Journal of Statistical
Mechanics: Theory and
Experiment

PAPER

Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices

To cite this article: Florent Krzakala *et al J. Stat. Mech.* (2012) P08009

View the article online for updates and enhancements.

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices

Florent Krzakala[1,5], **Marc Mézard**[2], **Francois Sausset**[2], **Yifan Sun**[1,3] **and Lenka Zdeborová**[4]

[1] CNRS and ESPCI ParisTech, 10 rue Vauquelin, UMR 7083 Gulliver, Paris F-75005, France
[2] Université Paris-Sud & CNRS, LPTMS, UMR8626, Bâtiment 100, F-91405 Orsay, France
[3] LMIB and School of Mathematics and Systems Science, Beihang University, 100191 Beijing, People's Republic of China
[4] Institut de Physique Théorique, IPhT, CEA Saclay, and URA 2306, CNRS, F-91191 Gif-sur-Yvette, France
E-mail: fk@espci.fr

Received 29 June 2012
Accepted 23 July 2012
Published 17 August 2012

Online at stacks.iop.org/JSTAT/2012/P08009
doi:10.1088/1742-5468/2012/08/P08009

Abstract. Compressed sensing is a signal processing method that acquires data directly in a compressed form. This allows one to make fewer measurements than were considered necessary to record a signal, enabling faster or more precise measurement protocols in a wide range of applications. Using an interdisciplinary approach, we have recently proposed in Krzakala *et al* (2012 *Phys. Rev.* X **2** 021005) a strategy that allows compressed sensing to be performed at acquisition rates approaching the theoretical optimal limits. In this paper, we give a more thorough presentation of our approach, and introduce many new results. We present the probabilistic approach to reconstruction and discuss its optimality and robustness. We detail the derivation of the message passing algorithm for reconstruction and expectation maximization learning of signal-model parameters. We further develop the asymptotic analysis of the corresponding phase diagrams with and without measurement noise, for different distributions of signals, and discuss the best possible reconstruction performances

[5] Author to whom any correspondence should be addressed.

regardless of the algorithm. We also present new efficient seeding matrices, test them on synthetic data and analyze their performance asymptotically.

Keywords: cavity and replica method, message-passing algorithms, error correcting codes, statistical inference

ArXiv ePrint: 1206.3953

Contents

1. Introduction

1.1. Background on compressed sensing

When acquiring a signal, one needs to perform as many measurements as the number of unknowns. For a continuous signal, for instance, this translates into Nyquist's law: in order to sample perfectly, the sampling rate must be at least twice the maximum frequency present in the signal. This conventional approach underlies virtually all signal acquisition protocols used in physics experiments, in audio and visual electronics, in medical imaging devices and so on. The compressed sensing (CS) approach is triggering a major evolution in signal acquisition that goes against this common wisdom: according to CS, one can recover signals and images perfectly using far fewer measurements, and this results in a gain of time, cost, and precision. To make this possible, CS relies on the fact that many signals of interest contain redundancy and thus are sparse in some basis (i.e. they contain many coefficients close to or equal to zero when represented in some domain). This is the same insight used in data compression: the pictures we take with our cameras can be strongly compressed in the wavelet basis (almost) without the loss of quality, and this idea is for instance behind the JPEG 2000 algorithm. It would thus be convenient to record signals directly in a compressed format (thus the origin of the name 'compressed sensing') to save both in memory space and in number of measurements. The CS approach aims to design measurement protocols that acquire only the necessary information about the signal, in some compressed form. In a second step, one uses computational power to reconstruct the original signal exactly [2, 3]. The inverse problem posed by this second step is in general highly non-trivial.

Mathematically, the CS problem can be posed as follows: given an N-component signal $\mathbf{s}$, one makes M measurements that are grouped into an M-component vector $\mathbf{y}$, obtained from $\mathbf{s}$ by a linear transformation using a $M \times N$ measurement matrix $\mathbf{F}$, given by $y_\mu = \sum_{i=1}^{N} F_{\mu i} s_i$ with $\mu = 1, 2, \ldots, M$. The observer has freedom in the choice of the measurement protocol, and he knows the results of the measure (the M values in vector $\mathbf{y}$) and the $M \times N$ matrix $\mathbf{F}$ (in general various kinds of noise are present, as we shall discuss below). The aim is then to reconstruct the signal $\mathbf{s}$ from the knowledge of $\mathbf{F}$ and $\mathbf{y}$. This amounts to inverting the linear system $\mathbf{y} = \mathbf{F}\mathbf{s}$. However, we want to have M as small as possible and when $M < N$ there are fewer equations than unknowns. The system is under-determined and the inverse problem is ill-defined. CS, however, deals with sparse signals $\mathbf{s}$, in the sense that only $K < N$ of the components are non-zero. In the noiseless case, an exact reconstruction is possible for such signals as soon as $M > K$, and this condition is also a necessary one for instance in the case where the non-zero component of the signal are independent identically distributed (iid) real variables, drawn from a distribution with a continuous part. This ability to recover signals using only a limited number of measurements is crucial in many fields ranging from experimental physics and image processing to astronomy or systems biology, making CS a very attractive concept.

The most widely used technique in CS is based on a development that took place six years ago thanks to the works of Candès *et al* [2]–[5]: they proposed to find the vector satisfying the constraints $\mathbf{y} = \mathbf{F}\mathbf{x}$ which has the smallest ℓ_1 norm, defined as $\|x\|_{\ell_1} = \sum_{i=1}^{N}|x_i|$. This optimization problem is convex and can be solved using efficient linear programming techniques. They have also suggested the use of a random measurement matrix $\mathbf{F}$ with iid entries. 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 function or a spike in the time domain is spread out in the frequency domain after a Fourier transform. These ideas have led to fast and efficient algorithms, and the ℓ_1-reconstruction is now widely used, and has been at the origin of the burst of interest in CS over the past few years. It is possible to compute exactly the performance of the ℓ_1 reconstruction in the limit $N \to \infty$, and the analytic study shows the appearance of a sharp phase transition [6]. For any signal with density $\rho = K/N$, the ℓ_1 reconstruction gives indeed the exact result $x = s$ with probability one only if $\alpha = M/N > \alpha_{\ell_1}(\rho)$, where $\alpha_{\ell_1}(\rho)$ is, however, larger than ρ. The ℓ_1 reconstruction is thus suboptimal: it requires more measurements than theoretically necessary.

J. Stat. Mech. (2012) P08009

1.2. Our main results

In this paper we analyze a probabilistic reconstruction of the signal in compressed sensing, which we have introduced in [1]. We provide here a more detailed presentation, and we include several new results. We use a simplification of the belief propagation (BP) algorithm, also known as approximate message passing (AMP) [7] or generalized approximate message passing (G-AMP) [8] in the context of CS. The probabilistic approach is combined with an expectation maximization type of learning of parameters as in [1] (which has been independently proposed in the context of G-AMP in [9]). We use the replica and cavity methods to analyze on one hand the asymptotic performance of the BP algorithm and on the other hand the information theoretic limits for signal reconstruction, and the associated phase transitions. For sensing matrices with iid entries there is a region of parameters (signal sparsity, undersampling rate and measurement noise) in which there is a gap between the BP reconstruction and the optimal reconstruction. In this hard region, BP iterations are blocked in a suboptimal fixed point. We also study in detail the phase diagram in the presence of measurement noise and observe that the region where BP is suboptimal persists, but becomes smaller and eventually disappears as the noise variance grows. Analyzing the origin of this algorithmic barrier and thinking about an analogy with crystal nucleation in [1] we designed and tested BP reconstruction with seeded measurement matrices for which this gap shrinks or entirely disappears. The implementation of our matrices and of the algorithm is available at http://aspics.krzakala.org/.

We now describe the organization of this paper and list its main contributions:

- *Optimality of the probabilistic reconstruction.* We review in section 2.2 the well-known fact that probabilistic inference is optimal when the signal model matches the actual signal distribution. For good performance one usually requires a signal model that is 'close enough' to the actual signal to be inferred. The unavailability of such a good signal model is often at the basis of the criticisms of this probabilistic—Bayesian—inference. The situation is much more favorable in the case of noiseless compressed sensing. We noticed, and proved, in [1] that, in the case of noiseless CS, probabilistic inference is optimal even if the signal model mismatches seriously the actual signal; details are in section 2.1. This property makes our approach very robust. In our numerical experiments we successfully use the Gauss–Bernoulli model even for signals that are far from having iid Gauss–Bernoulli components. Despite this result, in practice it turns out to be useful to incorporate expectation maximization learning of parameters of the signal model, as described in section 2.3.

- *The message passing reconstruction algorithm.* We derive in detail the reconstruction algorithm and discuss how it is related to the existing ones. In section 3.1 we give it in a form where the messages are being sent between signal components and measurement components and back—this being equivalent to the relaxed-BP algorithm [10, 11]. In section 3.2 we then derive a simplified form where messages 'live' only on the signal components and on the measurement component. This form is related to the seminal Thouless–Anderson–Palmer (TAP) [12] equations in spin glass theory, and is equivalent to the AMP algorithm in the context of CS [8, 13]. For measurement matrices with iid entries further simplifications of the algorithm exist, and are useful for a more efficient implementation, as is shown in section 3.3. We also derive the BP equations for expectation maximization learning of parameters in section 3.4.
- *Asymptotic analysis of the algorithm and of the probabilistic approach.* We use the cavity and replica methods to perform two types of asymptotic analysis. On one hand using the density evolution we describe the behavior of the belief propagation algorithm in the limit of large systems (section 4.1), on the other hand using the replica method we compute the theoretical limits for reconstruction (section 4.2), which are non-trivial in particular in the presence of noise and by definition do not depend on the algorithm. We derive the asymptotic evolution for measurement matrices having iid (or iid per block) entries of zero mean and variance $1/N$. The equations are independent of the other details of the distribution of matrix elements, and these predictions thus hold for many types of matrices (for instance, Gaussian or discrete binary ones). This makes our results very robust. In section 4.3 we then discuss the simplifications that appear in the Bayes-optimal case of matching signal model and signal distribution. In section 4.4 we derive the asymptotic evolution of the parameters in expectation maximization learning. Finally, in section 4.5 we summarize all these previous equations in the case of block measurement matrices.
- *Phase transitions, phase diagrams.* Using both the BP reconstruction algorithm and the asymptotic analysis we study the phase diagram and associated phase transitions for reconstruction of the signal. We study several settings: the optimal Bayesian inference when the signal model matches the signal distribution in section 5.1, the case when the signal model does not match the signal distribution and the phase diagram after expectation maximization learning in section 5.2, the phase diagram in the presence of measurement noise in section 5.3, and the reconstruction with seeding block matrices in section 6. Note that in doing the optimal Bayesian inference case, we thus study the best possible reconstruction performance, regardless of the algorithm.

- *Optimality achieving measurement matrices.* In [1] we introduced a new type of 'seeding' measurement matrix with which theoretically optimal reconstruction can be obtained using the BP algorithm. Such a 'threshold saturation' was later on proved for this type of matrix in [14] (called 'spatial coupling'). In section 6.1 we discuss again our motivation for the design of seeding matrices and show that there is relatively great freedom in implementing the concept of seeding. We give new examples of efficient seeding matrices, which are actually simpler and more efficient than the one we have introduced earlier. In section 6.2 we also show that these matrices are effective even when the model signal in the prior is different from the actual ones. In section 6.3 we illustrate that one can approach the optimal reconstruction limit, even in the case of noisy measurements.
- *Noise-sensitivity.* We discuss in detail the phase diagram and the performance of the algorithm in the presence of measurement noise in section 5.3. We show that there are two regions in the phase diagram. Either the BP approach is optimal, i.e. it provides the same mean-squared error as would be obtained by an intractable exhaustive search algorithm. Or BP is suboptimal due to an existence of a metastable state—in this case optimality can be restored using the seeding matrices, as we show in section 6.3. Overall this shows that the present approach has the best achievable noise stability.
- *Rigorous versus exact.* It is important to notice that the density evolution that we use for asymptotic analysis of BP was proved to be exact for the homogeneous matrices [15]. According to a private communication with the authors a proof for the block matrices also exists [16]. Therefore, our predictions on the behavior of the algorithm are exact. As far as our predictions for the optimal inference are concerned, although our presentation here is not rigorous, the predictions are exact in the context of the series of works [17]–[19].

Let us define here the block measurement matrices that we use in this paper to implement the seeding concept. Note however, that the seeding measurement matrices do not have to be block matrices. Other implementations are possible. We leave for future work an investigation into the optimal design for practical situations.

The block measurement matrices $F_{\mu i}$ are constructed as follows: the N variables are divided into L_c groups of N_p, $p = 1, \dots, L_c$, variables in each group. We denote $n_p = N_p/N$. And the M measurements are divided into L_r groups of M_q, $q = 1, \dots, L_r$, measurements in each group, define $\alpha_{qp} = M_q/N_p$. Then the matrix F is composed of $L_r \times L_c$ blocks and the matrix elements $F_{\mu i}$ are generated independently, in such a way that if μ is in group q and i in group p then $F_{\mu i}$ is a random number with zero mean and variance $J_{q,p}/N$. Thus we obtain a $L_r \times L_c$ coupling matrix $J_{q,p}$. For the asymptotic analysis we assume that $N_p \to \infty$, for all $p = 1, \dots, L_c$ and $M_q \to \infty$ for all $q = 1, \dots, L_r$. We define $I(\mu)$ $(I(i))$ to be the index of the block to which μ (i) belongs, B_q is the set of indices in block q. The case of a homogeneous matrix can easily be recovered by setting $L_c = L_r = 1$. Note that not all block matrices are good seeding matrices, the parameters have to be set in such a way that seeding is implemented (i.e. existence of the seed and interaction such that the seed grows). The choice of parameters is discussed in section 6.

Let us note that for both the homogeneous and the block matrices the results do not depend on the details of the distribution of its entries, as far as its mean and variance are fixed. In our simulations we mostly use Gaussian-distributed random entries, or $\pm 1/N$. The latter have the advantage of taking less memory space, since we can store them with bits and deal with the $\sqrt{N}$ separately (memory space to store the matrix is the main limitation of our simulations).

Note also that throughout the paper we use matrix entries of zero mean. Physical constraints might require the mean to be non-zero, but our algorithm would have to be modified for such cases. The problem, however, can be transformed rather easily to one of zero mean. Consider indeed the system $\mathbf{y} = \mathbf{F}\mathbf{s}$. Summing all M values of the vector $\mathbf{y}$ (and denoting $\bar{y} = (1/M)\sum_{\mu} y_{\mu}$ and $\overline{F_i} = (1/M)\sum_{\mu} F_{\mu i}$) one finds $M\bar{y} = \sum_{\mu}\sum_{i} F_{\mu i} x_i = \sum_i (\sum_{\mu} F_{\mu i}) s_i = M \sum_i \overline{F_i} s_i$. Denote $\bar{\mathbf{y}}$ the vector with all M components equal to $\bar{y}$ and $\bar{\mathbf{F}}$ the $M \times N$ matrix where the ith column is given by the values $\overline{F_i}$. Then the system $\mathbf{y} - \bar{\mathbf{y}} = (\mathbf{F} - \bar{\mathbf{F}})\mathbf{s}$ has a matrix with zero mean entries.

1.3. Related works

Here we discuss some interesting connections to other works on compressed sensing. It is important to realize that our main result, namely the joint design of an algorithm and a class of measurement matrices that lead to optimal CS reconstruction, and their analysis, is based on three main ingredients that were previously explored in the literature. These ingredients are the probabilistic approach, the use of a message passing algorithm for sampling from the probability distribution, and the design of seeding matrices. It is only the joint use of these three ingredients that achieves optimal reconstruction, and the understanding of the reasons owes a lot to accumulated knowledge from statistical physics of disordered systems (for instance, using seeding matrices with ℓ_1 reconstruction is useless, because ℓ_1 reconstruction is not limited by a glass transition, its limitation is intrinsic to the use of the ℓ_1 norm).

- The state-of-the-art method for signal reconstruction in CS is based on the minimization of the ℓ_1 norm of the signal under the linear constraint, for an overview of this technique see [2, 6]. A number of works also adopted a probabilistic or Bayesian approach [20]–[22]. Generically, one disadvantage of the probabilistic approach is that no exact algorithm is known for evaluation of the corresponding expectations. Whereas ℓ_1 minimization is done exactly using linear programming. In our approach, this problem is resolved with the use of belief propagation, which turns out to be an extremely efficient heuristic. Another issue of the Bayesian approach is the choice of the signal model. Whereas the performance of the ℓ_1 reconstruction is independent of the signal distribution, this is not the case for the Bayesian approach in general. We show that actually for the noiseless CS the optimal exact reconstruction is possible even if the signal model does not match the signal distribution.
- In the noiseless case of CS it is very intuitive that exact reconstruction of the signal is in principle possible if and only if the number of measurements is larger than the number of non-zero components of the signal, $\alpha > \rho_0$. In a more generic case, for instance in the presence of the measurement noise it is not straightforward to compute the best achievable mean-squared error in reconstruction. These theoretical optimality limits were analyzed rigorously in very general cases by [18, 19]. These results agree with the non-rigorous replica method as developed for CS e.g. in [23]–[25]. Here we analyze the theoretically optimal reconstruction using as well the replica method (and explicit its connection with the density evolution).

- Belief propagation (BP) is an inference algorithm that is exact on tree graphical models and that is a powerful heuristic also on loopy graphical models. It was discovered independently by several communities, in coding [26], in inference [27], or in statistical physics [12]. See [28, 29] for good overviews. Belief propagation was used for CS with sparse measurement matrices by several authors, see e.g. [22, 30, 31]. In the usual setting, however, CS corresponds to dense measurement matrices, hence a fully connected graphical model with continuous variables, the canonical form of BP iterations is intractable for such a case. However, neglecting only factors that go to zero in the large system size limit, the iterative equations can be written only for the means and variances of the corresponding probability distributions. Such a belief propagation algorithm was used in CS under the name relaxed BP (rBP) [10, 11]. Again, by neglecting only $o(1)$ terms rBP can be further simplified, as shown for ℓ_1 reconstruction in [7, 13] this version of message passing was called approximate message passing (AMP), it is equivalent to the Thouless–Anderson–Palmer (TAP) [12] equations in spin glass theory. Similar simplification of the BP algorithm was also used for the related problem of CDMA multiuser detection in [32]. The AMP was further generalized (G-AMP) to the case of a general signal model in [8, 13]. The algorithm that we use here is equivalent to G-AMP. We, however, provide an independent derivation.
- The performance of the belief propagation algorithm can be analyzed analytically in the large system limit. This can be done either using the replica method, as in [33], or using density evolution. The equivalence between the two was pointed out for a related problem—the CDMA multiuser detection—in [32]. An asymptotic density-evolution-like analysis of the AMP algorithm, called state evolution, was developed in [7], and more generally in [15]. State evolution is the analog of density evolution for dense graphs. General analysis of algorithmic phase transitions for G-AMP was presented in [34]. In this paper we perform the same density evolution analysis for other variants of the problem (with learning, where the signal model does not match the signal distribution, with noise, etc), without the rigorous proofs. Our main point is to analyze and understand the phase transitions that pose algorithmic barriers to the message passing reconstruction.
- In cases when the signal distribution is not known, we can use expectation maximization (EM) to learn the parameters of the signal model [35]. EM learning with the expectation step being done with BP was proposed in [36], recently it was applied also e.g. in [37]. In the context of CS, the EM was applied together with message passing reconstruction in [1]. An independent implementation along the same concept also appeared in [9] under the name EM-GAMP algorithm (where EM stands for expectation maximization). All the predictions made in the present paper thus also apply to the EM-GAMP algorithm.
- Based on our understanding of the properties of the algorithmic barrier encountered by the message passing reconstruction algorithm, we have designed special seeded measurement matrices for which reconstruction is possible even for close to optimal measurement rates. These matrices are based on the idea of spatial coupling that was developed first in error correcting codes [38, 39], see [40] for more transparent understanding and results. Several other applications of the same idea exist, in different contexts. For an overview see [41].

- The use of spatial coupling was first suggested for compressed sensing in [42], where the authors observed an improvement over the reconstruction with homogeneous measurement matrices (see figure 5 in [42]). They, however, did not combine all the key ingredients to achieve reconstruction up to close to the theoretical limit $\alpha = \rho_0$, as we did in [1]. Their implementation of belief propagation was also not using the simplification under which only the mean and variance of the messages are needed, hence the algorithm was not competitive speed-wise.

  We introduced seeded measurement matrices for CS in [1], and showed there, both numerically and using the density evolution, that with such matrices it is possible to achieve the information theoretic optimal measurement rates. The design was motivated by the idea of crystal nucleation and growth in statistical physics. Subsequent work [14] justified this threshold saturation rigorously in the special case when the signal model corresponds to the signal distribution, but also more generally using the concept of Rényi information dimension instead of sparsity, as in [17, 19]. Numerical experiments with seeded non-random (Gabor-type) matrices were also performed in [43].

2. Probabilistic reconstruction in compressed sensing

The definition of the compressed sensing problem as studied in this paper is as follows

$$y_\mu = \sum_{i=1}^{N} F_{\mu i} s_i + \xi_\mu \qquad \mu = 1, \dots, M, \tag{1}$$

where s_i are the signal elements, out of which only $K = \rho_0 N$ are non-zero, $0 < \rho_0 < 1$. We denote by ϕ_0 the asymptotic empirical distribution of the non-zero elements. $F_{\mu i}$ are the elements of a known measurement matrix, y_μ are the known results of measurements, and ξ_μ is Gaussian white noise on the measurement with variance Δ_μ. We denote by $\alpha = M/N$ the number of measurements per variable. The goal of CS is to find an approach (i.e. measurement matrix and a reconstruction algorithm) that allows reconstruction with as low values of α as possible.

In the asymptotic theoretical analysis we will be interested in the case of large signals $N \to \infty$, we will keep signal density ρ_0 and measurement rate α of order one. We also want to keep the components of the signal and of the measurements of order one, hence we consider the elements of the measurement matrix to have mean and variance of order $O(1/N)$.

We shall adopt a probabilistic inference approach to reconstruct the signal. The aim is to sample a vector $\mathbf{x}$ from the following probability measure

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

where Z, the partition function, is a normalization constant. Here we model the signal as stochastic with iid entries, the fraction of non-zero entries being $\rho > 0$ and their distribution being ϕ, we restrict ourselves to functions $\phi(x) < \infty$ with finite variance.

We stress that in general the signal properties are not known and hence (unless stated otherwise) we do not assume that the signal model matches the empirical signal distribution, $\rho = \rho_0$, $\Delta = \Delta_0$, nor $\phi = \phi_0$. Most previous approaches to reconstruction in CS can be stated in the form (2), e.g. the ℓ_1 minimization is equivalent to $\rho = 1$ and Laplace function ϕ. One crucial point in our approach is using $\rho < 1$, which includes the fact that one searches a sparse signal in the model of the signal.

Equation (2) can be seen as the Boltzmann measure on the disordered system with Hamiltonian

$$H(\mathbf{x}) = -\sum_{i=1}^{N} \log\left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] + \sum_{\mu=1}^{M} \frac{(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i)^2}{2\Delta_\mu}, \tag{3}$$

where the 'disorder' comes from the randomness of the measurement matrix $F_{\mu i}$ and the results y_μ. Stated this way, the problem is similar to a spin glass with N particles interacting with a long-range disordered potential. The signal $\mathbf{x} = \mathbf{s}$ is a very special configuration of these particles, that we can call 'planted', which was used to generate the problem (i.e. the value of the vector $\mathbf{y}$). In this sense all inference problems are equivalent to planted spin glass models.

2.1. Optimality in the noiseless case

In the noiseless case, $\Delta_\mu \to 0$, sampling from the measure $\hat{P}(\mathbf{x})$ leads to exact reconstruction as long as $\alpha > \rho_0$ and the support of the function ϕ contains all the non-zero elements of the signal (i.e. an arbitrary finite function of finite variance supported on real numbers for general real entry signals). In particular the density and the distribution of the true signal does not need to be known, i.e. $\rho \neq \rho_0$ and $\phi \neq \phi_0$. This is a strong optimality property that was proved in the large size limit $N \to \infty$ in [1] and that can be seen as follows.

Define an auxiliary partition function $Y(D)$ as the normalization of the measure $\hat{P}(\mathbf{x})$ restricted to configurations at a mean-squared distance D from the signal $\mathbf{s}$, i.e.

$$Y_\Delta(D) \equiv \int_{B_D(\mathbf{s})} \prod_{i=1}^{N} \mathrm{d}x_i \prod_{i=1}^{N} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right] \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta}} \mathrm{e}^{-(1/2\Delta)[\sum_{i=1}^{N} F_{\mu i}(x_i - s_i)]^2}, \tag{4}$$

where $B_D(\mathbf{s})$ is the sphere centered on $\mathbf{s}$, defined by $((1/N)\sum_{i=1}^{N}(x_i - s_i)^2 = D)$. When $D \to 0$ and $\Delta \to 0$, the N dimensional integral in (4) involves a product of $(1 - \rho_0 + \alpha)N$ Dirac delta functions. Hence as long as $\alpha > \rho_0$ the function $Y_\Delta(D)$ diverges as $D \to 0$, $\Delta \to 0$. This holds for every matrix F and every function ϕ as long as it is supported on all the elements of $\mathbf{s}$.

In a second part of the optimality argument one needs to show that $\lim_{\Delta\to 0} Y_\Delta(D)/Y_\Delta(0) = 0$ whenever $D > 0$. First, note that only configurations that solve all the M linear equations give a non-zero contribution to (4). Second, it is known that the signal $\mathbf{s}$ is the solution of the linear system with the largest number of zero elements [5], hence all the other solutions of the linear system have a negligible contribution to the integral (necessarily, a smaller number of Dirac delta functions remains after the integration).

Given this result, it then follows that for any ρ_0-dense original signal $\mathbf{s}$, and any $\alpha > \rho_0$, the probability $\hat{P}(\mathbf{s})$ of the original signal goes to one when $\Delta \to 0$. This result holds as long as the configuration minimizing the ℓ_0 norm equals the original signal $\mathbf{s}$. Remarkably this optimality holds independently of the distribution ϕ_0 of the original signal, which does not even need to be iid. Hence in the noiseless case, sampling $\mathbf{x}$ proportionally to the measure $\hat{P}(\mathbf{x})$ gives the exact reconstruction in the whole region $\alpha > \rho_0$.

2.2. The Bayesian optimality and the Nishimori conditions

The probabilistic approach can also be recovered from a Bayesian point of view. Indeed, given $\mathbf{F}$ and $\mathbf{y}$, from Bayes theorem, we have

$$P(\mathbf{x}|\mathbf{F},\mathbf{y}) = \frac{P(\mathbf{x}|\mathbf{F})P(\mathbf{y}|\mathbf{F},\mathbf{x})}{P(\mathbf{y}|\mathbf{F})}. \tag{5}$$

The value of measurements $\mathbf{y}$ given the knowledge of the matrix $\mathbf{F}$ and the signal $\mathbf{x}$ is, by definition of the problem, given by $P(\mathbf{y}|\mathbf{F},\mathbf{x}) = \prod_{\mu=1}^{M}\delta(y_\mu - \sum_{i=1}^{N}F_{\mu i}x_i)$ in the noiseless case, and by

$$P(\mathbf{y}|\mathbf{F},\mathbf{x}) = \prod_{\mu=1}^{M}\frac{1}{\sqrt{2\pi\Delta_\mu}}\mathrm{e}^{-(1/2\Delta_\mu)(y_\mu-\sum_{i=1}^{N}F_{\mu i}x_i)^2}, \tag{6}$$

with random Gaussian measurement noise of variance Δ_μ, for measurement μ. To express the probability $P(\mathbf{x}|\mathbf{F})$ we consider that the signal does not depend on the measurement matrix (which is true in all practical situations we are aware of). Further, in this paper, we do not aim to exploit possible correlations in signal entries (which could only improve the result of inference) and hence we model the signal as an iid:

$$P(\mathbf{x}|\mathbf{F}) = \prod_{i=1}^{N}\left[(1-\rho)\delta(x_i) + \rho\phi(x_i)\right]. \tag{7}$$

Thus the posterior probability of $\mathbf{x}$ after the measurement of $\mathbf{y}$ is given by

$$P(\mathbf{x}|\mathbf{F},\mathbf{y}) = \frac{1}{Z(\mathbf{y},\mathbf{F})}\prod_{i=1}^{N}[(1-\rho)\delta(x_i) + \rho\phi(x_i)]\prod_{\mu=1}^{M}\frac{1}{\sqrt{2\pi\Delta_\mu}}\mathrm{e}^{-(1/2\Delta_\mu)(y_\mu-\sum_{i=1}^{N}F_{\mu i}x_i)^2}, \tag{8}$$

where $Z(\mathbf{y},\mathbf{F}) = P(\mathbf{y}|\mathbf{F})$ is again the normalization constant. This is nothing else than the $\hat{P}(\mathbf{x})$ in equation (2).

We remind the reader that in the noiseless case, $\Delta_0 = \Delta_\mu = 0$, we have the optimality result for an arbitrary signal, as described in the previous section. However, for the case with noise, if the true density of the signal, ρ_0, the measurement noise, Δ_0, and the asymptotic empirical distribution of the signal, ϕ_0, are not known then sampling from (8) is in general not optimal.

However, if we assume knowledge of the true density of the signal, $\rho = \rho_0$, the measurement noise, $\Delta = \Delta_0$, and the asymptotic empirical distribution of the signal, $\phi = \phi_0$, then we just described the Bayes-optimal way to infer the signal $\mathbf{s}$ from the knowledge of the matrix $\mathbf{F}$ and the measurements $\mathbf{y}$. In particular, an estimator $\mathbf{x}^\star$ that minimizes the mean-squared error with respect to the original signal $\mathbf{s}$, defined as

$$\mathrm{MSE}(\mathbf{x}|\mathbf{F},\mathbf{y}) = \int \mathrm{d}\mathbf{s}\left[\sum_{i=1}^{N}(x_i - s_i)^2/N\right]P(\mathbf{s}|\mathbf{F},\mathbf{y}), \tag{9}$$

is then obtained from averages of s_i with respect to the probability measure $P(\mathbf{s}|\mathbf{F},\mathbf{y})$, i.e.,

$$x_i^\star = \int \mathrm{d}x_i \, x_i \, \nu_i(x_i), \tag{10}$$

where $\nu_i(x_i)$ is the marginal probability distribution of the variable i

$$\nu_i(x_i) \equiv \int_{\{x_j\}_{j\neq i}} P(\mathbf{x}|\mathbf{F},\mathbf{y}). \tag{11}$$

In the remainder of this article we will be using this estimator. To give another example, the optimal estimator that minimizes the mean 'absolute value' error $\mathrm{MAVE}(\mathbf{x}|\mathbf{F},\mathbf{y}) = \int \mathrm{d}\mathbf{s}\,[\sum_{i=1}^{N}|x_i - s_i|/N]P(\mathbf{s}|\mathbf{F},\mathbf{y})$ is given by the median of the marginal probability $\nu_i(x_i)$.

There are important identities that hold for the Bayes-optimal inference and that simplify many of the calculations that follow. In the physics of disordered systems these identities are known as the Nishimori conditions [44]–[46]. Basically, the Nishimori conditions follow from the fact that the planted configuration (i.e. the original signal) is an equilibrium configuration with respect to the Boltzmann measure (8). Hence many properties of the planted configuration can be computed without its knowledge by averaging over the distribution (8).

To derive the Nishimori conditions, consider the measurement matrix $\mathbf{F}$ fixed and for simplification let us drop the dependence on $\mathbf{F}$ from the notation. Consider a function $A(\mathbf{x})$ depending on a 'trial' configuration $\mathbf{x}$. We define the 'thermodynamic average' of A as

$$\langle A(\mathbf{x})\rangle \equiv \int \mathrm{d}\mathbf{x}\, A(\mathbf{x})P(\mathbf{x}|\mathbf{y}), \tag{12}$$

where $P(\mathbf{x}|\mathbf{y})$ is given by equation (8). Similarly, for a function $A(\mathbf{x}_1,\mathbf{x}_2)$ that depends on two trial configurations $\mathbf{x}_1$ and $\mathbf{x}_2$, we define

$$\langle\langle A(\mathbf{x}_1,\mathbf{x}_2)\rangle\rangle \equiv \int \mathrm{d}\mathbf{x}_1 \int \mathrm{d}\mathbf{x}_2\, A(\mathbf{x}_1,\mathbf{x}_2)P(\mathbf{x}_1|\mathbf{y})P(\mathbf{x}_2|\mathbf{y}). \tag{13}$$

For a function B that depends on the measurement $\mathbf{y}$ and on the signal $\mathbf{s}$ we define the 'disorder average' as

$$[B(\mathbf{s},\mathbf{y})] \equiv \int \mathrm{d}\mathbf{y} \int \mathrm{d}\mathbf{s}\, P(\mathbf{s})\, P(\mathbf{y}|\mathbf{s})B(\mathbf{s},\mathbf{y}), \tag{14}$$

where the signal distribution $P(\mathbf{s})$ is given by equation (7), and $P(\mathbf{y}|\mathbf{s})$ is the probability of a measurement $\mathbf{y}$ given the signal $\mathbf{s}$, as in equation (6). Note that if B does not explicitly depend on $\mathbf{s}$ then we have $[B(\mathbf{y})] \equiv \int \mathrm{d}\mathbf{y}\, Z(\mathbf{y})B(\mathbf{y})$, because $Z(\mathbf{y}) = \int \mathrm{d}\mathbf{s}\, P(\mathbf{s})\, P(\mathbf{y}|\mathbf{s})$. Using these definitions we obtain

$$\begin{aligned}[\langle A(\mathbf{x},\mathbf{s})\rangle] &= \int \mathrm{d}\mathbf{y} \int \mathrm{d}\mathbf{s}\, P(\mathbf{s})P(\mathbf{y}|\mathbf{s}) \int \mathrm{d}\mathbf{x}\, A(\mathbf{x},\mathbf{s})\, P(\mathbf{x}|\mathbf{y}) \\ &= \int \mathrm{d}\mathbf{y}\, Z(\mathbf{y}) \int \mathrm{d}\mathbf{s} \int \mathrm{d}\mathbf{x}\, A(\mathbf{x},\mathbf{s})\frac{P(\mathbf{s})P(\mathbf{y}|\mathbf{s})}{Z(\mathbf{y})}P(\mathbf{x}|\mathbf{y}) \\ &= \int \mathrm{d}\mathbf{y}\, Z(\mathbf{y}) \int \mathrm{d}\mathbf{x}_1 \int \mathrm{d}\mathbf{x}_2\, A(\mathbf{x}_1,\mathbf{x}_2)P(\mathbf{x}_2|\mathbf{y})P(\mathbf{x}_1|\mathbf{y}) = [\langle\langle A(\mathbf{x}_1,\mathbf{x}_2)\rangle\rangle], \end{aligned} \tag{15}$$

where in the third equality we renamed variables as $s = x_2$ and $x = x_1$. Equation (15) is the general form of the Nishimori condition.

We remind the reader that for many thermodynamic quantities the self-averaging property holds, i.e. for large system sizes the quantity $\langle A(\mathbf{x}, \mathbf{s})\rangle$ converges to the average over disorder of the same quantity. Equation (15) provides a rather general form of the Nishimori condition that holds for inference problems where the model for signal generation is known.

To give specific examples, let us define $m = \sum_{i=1}^{N} s_i x_i / N \equiv \mathbf{s} \cdot \mathbf{x}$ and $q = \mathbf{x}_1 \cdot \mathbf{x}_2$. Then we have in the thermodynamic limit $[\langle m\rangle] = [\langle q\rangle]$. Due to self-averaging we also have $m = q$ if $\mathbf{x}, \mathbf{x}_1$, and $\mathbf{x}_2$ were samples from the distribution $P(\mathbf{x}|\mathbf{y})$. Defining $Q = \mathbf{x} \cdot \mathbf{x}$, and using the Nishimori condition, we get $Q = \rho \,\mathrm{var}\phi$.

Note also that due to the self-averaging property we do not distinguish in what follows between the mean-squared error (9) and the squared error $E = \sum_{i=1}^{N} (\langle x_i\rangle - s_i)^2/N$.

2.3. Expectation maximization learning

In general, one does not know the true density of the signal, ρ_0, the measurement noise, Δ_0, nor the asymptotic empirical distribution of the signal, ϕ_0 (or its parameters). These parameters can be learned within the Bayesian approach, in a way similar to the expectation maximization algorithm [35, 45, 47]. Let us denote θ as the ensemble of these unknown parameters. Given the matrix $\mathbf{F}$ and measurement vector $\mathbf{y}$, the probability that the parameters take a given set of values θ is

$$P(\theta|\mathbf{F}, \mathbf{y}) = \frac{P(\theta|\mathbf{F})}{P(\mathbf{y}|F)} \int \mathrm{d}\mathbf{x}\, P(\mathbf{y}, \mathbf{x}|F, \theta) \propto P(\theta|\mathbf{F}) Z(\theta), \tag{16}$$

where $Z(\theta)$ is the normalization from (8) with a given set of parameters θ

$$Z(\rho, \bar{x}, \sigma, \Delta) = \int \prod_{i=1}^{N} \mathrm{d}x_i \prod_{i=1}^{N} \left[(1-\rho)\delta(x_i) + \frac{\rho}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x_i-\bar{x})^2/2\sigma^2}\right] \\ \times \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta}} \mathrm{e}^{-(1/2\Delta)(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i)^2}. \tag{17}$$

Considering that without knowing the measurements $\mathbf{y}$ we have no prior knowledge of θ, looking for the most probable value of parameters is equivalent to maximizing the partition function with respect to the parameters. Even if we do have a prior knowledge of θ, in the situations considered in this article the partition function scales exponentially in N and hence for large N and function $P(\theta|\mathbf{F})$ independent of N, and maximizing $Z(\theta)$ is still the right thing to do.

In what follows, in order to learn parameters θ we will hence derive stationary equations for the partition function $Z(\theta)$ (or its logarithm). Remarkably, in many settings this leads to simple iterative equations for learning of parameters.

3. The belief propagation reconstruction algorithm for compressed sensing

Exact computation of the averages (see equation (10)) requires exponential time and is thus intractable [48]. To approximate the expectations we will use a variant of the belief propagation (BP) algorithm [28, 29, 49]. Indeed, message passing has been shown very efficient in terms of both precision and speed for the CS problem. Our form of the message passing algorithm is closely related to the approximate message passing of [7] and is a special case of the generalized AMP of [8, 13]. We provide here an independent derivation of the algorithm.

3.1. Belief propagation recursion

The canonical BP equations for the probability measure $P(\mathbf{x}|\mathbf{F},\mathbf{y})$, equation (2), are expressed in terms of $2MN$ 'messages', $m_{j\to\mu}(x_j)$ and $m_{j\to\mu}(x_j)$, which are probability distribution functions. They read:

$$m_{\mu\to i}(x_i) = \frac{1}{Z^{\mu\to i}} \int \prod_{j\neq i} \mathrm{d}x_j\, \mathrm{e}^{-(1/2\Delta_\mu)(\sum_{j\neq i} F_{\mu j}x_j + F_{\mu i}x_i - y_\mu)^2} \prod_{j\neq i} m_{j\to\mu}(x_j), \quad (18)$$

$$m_{i\to\mu}(x_i) = \frac{1}{Z^{i\to\mu}} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \prod_{\gamma\neq\mu} m_{\gamma\to i}(x_i), \quad (19)$$

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

Using the Hubbard–Stratonovich transformation

$$\mathrm{e}^{-\omega^2/2\Delta} = \frac{1}{\sqrt{2\pi\Delta}} \int \mathrm{d}\lambda\, \mathrm{e}^{-\lambda^2/2\Delta + \mathrm{i}\lambda\omega/\Delta}, \quad (20)$$

for $\omega = (\sum_{j\neq i} F_{\mu j}x_j)$ we can simplify equation (18) as

$$m_{\mu\to i}(x_i) = \frac{1}{Z^{\mu\to i}\sqrt{2\pi\Delta}} \mathrm{e}^{-(1/2\Delta_\mu)(F_{\mu i}x_i - y_\mu)^2} \int \mathrm{d}\lambda\, \mathrm{e}^{-\lambda^2/2\Delta_\mu} \times \prod_{j\neq i} \left[\int \mathrm{d}x_j\, m_{j\to\mu}(x_j) \mathrm{e}^{(F_{\mu j}x_j/\Delta_\mu)(y_\mu - F_{\mu i}x_i + \mathrm{i}\lambda)} \right]. \quad (21)$$

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

$$a_{i\to\mu} \equiv \int \mathrm{d}x_i\, x_i\, m_{i\to\mu}(x_i), \quad (22)$$

$$v_{i\to\mu} \equiv \int \mathrm{d}x_i\, x_i^2\, m_{i\to\mu}(x_i) - a_{i\to\mu}^2, \quad (23)$$

we obtain

$$m_{\mu\to i}(x_i) = \frac{1}{Z^{\mu\to i}\sqrt{2\pi\Delta_\mu}} \mathrm{e}^{-(1/2\Delta_\mu)(F_{\mu i}x_i - y_\mu)^2} \int \mathrm{d}\lambda\, \mathrm{e}^{-\lambda^2/2\Delta_\mu} \times \prod_{j\neq i} [\mathrm{e}^{(F_{\mu j}a_{j\to\mu}/\Delta_\mu)(y_\mu - F_{\mu i}x_i + i\lambda) + (F_{\mu j}^2 v_{j\to\mu}/2\Delta_\mu^2)(y_\mu - F_{\mu i}x_i + i\lambda)^2}]. \quad (24)$$

Performing the Gaussian integral over λ, we obtain

$$m_{\mu\to i}(x_i) = \frac{1}{\tilde{Z}^{\mu\to i}} \mathrm{e}^{-(x_i^2/2)A_{\mu\to i}+B_{\mu\to i}x_i}, \qquad \tilde{Z}^{\mu\to i} = \sqrt{\frac{2\pi}{A_{\mu\to i}}} \mathrm{e}^{B_{\mu\to i}^2/2A_{\mu\to i}}, \tag{25}$$

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

$$A_{\mu\to i} = \frac{F_{\mu i}^2}{\Delta_\mu + \sum_{j\neq i} F_{\mu j}^2 v_{j\to\mu}}, \tag{26}$$

$$B_{\mu\to i} = \frac{F_{\mu i}(y_\mu - \sum_{j\neq i} F_{\mu j} a_{j\to\mu})}{\Delta_\mu + \sum_{j\neq i} F_{\mu j}^2 v_{j\to\mu}}. \tag{27}$$

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

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

$$m_{i\to\mu}(x_i) = \frac{1}{\tilde{Z}^{i\to\mu}}[(1-\rho)\delta(x_i) + \rho\phi(x_i)]\mathrm{e}^{-(x_i^2/2)\sum_{\gamma\neq\mu}A_{\gamma\to i}+x_i\sum_{\gamma\neq\mu}B_{\gamma\to i}}. \tag{28}$$

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

$$a_i \equiv \int \mathrm{d}x_i\, x_i\, m_i(x_i), \tag{29}$$

$$v_i \equiv \int \mathrm{d}x_i\, x_i^2\, m_i(x_i) - a_i^2, \tag{30}$$

where

$$m_i(x_i) = \frac{1}{\tilde{Z}^i}\,[(1-\rho)\delta(x_i) + \rho\phi(x_i)]\,\mathrm{e}^{-(x_i^2/2)\sum_\gamma A_{\gamma\to i}+x_i\sum_\gamma B_{\gamma\to i}}. \tag{31}$$

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

$$\mathcal{M}_\phi(\Sigma^2, R, x) = \frac{1}{\hat{Z}(\Sigma^2, R)}[(1-\rho)\delta(x) + \rho\phi(x)]\frac{1}{\sqrt{2\pi}\Sigma}\mathrm{e}^{-(x-R)^2/2\Sigma^2}, \tag{32}$$

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

$$f_a(\Sigma^2, R) \equiv \int \mathrm{d}x\, x\, \mathcal{M}(\Sigma^2, R, x), \tag{33}$$

$$f_c(\Sigma^2, R) \equiv \int \mathrm{d}x\, x^2\, \mathcal{M}(\Sigma^2, R, x) - f_a^2(\Sigma^2, R), \tag{34}$$

(where we do not write explicitly the dependence on ϕ). We give an explicit form for these two functions for the Gauss–Bernoulli signal model, equations (68)–(69), and for the mixture of Gaussians signal model in appendix C.

Notice that:

$$f_a(\Sigma^2, R) = R + \Sigma^2 \frac{\mathrm{d}}{\mathrm{d}R} \log \hat{Z}(\Sigma^2, R), \tag{35}$$

$$f_c(\Sigma^2, R) = \Sigma^2 \frac{\mathrm{d}}{\mathrm{d}R} f_a(\Sigma^2, R). \tag{36}$$

The closed form of the BP update is

$$a_{i\to\mu} = f_a\left(\frac{1}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}, \frac{\sum_{\gamma\neq\mu} B_{\gamma\to i}}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}\right), \qquad a_i = f_a\left(\frac{1}{\sum_{\gamma} A_{\gamma\to i}}, \frac{\sum_{\gamma} B_{\gamma\to i}}{\sum_{\gamma} A_{\gamma\to i}}\right), \tag{37}$$

$$v_{i\to\mu} = f_c\left(\frac{1}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}, \frac{\sum_{\gamma\neq\mu} B_{\gamma\to i}}{\sum_{\gamma\neq\mu} A_{\gamma\to i}}\right), \qquad v_i = f_c\left(\frac{1}{\sum_{\gamma} A_{\gamma\to i}}, \frac{\sum_{\gamma} B_{\gamma\to i}}{\sum_{\gamma} A_{\gamma\to i}}\right). \tag{38}$$

For a general signal model $\phi(x_i)$ the functions f_a and f_c can be computed using a numerical integration over x_i. In special cases, such as the case of Gaussian ϕ which we use in practice, these functions are easily computed analytically and are given in equations (68)–(69). Equations (22)–(23) together with (26)–(27) and (28) then lead to closed iterative message passing equations, which can be solved by iterations. There equations can be used for any signal $\mathbf{s}$, and any matrix $\mathbf{F}$. When a fixed point of the BP equations is reached, the elements of the original signal are estimated as $x_i^* = a_i$, and the corresponding variance v_i can be used to quantify the correctness of this estimate. Perfect reconstruction is found when the messages converge to a fixed point such that $a_i = s_i$ and $v_i = 0$.

A message passing algorithm equivalent to the one that we have just described was used in [11], where it was called 'relaxed belief propagation'. In [11], it was used as an approximate algorithm for the case of a sparse measurement matrix $\mathbf{F}$. In our case, the matrix is not sparse, and the use of mean and variances instead of the canonical BP messages is exact in the large N limit, thanks to the fact that the matrix is not sparse (a sum like $\sum_i F_{\mu_i} x_i$ contains of order N non-zero terms), and each element of the matrix F scales as $O(1/\sqrt{N})$.

3.2. The TAP form of the message passing algorithm

In the message passing form of BP described above, $2M\times N$ messages are sent, one between each variable component i and each measurement, in each iteration. In fact, it is possible to rewrite the BP equations in terms of $N+M$ messages instead of $2\ M\times N$, always within the assumption that the F matrix is not sparse, and that all its elements scale as $O(1/\sqrt{N})$. In statistical physics terms, this corresponds to the Thouless–Anderson–Palmer equations (TAP) [12] used in the study of spin glasses. In the large N limit, these are asymptotically equivalent (only $o(1)$ terms are neglected) to the BP equations. Going from BP to TAP is, in the compressed sensing literature, the step to go from the rBP [11] to the AMP [7] algorithm. Let us now show how to take this step.

In the large N limit, it is clear from (37) to (38) that the messages $a_{i\to\mu}$ and $v_{i\to\mu}$ are nearly independent of μ. However, one must be careful to keep the correcting 'Onsager reaction terms'. Let us define

$$\omega_\mu = \sum_i F_{\mu i} a_{i\to\mu}, \qquad V_\mu = \sum_i F_{\mu i}^2 v_{i\to\mu}, \tag{39}$$

$$\Sigma_i^2 = \frac{1}{\sum_\mu A_{\mu\to i}}, \qquad R_i = \frac{\sum_\mu B_{\mu\to i}}{\sum_\mu A_{\mu\to i}}. \tag{40}$$

Then we have

$$\Sigma_i^2 = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu - F_{\mu i}^2 v_{i\to\mu}}\right]^{-1} = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu}\right]^{-1}, \tag{41}$$

$$\begin{aligned} R_i &= \left[\sum_\mu \frac{F_{\mu i}(y_\mu - \omega_\mu + F_{\mu i} a_{i\to\mu})}{\Delta_\mu + V_\mu - F_{\mu i}^2 v_{i\to\mu}}\right]\left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu - F_{\mu i}^2 v_{i\to\mu}}\right]^{-1} \\ &= a_i + \frac{\sum_\mu F_{\mu i}(y_\mu - \omega_\mu)/(\Delta_\mu + V_\mu)}{\sum_\mu F_{\mu i}^2 1/(\Delta_\mu + V_\mu)}. \end{aligned} \tag{42}$$

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

$$a_{i\to\mu} = f_a\left(\frac{1}{\sum_\nu A_{\nu\to i} - A_{\mu\to i}}, \frac{\sum_\nu B_{\nu\to i} - B_{\mu\to i}}{\sum_\nu A_{\nu\to i} - A_{\mu\to i}}\right) = a_i - B_{\mu\to i}\Sigma^2 \frac{\partial f_a}{\partial R}\left(\Sigma_i^2, R_i\right). \tag{43}$$

Therefore

$$\omega_\mu = \sum_i F_{\mu i} a_i - \frac{(y_\mu - \omega_\mu)}{\Delta_\mu + V_\mu}\sum_i F_{\mu i}^2 v_i. \tag{44}$$

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

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

$$V_\mu^{t+1} = \sum_i F_{\mu i}^2\, v_i^t, \tag{45}$$

$$\omega_\mu^{t+1} = \sum_i F_{\mu i}\, a_i^t - \frac{(y_\mu - \omega_\mu^t)}{\Delta_\mu + V_\mu^t}\sum_i F_{\mu i}^2\, v_i^t, \tag{46}$$

$$(\Sigma_i^{t+1})^2 = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta_\mu + V_\mu^{t+1}}\right]^{-1}, \tag{47}$$

$$R_i^{t+1} = a_i^t + \frac{\sum_\mu F_{\mu i}(y_\mu - \omega_\mu^{t+1})/(\Delta_\mu + V_\mu^{t+1})}{\sum_\mu F_{\mu i}^2/(\Delta_\mu + V_\mu^{t+1})}, \tag{48}$$

$$a_i^{t+1} = f_a((\Sigma_i^{t+1})^2, R_i^{t+1}), \tag{49}$$

$$v_i^{t+1} = f_c((\Sigma_i^{t+1})^2, R_i^{t+1}). \tag{50}$$

We see that the signal model $P(x_i) = (1-\rho)\delta(x_i) + \rho\phi(x_i)$ assumed in the probabilistic approach appears only through the definitions (33) and (34) of the two functions f_a and f_c . In the case where the signal model is chosen as Gauss–Bernoulli, these functions are given explicitly by equations (68) and (69). Equations (45)–(50) are equivalent to the (generalized) approximate message passing of [7, 8].

A reasonable initialization of these equations is

$$a_i^{t=0} = \rho \int \mathrm{d}x\, x\, \phi(x), \tag{51}$$

$$v_i^{t=0} = \rho \int \mathrm{d}x\, x^2\, \phi(x) - (a_i^{t=0})^2, \tag{52}$$

$$\omega_\mu^{t=0} = y_\mu. \tag{53}$$

3.3. Further simplification for measurement matrices with random entries

For some special classes of random measurement matrices $\mathbf{F}$, the TAP equations (45)–(48) can be simplified further. Let us start with the case of a homogeneous matrix $\mathbf{F}$ with iid random entries of zero mean and variance $1/N$ (the distribution can be anything as long as the mean and variance are fixed). The simplification can be understood as follows. Consider for instance the quantity V_μ. Let us define $\bar{V}$ as the average of V_μ with respect to different realizations of the measurement matrix F.

$$\bar{V} = \sum_{i=1}^{N} \overline{F_{\mu i}^2} v_i = \frac{1}{N}\sum_{i=1}^{N} v_i. \tag{54}$$

The variance is

$$\begin{aligned}\mathrm{var}\, V \equiv \overline{(V_\mu - \bar{V})^2} &= \sum_{i\neq j} \overline{\left(F_{\mu i}^2 - \frac{1}{N}\right)\left(F_{\mu j}^2 - \frac{1}{N}\right)} v_i v_j + \sum_{i=1}^{N} \overline{\left(F_{\mu i}^2 - \frac{1}{N}\right)^2} v_i^2 \\ &= 0 + \frac{2}{N}\left(\frac{1}{N}\sum_{i=1}^{N} v_i^2\right) = O\left(\frac{1}{N}\right).\end{aligned} \tag{55}$$

Since the average is of order one and the variance of order $1/N$, in the limit of large N we can hence neglect the dependence on the index μ and consider all V_μ equal to their average. The same argument can be repeated for all the terms that contain $F_{\mu i}^2$. Hence for the homogeneous matrix $\mathbf{F}$ with iid random entries of zero mean and variance $1/N$, one can effectively 'replace' every $F_{\mu i}^2$ by $1/N$ in equations (47)–(48) and (45)–(46). The iteration equations then take the simpler form (assuming for simplicity that $\Delta_\mu = \Delta$)

$$V = \frac{1}{N}\sum_i v_i, \tag{56}$$

$$\omega_\mu = \sum_i F_{\mu i} a_i - \frac{(y_\mu - \omega_\mu)}{\Delta + V}\left[\frac{1}{N}\sum_i v_i\right], \tag{57}$$

$$\Sigma^2 = \frac{\Delta + V}{\alpha}, \tag{58}$$

$$R_i = a_i + \sum_\mu F_{\mu i}\frac{(y_\mu - \omega_\mu)}{\alpha}. \tag{59}$$

$$a_i = f_a(\Sigma^2, R_i), \tag{60}$$

$$v_i = f_c(\Sigma^2, R_i). \tag{61}$$

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

A second class of matrices for which a similar simplification exists is the case of the block matrices defined in section 1.2. For simplicity, we consider the case when the noise only depends on the block, i.e., $\Delta_\mu = \Delta_q$ for all μ in block q. For the block measurement matrix with random entries of variance $J_{q,p}/N$ the simplified TAP equations read

$$V_q = \frac{1}{N}\sum_{p=1}^{L_c} J_{q,p}\sum_{i\in B_p} v_i, \tag{62}$$

$$\omega_\mu = \sum_{p=1}^{L_c}\sum_{i\in B_p} F_{\mu i}a_i - \frac{y_\mu-\omega_\mu}{\Delta_{I(\mu)}+V_{I(\mu)}}\frac{1}{N}\sum_{p=1}^{L_c} J_{I(\mu),p}\sum_{i\in B_p} v_i, \tag{63}$$

$$\Sigma_p^2 = \left[n_p\sum_{q=1}^{L_r}\frac{\alpha_{qp}J_{q,p}}{\Delta_q+V_q}\right]^{-1}, \tag{64}$$

$$R_i = a_i + \frac{\sum_{q=1}^{L_r}\sum_{\mu\in B_q} F_{\mu i}(y_\mu-\omega_\mu)/(\Delta_q+V_q)}{n_{I(i)}\sum_{q=1}^{L_r}\alpha_{qI(i)}J_{q,I(i)}/(\Delta_q+V_q)}, \tag{65}$$

$$a_i = f_a(\Sigma_{I(i)}^2, R_i), \tag{66}$$

$$v_i = f_c(\Sigma_{I(i)}^2, R_i), \tag{67}$$

where $p=1,2,\ldots,L_c$, $q=1,2,\ldots,L_r$. $I(\mu)$ (and $I(i)$) is defined as the index of the block to which μ (i) belongs, B_q is the set of indices in block q. We remind that $\alpha_{qp}=M_q/N_p$ and $n_p=N_p/N$.

3.4. Parameter learning with expectation maximization

In our practical implementation, we use as signal model a Gauss–Bernoulli distribution. That is, the function $\phi(x)$ is Gaussian with mean $\bar{x}$ and variance σ^2. The functions f_a and f_c are in this case:

$$f_a(\Sigma^2,R) = \frac{\rho\,\mathrm{e}^{-(R-\bar{x})^2/2(\Sigma^2+\sigma^2)}(\Sigma/(\Sigma^2+\sigma^2)^{3/2})(\bar{x}\Sigma^2+R\sigma^2)}{(1-\rho)\mathrm{e}^{-R^2/2\Sigma^2}+\rho(\Sigma/(\sqrt{\Sigma^2+\sigma^2}))\mathrm{e}^{-(R-\bar{x})^2/2(\Sigma^2+\sigma^2)}}, \tag{68}$$

$$\begin{aligned} f_c(\Sigma^2,R) = {}& \left\{\rho\,(1-\rho)\mathrm{e}^{-(R^2/2\Sigma^2)-(R-\bar{x})^2/2(\Sigma^2+\sigma^2)}\frac{\Sigma}{(\Sigma^2+\sigma^2)^{5/2}}\right. \\ & \left.\times\,[\sigma^2\Sigma^2(\Sigma^2+\sigma^2)+(\bar{x}\Sigma^2+R\sigma^2)^2]+\rho^2\mathrm{e}^{-(R-\bar{x})^2/(\Sigma^2+\sigma^2)}\frac{\sigma^2\Sigma^4}{(\sigma^2+\Sigma^2)^2}\right\} \\ & \times\left\{\left[(1-\rho)\mathrm{e}^{-R^2/2\Sigma^2}+\rho\frac{\Sigma}{\sqrt{\Sigma^2+\sigma^2}}\mathrm{e}^{-(R-\bar{x})^2/2(\Sigma^2+\sigma^2)}\right]^2\right\}^{-1}. \end{aligned} \tag{69}$$

See also appendix C where we give the form of f_a and f_c for the signal model consisting of a mixture of Gaussians.

The most likely values of parameters $\rho, \bar{x}, \sigma, \Delta$ can be obtained via maximizing the partition function. Within the belief propagation approach this is equivalent to maximizing the Bethe free entropy $F(\rho,\bar{x},\sigma,\Delta)\equiv\log Z(\rho,\bar{x},\sigma,\Delta)$ expressed as [49]

$$F(\rho,\bar{x},\sigma,\Delta)=\sum_{\mu}\log Z^{\mu}+\sum_{i}\log Z^{i}-\sum_{(\mu i)}\log Z^{\mu i}, \tag{70}$$

where

$$Z^{i}=\int \mathrm{d}x_i\prod_{\mu}m_{\mu\to i}(x_i)\left[(1-\rho)\delta(x_i)+\frac{\rho}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x_i-\bar{x})^2/2\sigma^2}\right], \tag{71}$$

$$Z^{\mu i}=\int \mathrm{d}x_i m_{\mu\to i}(x_i)m_{i\to\mu}(x_i). \tag{72}$$

$$\begin{aligned}Z^{\mu}&=\int\prod_{i}\mathrm{d}x_i\prod_{i}m_{i\to\mu}(x_i)\frac{1}{\sqrt{2\pi\Delta_\mu}}\mathrm{e}^{-(y_\mu-\sum_i F_{\mu i}x_i)^2/2\Delta_\mu}\\ &=\frac{1}{\sqrt{2\pi(\Delta+V_\mu)}}\mathrm{e}^{-(y_\mu-\omega_\mu)^2/2(\Delta+V_\mu)}.\end{aligned} \tag{73}$$

The stationarity condition of Bethe free entropy (70) with respect to ρ leads to

$$\rho=\frac{\sum_i(1/\sigma^2+1/\Sigma_i^2)/(R_i/\Sigma_i^2+\bar{x}/\sigma^2)a_i}{\sum_i[1-\rho+\rho/\sigma(1/\sigma^2+1/\Sigma_i^2)^{1/2}\mathrm{e}^{(R_i/\Sigma_i^2+\bar{x}/\sigma^2)^2/2(1/\sigma^2+1/\Sigma_i^2)-\bar{x}^2/2\sigma^2}]^{-1}}. \tag{74}$$

Stationarity with respect to $\bar{x}$ and σ gives

$$\bar{x}=\frac{\sum_i a_i}{\rho\sum_i\left[\rho+(1-\rho)\sigma(1/\sigma^2+1/\Sigma_i^2)^{1/2}\mathrm{e}^{-(R_i/\Sigma_i^2+\bar{x}/\sigma^2)^2/2(1/\sigma^2+1/\Sigma_i^2)+\bar{x}^2/2\sigma^2}\right]^{-1}}, \tag{75}$$

$$\sigma^2=\frac{\sum_i(v_i+a_i^2)}{\rho\sum_i\left[\rho+(1-\rho)\sigma(1/\sigma^2+1/\Sigma_i^2)^{1/2}\mathrm{e}^{-(R_i/\Sigma_i^2+\bar{x}/\sigma^2)^2/2(1/\sigma^2+1/\Sigma_i^2)+\bar{x}^2/2\sigma^2}\right]^{-1}}-\bar{x}^2. \tag{76}$$

For simplicity, we consider that the noise is homogeneous, i.e., $\Delta_\mu=\Delta$, for all μ. The noise level Δ may be unknown, in which case one can learn it, like the other parameters, by maximizing the free entropy. The resulting condition for learning of the noise variance Δ reads:

$$\Delta=\frac{\sum_\mu(y_\mu-\omega_\mu)^2/(1+(1/\Delta)V_\mu)^2}{\sum_\mu 1/(1+(1/\Delta)V_\mu)}, \tag{77}$$

where ω_μ and V_μ are defined in equation (39).

Note that instead of using the steepest gradient descent in the Bethe free energy for the mean and variance (i.e. equations (75) and (76)) one can also use simpler expressions that are satisfied in the Bayes-optimal setting. In particular

$$\bar{x}=\frac{\sum_i a_i}{N\rho}, \tag{78}$$

$$\sigma^2=\frac{\sum_i(v_i+a_i^2)}{\rho N}-\bar{x}^2. \tag{79}$$

In our numerical implementations we use these simplified conditions. In the case where the matrix $\mathbf{F}$ is random with iid elements of zero mean and variance $1/N$, we can also use for learning the variance: $\sum_{\mu=1}^{M}y_\mu^2/N=\alpha\rho(\sigma^2+\bar{x}^2)$.

Equations (74)–(76) or (78) and (79) can be used for iterative learning of the parameters, in the spirit of expectation maximization. Equations (26), (27), (37), (38), (74), (78) and (79) altogether lead to the expectation maximization belief propagation (EM-BP) algorithm that we have first presented in [1]. In EM-BP one update of the BP messages is followed by an update of the parameters and this is repeated till convergence (of both BP messages and the parameters). In our implementations we initialize the parameters as follows

$$\rho^{t=0} = \alpha/10, \qquad \bar{x}^{t=0} = 0, \qquad \sigma^2_{t=0} = 1. \tag{80}$$

In the case the variance of the signal is not at all close to one, the sum rule $(1/M)\sum_\mu y_\mu^2 = (1/M)\sum_{\mu,i} F_{\mu i}^2 s_i^2$ suggests a more sensible initialization $\sigma^2_{t=0} = \sum_\mu y_\mu^2/(MN\mathrm{var}F\rho^{t=0})$. A new guess of parameters is obtained using equations (74), (78) and (79) except if the variance becomes negative, then the new variance is set to zero, or if the new value of ρ becomes larger than α, in which case α is taken as the new value for ρ. To obtain an updated guess for the parameters we also use 'damping'. The updated guess is obtained as 1/2 times the old value plus 1/2 times the newly computed value. Empirically this speeds up the convergence and prevents some numerical instabilities. If needed, such damping is also used to improve convergence for the BP messages themselves.

4. Asymptotic analysis: state evolution and replicas

Belief propagation is an efficient heuristic algorithm that is in some cases (such as the present one) amenable to asymptotic ($N \to \infty$) analytical analysis. This statistical analysis of BP iterations is known as the 'cavity method' (in statistical physics) [49, 50], the 'density evolution' in coding [51], and the 'state evolution' in the context of CS [7, 15]. The corresponding equations can also be derived using the replica method, that provides an exact asymptotic analysis of both the BP performance and the performance of an optimal (perhaps exponentially costly) reconstruction algorithm. In this section we first concentrate (parts A–D) on the case of 'homogeneous' measurement matrices with iid entries. We derive the density evolution equations in part A, and we detail the replica approach in part B. Part C shows the simplifications that take place in the Bayes-optimal case where the signal model gives the correct statistical properties of the underlying signal, and part D generalizes the density evolution equations to the case where one uses the learning procedure for the parameters of the signal model. Part E gives the density evolution equations in the more general case of block measurement matrices.

4.1. Density evolution of the message passing

We derive the density evolution equations in the case where the measurement matrix F has random entries that are iid, with mean 0 and variance $1/N$, and we assume that the parameters of the signal model are fixed.

The density evolution (or cavity method) uses a statistical analysis of the BP messages at iteration t, in the large N limit, in order to derive their distribution at iteration $t + 1$.

It turns out that these distributions are simply expressed in terms of two parameters:

$$V^t \equiv \frac{1}{N}\sum_{i=1}^{N} v_i^t \tag{81}$$

$$E^t \equiv \frac{1}{N}\sum_{i=1}^{N}(a_i^t - s_i)^2. \tag{82}$$

We remind the reader that s_i are the components of the original signal $\mathbf{s}$, and a_i^t, v_i^t are the mean and variance of the local beliefs defined in (29), at iteration t. V^t just measures the average variance of the local beliefs, and E^t is the mean-squared error achieved by BP, at a given iteration t.

Using the definition of the quantity R_i (40) and equation (27), we get

$$R_i^t = s_i + \frac{1}{\alpha}\left[\sum_\mu F_{\mu i}\xi_\mu + \sum_\mu F_{\mu i}\sum_{j\neq i} F_{\mu j}(s_j - a_{j\to\mu}^t)\right], \tag{83}$$

where ξ_μ is the measurement noise (as defined in (1)), a centered Gaussian variable with variance Δ_0. The variable $r_i^t = \sum_\mu F_{\mu i}\xi_\mu + \sum_\mu F_{\mu i}\sum_{j\neq i}F_{\mu j}(s_j - a_{j\to\mu}^t)$ is a random variable with respect to the distribution of the measurement matrix elements $F_{\mu i}$ (zero mean and $1/N$ variance matrix) and the noise ξ_i. Therefore r_i^t is a Gaussian random variable with mean and variance

$$\overline{r^t} = \sum_\mu \sum_{j\neq i} \overline{F_{\mu i}F_{\mu j}}(s_j - a_{j\to\mu}) = 0, \tag{84}$$

$$\begin{aligned}\overline{(r^t)^2} &= \sum_\mu \xi_\mu^2 F_{\mu i}^2 + \sum_\mu\sum_{j\neq i}\overline{F_{\mu i}^2 F_{\mu j}^2}(s_j - a_{j\to\mu})^2 \\ &= \alpha\Delta_0 + \frac{1}{N^2}\sum_{\mu=1}^{M}\sum_{j=1}^{N}(s_j - a_j)^2 = \alpha(E+\Delta_0).\end{aligned} \tag{85}$$

In the second inequality of (85) we neglected terms of $O(1/\sqrt{N})$.

Using the above results this leads us to the belief at iteration $t+1$, $m_i^{t+1}(x_i)$, being distributed as

$$m_i^{t+1}(x_i) \simeq \frac{1}{\hat{Z}^i}[(1-\rho)\delta(x_i) + \rho\phi(x_i)]\mathrm{e}^{-\alpha(x_i - s_i - z\sqrt{E+\Delta_0/\alpha})^2/2(\Delta+V)}, \tag{86}$$

where z is a random Gaussian variable with zero mean and unit variance, and $\hat{Z}^i$ is a normalization constant. Hence, using the definition of the BP order parameters given in (82), we get for a signal with iid elements

$$V^{t+1} = \int \mathrm{d}s\,[(1-\rho_0)\delta(s) + \rho_0\phi_0(s)]\int \mathcal{D}z\, f_c\left(\frac{\Delta+V^t}{\alpha}, s + z\sqrt{\frac{E^t+\Delta_0}{\alpha}}\right), \tag{87}$$

$$E^{t+1} = \int \mathrm{d}s\,[(1-\rho_0)\delta(s) + \rho_0\phi_0(s)]\int \mathcal{D}z\left[f_a\left(\frac{\Delta+V^t}{\alpha}, s + z\sqrt{\frac{E^t+\Delta_0}{\alpha}}\right) - s\right]^2, \tag{88}$$

where $\mathcal{D}z = \mathrm{d}z\,\mathrm{e}^{-z^2/2}/\sqrt{2\pi}$ is a Gaussian integration measure. For the special case of a Gauss–Bernoulli signal model, i.e. when the function ϕ is Gaussian with mean $\bar{x}$ and variance σ^2, the functions $f_a(\Sigma^2, R)$ and $f_c(\Sigma^2, R)$ are expressed explicitly in equations (68) and (69).

Equations (87) and (88) are the density evolution equations. They describe how the mean-squared error E and the variance order parameter V evolve during the iterations of the BP algorithm. Note that the density evolution equations are the same for the message passing and for the TAP equations, as indeed factors of $O(1/N)$ are neglected in the density evolution. If the messages are initialized as in (51)–(53), the initial conditions of the density evolution equations are:

$$E^{t=0} = \rho_0\overline{s^2} - 2\rho\rho_0\bar{s}\int \mathrm{d}x\, x\phi(x) + \rho^2\left[\int \mathrm{d}x\, x\phi(x)\right]^2, \tag{89}$$

$$V^{t=0} = \rho\int \mathrm{d}x\, x^2\phi(x) - \rho^2\left[\int \mathrm{d}x\, x\phi(x)\right]^2. \tag{90}$$

Figure 1 shows several examples of this mapping for the noiseless case $\Delta = \Delta_0 = 0$. We plot the evolution of the normalized vector $(V^{(t+1)} - V^{(t)}, E^{(t+1)} - E^{(t)})$. For a relatively high measurement density α, there is unique fixed point $E = V = 0$ corresponding to an exact reconstruction of the signal. When α is below some critical point, another attractive fixed point $E > 0, V > 0$ appears.

4.2. Replica analysis

The density evolution presented in the previous section can also be derived independently using the replica method [50]. The main advantage is that the replica computations give a physical meaning to all the fixed points of equations (87)–(88), even to those that are not reached by iterating the BP algorithm.

The thermodynamic properties of a disordered system given by the Hamiltonian defined in equation (3) are characterized by the average free entropy $\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}}(\log Z)$, where Z is the partition function defined in (2), $\mathbf{s}$ is the original signal and $\boldsymbol{\xi} = \{\xi_\mu\}_{\mu=1}^M$ are the measurement noise with zero mean and variance Δ_0 for $\mu = 1, 2, \ldots, M$. The free entropy is evaluated via the replica trick as

$$\Phi \equiv \frac{1}{N}\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}}(\log Z) = \frac{1}{N}\lim_{n\to 0}\frac{\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}}(Z^n) - 1}{n}. \tag{91}$$

Introducing n replicas, we get

$$\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}}(Z^n) = \int \prod_{i,a}\mathrm{d}x_i^a \prod_{i,a}[(1-\rho)\delta(x_i^a) + \rho\phi(x_i^a)] \\ \times \prod_{\mu}\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}}\frac{1}{\sqrt{2\pi\Delta}}\mathrm{e}^{-(1/2\Delta)\sum_{a=1}^{n}\left(\sum_{i=1}^{N}F_{\mu i}s_i + \xi_\mu - \sum_{i=1}^{N}F_{\mu i}x_i^a\right)^2}, \tag{92}$$

where $a, b, \ldots$ denote the replica indices, Δ is the assumed measurement noise and generally $\Delta \neq \Delta_0$.

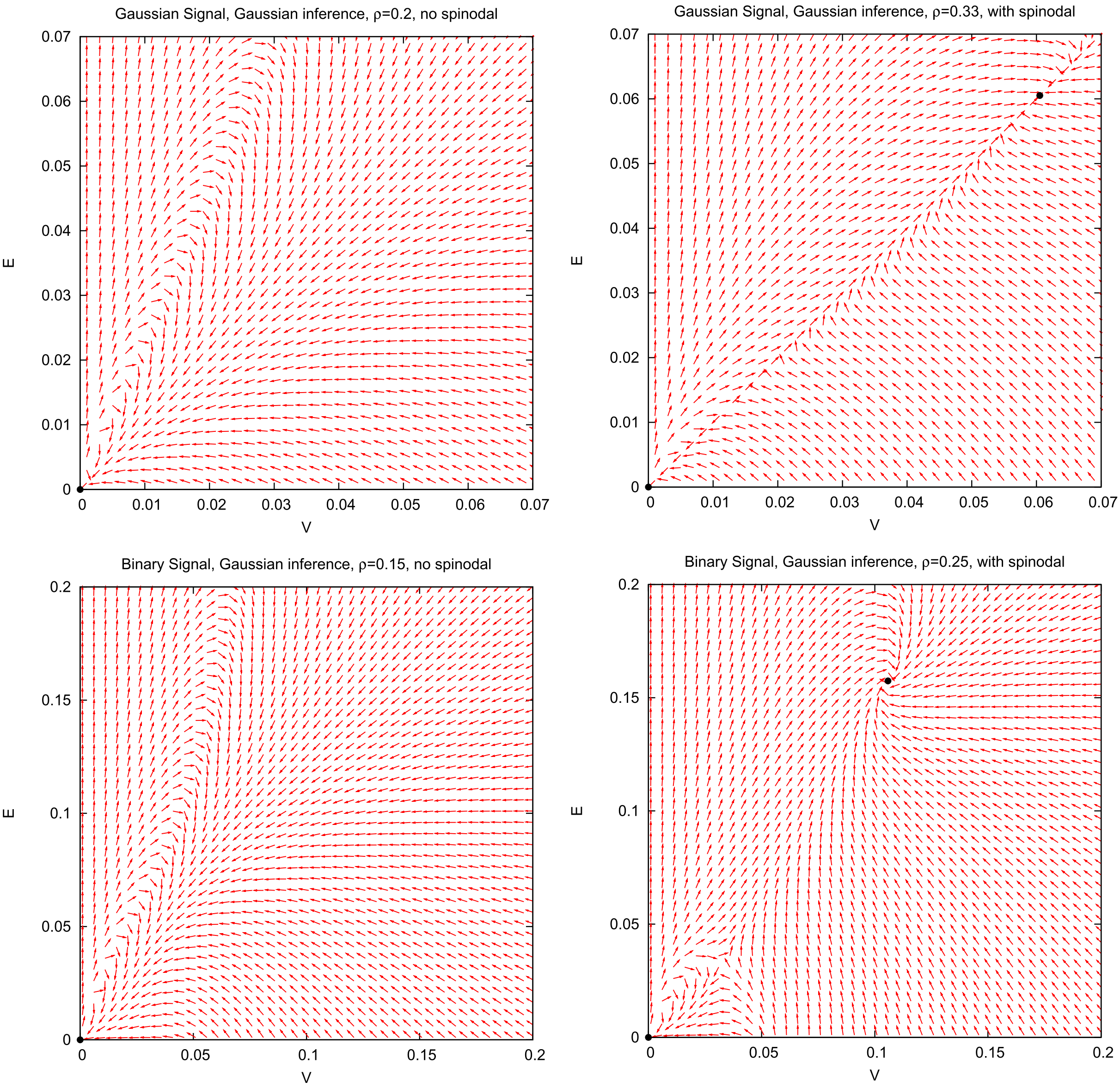

Figure 1. Examples of the BP density evolution, the y-axes give the mean-squared error of the current signal estimate $E = q - 2m + \rho_0 \overline{s^2}$, the x-axes give the average variance $V = Q - q$. Each arrow is a normalized vector $(V^{(t+1)} - V^{(t)}, E^{(t+1)} - E^{(t)})$. The signal model $\phi(x)$ is Gaussian with zero mean and unit variance, the signal distribution $\phi_0(x)$ is Gaussian on the top and $\{\pm 1\}$ on the bottom. The measurements are noiseless. On the left we show an example for relatively large measurement rate where there is a unique fixed point $E \to 0, V \to 0$. On the right there is another fixed point $E > 0, V > 0$, which is the attractive one for 'uninformed' initial conditions. Notice that on the top plots the line $V = E$ is stable: this is thanks to the Nishimori condition when the signal is described by the correct model ($\rho_0 = \rho$ and $\phi_0 = \phi$). In that case one can work in the $V = E$ sub-space.

In the case where the matrix $\mathbf{F}$ has iid elements with zero mean and variance $1/N$, we introduce the order parameters as follows

$$m^a = \frac{1}{N}\sum_{i=1}^{N} x_i^a s_i, \qquad a = 1, 2, \ldots, n, \tag{93}$$

$$Q^a = \frac{1}{N}\sum_{i=1}^{N} (x_i^a)^2, \qquad a = 1, 2, \ldots, n, \tag{94}$$

$$q^{ab} = \frac{1}{N}\sum_{i=1}^{N} x_i^a x_i^b, \qquad a < b. \tag{95}$$

We use a common trick of rewriting the identity

$$1 = \int \prod_a \mathrm{d}\hat{Q}_a \, \mathrm{d}Q_a \, \mathrm{d}\hat{m}_a \, \mathrm{d}m_a \\ \times \int \prod_{a<b} \mathrm{d}\hat{q}_{ab} \, \mathrm{d}q_{ab} \, \mathrm{e}^{\sum_a \hat{Q}_a [(N/2)Q_a - (1/2)\sum_j (x_j^a)^2] - \sum_{a<b} \hat{q}_{ab}(N q_{ab} - \sum_j x_j^a x_j^b) - \sum_a \hat{m}_a (N m_a - \sum_j x_j^a x_j^0)}.$$

When averaging Z^n, we first need to evaluate the quantity

$$X_\mu = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}[\mathrm{e}^{-(1/2\Delta)\sum_{a=1}^{n}(\sum_i F_{\mu i} s_i + \xi_\mu - \sum_i F_{\mu i} x_i^a)^2}] \tag{96}$$

at fixed signal $\mathbf{s}$ and configuration $\mathbf{x}$. In order to evaluate X_μ we first need to define $v_\mu^a = \sum_{i=1}^{N} F_{\mu i}(x_i^0 - x_i^a) + \xi_\mu$ with $a = \{0, 1, \ldots, n\}$, and where 0 corresponds to the index of the signal: $x_i^0 = s_i$. The v^a obeys a joint Gaussian distribution with covariance

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\left[\left(v_\mu^a\right)^2\right] = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}} \sum_i F_{\mu i}^2 \left(x_i^0 - x_i^a\right)^2 + \Delta_0 \\ = \frac{1}{N}\sum_i \left(x_i^0 - x_i^a\right)^2 + \Delta_0 = Q^a - 2m^a + \rho\overline{s^2} + \Delta_0 \tag{97}$$

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\left[v_\mu^a v_\mu^b\right] = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}} \sum_i F_{\mu i}^2 \left(x_i^0 - x_i^a\right)\left(x_i^0 - x_i^b\right) + \Delta_0 = q^{ab} - (m^a + m^b) + \rho\overline{s^2} + \Delta_0. \tag{98}$$

We shall use the so-called replica symmetric (RS) ansatz. This is consistent with using belief propagation, and it is known to be correct for the optimal Bayesian inference (i.e. when the signal model correspond to the empirical signal distribution) [46, 49]. In this ansatz the replicas are considered as equivalent, therefore:

$$m^a = m, \qquad q^{ab} = q, \qquad Q^a = Q. \tag{99}$$

Going back to X_μ, we now have

$$X_\mu = \mathbb{E}_{\mathbf{v}}\left[\mathrm{e}^{-(1/2\Delta)\sum_{a=1}^{n}\left(v_\mu^a\right)^2}\right] \tag{100}$$

with

$$P(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^n \det(G)}} \mathrm{e}^{-(1/2)\sum_{a,b} v_a (G^{-1})_{ab} v_b}, \tag{101}$$

where (under the RS hypothesis) the covariance matrix reads

$$G_{aa} = E_{\mathbf{v}}(v_\mu^a v_\mu^a) = Q + \rho\overline{s^2} - 2m + \Delta_0, \qquad a = 1, 2, \ldots, n, \tag{102}$$

$$G_{ab} = E_{\mathbf{v}}(v_\mu^a v_\mu^b) = q + \rho\overline{s^2} - 2m + \Delta_0, \qquad a < b. \tag{103}$$

Computing explicitly X_μ, one now finds

$$\begin{aligned} X_\mu &= \frac{1}{\sqrt{(2\pi)^n \det(G)}} \int D\mathbf{v}\, \mathrm{e}^{-(1/2)\sum\limits_{a,b} v^a \left[(G^{-1})_{ab} + (1/\Delta)\delta_{a,b}\right] v^b} \\ &= \frac{\int D\mathbf{v} \mathrm{e}^{-(1/2)\mathbf{v}^T (G^{-1} + \mathbb{1}/\Delta)\mathbf{v}}}{\int D\mathbf{v} \mathrm{e}^{-(1/2)\mathbf{v}^T G^{-1} \mathbf{v}}} = \frac{1}{\sqrt{\det(\mathbb{1} + G/\Delta)}}. \end{aligned} \tag{104}$$

We now compute this determinant. We have

$$G = (q + \rho\overline{s^2} - 2m + \Delta_0)\, \mathrm{II} + (Q - q)\mathbb{1}, \tag{105}$$

where II stands for the $n \times n$ matrix with elements all equal to one. The eigenvectors of G are (a) one eigenvector $(1, 1, \ldots, 1)$ with an eigenvalue $Q - q + n(q - 2m + \rho\overline{s^2} + \Delta_0)$, and (b) $n - 1$ eigenvectors of the type $(0, 0, 1, -1, 0, \ldots, 0)$ with eigenvalues $Q - q$. Therefore

$$\det\left(\mathbb{1} + \frac{G}{\Delta}\right) = \left[1 + \frac{1}{\Delta}(Q - q + n(q - 2m + \rho\overline{s^2} + \Delta_0))\right]\left[1 + \frac{Q - q}{\Delta}\right]^{n-1}. \tag{106}$$

To conclude the computation of X_μ we get

$$\lim_{n\to 0} X_\mu = \mathrm{e}^{-n/2[(q - 2m + \rho\overline{s^2} + \Delta_0)/(Q - q + \Delta) + \log(1 + (Q - q)/\Delta)]}. \tag{107}$$

We thus obtain

$$\begin{aligned} \mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}} Z^n = &\int \prod_a \mathrm{d}\hat{Q}_a\, \mathrm{d}Q_a\, \mathrm{d}\hat{m}_a\, \mathrm{d}m_a \\ &\times \int \prod_{ab} \mathrm{d}\hat{q}_{ab}\, \mathrm{d}q_{ab}\, \mathrm{e}^{N[(1/2)\sum_a \hat{Q}_a Q_a - \sum_{a<b} \hat{q}_{ab} q_{ab} - \sum_a \hat{m}_a m_a]} \prod_\mu \frac{X_\mu}{\sqrt{2\pi\Delta}} \\ &\times \left\{ \int \mathrm{d}x_0 \left[(1 - \rho_0)\delta(x_0) + \rho_0 \phi_0(x_0)\right] \right. \\ &\times \prod_a dx_a \left[(1 - \rho)\delta(x_a) + \rho\phi(x_a)\right] \\ &\times \left. \mathrm{e}^{-(1/2)\sum_a \hat{Q}_a x_a^2 + (1/2)\sum_{a\neq b} x_a x_b \hat{q}_{ab} + \sum_a \hat{m}_a x_a x_0} \right\}^N. \end{aligned} \tag{108}$$

Let us call Y the expression in the $\{\cdot\}$ in the last equation. Introducing the following transformation into the last equation

$$\mathrm{e}^{(1/2)\hat{q}_p \sum_{a\neq b} x^a x^b} = \int Dz \mathrm{e}^{z\sqrt{\hat{q}_p}\sum_{a=1}^n x_a} \mathrm{e}^{-(\hat{q}_p/2)\sum_{a=1}^n (x^a)^2}, \tag{109}$$

where $\mathcal{D}z$ is a Gaussian integration measure with zero mean and variance one, we obtain under the RS hypothesis

$$Y = \int \mathrm{d}x_0 \; [(1-\rho_0)\delta(x_0) + \rho_0\phi_0(x_0)] \\ \times \int Dz \left\{ \int \mathrm{d}x \, [(1-\rho)\delta(x) + \rho\phi(x)] \, \mathrm{e}^{-((\hat{Q}+\hat{q})/2)x^2 + \hat{m}xx_0 + z\sqrt{\hat{q}}x} \right\}^n . \qquad (110)$$

In the $n \to 0$ limit, one can write that $f(z)^n = 1 + n \log f(z)$ and thus $\int Dz f(z)^n = 1 + n \int Dz \log f(z) \approx \mathrm{e}^{n \int Dz \log f(z)}$. Grouping all terms together we finally get

$$\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}} Z^n = \int \mathrm{d}\hat{Q}\, \mathrm{d}Q\, \mathrm{d}\hat{q}\, \mathrm{d}q\, \mathrm{d}\hat{m}\, \mathrm{d}m\, \mathrm{e}^{nN\Phi(Q,q,m,\hat{Q},\hat{q},\hat{m})}, \qquad (111)$$

where Φ is the replica free energy function

$$\Phi(Q,q,m,\hat{Q},\hat{q},\hat{m}) = -\frac{\alpha}{2} \frac{q - 2m + \rho_0\overline{s^2} + \Delta_0}{\Delta + Q - q} - \frac{\alpha}{2}\log(\Delta + Q - q) + \frac{Q\hat{Q}}{2} - m\hat{m} + \frac{q\hat{q}}{2} \\ + \int \mathrm{d}s \, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z \log \\ \times \left\{ \int \mathrm{d}x \, \mathrm{e}^{-((\hat{Q}+\hat{q})/2)x^2 + \hat{m}xs + z\sqrt{\hat{q}}x} [(1-\rho)\delta(x) + \rho\phi(x)] \right\}. \qquad (112)$$

Note that $\mathcal{D}z$ is a Gaussian integration measure with zero mean and variance one, ρ_0 is the density of the signal, and $\phi_0(s)$ is the distribution of the signal components and $\overline{s^2} = \int \mathrm{d}s\, s^2\, \phi_0(s)$ is its second moment, Δ_0 is the true variance of the measurement noise.

The physical meaning of the order parameters is

$$Q = \frac{1}{N}\sum_i \langle x_i^2 \rangle, \qquad q = \frac{1}{N}\sum_i \langle x_i \rangle^2, \qquad m = \frac{1}{N}\sum_i s_i \langle x_i \rangle, \qquad (113)$$

in which the average is with respect to the measure $\hat{P}$ (2), whereas the other three $\hat{m}$, $\hat{q}$, $\hat{Q}$ are auxiliary parameters. Using the saddle point method and performing derivatives with respect to m, q, $Q-q$, $\hat{m}$, $\hat{q}$, and $\hat{Q}+\hat{q}$ we obtain the self-consistent equations

$$\hat{m} = \hat{Q} + \hat{q} = \frac{\alpha}{\Delta + Q - q}, \qquad \hat{q} = \frac{\alpha(q - 2m + \rho_0\overline{s^2} + \Delta_0)}{(\Delta + Q - q)^2}, \qquad (114)$$

$$m = \rho_0 \int \mathrm{d}s\, s\, \phi_0(s) \int \mathcal{D}z\, f_a\left(\frac{1}{\hat{m}}, s + z\frac{\sqrt{\hat{q}}}{\hat{m}}\right), \qquad (115)$$

$$Q - q = \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z\, f_c\left(\frac{1}{\hat{m}}, s + z\frac{\sqrt{\hat{q}}}{\hat{m}}\right), \qquad (116)$$

$$q = \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z\, f_a^2\left(\frac{1}{\hat{m}}, s + z\frac{\sqrt{\hat{q}}}{\hat{m}}\right). \qquad (117)$$

From the definition of the order parameters (113) we obtain

$$E = q - 2m + \rho_0\overline{s^2}, \qquad V = Q - q. \qquad (118)$$

It is easily seen that the set of stationary point equations (114)–(117) exactly reproduces the fixed-point condition of the density evolution equations (87)–(88): BP fixed points are stationary points of the free entropy (112).

The uniform sampling from the measure $\hat{P}$, equation (2), is described by the global maximum of Φ. We can use equation (112) in order to confirm (non-rigorously) our previous result about the optimality of the probabilistic approach for any $\phi(x)$ with a support that contains that of the signal ϕ_0, and finite second moment. Indeed the free entropy Φ, evaluated close to the signal, i.e. when $Q = q = m = \rho_0 \overline{s^2}$, diverges as $-(\alpha - \rho_0)\log(\Delta + Q - q)/2$. Therefore in the noiseless limit $\Delta \to 0$, Φ diverges when $E, V \to 0$, whenever $\alpha > \rho_0$.

It is useful to compute the free entropy restricted to configurations $\mathbf{x}$ at a fixed squared distance D from the signal, $D = \sum_i (x_i - s_i)^2/N$. When sampling from the probability $\hat{P} = P(\mathbf{x}|\mathbf{F}, \mathbf{y})$, in the limit of large N, the probability that the reconstructed signal $\mathbf{x}$ is at a squared distance $D = \sum_i (x_i - s_i)^2/N$ from the original signal $\mathbf{s}$ is proportional to $\mathrm{e}^{N\Phi(D)}$, where $\Phi(D)$ is the free entropy restricted to squared distance D. In order to compute $\Phi(D)$ we need to evaluate the following saddle point

$$\Phi(D) = \mathrm{SP}_{Q,q,\hat{Q},\hat{q},\hat{m}} \Phi(Q, q, (Q - D + \rho_0 \langle s^2 \rangle)/2, \hat{Q}, \hat{q}, \hat{m}), \tag{119}$$

which can be done using equations (115)–(117) and $\hat{q} = \alpha(q - 2m + \rho_0 \langle s^2 \rangle)/(Q - q)^2$, and $\hat{m} = \hat{Q} + \hat{q}$. The resulting free entropy $\Phi(D)$ is a useful quantity to visualize when the BP reconstruction fails. It will be shown and analyzed in section 4.3.

Let us, at this point, underline the difference between distance $D = \sum_i \langle (x_i - s_i)^2 \rangle / N = Q - 2m + \rho_0 \overline{s^2}$ and the mean-squared error $E = \sum_i (\langle x_i \rangle - s_i)^2/N = q - 2m + \rho_0 \overline{s^2}$. Clearly $D = E + V$, and one should not confuse the two definitions.

4.3. Analysis of Bayes-optimal inference

So far we have been discussing the general case when the signal is created using density ρ_0 and empirical distribution of the non-zero elements ϕ_0, and the belief propagation reconstruction algorithm is used with a signal model with density $\rho \neq \rho_0$ and entry distribution $\phi \neq \phi_0$. As we explained in section 2.2 the Bayes-optimal inference corresponds to the case when the statistical properties of the signal, and the distribution of the measurement noise are known. Then one can use a signal model with

$$\rho = \rho_0, \qquad \phi(x) = \phi_0(x), \qquad \Delta = \Delta_0. \tag{120}$$

In such a case exact sampling from the measure $\hat{P}$ (2) corresponds to the information theoretic optimal way of reconstructing the signal. This means that the predictions obtained in this case represent the best possible reconstruction performances *regardless of the algorithm used.*

The replica symmetric computation presented in the previous section becomes exact in this case, for reasons similar to those known in mean field spin glasses on the 'Nishimori line' [44]–[46], [49]. Hence in this Bayes-optimal case the above replica calculation can be used to study the information theoretic limits for reconstruction in CS. This is equivalent to what was rigorously established by [18, 19].

The density evolution and the free entropy can be simplified greatly in the Bayes-optimal case, since the Nishimori condition (15) gives the following equalities:

$$q = m, \qquad Q = \rho \overline{s^2}, \qquad E = V. \tag{121}$$

Hence in the Bayes-optimal case the density evolution is characterized by a single parameter, the mean-squared error $E = \rho \overline{s^2} - m$. Note that the mean-squared distance from the signal to a configuration sampled from the distribution $\hat{P}$ is $D = E + V = 2E$. The density evolution equations (87)–(88) or (114)–(117) reduce to:

$$E^{t+1} = \rho\overline{s^2} - \rho \int \mathrm{d}s\, s\, \phi(s) \int \mathcal{D}z\, f_a\left(\frac{\Delta + E^t}{\alpha}, s + z\frac{\sqrt{\Delta + E^t}}{\sqrt{\alpha}}\right). \tag{122}$$

(Note that the function f_a is defined in (33)). The initial condition of equation (89) is $E^{t=0} = \rho\overline{s^2} - \rho^2\bar{s}^2$.

The free entropy also becomes a function of the single variable E:

$$\begin{aligned}\Phi_{\mathrm{NL}}(E) = {} & -\frac{\alpha}{2} - \frac{\alpha}{2}\log(\Delta + E) - \frac{\alpha(\rho\overline{s^2} - E)}{2(\Delta + E)} \\ & + \int \mathrm{d}s\,[(1-\rho)\delta(s) + \rho\phi(s)] \int \mathcal{D}z \log \\ & \times \left\{\int \mathrm{d}x\, \mathrm{e}^{\alpha/(\Delta+E)x(s-(x/2))+zx(\sqrt{\alpha}/\sqrt{\Delta+E})}[(1-\rho)\delta(x) + \rho\phi(x)]\right\}.\end{aligned} \tag{123}$$

When the signal distribution is known, the value of the MSE E at the global maximum of this free entropy provides the Bayes-optimal reconstruction of the signal, i.e. the lowest achievable MSE given the knowledge of the measurement vector $\mathbf{y}$ and the measurement matrix $\mathbf{F}$. As we will see, depending on parameters α, ρ and $\phi(x)$, the BP algorithm where the MSE evolves according to (122) will either find this global maximum or it will get blocked in a local suboptimal maximum.

For completeness let us give the explicit form of the free entropy (123) for a Gauss–Bernoulli signal where ϕ_0 has zero mean and unit variance:

$$\begin{aligned}\Phi_{\mathrm{NL}}(E) = {} & -\frac{\alpha}{2}\left[\log(\Delta + E) + \frac{\Delta}{\Delta + E}\right] + (1-\rho)\frac{\alpha}{2(\alpha + \Delta + E)} \\ & + (1-\rho)\int \mathcal{D}z \log\left[(1-\rho)\mathrm{e}^{-z^2\alpha/2(\alpha+\Delta+E)} + \frac{\rho\sqrt{\Delta + E}}{\sqrt{\Delta + E + \alpha}}\right] \\ & + \rho\int \mathcal{D}z \log\left[(1-\rho)\mathrm{e}^{-z^2\alpha/2(\Delta+E)} + \frac{\rho\sqrt{\Delta + E}}{\sqrt{\Delta + E + \alpha}}\right].\end{aligned} \tag{124}$$

In this case, the condition of stationarity of the free entropy, giving also the fixed-point condition of density evolution, takes the simple form:

$$E = \rho - \frac{\rho^2}{\alpha + \Delta + E}\int \mathcal{D}z \frac{z^2}{\rho + (1-\rho)(\sqrt{\alpha + \Delta + E}/\sqrt{\Delta + E})\mathrm{e}^{-(z^2/2)(\alpha/\Delta+E)}}. \tag{125}$$

4.4. Density evolution with parameter learning

We study here the general case where the signal is created using a density ρ_0 and empirical distribution of the non-zero elements ϕ_0, and the belief propagation reconstruction algorithm is used with a different signal model, with density $\rho \neq \rho_0$ and distribution of the non-zero elements $\phi \neq \phi_0$. In this case, expectation maximization can be used to learn the parameters, as described in section 3.4. This modified BP procedure, including parameter learning, can also be studied with density evolution. We describe here the case that we use in our implementation, namely a model signal which is Gauss–Bernoulli, where

ϕ is Gaussian with mean $\bar{x}$ and variance σ^2. The learning conditions (74)–(76)) give the evolution of the parameters:

$$
\begin{aligned}
\rho^{(t+1)} = \rho^{(t)} \biggl\{ \int & \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \\
& \times \int \mathcal{D}z (g(\Sigma^2, s+zU)/(1-\rho^{(t)} + \rho^{(t)} g(\Sigma^2, s+zU))) \biggr\} \\
& \times \biggl\{ \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \\
& \times \int \mathcal{D}z (1/(1-\rho^{(t)} + \rho^{(t)} g(\Sigma^2, s+zU))) \biggr\}^{-1},
\end{aligned}
\tag{126}
$$

$$
\begin{aligned}
\bar{x}^{(t+1)} = & \biggl\{ \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z f_a(\Sigma^2, s+zU) \biggr\} \\
& \times \biggl\{ \rho^{(t)} \int \mathrm{d}s [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \\
& \times \int \mathcal{D}z (g(\Sigma^2, s+zU)/(1-\rho^{(t)} + \rho^{(t)} g(\Sigma^2, s+zU))) \biggr\}^{-1},
\end{aligned}
\tag{127}
$$

$$
\begin{aligned}
(\sigma^2)^{(t+1)} = & \biggl\{ \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z [f_a^2(\Sigma^2, s+zU) + f_c(\Sigma^2, s+zU)] \biggr\} \\
& \times \biggl\{ \rho^{(t)} \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \\
& \times \int \mathcal{D}z (g(\Sigma^2, s+zU)/(1-\rho^{(t)} + \rho^{(t)} g(\Sigma^2, s+zU))) \biggr\}^{-1} - [\bar{x}^{(t+1)}]^2,
\end{aligned}
\tag{128}
$$

where the function g is defined as

$$
g(\Sigma^2, R) = \frac{\Sigma}{\sqrt{\Sigma^2 + \sigma^2}} \mathrm{e}^{(R/\Sigma^2 + \bar{x}/\sigma^2)^2/(2(1/\Sigma^2 + 1/\sigma^2)) - \bar{x}^2/2\sigma^2}.
\tag{129}
$$

And we use

$$
\Sigma^2 = \frac{\Delta + V^t}{\alpha}, \qquad U \equiv \sqrt{\frac{\Delta_0 + E^t}{\alpha}}.
\tag{130}
$$

The density evolution for the simplified learning (78) and (79) reads

$$
\bar{x}^{(t+1)} = \frac{1}{\rho^{(t)}} \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z\, f_a(\Sigma^2, s+zU),
\tag{131}
$$

$$
\begin{aligned}
(\sigma^2)^{(t+1)} = \frac{1}{\rho^{(t)}} \int & \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \\
& \times \int \mathcal{D}z\, [f_a^2(\Sigma^2, s+zU) + f_c(\Sigma^2, s+zU)] - [\bar{x}^{(t+1)}]^2.
\end{aligned}
\tag{132}
$$

The density evolution equations now provide a mapping

$$
(E^{(t+1)}, V^{(t+1)}, \rho^{(t+1)}, \bar{x}^{(t+1)}, \sigma^{(t+1)}) = f(E^{(t)}, V^{(t)}, \rho^{(t)}, \bar{x}^{(t)}, \sigma^{(t)})
\tag{133}
$$

obtained by complementing the previous equations on V, and E (87)–(88) with the learning update equations (126)–(128). In our implementation we initialize $\rho^{t=0} = \alpha/10$, $\overline{x}^{t=0} = 0$, and $\sigma^2_{t=0} = 1$.

When a measurement noise is present the variance of the noise can be learned using equation (77), which in the density evolution becomes

$$\Delta^{(t)} = \frac{\Delta_0 + E^t}{1 + V/\Delta^{(t)}}. \tag{134}$$

4.5. Density evolution for block matrices

In the case of the block measurement matrices defined in section 1.2, one can easily generalize the above derivation of the density evolution and of the replica analysis. We just give the results here. For details of the derivation see appendix A.

The order parameters are now

$$Q_p \equiv \frac{1}{N_p}\sum_{i\in B_p}\langle x_i^2\rangle, \qquad q_p \equiv \frac{1}{N_p}\sum_{i\in B_p}\langle x_i\rangle^2, \qquad m_p \equiv \frac{1}{N_p}\sum_{i\in B_p} s_i\langle x_i\rangle \tag{135}$$

in each block $p \in \{1,\ldots,L_c\}$. The free entropy analogous to that in equation (112) becomes

$$\begin{aligned}
&\Phi(\{Q_p\}_{p=1}^{L_c},\{q_p\}_{p=1}^{L_c},\{m_p\}_{p=1}^{L_c},\{\hat{Q}_p\}_{p=1}^{L_c},\{\hat{q}_p\}_{p=1}^{L_c},\{\hat{m}_p\}_{p=1}^{L_c})\\
&\quad = -\frac{1}{2}\sum_{q=1}^{L_r} n_1\alpha_{q1}\left[\frac{\tilde{q}_q - 2\tilde{m}_q + \tilde{\rho}_q + \Delta_0}{\tilde{Q}_q - \tilde{q}_q + \Delta} + \log(\Delta + \tilde{Q}_q - \tilde{q}_q)\right]\\
&\qquad + \sum_{p=1}^{L_c} n_p\left(\frac{Q_p\hat{Q}_p}{2} - m_p\hat{m}_p + \frac{q_p\hat{q}_p}{2}\right)\\
&\qquad + \sum_{p=1}^{L_c} n_p\int \mathrm{d}s\,[(1-\rho_0)\delta(s) + \rho_0\phi_0(s)]\\
&\qquad \times \int \mathcal{D}z\log\left\{\int \mathrm{d}x\,\mathrm{e}^{-((\hat{Q}_p+\hat{q}_p)/2)x^2 + x(\hat{m}_p s + z\sqrt{\hat{q}_p})}[(1-\rho)\delta(x) + \rho\phi(x)]\right\},
\end{aligned} \tag{136}$$

where we introduced

$$\begin{aligned}
\tilde{\rho}_q &= \rho_0\overline{s^2}\sum_{p=1}^{L_c} J_{qp}n_p, \qquad &\tilde{m}_q &= \sum_{p=1}^{L_c} J_{qp}n_p m_p,\\
\tilde{q}_q &= \sum_{p=1}^{L_c} J_{qp}n_p q_p, \qquad &\tilde{Q}_q &= \sum_{p=1}^{L_c} J_{qp}n_p Q_p.
\end{aligned} \tag{137}$$

The equations corresponding to the stationarity condition for this free entropy read:

$$\hat{q}_p = n_p\sum_{q=1}^{L_r}\frac{\alpha_{qp}J_{qp}(\tilde{q}_q - 2\tilde{m}_q + \tilde{\rho}_q + \Delta_0)}{(\tilde{Q}_q - \tilde{q}_q + \Delta)^2}, \tag{138}$$

$$\hat{m}_p = n_p\sum_{q=1}^{L_r}\frac{\alpha_{qp}J_{qp}}{\tilde{Q}_q - \tilde{q}_q + \Delta}, \tag{139}$$

$$\hat{Q}_p = \hat{m}_p - \hat{q}_p, \tag{140}$$

$$m_p = \rho_0 \int \mathrm{d}s\, s\, \phi_0(s) \int \mathcal{D}z f_a\left(\frac{1}{\hat{m}_p}, s + z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right), \tag{141}$$

$$Q_p - q_p = \int \mathrm{d}s[(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z\, f_c\left(\frac{1}{\hat{m}}, s + z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right), \tag{142}$$

$$q_p = \int \mathrm{d}s[(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z\, f_a^2\left(\frac{1}{\hat{m}_p}, s + z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right). \tag{143}$$

When interpreted as a mapping (given the order parameters Q_p, q_p, m_p at time t, one computes $\hat{Q}_p$, $\hat{q}_p$, $\hat{m}_p$ form (138)–(140), and then finds the new order parameters Q_p, q_p, m_p at time $t{+}1$ using (141)–(143)), these equations are exactly the density evolution equations for the case of block matrices. These equations can be written in term of only $2L_c$ order parameters, the mean-squared error $E_p = q_p - 2m_p + \rho_0\overline{s^2}$ and the variance $V_p = Q_p - q_p$ in each block $p \in \{1,\ldots,L_c\}$. The explicit form of the density evolution equations in terms of these $2L_c$ order parameters is:

$$E_p^{(t+1)} = \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z \left[f_a\left(\frac{1}{\hat{m}_p}, s + z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right) - s\right]^2, \tag{144}$$

$$V_p^{(t+1)} = \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \int \mathcal{D}z f_c\left(\frac{1}{\hat{m}_p}, s + z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right), \tag{145}$$

where:

$$\hat{m}_p = n_p \sum_{q=1}^{L_r} \frac{\alpha_{qp} J_{qp}}{\Delta + \sum_{r=1}^{L_c} J_{qr} n_r V_r^{(t)}}, \tag{146}$$

$$\hat{q}_p = n_p \sum_{q=1}^{L_r} \left\{ \frac{\alpha_{qp} J_{qp}}{\left[\Delta + \sum_{r=1}^{L_c} J_{qr} n_r V_r^{(t)}\right]^2} \left[\Delta_0 + \sum_{s=1}^{L_c} J_{qs} n_s E_s^{(t)}\right] \right\}. \tag{147}$$

If one uses block measurement matrices together with expectation maximization learning of the parameters, for a Gauss–Bernoulli signal model, the density evolution equations for the parameters are:

$$\rho^{(t+1)} = \rho^{(t)} \Bigg(\frac{1}{L_c} \sum_{p=1}^{L_c} \int \mathcal{D}z \int \mathrm{d}s[(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \\ \times \frac{g(1/\hat{m}_p, s + z(\sqrt{\hat{q}_p}/\hat{m}_p))}{1 - \rho + \rho g(1/\hat{m}_p, s + z(\sqrt{\hat{q}_p}/\hat{m}_p))} \Bigg) \tag{148}$$

$$\times \left(\frac{1}{L_c} \sum_{p=1}^{L_c} \int \mathcal{D}z \int \mathrm{d}s\, [(1-\rho_0)\delta(s) + \rho_0\phi_0(s)] \\ \times \frac{1}{1 - \rho + \rho g(1/\hat{m}_p, s + z(\sqrt{\hat{q}_p}/\hat{m}_p))} \right)^{-1}, \tag{149}$$

$$\begin{aligned}\bar{x}^{(t+1)} = \frac{1}{\rho}\Bigg(&\frac{1}{L_c}\sum_{p=1}^{L_c}\int \mathcal{D}z\int \mathrm{d}s\,[(1-\rho_0)\delta(s)+\rho_0\phi_0(s)]\\&\times f_a\left(\frac{1}{\hat{m}_p}, s+z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right)\Bigg)\\&\times\left(\frac{1}{L_c}\sum_{p=1}^{L_c}\int \mathcal{D}z\int \mathrm{d}s\,[(1-\rho_0)\delta(s)+\rho_0\phi_0(s)]\right.\\&\left.\times\frac{g(1/\hat{m}_p, s+z(\sqrt{\hat{q}_p}/\hat{m}_p))}{1-\rho+\rho g(1/\hat{m}_p, s+z(\sqrt{\hat{q}_p}/\hat{m}_p))}\right)^{-1},\end{aligned} \tag{150}$$

$$\begin{aligned}(\sigma^2)^{(t+1)} = \frac{1}{\rho}\Bigg(&\frac{1}{L_c}\sum_{p=1}^{L_c}\int \mathcal{D}z\int \mathrm{d}s\,[(1-\rho_0)\delta(s)+\rho_0\phi_0(s)]\\&\times\left[f_a^2\left(\frac{1}{\hat{m}_p}, s+z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right)+f_c\left(\frac{1}{\hat{m}_p}, s+z\frac{\sqrt{\hat{q}_p}}{\hat{m}_p}\right)\right]\Bigg)\\&\times\left(\frac{1}{L_c}\sum_{p=1}^{L_c}\int \mathcal{D}z\int \mathrm{d}s\,[(1-\rho_0)\delta(s)+\rho_0\phi_0(s)]\right.\\&\left.\times\frac{g(1/\hat{m}_p, s+z(\sqrt{\hat{q}_p}/\hat{m}_p))}{1-\rho+\rho g(1/\hat{m}_p, s+z(\sqrt{\hat{q}_p}/\hat{m}_p))}\right)^{-1}-[\bar{x}^{(t+1)}]^2.\end{aligned} \tag{151}$$

As in the homogeneous case, the density evolution equation of the block measurement matrices simplify in the optimal Bayesian approach, when the correct distribution of the signal and its density are known $\rho_0 = \rho$, $\phi_0 = \phi$. In this case, the Nishimori conditions $m_p = q_p$ and $Q_p = \rho\overline{s^2}$ hold, hence $E_p = V_p$ holds for every block $p = 1, \ldots, L_c$. This leads to a single set of closed density evolution equations for the vector E_p, $p = 1, \ldots, L_c$, that reads

$$E_p^{(t+1)} = \int \mathrm{d}s\ [(1-\rho)\delta(s)+\rho\phi(s)]\int \mathcal{D}z\left[f_a\left(\frac{1}{\hat{m}_p}, s+z\frac{1}{\sqrt{\hat{m}_p}}\right)-s\right]^2, \tag{152}$$

$$\hat{m}_p = \sum_{q=1}^{L_r} n_p \frac{\alpha_{qp}J_{qp}}{\Delta+\sum_{r=1}^{L_c}J_{qr}n_r E_r^{(t)}}. \tag{153}$$

In the case where ϕ_0 is a centered Gaussian with unit variance, we get explicitly:

$$E_p^{(t+1)} = \rho - \frac{\rho^2\hat{m}_p}{\hat{m}_p+1}\int \mathcal{D}z\frac{z^2}{\rho+(1-\rho)\mathrm{e}^{-z^2\hat{m}_p/2}\sqrt{\hat{m}_p+1}}. \tag{154}$$

5. The phase diagrams

In this section we turn the equations from the previous section into phase diagrams to display the performance of belief propagation in CS reconstruction. We first discuss the noiseless case, with random homogeneous measurement matrices; this is a benchmark case that has been widely used to demonstrate the power of the ℓ_1 reconstruction. We use

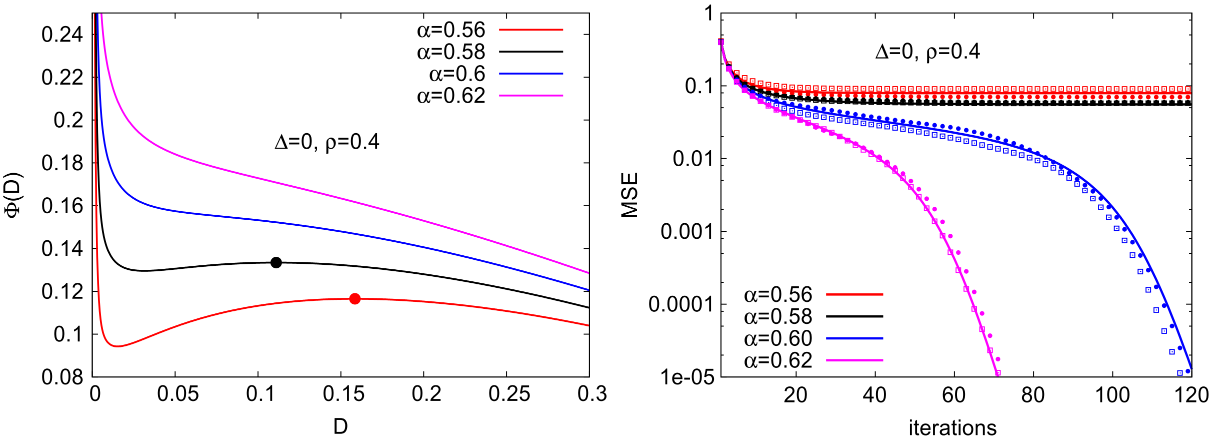

Figure 2. Left: the free entropy, $\Phi(D)$, is plotted as a function of $D = \langle \sum_i (x_i - s_i)^2/N \rangle$ for $\rho_0 = 0.4$ and several measurement rates α in the Bayesian approach (when both the signal and the signal model are described by a Gauss–Bernoulli distribution). The evolution of the BP algorithm is basically a steepest ascent in $\Phi(D)$, starting from a large value of D. Such ascent goes to the global maximum at $D = 0$ for large value of α, but is blocked in the local maximum that appears for $\alpha < \alpha_{\rm BP}(\rho_0 = 0.4) \approx 0.59$. For $\alpha < \rho_0$, the global maximum is not at $D = 0$ and exact inference is impossible. Right: using the same conditions as for the left figure, we show the evolution of the MSE measured experimentally during the iterations of BP for a signal of size $N = 15\,000$ (data points) compared to the theory using density evolution (line). For the two lower measurement rates, where $\alpha < 0.59$, the MSE saturates at a finite value. For the two higher ones it goes to zero. The full circles are for measurement matrices with iid Gaussian elements, the empty squares for matrices with iid ± 1 elements. We see small finite size corrections, but otherwise there is excellent agreement between the two cases, as expected from the theory, which states that only the mean and variance of the distribution of each matrix element matters.

measurement matrices with iid entries with zero mean and variance $1/N$ (note that our approach is independent of the distribution of the iid matrix elements and depends only on their mean and variance). Finally, we discuss the phase diagram for noisy measurements, which present several interesting features.

5.1. Noiseless measurements and the optimal Bayes case

In figure 2 we show the free entropy density at fixed squared distance, $\Phi(D)$, for the Bayes-optimal case in which both ϕ_0 and ϕ are Gaussian with zero mean and unit variance. The elements of the $M \times N$ measurement matrix $\mathbf{F}$ are independent random variables with zero mean and variance $1/N$.

The free entropy $\Phi(D)$ is computed using equations (112) and (119), which was derived using the replica method. The dynamics of the message passing algorithm (without learning) is a gradient dynamics leading to a maximum of the free entropy $\Phi(D)$, starting from high distance D. As expected, we see in figure 2 that $\Phi(D)$ has a global maximum at $D = 0$ if and only if $\alpha > \rho_0$, which confirms that the Bayesian optimal inference is in principle able to reach the theoretical limit $\alpha = \rho_0$ for exact reconstruction. The left-

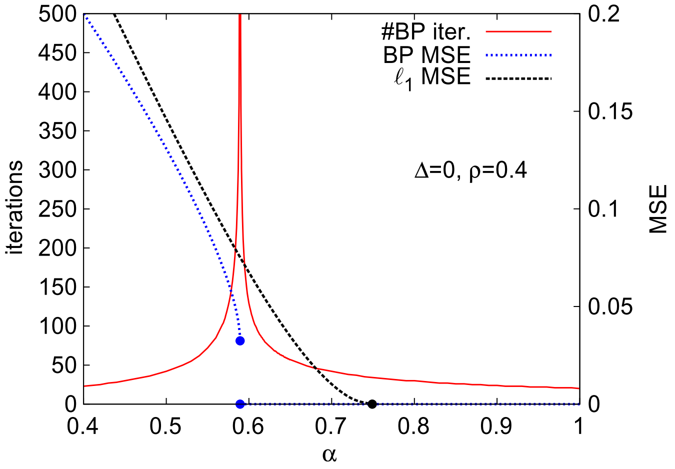

Figure 3. The full (red) line (left y-axis) is the convergence time of the BP algorithm, defined as the number of iterations needed such that the MSE obtained by the algorithm at a given iteration does not change more than by 10^{-7} in the next iteration. The data are obtained with the density evolution for a signal with density $\rho_0 = 0.4$, where the non-zero elements of the signal are Gaussian with zero mean and unit variance. Reconstruction is done in the Bayes-optimal case. The BP convergence time diverges as $\alpha \to \alpha_{\mathrm{BP}}$. The dotted lines (right y-axis) give the mean-squared error achieved by the BP algorithm (blue) and by the ℓ_1 minimization (black) for reconstruction of the same signal. Exact reconstruction is in principle possible in the whole region $\alpha > \rho_0$. The reconstruction with BP is exact for $\alpha > \alpha_{\mathrm{BP}}(\rho_0 = 0.4) \approx 0.59$, whereas the ℓ_1-reconstruction is exact only for $\alpha \gtrsim 0.75$. Note also in the regime $\alpha < \alpha_{\mathrm{BP}}$, where BP does not reconstruct exactly the signal, the MSE achieved by BP is always smaller than that of ℓ_1.

hand side of the figure shows the existence of a critical measurement rate $\alpha_{\mathrm{BP}}(\rho_0) > \rho_0$, below which a secondary local maximum of $\Phi(D)$ appears at $D > 0$. When this secondary maximum exists, the BP algorithm converges instead to it, and does not reach exact reconstruction. The threshold $\alpha_{\mathrm{BP}}(\rho_0)$ is obtained analytically as the smallest value of α such that $\Phi(D)$ is monotonic. The behavior of $\Phi(D)$ is typical of a first-order transition. The equilibrium transition appears at a number of measurements per unknown $\alpha = \rho_0$, which is the point where the global maximum of $\Phi(D)$ switches discontinuously from being at $D = 0$ (when $\alpha > \rho_0$) to a value $D > 0$. In this sense the value $\alpha = \alpha_{\mathrm{BP}}(\rho_0)$ appears like a spinodal point: it is the point below which the global maximum of $\Phi(D)$ is no longer reached by the dynamics. Instead, in the regime below the spinodal ($\alpha < \alpha_{\mathrm{BP}}(\rho_0)$), the dynamical evolution is attracted to a metastable non-optimal state with $D > 0$.

On the right-hand side of figure 2, we show the evolution of the MSE as predicted by the density evolution equations, as well as the MSE measured using the BP algorithm for a system with size $N = 15\,000$. Below the spinodal point $\alpha_{\mathrm{BP}}(\rho_0)$ the MSE does not converge to zero, because the system is trapped in a metastable state.

The spinodal transition is the physical reason that limits the performance of the BP algorithm. To illustrate this statement, we plot in figure 3 the BP convergence time as a function of the measurement rate α. As expected, the convergence time diverges around the spinodal transition α_{BP}. In the same figure 3 we also plot the MSE achieved by

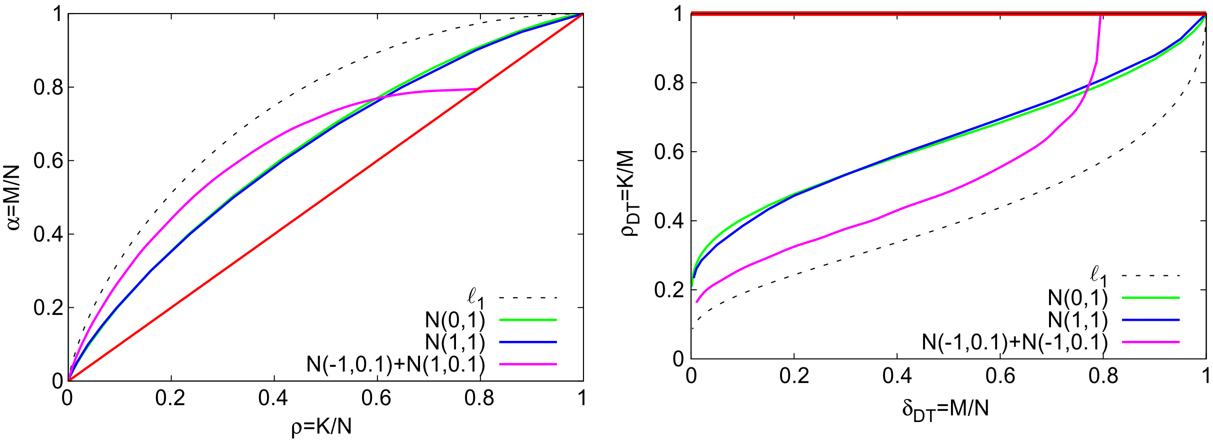

Figure 4. Phase diagram for the BP reconstruction in the optimal Bayesian case when the signal model is matching the empirical distribution of signal elements, i.e. $\phi(x) = \phi_0(x)$. The elements of the $M \times N$ measurement matrix $\mathbf{F}$ are iid variables with zero mean and variance $1/N$. The spinodal transition $\alpha_{\mathrm{BP}}(\rho_0)$ is computed with the asymptotic replica analysis and plotted for the following signal distributions: $\phi(x) = \mathcal{N}(0,1)$ (green), $\phi(x) = \mathcal{N}(1,1)$ (blue) $\phi(x) = [\mathcal{N}(-1,0.1) + \mathcal{N}(1,0.1)]/2$ (magenta, equations needed to obtain this curve are summarized in appendix C). Note that for some signals, e.g. the third case, there is a region of signal densities (here $\rho_0 \gtrsim 0.8$) for which the BP reconstruction is possible down to the optimal subsampling rates $\alpha = \rho_0$. The data are compared to the Donoho–Tanner phase transition $\alpha_{\ell_1}(\rho_0)$ (dashed) for ℓ_1 reconstruction that does not depend on the signal distribution, and to the theoretical limit for exact reconstruction $\alpha = \rho_0$ (red). The left-hand side represents the undersampling rate α as a function of the signal density ρ_0. The right-hand side shows the same data in the Donoho–Tanner notation, i.e. the number of non-zero elements in the signal per measurement is plotted as a function for the undersampling rate.

the BP reconstruction algorithm compared to the MSE achieved by the ℓ_1 minimization reconstruction for the same signal and the same measurement matrix as before. Note that here we are still in the favorable case when the signal model was equal to the signal distribution $\rho = \rho_0$, $\phi(x) = \phi_0(x)$.

Note that the ℓ_1 transition at α_{ℓ_1} is continuous (second order), whereas the spinodal transition is discontinuous (first order). The transition at α_{BP} is called a spinodal transition in the mean field theory of first-order phase transitions. It is similar to the one found in the cooling of liquids which go into a supercooled glassy state instead of crystallizing, and appears in the decoding of error correcting codes [49, 51] as well. This difference might seem formal, but it is absolutely essential as far as concerns the possibility of achieving the theoretically optimal reconstruction with the use of seeding measurement matrices (as discussed in section 5.2).

In figure 4 we show how the critical value α_{BP} depends on the signal density ρ and on the type of the signal, for several Gauss–Bernoulli signals. In this figure we still assume that the signal distribution is known, and hence $\rho_0 = \rho$ and $\phi_0 = \phi$. We compare it to the Donoho–Tanner phase transition α_{ℓ_1} that gives the limit for exact reconstruction with the

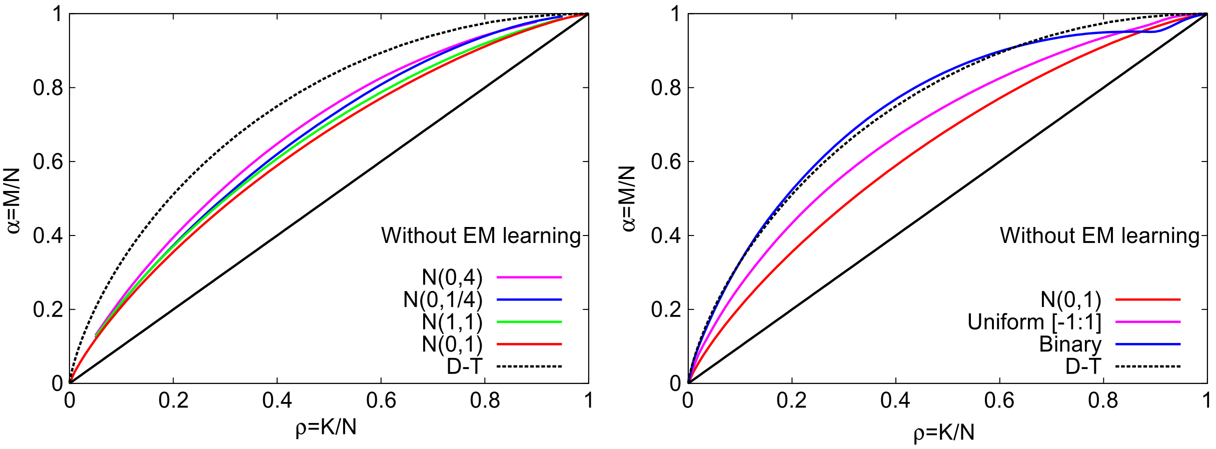

Figure 5. Phase diagram for the reconstruction with BP when the signal model does not match the empirical distribution of signal elements. The signal model is Gauss–Bernoulli with zero mean and unit variance. The measurement matrix is a homogeneous one with Gaussian iid entries. In this plot we assume that the signal density is known $\rho = \rho_0$. Different curves correspond to different distributions ϕ_0 of the signal. The dashed line gives the Donoho–Tanner transition line for ℓ_1 reconstruction, which is independent of the signal distribution.

ℓ_1 minimization [7, 24, 52], and to the information theoretic limit for exact reconstruction $\alpha = \rho$.

Note that for some signals, e.g. the mixture of Gaussians $\Phi(x) = [\mathcal{N}(-1, 0.1) + \mathcal{N}(1, 0.1)]/2$, there is a region of signal densities (here $\rho_0 \gtrsim 0.8$) for which the BP reconstruction is possible down to the optimal subsampling rates $\alpha = \rho_0$.

5.2. Noiseless measurements and the mismatching signal model

In this section we show the performance of BP reconstruction and the corresponding phase diagrams in the general case when the density of the signal and the distribution of the non-zero signal elements is not known

$$\rho \neq \rho_0, \qquad \phi(x) \neq \phi_0(x). \tag{155}$$

All the results we show are for the Gauss–Bernoulli model of the signal, i.e. $\phi(x) = \mathrm{e}^{-(x_i - \bar{x})^2/(2\sigma^2)}/(\sqrt{2\pi}\sigma)$. As we argued in section 2.1, for noiseless measurements the probabilistic reconstruction for CS is optimal as long as $\alpha > \rho_0$, even if the signal model is not the correct one, as in (155). This property can also be seen by analyzing the replica calculation of the free entropy (112) that close to exact reconstruction ($Q \to \rho_0 \overline{s^2}$, $q \to \rho_0 \overline{s^2}$, $m \to \rho_0 \overline{s^2}$) behaves as $\Phi \to -(\alpha - \rho_0) \log(Q - q)/2$. Unfortunately, in general, BP encounters a spinodal line (barrier) as in the case discussed in section 5.1. The position of this line (phase transition) depends on both the signal model $\phi(x)$ and the signal distribution $\phi_0(x)$.

In figure 5 we show the phase diagram for Gauss–Bernoulli signal model, i.e. the distribution of components being

$$P(x) = (1 - \rho)\delta(x) + \rho \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \tag{156}$$

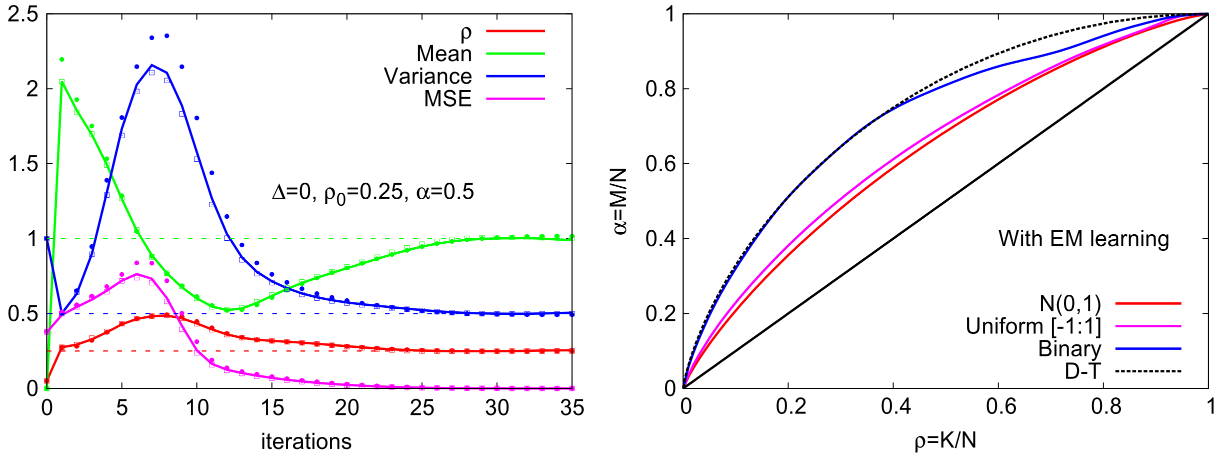

Figure 6. Left: learning of parameters for noiseless measurements. The signal is Gauss–Bernoulli with density $\rho_0 = 0.25$, mean $\bar{s} = 1$ and variance $\overline{s^2} - \bar{s}^2 = 0.5$. The measurement density is $\alpha = 0.5$. The EM-BP algorithm is initialized with $\rho = 0.05$, $\bar{x} = 0$, $\sigma^2 = 1$. In the figure we plot the evolution of the parameters and of the mean-squared error E. The full line is the analytic prediction using density evolution, the data points are for the EM-BP algorithm on an instance of $N = 12\,000$, the full points are for a measurement matrix with Gaussian elements, the empty points for a matrix with elements $\pm 1/N$. Right: phase diagram for the EM-BP reconstruction, that is when the signal model does not match the empirical distribution of signal elements, i.e. $\phi(x) \neq \phi_0(x)$. Different curves correspond to different distributions ϕ_0 of the signal. The dashed line gives the Donoho–Tanner transition line for ℓ_1 reconstruction, which is independent of the signal distribution.

and various signal components distributions $P_0(x) = (1-\rho_0)\delta(x) + \rho_0\phi_0(x)$. Here we assume $\rho = \rho_0$. We see that the performance of BP mostly slightly decreases. For some signal distributions (e.g. the binary case $\phi_0(x) = [\delta(x-1) + \delta(x+1)]/2$) there is a narrow region of parameters in which the ℓ_1-reconstruction becomes better than the probabilistic-BP approach.

In the case where the signal distribution and its sparsity are not known, the performance of BP can be improved by including the expectation maximization learning. We call this generalization EM-BP. In this paper we study the performance of EM-BP in the case where the signal model is Gauss–Bernoulli

$$P(x) = (1-\rho)\delta(x) + \rho \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\bar{x})^2/2\sigma^2}. \tag{157}$$

Expectation maximization is used to learn the three parameters $\rho, \bar{x}$ and σ. In EM-BP we do one update of BP messages followed by one update of the parameters. New values of parameters are computed using equations (74), (78) and (79). BP messages are then updated again using parameter values $\rho = [\rho_{\text{old}} + \min(\rho_{\text{new}}, \alpha)]/2$, $\bar{x} = (\bar{x}_{\text{old}} + \bar{x}_{\text{new}})/2$, $\sigma^2 = [\sigma^2_{\text{old}} + \max(\sigma^2_{\text{new}}, 0)]/2$. And this is repeated till convergence. The evolution of parameters under learning is illustrated in the left part of figure 6.

We observe that for the Gaussian-distributed signal elements (left part of figure 5) the correct mean and variance are always learned (even in the region where exact

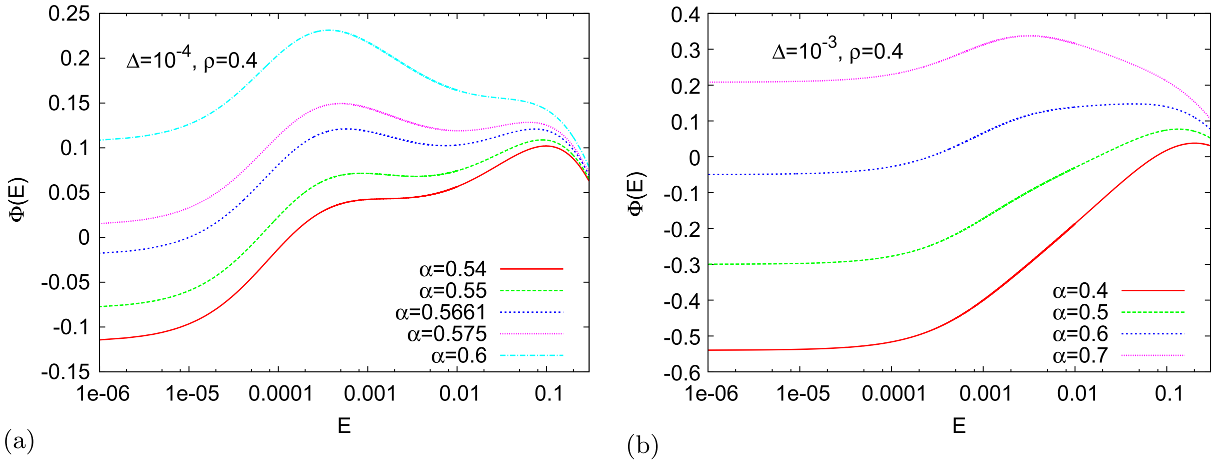

Figure 7. The free entropy $\Phi(E)$ in the presence of noise as a function of the MSE. (a) $\rho = 0.4$ and $\Delta = 10^{-4}$, there is a first-order phase transition and two local maxima do co-exist for region of subsampling rates $\alpha_d > \alpha > \alpha_s$. (b) For larger noise, $\Delta = 10^{-3}$, there is always only one maxima, in this case the EM-BP approach is always optimal, although the mean-squared error may be quite large.

reconstruction is not possible). In this case the spinodal line is always the same as in the case when the signal distribution was known, see figure 4. For signals with non-Gaussian distribution of elements, right part of figure 5, the spinodal line changes slightly, the lines with learning are shown in the right part of figure 6. We conclude that EM-BP improves on pure BP and on ℓ_1-reconstruction in many cases and hence it can be useful in practical situations. Of course if one has some knowledge of the signal distribution it is helpful to further include it in the signal model.

5.3. Phase diagram for noisy measurements

In this section we discuss compressed sensing with noisy measurements, $\Delta > 0$. We first describe the performance of the BP algorithm and the corresponding phase diagrams in the Bayes-optimal case when the signal model corresponds to the signal distribution. In a second part we then discuss the general noisy case with non-matching signal model and learning.

In figure 7 we plot the free entropy $\Phi(E)$, obtained from equation (124), as a function of the mean-squared error E, for a signal with non-zero elements being iid Gaussian variables with zero mean and unit variance, and a matching signal model. The main difference from the noiseless case, figure 2, is that the global maximum of the free entropy, that described the optimal achievable mean-squared error, is at non-zero values of the MSE. This indeed reflects the fact that with noisy measurements exact reconstruction is no longer possible.

Let us investigate whether the BP algorithm finds a configuration with the best achievable MSE or not. Again, BP is basically performing steepest ascent in the free entropy, starting from a large value of MSE. Depending on the value of the signal density ρ and the measurement noise variance Δ, we see two kinds of behavior as a function of the subsampling rate α. For some values of ρ, Δ, see figure 7(b), the global maximum of

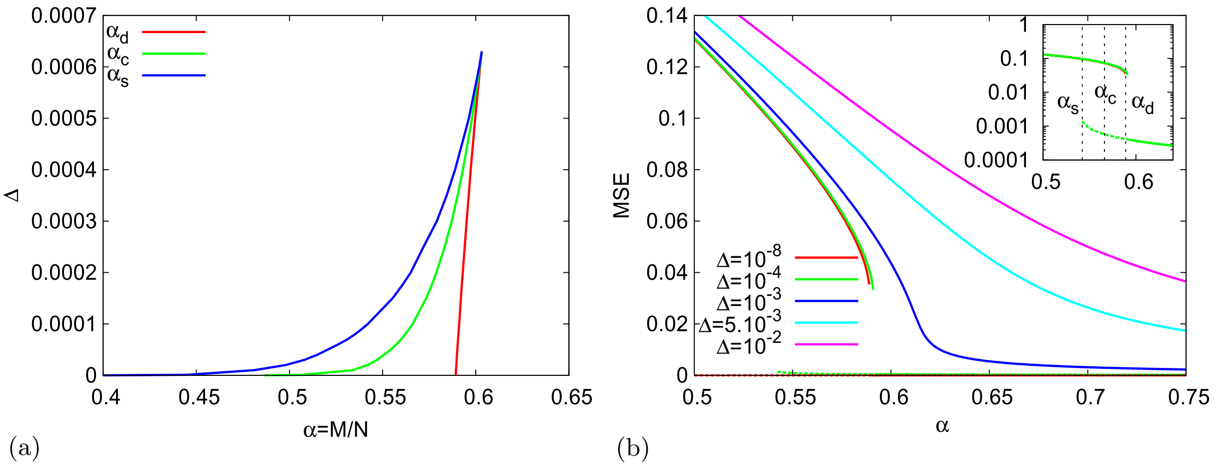

Figure 8. (a) The three phase transition lines in CS with noisy measurements for a Gauss–Bernoulli signal and matching signal model with density $\rho = 0.4$. The blue line is the spinodal line α_s, the red line is the dynamical line α_d, and the green line is the critical line α_c. For larger noise there is no such sharp threshold. A perfect sampling algorithm changes its behavior abruptly at α_c (the green line), where the quality of reconstructed signal would jump discontinuously from high MSE to low MSE. The BP algorithm (with the uninformed initialization) always converges to the local maxima of the free entropy corresponding to the largest MSE, hence its MSE jumps from a relative low value to a high value at α_d (the red line). BP is hence suboptimal for $\alpha_d > \alpha > \alpha_c$. (b) The MSE achieved by BP for several noise strengths. In the inset is the case of $\Delta = 10^{-4}$ with the three phase transitions depicted. For $\alpha_d > \alpha > \alpha_c$ the best achievable MSE corresponds to the lower part of the curve, whereas BP reconstruction achieves the MSE corresponding the upper part of the curve. Note that in this case the MSE achieved by ℓ_1 reconstruction would be much larger (non-zero for $\alpha \gtrsim 0.75$ even for the noiseless case, see figure 3).

$\Phi(E)$ is the only maximum for all α, and in that case BP will converge to it. For other values of ρ, Δ, see figure 7(a), the situation is similar to the noiseless case.

- For $\alpha > \alpha_d$ the free entropy has a single maximum at a small value of MSE comparable to Δ.
- For $\alpha_d > \alpha > \alpha_c$ the free entropy has two maxima, the one at lower MSE being the global one.
- For $\alpha_c > \alpha > \alpha_s$ the free entropy has two maxima, the one at higher MSE being the global one.
- For $\alpha < \alpha_s$ the free entropy has a single maximum at a value of MSE much larger than Δ.

The above result means that for a region of subsampling rates $\alpha_d > \alpha > \alpha_c$ the BP algorithm is suboptimal, as it converges to much higher MSE than the MSE corresponding to the optimal Bayes inference (global maximum of the free entropy). In the left part of figure 8 we plot the dependence of α_d, α_c, and α_s on the noise variance. In the right part we plot the MSE achieved by BP as a function of the subsampling rate. In cases where BP is suboptimal (for the two lowest noise variances) we compare to the optimal MSE. The

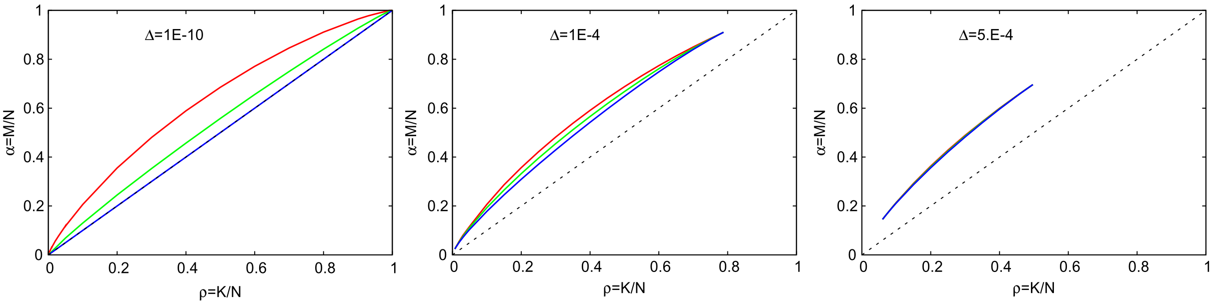

Figure 9. The three transition lines α_d, α_c, α_s shown in figure 8 for different values of the noise variance, growing from left to right: left: $\Delta = 10^{-10}$, middle $\Delta = 10^{-4}$, and right $\Delta = 5 \times 10^{-4}$.

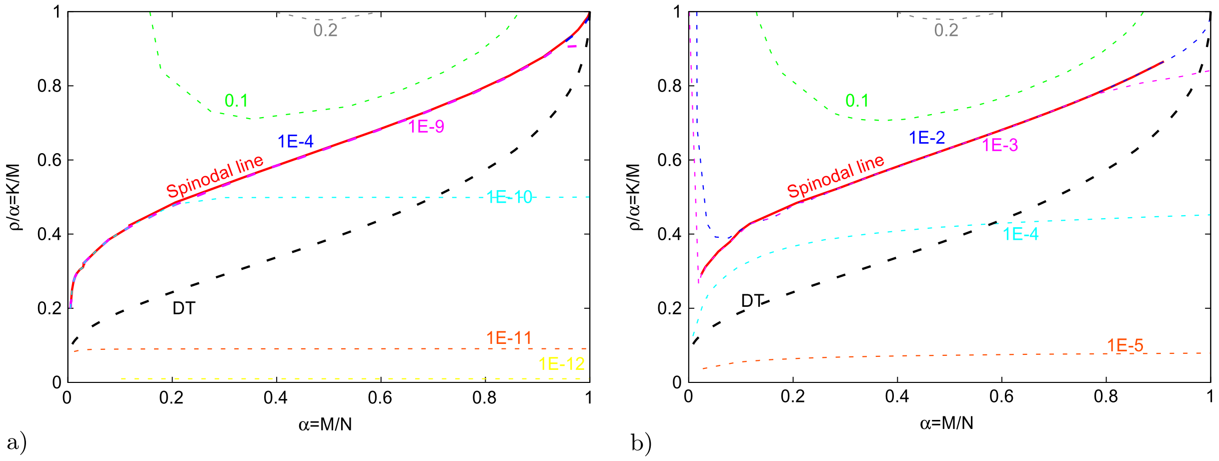

Figure 10. Phase diagram and level lines of the MSE for the BP algorithm in the presence of noise in the Donoho–Tanner convention. Left: using a noise with variance $\Delta = 10^{-10}$. Right: using a noise with variance $\Delta = 10^{-4}$. The Donoho–Tanner transition line for ℓ_1 is shown for comparison.

data presented in figures 7 and 8 are obtained from the density evolution, i.e. $N \to \infty$ limit of BP behavior. The behavior of BP for finite N agrees well with these results for system sizes of several thousands of elements and more.

In figure 9 we plot again the three phase transition lines for reconstruction with measurement noise. This time we plot the lines in the ρ–α phase diagram for several values of the variance Δ. As the noise increases the region of densities for which there is a sharp phase transition shrinks. For large enough values of $\Delta \gtrsim 0.000\,78$ there is no sharp phase transition for the inference of the Gauss–Bernoulli signal (with matching Gauss–Bernoulli signal model).

Another illustration of this phase diagram with noise is in figure 10, where we plot level lines following the MSE achieved by BP reconstruction. On the line α_d, the MSE of BP reconstructions increases discontinuously from values comparable to Δ to large values.

Of course in practical applications the noise level Δ_0 is often not known. In such cases learning of the noise level can be included in the EM-BP algorithm, using the noise

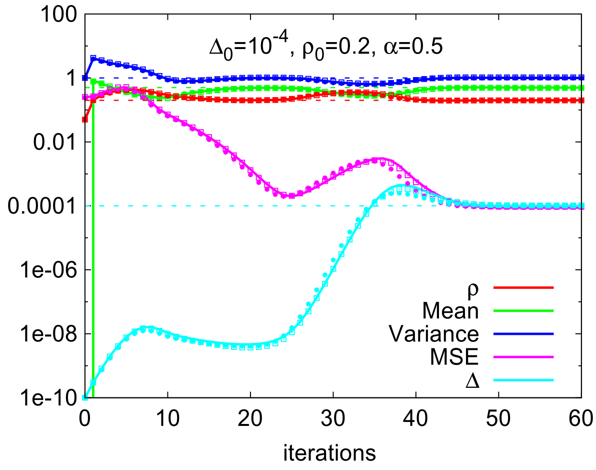

Figure 11. Learning of parameters for noisy measurements. The signal is Gauss–Bernoulli of density $\rho_0 = 0.2$, mean $\bar{s} = 0.5$ and variance $\overline{s^2} - \bar{s}^2 = 1$. The measurement rate is $\alpha = 0.5$ and the noise variance $\Delta_0 = 10^{-4}$. The EM-BP algorithm is initialized with $\rho = 0.05$, $\bar{x} = 0$, $\sigma^2 = 1$, $\Delta = 10^{-10}$. In the figure we plot the evolution of the parameters and of the mean-squared error E for three cases. The full line is the density evolution, the data points is the EM-BP algorithm on an instance of $N = 10\,000$, the full points are for a measurement matrix with Gaussian elements, the empty points for a matrix with elements $\pm 1/N$.

variance update equation (77). In figure 11 we illustrate the evolution of parameters and the mean-squared error E under such expectation maximization learning for a Gaussian signal of density $\rho_0 = 0.2$, with measurement rate $\alpha = 0.5$ and noise variance $\Delta_0 = 10^{-4}$.

6. Seeding matrices: a way to achieve optimality

In section 5.3 we exposed the reason why BP reconstruction for homogeneous measurement matrices $\mathbf{F}$ does not achieve subsampling rates down to the information theoretic limit $\alpha = \rho_0$. In [1] we developed a new type of measurement matrix—that we coined *seeding matrices*—for CS for which the limit $\alpha = \rho_0$ is achievable using the BP reconstruction. This was built on several results in the error correcting code community [38]–[41]. Here we shall explain further our motivations for the construction of the seeding matrices.

We shall give heuristic arguments why with these matrices it is possible to achieve theoretically optimal reconstruction and show, using the replica method (or equivalently, density evolution) that this is indeed the case. We want to point out that, while we use mostly the replica method/density evolution formalism, some rigorous results can be obtained. In particular, in the special Bayes-optimal case—when the signal model corresponds to the empirical distribution of the non-zero signal elements—it has been now proved rigorously in [14] that for the CS with seeding matrices the BP reconstruction is indeed able to achieve the information theoretic limit $\alpha = \rho_0$. Here, we shall show, using statistical physics tools, that seeding matrices allow close to optimal reconstruction also when the signal distribution is not known, which is even more appreciable.

6.1. Why and when does seeding work?

As exposed in the previous section, for homogeneous measurement matrices with iid entries, BP is able to reconstruct the signal correctly at $\alpha > \alpha_{\rm BP}$, below $\alpha_{\rm BP}$ a *metastable state* (i.e. a local maximum of $\Phi(D)$ at $D > 0$) appears in the measure $P(\mathbf{x}|\mathbf{F}, \mathbf{y})$. The iterations of the BP algorithm get 'trapped' in this state and BP is therefore unable to find the global maximum corresponding to the original signal (see figure 3). This is a situation well known in physics, that is typical for a system undergoing a first-order phase transition. A familiar example of a first-order phase transition being crystallization, i.e. the way a liquid changes into a solid. In physics, systems undergoing a first-order phase transition can be divided into two groups: (a) mean field systems, where the size of the boundary of a sphere of a (large) finite radius drawn around one particle (variable) is of the same order as the volume of this sphere. (b) Finite dimensional systems where the size of the boundary is much smaller than its volume. Typically in d dimensions, a sphere of radius r has surface $s_d r^{d-1}$ and volume $v_d r^d$ (s_d and v_d being the surface and volume of a sphere of radius one).

In mean field systems metastable states have exponentially large (in the size of the system) living time, meaning that is would take an exponential time to randomly find a fluctuation that would be able to overcome the barrier between the local maximum and the global one. Whereas in finite dimensional systems the living time of metastable states is always constant. A simplified argument leading to this conclusion uses the fact that maximization of the entropy is the driving force of system dynamics. Consider the system being in the metastable state (e.g. supercooled liquid), if a random fluctuation appears flipping a droplet of radius R into the equilibrium state (crystal) then this causes a free entropy increase of $\Delta\Phi v_d R^d$ and decrease because of the surface terms $\Gamma s_d R^{d-1}$ for R large enough $R > R^* = \Gamma s_d (d-1)/(\Delta\Phi v_d d)$; the gain is more important than the loss and such a randomly created droplet will start to grow. The crucial point is that the critical radius R^* does not depend on the system size N and hence such a fluctuation arises with a constant probability in the finite dimensional systems. The process we described here is on the basis of nucleation theory in physics which described the growth of crystal droplets close to a first-order phase transition [53, 54].

The whole idea of seeding matrices is to mimic the process of nucleation and crystal growth in the reconstruction of compressed sensing signal. This idea, together with the previous work on spatially coupled LDPC codes [40], also motivated the design of the seeding matrix in [1]. There are three key ingredients that need to be present in the system in order for the seeding to work.

(a) The free entropy driving force. To escape from a metastable state we need the existence of a higher maximum of the free entropy $\Phi(D)$. This ingredient is present in the BP reconstruction of the original signal as long as $\alpha > \rho_0$ (or $\alpha > \alpha_c$ for the nosy case). Let us note here that seeding does not improve the performance of the ℓ_1 reconstruction algorithms (see appendix B), because this 'driving force' is missing since the Donoho–Tanner transition is continuous (it is a second-order transition in the physics classification).

(b) The existence of a nucleus (seed). We need a part of the system to be already in the equilibrium state. This ingredient can be achieved by making the measurement matrix inhomogeneous and measuring at a much higher subsampling rate a small subpart of the signal—which we call a ‘seed’.

(c) An interaction between the seed and the rest of the signal that enables the growth of the seed. In [1] and [14] this was achieved via the so-called spatial coupling. The signal was divided into blocks and the measurements designed in such a way that only several neighboring blocks are measured at a time. Similar ideas have been used recently in the design of sparse coding matrices for error correcting codes [38]–[40], [55]. Here we also give an example of a seeded measurement matrix that does not have spatially coupled structure.

In this article we present several ways how to achieve points (b) and (c), and hence be able to do reconstruction in CS at yet lower subsampling rates. We, however, stress that there is relatively great freedom in the construction of these matrices, and their optimization and adaptation to physically constrained measurements is surely a promising area of future research.

The matrices we used are presented in figure 12. These are block matrices defined as follows: the N variables are divided into L_c groups of N_p, $p = 1, \dots, L_c$, variables in each group. We denote $n_p = N_p/N$. And the M measurements are divided into L_r groups of M_q, $q = 1, \dots, L_r$, measurements in each group, we define $\alpha_{qp} = M_q/N_p$. Then the matrix F is composed of $L_r \times L_c$ blocks and the matrix elements $F_{\mu i}$ are generated independently, in such a way that if μ is in group q and i in group p then $F_{\mu i}$ is a random number with zero mean and variance $J_{q,p}/N$. Thus we obtain a $L_r \times L_c$ coupling matrix $J_{q,p}$. For the asymptotic analysis we assume that $N_p \to \infty$, for all $p = 1, \dots, L_c$ and $M_q \to \infty$ for all $q = 1, \dots, L_r$. The total subsampling rate is then $\alpha = \sum_{q=1}^{L_r} M_q / (\sum_{p=1}^{L_c} N_p)$. The case of a homogeneous matrix can easily be recovered by setting $L_c = L_r = 1$. We define $I(\mu)$ or $I(i)$ to be the index of the block to which μ or i belongs, B_q is the set of indices in block q.

In all the examples of seeding matrices used in this article and presented in figure 12, the elements of the signal vector are split into L_c equally sized blocks ($N_p = N/L_c$). The first block of measurements has size M_1 and the other $L_r - 1$ measurement blocks have equal size $M_q = (M - M_1)/(L_r - 1)$ for $q > 1$. In all the examples here we achieve the seeding by taking $\alpha_{\text{seed}} = M_1 L_c/N$ larger than $\alpha_{\text{BP}} > \rho_0$, and $\alpha_{\text{bulk}} = M_q L_c/N$ for $q > 1$ that can be approaching ρ_0. The overall measurement rate is then

$$\alpha = \frac{\alpha_{\text{seed}} + (L_r - 1)\alpha_{\text{bulk}}}{L_c}. \tag{158}$$

Hence $\alpha \to \alpha_{\text{bulk}}$ as $L_c/L_r \to 1$, and $L_r \to \infty$. The matrix elements $F_{\mu i}$ are chosen as random i.i.d variables with variance $J_{q,p}/N$ if variable i is in the block p and measurement μ in the block q.

6.2. Seeding experiments for noiseless measurements

In figure 13 we demonstrate how BP reconstruction works for seeded measurement matrices. We generated signal elements of density $\rho_0 = 0.4$, the non-zero elements are Gaussian random variables with zero mean and unit variance. We obtained $\alpha = 0.5$ noiseless measurements per signal element using seeded matrices generated as described above. We plot the mean-squared error in every block (different lines) as a function of

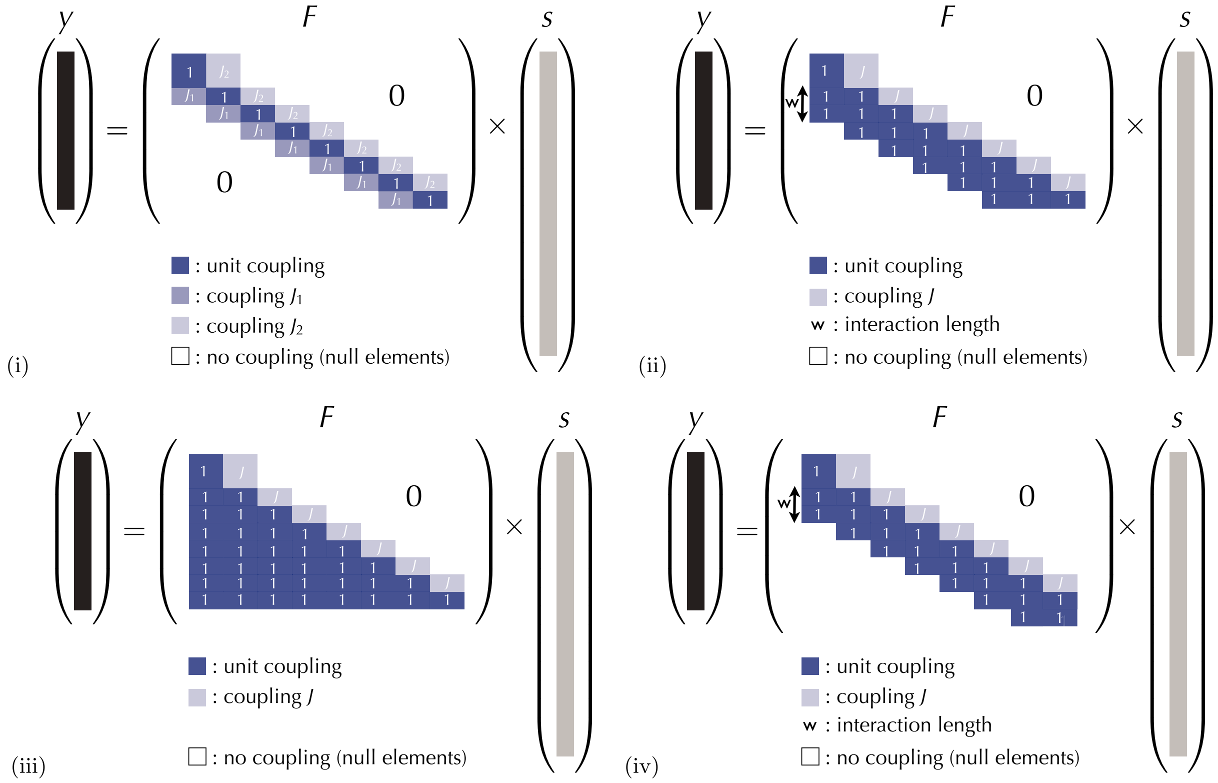

Figure 12. Examples of seeding measurement matrices $\mathbf{F}$ for CS. Here $L_c = 8$. (i) A band-diagonal matrix, already introduced in [1], where $L = L_c = L_r$, and $J_{p,q} = 0$, except for $J_{p,p} = 1$, $J_{p,p-1} = J_1$, and $J_{p-1,p} = J_2$. Good performance is typically obtained with large J_1 and small J_2. (ii) Another band-diagonal matrix where $L = L_c = L_r$, and $J_{p,q} = 0$, except for $J_{p,p} = 1$, $J_{p-1,p} = J$, and $J_{p,p-w} = 1$ with $w = 1, \ldots, W$. Good performance is typically obtained with small J and $W \geq 2$. Since all variances are lower than or equal to one, this matrix can be realized only having elements $(0, \pm 1)$. (iii) A lower triangular matrix, that can be viewed as the matrices of type (ii) when $W = L - 1$. Again, good performance is obtained with relatively small J. (iv) In some cases, we observed that the last block of variables was not recovered correctly. Adding a new line in the matrix $(L_r = L_c + 1)$, as in this example, cures the problem. All these matrices are motived by the same consideration: more measurements are made in the first block of the signal such that the information will first appear in this block, and then propagate into the whole vector.

BP iteration time. We compare a result from BP with its asymptotic density evolution behavior, obtaining excellent agreement. Note that in this case the BP reconstruction for standard homogeneous matrices would fail. Note that in both cases illustrated in figure 13 the first blocks are reconstructed fast, and by interaction with the subsequent blocks the reconstructed region is propagated to the following blocks.

Now that we have illustrated that the BP reconstruction for large systems indeed agrees with the asymptotic density evolution analysis, we plot in figure 14 two examples of the number of iterations (defined as the time when mean-squared error $E < 10^{-7}$) it takes to reconstruct exactly a signal of density ρ_0 with measurement rate $\alpha \to \rho_0$.

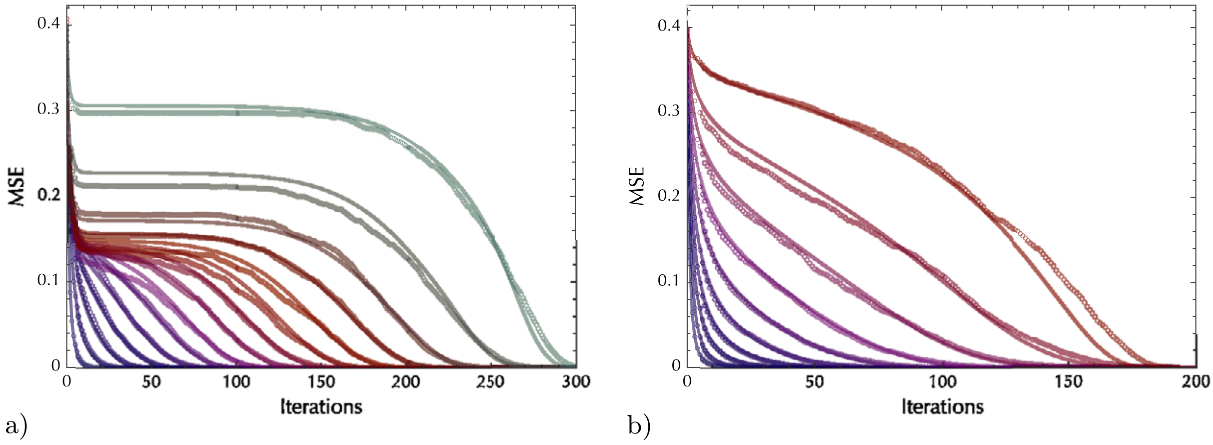

Figure 13. Reconstruction on the signal with seeded measurement matrices. The mean-squared error in every block is plotted as a function of the iteration time. We compare the numerical analysis of BP for a signal of $N = 40\,000$ elements with the analytic result obtained in the $N \to \infty$ limit using density evolution. The agreement is very good. The density of the signal is $\rho_0 = 0.4$, the non-zero elements are Gaussian with zero mean and unit variance. The measurement rate is $\alpha = 0.5$. The two cases are: (a) the seeding matrix of the type (ii) from figure 12 with i.i.d. $0, \pm 1$ random elements, $\alpha_{\text{seed}} = 0.7$, $\alpha_{\text{bulk}} = 0.485$, $L = 15$, $J = 0.01$ and $W = 2$. (b) The seeding matrix of the type (iii) from figure 12 with i.i.d. Gaussian random elements, $\alpha_{\text{seed}} = 0.68$, $\alpha_{\text{bulk}} = 0.48$, $L = 10$ and $J = 0.1$.

In figure 15 we show how the number of iterations needed for exact reconstruction depends on the number of blocks L for different signal densities ρ_0. We see that in case of the one-dimensional seeding matrix of type (ii) the number of iterations depends linearly on the number of blocks. The boundary of the reconstructed region is propagating as a kind of spatially localized wave at a constant speed, as illustrated in figure 16. On the other hand for the long-range triangular matrices of type (iii) the number of iterations grows only as the logarithm of the number of blocks, $\log L$, (at least for large L). The propagation of the reconstructed region does not really correspond to a localized traveling wave, as visible from figure 13(b). In both cases the speed of growth of the seed (i.e. reconstructed region) is proportional to the interaction strength between the first non-reconstructed block and the seed. In the case of one-dimensionally coupled matrices this strength does not depend on the position of the seed boundary. In the case of a triangular seeding matrix the strength is proportional to the size of the already reconstructed region, hence $\delta L/\delta t \sim L$, which gives the logarithmic dependence seen in figure 15 .

In figure 16 right, and figure 14 right we show that the BP reconstruction with seeding matrices works also in the case when the signal model does not correspond at all to the actual signal distribution. In the two figures the signal components are $0, \pm 1$, whereas the signal model was still Gauss–Bernoulli. Since the probabilistic approach is optimal for noiseless measurements even when the signal distribution is not known (as proven in section 2.1) the seeding strategy is able to approach the information theoretic limit $\alpha \to \rho_0$ also in this case.

We have not done expectation maximization learning in the data presented in this section, but this strategy is also useful with the seeding matrices and is included in our

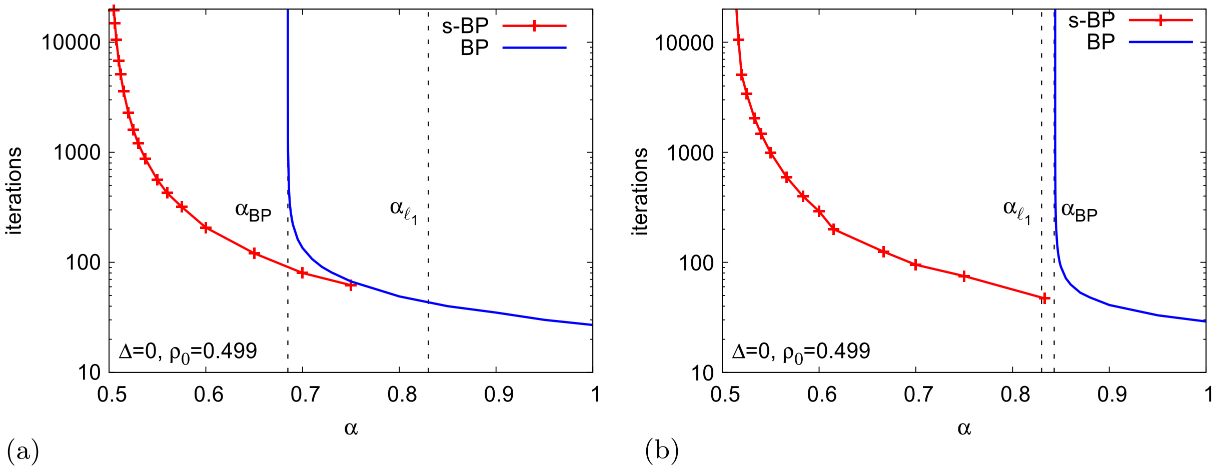

Figure 14. Reaching the $\alpha \to \rho_0$ limit. Number of iterations needed to find the original signal of density $\rho_0 = 0.499$ for (a) a Gauss–Bernoulli signal and (b) a 0, ±1 signal. In both cases, we used BP with a Gauss–Bernoulli signal model with $\rho = \rho_0$. The blue line shows the BP convergence time for homogeneous matrices, which diverges at the spinodal line $\alpha_{\rm BP}$. The red line shows the BP reconstruction done with type (iii) seeding matrices: (a) using $\alpha_{\rm seed} = 0.8$, $\alpha_{\rm bulk} = 0.5$, and $J = 2 \times 10^{-3}$, (b) using $\alpha_{\rm seed} = 1$, $\alpha_{\rm bulk} = 0.5$, and $J = 0.01$ (in this case we added one block of measurements, $L_r = L + 1$). As L increases in both cases, the total measurement rate α decreases and approaches $\alpha_{\rm bulk} = 0.5 \approx \rho_0 = 0.499$. The number of iterations needed for exact reconstruction then diverges with L. The difference between the reconstruction limits of BP and of ℓ_1 is striking.

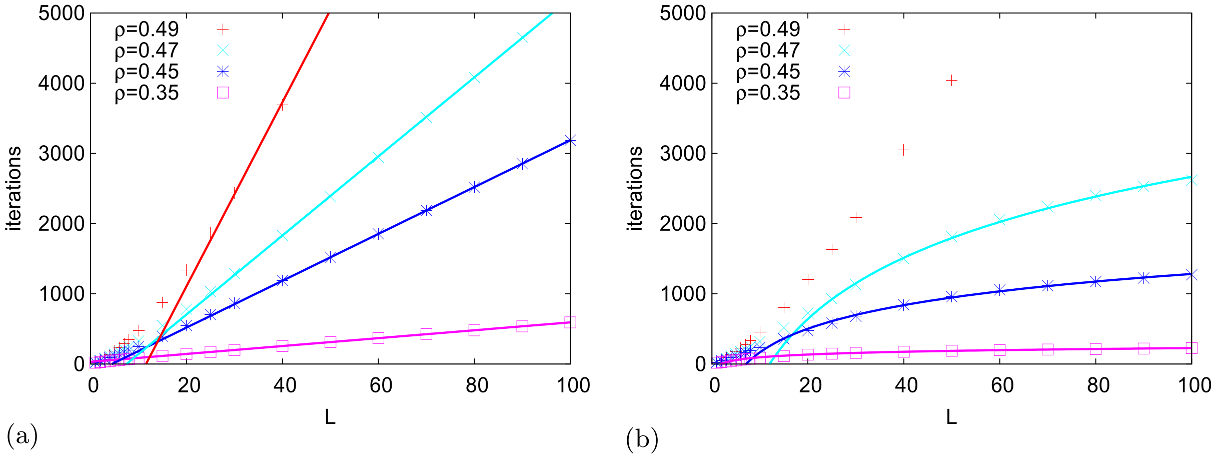

Figure 15. Number of iterations needed for reconstruction with type (ii) seeding matrices (on the left) and type (iii) seeding matrices (on the right). With type (ii) matrices, a wave is propagating in the system with a constant speed, while for type (iii) matrices with long-range interactions the speed is proportional to L, hence the total time scales as $\log L$. Left: we used $J = 0.02, W = 2$, $\alpha_{\rm seed} = 1.0$, $\alpha_{\rm bulk} = 0.5$. Right: we used $J = 0.01$, $\alpha_{\rm seed} = 1.0$, $\alpha_{\rm bulk} = 0.5$. We used $\rho = \rho_0$ to make these data.

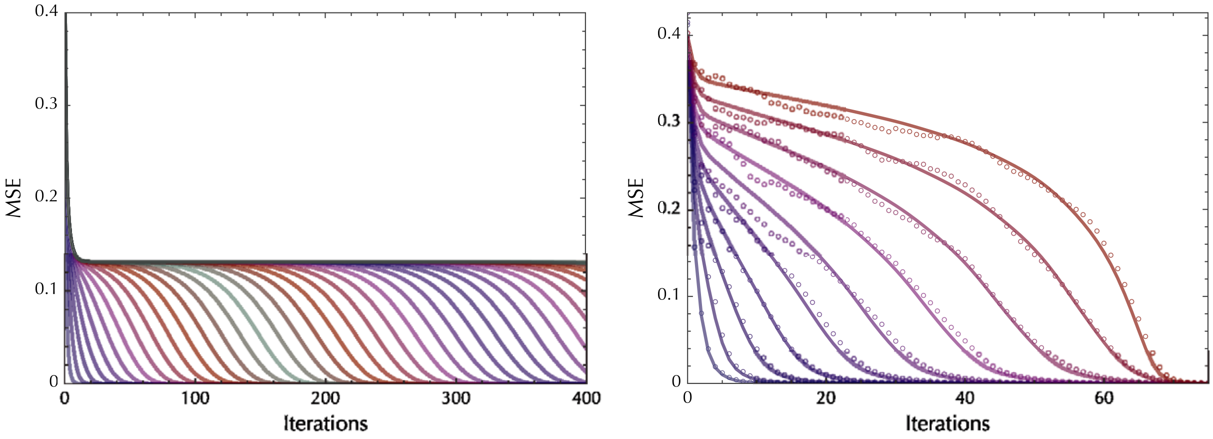

Figure 16. Left: evolution of the mean-squared error in each block as a function of iteration time. Here we used type (ii) seeding matrices with $W = 2$, $L = 50$, $\alpha_{\text{seed}} = 1.0$, $\alpha_{\text{bulk}} = 0.5$, and $J = 0.01$. With that type of matrix, the boundary of the reconstructed region is propagating as a localized wave. Right: same as figure 13 for a 'adversary-case' signal having components 0, ±1, with $N = 10\,000$. We used $\alpha = 0.6$ and $\rho_0 = 0.4$, with the seeding matrix of type (iv) with $L = 10$, $\alpha_{\text{seed}} = 1.0$, $\alpha_{\text{bulk}} = 0.5$, $J = 0.1$. Exact reconstruction is achieved even though the signal model (Gauss–Bernoulli) does not correspond to the empirical signal distribution.

implementations. Its behavior is analogous to the one in the case of homogeneous matrices, as discussed in section 5.2.

6.3. Seeding experiments for noisy measurements

Every CS method requires robustness with respect to the measurement noise. In section 5.3 we analyzed the phase diagram under measurement noise. In particular we showed the existence of two phase transitions $\alpha_c(\rho_0)$ and $\alpha_d(\rho_0)$ (see e.g. figure 8) such that for $\alpha \notin (\alpha_c, \alpha_d)$ and for the signal model matching the empirical signal distribution the belief propagation inference is as good as the optimal Bayesian inference. In other words the final MSE achieved by BP for $\alpha < \alpha_c$ or $\alpha > \alpha_d$ is the best achievable for a given measurement matrix $\mathbf{F}$. If a stronger noise robustness is required then one would have to use a different measurement protocol or much larger sampling rate α. The only region that is open to improvement is for measurement rates $\alpha_c < \alpha < \alpha_d$. With seeding we can indeed improve considerably the final MSE in this region. The noise stability of the seeding strategy was touched upon already in [1], see also [14] for a rigorous discussion.

The performance of the seeding strategy in the presence of noise can be again studied using the replica/density evolution equation. In figure 17 we illustrate the evolution of the MSE for CS with noisy measurements for subsampling rates $\alpha_c < \alpha < \alpha_d$ for which BP with the homogeneous matrices gives a MSE much larger than the noise variance Δ. Again, in order to have a working seeding mechanism, the free entropy associated with the fixed point of BP close to the solution must dominate the free entropy associated with the metastable state. This is the case for measurement rates $\alpha_c < \alpha_{\text{bulk}} < \alpha_d$ and can thus be exploited.

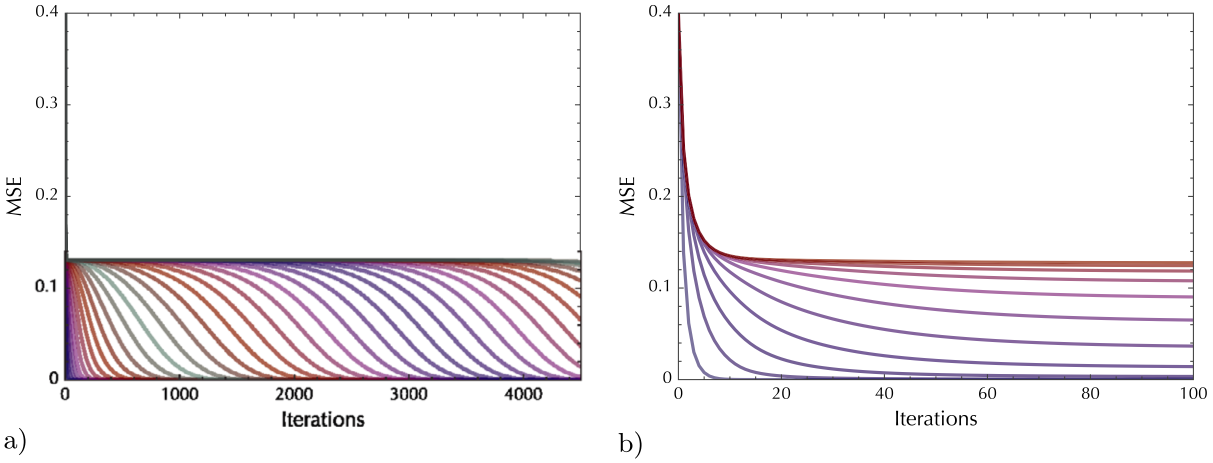

Figure 17. Seeding matrices with noise: evolution of the MSE in each block, as in figure 16, but in the noisy case. Here we used type (iv) seeding matrices with $W = 2$, $L = 100$, $\alpha_{\text{seed}} = 1.0$, $\alpha_{\text{bulk}} = 0.5$, and $J = 0.001$. Left: even with a large noise with standard deviation $\sqrt{\Delta} = 10^{-3}$, the front wave is still propagating with a finite speed, leading to a reconstruction with a final MSE of the order of Δ. This demonstrates the robustness of our approach with additive noise. Right: when the noise (here $\sqrt{\Delta} = 10^{-2}$) is too high (so that $\alpha_{\text{bulk}} < \alpha_c$, see text) there is no such propagation. Here, only the very first blocks (the first one having $\alpha_{\text{seed}} = 1$ and its close neighbors) go to low MSE while the rest of the system stays far from the solution. Essentially all the changes are done in the 100 first iterations shown here. In this case, the seeding matrices do not bring improvement (and no other method could in this case).

Note, however, that in the presence of noise the free energy difference between the global and local maxima is finite (whereas it was diverging for the noiseless case), this means that the seeding matrices need to be constructed with more care in order to saturate the threshold. In particular the interaction width W (see figure 12) has to grow when the threshold α_c is approached.

7. Conclusion

This paper presents a detailed analysis of the new strategy for compressed sensing that we introduced in [1]. With respect to this earlier work we have provided here a more detailed study of the phase diagrams and of the associated phase transition for BP reconstruction algorithm and for the (intractable) optimal reconstruction. We have treated in detail the case of noisy measurements and we have shown that our approach presents excellent stability with respect to noise, in the sense that the BP algorithm (with seeding if needed) is able to reconstruct the signal with mean-squared error as low as the optimal inference algorithm based on exhaustive enumeration (which is of course not computationally tractable). We have discussed reconstruction in the case of mismatching signal model and signal distribution and we have shown that in the noiseless case this mismatch does not pose a serious problem. We have introduced and studied new types of seeding measurement matrices with which we were also able to achieve reconstruction at almost optimal reconstruction rates.

Acknowledgments

This work has been supported in part by the EC Grant STAMINA, No. 265496, and by the Grant DySpaN of Triangle de la Physique. We thank Richard Morris for his comments on our manuscript.

Appendix A. Derivation of the replica analysis for block matrices

Here, we rederive the replica analysis for the seeding matrices described in the main text section 6. This follows closely the derivation presented in section 4.2, we only need add the block indices. To evaluate the average of the replicated partition function (92) for the block matrices $\mathbf{F}$ we introduce the order parameters per block

$$m_p^a = \frac{1}{N_p} \sum_{i \in B_p} x_i^a s_i, \qquad a = 1, 2, \ldots, n \tag{A.1}$$

$$Q_p^a = \frac{1}{N_p} \sum_{i \in B_p} (x_i^a)^2, \qquad a = 1, 2, \ldots, n \tag{A.2}$$

$$q_p^{ab} = \frac{1}{N_p} \sum_{i \in B_p} x_i^a x_i^b, \qquad a < b, \tag{A.3}$$

where B_p represents the index of the variables in block $p = 1, \ldots, L_c$.

We introduce a Dirac delta function that fixed the order parameters and we make use of the following integral representation for delta function:

$$1 = \int \prod_{a,p} \mathrm{d}\hat{Q}_p^a \, \mathrm{d}Q_p^a \prod_{a \neq b,p} \mathrm{d}\hat{q}_p^{ab} \, \mathrm{d}q_p^{ab} \prod_{a,p} \mathrm{d}\hat{m}_p^a \, \mathrm{d}m_p^a \, \mathrm{e}^{(1/2) \sum_{p=1}^{L_c} \sum_{a=1}^{n} \hat{Q}_p^a (N_p Q_p^a - \sum\limits_{i \in B_p} (x_i^a)^2)}$$
$$\times \, \mathrm{e}^{-(1/2) \sum_{p=1}^{L_c} \sum_{a \neq b} \hat{q}_p^{ab} (N_p q_p^{ab} - \sum_{i \in B_p} x_i^a x_i^b)} e^{-\sum_{p=1}^{L_c} \sum_{a=1}^{n} \hat{m}_p^a (N_p m_p^a - \sum_{i \in B_p} x_i^a x_i^0)}. \tag{A.4}$$

Inserting equation (A.4) into the expression of $\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}}(Z^n)$, equation (92), we get

$$\mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}}(Z^n) = \int \prod_{a,p} \mathrm{d}\hat{Q}_p^a \, \mathrm{d}Q_p^a \, \mathrm{d}\hat{m}_p^a \, \mathrm{d}m_p^a \, \mathrm{e}^{\sum_{p=1}^{L_c} N_p \left[(1/2) \hat{Q}_p^a Q_p^a - \hat{m}_p^a \, \mathrm{d}m_p^a \right]}$$
$$\times \prod_{a \neq b,p} \mathrm{d}\hat{q}_p^{ab} \, \mathrm{d}q_p^{ab} \, \mathrm{e}^{-(1/2) \sum_{p=1}^{L_c} N_p \hat{q}_p^{ab} q_p^{ab}} \int \prod_{i,a} \mathrm{d}x_i^a$$
$$\times \prod_{i,a} [(1-\rho)\delta(x_i^a) + \rho\phi(x_i^a)] \prod_i [(1-\rho_0)\delta(s_i) + \rho_0 \phi_0(s_i)]$$
$$\times \prod_\mu \frac{1}{\sqrt{2\pi\Delta_\mu}} \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}} [\mathrm{e}^{-(1/2\Delta_\mu) \sum_{a=1}^{n} (\sum_{i=1}^{N} F_{\mu i} s_i + \xi_\mu - \sum_{i=1}^{N} F_{\mu i} x_i^a)^2}]$$
$$\times \prod_{p=1}^{L_c} \mathrm{e}^{-(1/2) \sum_a \hat{Q}_p^q (\sum_{i \in B_p} x_i^a)^2 + (1/2) \sum_{a \neq b} \hat{q}_p^{ab} \sum_{i \in B_p} x_i^a x_i^b + \sum_a \hat{m}_p^a \sum_{i \in B_p} x_i^a s_i}.$$

When averaging Z^n, we again need to evaluate the quantity X_μ, defined in equation (96). We define $u_\mu^a = \sum_{i=1}^N F_{\mu i} x_i^a$, with $a = \{0, 1, \ldots, n\}$, where 0 corresponds to the index of the signal, $x_i^0 = s_i$. The quantities then obey a joint Gaussian distribution with $\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}(u_\mu^a) = 0$ and

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}(u_\mu^0 u_\mu^0) = \rho_0 \overline{s^2} \sum_{p=1}^{L_c} J_{I(\mu)p} n_p, \qquad \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}(u_\mu^a u_\mu^0) = \sum_{p=1}^{L_c} J_{I(\mu)p} n_p m_p^a, \tag{A.5}$$

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}(u_\mu^a u_\mu^a) = \sum_{p=1}^{L_c} J_{I(\mu)p} n_p Q_p^a, \qquad \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}(u_\mu^a u_\mu^b) = \sum_{p=1}^{L_c} J_{I(\mu)p} n_p Q_p^a. \tag{A.6}$$

Under the replica symmetric ansatz the replicas are considered equivalent, i.e.

$$m_p^a = m_p, \qquad q_p^{ab} = q_p, \qquad Q_p^a = Q_p. \tag{A.7}$$

We introduce $\tilde\rho_q, \tilde m_q, \tilde q_q, \tilde Q_q$ as follows

$$\begin{aligned} \tilde\rho_q &= \rho_0 \overline{s^2} \sum_{p=1}^{L_c} J_{qp} n_p, \qquad & \tilde m_q &= \sum_{p=1}^{L_c} J_{qp} n_p m_p, \\ \tilde q_q &= \sum_{p=1}^{L_c} J_{qp} n_p q_p, \qquad & \tilde Q_q &= \sum_{p=1}^{L_c} J_{qp} n_p Q_p. \end{aligned} \tag{A.8}$$

And thus $v_\mu^a = u_\mu^0 - u_\mu^a + \xi_\mu$, $a = 1, 2, \ldots, n$ are also joint Gaussian distributed with zero means and

$$G_{aa} = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}(v_\mu^a v_\mu^a) = \tilde Q_{I(\mu)} + \tilde\rho_{I(\mu)} - 2\tilde m_{I(\mu)} + \Delta_0, \qquad a = 1, 2, \ldots, n, \tag{A.9}$$

$$G_{ab} = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}(v_\mu^a v_\mu^b) = \tilde q_{I(\mu)} + \tilde\rho_{I(\mu)} - 2\tilde m_{I(\mu)} + \Delta_0, \qquad a < b, \tag{A.10}$$

where G is the inverse covariance matrix. For the block matrices we have

$$\det\left(\mathbb{1} + \frac{G}{\Delta}\right) = \mathrm{e}^{n[(\tilde q_{I(\mu)} - 2\tilde m_{I(\mu)} + \tilde\rho_{I(\mu)} + \Delta_0/\tilde Q_{I(\mu)} - \tilde q_{I(\mu)} + \Delta) + \log(1 + (\tilde Q_{I(\mu)} - \tilde q_{I(\mu)}/\Delta))]}. \tag{A.11}$$

From here, following the same steps as for derivation of equation (112), we obtain

$$\begin{aligned} \mathbb{E}_{\mathbf{F},\mathbf{s},\boldsymbol{\xi}} Z^n = \int \prod_p & \mathrm{d}\hat Q_p \, \mathrm{d}Q_p \, \mathrm{d}\hat q_p \, \mathrm{d}q_p \, \mathrm{d}\hat m_p \, \mathrm{d}m_p \\ & \times \exp\left(nN \left\{ \frac{1}{2} \sum_{q=1}^{L_r} n_1 \alpha_{q1} \left[\frac{\tilde q_q - 2\tilde m_q + \tilde\rho_q + \Delta_0}{\tilde Q_q - \tilde q_q + \Delta} + \log(\Delta + \tilde Q_q - \tilde q_q) \right] \right.\right. \\ & + \sum_{p=1}^{L_c} n_p \left(\frac{Q_p \hat Q_p}{2} - m_p \hat m_p + \frac{q_p \hat q_p}{2} \right) \\ & + \sum_{p=1}^{L_c} n_p \int \mathrm{d}s \, [(1 - \rho_0)\delta(s) + \rho_0 \phi_0(s)] \\ & \left.\left. \times \int \mathcal{D}z \log \int \mathrm{d}x \, \mathrm{e}^{-(\hat Q_p + \hat q_p/2)x^2 + x(\hat m_p s + z\sqrt{\hat q_p})} [(1 - \rho)\delta(x) + \rho\phi(x)] \right\} \right), \end{aligned} \tag{A.12}$$

where $\alpha_q = M_q/N$. From here we obtain the expression of free entropy in equation (136).

Appendix B. Phase diagram of the ℓ_1 reconstruction for seeding matrices

In this section, we apply the well-known ℓ_1 norm reconstruction for the seeding matrix (i) in figure 12, i.e., for the coupling matrix, $J_{p,p}=1, J_{p,p-1}=J_1, J_{p_1,p}=J_2$ and others are zeros, and see if the delicately designed matrix can also provide substantial improvement for the reconstruction limit.

In order to study the ℓ_1 norm, we use the large β limit of the problem defined by the partition function

$$Z=\int\prod_{i=1}^{N}(\mathrm{d}x_i\,\mathrm{e}^{-\beta|x_i|})\prod_{\mu=1}^{M}\delta\left(\sum_i F_{\mu i}(x_i-s_i)\right). \tag{B.1}$$

We can use all our previous replica computation with the substitution in the local measure of $(1-\rho)\delta(x_i)+\rho\phi_0(x_i)$ by $e^{-\beta|x_i|}$. In the case of our seeding matrix $\mathbf{F}$, this gives $Z=\int \mathrm{e}^{nN\Phi}$ with

$$\begin{aligned}
&\Phi(\{Q_p\}_{p=1}^L,\{q_p\}_{p=1}^L,\{m_p\}_{p=1}^L,\{\hat{Q}_p\}_{p=1}^L,\{\hat{q}_p\}_{p=1}^L,\{\hat{m}_p\}_{p=1}^L)\\
&\qquad= \ -\frac{1}{2}\sum_{p=1}^{L}\alpha_p\left[\frac{\tilde{q}_p-2\tilde{m}_p+\tilde{\rho}_p}{\tilde{Q}_p-\tilde{q}_p}+\log(\tilde{Q}_p-\tilde{q}_p)\right]\\
&\qquad\quad+\sum_{p=1}^{L}\left(\frac{Q_p\hat{Q}_p}{2}-m_p\hat{m}_p+\frac{q_p\hat{q}_p}{2}\right)\\
&\qquad\quad+\sum_{p=1}^{L}\int\mathcal{D}z\int\mathrm{d}s\,[(1-\rho)\delta(s)+\phi_0(s)]\\
&\qquad\quad\times\ \log\left\{\int\mathrm{d}x\,\mathrm{e}^{-((\hat{Q}_p+\hat{q}_p)/2)x^2+x(\hat{m}_p s+z\sqrt{\hat{q}_p})}\mathrm{e}^{-\beta|x|}\right\},
\end{aligned}\tag{B.2}$$

where we always use (137):

$$\tilde{\rho}_p=\rho_0\langle s^2\rangle\sum_{q=1}^{L}J_{pq},\qquad \tilde{m}_p=\sum_{q=1}^{L}J_{pq}m_q,\qquad \tilde{q}_p=\sum_{q=1}^{L}J_{pq}q_q,\qquad \tilde{Q}_p=\sum_{q=1}^{L}J_{pq}Q_q. \tag{B.3}$$

In the large β limit, we assume the scaling for the order parameters as follows:

$$\begin{aligned}
&\hat{Q}_p+\hat{q}_p=\beta\hat{R}_p,\qquad \hat{q}_p=\beta^2\hat{r}_p,\qquad \hat{m}_p=\beta\hat{\mu}_p\\
&Q_p=O(1),\qquad q_p=O(1),\qquad m_p=O(1),\qquad Q_p-q_p=O(1/\beta)
\end{aligned}\tag{B.4}$$

and we write specifically $r_p=\beta(Q_p-q_p)$. Therefore, the free entropy Φ scales linearly in β:

$$\begin{aligned}
\frac{\Phi}{\beta}=&-\frac{1}{2}\sum_{p=1}^{L}\alpha_p\left[\frac{\tilde{q}_p-2\tilde{m}_p+\tilde{\rho}_p}{\tilde{r}_p}\right]+\sum_{p=1}^{L}\left(\frac{q_p\hat{R}_p}{2}-m_p\hat{\mu}_p-\frac{r_p\hat{r}_p}{2}\right)\\
&-\sum_{p=1}^{L}\int\mathcal{D}z\int\mathrm{d}s\,[(1-\rho_0)\delta(s)+\phi_0(s)]\\
&\times\ \min_x\left(\frac{\hat{R}_p}{2}x^2-(\hat{\mu}_p s+\sqrt{\hat{r}_p}z)x+|x|\right).
\end{aligned}\tag{B.5}$$

doi:10.1088/1742-5468/2012/08/P08009

It is easy to check that, for $L = 1$, this gives back the free energy written e.g. by Kabashima *et al* [24].

In order to study the transition, we assume the following scaling when one is near to the regime of exact retrieval of the signal:

$$\forall\, p \in \{1, \dots, L\} : \qquad \hat{\mu}_p \to \infty, \qquad \hat{r}_p = 1/\lambda_p^2. \tag{B.6}$$

With the above scaling we find that the saddle point equations of order parameters $r_p, E_p = q_p - 2m_p + \rho_0\langle s^2\rangle, \hat{\mu}_p$ and $\hat{r}_p$ are independent of the distribution of the non-zero elements in signal $\phi_0(x)$, as long as $\phi_0(x) = \phi_0(-x)$. For $L = 1$, i.e., the canonical matrix $\mathbf{F}$, the saddle point equations are given as:

$$r = \frac{1}{\hat{\mu}}\left[\rho_0 + 2(1-\rho_0)\int_\lambda^\infty Dz\right], \tag{B.7}$$

$$E = \frac{1}{\lambda^2\hat{\mu}^2}\left[2(1-\rho_0)\int_\lambda^\infty Dz\,(z-\lambda)^2 + \rho_0(1+\lambda^2)\right], \tag{B.8}$$

$$\hat{\mu} = \frac{\alpha}{r} \tag{B.9}$$

$$\hat{r} = \frac{\alpha}{r^2}[q - 2m + \rho_0\langle s^2\rangle] = \frac{\hat{\mu}^2}{\alpha}[q - 2m + \rho_0\langle s^2\rangle]. \tag{B.10}$$

They can be simplified further as a closed system of two variables α, λ:

$$\begin{aligned} \alpha &= \rho_0 + 2(1-\rho_0)\int_\lambda^\infty Dz, \\ \alpha &= 2(1-\rho_0)\int_\lambda^\infty Dz(z-\lambda)^2 + \rho_0(1+\lambda^2). \end{aligned} \tag{B.11}$$

They give the critical value of α for a given value of ρ_0; these are the equations of [7, 24].

For the seeding matrix, $L \geq 2$, due to the fact that the final result does not depend on ϕ_0 (for symmetric ones), we thus take ϕ_0 as a centered Gaussian distribution of variance one. After some work we get:

$$r_p = \frac{2}{\hat{\mu}_p}\left[(1-\rho_0)H\left(\frac{1}{\sqrt{\hat{r}_p}}\right) + \rho_0 H\left(\frac{1}{\sqrt{\hat{r}_p + \hat{\mu}_p^2}}\right)\right], \tag{B.12}$$

$$\begin{aligned} E_p = \frac{1}{\hat{\mu}_p^2}\Bigg[&2(1-\rho_0)\hat{r}_p\psi\left(\frac{1}{\sqrt{\hat{r}_p}}, \frac{1}{\sqrt{\hat{r}_p}}\right) + 2\rho_0(\hat{r}_p + \hat{\mu}_p^2)\psi\left(\frac{1}{\sqrt{\hat{r}_p + \hat{\mu}_p^2}}, \frac{1}{\sqrt{\hat{r}_p + \hat{\mu}_p^2}}\right) \\ &- 4\rho_0\hat{\mu}_p^2 H\left(\frac{1}{\sqrt{\hat{r}_p + \hat{m}u_p^2}}\right) + \rho_0\hat{\mu}_p^2\Bigg], \end{aligned} \tag{B.13}$$

$$\hat{\mu}_p = \sum_q J_{qp}\frac{\alpha_q}{\sum_s J_{qs}r_s}, \tag{B.14}$$

$$\hat{r}_p = \sum_q J_{qp}\frac{\alpha_q}{(\sum_s J_{qs}r_s)^2}\left(\sum_s J_{qs}E_s\right), \tag{B.15}$$

Table B.1. Parameters used for the probabilistic-BP reconstruction of the Gaussian signal (left) and with the seeded ℓ_1.

ρ_0^{BP}	α	α_{seed}	α_{bulk}	J_1	J_2	L	$\rho_0^{\ell_1}$	α	α_{seed}	α_{bulk}	J_1	J_2	L	ρ_0 $(L=1)$
0.1	0.130	0.3	0.121	40	1.2	20	0.014	0.130	0.3	0.121	0.03	0.31	20	0.016
0.2	0.227	0.4	0.218	10	0.8	20	0.056	0.227	0.4	0.218	0.03	0.31	20	0.059
0.3	0.328	0.6	0.314	8	0.4	20	0.096	0.328	0.6	0.314	0.097	0.57	20	0.100
0.4	0.426	0.7	0.412	4	0.4	20	0.145	0.426	0.7	0.412	0.03	0.57	20	0.150
0.6	0.624	0.9	0.609	2	0.2	20	0.278	0.624	0.9	0.609	0.175	0.57	20	0.283
0.8	0.816	0.95	0.809	2	0.2	20	0.476	0.816	0.95	0.809	0.175	0.31	20	0.481

where

$$H(a) = \int_a^\infty Dz \tag{B.16}$$

and

$$\psi(a,b) = \int_a^\infty Dz\,(z-b)^2 = (1+b^2)H(a) + \frac{(a-2b)}{\sqrt{2\pi}}\mathrm{e}^{-a^2/2}. \tag{B.17}$$

Table B.1 gives the reconstruction threshold ρ_0 obtained for the same values of the α with probabilistic reconstruction (left) and ℓ_1 (right). The ℓ_1 results are obtained by optimizing over J_1, J_2 in the window $[0.03, 1]$. We notice that the results of ℓ_1 with optimal J_1, J_2 are slightly worse than the results of ℓ_1 with just one block $L = 1$.

Altogether, this demonstrates that the gain in performance using seeding matrices is really specific to the Bayes inference approach and hence it is the combination of the probabilistic approach, the message passing reconstruction with parameter learning, and the seeding design of the measurement matrix that is able to reach the best possible performance.

Appendix C. Equations for a mixture of Gaussians

We consider here the case when the signal model is a mixture of G Gaussians

$$\phi(x) = \sum_{a=1}^{G} w_a \mathcal{N}(\bar{x}_a, \sigma_a^2), \tag{C.1}$$

where w_a are non-negative weights $\sum_{a=1}^{G} w_a = 1$. The functions f_a and f_c (33)–(34) needed by the BP algorithm are then

$$f_a(\Sigma^2, R) = \frac{\rho\sum_{a=1}^{G} w_a \mathrm{e}^{-(R-\bar{x}_a)^2/2(\Sigma^2+\sigma_a^2)}(\Sigma/(\Sigma^2+\sigma_a^2)^{3/2})(\bar{x}_a\Sigma^2 + R\sigma_a^2)}{(1-\rho)\mathrm{e}^{-R^2/2\Sigma^2} + \rho\sum_{a=1}^{G} w_a(\Sigma/\sqrt{\Sigma^2+\sigma_a^2})\mathrm{e}^{-(R-\bar{x}_a)^2/2(\Sigma^2+\sigma_a^2)}}, \tag{C.2}$$

$$\begin{aligned} f_c(\Sigma^2, R) = {} & \left\{\rho\sum_{a=1}^{G} w_a \mathrm{e}^{-((R-\bar{x}_a)^2/2(\Sigma^2+\sigma_a^2))}(\Sigma/(\Sigma^2+\sigma_a^2)^{5/2})[\sigma_a^2\Sigma^2(\Sigma^2+\sigma_a^2)\right. \\ & \left. + (\bar{x}_a\Sigma^2 + R\sigma_a^2)^2]\right\}\left\{(1-\rho)\mathrm{e}^{-R^2/2\Sigma^2}\right. \\ & \left. + \rho\sum_{a=1}^{G} w_a(\Sigma/\sqrt{\Sigma^2+\sigma_a^2})\mathrm{e}^{-(R-\bar{x}_a)^2/2(\Sigma^2+\sigma_a^2)}\right\}^{-1} - f_a^2. \end{aligned} \tag{C.3}$$

For a signal that itself is a mixture of Gaussians

$$\phi_0(x) = \sum_{a=1}^{G_0} w_a^0 \mathcal{N}(\bar{x}_a^0, (\sigma_a^0)^2), \tag{C.4}$$

the density evolution equations (115)–(117) simplify into single-Gaussian-integral equations

$$\begin{aligned} E = \rho_0 \overline{s^2} &- 2\rho_0 \sum_{a=1}^{G_0} w_a^0 \bar{x}_a^0 \int \mathcal{D}z f_a \left(\frac{1}{\hat{m}}, z\sqrt{(\sigma_a^0)^2 + \frac{\hat{q}}{\hat{m}^2}} + \bar{x}_a^0 \right) \\ &- 2\rho_0 \sum_{a=1}^{G_0} w_a^0 \hat{m} (\sigma_a^0)^2 \int \mathcal{D}z f_c \left(\frac{1}{\hat{m}}, z\sqrt{(\sigma_a^0)^2 + \frac{\hat{q}}{\hat{m}^2}} + \bar{x}_a^0 \right) \\ &+ (1-\rho_0) \int \mathcal{D}z\, f_a^2 \left(\frac{1}{\hat{m}}, z\frac{\sqrt{\hat{q}}}{\hat{m}} \right) \\ &+ \rho_0 \sum_{a=1}^{G_0} w_a^0 \int \mathcal{D}z f_a^2 \left(\frac{1}{\hat{m}}, z\sqrt{(\sigma_a^0)^2 + \frac{\hat{q}}{\hat{m}^2}} + \bar{x}_a^0 \right), \end{aligned} \tag{C.5}$$

$$V = (1-\rho_0) \int \mathcal{D}z\, f_c \left(\frac{1}{\hat{m}}, z\frac{\sqrt{\hat{q}}}{\hat{m}} \right) + \rho_0 \sum_{a=1}^{G_0} w_a^0 \int \mathcal{D}z f_c \left(\frac{1}{\hat{m}}, z\sqrt{(\sigma_a^0)^2 + \frac{\hat{q}}{\hat{m}^2}} + \bar{x}_a^0 \right), \tag{C.6}$$

where we used integration by parts to obtain the simplification in the first equation. We took advantage of the fact that a double Gaussian integral of a function that depends only on a sum of the Gaussian variables can be written as a single Gaussian integral with variance being the sum of variances and mean being the sum of means. We note $1/\hat{m} = (\Delta + V)/\alpha$, and $\hat{q}/\hat{m}^2 = (\Delta_0 + E)/\alpha$.

Under the optimal Bayesian inference when $\phi_0(x) = \phi(x)$, $\rho_0 = \rho$, $\Delta = \Delta_0$ the system of two equations reduces into a single one, since $E = V$ and $\hat{q} = \hat{m}$.

References

[1] Krzakala F, Mézard M, Sausset F, Sun Y and Zdeborová L, *Statistical physics-based reconstruction in compressed sensing*, 2012 *Phys. Rev.* X **2** 021005

[2] Candès E J and Wakin M B, *An introduction to compressive sampling*, 2008 *IEEE Signal Process. Mag.* **25** 21

[3] Donoho D L, *Compressed sensing*, 2006 *IEEE Trans. Inform. Theory* **52** 1289

[4] Candès E J and Tao T, *Decoding by linear programming*, 2005 *IEEE Trans. Inform. Theory* **51** 4203

[5] Candès E, Romberg J and Tao T, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, 2006 *IEEE Trans. Inform. Theory* **52** 489

[6] Donoho D L and Tanner J, *Neighborliness of randomly projected simplices in high dimensions*, 2005 *Proc. Nat. Acad. Sci.* **102** 9452

[7] Donoho D L, Maleki A and Montanari A, *Message-passing algorithms for compressed sensing*, 2009 *Proc. Nat. Acad. Sci.* **106** 18914

[8] Rangan S, *Generalized approximate message passing for estimation with random linear mixing*, 2010 arXiv: arXiv:1010.5141v1 [cs.IT]

[9] Vila J P and Schniter P, *Expectation-maximization Bernoulli-Gaussian approximate message passing*, 2011 *Proc. Asilomar Conf. on Signals, Systems, and Computers (Pacific Grove, CA)*

[10] Guo D and Wang C-C, *Asymptotic mean-square optimality of belief propagation for sparse linear systems*, 2006 *Information Theory Workshop, 2006. ITW '06 Chengdu* pp 194–8

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

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

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

[14] Donoho D L, Javanmard A and Montanari A, *Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing*, 2011 arXiv:1112.0708v1 [cs.IT]

[15] Montanari A and Bayati M, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, 2011 *IEEE Trans. Inform. Theory* **57** 764

[16] Bayati M, Lelarge M and Montanari A, *Universality in message passing algorithms*, 2012 arXiv:1207.7321

[17] Wu Y and Verdu S, *Rényi information dimension: fundamental limits of almost lossless analog compression*, 2010 *IEEE Trans. Inform. Theory* **56** 3721

[18] Wu Y and Verdu S, *Optimal phase transitions in compressed sensing*, 2011 arXiv:1111.6822v1 [cs.IT]

[19] Wu Y and Verdu S, *MMSE dimension*, 2011 *IEEE Trans. Inform. Theory* **57** 4857

[20] Ji S, Xue Y and Carin L, *Bayesian compressive sensing*, 2008 *IEEE Trans. Signal Process.* **56** 2346

[21] Seeger M W and Nickisch H, *Compressed sensing and Bayesian experimental design*, 2008 *Proc. 25th Int. Conf. on Machine Learning, ICML '08* (New York, NY: ACM) pp 912–9

[22] Baron D, Sarvotham S and Baraniuk R, *Bayesian Compressive Sensing Via Belief Propagation*, 2010 *IEEE Trans. Signal Process.* **58** 269

[23] Rangan S, Fletcher A and Goyal V, *Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing*, 2009 arXiv:0906.3234v2

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

[25] Ganguli S and Sompolinsky H, *Statistical mechanics of compressed sensing*, 2010 *Phys. Rev. Lett.* **104** 188701

[26] Gallager R G, *Low-density parity check codes*, 1962 *IEEE Trans. Inform. Theory* **8** 21

[27] Pearl J, *Reverend Bayes on inference engines: a distributed hierarchical approach*, 1982 *Proc. American Association of Artificial Intelligence National Conf. on AI (Pittsburgh, PA)* pp 133–6

[28] Yedidia J, Freeman W and Weiss Y, *Understanding belief propagation and its generalizations*, 2003 *Exploring Artificial Intelligence in the New Millennium* (San Francisco, CA: Morgan Kaufmann) pp 239–6

[29] Kschischang F R, Frey B and Loeliger H-A, *Factor graphs and the sum-product algorithm*, 2001 *IEEE Trans. Inform. Theory* **47** 498

[30] Zhang F and Pfister H D, *On the iterative decoding of high rate LDPC codes with applications in compressed sensing*, 2008 *Proc. 47th Annual Allerton Conf. on Commun., Control, and Comp.*

[31] Kabashima Y and Wadayama T, *A signal recovery algorithm for sparse matrix based compressed sensing*, 2011 arXiv:1102.3220v1 [cs.IT]

[32] Kabashima Y, *A CDMA multiuser detection algorithm on the basis of belief propagation*, 2003 *J. Phys. A: Math. Gen.* **36** 11111

[33] Guo D, Baron D and Shamai S, *A single-letter characterization of optimal noisy compressed sensing*, 2009 *47th Annual Allerton Conf. on Communication, Control, and Computing, 2009 (Allerton 2009)* pp 52–9

[34] Donoho D L, Johnstone I and Montanari A, *Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising*, 2011 arXiv:1111.1041v1 [cs.IT]

[35] Dempster A, Laird N and Rubin D, *Maximum likelihood from incomplete data via the EM algorithm*, 1977 *J. R. Stat. Soc.* **38** 1

[36] Heskes T, Zoeter O and Wiegerinck W, *Approximate expectation maximization*, 2004 *Advances in Neural Information Processing Systems 16* ed S Thrun, L Saul and B Scholkopf (Cambridge, MA: MIT Press) pp 353–60

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

[38] Jimenez Felstrom A and Zigangirov K, *Time-varying periodic convolutional codes with low-density parity-check matrix*, 1999 *IEEE Trans. Inform. Theory* **45** 2181

[39] Lentmaier M and Fettweis G, *On the thresholds of generalized LDPC convolutional codes based on protographs*, 2010 *Information Theory Proceedings (ISIT)* pp 709–13

[40] Kudekar S, Richardson T and Urbanke R, *Threshold saturation via spatial coupling: why convolutional LDPC ensembles perform so well over the BEC*, 2010 *Information Theory Proceedings (ISIT)* pp 684–8

[41] Kudekar S, Richardson T and Urbanke R, *Spatially coupled ensembles universally achieve capacity under belief propagation*, 2012 arXiv:201.2999v1 [cs.IT]

[42] Kudekar S and Pfister H, *The effect of spatial coupling on compressive sensing*, 2010 *Communication, Control, and Computing (Allerton)* pp 347–53

[43] Javanmard A and Montanari A, *Subsampling at information theoretically optimal rates*, 2012 arXiv:1202.2525v1 [cs.IT]

[44] Opper M and Haussler D, *Generalization performance of Bayes optimal classification algorithm for learning a perceptron*, 1991 *Phys. Rev. Lett.* **66** 2677

[45] Iba Y, *The Nishimori line and Bayesian statistics*, 1999 *J. Phys. A: Math. Gen.* **32** 3875

[46] Nishimori H, 2001 *Statistical Physics of Spin Glasses and Information Processing* (Oxford: Oxford University Press)

[47] Decelle A, Krzakala F, Moore C and Zdeborová L, *Phase transition in the detection of modules in sparse networks*, 2011 *Phys. Rev. Lett.* **107** 065701

[48] Natarajan B K, *Sparse approximate solutions to linear systems*, 1995 *SIAM J. Comput.* **24** 227

[49] Mézard M and Montanari A, 2009 *Information, Physics, and Computation* (Oxford: Oxford University Press)

[50] Mézard M, Parisi G and Virasoro M A, 1987 *Spin-Glass Theory and Beyond* (*Lecture Notes in Physics* vol 9) (Singapore: World Scientific)

[51] Richardson T and Urbanke R, 2008 *Modern Coding Theory* (Cambridge: Cambridge University Press)

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

[53] Binder K, *Theory of first-order phase transitions*, 1987 *Rep. Prog. Phys.* **50** 783

[54] Krzakala F and Zdeborová L, *On melting dynamics and the glass transition. II. Glassy dynamics as a melting process*, 2011 *J. Chem. Phys.* **134** 034513

[55] Hassani S, Macris N and Urbanke R, *Coupled graphical models and their thresholds*, 2010 *Information Theory Workshop (ITW)* pp 1–5