Message-passing algorithms for compressed sensing

David L. Donoho[a,1], **Arian Maleki**[b], **and Andrea Montanari**[a,b,1]

Departments of [a]Statistics and [b]Electrical Engineering, Stanford University, Stanford, CA 94305

Contributed by David L. Donoho, September 11, 2009 (sent for review July 21, 2009)

Compressed sensing aims to undersample certain high-dimensional signals yet accurately reconstruct them by exploiting signal characteristics. Accurate reconstruction is possible when the object to be recovered is sufficiently sparse in a known basis. Currently, the best known sparsity–undersampling tradeoff is achieved when reconstructing by convex optimization, which is expensive in important large-scale applications. Fast iterative thresholding algorithms have been intensively studied as alternatives to convex optimization for large-scale problems. Unfortunately known fast algorithms offer substantially worse sparsity–undersampling tradeoffs than convex optimization. We introduce a simple costless modification to iterative thresholding making the sparsity–undersampling tradeoff of the new algorithms equivalent to that of the corresponding convex optimization procedures. The new iterative-thresholding algorithms are inspired by belief propagation in graphical models. Our empirical measurements of the sparsity–undersampling tradeoff for the new algorithms agree with theoretical calculations. We show that a state evolution formalism correctly derives the true sparsity–undersampling tradeoff. There is a surprising agreement between earlier calculations based on random convex polytopes and this apparently very different theoretical formalism.

combinatorial geometry | phase transitions | linear programming | iterative thresholding algorithms | state evolution

Compressed sensing refers to a growing body of techniques that "undersample" high-dimensional signals and yet recover them accurately (1). Such techniques make fewer measurements than traditional sampling theory demands: rather than sampling proportional to frequency bandwidth, they make only as many measurements as the underlying "information content" of those signals. However, compared with traditional sampling theory, which can recover signals by applying simple linear reconstruction formulas, the task of signal recovery from reduced measurements requires nonlinear and, so far, relatively expensive reconstruction schemes. One popular class of reconstruction schemes uses linear programming (LP) methods; there is an elegant theory for such schemes promising large improvements over ordinary sampling rules in recovering sparse signals. However, solving the required LPs is substantially more expensive in applications than the linear reconstruction schemes that are now standard. In certain imaging problems, the signal to be acquired may be an image with 10^6 pixels and the required LP would involve tens of thousands of constraints and millions of variables. Despite advances in the speed of LP, such problems are still dramatically more expensive to solve than we would like.

Here, we develop an iterative algorithm achieving reconstruction performance in one important sense identical to LP-based reconstruction while running dramatically faster. We assume that a vector y of n measurements is obtained from an unknown N-vector x_0 according to $y = Ax_0$, where A is the $n \times N$ measurement matrix $n < N$. Starting from an initial guess $x^0 = 0$, the first-order approximate message-passing (AMP) algorithm proceeds iteratively according to.

$$x^{t+1} = \eta_t(A^*z^t + x^t), \qquad [1]$$

$$z^t = y - Ax^t + \frac{1}{\delta} z^{t-1}\langle \eta'_{t-1}(A^*z^{t-1} + x^{t-1})\rangle. \qquad [2]$$

Here $\eta_t(\cdot)$ are scalar threshold functions (applied componentwise), $x^t \in \mathbb{R}^N$ is the current estimate of x_0, and $z^t \in \mathbb{R}^n$ is the current residual. A^* denotes transpose of A. For a vector $u = (u(1), \ldots, u(N))$, $\langle u \rangle \equiv \sum_{i=1}^{N} u(i)/N$. Finally $\eta'_t(s) = \frac{\partial}{\partial s}\eta_t(s)$.

Iterative thresholding algorithms of other types have been popular among researchers for some years (2), one focus being on schemes of the form

$$x^{t+1} = \eta_t(A^*z^t + x^t), \qquad [3]$$

$$z^t = y - Ax^t. \qquad [4]$$

Such schemes can have very low per-iteration cost and low storage requirements; they can attack very large-scale applications, much larger than standard LP solvers can attack. However, Eqs. **3** and **4** fall short of the sparsity–undersampling tradeoff offered by LP reconstruction (3).

Iterative thresholding schemes based on Eqs. **3** and **4** lack the crucial term in Eq. **2**, namely, $\frac{1}{\delta} z^{t-1}\langle \eta'_{t-1}(A^*z^{t-1} + x^{t-1})\rangle$ is not included. We derive this term from the theory of belief propagation in graphical models and show that it substantially improves the sparsity–undersampling tradeoff.

Extensive numerical and Monte Carlo work reported here shows that AMP, defined by Eqs. **1** and **2** achieves a sparsity–undersampling tradeoff matching the theoretical tradeoff which has been proved for LP-based reconstruction. We consider a parameter space with axes quantifying sparsity and undersampling. In the limit of large dimensions N, n, the parameter space splits in two phases: one where the AMP approach is successful in accurately reconstructing x_0 and one where it is unsuccessful. Refs. 4–6 derived regions of success and failure for LP-based recovery. We find these two ostensibly different partitions of the sparsity–undersampling parameter space to be identical. Both reconstruction approaches succeed or fail over the same regions (see Fig. 1).

Our finding has extensive empirical evidence and strong theoretical support. We introduce a state evolution (SE) formalism and find that it accurately predicts the dynamical behavior of numerous observables of the AMP algorithm. In this formalism, the mean squared error (MSE) of reconstruction is a state variable; its change from iteration to iteration is modeled by a simple scalar function, the MSE map. When this map has non-zero fixed points, the formalism predicts that AMP will not successfully recover the desired solution. The MSE map depends on the underlying sparsity and undersampling ratios and can develop non-zero fixed points over a region of sparsity/undersampling space. The region is evaluated analytically and found to coincide very precisely (i.e., within numerical precision) with the region over which LP-based methods are proved to fail. Extensive Monte Carlo testing of AMP

Author contributions: D.L.D. and A. Montanari designed research; D.L.D., A. Maleki, and A. Montanari performed research; D.L.D., A. Maleki, and A. Montanari analyzed data; and D.L.D., A. Maleki, and A. Montanari wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1] To whom correspondence may be addressed. E-mail: montanari@stanford.edu or donoho@stat.stanford.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0909892106/DCSupplemental.

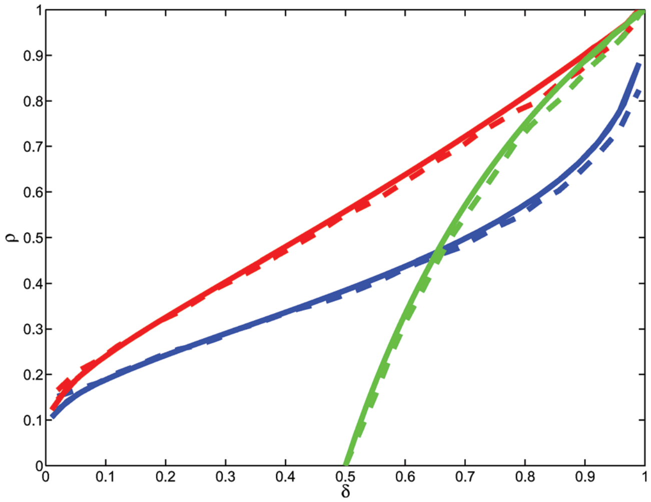

Fig. 1. The phase transition lines for reconstructing sparse nonnegative vectors (problem +, red), sparse signed vectors (problem ±, blue) and vectors with entries in $[-1, 1]$ (problem □, green). Continuous lines refer to analytical predictions from combinatorial geometry or the SE formalisms. Dashed lines present data from experiments with the AMP algorithm, with signal length $N = 1{,}000$ and $T = 1{,}000$ iterations. For each value of δ, we considered a grid of ρ values, at each value, generating 50 random problems. The dashed line presents the estimated 50th percentile of the response curve. At that percentile, the root MSE after T iterations obeys $\sigma_T \le 10^{-3}$ in half of the simulated reconstructions. The discrepancy between the observed PT for AMP and the theoretical curves is statistically significant (t-score 7). According to *Theorem 2* a discrepancy should be detectible at finite T. *SI Appendix*, Section 13 shows that *Theorem 2* accurately describes the evolution of MSE for the AMP algorithm.

reconstruction finds that the region where AMP fails is, to within statistical precision, the same region.

In short we introduce a fast iterative algorithm that is found to perform as well as corresponding LP-based methods on random problems. Our findings are supported from simulations and from a theoretical formalism.

Remarkably, the success/failure phases of LP reconstruction were previously found by methods in combinatorial geometry; we give here what amounts to a very simple formula for the phase boundary, derived using a very different and seemingly elegant theoretical principle.

Underdetermined Linear Systems. Let $x_0 \in \mathbb{R}^N$ be the signal of interest. We are interested in reconstructing it from the vector of measurements $y = Ax_0$, with $y \in \mathbb{R}^n$, for $n < N$. For the moment, we assume that the entries A_{ij} of the measurement matrix are independent and identically distributed normal $\mathsf{N}(0, 1/n)$.

We consider three canonical models for the signal x_0 and three reconstruction procedures based on LP.

+: x_0 is nonnegative, with at most k entries different from 0. Reconstruct by solving the LP: minimize $\sum_{i=1}^{N} x_i$ subject to $x \ge 0$, and $Ax = y$.

±: x_0 has as many as k non-zero entries. Reconstruct by solving the minimum ℓ_1 norm problem: minimize $||x||_1$, subject to $Ax = y$. This can be cast as an LP.

□: $x_0 \in [-1, 1]^N$, with at most k entries in the interior $(-1, 1)$. Reconstruction by solving the LP feasibility problem: find any vector $x \in [-1, +1]^N$ with $Ax = y$.

Despite the fact that the systems are underdetermined, under certain conditions on k, n, N these procedures perfectly recover x_0. This takes place subject to a sparsity–undersampling tradeoff, namely, an upper bound on the signal complexity k relative to n and N.

Phase Transitions. The sparsity–undersampling tradeoff can most easily be described by taking a large-system limit. In that limit, we fix parameters (δ, ρ) in $(0, 1)^2$ and let $k, n, N \to \infty$ with $k/n \to \rho$ and $n/N \to \delta$. The sparsity–undersampling behavior we study is controlled by (δ, ρ), with δ being the undersampling fraction and ρ being a measure of sparsity (with larger ρ corresponding to more complex signals).

The domain $(\delta, \rho) \in (0, 1)^2$ has two phases, a "success" phase, where exact reconstruction typically occurs, and a "failure" phase where exact reconstruction typically fails. More formally, for each choice of $\chi \in \{+, \pm, \square\}$ there is a function $\rho_{\rm CG}(\cdot; \chi)$ whose graph partitions the domain into two regions. In the upper region, where $\rho > \rho_{\rm CG}(\delta; \chi)$, the corresponding LP reconstruction $x_1(\chi)$ fails to recover x_0 in the following sense: as $k, n, N \to \infty$ in the large-system limit with $k/n \to \rho$ and $n/N \to \delta$, the probability of exact reconstruction $\{x_1(\chi) = x_0\}$ tends to zero exponentially fast. In the lower region, where $\rho < \rho_{\rm CG}(\delta; \chi)$, LP reconstruction succeeds to recover x_0 in the following sense: as $k, n, N \to \infty$ in the large-system limit with $k/n \to \rho$ and $n/N \to \delta$, the probability of exact reconstruction $\{x_1(\chi) = x_0\}$ tends to one exponentially fast. We refer to refs. 4–6 for proofs and precise definitions of the curves $\rho_{\rm CG}(\cdot; \chi)$.

The three functions $\rho_{\rm CG}(\cdot; +)$, $\rho_{\rm CG}(\cdot; \pm)$, $\rho_{\rm CG}(\cdot; \square)$ are shown in Fig. 1; they are the red, blue, and green curves, respectively. The ordering $\rho_{\rm CG}(\delta; +) > \rho_{\rm CG}(\delta; \pm)$ (red > blue) says that knowing that a signal is sparse and positive is more valuable than only knowing it is sparse. Both the red and blue curves behave as $\rho_{\rm CG}(\delta; +, \pm) \sim (2\log(1/\delta))^{-1}$ as $\delta \to 0$; surprisingly large amounts of undersampling are possible, if sufficient sparsity is present. In contrast, $\rho_{\rm CG}(\delta; \square) = 0$ for $\delta < 1/2$ (green curve) so the bounds $[-1, 1]$ are really of no help unless we use a limited amount of undersampling, i.e., by less than a factor of 2.

Explicit expressions for $\rho_{\rm CG}(\delta; +, \pm)$ are given in refs. 4 and 5; they are quite involved and use methods from combinatorial geometry. By *Finding 1* below, they agree within numerical precision to the following formula:

$$\rho_{\rm SE}(\delta; \chi) = \max_{z \ge 0} \left\{ \frac{1 - (\kappa_\chi/\delta)[(1+z^2)\Phi(-z) - z\phi(z)]}{1 + z^2 - \kappa_\chi[(1+z^2)\Phi(-z) - z\phi(z)]} \right\}, \tag*{[5]}$$

where $\kappa_\chi = 1, 2$ respectively for $\chi = +, \pm$. This formula, a principal result of this work, uses methods unrelated to combinatorial geometry.

Iterative Approaches. Mathematical results for the large-system limit correspond well to application needs. Realistic modern problems in spectroscopy and medical imaging demand reconstructions of objects with tens of thousands or even millions of unknowns. Extensive testing of practical convex optimizers in these problems (7) has shown that the large system asymptotic accurately describes the observed behavior of computed solutions to the above LPs. But the same testing shows that existing convex optimization algorithms run slowly on these large problems, taking minutes or even hours on the largest problems of interest.

Many researchers have abandoned formal convex optimization, turning to fast iterative methods instead (8–10).

The iteration (Eqs. **1** and **2**) is very attractive because it does not require the solution of a system of linear equations and because it does not require explicit operations on the matrix A; it only requires that one apply the operators A and A^* to any given vector. In a number of applications—for example magnetic resonance imaging—the operators A which make practical sense are not really Gaussian random matrices, but rather random sections of the Fourier transform and other physically inspired transforms (1). Such operators can be applied very rapidly using fast Fourier transforms, rendering the above iteration extremely fast. Provided the process stops after a limited number of iterations, the computations are very practical.

The thresholding functions $\{\eta_t(\cdot)\}_{t \ge 0}$ in these schemes depend on both iteration and problem setting. Here, we consider

$\eta_t(\cdot) = \eta(\cdot; \lambda\sigma_t, \chi)$, where λ is a threshold control parameter, $\chi \in \{+, \pm, \square\}$ denotes the setting, and $\sigma_t^2 = \mathrm{Ave}_j \mathbb{E}\{(x^t(j) - x_0(j))^2\}$ is the MSE of the current estimate x^t (in practice an empirical estimate of this quantity is used).

For instance, in the case of sparse signed vectors (i.e., problem setting $\pm$), we apply soft thresholding $\eta_t(u) = \eta(u; \lambda\sigma, \pm)$, where

$$\eta(u; \lambda\sigma, \pm) = \begin{cases} (u - \lambda\sigma) & \text{if } u \geq \lambda\sigma, \\ (u + \lambda\sigma) & \text{if } u \leq -\lambda\sigma, \\ 0 & \text{otherwise,} \end{cases} \qquad \textbf{[6]}$$

where we dropped the argument $\pm$ to lighten notation. Notice that η_t depends on the iteration number t only through the MSE σ_t^2.

Heuristics for Iterative Approaches. Why should the iterative approach work, i.e., converge to the correct answer x_0? We focus in this section on the popular case $\chi = \pm$. Suppose first that A is an orthogonal matrix, so $A^* = A^{-1}$. Then the iteration of Eqs. **1** and **2** stops in one step, correctly finding x_0. Next, imagine that A is an invertible matrix; using ref. 11 with clever scaling of A^* and clever choice of decreasing threshold, that iteration correctly finds x_0. Of course both these motivational observations assume $n = N$, i.e., no undersampling.

A motivational argument for thresholding in the undersampled case $n < N$ has been popular with engineers (1) and leads to a proper "psychology" for understanding our results. Consider the operator $H = A^*A - I$ and note that $A^*y = x_0 + Hx_0$. If A were orthogonal, we would of course have $H = 0$, and the iteration would, as we have seen, immediately succeed in one step. If A is a Gaussian random matrix and $n < N$, then of course A is not invertible and A^* is not A^{-1}. Instead of $Hx_0 = 0$, in the undersampled case Hx_0 behaves as a kind of noisy random vector, i.e., $A^*y = x_0 + \mathsf{noise}$. Now x_0 is supposed to be a sparse vector, and, as one can see, the noise term is accurately modeled as a vector with independent and identically distributed Gaussian entries with variance $n^{-1}\|x_0\|_2^2$.

In short, the first iteration gives us a "noisy" version of the sparse vector that we are seeking to recover. The problem of recovering a sparse vector from noisy measurements has been heavily discussed (12), and it is well understood that soft thresholding can produce a reduction in MSE when sufficient sparsity is present and the threshold is chosen appropriately. Consequently, one anticipates that x^1 will be closer to x_0 than A^*y.

At the second iteration, one has $A^*(y - Ax^1) = (x_0 - x_1) + H(x_0 - x^1)$. Naively, the matrix H does not correlate with x_0 or x^1, and so we might pretend that $H(x_0 - x^1)$ is again a Gaussian vector whose entries have variance $n^{-1}\|x_0 - x^1\|_2^2$. This "noise level" is smaller than at iteration zero, and so thresholding of this noise can be anticipated to produce an even more accurate result at iteration two, and so on.

There is a valuable digital communications interpretation of this process. The vector $w = Hx_0$ is the cross-channel interference or mutual access interference (MAI), i.e., the noiselike disturbance each coordinate of A^*y experiences from the presence of all the other "weakly interacting" coordinates. The thresholding iteration suppresses this interference in the sparse case by detecting the many "silent" channels and setting them a priori to zero, producing a putatively better guess at the next iteration. At that iteration, the remaining interference is proportional not to the size of the estimand, but instead to the estimation error; i.e., it is caused by the errors in reconstructing all the weakly interacting coordinates; these errors are only a fraction of the sizes of the estimands and so the error is significantly reduced at the next iteration.

SE. The above "sparse denoising/interference suppression" heuristic does agree qualitatively with the actual behavior one can observe in sample reconstructions. It is very tempting to take it literally. Assuming it is literally true that the MAI is Gaussian and independent from iteration to iteration, we can formally track the evolution, from iteration to iteration, of the MSE.

This gives a recursive equation for the formal MSE, i.e., the MSE which would be true if the heuristic were true. This takes the form

$$\sigma_{t+1}^2 = \Psi(\sigma_t^2), \qquad \textbf{[7]}$$

$$\Psi(\sigma^2) \equiv \mathbb{E}\left\{\left[\eta\left(X + \frac{\sigma}{\sqrt{\delta}} Z; \lambda\sigma\right) - X\right]^2\right\}. \qquad \textbf{[8]}$$

Here expectation is with respect to independent random variables $Z \sim \mathsf{N}(0, 1)$ and X, whose distribution coincides with the empirical distribution of the entries of x_0. We use soft thresholding (6) if the signal is sparse and signed, i.e. if $\chi = \pm$. In the case of sparse non-negative vectors, $\chi = +$, we will let $\eta(u; \lambda\sigma, +) = \max(u - \lambda\sigma, 0)$. Finally, for $\chi = \square$, we let $\eta(u; \square) = \mathrm{sign}(u)\min(|u|, 1)$. Calculations of this sort are familiar from the theory of soft thresholding of sparse signals; see *SI Appendix* for details.

We call $\Psi : \sigma^2 \mapsto \Psi(\sigma^2)$ the MSE map (see Fig. 2).

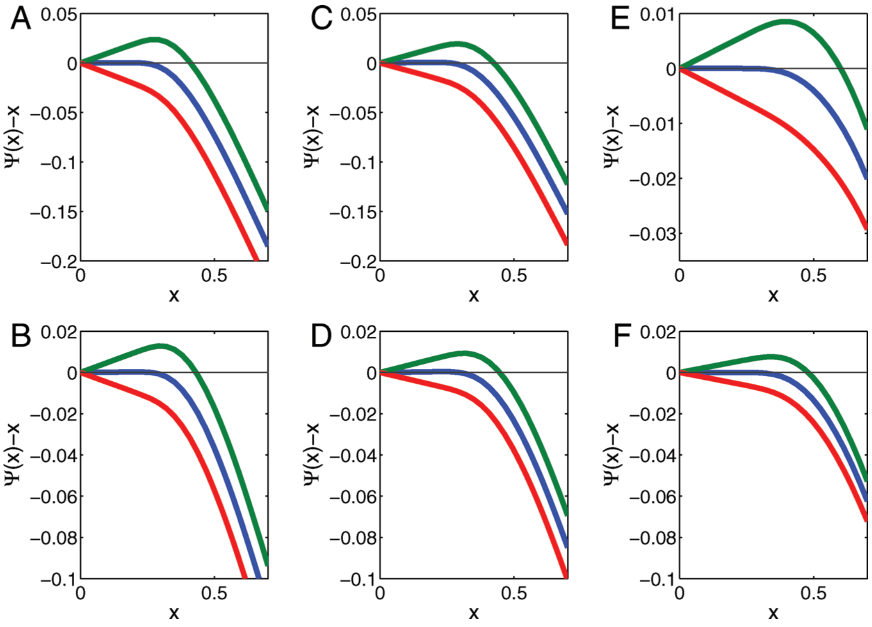

Fig. 2. Development of fixed points for formal MSE evolution. Here we plot $\Psi(\sigma^2) - \sigma^2$, where $\Psi(\cdot)$ is the MSE map for $\chi = +$ (left column), $\chi = \pm$ (center column), and $\chi = \square$ (right column) and where $\delta = 0.1$ (upper row, $\chi \in \{+, \pm\}$), $\delta = 0.55$ (upper row, $\chi = \square$), $\delta = 0.4$ (lower row, $\chi \in \{+, \pm\}$), and $\delta = 0.75$ (lower row, $\chi = \square$). A crossing of the y axis corresponds to a fixed point of Ψ. If the graphed quantity is negative for positive σ^2, Ψ has no fixed points for $\sigma > 0$. Different curves correspond to different values of ρ: where ρ is respectively less than, equal to, and greater than ρ_{SE}. In each case, Ψ has a stable fixed point at zero for $\rho < \rho_{\mathrm{SE}}$ and no other fixed points, an unstable fixed point at zero for $\rho = \rho_{\mathrm{SE}}$, and develops two fixed points at $\rho > \rho_{\mathrm{SE}}$. Blue curves correspond to $\rho = \rho_{\mathrm{SE}}(\delta; \chi)$, green corresponds to $\rho = 1.05 \cdot \rho_{\mathrm{SE}}(\delta; \chi)$, and red corresponds to $\rho = 0.95 \cdot \rho_{\mathrm{SE}}(\delta; \chi)$.

Definition 1. *Given implicit parameters $(\chi, \delta, \rho, \lambda, F)$, with $F = F_X$ the distribution of the random variable X, SE is the recursive map (one-dimensional dynamical system): $\sigma_t^2 \mapsto \Psi(\sigma_t^2)$.*

Implicit parameters $(\chi, \delta, \rho, \lambda, F)$ stay fixed during the evolution. Equivalently, the full state evolves by the rule

$$\left(\sigma_t^2; \chi, \delta, \rho, \lambda, F_X\right) \mapsto \left(\Psi\left(\sigma_t^2\right); \chi, \delta, \rho, \lambda, F_X\right).$$

Parameter space is partitioned into two regions:

Region (I): $\Psi(\sigma^2) < \sigma^2$ for all $\sigma^2 \in (0, \mathbb{E}X^2]$. Here $\sigma_t^2 \to 0$ as $t \to \infty$: the SE converges to zero.

Region (II): The complement of Region (I). Here, the SE recursion does not evolve to $\sigma^2 = 0$.

The partitioning of parameter space induces a notion of sparsity threshold, the minimal sparsity guarantee needed to obtain convergence of the formal MSE:

$$\rho_{\mathrm{SE}}(\delta; \chi, \lambda, F_X) \equiv \sup\{\rho : (\delta, \rho, \lambda, F_X) \in \text{Region (I)}\}. \qquad \textbf{[9]}$$

Of course, ρ_{SE} depends on the case $\chi \in \{+, \pm, \square\}$; it also seems to depend on the signal distribution F_X; however, an essential simplification is provided by the following.

Proposition 1. *For the three canonical problems $\chi \in \{+, \pm, \square\}$, any $\delta \in [0, 1]$, and any random variable X with the prescribed sparsity and bounded second moment, $\rho_{\mathrm{SE}}(\delta; \chi, \lambda, F_X)$ is independent of F_X.*

Independence from F allows us to write $\rho_{\mathrm{SE}}(\delta; \chi, \lambda)$ for the sparsity thresholds. For proof, see *SI Appendix*. Adopt the notation

$$\rho_{\mathrm{SE}}(\delta; \chi) = \sup_{\lambda \geq 0} \rho_{\mathrm{SE}}(\delta; \chi, \lambda). \qquad \textbf{[10]}$$

Finding 1. *For the three canonical problems $\chi \in \{+, \pm, \square\}$, and for any $\delta \in (0, 1)$*

$$\rho_{\mathrm{SE}}(\delta; \chi) = \rho_{\mathrm{CG}}(\delta; \chi). \qquad \textbf{[11]}$$

In short, the formal MSE evolves to zero exactly over the same region of (δ, ρ) phase space, as does the phase diagram for the corresponding convex optimization.

SI Appendix proves *Finding 1* rigorously in the case $\chi = \square$, all $\delta \in (0, 1)$. It also proves for $\chi \in \{+, \pm\}$, the weaker relation $\rho_{\mathrm{SE}}(\delta; \chi)/\rho_{\mathrm{CG}}(\delta; \chi) \to 1$ as $\delta \to 0$. Numerical evaluations of both sides of Eq. **11** are also observed to agree at all δ in a fine grid of points in $(0, 1)$.

Failure of Standard Iterative Algorithms. If we trusted that formal MSE truly describes the evolution of the iterative thresholding algorithm, *Finding 1* would imply that iterative thresholding allows undersampling just as aggressively in solving underdetermined linear systems as the corresponding LP.

Finding 1 gives new reason to hope for a possibility that has already inspired many researchers over the last five years: the possibility of finding a very fast algorithm that replicates the behavior of convex optimization in settings $+, \pm, \square$.

Unhappily the formal MSE calculation does not describe the behavior of iterative thresholding:

1. SE does not predict the observed properties of iterative thresholding algorithms.
2. Iterative thresholding algorithms, even when optimally tuned, do not achieve the optimal phase diagram.

Ref. 3 carried out an extensive empirical study of iterative thresholding algorithms. Even optimizing over the free parameter λ and the nonlinearity η, the phase transition was observed at significantly smaller values of ρ than those observed for LP-based algorithms. Even improvements over iterative thresholding such as CoSaMP and Subspace Pursuit (13, 14) did not achieve the transitions of LP-based methods (see also Fig. 3).

Numerical simulations also show very clearly that the MSE map does not describe the evolution of the actual MSE under iterative thresholding. The mathematical reason for this failure is quite simple. After the first iteration, the entries of x^t become strongly dependent, and SE does not predict the moments of x^t.

Message-Passing (MP) Algorithm. The main surprise of our work here is that this failure is not the end of the story. We now consider a modification of iterative thresholding inspired by MP algorithms for inference in graphical models (15), and graph-based error correcting codes (16). These are iterative algorithms, whose basic variables ("messages") are associated to directed edges in a graph that encodes the structure of the statistical model. The relevant graph here is a complete bipartite graph over N nodes on one side (variable nodes), and n on the others (measurement nodes). Messages are updated according to the rules

$$x^{t+1}_{i\to a} = \eta_t\left(\sum_{b\in[n]\setminus a} A_{bi} z^t_{b\to i}\right), \qquad \textbf{[12]}$$

$$z^t_{a\to i} = y_a - \sum_{j\in[N]\setminus i} A_{aj} x^t_{j\to a}, \qquad \textbf{[13]}$$

for each $(i, a) \in [N] \times [n]$. Just as in other areas where MP arises, the subscript $i \to a$ is vocalized "i sends to a," and $a \to i$ as "a sends to i."

This MP algorithm[†] has one important drawback with respect to iterative thresholding. Instead of updating N estimates, at each iteration it updates Nn messages, increasing significantly the algorithm complexity. On the other hand, the right-hand side of Eq. **12** depends weakly on the index a (only one out of n terms is excluded) and the right-hand side of Eq. **12** depends weakly on i. Neglecting altogether this dependence leads to the iterative thresholding in Eqs. **3** and **4**. A more careful analysis of this dependence leads to corrections of order one in the high-dimensional limit. Such corrections are however fully captured by the last term on the right-hand side of Eq. **2**, thus leading to the AMP algorithm. Statistical physicists would call this the "Onsager reaction term" (22).

SE is Correct for MP. Although AMP seems very similar to simple iterative thresholding in Eqs. **3** and **4**, SE accurately describes its properties but not those of the standard iteration. As a consequence of *Finding 1*, properly tuned versions of MP-based algorithms are asymptotically as powerful as LP reconstruction. We have conducted extensive simulation experiments with AMP and more limited experiments with MP, which is computationally more intensive (for details see *SI Appendix*). These experiments show that the performance of the algorithms can be accurately modeled using the MSE map. Let's be more specific.

According to SE, performance of the AMP algorithm is predicted by tracking the evolution of the formal MSE σ_t^2 via the recursion in Eq. **7**. Although this formalism is quite simple, it is accurate in the high-dimensional limit. Corresponding to the formal quantities calculated by SE are the actual quantities, so of course to the formal MSE corresponds the true MSE $N^{-1}\|x^t - x_0\|_2^2$. Other quantities can be computed in terms of the state σ_t^2 as well: for instance, the true false-alarm rate $(N - k)^{-1}\#\{i : x^t(i) \neq 0 \text{ and } x_0(i) = 0\}$ is predicted via the formal false-alarm rate $\mathbb{P}\{\eta_t(X + \delta^{-1/2}\sigma_t Z) \neq 0 | X = 0\}$. Analogously, the true missed-detection rate $k^{-1}\#\{i : x^t(i) = 0 \text{ and } x_0(i) \neq 0\}$ is predicted by the

[†] For earlier applications of MP to compressed sensing, see refs. 17–19. Relationships between MP and LP were explored in a number of papers, albeit from a different perspective (e.g., see refs. 20 and 21).

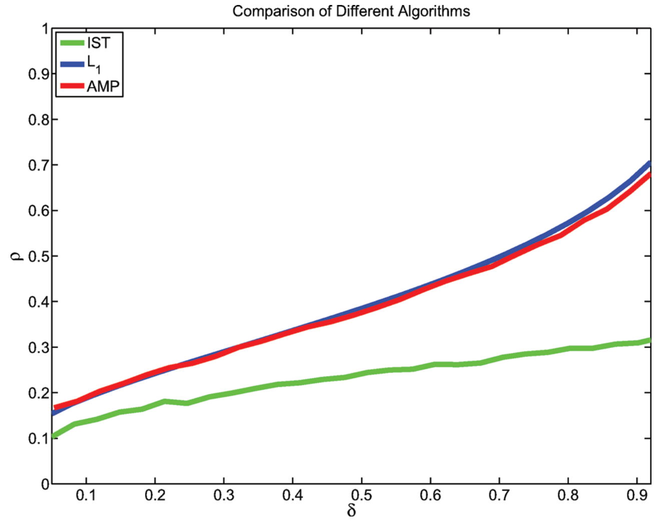

Fig. 3. Observed phase transitions of reconstruction algorithms. Red curve, AMP; green curve, iterative soft thresholding (IST); blue curve, theoretical ℓ_1 transition. Parameters of IST tuned for best possible phase transition (3). Reconstruction signal length $N = 1{,}000$. $T = 1{,}000$ iterations. Empirical phase transition is value of ρ at which success rate is 50%. Details are in *SI Appendix*.

formal missed-detection rate $\mathbb{P}\{\eta_t(X+\delta^{-1/2}\sigma_t Z) = 0|X \neq 0\}$, and so on.

Our experiments establish large N-agreement of actual and formal quantities. *SI Appendix* justifies the following.

Finding 2. *For the AMP algorithm, and large dimensions N, n, we observe*

I. *SE correctly predicts the evolution of numerous statistical properties of x^t with the iteration number t. The MSE, the number of non-zeros in x^t, the number of false alarms, the number of missed detections, and several other measures all evolve in way that matches the SE formalism to within experimental accuracy.*
II. *SE correctly predicts the success/failure to converge to the correct result. In particular, SE predicts no convergence when $\rho > \rho_{\rm SE}(\delta;\chi,\lambda)$, and convergence if $\rho < \rho_{\rm SE}(\delta;\chi,\lambda)$. This is indeed observed empirically.*

Analogous observations were made for MP.

Optimizing the MP Phase Transition. An inappropriately tuned version of MP/AMP will not perform well compared with other algorithms, for example LP-based reconstructions. However, SE provides a natural strategy to tune MP and AMP (i.e., to choose the free parameter λ): simply use the value achieving the maximum in Eq. **10**. We denote this value by $\lambda_\chi(\delta)$, $\chi \in \{+,\pm,\square\}$ and refer to the resulting algorithms as to optimally tuned MP/AMP (or sometimes MP/AMP for short). They achieve the SE phase transition:

$$\rho_{\rm SE}(\delta;\chi) = \rho_{\rm SE}(\delta;\chi,\lambda_\chi(\delta)).$$

An explicit characterization of $\lambda_\chi(\delta)$, $\chi \in \{+,\pm\}$ can be found in the next section. Optimally tuned AMP/MP has a formal MSE evolving to zero exponentially fast everywhere below phase transition.

Theorem 2. *For $\delta \in [0,1]$, $\rho < \rho_{\rm SE}(\delta;\chi)$, and any associated random variable X, the formal MSE of optimally tuned AMP/MP evolves to zero under SE. Viceversa, if $\rho > \rho_{\rm SE}(\delta;\chi)$, the formal MSE does not evolve to zero. Furthermore, for $\rho < \rho_{\rm SE}(\delta;\chi)$, there exists $b = b(\delta,\rho) > 0$ with the following property. If σ_t^2 denotes the formal MSE after t SE steps, then, for all $t \geq 0$*

$$\sigma_t^2 \leq \sigma_0^2 \exp(-bt). \qquad [14]$$

This rigorous result about evolution of formal MSE is complemented by empirical work showing that the actual MSE evolves the same way (see *SI Appendix*, which also offers formulas for the rate exponent b).

Details About the MSE Mapping

In this section, we sketch the proof of *Proposition 1*: the iterative threshold does not depend on the details of the signal distribution. Furthermore, we show how to derive the explicit expression for $\rho_{\rm SE}(\delta;\chi)$, $\chi \in \{+,\pm\}$, given in the Introduction.

Local Stability Bound. The SE threshold $\rho_{\rm SE}(\delta;\chi,\lambda)$ is the supremum of all ρ's such that the MSE map $\Psi(\sigma^2)$ lies below the σ^2 line for all $\sigma^2 > 0$. Since $\Psi(0) = 0$, for this to happen it must be true that the derivative of the MSE map at $\sigma^2 = 0$ is smaller than or equal to 1. We are therefore led to define the following "local stability" threshold:

$$\rho_{\rm LS}(\delta;\chi,\lambda) \equiv \sup\left\{\rho : \left.\frac{d\Psi}{d\sigma^2}\right|_{\sigma^2=0} < 1\right\}. \qquad [15]$$

The above argument implies that $\rho_{\rm SE}(\delta;\chi,\lambda) \leq \rho_{\rm LS}(\delta;\chi,\lambda)$.

Considering for instance $\chi = +$, we obtain the following expression for the first derivative of Ψ:

$$\frac{d\Psi}{d\sigma^2} = (\delta^{-1}+\lambda^2)\cdot\mathbb{E}\,\Phi\left(\frac{\sqrt{\delta}}{\sigma}(X-\lambda\sigma)\right) - \mathbb{E}\left\{\frac{(X+\lambda\sigma)}{\sigma\sqrt{\delta}}\,\phi\left(\frac{\sqrt{\delta}}{\sigma}(X-\lambda\sigma)\right)\right\},$$

where $\phi(z)$ is the standard Gaussian density at z, $\Phi(z) = \int_{-\infty}^{z}\phi(z')\,dz'$ is the Gaussian distribution, and $\xi = \delta^{-1}+\lambda^2$.

Evaluating this expression as $\sigma^2 \downarrow 0$, we get the local stability threshold for $\chi = +$:

$$\rho_{\rm LS}(\delta;\chi,\lambda) = \left.\frac{1-(\kappa_\chi/\delta)[(1+z^2)\Phi(-z)-z\phi(z)]}{1+z^2-\kappa_\chi[(1+z^2)\Phi(-z)-z\phi(z)]}\right|_{z=\lambda\sqrt{\delta}},$$

where κ_χ is the same as in Eq. **5**. Notice that $\rho_{\rm LS}(\delta;+,\lambda)$ depends on the distribution of X only through its sparsity (i.e., it is independent of F_X).

Tightness of the Bound and Optimal Tuning. We argued that $\frac{d\Psi}{d\sigma^2}|_{\sigma^2=0} < 1$ is necessary for the MSE map to converge to 0. This condition turns out to be sufficient because the function $\sigma^2 \mapsto \Psi(\sigma^2)$ is concave on $\mathbb{R}_+$. This indeed yields

$$\sigma_{t+1}^2 \leq \left.\frac{d\Psi}{d\sigma^2}\right|_{\sigma^2=0}\sigma_t^2, \qquad [16]$$

which implies exponential convergence to the correct solution (14). In particular we have

$$\rho_{\rm SE}(\delta;\chi,\lambda) = \rho_{\rm LS}(\delta;\chi,\lambda), \qquad [17]$$

whence $\rho_{\rm SE}(\delta;\chi,\lambda)$ is independent of F_X as claimed.

To prove $\sigma^2 \mapsto \Psi(\sigma^2)$ is concave, one computes its second derivative. In the case $\chi = +$, one needs to differentiate t the first derivative expression given above (*SI Appendix*). Two useful remarks follow. (*i*) The contribution due to $X = 0$ vanishes. (*ii*) Since a convex combination of concave functions is also concave, it is sufficient to consider the case in which $X = x_*$ deterministically.

As a by-product of this argument we obtain explicit expressions for the optimal tuning parameter, by maximizing the local stability threshold

$$\lambda_{+}(\delta) = \frac{1}{\sqrt{\delta}} \arg\max_{z \geq 0} \left\{ \frac{1 - (\kappa_{\chi}/\delta)[(1+z^2)\Phi(-z) - z\phi(z)]}{1 + z^2 - \kappa_{\chi}[(1+z^2)\Phi(-z) - z\phi(z)]} \right\}.$$

Before applying this formula in practice, please read the important notice in *SI Appendix*.

Discussion

Comparing Analytic Approaches. Refs. 10, 13, 14, and 23 analyzed iterative-thresholding-like algorithms and obtained rigorous results guaranteeing perfect recovery; the sparsity conditions they require are qualitatively correct but quantitatively are often considerably more stringent than what is truly necessary in practice. In contrast, we combine rigorous analysis of SE with extensive empirical work (documented in *SI Appendix*), to establish what really happens for our algorithm.

Relation with Minimax Risk. Let $\mathcal{F}_{\epsilon}^{\pm}$ be the class of probability distributions F supported on $(-\infty, \infty)$ with $\mathbb{P}\{X \neq 0\} \leq \epsilon$, and let $\eta(x; \lambda, \pm)$ denote the soft-threshold function (6) with threshold value λ. The minimax risk (12) is

$$M^{\pm}(\epsilon) \equiv \inf_{\lambda \geq 0} \sup_{F \in \mathcal{F}_{\epsilon}^{\pm}} \mathbb{E}_F\{[\eta(X + Z; \lambda, \pm) - X]^2\}, \qquad [18]$$

with $\lambda^{\pm}(\epsilon)$ the optimal λ. The optimal SE phase transition and optimal SE threshold obey

$$\delta = M^{\pm}(\rho\delta), \quad \rho = \rho_{\mathrm{SE}}(\delta; \pm). \qquad [19]$$

An analogous relation holds between the positive case $\rho_{\mathrm{SE}}(\delta; +)$, and the minimax threshold risk M^{+}, where F is constrained to be a distribution on $(0, \infty)$. Exploiting Eq. **19**, *SI Appendix* proves the high-undersampling limit:

$$\rho_{\mathrm{CG}}(\delta) = \rho_{\mathrm{SE}}(\delta)(1 + o(1)), \qquad \delta \to 0.$$

Other MP Algorithms. The nonlinearity $\eta(\cdot)$ in AMP Eqs. **1** and **2** might be chosen differently. For sufficiently regular such choices, the SE formalism might predict evolution of the MSE. One might hope to use SE to design better threshold nonlinearities. The threshold functions used render MSE maps $\sigma^2 \mapsto \Psi(\sigma^2)$ both monotone and concave. As a consequence, the phase transition line $\rho_{\mathrm{SE}}(\delta; \chi)$ for optimally tuned AMP is independent of the empirical distribution of the vector x_0. SE may be inaccurate without such properties.

Where SE is accurate, it offers limited room for improvement over the results here. If $\tilde{\rho}_{\mathrm{SE}}$ denotes a (hypothetical) phase transition derived by SE with any nonlinearity whatsoever, *SI Appendix* exploits Eq. **19** to prove

$$\tilde{\rho}_{\mathrm{SE}}(\delta; \chi) \leq \rho_{\mathrm{SE}}(\delta; \chi)(1 + o(1)), \qquad \delta \to 0, \quad \chi \in \{+, \pm\}.$$

In the limit of high undersampling, the nonlinearities studied here offer essentially unimprovable SE phase transitions. Our reconstruction experiments also suggest that other nonlinearities yield little improvement over thresholds used here.

Universality. The SE-derived phase transitions are not sensitive to the detailed distribution of coefficient amplitudes. Empirical results in *SI Appendix* find similar insensitivity of observed phase transitions for MP.

Gaussianity of the measurement matrix A can be relaxed; *SI Appendix* finds that other random matrix ensembles exhibit comparable phase transitions.

In applications, one often uses very large matrices A, which are never explicitly represented, but only applied as operators; examples include randomly undersampled partial Fourier transforms. *SI Appendix* finds that observed phase transitions for MP in the partial Fourier case are comparable to those for random A.

ACKNOWLEDGMENTS. We thank Iain Johnstone for helpful corrections and Microsoft Research New England for hospitality. A. Montanari was partially supported by the National Science Foundation CAREER Award CCF-0743978 and National Science Foundation Grant DMS-0806211. A. Maleki was partially supported by National Science Foundation Grant DMS-050530.

1. Baraniuk RG, Candès E, Nowak R, Vetterli M, eds (2008) Special Issue on Compressive Sampling. *IEEE Signal Processing Magazine* (IEEE, Los Alamitos, CA), Vol. 25, Issue 2.
2. Tropp J, Wright SJ (2009) Computational methods for sparse solution of linear inverse problems. arXiv:0907.3666v1.
3. Maleki A, Donoho DL (2009) Optimally tuned iterative thresholding algorithms for compressed sensing. arXiv:0909.0777.
4. Donoho DL (2006) High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Disc Comput Geom* 35:617–652.
5. Donoho DL, Tanner J (2005) Neighborliness of randomly-projected simplices in high dimensions. *Proc Natl Acad Sci USA* 102:9452–9457.
6. Donoho DL, Tanner J (2008) Counting faces of randomly projected hypercubes and orthants with applications. arXiv:0807.3590.
7. Donoho DL, Tanner J (2009) Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos Trans R Soc London Ser A* 367:4273–4293.
8. Herrity KK, Gilbert AC, Tropp JA (2006) Sparse approximation via iterative thresholding. *Proc IEEE Int Conf Acoust Speech Signal Proc* 3:624–627.
9. Tropp JA, Gilbert AC (2007) Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans Inf Theor* 53:4655–4666.
10. Indyk P, Ruzic M (2008) Near optimal sparse recovery in the l_1 norm. *Found Comput Sci* 199–207.
11. Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm Pure Appl Math* 75:1412–1457.
12. Donoho DL, Johnstone IM (1994) Minimax risk over l_p balls. *Prob Theor Rel Fields* 99:277–303.
13. Needel D, Tropp J (2008) CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl Comp Harm Anal* 26:301–321.
14. Dai W, Milenkovic O (2009) Subspace pursuit for compressive sensing signal reconstruction. arXiv:0803.0811v3.
15. Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Kaufmann, San Francisco).
16. Richardson TJ, Urbanke R (2008) *Modern Coding Theory* (Cambridge Univ Press, Cambridge, UK).
17. Lu Y, Montanari A, Prabhakar B, Dharmapurikar S, Kabbani A (2008) Counter braids: A novel counter architecture for per-flow measurement. *Proc 2008 ACM SIGMETRICS Int Conf Measur Model Comput Syst*, eds Liu Z, Misra V, Shenoy PJ (Assoc Comput Machinery, New York).
18. Sarvotham S, Baron D, Baraniuk R (2006) Compressed sensing reconstruction via belief propagation, preprint.
19. Zhang F, Pfister H (2009) On the iterative decoding of high-rate LDPC codes with applications in compressed sensing. arXiv:0903.2232v2.
20. Wainwright MJ, Jaakkola TS, Willsky AS (2005) MAP estimation via agreement on trees: Message-passing and linear programming. *IEEE Trans Inf Theor* 51:3697–3717.
21. Bayati M, Shah D, Sharma M (2008) Max-product for maximum weight matching: Convergence, correctness, and LP duality. *IEEE Trans Inf Theor* 54:1241–1251.
22. Thouless DJ, Anderson PW, Palmer RG (1977) Solution of 'Solvable model of a spin glass'. *Philos Mag* 35:593–601.
23. Jafarpour S, Xu W, Hassibi B, Calderbank AR (2008) Efficient and robust compressed sensing using high-quality expander graphs. *Comput Res Reposit* abs/0806.3802.