Variational Free Energies for Compressed Sensing

Florent Krzakala, Andre Manoel and Eric W. Tramel
Laboratoire de Physique Statistique, École Normale Supérieure
and Université Pierre et Marie Curie, Rue Lhomond Paris 75005 France
ESPCI and CNRS UMR 7083, 10 rue Vauquelin, Paris 75005 France

Lenka Zdeborová
Institut de Physique Théorique
CEA Saclay and URA 2306, CNRS
91191 Gif-sur-Yvette, France.

***Abstract*—We consider a variational free energy approach for compressed sensing. We first show that the naïve mean field approach performs remarkably well when coupled with a noise learning procedure. We also notice that it leads to the same equations as those used for iterative thresholding. We then discuss the Bethe free energy and how it corresponds to the fixed points of the approximate message passing algorithm. In both cases, we test numerically the direct optimization of the free energies as a converging sparse-estimation algorithm. We further derive the Bethe free energy in the context of generalized approximate message passing.**

I. Introduction

The last few years have witnessed spectacular advances in the application of message passing strategies to sparse estimation and compressed sensing (CS) [1–3]. However, these belief propagation (BP) based strategies often possess poor convergence properties in real applications. It is therefore interesting to look for alternative approaches with similar performance but better convergence properties. A standard alternative is the direct optimization of the Bethe free energy [4, 5]. The goal of the present contribution is to discuss the Bethe free energy in the context of CS, its relation with the iterative thresholding and the variational mean field approach.

In compressed sensing, we wish to estimate an unknown $N-$dimensional signal $\mathbf{x}$ from a set of $M-$dimensional measurements $\mathbf{y}$ where $M << N$, given the prior knowledge that only a small ratio ρ of the elements of $\mathbf{x}$ are non-zero. Here, $\rho = \frac{K}{N}$, where K is the number of non-zero coefficients of $\mathbf{x}$. The measurements are obtained through a linear transformation followed by a component-wise output map. We first concentrate on the case, $\mathbf{y} = \Phi\mathbf{x} + \xi$, where $\xi \sim \mathcal{N}(0, \Delta_0 \mathbf{I}_M)$ is *i.i.d.* white Gaussian noise and Φ is the $M \times N$ measurement matrix. The case of a more general output $P_{\text{out}}(\mathbf{y}|\mathbf{\Phi x})$ will be treated in Sec. IV. Graphical-models [6, 7] are a natural tool to use when discussing such problems in a probabilistic setting. Here we shall assume (although it is not strictly necessary, as shown in [3]) the knowledge of the empirical distribution of $\mathbf{x}$,

$$P_0(\mathbf{x}) = \prod_i \left[\rho\mathcal{N}(0,1) + (1-\rho)\delta(x_i)\right], \quad (1)$$

which leads to the posterior distribution

$$P(\mathbf{x}|\Phi,\mathbf{y}) = \frac{P_0(\mathbf{x})P(\mathbf{y}|\Phi,\mathbf{x})}{Z(\mathbf{y},\Phi)}, \quad (2)$$

$$= \frac{1}{Z(\mathbf{y},\Phi)} \prod_{i=1}^{N} P_0(x_i) \prod_{\mu=1}^{M} \frac{e^{-\frac{(y_\mu - \sum_{i=1}^{N}\Phi_{\mu i}x_i)^2}{2\Delta}}}{\sqrt{2\pi\Delta}}. \quad (3)$$

Our goal is to perform probabilistic inference and estimate the posterior distribution by minimizing the Gibbs free energy $\mathcal{F}$ over a trial distribution, P_{var}, with

$$\mathcal{F}(\{P_{\text{var}}\}) = D_{KL}(P_{\text{var}}||P_0(\mathbf{x})) - \langle \log P(\mathbf{y}|\mathbf{x})\rangle_{\{P_{\text{var}}\}}, \quad (4)$$

where $\langle \cdot \rangle_{\{P_{\text{var}}\}}$ denotes the average over distribution P_{var} and D_{KL} is the Kullback-Leibler divergence.

A. Outline and Main Results

We first discuss in Sec. II the naïve mean field approach to the problem. It turns out that this provides remarkably good results if one couples it with the estimation of the noise variance, Δ. We find that noise estimation for the naïve mean field, which was first considered in [8], is indeed crucial to the performance of the naïve mean field for CS. We discuss minimization of the mean field free energy as an alternative algorithm. We also show, perhaps surprisingly, that the mean field approach leads to the *same equations* as those utilized for iterative thresholding and demonstrate how the two approaches are, in fact, formally related within a Bayesian framework.

We then consider in Sec. III the Bethe free energy and show how it corresponds to the fixed point of the approximate message passing (AMP) [1–3] algorithm. We will show through an explicit minimization that the direct optimization of this free energy is a promising alternative approach to AMP. Interestingly, there is a very close relationship between the minimization of the mean field and the Bethe free energy.

Finally, in Sec. IV we derive the Bethe free energy in the case of generic output distribution P_{out}. In a recent work [9] the authors have shown how a fixed point of the generalized-AMP corresponds to a stationary point of a function. Perhaps unsurprisingly, we show that this function is nothing but the Bethe free energy itself.

II. The mean field approach

A. A Separable Ansatz

It is instructive to first review the simplest variational solution to the CS problem, namely, the mean field one where $P_{\text{var}} = \prod_i Q_i(x_i)$. In such a case, the minimum of the free energy is achieved for $Q_i(x_i) \propto \exp\langle \log\left(P(\mathbf{y}|\mathbf{x})P_0(\mathbf{x})\right)\rangle_{P_{\text{var}}(\mathbf{x}_{\setminus i})}$ where we denote $\mathbf{x}_{\setminus i}$ to be all entries of $\mathbf{x}$ which *are not* x_i. We thus observe that the variational distribution is a product of the prior and a Gaussian which defines the distribution for x_i,

$$Q(x_i;R_i,\Sigma_i) \triangleq \frac{1}{Z(R_i,\Sigma_i)} P_0(x_i) e^{-\frac{(x_i-R_i)^2}{2\Sigma_i^2}}, \quad (5)$$

with the normalization $Z(R,\Sigma)=\int \mathrm{d}x P_0(x)e^{-\frac{(x-R)^2}{2\Sigma^2}}$. Note that since we consider (5) at a single coefficient, we drop i from the notation. We denote the mean and the variance of (5) by the functions f_a and f_c, respectively,

$$f_\mathrm{a}(R,\Sigma) \triangleq \int \mathrm{d}x \frac{x}{Z(R,\Sigma)} P_0(x) e^{-\frac{(x-R)^2}{2\Sigma^2}}, \tag{6}$$

$$f_\mathrm{c}(R,\Sigma) \triangleq \int \mathrm{d}x \frac{x^2}{Z(R,\Sigma)} P_0(x) e^{-\frac{(x-R)^2}{2\Sigma^2}} - f_\mathrm{a}^2(R,\Sigma). \tag{7}$$

The following identities will be useful in the sequel,

$$\frac{\partial}{\partial R}\log Z(R,\Sigma) = \frac{f_\mathrm{a}(R,\Sigma^2)-R}{\Sigma^2}, \tag{8}$$

$$\frac{\partial}{\partial \Sigma^2}\log Z(R,\Sigma) = \frac{(f_\mathrm{a}(R,\Sigma^2)-R)^2+f_\mathrm{c}(R,\Sigma)}{2\Sigma^4}. \tag{9}$$

With this separable ansatz, we can now compute the expression for the Gibbs free energy, i.e. $\mathcal{F}$ evaluated at the distribution $\mathcal{Q}$, using the short-hand notations $a_i \triangleq f_\mathrm{a}(R_i,\Sigma_i)$ and $c_i \triangleq f_\mathrm{c}(R_i,\Sigma_i)$, for every pair (R_i,Σ_i),

$$\mathcal{F}_{MF}(\{R_i\},\{\Sigma_i\}) \triangleq \frac{M}{2}\log(2\pi\Delta) + D_{KL}(Q||P_0) + \frac{1}{2\Delta}\sum_\mu \left[(y_\mu - \sum_i \Phi_{\mu i} a_i)^2 + \sum_i \Phi_{\mu i}^2 c_i \right], \tag{10}$$

where $i \in \{1,2,\dots,N\}$, $\mu \in \{1,2,\dots,M\}$, and the Kullback-Leibler divergence between the variational ansatz and the prior is given by

$$D_{KL}(Q||P_0) = -\sum_i \left[\log Z(R_i,\Sigma_i) + \frac{c_i + (a_i - R_i)^2}{2\Sigma_i^2} \right]. \tag{11}$$

B. Stationary Points and Iterative Thresholding

We now investigate the stationary points of (10). In order to do so, we shall first consider a slightly different free energy that we shall call the "unconstrained" mean field free energy wherein we treat the a_i and c_i in (10) and (11) as free variables independent from R_i and Σ_i. At the stationary points of this unconstrained free energy we find

$$\frac{1}{\Sigma_i^2} = \sum_\mu \frac{\Phi_{\mu i}^2}{\Delta}, \tag{12}$$

$$R_i = a_i + \frac{\Sigma_i^2}{\Delta}\sum_\mu \Phi_{\mu i}(y_\mu - \sum_j \Phi_{\mu j} a_j), \tag{13}$$

$$a_i = f_\mathrm{a}(R_i,\Sigma_i), \tag{14}$$

$$c_i = f_\mathrm{c}(R_i,\Sigma_i). \tag{15}$$

From (14) and (15) we see that stationary points of the unconstrained and constrained free energy are equivalent. These equations are, in fact, nothing more than the iterative mean field method, where one updates the distribution (5) at each iteration. In [8] this method was applied, albeit with different notations, sequentially to each element in $\mathbf{x}$ in order to minimize the free energy. Properly rescaled, these equations lead to the following property.

Proposition 1: The fixed points of iterative thresholding using a given thresholding function η_Δ are identical to the (properly rescaled) stationary points of the mean field free energy (5).

Proof: If one rescales Φ such that $\sum_\mu \Phi_{\mu i}^2 = 1$, when the fixed-point equations are updated in parallel, we see that

$$\mathbf{a}^{t+1} = \eta_\Delta(\Phi^* \mathbf{z}^t + \mathbf{a}^t) \quad \text{where} \quad \mathbf{z}^t = \mathbf{y} - \Phi \mathbf{a}^t, \tag{16}$$

where $\eta_\Delta(R) = f_\mathrm{a}(R,\Delta)$. This is *exactly* iterative thresholding (see [10]). ∎

This is an interesting and, perhaps, unexpected connection which was also noticed in the context of AMP [7]. If one performs a mean field variational Bayesian learning with an ℓ_1 or ℓ_0 type "prior", then the resulting update equations are nothing more than soft and hard iterative thresholding, respectively.

C. Numerical Investigation

In order to study the performance of the mean field approach, we have performed a numerical optimization of the mean field free energy, as shown in Fig. 1, using the knowledge of both the prior distribution and the value of the true noise variance, Δ_0. Surprisingly, the results are rather poor. Since the free energy is not convex, it may possess many minima. Because of this, the correct solution is almost never found in any setting we tested.

Motivated by the results of [8] and by the strong connection between the AMP fixed points and noise estimation, which we discuss in the sequel, we thus consider Δ as a further variable to optimize over rather than a parameter. This modification is also favorable because it allows for the inference of the noise variance, a value which is generally unknown *a priori.* The estimate of the noise variance is given by the zero of the partial derivative of (10) w.r.t. Δ,

$$\Delta^* = \frac{1}{M}\|\mathbf{y} - \Phi\mathbf{a}\|_2^2 + \frac{1}{M}\|\Phi^2 \mathbf{c}\|_1, \tag{17}$$

which shows that Δ^* is a function of the proximity of the means, $\mathbf{a}$, to the measurements in the projected domain and the estimation of the variances, $\mathbf{c}$, where the square in the second term is taken element-wise.

As shown in Fig. 1, when the noise variance is learned the performance of the mean field approach improves dramatically and displays a much better phase transition in reconstruction performance than convex optimization (which does not use the prior knowledge of $P_0(x)$). This transition is very close to the one obtained by AMP when the signal is very sparse (ρ small). As noted in [8], the sequential update of (12)–(15) is guaranteed to converge to a local minimum of the mean field free energy. This algoritmh (called SOBAP in [8]) gives very similar performances and in particular the same transition as observed in the center panel of Fig. 1. Our goal in the next section is to have now a similar guarantee while matching AMP performance.

III. The Bethe Approach

AMP has been shown to be a very powerful algorithm for CS signal recovery. The algorithm is obtained by a Gaussian approximation of the BP algorithm when the measurement matrix Φ has *iid* elements of mean and variance of $O(1/N)$. We refer the reader to [1–3] and in particular to [11] for the present notation and the derivation of AMP from BP. Here, we give the iterative form of the algorithm:

$$V_\mu^{t+1} = \sum_i \Phi_{\mu i}^2 c_i^t \,, \tag{18}$$

$$\omega_\mu^{t+1} = \sum_i \Phi_{\mu i} a_i^t - (y_\mu - \omega_\mu^t)\frac{V_\mu^{t+1}}{\Delta + V_\mu^t} \,, \tag{19}$$

$$(\Sigma_i^{t+1})^2 = \left[\sum_\mu \frac{\Phi_{\mu i}^2}{\Delta + V_\mu^{t+1}}\right]^{-1} \,, \tag{20}$$

$$R_i^{t+1} = a_i^t + (\Sigma_i^{t+1})^2 \sum_\mu \Phi_{\mu i} \frac{(y_\mu - \omega_\mu^{t+1})}{\Delta + V_\mu^{t+1}} \,, \tag{21}$$

together with the consistency equations (14) and (15).

A. AMP vs. Mean Field

We now investigate the fixed points of AMP. In this case, one can solve for ω in (19) and remove this variable from all the other equations. Then, at the fixed points, we obtain

$$\frac{1}{\Sigma_i^2} = \sum_\mu \frac{\Phi_{\mu i}^2}{\Delta + \sum_j \Phi_{\mu j}^2 c_j}, \tag{22}$$

$$R_i = a_i + \frac{\Sigma_i^2}{\Delta} \sum_\mu \Phi_{\mu i}(y_\mu - \sum_j \Phi_{\mu j} a_j). \tag{23}$$

These are exactly the same equations as the mean-field ones, (12) and (13), *except* for Σ_i^2, where the Δ term has been replaced by $\Delta + \sum_j \Phi_{\mu j}^2 c_j$. This difference is crucial to the performance of AMP over the mean field iteration given in (12)–(15). The key in AMP is that the variance-like term Σ_i^2 is computed consistently with the present estimations as it incorporates the effect of all c_i. This, *a posteriori*, admits the interpretation of noise learning in the mean field approach as a method of approximating (22) by using Δ^* in (12). Hence, the similarity in performance between AMP and the mean field approach with noise learning seen in Fig. 1 for small ρ.

B. The Bethe Free Energy

While AMP is a powerful method, it does not always converge to a solution, especially if the entries of the sensing matrix are not *iid* randomly distributed. A simple modification of the mean field free energy (10) leads to what is called the Bethe free energy [4, 15],

$$\mathcal{F}^{\text{Bethe}}(\{R_i\},\{\Sigma_i\}) \triangleq \sum_\mu \frac{(y_\mu - \sum_i \Phi_{\mu i} a_i)^2}{2\Delta} + \frac{M}{2}\log 2\pi\Delta + \sum_\mu \frac{1}{2}\log\left[1 + \sum_i \Phi_{\mu i}^2 c_i/\Delta\right] + D_{KL}(Q||P_0), \tag{24}$$

where the KL distance is given by (11). This free energy is derived in Sec. IV. For now, let us accept this expression and investigate its properties.

Proposition 2: The Bethe free energy $\mathcal{F}^{\text{Bethe}}(R_i, \Sigma_i)$ in (24) has at least one minimum and is strictly upper bounded by the mean field free energy.

Proof: The proof follows from the fact that (24) is the sum of two terms which both possess a lower bound: a "cost-like" term bounded by $\frac{M}{2}\log 2\pi\Delta$ and the non-negative Kullback-Leibler term. Moreover, since $\log(1+x) \leq x$ for $x \geq 0$, one can see that $\frac{M}{2}\log 2\pi\Delta \leq \mathcal{F}^{\text{Bethe}}(\{R_i\},\{\Sigma_i\}) \leq \mathcal{F}_{MF}(\{R_i\},\{\Sigma_i\})$. ■

We shall now connect this minimum, and the other possible stationary points, to the fixed point(s) of the AMP recursion.

C. Equivalence with AMP

Theorem 1: All stationary points of the Bethe free energy correspond to fixed points of AMP.

Proof: The proof follows the same outline as in the case of the mean field free energy. Define the "unconstrained" Bethe free energy where a_i and c_i are free variables. Stationarity with respect to c_i and a_i leads to (22) and (23), and stationarity w.r.t. R_i and Σ_i to the consistency equations, (14) and (15). This demonstrates the correspondence between AMP fixed points and the stationary points of the "unconstrained" free energy. Since, at the stationarity points, the consistency equations are satisfied, then all stationarity points of the "unconstrained" free energy are stationarity points of the normal Bethe free energy (24) and vice-versa. ■

Note the difference between the "unconstrained" and the "constrained" Bethe free energy (24). While the former allows one to easily generate the AMP fixed points (a classical property, see [15]), it is not bounded and cannot generically be interpreted as a variational free energy. Only the latter "constrained" form should be considered a proper variational functional, as indicated by Proposition 2. Indeed, while we should look for a minimum of the constrained free energy, all stationary points of the unconstrained functional appear, instead, as saddle points. In fact, for a given variable i, the sign of the second derivative shows that the unconstrained free energy is a minimum for a_i and a maximum for c_i, R_i and Σ_i^2. This is, again, reminiscent of the known phenomena that the fixed points of the Bethe free energy are, in general, only saddle points unless some consistency conditions are imposed (see [16]).

D. Numerical Investigation

There are a number of ways the Bethe free energy (24) can be used. For instance, one can utilize it to damp, self-consistently, the AMP iteration to ensure a strict minimization, or at least a minimizing trend, at each AMP step. We have empirically observed that this method significantly increases the convergence properties of the AMP approach[1]. Some authors [2, 17] have, instead, used the mean field free energy to the same effect. This approach, however, does not seem well justified as it is truly the Bethe free energy which is optimized by AMP.

We have also performed numerical optimization of the Bethe free energy using the same approach as we used for the mean field approach. As shown in Fig. 1, direct minimization gives the same performance as iterating the AMP equations and, in fact, reaches the usual AMP limit obtained

[1] See our recent implementation of the AMP algorithm at http://aspics.krzakala.org/ or on GitHub at https://github.com/jeanbarbier/BPCS_common

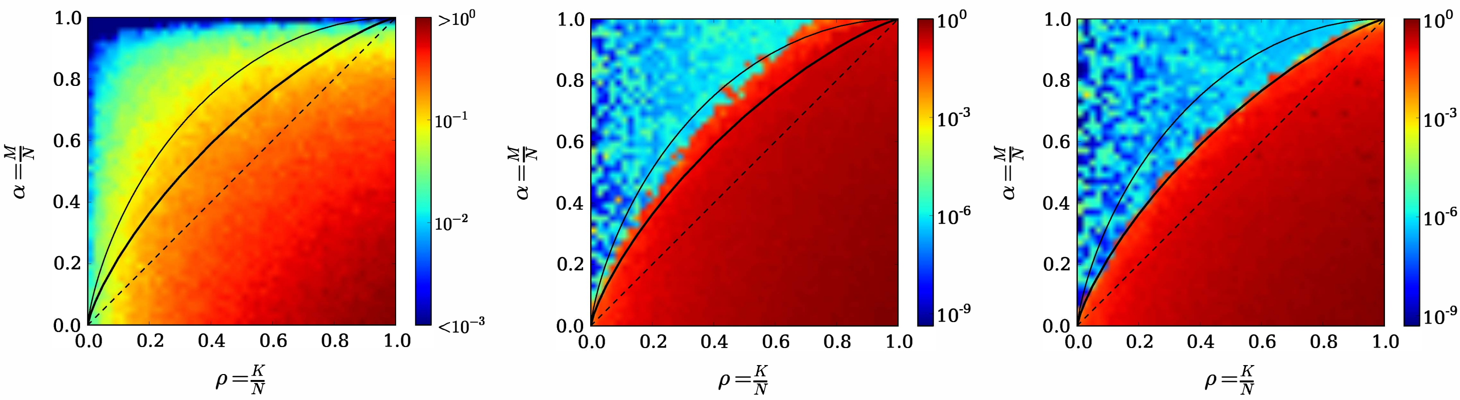

Fig. 1. Phase diagram for the performance of (from left to right) mean field, mean field with noise learning, and Bethe in the α–ρ plane using Gaussian noise with $\Delta_0 = 10^{-8}$, $N = 1024$, $\alpha = M/N$, and $\rho = K/N$. The measurements $\mathbf{y}$ are generated using matrix Φ with *iid* Gaussian elements of zero mean and unit variance. These numerical results were obtained using the function `fmin_l_bfgs_b` of the SciPy package [12] to minimize the free energy (10) (left and center) and (24) (right). The lines denote (from bottom to top) the optimal threshold for noiseless compressed sensing (straight dashed line, e.g. [13]), the Bayesian AMP phase transition for a Gauss-Bernoulli signal (reached in the right panel, see [3]) and the Donoho-Tanner transition for convex ℓ_1 reconstruction [14]. **Left:** In the pure mean field case, reconstruction is always mediocre. **Center:** With noise learning, the performance greatly improves and, in particular, outperforms convex optimization. **Right:** The best results are obtained by the minimization of the Bethe free energy which gives the same results as the AMP algorithm.

by the (rigorous [18]) state evolution analysis [3, 11]. Direct minimization of the Bethe free energy, a task that can be adressed in many different ways guaranties to find at least a local minima, is therefore a promising alternative to the AMP when convergence problems are encountered.

IV. Derivation of the Bethe Free Energy

Given the probabilistic model defined in Sec. I and that the measurement matrix Φ possesses *i.i.d.* elements of mean and variance $O(1/N)$, the fixed point of the BP equations can be used to estimate the posterior likelihood. The logarithm of this normalization is, up to a sign, called the Bethe free energy [4, 15]. In BP, one utilizes a graphical model by updating messages from constraints to variables, $m_{\mu\to i}(x_i)$, and from variables to constraints, $m_{i\to\mu}(x_i)$. Following [11], one can write the Bethe free energy for a given instance as

$$-\mathcal{F}^{\text{Bethe}} = \sum_{\mu} \log \mathcal{Z}^{\mu} + \sum_{i} \log \mathcal{Z}^{i} - \sum_{\mu i} \log \mathcal{Z}^{\mu i} \quad (25)$$

$$\text{where } \mathcal{Z}^{i} = \int \mathrm{d}x_i \prod_{\mu} m_{\mu\to i}(x_i) P_0(x_i)\,, \quad (26)$$

$$\mathcal{Z}^{\mu i} = \int \mathrm{d}x_i m_{\mu\to i}(x_i) m_{i\to\mu}(x_i)\,, \quad (27)$$

$$\mathcal{Z}^{\mu} = \int \mathrm{d}z \frac{e^{-\frac{(\omega_\mu - z)^2}{2V_\mu}}}{\sqrt{2\pi V_\mu}} P_{\text{out}}(y_\mu|z)\,. \quad (28)$$

Note that here we follow the framework of GAMP [2] and consider the context where the observations of the sparse signal are given by element-wise measurements, y_μ, specified by some known probability distribution function $P_{\text{out}}(y_\mu|z_\mu)$, where $z_\mu = \sum_i \Phi_{\mu i} x_i$. Following the notation of [2], we define the output function as

$$g_{\text{out}}(\omega, y, V) \triangleq \frac{\int \mathrm{d}z P_{\text{out}}(y|z)\,(z-\omega)\, e^{-\frac{(z-\omega)^2}{2V}}}{V \int \mathrm{d}z P_{\text{out}}(y|z) e^{-\frac{(z-\omega)^2}{2V}}}\,. \quad (29)$$

The integrals in the evaluation of the free energy are not algorithmically tractable in their general form. Using the same notations as in [11] and the same approximations used to go from BP to AMP, which are valid in the leading order when $N \to \infty$, we shall obtain a tractable form for the free energy. First, we use the properties of the BP messages [11] to rewrite (25),

$$-\mathcal{F}^{\text{Bethe}} = \sum_{\mu} \log \mathcal{Z}^{\mu} + \sum_{i} \log \mathcal{X}^{i} + \sum_{i} \log \mathcal{Y}^{i}, \quad (30)$$

where

$$\mathcal{X}^{i} = \int \mathrm{d}x_i P_0(x_i) e^{-\frac{x_i^2}{2\Sigma_i^2} + x_i \frac{R_i}{\Sigma_i^2}}\,, \quad (31)$$

$$\log \mathcal{Y}^{i} = -\frac{R_i}{\Sigma_i^2} a_i + \frac{1}{2\Sigma_i^2}(c_i + a_i^2) + \frac{1}{2} c_i \sum_{\mu=1}^{M} \Phi_{\mu i}^2 g_{\text{out}}^2\,. \quad (32)$$

Then, we replace g_{out}^2 by its fixed-point value, $g_{\text{out}}^2 = (\sum_i \Phi_{\mu i} a_i - \omega_\mu)/V_\mu$, to obtain the following expression which gives the (negative) posterior likelihood given a *fixed point* of the GAMP equations,

$$\begin{aligned} \mathcal{F}_{\text{GAMP}}^{\text{Bethe}}\left(\{R_i\}, \{\Sigma_i\}, \{\omega_\mu\}, \{a_i\}, \{c_i\}\right) = -\sum_{\mu} \log \mathcal{Z}_\mu \\ -\sum_{i} \frac{c_i + (a_i - R_i)^2}{2\Sigma_i^2} - \sum_{\mu} \frac{(\omega_\mu - \sum_i \Phi_{\mu i} a_i)^2}{2V_\mu} \\ -\sum_{i} \log Z(R_i, \Sigma_i) \quad \text{with } V_\mu = \sum_i \Phi_{\mu i}^2 c_i\,, \end{aligned} \quad (33)$$

where $Z(R, \Sigma)$ is the same as in (5). In its present form, the Bethe free energy can be easily computed and satisfies the following theorem.

Theorem 2: (Bethe/GAMP correspondence) The fixed points of the GAMP message passing equations are the stationary points of the cost function $\mathcal{F}_{\text{GAMP}}^{\text{Bethe}}$ (33).

Proof: By setting the derivatives of (33) with respect to R_i, Σ_i, ω_μ, a_i, and c_i to zero we obtain (14), (15), (19), (21), and (20), respectively. Or, more precisely, the GAMP analogs of the equations using the generic output function $g_{\rm out}$. ■

While the fixed points of the message passing equations are the stationary points of $\mathcal{F}^{\rm Bethe}_{\rm GAMP}$, they have no reason to minimize (33). Indeed, they are only saddle points of this expression. This is no surprise: we are not only optimizing the free energy with respect to a given distribution, we also have to satisfy the consistency conditions between, for instance, the parameters Σ_i and R_i and the values a_i and c_i. Only when $a_i = f_{\rm a}(R_i, \Sigma_i)$ is there consistency between these variables. In fact, as is always the case with the Bethe free energy, only at a fixed point can it be interpreted as an estimation of the posterior. It is thus practical to return to a variational form of the free energy that one should simply minimize. To do this, we impose the consistency conditions and express the free energy as a function of the parameters of our trial distributions for the two matrices,

$$\mathcal{F}^{\rm B}_{\rm var}\left(\{R_i\},\{\Sigma_i\}\right) = \mathcal{F}^{\rm Bethe}_{\rm GAMP}\left(\{R_i\},\{\Sigma_i\},\{\omega^*_\mu\},\{a^*_i\},\{c^*_i\}\right),$$

where the $*$ variables are given in terms of the fixed points as function of R_i and Σ_i only. In order to write this variational expression in a nicer form, let us define the distribution

$$\mathcal{M}(z,\omega,V) \triangleq \frac{1}{\mathcal{Z}^\mu} P_{\rm out}(y|z) \frac{1}{\sqrt{2\pi V}} e^{-\frac{(z-\omega)^2}{2V}}. \tag{34}$$

Then, one has

$$-D_{\rm KL}(\mathcal{M}||P_{\rm out}) = \log \mathcal{Z}^\mu + \frac{\log 2\pi V + 1 + V(\partial_{\omega_\mu} g_{\rm out} + g^2_{\rm out})}{2},$$

where $\partial_{\omega_\mu} g_{\rm out}$ is the partial derivative of $g_{\rm out}$ w.r.t. ω_μ. Finally, we obtain

$$\mathcal{F}^{\rm B}_{\rm var}\left(\{R_i\},\{\Sigma_i\}\right) = \sum_i D_{\rm KL}(Q||P_0) + \sum_\mu D_{\rm KL}(\mathcal{M}||P_{\rm out}) + \frac{1}{2}\sum_\mu \left(\log 2\pi V^*_\mu + 1 + V^*_\mu \partial_{\omega_\mu} g_{\rm out}\right), \tag{35}$$

with V^*_μ and ω^*_μ satisfying their respective fixed-point conditions. Note that this is in the same form as the expression in [9]. Thus, we observe that (35) is the Bethe estimation of the posterior. An important difference with [9] is that we use the Bethe free energy in Sec. III in order to obtain a bounded variational expression which is in turn used to recast AMP and GAMP as cost minimization problems; however, [9] discusses a promising ADMM-like strategy. Lastly, one can observe that, in the case of CS with Gaussian noise corrupted measurements, $g_{\rm out}(\omega_\mu, y_\mu, V_\mu) = (y_\mu - \omega_\mu)/(\Delta + V_\mu)$. This output function can be used to obtain (24). Note that the computation of the Bethe free energy sketched here is very general and can be repeated for instance for Bilinear generalized approximated message passing problems, see [19] for a detailed presentation.

V. CONCLUSION

We have considered the variational free energy approach for CS and discussed the properties of the resulting mean field and Bethe functional. We also demonstrate how the mean field approach paired with noise learning serves as an approximation of the AMP algorithm. Most interestingly, AMP has been recast in a form equivalent to a cost function minimization. One possible avenue for future work is to investigate efficient ways of minimizing this cost function with convergence guarantees.

ACKNOWLEDGMENT

This work has been supported in part by the ERC under the European Unions 7th Framework Programme Grant Agreement 307087-SPARCS, by the Grant DySpaN of Triangle de la Physique, and by FAPESP under grant 13/01213-8.

REFERENCES

[1] D. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *Proc. IEEE Information Theory Workshop*, Dublin, Ireland, August 2010, pp. 1 –5.

[2] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symposium on Information Theory*, St. Petersburg, Russia, July 2011, pp. 2168 –2172.

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

[4] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring Artificial Intelligence in the New Millennium*, vol. 8, pp. 236–239, 2003.

[5] V. Chandrasekaran, M. Chertkov, D. Gamarnik, D. Shah, and J. Shin, "Counting independent sets using the Bethe approximation," *SIAM Journal on Discrete Mathematics*, vol. 25, no. 2, pp. 1012–1034, 2011.

[6] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[7] A. Montanari, "Graphical models concepts in compressed sensing," *Compressed Sensing: Theory and Applications*, pp. 394–438, 2012.

[8] A. Drémeau, C. Herzet, and L. Daudet, "Boltzmann machine and mean-field approximation for structured sparse decompositions," *Signal Processing, IEEE Transactions on*, vol. 60, no. 7, pp. 3425–3438, 2012.

[9] S. Rangan, P. Schniter, E. Riegler, A. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," in *Proc. IEEE Int. Symposium on Information Theory*, Istanbul, Turkey, July 2013.

[10] A. Maleki and D. L. Donoho, "Optimally tuned iterative reconstruction algorithms for compressed sensing," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 330–341, 2010.

[11] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *J. Stat. Mech.*, 2012.

[12] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org/

[13] Y. Wu and S. Verdu, "Optimal phase transitions in compressed sensing," *Information Theory, IEEE Transactions on*, vol. 58, no. 10, pp. 6241–6263, October 2012.

[14] D. L. Donoho and J. Tanner, "Sparse nonnegative solution of underdetermined linear equations by linear programming," *Proc. National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.

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

[16] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *Information Theory, IEEE Transactions on*, vol. 51, no. 7, pp. 2282–2312, 2005.

[17] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing," *arXiv preprint arXiv:1310.2632*, 2013.

[18] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *Information Theory, IEEE Transactions on*, vol. 57, no. 2, pp. 764–785, 2011.

[19] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, "Phase transitions and sample complexity in bayes-optimal matrix factorization," *arXiv preprint arXiv:1402.1298.*