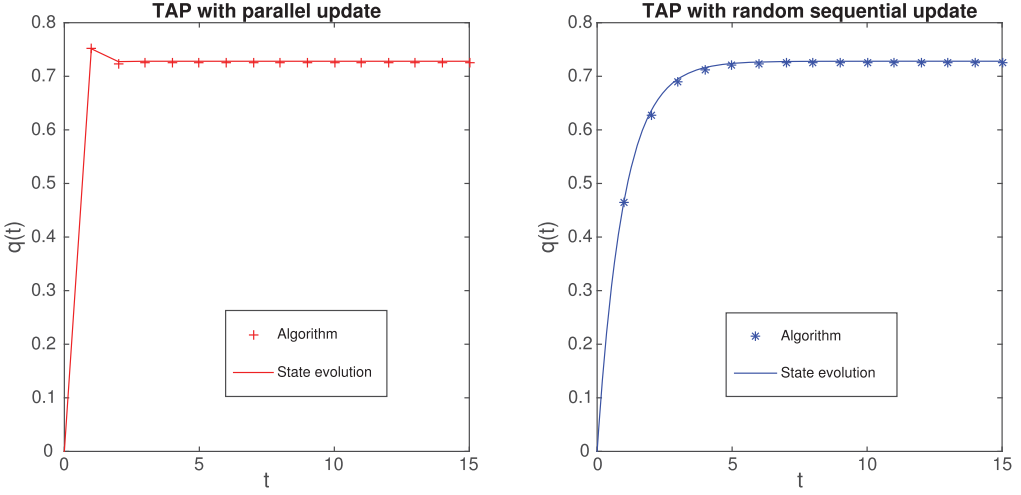


Figure 9. State evolution versus actual simulation for the TAP equations in the Sherrington–Kirkpatrick with parallel update (left) and random sequential one (right), for a SG system with  $N = 5000$ ,  $1/\beta = T = 3$ ,  $\tilde{J} = 1$  and  $h = 4$ . The variance  $q$  of the distribution of messages is shown. In both cases, the state evolution equations (with the iterative form on the left and the differential one on the right) perfectly describe the evolution of the algorithm.

the tanh is a sum of independent variables (by the assumptions of BP) and thus follows a Gaussian distribution. Denoting  $m$  and  $q$  the mean and the variance of this distribution of messages, two self-consistent equations can be derived

$$m' = \frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} \tanh(\beta z \tilde{J} \sqrt{q^{t-1}} + \beta h) dz, \quad (101)$$

$$q' = \frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} \tanh^2(\beta z \tilde{J} \sqrt{q^{t-1}} + \beta h) dz. \quad (102)$$

The fixed point of these equations provides the well-known replica symmetric solution for the magnetization and the spin overlap in the Sherrington–Kirkpatrick model, as was first derived in [108]. Since this is the correct solution outside the SG phase (above the so-called dAT line [170]), it is reassuring to see that we have a convergent algorithm that will find the marginals fast in the densely connected case. We will return to this observation in Section 6.3 when we will discuss inference in densely connected models.

Note also that Equations (101)–(102) correspond to a parallel update where all magnetizations are updated at the same time. Very often, it has been found that such update might be problematic for convergence and that one should update spins one at a time for better convergence properties. This will also be discussed in Section 6.3 when discussing the AMP algorithm [171]. In that case one chooses randomly at each time step which spin should be updated, then one can still use Equation (100) with a careful book-keeping of the time indices. The state evolution equation (102) become in that case differential equations in time. Instead of  $q^{t+1} = f(q^t)$  as in Equation (102) we have now  $dq/dt = f(q^t) - q^t$  in the continuous time limit. We show an example of the behavior of these two approaches in Figure 9. While in this case the parallel update was faster, we will see that sometimes the other one could be beneficial because it is smoother and avoids, for example, the overshoot we see in the first iteration of the parallel update.



These considerations stems from the derivation of TAP equations starting from the BP equations for random interaction. To finish this section, we point out that there have been a large body of work in trying to write TAP equations for interaction that are not random. We refer the interested reader to the book collection [38] for detailed discussion of the TAP equations in this context. Along this line of thoughts, the adaptive TAP approach [172] allows to learn the structure of the matrix  $J$ , while the structure of the Onsager term can be exactly computed depending on the structure of the correlations (see, e.g. [172–174]).

Since these approaches do not lead to a natural time ordering, it is worth mentioning that, very recently, these have been revisited [175] in the light of the “time indices” subtlety, leading a much more satisfying theory of the dynamics of the TAP equations on general matrices.

## 5. Clustering of networks and community detection

Clustering is one of the most commonly studied problems in machine learning. The goal is to assign data points to  $q$  clusters in a way that the cluster is somehow representative of the properties of the points inside it. About 15 years ago, a large number of scientists got interested in the research on so-called complex networks, that is sets of nodes where some pairs are connected by edges, arising in a wide range of settings. A natural question followed: How to cluster networks? That is, how to assign nodes of a graph in groups such that the largest possible part of the edges can be explained by a set of affinities between the clusters (instead of between the nodes themselves)? Typically, one considers a number of clusters much smaller than the number of nodes. Part of the motivation for studying this question came from studies of social networks which resulted in the name for this problem: *community detection*.

### 5.1. Biased overview of community detection

The literature on community detection is vast and there are a number of reviews that make a great effort in putting the existing works in chronological and merits perspective, for example [176]. Here we give a perspective biased toward the approach of statistical physics of disordered systems.

In community detection we are given a graph  $G(V, E)$  of  $N$  nodes and  $M$  edges. We search to assign the nodes into  $q$  groups, calling the assignment  $s_i \in \{1, \dots, q\}$ ,  $i = 1, \dots, N$ . In the most commonly considered assortative setting, the goal is that there are many edges between nodes of the same group and few edges among the groups. Traditionally, in statistics or machine learning, this problem would be solved by a spectral relaxation. One takes a matrix associated to the graph, its adjacency, its Laplacian or even something else, and compute its largest (or smallest) eigenvalues and associated eigenvectors. The elements of these eigenvectors would then reveal in a *visible* way the community structure [177, 178], for more details, see the nice review [179].

An influential set of works on community detection is due to Mark Newman and collaborators who introduced the so-called *modularity* function [180–182]

$$Q = \frac{1}{2M} \sum_{(i,j)} \left( A_{ij} - \frac{d_i d_j}{2M} \right) \delta_{s_i, s_j}, \quad (103)$$

where the sum is over all distinct pairs of nodes,  $d_i$  is the degree of node  $i$ . Community detection is performed by finding an assignment of nodes that maximizes this modularity. Optimizing the modularity is a non-deterministic polynomial -hard problem. In practice, however, many efficient heuristic algorithms were suggested for this task, see, for example [176].

Modularity-based community detection has nice connections to statistical mechanics and SG. Reichardt and Bornholdt [183] pointed out that modularity maximization can be seen as a case of finding a ground state of a particular Potts SG. In order to assess the statistical significance of community structure found by maximizing the modularity, Ref. [183] estimated the maximum value of modularity on a random graph, as the null model. Indeed, on sparse random graphs the maximum value of modularity can be very large even when no community structure whatsoever is present. To give an example, results on random graph partitioning [184] teach us that a random three-regular graph can be divided into two groups of roughly equal size in such a way that only about 11% of edges are between the groups. This visually looks like a very good community structure with high modularity. Yet there were no communities in the process that created the random graph. The need to compare the value of modularity to a null model is not very satisfying. Ideally, one would like to have a method able to tell that the underlying graph does not contain any communities if it did not.

Let us assume that we believe there were  $q$  communities respectively taking a fraction  $n_a$  of nodes,  $a = 1, \dots, q$ ,  $\sum_a n_a = 1$ , and the probability that a given random node from community  $a$  is connected to a given random node from community  $b$  is  $c_{ab}/N$ , where  $N$  is the total number of nodes and  $c_{ab}$  is a system size independent  $q \times q$  matrix. There will be roughly  $c_{ab}n_a n_b N$  edges between groups  $a$  and  $b$ , and  $c_{aa}n_a^2 N/2$  edges inside group  $a$ . The above very natural assumption is called the sparse SBM [185,186], and is commonly considered in this setting. As Bayesian probability theory teaches us, the best possible way of inferring the hidden communities is to enumerate all assignments of nodes into communities that have roughly the above sizes and respective number of edges between groups. Subsequently one averages over all such assignments (factoring away the global permutational symmetry between groups). This way one ends up with probabilities for a given node to belong to a given group. For a truly random graph these probabilities will not be different from the fractions  $n_a$ .

From a computationally complexity point of view, exactly enumerating assignments as above is even harder than finding optimal values of a cost function such as the modularity. However, from a statistical physics point of view evaluating the above averages is equivalent to the computation of local magnetizations in the corresponding Potts glass model. Notice, for instance, that planted coloring [31] is simply a special case of the SBM ( $c_{aa} = 0$  and  $c_{ab} = c_{cd}$  for all  $a \neq b$ ,  $c \neq d$ ). We shall summarize in the rest of this section the asymptotically exact analysis of the SBM [125,187].

## 5.2. Asymptotically exact analysis of the SBM

We explained all the methodology needed for the asymptotic analysis of the SBM on the example of the planted SG in Section 2. In this section, we consider again the teacher–student scenario where the teacher generates the observed graph from the SBM. We denote  $\theta = \{n_a, c_{ab}\}$  the set of parameters of the SBM. We remind that  $n_a$  is the fraction of nodes taken by group  $a$ ,  $a = 1, \dots, q$ , and  $c_{ab}/N$  is the probability that a given node in group  $a$  connects to a given node in group  $b$ . Again we denote by  $\theta^*$  the actual values that the teacher used to generate the graph  $G$ . The graph is represented by its adjacency matrix  $A_{ij}$ , with  $N$  nodes and  $M$  edges. Further  $s_i \in \{1, \dots, q\}$  are the Potts spins denoting the assignments to the respective groups.

The posterior likelihood of the SBM reads

$$P(\mathbf{s} | G, \theta) = \frac{1}{Z(G, \theta)} \prod_{i=1}^N n_{s_i} \prod_{\langle i, j \rangle} (c_{s_i s_j})^{A_{ij}} \left(1 - \frac{c_{s_i s_j}}{N}\right)^{1-A_{ij}}, \quad (104)$$

where the second product is over all distinct pairs, and we included a constant  $N^{-M}$  into the partition function. We need to evaluate the MMO estimator (22) in order to find a configuration that

maximizes the number of nodes that are assigned into the same group as the planted configuration. In a physics language, we would rather think of the corresponding Potts model Hamiltonian

$$\mathcal{H}(G, \theta, \mathbf{s}) = \sum_{i=1}^N \log n_{s_i} + \sum_{\langle i,j \rangle} \left[ A_{ij} \log c_{s_i s_j} + (1 - A_{ij}) \log \left( 1 - \frac{c_{s_i s_j}}{N} \right) \right] \quad (105)$$

and the MMO estimator turns out to be computed from the local magnetizations of this Potts model.

When the parameters  $\theta^*$  are not known, we need to learn them according to the lines of Section 2.4. We write the posterior distribution on the parameters  $P(\theta | G)$  and notice that it is concentrated around the maximum of the partition function  $Z(G, \theta)$ . We then iteratively maximize this partition function by imposing the Nishimori conditions

$$n_a = \frac{1}{N} \sum_{i=1}^N \mathbb{E}(\delta_{s_i, a}), \quad (106)$$

$$c_{ab} = \frac{1}{N n_a n_b} \sum_{(ij) \in E} \mathbb{E}(\delta_{a, s_i} \delta_{b, s_j}) \quad \text{when } a \neq b, \quad (107)$$

$$c_{aa} = \frac{2}{N n_a^2} \sum_{(ij) \in E} \mathbb{E}(\delta_{a, s_i} \delta_{a, s_j}). \quad (108)$$

This learning algorithm was called expectation maximization learning in [125]. The expectations were computed using BP or Gibbs sampling as will be discussed subsequently. But even if the expectations were computed exactly, this parameter learning algorithm depends strongly on the initial choice for the parameters  $\theta$ . It is easily blocked in a local maximum of the partition function. The size of the space of initial conditions that need to be explored to find the global maximum is independent of  $N$ , but in practice this is a problem, and completely general models  $c_{ab}$  are not so easy to learn. A considerable improvement can be obtained when the learning is initialized based on the result of spectral clustering. This procedure was suggested in [188] and works nicely, in particular in conjunction with the best known spectral clustering techniques [140] as will be discussed in Section 5.4.

Two algorithms that are (in the Bayes-optimal setting) asymptotically exact for computing the marginals and the expectations in Equations (106)–(108) were proposed in [125, 187]. One is the Gibbs sampling (or Markov chain Monte Carlo), another one is BP. We present here the BP algorithm, as we have introduced it in Section 3. It can also be used very directly to understand the behavior of the system and the phase transitions in the thermodynamic limit. The BP algorithm for the SBM was first written by Hastings [189] but in that work the parameters were fixed ad hoc, not in the Bayes-optimal way, nor by learning. Moreover, the algorithm was tested on a network of only 128 nodes, which is too small to observe any phase transition.

Repeating the steps of Section 3, we now derive BP. For the posterior distribution (104) BP is written in terms of the so-called messages  $\psi_{s_i}^{i \rightarrow j}$  that are defined as probabilities that node  $i$  is assigned to group  $s_i$  conditioned on the absence of edge  $(ij)$ . Assuming asymptotic conditional independence of the probabilities  $\psi_{s_k}^{k \rightarrow i}$  for  $k \neq j, i$ , the iterative equations read

$$\psi_{s_i}^{i \rightarrow j} = \frac{1}{Z^{i \rightarrow j}} n_{s_i} \prod_{k \neq i, j} \left[ \sum_{s_k} c_{s_i s_k}^{A_{ik}} \left( 1 - \frac{c_{s_i s_k}}{N} \right)^{1-A_{ik}} \psi_{s_k}^{k \rightarrow i} \right], \quad (109)$$

where  $Z^{i \rightarrow j}$  is a normalization constant ensuring that  $\sum_{s=1}^q \psi_s^{i \rightarrow j} = 1$ . We further notice that because of the factor  $1/N$ , messages on non-edges  $(ij) \notin E$  depend only weakly on the target

node and hence at the leading order the BP equations simplify to

$$\psi_{s_i}^{i \rightarrow j} = \frac{1}{Z^{i \rightarrow j}} n_{s_i} e^{-h_{s_i}} \prod_{k \in \partial i \setminus j} \left[ \sum_{s_k} c_{s_k s_i} \psi_{s_k}^{k \rightarrow i} \right], \quad (110)$$

where we denote by  $\partial i$  the set of neighbors of  $i$ , and define an auxiliary external field as

$$h_{s_i} = \frac{1}{N} \sum_{k=1}^N \sum_{s_k} c_{s_k s_i} \psi_{s_k}^k, \quad (111)$$

where  $\psi_{s_i}^i$  is the BP estimate of the marginal probability of node  $i$

$$\psi_{s_i}^i = \frac{1}{Z^i} n_{s_i} e^{-h_{s_i}} \prod_{j \in \partial i} \left[ \sum_{s_j} c_{s_j s_i} \psi_{s_j}^{j \rightarrow i} \right], \quad (112)$$

with  $Z^i$  being again a normalization. In order to find a fixed point of Equation (110) we update the messages  $\psi^{i \rightarrow j}$ , recompute  $\psi^j$ , update the field  $h_{s_i}$  by adding the new contribution, subtracting the old one, and repeat.

Once a fixed point of the BP equations is reached, its associated free energy (log-likelihood)  $f_{\text{BP}} = -\log Z(G, \theta)/N$  reads

$$f_{\text{BP}}(G, \theta) = -\frac{1}{N} \sum_i \log Z^i + \frac{1}{N} \sum_{(i,j) \in E} \log Z^{ij} - \frac{c}{2}, \quad (113)$$

where

$$Z^{ij} = \sum_{a < b} c_{ab} (\psi_a^{i \rightarrow j} \psi_b^{j \rightarrow i} + \psi_b^{i \rightarrow j} \psi_a^{j \rightarrow i}) + \sum_a c_{aa} \psi_a^{i \rightarrow j} \psi_a^{j \rightarrow i} \quad \text{for } (i, j) \in E, \quad (114)$$

$$Z^i = \sum_{s_i} n_{s_i} e^{-h_{s_i}} \prod_{j \in \partial i} \sum_{s_j} c_{s_j s_i} \psi_{s_j}^{j \rightarrow i}. \quad (115)$$

In [125, 187] the author argued that the above equations can be used to analyze the asymptotic performance of Bayes-optimal inference in the SBM,  $\theta = \theta^*$ . The analysis goes as follow: let us iterate BP (110) from two different initializations:

- Random initialization, for each  $(ij) \in E$  we chose  $\psi_s^{i \rightarrow j} = n_a + \epsilon_s^{i \rightarrow j}$  where  $\epsilon_s^{i \rightarrow j}$  is a small zero-sum perturbation,  $\sum_s \epsilon_s^{i \rightarrow j} = 0$ .
- Planted initialization, each message is initialized as  $\psi_s^{i \rightarrow j} = \delta_{s, s_i^*}$ , where  $s_i^*$  is the actual assignment of node to groups (that is known only for the purpose of analysis).

The free energy (113) is computed for the two corresponding fixed points. The fixed point with smaller free energy (larger likelihood) corresponds asymptotically to the performance of the Bayes-optimal inference. Here “asymptotically” is meant in the sense that the difference between the Bayes-optimal and the BP marginals goes to zero as  $N \rightarrow \infty$ . This is an important and non-trivial property, because in general the analysis of Bayes-optimal inference in such a high-dimensional setting is very difficult (sharp-P hard). Note that the asymptotic exactness of BP is so far only a conjecture. Rigorous results showing the asymptotic exactness of BP are so far only partial, either at relatively high average degrees [190], or in related problems for very low average degrees [191], or, for example, in ferromagnets [192].

Another property of BP that makes it very convenient for the analysis of phase transitions and asymptotic properties is the fact that BP, in a sense, ignores some part of the finite size effects. Let us compare BP to Markov chain Monte Carlo methods. In the large size limit the performance and behavior of BP and MCMC are perfectly comparable, as demonstrated for the SBM in [125]. But for finite sizes there are several remarkable differences. Consider a set of parameters where the two initializations of BP do not give the same fixed point. Even if we iterate BP for infinite time we will have two different fixed points. On the other hand, in Monte Carlo Gibbs sampling for a system of size  $N$ , we will eventually find the lower free energy state (just as for every ergodic Markov chain satisfying detailed balance). The time needed is in general exponentially large in  $N$ , but decreases as the free energy barrier gets smaller and therefore locating the spinodal threshold precisely from finite system size simulations is computationally more demanding compared to BP. Another manifestation of this property is that in the paramagnetic phase on a graph with  $N$  nodes and  $M$  edges the free energy given by BP is always the same number. Whereas the exact (finite  $N$ ) free energy typically has some fluctuations.

In the SBM, BP provides an asymptotically exact analysis of Bayes-optimal inference. In the case where the model or its parameters  $\theta$  are mismatching the ones that generated the graph, this is not true in general. The most striking example here is the MAP estimator, in other words the ground state of the corresponding Potts model. Modularity maximization is a special case of this setting. The ground state could also be estimated by BP if we introduced a temperature-like parameter  $1/\beta$  in the posterior and tuned this temperature to zero. A related algorithm was discussed recently in the context of modularity-based community detection in [193]. However, as we decrease this temperature we will typically encounter a point  $\beta_{\text{SG}} > 1$  above which the replica symmetric approximation fails and the corresponding BP marginals are not any longer asymptotically exact. When this phase transition at  $\beta_{\text{SG}}$  is continuous this will manifest itself as non-convergence of the BP algorithm. Note that depending on the set of parameters, this SG phase can be correlated to the original assignment (mixed phase) or not (pure SG phase). This behavior is qualitatively the same as the one seen in the phase diagram presented in Figure 2. In the above sense it is hence easier to compute the marginals at the Bayes-optimal value of the parameters (that may need to be learned as described above) than to compute the ground state (or maximize the modularity) that typically is glassy.

### 5.3. Examples of phase transitions in the SBM

The asymptotic analysis of the SBM as described above leads to the discovery of phase transitions. The most striking picture arises in the case when the parameters of the model are such that the average degree is the same in every group. In a sense this is the hardest case for inference, because if groups differ in their average degree, then a simple histogram of degrees gives some information about the correct assignment into groups.

Mathematically, the same average degree condition is expressed as

$$\sum_b c_{ab} n_b = c \quad \forall b = 1, \dots, q, \quad (116)$$

where  $c$  is the average degree. It is straightforward to see that when Equation (116) holds, then

$$\psi_s^{i \rightarrow j} = n_s \quad (117)$$

is always a (so-called *uniform*) fixed point (in general not the only one) of BP (110).

Let us now return to the concept of quiet planting explained for the planted SG in Section 2.5.3 and note that all we needed for quiet planting to work was precisely the existence of such a uniform fixed point. It is hence straightforward that the phase transitions described in Sections 2.3

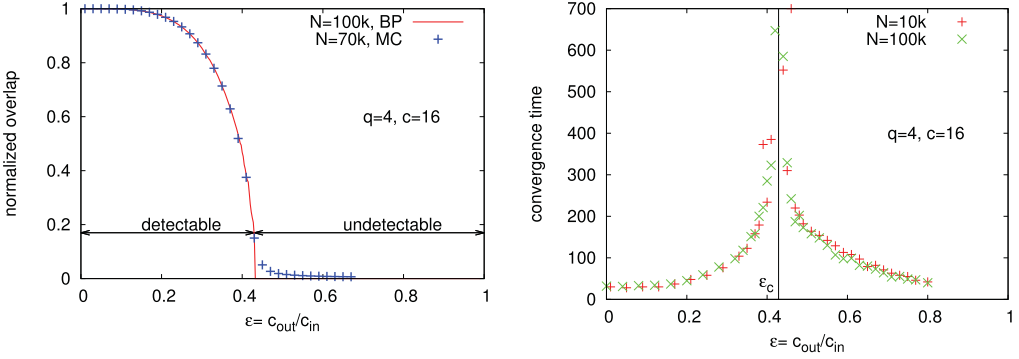


Figure 10. Left: Normalized overlap between the MMO estimator and the original assignment to groups. We compare the BP result to Monte Carlo Gibbs sampling. Both agree well, but we see the absence of finite size corrections in BP in the undetectable phase. Right: The BP convergence time, illustrating the critical slowing down around the phase transition. Figures are both taken from [125].

and 3.5 also exist in Bayes-optimal inference in the SBM. Let us give the example of assortative community structure with  $q=4$  equally sized groups,  $n_a = 1/q$  for all  $a = 1, \dots, q$ , average degree  $c = 16$ , and affinity parameters such that  $c_{aa} = c_{\text{in}}$  and  $c_{ab} = c_{\text{out}}$  for all  $a \neq b$ . We define a noise-like parameter  $\epsilon = c_{\text{out}}/c_{\text{in}}$ . When  $\epsilon = 0$  the model generates four entirely disconnected groups, and when  $\epsilon = 1$  the generation process corresponds to an ER graph with no information on the community structure. Without having any experience with phase transition and quiet planting, one might anticipate that whenever  $\epsilon < 1$  the Bayes-optimal inference will give a better overlap than a random assignment of nodes into groups. This expectation is wrong and the phase transition that happens instead is illustrated in Figure 10.

For other values of parameters we can observe the triplet of discontinuous phase transition discussed in Section 3.5. Actually, the planted coloring example, that was presented in Figure 6, is a case of the SBM with  $q=5$  equally sized groups and extremely dis-assortative structure  $c_{\text{in}} = 0$ .

For the general case of the sparse SBM, the order of the phase transition, and the locations of the dynamical  $c_d$  and detectability  $c_c$  phase transitions, have to be analyzed case by case mostly numerically. However, the spinodal transition has a neat analytical expression in the case of the same average degree in every group (116). This expression is obtained as in Section 3.3 by analyzing the local stability of the uniform fixed point. BP (110) linearized around this fixed point reads

$$\epsilon_a^{i \rightarrow j} = \sum_{b=1}^q \sum_{k \in \partial i \setminus j} T_{ab} \epsilon_b^{k \rightarrow i}, \quad (118)$$

where the  $q \times q$  matrix  $T_{ab}$  is defined as

$$T^{ab} = n_a \left( \frac{c_{ab}}{c} - 1 \right). \quad (119)$$

The phase transition where the paramagnetic fixed point ceases to be locally stable happens when

$$c\lambda_T^2 = 1, \quad (120)$$

where  $\lambda_T$  is the largest eigenvalue of the matrix  $T$ .

In case of equally sized groups in the *assortative-diagonal* case where  $c_{aa} = c_{\text{in}}$  and  $c_{ab} = c_{\text{out}}$  for  $a \neq b$ , the largest eigenvalue of the  $T^{ab}$  matrix (119) is  $\lambda_T = (c_{\text{in}} - c_{\text{out}})/(cq)$  from which Equation (120) implies for the spinodal phase transition [187]

$$|c_{\text{in}} - c_{\text{out}}| = q\sqrt{c}. \quad (121)$$

When the difference between  $c_{\text{in}}$  and  $c_{\text{out}}$  is larger, BP detects communities and gives asymptotically the Bayes-optimal overlap. When the difference is smaller BP fails. Depending on the parameters and the order of the phase transition the detectability phase transition itself can either be given by Equation (120) or happen at strictly smaller  $|c_{\text{in}} - c_{\text{out}}|$ , giving rise to the hard phase discussed in Section 4.2.

For the planted graph coloring (i.e. the extremely dis-assortative SBM,  $c_{\text{in}} = 0$ ), the detectability phase transition is of second order for  $q \leq 3$  and of first order for  $q \geq 4$ . The spinodal transition happens at  $c_s = (q - 1)^2$ . In the large  $q$  limit the dynamical and detectability phase transitions can be deduced from [30] while keeping in mind the concept of quiet planting

$$c_d = q(\log q + \log \log q) + O(1), \quad (122)$$

$$c_c = 2q \log q - \log q - 2 \log 2 + o(1). \quad (123)$$

For the assortative case  $c_{\text{in}} > c_{\text{out}}$  the phase transition is of second order for  $q \leq 4$  and of first order (although a very weak one for small  $q$ ) for  $q \geq 5$ . The limit of dense graphs (average probability for an edge to exist is  $p$ ,  $p_{\text{in/out}}$  are probabilities of connection within/between groups), and large number of groups  $q \rightarrow \infty$  was analyzed in [194] and is related to critical temperatures in the dense Potts glass [195,196]. The regime in which both the dynamical and detectability phase transitions appear is

$$N(p_{\text{in}} - p_{\text{out}})^2 = \gamma(q)p(1 - p), \quad (124)$$

with the dynamical transitions at  $\gamma_d(q) = 2q \log q[1 + o(1)]$ , the detectability transition at  $\gamma_c(q) = 4q \log q[1 + o(1)]$ , and the spinodal threshold at  $\gamma_s(q) = q^2$ . In this limit we see that the hard phase between  $\gamma_s(q)$  and  $\gamma_c(q)$  is very wide (compared to the overall scale).

The prediction and location of these phase transitions in the SBM from [125,187] was followed by a remarkable mathematical development where the location of the threshold was made rigorous for  $q=2$  [78,197,198]. The detectable side was proven for generic  $q$  in [146].

To briefly comment on the case where the average degree of every group is not the same: BP is still conjectured to be asymptotically exact, but the phase transitions get smeared. As recently studied in [199], the second-order phase transitions disappear, and the first-order phase transitions shrink into a multi-critical point beyond which they also disappear. The situation is similar to the semi-supervised clustering, where the assignment of a fraction of nodes is known [200]. From a physics point of view it is relevant to note that this problem is an inference interpretation of the particle pinning that was recently used quite extensively for studies of structural glasses [201].

#### 5.4. Spectral redemption for clustering sparse networks

Any reader that did not skip Section 4.1 is now clearly anticipating the use of the non-backtracking matrix for spectral clustering of networks. Indeed, in [140] the following spectral clustering method was proposed:

- Construct the non-backtracking matrix

$$B_{i \rightarrow j, k \rightarrow l} = \delta_{il}(1 - \delta_{jk}), \quad (125)$$

and compute its largest (in module) eigenvalues until they start to have a non-zero imaginary part. Denote  $k$  the number of those large real eigenvalues.



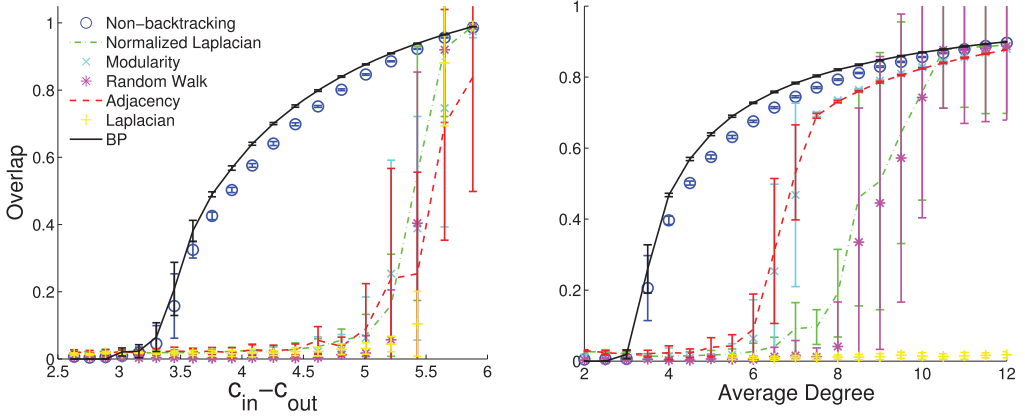


Figure 11. The accuracy of spectral algorithms based on different linear operators, and of BP, for two groups of equal size. On the left, we vary  $c_{\text{in}} - c_{\text{out}}$  while fixing the average degree  $c = 3$ ; the detectability transition given occurs at  $c_{\text{in}} - c_{\text{out}} = 2\sqrt{3} \approx 3.46$ . On the right, we set  $c_{\text{out}}/c_{\text{in}} = 0.3$  and vary  $c$ ; the detectability transition is at  $c \approx 3.45$ . Each point is averaged over 20 instances with  $N = 10^5$ . The spectral algorithm based on the non-backtracking matrix  $B$  achieves an accuracy close to that of BP, and both remain large all the way down to the transition. Standard spectral algorithms based on the adjacency matrix, modularity matrix, Laplacian, normalized Laplacian, and the random walk matrix all fail well above the transition, giving a regime where they do no better than chance. Figure is taken from [140].

- Consider the  $k$  corresponding  $2M$ -dimensional eigenvectors  $u_{i \rightarrow j}$  and create a set of  $N$ -dimensional vectors by resuming  $u_i = \sum_{k \in \partial i} u_{k \rightarrow i}$ . Then view the  $k$  vectors  $u_i$  as  $N$  points in  $k$  dimensional space and cluster those points into  $k$  groups using some off-the-shell clustering algorithm such as  $k$ -means.

Note in particular that  $k$  here is the estimator for the number of groups. Up to this point we were assuming that the number of groups was known. The performance of this non-backtracking-based spectral algorithms compared to BP, and five other spectral algorithms on clustering of sparse graphs, is illustrated in Figure 11. All these spectral algorithms look for the largest (or smallest) eigenvalue of the associated matrix and then assign nodes into groups based on the sign of the corresponding eigenvector. The five matrices that are most commonly used for clustering are the adjacency matrix  $A$ , the Laplacian  $L = D - A$  (where  $D$  is a diagonal matrix with the node degrees  $d_i$  on the diagonal), normalized Laplacian  $\tilde{L} = D^{-1/2} L D^{-1/2}$ , random walk matrix  $W_{ij} = A_{ij}/d_i$ , and the modularity matrix  $Q_{ij} = A_{ij} - d_i d_j / (2M)$ . The spectra of all these matrices on sparse graphs is affected by the existence of large eigenvalues associated to localized eigenvectors. This fact deteriorated their performance with respect to both BP and the non-backtracking-based algorithm.

Reference [140] analyzed the performance of the above algorithm on networks generated from the SBM with equal average degree in every group. A large part of results that were not rigorous in [140] were very remarkably made rigorous in [146].

On a random graph taken from the configurational model, that is, at random with a given degree distribution  $P(d)$ , in the large size limit, the spectrum of the non-backtracking matrix is (except one eigenvalue) included inside a circle of radius  $\sqrt{\tilde{c}}$ . Here  $\tilde{c}$  is the average excess degree of the graph computed as

$$\tilde{c} = \frac{\sum_d d^2 P(d)}{\sum_d d P(d)} - 1. \quad (126)$$

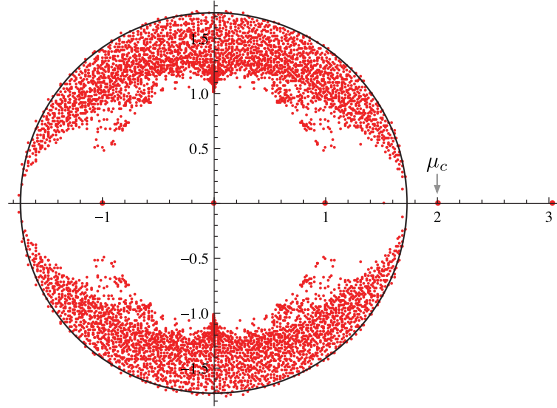


Figure 12. The spectrum of the non-backtracking matrix  $B$  for a network generated by the block model with  $N=4000$ ,  $c_{\text{in}} = 5$  and  $c_{\text{out}} = 1$ . The leading eigenvalue is at  $c=3$ , the second eigenvalue is close to  $\mu_c = (c_{\text{in}} - c_{\text{out}})/2 = 2$ , and the bulk of the spectrum is confined to the disk of radius  $\sqrt{c} = \sqrt{3}$ . The spectral algorithm that labels vertices according to the sign of  $B$ 's second eigenvector (summed over the incoming edges at each vertex) labels the majority of vertices correctly. Figure is taken from [140].

The one eigenvalue that is not inside this circle is real and has value  $\tilde{c}$ . This random graph spectrum should be contrasted with the spectrum of the adjacency matrix (and other related ones) which for graphs with constant average and unbounded maximum degree is not even bounded in the thermodynamic limit.

Now, consider a graph generated by the SBM with equal average degree for every group. Denote  $\lambda_T$  an eigenvalue of the matrix  $T$ , Equation (119), then as long as  $\lambda_T > 1/\sqrt{\tilde{c}}$  the non-backtracking matrix will have purely real eigenvalues at positions  $c\lambda_T$ . This way we obtain up to  $q-1$  additional eigenvalues on the real axes out of the circle of radius  $\sqrt{\tilde{c}}$ . Note that in the *assortative-diagonal* case the matrix  $T$  has  $q-1$  degenerate eigenvalues  $\lambda_T = (c_{\text{in}} - c_{\text{out}})/(qc)$ , and therefore the non-backtracking matrix has a  $(q-1)$ -times degenerate eigenvalue at  $(c_{\text{in}} - c_{\text{out}})/q$  which merges into the bulk below the spinodal phase transition. The spectrum of the non-backtracking operator is illustrated in Figure 12.

Consequently, when the average degree of every group is the same, condition (116), the phase transition of the spectral algorithm coincides with the spinodal phase transition of the BP. In other words, as long as the average degree is larger than given by Equation (120) the eigenvectors corresponding to the eigenvalues on the real axes out of the circle contain information about the ground-truth communities.

We need to stress that the agreement between the spectral and BP thresholds arises only for the case of equal average degree in every group. When the condition (116) is not satisfied there is information about the groups on the degree and BP explores this information in an optimal way. What concerns the phase transitions, they tend to be smeared (second-order phase transitions disappear, and first-order ones become weaker) [199]. The spectral methods, on the other hand, always have a threshold where the informative eigenvalues merge back into the bulk. At this spectral phase transition, when Equation (116) does not hold, BP does not witness any particular change in performance.

Let us compare properties of BP and the non-backtracking-based spectral clustering. The positive point for BP is that it gives a slightly better overlap. The positive points for NB-spectral clustering, that are shared with other spectral methods, for example, all those tested in Figure 11, are that it does not need to get any information about the model parameters nor the number of

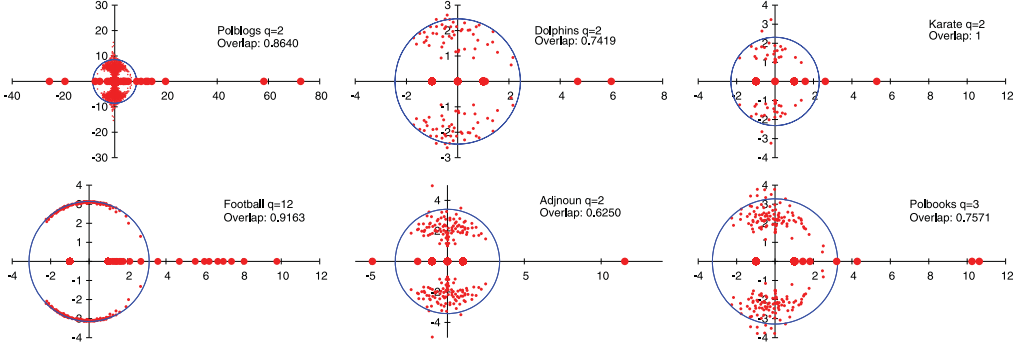


Figure 13. Spectrum of the non-backtracking matrix in the complex plane for some real networks commonly used as benchmarks for community detection, taken from [180,229–233]. The radius of the circle is the square root of the largest eigenvalue, which is a heuristic estimate of the bulk of the spectrum. The overlap is computed using the signs of the second eigenvector for the networks with two communities, and using  $k$ -means for those with three and more communities. The non-backtracking operator detects communities in all these networks, with an overlap comparable to the performance of other spectral methods. As in the case of synthetic networks generated by the SBM, the number of real eigenvalues outside the bulk appears to be a good indicator of the number  $q$  of communities. Figure taken from [140].

groups. It works in time linear in the number of groups, as opposed to quadratic for BP. It is not expected to be affected by the presence of small loops which usually cause the failure of the Bethe approximation. An inconvenience that is shared by all the considered spectral methods is that the performance can deteriorate considerably upon small adversarial changes to the graph (such as planting a small dense subgraph). Approaches based on some global optimization, such as BP, are expected to be robust toward small local changes.

Another inconvenient property of the non-backtracking matrix is that it can be rather large. However, this is no obstacle at all because the spectrum (different from  $\pm 1$ ) of  $B$  can be computed from the spectrum of the following  $2N \times 2N$  matrix  $B'$  [140]

$$B' = \begin{pmatrix} 0 & D - \mathbb{1} \\ -\mathbb{1} & A \end{pmatrix}, \quad (127)$$

where  $A$  is the adjacency matrix and  $D$  is the diagonal matrix with  $D_{ii} = d_i$  being the degree of node  $i$ .

The matrix  $B'$  is smaller but still non-symmetric. The power method for computing the largest eigenvalues works much better for symmetric matrices. In [202], it was suggested to use the so-called Bethe Hessian matrix defined as

$$\text{BH}(r) = (r^2 - 1)\mathbb{1} - rA + D, \quad (128)$$

with  $r = \sqrt{\tilde{c}}$  (or more general the squared root of the leading eigenvalue of Equation (125) which is usually not problematic to compute) as a basis for spectral clustering. In particular, to take eigenvectors of all of its negative eigenvalues. Reference [202] showed that this has all the advantages of the non-backtracking matrix while being  $N \times N$  and symmetric. The name Bethe Hessian is motivated by the relation between the Hessian of the Bethe free energy and this matrix [203,204], which is particularly apparent when generalized to the weighted case.

### 5.5. More on the non-backtracking operator

The non-backtracking operator as defined above is relatively well known in mathematics, often called the Hashimoto directed edges adjacency matrix [205]. It is used in a number of works as a tool in proofs, see, for example [206–211]. However, using it as a basis for a spectral algorithm was first suggested in [140]. The idea of using the linearization of BP was inspired by the earlier work of Coja-Oghlan et al. [212]. Since then, a considerable number of papers on promising algorithmic applications of the non-backtracking matrix on sparse networks started appearing, see [141,213,214]. Here, we briefly discuss selected contributions in this direction.

Looking at Figure 12, we observe a very interesting structure in the bulk spectral density of the non-backtracking operator. Motivated by this, Ref. [215] applied the cavity method designed to compute spectral densities of non-Hermitian random matrices [216] to study the non-backtracking spectral density. A phase transition in the spectral density was observed to happen on the boundary of the circle of radius  $\sqrt{c}$ . Such a phase transition is absent in the spectral density of traditional matrices such as the adjacency, Laplacian, random walks, etc. This is no proof since we are only concerned with spectral density not with sub-dominantly numerous eigenvalues, but from a physics point of view a phase transition is a strong indication that the properties of the non-backtracking matrix on a random graph are fundamentally different from those of traditional matrices. Reference [215] also observed that the eigenvalues that are close to the center of the circle are not visible in the asymptotic calculation of the spectral density. Numerical investigations revealed that eigenvectors corresponding to these eigenvalues are localized. More detailed understanding of this localization and its physical consequences is an interesting direction of future work.

In [217], the non-backtracking spectral clustering method was generalized to clustering of sparse hypergraphs. These are graphs where instead of connected pairs of nodes, one connects  $k$ -uples of nodes. Interestingly, using the non-backtracking-based spectral method detection can be done also, for example, in planted CSP without even knowing the form of the corresponding constraints. Moreover, for the same reasons as in clustering, the method works down to previously computed spinodal thresholds [138]. It was also observed that the overlap (performance) of the spectral method is always a continuous function of the noise parameter, even if the Bayes-optimal BP has a strongly discontinuous transition. This further supports our previous suggestion that the spectral methods are extremely useful to point toward the right direction, but message-passing-based methods should then be used to improve the precision. Note that the hypergraph clustering can also be formulated as tensor factorization that became recently popular in connection with interdependent networks or networks with different kinds of edges that are referred to as multiplexes in the literature on complex networks.

Another exciting application of the non-backtracking matrix, or the Bethe Hessian, is the matrix completion problem [218]. In matrix completion one aims to estimate a low-rank matrix from a small random sample of its entries. This problem became popular in connection to collaborative filtering and related recommendation systems. Of course, the fewer entries are observed, the harder the problem is. Reference [218] shows that with the use of the non-backtracking matrix the number of entries from which the low-rank structure is detectable is smaller than with other known algorithms. The way the matrix completion problem is approached is by realizing that a Hopfield model with a Hebb learning rule can be viewed as a rank  $r$  structure that is able to recover the  $r$  patterns as distinct thermodynamic phases [219]. One views the observed entries of the low-rank matrix to complete as a sparse Hopfield model and estimate the patterns from the leading eigenvectors of the corresponding Bethe Hessian. In a second stage, we use this as an initialization for more standard matrix completion algorithms in the spirit of Keshavan et al. [220], and achieve overall better performance.

Another problem where the leading eigenvalue of the non-backtracking matrix is relevant is percolation on networks [221]. In percolation, we start with a graph  $G$  and remove each of its

edges with probability  $1 - p$ . Then we look at the largest component of the remaining graph and ask whether it covers an extensive fraction of all nodes. The value of  $p_c$  below which the giant component does not exist anymore with high probability is called the percolation threshold. Percolation thresholds were computed for many geometries including random graphs. A remarkable result concerns any sequence of dense graphs for which it was proven that the threshold is given by the inverse of the leading eigenvalue of the adjacency matrix [222]. This result does not extend to sparse graphs, where a counter-example is given by random regular graphs in which the leading eigenvalue is the degree whereas the percolation threshold is the inverse of the degree minus one.

Based on a algorithm for percolation, Karrer et al. [221] argues that on a large class of sparse locally tree-like graphs the percolation threshold is given by the inverse of the leading eigenvalue of the non-backtracking matrix. Specifying for what exact class of graphs this result holds is still an open problem. Reference [221] also proved that the non-backtracking matrix provides a lower bound on the percolation threshold for a generic graph (with the exception of trees). An analogous result was obtained independently by Hamilton and Pryadko [223]. Due to these results, the non-backtracking matrix was used to study percolation on real networks [214,224,225].

### 5.6. *The deceptiveness of the variational Bayes method*

Both in physics and in computer science, a very common and popular method for approximate Bayesian inference is the so-called naive mean-field method, often called variational Bayesian inference [38,226]. For clustering of sparse networks generated by the SBM this method was studied, for example, in [227,228]. Away from the detectability threshold the variational Bayes method gives very sensible results. Deep in the undetectable phase (very small  $|c_{\text{in}} - c_{\text{out}}|$ ) the predicted magnetizations are close to zero, reflecting the undetectability. Deep in the detectable phase the variational Bayesian method is able to estimate the correct group assignment, and is statistically consistent as proven in [228]. The performance of variational Bayes in the neighborhood of the detectability transition was studied in [188]. The conclusions of the comparison to BP are quite interesting. It is not surprising that BP is more precise and converges faster on sparse graphs. What is surprising is that the naive mean-field inference fails in a very deceptive way: the marginals it predicts are very biased toward a configuration that is not correlated with the planted (ground-truth) configuration. Reference [188] quantifies this by introducing the so-called illusive overlap and comparing it to the actual one. This means that it can be rather dangerous to use naive mean-field for testing statistical significance of the results it provides and this could have profound consequences beyond the problem of clustering networks. Especially when trying to use the variational Bayes method in the vicinity of the corresponding detectable threshold, that is, in regimes where the number of samples is barely sufficient for estimation. Monte Carlo methods do not have this problem, but are in general slower.

### 5.7. *Real networks are not generated by the SBM*

Our study of clustering of networks is asymptotically exact for the case when the networks were generated by the SBM. In view of actual data sets and realistic applications this is never the case.

One important discrepancy between the SBM and real networks is that the latter often have communities with a broad degree distribution, whereas the degree distribution of every group generated by the SBM is Poissonian. A variant of the SBM that takes the degree distribution into account is the so-called degree corrected SBM [234]. An asymptotically exact BP can also easily be written for this model, see [235]. In Ref. [235] it was also shown that the corresponding Bethe free energy is a great tool for model selection between the canonical and degree corrected SBMs, and that classical model selection criteria do not apply to this problem because of the limited

information per degree of freedom. Interestingly, BP for the degree corrected SBM also leads to the same non-backtracking matrix upon linearization.

There are of course many more discrepancies between real networks and those generated by the (degree corrected) SBM. However, when observing the spectrum of the non-backtracking operator on some of the commonly used benchmarks for community detection, we conclude that the properties derived for the SBM basically stay conserved also in these examples, see Figure 13. To conclude this section, let us quote the British mathematician and Bayesian statistician George E. P. Box (1919–2013): “Essentially, all models are wrong, but some are useful”.

## 6. Linear estimation and compressed sensing

The inference problem addressed in this section is the following. We observe a  $M$  dimensional vector  $y_\mu$ ,  $\mu = 1, \dots, M$  that was created component-wise depending on a linear projection of an unknown  $N$  dimensional vector  $x_i$ ,  $i = 1, \dots, N$  by a known matrix  $F_{\mu i}$

$$y_\mu = f_{\text{out}} \left( \sum_{i=1}^N F_{\mu i} x_i \right), \quad (129)$$

where  $f_{\text{out}}$  is an output function that includes noise. Examples are the additive white Gaussian noise (AWGN) where the output is given by  $f_{\text{out}}(x) = x + w$  with  $w$  being random Gaussian variables, or the linear threshold output where  $f_{\text{out}}(x) = \text{sign}(x - \kappa)$ , with  $\kappa$  being a threshold value. The goal in linear estimation is to infer  $\mathbf{x}$  from the knowledge of  $\mathbf{y}$ ,  $F$  and  $f_{\text{out}}$ . An alternative representation of the output function  $f_{\text{out}}$  is to denote  $z_\mu = \sum_{i=1}^N F_{\mu i} x_i$  and talk about the likelihood of observing a given vector  $\mathbf{y}$  given  $\mathbf{z}$

$$P(\mathbf{y} | \mathbf{z}) = \prod_{\mu=1}^M P_{\text{out}}(y_\mu | z_\mu). \quad (130)$$

### 6.1. Some examples of linear estimation

To put the immense number of important applications of the above setting into perspective, let us give an illustrative example from medical diagnostics. Consider we have data about  $M$  cancer-patients that took a given treatment. Then  $y_\mu$  is a variable rating the effectiveness of the treatment for a patient  $\mu$ . Our goal is to predict the outcome of the treatment for a new patient that did not undergo treatment yet. To do that, we collect medical and personal data about each of the  $M$  patients, their medical history, family situation, but also for instance Facebook feeds, where they worked, what sport they did in their life, what education they have, how they rate their relationship with their wife, and all kinds of other information that is on first sight seems totally unrelated to the eventual outcome of the treatment. This information is collected in the vector  $\vec{F}_\mu$ . We aim to find a set of coefficients  $x_i$  such that the observations  $y_\mu$  can be well explained by  $z_\mu = \sum_{i=1}^N F_{\mu i} x_i$  for some  $P_{\text{out}}(y_\mu | z_\mu)$ .

In the way we set the problem, it is reasonable to think that most patient-attributes will have no effect on the outcome of the treatment, that is, for many  $i$ 's the  $x_i$  will be close to zero. In a method, called *least absolute shrinkage and selection operator* (LASSO) [236,237], that is used everyday for similar types of problems (in banking or insurance industry for instance) the coefficients  $x_i$  are such that

$$\text{LASSO : } \min_{\mathbf{x}} \left[ \sum_{\mu=1}^M \left( y_\mu - \sum_{i=1}^N F_{\mu i} x_i \right)^2 + \lambda \sum_{i=1}^N |x_i| \right], \quad (131)$$



where  $\lambda$  is a regularization parameter that controls the trade-off between a good fit and the sparsity of the vector  $\mathbf{x}$ . The advantage of LASSO is that it is a convex problem that can be solved efficiently and with strong guarantees using linear programming. LASSO and related linear regression problems are well studied in statistics and machine learning, instead of reviewing the literature we refer the reader to the excellent textbook [10].

LASSO can be framed in the Bayesian probabilistic setting by considering

$$P_{\text{out}}(y_\mu | z_\mu) = \frac{1}{\sqrt{2\pi}\Delta} e^{-(y_\mu - \sum_{i=1}^N F_{\mu i} x_i)^2 / 2\Delta}, \quad (132)$$

and the prior of  $x_i$  being the Laplace distribution

$$P_X^{\ell_1}(x_i) = e^{-\lambda|x_i|}/2. \quad (133)$$

Another sparsity inducing prior is a combination of strict zeros and fraction  $\rho$  of non-zeros

$$P_X(x_i) = \rho\phi(x_i) + (1 - \rho)\delta(x_i), \quad (134)$$

where  $\phi(x_i)$  is the distribution of non-zero elements. Linear programming does not work for priors such as Equation (134), but the statistical physics framework is more general in this aspect.

In many applications, the output variables  $y_\mu$  are binary (interpreted in the above example as the treatment worked or did not), then the most commonly considered output distribution is the logistic function. The conditional dependence then is

$$P_{\text{out}}^{\text{logistic}}(y_\mu = 1 | z_\mu) = \frac{1}{1 + e^{-z_\mu}}. \quad (135)$$

The example of this linear estimation setting that is the best known in statistical physics is the perceptron [238], as introduced briefly in Section 1.2.2. In the perceptron setting, components of the unknown vector  $\mathbf{x}$  are interpreted as synaptic weights (that are in general not sparse). The rows of the matrix  $F$  are patterns presented to the perceptron. The measurements  $y_\mu$  are binary and correspond to classifications of the patterns (think of some patterns being pictures of cats  $y_\mu = +1$  and the others being pictures of dogs  $y_\mu = -1$ ). The goal is to learn the synaptic weights  $x_i$  in such a way that  $\sum_i F_{\mu i} x_i > 0$  when  $y_\mu = +1$ , and  $\sum_i F_{\mu i} x_i < 0$  when  $y_\mu = -1$ . The random version of the problem has been considerably studied by physicists, see for instance [9,17,239–241].

Another example well known and studied in the statistical physics literature is the code-division multiple access (CDMA) problem in which  $N$  users simultaneously aim to transmit their binary signal  $x_i \in \{\pm 1\}$  to one base station. In order to be able to deal with interference one designs a so-called spreading code associated to each user  $F_{\mu i} \in \{\pm 1\}/\sqrt{N}$ ,  $\mu = 1, \dots, M$ , that spreads the signal of each user among  $M$  samples. The base station then receives a noisy linear combination of all the signals  $y_\mu$ , Equation (129). The goal is to reconstruct the signal of every user  $x_i$ .

Interestingly, the two examples above are the first inference problems on which, to our knowledge, the strategy discussed in the present chapter has been introduced in a statistical physics. The seminal work of Marc Mézard derived the TAP equations for the perceptron problem as early as in 1989 [242]. It has then been generalized to many other cases, mainly by Kabashima and co-workers (see [169,243] again for the perceptron and [164] for randomly spreading CDMA).

In this chapter, we shall follow these pioneering works and adopt the Bayesian probabilistic framework for the above generalized linear estimation problem. In general this is algorithmically much more involved than the convex relaxations of the problem, and consequently less



widely studied. The cases where interesting theoretical analysis can be done are limited to special classes of matrices  $F_{\mu i}$  (mostly with random iid elements, but see the discussion in Section 6.6). Nevertheless the resulting algorithms can (as always) be used as heuristic tools for more general data  $F_{\mu i}$ . One problem where the above linear estimation problem truly involves a matrix  $F_{\mu i}$  with random iid elements is *compressed sensing* [244] to which we dedicate the next sections.

## 6.2. Compressed sensing

Compressed sensing, a special case of the above linear estimation setting, triggered a major evolution in signal acquisition. Most signals of interest are compressible and this fact is widely used to save storage place. But so far compression is made only once the full signal is measured. In many applications (e.g. medical imaging using magnetic resonance imaging or tomography) it is desirable to reduce the number of measurements as much as possible (to reduce costs, time, radiation exposition dose, etc.). It is hence desirable to measure the signal directly in the compressed format. Compressed sensing is a concept implementing exactly this idea by designing new measurement protocols. In compressed sensing one aims to infer a sparse  $N$ -dimensional vector  $\mathbf{x}$  from the smallest possible number  $M$  of linear projections  $\mathbf{y} = F\mathbf{x}$ . The  $M \times N$  measurement matrix  $F$  is to be designed.

Most commonly used techniques in compressed sensing are based on a development that took place in 2006 thanks to the works of Donoho [244], Candes and Tao [245], and Candès and Wakin [246]. They proposed to find the vector  $\mathbf{x}$  satisfying the constraints  $\mathbf{y} = F\mathbf{x}$  which has the smallest  $\ell_1$  norm (i.e. the sum of the absolute values of its components). This  $\ell_1$  minimization problem can be solved efficiently using linear programming techniques. They also suggested using a random matrix for  $F$ . This is a crucial point, as it makes the  $M$  measurements random and incoherent. Incoherence expresses the idea that objects having a sparse representation must be spread out in the domain in which they are acquired, just as a Dirac or a spike in the time domain is spread out in the frequency domain after a Fourier transform. Rather strong and generic theoretical guarantees can be derived for the  $\ell_1$ -based sparse reconstruction algorithms [237]. These ideas have led to a fast, efficient and robust algorithm, and this  $\ell_1$ -reconstruction is now widely used, and has been at the origin of the burst of interest in compressed sensing over the last decade.

It is possible to compute exactly the performance of  $\ell_1$ -reconstruction in the limit  $N \rightarrow \infty$ , and the analytic study shows the appearance of a sharp phase transition [247]. For any signal with density  $\rho = K/N$ , where  $K$  is the number of non-zero elements, the  $\ell_1$ -reconstruction gives the exact signal with probability one if the measurement rate,  $\alpha = M/N$ , is above the so-called Donoho–Tanner line,  $\alpha > \alpha_{\ell_1}(\rho) > \rho$ , see Figure 15 (black dashed line). Interestingly, the very same transition can be found within the replica method [248,249] but can also be computed, rigorously, with the AMP technics that we discuss in this chapter [165].

If one is willing to forget the (exponentially large) computational cost, sparse signal reconstruction from random linear projections is possible whenever  $\alpha > \rho$  (this can be done for instance by exhaustively checking all possible positions of the non-zero elements and trying to invert the reduced system of linear equations). A challenging question is whether there is a compressed sensing design and a computationally tractable algorithm that can achieve reconstruction down to such low measurement rates. In a series of works [250,251], complemented by an impressive rigorous proof of Donoho et al. [252], this question was answered positively under the assumption that the empirical distribution of the non-zero elements of the signal is known. This contribution to the theory of compressed sensing is explained below and uses many of the statistical physics elements from Section 2.

### 6.3. Approximate message passing

AMP is the algorithmic procedure behind the aforementioned work of Donoho et al. [165] on linear estimation and compressed sensing. Origins of AMP trace back to statistical physics works on the perceptron problem [242] and on the CDMA [164].

What nowadays is known as the generalized approximate message passing (G-AMP) algorithm, i.e. the generalization to arbitrary element-wise output functions, was derived for the first time, to our knowledge, in connection to perceptron neural network in [169,242]. The algorithm presented in [169] had the correct time indices and good performance was illustrated on the CDMA problem. However, for an unknown reasons, that paper did not have much impact and many of its results waited to be rediscovered in the connection to compressed sensing much later. The opinion that this type of algorithm has invincible convergence problems prevailed in the literature until the authors of Braunstein and Zecchina [253] demonstrated impressive performance of a related message passing algorithm for learning in binary perceptron. AMP is closely related to relaxed-BP (r-BP) [254]. The algorithm of Braunstein and Zecchina [253] can actually be seen as a special case of r-BP.

Donoho et al. introduced AMP for compressed sensing [165], they also established the name. Very remarkably, the rigorous proof of its performance for random matrices  $F$  was done in [167,255]. A comprehensive generalization to arbitrary priors and element-wise output function is due to Rangan [256], who established the name *G-AMP* for the version with a general output channel. Many variants of this approach have then been considered (see for instance, without being exhaustive [251,257–259]).

As is clear from the earlier papers [164], AMP is actually very closely related to the TAP equations [46] presented in Section 4.5. The main difference between TAP and AMP is that TAP was originally derived for a model with Ising spins and pairwise interactions, whereas AMP is for continuous variables and  $N$ -body interactions.

#### 6.3.1. Relaxed-belief propagation

A detailed statistical physics derivation of the AMP algorithm was presented in [251] for the Gaussian output channel, and for general output as a part of the yet more general problem of matrix factorization in [260]. Here we shall describe the main elements necessary for the reader to understand where the resulting algorithm comes from.

We are given the posterior distribution

$$P(\mathbf{x} | \mathbf{y}, F) = \frac{1}{Z} \prod_{\mu=1}^M P_{\text{out}}(y_{\mu} | z_{\mu}) \prod_{i=1}^N P_X(x_i), \quad \text{where } z_{\mu} = \sum_{i=1}^N F_{\mu i} x_i, \quad (136)$$

where the matrix  $F_{\mu i}$  has independent random entries (not necessarily identically distributed) with mean and variance  $O(1/N)$ . This posterior probability distribution corresponds to a graph of interactions  $y_{\mu}$  between variables (spins)  $x_i$  called the *graphical model* as depicted in Figure 14.

A starting point in the derivation of AMP is to write the BP algorithm corresponding to this graphical model, as done originally in [261]. The matrix  $F_{\mu i}$  plays the role of randomly quenched disorder, the measurements  $y_{\mu}$  are planted disorder. As long as the elements of  $F_{\mu i}$  are independent and their mean and variance of order  $O(1/N)$  the corresponding system is a mean-field SG. In the Bayes-optimal case (i.e. when the prior is matching the true empirical distribution of the signal) the fixed point of BP with lowest free energy then provides the asymptotically exact marginals of the above posterior probability distribution.

For model such as Equation (134), where the interaction are not pairwise, the BP equation we presented in Section 3.2 need to be slightly modified, for examples, see [5], but the logic is very

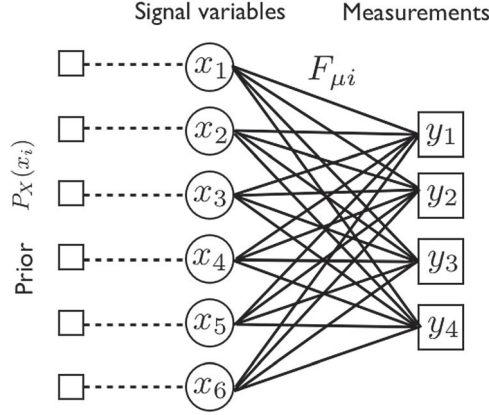


Figure 14. Factor graph of the linear estimation problem corresponding to the posterior probability (134). Circles represent the unknown variables, whereas squares represent the interactions between variables.

similar. BP still implements a message passing scheme between nodes in the graphical model of Figure 14, ultimately allowing one to compute approximations of the posterior marginals. Messages  $m_{i \rightarrow \mu}$  are sent from the variable-nodes (circles) to the factor-nodes (squared) and subsequent messages  $m_{\mu \rightarrow i}$  are sent from factor-nodes back to variable-nodes that corresponds to algorithm's current “beliefs” about the probabilistic distribution of the variables  $x_i$ . It reads

$$m_{i \rightarrow \mu}(x_i) = \frac{1}{z_{i \rightarrow \mu}} P_X(x_i) \prod_{\gamma \neq \mu} m_{\gamma \rightarrow i}(x_i), \quad (137)$$

$$m_{\mu \rightarrow i}(x_i) = \frac{1}{z_{\mu \rightarrow i}} \int \prod_{j \neq i} [dx_j m_{j \rightarrow \mu}(x_j)] P_{\text{out}} \left( y_\mu \left| \sum_l F_{\mu l} x_l \right. \right). \quad (138)$$

While being easily written, this BP is not computationally tractable, because every interaction involves  $N$  variables, and the resulting BP equations involve  $(N - 1)$ -uple integrals.

Two facts enable the derivation of a tractable BP algorithm: the central limit theorem, on the one hand, and a projection of the messages to only two of their moments (also used in algorithms such as Gaussian BP [262] or non-parametric BP [263]). This results in the so-called r-BP [254]: a form of equations that is tractable and involves a pair of means and variances for every pair variable-interaction.

Let us go through the steps that allow to go from BP to r-BP and first concentrate our attention to Equation (138). Consider the variable  $z_\mu = F_{\mu i} x_i + \sum_{j \neq i} F_{\mu j} x_j$ . In the summation in Equation (138), all  $x_j$ s are independent, and thus we expect, according to the central limit theorem,  $\sum_{j \neq i} F_{\mu j}^2 x_j^2$  to be a Gaussian random variable with mean  $\omega_{\mu \rightarrow i} = \sum_{j \neq i} F_{\mu j} a_{j \rightarrow \mu}$  and variance  $V_{\mu \rightarrow i} = \sum_{j \neq i} F_{\mu j}^2 v_{j \rightarrow \mu}$ , where we have denoted

$$a_{i \rightarrow \mu} \equiv \int dx_i x_i m_{i \rightarrow \mu}(x_i) \quad \text{and} \quad v_{i \rightarrow \mu} \equiv \int dx_i x_i^2 m_{i \rightarrow \mu}(x_i) - a_{i \rightarrow \mu}^2. \quad (139)$$

We thus can replace the multi-dimensional integral in Equation (138) by a scalar Gaussian one over the random variable  $z$ :

$$m_{\mu \rightarrow i}(x_i) \propto \int dz_\mu P_{\text{out}}(y_\mu | z_\mu) e^{-(z_\mu - \omega_{\mu \rightarrow i} - F_{\mu i} x_i)^2 / 2V_{\mu \rightarrow i}}. \quad (140)$$

We have replaced a complicated multi-dimensional integral involving full probability distributions by a single scalar one that involves only first two moments.

One can further simplify the equations: the next step is to rewrite  $(z - \omega_{\mu \rightarrow i} - F_{\mu i} x_i)^2 = (z - \omega_{\mu \rightarrow i})^2 + F_{\mu i}^2 x_i^2 - 2(z - \omega_{\mu \rightarrow i})F_{\mu i} x_i$  and to use the fact that  $F_{\mu i}$  is  $O(1/\sqrt{N})$  to expand the exponential

$$m_{\mu \rightarrow i}(x_i) \propto \int dz_{\mu} P_{\text{out}}(y_{\mu} | z_{\mu}) e^{-(z - \omega_{\mu \rightarrow i})^2 / 2V_{\mu \rightarrow i}} \times \left( 1 + F_{\mu i}^2 x_i^2 - 2(z - \omega_{\mu \rightarrow i})F_{\mu i} x_i + \frac{1}{2}(z - \omega_{\mu \rightarrow i})^2 F_{\mu i}^2 x_i^2 + o(1/N) \right). \quad (141)$$

At this point, it is convenient to introduce the output function  $g_{\text{out}}$ , defined via the output probability  $P_{\text{out}}$  as

$$g_{\text{out}}(\omega, y, V) \equiv \frac{\int dz P_{\text{out}}(y | z)(z - \omega) e^{-(z - \omega)^2 / 2V}}{V \int dz P_{\text{out}}(y | z) e^{-(z - \omega)^2 / 2V}}. \quad (142)$$

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

$$\frac{\int dz P_{\text{out}}(y | z)(z - \omega)^2 e^{-(z - \omega)^2 / 2V}}{V^2 \int dz P_{\text{out}}(y | z) e^{-(z - \omega)^2 / 2V}} = \frac{1}{V} + \partial_{\omega} g_{\text{out}}(\omega, y, V) + g_{\text{out}}^2(\omega, y, V). \quad (143)$$

Using definition (142), and re-exponentiating the  $x_i$ -dependent terms while keeping all the leading order terms, we obtain (after normalization) that the iterative form of Equation (138) reads:

$$m_{\mu \rightarrow i}(t, x_i) = \sqrt{\frac{A_{\mu \rightarrow i}^t}{2\pi N}} e^{-(x_i^2 / 2N)A_{\mu \rightarrow i}^t + B_{\mu \rightarrow i}^t(x_i / \sqrt{N}) - (B_{\mu \rightarrow i}^t)^2 / 2A_{\mu \rightarrow i}^t}, \quad (144)$$

with

$$B_{\mu \rightarrow i}^t = g_{\text{out}}(\omega_{\mu \rightarrow i}^t, y_{\mu}, V_{\mu \rightarrow i}^t) F_{\mu i} \quad \text{and} \quad A_{\mu \rightarrow i}^t = -\partial_{\omega} g_{\text{out}}(\omega_{\mu \rightarrow i}^t, y_{\mu}, V_{\mu \rightarrow i}^t) F_{\mu i}^2.$$

Given that  $m_{\mu \rightarrow i}(t, x_i)$  can be written with a quadratic form in the exponential, we just write it as a Gaussian distribution. We can now finally close the equations by writing Equation (137) as a product of these Gaussians with the prior

$$m_{i \rightarrow \mu}(x_i) \propto P_X(x_i) e^{-(x_i - R_{i \rightarrow \mu})^2 / 2\Sigma_{i \rightarrow \mu}}, \quad (145)$$

so that we can finally give the iterative form of the r-BP algorithm Algorithm 1 where we denote by  $\partial_{\omega}$  (respectively  $\partial_R$ ) the partial derivative with respect to variable  $\omega$  (respectively  $R$ ) and we define the “input” functions as

$$f_a(\Sigma, R) = \frac{\int dx x P_X(x) e^{-(x - R)^2 / 2\Sigma}}{\int dx P_X(x) e^{-(x - R)^2 / 2\Sigma}}, \quad f_v(\Sigma, R) = \Sigma \partial_R f_a(\Sigma, R). \quad (146)$$

### 6.3.2. From r-BP to G-AMP

While one can implement and run the r-BP, it can be simplified further without changing the leading order behavior of the marginals by realizing that the dependence of the messages on the

**Algorithm 1** relaxed-belief propagation (r-BP)**Input:**  $\mathbf{y}$ *Initialize:*  $\mathbf{a}_{i \rightarrow \mu}(t = 0), \mathbf{v}_{i \rightarrow \mu}(t = 0), t = 1$ **repeat**  r-BP Update of  $\{\omega_{\mu \rightarrow i}, V_{\mu \rightarrow i}\}$ 

$$V_{\mu \rightarrow i}(t) \leftarrow \sum_{j \neq i} F_{\mu j}^2 v_{j \rightarrow \mu}(t-1), \quad (147)$$

$$\omega_{\mu \rightarrow i}(t) \leftarrow \sum_{j \neq i} F_{\mu j} a_{j \rightarrow \mu}(t-1). \quad (148)$$

  r-BP Update of  $\{A_{\mu \rightarrow i}, B_{\mu \rightarrow i}\}$ 

$$B_{\mu \rightarrow i}(t) \leftarrow g_{\text{out}}(\omega_{\mu \rightarrow i}(t), y_{\mu}, V_{\mu \rightarrow i}(t)) F_{\mu i}, \quad (149)$$

$$A_{\mu \rightarrow i}(t) \leftarrow -\partial_{\omega} g_{\text{out}}(\omega_{\mu \rightarrow i}(t), y_{\mu}, V_{\mu \rightarrow i}(t)) F_{\mu i}^2. \quad (150)$$

  r-BP Update of  $\{R_{\mu \rightarrow i}, \Sigma_{\mu \rightarrow i}\}$ 

$$\Sigma_{i \rightarrow \mu}(t) \leftarrow \frac{1}{\sum_{v \neq \mu} A_{v \rightarrow i}(t)}, \quad (151)$$

$$R_{i \rightarrow \mu}(t) \leftarrow \Sigma_{i \rightarrow \mu}(t) \sum_{v \neq \mu} B_{v \rightarrow i}(t). \quad (152)$$

  AMP Update of the estimated partial marginals  $\mathbf{a}_{i \rightarrow \mu}(t), \mathbf{v}_{i \rightarrow \mu}(t)$ 

$$a_{i \rightarrow \mu} \leftarrow f_a(\Sigma_{i \rightarrow \mu}(t), R_{i \rightarrow \mu}(t)), \quad (153)$$

$$v_{i \rightarrow \mu} \leftarrow f_v(\Sigma_{i \rightarrow \mu}(t), R_{i \rightarrow \mu}(t)). \quad (154)$$

 $t \leftarrow t + 1$ **until** Convergence on  $\mathbf{a}_{i \rightarrow \mu}(t), \mathbf{v}_{i \rightarrow \mu}(t)$ **output:** Estimated marginals mean and variances:

$$a_i \leftarrow f_a\left(\frac{1}{\sum_v A_{v \rightarrow i}}, \frac{\sum_v B_{v \rightarrow i}}{\sum_v A_{v \rightarrow i}}\right), \quad (155)$$

$$v_i \leftarrow f_v\left(\frac{1}{\sum_v A_{v \rightarrow i}}, \frac{\sum_v B_{v \rightarrow i}}{\sum_v A_{v \rightarrow i}}\right). \quad (156)$$

target node is weak. This is exactly what we have done to go from the standard BP to the TAP equations in Section 4.5.

After the corresponding Taylor expansion the corrections add up into the so-called *Onsager reaction terms* [46]. The final G-AMP iterative equations are written in terms of means  $a_i$  and their variances  $c_i$  for each of the variables  $x_i$ . The whole derivation is done in a way that the

leading order of the marginal  $a_i$  are conserved. Given the BP was asymptotically exact for the computations of the marginals so is the G-AMP in the computations of the means and variances. Let us define

$$\omega_\mu^{t+1} = \sum_i F_{\mu i} a_{i \rightarrow \mu}^t, \quad (157)$$

$$V_\mu^{t+1} = \sum_i F_{\mu i}^2 v_{i \rightarrow \mu}^t, \quad (158)$$

$$\Sigma_i^{t+1} = \frac{1}{\sum_\mu A_{\mu \rightarrow i}^{t+1}}, \quad (159)$$

$$R_i^{t+1} = \frac{\sum_\mu B_{\mu \rightarrow i}^{t+1}}{\sum_\mu A_{\mu \rightarrow i}^t}. \quad (160)$$

First we notice that the correction to  $V$  are small so that

$$V_\mu^{t+1} = \sum_i F_{\mu i}^2 v_{i \rightarrow \mu}^t \approx \sum_i F_{\mu i}^2 v_i^t. \quad (161)$$

From now on we will use the notation  $\approx$  for equalities that are true only in the leading order.

From the definition of  $A$  and  $B$  we get

$$\begin{aligned} \Sigma_i &= \left[ - \sum_\mu F_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_{\mu \rightarrow i}^{t+1}, y_\mu, V_{\mu \rightarrow i}^{t+1}) \right]^{-1} \approx \left[ - \sum_\mu F_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}) \right]^{-1}, \\ R_i &= \left[ - \sum_\mu F_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}) \right]^{-1} \times \left[ \sum_\mu F_{\mu i} g_{\text{out}}(\omega_{\mu \rightarrow i}^{t+1}, y_\mu, V_{\mu \rightarrow i}^{t+1}) \right]. \end{aligned}$$

Then we write that

$$g_{\text{out}}(\omega_{\mu \rightarrow i}^{t+1}, y_\mu, V_{\mu \rightarrow i}^{t+1}) \approx g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}) - F_{\mu i} a_{i \rightarrow \mu}^t \partial_\omega g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}) \quad (162)$$

$$\approx g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}) - F_{\mu i} a_i^t \partial_\omega g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}). \quad (163)$$

So that finally

$$(\Sigma_i)^{t+1} = \left[ - \sum_\mu F_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}) \right]^{-1}, \quad (164)$$

$$R_i^{t+1} = ((\Sigma_i)^{t+1})^{-1} \times \left[ \sum_\mu F_{\mu i} g_{\text{out}}(\omega_\mu, y_\mu, V_\mu) - F_{\mu i}^2 a_i \partial_\omega g_{\text{out}}(\omega_\mu, y_\mu, V_\mu) \right] \quad (165)$$

$$= a_i^t + ((\Sigma_i)^{t+1})^{-1} \sum_\mu F_{\mu i} g_{\text{out}}(\omega_\mu^{t+1}, y_\mu, V_\mu^{t+1}). \quad (166)$$

Now let us consider  $a_{i \rightarrow \mu}$ :

$$a_{i \rightarrow \mu}^t = f_a(R_{i \rightarrow \mu}^t, \Sigma_{i \rightarrow \mu}) \approx f_a(R_{i \rightarrow \mu}^t, \Sigma_i) \quad (167)$$

$$\approx f_a(R_i^t, \Sigma_i) - B_{\mu \rightarrow i}^t f_v(R_i^t, \Sigma_i) \quad (168)$$

$$\approx a_i^t - g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t) F_{\mu i} v_i^t. \quad (169)$$

So that

$$\omega_\mu^{t+1} = \sum_i F_{\mu i} a_i^t - \sum_i g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t) F_{\mu i}^2 v_i^t = \sum_i F_{\mu i} a_i^t - V_\mu^t g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t). \quad (170)$$

---

**Algorithm 2** Generalized approximate message passing (G-AMP)

---

**Input:**  $\mathbf{y}$

*Initialize:*  $\mathbf{a}^0, \mathbf{v}^0, t = 1, g_{\text{out}, \mu}^0$

**repeat**

AMP Update of  $\omega_\mu, V_\mu$

$$V_\mu^t \leftarrow \sum_i F_{\mu i}^2 v_i^{t-1}, \quad (171)$$

$$\omega_\mu^t \leftarrow \sum_i F_{\mu i} a_i^{t-1} - V_\mu^t g_{\text{out}}^{t-1}. \quad (172)$$

AMP Update of  $\Sigma_i, R_i$  and  $g_{\text{out}, \mu}$

$$\Sigma_i^t \leftarrow \left[ - \sum_\mu F_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t) \right]^{-1}, \quad (173)$$

$$R_i^t \leftarrow a_i^{t-1} + (\Sigma_i^{t-1})^{-1} \sum_\mu F_{\mu i} g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t). \quad (174)$$

AMP Update of the estimated marginals  $a_i, v_i$

$$a_i^{t+1} \leftarrow f_a(\Sigma_i, R_i^{t+1}), \quad (175)$$

$$v_i^{t+1} \leftarrow f_v(\Sigma_i, R_i^{t+1}). \quad (176)$$

$t \leftarrow t + 1$

**until** Convergence on  $\mathbf{a}, \mathbf{v}$

**output:**  $\mathbf{a}, \mathbf{v}$

---

An important aspect of these equations to note is the index  $t - 1$  in the Onsager reaction term in Equation (172) that is crucial for convergence and appears for the same reason as in the TAP equations in Section 4.5. Note that the whole algorithm is comfortably written in terms of matrix multiplications only, this is very useful for implementations where fast linear algebra can be used.

### 6.3.3. The potential

Using the usual Bethe approach, one can also compute the corresponding potential, and show that these message passing equations actually aim to minimize the following Bethe free energy [264–266]:

$$\begin{aligned} \mathcal{F}^{\text{Bethe}}(\{\Sigma_i\}, \{R_i\}) &= \sum_i D_{\text{KL}}(Q_i \parallel P_X) + \sum_\mu D_{\text{KL}}(\mathcal{M}_\mu \parallel P_{\text{out}}) \\ &\quad + \frac{1}{2} \sum_\mu [\log 2\pi V_\mu^* + 1 + V_\mu^* \partial_\omega g_{\text{out}}(\omega_\mu^*, y_\mu, V_\mu^*)], \end{aligned} \quad (177)$$



with  $V_\mu^*$  and  $\omega_\mu^*$  satisfying their respective fixed-point conditions, and  $D_{\text{KL}}(Q||P) = \int [\log Q(x) - \log P(x)] dQ(x)$  being the Kullback–Leibler divergences between two probability distributions. The two Kullback–Leibler divergences above are taken with respect to probability distributions

$$\mathcal{M}_\mu(z; \omega_\mu^*, y_\mu, V_\mu^*) = \frac{1}{\mathcal{Z}_\mathcal{M}} P_{\text{out}}(y_\mu | z) \frac{1}{\sqrt{2\pi V_\mu^*}} e^{-(z - \omega_\mu^*)^2 / 2V_\mu^*}, \quad (178)$$

$$Q_i(x; T_i, \Sigma_i) = \frac{1}{\mathcal{Z}_\mathcal{Q}} P_X(x) e^{-(x - T_i)^2 / 2\Sigma_i}, \quad (179)$$

where  $\mathcal{Z}_\mathcal{M}$  and  $\mathcal{Z}_\mathcal{Q}$  are the corresponding normalizations. Clearly these two probability distributions are related to the denominators in Equations (146) and (142).

#### 6.3.4. Note on the convergence and parameter learning

When G-AMP converges, its performance in terms of mean-squared error is usually considerably better than the performance of other existing algorithms. The AMP equations are proven to converge in the idealized setting of large systems, random iid matrices  $F$  with zero mean entries, and when the prior  $P_X$  either corresponds to the empirical distribution of the actual signal  $\mathbf{x}^*$  or to the convex relaxation of the problem [167,255]. However, for mismatched signal distributions, for non-random matrices, positive mean random matrices or small system-sizes they might have unpleasant convergence issues.

Justified by the promise of a strikingly better performance, it is an important subject of current research how to make the G-AMP algorithm the most robust as possible, while keeping its performance and speed. Several recent works followed this line of research. In [267] it was identified why AMP has convergence problems for the simple case of iid matrices with elements of non-zero mean. For means of the matrix elements larger than some critical value, there is an antiferromagnetic-like (i.e. corresponding to a negative eigenvalue of the Hessian matrix) instability of the Nishimori line. Finite size effects always cause a slight mismatch between the true and the empirical distribution of the signal, and hence a slight deviation from the Nishimori line. The instability amplifies the deviation and causes the iteration of the AMP equation to fail and go far away from the Nishimori line. Even though we know that in the Bayes-optimal case the correct fixed point must satisfy the Nishimori conditions.

The simplest way to fix this problem for matrices of non-zero mean is the so-called mean removal as used in many of the early implementations and discussed in [266]. However, the instability of the Nishimori line is a problem that can cause failure of the algorithm in a wider range of settings than the non-zero mean matrices, for example, for some of the structured matrices. It is therefore useful to study other strategies to stabilize the Nishimori line in order to improve the convergence. We observed that randomization of the update is doing just that. Combining the sequential update while keeping track of the correct time indices, Ref. [171] designed the so-called Swept-AMP (SwAMP). Reference [251] studied the expectation–maximization learning of parameters for the compressed sensing problem when the number of parameters is small, typically finite while the system size grows. Since the parameter learning is precisely imposing the validity of the Nishimori conditions, this is another way that greatly stabilizes the Nishimori line and improves the convergence.

An additional approach to stabilize the Nishimori line is to slow down the iterations via damping, that is, taking a linear combination of the new and old values when updating the messages. Such damping can, however, slow down the iteration-process considerably. In order to avoid a drastic slow down we take advantage of the following related development.

The authors of [265] studied the Bethe free energy associated to G-AMP and wrote a variational form of the Bethe free energy (177) for which the G-AMP fixed points are local minima (not only generic stationary points such as saddles). This free energy can be optimized directly leading to a provably converging (but considerably slower) method to find the fixed points. This variational Bethe free energy can be used to slow down the iterations in a way to keep decreasing the free energy – this we call adaptive damping in [266]. Other groups suggested several closely related convergence improvements [264,268].

Combining the above implementation features – adaptive damping, sequential update, mean removal and parameter learning – leads to the current state-of-the art implementation of the G-AMP algorithm for linear estimation problems [269].

### 6.3.5. Examples of priors and outputs

The G-AMP algorithm is written for a generic prior on the signal  $P_X$  (as long as it is factorized over the elements) and a generic element-wise output channel  $P_{\text{out}}$ . The algorithm depends on their specific form only through the function  $f_a$  and  $g_{\text{out}}$  defined by Equations (146) and (142). It is useful to give a couple of explicit examples.

The sparse prior that is most commonly considered in probabilistic compressed sensing is the Gauss–Bernoulli prior, that is when in Equation (134) we have  $\phi(x) = \mathcal{N}(\bar{x}, \sigma)$  Gaussian with mean  $\bar{x}$  and variance  $\sigma$ . For this prior the input function  $f_a$  reads

$$f_a^{\text{Gauss–Bernoulli}}(\Sigma, T) = \frac{\rho e^{-(T-\bar{x})^2/2(\Sigma+\sigma)} \frac{\sqrt{\Sigma}}{(\Sigma+\sigma)^{3/2}} (\bar{x}\Sigma + T\sigma)}{(1-\rho) e^{-T^2/2\Sigma} + \rho \frac{\sqrt{\Sigma}}{\sqrt{\Sigma+\sigma}} e^{-(T-\bar{x})^2/2(\Sigma+\sigma)}}. \quad (180)$$

The most commonly considered output channel is simply AWGN (132). The output function then reads

$$g_{\text{out}}^{\text{AWGN}}(\omega, y, V) = \frac{y - \omega}{\Delta + V}. \quad (181)$$

As we anticipated above, the example of linear estimation that was most broadly studied in statistical physics is the case of the perceptron problem discussed in detail, for example, in [9]. In the perceptron problem each of the  $M$   $N$ -dimensional patterns  $F_\mu$  is multiplied by a vector of synaptic weights  $x_i$  in order to produce an output  $y_\mu$  according to

$$y_\mu = 1 \quad \text{if} \quad \sum_{i=1}^N F_{\mu i} x_i > \kappa, \quad (182)$$

$$y_\mu = -1 \quad \text{otherwise}, \quad (183)$$

where  $\kappa$  is a threshold value independent of the pattern. The perceptron is designed to classify patterns, that is, one starts with a training set of patterns and their corresponding outputs  $y_\mu$  and aims to learn the weights  $x_i$  in such a way that the above relation between patterns and outputs is satisfied. To relate this to the linear estimation problem above, let us consider the perceptron problem in the teacher–student scenario where the teacher perceptron generated the output  $y_\mu$  using some ground-truth set of synaptic weights  $x_i^*$ . The student perceptron knows only the patterns and the outputs and aims to learn the weights. How many patterns are needed for the student to be able to learn the synaptic weights reliably? What are efficient learning algorithms?

In the simplest case where the threshold is zero,  $\kappa = 0$  one can redefine the patterns  $F_{\mu i} \leftarrow F_{\mu i} y_\mu$  in which case the corresponding redefined output is  $y_\mu = 1$ . The output function in that

case reads

$$g_{\text{out}}^{\text{perceptron}}(\omega, V) = \frac{1}{\sqrt{2\pi}V} \frac{e^{-\omega^2/2V}}{H(-\frac{\omega}{\sqrt{V}})}, \quad (184)$$

where

$$H(x) = \int_x^\infty \frac{dt}{\sqrt{2\pi}} e^{-t^2/2}. \quad (185)$$

Note also that the perceptron problem formulated this way is closely related to what is called 1-bit compressed sensing in the signal processing literature [270,271].

In physics a case of a perceptron that was studied in detail is that of binary synaptic weights  $x_i \in \{\pm 1\}$  [17,239,272]. To take that into account in the G-AMP we consider the binary prior  $P_X(x) = [\delta(x-1) + \delta(x+1)]/2$  which leads to the input function

$$f_a^{\text{binary}}(\Sigma, T) = \tanh\left(\frac{T}{\Sigma}\right). \quad (186)$$

#### 6.4. Replica free energy, state evolution and phase transitions

##### 6.4.1. State evolution

The performance of the G-AMP algorithm above can be analyzed in the large size limit, for matrices with independent entries. The corresponding analysis was baptized *state evolution* in [168]. Rigorous results on the state evolution are provided in [167,255,273] for both AMP and G-AMP.

A statistical physics derivation of the state evolution in the spirit of the cavity method is given in [251] and is formally very related to the state evolution of the TAP equation in Section 4.5. The main point in the derivation of the state evolution is not to start from the G-AMP algorithm but from its message passing form, Equations (148)–(154). One then recalls that in BP we assumed conditional independence of incoming messages which translates into independence of terms present in sums of an extensive number of terms in the message passing. One then utilizes the central limit theorem, enabling a reduction to the mean and variance of these terms matters. One simply needs to keep track of the evolution of this mean and variance.

In statistical physics terms, the state evolution exactly corresponds to the derivation of the replica symmetric equations via the cavity method. Since the cavity and replica methods always yield the same result, the state evolution equations are exactly equivalent to the equations stemming from the replica method for the corresponding problem. For compressed sensing this is illustrated in [251]. For the perceptron problem all the replica symmetric equations and related results, for example, in [9], can be obtained as a special case of the G-AMP state evolution using the perceptron output.

Let us sketch briefly the derivation of state evolution for G-AMP. We follow the derivation of Kabashima et al. [260] for the specific case of linear estimation. In order to write the equations for the MMSE for a general output functions, we need to rewrite the output function using  $y = h(z, w)$  where  $w$  is a random variable distributed according to  $P_W(w)$ . Then

$$P_{\text{out}}(y | z) = \int dw P_W(w) \delta[y - h(z, w)]. \quad (187)$$

Now we consider the random variable  $V_\mu = \sum_i F_{\mu i}^2 v_i$ . On average its value is  $\sum_i v_i / N$  whereas its variance is  $o(1)$ . It thus concentrates around its mean as  $N \rightarrow \infty$  and therefore

$$V^{t+1} = \frac{1}{N} \sum_i v_i. \quad (188)$$

Let us now turn our attention to  $\omega_{\mu \rightarrow i} = \sum_{j \neq i} F_{\mu i} a_{i \rightarrow j}$  and  $z_{\mu \rightarrow i} = \sum_{j \neq i} F_{\mu i} s_i$ . These quantities are, according to the assumptions of the cavity method, sums of uncorrelated variables, so they must converge to (correlated) Gaussian variables with co-variances reading

$$\mathbb{E}[\omega^2] = \mathbb{E}[a^2] = q, \quad (189)$$

$$\mathbb{E}[z\omega] = \mathbb{E}[sa] = m, \quad (190)$$

$$\mathbb{E}[z^2] = \mathbb{E}[s^2] = q_0, \quad (191)$$

where we defined the order parameters  $m$  and  $q$  as the average correlation between the estimator  $a$  and the ground-truth signal  $s$ , and as a self-correlation for the estimator  $a$ , respectively.

Let us now see how  $R_i$  behaves (note the difference between the letters  $\omega$  and  $w$ ):

$$\frac{R_i}{\Sigma_i} = \sum_{\mu} B_{\mu \rightarrow i} = \sum_{\mu} F_{\mu i} g_{\text{out}}(\omega_{\mu \rightarrow i}, y_{\mu}, V_{\mu \rightarrow i}) \quad (192)$$

$$= \sum_{\mu} F_{\mu i} g_{\text{out}} \left( \omega_{\mu \rightarrow i}, h \left( \sum_{j \neq i} F_{\mu j} s_j + F_{\mu i} s_i, w_{\mu} \right), V \right) \quad (193)$$

$$= \sum_{\mu} F_{\mu i} g_{\text{out}} \left( \omega_{\mu \rightarrow i}, h \left( \sum_{j \neq i} F_{\mu j} s_j, w_{\mu} \right), V \right) + s_i \alpha \hat{m}, \quad (194)$$

where we define

$$\hat{m} = E_{\omega, z, w}[\partial_z g_{\text{out}}(\omega, h(z, w), V)]. \quad (195)$$

We further write

$$\frac{R_i}{\Sigma_i} = \mathcal{N}(0, 1) \sqrt{\alpha \hat{q}} + s_i \alpha \hat{m}, \quad (196)$$

with  $\mathcal{N}(0, 1)$  being a random Gaussian variables of zero mean and unit variance, and where

$$\hat{q} = \mathbb{E}_{\omega, z, w}[g_{\text{out}}^2(\omega, h(z, w), V)], \quad (197)$$

while the computation of  $\Sigma$  simply gives, thanks to concentration,

$$\frac{1}{\Sigma} = \alpha \hat{q}. \quad (198)$$

Now we can close all equations by computing  $q$  and  $m$  from their definitions (189)–(190) as

$$q = \mathbb{E}_s[\mathbb{E}_{R, \Sigma} f_a^2(\Sigma, R)], \quad (199)$$

$$m = \mathbb{E}_s[\mathbb{E}_{R, \Sigma} s f_a(\Sigma, R)]. \quad (200)$$

In the Bayes-optimal case, the Nishimori condition imposes  $q = m$  and thus provides further simplifications. In fact, one can show that  $\hat{q}$  is also equal to  $\hat{m}$  [260]. In this case, the state evolution leads to the asymptotic Bayes-optimal MMSE for the corresponding inference problem. Just as it did for the planted SG or the clustering of networks above. Since in linear estimation we are working with a fully connected graphical model the asymptotic analysis can be written in terms of only two scalar parameters [251].

Finally, writing explicitly the integrals, the state evolution reads

$$m^{t+1} = \int dx P_X(x) \int d\xi \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} f_a^2 \left( \frac{1}{\alpha \hat{m}^t}, x + \frac{\xi}{\sqrt{\alpha \hat{m}^t}} \right), \quad (201)$$

$$\hat{m}^t = - \int dw P_W(w) \int dp dz \frac{e^{-p^2/2m^t} e^{-(z-p)^2/2(\rho_0-m^t)}}{2\pi \sqrt{m^t(\rho_0-m^t)}} \partial_p g_{\text{out}}(p, h(z, w), \rho_0 - m^t), \quad (202)$$

where we denote  $\rho_0 = NE_F(F^2)E_{P_X}(x^2)$ . For the sparse prior (134) where the non-zero elements have the second moment equal to  $\bar{x}^2$  and the matrix  $F$  having iid elements of zero mean and variance  $1/N$  we get  $\rho_0 = \rho \bar{x}^2$ . For the usual AWGN output (132) the second equation simplifies considerably into

$$\hat{m}^t = \frac{1}{\Delta + \rho \bar{x}^2 - m^t}. \quad (203)$$

The mean-squared error of the AMP estimator is then  $\text{MSE} = \rho \bar{x}^2 - m^t$ . To evaluate the performance of AMP initialized in a way unrelated to the unknown signal, the state evolution is iterated until convergence starting from  $m^{t=0} = \rho^2 \bar{x}^2$ , where  $\bar{x}$  is the mean of the distribution  $\phi(x)$  in Equation (134).

#### 6.4.2. Free energy

The same procedure we use to derive the state evolution can be used to write a scalar expression for the Bethe free energy (177). Alternatively, the same expression can be obtained using the replica method, as, for example, in [260]. The average Bethe free energy for the generalized linear estimation problem can be written as  $f^{\text{Bethe}} = \min_{m, \hat{m}} \Phi(m, \hat{m})$  where

$$\begin{aligned} \Phi(m, \hat{m}) = & \alpha \int dy \mathcal{D}\xi \mathcal{D}s P_{\text{out}}(y | s\sqrt{\rho_0 - m} + \xi\sqrt{m}) \log \left[ \int dx P_{\text{out}}(y | x\sqrt{\rho_0 - m} + \xi\sqrt{m}) \right] \\ & - \frac{m\hat{m}}{2} + \int \mathcal{D}\xi \mathcal{D}s P_X(s) e^{-\hat{m}s^2/2 + \xi s\sqrt{\hat{m}}} \log [dx P_X(s) e^{-\hat{m}x^2/2 + \xi x\sqrt{\hat{m}}}], \end{aligned} \quad (204)$$

The free energy can be also written as a function of the MSE  $E$ . In the case of the AWGN channel with variance  $\Delta$ , for instance, it reads  $f^{\text{Bethe}} = \min_E \Phi(E)$  where [251]

$$\begin{aligned} \Phi(E) = & -\frac{\alpha}{2} - \frac{\alpha}{2} \log(\Delta + E) - \frac{\alpha(\rho \bar{x}^2 - E)}{2(\Delta + E)} \\ & + \int ds P_X(s) \int \mathcal{D}\xi \log \left\{ \int dx e^{\alpha/(\Delta+E)x(s-x/2) + \xi x(\sqrt{\alpha}/\sqrt{\Delta+E})} P_X(x) \right\}. \end{aligned} \quad (205)$$

Actually, this formula is, again, very close to the one discussed in the context of CDMA by Tanaka [274] and Guo and Verdú [275]. An equivalent form has been derived for compressed sensing by Wu and Verdú [276]. Much more can be done with the replica method in this context (see, e.g. [277, 278]).

#### 6.4.3. Phase transitions

What we want to discuss here is again the appearance of the very generic phenomena we have discussed already in the first chapter, and illustrated in Figure 1: the presence of two different transitions when trying to solve the Bayesian optimal problem.

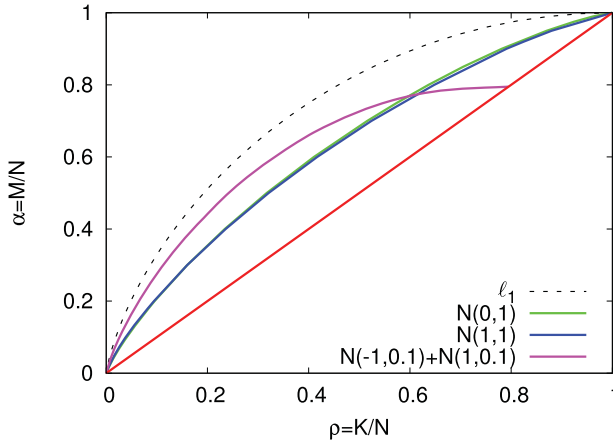


Figure 15. Phase diagram for the AMP reconstruction in the optimal Bayesian case when the signal model is matching the empirical distribution of signal elements. The elements of the  $M \times N$  measurement matrix  $F$  are iid variables with zero mean and variance  $1/N$ . The spinodal transition  $\alpha_s(\rho)$  is computed with the state evolution and plotted for the following signal distributions:  $\phi(x) = \mathcal{N}(0, 1)$  (green),  $\phi(x) = \mathcal{N}(1, 1)$  (blue),  $\phi(x) = [\mathcal{N}(-1, 0.1) + \mathcal{N}(1, 0.1)]/2$  (magenta). The data are compared to the Donoho–Tanner phase transition  $\alpha_{\ell_1}(\rho)$  (dashed) for  $\ell_1$  reconstruction that does not depend on the signal distribution, and to the theoretical limit for exact reconstruction  $\alpha = \rho$  (red). Figure taken from [251].

In Figure 15 we plot down to what measurement rate  $\alpha = M/N$  AMP recovers a signal of density  $\rho$  for several distributions of non-zero signal elements  $\phi(x)$ . We observe a first-order phase transition that depends on the distribution of non-zeros.

Interestingly, the corresponding line is always better than the  $\ell_1$  transition, as shown in [279], where the distribution maximizing the spinodal line for Bayes-optimal inference was computed. This is to be expected, since we might hope that the spinodal for the Bayesian optimal problem is the “best” spinodal, though there is no generic proof of this.

One might wonder how a measurement noise is affecting this diagram? Reference [251] further studied how the optimal inference performance is influenced by measurement noise, mismatch between the prior and the actual signal distribution, how the performance changes under expectation maximization learning of the parameters, such as  $\rho$  or the mean  $\bar{x}$  and the variance  $\sigma$  of the signal distribution. Reference [280] explored the role of approximately sparse signals when the  $\delta(x_i)$  in Equation (132) is replaced by a narrow Gaussian. The influence of an uncertainty in the elements of the matrix  $F$  was studied in [281]. As expected from statistical physics, where in the presence of disorder first-order phase transitions become weaker, second order or disappear entirely, the same happens to the first-order phase transition in compressed sensing under presence of various kinds of noise. Qualitative results are illustrated in [251, 280, 281].

Existence of phase transition is not restricted to compressed sensing. In fact, the very same phenomenology has been discussed and identified as early as in the 1990s in the context of learning a rule via a perceptron with binary weights [8, 9, 17, 242]. In the present setting, this corresponds to a binary prior for the signal, and a perceptron channel given by Equation (166). In this case, it becomes possible to identify perfectly the signal when the ratio between the number of examples  $M$  and the number of variables  $N$  is larger than  $\alpha \approx 1.245$ , a result first derived by Derrida and Gardner [17]. It is only beyond  $\alpha \approx 1.49$ , however, that the G-AMP algorithm actually succeed in finding the hidden rule (i.e. the planted weights). All these classic results can thus be recovered by the present, unifying, approach. Note that, for this problem, the G-AMP

$$\begin{pmatrix} y \\ \vdots \end{pmatrix} = \begin{pmatrix} F & 0 \\ 0 & \end{pmatrix} \times \begin{pmatrix} s \\ \vdots \end{pmatrix}$$

: unit coupling  
 : coupling  $f_1$   
 : coupling  $f_2$   
 : no coupling (null elements)

Figure 16. Example of a spatially coupled measurement matrix  $F$  for compressed sensing. The elements of the matrix are taken independently at random, their variance has the block diagonal structure as illustrated in the figure. The figure is taken from [251].

(or TAP) equations has been written and proposed as early as in 1989 in [242]. The more recent application of either r-BP [253] or G-AMP to actual data [282] has also been considered.

Another classical example of this approach with a similar phenomenology is given by the classical work of Tanaka and Kabashima in the context of CDMA [164,274].

### 6.5. Spatial coupling

As discussed in Section 4.2, if nothing is changed about the graphical model corresponding to compressed sensing, the first-order phase transition presented in Figure 15 causes a barrier that is conjectured to be unbeatable by any polynomial algorithm. In an idealized setting of compressed sensing, however, the design of the measurement matrix  $F_{\mu i}$  is entirely up to us. Thanks to this, compressed sensing belongs to the class of problems where spatial coupling can be successfully applied. This was the main point of ref. [250], where spatially coupled measurement matrices were designed that allowed AMP to reconstruct the signal in compressed sensing anytime the measurement rate  $\alpha$  is larger than the density of non-zero elements in the signal  $\rho$ , i.e. all the way above the red line in Figure 15. This *threshold saturation* was soon after proven rigorously in [252]. We should note that spatial coupling for compressed sensing was considered even before in [283], but in that work the observed improvement was limited and it did not seem possible to achieve the information theoretical limit  $\alpha = \rho$ .

A spatially coupled matrix suggested in [250] is illustrated in Figure 16. The physical mechanism of why the spinodal transition disappears with such spatially coupled measurement matrices is exactly the same as for the spatially coupled Currie–Weiss model discussed in Section 4.3. The first block of the matrix has a larger effective measurement rate and plays the role of a nucleation seed. The interactions in the blocks next to the diagonal then ensure that the seed grows into the whole system. Reference [251] further discusses the choice of parameters of the spatially coupled measurement matrices, other related designs and the effect of noise. With the goal of optimizing the spatially coupled matrices the properties of the nucleation and necessary properties of the seed in compressed sensing were studied in [284].

There are several problems in the literature closely formally related to compressed sensing to which spatial coupling was applied. One of them is, again, the CDMA defined in Section 6.1.



When the spreading code is not random iid but correspond to a spatially coupled random matrix then the performance of the CDMA system can be further improved [285,286].

Another problem formally very closely related to compressed sensing is the real-valued error correction where the codeword is corrupted by gross errors on a fraction of entries and a small noise on all the entries. In [287] it was shown that the error correction and its robustness toward noise can be enhanced considerably using AMP and spatial coupling.

A similar application of these ideas has allowed for the development of the spatially coupled sparse superposition codes [288–292]. These are efficient coding schemes using exactly the compressed sensing setting with a twist: the variables are  $L$ -dimensional and are forced to point into one corner of the  $L$ -hypercube. These codes, with the associated AMP decoder, have been shown to be fast and reliable, and actually reach the Shannon limit for the AWGN channel in the large  $L$  limit. It is actually expected that they are Shannon achieving for any memoryless channel.

## 6.6. Non-random matrices and non-separable priors

### 6.6.1. Structured priors

AMP as described above uses the empirical distribution of signal elements. A natural question is whether its performance can be improved further if more information is known about the signal. There are works that have investigated this question and combined AMP with more complex, structured priors. Utilizing such structured priors is the key to leveraging many of the advancements recently seen in statistical signal representation. For instance techniques such as hybrid AMP [293,294] have shown promising results. In those works, an additional graphical model is introduced in order to take into account the correlation between variables. The improvement, for instance for image denoising, can be spectacular.

Another possible approach to structured priors is not to model the correlations directly, but instead to utilize a bipartite construction via hidden variables, as in the restricted Boltzmann machine (RBM) [295,296]. If a binary RBM can be trained to model the support pattern of a given signal class, then the statistical description of the RBM admits its use within AMP, as was shown recently in [297]. This is particularly interesting since RBMs are the building blocks of “deep belief networks” [298] and have recently sparked a surge of interest, partly because of the efficient algorithms developed to train them (e.g. contrastive divergence (CD) [296]). The possibility, shown in [297], to incorporate deep-learned priors into generalized linear problems such as compressed sensing appears very promising.

### 6.6.2. Orthogonal matrices

Using non-random matrices for the linear projection is crucial to explore the range of applicability. The  $\ell_1$ -minimization-based approaches provably work well for a much larger class of matrices than random ones [237]. It is therefore important to explore performance of the approach stemming from statistical physics for more general matrices. There has been a collection of works in this direction, for instance in the analysis of the so-called perceptron learning with correlated pattern (which form the sensing matrix) [241,243].

As we have mentioned in Section 4.5, there are many works in the context of the Ising model that have extended the TAP approach for more generic random matrices. Similar contributions have recently emerged in the context of AMP as well: the recent work on the so-called S-AMP algorithm that extends the AMP algorithm to general matrix ensembles when there is a certain well-defined large system size limit, based on the so-called  $S$ -transform of the spectrum of the measurement matrix [299]. Older works, in fact, already used a similar approach for the CDMA problem [300]. Analysis of performances in this direction is also doable [301,302]. This shows

that there are still interesting open problems in this direction and we feel that there is still much more to be done in the context of AMP beyond random matrices.

### 6.6.3. Reconstruction in discrete tomography

Signal reconstruction in X-ray computed tomography (CT) is another problem that belongs to the class of linear estimation. The matrix  $F_{\mu i}$  corresponding to CT has a very particular structure. Classical reconstruction algorithms such as filtered back-projection require the corresponding linear system to be sufficiently determined. Therefore, the number of measurements needed is generally close to the number of pixels to be reconstructed. Nevertheless, many applications would benefit from being able to reconstruct from a smaller number of angles. Reference [303] designed and tested a message passing algorithm that is adapted to the measures performed in CT and is able to provide a reliable reconstruction with a number of measurements that is smaller than what was required by previous techniques. This algorithm is another example of how inspiration from physics can provide a promising algorithmic approach.

Each measurement in X-ray CT is well approximated by a sum of absorption coefficient of the material under investigation along a straight line of the corresponding X-ray. Let us discretize the object under study into small elements, and assign a discrete value  $x_i \in \{1, \dots, q\}$  of the absorption coefficient to every small element  $i$ . One component of the measurement in CT is then

$$y_\mu = \sum_{i \in \mu} x_i, \quad (206)$$

where the sum is over all elements that lie on the line  $\mu$ . The structural information that in general allows us to decrease the number of measurement is that two pixels that lie next to each other mostly have the same value. This expectation is represented in the following posterior distribution

$$P(\mathbf{x} | \mathbf{y}) = \frac{1}{Z} \prod_{\mu=1}^M \left[ \delta \left( y_\mu - \sum_{i \in \mu} x_i \right) e^{J_\mu \sum_{(ij) \in \mu} \delta_{x_i, x_j}} \right], \quad (207)$$

where the notation  $(ij) \in \mu$  means the elements  $i$  and  $j$  are next to each other on the line corresponding to measurement  $\mu$ ,  $J_\mu$  is the interaction constant.

Reference [303] writes the BP algorithm for estimating marginals of this posterior distribution. However, one update of this BP involves intractable sums over all configurations of variables on one line. We realize that these sums correspond to the partition functions of one-dimensional Potts model with fixed magnetization. A one-dimensional line is a tree and for trees partition functions can be computed exactly with BP. We use this to design an algorithm that uses BP within BP. The resulting implementation is comparably fast to the competing convex relaxation-based algorithms and provides better reconstruction and robustness to noise. We believe that overall the algorithm suggested in [303] provides a promising perspective for reconstruction in X-ray CT.

## 6.7. An application in physics: optics in complex media

We would like to close the loop with physics and mention one explicit physics experiment in the field of compressed optics where AMP and related tools have been especially useful.

Wave propagation in complex media is a fundamental problem in physics, be it in acoustics, optics, or electromagnetism. In optics, it is particularly relevant for imaging applications. Indeed, when light passes through a multiple scattering medium, such as a biological tissue or a layer of paint, ballistic light is rapidly attenuated, preventing conventional imaging techniques, and

random scattering events generate a so-called speckle pattern that is usually considered useless for imaging.

Recently, however, wavefront shaping using spatial light modulators has emerged as a unique tool to manipulate multiply scattered coherent light, for focusing or imaging in scattering media [304]. This makes it possible to measure the so-called transmission matrix of the medium [305], which fully describes light propagation through the linear medium, from the modulator de vice to the detector. Interestingly, these transmission matrices, due to the fact that they account for the multiple scattering and interference, are essentially iid random matrices [305], which is exactly what AMP techniques have been designed for.

An important issue in most such recent experiments lies in accessing the amplitude and phase of the output field, that in optics usually requires a holographic measurement, i.e. the use of a second reference beam. The phase and amplitude of the measured field can then be extracted by simple linear combinations of interference patterns with a phase shifted or off-axis reference. This however poses the unavoidable experimental problem of the interferometric stability of the reference arm. In order to be able to avoid entirely the use of the reference beam, we need to solve a problem that looks like compressed sensing: finding  $\mathbf{x}$  (sparse or binary) such that  $\mathbf{y} = |\mathbf{F}\mathbf{x}|$  where  $\mathbf{F}$  is complex and iid. Combining the G-AMP and the SwAMP approach for phase retrieval [171,306], it has been possible to precisely do this, to calibrate efficiently, and to perform a reference-less measurement of the transmission matrix of a highly scattering material. Within this framework, the work of Drémeau et al. [307] showed that the transfer matrix can still be retrieved for an opaque material (a thin layer of white paint), and can then be used for imaging or focusing, using the opaque complex material as a lens. This was experimentally validated.

As a result the full complex-valued transmission matrix of a strongly scattering material can be estimated, up to a global phase, with a simple experimental setup involving only real-valued inputs and outputs in combination with the techniques presented in this review. We believe this illustrates the power of the techniques presented in this review, and shows that statistical physics does not only provide interesting perspective to inference problems, but also that inference techniques can be useful for physics experiments in return.

### 6.8. Beyond linear estimation: matrix factorization

The technics that have been used in the context of (generalized) linear estimation are not limited to this setting. In fact, AMP can also be applied to the more challenging setting of matrix factorization.

Decomposing a matrix into a product of two matrices of given properties is another problem with an huge range of applications. Mathematically stated, we are given a noisy measurement of a matrix  $\mathbf{Y}$  of dimension  $M \times N$  and we aim to decompose it into a product  $\mathbf{Y} = \mathbf{D}\mathbf{X}$ , where  $\mathbf{D}$  is a  $M \times R$  matrix, and  $\mathbf{X}$  is a  $R \times N$  matrix. Depending on the application the problem is subject to various constraints such as sparsity of  $\mathbf{D}$  or  $\mathbf{X}$ , low rank  $R$ , or specific models for the noise. Such constraints turn the matrix factorization into an algorithmically extremely challenging problem. Let us list several examples of potential applications.

- *Dictionary learning*: Many modern signal processing devices take a great advantage of the fact that almost every class of signals of interest has a sparse representation, i.e. there exists a basis such that the observed signal can be written as a sparse linear combination of atoms (i.e. columns of the corresponding matrix) from this basis. Widely known, studied and used examples include the wavelet basis for images, of Fourier basis for acoustic signals. For other examples of interest, especially those which utilize uncommon or novel data types, a suitable basis for sparse representation may not yet be known. The goal of dictionary

learning [308–310] is to learn such a basis purely from samples of the data. Naturally the more samples are available the easier is this task. The challenge is hence to design an algorithm that would be able to learn the dictionary efficiently from the smallest possible number of samples. In the mathematical statement of the problem the  $N$  columns of matrix  $Y$  represent the  $N$  samples of the data for which we aim to find a sparse representation.  $M$  is then the dimension of each of the data points. The dictionary learning problem is a decomposition of the matrix  $Y$  into a product  $Y = DX + W$ , where  $D$  is the dictionary of  $R$  so-called atoms, and  $X$  is the sparse representation of the data, having only a fraction  $\rho$  of non-zero elements in each column.  $W$  is a possible noise that may be taking into account the approximative nature of the sparse representation.

- *Feature learning in neural network:* Arguably one of the most impressive engineering success of the last decade is the design of the so-called *deep-learned* neural networks [311–313]. Using these, computers are now able to recognize people on video, to tell a dog from a cat on a picture, to process speech efficiently and even to answer complicated questions. At the roots of this efficiency is the ability to learn features. The simplest building block of deep neural networks can be seen as a matrix factorization problem with an output  $P_{\text{out}}(y | z)$  corresponding to the activation functions.
- *Blind source separation:* A typical example of blind source separation [314,315] are  $M$  sensors recording noise at a party during time  $N$ . We then aim to separate the speech of the  $R$  participants of the party based on these recordings. Recording at one of the sensors is an unknown linear combination of the speech of all the present people. A particularly challenging, but also very interesting application-wise, is when the number of sensors is smaller than the number of people,  $M < R$ . The individual speeches may then be in principle recovered only if they can be represented as sparse in some known basis.
- *Sparse PCA:* In PCA the aim is to approximate the matrix  $Y$  by a low-rank matrix, that is, columns of  $Y$  are written as a linear combination of only few elements. This reduces the information needed to describe the data contained in  $Y$ . This is usually done via the singular value decomposition (SVD). In some applications it is desirable that the linear combination to be sparse [316,317], thus further reducing the necessary information. The sparsity constraint cannot be straightforwardly included in the SVD and hence alternative algorithmic approaches are needed.

### 6.8.1. Learning random sparse dictionaries

The matrix factorization problem can again be studied within the teacher–student scenario: the teacher generates the matrices  $F$  and  $X$  with random iid entries, from some probability distribution (possibly sparse) and provides to the student element-wise measurement of the product  $FX$  via an output function  $P_{\text{out}}(Y | FX)$ . Let us consider the dimensions  $M, N, R$  to be all of the same order. Both the AMP [260,318–320] and the replica approach [260,318,321,322] can be applied to such problems. In this the problem of identifiability of  $F$  and  $X$  from  $Y$ , the corresponding sample complexity and the associated Bayes-optimal mean-squared error can be derived. We refer to these works for the phase diagrams for special cases of the problem such as the sparse dictionary learning, blind source separation, sparse PCA, robust PCA or matrix completion.

These results are very striking because in a number of relevant cases (including the dictionary learning and source separation) it is observed that the ideal performance predicted by this approach is far better than the one achieved by current algorithms. This suggests that there is a lot to gain in algorithmic efficiency in such problems. Making the algorithm of [260,320] more

robust along the lines described above for AMP for linear estimation and applying it to more realistic data is a very promising subject of future work.

### 6.8.2. Low-rank decomposition

In some part of the applications mentioned above (e.g. the PCA) the dimension  $R$ , corresponding to the rank of matrix  $Y$ , is considered as very small. In those cases it is reasonable to consider the limit where  $M$  and  $N$  are comparable and both  $N, M \rightarrow \infty$  whereas  $R = O(1)$ . The graphical model corresponding to this case has  $R$ -dimensional variables and pairwise interactions. Problems that can be represented via this graphical model include variations of PCA, submatrix localization problem, biclustering, clustering.

Again AMP type of algorithm can be written for a general element-wise output function  $P_{\text{out}}(Y | FX)$ , and the associated state evolution leads to an analysis of the phase diagram, phase transitions and optimal mean-squared errors. Interestingly in this approach the prior distribution can be any  $R$ -variate distribution, thus allowing for instance for priors representing membership to one or more clusters.

A problem corresponding to the rank one matrix factorization, clustering in the Gaussian mixture model, was studied long time ago using the replica method [323–325]. It was presented as a prototypical example of unsupervised learning, see, for example, Chapter 8 in the textbook [4]. Note, however, that many more settings closer to current unsupervised learning architectures should be analyzed. The revival of interest in low-rank matrix factorization was linked with the development of the corresponding AMP algorithm by Rangan and Fletcher for rank  $R = 1$  [326], and Matsushita and Tanaka for general rank [327] followed by rigorous work of [328]. Recently, a version for generic rank and generic output (as in G-AMP) has been presented and analyzed in [194,329]. The resulting algorithm when tested numerically has good convergence properties and these works hence open the stage to further studies and applications along the line discussed in this review.

As a final comment, it is to be noted that the phase transitions observed in the matrix factorization setting are again of the same kind that we have described all along this review, schematically in Figure 1. In many problems, the presence of distinct algorithmic and information theoretic thresholds is observed. In this setting, it is particularly interesting to characterize when spectral methods [330], the most widely used approach for low rank factorization, are optimal, and when they are not. As shown in, for example [149,194,329,331], there are a number of situations where they are sub-optimal. This is a valuable information one can get from the physics approach, that should motivate the creation of new algorithms approaching the best possible performance.

## 7. Perspectives

This review discusses recent progress in the statistical physics approach to understanding of different problems of statistical inference. Just as theoretical physics often focuses on understanding of idealized models that represent in a simplified way the salient features of realistic settings, we focus on the teacher–student setting under which the Bayes-optimal inference can be analyzed without ambiguity, and the phase transition can be identified and discussed.

The main concept that we presented in detail is the mathematical relation between the Bayes-optimal inference and properties of disordered systems on the co-called Nishimori line. We discussed in detail various phase transitions that arise in the Bayes-optimal inference and related them to statistical and computational thresholds in the corresponding inference problems. The resulting picture, that we explain in detail on the simple example of the *planted* SG, is relevant for a large class of inference problems. In later sections we illustrate this picture on the problem

of clustering of sparse networks and on the problem of generalized linear estimation. In today's scientific landscape the number of very interesting data-driven problems is considerable and, in the authors' opinion, in a considerable number of them the strategy presented in this review will bring exciting new results.

Let us recall here the main theme of the present review. For all the presented problems, the central scientific questions that one can attempt to answer are: (1) Under what conditions is the information contained in the measurements sufficient for a satisfactory inference to be possible? (2) What are the most efficient algorithms for this task? In the probabilistic setting discussed here, what is striking is that all of these problems have a very similar phenomenology, that also appeared earlier when physicists studied error correcting codes and perceptron learning, and that often these two questions can be precisely answered. When the amount of information is too low a successful inference of the signal is not possible *for any algorithm*: the corresponding information is simply insufficient. On the other hand, for large enough amount of data, inference is possible, and the two regime are separated by a sharp phase transition. Finally, and perhaps more importantly, there is often (as in first-order transition in physics), an intermediate regime where a successful inference is in principal possible but algorithmically hard. We have shown many example where these transitions can be computed. A first-order phase transition is always linked to appearance of computational hardness, whereas a second-order phase transition is not.

There is a deep value in this knowledge: if these transitions were not known, one could not know how good are the state-of-the-art algorithms. To give an example, it is the very knowledge of the precise Shannon bound that is driving research in information theory: if we know that we could be better in principle, when it is worth working on better algorithms. As we have discussed in the very last chapter, for instance, present day matrix factorization algorithms are often very far from optimality.

It is thus natural that, next to the theoretical understanding, an important outcome of the research reviewed in this manuscript are new algorithmic ideas that are applicable beyond the idealized setting for which the presented theoretical results hold. This translation of theoretical physics calculations into algorithmic strategies is very fruitful, and promising direction of research. What is, perhaps, more challenging is to push such algorithmic ideas into actual applications. From a point of view of a physicist, some of the problems presented here may seem very applied, but in reality the research presented here is positioned on the theoretical side of the fields of statistics, machine learning or signal processing. There is still a large step toward actual applications. Surely this is a very exciting direction to explore and many of the researchers active in this field do explore it.

Let us now discuss some of the limitations and challenges related to the results and concepts presented in this review.

On the algorithmic side a large part of the algorithms that stem from the research presented in this review involve some kind of message passing derived under various assumptions of correlation decay. These assumptions rarely hold of real datasets and hence sometimes this causes a failure of the corresponding algorithm. In our opinion the theoretical potential of these message passing algorithms is very large, but much work needs to be done in order to render these algorithms more robust and suitable for a generic problem.

Computer science and related fields mostly aim at understanding the worst possible setting or to prove something under the weakest possible conditions. Analysis in a setting such as the teacher–student scenario usually does not seem generic enough. In machine learning and more applied parts of computer science the focus is on a couple of benchmark-datasets (such as, e.g. the MNIST database). Simple models that aim to represent reasonably well the generic situation are a domain of physical sciences. This is how progress in physics is made, and in our opinion this is transferable to other fields as well.



Physicists can use with large effectiveness tools and methods from statistical physics of disordered systems such as glasses and SG. The advantage of these methods is that, unlike any other existing generic method, they are able to describe the Bayes-optimal probabilistic inference. The disadvantage is that they are not fully rigorous and are therefore sometimes viewed with dismissal or skepticism by mathematically oriented researchers. On the other hand, rigorous proofs of results obtained with the cavity and replica method are a great mathematical challenge and many partial results were obtained that have led to new mathematical tools making the whole probability theory more powerful. One of the goals of the presented research is to transfer the physics reasoning and insight to mathematicians in order to help them to craft plausible proof strategies.

## Acknowledgments

We would like to express our gratitude to all our colleagues with whom many of the results presented here have been obtained. In particular, we wish to thank Maria-Chiara Angelini, Jean Barbier, Francesco Caltagirone, T. Castellani, Michael Chertkov, Andrea Crisanti, Leticia F. Cugliandolo, Laurent Daudet, Aurelien Decelle, Angelique Drémeau, Laura Foini, Silvio Franz, Sylvain Gigan, Emmanuelle Gouillart, Jacob E. Jensen, Yoshiyuki Kabashima, Brian Karrer, Lukas Kroc, Srinivas Kudekar, Jorge Kurchan, Marc Lelarge, Antoine Liutkus, Thibault Lesieur, Luca Leuzzi, André Manoel, David Martina, Marc Mézard, Andrea Montanari, Cristopher Moore, Richard G. Morris, Elchanan Mossel, Joe Neeman, Mark Newman, Hidetoshi Nishimori, Boshra Rajaei, Sundeep Rangan, Joerg Reichardt, Federico Ricci-Tersenghi, Alaa Saade, David Saad, Ayaka Sakata, Francois Sausset, Christian Schmidt, Christophe Schülke, Guilhem Semerjian, Cosma R. Shalizi, David Sherrington, Alan Sly, Phil Schniter, Eric W. Tramel, Rüdiger Urbanke, Massimo Vergassola, Jeremy Vila, Xiaoran Yan, Sun Yifan, Francesco Zamponi, Riccardo Zecchina and Pan Zhang.

A part of this manuscript has been written during a visit to UC Berkeley as part of the semester long program on “Counting complexity and phase transitions”, at the Simons Institute for the Theory of Computing, which we thank for the kind hospitality.

Finally, we also wish to thank Alena and Julie for their continuous support and for letting us work, if only from time to time, on this review.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

Part of the research leading to the presented results has received funding from the European Research Council under the European Union’s 7th Framework Programme (FP/2007-2013/ERC Grant Agreement 307087-SPARCS).

## References

- [1] S.B. McGrayne, *The Theory That Would Not Die: How Bayes Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press, New Haven, 2011.
- [2] P. Grassberger and J.P. Nadal, *From Statistical Physics to Statistical Inference and Back*, Kluwer Academic, Berlin, 1994.
- [3] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, Oxford University Press, Oxford, 2001.



- [4] A. Engel and C.P.L.V.D. Broeck, *Statistical Mechanics of Learning*, Cambridge University Press, New York, NY, 2001.
- [5] M. Mézard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, Oxford, 2009.
- [6] F. Krzakala, F. Ricci-Tersenghi, L. Zdeborová, R. Zecchina, E.W. Tramel and L.F. Cugliandolo, *Statistical Physics, Optimization, Inference and Message-passing Algorithms*, Oxford University Press, Oxford, 2015.
- [7] M. Mézard, G. Parisi and M.A. Virasoro, *Spin-glass Theory and Beyond*, Lecture Notes in Physics, Vol. 9, World Scientific, Singapore, 1987.
- [8] H.S. Seung, H. Sompolinsky and N. Tishby, Phys. Rev. A 45 (1992) p.6056.
- [9] T.L.H. Watkin, A. Rau and M. Biehl, Rev. Mod. Phys. 65 (1993) p.499.
- [10] T. Hastie, R. Tibshirani, J. Friedman and J. Franklin, Math. Intelligencer 27 (2005) p.83.
- [11] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, Cambridge, 2012.
- [12] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, 2003.
- [13] N. Surlas, Nature 339 (1989) p.693.
- [14] A. Montanari and R. Urbanke, *Complex Systems*, Les Houches Lecture Notes, Elsevier, Amsterdam, 2007, p.69.
- [15] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer Science & Business Media, Berlin, 2013.
- [16] B. De Finetti, *Theory of Probability. A Critical Introductory Treatment*, John Wiley & Sons, Hoboken, 1979.
- [17] E. Gardner and B. Derrida, J. Phys. A: Math. Gen. 22 (1989) p.1983.
- [18] E.T. Jaynes, Phys. Rev. 106 (1957) p.620.
- [19] C.E. Shannon, Bell Syst. Tech. J. 27 (1948) p.379, 623–655.
- [20] S. Kirkpatrick, C.D. Gelatt Jr. and M.P. Vecchi, Science 220 (1983) p.671.
- [21] M. Mézard and G. Parisi, Eur. Phys. J. B 20 (2001) p.217.
- [22] M. Mézard and G. Parisi, J. Stat. Phys. 111 (2003) p.1.
- [23] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. USA 84 (1987) p.7524.
- [24] N. Surlas, *Statistical mechanics and error-correcting codes*, in *From Statistical Physics to Statistical Inference and Back*, P. Grassberger and J.P. Nadal, eds., Kluwer, Alphen aan den Rijn, 1994, p.195.
- [25] N. Surlas, Europhys. Lett. 25 (1994) p.159.
- [26] M. Jerrum, Random Struct. Algorithms 3 (1992) p.347.
- [27] M. Jerrum and G.B. Sorkin, *Simulated annealing for graph bisection*, *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, Palo Alto, 1993, p.94.
- [28] D. Achlioptas and A. Coja-Oghlan, *Algorithmic barriers from phase transitions*, *IEEE 49th Annual IEEE Symposium on Foundations of Computer Science, FOCS'08*, Philadelphia, 2008, p.793.
- [29] R. Mulet, A. Pagnani, M. Weigt and R. Zecchina, Phys. Rev. Lett. 89 (2002) p.268701.
- [30] L. Zdeborová and F. Krzakala, Phys. Rev. E 76 (2007) p.031131.
- [31] F. Krzakala and L. Zdeborová, Phys. Rev. Lett. 102 (2009) p.238701.
- [32] J.J. Hopfield, Proc. Natl. Acad. Sci. USA 79 (1982) p.2554.
- [33] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) p.721.
- [34] A.E. Gelfand and A.F.M. Smith, J. Am. Stat. Assoc. 85 (1990) p.398.
- [35] R.M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Tech. Rep., Department of Computer Science, University of Toronto Toronto, Ontario, 1993.
- [36] C. Peterson and J.R. Anderson, Complex Syst. 1 (1987) p.995.
- [37] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola and L.K. Saul, Mach. Learn. 37 (1999) p.183.
- [38] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice*, MIT Press, 2001.
- [39] M.J. Wainwright and M.I. Jordan, Found. Trends<sup>ASCS</sup> Mach. Learn. 1 (2008) p.1.
- [40] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, J. Chem. Phys. 21 (1953) p.1087.
- [41] W.K. Hastings, Biometrika 57 (1970) p.97.

- [42] J.D. Van der Waals, Nobel Lect. Phys. 1 (1910) p.254.
- [43] P. Weiss and G. Foëx, Archives des sciences physiques et naturelles 31 (1911) p.5.
- [44] H.A. Bethe, Proc. R. Soc. Lond. Ser. A: Math. Phys. Sci. 150 (1935) p.552.
- [45] R. Peierls, Proc. R. Soc. Lond. Ser. A: Math. Phys. Eng. Sci. 154 (1936), p.207.
- [46] D.J. Thouless, P.W. Anderson and R.G. Palmer, Phil. Mag. 35 (1977) p.593.
- [47] T. Plefka, J. Phys. A: Math. Gen. 15 (1982) p.1971.
- [48] A. Georges and J.S. Yedidia, J. Phys. A: Math. Gen. 24 (1991) p.2173.
- [49] R. Kikuchi, Phys. Rev. 81 (1951) p.988.
- [50] R.G. Gallager, IEEE Trans. Inform. Theory 8 (1962) p.21.
- [51] J. Pearl, *Reverend Bayes on inference engines: a distributed hierarchical approach*, Proceedings American Association of Artificial Intelligence National Conference on AI, Pittsburgh, PA, 1982, p.133.
- [52] J. Yedidia, W. Freeman and Y. Weiss, *Understanding belief propagation and its generalizations*, in *Exploring Artificial Intelligence in the New Millennium*, Bernhard Nebel and Gerhard Lakemeyer, eds., Morgan Kaufmann Publishers, Burlington, 2003, p.239.
- [53] P. Erdős and A. Rényi, Publ. Math. Debrecen 6 (1959) p.290.
- [54] P. Erdős and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. 5 (1960) p.17.
- [55] E. Abbe, A.S. Bandeira, A. Bracher and A. Singer, IEEE Trans. Netw. Sci. Eng. 1 (2014) p.10.
- [56] E. Abbe and A. Montanari, *Conditional random fields, planted constraint satisfaction and entropy concentration*, in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, M. Goemans, K. Jansen, J.D.P. Rolim and L. Trevisan, eds., Springer, Berlin, 2013, p.332.
- [57] A. Saade, M. Lelarge, F. Krzakala and L. Zdeborová, *Spectral detection in the censored block model*, 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, 2015, p.1184.
- [58] D.C. Mattis, Phys. Lett. A 56 (1976) p.421.
- [59] H. Nishimori, J. Phys. C: Solid State Phys. 13 (1980) p.4071.
- [60] H. Nishimori, Prog. Theor. Phys. 66 (1981) p.1169.
- [61] C. Kwon and D.J. Thouless, Phys. Rev. B 37 (1988) p.7649.
- [62] J.M. Carlson, J.T. Chayes, L. Chayes, J. Sethna and D.J. Thouless, J. Stat. Phys. 61 (1990) p.987.
- [63] J.M. Carlson, J.T. Chayes, J.P. Sethna and D. Thouless, J. Stat. Phys. 61 (1990) p.1069.
- [64] F. Krzakala and L. Zdeborová, J. Chem. Phys. 134 (2011) p.034513.
- [65] T. Castellani, F. Krzakala and F. Ricci-Tersenghi, Eur. Phys. J. B: Condens. Matter Complex Syst. 47 (2005) p.99.
- [66] D.J. Thouless, Phys. Rev. Lett. 56 (1986) p.1082.
- [67] S. Heimlicher, M. Lelarge and L. Massoulié, arXiv preprint, arXiv:1209.2910 (2012).
- [68] M. Lelarge, L. Massoulié and J. Xu, *Reconstruction in the labeled stochastic block model*, 2013 IEEE Information Theory Workshop (ITW), Sevilla, September 2013, p.1.
- [69] A.P. Dempster, N.M. Laird and D.B. Rubin, J. R. Stat. Soc. Ser. B (Methodol.) 39 (1977) p.1.
- [70] A. Georges, D. Hansel, P. Le Doussal and J.-P. Bouchaud, J. Phys. 46 (1985) p.1827.
- [71] Y. Iba, J. Phys. A: Math. Gen. 32 (1999) p.3875.
- [72] M. Mézard, G. Parisi and R. Zecchina, Science 297 (2002) p.812.
- [73] P.M. Goldbart, N. Goldenfeld and D. Sherrington, *Stealing the Gold-A Celebration of the Pioneering Physics of Sam Edwards*, Oxford University Press, Oxford, 2005.
- [74] J. Wehr and M. Aizenman, J. Stat. Phys. 60 (1990) p.287.
- [75] F. Guerra and F.L. Toninelli, Commun. Math. Phys. 230 (2002) p.71.
- [76] C. Moore and S. Mertens, *The Nature of Computation*, Oxford University Press, Oxford, 2011.
- [77] J.P. Bouchaud, L. Cugliandolo, J. Kurchan and M. Mézard, *Out of equilibrium dynamics in spin glasses and other glassy systems*, in *Spin Glasses and Random Fields*, A.P. Young, ed., World Scientific, Singapore, 1998, p.161.
- [78] E. Mossel, J. Neeman and A. Sly, arXiv preprint, arXiv:1202.1499 (2012).
- [79] H. Sompolinsky and A. Zippelius, Phys. Rev. B 25 (1982) p.6860.
- [80] F. Krzakala and L. Zdeborová, J. Chem. Phys. 134 (2011) p.034512.
- [81] A. Montanari, Eur. Trans. Telecommun. 19 (2008) p.385.

- [82] L. Berthier and G. Biroli, *Rev. Mod. Phys.* 83 (2011) p.587.
- [83] A. Montanari and G. Semerjian, *J. Stat. Phys.* 125 (2006) p.23.
- [84] A. Montanari and G. Semerjian, *J. Stat. Phys.* 124 (2006) p.103.
- [85] R. Mari and J. Kurchan, *J. Chem. Phys.* 135 (2011) p.124504.
- [86] L. Foini, F. Krzakala and F. Zamponi, *J. Stat. Mech.: Theory Exp.* 2012 (2012) p.P06013.
- [87] V. Bapst, G. Semerjian and F. Zamponi, *Phys. Rev. A* 87 (2013) p.042322.
- [88] P. Charbonneau, Y. Jin, G. Parisi and F. Zamponi, *Proc. Natl. Acad. Sci. USA* 111 (2014) p.15025.
- [89] M.S. Mariani, G. Parisi and C. Rainone, *Proc. Natl. Acad. Sci. USA* 112 (2015) p.2361.
- [90] P. Charbonneau, Y. Jin, G. Parisi, C. Rainone, B. Seoane and F. Zamponi, *Phys. Rev. E* 92 (2015) p.012316.
- [91] G.M. Hocky, L. Berthier and D.R. Reichman, *J. Chem. Phys.* 141 (2014) p.224503.
- [92] S. Franz and G. Parisi, *Phys. Rev. Lett.* 79 (1997) p.2486.
- [93] D.J. Gross and M. Mézard, *Nucl. Phys. B* 240 (1984) p.431.
- [94] F. Ricci-Tersenghi, M. Weigt and R. Zecchina, *Phys. Rev. E* 63 (2001) p.026702.
- [95] M. Mézard, F. Ricci-Tersenghi and R. Zecchina, *J. Stat. Phys.* 111 (2003) p.505.
- [96] F. Krzakala and L. Zdeborová, *Europhys. Lett.* 90 (2010) p.66002.
- [97] L. Zdeborová and F. Krzakala, *Phys. Rev. B* 81 (2010) p.224205.
- [98] L.F. Cugliandolo and J. Kurchan, *Phys. Rev. Lett.* 71 (1993) p.173.
- [99] A. Crisanti, H. Horner and H.-J. Sommers, *Z. Phys. B: Condens. Matter* 92 (1993) p.257.
- [100] F. Krzakala and J. Kurchan, *Phys. Rev. E* 76 (2007) p.021122.
- [101] A. Barrat, S. Franz and G. Parisi, *J. Phys. A* 30 (1997) p.5593.
- [102] F. Krzakala and L. Zdeborová, *J. Phys.: Conf. Ser.* 473 (2013), p.012022.
- [103] Y. Sun, A. Crisanti, F. Krzakala, L. Leuzzi and L. Zdeborová, *J. Stat. Mech.: Theory Exp.* 2012 (2012) p.P07002.
- [104] S. Franz and G. Parisi, *J. Stat. Mech.: Theory Exp.* 2013 (2013) p.P02003.
- [105] T.R. Kirkpatrick and P.G. Wolynes, *Phys. Rev. A* 35 (1987) p.3072.
- [106] T.R. Kirkpatrick and P.G. Wolynes, *Phys. Rev. B* 36 (1987) p.8552.
- [107] G. Biroli and J.P. Bouchaud, *The random first-order transition theory of glasses: A Critical Assessment*, in *Structural Glasses and Supercooled Liquids: Theory, Experiment, and Applications*. Peter G. Wolynes and Vassiliy Lubchenko, eds., John Wiley & Sons, Hoboken, 2012, p.31.
- [108] D. Sherrington and S. Kirkpatrick, *Phys. Rev. Lett.* 35 (1975) p.1792.
- [109] M. Mézard, G. Parisi and M.A. Virasoro, *Europhys. Lett.* 1 (1986) p.77.
- [110] D.J. Gross, I. Kanter and H. Sompolinsky, *Phys. Rev. Lett.* 55 (1985) p.304.
- [111] T.R. Kirkpatrick and D. Thirumalai, *Phys. Rev. Lett.* 58 (1987) p.2091.
- [112] T.R. Kirkpatrick and D. Thirumalai, *Phys. Rev. B* 36 (1987) p.5388.
- [113] T.R. Kirkpatrick, D. Thirumalai and P.G. Wolynes, *Phys. Rev. A* 40 (1989) p.1045.
- [114] L. Viana and A.J. Bray, *J. Phys. C* 18 (1985) p.3037.
- [115] H.A. Bethe, *Proc. R. Soc. Lond. A* 150 (1935) p.552.
- [116] L. Onsager, *Phys. Rev.* 65 (1944) p.117.
- [117] R.G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, New York, NY, 1968.
- [118] Y. Kabashima and D. Saad, *Europhys. Lett.* 44 (1998) p.668.
- [119] S. Janson, T. Luczak and A. Rucinski, *Random Graphs*, Wiley, New York, NY, 2000.
- [120] B. Bollobás, *Random Graphs*, 2nd ed., Cambridge University Press, Cambridge, 2001.
- [121] D. Bowman and K. Levin, *Phys. Rev. B* 25 (1982) p.3438.
- [122] A. Pagnani, G. Parisi and M. Ratiéville, *Phys. Rev. E* 68 (2003) p.046706.
- [123] H. Nishimori and K.Y. M. Wong, *Phys. Rev. E* 60 (1999) p.132.
- [124] S. Franz, M. Mézard, F. Ricci-Tersenghi, M. Weigt and R. Zecchina, *Europhys. Lett.* 55 (2001) p.465.
- [125] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, *Phys. Rev. E* 84 (2011) p.066106.
- [126] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian and L. Zdeborová, *Proc. Natl. Acad. Sci. USA* 104 (2007) p.10318.
- [127] G. Jaeger, *Arch. Hist. Exact Sci.* 53 (1998) p.51.
- [128] T. Mora, *Géométrie et inférence dans l'optimisation et en théorie de l'information*, Ph.D. thesis, Université Paris-Sud, 2007. Available at <http://tel.archives-ouvertes.fr/tel-00175221/en/>

- [129] R. Monasson and R. Zecchina, Phys. Rev. Lett. 76 (1996) p.3881.
- [130] F. Krzakala, M. Mézard and L. Zdeborová, J. Satisfiab. Boolean Model. Comput. 8 (2014) p.149.
- [131] D. Achlioptas and C. Moore, *The asymptotic order of the random  $k$ -SAT threshold*, *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, Vancouver, 2002, p.779.
- [132] D. Achlioptas and Y. Peres, J. Am. Math. Soc. 17 (2004) p.947.
- [133] W. Barthel, A.K. Hartmann, M. Leone, F. Ricci-Tersenghi, M. Weigt and R. Zecchina, Phys. Rev. Lett. 88 (2002) p.188701.
- [134] H. Jia, C. Moore and D. Strain, J. Artif. Intell. Res. 28 (2007) p.107 .
- [135] J. Barbier, F. Krzakala, L. Zdeborová and P. Zhang, J. Phys.: Conf. Ser. 473 (2013), p.012021.
- [136] V. Feldman, W. Perkins and S. Vempala, *On the complexity of random satisfiability problems with planted solutions*, in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, Portland, 2015, p.77.
- [137] E. Richard and A. Montanari, *A statistical model for tensor PCA*, in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger, eds., Neural Information Processing Systems Foundation, Montreal, 2014, p.2897.
- [138] L. Zdeborová and F. Krzakala, SIAM J. Discrete Math. 25 (2011) p.750.
- [139] I. Hen and A.P. Young, Phys. Rev. E 84 (2011) p.061152.
- [140] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová and P. Zhang, Proc. Natl. Acad. Sci. 110 (2013) p.20935.
- [141] P. Zhang, Phys. Rev. E 91 (2015) p.042120.
- [142] A. Coja-Oghlan, Combin. Probab. Comput. 19 (2010) p.227.
- [143] A. Coja-Oghlan and A. Lanka, SIAM J. Discrete Math. 23 (2009) p.1682.
- [144] L. Laloux, P. Cizeau, J.-P. Bouchaud and M. Potters, Phys. Rev. Lett. 83 (1999) p.1467.
- [145] J.P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*, Cambridge university press, Cambridge, 2003.
- [146] C. Bordenave, M. Lelarge and L. Massoulié, *Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs*, *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, Berkeley, 2015, p.1347.
- [147] N. Alon, M. Krivelevich and B. Sudakov, Random Struct. Algorithms 13 (1998) p.457.
- [148] A. Juels and M. Peinado, Designs Codes Cryptogr. 20 (2000) p.269.
- [149] Y. Deshpande and A. Montanari, Found. Comput. Math. 15 (2015) p.1.
- [150] Q. Berthet and P. Rigollet, arXiv preprint, arXiv:1304.0828 (2013).
- [151] B. Hajek, Y. Wu and J. Xu, arXiv preprint, arXiv:1406.6625 (2014).
- [152] Z. Ma and Y. Wu, Ann. Stat. 43 (2015) p.1089.
- [153] T.T. Cai, T. Liang and A. Rakhlin, arXiv preprint, arXiv:1502.01988 (2015).
- [154] P. Cheeseman, B. Kanefsky and W.M. Taylor, *Where the Really Hard Problems Are*, *Proceedings of the 12th IJCAI*, Morgan Kaufmann, San Mateo, CA, 1991, p.331.
- [155] B. Selman, D.G. Mitchell and H.J. Levesque, Artif. Intell. 81 (1996) p.17.
- [156] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky, Nature 400 (1999) p.133.
- [157] L. Zdeborová, Acta Physica Slovaca 59 (2009) p.169.
- [158] M. Lentmaier, A. Sridharan, D.J. Costello and K.S. Zigangirov, IEEE Trans. Inf. Theory 56 (2010) p.5274.
- [159] S. Kudekar, T.J. Richardson and R.L. Urbanke, IEEE Trans. Inform. Theory 57 (2011) p.803.
- [160] A.J. Felstrom and K. Zigangirov, IEEE Trans. Inf. Theory 45 (1999) p.2181.
- [161] S.H. Hassani, N. Macris and R. Urbanke, J. Stat. Mech.: Theor. Exp. 2012 (2012) p.P02011.
- [162] F. Caltagirone, S. Franz, R.G. Morris and L. Zdeborová, Phys. Rev. E 89 (2014) p.012102.
- [163] Y. Kabashima, J. Phys. Soc. Japan 72 (2003) p.1645.
- [164] Y. Kabashima, J. Phys. A: Math. Gen. 36 (2003) p.11111.
- [165] D.L. Donoho, A. Maleki and A. Montanari, Proc. Natl. Acad. Sci. USA 106 (2009) p.18914.
- [166] E. Bolthausen, Commun. Math. Phys. 325 (2014) p.333.
- [167] M. Bayati and A. Montanari, IEEE Trans. Inform. Theory 57 (2011) p.764.
- [168] D.L. Donoho, A. Maleki and A. Montanari, Proc. Natl. Acad. Sci. USA 106 (2009) p.18914.

- [169] Y. Kabashima and S. Uda, *A BP-based algorithm for performing Bayesian inference in large perceptron-type networks*, in *Algorithmic Learning Theory*, S.B. David, J. Case and A. Maruoka, eds., Springer, Berlin, 2004, p.479.
- [170] J.R.L. de Almeida and D.J. Thouless, *J. Phys. A* 11 (1978) p.983.
- [171] A. Manoel, F. Krzakala, E. Tramel and L. Zdeborová, *Swept approximate message passing for sparse estimation*, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, Lille, 2015, p.1123.
- [172] M. Oppen and O. Winther, *Phys. Rev. Lett.* 86 (2001) p.3695.
- [173] G. Parisi and M. Potters, *J. Phys. A: Math. Gen.* 28 (1995) p.5267.
- [174] M. Shamir and H. Sompolinsky, *Phys. Rev. E* 61 (2000) p.1839.
- [175] M. Oppen, B. Cakmak and O. Winther, *J. Phys. A: Math. Theor.* 49 (2016) p.114002.
- [176] S. Fortunato, *Phys. Rep.* 486 (2010) p.75.
- [177] W.E. Donath and A.J. Hoffman, *IBM J. Res. Dev.* 17 (1973) p.420.
- [178] M. Fiedler, *Czechoslovak Math. J.* 23 (1973) p.298.
- [179] U. Von Luxburg, *Stat. Comput.* 17 (2007) p.395.
- [180] M. Girvan and M.E.J. Newman, *Proc. Natl. Acad. Sci. USA* 99 (2002) p.7821.
- [181] A. Clauset, M.E.J. Newman and C. Moore, *Phys. Rev. E* 70 (2004) p.066111.
- [182] M.E.J. Newman, *Proc. Natl. Acad. Sci. USA* 103 (2006) p.8577.
- [183] J. Reichardt and S. Bornholdt, *Phys. Rev. E* 74 (2006) p.016110.
- [184] L. Zdeborová and S. Boettcher, *J. Stat. Mech.: Theory Exp.* 2010 (2010) p.P02020.
- [185] H.C. White, S.A. Boorman and R.L. Breiger, *Am. J. Sociol.* 81 (1976) p.730.
- [186] P.W. Holland, K.B. Laskey and S. Leinhardt, *Social Netw.* 5 (1983) p.109.
- [187] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, *Phys. Rev. Lett.* 107 (2011) p.065701.
- [188] P. Zhang, F. Krzakala, J. Reichardt and L. Zdeborová, *J. Stat. Mech.: Theory Exp.* 2012 (2012) p.P12021.
- [189] M.B. Hastings, *Phys. Rev. E* 74 (2006) p.035102.
- [190] E. Mossel, J. Neeman and A. Sly, *Belief propagation, robust reconstruction and optimal recovery of block models*, *Proceedings of the 27th Conference on Learning Theory*, Barcelona, 2014, p.356.
- [191] A. Montanari and D. Shah, *Counting good truth assignments of random k-SAT formulae*, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, 2007, p.1255.
- [192] A. Dembo and A. Montanari, *Ann. Appl. Probab.* 20 (2010) p.565.
- [193] P. Zhang and C. Moore, *Proc. Natl. Acad. Sci.* 111 (2014) p.18144.
- [194] T. Lesieur, F. Krzakala and L. Zdeborová, *MMSE of probabilistic low-rank matrix estimation: universality with respect to the output channel*, *2015 53rd Annual Allerton Conference on Communication, Control, and Computing*, Allerton, 2015, p.680.
- [195] D.J. Gross, I. Kanter and H. Sompolinsky, *Phys. Rev. Lett.* 55 (1985) p.304.
- [196] F. Caltagirone, G. Parisi and T. Rizzo, *Phys. Rev. E* 85 (2012) p.051504.
- [197] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, New York, 2014, p.694.
- [198] E. Mossel, J. Neeman and A. Sly, *arXiv preprint*, arXiv:1311.4115 (2013).
- [199] P. Zhang, C. Moore and M. E.J. Newman, *Phys. Rev. E* 93 (2016) p.012303.
- [200] P. Zhang, C. Moore and L. Zdeborová, *Phys. Rev. E* 90 (2014) p.052802.
- [201] C. Cammarota and G. Biroli, *Proc. Natl. Acad. Sci.* 109 (2012) p.8850.
- [202] A. Saade, F. Krzakala and L. Zdeborová, *Spectral clustering of graphs with the Bethe Hessian*, in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger, eds., Neural Information Processing Systems Foundation, Montreal, 2014, p.406.
- [203] J.M. Mooij and H.J. Kappen, *Validity Estimates for Loopy Belief Propagation on Binary Real-world Networks*, in *Advances in neural information processing systems*, L.K. Saul, Y. Weiss and L. Bottou, eds., 2004, p.945.
- [204] F. Ricci-Tersenghi, *J. Stat. Mech.: Theory Exp.* (2012) p.P08015.



- [205] K.I. Hashimoto, *Automorphic Forms and Geometry of Arithmetic Varieties*, in *Zeta Functions of Finite Graphs and Representations of  $p$ -adic Groups*, K. Hashimoto and Y. Namikawa, eds., Academic Press, London, 1989, p.211.
- [206] H. Bass, *Int. J. Math.* 3 (1992) p.717.
- [207] N. Alon, I. Benjamini, E. Lubetzky and S. Sodin, *Commun. Contemp. Math.* 9 (2007) p.585.
- [208] S. Sodin, *J. Math. Phys.* 48 (2007) p.123503.
- [209] J. Friedman, *A Proof of Alon's Second Eigenvalue Conjecture and Related Problems*, American Mathematical Society, Providence, 2008.
- [210] D.R. Karger, S. Oh and D. Shah, *Iterative learning for reliable crowdsourcing systems*, in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira and K.Q. Weinberger, eds., Neural Information Processing Systems Foundation, Granada, 2011, p.1953.
- [211] H.D. Pfister and P.O. Vontobel, *On the relevance of graph covers and zeta functions for the analysis of SPA decoding of cycle codes*, *2013 IEEE International Symposium on Information Theory Proceedings (ISIT)*, Istanbul, 2013, p.3000.
- [212] A. Coja-Oghlan, E. Mossel and D. Vilenchik, *Combin. Probab. Comput.* 18 (2009) p.881.
- [213] T. Martin, X. Zhang and M.E.J. Newman, *Phys. Rev. E* 90 (2014) p.052808.
- [214] F. Morone and H.A. Makse, *Nature* 524 (2015) p.65.
- [215] A. Saade, F. Krzakala and L. Zdeborová, *Europhys. Lett.* 107 (2014) p.50005.
- [216] T. Rogers and I.P. Castillo, *Phys. Rev. E* 79 (2009) p.012101.
- [217] M.C. Angelini, F. Caltagirone, F. Krzakala and L. Zdeborová, *arXiv preprint*, arXiv:1507.04113 (2015).
- [218] A. Saade, F. Krzakala and L. Zdeborová, *Matrix completion from fewer entries: spectral detectability and rank estimation*, in *Advances in Neural Information Processing Systems*, C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett, eds., Neural Information Processing Systems Foundation, Montreal, 2015, p.1261.
- [219] J.J. Hopfield, *Proc. Natl. Acad. Sci. USA* 79 (1982) p.2554.
- [220] R.H. Keshavan, A. Montanari and S. Oh, *IEEE Trans. Inform. Theory* 56 (2010) p.2980.
- [221] B. Karrer, M.E.J. Newman and L. Zdeborová, *Phys. Rev. Lett.* 113 (2014) p.208702.
- [222] B. Bollobás, C. Borgs, J. Chayes and O. Riordan, *Ann. Probab.* 38 (2010) p.150.
- [223] K.E. Hamilton and L.P. Pryadko, *Phys. Rev. Lett.* 113 (2014) p.208701.
- [224] T. Rogers, *Europhys. Lett.* 109 (2015) p.28005.
- [225] F. Radicchi, *Phys. Rev. E* 91 (2015) p.010801.
- [226] M.J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, University of London, London, 2003.
- [227] J.M. Hofman and C.H. Wiggins, *Phys. Rev. Lett.* 100 (2008) p.258701.
- [228] A. Celisse, J.-J. Daudin and L. Pierre, *Electron. J. Stat.* 6 (2012) p.1847.
- [229] L.A. Adamic and N. Glance, *The political blogosphere and the 2004 US election: divided they blog*, *Proceedings of the 3rd International Workshop on Link Discovery*, Chicago, 2005, p.36.
- [230] W.W. Zachary, *J. Anthropol. Res.* 33 (1977) p.452.
- [231] M.E.J. Newman, *Phys. Rev. E* 74 (2006) p.036104.
- [232] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten and S.M. Dawson, *Behav. Ecol. Sociobiol.* 54 (2003) p.396.
- [233] V. Krebs, unpublished. Available at <http://www.orgnet.com>
- [234] B. Karrer and M.E.J. Newman, *Phys. Rev. E* 83 (2011) p.016107.
- [235] X. Yan, C. Shalizi, J.E. Jensen, F. Krzakala, C. Moore, L. Zdeborová, P. Zhang and Y. Zhu, *J. Stat. Mech.: Theory Exp.* 2014 (2014) p.P05007.
- [236] R. Tibshirani, *J. R. Stat. Soc. Ser. B (Methodol.)* 58 (1996) p.267.
- [237] T. Hastie, R. Tibshirani and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, 2015.
- [238] F. Rosenblatt, *Psychol. Rev.* 65 (1958) p.386.
- [239] E. Gardner and B. Derrida, *J. Phys. A: Math. Gen.* 21 (1988) p.271.
- [240] W. Krauth and M. Mézard, *J. Phys.* 50 (1989) p.3057.
- [241] T. Shinzato and Y. Kabashima, *J. Phys. A: Math. Theor.* 41 (2008) p.324013.

- [242] M. Mézard, J. Phys. A: Math. Gen. 22 (1989) p.2181.
- [243] Y. Kabashima, J. Phys.: Conf. Ser. 95 (2008), p.012001.
- [244] D.L. Donoho, IEEE Trans. Inform. Theory 52 (2006) p.1289.
- [245] E.J. Candes and T. Tao, IEEE Trans. Inform. Theory 52 (2006) p.5406.
- [246] E.J. Candè and M.B. Wakin, IEEE Signal Process. Mag. 25 (2008) p.21.
- [247] D.L. Donoho and J. Tanner, Proc. Natl. Acad. Sci. USA 102 (2005) p.9446.
- [248] Y. Kabashima, T. Wadayama and T. Tanaka, J. Stat. Mech.: Theory Exp. 2009 (2009) p.L09003.
- [249] S. Ganguli and H. Sompolinsky, Phys. Rev. Lett. 104 (2010) p.188701.
- [250] F. Krzakala, M. Mézard, F. Sausset, Y. Sun and L. Zdeborová, Phys. Rev. X 2 (2012) p.021005.
- [251] F. Krzakala, M. Mézard, F. Sausset, Y. Sun and L. Zdeborová, J. Stat. Mech.: Theory Exp. 2012 (2012) p.P08009.
- [252] D.L. Donoho, A. Javanmard and A. Montanari, IEEE Trans. Inform. Theory 59 (2013) p.7434.
- [253] A. Braunstein and R. Zecchina, Phys. Rev. Lett. 96 (2006) p.030201.
- [254] S. Rangan, *Estimation with random linear mixing, belief propagation and compressed sensing*, 2010 44th Annual Conference on Information Sciences and Systems (CISS), Princeton, 2010, p.1.
- [255] M. Bayati, M. Lelarge and A. Montanari, Ann. Appl. Probab. 25 (2015) p.753.
- [256] S. Rangan, *Generalized approximate message passing for estimation with random linear mixing*, 2011 IEEE International Symposium on Information Theory Proceedings (ISIT), St. Petersburg, 2011, p.2168.
- [257] U. Kamilov, S. Rangan, M. Unser and A.K. Fletcher, *Approximate message passing with consistent parameter estimation and applications to sparse learning*, in *Advances in Neural Information Processing Systems*, F. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, eds., Neural Information Processing Systems Foundation, Lake Tahoe, 2012, p.2438.
- [258] A.K. Fletcher, S. Rangan, L.R. Varshney and A. Bhargava, *Neural reconstruction with approximate message passing (NeuRAMP)*, in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira and K.Q. Weinberger, eds., Neural Information Processing Systems Foundation, Granada, 2011, p.2555.
- [259] S. Som and P. Schniter, IEEE Trans. Signal Process. 60 (2012) p.3439.
- [260] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata and L. Zdeborová, IEEE Trans. Inf. Theory, preprint arxiv:1402.1298 (2016).
- [261] D. Baron, S. Sarvotham and R.G. Baraniuk, IEEE Trans. Signal Process. 58 (2010) p.269.
- [262] O. Shental, P.H. Siegel, J.K. Wolf, D. Bickson and D. Dolev, *Gaussian belief propagation solver for systems of linear equations*, in *IEEE International Symposium on Information Theory, 2008 (ISIT 2008)*, Toronto, 2008, p.1863.
- [263] E.B. Sudderth, A.T. Ihler, M. Isard, W.T. Freeman and A.S. Willsky, Commun. ACM 53 (2010) p.95.
- [264] S. Rangan, P. Schniter and A. Fletcher, *On the convergence of approximate message passing with arbitrary matrices*, 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, 2014, p.236.
- [265] F. Krzakala, A. Manoel, E.W. Tramel and L. Zdeborová, *Variational free energies for compressed sensing*, 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, 2014, p.1499.
- [266] J. Vila, P. Schniter, S. Rangan, F. Krzakala and L. Zdeborová, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, 2015, p.2021.
- [267] F. Caltagirone, L. Zdeborová and F. Krzakala, *On convergence of approximate message passing*, 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, 2014, p.1812.
- [268] S. Rangan, A.K. Fletcher, P. Schniter and U.S. Kamilov, *Inference for generalized linear models via alternating directions and Bethe free energy minimization*, 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, 2015, p.1640.
- [269] Gamp in matlab. Available at [http://gampmatlab.wikia.com/wiki/Generalized\\_Approximate\\_Message\\_Passing](http://gampmatlab.wikia.com/wiki/Generalized_Approximate_Message_Passing)
- [270] P.T. Boufounos and R.G. Baraniuk, *1-bit compressive sensing*, 42nd Annual Conference on Information Sciences and Systems, 2008 (CISS 2008), Princeton, 2008, p.16.
- [271] Y. Xu, Y. Kabashima and L. Zdeborová, J. Stat. Mech.: Theory Exp. 2014 (2014) p.P11015.
- [272] G. Györfyi, Phys. Rev. A 41 (1990) p.7097.



- [273] A. Javanmard and A. Montanari, Inform. Inference (2013) p.iat004. Available at <http://imaiai.oxfordjournals.org/content/early/2013/10/16/imaiai.iat004.short>
- [274] T. Tanaka, IEEE Trans. Inform. Theory 48 (2002) p.2888.
- [275] D. Guo and S. Verdú, IEEE Trans. Inform. Theory 51 (2005) p.1983.
- [276] Y. Wu and S. Verdú, IEEE Trans. Inform. Theory 58 (2012) p.6241.
- [277] S. Rangan, V. Goyal and A.K. Fletcher, *Asymptotic analysis of map estimation via the replica method and compressed sensing*, in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams and A. Culotta, eds., Neural Information Processing Systems Foundation, Vancouver, 2009, p.1545.
- [278] K. Takeda and Y. Kabashima, *Statistical mechanical analysis of compressed sensing utilizing correlated compression matrix*, 2010 IEEE International Symposium on Information Theory Proceedings (ISIT), Austin, 2010, p.1538.
- [279] D.L. Donoho, I. Johnstone and A. Montanari, IEEE Trans. Inform. Theory 59 (2013) p.3396.
- [280] J. Barbier, F. Krzakala, M. Mézard and L. Zdeborová, *Compressed sensing of approximately-sparse signals: phase transitions and optimal reconstruction*, 2012 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton, 2012, p.800.
- [281] F. Krzakala, M. Mézard and L. Zdeborová, *Compressed sensing under matrix uncertainty: optimum thresholds and robust approximate message passing*, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, 2013, p.5519.
- [282] J. Ziniel, P. Schniter and P. Sederberg, *Binary linear classification and feature selection via generalized approximate message passing*, 2014 48th Annual Conference on Information Sciences and Systems (CISS), Princeton, 2014, p.1.
- [283] S. Kudekar and H. Pfister, *The effect of spatial coupling on compressive sensing*, Communication, Control, and Computing, Allerton, 2010, p.347.
- [284] F. Caltagirone and L. Zdeborová, arXiv preprint, arXiv:1401.6380 (2014).
- [285] K. Takeuchi, T. Tanaka and T. Kawabata, *Improvement of BP-based CDMA multiuser detection by spatial coupling*, IEEE International Symposium on Information Theory Proceedings (ISIT), St. Petersburg, 2011, p.1489.
- [286] C. Schlegel and D. Truhachev, *Multiple access demodulation in the lifted signal graph with spatial coupling*, IEEE International Symposium on Information Theory Proceedings (ISIT), St. Petersburg, 2011, p.2989.
- [287] J. Barbier, F. Krzakala, L. Zdeborová and P. Zhang, *Robust error correction for real-valued signals via message-passing decoding and spatial coupling*, 2013 IEEE Information Theory Workshop (ITW), Seville, 2013, p.1.
- [288] A. Joseph and A.R. Barron, IEEE Trans. Inform. Theory 58 (2012) p.2541.
- [289] J. Barbier and F. Krzakala, *Replica analysis and approximate message passing decoder for superposition codes*, 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, 2014, p.1494.
- [290] J. Barbier, C. Schülke and F. Krzakala, J. Stat. Mech.: Theory Exp. 2015 (2015) p.P05013.
- [291] C. Rush, A. Greig and R. Venkataramanan, *Capacity-achieving sparse regression codes via approximate message passing decoding*, 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, 2015, p.2016.
- [292] J. Barbier and F. Krzakala, arXiv preprint, arXiv:1503.08040 (2015).
- [293] P. Schniter, *Turbo reconstruction of structured sparse signals*, 2010 44th Annual Conference on Information Sciences and Systems (CISS), Princeton, 2010, p.1.
- [294] S. Rangan, A.K. Fletcher, V.K. Goyal and P. Schniter, arXiv preprint, arXiv:1111.2581 (2011).
- [295] G.E. Hinton, Neural Comput. 14 (2002) p.1771.
- [296] G.E. Hinton, *A Practical Guide to Training Restricted Boltzmann Machines*, Tech. Rep. UTML TR 2010-003, University of Toronto, Toronto, 2010.
- [297] E.W. Tramel, A. Drémeau and F. Krzakala, arXiv preprint, arXiv:1502.06470 (2015).
- [298] Y. Bengio, Found. Trends<sup>ASG</sup> Mach. Learn. 2 (2009) p.1.
- [299] B. Cakmak, O. Winther and B.H. Fleury, *S-AMP: approximate message passing for general matrix ensembles*, 2014 IEEE Information Theory Workshop (ITW), 2014, p.192.

- [300] K. Takeda, S. Uda and Y. Kabashima, *Europhys. Lett.* 76 (2006) p.1193.
- [301] Y. Kabashima and M. Vehkaperä, *Signal recovery using expectation consistent approximation for linear observations*, 2014 *IEEE International Symposium on Information Theory (ISIT)*, Honolulu, 2014, p.226.
- [302] T. Liu, C.K. Wen, S. Jin and X. You, arXiv preprint, arXiv:1512.04833 (2015).
- [303] E. Gouillart, F. Krzakala, M. Mézard and L. Zdeborová, *Inverse Probl.* 29 (2013) p.035003.
- [304] A.P. Mosk, A. Lagendijk, G. Lerosey and M. Fink, *Nat. Photonics* 6 (2012) p.283.
- [305] S.M. Popoff, G. Lerosey, R. Carminati, M. Fink, A.C. Boccarda and S. Gigan, *Phys. Rev. Lett.* 104 (2010) p.100601.
- [306] P. Schniter and S. Rangan, *IEEE Trans. Signal Process.* 63 (2015) p.1043.
- [307] A. Drémeau, A. Liutkus, D. Martina, O. Katz, C. Schülke, F. Krzakala, S. Gigan and L. Daudet, *Opt. Exp.* 23 (2015) p.11898.
- [308] S.G. Mallat and Z. Zhang, *IEEE Trans. Signal Process.* 41 (1993) p.3397.
- [309] M. Elad and M. Aharon, *IEEE Trans. Image Process.* 15 (2006) p.3736.
- [310] J. Mairal, F. Bach, J. Ponce and G. Sapiro, *Online dictionary learning for sparse coding*, *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, 2009, p.689.
- [311] G.E. Hinton, S. Osindero and Y.-W. Teh, *Neural Comput.* 18 (2006) p.1527.
- [312] G.E. Hinton and R.R. Salakhutdinov, *Science* 313 (2006) p.504.
- [313] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, *Adv. Neural Inform. Process. Syst.* 19 (2007) p.153.
- [314] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso and E. Moulines, *IEEE Trans. Signal Process.* 45 (1997) p.434.
- [315] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M.J. McKeown, V. Iragui and T.J. Sejnowski, *Psychophysiology* 37 (2000) p.163.
- [316] I.M. Johnstone and A.Y. Lu, *J. Am. Stat. Assoc.* 104 (2009) p.682.
- [317] H. Zou, T. Hastie and R. Tibshirani, *J. Comput. Graph. Stat.* 15 (2006) p.265.
- [318] F. Krzakala, M. Mézard and L. Zdeborová, *Phase diagram and approximate message passing for blind calibration and dictionary learning*, 2013 *IEEE International Symposium on Information Theory Proceedings (ISIT)*, Istanbul, 2013, p.659.
- [319] J. Vila, P. Schniter and J. Meola, *Hyperspectral image unmixing via bilinear generalized approximate message passing*, in *SPIE Defense, Security, and Sensing*, Sylvia S. Shen and Paul E. Lewis, eds., Baltimore, 2013, p.87430Y. Available at <http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1690258>
- [320] J.T. Parker, P. Schniter and V. Cevher, *IEEE Trans. Signal Process.* 62 (2014) p.5839.
- [321] A. Sakata and Y. Kabashima, *Europhys. Lett.* 103 (2013) p.28008.
- [322] A. Sakata and Y. Kabashima, *Sample complexity of Bayesian optimal dictionary learning*, 2013 *IEEE International Symposium on Information Theory Proceedings (ISIT)*, Istanbul, 2013, p.669.
- [323] T.L.H. Watkin and J.-P. Nadal, *J. Phys. A: Math. Gen.* 27 (1994) p.1899.
- [324] M. Biehl and A. Mietzner, *J. Phys. A: Math. Gen.* 27 (1994) p.1885.
- [325] N. Barkai and H. Sompolinsky, *Phys. Rev. E* 50 (1994) p.1766.
- [326] S. Rangan and A.K. Fletcher, *Iterative estimation of constrained rank-one matrices in noise*, 2012 *IEEE International Symposium on Information Theory Proceedings (ISIT)*, Cambridge, 2012, p.1246.
- [327] R. Matsushita and T. Tanaka, *Low-rank matrix reconstruction and clustering via approximate message passing*, in *Advances in Neural Information Processing Systems*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds., Neural Information Processing Systems Foundation, Lake Tahoe, 2013, p.917.
- [328] Y. Deshpande and A. Montanari, *Information-theoretically optimal sparse PCA*, 2014 *IEEE International Symposium on Information Theory (ISIT)*, Honolulu, 2014, p.2197.
- [329] T. Lesieur, F. Krzakala and L. Zdeborová, *Phase transitions in sparse PCA*, 2015 *IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, 2015, p.1635.
- [330] J. Baik, G. BenArous and S. Péché, *Ann. Probab.* 33 (2005) p.1643.

- [331] A. Montanari, D. Reichman and O. Zeitouni, *On the limitation of spectral methods: from the Gaussian hidden clique problem to rank-one perturbations of Gaussian tensors*, in *Advances in Neural Information Processing Systems*, C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett, eds., Neural Information Processing Systems Foundation, Montreal, 2015, p.217.