Supporting Information to: Message Passing Algorithms for Compressed Sensing

David L. Donoho[*] Arian Maleki[†] and Andrea Montanari[*,†]

July 20, 2009;
Revised after Referee Comments August 16, 2009

Abstract

This document presents details concerning analytical derivations and numerical experiments that support the claims made in the main text 'Message Passing Algorithms for Compressed Sensing', submitted for publication in the *Proceedings of the National Academy of Sciences*, USA. Hereafter 'main text'.

One can find here:

- Derivations of explicit Formulas for the MSE Map, and the optimal thresholds; see Section 3 below.
- Proof of Theorem 1; see Section 3 below.
- Proof of concavity of the MSE Map, see again Section 3 below.
- Explanation of the connection between Minimax Thresholding, Minimax Risk, and rigorous proof of formula [19] in the main text; see Section 4 below, Theorem 4.2.
- Formulas for the rate exponent b of Theorem 2 in the main text, expressed in terms of the minimax threshold risk; see Section 5 below.
- Proof of Finding 1 of the main text in the case $\chi = \square$, as referred to after Finding 1, Page 3 in the main text. See Theorem 3.1 below.
- Proof of Finding 1 of the main text in the case $\chi = +, \pm$, at least in the asymptotic sense $\delta \to 0$, as referred to after Finding 1, Page 3, in the main text, and after equation 19 in the main text. See Section 6 below.
- Proof that other nonlinearities do not improve the State Evolution phase transition, as claimed in subsection "Other Message Passing Algorithms" on Page 6 of the main text; see Section 7 below.
- The experimental protocol that we followed – Section 8 below.
- The technique we used for estimating phase transitions; Section 9 below.
- Our method for tuning of competing algorithms; Section 10 below.
- Presentation of Phase Transition Results – Section 11.
- Examples Documenting the Interference Heuristic discussed in section "Heuristics for Iterative Approaches" of the main text – Section 12.
- Examples Documenting Finding 2 of the Main text – Section 13.
- Results on Universality to Coefficient Ensembles – Section 14.
- Results on Universality across Matrix Ensembles – Section 15.
- Timing Comparisons – Section 16.
- Postscript: Relationship with Stojnic's Bounds – Section 17.

[*]Department of Statistics, Stanford University

[†]Department of Electrical Engineering, Stanford University

1 Important Notice

Readers familiar with the literature of thresholding of sparse signals will want to know that an implicit rescaling is needed to match equations from that literature with equations here. Specifically, in the traditional literature, one is used to seeing expressions $\eta(x;\lambda\sigma)$ in cases where σ is the standard deviation of an underlying normal distribution. This means the threshold λ is specified in standard deviations, so many people will immediately understand values like of $\lambda = 2, 3$ etc in terms of their false alarm rates. In the main text, the expression $\eta(x;\lambda\sigma)$ appears numerous times, but note that σ is not the standard deviation of the relevant normal distribution; instead, the standard deviation of that normal is $\tau = \sigma/\sqrt{\delta}$. It follows that λ in the main text is calibrated differently from the way λ would be calibrated in other sources, differing by a δ-dependent scale factor. If we let λ_{SE}^{sd} denote the quantity λ_{SE} appropriately rescaled so that it is in units of standard deviations of the underlying normal distribution, then the needed conversion to sd units is

$$\lambda_{SE}^{sd} = \lambda_{SE} \cdot \sqrt{\delta}. \tag{1}$$

2 A summary of notation

The main paper will be referred as DMM throughout this note. All the notations are consistent with the notations used in DMM. We will use repeatedly the notation $\epsilon = \delta\rho$.

3 State Evolution Formulas

In the main text we mentioned $\rho_{\text{SE}}(\delta;\chi,\lambda,F_X)$ is independent of F_X. We also mentioned a few formulas for $\rho_{\text{SE}}(\delta;\chi)$. The goal of this section is to explain the calculations involved in deriving these results. First, recall the expression for the MSE map

$$\Psi(\sigma^2) = \mathbb{E}\Big\{\big(\eta(X+\frac{\sigma}{\sqrt{\delta}}Z;\lambda\sigma,\chi)-X\big)^2\Big\}. \tag{2}$$

We denote by $\partial_1\eta$ and $\partial_2\eta$ the partial derivatives of η with respect to its first and second arguments. Using Stein's lemma, we get

$$\begin{aligned}\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2} = \frac{1}{\delta}\,\mathbb{E}\Big\{\partial_1\eta(X+\frac{\sigma}{\sqrt{\delta}}Z;\lambda\sigma)^2\Big\} + \frac{1}{\delta}\mathbb{E}\Big\{\big[\eta(X+\frac{\sigma}{\sqrt{\delta}}Z;\lambda\sigma)-X\big]\partial_1^2\eta(X+\frac{\sigma}{\sqrt{\delta}}Z;\lambda\sigma)\Big\} \\ + \frac{\lambda}{\sigma}\mathbb{E}\Big\{\big[\eta(X+\frac{\sigma}{\sqrt{\delta}}Z;\lambda\sigma)-X\big]\partial_2\eta(X+\frac{\sigma}{\sqrt{\delta}}Z;\lambda\sigma)\Big\},\end{aligned} \tag{3}$$

where we dropped the dependence of $\eta(\,\cdot\,)$ on the constraint χ to simplify the formula. Notice that, for our choices of η, $\partial_1\eta(\,\cdot\,)$, $\partial_2\eta(\,\cdot\,)$ exist almost everywhere. Since Z has a density, this is sufficient to make their expectations well defined. On the other hand $\partial_1^2\eta(\,\cdot\,)$ is interpreted in the sense of distributions.

3.1 Case $\chi = +$

In this case we have $X \geq 0$ almost surely and the threshold function is

$$\eta(x;\lambda\sigma) = \begin{cases} (x-\lambda\sigma) & \text{if } x \geq \lambda\sigma, \\ 0 & \text{otherwise.}\end{cases}$$

As a consequence $\partial_1\eta(x;\lambda\sigma) = -\partial_2\eta(x;\lambda\sigma) = \mathbb{I}(x \geq \lambda\sigma)$ (almost everywhere). Further $\partial_1^2\eta(\,\cdot\,;\lambda\sigma) = \delta_{\lambda\sigma}$. (Recall that this means $\int \partial_1^2\eta(x;\lambda\sigma)\, f(x)\,\mathrm{d}x = f(\lambda\sigma)$ for any smooth f.) This yields

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

As $\sigma \downarrow 0$, we have $\Phi\Big(\frac{\sqrt{\delta}}{\sigma}(X - \lambda\sigma)\Big) \to 1$ and $\phi\Big(\frac{\sqrt{\delta}}{\sigma}(X - \lambda\sigma)\Big) \to 0$ if $X > 0$. Therefore,

$$\left.\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2}\right|_0 = \left(\frac{1}{\delta} + \lambda^2\right)\rho\delta + \left(\frac{1}{\delta} + \lambda^2\right)\,(1 - \rho\delta)\,\Phi(-\lambda\sqrt{\delta}) - \frac{\lambda}{\sqrt{\delta}}(1 - \rho\delta)\,\phi(-\lambda\sqrt{\delta})\,.$$

The local stability threshold $\rho_{\mathrm{LS}}(\delta;+,\lambda)$ is obtained by setting $\left.\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2}\right|_0 = 1$.

In order to prove the concavity of $\sigma^2 \mapsto \Psi(\sigma^2)$ first notice that a convex combination of concave functions is concave and so it is sufficient to show the concavity in the case $X = x \geq 0$ deterministically. Next notice that, in the case $x = 0$, $\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2}$ is independent of σ^2. A a consequence, it is sufficient to prove $\frac{\mathrm{d}^2\Psi_x}{\mathrm{d}(\sigma^2)^2} \leq 0$ where

$$\delta\,\frac{\mathrm{d}\Psi_x}{\mathrm{d}\sigma^2} = \big(1 + \lambda^2\delta\big)\,\Phi\Big(\frac{\sqrt{\delta}}{\sigma}(x - \lambda\sigma)\Big) - \frac{\sqrt{\delta}}{\sigma}(x + \lambda\sigma)\,\phi\Big(\frac{\sqrt{\delta}}{\sigma}(x - \lambda\sigma)\Big)\,.$$

Using $\Phi'(u) = \phi(u)$ and $\phi'(u) = -u\phi(u)$, we get

$$\begin{aligned}
\delta\,\frac{\mathrm{d}^2\Psi_x}{\mathrm{d}(\sigma^2)^2} &= -\frac{x\sqrt{\delta}}{2\sigma^3}\left\{(1 + \lambda^2\delta)\,\phi\Big(\frac{\sqrt{\delta}}{\sigma}(x - \lambda\sigma)\Big) + \frac{\delta}{\sigma^2}(x^2 - \lambda^2\sigma^2)\,\phi\Big(\frac{\sqrt{\delta}}{\sigma}(x - \lambda\sigma)\Big) - \phi\Big(\frac{\sqrt{\delta}}{\sigma}(x - \lambda\sigma)\Big)\right\} \\
&= -\frac{x\sqrt{\delta}}{2\sigma^3}\left\{\frac{\delta}{\sigma^2}\,x^2\right\}\phi\Big(\frac{\sqrt{\delta}}{\sigma}(x - \lambda\sigma)\Big) < 0 \qquad (4)
\end{aligned}$$

for $x > 0$.

3.2 Case $\chi = \pm$

Here X is supported on $(-\infty, \infty)$ with $\mathbb{P}\{X \neq 0\} \leq \epsilon = \rho\delta$. Recall the definition of soft threshold

$$\eta(x;\lambda\sigma) = \begin{cases} (x - \lambda\sigma) & \text{if } x \geq \lambda\sigma, \\ (x + \lambda\sigma) & \text{if } x \leq -\lambda\sigma, \\ 0 & \text{otherwise.} \end{cases}$$

As a consequence $\partial_1\eta(x;\lambda\sigma) = \mathbb{I}(|x| \geq \lambda\sigma)$ and $\partial_2\eta(x;\lambda\sigma) = -\mathrm{sign}(x)\mathbb{I}(|x| \geq \lambda\sigma)$. Further $\partial_1^2\eta(\,\cdot\,;\lambda\sigma) = \delta_{\lambda\sigma} - \delta_{-\lambda\sigma}$. This yields

$$\begin{aligned}
\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2} &= \left(\frac{1}{\delta} + \lambda^2\right)\mathbb{E}\Big\{\Phi\Big(\frac{\sqrt{\delta}}{\sigma}(X - \lambda\sigma)\Big) + \Phi\Big(-\frac{\sqrt{\delta}}{\sigma}(X + \lambda\sigma)\Big)\Big\} \\
&\quad -\frac{\lambda}{\sqrt{\delta}}\,\mathbb{E}\Big\{\phi\Big(\frac{\sqrt{\delta}}{\sigma}(X - \lambda\sigma)\Big) + \phi\Big(\frac{\sqrt{\delta}}{\sigma}(X + \lambda\sigma)\Big)\Big\} \\
&\quad -\frac{1}{\sigma\sqrt{\delta}}\,\mathbb{E}\Big\{X\Big[\phi\Big(\frac{\sqrt{\delta}}{\sigma}(X - \lambda\sigma)\Big) - \phi\Big(\frac{\sqrt{\delta}}{\sigma}(X + \lambda\sigma)\Big)\Big]\Big\}\,.
\end{aligned}$$

By letting $\sigma \downarrow 0$ we get

$$\left.\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2}\right|_0 = \left(\frac{1}{\delta} + \lambda^2\right)\rho\delta + \left(\frac{1}{\delta} + \lambda^2\right)\,(1 - \rho\delta)\,2\,\Phi(-\lambda\sqrt{\delta}) - \frac{\lambda}{\sqrt{\delta}}(1 - \rho\delta)\,2\,\phi(-\lambda\sqrt{\delta})\,,$$

which yields the local stability threshold $\rho_{\mathrm{LS}}(\delta;\pm,\lambda)$ by $\left.\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2}\right|_0 = 1$.

Finally the proof of the concavity of $\sigma^2 \mapsto \Psi(\sigma^2)$ is completely analogous to the case $\chi = +$.

3.3 Case $\chi = \square$

Finally consider the case of X supported on $[-1,+1]$ with $\mathbb{P}\{X \notin \{+1,-1\}\} \le \epsilon$. In this case we proposed the following nonlinearity,

$$\eta(x) = \begin{cases} +1 & \text{if } x > +1, \\ x & \text{if } -1 \le x \le +1, \\ -1 & \text{if } x \le -1. \end{cases}$$

Notice that the nonlinearity does not depend on any threshold parameter. We have $\partial_1\eta(x) = \mathbb{I}(x \in [-1,+1])$ and $\partial_1^2\eta(x) = -\delta_1 + \delta_{-1}$. A simple calculation yields

$$\begin{aligned} \frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2} &= \frac{1}{\delta}\,\mathbb{P}\Big\{X + \frac{\sigma}{\sqrt{\delta}}Z \in [-1,+1]\Big\} \\ &= \frac{1}{\delta}\,\mathbb{E}\Big\{\Phi\Big(\frac{\sqrt{\delta}}{\sigma}(1-X)\Big) - \Phi\Big(-\frac{\sqrt{\delta}}{\sigma}(1+X)\Big)\Big\}. \end{aligned}$$

As $\sigma \downarrow 0$ we get

$$\left.\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2}\right|_0 = \frac{1}{2\delta}(1+\rho\delta)\,,$$

whence the local stability condition $\left.\frac{\mathrm{d}\Psi}{\mathrm{d}\sigma^2}\right|_0 < 1$ yields $\rho_{\mathrm{LS}}(\delta;\square) = (2-\delta^{-1})_+$.

Concavity of $\sigma^2 \mapsto \Psi(\sigma^2)$ immediately follows from the fact that $\Phi(\frac{\sqrt{\delta}}{\sigma}(1-x))$ is non-increasing in σ for $x \le 1$ and $\Phi(-\frac{\sqrt{\delta}}{\sigma}(1+x))$ is non-decreasing for $x \ge -1$. Using the combinatorial geometry result of [1] we get

Theorem 3.1. *For any $\delta \in [0,1]$,*

$$\rho_{\mathrm{CG}}(\delta;\square) = \rho_{\mathrm{SE}}(\delta;\square) = \rho_{\mathrm{LS}}(\delta;\square) = \max\left\{0, 2-\delta^{-1}\right\}. \tag{5}$$

4 Relation to Minimax Thresholding

4.1 Minimax Thresholding Policy

We denote by $\mathcal{F}_\epsilon^+$ the collection of all CDF's supported in $[0,\infty)$ and with $F(0) \ge 1-\epsilon$, and by $\mathcal{F}_\epsilon^\pm$ the collection of all CDF's supported in $(-\infty,\infty)$ and with $F(0+) - F(0-) \ge 1-\epsilon$. For $\chi \in \{+,\pm\}$, define the minimax threshold MSE

$$M^*(\epsilon;\chi) = \inf_\lambda \sup_{F \in \mathcal{F}_\epsilon^\chi} \mathbb{E}_F\left\{\eta(X+Z;\lambda,\chi) - X)^2\right\}, \tag{6}$$

where $\mathbb{E}_F$ denote expectation with respect to the random variable X with distribution F, and $\eta(x;\lambda) = \mathrm{sign}(x)(|x|-\lambda)_+$ for $\chi = \pm$ and $\eta(x;\lambda) = (x-\lambda)_+$ for $\chi = +$. *Minimax Thresholding* was discussed for the case $\chi = +$ in [2] and for $\chi = \pm$ in [3, 4].

This machinery gives us a way to look at the results derived above in very commonsense terms. Suppose we know δ and ρ but *not* the distribution F of X. Let's consider what threshold one might use, and ask at each given iteration of SE, the threshold which gives us the best possible control of the resulting formal MSE. That best possible threshold λ^t is by definition the minimax threshold at nonzero fraction $\epsilon = \rho\delta$, appropriately scaled by the effective noise level $\tau = \sigma/\sqrt{\delta}$,

$$\lambda^t = \lambda^*(\rho\cdot\delta;\chi)\cdot\sigma/\sqrt{\delta},$$

where $\chi \in \{+, \pm\}$ depending on the case at hand. Note that this threshold does not depend on F. It depends on iteration only through the effective noise level at that iteration. The guarantee we then get for the formal MSE is the minimax threshold risk, appropriately scaled by the square of the effective noise level:

$$\mathsf{MSE} \le M^*(\rho\delta;\chi)\cdot\tau^2 = M^*(\rho\delta;\chi)\frac{\sigma^2}{\delta},\ . \tag{7}$$

for $\chi \in \{+, \pm\}$. This guarantee gives us a reduction in MSE over the previous iteration if and only if the right-hand side in Eq. (7) is smaller than σ^2, i.e. if and only if

$$M^*(\rho\delta;\chi) < \delta\,, \qquad \chi \in \{+, \pm\}.$$

In short, we can use state evolution with the minimax threshold, appropriately scaled by effective noise level, and we get a guaranteed fractional reduction in MSE at each iteration, with fractional improvement

$$\omega_{\mathrm{MM}}(\delta,\rho;\chi) = (1 - M^*(\rho\delta;\chi)/\delta); \tag{8}$$

hence the formal SE evolution is bounded by:

$$\sigma_t^2 \le \omega_{\mathrm{MM}}(\delta,\rho;\chi)^t \cdot \mathbb{E}X^2, \qquad t = 1, 2, \ldots. \tag{9}$$

Results analogous to those of the main text hold for this minimax thresholding policy. That is, we can define a minimax thresholding phase transition such that below that transition, state evolution with minimax thresholding converges:

$$\rho_{\mathrm{MM}}(\delta;\chi) = \sup\{\rho : M^*(\rho\delta;\chi) < \delta\}; \qquad \chi \in \{+, \pm\}.$$

Theorem 4.1. *Under SE with the minimax thresholding policy described above, for each (δ,ρ) in $(0,1)^2$ obeying $\rho < \rho_{\mathrm{MM}}(\delta;\chi)$, and for every marginal distribution $F \in \mathcal{F}_\epsilon^\chi$, the formal MSE evolves to zero, with dynamics bounded by (8)- (9).*

4.2 Relating Optimal Thresholding to Minimax Thresholding

An important difference between the optimal threshold defined in the main text and the minimax threshold is that $\lambda_\chi = \lambda_\chi(\delta)$ depends only on the assumed δ – no specific ρ need be chosen while minimax thresholding as defined above requires that one specify both δ and ρ. However, since the methodology is seemingly pointless above the minimax phase transition, one might think to specify $\rho = \rho_{\mathrm{MM}}(\delta;\chi)$. This new threshold $\lambda_{\mathrm{MM}}(\delta;\chi) = \lambda^*(\delta\rho_{\mathrm{MM}}(\delta);\chi)$ then requires no specification of ρ. As it turns out, the SE threshold coincides with this new threshold.

Theorem 4.2. *For $\chi \in \{+, \pm\}$ and $\delta \in [0,1]$*

$$M^*(\rho\delta;\chi) = \delta \quad \text{ if and only if } \quad \rho = \rho_{\mathrm{SE}}(\delta;\chi)\,. \tag{10}$$

Let $\lambda_\chi(\delta)$ denote the minimax threshold defined in the main text, and let $\lambda_\chi^{sd}(\delta)$ denote the same quantity expressed in sd units (1). Then

$$\lambda_\chi^{sd}(\delta) = \lambda^\chi(\rho\delta), \qquad \rho = \rho_{\mathrm{SE}}(\delta;\chi), \qquad \chi \in \{+, \pm\}$$

Proof. It is convenient to introduce the following explicit notation for the MSE map:

$$\Psi(\sigma^2;\delta,\lambda,F) = \mathbb{E}_F\Big\{\big(\eta(X+\frac{\sigma}{\sqrt{\delta}}Z;\lambda\sigma)-X\big)^2\Big\}, \tag{11}$$

where $Z\sim\mathsf{N}(0,1)$ is independent of X, and $X\sim F$. As above, we drop the dependency of the threshold function on $\chi\in\{+,\pm\}$ Since $\eta(ax;a\lambda)=a\,\eta(x;\lambda)$ for any positive a, we have the scale invariance

$$\Psi(\sigma^2;\delta,\lambda,F,\chi) = \frac{\sigma^2}{\delta}\Psi(1;1,\lambda\sqrt{\delta},S_{\delta^{1/2}/\sigma}F), \tag{12}$$

where $(S_aF)(x)=F(x/a)$ is the operator that takes the CDF of a random variable X and returns the CDF of the random variable aX.

Define

$$J(\delta,\rho;\chi) = \inf_{\lambda\ge 0}\sup_{F\in\mathcal{F}_\epsilon^\chi}\sup_{\sigma^2\in(0,\mathbb{E}_F\{X^2\}]}\frac{1}{\sigma^2}\,\Psi(\sigma^2;\delta,\lambda,F,\chi)\,, \tag{13}$$

where $\epsilon\equiv\rho\delta$. It follows from the definition of SE threshold that $\rho<\rho_{\rm SE}(\delta;\chi)$ if and only if $J(\delta,\rho;\chi)<1$. We first notice that by concavity of $\sigma^2\mapsto\Psi(\sigma^2;\delta,\lambda,F,\chi)$, we have

$$\begin{aligned}
J(\delta,\rho;\chi) &= \inf_{\lambda}\sup_{F\in\mathcal{F}_\epsilon^\chi}\sup_{\sigma^2>0}\frac{1}{\sigma^2}\,\Psi(\sigma^2;\delta,\lambda,F,\chi) &(14)\\
&= \frac{1}{\delta}\inf_{\lambda}\sup_{F\in\mathcal{F}_\epsilon^\chi}\sup_{\sigma^2>0}\Psi(1;1,\lambda\sqrt{\delta},S_{\delta^{1/2}/\sigma}F) &(15)\\
&= \frac{1}{\delta}\inf_{\lambda}\sup_{F\in\mathcal{F}_\epsilon^P}\Psi(1;1,\lambda,F) &(16)
\end{aligned}$$

where the second identity follows from the invariance property and the third from the observation that $S_a\mathcal{F}_\epsilon^\chi=\mathcal{F}_\epsilon^\chi$ for any $a>0$. Comparing with the definition (6), we finally obtain

$$J(\delta,\rho;\chi) = \frac{1}{\delta}\,M^*(\delta\rho;\chi)\,. \tag{17}$$

Therefore $\rho<\rho_{\rm SE}(\delta;\chi)$ if and only $\delta>M^*(\delta\rho;\chi)$, which implies the thesis. □

5 Convergence Rate of State Evolution

The optimal thresholding policy described in the main text is the same as using the minimax thresholding policy but instead assuming the most pessimistic possible choice of ρ – the largest ρ that can possibly make sense. In contrast minimax thresholding is ρ-adaptive, and can use a smaller threshold where it would be valuable. Below the SE phase transition, both methods will converge, so what's different?

Note that $\lambda_{\rm SE}(\delta;\chi)$ and $\lambda_{MM}(\delta,\rho;\chi)$ are dimensionally different; λ_{MM} is in standard deviation units. Converting $\lambda_{\rm SE}$ into sd units by (1), we have $\lambda^{sd}_{\rm SE}=\lambda_{\rm SE}\cdot\delta^{1/2}$. Even after this calibration, we find that methods will generally use different thresholds, i.e. if $\rho<\rho_{\rm SE}$,

$$\lambda_{\rm MM}(\delta,\rho;\chi)\neq\lambda^{sd}_{\rm SE}(\delta;\chi),\qquad \chi\in\{+,\pm\}.$$

In consequence, the methods may have different rates of convergence. Define the worst-case threshold MSE

$$\mathsf{MSE}(\epsilon,\lambda;\chi) = \sup_{F\in\mathcal{F}_\epsilon^\chi}\mathbb{E}_F\left\{\eta(X+Z;\lambda)-X\right)^2\}$$

and set

$$M_{\mathrm{SE}}(\delta, \rho; \chi) = \mathsf{MSE}(\delta\rho, \lambda^{sd}_{\mathrm{SE}}(\delta, \chi); \chi).$$

This is the MSE guarantee achieved by using $\lambda^{sd}_{\mathrm{SE}}(\delta)$ when in fact (δ, ρ) is the case. Now by definition of minimax threshold MSE,

$$M_{\mathrm{SE}}(\delta, \rho; \chi) \geq M^*(\delta\rho; \chi); \tag{18}$$

the inequality is generally strict. The convergence rate of optimal AMP under SE was described implicitly in the main text. We can give more precise information using this notation. Define

$$\omega_{\mathrm{SE}}(\delta, \rho; \chi) = (1 - M_{\mathrm{SE}}(\delta, \rho; \chi)/\delta);$$

Then we have for the formal MSE of AMP

$$\sigma_t^2 \leq \omega_{\mathrm{SE}}(\delta, \rho; \chi)^t \cdot \mathbb{E}X^2, \qquad t = 1, 2, 3, \ldots$$

In the main text, the same relation was written in terms of $\exp(-bt)$, with $b > 0$; here we see that we may take $b(\delta, \rho) = -\log(\omega_{\mathrm{SE}}(\delta, \rho))$. Explicit evaluation of this b requires evaluation of the worst-case thersholding risk $\mathsf{MSE}(\epsilon, \lambda)$. Now by (18) we have

$$\omega_{\mathrm{SE}}(\delta, \rho; \chi) \geq \omega_{\mathrm{MM}}(\delta, \rho; \chi),$$

generally with strict inequality; so by using the ρ-adaptive threshold one gets better convergence rates.

6 Rigorous Asymptotic Agreement of SE and CG

In this section we prove

Theorem 6.1. *For $\chi \in \{+, \pm\}$*

$$\lim_{\delta \to 0} \frac{\rho_{\mathrm{CG}}(\delta; \chi)}{\rho_{\mathrm{SE}}(\delta; \chi)} = 1. \tag{19}$$

In words, $\rho_{\mathrm{CG}}(\delta; \chi)$ is the phase transition computed by combinatorial geometry (polytope theory) and $\rho_{\mathrm{SE}}(\delta, \chi)$ obtained by state evolution: they are rigorously equivalent in the highly undersampled limit (i.e. $\delta \to 0$ limit). In the main text, we only can make the observation that they agree numerically.

6.1 Properties of the minimax threshold

We summarize here several known properties of the minimax threshold (6), which provide useful information about the behavior of SE.

The extremal F achieving the supremum in Eq. (6) is known. In the case $\chi = +$, it is a two-point mixture

$$F_\epsilon^+ = (1 - \epsilon)\, \delta_0 + \epsilon\, \delta_{\mu^+(\epsilon)}\,. \tag{20}$$

In the signed case $\chi = \pm$, it is a three-point symmetric mixture

$$F_\epsilon^\pm = (1 - \epsilon)\, \delta_0 + \frac{\epsilon}{2}\, (\delta_{\mu^\pm(\epsilon)} + \delta_{-\mu^\pm(\epsilon)})\,. \tag{21}$$

Precise asymptotic expressions for $\mu^\chi(\epsilon)$ are available. In particular, for $\chi \in \{+, \pm\}$,

$$\mu^\chi(\epsilon) = \sqrt{2\log(1/\epsilon)}(1 + o(1)) \qquad \text{as } \epsilon \to 0\,. \tag{22}$$

We also know that

$$M^*(\epsilon; \chi) = 2\log(\epsilon^{-1})(1 + o(1)) \qquad \text{as } \epsilon \to 0\,. \tag{23}$$

6.2 Proof of Theorem 6.1

Combining Theorem 4.2 and Eq. (23), we get

$$\rho_{\mathrm{SE}}(\delta;\rho) \sim \frac{1}{2\log(\delta^{-1})}, \qquad \delta \to 0\,. \tag{24}$$

(correction terms that can be explicitly given). Now we know rigorously from [5] that the LP-based phase transitions satisfy a similar relationship:

Theorem 6.2 (Donoho and Tanner [5])**.** *For* $\chi \in \{+,\pm\}$

$$\rho_{\mathrm{CG}}(\delta,\chi) \sim \frac{1}{2\log(\delta^{-1})}, \qquad \delta \to 0. \tag{25}$$

Combining now with eqn. 24 we get Theorem 6.1. □

7 Rigorous Asymptotic Optimality of Soft Thresholding

The discussion in the main text alluded to the possibility of improving on soft thresholding. Here we give a more formal discussion. We work in the situations $\chi \in \{+,\pm\}$. Let $\widetilde{\eta}$ denote some arbitrary nonlinearity with tuning parameter λ. (For a concrete example, think of hard thresholding). We can define the minimax MSE for this nonlinearity in the natural way

$$\widetilde{M}(\epsilon;\chi) = \inf_{\lambda} \sup_{F \in \mathcal{F}_\epsilon^\chi} \mathbb{E}_F \left\{\widetilde{\eta}(X+Z;\lambda) - X)^2\right\}\,, . \tag{26}$$

there is a corresponding minimax threshold $\widetilde{\lambda}(\epsilon;\chi)$. We can deploy the minimax threshold in AMP by setting $\epsilon = \rho\delta$ and rescaling the threshold by the effective noise level $\tau = \sigma/\sqrt{\delta}$:

$$\begin{aligned} \text{actual threshold at iteration } t &= \widetilde{\lambda}(\epsilon;\chi)\cdot\tau \\ &= \widetilde{\lambda}(\rho\delta;\chi)\cdot\sigma_t/\sqrt{\delta}. \end{aligned}$$

Under state evolution, this is guaranteed to reduce the MSE provided

$$\widetilde{M}(\rho\delta;\chi) < \delta.$$

In that case we get full evolution to zero. It makes sense to define the minimax phase transition:

$$\widetilde{\rho}_{\mathrm{SE}}(\delta;\chi) = \sup\{\rho \;:\; \widetilde{M}(\rho\delta;\chi) < \delta\}; \qquad \chi \in \{+,\pm\}.$$

Whatever be F, for (δ,ρ) with $\rho < \widetilde{\rho}_{\mathrm{SE}}(\delta)$, SE evolves the formal MSE of $\widetilde{\eta}$ to zero.

It is tempting to hope that some very special nonlinearity can do substantially better than soft thresholding. At least for the minimax phase transition, this is not so:

Theorem 7.1. *Let* $\widetilde{\rho}_{\mathrm{MM}}(\delta;\chi)$ *be a minimax phase transition computed under the State Evolution formalism for the cases* $\chi \in \{+,\pm)$ *with some scalar nonlinearity* $\widetilde{\eta}$*. Let* $\rho_{\mathrm{SE}}(\delta;\chi)$ *be the phase transition calculated in the main text for soft thresholding with corresponding optimal* λ*. Then for* $\chi \in \{+,\pm\}$

$$\lim_{\delta\to 0} \frac{\widetilde{\rho}_{\mathrm{SE}}(\delta;\chi)}{\rho_{\mathrm{SE}}(\delta;\chi)} \le 1.$$

In words, no other nonlinearity can outperform soft thresholding in the limit of extreme undersampling – in the sense of minimax phase transitions. This is best understood using a notion from the main text. We there said that the parameter space (δ,ρ,λ,F) can be partitioned into two regions. Region (I), where zero is the unique fixed point of the MSE map, and is a stable fixed point; and its complement, Region (II). Theorem 7.1 says that the range of ρ guaranteeing membership in Region (I) cannot be dramatically expanded by using a different nonlinearity.

7.1 Some results on Minimax Risk

The proof depends on some known results about minimax MSE, where we are allowed to choose not just the threshold, but also the nonlinearity. For $\chi \in \{+, \pm\}$, define the minimax MSE

$$M^{\star\star}(\epsilon;\chi) = \inf_{\widetilde{\eta}} \sup_{F \in \mathcal{F}_\epsilon^\chi} \mathbb{E}_F \left\{ \widetilde{\eta}(X+Z) - X)^2 \right\} , \tag{27}$$

Here the minimization is over *all measurable functions* $\widetilde{\eta} : \mathbf{R} \mapsto \mathbf{R}$. *Minimax MSE* was discussed for the case $\chi = +$ in [2] and for $\chi = \pm$ in [6, 3, 4]. It is known that

$$M^{\star\star}(\epsilon;\chi) \sim 2\log(\epsilon^{-1}). \qquad \epsilon \to 0. \tag{28}$$

7.2 Proof of Theorem 7.1

Evidently, any specific nonlinearity cannot do better than the minimax risk:

$$\widetilde{M}^*(\epsilon) \geq M^{**}(\epsilon;\chi).$$

Consequently, if we put

$$\rho^{**}(\delta;\chi) = \sup\{\rho : \ M^{**}(\delta\rho;\chi) < \delta\}$$

then

$$\widetilde{\rho}^*(\delta,\chi) \leq \rho^{\star\star}(\delta,\chi).$$

From (28) and the last two displays we conclude

$$\widetilde{\rho}^*(\delta;\chi) \leq \frac{1}{2\log(1/\delta)} \sim \rho_{\mathrm{SE}}(\delta,\chi), \qquad \delta \to 0.$$

Theorem 7.1 is proven. □

8 Data Generation

For a given algorithm with a fully specified parameter vector, we conduct one phase transition measurement experiment as follows. We fix a *problem suite*, i.e. a matrix ensemble and a coefficient distribution for generating problem instances (A, x_0). We also fix a grid of δ values in $[0, 1]$, typically 30 values equispaced between 0.02 and 0.99. Subordinate to this grid, we consider a series of ρ values. Two cases arise frequently:

- *Focused Search design.* 20 values between $\rho_{\mathrm{CG}}(\delta;\chi) - 1/10$ and $\rho_{\mathrm{CG}}(\delta;\chi) + 1/10$, where ρ_{CG} is the theoretically expected phase transition deriving from combinatorial geometry (according to case $\chi \in \{+, \pm, \square\}$).
- *General Search design.* 30 values equispaced between 0 and 1.

We then have a (possibly non-cartesian) grid of δ, ρ values in parameter space $[0, 1]^2$. At each (δ, ρ) combination, we will take M problem instances; in our case $M = 20$. We also fix a measure of success; see below.

Once we specify the problem size N, the experiment is now fully specified; we set $n = \lceil \delta N \rceil$ and $k = \lceil \rho n \rceil$, and generate M problem instances, and obtain M algorithm outputs $\hat{x}_i$, and M success indicators S_i, $i = 1, \ldots M$.

A problem instance (y, A, x_0) consists of $n \times N$ matrix A from the given matrix ensemble and a k-sparse vector x_0 from the given coefficient ensemble. Then $y = Ax_0$. The algorithm is called with problem instance (y, A) and it produces a result $\hat{x}$. We declare success if

$$\frac{\|x_0 - \hat{x}\|_2}{\|x_0\|_2} \leq \texttt{tol},$$

where `tol` is a given parameter; in our case 10^{-4}; the variable S_i indicates success on the i-th Monte Carlo realization. To summarize all M Monte Carlo repetitions, we set $S = \sum_i S_i$.

The result of such an experiment is a dataset with tuples (N, n, k, M, S); each tuple giving the results at one combination (ρ, δ). The meta-information describing the experiment is the specification of the algorithm with all its parameters, the problem suite, and the success measure with its tolerance.

9 Estimating Phase Transitions

From such a dataset we find the location of the phase transition as follows. Corresponding to each fixed value of δ in our grid, we have a collection of tuples (N, n, k, M, S) with $n/N = \delta$ and varying k. Pretending that our random number generator makes truly independent random numbers, the result S at one experiment is binomial $\mathsf{Bin}(\pi, M)$, where the success probability $\pi \in [0, 1]$. Extensive prior experiments show that this probability varies from 1 when ρ is well below ρ_{CG} to 0 when ρ is well above ρ_{CG}. In short, the success probability

$$\pi = \pi(\rho|\delta; N).$$

We define the *finite-N phase transition* as the value of ρ at which success probability is 50%:

$$\pi(\rho|\delta; N) = \frac{1}{2} \quad \text{at} \quad \rho = \rho(\delta).$$

This notion is well-known in biometrics where the 50% point of the dose-response is called the LD50. (Actually we have the implicit dependence $\rho(\delta) \equiv \rho(\delta|N, \texttt{tol})$; the tolerance in the success definition has a (usually slight) effect, as well as the problem size N)

To estimate the phase transition from data, we model dependence of success probability on ρ using generalized linear models (GLMs). We take a δ-constant slice of the dataset obtaining triples $(k, M, S(k, n, N))$, and model $S(k, n, N) \sim \mathsf{Bin}(\pi_k; M)$ where the success probabilities obeys a generalized linear model with logistic link

$$\mathrm{logit}(\pi) = a + b\rho$$

where $\rho = k/n$; in biometric language, we are modeling that the dose-response probability, where ρ is the 'complexity-dose', follows a logistic curve.

In terms of the fitted parameters $\hat{a}$,$\hat{b}$, we have the estimated phase transition

$$\hat{\rho}(\delta) = -\hat{a}/\hat{b},$$

and the estimated transition width is

$$\hat{w}(\delta) = 1/\hat{b}.$$

Note that, actually,

$$\hat{\rho}(\delta) = \hat{\rho}(\delta|N, \texttt{tol}), \qquad \hat{w}(\delta) = \hat{w}(\delta|N, \texttt{tol})\,.$$

We may be able to see the phase transition and its width varying with N and with the success tolerance.

Because we make only M measurements in our Monte Carlo experiments, these results are subject to sampling fluctuations. Confidence statements can be made for $\hat{\rho}$ using standard statistical software.

10 Tuning of Algorithms

The procedure so far gives us, for each fully-specified combination of algorithm parameters Λ and each problem suite $\mathcal{S}$, a dataset $(\Lambda, \mathcal{S}, \delta, \hat{\rho}(\delta; \Lambda, S))$. When an algorithm has such parameters, we can define, for each fixed δ, the value of the parameters which gives the highest transition:

$$\hat{\rho}^{opt}(\delta; \mathcal{S}) = \max_{\Lambda} \hat{\rho}(\delta; \Lambda, \mathcal{S});$$

with associated optimal parameters $\Lambda^{opt}(\delta; \mathcal{S})$. When the results of the algorithm depend strongly on problem suite as well, we can also tune to optimize worst-case performance across suites, getting the minimax transition

$$\hat{\rho}^{\mathrm{MM}}(\delta) = \max_{\Lambda} \min_{\mathcal{S}} \hat{\rho}(\delta; \Lambda, \mathcal{S}).$$

and corresponding minimax parameters $\Lambda^{\mathrm{MM}}(\delta)$. This procedure was followed in [7] for a wide range of popular algorithms. Figure 3 of the main text presents the observed minimax transitions.

11 Results: Empirical Phase Transition

Figure S1 (which is a complete version of Figure 3 in the main text) compares observed phase transitions of several algorithms including AMP. We considered what was called in [7] the *standard suite*, with these choices

- Matrix ensemble: Uniform spherical ensemble(USE); each column of A is drawn uniformly at random from the unit sphere in $\mathbb{R}^n$.
- Coefficient ensemble: The vector x_0 has k nonzeros in random locations, with constant amplitude of nonzeros. If $\chi = +$, $x_0(i) \in \{0, +1\}$; if $\chi \in \{\pm, \square\}$, $x_0(i) \in \{+1, 0, -1\}$ (with equiprobable positive and negative entries).

For each algorithm we generated an appropriate grid of (δ, ρ) and created $M = 20$ independent problem instances at each gridpoint, i.e. independent realizations of vector x and measurement matrix A.

For AMP we used a focused search design, focused around $\rho_{\mathrm{CG}}(\delta)$. We used $N = 1000$. To reconstruct x, we run $T = 1000$ AMP iterations and report the mean square error at the final iteration. This choice of T is surely more than one would use in practice, however, we find (just as Theorem 2 of the Main Text would predict) that, for a given tolerance in defining 'Success', the measured phase transition we obtain with larger T is closer to the State Evolution transition. Later sections show that choosing T much smaller, eg 50 or less, is quite reasonable and practical. The only exception is the obvious one, predicted by State Evolution, namely that since the rate exponent $b(\delta, \rho)$ approaches 0 as we approach phase transition, the number of iterations should be larger very close to phase transition. For other algorithms, we used the general search design as described above. The computations for the other algorithms were actually performed for another project that has been carefully documented in [7]. We did not do fresh computations on those algorithms for this paper. For more details about those observed phase transitions we refer the reader to [7]. Admittedly, it would have been a more direct comparison to match AMP not at $N = 1000$ but instead at $N = 800$ in comparing with the other algorithms. However, the State Evolution Formalism predicts that results stabilize for all large N, so there is not much difference between the AMP at N=800 and N=1000; and extensive empirical work confirms this prediction - see Section 13 below. Consequently, it does make sense to plot these on the same scale and make such direct comparisons.

The calculation of the phase transition curve of AMP takes around 36 hours on a single Pentium 4 processor.

Observed Phase transitions for other coefficient ensembles and matrix ensembles are discussed below in sections 14 and 15.

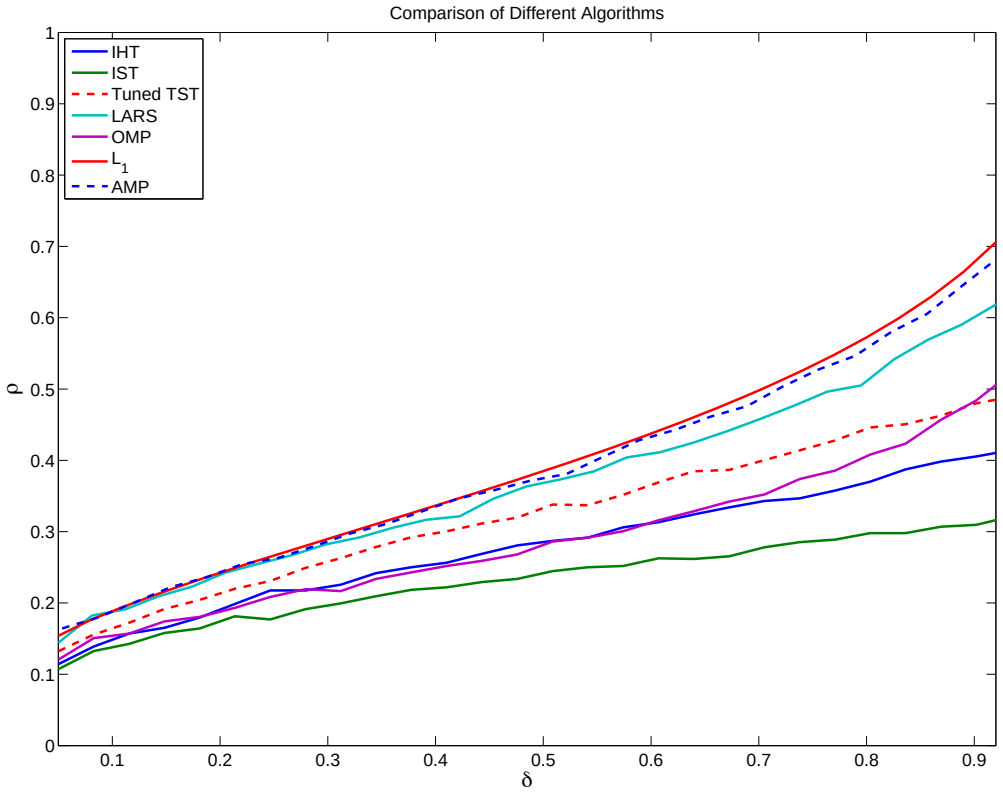

Figure S 1: Observed Phase Transitions for 6 Algorithms, and ρ_{SE}. AMP: method introduced in main text. IST: Iterative Soft Thresholding. IHT: Iterative Hard Thresholding. TST: a class of two-stage thresholding algorithms including subspace pursuit and CoSamp. OMP: Orthogonal Matching Pursuit. Note that the ℓ_1 curve coincides with the state evolution transition ρ_{SE}, a theoretical calculation. The other curves show empirical results. Note that IST, IHT, and TST were optimally tuned according to the method in [7]. In this set of experiments, $N = 800$ for all algorithms except AMP; results were obtained by an earlier project [7] and no new experiments were done here. $N = 1000$ for AMP; these computations were done explicitly for this project.

12 Example of the Interference Heuristic

In the main text, our motivation of the SE formalism used the assumption that the mutual access interference term $\mathrm{MAI}_t = (A^*A - I)(x^t - x_0)$ is marginally nearly Gaussian – i.e. the distribution function of the entries in the MAI vector is approximately Gaussian.

As we mentioned, this heuristic motivates the definition of the MSE map. It is easy to prove that the heuristic is valid at the first iteration; but for the validity of SE, it must continue to be true at every iteration until the algorithm stops. Figure S2 presents a typical example. In this example we have considered USE matrix ensemble and Rademacher coefficient ensemble. Also N is set to a small size problem 2000 and $(\delta, \rho) = (0.9, 0.52)$. The algorithm is tracked across 90 iterations. Each panel exhibits a linear trend, indicating approximate Gaussianity. The slope is decreasing with iteration count. The slope is the square root of the MSE, and its decrease indicates that the MSE is evolving towards zero. More interestingly,

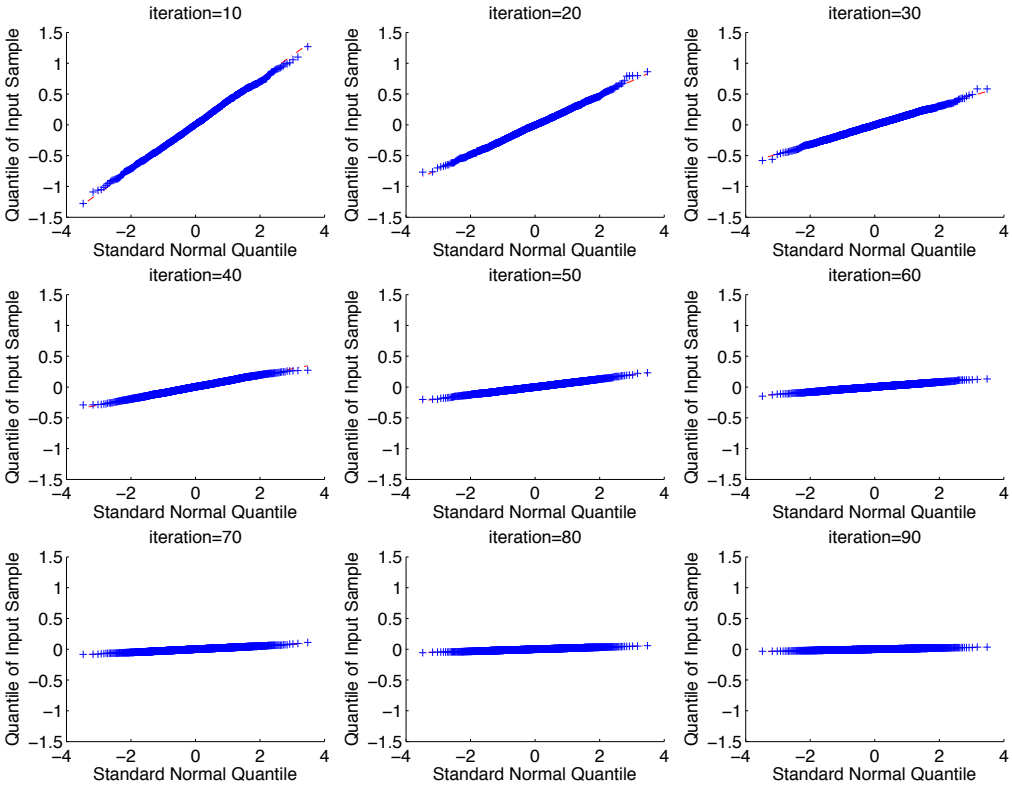

Figure S 2: QQ Plots tracking marginal distribution of mutual access interference (MAI): $(A^*A-I)(x^t-x_0)$. Panels (a)-(i): iterations $10, 20, \ldots, 90$. Each panel shows a standard quantile-quantile plot of MAI values versus normal distribution in blue. The sorted MAI values are plotted against the expected order statistics at the standard normal distribution. If the MAI values were normally distributed with mean μ and standard deviation τ the points would lie along a straight line with intercept μ and slope τ. Also shown: points in red (mostly obscured) along a straight line with intercept $AveMAI$ and slope $SD[MAI]$. Approximate linearity indicates approximate normality. Decreasing slope with increasing iteration number indicates decreasing standard deviation as iterations progress. For this experiment $N = 2000$, $\delta = .9$ and $\rho = 0.52$.

figure S3 shows the QQplot of the MAI for the partial Fourier matrix ensemble. Nonzero Coefficients here are again from the Rademacher ensemble and $(N, \delta, \rho) = (16384, 0.5, 0.35)$.

We have made many similar plots in the research for this paper and observed time and again the same patterns. Weak exceptions have been noticeable when N has been very small, particularly in a few cases where k has been very small.

13 Testing Predictions of State Evolution

The last section gave an illustration tracking the actual evolution of the AMP algorithm, it showed that the State Evolution heuristic is qualitatively correct.

We now consider predictions made by SE and their quantitative match with empirical observations.

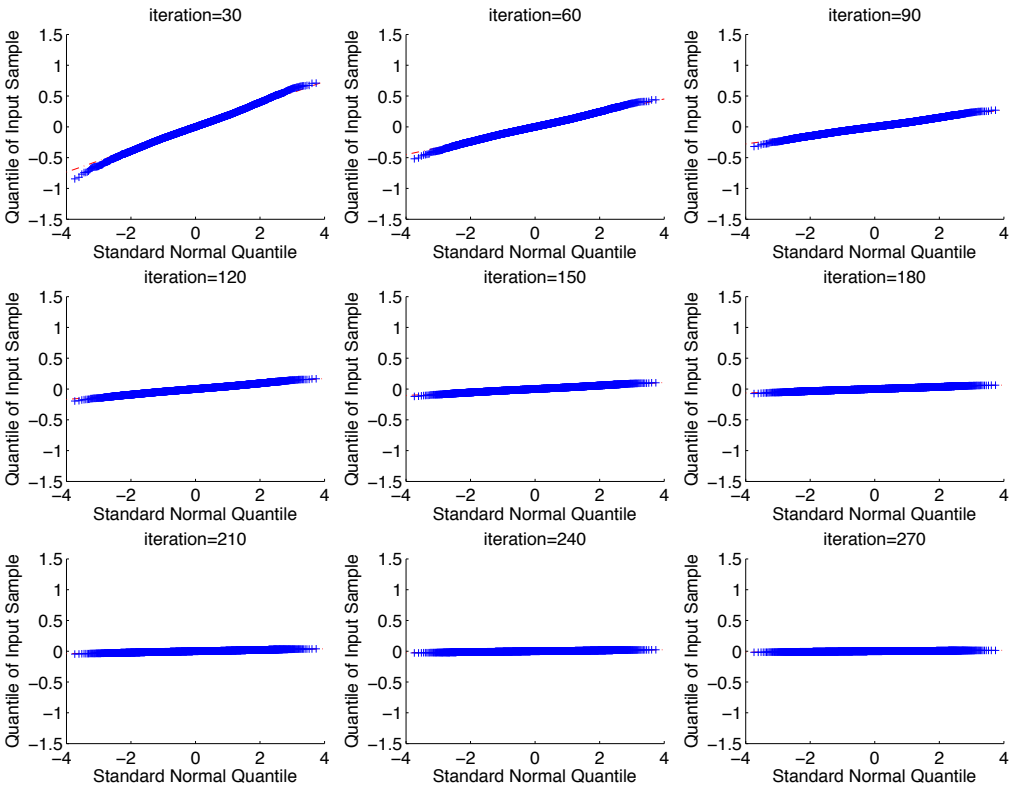

Figure S 3: QQ Plots tracking marginal distribution of mutual access interference (MAI). Matrix Ensemble: partial Fourier. Panels (a)-(i): iterations 30,60,..., 270. For this experiment, $N = 16384$, $\delta = 1/2$ and $\rho = 0.35$. For other details, see Fig. 2.

We consider predictions of four observables:

- MSE on zeros and MSE on non-zeros:

$$\texttt{MSEZ} = \mathbb{E}[\hat{x}(i)^2 | x_0(i) = 0], \qquad \texttt{MSENZ} = \mathbb{E}[(\hat{x}(i) - x_0(i))^2 | x_0(i) \neq 0]$$

- Missed detection rate and False alarm rate:

$$\texttt{MDR} = \mathbb{P}[\hat{x}(i) = 0 | x_0(i) \neq 0], \qquad \texttt{FAR} = \mathbb{P}[\hat{x}(i) \neq 0 | x_0(i) = 0]$$

We illustrate the calculation of MDR. Other quantities are computed similarly. Let $\epsilon = \delta\rho$, and suppose that entries in $x_0(i)$ are either 0, 1, or -1, with $\mathbb{P}\{x_0(i) = \pm 1\} = \epsilon/2$. Then, with $Z \sim N(0,1)$,

$$\begin{aligned} \mathbb{P}[\hat{x}(i) = 0 | x_0(i) \neq 0] &= \mathbb{P}[\eta(1 + \frac{\sigma}{\sqrt{\delta}} Z) \neq 0] \\ &= \mathbb{P}[1 + \frac{\sigma}{\sqrt{\delta}} Z \notin (-\lambda\sigma, \lambda\sigma)] \\ &= \mathbb{P}[Z \notin (a, b)] \end{aligned} \tag{29}$$

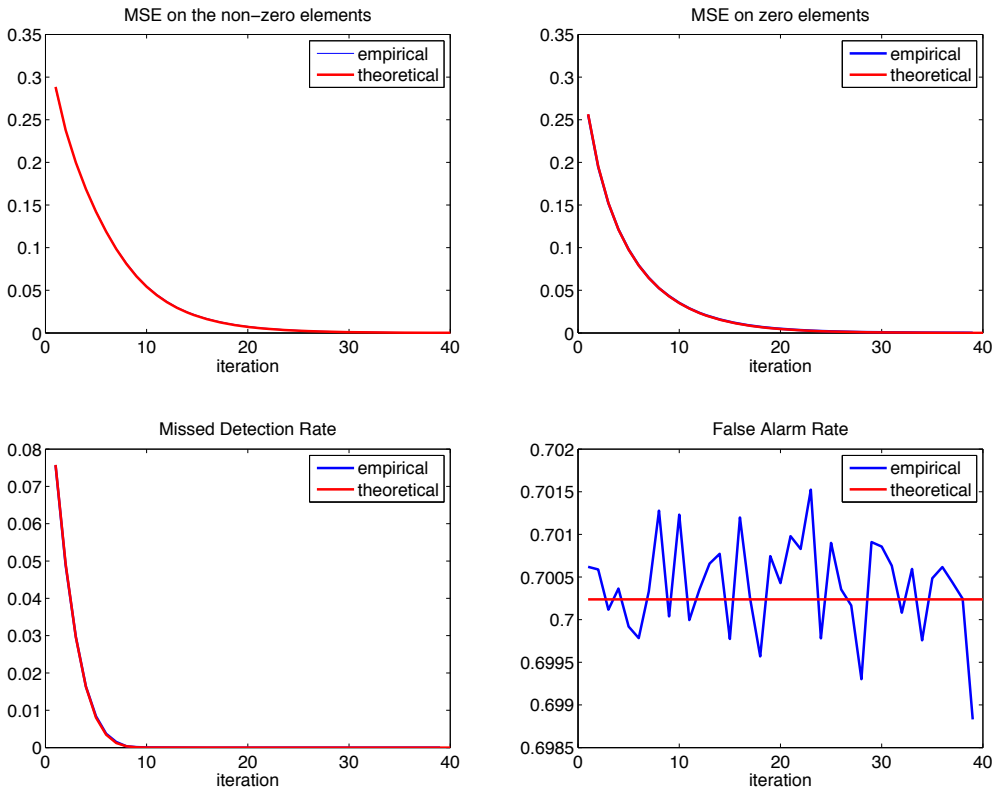

Figure S 4: Comparison of State Evolution predictions against observations. Panels (a)-(d): MSENZ, MSE, MDR, FAR. Curve in red: theoretical prediction. Curve in blue: mean observable. Each panel shows the evolution of a specific observable as iterations progress. Two curves are present in each panel, however, except for the lower left panel, the blue curve (empirical data) is obscured by the presence of the red curve. The two curves are in close agreement in all panels. For this experiment $N = 5000$, $\delta = .3$ and $\rho = 0.15$.

with $a = ((-\lambda - 1/\sigma) \cdot \sqrt{\delta}$, $b = (\lambda - 1/\sigma) \cdot \sqrt{\delta}$.

In short, the calculation merely requires classical properties of the normal distribution. The three other quantities simply require other similar properties of the normal. As discussed in the main text, SE evolution makes an iteration-by-iteration prediction of σ_t; in order to calculate predictions of `MDR`, `FAR`, `MSENZ` and `MSEZ`, the parameters ϵ and λ are also needed.

We compared the state evolution predictions with the actual values by a Monte Carlo experiment. We chose these triples (δ, ρ, N): $(0.3, 0.15, 5000)$, $(0.5, 0.2, 4000)$, $(0.7, 0.36, 3000)$. We again used the standard problem suite (USE matrix and unit amplitude nonzero). At each combination of (δ, ρ, N), we generated $M = 200$ random problem instances from the standard problem suite, and ran the AMP algorithm for a fixed number of iterations. We computed the observables at each iteration. For example, the empirical missed detection rate is estimated by

$$\texttt{eMDR}(t) = \frac{\#\{i : x^t(i) = 0 \quad \text{and} \quad x_0(i) \neq 0\}}{\#\{i : x_0(i) \neq 0\}}.$$

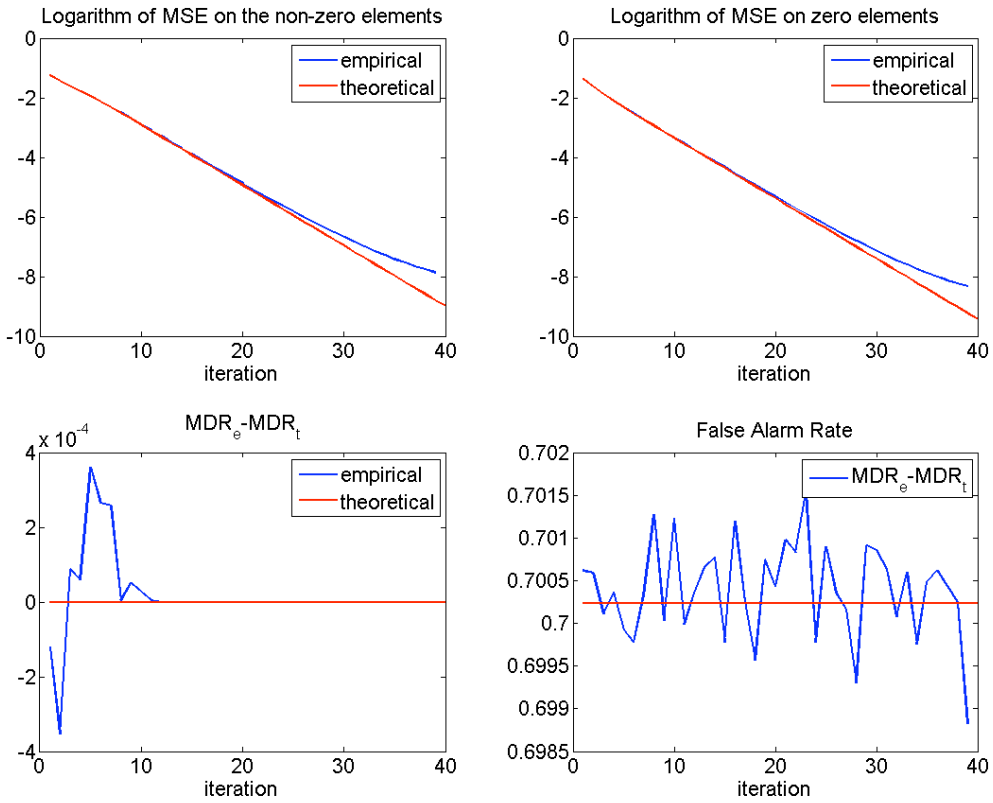

Figure S 5: Comparison of State Evolution predictions against observations. Same underlying situation and results as Figure S4. Plots in top row: Y axis is on a logarithmic plotting scale, i.e. we are view log(MSE) versus iteration. Lower left panel: Y axis is now the difference between Missed Detection Rate and SE prediction.

We averaged the observable trajectories across the M Monte Carlo realizations, producing empirical averages.

The results for the three cases are presented in Figures S4, S6, S7. Shown on the display are curves indicating both the theoretical prediction and the empirical averages. In the case of the upper row and the lower left panel, the two curves are so close that one cannot easily tell that two curves are, in fact, being displayed.

We have made many similar plots in the research for this paper and observed time and again the same level of agreement between SE predictions of observables and actual results.

14 Coefficient Universality

SE displays invariance of the evolution results with respect to the coefficient distribution of the nonzeros. What happens in practice?

We studied invariance of AMP results as we varied the distributions of the nonzeros in x_0. We consider the problem $\chi = \pm$ and used the following distributions for the non-zero entries of x_0:

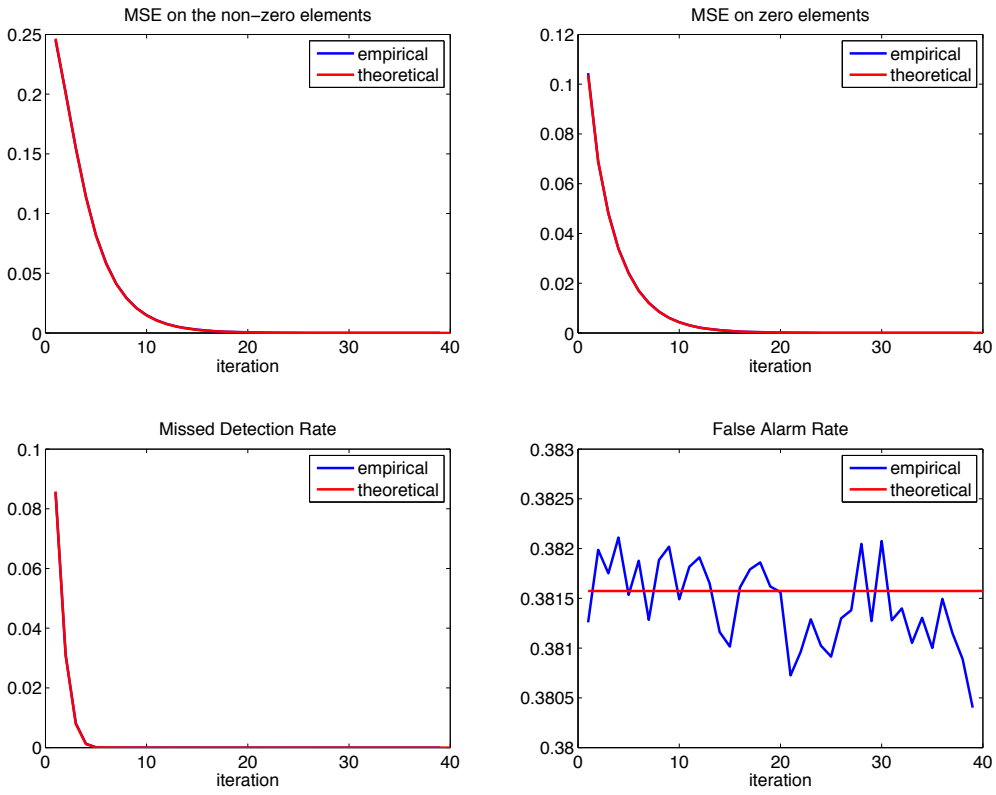

Figure S 6: Comparison of State Evolution predictions against observations. For this experiment $N = 4000$, $\delta = 0.5$ and $\rho = 0.20$. For other details, see Figure S4.

- Uniform in $[-1, +1]$;
- Rademacher (uniform in $\{+1, -1\}$);
- Gaussian;
- Cauchy.

In this study, $N = 2000$, and we considered $\delta = 0.1$, 0.3. For each value of δ we considered 20 equispaced values of ρ in the interval $[\rho_{\mathrm{CG}}(\delta;\pm) - 1/10, \rho_{\mathrm{CG}}(\delta;\pm) + 1/10]$, running each time $T = 1000$ AMP iterations. Data are presented in the two panels of Figures S8.

Each plot displays the fraction of success (S/M) as a function of ρ and a fitted success probability i.e. in terms of success probabilities, the curves display $\pi(\rho)$. In each case 4 curves and 4 sets of data points are displayed, corresponding to the 4 ensembles. The four datasets are visually quite similar, and it is apparent that indeed a considerable degree of invariance is present.

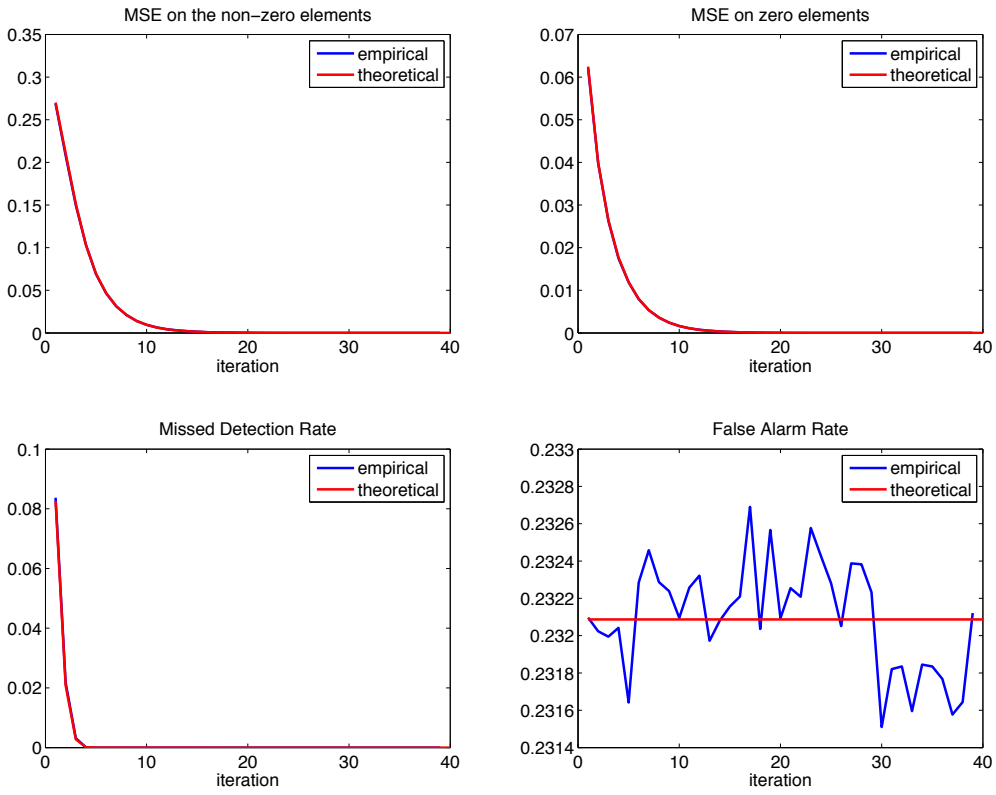

Figure S 7: Comparison of State Evolution predictions against observations for $N = 4000$, $\rho = 0.36$, $\delta = 0.70$. For other details, see Figure S4.

15 Matrix Universality

The Discussion section in the main text referred to evidence that our results are not limited to the Gaussian distribution.

We conducted a study of AMP where everything was the same as in Figure S1 above, however, the matrix ensemble could change. We considered three such ensembles: USE (columns iid uniformly distributed on the unit sphere), Rademacher (random entries iid ± 1 equiprobable), and Partial Fourier, (randomly select n rows from $N \times N$ fourier matrix.) We only considered the case $\chi = \pm$. Results are shown in Fig. 9, and compared to the theoretical phase transition for ℓ_1.

16 Timing Results

In actual applications, AMP runs rapidly.

We first describe a study comparing AMP to the LARS algorithm [8]. LARS is appropriate for comparison because, among the iterative algorithms previously proposed, its phase transition is closest to the ℓ_1 transition. So it comes closest to duplicating the AMP sparsity-undersampling tradeoff.

Each algorithm proceeds iteratively and needs a stopping rule. In both cases, we stopped calculations

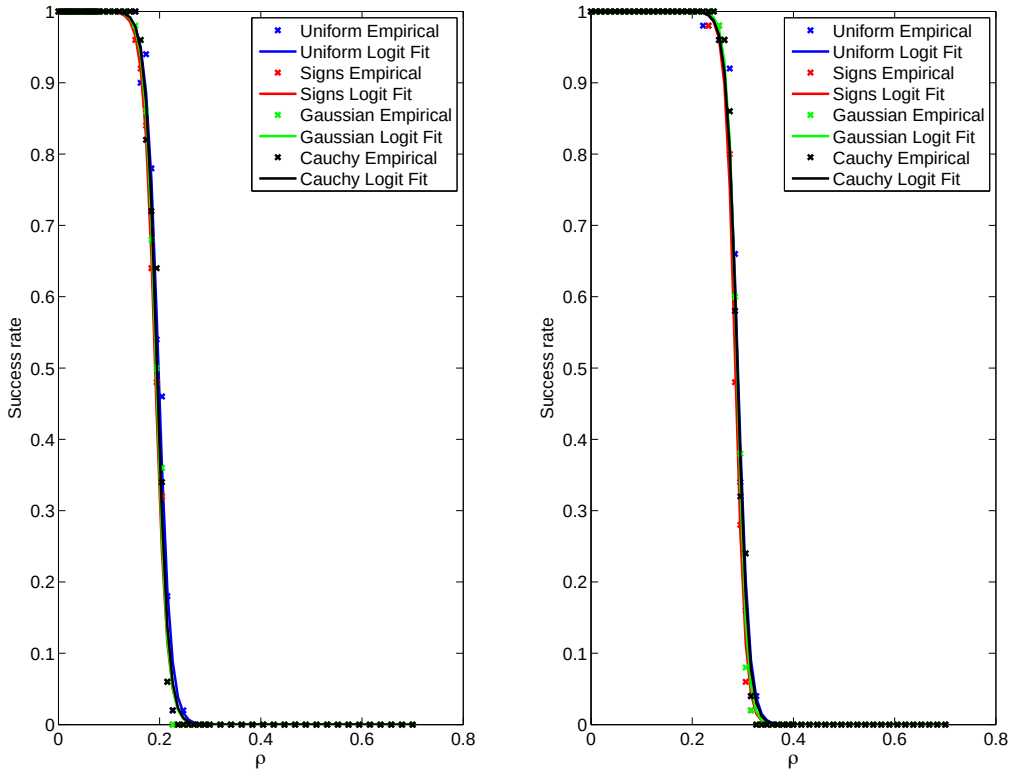

Figure S 8: Comparison of Success probabilities for different coefficient ensembles. Left panel: $\delta = 0.10$; Right panel: $\delta = 0.3$. Red: unit-amplitude coefficients. Blue: uniform $[-1, 1]$. Green: Gaussian. Black: Cauchy. Points: observed success fractions. Curves: Logistic fit.

when the relative fidelity measure exceeded 0.999, ie when $\|y - Ax^t\|_2/\|y\|_2 < 0.001$.

In our study, we used the partial Fourier matrix ensemble with unit amplitude for nonzero entries in the signal x_0. We considered a range of problem sizes (N, n, k) and in each case averaged timing results over $M = 20$ problem instances. Table S 1 presents timing results.

In all situations studied, AMP is substantially faster than LARS. There are a few very sparse situations – i.e. where k is in the tens or few hundreds – where LARS performs relatively well, losing the race by less than a factor 3. However, as the complexity of the objects increases, so that k is several hundred or even one thousand, LARS is beaten by factors of 10 or even more.

(For very large k, AMP has a decisive advantage. When the matrix A is dense, LARS requires at least $c_1 \cdot k \cdot n \cdot N$ operations, while AMP requires at most $c_2 \cdot n \cdot N$ operations. Here $c_2 = \log((\mathbb{E}X^2)/\sigma_T^2)/b$ is a bound on the number of iterations, and $(\mathbb{E}X^2)/\sigma_T^2$ is the relative improvement in MSE in T iterations. Hence in terms of flops we have

$$\frac{\mathsf{flops}(\mathrm{LARS})}{\mathsf{flops}(\mathrm{AMP})} \geq \frac{kb(\delta, \rho)}{\log((\mathbb{E}X^2)/\sigma_T^2)} .$$

This logarithmic dependence of the denominator is very weak, and very roughly this ratio scales directly with k.)

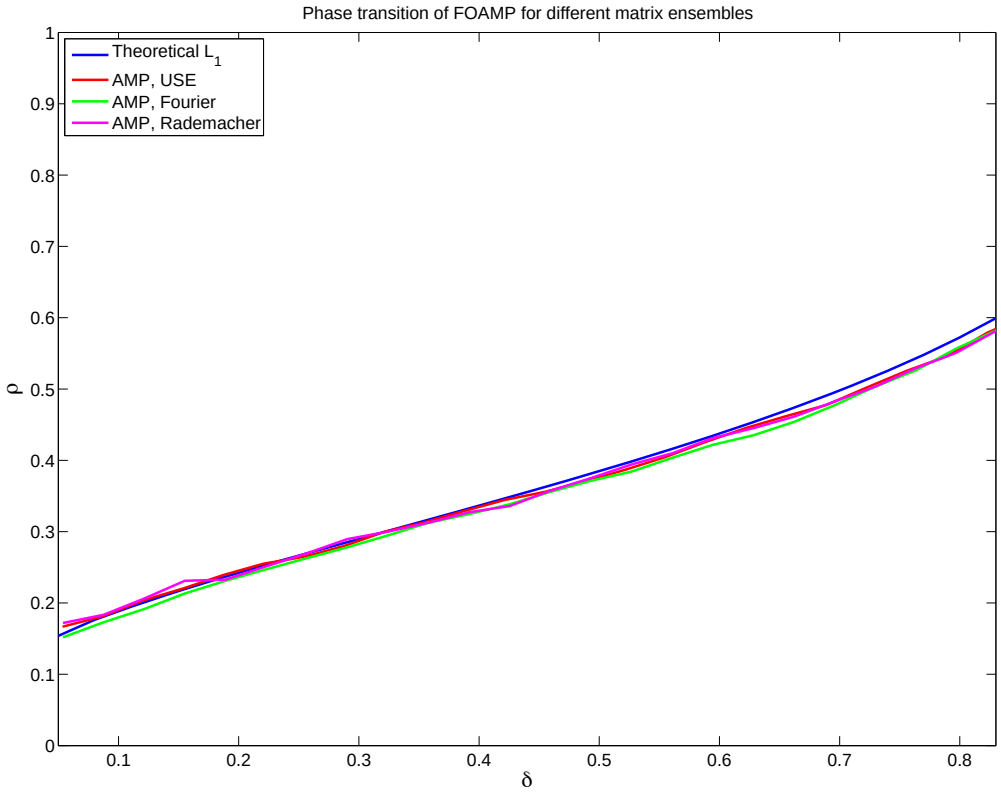

Figure S 9: Observed Phase Transitions at different matrix ensembles. Case $\chi = \pm$. Red: Uniform Spherical Ensemble (Gaussian with normalize column lengths). Magenta: Rademacher (± 1 equiprobable). Green: partial Fourier. Blue: ρ_{ℓ_1}.

We also studied AMP's ability to solve very large problems.

We conducted a series of trials with increasing N in a case where A and A^* can be applied rapidly, without using ordinary matrix storage and matrix operations; specifically, the partial Fourier ensemble. For nonzeros of the signal x_0. we chose unit amplitude nonzeros.

We considered the fixed choice $(\delta, \rho) = (1/6, 1/8)$ and N ranging from $1K$ to $(K = 1024)$ to $256K$ in powers of 2. At each signal length N we generated $M = 10$ random problem instances and measured CPU times (on a single Pentium 4 processor) and iteration counts for AMP in each instance. We considered four stopping rules, based on MSE σ^2, $\sigma^2/2$, $\sigma^2/4$, and $\sigma^2/8$, where $\sigma^2 = 2^{-13}$. We then averaged timing results over the $M = 10$ randomly generated problem instances

Figure S10 presents the number of iterations as a function of the problem size and accuracy level. According to the SE formalism, this should be a constant independent of N at each fixed (δ, ρ) and we see indeed that this is the case for AMP: the number of iterations is close to constant for all large N. Also according to the SE formalism, each additional iteration produces a proportional reduction in formal MSE, and indeed in practice each increment of 5 AMP iterations reduces the actual MSE by about half.

Figure S11 presents CPU time as a function of the problem size and accuracy level. Since we are using the partial Fourier ensemble, the cost of applying A and A^* is proportional to $N \log(N)$; this is much less

Table S 1: Timing Comparison of AMP and LARS. Average Times in CPU seconds.

N	n	k	AMP	LARS
4096	820	120	0.19	0.7
8192	1640	240	0.34	3.45
16384	3280	480	0.72	19.45
32768	1640	160	2.41	7.28
16384	820	80	1.32	1.51
8192	820	110	0.61	1.91
16384	1640	220	1.1	5.5
32768	3280	440	2.31	23.5
4096	1640	270	0.12	1.22
8192	3280	540	0.22	5.45
16384	6560	1080	0.45	27.3
32768	1640	220	6.95	17.53

than what we would expect for the cost of applying a general dense matrix. We see that indeed AMP execution time scales very favorably with N in this case – to the eye, the timing seems practically linear with N. The timing results show that each doubling of N produces essentially a doubling of execution time. iteration produces a proportional reduction in formal MSE, and indeed in practice each increment of 5 AMP iterations reduces the MSE by about half.

17 Postscript: Relation with Stojnic's bounds

In [9], Mihalo Stojnic proved bounds on the reconstrtion phase transition for both problems $\chi = \pm$ and $\chi = +$, under LP reconstruction (respectively Theorem 4 and Theorem 7 in that paper). The bounds appear to match the prediction of [10], although there is no proof of this fact.

The paper by Stojnic appeared as a preprint at ArXiV on the same day as the ArXiv version of this paper [11]. The techniques are completely different (in particular, [9] does not provide a fast reconstruction algorithm). It is nevertheless interesting to compare our results with the ones of [9].

It is a lengthy but straightforward exercise of calculus to check that the bounds of [9] coincide with the SE prediction, cf. eqn [5] in the main manuscript. For the reader who is willing to repeat the same exercise, we provide some hints.

First of all, it seems convenient to compare both results with the following parametric expressions for the SE phase transitions.

For $\chi = +$:

$$\delta = \frac{\phi(z)}{\phi(z) + z\,\Phi(z)}\,, \tag{30}$$

$$\rho = 1 - \frac{z(1-\Phi(z))}{\phi(z)}\,, \tag{31}$$

with $z \in [0, \infty)$.

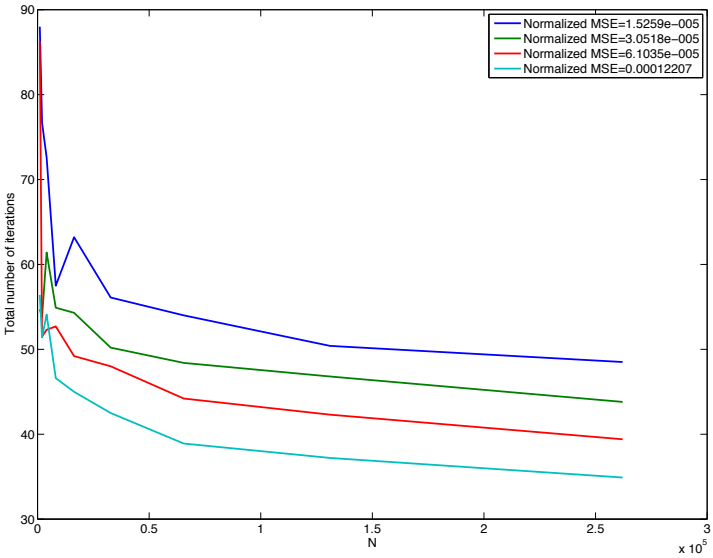

Figure S 10: Iteration Counts versus Signal Length N. Different curves show results for different stopping rules. Horizontal axis: signal length N. Vertical axis: Number of iterations, T. Blue, Green, Red, Aqua curves depict results when stopping thresholds are set at $2^{-13} \cdot 2^{4-\ell}$, with $\ell = 0, 1, 2, 3$ Each doubling of accuracy costs about 5 iterations. The number of iterations T needed to reach a specified MSE does *not* increase with N; this agrees with the SE formalism.

For $\chi = \pm$:

$$\delta = \frac{\phi(z)}{\phi(z) + z\,(\Phi(z) - 1/2)}\,, \tag{32}$$

$$\rho = 1 - \frac{z(1-\Phi(z))}{\phi(z)}\,, \tag{33}$$

with $z \in [0, \infty)$.

As z varies in its domain, $(\delta(z), \rho(z))$ describe the $(\delta, \rho_{\mathrm{SE}}(\delta; \chi))$ curve

These expressions are obtained eqn [5] in our main text by simple calculus. In order to compare with Stojnic's variables α, β_w (resp. β_w^+ for $\chi = +$) and θ_w (resp. $\beta_w^\pm$ for $\chi = \pm$), the following notations dictionary should be used.

For $\chi = +$ (on the right are the notations in our work)

$$\alpha = \delta\,, \tag{34}$$

$$\beta_w^+ = \rho\delta\,, \tag{35}$$

$$\mathrm{erfinv}\Big(2\frac{1-\theta_w^+}{1-\beta_w^+} - 1\Big) = \frac{z}{\sqrt{2}}\,. \tag{36}$$

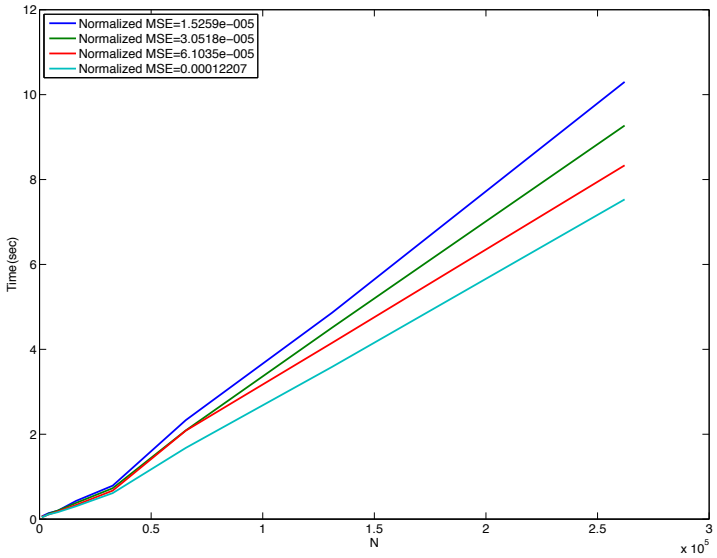

Figure S 11: CPU Time Scaling with N . Different curves show results for different stopping rules. Horizontal axis: signal length N. Vertical axis: CPU time(seconds). Blue, Green, Red, Aqua curves depict results when stopping thresholds are set at $2^{-13} \cdot 2^{4-\ell}$, with $\ell = 0, 1, 2, 3$ CPU time grows linearly with problem size, and scales logarithmically with final MSE. This agrees with the SE formalism.

For $\chi = \pm$ (on the right are the notations in our work)

$$\alpha = \delta\,, \tag{37}$$

$$\beta_w = \rho\delta\,, \tag{38}$$

$$\mathrm{erfinv}\Big(\frac{1-\theta_w^+}{1-\beta_w^+}\Big) = \frac{z}{\sqrt{2}}\,. \tag{39}$$

References

[1] D. L. Donoho and J. Tanner. Counting faces of randomly projected hypercubes and orthants with applications. *ArXiv*, 2008.

[2] D.L. Donoho, I.M. Johnstone, J.C. Hoch, and A.S. Stern. Maximum entropy and the nearly black object. *JRSS B*, 54(1):41–81, 1992.

[3] D. L. Donoho and I. M. Johnstone. Minimax risk over l_p balls. *Prob. Thry. Rel. Fields*, 99:277–303, 1994.

[4] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[5] D. L. Donoho and J. Tanner. Precise undersampling theorems. *Proc. IEEE.* in submission.

[6] P. J. Bickel. Minimax estimation of the mean of a normal distribution subject to doing well at a point. In S. Zacks et al, editor, *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, pages 511–528. Academic Press, 1983.

[7] A. Maleki and D. L. Donoho. Optimally tuned iterative thresholding algorithms for compressed sensing. *IEEE J. Sel. Areas Sig. Proc.*, 2009. submitted.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32:407–499, 2004.

[9] M. Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. arXiv:0907.3666, 2009.

[10] D. L. Donoho and J. Tanner. Neighborliness of randomly-projected simplices in high dimensions. *PNAS*, 102(27):9452–9457, 2005.

[11] David L. Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing. arXiv:0907.3574, 2009.