Information-Theoretically Optimal Compressed Sensing via Spatial Coupling and Approximate Message Passing

David L. Donoho
Department of Statistics
Stanford University

Adel Javanmard
Department of Electrical Engineering
Stanford University

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

***Abstract*—We study the compressed sensing reconstruction problem for a broad class of random, band-diagonal sensing matrices. This construction is inspired by the idea of spatial coupling in coding theory. As demonstrated heuristically and numerically by Krzakala et al. [11], message passing algorithms can effectively solve the reconstruction problem for spatially coupled measurements with undersampling rates close to the fraction of non-zero coordinates.**

We use an approximate message passing (AMP) algorithm and analyze it through the state evolution method. We give a rigorous proof that this approach is successful as soon as the undersampling rate δ exceeds the (upper) Rényi information dimension of the signal, $\overline{d}(p_X)$. More precisely, for a sequence of signals of diverging dimension n whose empirical distribution converges to p_X, reconstruction is with high probability successful from $\overline{d}(p_X)\, n+o(n)$ measurements taken according to a band diagonal matrix.

For sparse signals, i.e. sequences of dimension n and $k(n)$ non-zero entries, this implies reconstruction from $k(n)+o(n)$ measurements. For 'discrete' signals, i.e. signals whose coordinates take a fixed finite set of values, this implies reconstruction from $o(n)$ measurements. The result is robust with respect to noise, does not apply uniquely to random signals, but requires the knowledge of the empirical distribution of the signal p_X.

I. Introduction

A. Background and contributions

Assume that m linear measurements are taken of an unknown n-dimensional signal $x \in \mathbb{R}^n$, according to the model

$$y = Ax\,. \tag{1}$$

The reconstruction problem requires to reconstruct x from the measured vector $y \in \mathbb{R}^m$, and the sensing matrix $A \in \mathbb{R}^{m\times n}$.

It is an elementary fact of linear algebra that the reconstruction problem will not have a unique solution unless $m \geq n$. This observation is however challenged within compressed sensing. A large corpus of research shows that, under the assumption that x is sparse, a dramatically smaller number of measurements is sufficient [6], [2]. Namely, if only k entries of x are non-vanishing, then roughly $m \gtrsim 2k\log(n/k)$ measurements are sufficient for A random, and reconstruction can be solved efficiently by convex programming. Deterministic sensing matrices achieve similar performances, provided they satisfy a suitable restricted isometry condition [4]. On top of this, reconstruction is robust with respect to the addition of noise [3], [5], i.e. under the model

$$y = Ax + w\,, \tag{2}$$

with –say– $w \in \mathbb{R}^m$ a random vector with i.i.d. components $w_i \sim \mathsf{N}(0,\sigma^2)$. In this context, the notions of 'robustness' or 'stability' refers to the existence of universal constants C such that the per-coordinate mean square error in reconstructing x from noisy observation y is upper bounded by $C\,\sigma^2$.

From an information-theoretic point of view it remains however unclear why we cannot achieve the same goal with far fewer than $2\,k\log(n/k)$ measurements. Indeed, we can interpret Eq. (1) as describing an analog data compression process, with y a compressed version of x. From this point of view, we can encode all the information about x in a single real number $y \in \mathbb{R}$ (i.e. use $m = 1$), because the cardinality of $\mathbb{R}$ is the same as the one of $\mathbb{R}^n$. Motivated by this puzzling remark, Wu and Verdú [15] introduced a Shannon-theoretic analogue of compressed sensing, whereby the vector x has i.i.d. components $x_i \sim p_X$. Crucially, the distribution p_X is available to, and may be used by the reconstruction algorithm. Under the mild assumptions that sensing is linear (as per Eq. (1)), and that the reconstruction mapping is Lipschitz continuous, they proved that compression is asymptotically lossless if and only if

$$m \geq n\,\overline{d}(p_X) + o(n)\,. \tag{3}$$

Here $\overline{d}(p_X)$ is the (upper) Rényi information dimension of the distribution p_X. We refer to Section II for a definition of this quantity. Suffices to say that, if p_X is ε-sparse (i.e. if it puts mass at most ε on nonzeros) then $\overline{d}(p_X) \leq \varepsilon$. Also, if p_X is the convex combination of a discrete part (sum of Dirac's delta) and an absolutely continuous part (with a density), then $\overline{d}(p_X)$ is equal to the weight of the absolutely continuous part.

This result is quite striking. For instance, it implies that, for random k-sparse vectors, $m \geq k + o(n)$ measurements are sufficient. Also, if the entries of x are random and take values in –say– $\{-10,-9,\dots,-9,+10\}$, then a sublinear number of measurements $m = o(n)$, is sufficient! At the same time, the result of Wu and Verdú presents two important limitations. First, it does not provide robustness guarantees of the type described above. It therefore leaves open the possibility that reconstruction is highly sensitive to noise when m is significantly smaller than the number of measurements required in classical compressed sensing, namely $\Theta(k\log(n/k))$ for k-sparse vectors. Second, it does not provide any computationally practical algorithms for reconstructing x from measurements y.

In an independent line of work, Krzakala et al. [11] de-

veloped an approach that leverages on the idea of *spatial coupling*. This idea was introduced for the compressed sensing literature by Kudekar and Pfister [12]. Spatially coupled matrices are –roughly speaking– random sensing matrices with a band-diagonal structure. The analogy is, this time, with channel coding. In this context, spatial coupling, in conjunction with message-passing decoding, allows to achieve Shannon capacity on memoryless communication channels. By analogy, it is reasonable to hope that a similar approach might enable to sense random vectors x at an undersampling rate m/n close to the Rényi information dimension of the coordinates of x, $\overline{d}(p_X)$. Indeed, the authors of [11] evaluate this approach numerically on a few classes of random vectors and demonstrate that it indeed achieves rates close to the fraction of non-zero entries. They also support this claim by insightful statistical physics arguments.

Finally, let us mention that robust sparse recovery of k-sparse vectors from $m = O(k \log\log(n/k))$ measurement is possible, using suitable 'adaptive' sensing schemes [10].

B. Our Contribution

In this paper, we fill the gap between the above works, and present the following contributions:

• **Construction.** We describe a construction for spatially coupled sensing matrices A that is somewhat broader than the one of [11] and give precise prescriptions for the asymptotics of various parameters. We also use a somewhat different reconstruction algorithm from the one in [11], by building on the approximate message passing (AMP) approach of [8], [9]. AMP algorithms have the advantage of smaller memory complexity with respect to standard message passing, and of smaller computational complexity whenever fast multiplication procedures are available for A.

• **Rigorous proof of convergence.** Our main contribution is a rigorous proof that the above approach indeed achieves the information-theoretic limits set out by Wu and Verdú [15]. Indeed, we prove that, for sequences of spatially coupled sensing matrices $\{A(n)\}_{n\in\mathbb{N}}$, $A(n) \in \mathbb{R}^{m(n)\times n}$ with asymptotic undersampling rate $\delta = \lim_{n\to\infty} m(n)/n$, AMP reconstruction is with high probability successful in recovering the signal x, provided $\delta > \overline{d}(p_X)$.

• **Robustness to noise.** We prove that the present approach is robust[1] to noise in the following sense. For any signal distribution p_X and undersampling rate δ, there exists a constant C such that the output $\widehat{x}(y)$ of the reconstruction algorithm achieves a mean square error per coordinate $n^{-1}\mathbb{E}\{\|\widehat{x}(y) - x\|_2^2\} \le C\,\sigma^2$. This result holds under the noisy measurement model (2) for a broad class of noise models for w, including i.i.d. noise coordinates w_i with $\mathbb{E}\{w_i^2\} = \sigma^2 < \infty$.

• **Non-random signals.** Our proof does not apply uniquely to random signals x with i.i.d. components, but indeed to more general sequences of signals $\{x(n)\}_{n\in\mathbb{N}}$, $x(n) \in \mathbb{R}^n$ indexed by their dimension n. The conditions required are: (1) that the empirical distribution of the coordinates of $x(n)$ converges (weakly) to p_X; and (2) that $\|x(n)\|_2^2$ converges to the second moment of the asymptotic law p_X. Interestingly, the present framework changes the notion of 'structure' that is relevant for reconstructing the signal x. Indeed, the focus is shifted from the *sparsity* of x to the *information dimension* $\overline{d}(p_X)$.

[1] This robustness bound holds for all $\delta > \overline{D}(p_X)$, where $\overline{D}(p_X) = \overline{d}(p_X)$ for a broad class of distributions p_X (including distributions without *singular continuous component*). When $\overline{d}(p_X) < \overline{D}(p_X)$, a somewhat weaker robustness bound holds for $\overline{d}(p_X) < \delta \le \overline{D}(p_X)$.

C. Organization

In the next section we state formally our results, and discuss their implications as well as the basic intuition behind them. Section III provides a precise description of the matrix construction and the reconstruction algorithm. Due to space limitations, the proofs of the theorems are removed from this version of the paper and can be found in [7].

II. Formal statement of the results

We consider the noisy model (2). An instance of the problem is therefore completely specified by the triple (x, w, A). We will be interested in the asymptotic properties of sequence of instances indexed by the problem dimensions $\mathcal{S} = \{(x(n), w(n), A(n))\}_{n\in\mathbb{N}}$. We recall a definition from [1]. (More precisely, [1] introduces the $B = 1$ case of this definition.)

Definition II.1. *The sequence of instances $\mathcal{S} = \{x(n), w(n), A(n)\}_{n\in\mathbb{N}}$ indexed by n is said to be a* B-converging sequence *if $x(n) \in \mathbb{R}^n$, $w(n) \in \mathbb{R}^m$, $A(n) \in \mathbb{R}^{m\times n}$ with $m = m(n)$ is such that $m/n \to \delta \in (0, \infty)$, and in addition the following conditions hold:*

(a) *The empirical distribution of the entries of $x(n)$ converges weakly to a probability measure p_X on $\mathbb{R}$ with bounded second moment. Further $n^{-1}\sum_{i=1}^n x_i(n)^2 \to \mathbb{E}_{p_X}\{X^2\}$.*

(b) *The empirical distribution of the entries of $w(n)$ converges weakly to a probability measure p_W on $\mathbb{R}$ with bounded second moment. Further $m^{-1}\sum_{i=1}^m w_i(n)^2 \to \mathbb{E}_{p_W}\{W^2\} \equiv \sigma^2$.*

(c) *If $\{e_i\}_{1\le i\le n}$, $e_i \in \mathbb{R}^n$ denotes the canonical basis, then*
$$\limsup_{n\to\infty} \max_{i\in[n]} \|A(n)e_i\|_2 \le B,$$
$$\liminf_{n\to\infty} \min_{i\in[n]} \|A(n)e_i\|_2 \ge 1/B.$$

We further say that $\mathcal{S}$ is a converging sequence *if it is B-converging for some B. We say that $\{A(n)\}_{n\ge 0}$ is a converging sequence of sensing matrices if they satisfy condition (c) above for some B.*

Finally, if the sequence $\{(x(n), w(n), A(n))\}_{n\ge 0}$ is random, the above conditions are required to hold almost surely.

Given a sensing matrix A, and a vector of measurements y, a reconstruction algorithm produces an estimate $\widehat{x}(A; y) \in \mathbb{R}^n$ of x. In this paper we assume that the empirical distribution p_X, and the noise level σ^2 are known to the estimator, and hence the mapping $\widehat{x} : (A, y) \mapsto \widehat{x}(A; y)$ implicitly depends on p_X and σ^2. Since however p_X, σ^2 are fixed throughout, we avoid the cumbersome notation $\widehat{x}(A, y, p_X, \sigma^2)$.

Given a converging sequence of instances $\mathcal{S} = \{x(n), w(n), A(n)\}_{n\in\mathbb{N}}$, and an estimator $\widehat{x}$, we define the asymptotic per-coordinate reconstruction mean square error as

$$\mathsf{MSE}(\mathcal{S};\widehat{x}) = \limsup_{n\to\infty} \frac{1}{n} \left\|\widehat{x}\big(A(n); y(n)\big) - x(n)\right\|^2 . \qquad (4)$$

Notice that the quantity on the right hand side depends on the matrix $A(n)$, which will be random, and on the signal and noise vectors $x(n)$, $w(n)$ which can themselves be random. Our results hold almost surely with respect to these random variables.

In this paper we study a specific low-complexity estimator, based on the AMP algorithm first proposed in [8]. This proceed by the following iteration (initialized with $x_i^1 = \mathbb{E}_{p_X} X$ for all $i \in [n]$).

$$x^{t+1} = \eta_t(x^t + (Q^t \odot A)^* r^t) , \qquad (5)$$

$$r^t = y - Ax^t + \mathsf{b}_t \odot r^{t-1} . \qquad (6)$$

Here, for each t, $\eta_t : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable nonlinear function that depends on the input distribution p_X. Further, η_t is separable, namely, for a vector $v \in \mathbb{R}^n$, we have $\eta_t(v) = (\eta_{1,t}(v_1), \dots, \eta_{n,t}(v_n))$. The matrix $Q^t \in \mathbb{R}^{m\times n}$ and the vector $\mathsf{b}_t \in \mathbb{R}^m$ can be efficiently computed from the current state x^t of the algorithm, $\odot$ indicates Hadamard (entrywise) product and X^* denotes the transpose of matrix X. Further Q^t does not depend on the problem instance and hence can be precomputed. Both Q^t and b_t are block-constants. This property makes their evaluation, storage and manipulation particularly convenient. We refer to the next section for explicit definitions of these quantities. In particular, the specific choice of $\eta_{i,t}$ is dictated by the objective of minimizing the mean square error at iteration $t+1$, and hence takes the form of a Bayes optimal estimator for the prior p_X. In order to stress this point, we will occasionally refer to this as to the Bayes optimal AMP algorithm.

We denote by $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ the mean square error achieved by the Bayes optimal AMP algorithm, where we made explicit the dependence on σ^2. Since the AMP estimate depends on the iteration number t, the definition of $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ requires some care. The basic point is that we need to iterate the algorithm only for a constant number of iterations, as n gets large. Formally, we let

$$\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2) \equiv \lim_{t\to\infty} \limsup_{n\to\infty} \frac{1}{n} \left\|x^t\big(A(n); y(n)\big) - x(n)\right\|^2 .$$

As discussed above, limits will be shown to exist almost surely, when the instances $(x(n), w(n), A(n))$ are random, and almost sure upper bounds on $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ will be proved. (Indeed $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ turns out to be deterministic.)

We will tie the success of our compressed sensing scheme to the fundamental information-theoretic limit established in [15]. The latter is expressed in terms of the (upper) Rényi information dimension of the probability measure p_X [13], denoted by $\overline{d}(p_X)$. Further, our 'stability' result is expressed in terms of the (upper) MMSE dimension of the probability measure p_X [16], denoted by $\overline{D}(p_X)$. It is convenient to recall the following result in this regard.

Proposition II.2 ([13], [16])**.** *Consider the Lebesgue's decomposition of probability p_X as $p_X = (1-\alpha)\nu_{\rm d} + \alpha_1\nu_{\rm ac} + \alpha_2\nu_{\rm sc}$, where $\nu_{\rm d}$ is a discrete distribution, $\nu_{\rm ac}$ is an absolutely continuos measure with respect to Lebesgue measure, and $\nu_{\rm sc}$ is a singular continuos measure with respect to Lebesgue measure. Then, $\overline{d}(p_X) \le \overline{D}(p_X) \le \alpha = \alpha_1 + \alpha_2$, and if $\alpha_2 = 0$, then $\overline{d}(p_X) = \overline{D}(p_X) = \alpha$.*

We are now in position to state our main results.

Theorem II.3. *Let p_X be a probability measure on the real line and assume $\delta > \overline{d}(p_X)$. Then there exists a random converging sequence of sensing matrices $\{A(n)\}_{n\ge 0}$, $A(n) \in \mathbb{R}^{m\times n}$, $m(n)/n \to \delta$, for which the following holds. For any $\varepsilon > 0$, there exists σ_0 such that for any converging sequence of instances $\{(x(n), w(n))\}_{n\ge 0}$ with parameters (p_X, σ^2, δ) and $\sigma \in [0, \sigma_0]$, we have, almost surely*

$$\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2) \le \varepsilon . \qquad (7)$$

Theorem II.4. *Let p_X be a probability measure on the real line and assume $\delta > \overline{D}(p_X)$. Then there exists a random converging sequence of sensing matrices $\{A(n)\}_{n\ge 0}$, $A(n) \in \mathbb{R}^{m\times n}$, $m(n)/n \to \delta$ and a finite stability constant $C = C(p_X, \delta)$, such that the following is true. For any converging sequence of instances $\{(x(n), w(n))\}_{n\ge 0}$ with parameters (p_X, σ^2, δ), we have, almost surely*

$$\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2) \le C\,\sigma^2 . \qquad (8)$$

Notice that, by Proposition II.2, $\overline{D}(p_X) \ge \overline{d}(p_X)$, and $\overline{D}(p_X) = \overline{d}(p_X)$ for a broad class of probability measures p_X. Further, the noiseless model (1) is covered as a special case of Theorem II.3 by taking $\sigma^2 \downarrow 0$.

A. Discussion

It is instructive to spell out in detail a specific example.

Example (Mixture signal with a point mass). Consider a mixture distribution of the form

$$p_X = (1-\alpha)\,\delta_0 + \alpha\, q , \qquad (9)$$

where q is a measure that is absolutely continuous with respect to Lebesgue measure, i.e. $q(\mathrm{d}x) = f(x)\,\mathrm{d}x$ for some measurable function f. Then, by Proposition II.2, we have $\overline{d}(p_X) = \overline{D}(p_X) = \alpha$. Now let $\{x(n)\}_{n\ge 0}$ be a sequence of vectors with i.i.d. components $x(n)_i \sim p_X$. Denote by $k(n)$ the number of nonzero entries in $x(n)$. Then, almost surely as $n \to \infty$, Bayes optimal AMP recovers the signal $x(n)$ from $m(n) = k(n) + o(n)$ spatially coupled measurements.

Under the regularity hypotheses of [15], no scheme can do substantially better, i.e. reconstruct $x(n)$ from $m(n)$ measurements if $\limsup_{n\to\infty} m(n)/k(n) < 1$.

One way to think about this result is the following. If an oracle gave us the support of $x(n)$, we would still need $m(n) \ge k(n) - o(n)$ measurements to reconstruct the signal. Indeed, the entries in the support have distribution q, and

$\overline{d}(q) = 1$. Theorem II.3 implies that the measurements overhead for estimating the support of $x(n)$ is sublinear, $o(n)$, even when the support is of order n.

In the next section we describe the basic intuition behind the surprising phenomenon in Theorems II.3 and II.4, and why are spatially-coupled sensing matrices so useful.

B. How does spatial coupling work?

Spatially-coupled sensing matrices A are –roughly speaking– band diagonal matrices. It is convenient to think of the graph structure that they induce on the reconstruction problem. Associate one node (a *variable node* in the language of factor graphs) to each coordinate i in the unknown signal x. Order these nodes on the real line $\mathbb{R}$, putting the i-th node at location $i \in \mathbb{R}$. Analogously, associate a node (a *factor node*) to each coordinate a in the measurement vector y, and place the node a at position a/δ on the same line. Connect this node to all the variable nodes i such that $A_{ai} \neq 0$. If A is band diagonal, only nodes that are placed close enough will be connected by an edge. See Figure 1 for an illustration.

In a spatially coupled matrix, additional measurements are associated to the first few coordinates of x, say coordinates $x_1, \dots, x_{n_0}$ with n_0 much smaller than n. This has a negligible impact on the overall undersampling ratio as $n/n_0 \to \infty$. Although the overall undersampling remains $\delta < 1$, the coordinates $x_1, \dots, x_{n_0}$ are oversampled. This ensures that these first coordinates are recovered correctly (up to a mean square error of order σ^2). As the algorithm is iterated, the contribution of these first few coordinates is correctly subtracted from all the measurements, and hence we can effectively eliminate those nodes from the graph. In the resulting graph, the first few variables are effectively oversampled and hence the algorithm will reconstruct their values, up to a mean square error of order σ^2. As the process is iterated, variables are progressively reconstructed, proceeding from left to right along the node layout.

While the above explains the basic dynamics of AMP reconstruction algorithms under spatial coupling, a careful consideration reveals that this picture leaves open several challenging questions. In particular, why does the overall undersampling factor δ have to exceed $\overline{d}(p_X)$ for reconstruction to be successful? Our proof is based on a potential function argument. We will prove that there exists a potential function for the AMP algorithm, such that, when $\delta > \overline{d}(p_X)$, this function has its global minimum close to exact reconstruction. Further, we will prove that, unless this minimum is essentially achieved, AMP can always decrease the function.

III. Matrix and algorithm construction

In this section, we define an ensemble of random matrices, and the corresponding choices of Q^t, b_t, η_t that achieve the reconstruction guarantees in Theorems II.3 and II.4.

A. General matrix ensemble

The sensing matrix A will be constructed randomly, from an ensemble denoted by $\mathcal{M}(W, M, N)$. The ensemble depends on two integers $M, N \in \mathbb{N}$, and on a matrix with non-negative entries $W \in \mathbb{R}_+^{\mathsf{R}\times\mathsf{C}}$, whose rows and columns are indexed by the finite sets R, C (respectively 'rows' and 'columns'). The matrix is *roughly row-stochastic*, i.e.

$$\frac{1}{2} \le \sum_{c\in\mathsf{C}} W_{r,c} \le 2\,, \qquad \text{for all } r \in \mathsf{R}\,. \tag{10}$$

We will let $|\mathsf{R}| \equiv L_r$ and $|\mathsf{C}| \equiv L_c$ denote the matrix dimensions. The ensemble parameters are related to the sensing matrix dimensions by $n = NL_c$ and $m = ML_r$.

In order to describe a random matrix $A \sim \mathcal{M}(W, M, N)$ from this ensemble, partition the columns and rows indices in, respectively, L_c and L_r groups of equal size. Explicitly

$$[n] = \cup_{s\in\mathsf{C}} C(s)\,, \quad |C(s)| = N\,,$$
$$[m] = \cup_{r\in\mathsf{R}} R(r)\,, \quad |R(r)| = M\,.$$

Here and below we use $[k]$ to denote the set of first k integers $[k] \equiv \{1, 2, \dots, k\}$. Further, if $i \in R(r)$ or $j \in C(s)$ we will write, respectively, $r = \mathsf{g}(i)$ or $s = \mathsf{g}(j)$. In other words $\mathsf{g}(\,\cdot\,)$ is the operator determining the group index of a given row or column.

With this notation we have the following concise definition of the ensemble.

Definition III.1. *A random sensing matrix A is distributed according to the ensemble $\mathcal{M}(W, M, N)$ (and we write $A \sim \mathcal{M}(W, M, N)$) if the entries $\{A_{ij},\ \ i \in [m], j \in [n]\}$ are independent Gaussian random variables with*

$$A_{ij} \sim \mathsf{N}\Big(0, \frac{1}{M}\, W_{\mathsf{g}(i),\mathsf{g}(j)}\Big)\,. \tag{11}$$

B. State evolution

State evolution allows an exact asymptotic analysis of AMP algorithms in the limit of a large number of dimensions. As indicated by the name, it bears close resemblance to the density evolution method in iterative coding theory [14]. Somewhat surprisingly, this analysis approach is asymptotically exact despite the underlying factor graph being far from locally tree-like.

State evolution recursion is used in defining the parameters Q^t, b_t, η_t and also plays a crucial role in the algorithm analysis [7]. In the present case, state evolution takes the following form.

Definition III.2. *Given $W \in \mathbb{R}_+^{L_r\times L_c}$ roughly row-stochastic, the corresponding* state evolution sequence *is the sequence of*

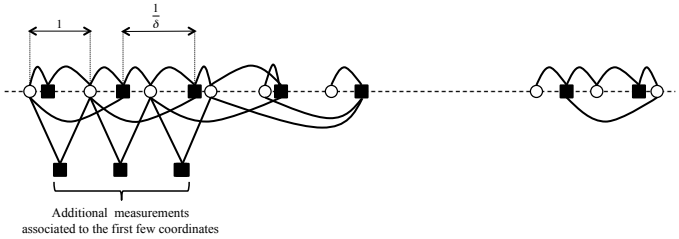

Fig. 1. Graph structure of a spatially coupled matrix. Variable nodes are shown as circle and check nodes are represented by square.

vectors $\{\phi(t), \psi(t)\}_{t\geq 0}$, $\phi(t) = (\phi_a(t))_{a\in \mathsf{R}} \in \mathbb{R}_+^{\mathsf{R}}$, $\psi(t) = (\psi_i(t))_{i\in \mathsf{C}} \in \mathbb{R}_+^{\mathsf{C}}$, *defined recursively by*

$$\begin{aligned} \phi_a(t) &= \sigma^2 + \frac{1}{\delta}\sum_{i\in \mathsf{C}} W_{a,i}\, \psi_i(t)\,, \\ \psi_i(t+1) &= \mathsf{mmse}\Big(\sum_{b\in \mathsf{R}} W_{b,i}\phi_b^{-1}(t)\Big)\,, \end{aligned} \tag{12}$$

for all $t \geq 0$, *with initial condition* $\psi_i(0) = \infty$ *for all* $i \in \mathsf{C}$.

C. General algorithm definition

In order to fully define the AMP algorithm (5), (6), we need to provide constructions for the matrix Q^t, the nonlinearities η_t, and the vector b_t. In doing this, we exploit the fact that the state evolution sequence $\{\phi(t)\}_{t\geq 0}$ can be precomputed.

We define the matrix Q^t by

$$Q^t_{ij} \equiv \frac{\phi_{\mathsf{g}(i)}(t)^{-1}}{\sum_{k=1}^{L_r} W_{k,\mathsf{g}(j)}\phi_k(t)^{-1}}\,. \tag{13}$$

Notice that Q^t is block-constant: for any $r \in \mathsf{R}$, $s \in \mathsf{C}$, the block $Q^t_{R(r),C(s)}$ has all its entries equal.

As mentioned in Section I, the function $\eta_t : \mathbb{R}^n \to \mathbb{R}^n$ is chosen to be separable, i.e. for $v \in \mathbb{R}^N$:

$$\eta_t(v) = \big(\eta_{t,1}(v_1), \eta_{t,2}(v_2),\, \dots\, , \eta_{t,N}(v_N)\big)\,. \tag{14}$$

We take $\eta_{t,i}$ to be a conditional expectation estimator for $X \sim p_X$ in gaussian noise:

$$\begin{aligned} \eta_{t,i}(v_i) &= \mathbb{E}\big\{X \,\big|\, X + \, s_{\mathsf{g}(i)}(t)^{-1/2}Z = v_i\,\big\}\,, \\ s_r(t) &\equiv \sum_{u\in \mathsf{R}} W_{u,r}\phi_u(t)^{-1}\,, \end{aligned} \tag{15}$$

where $Z \sim \mathsf{N}(0,1)$ and independent of X. Finally, in order to define the vector b_i^t, let us introduce the quantity

$$\langle \eta_t'\rangle_u = \frac{1}{N}\sum_{i\in C(u)} \eta'_{t,i}\big(x_i^t + ((Q^t \odot A)^* r^t)_i\big)\,. \tag{16}$$

The vector b^t is then defined by

$$\mathsf{b}_i^t \equiv \frac{1}{\delta}\sum_{u\in \mathsf{C}} W_{\mathsf{g}(i),u}\widetilde{Q}^{t-1}_{\mathsf{g}(i),u}\,\langle \eta'_{t-1}\rangle_u\,, \tag{17}$$

where we defined $Q^t_{i,j} = \widetilde{Q}^t_{r,u}$ for $i \in R(r)$, $j \in C(u)$. Again b_i^t is block-constant: the vector $\mathsf{b}^t_{C(u)}$ has all its entries equal.

This completes our definition of the AMP algorithm.

D. Choices of parameters

In order to prove our main Theorem II.3, we use a sensing matrix from the ensemble $\mathcal{M}(W, M, N)$ for a suitable choice of the matrix $W \in \mathbb{R}^{\mathsf{R}\times\mathsf{C}}$. Our construction depends on parameters $\rho \in \mathbb{R}_+$, $L, L_0 \in \mathbb{N}$, and on the 'shape function' $\mathcal{W}$. As explained below, ρ will be taken to be small, and hence we will treat $1/\rho$ as an integer to avoid rounding (which introduces in any case a negligible error).

Definition III.3. *A* shape function *is a function* $\mathcal{W} : \mathbb{R} \to \mathbb{R}_+$ *continuously differentiable, with support in* $[-1, 1]$ *and such that* $\int_{\mathbb{R}} \mathcal{W}(u)\,\mathrm{d}u = 1$, *and* $\mathcal{W}(-u) = \mathcal{W}(u)$.

We let $\mathsf{C} \cong \{-2\rho^{-1}, \dots, 0, 1, \dots, L-1\}$, so that $L_c = L + 2\rho^{-1}$. The rows are partitioned as follows:

$$\mathsf{R} = \mathsf{R}_0 \cup \Big\{\cup_{i=-2\rho^{-1}}^{-1} \mathsf{R}_i\Big\}\,,$$

where $\mathsf{R}_0 \cong \{-\rho^{-1}, \dots, 0, 1, \dots, L-1+\rho^{-1}\}$, and $|\mathsf{R}_i| = L_0$. Hence $L_r = L_c + 2\rho^{-1}L_0$.

Finally, we take N so that $n = NL_c$, and let $M = N\delta$ so that $m = ML_r = N(L_c + 2\rho^{-1}L_0)\delta$. Notice that $m/n = \delta(L_c + 2\rho^{-1}L_0)/L_c$. Since we will take L_c much larger than L_0/ρ, we in fact have m/n arbitrarily close to δ.

Given these inputs, we construct the corresponding matrix $W = W(L, L_0, \mathcal{W}, \rho)$ as follows.

1) For $i \in \{-2\rho^{-1}, \dots, -1\}$, and each $a \in \mathsf{R}_i$, we let $W_{a,i} = 1$. Further, $W_{a,j} = 0$ for all $j \in \mathsf{C} \setminus \{i\}$.

2) For all $a \in \mathsf{R}_0 \cong \{-\rho^{-1}, \dots, 0, \dots, L-1+\rho^{-1}\}$, we let $W_{a,i} = \rho\,\mathcal{W}\big(\rho\,(a-i)\big)$ for $i \in \{-2\rho^{-1}, \dots, L-1\}$.

It is not hard to check that W is roughly row-stochastic.

Acknowledgment

A.J. is supported by a Caroline and Fabian Pease Stanford Graduate Fellowship. Partially supported by NSF CAREER award CCF- 0743978 and AFOSR grant FA9550-10-1-0360. The authors thank the reviewers for their insightful comments.

References

[1] M. Bayati and A. Montanari. The LASSO risk for gaussian matrices. *IEEE Trans. on Inform. Theory*, 2011. arXiv:1008.2581.

[2] E. Candes, J. K. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Inform. Theory*, 52:489 – 509, 2006.

[3] E. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2006.

[4] E. J. Candés and T. Tao. Decoding by linear programming. *IEEE Trans. on Inform. Theory*, 51:4203–4215, 2005.

[5] D. Donoho, A. Maleki, and A. Montanari. The Noise Sensitivity Phase Transition in Compressed Sensing. *IEEE Trans. on Inform. Theory*, 57:6920–6941, 2011.

[6] D. L. Donoho. Compressed sensing. *IEEE Trans. on Inform. Theory*, 52:489–509, April 2006.

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

[8] D. L. Donoho, A. Maleki, and A. Montanari. Message Passing Algorithms for Compressed Sensing. *PNAS*, 106:18914–18919, 2009.

[9] D. L. Donoho, A. Maleki, and A. Montanari. Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction. In *Proc. of ITW*, Cairo, 2010.

[10] P. Indyk, E. Price, and D. Woodruff. On the Power of Adaptivity in Sparse Recovery. In *FOCS*, October 2011.

[11] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborova. Statistical physics-based reconstruction in compressed sensing. arXiv:1109.4424, 2011.

[12] S. Kudekar and H. Pfister. The effect of spatial coupling on compressive sensing. In *48th Annual Allerton Conference*, pages 347 –353, 2010.

[13] A. Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Hungarica*, 10:193–215, 1959.

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

[15] Y. Wu and S. Verdú. Rényi Information Dimension: Fundamental Limits of Almost Lossless Analog Compression. *IEEE Trans. on Inform. Theory*, 56:3721–3748, 2010.

[16] Y. Wu and S. Verdú. MMSE dimension. *IEEE Trans. on Inform. Theory*, 57(8):4857–4879, 2011.