

# Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising

David Donoho\*, Iain Johnstone\* and Andrea Montanari\*,†

October 31, 2012

## Abstract

Compressed sensing posits that, within limits, one can undersample a sparse signal and yet reconstruct it accurately. Knowing the precise limits to such undersampling is important both for theory and practice.

We present a formula that characterizes the allowed undersampling of generalized sparse objects. The formula applies to Approximate Message Passing (AMP) algorithms for compressed sensing, which are here generalized to employ denoising operators besides the traditional scalar soft thresholding denoiser. This paper gives several examples including scalar denoisers not derived from convex penalization – the *firm* shrinkage nonlinearity and the *minimax* nonlinearity – and also nonscalar denoisers – block thresholding, monotone regression, and total variation minimization.

Let the variables  $\varepsilon = k/N$  and  $\delta = n/N$  denote the generalized sparsity and undersampling fractions for sampling the  $k$ -generalized-sparse  $N$ -vector  $x_0$  according to  $y = Ax_0$ . Here  $A$  is an  $n \times N$  measurement matrix whose entries are iid standard Gaussian. The formula states that the phase transition curve  $\delta = \delta(\varepsilon)$  separating successful from unsuccessful reconstruction of  $x_0$  by AMP is given by:

$$\delta = M(\varepsilon|\text{Denoiser}),$$

where  $M(\varepsilon|\text{Denoiser})$  denotes the per-coordinate minimax mean squared error (MSE) of the specified, optimally-tuned denoiser in the *directly observed* problem  $y = x + z$ . In short, the phase transition of a noiseless *undersampling* problem is identical to the minimax MSE in a *denoising* problem. We prove that this formula follows from state evolution and present numerical results validating it in a wide range of settings.

The above formula generates numerous new insights, both in the scalar and in the nonscalar cases.

**Key Words:** Approximate Message Passing. Lasso. Group Lasso, Joint Sparsity, James-Stein, Minimax Risk over Nearly-Black Objects. Minimax Risk of Soft Thresholding. Minimax Risk of Firm Thresholding. Minimax Shrinkage. Nonconvex penalization. State Evolution. Total Variation Minimization. Monotone Regression.

**Acknowledgements.** NSF DMS-0505303, NSF DMS-0906812, NSF CAREER CCF-0743978.

---

\*Department of Statistics, Stanford University

†Department of Electrical Engineering, Stanford University

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Signal models . . . . .	4
1.2	Denoising and minimax MSE . . . . .	5
1.3	Compressed sensing and AMP reconstruction . . . . .	7
1.4	Phase transition for AMP . . . . .	7
1.5	This Paper . . . . .	9
1.6	Contributions . . . . .	10
1.7	Related literature . . . . .	12
<b>2</b>	<b>Scalar-separable denoisers</b>	<b>13</b>
2.1	Minimax MSE of a separable denoiser . . . . .	13
2.2	Firm shrinkage . . . . .	14
2.3	Minimax shrinkage . . . . .	15
2.4	Empirical phase transition behavior . . . . .	18
<b>3</b>	<b>Block-separable denoisers</b>	<b>20</b>
3.1	Block soft thresholding . . . . .	21
3.2	Block James-Stein . . . . .	24
3.3	Empirical phase transition behavior . . . . .	26
<b>4</b>	<b>Monotone regression</b>	<b>26</b>
4.1	Minimax MSE . . . . .	28
4.2	Empirical phase transition behavior . . . . .	31
<b>5</b>	<b>Total variation minimization</b>	<b>32</b>
5.1	Minimax MSE . . . . .	32
5.2	Empirical phase transition behavior . . . . .	35
<b>6</b>	<b>Characterization of the phase transition using state evolution</b>	<b>37</b>
6.1	State Evolution . . . . .	37
6.2	State Evolution Phase Transition . . . . .	39
6.3	Non-convergence of state evolution . . . . .	40
<b>7</b>	<b>Phase transitions for other algorithms</b>	<b>41</b>
7.1	Interpretation of separable denoisers in terms of penalized least-squares . . . . .	42
<b>A</b>	<b>Classical cases</b>	<b>48</b>
<b>B</b>	<b>Calculation of minimax MSE</b>	<b>49</b>
<b>C</b>	<b>Convergence properties of AMP</b>	<b>53</b>
<b>D</b>	<b>Calculation of minimax MSE for block soft thresholding</b>	<b>54</b>
D.1	Proof of Lemma 3.2 . . . . .	54
D.2	Proof of Lemma 3.3 . . . . .	54

<b>E</b>	<b>Non-convergence of state evolution</b>	<b>55</b>
<b>F</b>	<b>Monotone regression</b>	<b>56</b>
	F.1 Proof of Lemma 4.1, part (b) . . . . .	56
	F.2 Proof of Lemma 4.2 . . . . .	58
<b>G</b>	<b>Further computational details</b>	<b>61</b>
<b>H</b>	<b>Finite-<math>N</math> scaling and error analysis</b>	<b>61</b>
	H.1 Offsets decay toward zero . . . . .	61
	H.2 Transitions sharpen . . . . .	62
<b>I</b>	<b>Further properties of the risk function</b>	<b>64</b>
	I.1 Block soft thresholding . . . . .	64
	I.2 Positive-part James-Stein denoiser . . . . .	65
	I.2.1 Risk at 0 . . . . .	65
	I.2.2 Monotonicity of $R(\mu)$ . . . . .	66
	I.3 Proof of Lemma 6.2 . . . . .	66

# 1 Introduction

In the noiseless compressed sensing problem, we are given a collection of linear measurements of an unknown vector  $x_0$ :

$$y = Ax_0. \tag{1.1}$$

Here the measurement matrix  $A$  is  $n$  by  $N$ ,  $n < N$ , and the  $N$ -vector  $x_0$  is the object we wish to recover. Both  $y$  and  $A$  are known, while  $x_0$  is unknown and we seek to recover an approximation to  $x_0$ .

Since  $n < N$ , the equations are underdetermined. It seems hopeless to recover  $x_0$  in general, but in compressed sensing one also assumes that the object is *sparse* in the appropriate sense. Suppose that the object is known to be  $k$ -sparse, i.e. to have  $k$  nonzero entries. If the problem dimensions  $(k, n, N)$  are large, many recovery algorithms exhibit the phenomenon of *phase transition*.

Explicitly, let  $\varepsilon = k/N$  and  $\delta = n/N$  denote the sparsity and undersampling parameters, respectively. Hence  $(\varepsilon, \delta) \in [0, 1]^2$  defines a phase space for the different kinds of limiting situations we may encounter as  $(k, n, N)$  grow large. For a variety of algorithms and Gaussian matrices  $A$  with iid entries, one finds that this phase space can be partitioned into two phases: “success” and “failure”. Namely, for a given algorithm  $\mathcal{A}$  and given sparsity fraction  $\varepsilon$ , there exists a critical fraction  $\delta(\varepsilon|\mathcal{A})$  such that if the sampling rate  $\delta$  is larger than the critical value,  $\delta > \delta(\varepsilon|\mathcal{A})$ , then the algorithm is successful in recovering the underlying object  $x_0$  with high probability<sup>1</sup>, while if  $\delta < \delta(\varepsilon|\mathcal{A})$  the algorithm is unsuccessful, also with high probability. In particular,  $\delta(\varepsilon|\mathcal{A}) < 1$  means that it is indeed possible to undersample and still recover the unknown signal. In fact  $\delta(\varepsilon|\mathcal{A})$  shows precisely the limits of allowable undersampling. By now a large amount of empirical and theoretical knowledge has been compiled about the phase transitions exhibited by different algorithms: we refer the reader to [DT10b, Don06, DT09a, XH10, BCT11, Sto10, KWT09, DMM09, DMM11, Wai09]. In a parallel line of work, a number of sufficient conditions under which undersampling is possible using deterministic matrices have been studied, see e.g. [CT05, BRT09, BGI<sup>+</sup>08, Can06].

It is however fair to say that the research focused so far on ‘unstructured’ notions of sparsity whereby  $k$  simply counts the number of non-zero entries in  $x_0$ . (We refer to Section 1.7 for an overview of related literature.) On the other hand, applications naturally lead to ‘structured’ notions of sparsity. This paper applies an algorithm framework - *Approximate Message Passing* (AMP) - to construct specific algorithms applicable to a variety of compressed sensing settings, including block and structured sparsity, convex and nonconvex penalization, and develops a single unifying formula that, specialized to each instance, gives the actual phase transition that we observe in practice. To give a preview of our results, we first recall some facts about statistical decision theory and AMP reconstruction. For the sake of illustration, the classical case of simple sparse vectors will be used as a running example.

## 1.1 Signal models

Throughout this paper, we will consider estimation of unknown structured signals  $x \in \mathbf{R}^N$  from a minimax point of view. Various notions of structures can be formalized by considering a family  $\mathcal{F}_N$  of probability measures over  $\mathbf{R}^N$ . One such probability measure will be denoted by  $\nu_N \in \mathcal{F}_N$  and a signal with distribution  $\nu_N$  will often be denoted as  $\mathbf{X} \sim \nu_N$ .

---

<sup>1</sup>Throughout the paper, we will write that an event holds with high probability (w.h.p.) if its probability converges to 1 in the large system limit  $N, n \rightarrow \infty$  with  $\delta = n/N$  and  $\varepsilon = k/N$  fixed.

The family  $\mathcal{F}_N$  will typically include degenerate distributions, i.e. point masses  $\nu_N = \delta_{x_0}$  for some  $x_0 \in \mathbf{R}^N$ .

**Example 1.1.** The case of simple sparse vectors corresponds to the family

$$\mathcal{F}_{N,\varepsilon} \equiv \left\{ \nu_N \in \mathcal{P}(\mathbf{R}^N) : \mathbb{E}_{\nu_N} \{ \|\mathbf{X}\|_0 \} \leq N\varepsilon \right\}, \quad (1.2)$$

where  $\mathcal{P}(\mathbf{R}^N)$  is the set of Borel probability measures on  $\mathbf{R}^N$  and, as usual,  $\|v\|_0$  denotes the number of nonzero entries of the vector  $v$ . ■

As exemplified in this case  $\mathcal{F}_N$  is often indexed by a sparsity parameter  $\varepsilon$ , with  $N\varepsilon$  corresponding to the number of non-zero entries. We will sometimes use the notation  $\mathcal{F}_{N,\varepsilon}$  to indicate this dependency, also beyond the last example. Two further common properties that will always hold unless otherwise stated are the following.

1. *Nestedness.* If  $\varepsilon_1 \leq \varepsilon_2$  then  $\mathcal{F}_{N,\varepsilon_1} \subseteq \mathcal{F}_{N,\varepsilon_2}$ .
2. *Scale invariance.* If  $\nu_N \in \mathcal{F}_{N,\varepsilon}$  then any scaled version of  $\nu_N$  (defined by letting  $\nu_N^a(B) = \nu(aB)$  for some  $a > 0$ ) is also in  $\mathcal{F}_{N,\varepsilon}$ .

We will often omit the subscript  $N$  if  $N = 1$ .

## 1.2 Denoising and minimax MSE

The denoising problem requires to reconstruct a signal  $x \in \mathbf{R}^N$  from observations  $\mathbf{Y} = x + \mathbf{Z}$  whereby  $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{N \times N})$  is a noise vector of known variance. (Here and below  $\mathbf{I}_{m \times m}$  denoted the identity matrix in  $m$  dimensions.) A denoiser is a mapping

$$\begin{aligned} \eta(\cdot; \tau, \sigma) : \mathbf{R}^N &\rightarrow \mathbf{R}^N, \\ y &\mapsto \eta(y; \tau, \sigma), \end{aligned}$$

that returns an estimate of  $x$  when applied to observations  $y = \mathbf{Y}$ . The denoiser is parametrized by the noise scale  $\sigma$  and additional tuning parameters  $\tau \in \Theta$ . Often denoisers have the property  $\|\eta(y; \tau, \sigma)\|_2 \leq \|y\|_2$  and are hence called ‘shrinkers’. We will often have  $\Theta = \mathbf{R}_+$ , i.e. the denoiser depends on a single non-negative parameter, but more complex choices of the parameter space  $\Theta$  fit in the formalism as well.

Following the minimax formulation in the previous section, we evaluate denoisers on signals  $\mathbf{X} \sim \nu_N \in \mathcal{F}_{N,\varepsilon}$ , for specific class of distributions  $\mathcal{F}_{N,\varepsilon}$ . Because of the scale invariance property of  $\mathcal{F}_{N,\varepsilon}$ , it is sufficient to consider scale invariant denoisers:

$$\eta(y; \tau, \sigma) = \sigma \eta\left(\frac{y}{\sigma}; \tau, 1\right) \equiv \sigma \eta\left(\frac{y}{\sigma}; \tau\right), \quad (1.3)$$

Hence we omit the last argument when this is  $\sigma = 1$ . We evaluate a denoiser  $\eta$  through its minimax mean square error (MSE) per coordinate

$$M(\mathcal{F}_{N,\varepsilon}|\eta) = \frac{1}{N} \inf_{\tau \in \Theta} \sup_{\nu_N \in \mathcal{F}_{N,\varepsilon}} \mathbb{E}_{\nu_N} \left\{ \|\mathbf{X} - \eta(\mathbf{Y}; \tau)\|_2^2 \right\}, \quad (1.4)$$

where expectation is taken with respect to  $\mathbf{X} \sim \nu_N$  and  $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$ ,  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{N \times N})$ . In words, we tune the denoiser optimally to control the (per-coordinate) mean square error for typical signals from even the most unfavorable choice within our class  $\mathcal{F}_{N,\varepsilon}$ .

In the following we will be particularly interested in the high-dimensional limit of the minimax MSE. It will be implicitly understood that we are given a sequence of probability distributions classes  $\{\mathcal{F}_{N,\varepsilon}\}_{N \geq 1}$  indexed by the dimension and a sequence of denoisers  $\eta = \{\eta_N\}_{N \geq 1}$  also indexed by the dimension (the subscript  $N$  will be omitted on  $\eta$ ). We define the asymptotic minimax MSE through the following limit (whenever it exists)

$$M(\varepsilon|\eta) = \lim_{N \rightarrow \infty} \frac{1}{N} \inf_{\tau \in \Theta} \sup_{\nu_N \in \mathcal{F}_{N,\varepsilon}} \mathbb{E}_{\nu_N} \left\{ \|\mathbf{X} - \eta(\mathbf{Y}; \tau)\|_2^2 \right\}. \quad (1.5)$$

We say that a denoiser is *separable* if, for  $v = (v_1, \dots, v_N) \in \mathbf{R}^N$ , we have

$$\eta_N(v; \tau) = (\eta_1(v_1; \tau), \eta_1(v_2; \tau), \dots, \eta_1(v_N; \tau)). \quad (1.6)$$

**Example 1.2.** A well studied denoiser is coordinatewise soft-thresholding, that we will denote by  $\eta^{soft}$ . This is a separable denoiser with a unique parameter  $\tau \in \Theta = \mathbf{R}_+$  (the threshold). On each coordinate  $y \in \mathbf{R}$  this acts as

$$\eta^{soft}(y; \tau) = \begin{cases} y - \tau & \text{if } \tau \leq y, \\ 0 & \text{if } -\tau \leq y \leq \tau, \\ y + \tau & \text{if } y < -\tau. \end{cases}$$

Soft thresholding is well suited for sparse signals from the class  $\mathcal{F}_{N,\varepsilon}$  defined in Eq. (1.2). It turns out that the resulting minimax MSE  $M(\mathcal{F}_{N,\varepsilon}|\eta)$  can be characterized in terms of a scalar estimation problem, namely for all  $N$ ,  $M(\varepsilon|\eta^{soft}) = M(\mathcal{F}_{N,\varepsilon}|\eta^{soft}) = M(\mathcal{F}_{1,\varepsilon}|\eta^{soft})$ . Explicitly, all these quantities are given by

$$M(\varepsilon|\text{Soft}) = \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_\varepsilon} \mathbb{E}_\nu \left\{ [X - \eta^{soft}(X + Z; \tau)]^2 \right\},$$

where expectation is taken with respect to  $X \sim \nu$  and  $Z \sim \mathbf{N}(0, 1)$  independent of  $X$ . We refer to [DJ94, DMM09, DMM11] for an explicit characterization of this quantity (a summary being provided in Section 2). In particular  $M(\varepsilon|\text{Soft})$  can be explicitly evaluated. ■

In several other examples  $M(\mathcal{F}_{N,\varepsilon}|\eta)$  has been explicitly evaluated (see [DMM09], Supplementary Information).

**Example 1.3.** The positive-constrained case, where  $x_0 \geq 0$  can be modeled by considering the family of probability distributions  $\mathcal{F}_{N,\varepsilon,+} \equiv \{\nu_N \in \mathcal{P}(\mathbf{R}_+^N) : \mathbb{E}_{\nu_N} \{\|\mathbf{X}\|_0\} \leq N\varepsilon\}$  supported in the positive orthant. A natural denoiser is positive soft-thresholding  $\eta^{softpos}$ . This is again separable with, for  $y \in \mathbf{R}$ ,  $\eta^{softpos}(y; \tau) = (y - \tau)_+ \equiv \max(y - \tau, 0)$ . We have, again,  $M(\varepsilon|\eta^{softpos}) = M(\mathcal{F}_{N,\varepsilon,+}|\eta^{softpos}) = M(\mathcal{F}_{1,\varepsilon,+}|\eta^{softpos})$ . These quantities will be denoted by  $M(\varepsilon|\text{SoftPos})$ . ■

**Example 1.4.** The box-constrained case where  $x_0 \in [0, 1]^N$ , can be modeled through the class  $\mathcal{F}_{N,\varepsilon,\square} \equiv \{\nu_N \in \mathcal{P}([0, 1]^N) : \mathbb{E}_{\nu_N} \{\sum_{i=1}^N \mathbf{1}_{\{x_i \in (0,1)\}}\} \leq N\varepsilon\}$ . A natural denoiser is coordinatewise capping. Namely, for  $y \in \mathbf{R}$   $\eta^{cap}(y) = \min(1, \max(y, 0))$  (in this case there is no tuning parameter,  $\Theta = \emptyset$ ). Notice that, in this case, the signal class is not scale invariant, and hence the present framework does not apply directly. We discuss in Appendix A how to modify it. ■

In this paper we will give several other calculations of  $M(\mathcal{F}_{N,\varepsilon}|\eta)$ , for signal structures and denoisers going considerably beyond these examples.

### 1.3 Compressed sensing and AMP reconstruction

Consider now the noiseless compressed sensing problem, i.e. the problem of recovering a signal  $x_0 \in \mathbf{R}^N$  from  $n < N$  linear observations  $y = Ax_0$ , cf. Eq. (1.1). The key intuition is that this can be done exploiting the structure of  $x_0$ , sparsity being a special example. Approximate Message Passing (AMP) is an iterative scheme that allows to exploit richer types of structure in a flexible way. Given a denoiser  $\eta(\cdot; \tau, \sigma) : \mathbf{R}^N \rightarrow \mathbf{R}^N$  that is well suited for reconstructing  $x_0$  from observations  $x_0 + \mathbf{Z}$ , the AMP framework turns it into a scheme for solving the compressed sensing problem.

The AMP iteration starts from  $x^0 = 0$ , and proceeds for iterations  $t = 1, 2, \dots$  by maintaining a *current reconstruction*  $x^t \in \mathbf{R}^N$  and a *current working residual*  $z^t \in \mathbf{R}^n$ , and adjusting these iteratively. At iteration  $t$ , it forms a vector of *current pseudo-data*  $y^t = x^t + A^\top z^t$  and the next iteration's estimate is obtained by applying  $\eta$  to the current pseudo-data:

$$y^t = x^t + A^\top z^t, \quad (1.7)$$

$$x^{t+1} = \eta(y^t; \tau, \sigma_t), \quad (1.8)$$

$$z^{t+1} = y - Ax^{t+1} + \mathbf{b}_t z^t. \quad (1.9)$$

Here  $\mathbf{b}_t$  is a scalar determined by

$$\mathbf{b}_t \equiv \frac{1}{n} \operatorname{div} \eta(y; \tau, \sigma_{t-1}) \Big|_{y=y^{t-1}}. \quad (1.10)$$

The rationale for this specific choice of  $\mathbf{b}_t$  is discussed in [DMM09, DMM10, BM11a]: a justification goes beyond the scope of the present paper. The parameter  $\sigma_t$  can be interpreted as the noise standard deviation for the pseudo-data  $y^t$ . This can be estimated from  $y^t$  or  $z^t$  as explained in Appendix G.

Conceptually, AMP constructs an artificial denoising problem at each iteration and solves it using the denoising defined by  $\eta$ . In other words, it solves a compressed sensing problem by successive denoising. For the purpose of this paper, this description should be sufficient, save for two remarks.

*First*, the specifics of the construction are absolutely crucial for the results of this paper. These are embedded in the specification of the scale factors  $\mathbf{b}_t$  and  $\sigma_t$ .

*Second*, the above algorithm framework was originally proposed in [DMM09, DMM10] in the case of a *separable* denoiser  $\eta$ , i.e. a denoiser acting independently on each coordinate. In that paper the algorithm was derived by constructing a proper belief propagation message passing algorithm, and then obtaining the above algorithm as a first-order approximation. Specific separable denoisers corresponded to different choices of the prior in belief propagation.

A central point of this paper is that the form of the algorithm (1.7), (1.8), (1.9) is really more general and can be used in settings outside the original definition.

### 1.4 Phase transition for AMP

A recurring property of AMP algorithms is that they undergo a *phase transition*. When the undersampling ratio  $\delta$  decreases below a certain threshold (that depends on the signal class  $\mathcal{F}_{\varepsilon,N}$

and the denoiser  $\eta$ ), the algorithm behavior changes from being successful most of the times, to failing most of the times. In order to formalize this notion, we introduce the following terminology.

**Definition 1.1.** We say that AMP succeeds with high probability for the signal class  $\mathcal{F}_{N,\varepsilon}$ , and denoising procedure  $\eta$  if there exist a choice of the tuning parameter  $\tau \in \Theta$  such that the following happens. For each  $\xi > 0$ , there exists a function  $o(t)$  with  $\lim_{t \rightarrow \infty} o(t) = 0$  such that, for any  $\nu_N \in \mathcal{F}_{N,\varepsilon}$ ,

$$\limsup_{N \rightarrow \infty} \mathbb{P}_{\nu_N} \{ \|x^t - x_0\|_2^2 \geq N\xi \} \leq o(t). \quad (1.11)$$

Here probability is taken with respect to  $x_0 = \mathbf{X} \sim \nu_N$  and the sensing matrix  $A$ . Further, the limit  $N \rightarrow \infty$  is taken with  $n/N \rightarrow \delta$ .

Viceversa we say that AMP fails with high probability for the signal class  $\mathcal{F}_{N,\varepsilon}$ , and denoising procedure  $\eta$  if for any  $\tau \in \Theta$  the following happens. There exists  $\xi > 0$  and a sequence  $\nu_N \in \mathcal{F}_{N,\varepsilon}$  such that, for all  $t \geq 0$

$$\limsup_{N \rightarrow \infty} \mathbb{P}_{\nu_N} \{ \|x^t - x_0\|_2^2 \geq N\xi \} = 1. \quad (1.12)$$

Note that we could have chosen other, slightly different, notions of convergence, e.g. requiring  $\mathbb{P}_{\nu_N} \{ \lim_{t \rightarrow \infty} \|x^t - x_0\|_2 \} \rightarrow 1$  as  $N \rightarrow \infty$ . The notion of convergence adopted here corresponds instead to achieving arbitrarily small MSE per coordinate *in a constant number of iterations* (independent of  $N$ ). This notion is more appropriate for practical applications and better suited to the theory of AMP algorithms (see Section 6).

Our main result is the following general relation between denoising and compressed sensing.

**Phase Transition Formula for AMP.** Consider compressed sensing reconstruction over the signal class  $\mathcal{F}_{N,\varepsilon}$ , using AMP with the denoiser  $\eta$ . Denote by  $M(\varepsilon|\eta)$  the asymptotic minimax MSE per coordinate using denoiser  $\eta$ .

Then AMP succeeds with high probability if

$$\delta > M(\varepsilon|\eta). \quad (1.13)$$

Viceversa AMP fails with high probability for  $\delta < M(\varepsilon|\eta)$ .

**Example 1.5.** Let  $\mathcal{F}_{N,\varepsilon}$  be the class of signals with at most  $N\varepsilon$  non-zero entries (in expectation) and consider AMP with soft thresholding  $\eta^{soft}(\cdot; \tau)$ . Then the above formula states that reconstruction will succeed if  $\delta > M(\varepsilon|\text{Soft})$  and fail for  $\delta < M(\varepsilon|\text{Soft})$ . This result was first proved in [DMM09] to follow from state evolution. State evolution was subsequently established as a rigorous tool in [BM11a].

The same paper [DMM09] studied AMP with positive soft thresholding and showed that it succeeds for  $\delta > M(\varepsilon|\text{SoftPos})$ , AMP with capping, proving that it succeeds for  $\delta > M(\varepsilon|\text{Cap})$ .

Appendix A spells out how these existing results fall under the aegis of Eq. (1.13). ■

*Comparison to  $(\rho, \delta)$  phase diagrams.* In prior literature on phase transitions in compressed sensing, [DT10b, Don06, DT09a, BCT11, DMM09, DMM11], the authors considered a different phase diagram, based on variables  $\delta$  and  $\rho = \varepsilon/\delta$ . The relation  $\varepsilon = \rho\delta$  makes for a 1-1 relationship between the diagrams, so all information in the two diagrams can be presented in either format.



## 1.5 This Paper

Our aim in this paper is to show that formula (1.13) is correct in settings extending far beyond the three cases just mentioned in Example 1.5. We lay out several denoising problems, and in each one verify the general formula. This requires in each case (a) calculating the minimax MSE for a problem of statistical decision theory; (b) implementing AMP for compressed sensing with the given denoising family; and (c) verifying empirically that the phase transition does indeed occur at the precise sparsity/undersampling tradeoff indicated by the formula.

In particular, we consider the following denoising tasks, and corresponding compressed sensing problems.

**Firm Shrinkage.** Again we consider the class of sparse vectors  $\mathcal{F}_{N,\varepsilon}$  but instead of soft-thresholding, we use the *firm shrinkage* denoiser  $\eta^{firm}(\cdot; \tau)$ . This is again a separable denoiser with two tuning parameters  $\tau = (\tau_1, \tau_2)$  with  $\tau \in \Theta \equiv \{(\tau_1, \tau_2) : 0 \leq \tau_1 < \tau_2 < \infty\}$ . It acts on each coordinate by setting  $\eta^{firm}(y; \tau) = 0$  for  $|y| < \tau_1$ ,  $\eta^{firm}(y; \tau) = y$  for  $|y| > \tau_2$  and interpolating linearly.

Denoting by  $M(\varepsilon|\text{Firm})$  the associated asymptotic minimax MSE, we will show that  $M(\varepsilon|\text{Firm}) < M(\varepsilon|\text{Soft})$  strictly. By verifying the general formula, we show that the phase transition curve for optimally-tuned AMP firm shrinkage is slightly better than the phase transition for optimally tuned AMP soft shrinkage.

**Minimax Shrinkage.** For the same class of sparse vectors  $\mathcal{F}_{N,\varepsilon}$ , we consider the separable denoiser  $\eta$  applies coordinatewise shrinkage using a *minimax* shrinkage. In other words implicitly we are optimizing the mean square error over  $\Theta \equiv \{\text{all scalar nonlinearities}\}$ . We calculate the minimax MSE function  $M(\varepsilon|\text{Minimax})$ , and show that  $M(\varepsilon|\text{Minimax}) < M(\varepsilon|\text{Firm})$  strictly. By verifying the general formula we show that the phase transition curve for AMP minimax shrinkage is slightly better than the phase transition for both AMP soft or firm shrinkage.

**Block Thresholding.** Here we consider the class of block sparse vectors  $\mathcal{F}_{N,\varepsilon,B}$  (see Section 3 for a formal definition). We use two block-separable denoisers: either block soft thresholding (for block length  $B$ , the  $B$ -variate nonlinearity obeys  $\eta_{B,\lambda}(y) = y \cdot (1 - \|y\|_2/\lambda)_+$ ) or block James-Stein denoiser. We will compute the minimax MSE function  $M_B(\varepsilon|\text{BlockSoft})$ , and bound the minimax MSE function  $M_B(\varepsilon|\text{JamesStein})$ . We will verify that the phase transition curve for optimally-tuned AMP with block-separable denoisers follows the general formula.

Notice that, as demonstrated numerically in [DMM09], and proved in [BM11b] in the case of Gaussian sensing matrices, soft-thresholding AMP reconstruction coincides with LASSO reconstruction (in the large system limit). By the above results, firm-shrinkage AMP and minimax AMP both outperform LASSO reconstruction. Correspondingly, it can be argued that blocksoft thresholding AMP coincides asymptotically with group LASSO, and hence James-Stein AMP outperforms the latter.

In all of the above examples, the denoisers are coordinatewise or at least blockwise separable. We next consider examples where the denoiser has more subtle structure. We find that formula (1.13) applies more generally.

**Monotone Regression.** We consider the class  $\mathcal{F}_{N,\varepsilon,\text{mono}}$  of vectors that are monotone with at most  $N\varepsilon$  points of increase. As denoiser, we use the least-squares projection  $\eta$  onto the cone of monotone increasing functions.

**Total variation minimization.** We consider the class  $\mathcal{F}_{N,\varepsilon,TV}$  of vectors that have at most  $N\varepsilon$  points of change. The denoiser  $\eta$  minimizes the residual sum of squares penalized by  $\tau$  times the total variation of the signal.

In these cases, evaluating the asymptotic minimax MSE is more challenging than for separable denoisers and simpler classes of signals. Nevertheless, we will show that it can be done quite explicitly. We find well-defined phase transitions for AMP reconstruction, precisely at the location predicted by the general formula (1.13).

## 1.6 Contributions

We list eight contributions, beginning with the two most obvious:

1. *Application of the AMP framework to a wider range of shrinkers  $\eta(\cdot)$ .* We implement and study AMP algorithms that don't correspond to  $\ell_1$  penalization (e.g. the Firm and Minimax scalar shrinkers) and also that don't correspond to scalar separable nonlinearities: both the block separable case and the general non-separable cases.
2. *A formula for phase transitions of AMP algorithms.* We confirm that formula (1.13) accurately describes the sparsity-undersampling tradeoff under which AMP algorithms successfully recover a sparse structured signal from underdetermined measurements. We prove that this relation follows from the state evolution formalism.
3. *A formula predicting the phase transitions of many convex optimization problems.* Much work on compressed sensing establishes the possibility of recovery under sparsity by solving convex optimization problems. Unfortunately, considerable work was required to obtain sharp phase transition results for *one* convex optimization algorithm:  $\ell_1$  minimization [Don06, DT05, DT09a, DT10b]. The arguments needed to attack –for example– the block-sparsity case seemed to be quite different [Sto10].

As demonstrated in [DMM11] and proved in [BM11a] in the case of the LASSO, there exists a correspondence between convex optimization methods and specific AMP algorithms. We will show that this correspondence is considerably more general. This provides a unified approach which yields sharp phase transition predictions in numerous cases.

4. *Reconstruction approaches not based on convex penalization, with sharp guarantees.* We introduced three new AMP algorithms, based on Firm, Minimax, and James-Stein shrinkage, which do not correspond to any obvious convex penalization. These methods have better phase transitions than the corresponding convex optimization problems, in their domains (e.g. Firm and Minimax outperform  $\ell_1$  minimization, while James-Stein outperforms block soft shrinkage for large  $B$ ). We show that these algorithms are in correspondence with penalized least square problems, but that the implied penalties are nonconvex<sup>2</sup>. Nevertheless,

---

<sup>2</sup>Fornasier and Rauhut [FR08] show that some denoisers corresponding to non-convex penalties can be implemented via convex optimization by adding suitable auxiliary variables. This is the case, for instance, for firm thresholding

AMP appear to converge to the correct solution with high probability, as long as the under-sampling ratio is above the phase transition boundary. In the interior of the success phase, these methods typically converge exponentially fast.

5. *Limited benefit of nonconvex penalization for ordinary sparsity.* Within the class of scalar separable AMP algorithms, the best achievable phase transition is obtained by the minimax shrinker. Unfortunately the improvement in the transition is relatively small.
6. *Existence of algorithms for the block sparse case approaching “ideal” behavior.* Most attention in the group sparse case concerns block soft thresholding and the corresponding  $\ell_2 - \ell_1$  penalized regression, a.k.a. group LASSO [YL10]. We show here that the phase transitions for block thresholding do not tend as  $B \rightarrow \infty$  to the ideal transition, i.e. that compressed sensing reconstruction is possible as soon as  $\delta > \varepsilon$  (i.e. from, roughly, as many measurements as nonzeros). On the other hand, we show that positive-part James-Stein shrinkage *does* tend to such an ideal limit.
7. *Identification of combinatorial geometry phase transitions with minimax Mean-Square Error.* The phase transitions for  $\ell_1$  optimization, and for positivity-constrained  $\ell_1$  optimization are determined by combinatorial geometry, see [DT09a]. By our general formula (1.13), these transitions are the same as the minimax MSE in problems of scalar denoising.
8. *Calculation of the minimax MSE of monotone regression and total variation denoising.* We are not aware of any previous work computing the minimax MSE of these denoising procedures under the condition of  $\varepsilon$ -sparse first differences. We prove here a characterization for each of these cases and show that it agrees with the phase transition of both AMP and convex optimization algorithms.

A conjectures flow naturally from this work:

*State Evolution accurately describes the behavior of a wide range of AMP algorithms, for large system sizes  $N$ .* State evolution is a formalism that allows to characterize the asymptotic behavior of AMP as the number of dimension tend to infinity [DMM09]. We show in Section 6 that the general relation (1.13) can be proved by assuming state evolution to hold.

In the case of separable denoisers, under suitable regularity conditions, the correctness of state evolution as a description of AMP is proved by [BM11a]. Since formula (1.13) is apparently successful beyond the separable case, it is natural to conjecture that state evolution applies much more generally than to the cases proven so far.

Our study supports the general conclusion that AMP provides a general tool in compressed sensing, that is applicable beyond simple sparse signal. If one knows that a certain shrinker is appropriate for denoising a certain type of signal, then the corresponding AMP algorithm provides an efficient reconstruction method for the associate compressed sensing problem. The denoising minimax MSE then maps to the sparsity undersampling tradeoff.

An interesting research direction is the study of the noisy linear model  $y = Ax_0 + w$ , whereby  $w$  is a noise vector (e.g.  $w \sim \mathbf{N}(0, \sigma^2 \mathbf{I}_{m \times m})$ ). In analogy [DMM11], we expect reconstruction to be stable with respect to noise for  $\delta < M(\varepsilon|\eta)$  and instable for  $\delta > M(\varepsilon|\eta)$ .

---

denoising. Unfortunately, the same method does not apply –in general– to the compressed sensing problem (i.e. for non-diagonal matrices  $A$ ).

## 1.7 Related literature

Approximate message passing algorithms for compressed sensing reconstruction were introduced in [DMM09]. They were largely motivated by the connection with message passing algorithms in iterative decoding systems [RU08], and with mean field methods in statistical physics [MM09] (in particular the cavity method and TAP equations). We refer to [DMM11] for a discussion of these connections.

The original AMP framework [DMM09, DMM10] included iterations of the form defined in Eqs. (1.7), (1.8), (1.9) whereby the denoiser is separable. While this covers the Firm and Minimax shrinkage rules studied in this paper, it did not include the various non-separable denoisers we discuss below, namely the block, monotone and total variation denoisers. Further, in [DMM09], the phase transition behavior was validated numerically only for Soft, SoftPos and Cap denoisers, that are in correspondence with well-studied convex optimization problems. The extension to a noisy linear model  $y = Ax_0 + w$ , with  $w \in \mathbf{R}^n$  a vector of iid random entries was carried out in [DMM11]. We also refer to [Mon12] for an overview of this work.

Several papers investigate generalizations of the original framework put forward in [DMM09]. The paper [BM11a] defines a general class of approximate message passing algorithms for which the state evolution was proved to be correct. This include in particular all separable Lipschitz-continuous denoisers. Generalizations of this result were proved in [BLM12, JM12]. Notice that all the separable denoisers treated in this papers are Lipschitz continuous with the exception of hard thresholding. While the last case is not covered by [BM11a], we expect state evolution to hold for hard thresholding AMP as well, by a suitable approximation argument.

Rangan [Ran11] introduces a class of generalized approximate message passing (G-AMP) algorithms that cope with –roughly– two extensions of the basic noisy linear model. First, the noisy measurement vector  $y$  can be a non-linear (random) function of the noiseless measurement  $Ax_0$ . Second, each of the ‘coordinates’ of  $x_0$  can itself be a –low dimensional– vector. Interesting applications of this framework were developed in [KGR11, Sch11]. Let us notice that G-AMP does not cover any of the non-separable cases treated here (even the block sparse example), and hence provides a generalization in an ‘orthogonal’ direction.

In a parallel line of work, Schniter applied AMP to a number of examples in which the signal  $x_0$  has a structured prior [Sch10, SSS10, SPS10]. Inference with respect to the prior is carried out using belief propagation, and this is combined with AMP to compute *a posteriori* estimates. This type of application fits within the class of problems studied here, by choosing the denoiser  $\eta_t$  in Eq. (1.8) be given by the appropriate conditional expectation with respect to the signal prior. Note however that the general scheme provided by Eqs. (1.7), (1.8), (1.9) encompasses cases in which the denoiser is not the Bayes estimator for a specific prior.

A special case of known prior is the one in which  $x_0 = \mathbf{X} \sim \nu_N$  is distributed according to the (known) product measure  $\nu_N = \nu \times \dots \times \nu$  (i.e. the coordinates of  $\mathbf{X}$  are iid with known distribution  $\nu$ ). The fundamental limits for compressed sensing reconstruction were established in [WV10]. The natural AMP algorithm uses in this case a posterior expectation denoiser [DMM10]. It was proved in [DJM11] that, for suitable sensing matrices with heteroscedastic entries, this approach achieves the fundamental limits of [WV10] (this approach was put forward in [KMS<sup>+</sup>12] on the basis of a statistical physics argument). This case fits within the general philosophy of the present paper whereby the class  $\mathcal{F}_{N,\varepsilon}$  consists of a single distribution, namely  $\mathcal{F}_{N,\varepsilon} = \{\nu_N\}$ . However, we prefer not treating this example in the present paper because it is a degenerate case, and the fact that

$\mathcal{F}_{N,\varepsilon}$  is not scale invariant leads to some technical differences. We refer instead to [DJM11].

Statistical physics methods were also used to study  $\ell_1$ -based reconstruction methods in [RFG09, KWT09].

Maleki, Anitori, Yang and Baraniuk [MAYB11] used methods analogous to the ones developed here to study phase transitions for compressed sensing with complex vectors. This is closely related to the block-separable setting considered in Section 3 (there is however some difference in the structure of the sensing matrix).

Structured sparsity models are studied from a different point of view in [BCDH10, CHDB08, CICB10]. Those works focus on deriving sparsity models that capture a variety applications, and of convex relaxations that promote the relevant sparsity patterns. Reconstruction guarantees are proved under suitable ‘isometry’ assumptions on the sensing matrix.

Closer to our approach is a recent series of papers [CRPW11, RRN11, RRN11], considering general classes of structured signals under random measurements. Let us emphasize two important differences with respect to our work. First, these papers only deal with convex reconstruction methods, while we shall analyze several approaches that are not derived from convex optimization and demonstrate improvements. Second, they establish reconstruction guarantees using concentration-of-measure arguments, while we propose *exact* asymptotics (essentially based on weak convergence), which enables us to unveil the key relation (1.13) between denoising and the compressed sensing phase transition.

## 2 Scalar-separable denoisers

In this section we study scalar-separable denoisers, cf. Eq. (1.6), that further satisfy the scaling relation (1.3). Unless stated otherwise, we will assume that signals belong to the simple sparsity class introduced in Eq. (1.2), to be denoted as  $\mathcal{F}_{N,\varepsilon}$ .

### 2.1 Minimax MSE of a separable denoiser

As mentioned in the previous section, the computation of the minimax MSE is greatly simplified for separable denoisers. We state and prove the following elementary result in greater generality than necessary for this section. (In particular  $\mathcal{F}_{N,\varepsilon}$  is here a general family of probability distributions.)

**Lemma 2.1.** *Let  $\mathcal{F}_{N,\varepsilon} \subseteq \mathcal{P}(\mathbf{R}^N)$  be any family of probability distributions satisfies the following conditions: (i) If  $\nu_1 \in \mathcal{F}_{1,\varepsilon}$ , then defining  $\nu_N \equiv \nu_1 \times \cdots \times \nu_1$  ( $N$  times), we have  $\nu_N \in \mathcal{F}_{N,\varepsilon}$ ; (ii) Viceversa, if  $\nu_N \in \mathcal{F}_N$ , then letting  $\nu_{N,i}$  denote the  $i$ -th marginal of  $\nu_N$ , we have  $\bar{\nu}_N \equiv N^{-1} \sum_{i=1}^N \nu_{N,i} \in \mathcal{F}_{1,\varepsilon}$ .*

*Then, for any separable denoiser  $\eta$ , and for any  $N$ ,*

$$M(\varepsilon|\eta) = M(\mathcal{F}_{N,\varepsilon}|\eta) = M(\mathcal{F}_{1,\varepsilon}|\eta).$$

*Proof.* Fix  $\tau \in \Theta$  and define, for  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{N \times N})$ ,

$$M(\mathcal{F}_{N,\varepsilon}|\eta, \tau) \equiv \frac{1}{N} \sup_{\nu_N \in \mathcal{F}_{N,\varepsilon}} \mathbb{E}_{\nu_N} \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \mathbf{Z}; \tau)\|_2^2 \right\}.$$

The lemma then follows immediately if we prove that, for any  $N$ ,  $M(\mathcal{F}_{N,\varepsilon}|\eta, \tau) = M(\mathcal{F}_{1,\varepsilon}|\eta, \tau)$ . In order to prove the last statement, first notice that, by property (i):

$$\begin{aligned} M(\mathcal{F}_{N,\varepsilon}|\eta, \tau) &\geq \frac{1}{N} \sup_{\nu_1 \in \mathcal{F}_{1,\varepsilon}} \mathbb{E}_{\nu_1 \times \dots \times \nu_1} \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \mathbf{Z}; \tau)\|_2^2 \right\} \\ &= \sup_{\nu_1 \in \mathcal{F}_{1,\varepsilon}} \mathbb{E}_{\nu_1} \left\{ \|X_1 - \eta(X_1 + Z; \tau)\|_2^2 \right\} = M(\mathcal{F}_{1,\varepsilon}|\eta, \tau). \end{aligned}$$

The proof is finished by property (ii), since

$$\begin{aligned} M(\mathcal{F}_{N,\varepsilon}|\eta, \tau) &= \frac{1}{N} \sup_{\nu_N \in \mathcal{F}_{N,\varepsilon}} \sum_{i=1}^N \mathbb{E}_{\nu_N} \left\{ \|X_i - \eta(X_i + Z; \tau)\|_2^2 \right\} \\ &= \sup_{\nu_N \in \mathcal{F}_{N,\varepsilon}} \mathbb{E}_{\bar{\nu}_N} \left\{ \|X - \eta(X + Z; \tau)\|_2^2 \right\} \leq M(\mathcal{F}_{N,\varepsilon}|\eta, \tau). \end{aligned}$$

□

This lemma reduces the problem to solving a minimax scalar estimation problem. This problem was characterized before for soft thresholding  $\eta = \eta^{soft}$ , positive soft thresholding  $\eta = \eta^{softpos}$ , and also hard thresholding  $\eta(y; \tau) = y1_{\{|y| > \tau\}}$  [DDS92, DJ94]. Plots of the minimax soft threshold  $\tau^*(\varepsilon)$  and the minimax MSE are available in [DMM09, DMM11]. Such plots also appear later in this paper as baselines for comparison of interesting new families, namely Firm and Minimax shrinkage.

## 2.2 Firm shrinkage

A frequently-voiced criticism of  $\ell_1$  minimization and soft thresholding is the tendency to shrink large values by more than warranted. In the mid 1990's, *firm* shrinkage was introduced to correct this tendency by Gao and Bruce [GB97]. As suggested by the name, this denoiser is intermediate between soft and hard thresholding: it is continuous like soft thresholding, but does not shrink large values, like hard thresholding. Formally, for  $\tau = (\tau_1, \tau_2) \in \Theta$ ,  $\Theta \equiv \{(\tau_1, \tau_2) : 0 \leq \tau_1 < \tau_2 < \infty\}$ , we have

$$\eta^{firm}(y; \tau_1, \tau_2) = \begin{cases} 0 & |y| < \tau_1, \\ \text{sgn}(y) (|y| - \tau_1) \tau_2 / (\tau_2 - \tau_1) & \tau_1 < |y| < \tau_2, \\ y & |y| > \tau_2. \end{cases}$$

Soft and hard thresholding can be recovered as limiting cases:

$$\eta^{soft}(y; \tau_1) = \lim_{\tau_2 \rightarrow \infty} \eta^{firm}(y; \tau_1, \tau_2), \quad \eta^{hard}(y; \tau_1) = \lim_{\tau_2 \rightarrow \tau_1} \eta^{firm}(y; \tau_1, \tau_2). \quad (2.1)$$

Lemma 2.1 yields the following formula for the minimax MSE of firm shrinkage:

$$M(\varepsilon|\text{Firm}) = M(\mathcal{F}_{1,\varepsilon}|\eta^{firm}) = \inf_{0 \leq \tau_1 < \tau_2} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}} \mathbb{E}_{\nu} \left\{ [X - \eta^{firm}(X + Z; \tau_1, \tau_2)]^2 \right\}. \quad (2.2)$$

Notice that the supremum over  $\nu$  is necessarily achieved at a probability distribution of the form  $\nu = \nu_{\varepsilon, \mu} = (\varepsilon/2)\delta_{\mu} + (1 - \varepsilon)\delta_0 + (\varepsilon/2)\delta_{-\mu}$ , for some  $\mu \in [0, \infty]$  (indeed the risk is an even function of  $X$ , and the extreme points of the even subset of  $\mathcal{F}_{1,\varepsilon}$  take this form). Explicit expressions under this distribution are given in [GB97]. The computation of  $M(\varepsilon|\text{Firm})$  is therefore reduced to the

$\varepsilon$	$M(\varepsilon \text{Hard})$	$M(\varepsilon \text{Soft})$	$M(\varepsilon \text{Firm})$	$M(\varepsilon \text{Minimax})$
0.010	0.0729	0.0612	0.0552	0.0533
0.025	0.1547	0.1231	0.1137	0.1093
0.050	0.2676	0.2039	0.1921	0.1841
0.100	0.4497	0.3288	0.3165	0.3025
0.150	0.5960	0.4279	0.4171	0.3983
0.200	0.7161	0.5111	0.5024	0.4802
0.250	0.8141	0.5829	0.5763	0.5516

Table 1: Minimax MSE of various separable denoisers applied to sparse vectors from the class  $\mathcal{F}_{N,\varepsilon}$ , across sparsity levels

calculus problem of finding the saddle point of  $f(\mu, \tau_1, \tau_2) \equiv \mathbb{E}_{\nu_{\varepsilon,\mu}} \{ [X - \eta^{\text{firm}}(X + Z; \tau_1, \tau_2)]^2 \}$ . This can be efficiently solved numerically, yielding the minimax risk and the optimal thresholds  $\tau_1^*(\varepsilon)$  and  $\tau_2^*(\varepsilon)$ . It follows immediately from the definition that  $0 \leq M(\varepsilon|\text{Firm}) \leq 1$ , that  $M$  is monotone increasing with  $\varepsilon$ , that  $M(\varepsilon|\text{Firm}) \rightarrow 1$  as  $\varepsilon \rightarrow 1$  and that  $M(\varepsilon|\text{Firm}) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ .

Figure 1 and Table 1 show the minimax MSE for firm shrinkage as resulting from this calculation. The figure also shows similar results for soft and hard thresholding, for comparative purposes. Over the range presented, the minimax MSE for firm thresholding is *strictly smaller* than the MSE for hard or soft thresholding. Namely, over this range of  $\varepsilon$ ,

$$M(\varepsilon|\text{Firm}) < M(\varepsilon|\text{Soft}) < M(\varepsilon|\text{Hard}).$$

This validates the criticisms of soft thresholding, which is often said to shrink large values too heavily<sup>3</sup>.

Figure 2.2 shows the minimax thresholds. At least for  $\varepsilon < 1/3$  we see clearly that  $\tau_1^*(\varepsilon) < \tau_2^*(\varepsilon) < \infty$ , so firm thresholding is preferred over the limiting cases of hard and soft thresholding<sup>4</sup>. Figure 2.2 shows the corresponding minimax denoisers for specific values of  $\varepsilon$ . Finally, Figure 2.2 plots the minimax value of  $\mu$  as a function of  $\varepsilon$  (corresponding to the minimax probability distribution  $\nu_{\varepsilon,\mu}$ ).

### 2.3 Minimax shrinkage

The previous example showed that a parametric family of shrinkers can improve on soft thresholding, and hence improve the predicted phase transition according to (1.13). The ultimate improvement one could make in this direction is to use the *globally* minimax nonlinear shrinker. This is the separable denoiser  $\eta$  that is minimax not within some parametric family, such as the soft thresholding or the broader firm thresholding family, but minimax over *all* measurable nonlinearities  $\eta : \mathbf{R} \rightarrow \mathbf{R}$ . While this notion might appear somewhat abstract, it can be in fact implemented in practice as illustrated in Figure 2.2, that present plots of the more familiar denoisers (hard, soft, and firm) together with the minimax denoiser.

<sup>3</sup>Note, however, that the use of hard thresholding instead of soft thresholding leads to a larger worst case mean square error.

<sup>4</sup>These are numerical results. It is an open question whether, for  $\varepsilon > 1/3$  the minimax firm threshold have parameter  $\tau_2(\varepsilon) = \infty$  reducing it to soft thresholding.

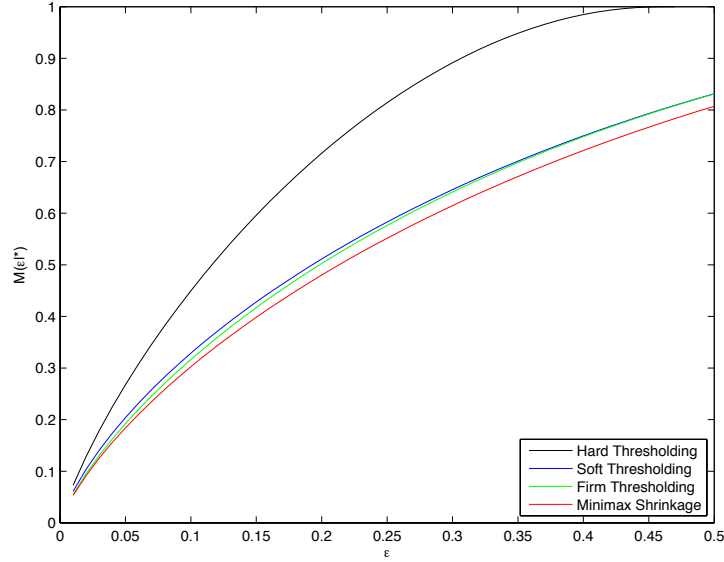


Figure 1: Minimax MSE of various separable denoisers, as a function of sparsity parameter  $\varepsilon$ . The minimax MSE of firm thresholding is very close to that of soft thresholding for  $\varepsilon > 1/3$ . From top to bottom the curves refer to hard thresholding, soft thresholding, firm thresholding and minimax denoiser.

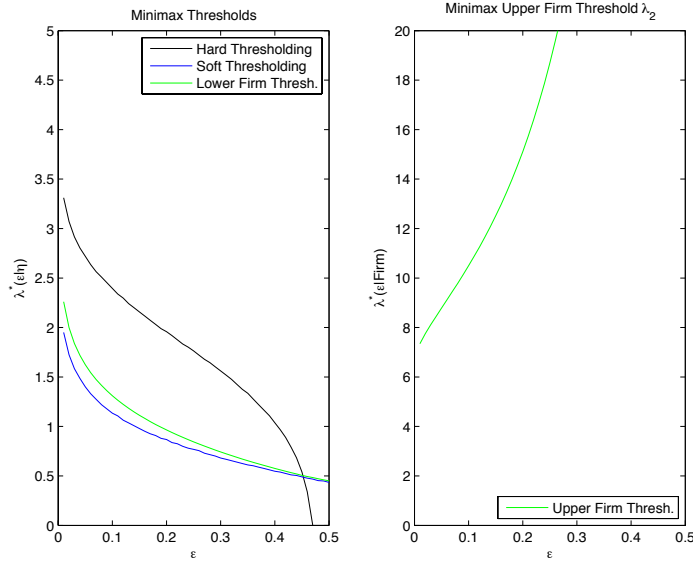


Figure 2: Minimax thresholds of various separable denoisers, as a function of sparsity  $\varepsilon$ . The sudden drop in the value of the hard threshold to zero near 0.45 coincides with the value of the minimax MSE in Figure 1 reaching 1. The numerical approximation to the minimax lower firm threshold approaches the minimax soft threshold as  $\varepsilon$  increases beyond 0.3, and the upper firm threshold increases rapidly.



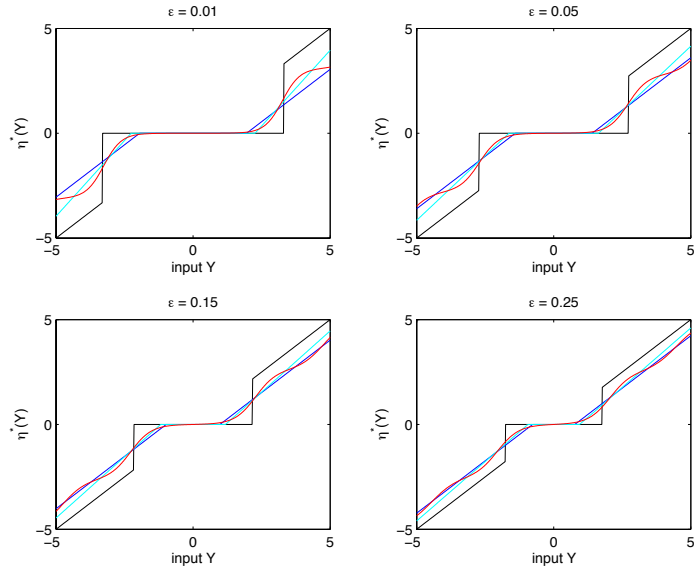


Figure 3: Minimax scalar denoisers of various types, at sparsity  $\varepsilon = 0.01, 0.05, 0.15, 0.25$ . Black: hard thresholding; Blue: soft; Aqua: firm; Red: minimax shrinkage.

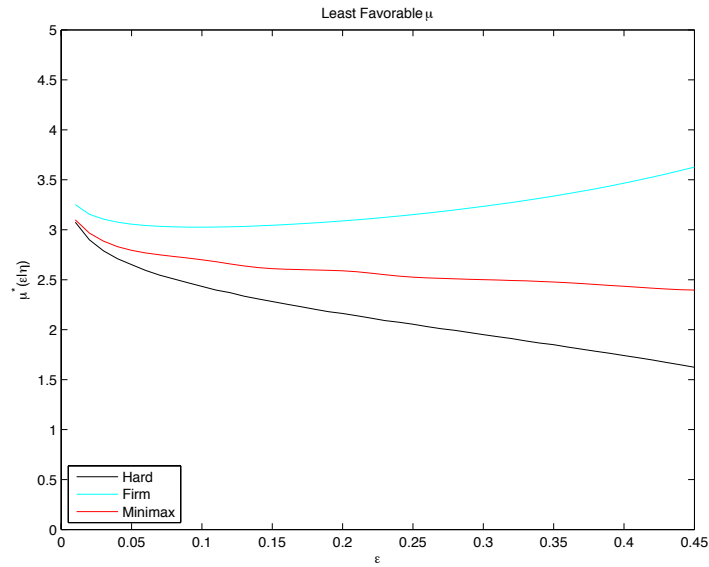


Figure 4: Least favorable  $\mu$  for various separable denoisers, as a function of sparsity  $\varepsilon$ . From top to bottom the curves refer to firm thresholding, minimax denoiser and hard thresholding. Soft thresholding has minimax  $\mu = \pm\infty$ , not shown.

Formally, let  $\Theta \equiv \mathcal{L}(\mathbf{R})$  be the set of all measurable functions  $\tau : \mathbf{R} \rightarrow \mathbf{R}$ , for such a  $\tau$ , set  $\eta^{all}(x; \tau) \equiv \tau(x)$ . The minimax MSE over this class is

$$M(\varepsilon|\text{Minimax}) = \inf_{\eta \in \mathcal{L}(\mathbf{R})} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}} \mathbb{E}\{[X - \eta(X + Z)]^2\}. \quad (2.3)$$

The calculation of this quantity uses a variety of ideas from minimax decision theory, developed through several papers [Bic81, CS81, BC83, DDS92, Joh94b, Joh94a, DJ94]: details are given in Appendix B. A key point of this computation is the characterization of the minimax nonlinearity as the minimal MSE Bayes rule (that is, the conditional expectation) for the so-called least-favorable prior. The least-favorable prior is the solution of Mallows’ classical Fisher information problem [Mal78], for which we compute numerical upper and lower bounds that coincide within the stated precision.

Table 1 and Figure 1 present numerical values associated with the solution of the minimax problem. As expected,  $M(\varepsilon|\text{Minimax}) < M(\varepsilon|\text{Firm}) \leq M(\varepsilon|\text{Soft})$ , i.e. optimizing over all nonlinearities yields a smaller mean square error than soft or firm thresholding. On the other hand, Table 1 shows that the improvements are typically of size 0.01 or smaller over the range  $\varepsilon \in (0.01, 0.25)$ . For very small  $\varepsilon$  it was pointed out in [DMM09] that [DJ94] implies

$$\lim_{\varepsilon \rightarrow 0} \frac{M(\varepsilon|\text{Minimax})}{M(\varepsilon|\text{Soft})} = 1.$$

In the limit of extreme sparsity, there is nothing to be gained by completely general nonlinearities over soft thresholding. The improvement is non-vanishing, but moderate for  $\varepsilon$  non-vanishing.

## 2.4 Empirical phase transition behavior

The research hypothesis driving this paper is that Eq. (1.13) describes the phase transition of AMP algorithms. In order to be completely explicit, we need to check the following predictions, for each nonlinearity  $\eta$  of interest

1. There exists a curve  $\varepsilon \mapsto \delta(\varepsilon|\eta)$  such that for  $\delta > \delta(\varepsilon|\eta)$  the corresponding AMP algorithm will typically succeed in reconstructing the unknown signal  $x_0$ , and for  $\delta < \delta(\varepsilon|\eta)$  the algorithm will typically fail.
2. The curve is related to the corresponding scalar denoising problem by  $\delta(\varepsilon|\eta) = M(\varepsilon|\eta)$ .

We now test this hypothesis for the firm and globally minimax nonlinearities  $\eta \in \{\eta^{firm}, \eta^{all}\}$ .

Our experiment was conducted along the same lines as [MD10, DT09b, DMM09, DMM11, BT10]. We considered a range of problem sizes  $N \in \{1000, 2000, 4000\}$  and a range of sparsity parameters  $\varepsilon \in \{0.01, 0.02, 0.05, .10, 0.15, 0.20, 0.25\}$ , and a grid of  $\delta$  values surrounding the predicted phase transition  $\delta(\varepsilon|\eta)$ . We ran  $N_{\text{sample}} = 1000$  Monte Carlo reconstructions at each parameter combination. We declared “success” when the relative mean-squared error was below 1%:

$$\frac{\|\widehat{x}^t - x_0\|_2^2}{\|x_0\|_2^2} < 0.01.$$

We used  $t = 300$  iterations of AMP<sup>5</sup>. We repeated a subset of our simulations with different requirements on  $\|\widehat{x}^t - x_0\|_2^2/\|x_0\|_2^2$  and different number of iterations, without significant changes in the

---

<sup>5</sup>In most cases the mentioned convergence criterion is reached after a much smaller number of iterations (roughly 20)

threshold location. This point is further justified in Appendix C. In the interest of reproducibility, a suite of Java classes for carrying out these and other simulations in the paper is made available as [DJM12].

Firm Thresholding with Minimax Tuning							
$\varepsilon$	Pred	off.1000	off.2000	ci.1000.lo	ci.1000.hi	ci.2000.lo	ci.2000.hi
0.01	0.0550	0.0057	0.0040	0.0054	0.0060	0.0038	0.0043
0.02	0.0960	0.0065	0.0039	0.0062	0.0069	0.0036	0.0041
0.05	0.1920	0.0064	0.0047	0.0060	0.0068	0.0043	0.0050
0.10	0.3170	0.0067	0.0050	0.0063	0.0071	0.0047	0.0054
0.15	0.4190	0.0070	0.0049	0.0065	0.0074	0.0045	0.0052
0.20	0.5070	0.0068	0.0052	0.0064	0.0073	0.0048	0.0055
0.25	0.5840	0.0069	0.0055	0.0064	0.0073	0.0051	0.0058
Minimax Denoiser							
$\varepsilon$	Pred	off.1000	off.2000	ci.1000.lo	ci.1000.hi	ci.2000.lo	ci.2000.hi
0.01	0.0530	0.0075	0.0052	0.0072	0.0078	0.0050	0.0055
0.05	0.1840	0.0126	0.0099	0.0122	0.0130	0.0096	0.0102
0.10	0.3020	0.0183	0.0164	0.0178	0.0188	0.0160	0.0167
0.15	0.3980	0.0147	0.0127	0.0142	0.0151	0.0124	0.0131
0.20	0.4800	0.0148	0.0114	0.0143	0.0152	0.0110	0.0118
0.25	0.5510	0.0155	0.0124	0.0151	0.0160	0.0120	0.0127
Soft Thresholding with Minimax Tuning							
$\varepsilon$	Pred	off.1000	off.2000	ci.1000.lo	ci.1000.hi	ci.2000.lo	ci.2000.hi
0.01	0.0610	0.0061	0.0045	0.0058	0.0064	0.0043	0.0048
0.02	0.1040	0.0065	0.0052	0.0062	0.0069	0.0049	0.0055
0.05	0.2030	0.0089	0.0072	0.0085	0.0093	0.0068	0.0075
0.15	0.4270	0.0117	0.0100	0.0112	0.0122	0.0096	0.0104
0.20	0.5110	0.0118	0.0100	0.0113	0.0124	0.0096	0.0104
0.25	0.5830	0.0127	0.0106	0.0122	0.0132	0.0102	0.0110

Table 2: Empirical phase transition studies for AMP algorithms on the class of simple sparse vectors  $\mathcal{F}_{N,\varepsilon}$  based on three separable denoisers. At each value of  $\delta, \varepsilon$  we carried out 1000 Monte Carlo repetitions at  $N = 1000, 2000$ . Column *Pred* corresponds to general formula (1.13) yielding the critical value of  $\delta$ , and is thought to be accurate for large  $N$ . The columns *off.1000* and *off.2000* report values of the offset  $\widehat{\text{PT}}$  estimated by logistic regression, using Eqs. (2.4) and (2.5). (Note that offsets are systematically smaller at  $N = 2000$  than at  $N = 1000$ , consistent with hypothesis that they vanish as  $N \rightarrow \infty$ .) Columns *ci.1000.lo*, *ci.1000.hi*, *ci.2000.lo*, *ci.2000.hi* give lower and upper endpoints of formal 95% confidence intervals for the offset  $\widehat{\text{PT}}$ .

We proceeded to analyse the outcomes of these numerical simulations as follows, see also Appendix H (a similar analysis was already carried out in [DMM09, DT09b]). The simulations generated a data set, containing, for each algorithm and each fixed  $\varepsilon$ , a list of values  $\delta_i$  and empirical success fractions  $\widehat{p}_i$ . The success fractions observed at  $\delta > M(\varepsilon|\eta)$  were indeed typically better than 50% and at  $\delta < M(\varepsilon|\eta)$  were typically worse than 50%.

To quantify this tendency, we fit a logistic regression

$$\log \frac{\widehat{p}_i}{1 - \widehat{p}_i} = \alpha + \beta(\delta_i - \delta(\varepsilon|\eta)), \quad (2.4)$$

where  $\delta(\varepsilon|\eta) = M(\varepsilon|\eta)$  was computed analytically using ideas mentioned earlier. The choice of the model (2.4) is motivated by the observation that the success probability increases rapidly around the phase transition, and by the common statistical use of logistic models. Also, similar models have been proved to be asymptotically correct in analogous phase transition phenomena [DM08].

For each set of data corresponding to given  $(N, \varepsilon)$  and each non-linearity, we estimate  $\alpha$  and  $\beta$  from the logit fit, leading to values  $\widehat{\alpha}, \widehat{\beta}$ . Using these quantities, we estimate the phase transition location as the value at which the probability  $\widehat{p}$  of success is 50%. Using Eq. (2.4) this corresponds to  $\widehat{\alpha} + \widehat{\beta}(\delta - \delta(\varepsilon|\eta)) = 0$ , i.e.  $\delta = \delta(\varepsilon|\eta) - (\widehat{\alpha}/\widehat{\beta})$ . We are therefore led to define the offset between the empirical phase transition and the prediction  $\delta(\varepsilon|\eta) = M(\varepsilon|\eta)$  as

$$\widehat{\text{PT}} \equiv -\frac{\widehat{\alpha}}{\widehat{\beta}}. \quad (2.5)$$

In order to check the general relation provided by Eq. (1.13) we need to show that  $\widehat{\text{PT}}(N, \varepsilon)$  tends to zero as  $N$  gets large, to within the statistical uncertainty. In Table 2 we report our results on the empirical phase transition, confirming that indeed the offset is small and decreasing with  $N$ .

A few additional remarks on these data are of interest:

- We calculated formal 95% confidence intervals for  $\widehat{\text{PT}}$ , indicating the tight control we have of the correct value.
- As in earlier studies [DT09b], we expect that  $\widehat{\text{PT}}(N, \varepsilon)$  tend 0 at a rate that is inversely proportional to a power of  $N$ . Namely

$$\widehat{\text{PT}}(N, \varepsilon) \approx \frac{\text{const}}{N^\gamma} + \text{sampling error},$$

for some  $\gamma \in (0, 1]$ . Our data supports this relationship, with  $\gamma \approx 1/3$ . See Appendix H.

- Denoting by  $\widehat{\beta}_N$  the fitted slope coefficient at dimension  $N$ , evidence that  $\widehat{\beta}_N$  is increasing with larger  $N$  indicates that a sharpening of the phase transition is indeed occurring. Appendix H shows that  $\widehat{\beta}_N \sim \sqrt{N}$  is consistent with our data.

We refer to Appendices G and H for further details.

### 3 Block-separable denoisers

We now turn to the case of block-structured sparsity, first introducing some notational conventions. We partition the vector  $x = (x_1, x_2, \dots, x_N)$  into  $M$  blocks each of size  $B$ . Denoting by  $\text{block}_m(x) = (x_{(m-1)B+1}, \dots, x_{mB})$  the  $m$ -th block, we hence write

$$x = (\text{block}_1(x), \dots, \text{block}_M(x)),$$

with, of course,  $N = MB$ .

A block-separable denoiser is a mapping  $\eta(\cdot; \tau) : \mathbf{R}^N \rightarrow \mathbf{R}^N$  that decomposes according to the above partition:

$$\eta(x; \tau) = (\eta(\text{block}_1(x); \tau), \dots, \eta(\text{block}_M(x); \tau)), \quad (3.1)$$

where, with an abuse of notation, we use the same symbol to denote the single-block denoiser  $\eta(\cdot; \tau) : \mathbf{R}^B \rightarrow \mathbf{R}^B$ . The last equation replaces Eq. (1.6) which correspond to the simple separable case. The above form applies to noise with variance  $\sigma^2 = 1$ . For general variance we adopt again the scaling relation (1.3).

We will apply these denoiser to signals from the block-sparse class  $\mathcal{F}_{N,\varepsilon,B}$  defined as follows for  $\varepsilon \in [0, 1]$ ,  $B \in \mathbf{N}$ ,  $M \equiv N/B$ ,

$$\mathcal{F}_{N,\varepsilon,B} \equiv \left\{ \nu_N \in \mathcal{P}(\mathbf{R}^N) : \mathbb{E}_{\nu_N} [\#\{i \in [M] : \text{block}_i(\mathbf{X}) \neq 0\}] \leq M\varepsilon \right\}. \quad (3.2)$$

In words, this is the class of (random) vectors  $\mathbf{X}$  that have (in expectation) at most  $M\varepsilon$  blocks different from 0. For simplicity, we will write  $\mathcal{F}_{\varepsilon,B}$  for the  $M = 1$  case,  $\mathcal{F}_{B,\varepsilon,B}$

The same simplifications described in Section 2.1 applies, with obvious modifications, to the present context.

**Lemma 3.1.** *Let  $\mathcal{F}_{N,\varepsilon} \subseteq \mathcal{P}(\mathbf{R}^N)$  be any family of probability distributions satisfies the following conditions: (i) If  $\nu_B \in \mathcal{F}_{B,\varepsilon}$ , then defining  $\nu_N \equiv \nu_B \times \dots \times \nu_B$  ( $M = N/B$  times), we have  $\nu_N \in \mathcal{F}_{N,\varepsilon}$ ; (ii) Viceversa, if  $\nu_N \in \mathcal{F}_N$ , then letting  $\nu_{N,i}$  denote the marginal of the  $i$ -th block under  $\nu_N$ , we have  $\bar{\nu}_N \equiv M^{-1} \sum_{i=1}^M \nu_{N,i} \in \mathcal{F}_{B,\varepsilon}$ .*

*Then, for any block-separable denoiser  $\eta$ , and for any  $N$  multiple of  $B$*

$$M(\varepsilon|\eta) = M(\mathcal{F}_{N,\varepsilon}|\eta) = M(\mathcal{F}_{B,\varepsilon}|\eta).$$

The proof is omitted since it is an immediate generalization of the one of Lemma 3.1. The class  $\mathcal{F}_{N,\varepsilon,B}$  to be studied in the rest of this section clearly satisfy the assumption of this lemma.

### 3.1 Block soft thresholding

Block-soft thresholding  $\eta^{\text{soft}}(\cdot; \tau) : \mathbf{R}^B \rightarrow \mathbf{R}^B$  is the nonlinear shrinker defined by letting, for  $y \in \mathbf{R}^B$ , and  $\tau \in \mathbf{R}_+$ ,

$$\eta^{\text{soft}}(y; \tau) = \left(1 - \frac{\tau}{\|y\|_2}\right)_+ \cdot y, \quad (3.3)$$

where  $(z)_+ \equiv \max(z, 0)$ . The case  $B = 1$  reduces to traditional soft thresholding of Example 1.2. More generally,  $\eta^{\text{soft}}(y; \tau)$  shrinks its argument  $y$  to 0 if  $\|y\|_2 \leq \tau$  and moves it by an amount  $\tau$  towards the origin otherwise. It can also be regarded as the solution of a penalized least squares problem, namely

$$\eta^{\text{soft}}(y; \tau) \equiv \arg \min_{x \in \mathbf{R}^B} \left\{ \frac{1}{2} \|y - x\|_2^2 + \tau \|x\|_2 \right\}.$$

Block thresholding has previously been considered by Hall, Kerkyacharian and Picard [HKP98] and by Cai [Cai99] although in specific ‘wavelet’ applications.

In view of Lemma 3.1, computing the minimax risk reduce to solve the block minimax problem (in this section we add the subscript  $B$  for greater clarity)

$$M_B(\varepsilon|\text{BlockSoft}) \equiv \frac{1}{B} \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \mathbb{E}_\nu \{ [\mathbf{X} - \eta^{\text{soft}}(\mathbf{X} + \mathbf{Z}; \tau)]^2 \}, \quad (3.4)$$

where expectation is taken with respect to  $X \sim \nu$  independent of  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{B \times B})$ . Notice that the condition  $\nu \in \mathcal{F}_{\varepsilon, B}$  simply amount to saying that  $\nu$  is a probability measure on  $\mathbf{R}^B$  with  $\nu(\{0\}) \geq 1 - \varepsilon$ . The calculation of  $M_B(\varepsilon|\text{BlockSoft})$  can be reduced to a calculus problem. We state the results below deferring calculations to Appendix D.

**Lemma 3.2.** *Let  $X_B$  be a chi-square random variable with  $B$  degrees of freedom and define the functions  $g, h : \mathbf{R}_+ \rightarrow \mathbf{R}$  as follows*

$$h(\tau^2) \equiv \frac{\tau}{\mathbb{E} \left\{ (\sqrt{X_B} - \tau) \mathbb{I}_{X_B \geq \tau^2} \right\}}, \quad g(\tau^2) \equiv \frac{\tau \mathbb{E} \left\{ (\sqrt{X_B} - \tau)^2 \mathbb{I}_{X_B \geq \tau^2} \right\}}{\mathbb{E} \left\{ (\sqrt{X_B} - \tau) \mathbb{I}_{X_B \geq \tau^2} \right\}}.$$

The minimax risk of block soft thresholding over the class  $\mathcal{F}_{N, \varepsilon, B}$  is given by

$$M_B(\varepsilon|\text{BlockSoft}) = \frac{B + \tau^2 + g(\tau^2)}{B(1 + h(\tau^2))}, \quad \varepsilon = \frac{1}{1 + h(\tau^2)}. \quad (3.5)$$

This is a parametric expression for  $\tau \in [0, \infty)$ . The parameter corresponds to the minimax threshold  $\tau$ .

In Figure 3.1 we present graphs of  $M(\varepsilon) = M_B(\varepsilon|\text{BlockSoft})$  as a function of  $\varepsilon$ . It is immediate to prove the following structural properties: (i)  $0 \leq M(\varepsilon) \leq 1$  (the upper bound follows from taking  $\tau = 0$ ); (ii)  $M(\varepsilon)$  is monotone increasing and concave (monotonicity is a consequence of  $\mathcal{F}_{B, \varepsilon} \subseteq \mathcal{F}_{B, \varepsilon'}$  for  $\varepsilon \leq \varepsilon'$ , and concavity follows since any measure in  $\mathcal{F}_{q\varepsilon_1 + (1-q)\varepsilon_2, B}$  can be written as convex combination of measures in  $\mathcal{F}_{\varepsilon_1, B}$  and in  $\mathcal{F}_{\varepsilon_2, B}$ ); (iii)  $M(\varepsilon) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ ;  $M(\varepsilon) \rightarrow 1$  as  $\varepsilon \rightarrow 1$ . (Recall that we are considering the MSE per coordinate.) Associated with the minimax problem is also an optimal threshold value  $\tau^*(\varepsilon|B)$ , that we plot in Figure 3.1.

A particularly interesting case is the one of large blocks. As  $B \rightarrow \infty$  the minimax MSE has a well defined, and particularly explicit limit.

**Lemma 3.3.** *For large  $B$  we have*

$$\lim_{B \rightarrow \infty} M_B(\varepsilon|\text{BlockSoft}) = M_\infty(\varepsilon|\text{BlockSoft}) \equiv 2\varepsilon - \varepsilon^2.$$

Further, when properly normalized, the minimax threshold converges with increasing block size:

$$\lim_{B \rightarrow \infty} \tau^*(\varepsilon|B)/\sqrt{B} = \tau^*(\varepsilon|\infty) \equiv (1 - \varepsilon).$$

This lemma is proved in Appendix D.2.

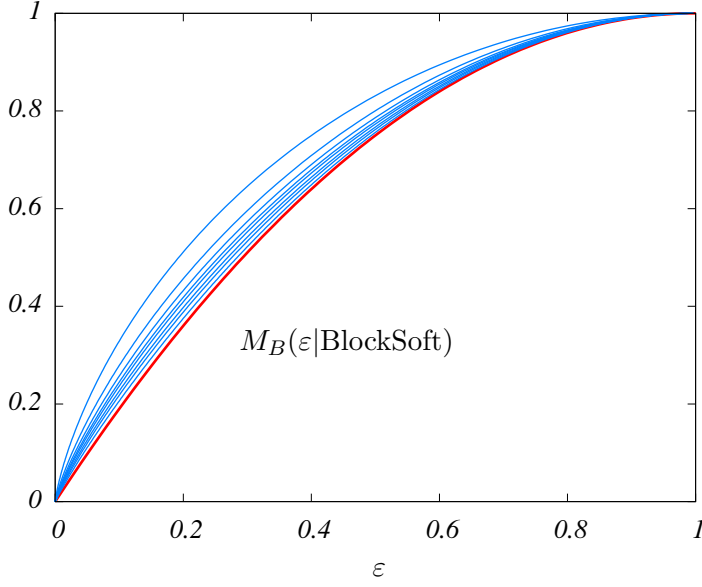


Figure 5: Minimax MSE  $M_B(\varepsilon|\text{BlockSoft})$  of block soft-thresholding for block sparse vectors from the class  $\mathcal{F}_{N,\varepsilon,B}$ , cf. Eq. (3.4). Thin blue curves refer (from top to bottom) to block sizes  $B = 1, 2, 3, 5, 7, 10, 20$ . Thick red line is the large block size limit  $M_\infty(\varepsilon|\text{BlockSoft}) = 2\varepsilon - \varepsilon^2$ .

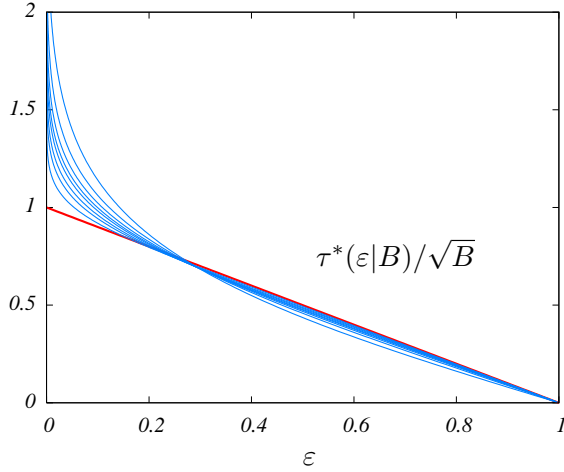


Figure 6: Minimax threshold  $\tau^*(\varepsilon|B)$  of block soft-thresholding for block sparse vectors from the class  $\mathcal{F}_{N,\varepsilon,B}$ . Thin blue lines are the normalized threshold  $\tau^*(\varepsilon|B)/\sqrt{B}$  as a function of  $\varepsilon$ , for  $B = 1, 2, 3, 4, 5, 7, 10, 20$  (from top to bottom on the left side of the plot, and from bottom to top on the right). The thick red line is the large block size limit  $\tau^*(\varepsilon|\text{BlockSoft}) = 1 - \varepsilon$ .

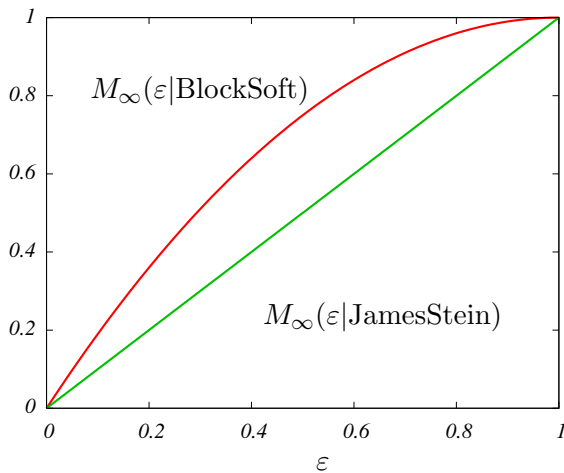


Figure 7: Large block sizes ( $B \rightarrow \infty$ ) MSE of block soft (upper curve) and James-Stein (lower curve) denoisers. The limit soft thresholding minimax MSE is  $M_\infty(\varepsilon|\text{BlockSoft}) = 2\varepsilon - \varepsilon^2$ . The limit James-Stein minimax MSE coincides with the ideal MSE  $M_{\text{Oracle}}(\varepsilon) = \varepsilon$ .

### 3.2 Block James-Stein

From the point of view of  $\ell_1$ -penalized estimation, and compressed sensing, block soft thresholding seems very natural. However, the limiting relationship in Lemma 3.3 shows that this approach leaves room for improvement at large block sizes. A simple upper bound on the MSE is the one achieved by a denoiser which utilizes a special oracle that tells us without error which blocks are zero and which are nonzero. We refer to this as to the *ideal* (or *oracle*) MSE. It is easy to see that the minimax ideal MSE is  $M_B(\varepsilon|\text{Oracle}) = \varepsilon$ . This is considerably smaller than  $M_B(\varepsilon|\text{BlockSoft})$ , even in the  $B \rightarrow \infty$  limit characterized by Lemma 3.3. On the other hand, the denoising/compressed sensing problems become easier as  $B \rightarrow \infty$ . Can we hope to achieve  $M_B(\varepsilon|\text{Oracle})$ ?

In order to approach oracle MSE for large  $B$ , we propose to use the positive-part James-Stein shrinkage estimator [JS10]. This is again a block-separable denoiser that acts as follows on a block  $y \in \mathbf{R}^B$ :

$$\eta^{JS}(y) = \left(1 - \frac{(B-2)}{\|y\|_2^2}\right)_+ y.$$

Analogously to block soft thresholding, this estimator shrinks to 0 blocks with small norms. On the other hand, its bias vanishes as  $\|y\|_2 \rightarrow \infty$ . Using once more Lemma 3.1 we have (notice that in this case there is no tuning parameter)

$$M_B(\varepsilon|\text{JamesStein}) \equiv \frac{1}{B} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \mathbb{E}_\nu \{[\mathbf{X} - \eta^{JS}(\mathbf{X} + \mathbf{Z})]^2\}.$$

Remarkably, the limiting  $B \rightarrow \infty$  behavior of this denoiser is ideal, and noticeably better than block soft thresholding, as shown by Figure 3.2 and formally in the next lemma.

**Lemma 3.4.** *Let  $M_B(\varepsilon|\text{JamesStein})$  denote the minimax MSE for  $\eta^{JS}$  over the class of  $\varepsilon$ -block*



sparse sequences. For any  $B > 2$ , we have:

$$M_B(\varepsilon|\text{JamesStein}) \leq \varepsilon + \frac{2}{B}.$$

*Proof.* Consider temporarily the case where the observation is  $\mathbf{Y} = \mu + \mathbf{Z}$  with  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{B \times B})$ , and  $\mu \in \mathbf{R}^B$  nonrandom and known. A simple calculation shows that the optimal linear estimator of the form  $\eta(y) = cy$ ,  $c \in \mathbf{R}$ , is given by

$$\eta^{IDL}(y) = c(\mu) \cdot y, \quad c(\mu) \equiv \frac{\|\mu\|_2^2}{\|\mu\|_2^2 + B}.$$

This estimator uses information about  $\|\mu\|_2$  (which could only be supplied by an oracle) to choose the constant  $c$  as a function of  $\|\mu\|_2$ . Note in particular that the risk of this estimator is

$$R(\mu; \eta^{IDL}) \equiv \mathbb{E}\{\|\eta^{IDL}(\mu + Z)\|_2^2\} = \frac{B\|\mu\|_2^2}{\|\mu\|_2^2 + B},$$

and, in particular,

$$0 = R(\mu = 0, \eta^{IDL}) \leq \sup_{\mu \in \mathbf{R}^B} R(\eta^{IDL}, \mu) = B. \quad (3.6)$$

Define the ideal worst-case MSE by

$$M_B(\varepsilon|\eta^{IDL}) \equiv \frac{1}{B} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \mathbb{E}_\nu\{R(\mu = X, \eta^{IDL})\}.$$

Applying (3.6), and keeping in mind that, for  $\nu \in \mathcal{F}_{\varepsilon, B}$ ,  $\nu(\{X = 0\}) \geq (1 - \varepsilon)$ , we have

$$M_B(\varepsilon|\eta^{IDL}) = \varepsilon. \quad (3.7)$$

The oracle inequality [DJ95, Theorem 5] shows that for  $B > 2$ , and for every vector  $\mu \in \mathbf{R}^B$ , if  $Y \sim \mathbf{N}(\mu, \mathbf{I}_{B \times B})$ , then

$$R(\mu; \eta^{JS}) \leq R(\mu; \eta^{IDL}) + 2.$$

Combined with the previous display this proves the Lemma.  $\square$

The argument in the proof leads in fact to a convenient expression for  $M_B(\varepsilon|\text{JamesStein})$ . With the notations introduced there, we have

$$M_B(\varepsilon|\text{JamesStein}) = \frac{1 - \varepsilon}{B} R(0; \eta^{JS}) + \frac{\varepsilon}{B} \sup_{\mu \in \mathbf{R}^B} R(\mu; \eta^{JS}).$$

Now  $\eta^{JS}$  is known to be minimax for the unconstrained problem of estimating a non-sparse vector  $\mu$ , i.e.  $\sup_{\mu \in \mathbf{R}^B} R(\mu; \eta^{JS}) = B$  yielding

$$M_B(\varepsilon|\text{JamesStein}) = \frac{1 - \varepsilon}{B} R(0; \eta^{JS}) + \varepsilon.$$

Therefore computing the minimax MSE for  $\eta^{JS}$  reduces to computing the single quantity  $R(0; \eta^{JS})$ , that can be estimated through numerical integration. A good approximation for large  $B$  is provided

by the following formula  $R(0; \eta^{JS}) = B^{-1} + \kappa B^{-3/2} + O(B^{-2})$  with  $\kappa \approx 0.752$  (cf. Appendix I.2.1). Hence we have

$$M_B(\varepsilon|\text{JamesStein}) = (1 - \varepsilon) \left\{ \frac{1}{B} + \frac{\kappa}{B^{3/2}} + O(B^{-2}) \right\} + \varepsilon. \quad (3.8)$$

In the next section will use this formula (neglecting  $O(B^{-2})$  terms) in comparing the general prediction of Eq. (1.13) with the empirical results for the James-Stein AMP algorithm. Numerical integration reveals that this formula is accurate enough for such comparison.

### 3.3 Empirical phase transition behavior

We now turn to the compressed sensing reconstruction problem whereby the block-sparse vector  $x_0$  is reconstructed from observed data  $y = Ax_0$  using the AMP algorithm. We want to test the hypothesis that Eq. (1.13) describes the phase transition of the two block shrinkage AMP algorithms, corresponding to the block soft thresholding, and block James-Stein.

We conducted a set of experiments similar to those described in Section 2.4 We constructed block-sparse signals at different undersampling and sparsity levels and ran tests of block thresholding AMP. More precisely, we used the update equations (1.7) to (1.9) with  $\eta = \eta^{soft}$  (block soft thresholding AMP) or  $\eta = \eta^{JS}$  (James-Stein AMP).

It is a straightforward calculus exercise to compute an explicit expression for the memory term  $\mathbf{b}_t$ . For block soft thresholding AMP we get

$$\mathbf{b}_t \equiv \frac{1}{n} \operatorname{div} \eta^{soft}(y; \tau, \sigma_{t-1}) \Big|_{y=y^{t-1}} = \frac{1}{n} \sum_{\ell=1}^M \left( B - \frac{(B-1)\tau\sigma_{t-1}}{\|block_{\ell}(y^{t-1})\|_2} \right) \mathbb{I}_{\{\|block_{\ell}(y^{t-1})\|_2 > \tau\sigma_{t-1}\}}. \quad (3.9)$$

For James-Stein AMP we have

$$\mathbf{b} \equiv \frac{1}{n} \operatorname{div} \eta^{JS}(y; \sigma_{t-1}) \Big|_{y=y^{t-1}} = \frac{1}{n} \sum_{\ell=1}^M \left( B - \frac{(B-2)^2}{\|block_{\ell}(y^{t-1})\|_2^2} \right) \mathbb{I}_{\{\|block_{\ell}(y^{t-1})\|_2^2 > B-2\}}. \quad (3.10)$$

Our results show that the curve  $\delta = M_B(\varepsilon|\text{BlockSoft})$  correctly separates two phases of performance: below this curve success in AMP recovery is atypical and above it is typical. Similarly, the curve  $\delta = M_B(\varepsilon|\text{JamesStein})$  correctly describes the phase transition for block James-Stein shrinkage. The empirical results are presented in Figure 8 (for block soft thresholding) and Figure 9 (for block James-Stein). We refer to Appendix G for further details.

## 4 Monotone regression

In this section and the next, we show that, quite surprisingly, the formula (1.13) can be applied also to some highly nontrivial non-separable denoisers.

In this section we consider vectors that are monotone, and mostly constant. Let  $\mathcal{M}$  denote the cone of nondecreasing sequences:

$$\mathcal{M}_N \equiv \{x \in \mathbf{R}^N \quad : \quad x_{t+1} \geq x_t \quad \text{for all } t \in \{1, 2, \dots, N-1\}\}.$$

We then define the class of mostly constant non-decreasing vectors

$$\mathcal{F}_{N,\varepsilon,\text{mono}} \equiv \left\{ \nu_N \in \mathcal{P}(\mathbf{R}^N) \quad : \quad \operatorname{supp}(\nu_N) \subseteq \mathcal{M}_N, \quad \mathbb{E}_{\nu_N} [\#\{t \in [N-1] : x_{t+1} > x_t\}] \leq N\varepsilon \right\}.$$

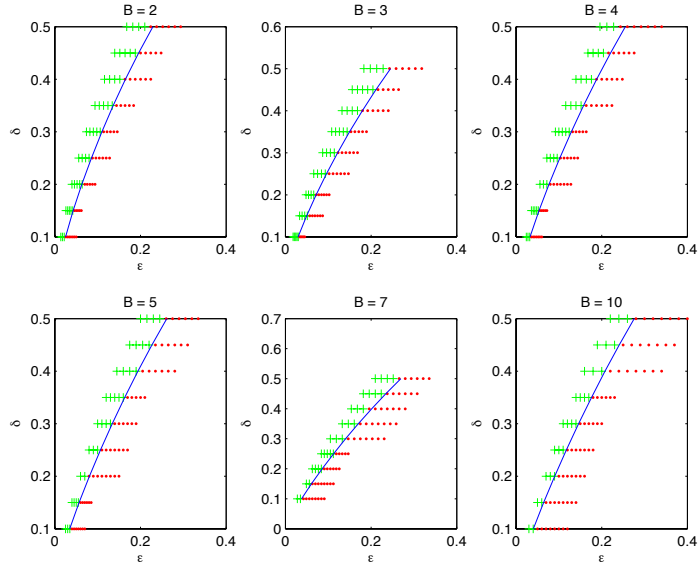


Figure 8: Phase transition results for block soft thresholding AMP with minimax threshold. Here the signal dimension is  $N = 1000$ ,  $\delta = n/N$  is the undersampling fraction, and  $\varepsilon$  is the sparsity parameter (fraction of non-zero entries). Red: less than 50% fraction of correct recovery. Green: greater than 50% fraction of successful recovery. Blue Curve:  $\delta = M_B(\varepsilon|\text{BlockSoft})$ .

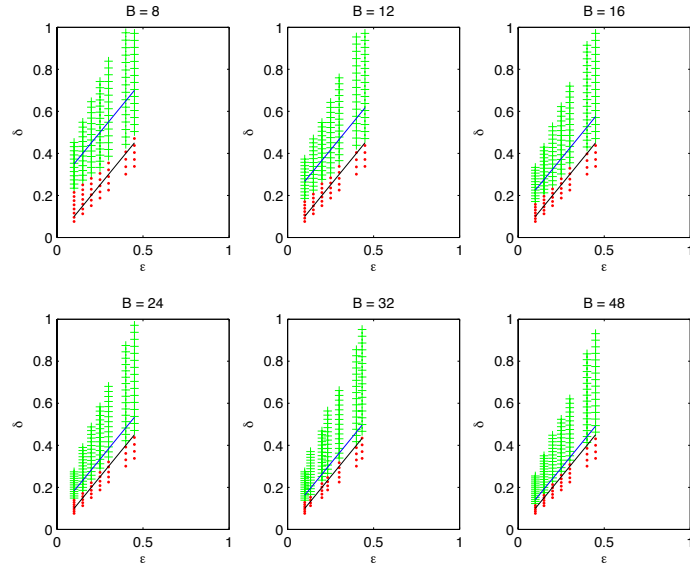


Figure 9: Phase transition results for block James-Stein AMP. Here the signal dimension is  $N = 1000$ ,  $\delta = n/N$  is the undersampling fraction, and  $\varepsilon$  is the sparsity parameter (fraction of non-zero entries). Red: less than 50% fraction of correct recovery. Green : greater than 50% fraction of successful recovery. Blue Curve: asymptotic (large- $B$ ) formula for  $\delta = M_B(\varepsilon|\text{JamesStein})$  (3.8). Black Curve:  $\delta = \varepsilon$  (lower bound on minimax risk of James-Stein Shrinkage).

Since vectors from this class are –in general– not sparse, we will occasionally refer to the parameter  $\varepsilon$  as to the ‘simplicity’ parameter.

For this problem we will consider the denoiser  $\eta^{mono} : \mathbf{R}^N \rightarrow \mathbf{R}^N$ , that solves the *monotone regression* problem

$$\eta^{mono}(y) = \operatorname{argmin}_{x \in \mathcal{M}_N} \|y - x\|_2^2. \quad (4.1)$$

In other words,  $\eta^{mono}$  is the (Euclidean) projection on the cone of monotone sequences. This denoiser is highly non-separable, as one can understand most clearly by studying the standard pool-adjacent-violators algorithm for implementing it (see [BC90] for a recent reference).

#### 4.1 Minimax MSE

In order to apply formula (1.13), we need to calculate  $M(\varepsilon|\text{MonoReg})$ , which requires in particular determining the least favorable distribution  $\nu_N \in \mathcal{F}_{\varepsilon, N, \text{mono}}$  and proving that the limit  $N \rightarrow \infty$  of the minimax MSE exists. We present here the main ideas, deferring details to Appendix F.

It is convenient to introduce the risk at  $\mu \in \mathcal{M}_N$ :

$$R_N(\mu) \equiv \mathbb{E}\{\|\eta^{mono}(\mu + \mathbf{Z}) - \mu\|_2^2\}, \quad (4.2)$$

where expectation is taken with respect to  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{N \times N})$ . It is further useful to introduce a specific notation for the risk at 0, namely

$$r(N) \equiv R_N(\mu = 0) = \mathbb{E}\{\|\eta^{mono}(\mathbf{Z})\|_2^2\}. \quad (4.3)$$

**Lemma 4.1.** *The risk of monotone regression satisfies the following properties*

- (a) *The function  $t \mapsto R_N(t\mu)$  is monotone increasing for  $t \in \mathbf{R}_+$ .*
- (b) *Let  $I_+(\mu) \equiv \{i \in [N-1] : \mu_i < \mu_{i+1}\}$  be the set of increase points of  $\mu$ . Denoting them by  $I_+(\mu) \equiv \{i_1, i_2, \dots, i_{K(\mu)}\}$ ,  $i_k \leq i_{k+1}$ , let  $J_k \equiv \{i_k + 1, i_k + 2, \dots, i_{k+1}\}$  for  $k \in \{0, \dots, K(\mu)\}$  (with, by convention,  $i_0 = 0$ ,  $i_{K(\mu)+1} = N$ ). Then, for any  $\mu \in \mathcal{M}_N$ ,*

$$\lim_{t \rightarrow \infty} R_N(t\mu) = \sum_{k=0}^{K(\mu)} r(|J_k(\mu)|). \quad (4.4)$$

*Proof of part (a).* For a non-empty closed convex  $\mathcal{S} \subseteq \mathbf{R}^N$ , we let  $\mathbf{P}_{\mathcal{S}} : \mathbf{R}^N \rightarrow \mathbf{R}^N$  denote the Euclidean projector to  $\mathcal{S}$ , i.e.  $\mathbf{P}_{\mathcal{S}}(y) \equiv \operatorname{argmin}_{x \in \mathcal{S}} \|x - y\|_2$ . Further, for  $v \in \mathbf{R}^N$ ,  $\mathcal{S} + v \equiv \{x + v : x \in \mathcal{S}\}$ .

Note that it is sufficient to show that, letting  $D(\mu; z) \equiv \|\eta^{mono}(\mu + z) - \mu\|_2^2$ ,  $t \mapsto D(t\mu; z)$  is monotone increasing in  $t \in \mathbf{R}_+$ . By continuity of the projection operator, it follows that  $\mu \mapsto D(\mu; z)$  is continuous. Further, notice that  $\mathcal{M}_N$  is a cone obtained as the intersection of  $N - 1$  half-spaces:

$$\mathcal{M}_N = \bigcap_{i=1}^{N-1} \mathcal{H}_i, \quad \mathcal{H}_i \equiv \{x \in \mathbf{R}^N : x_i \leq x_{i+1}\}.$$

Let  $\mathcal{V}_i = \{x \in \mathbf{R}^N : x_i = x_{i+1}\}$  be the separating hyperplane for  $\mathcal{H}_i$  and, for  $B \subseteq [N - 1]$ , define

$$\mathcal{V}_B \equiv \bigcap_{i \in B} \mathcal{V}_i = \{x \in \mathbf{R}^N : x_i = x_{i+1} \forall i \in B\},$$

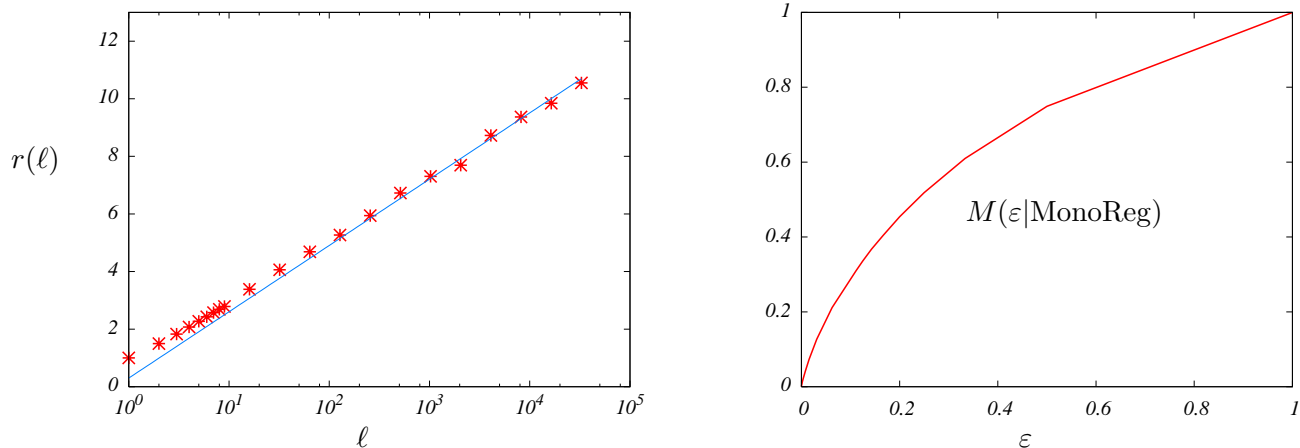


Figure 10: Left: risk at 0 for monotone regression. Here we plot the mean square error  $r(\ell) \equiv R_\ell(\mu = 0)$  as a function of the signal dimension  $\ell$ . Symbols were obtained through Monte Carlo simulations for  $\ell \in \{1, 2, 3, \dots, 10, 16, 32, \dots, 32, 768 = 2^{15}\}$ . The continuous line is the fit  $r(\ell) = \log(\ell) + 0.3$ . Right: Minimax mean square error of monotone regression, as per Theorem 4.1.

with, by convention  $\mathcal{V}_\emptyset \equiv \mathbf{R}^N$ . Since  $\eta^{\text{mono}} = \mathcal{P}_{\mathcal{M}_N}$ , we have that  $\eta^{\text{mono}}$  is continuous and piecewise linear and equal one of the projectors  $\mathcal{P}_{\mathcal{V}_B}$ , that we will denote by  $\mathcal{P}_B$  for  $B \subseteq [N-1]$ . It is therefore sufficient to show that, defining for  $B \subseteq [N-1]$ ,

$$D_B(\mu; z) \equiv \|\mathcal{P}_B(\mu + z) - \mu\|_2^2,$$

the function  $t \mapsto D_B(t\mu, z)$  is monotone increasing for  $t \in \mathbf{R}_+$ .

Let  $B \equiv \cup_{k=1}^K B_k$  where each  $B_k$  is a contiguous segment (in the sense of point (b)), and  $\overline{B}_k \equiv \{i \in [N] : i \in B_k \vee (i-1) \in B_k\}$ . Further, for  $x \in \mathbf{R}^N$ , let  $\overline{x}_S \equiv |S|^{-1} \sum_{i \in S} x_i$ . Then, for any  $x \in \mathbf{R}^N$ ,

$$\mathcal{P}_B(x)_i = \begin{cases} \overline{x}_{\overline{B}_k} & \text{if } i \in B_k, k \in \{1, \dots, K\}, \\ x_i & \text{otherwise.} \end{cases}$$

Hence

$$D_B(t\mu; z) = \sum_{k=1}^K |\overline{B}_k| \left[ \frac{t^2}{|\overline{B}_k|} \sum_{i \in \overline{B}_k} (\mu_i - \overline{\mu}_{\overline{B}_k})^2 + \overline{z}_{\overline{B}_k}^2 \right] + \sum_{i \in [N] \setminus \cup_{k=1}^K \overline{B}_k} z_i^2,$$

which is clearly increasing in  $t \in \mathbf{R}_+$ . □

The proof of part (b) is deferred to Appendix F.

The last Lemma shows that the least favorable signal  $\mu$  is constant on  $N(1-\varepsilon)$  positions of the interval  $\{1, 2, \dots, N\}$  and has large (going to infinity) jumps at the remaining  $N\varepsilon$  increase points. The resulting risk only depends on the distribution of the lengths of the intervals over which  $\mu$  is constant.

The next Lemma provides some useful insight on the behavior of the risk at 0. This is crucial since it determines the minimax risk though Eq. (4.4).

**Lemma 4.2.** *The monotone regression risk at zero, defined through Eq. (4.3) satisfies  $r(1) = 1$  and, for any  $N \geq 10$*

$$r(N) \leq 20 (\log N)^2. \quad (4.5)$$

Further  $r(N) < 2N[\log N + 1]$  for all  $N$ .

The proof of this Lemma can be found in Appendix F.2.

For moderate values of  $N$ ,  $r(N)$  can be computed numerically through Monte Carlo simulations. Figure 10 presents the results of such a simulation. It appears that  $r(N) = \Theta(\log N)$  as  $N \rightarrow \infty$  suggesting that the last lemma is loose by a logarithmic factor.

We can finally establish our main result on the minimax MSE of monotone regression over the class  $\mathcal{F}_{N,\varepsilon,\text{mono}}$ . Remarkably, we are able to characterize the least favorable distribution  $\nu_N \in \mathcal{F}_{N,\varepsilon,\text{mono}}$ .

**Theorem 4.1.** *The asymptotic minimax MSE of monotone regression*

$$M(\varepsilon|\text{MonoReg}) = \lim_{N \rightarrow \infty} M(\mathcal{F}_{N,\varepsilon,\text{mono}}|\eta^{\text{mono}})$$

exists and is given by

$$M(\varepsilon|\text{MonoReg}) = \max \left\{ \varepsilon \sum_{\ell \geq 1} \pi_\ell r(\ell) : \sum_{\ell \geq 1} \ell \pi_\ell = 1/\varepsilon \right\}, \quad (4.6)$$

where the maximization is over the probability distribution  $\{\pi_\ell\}_{\ell \in \mathbf{N}}$ . Equivalently, the curve  $(1/\varepsilon, M(\varepsilon|\text{MonoReg})/\varepsilon)$  is the least concave envelope of the point set  $\{(\ell, r(\ell))\}_{\ell \in \mathbf{N}}$ . Further, there exists  $\varepsilon_0 > 0$  such that, for all  $\varepsilon \in [0, \varepsilon_0]$ ,

$$M(\varepsilon|\text{MonoReg}) \leq 20 \varepsilon (\log 1/\varepsilon)^2.$$

Finally, for any  $\xi > 0$ ,  $\varepsilon \in (0, 1)$ , the following distribution  $\nu_N^{(\varepsilon, \xi)} \in \mathcal{F}_{N,\varepsilon,\text{mono}}$  has risk larger than  $M(\varepsilon|\text{MonoReg}) - \xi$  for all  $N$  large enough. A signal  $\mathbf{X} \sim \nu_N^{(\varepsilon, \xi)}$  has  $X_{i+1} - X_i = \Delta > 0$  at all increase points  $i \in [N - 1]$  for some  $\Delta = \Delta(\xi)$  large enough, and the lengths of intervals between increase points have distribution  $\pi$  achieving the max in Eq. (4.6).

*Proof.* With a slight abuse of notation we define, for  $\nu_N \in \mathcal{F}_{N,\varepsilon,\text{mono}}$ , the expected risk of monotone regression as  $R_N(\nu_N) = \mathbb{E}_{\nu_N} \{ \|\eta^{\text{mono}}(\mathbf{X} + \mathbf{Z}) - \mathbf{X}\|_2^2 \}$  where  $\mathbf{X} \sim \nu_N$ . Further, for  $t \in \mathbf{R}_+$ , let  $S_t \nu_N$  be the distribution obtained by rescaling  $\nu_N$ :  $S_t \nu_N((a, b]) = \nu_N((a/t, b/t])$ . Further let  $\{\pi_\ell^{\mathbf{X}}\}_\ell$  be the empirical distribution of the lengths of the constant intervals  $J_k$  (as per Lemma 4.1.(b)), and  $K(\mathbf{X})$  be their number. We then have, by Lemma 4.1,

$$R_N(\nu_N) \leq \lim_{t \rightarrow \infty} R_N(S_t \nu_N) = \mathbb{E}_{\nu_N} \left\{ K(\mathbf{X}) \sum_{\ell=1}^{\infty} \pi_\ell^{\mathbf{X}} r(\ell) \right\}.$$

Further, by definition

$$\mathbb{E} \left\{ \frac{K(\mathbf{X})}{N} \right\} = \mathbb{E} \left\{ \frac{1}{\sum_{\ell=1}^{\infty} \pi_\ell^{\mathbf{X}} \ell} \right\} = \varepsilon.$$

Hence  $M(\varepsilon|\text{MonoReg})$  is immediately upper bounded by the right hand side of Eq. (4.6). The matching lower bound is obtained by evaluating the above expressions for the distribution  $\nu_N^{(\varepsilon, \xi)}$ .

Finally Eq. (4.7) follows by using Lemma 4.2 in Eq. (4.6).  $\square$

The resulting curve  $M(\varepsilon|\text{MonoReg})$  is presented in Fig. 10.

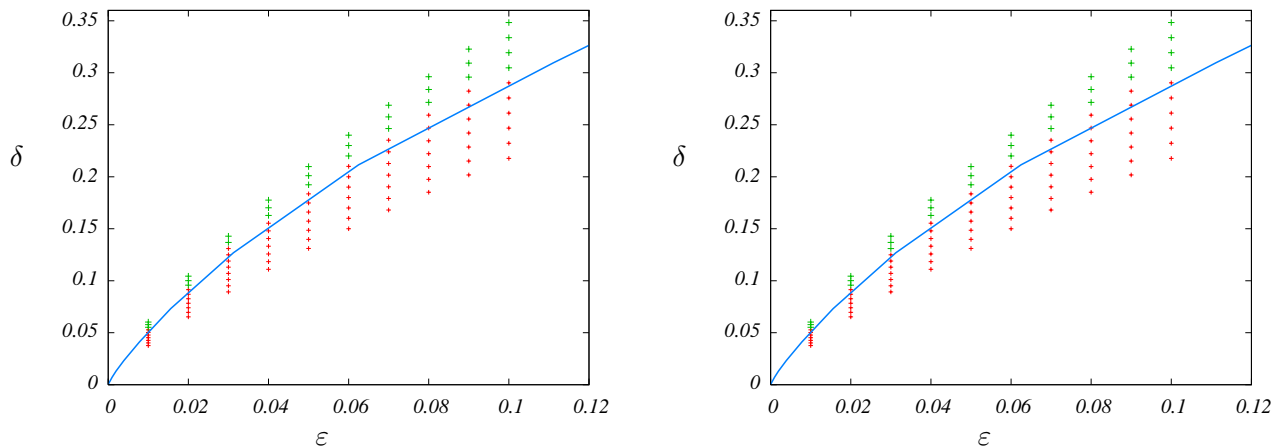


Figure 11: Results for monotone regression AMP. Here  $\delta = n/N$  undersampling fraction and  $\varepsilon$  is the sparsity measure (i.e. the fraction of increase points in the signal to be reconstructed). Signals were generated according to the least favorable distribution in Theorem 4.1. Small red crosses: less than 50% fraction of correct recovery. Large green crosses: greater than 50% fraction of correct recovery. Curve: minimax MSE  $\delta = M(\varepsilon|\text{MonoReg})$ . Left frame:  $N = 1000$ . Right frame:  $N = 2000$ .

## 4.2 Empirical phase transition behavior

We next consider the compressed sensing problem. We programmed the AMP iteration (1.7)-(1.9), with  $\eta(\cdot; \tau, \sigma) = \eta^{\text{mono}}(\cdot)$  the monotone regression denoiser. The denoiser itself was implemented using the standard pool adjacent violators algorithm.

It is a simple exercise to obtain an explicit formula for the memory term  $\mathbf{b}_t$ . As in Lemma 4.1, let  $K(\mu)$  denote the number of increase points in the signal  $\mu \in \mathbf{R}^N$ . We then have

$$\mathbf{b}_t \equiv \frac{1}{n} \operatorname{div} \eta^{\text{mono}}(y) \Big|_{y=y^{t-1}} = \frac{1}{n} \left\{ K(\eta^{\text{mono}}(y^{t-1})) + 1 \right\}. \quad (4.7)$$

We will refer to this specific version of AMP as to *monoreg AMP*.

Our numerical simulations are summarized in Figure 11. Each data point corresponds to the empirical success probability over 100 independent reconstruction experiments, using approximately least favorable signals. More precisely, we used piecewise constant signals, increasing, with equal jump sizes  $\mu$  and constant intervals distributed according to the minimax law  $\{\pi_\ell\}$ . The signals start with  $x_1 = \mu$ , and we took  $\mu = 10$  (the results were statistically independent of  $\mu \gtrsim 5$ ).

In the present case we evaluated success probability using the following (Hamming-like) distance

$$H_\alpha(x^t, x_0) \equiv \frac{1}{n} \left| \left\{ i \in [N] : |x_i^t - x_{0,i}| \geq \alpha \right\} \right|, \quad (4.8)$$

and declared a success when  $H(x^t, x_0) \leq \beta$ . In Figure 11 we used  $t = 300$  and  $\alpha = \beta = 0.01$ , but very similar results are obtained with other values of the parameters. The rationale for using  $H_\alpha(x^t, x_0)$  instead of the normalized mean square error lies in the structure of the signals  $x_0$ . Since the least favorable  $x_0$  is monotone with large jumps, its norm is very large, concentrated at

endpoints, and depends strongly on  $N$ . This leads to subtle normalization issues across different  $N$ .

The agreement between the empirical phase transition and the general prediction  $\delta = M(\varepsilon|\text{MonoReg})$  in Fig. 11 is satisfactory and improves with the signal's length.

## 5 Total variation minimization

In this section we consider vectors  $x \in \mathbf{R}^N$  that are mostly constant, with a few change points. In order to model this problem, we introduce the class of probability distributions

$$\mathcal{F}_{N,\varepsilon,TV} \equiv \left\{ \nu_N \in \mathcal{P}(\mathbf{R}^N) : \mathbb{E}_{\nu_N} [\#\{t \in [N-1] : x_{t+1} \neq x_t\}] \leq N\varepsilon \right\}.$$

Again  $\varepsilon \in (0, 1)$  is a ‘simplicity’ parameter. Note that this class is quite similar to the class  $\mathcal{F}_{N,\varepsilon,\text{mono}}$  studied in the previous section, the ‘only’ difference being that change points can be either points of increase or points of decrease.

A convenient denoiser for this setting is the *total variation penalized least-squares* [ROF92], also called fused LASSO [TSR<sup>+</sup>05], that we will denote by  $\eta^{tv}(\cdot; \tau) : \mathbf{R}^N \rightarrow \mathbf{R}^N$ . This depends on  $\tau \in \mathbf{R}_+$  and, for  $y \in \mathbf{R}^N$  and noise variance  $\sigma^2 =$ , it returns

$$\eta^{tv}(y; \tau) \equiv \operatorname{argmin}_{x \in \mathbf{R}^N} \left\{ \frac{1}{2} \|y - x\|_2^2 + \tau \|x\|_{TV} \right\}, \quad (5.1)$$

$$\|x\|_{TV} \equiv \sum_{i=1}^{N-1} |x_{i+1} - x_i|. \quad (5.2)$$

An extensive literature is devoted to solving this denoising problem, see for example [VO96]. For general noise variance  $\sigma^2$ , the above expression is generalized through the usual scaling relationship (1.3).

Much of the analysis in this section is analogous to the one of monotone regression. We will therefore present several arguments in synthetic form to limit redundancy.

### 5.1 Minimax MSE

In this section we outline the computation of the asymptotic minimax MSE of the total variation denoiser over the class  $\mathcal{F}_{N,\varepsilon,TV}$ , to be denoted by  $M(\varepsilon|TV)$ .

We start by defining a generalization of the problem (5.1). For  $s = (s_1, s_2) \in \{+1, -1\}^2$ , we let

$$\eta_s^{tv}(y; \tau) \equiv \operatorname{argmin}_{x \in \mathbf{R}^N} \left\{ \frac{1}{2} \|y - x\|_2^2 + \tau \|x\|_{TV} + \tau (s_1 x_1 + s_2 x_N) \right\}.$$

For economy of notation, we will write  $s = ++, +0, +-, \dots$  instead of, respectively,  $s = (+1, +1), (+1, 0), (+1, -1), \dots$ . We further omit the subscript for the standard case  $s = 00$ .

We define the risk at  $\mu \in \mathbf{R}^N$  as

$$R_{N,s}(\mu; \tau) \equiv \mathbb{E} \left\{ \left\| \eta_s^{tv}(\mu + \mathbf{Z}; \tau) - \mu \right\|_2^2 \right\}, \quad (5.3)$$

where  $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_{N \times N})$ . We denote the risk at 0 by

$$r_s(N; \tau) \equiv R_{N,s}(\mu = 0; \tau) = \mathbb{E} \left\{ \left\| \eta_s^{tv}(\mathbf{Z}; \tau) \right\|_2^2 \right\}. \quad (5.4)$$



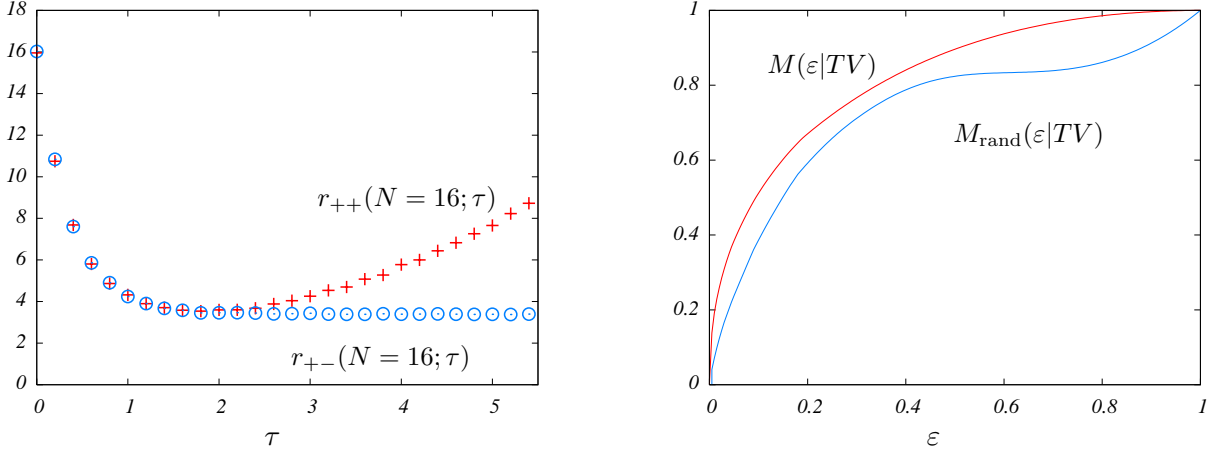


Figure 12: Left: Risk at zero of total variation denoising, as a function of the regularization parameter  $\tau \in \mathbf{R}_+$ . The risk was estimated through Mont Carlo integration. Right: Upper curve: minimax MSE of total variation denoiser over the class  $\mathcal{F}_{M,\varepsilon,TV}$  as per Theorem 5.1. Lower curve: minimax MSE over signals with random change points.

Notice that, by symmetry,  $r_s(N; \tau) = r_{-s}(N; \tau)$  and  $r_{s_1, s_2}(N; \tau) = r_{s_2, s_1}(N; \tau)$ . We then have the following analogous of Lemma 4.1.

**Lemma 5.1.** *The risk of total variation regression satisfies the following properties*

- (a) *The function  $t \mapsto R_N(t\mu; \tau)$  is monotone increasing for  $t \in \mathbf{R}_+$ .*
- (b) *Let  $I_{\neq}(\mu) \equiv \{i \in [N-1] : \mu_i \neq \mu_{i+1}\}$  be the set of change points of  $\mu$ . Denoting them by  $I_{\neq}(\mu) \equiv \{i_1, i_2, \dots, i_{K(\mu)}\}$ ,  $i_k \leq i_{k+1}$ , let  $J_k \equiv \{i_k + 1, i_k + 2, \dots, i_{k+1}\}$  for  $k \in \{0, \dots, K\}$  (with, by convention,  $i_0 = 0$ ,  $i_{K+1} = N$ ). Further, for  $K \geq 1$ , let  $s(k) = [\text{sgn}(\mu_{i_k} - \mu_{i_k+1}), \text{sgn}(\mu_{i_{k+1}+1} - \mu_{i_{k+1}})]$  for  $k \in \{1, \dots, K-1\}$ ,  $s(0) = [0, \text{sgn}(\mu_{i_1+1} - \mu_{i_1})]$ ,  $s(K) = [0, \text{sgn}(\mu_{i_K} - \mu_{i_K+1})]$ . For  $K = 0$  we let  $s(0) = (0, 0)$ .*

*Then, for any  $\mu \in \mathbf{R}^N$ ,*

$$\lim_{t \rightarrow \infty} R_N(t\mu; \tau) = \sum_{k=0}^{K(\mu)} r_{s(k)}(|J_k(\mu)|; \tau). \quad (5.5)$$

*Proof.* The argument in part (b) is essentially the same as in part (b) of Lemma 4.1 and we will therefore omit it.

For proving part (a), we will prove that, letting  $D(\mu; z) \equiv \|\eta^{tv}(\mu + z; \tau) - \mu\|_2^2$ , the function  $t \mapsto D(t\mu; z)$  is increasing for  $t \in \mathbf{R}_+$ . First notice that the stationarity condition for the minimum in Eq. (5.1) reads

$$x_i - y_i = \tau v_i - \tau v_{i-1}, \quad (5.6)$$

$$v_i = \begin{cases} +1 & \text{if } x_{i+1} > x_i, \\ -1 & \text{if } x_{i+1} < x_i, \\ v_i \in [-1, +1] & \text{otherwise,} \end{cases} \quad (5.7)$$

with the convention that  $v_0 = v_N = 0$ . Let  $I_{\neq} = I_{\neq}(x)$  and  $J_k, s(k)$  be defined as in part (b) of the statement. Then, summing Eq. (5.6) over  $i \in J_k$ , we get

$$x_i = \frac{1}{|J_k|} \sum_{i \in J_k} y_i + \tau \bar{s}(k), \quad \text{for all } i \in J_k, \quad (5.8)$$

where  $\bar{s}(k) = s_1(k) + s_2(k)$ . Hence  $\eta^{tv}(\cdot; \tau)$  is piecewise affine with components indexed by  $\mathcal{J} = \{J_k, s(k)\}_{k \in [K]}$ . Within each component, we have  $\eta^{tv}(y; \tau) = F_{\mathcal{J}}(y)$  with  $F_{\mathcal{J}}$  defined as per Eq. (5.8).

Since  $y \mapsto \eta^{tv}(y; \tau)$  is continuous (and hence  $t \mapsto D(t\mu; z)$  is), it is sufficient to prov that, letting

$$D_{\mathcal{J}}(\mu; z) \equiv \|F_{\mathcal{J}}(\mu + z) - \mu\|_2^2,$$

the function  $t \mapsto D_{\mathcal{J}}(\mu; z)$  is monotone increasing for  $t \in \mathbf{R}_+$ . Using Eq. (5.8) we obtain

$$D_{\mathcal{J}}(\mu; z) = \sum_{k=0}^K |J_k| \left[ \frac{1}{|J_k|} \sum_{i \in J_k} (\mu_i - \bar{\mu}_{J_k})^2 + (\bar{z}_{J_k} + \bar{s}(k)\tau)^2 \right],$$

where  $\bar{x}_{J_k}$  denotes the average of vector  $x$  over  $J_k$ . It follows that  $t \mapsto D_{\mathcal{J}}(\mu; z)$  is increasing as claimed.  $\square$

The risk at 0,  $r_s(N; \tau)$ , can be computed numerically for moderate values of  $N$ . Notice that the cases  $r_{00}(N; \tau)$ ,  $r_{0\pm}(N; \tau)$ , and  $r_{\pm 0}(N; \tau)$  are only relevant for the boundary intervals  $J_0(\mu)$  and  $J_K(\mu)$  and turn out to be immaterial for the asymptotic minimax risk. Thanks to symmetries, the only relevant cases are  $r_{++}(N; \tau)$  and  $r_{+-}(N; \tau)$ . The results of a numerical computation for these quantities is shown in Figure 12. These calculations suggest  $r_{++}(N; \tau) \geq r_{+-}(N; \tau)$ , which is indeed consistent with intuition as boundary conditions  $++$  induce a larger bias. Also, it is easy to prove that  $r_{+-}(N; \tau) \rightarrow 0$  as  $\tau \rightarrow \infty$  (as  $\tau \rightarrow \infty$ ,  $\eta^{tv}(y; \tau)$  converges to a constant vector).

Using the last Lemma, and proceeding as in the proof of Theorem 4.1, it is immediate to obtain a characterization of the minimax MSE of the total variation denoiser. For technical reasons, we need to introduce the class  $\mathcal{F}_{N,\varepsilon,TV}(L)$  of vectors in  $\mathcal{F}_{N,\varepsilon,TV}$  with distance at most  $L$  between changepoints.

**Theorem 5.1.** *The asymptotic minimax MSE of total variation denoiser*

$$M_L(\varepsilon|TV) \equiv \lim_{N \rightarrow \infty} M(\mathcal{F}_{N,\varepsilon,TV}(L)|TV)$$

*exists and is given by*

$$M_L(\varepsilon|TV) = \inf_{\tau \in \mathbf{R}_+} \max_{\pi} \left\{ \varepsilon \sum_{\ell, s \in \{++,+-\}} \pi_{\ell,s} r_s(\ell; \tau) : \sum_{\ell, s \in \{++,+-\}} \ell \pi_{\ell,s} = 1/\varepsilon \right\}, \quad (5.9)$$

*where the maximization is over the probability distribution  $\{\pi_{\ell,s}\}_{1 \leq \ell \leq L, s \in \{++,+-\}}$ .*

*For any  $\xi > 0$ ,  $\varepsilon \in (0, 1)$ , the following distribution  $\nu_N^{(\varepsilon, \xi)} \in \mathcal{F}_{N,\varepsilon,TV}(L)$  has risk larger than  $M_L(\varepsilon|TV) - \xi$  for all  $N$  large enough. A signal  $\mathbf{X} \sim \nu_N^{(\varepsilon, \xi)}$  has  $|X_{i+1} - X_i| = \Delta > 0$  at all change points  $i \in [N - 1]$  for some  $\Delta = \Delta(\xi)$  large enough. Further, for an interval  $J$  between change points, let the type of  $J$  be  $(\ell, s)$  with  $\ell \in \mathbf{N}$  its length and  $s \in \{++,+-\}$  depending whether the adjacent change points are both increase or decrease points  $(+-)$  or not  $(++)$ . Then the empirical distribution of types under  $\nu_N^{(\varepsilon, \xi)}$  is given by  $\pi$  solving the saddle point problem in Eq. (4.6).*

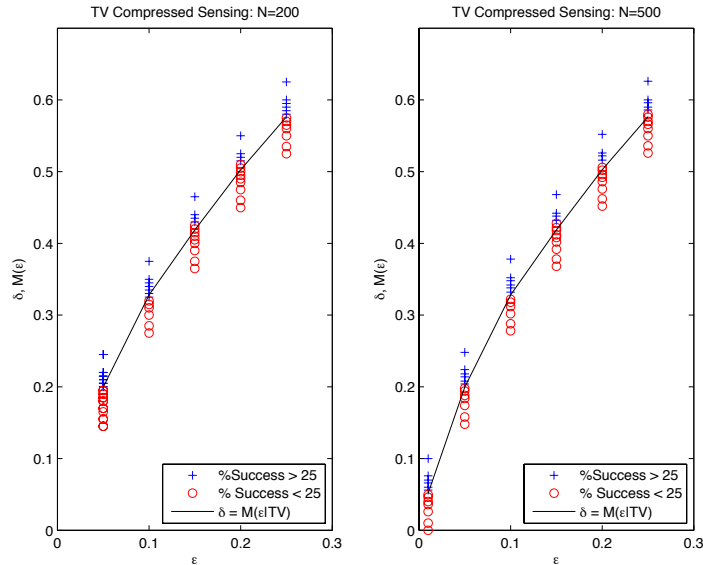


Figure 13: Results for TV AMP under the random changepoint signal prior. Here  $\delta = n/N$  is the undersampling fraction  $\varepsilon$  is the generalized sparsity measure (i.e. the fraction of change points in the signal to be reconstructed). Circles: less than 25% fraction of correct recovery. Crosses: greater than 25% fraction of correct recovery. Curve: minimax MSE curve for random changepoints  $\delta = M_{\text{random}}(\varepsilon|TV)$ . Left frame:  $N = 200$ . Right frame:  $N = 500$ .

We omit this proof since it is an immediate generalization of the one of Theorem 4.1. Notice that  $M_L(\varepsilon)$  is monotone increasing in  $L$  and hence admit a limit as  $L \rightarrow \infty$ . We expect that

$$M(\varepsilon|TV) = \lim_{L \rightarrow \infty} M_L(\varepsilon|TV).$$

This limit can be evaluated numerically and in Figure 12 we plot the resulting minimax risk.

Notice that, by properly modifying Eq. (5.9), one obtains the minimax risk over subsets of  $\mathcal{F}_{N,\varepsilon}$  with constrained change point distributions. For instance, we can consider the case in which the lengths between change points are distributed as for uniformly random change points, and increase/decrease points are alternating. We then get

$$\pi_{\ell,++}^\varepsilon = \varepsilon(1 - \varepsilon)^{\ell-1}, \quad \pi_{\ell,+ -}^\varepsilon = 0, \quad \ell \geq 1.$$

We consequently define the random changepoint minimax risk as

$$M_{\text{rand}}(\varepsilon|TV) = \inf_{\tau \in \mathbf{R}_+} \left\{ \varepsilon \sum_{\ell} \pi_{\ell,s}^\varepsilon r_s(\ell; \tau) \right\}. \quad (5.10)$$

This curve is plotted in Figure 12 for comparison.

## 5.2 Empirical phase transition behavior

We implemented the the AMP iteration (1.7)-(1.9), using the total variation denoiser  $\eta(\cdot; \tau, \sigma) = \eta^{tv}(\cdot; \tau, \sigma)$ . For the latter, we used the software package `tvdp` (in the Matlab implementation) or the projected Newton method [VO96] (in the Java implementation).

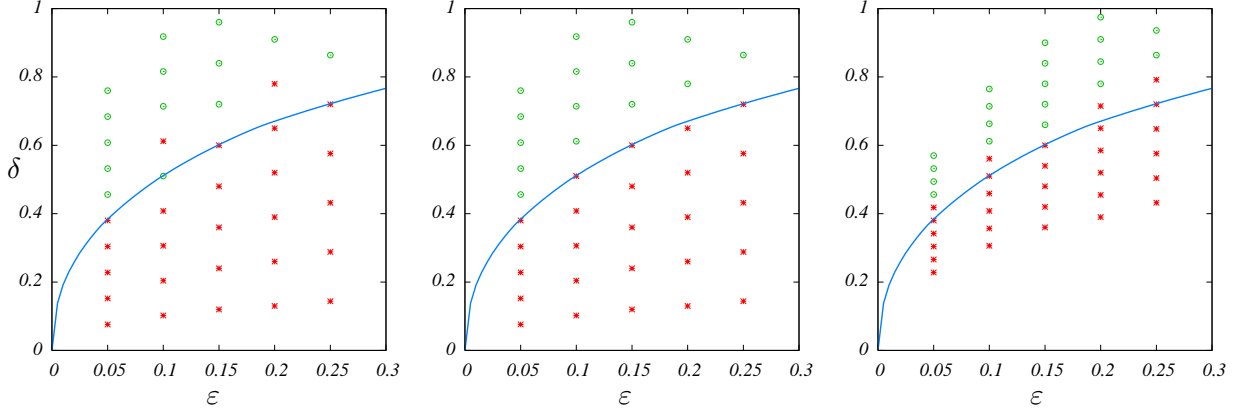


Figure 14: Results for TV AMP under the least favorable signal prior. Here  $\delta = n/N$  is the undersampling fraction  $\varepsilon$  is the generalized sparsity measure (i.e. the fraction of change points in the signal to be reconstructed). Red stars: less than 50% fraction of correct recovery. Green circles: greater than 50% fraction of correct recovery. Curve: minimax MSE curve  $\delta = M(\varepsilon|TV)$ . The three frames correspond to (from left to right)  $N = 100, 250, 500$ .

For  $x \in \mathbf{R}^N$  let  $K_0(x)$  denote the number of constant segments in  $x$  or, equivalently, the number of change points in  $x$ , plus one. We then have the following expression for the memory term in Eq. (1.7)-(1.9):

$$\mathbf{b}_t \equiv \frac{1}{n} \operatorname{div} \eta^{tv}(y; \tau, \sigma_{t-1}) \Big|_{y=y^{t-1}} = \frac{1}{n} K_0(\eta^{tv}(y^{t-1}; \tau, \sigma^{t-1})). \quad (5.11)$$

We will refer to this specific version of AMP as to *TV-AMP*.

We carried out two types of experiments. In the first class of experiments we considered signals  $x$  with distances between change points distributed as

$$\pi_{\ell,++}^\varepsilon = \pi_{\ell,+ -}^\varepsilon = \frac{1}{2} \varepsilon (1 - \varepsilon)^{\ell-1}.$$

This is the same distribution as if each position is independently an increase point or a decrease point, each with probability  $\varepsilon/2$ . The predicted phase transition curve is given by Eq. (5.10) and the minimax value of  $\tau$  is used in the AMP implementation. The simulation results are presented in Figure 13 for  $N = 200, 500$  and show good agreement between predictions and observations. In this case we used the Hamming metric (4.8), because the norm of the typical signal  $\|x_0\|_2^2$  scales super-linearly in  $N$ .

In the second set of experiments we used the distribution  $\{\pi_{\ell,s}\}_{s \in \{++,+-\}, \ell \in \mathbf{N}}$  achieving the sup in Eq. (5.9) for  $L$  large. More precisely, we fixed  $L = 25, 50, 100$  and solved numerically the optimization problem (5.9). The solution appears to be independent of  $L$  and put small weight on large lengths  $\ell$ . The least favorable distribution for  $L = 50$  was used to generate signals. The success probability is compared with the predicted phase transition curve  $M(\varepsilon|TV)$  in Figure 14 for  $N = 100, 250, 500$ . In this case we used the MSE metric and declared success if  $\|\hat{x}^5 - x_0\|_2^2 / \|x_0\|_2^2 \leq 0.001$ .

## 6 Characterization of the phase transition using state evolution

In this section we prove the basic relation (1.13) between minimax mean square error in denoising and the phase-transition boundary in the sparsity-undersampling plane. Our proof assumes that the state evolution formalism developed in [DMM09, DMM10, DMM11] holds, in the precise terms stated below. This formalism was established rigorously for separable denoisers (under additional regularity assumptions) in [BM11a].

A crucial observation for state evolution is that the mean squared error of the AMP reconstruction  $x^t$  at iteration  $t$  is practically non-random for large system sizes  $N$ , and has a well-defined limit as  $N \rightarrow \infty$ . In particular, the limit

$$m_t = \lim_{N \rightarrow \infty} \frac{1}{N} \|x_0 - x^t\|_2^2,$$

exists almost surely (here we assume  $n \rightarrow \infty$ , while  $n/N \rightarrow \delta$ ). Moreover, the evolution of  $m_t$  with increasing  $t$  is dictated by a formula  $m_{t+1} = \Psi(m_t)$  which is explicitly computable, and defined below. We will use the term *state evolution* to refer both to the mapping  $m \mapsto \Psi(m)$  and to the sequence  $\{m_t\}_{t \geq 0}$  with appropriate initial condition. State evolution allows to determine whether AMP recovers the signal  $x_0$  correctly, by simply checking whether  $m_t \rightarrow 0$  as  $t \rightarrow \infty$  (in which case the MSE vanishes asymptotically) or not. The latter problem does in turn reduce to a problem in real analysis.

The papers [DMM09, DMM10, DMM11] developed the state evolution framework for separable denoisers and verified its predictions numerically for three specific examples (namely for the shrinkers Soft, SoftPos, and Cap). However, the heuristic argument presented in those papers was much more general. Indeed, [BM11a] proved that state evolution holds, in a precise asymptotic sense, for Gaussian measurement matrices  $A$  with iid entries and generic separable denoisers, under mild regularity assumptions. A generalization to non-gaussian entries was subsequently proved in [BLM12].

Here we generalize this approach to non-separable denoisers  $\eta(\cdot; \tau, \sigma) : \mathbf{R}^N \mapsto \mathbf{R}^N$ . This framework covers all the shrinkers discussed in Sections 2 to 5, and yields a formal proof of the main formula (1.13), under the assumption that indeed state evolution is correct in this broader context. We will throughout assume the scaling relation (1.3).

In the next sections we will first introduce some basic notations and facts about state evolution. Then we will prove the phase transition expression (1.13) by establishing first a lower bound and then a matching upper bound, both given by the minimax MSE.

### 6.1 State Evolution

The next definition provides the suitable generalization of the state evolution mapping to the present setting.

**Definition 6.1.** For given  $\delta, \tau \geq 0$ , and  $\nu = \{\nu_N\}_{N \in \mathbf{N}}$  a sequence of probability distributions over  $\mathbf{R}^N$ , define the state evolution mapping  $\Psi(\cdot; \delta, \tau, \nu) : \mathbf{R} \mapsto \mathbf{R}$  by

$$\Psi(m; \delta, \tau, \nu) \equiv \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\nu_N} \left\{ \left\| \mathbf{X} - \eta \left( \mathbf{X} + \sqrt{\frac{m}{\delta}} \mathbf{Z}; \tau, \sqrt{\frac{m}{\delta}} \right) \right\|_2^2 \right\}, \quad (6.1)$$

whenever the limit on the right-hand side exists. Here, as before,  $\mathbf{X}$  and  $\mathbf{Z}$  are independent vectors,  $\mathbf{X} \sim \nu_N$ ,  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{N \times N})$ . In other words,  $\Psi(m; \delta, \tau, \nu)$  is the per-coordinate MSE of denoiser  $\eta$  at noise level  $\sigma^2 \equiv m/\delta$ .

Given implicit fixed parameters  $(\delta, \tau, \nu)$ , state evolution is the one-dimensional dynamical system:  $m_{t+1} = \Psi(m_t)$  starting from  $m_0 = \lim_{N \rightarrow \infty} n^{-1} \mathbb{E}_{\nu_N} \|\mathbf{X}\|_2^2$ .

The fixed points of the mapping  $m \mapsto \Psi(m)$  play of course a crucial role in the analysis of state evolution.

**Definition 6.2.** The highest fixed point of the mapping  $\Psi(\cdot) = \Psi(\cdot; \delta, \tau, \nu)$  is defined as  $\text{HFP}(\Psi) \equiv \sup\{m : \Psi(m) \geq m\}$ .

The importance and applicability this notion is underscored by the next two observations. Here and below we say that a function  $f : \mathbf{R}_+ \rightarrow \mathbf{R}$  is *starshaped* if  $x \mapsto f(x)/x$  is decreasing.

**Lemma 6.1.** Suppose that  $m_0 > 0$  and any one of these three conditions holds:

- (a)  $\Psi(m)$  is an increasing function of  $m$ , and the initial condition of state evolution satisfies  $m_0 \geq \text{HFP}(\Psi)$ ;
- (b)  $\Psi(m)$  is an increasing starshaped function of  $m$ .
- (c)  $\text{HFP}(\Psi) = 0$ .

Then state evolution converges to the highest fixed point:

$$\lim_{t \rightarrow \infty} m_t = \text{HFP}(\Psi).$$

Further, if  $\text{HFP}(\Psi) > 0$  and  $\Psi(m)$  is a starshaped function of  $m$ , then  $\liminf_{t \rightarrow \infty} m_t > 0$ .

The proof is a standard calculus exercise; we omit it.

**Lemma 6.2.** The function  $m \mapsto \Psi(m)$  is starshaped for all of the following choices of the denoiser  $\eta$ :

- Soft thresholding.
- Positive soft thresholding.
- Block soft thresholding.
- James-Stein shrinkage.
- Monotone regression denoiser.
- Total variation denoiser.

The proof of this Lemma is deferred to Appendix I.

## 6.2 State Evolution Phase Transition

Consider a collection  $\mathcal{F}_{N,\varepsilon}$  of probability distributions over  $\mathbf{R}^N$ , indexed by  $\varepsilon \in [0, 1]$  as per Section 1.1 (these do not need to be simple sparse signals). For a sequence of probability distributions  $\nu = \{\nu_N\}_{N \in \mathbf{N}}$  we write  $\nu \in \mathcal{F}_\varepsilon$  if  $\nu_N \in \mathcal{F}_{N,\varepsilon}$  for all  $N$  and the limit on the Eq. (6.1) exists for each  $m \in \mathbf{R}_+$ . Letting  $\text{HFP}(\delta, \tau, \nu) = \text{HFP}(\Psi(\cdot; \delta, \tau, \nu))$ , we consider the minimax value:

$$\text{HFP}^*(\varepsilon, \delta) = \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_\varepsilon} \text{HFP}(\delta, \tau, \nu).$$

**Definition 6.3.** For  $\varepsilon \in [0, 1]$ , define the state evolution phase transition as

$$\delta_{SE}(\varepsilon|\eta) \equiv \inf \{ \delta \geq 0 : \text{HFP}^*(\varepsilon, \delta) = 0 \}.$$

Note that  $\text{HFP}^*(\varepsilon, \delta)$  is monotone decreasing as a function of  $\delta$ , by definition of the state evolution mapping  $\Psi$ , cf. Eq. (6.1). It follows that  $\text{HFP}^*(\varepsilon, \delta) = 0$  for  $\delta > \delta_{SE}(\varepsilon)$  and  $\text{HFP}^*(\varepsilon, \delta) > 0$  for  $\delta < \delta_{SE}(\varepsilon)$ . Further, by nestedness, it is monotone increasing as a function of  $\varepsilon$ , which implies that  $\varepsilon \mapsto \delta_{SE}(\varepsilon)$  is monotone increasing. The rationale for this definition is that, for  $\delta > \delta_{SE}(\varepsilon)$  and under any of the assumptions of Lemma 6.1, state evolution predicts that AMP will correctly recover the signal  $x_0$ .

**Theorem 6.1.** Let  $\mathcal{F}_{N,\varepsilon}$  be a nested, scale-invariant collection of probability distributions, and assume that the shrinker  $\eta(\cdot; \tau, \sigma)$  obeys the scaling relation (1.3). Define the minimax MSE  $M(\varepsilon|\eta)$  as per Eq. (1.5). Then

$$\delta_{SE}(\varepsilon|\eta) = M(\varepsilon|\eta).$$

In order to prove this result, we shall first establish a more general fact. Given a sequence of distributions  $\nu = \{\nu_N\}_{N \in \mathbf{N}}$  and, for any  $\tau \in \Theta$ , we let

$$\delta_{SE}(\tau, \nu|\eta) \equiv \inf \{ \delta \geq 0 \text{ such that } \text{HFP}(\delta, \tau, \nu) = 0 \}. \quad (6.2)$$

The rationale for this definition is clear. Under the assumption that state evolution holds, for a signal  $x_0$  sampled from distribution  $\nu_N$ , AMP (with tuning parameter  $\tau$ ) is guaranteed to reconstruct  $x_0$  if and only if  $\delta > \delta_{SE}(\tau, \nu|\eta)$ .

Analogously, we let

$$M(\tau, \nu|\eta) = \sup_{\sigma > 0} \lim_{N \rightarrow \infty} \frac{1}{N\sigma^2} \mathbb{E}_{\nu_N} \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \sigma\mathbf{Z}; \tau, \sigma)\|_2^2 \right\}. \quad (6.3)$$

the normalized MSE for denoising with worst case case signal-to-noise ratio. With these definitions we have the following.

**Lemma 6.3.** For any sequence of probability measures  $\nu = \{\nu_N\}_{N \geq 1}$ , and any  $\tau \in \Theta$ , we have

$$\delta_{SE}(\tau, \nu|\eta) = M(\tau, \nu|\eta). \quad (6.4)$$

*Proof.* By definition

$$\begin{aligned}
\delta_{SE}(\tau, \nu | \eta) &= \inf \left\{ \delta > 0 \text{ such that } \Psi(m; \delta, \tau, \nu) < m \text{ for all } m > 0 \right\} \\
&= \inf \left\{ \delta > 0 \text{ such that } \sup_{m>0} \frac{1}{m} \Psi(m; \delta, \tau, \nu) < 1 \right\} \\
&= \inf \left\{ \delta > 0 \text{ such that } \sup_{m>0} \lim_{N \rightarrow \infty} \frac{1}{mN} \mathbb{E} \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \sqrt{m/\delta} \mathbf{Z}; \tau, \sqrt{m/\delta})\|^2 \right\} < 1 \right\} \\
&= \inf \left\{ \delta > 0 \text{ such that } \sup_{\sigma>0} \lim_{N \rightarrow \infty} \frac{1}{\delta \sigma^2 N} \mathbb{E} \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \sigma \mathbf{Z}; \tau, \sigma)\|^2 \right\} < 1 \right\} \\
&= \inf \left\{ \delta > 0 \text{ such that } \frac{1}{\delta} M(\tau, \nu | \eta) < 1 \right\} = M(\tau, \nu | \eta).
\end{aligned}$$

□

We are now in position to prove Theorem 6.1.

*Proof of Theorem 6.1.* Throughout the proof we drop the argument  $\eta$ , since this is kept constant.

Define  $\delta_*(\varepsilon) \equiv \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_\varepsilon} \delta(\tau, \nu)$ . We claim that  $\delta_*(\varepsilon) = \delta_{SE}(\varepsilon)$ . Indeed choose  $\delta \in [0, \delta_{SE}(\varepsilon))$ . Then by definition there exists  $m > 0$  such that, for all  $\tau \in \Theta$ , there exists  $\nu \in \mathcal{F}_\varepsilon$  with  $\text{HFP}(\delta, \tau, \nu) > m$ . Hence, for all  $\tau \in \Theta$ , there exists  $\nu \in \mathcal{F}_\varepsilon$  with  $\delta(\tau, \nu) \geq \delta$ . This implies that, for all  $\tau \in \Theta$ ,  $\sup_{\nu \in \mathcal{F}_\varepsilon} \delta(\tau, \nu) > \delta$ , i.e.  $\delta_*(\varepsilon) \geq \delta$ . Proceeding in the same way, it is immediate to prove that, for any  $\delta \in [0, \delta_*(\varepsilon))$ , we have  $\delta_{SE}(\varepsilon) > \delta$ . Hence, we conclude that  $\delta_{SE}(\varepsilon) = \delta_*(\varepsilon)$  as claimed.

To conclude the proof, we note that, by Lemma 6.3, we have

$$\begin{aligned}
\delta_*(\varepsilon) &= \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} M(\tau, \nu) \\
&= \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \sup_{\sigma > 0} \lim_{N \rightarrow \infty} \frac{1}{N \sigma^2} \mathbb{E}_{\nu_N} \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \sigma \mathbf{Z}; \tau, \sigma)\|_2^2 \right\} \\
&= \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_\varepsilon} \frac{1}{N} \mathbb{E}_\nu \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \mathbf{Z}; \tau)\|_2^2 \right\},
\end{aligned}$$

where the last equality follows from the scale invariant property of  $\mathcal{F}_{N, \varepsilon}$ . The last quantity is nothing but  $M(\varepsilon)$ . □

### 6.3 Non-convergence of state evolution

By Definition 6.3 and Lemma 6.1.(c) it follows that, for all  $\delta < \delta_{SE}(\varepsilon | \eta)$ , all probability distributions  $\nu \in \mathcal{F}_\varepsilon$ , and all initial conditions  $m_0$ , state evolution converges to the zero error fixed point, namely  $\lim_{t \rightarrow \infty} m_t = 0$ . Viceversa, for  $\delta > \delta_{SE}(\varepsilon | \eta)$ , there exists  $\nu \in \mathcal{F}_\varepsilon$ , and an initial condition  $m_0$ , such that  $\lim_{t \rightarrow \infty} m_t = m_\infty > 0$ . The reader might wonder whether this conclusion (non-convergence to 0) also holds if we use the initial condition that is relevant for AMP, i.e. if, for any  $\delta > \delta_{SE}(\varepsilon | \eta)$ , there exists  $\nu \in \mathcal{F}_\varepsilon$  such that, taking  $m_0 = \lim_{N \rightarrow \infty} N^{-1} \mathbb{E}_{\nu_N} \{\|\mathbf{X}\|_2^2\}$ , we have  $\lim_{t \rightarrow \infty} m_t = m_\infty > 0$ .

Lemmas 6.1.(b) and 6.2 immediately imply that the answer is positive for soft thresholding, positive soft thresholding, block soft thresholding, James-Stein shrinkage, monotone regression and total variation denoising. It turns out that the answer is positive also for firm thresholding and the global minimax denoiser. In Appendix E we describe the argument for these cases.



## 7 Phase transitions for other algorithms

Formula (1.13) connects an algorithmic property – phase transitions of AMP recovery algorithms – with a property from statistical decision theory – minimax mean squared error in denoising.

We want to explore a further connection, relating the behavior of convex optimization with that of certain AMP algorithms. As proved in [BM11b], in the large system limit AMP with soft thresholding denoiser effectively computes the solution to

$$(P_{1,\lambda}) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1, \quad x \in \mathbf{R}^N.$$

for an appropriately calibrated  $\lambda = \lambda(\tau)$ .

More generally, consider a generalized reconstruction method of the form

$$(P_J) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda J(x), \quad x \in \mathbf{R}^N, \quad (7.1)$$

where  $J : \mathbf{R}^N \rightarrow \mathbf{R}$  is a convex penalization. To this reconstruction problem, we can associate an AMP algorithm, by using the denoiser  $\eta^J(\cdot; \tau)$  in Eq. (1.7), (1.8), (1.9), whereby

$$\eta^J(y; \tau) \equiv \arg \min_{x \in \mathbf{R}^N} \left\{ \frac{1}{2}\|y - x\|_2^2 + \tau J(x) \right\}. \quad (7.2)$$

(we also let  $\eta^J(\cdot; \tau, \sigma) = \eta^J(\cdot; \tau \cdot \sigma)$ ). In other words  $\eta^J$  is the proximal operator of the penalization  $J(\cdot)$ . We will refer to this algorithm as to AMP- $J$ .

We then have the following general correspondence, which follows immediately by writing the stationarity condition of problem  $P_J = P_J(\lambda)$  and the fixed points of AMP- $J$  (see [Mon12]).

**Proposition 7.1.** *Any fixed point  $x^\infty$  of AMP- $J$  with fixed point parameters  $\tau_\infty$ ,  $\mathbf{b}_\infty$ ,  $\sigma_\infty$  corresponds to a stationary point of  $P_J(\lambda)$  with  $\lambda$  given by*

$$\lambda = \tau_\infty \sigma_\infty (1 - \mathbf{b}_\infty), \quad \mathbf{b}_\infty = \frac{1}{n} \operatorname{div} \eta(y; \tau^\infty) \Big|_{y=y^\infty} \quad (7.3)$$

*In particular, if the regularizer  $J$  is convex, then fixed points correspond to minimizers.*

**Example 7.1.** The fixed points of AMP with positive soft thresholding denoiser are solutions of

$$(P_{1,\lambda}^+) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda \langle 1, x \rangle, \quad x \in \mathbf{R}_+^N,$$

where  $\langle \cdot, \cdot \rangle$  denotes the standard scalar product over  $\mathbf{R}^N$ , and  $1$  is the all-one vector.

**Example 7.2.** The fixed points of AMP with capping denoiser effectively are solutions of

$$(P_\square) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2, \quad x \in [0, 1]^N.$$

For noiseless compressed sensing reconstruction, the correspondence involves the limit  $\lambda \rightarrow 0$  of the above problem, that is

$$\begin{aligned} & \text{minimize} && J(x) && x \in \mathbf{R}^N, \\ & \text{subject to} && y = Ax. \end{aligned}$$

Phase transitions for such convex programs were characterized in [Don06, DT09a, DT10a] for the three examples mentioned above, using methods from combinatorial geometry. Thus, whenever AMP converges with high probability, formula (1.13) connects fundamental problems of minimax decision theory to fundamental problems in high dimensional combinatorial geometry.

We will next verify numerically this correspondence beyond the three classical examples (again focusing on noiseless measurements). In the block-structured case, we can compare block-soft AMP to the following convex optimization problem:

$$(P_{2,1}) \quad \begin{aligned} &\text{minimize} && \|x\|_{2,1} \equiv \sum_{m=1}^M \|block_m(x)\|_2, \\ &\text{subject to} && y = Ax. \end{aligned}$$

The norm  $\|x\|_{2,1}$  is called the mixed  $\ell_{2,1}$  norm. Figure 15 presents empirical reconstruction results obtained by solving  $(P_{2,1})$ . These experiments verify that the phase transition occurs around the location predicted by (1.13), just as with block soft thresholding AMP.

Analogously, we can compare monoreg AMP to the following convex optimization problem:

$$(P_{\text{mono}}) \quad \begin{aligned} &\text{minimize} && \langle 1, \Delta x \rangle \\ &\text{subject to} && y = Ax, \quad \Delta x \geq 0, \end{aligned}$$

where  $\Delta x = (\Delta x_1, \dots, \Delta x_N)$ ,  $\Delta x_i = (x_{i+1} - x_i)$ . Figure 16 verifies that the phase transition occurs around the location predicted by (1.13), just as with monoreg AMP.

Finally, we can compare TV AMP to the following convex optimization problem:

$$(P_{TV}) \quad \begin{aligned} &\text{minimize} && \|x\|_{TV}, \\ &\text{subject to} && y = Ax, \end{aligned}$$

where again  $\|x\|_{TV} \equiv \sum_{i=1}^{N-1} |\Delta x_i|$ . Figure 17 verifies that the phase transition occurs at the location predicted by (1.13), just as with TV AMP.

## 7.1 Interpretation of separable denoisers in terms of penalized least-squares

Scalar soft thresholding  $\eta^{soft}$  can be interpreted as the solution of a  $\ell_1$ -penalized least-squares problem

$$(P_J) \quad \text{minimize} \quad \frac{1}{2}(y - x)^2 + J(x),$$

where  $J(x) = \lambda|x|$  (here we redefined  $J$  to absorb the factor  $\tau$ ).

It turns out that a similar interpretation can be given for any scalar denoiser (and hence for any separable denoiser as well). In particular, the minimax and firm thresholding rules  $\eta^{all}$  and  $\eta^{firm}(\cdot; \tau_1, \tau_2)$  are optimizers of penalization schemes of the same form as above, with  $J(x)$  *non-convex*.

We can construct the penalty  $J(\cdot)$  corresponding to a denoiser  $\eta(\cdot)$  by observing that  $x + J'(x) = y$  at the solution  $x = \eta(y)$ . Defining the residual  $\Delta(y) = y - \eta(y)$ , and noting that  $\eta(y) + \Delta(y) = y$ , we obtain that  $\Delta(y)$  and  $J'(x)$  are related through the change of variables

$$J'(\eta(y)) = \Delta(y).$$

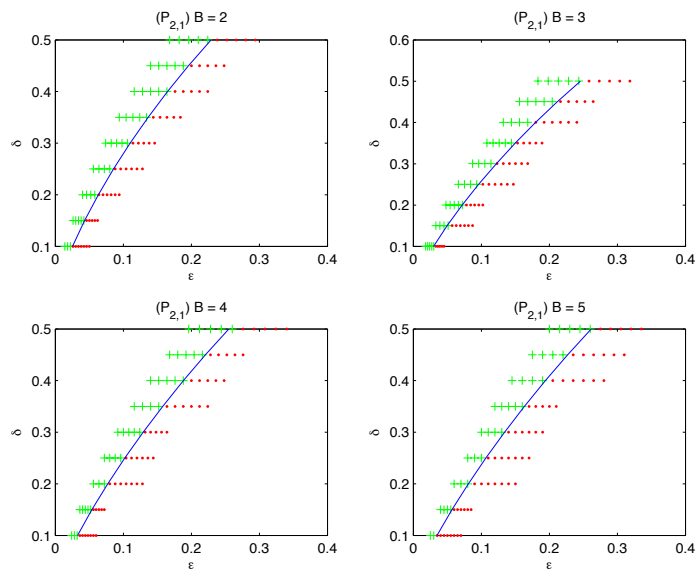


Figure 15: Phase transition results for reconstructing block-sparse signals using the convex program  $(P_{2,1})$ . Here  $\delta = n/N$  is the undersampling fraction and  $\varepsilon = k/N$  is the sparsity measure. Here the signal dimension is  $N = 1000$ ,  $\delta = n/N$  is the undersampling fraction, and  $\varepsilon$  is the sparsity parameter (fraction of non-zero entries). Red: probability of correct recovery smaller than 50%. Green: probability of correct recovery larger than 50%. The problem  $(P_{2,1})$  was solved using Sedumi. The curve separating success from failure is the minimax MSE curve  $\delta = M(\varepsilon|\text{BlockSoft})$ .

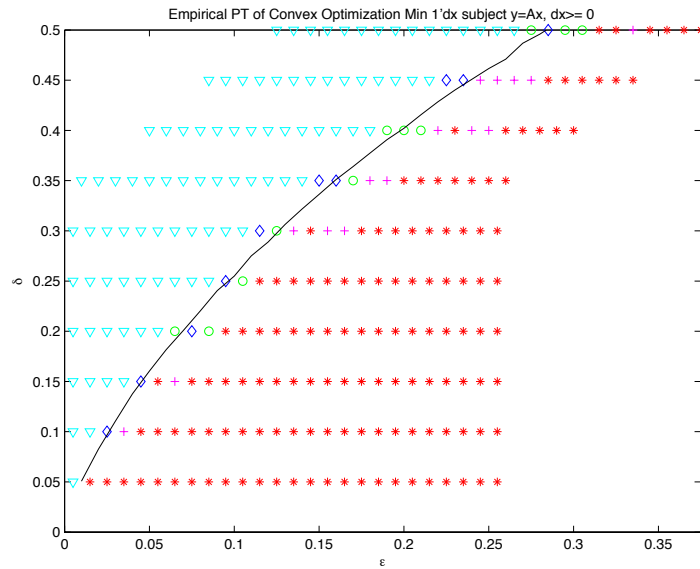


Figure 16: Phase transition results for reconstructing monotone signals using the convex program ( $P_{\text{mono}}$ ). Here the signal dimension is  $N = 200$ ,  $\delta = n/N$  is the undersampling fraction, and  $\varepsilon$  is the sparsity parameter (fraction of change points). Red: fraction of correct recovery smaller than 50%. Aqua: fraction of correct recovery larger than 50%. Curve: minimax MSE curve  $\delta = M(\varepsilon|\text{MonoReg})$ . Compare with figure 11.

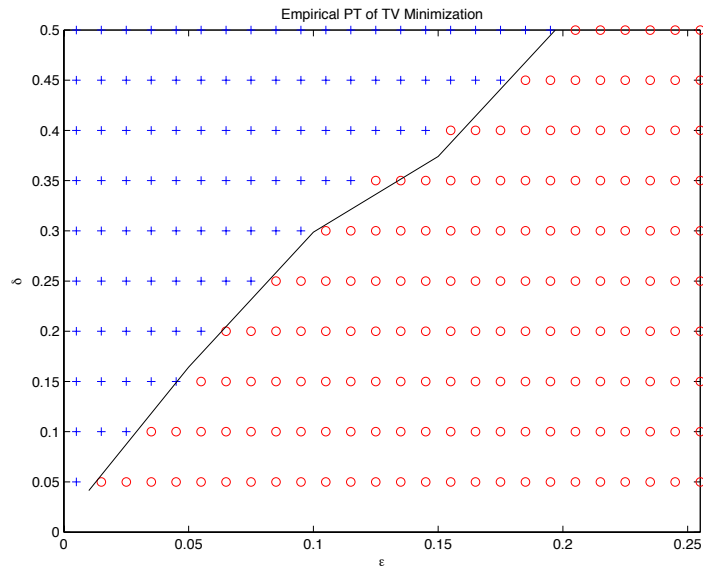


Figure 17: Phase transition results for reconstructing piecewise constant signals using the convex program ( $P_{TV}$ ). Here,  $N = 200$ ,  $\delta = n/N$  is the undersampling fraction and  $\epsilon$  is the fraction of change points. Here random changepoints were used. Red: fraction of correct recovery smaller than 50%. Aqua: fraction of correct recovery larger than 50%. Curve: minimax MSE curve  $\delta = M_{\text{random}}(\epsilon|TV)$ . Compare with Figure 13.

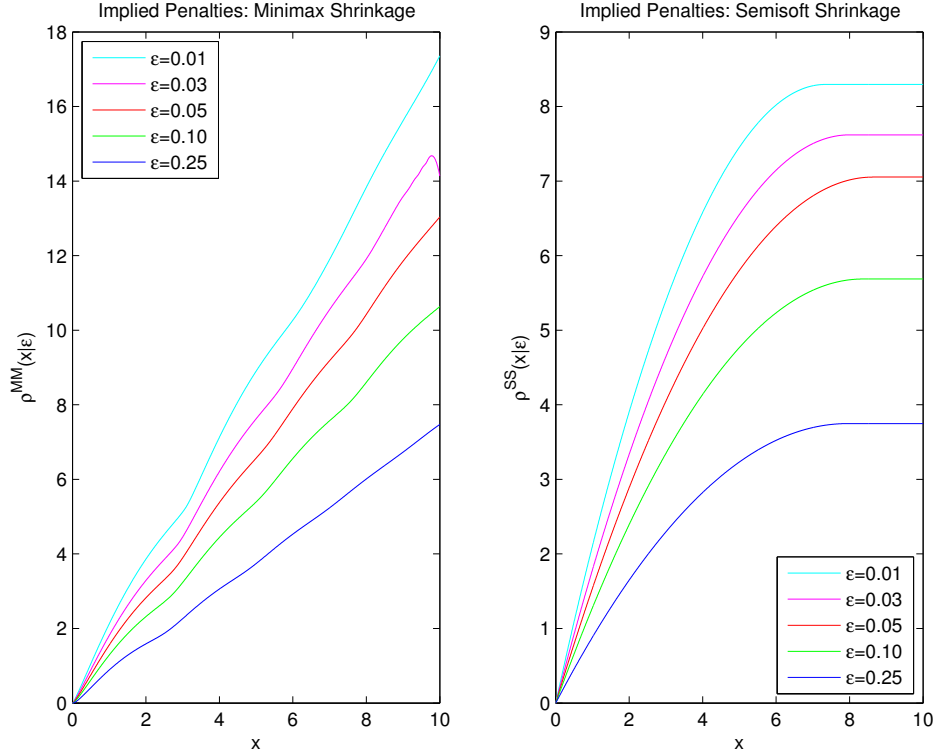


Figure 18: Implied penalties  $J(x)$  for two denoisers at  $\varepsilon = 0.01, 0.025, 0.05, 0.10, 0.25$  (from top to bottom). Left: Globally minimax shrinkage. Right: Minimax firm shrinkage. We plot only results for positive arguments  $x > 0$ ; results for negative arguments are obtained by symmetry  $J(-x) = J(x)$ .

In other words,

$$J(x) = \int_0^x \Delta(\eta^{-1}(u)) du.$$

Figure 18 shows the implied penalties for minimax firm and global minimax shrinkage, respectively. In both cases, the optimal penalizations are nonconvex, in accord with the commonly-held belief that nonconvex penalization is “superior” to  $\ell_1$  penalization [CY08, KXAH10]. However, the optimal penalization is seemingly very close to the  $\ell_1$  penalization, so that the prejudice in favor of nonconvex penalization must be re-examined.<sup>6</sup>

A similar analysis can be carried out for block separable denoisers that are covariant under rotation, i.e. if  $\eta(\cdot; \tau) : \mathbf{R}^B \rightarrow \mathbf{R}^B$  satisfies  $\eta(Rx; \tau) = R\eta(x; \tau)$  for any rotation  $R$ . We already mentioned that the block soft denoiser can be written as

$$\eta^{soft}(y; \tau) = \operatorname{argmin}_{x \in \mathbf{R}^B} \left\{ \frac{1}{2} \|y - x\|_2^2 + \tau \|x\|_2 \right\}.$$

<sup>6</sup>The minimax nonlinearity is somewhat complicated to implement – see Appendix B. Figure 18 suggests to us that among piecewise linear rules, close to the best phase transition behavior is obtained by a rule which obeys  $\eta(y) < y - \text{const}$  for large  $y$ , a possibility that is not explored in the firm family, or in any other conventional proposal.

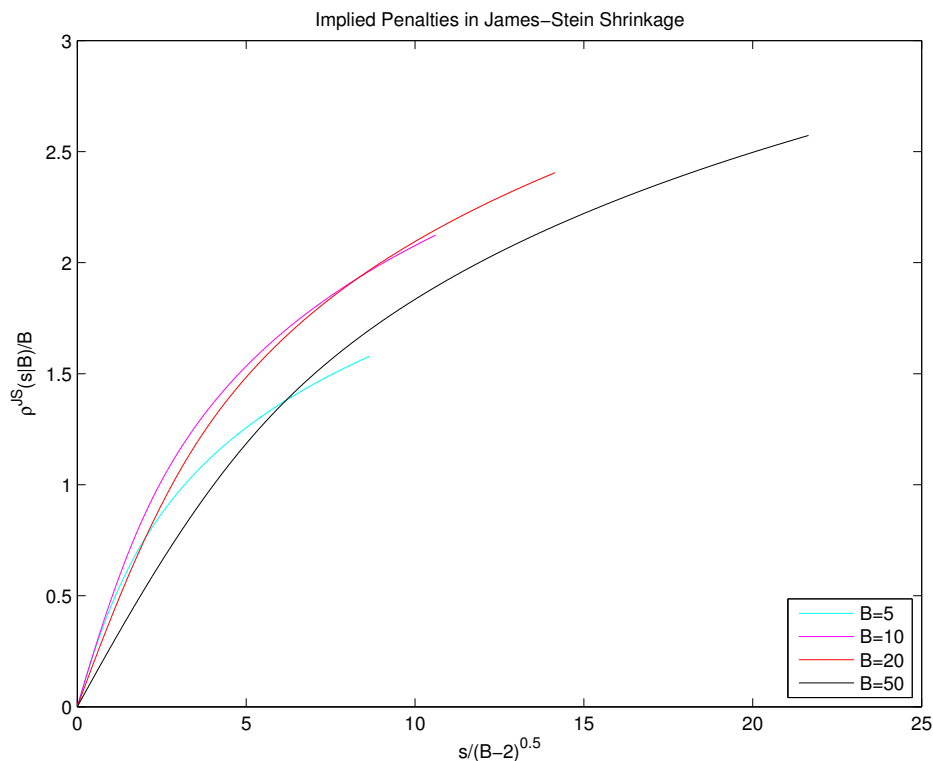


Figure 19: Implied penalties  $J(s|B)$  for James-Stein positive-part estimator for block sizes  $B \in \{5, 10, 20, 50\}$ . Note that both axes are scaled with  $B$  dependent, so that all plots fit on a common scale. More precisely, the  $y$ -axis presents  $J(s|B)/B$  and the  $x$ -axis presents  $s/\sqrt{B-2}$ .

An implied penalty can also be derived for block James-Stein shrinkage  $\eta^{JS}$ . Due to the covariance under rotation, the corresponding penalty only depends on the modulus of  $x \in \mathbf{R}^B$ . Figure 19 shows the implied penalty  $J(s|B)$  as a function of  $s = \|x\|_2$ . Namely, the penalty  $J(\cdot|B) : \mathbf{R}_+ \rightarrow \mathbf{R}$  is such that, for  $y \in \mathbf{R}^B$ ,

$$\eta^{JS}(y) = \operatorname{argmin}_{x \in \mathbf{R}^B} \min \left\{ \frac{1}{2} \|y - x\|_2^2 + J(\|x\|_2|B) \right\}$$

coincides with the positive-part James-Stein estimator. The implicit penalization is again nonconvex.

## A Classical cases

The general formula (1.13) was already validated in [DMM09] for the following three classical cases:

- (i) Simple sparse signals from the class  $\mathcal{F}_{N,\varepsilon}$  (cf. Eq. (1.2)) and soft thresholding denoiser.
- (ii) Non-negative sparse signals with soft positive thresholding denoiser, cf. Example 1.3.
- (iii) Box constrained signals with capping denoiser, cf. Example 1.3.

Analytical expressions for the phase transition curves were computed using state evolution in the Online Supplement to [DMM09]. We review the results here since they provide a useful stepping stone for understanding more complicate cases (cf., e.g.. Section 6).

Notice that the last of the examples above (box-constrained signals) is not scale invariant, according to the general definition of Section 1.1. For non scale-invariant classes, the definitions (1.4) and (1.5) are generalized by taking the supremum over the noise covariance as well. Namely, for a generic class  $\mathcal{F}_{N,\varepsilon}$ , we let

$$M(\mathcal{F}_{N,\varepsilon}|\eta) = \inf_{\tau \in \Theta} \sup_{\sigma \in \mathbf{R}_+} \sup_{\nu_N \in \mathcal{F}_{N,\varepsilon}} \frac{1}{N\sigma^2} \mathbb{E}_{\nu_N} \left\{ \|\mathbf{X} - \eta(\mathbf{X} + \sigma \mathbf{Z}; \tau\sigma)\|_2^2 \right\}, \quad (\text{A.1})$$

$$M(\varepsilon|\eta) = \lim_{N \rightarrow \infty} M(\mathcal{F}_{N,\varepsilon}|\eta). \quad (\text{A.2})$$

It is easy to check that this definition coincides with the earlier one for scale-invariant classes. We will write  $M(\varepsilon|\text{Cap})$  instead of  $M(\varepsilon|\eta^{\text{cap}})$  for box constrained signals with capping denoiser, i.e. case (iii) above. The results for case (iii) is further elucidated by comparing it with the following scale invariant problem:

- (iv) We consider sparse non-negative signals  $x_0 \geq 0$ , modeled through the class  $\mathcal{F}_{N,\varepsilon,+}$ , and the simple positive part denoiser  $\eta^+(y) \equiv \max(y, 0)$ . We denote corresponding minimax risk by  $M(\varepsilon|\text{Pos})$ .

The minimax risk for examples (i), (ii), (iv) can be computed explicitly. Indeed in both cases we have

$$\begin{aligned} M(\varepsilon|\eta) &= \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}/\nu \in \mathcal{F}_{1,\varepsilon,+}} \mathbb{E} \left\{ (\eta(X + Z; \tau) - X)^2 \right\} \\ &= \inf_{\tau \in \mathbf{R}_+} \sup_{\nu = (1-\varepsilon)\delta_0 + \varepsilon\delta_\mu} \mathbb{E} \left\{ (\eta(X + Z; \tau) - X)^2 \right\}. \end{aligned}$$

Here  $X \sim \nu$  and  $Z \sim \mathbf{N}(0, 1)$  are independent random variables. The reduction second equality follows from the remark that the extremal distributions in  $\mathcal{F}_{1,\varepsilon}$  and  $\mathcal{F}_{1,\varepsilon,+}$  are mixtures of two point distributions. This remark reduces the calculation of the minimax risk to a simple calculation [DMM09], whose results are summarized below.



**Proposition A.1.** *The minimax risk for the problems stated above is*

$$M(\varepsilon|\text{Soft}) = \frac{2\phi(\tau)}{\tau + 2\phi(\tau) - 2\tau\Phi(-\tau)}, \quad \varepsilon = \frac{2\phi(\tau) - 2\tau\Phi(-\tau)}{\tau + 2\phi(\tau) - 2\tau\Phi(-\tau)}, \quad (\text{A.3})$$

$$M(\varepsilon|\text{SoftPos}) = \frac{\phi(\tau)}{\tau + \phi(\tau) - \tau\Phi(-\tau)}, \quad \varepsilon = \frac{\phi(\tau) - \tau\Phi(-\tau)}{\tau + \phi(\tau) - \tau\Phi(-\tau)}, \quad (\text{A.4})$$

$$M(\varepsilon|\text{Cap}) = \frac{1}{2}(1 + \varepsilon), \quad (\text{A.5})$$

$$M(\varepsilon|\text{Pos}) = \frac{1}{2}(1 + \varepsilon). \quad (\text{A.6})$$

(The first two are parametric expressions in  $\tau$ , which is the optimal threshold at the given sparsity level.)

Also, the AMP phase transition for the noiseless reconstruction problem is given in all of these cases by  $\delta = M(\varepsilon|\eta)$ . In other words AMP succeeds with high probability of  $\delta > M(\varepsilon|\eta)$  and fails with high probability if  $\delta < M(\varepsilon|\eta)$ .

As mentioned above, the calculation of the minimax risk is a calculus exercise, and follows the same lines as in [DMM09]. This coincide with the AMP threshold by the general analysis of Section 6 for cases (i), (ii), (iv). For the non-scale invariant case, we refer, once more, to [DMM09].

Notice that problem (iii) and (iv) have the same minimax risk. This identity mirrors a result in [DT10a] that characterizes the phase transition threshold for reconstructing  $x_0 \in \mathcal{S}_N \subseteq \mathbf{R}^N$  from noiseless linear measurements  $y = Ax_0$ , with  $\mathcal{S}_N = [0, 1]^N$  or  $\mathcal{S}_N = \mathbf{R}_+^N$ . If a simple feasibility linear program is used (namely, find any  $x \in \mathcal{S}_N$  with  $y = Ax$ ), then the undersampling threshold for both problems is given by  $\delta = (1 + \varepsilon)/2$ .

## B Calculation of minimax MSE

In this Appendix we describe the calculation of the global minimax risk over the class  $\mathcal{F} \equiv \mathcal{F}_{1,\varepsilon}$ , as defined per Eq. (2.3). In particular, we will explain how the values in Table 1 and Figure 1 for  $M(\varepsilon|\text{Minimax})$  have been computed.

Throughout this section we let  $\text{MSE}(\nu, \eta) \equiv \mathbb{E}\{(\eta(X + Z) - X)^2\}$  where  $X \sim \nu$  and  $Z \sim \mathbf{N}(0, 1)$  are independent random variables. From standard minimax theory [Joh12], and using an identity attributed to Brown, we have, for  $\mathcal{F}$  convex and weakly compact

$$\inf_{\eta} \sup_{\nu \in \mathcal{F}} \text{MSE}(\nu, \eta) = \sup_{\nu \in \mathcal{F}} \inf_{\eta} \text{MSE}(\nu, \eta) = \sup_{\nu \in \mathcal{F}} \text{MSE}(\nu, \eta^\nu) = 1 - \inf_{\nu \in \mathcal{F}} I(\gamma \star \nu). \quad (\text{B.1})$$

Here  $\eta^\nu$  is the posterior mean estimator for prior  $\nu$ ,  $\gamma$  is the Gaussian measure  $\gamma(dx) = \phi(x)dx$ ,  $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ ,  $\gamma \star \nu$  denotes the convolution of measures, and  $I$  denotes the Fisher information. For a probability measure  $\nu_f(dx) = f(x)dx$ , with density  $f$  with respect to the Lebesgue measure this is defined as

$$I(\nu_f) = \int \frac{(f'(x))^2}{f(x)} dx.$$

Further, if  $\mathcal{F}$  is convex and weakly compact, the set of probability measures  $\{\gamma \star \nu : \nu \in \mathcal{F}\}$  is also convex and weakly compact. It follows from [Hub64, Theorem 4] that the inf in Eq. (B.1) is

achieved at a unique point  $\nu = \nu^*$ . Hence

$$M(\varepsilon|\text{Minimax}) = 1 - \min_{\nu \in \mathcal{F}} I(\gamma \star \nu) = 1 - I(\gamma \star \nu^*) \equiv 1 - I^*, \quad (\text{B.2})$$

where we defined  $I^* = I^*(\varepsilon) = I(\gamma \star \nu^*)$ . The unique minimizer  $\nu^*$  is known as the *least favorable* distribution. The minimax optimal denoiser (achieving the inf over  $\eta$  in Eq. (B.1)) is the posterior expectation with respect to the prior  $\nu^*$ .

Bickel and Collins [BC83, Theorem 1], prove that, under suitable assumption on the class  $\mathcal{F}$ , the least favorable distribution is a mixture of point masses

$$\nu^* = \sum_{i=-\infty}^{\infty} \alpha_i \delta_{\mu_i},$$

where  $\sum_i \alpha_i = 1$ ,  $\alpha_i > 0$  and the sequence  $\{\mu_i\}_{i \in \mathbf{Z}}$  has no accumulation points except, possibly, at  $\pm\infty$ . As mentioned above, the minimax denoiser is the posterior expectation associate to the prior  $\nu^*$ . By Tweedie's formula, this takes the form

$$\eta^*(y) = y - \psi^*(y),$$

where  $\psi^*$  is the so-called score function

$$\psi^*(y) = -\frac{d}{dy} \log f^*(y),$$

and  $f^*$  is the density of  $\nu^* \star \gamma$ .

Focusing now specifically on the class  $\mathcal{F} = \mathcal{F}_{1,\varepsilon}$ . This case is covered by the general theory of [BC83], and corresponds to their example (ii). Without loss of generality we can assume that the  $\mu_i$  are monotone increasing, with  $\mu_{-i} = \mu_i$ , and that  $\mu_0 = 0$ , with  $\alpha_0 = (1 - \varepsilon)$ . A conjecture of Mallows [Mal78] states that in fact we may take

$$\mu_i = c \cdot i \quad \alpha_i = \frac{1}{2}(1 - \varepsilon)(1 - \lambda)\lambda^{|i|-1}, \quad i \in \mathbf{Z} \setminus \{0\}.$$

In other words, the conjectured least favorable distribution has the form of a two-sided geometric distribution on a scaled copy of  $\mathbf{Z}$ . While the conjecture has not been proved, Mallows [Mal78] provided an argument (based on an analogous problem in robust estimation [Hub64]) that suggesting that indeed it captures the correct tail behavior.

For estimating the minimax risk numerically, we chose a large parameter  $K$  and assume a generalized ‘‘Mallows form’’ for  $|i| > K$ . More precisely, we assume an equispaced grid, and geometrically decaying weights. This is a little more general than what Mallows proposed, having 3 total degrees of freedom (spacing, total weight and rate of decay), rather than two. For  $-K \leq i \leq K$ , we allow the parameters  $\alpha_i$  and  $\mu_i$  to vary freely. In this way we obtained a parametric family  $\nu_\theta$  of probability distributions, with parameter  $\theta = ((\alpha_i)_{i \in [K]}, (\mu_i)_{i \in [K]}, c_0, c_1, \lambda)$ , with

$$\nu_\theta = (1 - \varepsilon)\delta_0 + \frac{\varepsilon}{2} \cdot \sum_{k=1}^K \alpha_k \cdot (\delta_{\mu_k} + \delta_{-\mu_k}) + \frac{c_0}{2} \sum_{k>K} \lambda^k (\delta_{c_1 k} + \delta_{-c_1 k}).$$

Define

$$I^+ \equiv \min\{I(\gamma \star \nu_\theta) : \theta \in \mathbf{R}^{2K+3}\}. \quad (\text{B.3})$$

$\varepsilon$	$K$	$1 - I^+(\varepsilon)$	$\max_{\nu} \text{MSE}^*(\nu, \eta^+)$
0.01	2	0.0533	0.0533
0.05	2	0.1841	0.1841
0.10	2	0.3026	0.3026
0.15	2	0.3983	0.3984
0.20	4	0.4802	0.4803
0.25	4	0.5516	0.5516

Table 3: Numerical values of the lower bound on minimax MSE ( $1 - I^+(\varepsilon)$ ) and of upper bound on minimax MSE ( $1 - I_-(\varepsilon)$ ). The upper bound is the numerically computed worst-case MSE of the corresponding denoiser  $\eta^+$ . In all of these cases the bounds agree except possibly in the fourth decimal place.

The quantity  $I^+ = I^+(\varepsilon)$  can be estimated numerically, and provides an upper bound on  $I^*$ . We used up to  $K = 50$  and checked that the resulting  $I^+$  is insensitive to this choice. Notice that choice of the Mallows form for  $i > K$  is immaterial for two reasons: (i) As a consequence [BC83] and of the weak continuity of Fisher information,  $I^+$  should be insensitive to the tail behavior of the distribution  $F$ ; (ii) We are only using it to derive an upper bound.

In order to get a lower bound on  $I^*$ , we use Huber’s minimax theorem [Hub64, HR09], which implies that, for any  $\psi : \mathbf{R} \rightarrow \mathbf{R}$  differentiable in measure,

$$I(\gamma \star \nu^*) \geq \sup_g J(\psi, g) \equiv \sup_g \frac{(\int \psi' g_\psi)^2}{\int \psi^2 g_\psi}, \quad (\text{B.4})$$

where in the supremum  $g$  ranges over all densities of probability measures  $\gamma \star \nu$  with  $\nu \in \mathcal{F}_{1,\varepsilon}$ .

Let  $\nu^+$  denote the probability measure corresponding to the optimum of the parametric optimization (B.3). Denote by  $\psi^+$  denote the corresponding score function

$$\psi^+(y) = -\frac{d}{dy} \log f^+(y),$$

where  $f^+ = \nu^+ \star \gamma$ . This corresponds to a denoiser  $\eta^+(y) = y - \psi^+(y)$ . Let  $g^+$  denote a maximizing density  $g$  for  $J(\psi^+, g)$ . By Huber’s theory, this can be chosen to be two-point mixture  $(1 - \varepsilon)\delta_0 + \varepsilon\delta_{\mu^+}$  where  $\mu^+$  is chosen to achieve the worst case value on the right side of (B.4) and set  $I_- = I_-(\varepsilon) = J(\psi^+, g^+)$ .

We have the bounds  $I_- \leq I^* \leq I^+$ . Numerically, we compute integrals and extrema over fine grids with at least 100 samples per unit of range, getting not  $I^+$  and  $I_-$  but instead numerical approximations  $\tilde{I}^+$  and  $\tilde{I}_-$ . Table 3 presents some information about numerical approximation results, which may help the reader assess its accuracy for small values of  $K$ . Some minimizing distributions obtained in this way are shown in Figure 20; the mass points ( $\mu_i$ ) are displayed in Figure 21.

Our numerical results, showing that  $\tilde{I}_- \approx \tilde{I}_+$  allows us to infer that the Mallows form is approximately correct. The denoiser that we actually apply in our estimation and compressed sensing experiments is:

$$\eta^+(y) = y - \psi^+(y).$$

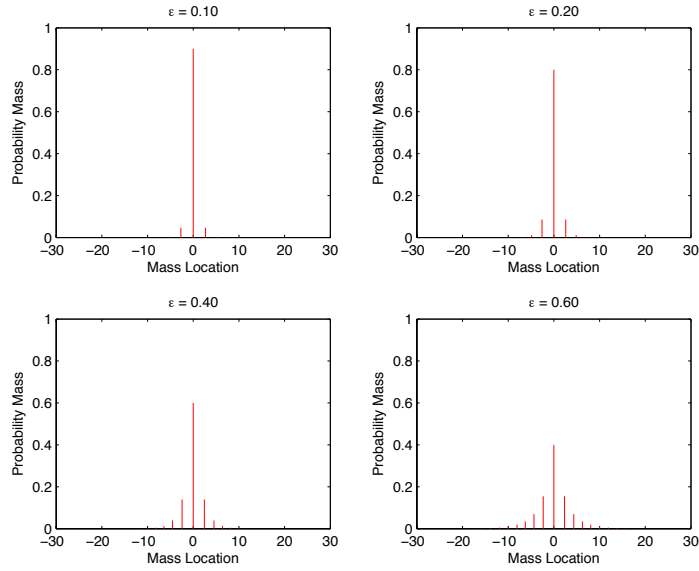


Figure 20: Least-favorable distributions over the class  $\mathcal{F}_{1,\epsilon}$  obtained by numerically minimizing Fisher information.

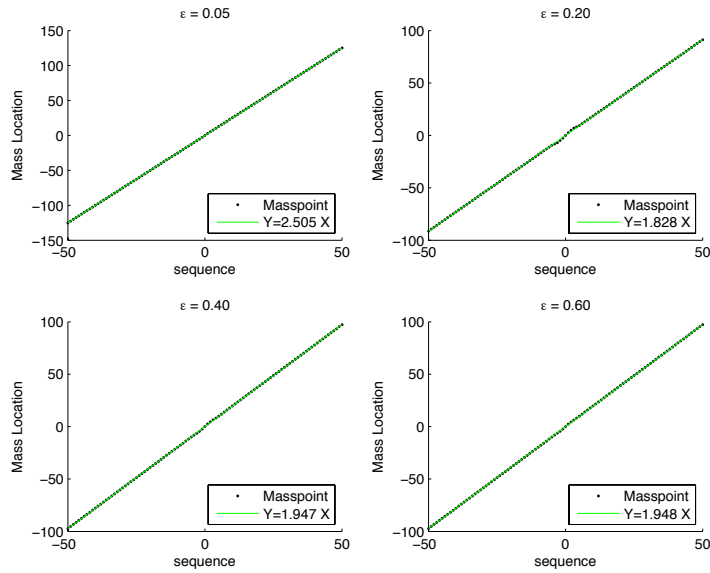


Figure 21: Locations of mass points in approximate least-favorable distribution ( $\mu_i$ ) versus index  $i$ , at various  $\epsilon$ , Note the approximate linearity; least-squares linear fit has slope parameter given in legend. The ‘wiggles’ away from the linear behavior occur near the origin.

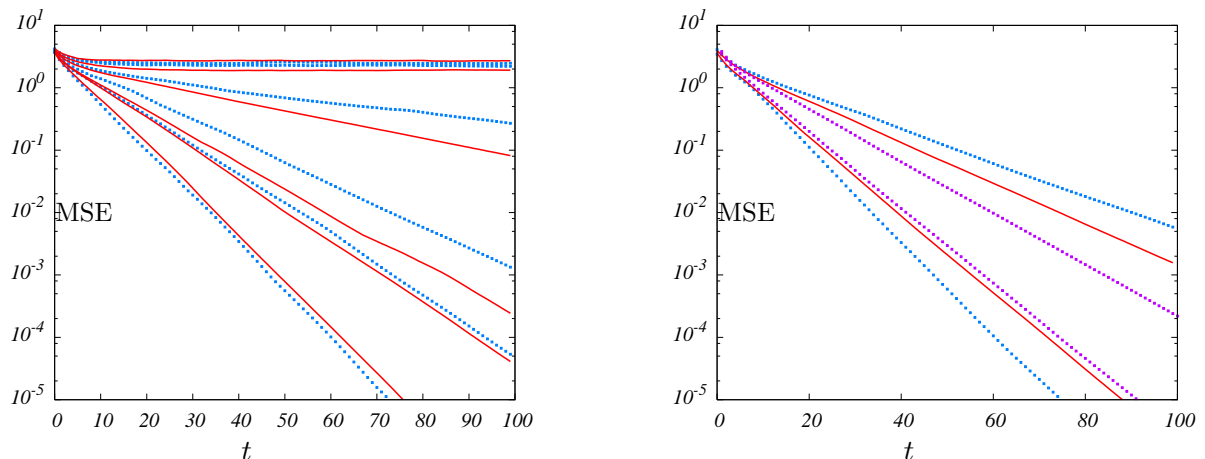


Figure 22: Convergence of AMP. Here we consider soft thresholding AMP for simple sparse signals with  $\varepsilon = 0.05$  (predicted threshold  $M(\varepsilon|\eta) \approx 0.2039$ ). Left frame: Median MSE after  $t$  iterations over 50 independent realizations of the sensing matrix. Blue dotted lines correspond to  $n = 2000$  and red continuous lines to  $n = 4000$ . From top to bottom:  $\delta = 0.19, 0.20, 0.21, 0.22, 0.23, 0.24$ . Right frame: 25% (lower tree curves) and 75% (upper tree curves) percentile curves for the case  $\delta = 0.23$ . The external (dotted), middle (continuous) and internal (dotted) pair refer, respectively, to  $n = 2000, 4000$ , and  $8000$ .

## C Convergence properties of AMP

Throughout the paper we checked convergence of AMP by imposing a threshold on the reconstruction accuracy and the number of iterations. For instance, in the case of separable denoisers, cf. Section 2.4, we declared the reconstruction successful if

$$\frac{\|\hat{x}^t - x_0\|_2^2}{\|x_0\|_2^2} < \gamma$$

for a certain choice of  $t$  and  $\gamma > 0$ . In particular, the results presented correspond to  $\gamma = 0.01$  and  $t = 300$ .

It is natural to ask how to choose  $\gamma$  and  $t$ , and whether different choices of  $\gamma$  and  $t$  would lead to significantly different estimates of the phase transition boundary. It turns out that the empirical phase transition is fairly insensitive to these choices for the cases considered here, as soon as  $t \gtrsim 100$  is sufficiently large and  $\gamma \lesssim 0.05$ . This insensitivity is related to the convergence properties of AMP. Indeed both theory and empirical evidence [DMM09, DMM11] indicate exponential convergence. Namely, for all  $\delta > M(\varepsilon|\eta)$ , there exist dimension independent constants  $C = C(\delta)$ ,  $b = b(\delta) > 0$  such that, with high probability,

$$\|\hat{x}^t - x_0\|_2^2 \leq C\|x_0\|_2^2 e^{-bt}.$$

On the other hand, for  $\delta < M(\varepsilon|\eta)$ ,  $\|\hat{x}^t - x_0\|_2^2 \geq c(\delta)n$ , with high probability.

Figure 22 presents data that confirm this behavior (further numerical evidence can be found in [DMM09]). The data refer to simple sparse signals with  $\varepsilon = 0.05$ , and soft thresholding denoising.

The curves correspond to several values of  $\delta$  close to the predicted phase transition location  $\delta = M(\varepsilon|\eta) \approx 0.0239$ . Notice the clear exponential decay of the error for  $\delta > M(\varepsilon|\eta)$  and a large constant mean square error for  $\delta < M(\varepsilon|\eta)$ .

If the phase transition has to be determined with relative accuracy  $\Delta$ , this suggests the rule of thumb  $\exp\{-b(\delta_* + \Delta)t\} \leq \gamma$  and  $c(\delta_* - \Delta) \geq \gamma$ . We verified that these conditions are satisfied by our choices of  $t$  and  $\gamma$ .

## D Calculation of minimax MSE for block soft thresholding

### D.1 Proof of Lemma 3.2

The argument is analogous to the one for the scalar case (corresponding to  $B = 1$ ) treated in Appendix A. For  $\mu \in \mathbf{R}^B$  and  $\tau \in \mathbf{R}_+$ , define the risk at  $\mu$  as

$$R(\mu; \tau) \equiv \mathbb{E}\{\|\eta^{soft}(\mu + \mathbf{X}; \tau) - \mu\|_2^2\}. \quad (\text{D.1})$$

where  $\mathbf{X} \sim \nu$  and  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{B \times B})$ . Since the two point mixtures are the extremal distributions in  $\mathcal{F}_{\varepsilon, B}$ , we have

$$\begin{aligned} M_B(\varepsilon|\text{BlockSoft}) &= \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_{B, \varepsilon}} \frac{1}{B} \mathbb{E}\{\|\eta(\mathbf{X} + \mathbf{Z}; \tau) - \mathbf{X}\|^2\} \\ &= \inf_{\tau \in \mathbf{R}_+} \sup_{\substack{\nu = (1 - \varepsilon)\delta_0 + \varepsilon\delta_\mu \\ \mu \in \mathbf{R}^B}} \frac{1}{B} \mathbb{E}\{\|\eta(\mathbf{X} + \mathbf{Z}; \tau) - \mathbf{Z}\|^2\} \\ &= \frac{1}{B} \inf_{\tau \in \mathbf{R}_+} \sup_{\mu \in \mathbf{R}^B} \left\{ (1 - \varepsilon)R(0; \tau) + \varepsilon R(\mu; \tau) \right\}. \end{aligned}$$

By the definition of chi-square distribution, we have

$$R(0; \tau) = \mathbb{E}\{\|\eta(\mathbf{Z}; \tau)\|^2\} = \mathbb{E}\{(\sqrt{X_B} - \tau)_+^2\}.$$

It follows from the invariance of the distribution of  $\mathbf{Z}$  under rotations that  $R(\mu; \tau)$  only depends on  $\mu$  through its norm  $\|\mu\|$ . Further, as proved in Appendix I.1  $R(\mu; \tau)$  is increasing in  $\|\mu\|$ , and

$$R(\infty; \tau) \equiv \lim_{\|\mu\| \rightarrow \infty} R(\mu; \tau) = B + \tau^2.$$

We therefore obtain

$$M_B(\varepsilon|\text{BlockSoft}) = \frac{1}{B} \inf_{\tau \in \mathbf{R}_+} \left\{ (1 - \varepsilon)\mathbb{E}\{(\sqrt{X_B} - \tau)_+^2\} + \varepsilon(B + \tau^2) \right\}. \quad (\text{D.2})$$

At this point the problem is reduced to a calculus exercise.

### D.2 Proof of Lemma 3.3

In this appendix we consider the asymptotics for large block size  $B \rightarrow \infty$ . It is easy to show that the minimax threshold level  $\tau$  must be of order  $\sqrt{B}$ . By a compactness argument, we can assume

$\tau = c\sqrt{B}$  for some  $c$  to be determined. Define the risk as in Eq. (D.1) (note that this depends implicitly on  $B$ ) and the normalized risk as  $\tilde{R}_B(\mu; \tau) = R(\mu; \tau)/B$ . We claim that

$$\lim_{B \rightarrow \infty} \tilde{R}_B(0; c\sqrt{B}) = (1 - c)_+^2, \quad (\text{D.3})$$

$$\lim_{B \rightarrow \infty} \tilde{R}_B(\infty; c\sqrt{B}) = 1 + c^2. \quad (\text{D.4})$$

Assuming these claim to hold, we have, by Eq. (D.2),

$$\lim_{B \rightarrow \infty} M_B(\varepsilon; \text{BlockSoft}) = (1 - \varepsilon)(1 - c)_+^2 + \varepsilon(1 + c^2).$$

Calculus shows that the minimum is achieved at  $c^* = 1 - \varepsilon$ , whence

$$\lim_{B \rightarrow \infty} M_B(\varepsilon | \text{BlockSoft}) = (1 - \varepsilon)\varepsilon^2 + \varepsilon(1 + (1 - \varepsilon)^2),$$

which coincides with the statement of Lemma 3.3.

In order to complete the proof, we have to prove claims (D.3) and (D.4). The second one is immediate because of Lemma I.1 that implies indeed  $\tilde{R}_B(\infty; c\sqrt{B}) = 1 + c^2$ .

The limit (D.3) follows instead from the central limit theorem. Indeed, let  $X_B$  denote a central chi-squared with  $B$  degrees of freedom. Its square root is the norm of a standard Gaussian random vector in dimension  $B$ , and concentrates around  $\sqrt{B}$ . Indeed by the central limit theorem  $\sqrt{2}(\sqrt{X_B} - \sqrt{B}) \Rightarrow_D \mathbf{N}(0, 1)$  as  $B \rightarrow \infty$ . Therefore, we have

$$\begin{aligned} \lim_{B \rightarrow \infty} \tilde{R}_B(0; c\sqrt{B}) &= \lim_{B \rightarrow \infty} \frac{1}{B} \mathbb{E}\{(\sqrt{X_B} - c\sqrt{B})_+^2\} \\ &= \lim_{B \rightarrow \infty} \frac{1}{B} \mathbb{E}\{(\sqrt{B} + \frac{1}{\sqrt{2B}}Z - c\sqrt{B})_+^2\} \\ &= \lim_{B \rightarrow \infty} \mathbb{E}\left\{\left(1 - c + \frac{1}{\sqrt{2B}}Z\right)_+^2\right\} \end{aligned}$$

and the latter converges to  $(1 - c)_+^2$  by dominated convergence.

## E Non-convergence of state evolution

In this appendix we show non-convergence of state evolution for firm-thresholding AMP and globally minimax AMP below  $\delta_{SE}(\varepsilon)$ . More precisely, we show that, for  $\delta < \delta_{SE}(\varepsilon)$  there exists a probability distribution  $\nu$  such that the state evolution sequence  $\{m_t\}_{t \geq 0}$ ,  $m_0 = \mathbb{E}_\nu\{X^2\}$  does not converge to 0.

We begin by developing a lower bound that holds for all the denoisers  $\eta$  studied in this paper. For notational simplicity, we consider in fact separable denoisers, but the result is easily see to hold in general, provided that the signal class is scale-invariant.

Let  $\nu_{\varepsilon, \mu}$  denote the mixture  $(1 - \varepsilon)\delta_0 + \varepsilon\delta_\mu$  where  $\mu \in \mathbf{R}_+ \cup \{\infty\}$  can be either finite or infinite. Define the risk at  $\mu$  as

$$R(\mu; \tau) = \mathbb{E}_{\nu_{\varepsilon, \mu}} \left\{ [\eta(X + Z; \tau) - X]^2 \right\}.$$

where  $X \sim \nu_{\varepsilon, \mu}$  is independent of  $Z \sim \mathbf{N}(0, 1)$ .

**Definition E.1.** We say that the risk function  $R$  is super-quadratic on  $[0, \mu_*)$  if, for any  $\mu \in [0, \mu_*)$ ,

$$R(\mu_*) \geq \left(\frac{\mu}{\mu_*}\right)^2 \cdot R(\mu_*; \tau). \quad (\text{E.1})$$

The next result shows that superquadratic behavior of the risk function implies non convergence of state evolution for signal distribution  $\nu_{\varepsilon, \mu}$ .

**Lemma E.1. (State Evolution Non-Convergence)** Fix any  $\tau \in \Theta$ , and assume there exists  $\mu_0 > 0$  such that: (i) The risk function  $R(\mu; \tau)$  is superquadratic on  $[0, \mu_*)$ , (ii)  $R(\mu_*) \geq \delta$ , and (iii)  $\delta \geq \varepsilon$ .

Let  $\Psi(m) = \Psi(m; \delta, \tau, \nu_{\varepsilon, \mu})$ . Then there is  $m_{\text{fp}} > 0$  such that

$$m \geq m_{\text{fp}} \Rightarrow \Psi(m) \geq m_{\text{fp}}.$$

In particular, if  $m_{\text{fp}} < m_0 = \mathbb{E}_{\nu_{\varepsilon, \mu}} \|X\|^2 = \varepsilon \mu^2$  then, for all  $t \geq 0$ ,

$$m_t \geq m_{\text{fp}} > 0. \quad (\text{E.2})$$

*Proof.* By the scaling relation (1.3).

$$\Psi(m) \equiv \frac{1}{\delta} \mathbb{E}_{\nu_{\varepsilon, \mu}} \left\{ \|X - \eta(X + \sqrt{m/\delta}Z; \tau, \sqrt{m/\delta})\|_2^2 \right\} = \frac{m}{\delta} R\left(\frac{\mu}{\sqrt{m}}; \tau\right).$$

Define  $m_{\text{fp}} = (\mu/\mu_*)^2$  and assume  $m \geq m_{\text{fp}}$ . This implies  $\mu/\sqrt{m} \leq \mu_*$ , and, since  $R$  is superquadratic by assumption,

$$\Psi(m) \geq \frac{m}{\delta} \left(\frac{\mu}{\sqrt{m}} \frac{1}{\mu_*}\right)^2 R(\mu_*; \tau) = \left(\frac{\mu}{\mu_*}\right)^2 = m_{\text{fp}},$$

which concludes the proof.  $\square$

For firm and minimax denoisers  $\eta \in \{\eta^{\text{firm}}, \eta^{\text{all}}\}$ , we took a fine grid of  $\varepsilon$  and at each fixed  $\varepsilon$  evaluated  $R(\mu; \tau)$  on a fine grid of  $\mu$ , checking the inequality (E.1). In the case of firm thresholding we used the minimax threshold values  $\tau = \tau^*(\varepsilon)$ . We further used the least favorable  $\mu$ ,  $\mu = \mu^*(\varepsilon)$ . Sample results are presented in Figures 23 and 24.

These computations show that the risk function  $R(\mu; \tau^*(\varepsilon))$  is superquadratic on  $(0, \mu^*(\varepsilon))$ .

## F Monotone regression

### F.1 Proof of Lemma 4.1, part (b)

The risk at  $\mu$  can be equivalently be written as  $R_N(t\mu) = \mathbb{E}\{\|v(t\mu, \mathbf{Z})\|_2^2\}$ ,  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{N \times N})$ , where  $v = v(t\mu, \mathbf{Z})$  solves the optimization problem

$$\begin{aligned} & \text{minimize} && \|v - z\|_2^2, \\ & \text{subject to} && \Delta v_i \geq -t\Delta \mu_i, \text{ for all } i \in \{1, \dots, N-1\}, \end{aligned}$$



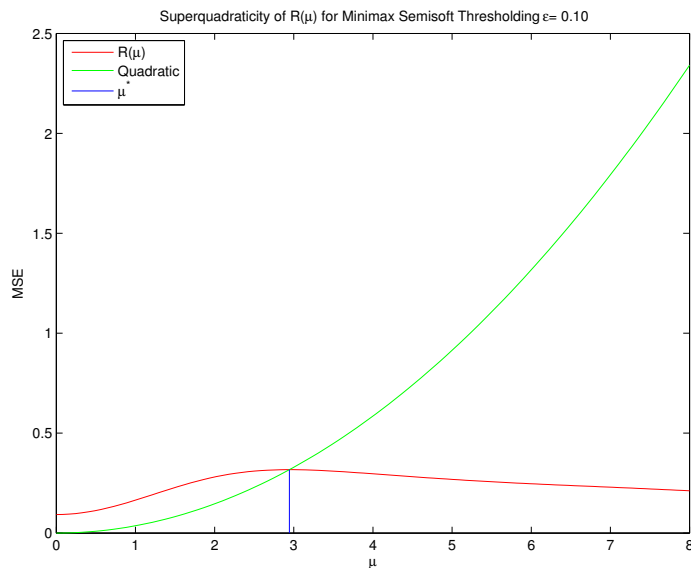


Figure 23: Numerical verification of superquadraticity for firm shrinkage. The green parabola depicts the quadratic function  $R(\mu^*; \tau^*) \cdot (\mu/\mu^*)^2$ . The vertical line depicts the position of  $\mu^*$ . The red curve depicts  $R(\mu; \tau^*)$ . The fact that the latter is above the parabola throughout the interval  $[0, \mu^*)$  verifies the superquadratic condition (E.1).

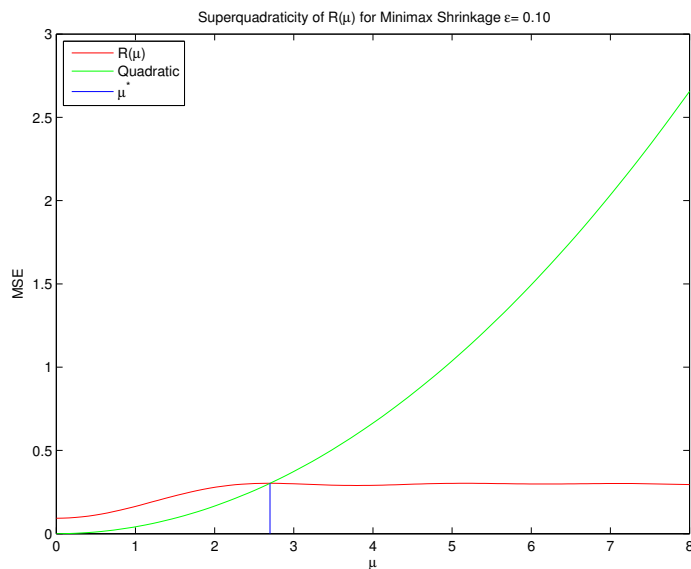


Figure 24: Numerical verification of superquadraticity for minimax shrinkage. The green parabola depicts the quadratic function  $R(\mu^*; \tau^*) \cdot (\mu/\mu^*)^2$ . The vertical line depicts the position of  $\mu^*$ . The red curve depicts  $R(\mu; \tau^*)$ . The fact that the latter is above the parabola throughout the interval  $[0, \mu^*)$  verifies the superquadratic condition (E.1).

where we recall that  $\Delta v_i = v_{i+1} - v_i$  is the discrete derivative. (Of course this problem does not provide an algorithm since it requires to know  $\Delta\mu_i$ , but here we are interested in it only for analysis purposes.)

As  $t \rightarrow \infty$ , all the constraints  $\Delta v_i \geq -t\Delta\mu_i$  for which  $\Delta\mu_i > 0$  become irrelevant. We are naturally led to defining the following *localized* monotone regression problem

$$\begin{aligned} (Q_{lmono}) \quad & \text{minimize} && \|v - z\|_2^2, \\ & \text{subject to} && \Delta v_i \geq 0, \text{ for all } i \in I_0 \equiv [N-1] \setminus I_+. \end{aligned}$$

(Here and below omit the dependence of  $I_0, I_+$  on  $\mu$ .) Let  $\eta^{lmono}(z; I_0)$  denote the solution of  $(Q_{lmono})$  with data  $z, I_0$ . The above discussion implies that, for  $z = \mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{N \times N})$ , we have the following limit in probability:

$$\lim_{t \rightarrow \infty} (\eta^{mono}(t\mu + \mathbf{Z}) - \mu) = \eta^{lmono}(\mathbf{Z}; I_0).$$

As a consequence (and using the fact that the higher moments of  $(\eta^{mono}(t\mu + \mathbf{Z}) - \mu)$  are bounded uniformly in  $t$ ) defining  $R^{loc}(I_0) = \mathbb{E}\{\|\eta^{lmono}(\mathbf{Z}; I_0)\|_2^2\}$ , we have

$$\lim_{t \rightarrow \infty} R_N(t\mu) = R^{loc}(I_0).$$

In words, the risk ‘at infinity’ of monotone regression is simply given by the local risk.

In order to conclude the proof, it is sufficient to show that  $R^{loc}(I_0)$  is given by the right-hand side of Eq. (4.4). It is easy to check that the problem  $(Q_{lmono})$  separates into independent optimization problem for each  $J_k$ . Namely, for  $i \in J_k$ ,  $v_i$  can be found by solving the following smaller problem

$$\begin{aligned} (Q_{lmono, J_k}) \quad & \text{minimize} && \|v_{J_k} - z_{J_k}\|_2^2, \\ & \text{subject to} && \Delta v_i \geq 0, \quad \text{if both } i, i+1 \in J_k, \end{aligned}$$

where  $v_{J_k} = (v_{i_k+1}, \dots, v_{i_{k+1}})$  and, if the segment  $J_k$  is a singleton, the constraint disappears. Let  $v_{J_k}(z_{J_k})$  be the solution of this problem. Then,

$$R^{loc}(I_0) = \sum_{k=0}^K \mathbb{E}\{\|v_{J_k}(\mathbf{Z}_{J_k})\|^2\}.$$

On the other hand  $(Q_{lmono, J_k})$  is simply the monotone regression problem on the segment  $J_k$  with data  $y_{J_k} = 0 + z_{J_k}$ , and hence  $\mathbb{E}\{\|v_{J_k}(\mathbf{Z}_{J_k})\|^2\} = r(|J_k|)$ , which implies immediately the desired claim.

## F.2 Proof of Lemma 4.2

Throughout the proof, we denote by  $x = x(\mathbf{Z}) = \eta^{mono}(z = \mathbf{Z})$  the solution of the monotone regression problem

$$\begin{aligned} & \text{minimize} && \|x - z\|_2^2, \\ & \text{subject to} && x_i \leq x_{i+1}, \quad \text{for all } i \in \{1, \dots, N-1\}, \end{aligned}$$

with data  $z = \mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{N \times N})$ . We then have  $r(N) = \mathbb{E}\{\|x(\mathbf{Z})\|_2^2\}$ .

Clearly  $r(1) = 1$  since in this case there is no monotonicity constraint and the solution of the regression problem is simply  $x = z$ . In order to prove Eq. (4.5), let  $k \in I_+(x)$  (i.e. an increase point:  $x_k < x_{k+1}$ ) and define  $r^{(k)} \in \mathbf{R}^N$  by letting  $r_i^{(k)} = 1_{\{i > k\}}$ , and  $l^{(k)} \in \mathbf{R}^N$  by letting  $l_i^{(k)} = 1_{\{i \leq k\}}$ . Then,  $x + \xi r^{(k)} \in \mathcal{M}_N$ ,  $x + \xi l^{(k)} \in \mathcal{M}_N$  for all  $\xi$  small enough. Hence, we must have  $\|x + \xi r^{(k)} - z\|_2^2 \geq \|x - z\|_2^2$ , and  $\|x + \xi l^{(k)} - z\|_2^2 \geq \|x - z\|_2^2$  for all  $\xi$  small enough. Expanding to linear order in  $\xi$ , we conclude that, for all  $k \in I_+(x)$ :

$$\sum_{i=1}^k x_i = \sum_{i=1}^k z_i, \quad \sum_{i=k+1}^N x_i = \sum_{i=k+1}^N z_i. \quad (\text{F.1})$$

Further, if  $r^{(0)}$  is the all 1 vector,  $x + \xi r^{(0)} \in \mathcal{M}_N$  for all  $\xi \in \mathbf{R}$ . Minimizing with respect to  $\xi$ , we get

$$\sum_{i=1}^N x_i = \sum_{i=1}^N z_i. \quad (\text{F.2})$$

Define the events (for  $k \in [N-1]$ )

$$\mathcal{E}_k \equiv \left\{ k \in I_+(x); x_k \leq -\sqrt{\frac{6 \log N}{k}} \text{ or } x_{k+1} \geq \sqrt{\frac{6 \log N}{N-k}} \right\}.$$

By virtue of Eq. (F.1), and using the fact that  $x$  is monotone, we have

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_k\} &\leq \mathbb{P}\left\{ \sum_{i=1}^k Z_i \leq -\sqrt{6k \log N} \right\} + \mathbb{P}\left\{ \sum_{i=k+1}^N Z_i \geq \sqrt{6(N-k) \log N} \right\} \\ &\leq \frac{1}{2} \exp\left\{ -\frac{6k \log N}{2k} \right\} + \frac{1}{2} \exp\left\{ -\frac{6(N-k) \log N}{2(N-k)} \right\} = \frac{1}{N^3}, \end{aligned} \quad (\text{F.3})$$

where we used the fact that, for  $Z \sim \mathbf{N}(0, 1)$  and  $a \geq 0$ ,  $\mathbb{P}\{Z \geq a\} \leq (1/2) \exp(-a^2/2)$ .

Next define

$$\mathcal{E}_\emptyset \equiv \left\{ x_1 \geq \sqrt{\frac{6 \log N}{N}} \text{ or } x_N \leq -\sqrt{\frac{6 \log N}{N}} \right\}.$$

By Eq. (F.2) and monotonicity we have

$$\mathbb{P}\{\mathcal{E}_\emptyset\} \leq \mathbb{P}\left\{ \left| \sum_{i=1}^N Z_i \right| \geq \sqrt{6N \log N} \right\} \leq \frac{1}{N^3}. \quad (\text{F.4})$$

Finally, consider the event

$$\mathcal{E} \equiv \left\{ -\sqrt{\frac{6 \log N}{i}} \leq x_i \leq \sqrt{\frac{6 \log N}{N-i+1}} \text{ for all } i \in [N] \right\}.$$

It is then easy to check that, letting  $\mathcal{E}^c$  denote the complement of  $\mathcal{E}$ ,

$$\mathcal{E}^c \subseteq \mathcal{E}_\emptyset \cup \mathcal{E}_1 \cup \dots \cup \mathcal{E}_{N-1}.$$

Indeed define, for  $i \in [N]$ ,  $k(i) \equiv \max\{k : k < i, k \in I_+(x)\}$  with, by convention,  $k(i) = 0$  if the set  $\{k < i, k \in I_+(x)\}$  is empty. Then, by definition of  $I_+(x)$ ,

$$x_i = x_{k(i)+1} \leq \sqrt{\frac{6 \log N}{N - k(i)}} \leq \sqrt{\frac{6 \log N}{N - i + 1}},$$

where the first inequality follows by definition of the events  $\mathcal{E}_\emptyset, \mathcal{E}_1, \dots, \mathcal{E}_{N-1}$  and the second since  $i \geq k(i) + 1$ . The inequality  $x_i \geq -\sqrt{(6 \log N)/i}$  follows essentially by the same argument. By union bound we therefore obtain

$$\mathbb{P}\{\mathcal{E}\} \geq 1 - \frac{1}{N^2}. \quad (\text{F.5})$$

We can therefore bound the risk as follows

$$r(N) = \mathbb{E}\{\|x(\mathbf{Z})\|_2^2\} = \mathbb{E}\{\|x(\mathbf{Z})\|_2^2; \mathcal{E}\} + \mathbb{E}\{\|x(\mathbf{Z})\|_2^2; \mathcal{E}^c\}.$$

By definition of  $\mathcal{E}$ , we have

$$\begin{aligned} \mathbb{E}\{\|x(\mathbf{Z})\|_2^2; \mathcal{E}\} &\leq \sum_{i=1}^N \max\left\{\frac{6 \log N}{i}, \frac{6 \log N}{N - i + 1}\right\} \\ &\leq 12 \log N \sum_{i=1}^{\lceil N/2 \rceil} \frac{1}{i} \leq 12 \log N (\log N + 1). \end{aligned} \quad (\text{F.6})$$

On the other hand it is easy to see that, defining  $Z_{\max} = \max_{i \in [N]} |Z_i|$ , we necessarily have  $|x_i(\mathbf{Z})| \leq Z_{\max}$  for all  $i \in [N]$ , whence

$$\mathbb{E}\{\|x(\mathbf{Z})\|_2^2; \mathcal{E}^c\} \leq N \mathbb{E}\{Z_{\max}^2; \mathcal{E}^c\} \leq N \mathbb{E}\{Z_{\max}^4\}^{1/2} \mathbb{P}\{\mathcal{E}^c\}^{1/2} \leq \mathbb{E}\{Z_{\max}^4\}^{1/2}. \quad (\text{F.7})$$

Here the second inequality follows from Cauchy-Schwarz and the last one from Eq. (F.5). By union bound,  $\mathbb{P}\{Z_{\max} \geq z\} \leq N \mathbb{P}\{|Z_1| \geq z\} \leq N \exp\{-z^2/2\}$ . Hence

$$\begin{aligned} \mathbb{E}\{Z_{\max}^4\} &= 4 \int_0^\infty z^3 \mathbb{P}\{Z_{\max} \geq z\} dz \\ &\leq 4 \int_0^{\sqrt{2 \log N}} z^3 dz + 4N \int_{\sqrt{2 \log N}}^\infty z^3 e^{-z^2/2} dz \\ &\leq (2 \log N)^2 + 4 \int_0^\infty (\sqrt{2 \log N} + u)^3 e^{-\sqrt{2 \log N} u} u du \\ &\leq (2 \log N)^2 + 4(2 \log N) + 20 \leq (4 \log N)^2, \end{aligned}$$

where the last inequality holds for  $N \geq 10$ . Using this bound in conjunction with Eq. (F.6), we finally get

$$r(N) \leq 12(\log N)^2 + 16 \log N \leq 20(\log N)^2.$$

Finally notice that for arbitrary  $N$  the simple bound  $r(N) \leq N \mathbb{E}\{Z_{\max}^2\} \leq 2N[\log N + 1]$  can be proved by controlling  $Z_{\max}$  along the same lines as above.

## G Further computational details

We record here a few details that have been omitted from the main text.

**Estimate of the effective noise level.** The AMP iteration, cf. Eqs. (1.8), (1.8), (1.9), requires to estimate the variance  $\sigma_t^2$  of the effective observation at time  $t$ . According to the the general theory of state evolution [DMM09, BM11a], the empirical distribution of the coordinates of  $z^t$  is asymptotically Gaussian with mean 0 and variance  $\sigma_t^2$ . This motivates the following estimator, first proposed in [DMM09]:

$$\hat{\sigma}^t = \frac{1}{\Phi^{-1}(0.75)} \text{median}\{|z^t|\},$$

where  $\Phi(z)$  is the normal distribution function. This is known as the 25% pseudo-variance in robust estimator and has the advantage of being insensitive to a small fraction of large outliers.

**Computation environment.** Computations were done in partly using Matlab, and partly through a Java program . In the spirit of reproducible research, a suite of Java classes that allow to repeat our simulations is available through an open code repository [DJM12].

**Numerical computation of minimax risk.** The plots of minimax risk were obtained by evaluating numerically the expression in the main text. For separable and block-separable denoisers (with the exception of the global minimax denoiser  $\eta^{all}$ ), the integrals can be expressed in terms of the Gaussian distribution function or incomplete beta functions. For the global minimax denoiser, integration was performed numerically using the standard MATLAB routines.

Evaluation of the minimax risk required searching the least favorable distribution among two point mixtures of the form  $(1 - \varepsilon)\delta_0 + \delta_\mu$ , in the separable case and  $(-\varepsilon)\delta_0 + \varepsilon\delta_{\mu e_1}$  in the block separable one. Optimization over  $\mu \in \mathbf{R}_+$  was performed by brute force search over a grid, with recursive refinement of the grid.

For the non-separable cases, the procedure for approximating the minimax MSE was explained Section 4 and 5.

## H Finite- $N$ scaling and error analysis

The empirical phase transitions observed in this paper admit further analysis, to verify whether the following expected behavior take place, namely: (a) the offsets tend towards zero with increasing  $N$ ; (b) the steepness of the phase transition increases with increasing  $N$ .

### H.1 Offsets decay toward zero

As described in Section 2.4, at each fixed value  $\varepsilon$  of the sparsity parameter, we gathered data at several different values of  $\delta$ , and obtained the empirical phase transition parameter  $\widehat{\text{PT}}(N, \varepsilon, \eta)$ , recorded as *offset* from prediction, so that  $\widehat{\text{PT}}(N, \varepsilon, \eta) = 0$  means that the 50% success location fitted to the  $\varepsilon$ -fixed,  $\delta$ -varying dataset is exactly at the predicted location  $M(\varepsilon|\eta)$ . Our analysis gave not only the empirical phase transition location, but also its formal standard error  $SE(\widehat{\text{PT}})$ . (Here we make explicit the dependence of  $\widehat{\text{PT}}$  on the specific denoiser.)

$\gamma$	$R^2$
1/3	0.99521
1/2	0.98724
2/3	0.9666811
3/4	0.950506
1	0.893454

Table 4: Powers  $\gamma$  and resulting  $R^2$  for fitting model (H.1) to the empirical offsets of various phase transitions.

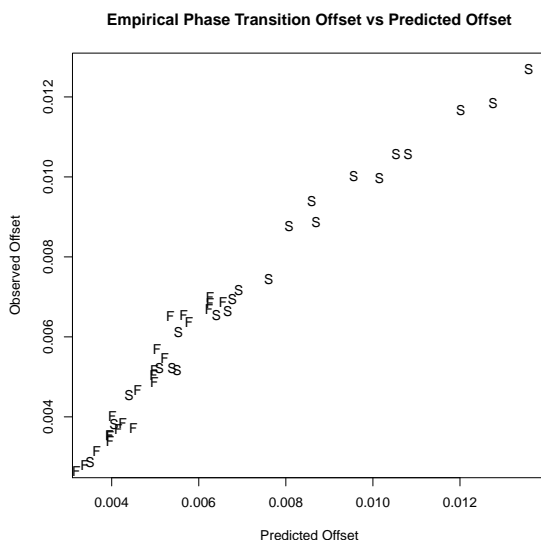


Figure 25: Fitted offsets versus predicted offsets in model (H.1) for various values of  $\varepsilon$  and soft and firm thresholding (respectively S and F).

We fit a linear model to the dataset including all the phase transition results for soft and firm thresholding. We considered several exponents  $\gamma$  that might be describing the decay of the offset with increasing  $N$ :

$$\widehat{\text{PT}}(N, \varepsilon, \eta) = c(\varepsilon, \eta)N^{-\gamma} + SE(\widehat{\text{PT}}) \cdot N(0, 1), \quad (\text{H.1})$$

Table 4 shows that  $\gamma = 1/3$  provides an adequate description of the offsets, with an  $R^2$  exceeding 0.995. A plot of raw  $\widehat{\text{PT}}$ 's versus the predictions of model (H.1) is given in Figure 25.

## H.2 Transitions sharpen

In addition to an empirical phase transition parameter  $\widehat{\text{PT}}(N, \varepsilon, \eta)$  we also fitted an empirical steepness parameter  $\widehat{\beta}(N, \varepsilon, \eta)$ , according to the logistic model:

$$\log \frac{\widehat{p}_i}{1 - \widehat{p}_i} = \beta(\delta_i - \widehat{\text{PT}}),$$

$\gamma$	$R^2$
1/3	0.99286
1/2	0.99927
2/3	0.991064
3/4	0.982350
1	0.947641

Table 5: Powers  $\gamma$  and resulting  $R^2$  for fitting model (H.2) to the empirical slope parameters.

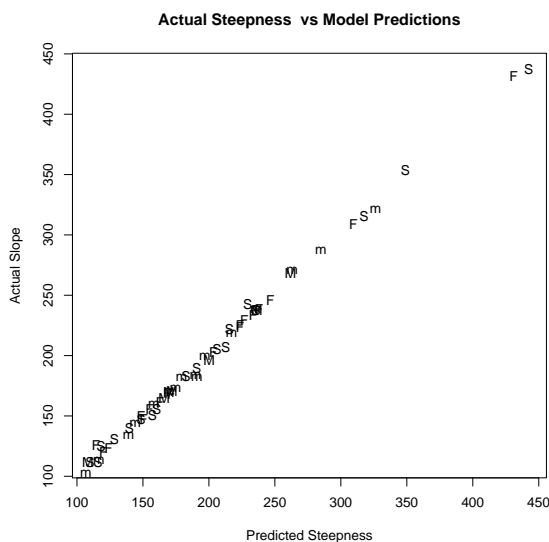


Figure 26: Fitted steepness versus predicted steepness in model (H.2) for soft (S), firm (F) and minimax (m) denoisers.

where  $\hat{p}_i$  is the empirical success probability for the  $i$ -th experiment.

We expect  $\hat{\beta}(N, \varepsilon, \eta)$  to grow with increasing  $N$ , corresponding to increasingly abrupt transitions from complete failure at  $\delta \ll \widehat{\text{PT}}(N, \varepsilon, \eta)$  to complete success at  $\delta \gg \widehat{\text{PT}}(N, \varepsilon, \eta)$ . In order to test this behavior, we fitted a linear model to the values of  $\hat{\beta}$  computed for multiple values of  $N$ ,  $\varepsilon$ , and denoisers  $\eta$ . We considered a range of powers  $\tilde{\gamma}$  that might be describing the growth of the steepness with increasing  $N$ :

$$\hat{\beta}(N, \varepsilon, \eta) = c(\varepsilon, \eta) N^{\tilde{\gamma}} + \text{Error}. \quad (\text{H.2})$$

Table 5 shows that  $\gamma = 1/2$  provides an adequate description of the steepnesses, with an  $R^2$  exceeding 0.999. A plot of raw  $\hat{\beta}$ 's versus the predictions of model (H.2) is given in Figure 26.

## I Further properties of the risk function

In this appendix we prove several useful properties of the risk function of the block soft and James-Stein denoisers. Throughout this section, we define the risk at  $\mu$  as

$$R(\mu; \eta) = \mathbb{E}\{\|\eta(\mu + \mathbf{Z}) - \mu\|_2^2\}. \quad (\text{I.1})$$

The argument  $\eta$  will be dropped or replaced by the threshold level  $\tau$  whenever clear from the context. Since we only consider denoisers that are equivariant under rotation,  $R(\mu; \eta)$  depends on the vector  $\mu$  only through its norm  $\|\mu\|_2$ . With a slight abuse of notation, we will use  $\mu$  to denote the norm as well. In other words, the reader can assume  $\mu = \mu e_1$ .

### I.1 Block soft thresholding

In this section we consider the block soft denoiser  $\eta^{soft}(\cdot; \tau) : \mathbf{R}^B \rightarrow \mathbf{R}^B$ . We will write  $R(\mu; \tau)$  for  $R(\mu; \eta^{soft}(\cdot; \tau))$ .

**Lemma I.1.** *For block soft thresholding the risk function  $\mu \mapsto R(\mu; \tau)$  has these properties:*

$$\mu \rightarrow R(\mu; \tau) \text{ is monotone increasing,} \quad (\text{I.2})$$

$$\frac{\partial}{\partial \mu^2} R(\mu; \mu) \leq 1, \quad (\text{I.3})$$

$$R(\mu; \tau) \leq \min\{R(0; \tau) + \mu; B + \tau^2\}, \quad (\text{I.4})$$

$$\lim_{\mu \rightarrow \infty} R(\mu; \tau) = B + \tau^2. \quad (\text{I.5})$$

*Proof.* It will be convenient within the proof to set  $d \equiv B$  and  $\xi \equiv \mu^2$ . Let  $S^2 = \|\mu + \mathbf{Z}\|_2^2$ , so that  $S^2$  is distributed as a non-central chi-square  $\chi_d^2(\xi)$  with, in particular,  $\mathbb{E}S^2 = d + \xi$ . Since the block soft thresholding rule is weakly differentiable, Stein's unbiased estimate of risk yields  $R(\mu; \tau) = \mathbb{E}U(S)$ , with

$$U(S) = \begin{cases} S^2 - d & \text{for } S < \tau, \\ d + \tau^2 - 2(d-1)\tau S^{-1} & \text{for } S \geq \tau. \end{cases}$$

Let  $f_{\xi, d}(w)$  be the density function of  $S^2 \sim \chi_d^2(\xi)$ . This satisfies

$$\frac{\partial}{\partial \xi} f_{\xi, d}(w) = -\frac{\partial}{\partial w} f_{\xi, d+2}(w) = \frac{1}{2} [f_{\xi, d+2}(w) - f_{\xi, d}(w)].$$

Applying the first identity, integrating by parts, canceling terms, and then using the second identity, we obtain

$$\frac{\partial}{\partial \mu^2} R(\mu^2; \tau) = \int_0^{\tau^2} f_{\xi, d}(w) dw + (d-1)\tau \int_{\tau^2}^{\infty} w^{-3/2} f_{\xi, d+2}(w) dw \geq 0. \quad (\text{I.6})$$

For the upper bound (I.3), use the Poisson mixture representation  $f_{\xi, d+2}(w) = \sum_{j=0}^{\infty} p_{\xi/2}(j) f_{d+2+2j}(w)$  with  $p_{\lambda}(j) \equiv \lambda^j e^{-\lambda} / j!$  and an identity for the central  $\chi^2$  density family to obtain

$$\frac{f_{d+2+2j}(w)}{w} = \frac{f_{d+2j}(w)}{d+2j} \leq \frac{f_{d+2j}(w)}{d}$$



and so to conclude that  $f_{\xi,d+2}(w) \leq (w/d)f_{\xi,d}(w)$ . Hence the second term in (I.6) is bounded by  $(1 - 1/\tau) \int_{\tau^2}^{\infty} f_{\xi,d}(w) dw$ , whence follows the conclusion  $(\partial R/\partial \mu^2) \leq 1$ . Property (I.4) is immediate from (I.2) and (I.3) and the large- $\mu$  limit of  $R$ .

To obtain the bound in (I.5), write the risk function using the unbiased risk formula as

$$R(\mu) = \xi + \mathbb{E}[2d + \tau^2 - S^2 - 2(d-1)\tau S^{-1}, S > \tau]$$

and observe that the integrand is decreasing in  $S$ , and bounded above by 2 for  $S > \tau$ , so that  $R(0; \tau) \leq 2\mathbb{P}(\chi_d^2 \geq \tau^2)$ . This yields (I.5).  $\square$

## I.2 Positive-part James-Stein denoiser

As in the previous section, we set  $\xi \equiv \mu^2$  and  $d = B$ . Here we consider the James-Stein denoiser  $\eta^{JS} : \mathbf{R}^d \rightarrow \mathbf{R}^d$  defined by

$$\eta^{JS}(y) = \left(1 - \frac{(d-2)}{\|y\|^2}\right)_+ y, \quad (\text{I.7})$$

and we will write  $R(\mu) = R(\mu; \eta^{JS})$ .

We again let  $S^2 = \|\mu + \mathbf{Z}\|_2^2$ , with  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{d \times d})$ . We have the noncentral chi-squared distribution  $S \sim \chi_d^2(\xi)$  with noncentrality  $\xi$ . Stein's unbiased estimate of risk is

$$U(S) = \begin{cases} S^2 - d & \text{for } S^2 < d - 2, \\ d - (d-2)^2/S^2 & \text{for } S^2 > d - 2, \end{cases}$$

and  $R(\mu) = \mathbb{E}U(S)$ .

### I.2.1 Risk at 0

We will first develop an approximation of the risk at 0 that was used in Sections (3.2) and (3.3).

Let  $f_d(w)$  denote the density of a central chi-squared with  $d$  degrees of freedom. We then have the density satisfies

$$\begin{aligned} R(0) &= \mathbb{E}\left\{\left\|\left(1 - \frac{d-2}{S^2}\right)_+ \mathbf{Z}\right\|^2\right\} \\ &= \int_{d-2}^{\infty} \left[w - 2(d-2) - \frac{(d-2)^2}{w}\right] f_d(w) dw. \end{aligned}$$

Using the identity  $wf_{d-2}(w) = (d-2)f_d(w)$  and letting  $D = d-2$ , we can rewrite the last expression as

$$R(0) = \frac{1}{(d-2)} \int_{d-2}^{\infty} [w - (d-2)]^2 f_{d-2}(w) dw = D^{-1} \mathbb{E}(\chi_D^2 - D)_+^2. \quad (\text{I.8})$$

We have the convergence in distribution,

$$D^{-1/2}(\chi_D^2 - D) \Rightarrow \mathbf{N}(0, 1), \quad D \rightarrow \infty.$$

By a standard tightness argument, this implies that that

$$\lim_{d \rightarrow \infty} R(0) = 2\mathbb{E}Z_+^2 = 1,$$

where  $Z \sim \mathbf{N}(0, 1)$ .

An Edgeworth series leads to the expansion  $R(0) = 1 + R_1 d^{-1/2} + \Theta(d^{-2})$ . Indeed, one can integrate the expression (I.8) numerically, and the numerical values are consistent with  $R(0) \approx 1 + 0.752/\sqrt{d}$  for large  $d$ .

### I.2.2 Monotonicity of $R(\mu)$ .

We use the variation-diminishing version of total positivity, developed in [BJM81].

**Theorem I.1** (Brown, Johnstone, and MacGibbon). *The non-central  $\chi^2$  family is strictly variation diminishing of all orders*

For  $g : [0, \infty) \rightarrow \mathbf{R}$ , let  $S^-(g)$  and  $S^+(g)$  denote the number of sign changes and strict sign changes of  $g$ , and let  $IS(g)$  denote the sign of  $g(0)$  (assuming that  $g(0) \neq 0$ , the more general definition being given in [BJM81]). Further define the function  $\gamma : [0, \infty) \mapsto \mathbf{R}$  by

$$\gamma(\xi) = \int_0^\infty g(w) f_{d,\xi}(w) dw, \quad (\text{I.9})$$

where  $f_{d,\xi}(\cdot)$  is the density of the noncentral chi-square with  $d$  degrees of freedom and noncentrality  $\xi$ . By the SVR property we have that that  $S^+(\gamma) \leq S^-(g)$  and that if  $S^+(\gamma) = S^-(g)$  then necessarily  $IS(\gamma) = IS(g)$ . In particular this implies that, if  $g$  is strictly increasing, then  $\gamma$  is strictly increasing as well. Indeed this follows by letting  $g_a(w) = g(w) - a$  for  $a \in \mathbf{R}$  and

$$\gamma_a(\xi) \equiv \int_0^\infty g_a(w) f_{d,\xi}(w) dw = \gamma(\theta) - a.$$

If  $g$  is strictly increasing, then  $S^-(g_a) \leq 1$  for all  $a \in \mathbf{R}$ , whence  $S^+(\gamma_a) \leq 1$  for all  $a$ , with  $IS(\gamma) = IS(g)$  whenever  $S^+(\gamma_a) = 1$ . This in turns implies that  $\gamma$  is increasing.

We now verify that the risk  $R(\mu)$  of  $\eta^{JS}$  is monotone increasing in  $\xi = \|\mu\|^2 \in [0, \infty)$ . Let

$$g(w) = U(w^{1/2}) = \begin{cases} w - d & \text{for } w < d - 2, \\ d - (d - 2)^2/w & \text{for } w > d - 2, \end{cases}$$

and define  $\gamma(\xi)$  using Eq. (I.9). Note that  $g$  is strictly increasing and hence  $\xi \mapsto \gamma(\xi)$  is increasing as well by the above argument. But  $U(S)$  is Stein's unbiased risk estimator and hence  $R(\mu) = \gamma(\xi = \|m\mu\|^2)$ , which implies the claim.

### I.3 Proof of Lemma 6.2

Since any probability distribution is written as a convex combination of point masses, it is sufficient to prove the claim for  $\nu = \delta_\mu$ . In this case, using the scaling relation (1.3), we have

$$\Psi(m; \delta, \tau, \delta_\mu) = \frac{m}{\delta} R(\mu\sqrt{\delta/m}), \quad (\text{I.10})$$

with  $R(\cdot)$  the risk function. Therefore the state evolution mapping is starshaped for all distributions  $\nu$  if and only if  $\mu \mapsto R(\mu)$  is monotone increasing.

The monotonicity of the risk function was proved in [DMM09] for soft thresholding and positive soft thresholding. It is proved in Section 4 and 5 for monotone regression and total variation denoising. Finally, it is proved in Section I.1 and I.2 for block soft and James-Stein denoisers.

## References

- [BC83] P. Bickel and J. Collins, *Minimizing fisher information over mixtures of distributions*, Sankhya A **45** (1983), 1–19.
- [BC90] M.J. Best and N. Chakravarti, *Active set algorithms for isotonic regression; a unifying framework*, Mathematical Programming **47** (1990), 425–439.
- [BCDH10] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, *Model-Based Compressive Sensing*, IEEE Trans. on Inform. Theory **56** (2010), 1982–2001.
- [BCT11] J. D. Blanchard, C. Cartis, and J. Tanner, *Compressed sensing: How sharp is the restricted isometry property?*, SIAM Review **53** (2011), no. 1, 105–125.
- [BGI<sup>+</sup>08] R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss, *Combining geometry and combinatorics: A unified approach to sparse signal recovery*, 47th Annual Allerton Conference (Monticello, IL), September 2008, pp. 798 – 805.
- [Bic81] P. Bickel, *Minimax estimation of the mean of a normal distribution when the parameter space is restricted*, Annals of Statistics **9** (1981), 1301–1309.
- [BJM81] L.D. Brown., I.M. Johnstone, and K.B. MacGibbon, *Variation diminishing transformations: A direct approach to total positivity and its statistical applications*, Journal of the American Statistical Association **76** (1981), no. 376, 824–832.
- [BLM12] M. Bayati, M. Lelarge, and A. Montanari, *Universality in Polytope Phase Transitions and Message Passing Algorithms*, arXiv:1207.7321, 2012.
- [BM11a] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory **57** (2011), 764–785.
- [BM11b] ———, *The LASSO risk for gaussian matrices*, IEEE Trans. on Inform. Theory **58** (2011), 1997–2017.
- [BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Amer. J. of Mathematics **37** (2009), 1705–1732.
- [BT10] J. Blanchard and J. Tanner, *Large scale iterative hard thresholding on a graphical processing unit*, ICNAAM 2010: International Conference of Numerical Analysis and Applied Mathematics **1281** (2010), 1730–1734.
- [Cai99] T.T. Cai, *Adaptive wavelet estimation: A block thresholding and oracle inequality approach*, The Annals of Statistics **27** (1999), no. 3, pp. 898–924.

- [Can06] E. Candès, *Compressive sampling*, Proceedings of the International Congress of Mathematicians (Madrid, Spain), 2006.
- [CHDB08] V. Cevher, C. Hegde, M. F. Duarte, and R. G. Baraniuk, *Sparse Signal Recovery Using Markov Random Fields*, Neural Information Processing Systems (Vancouver), December 2008.
- [CICB10] V. Cevher, P. Indyk, L. Carin, and R.G. Baraniuk, *Sparse Signal Recovery and Acquisition with Graphical Models*, IEEE Signal Processing Magazine **27** (2010), 92–103.
- [CRPW11] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A.S. Willsky, *The convex geometry of linear inverse problems*, arXiv:1012.0621, 2011.
- [CS81] G. Casella and W. E. Strawderman, *Estimating a bounded normal mean*, Annals of Statistics **9** (1981), 870–878.
- [CT05] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. on Inform. Theory **51** (2005), 4203–4215.
- [CY08] R. Chartrand and W. Yin, *Iteratively reweighted algorithms for compressive sensing*, ICASSP, IEEE, 2008, pp. 3869–3872.
- [DDS92] J.C. Hoch D.L. Donoho, I.M. Johnstone and A.S. Stern, *Maximum entropy and the nearly black object (with discussion)*, J. Roy. Statist. Soc. B **54** (1992), 41–81.
- [DJ94] D. L. Donoho and I. M. Johnstone, *Minimax risk over  $l_p$  balls*, Prob. Th. and Rel. Fields **99** (1994), 277–303.
- [DJ95] D.L. Donoho and I.M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assn. **90** (1995), 1200–1224.
- [DJM11] D. Donoho, A. Javanmard, and A. Montanari, *Information-Theoretically optimal compressed sensing via spatial coupling and approximate message passing*, arXiv:1112.0708, 2011.
- [DJM12] D.L. Donoho, I. Johnstone, and A. Montanari, *Amp public svn repository*, <https://subversion.assembla.com/svn/amp-public/>, November 2012.
- [DM08] A. Dembo and A. Montanari, *Finite size scaling for the core of large random hypergraphs*, Ann. Appl. Prob. **18** (2008), 1993–2040.
- [DMM09] D. L. Donoho, A. Maleki, and A. Montanari, *Message passing algorithms for compressed sensing*, Proceedings of the National Academy of Sciences **106** (2009), 18914–18919.
- [DMM10] ———, *Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction*, Proceedings of IEEE Inform. Theory Workshop (Cairo), 2010.
- [DMM11] D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, IEEE Trans. on Inform. Theory **57** (2011), 6920–6941.

- [Don06] D. L. Donoho, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete Comput. Geometry **35** (2006), 617–652.
- [DT05] D. L. Donoho and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proceedings of the National Academy of Sciences **102** (2005), no. 27, 9452–9457.
- [DT09a] ———, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, Journal of American Mathematical Society **22** (2009), 1–53.
- [DT09b] D.L. Donoho and J. Tanner, *Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing*, Phil. Trans. R. Soc. A **367** (2009), 4273–4293.
- [DT10a] D. L. Donoho and J. Tanner, *Counting the faces of randomly-projected hypercubes and orthants, with applications*, Discrete & Computational Geometry **43** (2010), no. 3, 522–541.
- [DT10b] D. L. Donoho and J. Tanner, *Precise undersampling theorems*, Proceedings of the IEEE **98** (2010), 913–924.
- [FR08] M. Fornasier and H. Rauhut, *Iterative thresholding algorithms*, Appl. and Comp. Harmonic Analysis **25** (2008), 187–208.
- [GB97] H.-Y. Gao and A.G. Bruce, *Waveshrink with firm shrinkage*, Statistica Sinica **7** (1997), 855–874.
- [HKP98] P. Hall, G. Kerkycharian, and D. Picard, *Block threshold rules for curve estimation using kernel and wavelet methods*, Ann. Statist. **26** (1998), 922–942.
- [HR09] P.J. Huber and E. Ronchetti, *Robust statistics (second edition)*, J. Wiley and Sons, 2009.
- [Hub64] P.J. Huber, *Robust estimation of a location parameter*, The Annals of Mathematical Statistics **35** (1964), no. 1, 73–101.
- [JM12] A. Javanmard and A. Montanari, *State Evolution for Vector Approximate Message Passing*, In preparation, 2012.
- [Joh94a] I. M. Johnstone, *Minimax bayes, asymptotic minimax and sparse wavelet priors*, Statistical Decision Theory and Related Topics, V (S.S. Gupta and eds. J.O. Berger, eds.), Springer-Verlag, 1994, p. 303326.
- [Joh94b] I.M. Johnstone, *On minimax estimation of a sparse normal mean vector*, Annals of Statistics **22** (1994), 271289.
- [Joh12] ———, *Gaussian estimation: Sequence and wavelet models*, 2012, <http://www-stat.stanford.edu/~imj/GE12-27-11.pdf>.
- [JS10] W. James and C. Stein, *Estimation with quadratic loss*, Proc. Fourth Berkeley Symp. Math. Statist. Prob (Berkeley), 2010.

- [KGR11] U.S. Kamilov, V.K. Goyal, and S. Rangan, *Message-Passing Estimation from Quantized Samples*, arXiv:1105.6368, 2011.
- [KMS<sup>+</sup>12] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborova, *Statistical physics-based reconstruction in compressed sensing*, Phys.Rev X **2** (2012), 021005.
- [KWT09] Y. Kabashima, T. Wadayama, and T. Tanaka, *A typical reconstruction limit for compressed sensing based on lp-norm minimization*, J.Stat. Mech. (2009), L09003.
- [KXAH10] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi, *Improved sparse recovery thresholds with two-step reweighted  $\ell_1$  minimization*, IEEE Intl. Symp. on Inform. Theory (Austin, TX), June 2010.
- [Mal78] C. L. Mallows, *Minimizing an integral*, SIAM Review **20** (1978), 183.
- [MAYB11] A. Maleki, L. Anitori, A. Yang, and R. Baraniuk, *Asymptotic Analysis of Complex LASSO via Complex Approximate Message Passing (CAMP)*, arXiv:1108.0477, 2011.
- [MD10] A. Maleki and D. L. Donoho, *Optimally tuned iterative reconstruction algorithms for compressed sensing*, IEEE Journal of Selected Topics in Signal Processing **4** (2010), 330–341.
- [MM09] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford University Press, Oxford, 2009.
- [Mon12] A. Montanari, *Graphical models concepts in compressed sensing*, Compressed Sensing (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012.
- [Ran11] S. Rangan, *Generalized Approximate Message Passing for Estimation with Random Linear Mixing*, IEEE Intl. Symp. on Inform. Theory (St. Petersburg), August 2011.
- [RFG09] S. Rangan, A. K. Fletcher, and V. K. Goyal, *Asymptotic analysis of map estimation via the replica method and applications to compressed sensing*, Neural Information Processing Systems (NIPS) (Vancouver), 2009.
- [ROF92] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), 259–268.
- [RRN11] N. Rao, B. Recht, and R. Nowak, *Tight measurement bounds for exact recovery of structured sparse signals*, arXiv:1106.4355, 2011.
- [RU08] T.J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, 2008.
- [Sch10] P. Schniter, *Turbo Reconstruction of Structured Sparse Signals*, Proceedings of the Conference on Information Sciences and Systems (Princeton), 2010.
- [Sch11] ———, *A message-passing receiver for BICM-OFDM over unknown clustered-sparse channels*, arXiv:1101.4724, 2011.

- [SPS10] S. Som, L. C. Potter, and P. Schniter, *Compressive Imaging using Approximate Message Passing and a Markov-Tree Prior*, Proc. Asilomar Conf. on Signals, Systems, and Computers, November 2010.
- [SSS10] L.C. Potter S. Som and P. Schniter, *On Approximate Message Passing for Reconstruction of Non-Uniformly Sparse Signals*, Proceedings of the National Aerospace and Electronics Conference (Dayton, OH), 2010.
- [Sto10] M. Stojnic, *Block-length dependent thresholds for 2/1-optimization in block-sparse compressed sensing*, ICASSP, IEEE, 2010, pp. 3918–3921.
- [TSR<sup>+</sup>05] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, J. R. Statist. Soc. B **67** (205), 91–108.
- [VO96] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput **17** (1996), 227–238.
- [Wai09] M.J. Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (lasso)*, IEEE Trans. on Inform. Theory **55** (2009), 2183–2202.
- [WV10] Y. Wu and S. Verdú, *Rényi Information Dimension: Fundamental Limits of Almost Lossless Analog Compression*, IEEE Trans. on Inform. Theory **56** (2010), 3721–3748.
- [XH10] W. Xu and B. Hassibi, *Compressive sensing over the grassmann manifold: a unified geometric framework*, arXiv:1005.3729, 2010.
- [YL10] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society B **68** (2010), 49–67.