Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction

David L. Donoho
Department of Statistics
Stanford University

Arian Maleki
Department of Electrical Engineering
Stanford University

Andrea Montanari
Department of Electrical Engineering
and Department of Statistics
Stanford University

***Abstract*—In a recent paper, the authors proposed a new class of low-complexity iterative thresholding algorithms for reconstructing sparse signals from a small set of linear measurements [1]. The new algorithms are broadly referred to as AMP, for *approximate message passing*. This is the first of two conference papers describing the derivation of these algorithms, connection with the related literature, extensions of the original framework, and new empirical evidence.**

In particular, the present paper outlines the derivation of AMP from standard sum-product belief propagation, and its extension in several directions. We also discuss relations with formal calculations based on statistical mechanics methods.

I. Introduction

Let s_o be a vector in $\mathbb{R}^N$. We observe $n < N$ linear measurements of this vector through the matrix A, $y = As_o$. The goal is to recover s_o from (y, A). Although the system of equations is underdetermined, the underlying signal can still be recovered exactly or approximately if it is 'simple' or 'structured' in an appropriate sense. A specific notion of 'simplicity' postulates that s is exactly or approximately sparse.

The ℓ_1 minimization, also known as the basis pursuit [2], has attracted attention for its success in solving such underdetermined systems. It consists in solving the following optimization problem:

$$\text{minimize } \|s\|_1 , \qquad \text{subject to} \quad As = y . \tag{1}$$

The solution of this problem can be obtained through generic linear programming (LP) algorithms. While LP has polynomial complexity, standard LP solvers are too complex for use in large scale applications, such as magnetic resonance imaging and seismic data analysis. Low computational complexity of iterative thresholding algorithms has made them an appealing choice for such applications. Many variations of these approaches have been proposed. The interested reader is referred to [3] for a survey and detailed comparison. The final conclusion of that paper is rather disappointing: optimally tuned iterative thresholding algorithms have a significantly worse sparsity-undersampling tradeoff than basis pursuit.

Recently [1], we proposed an algorithm that appears to offer the best of both worlds: the low complexity of iterative thresholding algorithm, and the reconstruction power of the basis pursuit [1]. This algorithm is in fact an instance of a broader family of algorithms, that was called AMP, for approximate message passing, in [1]. The goal of this paper is to justify AMP by applying sum-product belief propagation for a suitable joint distribution over the variables $s_1, s_2, \dots, s_N$.

The paper is organized as follows: In Section II we explain the notations used in this paper. We then derive the AMP algorithm associated the basis pursuit problem in Section III. In Section IV, we consider the AMP for the basis pursuit denoising (BPDN) or Lasso problem. We will also generalize the algorithm to the Bayesian setting where the distribution of the elements of s_o is known, in Section V. Finally we will explain the connection with formal calculations based on non-rigorous statistical mechanics methods in Section VI.

Due to space limitations, proofs are omitted and can be found in a longer version of this paper [4].

II. Notations

The letters $a, b, c, \dots$ denote indices in $[n] \equiv \{1, \dots, n\}$ and $i, j, k, \dots$ represent indices in $[N] \equiv \{1, \dots, N\}$. The a, i element of the matrix A will be indicated as A_{ai}. The elements of the vectors y, s, x, and s_o are indicated by y_a, s_i, x_i, and $s_{o,i}$ respectively.

The ratio $\delta = n/N$ is a measure of indeterminacy of the system of equations. Whenever we refer to the large system limit we consider the case where $N, n \to \infty$ with δ fixed. In this limit the typical entry of A should scale as $1/\sqrt{n}$. In the concrete derivation, for the sake of simplicity we assume that $A_{ai} \in \{+1/\sqrt{n}, -1/\sqrt{n}\}$. This assumption is not crucial, and only simplifies the calculations. Although the algorithms are developed from the large system limit, in practice, they perform well even in the medium size problems with 'just' thousands of variables and hundreds of measurements [5].

III. AMP for the Basis Pursuit

In this section we consider the the basis pursuit problem as defined in Eq. (1). The derivation of AMP proceeds in 4 steps:
(1) Construct a joint distribution over $(s_1, \dots, s_N)$, parameterized by $\beta \in \mathbb{R}_+$, associated with the problem of interest and write down the corresponding sum-product algorithm.
(2) Show, by a central limit theorem argument, that for the large system limit, the sum-product messages can well be approximated by the families with two scalar parameters. Derive the update rules for these parameters.
(3) Take the limit $\beta \to \infty$ and get the appropriate rules for basis pursuit problem.
(4) Approximate the message passing rules for the large system limit. The resulting algorithm is AMP.

A. Construction of the graphical model

We consider the following joint probability distribution over the variables $s_1, s_2, \dots s_N$

$$\mu(\mathrm{d}s) = \frac{1}{Z}\prod_{i=1}^{N}\exp\left(-\beta|s_i|\right)\prod_{a=1}^{n}\delta_{\{y_a=(As)_a\}}\,. \qquad (2)$$

Here $\delta_{\{y_a=(As)_a\}}$ denotes a Dirac distribution on the hyperplane $y_a = (Ax)_a$. Products of such distributions associated with distinct hyperplanes yield a well defined measure. As we let $\beta \to \infty$, the mass of μ concentrates around the solution of (1). If the minimizer is unique and we have access to the marginals of μ, we can therefore solve (1). Belief propagation provides low-complexity heuristics for approximating such marginals.

In order to introduce belief propagation, consider the factor graph $G = (V, F, E)$ with variable nodes $V = [N]$, factor nodes $F = [n]$ and edges $E = [N] \times [n] = \{(i, a) : i \in [N], a \in [n]\}$. Hence G is the complete bipartite graph with N variable nodes and n factor nodes. It is clear that the joint distribution (2) is structured according to this factor graph. Associated with the edges of this graph are the belief propagation messages $\{\nu_{i\to a}\}_{i\in V, a\in C}$ and $\{\hat{\nu}_{a\to i}\}_{i\in V, a\in C}$. In the present case, messages are probability measures over the real line. The update rules for these densities are

$$\nu^{t+1}_{i\to a}(s_i) \;\cong\; e^{-\beta|s_i|}\prod_{b\neq a}\hat{\nu}^{t}_{b\to i}(s_i)\,, \qquad (3)$$

$$\hat{\nu}^{t}_{a\to i}(s_i) \;\cong\; \int\prod_{j\neq i}\nu^{t}_{j\to a}(s_i)\;\delta_{\{y_a-(As)_a\}}\mathrm{d}s\,, \qquad (4)$$

where a superscript denotes the iteration number and the symbol $\cong$ denotes identity between probability distributions up to a normalization constant[1].

B. Large system limit

A key remark is that in the large system limit, the messages $\hat{\nu}^t_{a\to i}(\,\cdot\,)$ are approximately Gaussian densities with variances of order N, and the messages $\nu^t_{i\to a}(\,\cdot\,)$ are accurately approximated by the product of a Gaussian and a Laplace density. We state this fact formally below. Recall that, given two measure μ_1 and μ_2 over $\mathbb{R}$, their Kolmogorov distance is given by $||\mu_1 - \mu_2||_{\mathrm{K}} \equiv \sup_{a\in\mathbb{R}}|\mu_1(-\infty, a] - \mu_2(-\infty, a]|$.

The first Lemma is an estimate of the messages $\hat{\nu}^t_{a\to i}$.

Lemma III.1. *Let $x^t_{j\to a}$ and $(\tau^t_{j\to a}/\beta)$ be, respectively, the mean and the variance of the distribution $\nu^t_{j\to a}$. Assume further $\int|s_j|^3\mathrm{d}\nu^t_{j\to a}(s_j) \le C_t$ uniformly in N, n. Then there exists a constant C'_t such that*

$$||\hat{\nu}^t_{a\to i} - \hat{\phi}^t_{a\to i}||_{\mathrm{K}} \le \frac{C'_t}{N^{1/2}(\hat{\tau}^t_{a\to i})^3}\,,$$

$$\hat{\phi}^t_{a\to i}(\mathrm{d}s_i) \equiv \sqrt{\frac{\beta A_{ai}^2}{2\pi\hat{\tau}^t_{a\to i}}}\;\exp\left\{\frac{\beta}{2\hat{\tau}^t_{a\to i}}(A_{ai}s_i - z^t_{a\to i})^2\right\}\,\mathrm{d}s_i\,, \qquad (5)$$

where the distribution parameters are given by

$$z^t_{a\to i} \equiv y_a - \sum_{j\neq i}A_{aj}x^t_{j\to a}, \qquad \hat{\tau}^t_{a\to i} \equiv \sum_{j\neq i}A^2_{aj}\tau^t_{j\to a}. \qquad (6)$$

Motivated by this lemma, we consider the computation of the means and the variances of the messages $\nu^{t+1}_{i\to a}(s_i)$. It is convenient to introduce a family of densities

$$f_\beta(s; x, b) \equiv \frac{1}{z_\beta(x, b)}\;\exp\left\{-\beta|s| - \frac{\beta}{2b}(s - x)^2\right\}. \qquad (7)$$

Also let F_β and G_β denote its mean and variance, i.e.,

$$\mathsf{F}_\beta(x; b) \equiv \mathbb{E}_{f_\beta(\,\cdot\,;x,b)}(Z)\,, \quad \mathsf{G}_\beta(x; b) \equiv \mathrm{Var}_{f_\beta(\,\cdot\,;x,b)}(Z)\,. \qquad (8)$$

¿From Eq. (6), we expect $\hat{\tau}^t_{i\to a}$ to concentrate tightly. Therefore we assume that it is independent of the edge (i, a).

Lemma III.2. *Suppose that at iteration t, the messages from the factor nodes to the variable nodes are $\hat{\nu}^t_{a\to i} = \hat{\phi}^t_{a\to i}$, with $\hat{\phi}^t_{a\to i}$ defined as in Eq. (5) with parameters $z^t_{a\to i}$ and $\hat{\tau}^t_{a\to i} = \hat{\tau}^t$. Then at the next iteration we have*

$$\nu^{t+1}_{i\to a}(s_i) = \phi^{t+1}_{i\to a}(s_i)\,\{1 + O(s_i^2/n)\}\,,$$
$$\phi^{t+1}_{i\to a}(s_i) \equiv f_\beta(s_i; \sum_{b\neq a}A_{bi}z^t_{b\to i}, \hat{\tau}^t)\,.$$

The mean and the variances of these messages are given by

$$x^{t+1}_{i\to a} = \mathsf{F}_\beta(\sum_{b\neq a}A_{bi}z^t_{b\to i}; \hat{\tau}^t),$$
$$\tau^t_{i\to a} = \beta\,\mathsf{G}_\beta\Big(\sum_{b\neq a}A_{bi}z^t_{b\to i}; \hat{\tau}^t\Big).$$

C. Large β limit

In the limit $\beta \to \infty$, we can simplify the functions F_β and G_β. Consider the soft thresholding function $\eta(x; b) = \mathrm{sign}(x)(|x| - b)_+$. It is well known that this admits the alternative characterization

$$\eta(x; b) = \mathrm{argmin}_{s\in\mathbb{R}}\left\{|s| + \frac{1}{2b}(s - x)^2\right\}. \qquad (9)$$

In the $\beta \to \infty$ limit, the integral that defines $\mathsf{F}_\beta(x; b)$ is dominated by the maximum value of the exponent, that corresponds to $s_* = \eta(x; b)$ and therefore $\mathsf{F}_\beta(x; b) \to \eta(x; b)$. The variance (and hence the function $\mathsf{G}_\beta(x; b)$) can be estimated by approximating the density $f_\beta(s; x, b)$ near s_*. Two cases can occur. If $s_* \neq 0$, then a Gaussian approximation holds and $\mathsf{G}_\beta(x; b) = \Theta(1/\beta)$. On the other hand, if $s_* = 0$, $f_\beta(s; x, b)$ can be approximated by a Laplace distribution, leading to $\mathsf{G}_\beta(x; b) = \Theta(1/\beta^2)$ (which is negligible). We summarize this discussion in the following.

Lemma III.3. *For bounded x, b, we have*

$$\lim_{\beta\to\infty}\mathsf{F}_\beta(x; \beta) = \eta(x; b)\,, \qquad \lim_{\beta\to\infty}\beta\,\mathsf{G}_\beta(x; \beta) = b\,\eta'(x; b)\,.$$

[1] More precisely, given two non-negative functions $p, q : \Omega \to \mathbb{R}$ over the same space, we write $p(s) \cong q(s)$ if there exists a positive constant a such that $p(s) = a\,q(s)$ for every $s \in \Omega$.

Lemmas III.1,III.2, and III.3 suggest the following equivalent form for the message passing algorithm (for large β):

$$x^{t+1}_{i\to a} = \eta\Big(\sum_{b\neq a} A_{bi} z^t_{b\to i}; \hat{\tau}^t\Big)\,, \tag{10}$$

$$z^t_{a\to i} \equiv y_a - \sum_{j\neq i} A_{aj} x^t_{j\to a}, \tag{11}$$

$$\hat{\tau}^{t+1} = \frac{\hat{\tau}^t}{N\delta} \sum_{i=1}^{N} \eta'\Big(\sum_b A_{bi} z^t_{b\to i}; \hat{\tau}^t\Big)\,. \tag{12}$$

D. From message passing to AMP

The updates in Eqs. (11), (12) are easy to implement but nevertheless the overall algorithm is still rather complex because it requires to track $2nN$ messages. The goal of this section is to further simplify the update equations. In order to justify the approximation we assume that the messages can be approximated as $x^t_{i\to a} = x^t_i + \delta x^t_{i\to a} + O(1/N)$, $z^t_{a\to i} = z^t_a + \delta z^t_{a\to i} + O(1/N)$, with $\delta x^t_{i\to a}, \delta z^t_{a\to i} = O(\frac{1}{\sqrt{N}})$ (here the $O(\,\cdot\,)$ errors are assumed uniform in the choice of the edge). We also consider a general message passing algorithms of the form

$$x^{t+1}_{i\to a} = \eta_t\Big(\sum_{b\neq a} A_{bi} z^t_{b\to i}\Big)\,,\ \ z^t_{a\to i} \equiv y_a - \sum_{j\neq i} A_{aj} x^t_{j\to a}, \tag{13}$$

with $\{\eta_t(\,\cdot\,)\}_{t\in\mathbb{N}}$ a sequence of differendiable nonlinear functions with bounded derivatives. Notice that the algorithm derived at the end of the previous section, cf. Eqs. (11), Eqs. (12), is of this form, albeit with η_t non-differentiable at 2 points. But this does not change the result, as long as the nonlinear functions are Lipschitz continuous. In the interest of simplicity, we just discuss the differentiable model.

Lemma III.4. *Suppose that the asymptotic behavior described in the paragraph above holds for the message passing algorithm (13). Then x^t_i and z^t_a satisfy the following equations*

$$x^{t+1}_i = \eta_t\Big(\sum_a A_{ia} z^t_a + x^t_i\Big) + o_N(1),$$

$$z^t_a = y_a - \sum_j A_{aj} x^t_j + \frac{1}{\delta} z^{t-1}_a \langle \eta'_{t-1}(A^* z^{t-1} + x^{t-1})\rangle + o_N(1),$$

where the $o_N(1)$ terms vanish as $N, n \to \infty$.

As a consequence, the resulting algorithm can be written in the vector notation as follows:

$$x^{t+1} = \eta(A^* z^t + x^t; \hat{\tau}^t)\,, \tag{14}$$

$$z^t = y - Ax^t + \frac{1}{\delta} z^{t-1} \langle \eta'(A^* z^{t-1} + x^{t-1}_i; \hat{\tau}^{t-1})\rangle\,, \tag{15}$$

where $\langle\,\cdot\,\rangle$ denotes the average of a vector.

We also get the following recursion for $\hat{\tau}$:

$$\hat{\tau}^t = \frac{\hat{\tau}^{t-1}}{\delta} \langle \eta'(A^* z^{t-1} + x^t; \hat{\tau}^{t-1})\rangle\,. \tag{16}$$

E. Comments

Threshold level. The derivation presented here yields a 'parameter free' algorithm. The threshold level $\hat{\tau}^t$ is updated by the recursion in Eq. (16). One could take the alternative point of view that $\hat{\tau}^t$ is a parameter to be optimized. This point of view was adopted in [1], [5]. It is expected that the two points of view coincide in the large system limit, but it might be advantageous to consider a general sequence of thresholds.

Mathematical derivation of AMP. We showed that in a specific limit (large systems, and large β) the sum-product update rules can be significantly simplified to (14), (15). We should emphasize that our results concern just a single step of the iterative procedure. As such they do not prove that the sum-product messages are carefully tracked by Eqs. (14), (15). In principle it could be that the error terms in our approximation, while negligible at each step, conjure up to become large after a finite number of iterations. We do not expect this to be the case, but it is nevertheless an open mathematical problem.

IV. AMP FOR BPDN/LASSO

Another popular reconstruction procedure in compressed sensing is the following problem

$$\text{minimize } \lambda\|s\|_1 + \frac{1}{2}\|y - As\|_2^2. \tag{17}$$

The derivation of the corressponding AMP is similar to the one in the previous section. We therefore limit ourself to mentioning a few differences.

As before we define a joint density distribution on the variables $s = (s_1, \dots, s_N)$

$$\mu(\mathrm{d}s) = \frac{1}{Z}\prod_{i=1}^{N} \exp(-\beta\lambda|s_i|) \prod_{a=1}^{n} \exp\Big\{-\frac{\beta}{2}(y_a - (As)_a)^2\Big\}\, \mathrm{d}s\,.$$

The mode of this distribution coincides with the solution of the problem (17) and the distribution concentrates on its mode as $\beta \to \infty$. The sum-product algorithm is

$$\nu^{t+1}_{i\to a}(s_i) \cong \exp(-\beta\lambda|s_i|) \prod_{b\neq a} \nu^t_{b\to i}(s_i),$$

$$\hat{\nu}^t_{a\to i}(s_i) \cong \int \exp\Big\{-\frac{\beta}{2}(y_a - (As)_a)^2\Big\} \prod_{j\neq i} \mathrm{d}\nu^t_{j\to a}(s_j)\,.$$

Proceeding as above, we derive an asymptotically equivalent form of the belief propagation for $N \to \infty$ and $\beta \to \infty$. We get the following algorithm in the vector notation:

$$x^t = \eta(x^t + A^* z^t; \lambda + \gamma^t)\,, \tag{18}$$

$$z^{t+1} = y - Ax^t + \frac{1}{\delta} z^t \langle \eta'(x^{t-1} + A^* z^{t-1}),\rangle \tag{19}$$

which generalize Eqs. (14) and (15). The threshold level is computed iteratively as follows

$$\gamma^{t+1} = \frac{\lambda + \gamma^t}{\delta} \langle \eta'(Az^t + x^t; \gamma^t + \lambda)\rangle\,. \tag{20}$$

Notice that the only deviation from the algorithm in the previous section is in the recursion for the threshold level.

V. AMP FOR RECONSTRUCTION WITH PRIOR INFORMATION

In the two cases we discussed so far, the distribution of the signal s_o was not known. This is a very natural and practical assumption. Nevertheless, it might be possible in specific scenarios to estimate the input distribution. This extra information may be used to improve the recovery algorithms. Also, the case of known signal distribution provides a benchmark for the other approaches. In this section we define a very simple iterative theresholding algorithm for these situations.

Let $\alpha = \alpha_1 \times \alpha_2 \cdots \times \alpha_N$ be the joint probability distribution of the variables $s_1, s_2, \dots, s_N$. It is then natural to consider the distribution

$$\mu(\mathrm{d}s) = \frac{1}{Z} \prod_{a=1}^{n} \exp\Big\{ -\frac{\beta}{2}(y_a - (As)_a)^2 \Big\} \prod_{i=1}^{N} \alpha_i(\mathrm{d}s_i)\,,$$

since μ is the *a posteriori* distribution of s, when $y = As + w$ is observed. Here, w is a noise vector with i.i.d. normal entries and is independent of s. The sum-product update rules are

$$\nu^{t+1}_{i\to a}(\mathrm{d}s_i) \cong \prod_{b\neq a} \hat{\nu}^{t}_{b\to i}(s_i)\, \alpha_i(\mathrm{d}s_i)\,,$$

$$\nu^{t}_{a\to i}(s_i) \cong \int \exp\Big\{ -\frac{\beta}{2}(y_a - (As)_a)^2 \Big\} \prod_{j\neq i} \nu^{t}_{j\to a}(\mathrm{d}s_j)\,.$$

Notice that the above update rules are well defined. At each iteration t, the message $\nu^{t+1}_{i\to a}(\mathrm{d}s_i)$ is a probability measure on $\mathbb{R}$, and the first equation gives its density with respect to α_i. The message $\nu^{t}_{a\to i}(s_i)$ is instead a non-negative measurable function (equivalently, a density) given by the second equation. Clearly the case studied in the previous section corresponds to $\alpha_i \cong \exp(-\beta|s_i|)$.

In order to derive the simplified version of the message passing algorithm, we introduce the following family of measures over $\mathbb{R}$

$$f_i(\mathrm{d}s; x, b) \equiv \frac{1}{z_\beta(x,b)}\, \exp\Big\{ -\frac{\beta}{2b}(s-x)^2 \Big\}\, \alpha_i(\mathrm{d}s)\,, \quad (21)$$

indexed by $i \in [N]$, $x \in \mathbb{R}$, $b \in \mathbb{R}_+$ (β is fixed here). The mean and the variance of this distribution define the functions (here $Z \sim f_i(\,\cdot\,; x, b)$)

$$\mathsf{F}_i(x;b) \equiv \mathbb{E}_{f_i(\,\cdot\,;x,b)}(Z)\,, \quad \mathsf{G}_i(x;b) \equiv \mathrm{Var}_{f_i(\,\cdot\,;x,b)}(Z)\,. \quad (22)$$

These functions have a natural estimation-theoretic interpretation. Let X_i be a random variable with distribution α_i, and assume that $\widetilde{Y}_i = X_i + W_i$ is observed with W_i gaussian noise with variance b/β. The above functions are –respectively– the conditional expectation and conditional variance of X_i, given that $\widetilde{Y}_i = x$:

$$\mathsf{F}_i(x;b) = \mathbb{E}(X_i|\widetilde{Y}_i = x)\,, \quad \mathsf{G}_i(x;b) = \mathrm{Var}(X_i|\widetilde{Y} = x)\,.$$

The approach described in Section III yields the following AMP (in vector notation)

$$x^t = \mathsf{F}(x^t + A^* z^t; \lambda + \gamma^t), \quad (23)$$

$$z^{t+1} = y - Ax^t + \frac{1}{\delta} z^t \langle \mathsf{F}'(x^{t-1} + A^* z^{t-1}) \rangle\,. \quad (24)$$

Here, if $x \in \mathbb{R}^N$, $\mathsf{F}(x;b) \in \mathbb{R}^N$ is the vector $\mathsf{F}(x;b) = (\mathsf{F}_1(x_i;b), \mathsf{F}_2(x_2;b), \dots, \mathsf{F}_N(x_N;b))$. Analogously $\mathsf{F}'(x) = (\mathsf{F}'_1(x_i;b), \mathsf{F}'_2(x_2;b), \dots, \mathsf{F}'_N(x_N;b))$ (derivative being taken with respect to the first argument). Finally, the threshold level is computed iteratively as follows

$$\gamma^{t+1} = \frac{1}{\delta} \langle \mathsf{G}(Az^t + x^t; \gamma^t + \lambda) \rangle\,. \quad (25)$$

A. Comments

The AMP algorithm described in this section is marginally more complex than the ones in the previous sections. The main difference is that the soft thresholding function $\eta(\,\cdot\,)$ is replaced with the conditional expectation $\mathsf{F}(\,\cdot\,)$. While the latter does not admit, in general, a closed form expression, it is not hard to construct accurate approximations that are easy to evaluate.

VI. RELATED WORK

In this section we would like to clarify the relation of the present approach with earlier results in the literature. Each of these lines of work evolved from different motivations, and there was so far little – if any – contact between them.

A. Other message passing algorithms

The use of message passing algorithms for compressed sensing problems was suggested before, see for instance [6]. However such a proposal faced two major difficulties.

(1) According to the standard prescription, messages used in the the sum-product algorithm should be probability measures over the real line $\mathbb{R}$, cf. Eqs. (3), (4). This is impractical from a computational point of view. (A low complexity message-passing algorithm for a related problem was used in [7]).

(2) The factor graph on which the sum-product algorithm is run is the complete bipartite graph with N variable nodes, and n function nodes. In other words, unless the underlying matrix is sparse [8], the graphical model is very dense. This requires to update Nn messages per iteration, and each message update depend on N or n input messages. Again this is very expensive computationally.

The previous pages show that problem (2) does not add to (1), but in fact solves it! Indeed, the high density of the graph leads to approximately Gaussian messages from factor nodes to variable nodes, via central limit theorem. Gaussian messages are in turn parametrized by two numbers: mean and variance. It turns out that is is sufficient to keep track only of the means, again because of the high density.

Problem (2) is also solved by the high density nature of the graph, since all the messages departing from the same node of the graph are very similar with each other.

One last key difficulty with the use of belief propagation in compressed sensing was

(3) The use of belief propagation requires to define a prior on the vector s_o. For most applications, no good prior is available.

The solution of this problem lies in using a Laplace prior as in Eq. (2). A first justification of this choice lies in the fact that, as $\beta \to \infty$, the resulting probability measure concentrates around its mode, that is the solution of the basis pursuit problem (1). A deeper reason for this choice is that it is intimately related to the soft threshold non-linearity $\eta(x;\theta)$, which is step-by-step optimal in a minimax sense [1], [5].

B. Historical background and statistical physics

There is a well studied connection between statistical physics techniques and message passing algorithms [9]. In particular, the sum-product algorithm corresponds to the Bethe-Peierls approximation in statistical physics, and its fixed points are stationary points of the Bethe free energy. In the context of spin glass theory, the Bethe-Peierls approximation is also referred to as the 'replica symmetric cavity method'.

The Bethe-Peierls approximation postulates a set of non-linear equations on quantities that are equivalent to the sum-product messages, and which are in correspondence with local marginals. In the special cases of spin glasses on the complete graph (the celebrated Sherrington-Kirkpatrick model), these equations reduce to the so-called TAP equations, named after Thouless, Anderson and Palmer who first used them [10].

The original TAP equations where a set of non-linear equations for local magnetizations (i.e. expectations of a single variable). Thouless, Anderson and Palmer first recognized that naive mean field is not accurate enough in the spin glass model, and corrected it by adding the so called Onsager reaction term that is analogous to the last term in Eq. (15). More than 30 years after the original paper, a complete mathematical justification of the TAP equations remains an open problem in spin glass theory, although important partial results exist [11]. While the connection between belief propagation and Bethe-Peierls approximation stimulated a considerable amount of research [12], the algorithmic uses of TAP equations have received only sparse attention. Remarkable exceptions include [13], [14], [15].

C. State evolution and replica calculations

In the context of coding theory, message passing algorithms are analyzed through density evolution [16]. The common justification for density evolution is that the underlying graph is random and sparce, and hence converges locally to a tree in the large system limit. In the case of trees density evolution is exact, hence it is asymptotically exact for sparse random graphs.

State evolution is the analog of density evolution in the case of dense graphs. For definitions and results on state evolution we refer to the [1], [5]. The success of state evolution cannot be ascribed to the locally tree-like structure of the graph, and calls for new mathematical ideas.

The fixed points of state evolution describe the output of the corresponding AMP, when the latter is run for a sufficiently large number of iterations (independent of the dimensions n, N). It is well known, within statistical mechanics [9], that the fixed point equations do indeed coincide with the equations obtained through a completely different non-rigorous approach, the *replica method* (in its replica-symmetric form). This is indeed an instance of a more general equivalence between replica and cavity methods.

During the last few months, several papers investigated compressed sensing problems using the replica method [17], [18], [19]. In view of the discussion above, it is not surprising that these results can be recovered from the state evolution formalism put forward in [1]. Let us mention that the latter has several advantages over the replica method: (1) It is more concrete, and its assumptions can be checked quantitatively through simulations; (2) It is intimately related to efficient message passing algorithms; (3) It actually allows to predict the performances of these algorithms.

Acknowledgment

This work was partially supported by a Terman fellowship, the NSF CAREER award CCF-0743978 and the NSF grant DMS-0806211.

References

[1] D. L. Donoho, A. Maleki and A. Montanari, "Message Passing Algorithms for Compressed Sensing," Proc. Natl. Acad. Sci. (2009),

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," SIAM J. on Scientific Comput., 20 (1998) 33-61

[3] A. Maleki and D. L. Donoho, "Optimally Tuned Iterative Reconstruction Algorithms for Compressed Sensing," submitted to IEEE journal of selected areas in signal processing, accepted for publication, 2009.

[4] D. L. Donoho, A. Maleki and A. Montanari, "How to Design Message Passing Algorithms for Compressed Sensing," *In preparation* (2010)

[5] D. L. Donoho, A. Maleki and A. Montanari, "Message Passing Algorithms for Compressed Sensing: II. Analysis and Validation," IEEE Inform. Theory Workshop, Cairo, January 2010

[6] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian Compressive Sensing via Belief Propagation," accepted to IEEE Transactions on Signal Processing (2009)

[7] Y. Lu, A. Montanari, B. Prabhakar, S. Dharmapurikar and A. Kabbani, "Counter Braids: A Novel Counter Architecture for Per-Flow Measurement", SIGMETRICS, Annapolis, June 2008

[8] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff and M. J. Strauss, "Combining geometry with combinatorics: a unified approach to sparse signal recovery," Proc. of the 46th Annual Allerton Conference, Monticello, IL, September 2008

[9] M. Mézard and A. Montanari, *Information, physics, and computation,* Oxford University Press, Oxford, 2009

[10] D. J. Thouless and P. W. Anderson and R. G. Palmer", "Solution of 'Solvable model of a spin glass'," Phil. Mag., 35 (1977) 593-601

[11] M. Talagrand, *Spin Glasses: A Challenge for Mathematicians*, Springer-Verlag, Berlin, 2003

[12] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," IEEE Trans. Inf. Theory, 51 (2005) 2282–2313.

[13] M. Opper and O. Winther, "From Naive Mean Field Theory to the TAP Equations," in *Advanced mean field methods: theory and practice*, MIT Press, 2001

[14] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," J. Phys. A, 36 (2003) 11111-11121

[15] J. P. Neirotti and D. Saad, "Improved message passing for inference in densely connected systems", Europhys. Lett. 71 (2005) 866-872

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

[17] S. Rankan, A. K. Fletcher, and V. K. Goyal "Asymptotic Analysis of MAP Estimation via the Replica Method and Applications to Compressed Sensing", arXiv:0906.3234 (2009)

[18] Y. Kabashima, T. Wadayama, and T. Tanaka "A typical reconstruction limit for compressed sensing based on Lp-norm minimization," J.Stat. Mech. (2009) L09003

[19] D. Baron, D. Guo, and S. Shamai, ""A Single-letter Characterization of Optimal Noisy Compressed Sensing", Proc. of the 47th Annual Allerton Conference, Monticello, IL, September 2009