

# High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing

David Donoho \* and Andrea Montanari †

November 14, 2013

## Abstract

In a recent article (*Proc. Natl. Acad. Sci.*, 110(36), 14557-14562), El Karoui et al. study the distribution of robust regression estimators in the regime in which the number of parameters  $p$  is of the same order as the number of samples  $n$ . Using numerical simulations and ‘highly plausible’ heuristic arguments, they unveil a striking new phenomenon. Namely, the regression coefficients contain an extra Gaussian noise component that is not explained by classical concepts such as the Fisher information matrix.

We show here that that this phenomenon can be characterized rigorously techniques that were developed by the authors for analyzing the Lasso estimator under high-dimensional asymptotics. We introduce an *approximate message passing* (AMP) algorithm to compute M-estimators and deploy *state evolution* to evaluate the operating characteristics of AMP and so also M-estimates. Our analysis clarifies that the ‘extra Gaussian noise’ encountered in this problem is fundamentally similar to phenomena already studied for regularized least squares in the setting  $n < p$ .

## 1 M-Estimation under high dimensional asymptotics

Consider the traditional linear regression model

$$Y = \mathbf{X} \theta_0 + W, \quad (1)$$

with  $Y = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$  a vector of responses,  $\mathbf{X} \in \mathbb{R}^{n \times p}$  a known design matrix,  $\theta_0 \in \mathbb{R}^p$  a vector of parameters, and  $W \in \mathbb{R}^n$  random noise having zero-mean components  $W = (W_1, \dots, W_n)^\top$  i.i.d. with distribution  $F = F_W$  having finite second moment <sup>1</sup>.

We are interested in estimating  $\theta_0$  from observed data<sup>2</sup>  $(Y, \mathbf{X})$  using a traditional M-estimator, defined by a non-negative convex function  $\rho : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ :

$$\hat{\theta}(Y; \mathbf{X}) \equiv \arg \min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta; Y, \mathbf{X}), \quad \mathcal{L}(\theta; Y, \mathbf{X}) \equiv \sum_{i=1}^n \rho(Y_i - \langle X_i, \theta \rangle), \quad (2)$$

---

\*Department of Statistics, Stanford University

†Department of Electrical Engineering and Department of Statistics, Stanford University

<sup>1</sup>With a slight abuse of notation, we shall use  $W$  to denote a random variable with the same distribution  $F_W$ .

<sup>2</sup>We denote by  $X_1, \dots, X_n$  the rows of  $\mathbf{X}$ . We often omit the arguments  $Y, \mathbf{X}$  as this dependency will hold throughout. Without loss of generality, we assume that the columns of  $\mathbf{X}$  are normalized so that  $\|\mathbf{X} e_i\|_2 \approx 1$ . (A more precise assumption will be formulated below.)

where  $\langle u, v \rangle = \sum_{i=1}^m u_i v_i$  is the standard scalar product in  $\mathbb{R}^m$ , and  $\hat{\theta}$  is chosen arbitrarily if there is multiple minimizers.

Although this is a completely traditional problem, we consider it under *high-dimensional asymptotics* where the number of parameters  $p$  and the number of observations  $n$  are both tending to infinity, at the same rate. This is becoming a popular asymptotic model owing to the modern awareness of ‘big data’ and ‘data deluge’; but also because it leads to entirely new phenomena.

### 1.1 Extra Gaussian noise due to high-dimensional asymptotics

Classical statistical theory considered the situation where the number of regression parameters  $p$  is fixed and the number of samples  $n$  is tending to infinity. The asymptotic distribution was found by Huber [Hub73, Bic75] to be normal  $\mathbf{N}(0, \mathbf{V})$  where the asymptotic variance matrix  $\mathbf{V}$  is given by

$$\mathbf{V} = V(\psi, F_W)(\mathbf{X}^\top \mathbf{X})^{-1} \tag{3}$$

here  $\psi = \rho'$  is the score function of the M-estimator and  $V(\psi, F) = (\int \psi^2 dF) / (\int \psi' dF)^2$  the asymptotic variance functional of [Hub64], and  $(\mathbf{X}^\top \mathbf{X})$  the usual Gram matrix associated with the least-squares problem. Importantly, it was found that for efficient estimation – i.e. the smallest possible asymptotic variance – the optimal M-estimator depended on the probability distribution  $F_W$  of the errors  $W$ . Choosing  $\psi(x) = (\log f_W(x))'$  (with  $f_W$  the density of  $W$ ), the asymptotic variance functional yields  $V(\psi, F_W) = 1/I(F_W)$ , with  $I(F)$  denoting the Fisher information. This achieves the fundamental limit on the accuracy of M-estimators [Hub73].

In modern statistical practice there is increasing interest in applications where the number of explanatory variables  $p$  is very large, and comparable to  $n$ . Examples of this new regime can be given, spanning bioinformatics, machine learning, imaging, and signal processing (a few research areas in the last domains include [LDSP08, Sca97, Ric05, Cha03]).

This paper considers the properties of M-estimators in the high-dimensional asymptotic  $n \rightarrow \infty$ ,  $n/p(n) \rightarrow \delta \in (1, \infty)$  In this regime, the asymptotic distribution of M-estimators no longer needs to obey the classical formula (3) in widespread use. We make a random-design assumption on the  $\mathbf{X}$ ’s detailed below. We show that the asymptotic covariance matrix of the parameters is now of the form

$$\mathbf{V} = V(\tilde{\Psi}, \tilde{F}_W)(\mathbb{E}\{\mathbf{X}^\top \mathbf{X}\})^{-1}, \tag{4}$$

where  $V$  is still Huber’s asymptotic variance functional, but  $\tilde{\Psi}$  is the *effective score function*, which is different from  $\psi$  under high-dimensional asymptotics and  $\tilde{F}_W$  is the *effective error distribution*, which is different from  $F_W$  under high-dimensional asymptotics. In the limit  $\delta \rightarrow \infty$ , the effective score and the effective error distribution both tend to their classical counterparts, and one recovers  $V(\psi, F_W)$ .

The effective error distribution  $\tilde{F}_W$  is a convolution of the noise distribution with an extra Gaussian noise component, not seen in the classical setting (here  $\star$  denotes convolution):

$$\tilde{F}_W \equiv F_W \star \mathbf{N}(0, \tau_\star^2(\psi, F_W, \delta)). \tag{5}$$

The extra Gaussian noise depends in a complex way on  $\psi$ ,  $F_W$ ,  $\delta$ , which we characterize fully below in Corollary 4.2.

Several important insights follow immediately:

1. Existing formulas are inadequate for confidence statements about M-estimates under high dimensional asymptotics, and will need to be systematically broadened.
2. Classical maximum likelihood estimates are inefficient under high-dimensional asymptotics. The idea dominating theoretical statistics since R.A. Fisher to use  $\psi = (-\log f_W)'$  as a scoring rule, does not yield the efficient estimator.
3. The usual Fisher Information bound is not necessarily attainable in the high-dimensional asymptotic, as  $I(\tilde{F}_W) < I(F_W)$ .

M-estimation in this high-dimensional asymptotic setting was considered in a recent article by El Karoui, Bean, Bickel, Lim, and Yu [EKBBL13], who studied the distribution of  $\hat{\theta}$  for Gaussian design matrices  $\mathbf{X}$ . In short they observed empirically the basic phenomenon of extra Gaussian noise appearing in high-dimensional asymptotics and rendering classical inference incorrect. The dependence of the additional variance  $\tau_*^2$  on  $\delta$ ,  $\psi$  and  $F$  was characterized by [EKBBL13] through a non-rigorous heuristics<sup>3</sup> that the authors describe as ‘highly plausible and buttressed by simulations.’<sup>4</sup> (We refer to Section 5 for further discussion of related work.)

## 1.2 Proof Strategy: Approximate Message Passing

In the present paper, we show that this important statistical phenomenon can be *characterized rigorously*, in a way that we think fully explains the main new concepts of extra Gaussian noise, effective noise and the effective score. Our proof strategy has three steps

- Introduce an *Approximate Message Passing* (AMP) algorithm for M-estimation; an iterative procedure with the M-estimator as a fixed point, and having the effective score function  $\tilde{\Psi}$  as its score function at algorithm convergence.
- Introduce *State Evolution* for calculating properties of the AMP algorithm iteration by iteration. We show that these calculations are exact at each iteration in the large- $n$  limit where we freeze the iteration number and let  $n \rightarrow \infty$ .

At the center of the State Evolution calculation is precisely an extra Gaussian noise term that is tracked from iteration to iteration, and which is shown to converge to a nonzero noise level. In this way, State Evolution makes very explicit that AMP faces at each iteration and even in the limit, an effective noise that differs from the noise  $W$  by addition of an appreciable extra independent Gaussian noise.

- Show that the AMP algorithm converges to the solution of the M-estimation problem in mean square, from which it follows that the asymptotic variance of the M-estimator is identical to the asymptotic variance of the AMP algorithm. More specifically, the asymptotic variance of the M-estimator is given by a formula involving the effective score function and the effective noise.

---

<sup>3</sup>To the reader familiar with the mathematical theory of spin glasses, the argument of [EKBBL13] appears analogous to the cavity method from statistical physics [MPV87, MM09, Tal10]

<sup>4</sup>After the first version of our manuscript was posted on ArXiv, Nouredine El Karoui announced an independent proof of related results, using a completely different approach.

As it turns out, our formula for the asymptotic variance coincides with the one derived heuristically in [EKBBL13, Corollary 1] although our technique is remarkably different, and our proof provides a very clear understanding of the operational significance of the terms appearing in the asymptotic variance. It also allows explicit calculation of many other operating characteristics of the M-estimator, for example when used as an outlier detector<sup>5</sup>.

### 1.3 Underlying tools

At the heart of our analysis, we are simply applying an approach developed in [BM11, BM12] for rigorous analysis of solutions to convex optimization problems under high-dimensional asymptotics.

That approach grew out of a series of earlier papers studying the compressed sensing problem [DMM09, DMM11, DJMM11, BM12]. From the perspective of this paper, those papers considered the same regression model (1) as here; however, they emphasized the challenging asymptotic regime where there are fewer observations than predictors, (i.e.  $n/p(n) \rightarrow \delta \in (0, 1)$ ) so that even in the noiseless case, the equations  $Y = \mathbf{X}\theta$  would be underdetermined. In the  $p > n$  setting, it became popular to use  $\ell_1$ -penalized least squares (Lasso, [Tib96, CD95]). That series of papers considered the Lasso convex optimization problem in the case of  $\mathbf{X}$  with iid  $\mathcal{N}(0, 1/n)$  entries (just as here) and followed the same 3-step strategy we use here; namely, 1. Introducing an AMP algorithm; 2. Obtaining the asymptotic distribution of AMP by State Evolution; and 3. Showing that AMP agrees with the Lasso solution in the large- $n$  limit. This procedure proved that the Lasso solution has the asymptotic distribution

$$\hat{\theta}^u \sim \mathbf{N}(\theta_0, (\sigma^2 + \tau_{\text{Lasso}}^2)\mathbf{I}_{p \times p}) \quad (6)$$

where  $\sigma^2$  is the variance of the noise in the measurements, and  $\tau_{\text{Lasso}}^2$  is the variance of an extra Gaussian noise, not appearing in the classical setting where  $p(n)/n \rightarrow 0$ . The variance of this extra Gaussian noise was obtained by state evolution and shown to depend on the distribution of the coefficients being recovered, and on the noise level in a seemingly complicated way that can be characterized by a fixed-point relation, see [DMM11, BM12]. At the center of the rigorous analysis stand the papers [BM11, BM12] which analyze recurrences of the type used by AMP and establish the validity of State Evolution in considerable generality. Those same papers stand at the center of our analysis in this paper.

Apart from allowing a simple treatment, this provides a unified understanding of the phenomenon of high-dimensional extra Gaussian noise.

### 1.4 The role of AMP

This paper introduces a new first-order algorithm for computing the M-estimator  $\hat{\theta}$  which is uniquely appropriate for the random-design case. This algorithm fits within the class of approximate message passing (AMP) algorithms introduced in [DMM09, BM11] (see also [Ran11] for extensions). This algorithm is of independent interest because of its low computational complexity.

AMP has a deceptive simplicity. As an iterative procedure for convex optimization, it looks almost the same as the ‘standard’ application of simple fixed-stepsize gradient descent. However, it is

---

<sup>5</sup>The slightly more general [EKBBL13, Result 1] covers heteroscedastic noise is not covered by the analysis of this paper, but should be provable by adapting our argument.

intended for use in the random-design setting, and it has an extra memory term (aka reaction term) that modifies the iteration in a profound and beneficial way. In the Lasso setting, AMP algorithms have been shown to have remarkable fast convergence properties [DMM09], far outperforming more complex-looking iterations like Nesterov and FISTA.

In the present paper, AMP has a second important wrinkle – it solves a convex optimization problem associated to minimizing  $\rho$  with iterations based on gradient descent with an objective  $\rho_{b_t}$  which varies from one iteration to the next, as  $b_t$  changes, but which does not tend to  $\rho$  in the limit.

In the present paper, AMP is mainly used as a proof device, one component of the three-part strategy outlined earlier. However, a key benefit produced by the curious features of AMP is strong heuristic insight, which would not be available for a ‘standard’ gradient-descent algorithm.

The AMP proof strategy makes visible the extra Gaussian noise appearing in the M-estimator  $\hat{\theta}$ . Elementary considerations show that such extra noise is present at iteration zero of AMP. State Evolution faithfully tracks the dynamics of this extra noise across iterations. State Evolution proves that the extra noise level does not go to zero asymptotically with increasing iterations, but instead that the extra noise level tends to a fixed nonzero value. Because AMP is solving the M-estimation problem, the M-estimator must be infected by this extra noise.

The AMP algorithm and its State Evolution analysis shows that the extra noise in parameter  $\hat{\theta}_i^t$  at iteration  $t$  is due to cross-parameter estimation noise leakage, where errors in the estimation of all other parameters at the previous iteration ( $t - 1$ ) cause extra noise to appear in  $\hat{\theta}_i^t$ . In the classical setting no such effect is visible. One could say that the central fact about the high-dimensional setting revealed here as well as in our earlier work [DMM09, DMM11, DJMM11, BM12], is that when there are so many parameters to estimate, one cannot really insulate the estimation of any one parameter from the errors in estimation of all the other parameters.

## 2 Approximate Message Passing (AMP)

### 2.1 A family of score functions

For the rest of the paper, we make the following smoothness assumption on  $\rho$ :

**Definition 2.1.** *We call the loss function  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  smooth if it is continuously differentiable, with absolutely continuous derivative  $\psi = \rho'$  having an a.e. derivative  $\psi'$  that is bounded:  $\sup_{u \in \mathbb{R}} \psi'(u) < \infty$ .*

Our assumption excludes some interesting cases, such as  $\rho(u) = |u|$ , but includes for instance the Huber loss <sup>6</sup>

$$\rho_{\text{H}}(z; \lambda) = \begin{cases} z^2/2 & \text{if } |z| \leq \lambda, \\ \lambda|z| - \lambda^2/2 & \text{otherwise.} \end{cases} \quad (7)$$

Associated to  $\rho$ , we introduce the family  $\rho_b$  of regularizations of  $\rho$ :

$$\rho_b(z) \equiv \min_{x \in \mathbb{R}} \left\{ b\rho(x) + \frac{1}{2}(x - z)^2 \right\}, \quad (8)$$

---

<sup>6</sup>We expect that the proof technique developed in this paper should be generalizable to a broader class of functions  $\rho$ , at the cost of additional technical complications.

in words, this is the min-convolution of the original loss with a square loss. Each  $\rho_b$  has a corresponding score function

$$\Psi(z; b) = \rho'_b(z).$$

The effective score of the M-estimator belongs to this family, for a particular choice of  $b$ , explained below.

In the classical M-estimation literature [HR09], monotonicity and differentiability of the score function  $\psi$  is frequently useful; our assumptions on  $\rho$  guarantee these properties for the nominal score function  $\psi$ . The score family  $\Psi(\cdot; b)$  has such properties as well: for any  $b$ ,  $\Psi(\cdot; b)$  is a strictly monotone increasing function; second, for any  $b > 0$ ,  $\Psi(\cdot; b)$  is a contraction. With  $\Psi'$  denoting differentiation with respect to the first variable, we have  $\Psi'(z; b) \in (0, 1)$ . For proof and further discussion, see Appendix A.

Before proceeding, we give an example. Consider the Huber loss  $\rho_H(z; \lambda)$ , with score function  $\psi(z; \lambda) = \min(\max(-\lambda, z), \lambda)$ . We have

$$\Psi(z; b) = b\psi\left(\frac{z}{1+b}; \lambda\right).$$

In particular the shape of each  $\Psi$  is similar to  $\psi$ , but the slope of the central part is now  $\|\Psi'(\cdot; b)\|_\infty = \frac{b}{1+b} < 1$ .

## 2.2 AMP algorithm

Our proposed approximate message passing (AMP) algorithm for the optimization problem (2) is iterative, starting at iteration 0 with an initial estimate  $\hat{\theta}^0 \in \mathbb{R}^p$ . At iteration  $t = 0, 1, 2, \dots$  it applies a simple procedure to update its estimate  $\hat{\theta}^t \in \mathbb{R}^p$ , producing  $\hat{\theta}^{t+1}$ . The procedure involves three steps.

**Adjusted residuals.** Using the current estimate  $\hat{\theta}^t$ , we compute the vector of *adjusted residuals*  $R^t \in \mathbb{R}^n$ ,

$$R^t = Y - \mathbf{X}\hat{\theta}^t + \Psi(R^{t-1}; b_{t-1}); \tag{9}$$

where to the ordinary residuals  $Y - \mathbf{X}\hat{\theta}^t$  we here add the extra term<sup>7</sup>  $\Psi(R^{t-1}; b_{t-1})$ .

**Effective Score.** We choose a scalar  $b_t > 0$ , so that the effective score  $\Psi(\cdot; b_t)$  has empirical average slope  $p/n \in (0, 1)$ . Setting  $\delta = \delta(n) = n/p > 1$ , we take any solution<sup>8</sup> (for instance the smallest solution) to<sup>9</sup>:

$$\frac{1}{\delta} = \frac{1}{n} \sum_{i=1}^n \Psi'(R_i^t; b). \tag{10}$$

<sup>7</sup>Here and below, given  $f : \mathbb{R} \rightarrow \mathbb{R}$  and  $v = (v_1, \dots, v_m)^\top \in \mathbb{R}^m$ , we define  $f(v) \in \mathbb{R}^m$  by applying  $f$  coordinate-wise to  $v$ , i.e.  $f(v) \equiv (f(v_1), \dots, f(v_m))^\top$ .

<sup>8</sup>This equation always admits at least one solution since  $b \mapsto \Psi'(r; b)$  is continuous in  $b \geq 0$ , with  $\Psi'(r; 0) = 0$  and (for  $\rho$  strictly convex)  $\Psi'(r; \infty) = 1$ , cf. Proposition A.1.

<sup>9</sup>Under this prescription, the sequence  $b_t$  depends on the instance  $(Y, \mathbf{X})$ . As explained in the next section, for the proof of our main result we will use a slightly different prescription, that is independent of the problem instance.

**Scoring.** We apply the effective score function  $\Psi(R^t; b_t)$ :

$$\hat{\theta}^{t+1} = \hat{\theta}^t + \delta \mathbf{X}^\top \Psi(R^t; b_t). \quad (11)$$

The Scoring step of the AMP iteration (11) is similar to traditional iterative methods for M-estimation, compare [Bic75]. Indeed, using the traditional residual  $z^t = Y - \mathbf{X}\theta^t$ , the traditional method of scoring at iteration  $t$  would read

$$\hat{\theta}^{t+1} = \hat{\theta}^t + \frac{1}{\frac{1}{n} \sum_{i=1}^n \psi'(z_i^t)} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \psi(z^t), \quad (12)$$

and one can see correspondences of individual terms to the method of scoring used in AMP. Of course the traditional term  $[\sum_{i=1}^n \psi'(z_i^t)/n]^{-1}$  corresponds to AMP's  $[\sum_{i=1}^n \Psi'(R_i^t; b_t)/n]^{-1} \equiv \delta$  (because of step (10)), while the traditional term  $(\mathbf{X}^\top \mathbf{X})^{-1}$  corresponds to AMP's implicit  $\mathbf{I}_{p \times p}$  – which is appropriate in the present context because our random-design assumption below makes  $\mathbf{X}^\top \mathbf{X}$  behave approximately like the identity matrix.

### 2.3 Relation to M-estimation

The next lemma explains the reason for using the effective score  $\Psi(\cdot; b_t)$  in the AMP algorithm: this is what connects the AMP iteration to M-estimation (2).

**Lemma 2.2.** *Let  $(\hat{\theta}_*, R_*, b_*)$  be a fixed point of the AMP iteration (9), (10), (11) having  $b_* > 0$ . Then  $\hat{\theta}_*$  is a minimizer of the problem (2). Viceversa, any minimizer  $\hat{\theta}_*$  of the problem (2) corresponds to one (or more) AMP fixed points of the form  $(\hat{\theta}_*, R_*, b_*)$ .*

*Proof.* By differentiating Eq. (2), and omitting the arguments  $Y, \mathbf{X}$  for simplicity from  $\mathcal{L}(\theta; Y, \mathbf{X})$ , we get

$$\nabla_\theta \mathcal{L}(\theta) = - \sum_{i=1}^n \rho'(Y_i - \langle X_i, \theta \rangle) X_i = -\mathbf{X}^\top \rho'(Y - \mathbf{X}\theta), \quad (13)$$

where as usual  $\rho'$  is applied component-wise to vector arguments. The minimizers of  $\mathcal{L}(\theta)$  are all the vectors  $\theta$  for which the right hand side vanishes.

Consider then a fixed point  $(\hat{\theta}_*, R_*, b_*)$ , of the AMP iteration (9), (11). This satisfies the equations

$$R_* = Y - \mathbf{X}\hat{\theta}_* + \Psi(R_*; b_*), \quad (14)$$

$$0 = \delta \mathbf{X}^\top \Psi(R_*; b_*). \quad (15)$$

The first equation can be written as

$$Y - \mathbf{X}\hat{\theta}_* = R_* - \Psi(R_*; b_*), \quad (16)$$

Using Proposition A.2 below, (16) implies that  $\Psi(R_*; b_*) = b_* \rho'(Y - \mathbf{X}\hat{\theta}_*)$ . Hence the second equation reads

$$0 = \delta b_* \mathbf{X}^\top \rho'(Y - \mathbf{X}\hat{\theta}_*), \quad (17)$$

which coincides with the stationarity condition (13) for  $b_* > 0$ . This concludes the proof.  $\square$

## 2.4 Example

To make the AMP algorithm concrete, we consider an example with  $n = 1000$ ,  $p = 200$ , so  $\delta = 5$ . For design matrix we let  $X_{i,j} \sim \mathbf{N}(0, \frac{1}{n})$ , and we draw  $\theta_0$  a random vector of norm  $\|\theta_0\|_2 = 6\sqrt{p}$ . For the distribution  $F = F_W$  of errors, we use Huber's contaminated normal distribution  $\text{CN}(0.05, 10)$ , so that  $F = 0.95\Phi + 0.05H_{10}$ , where  $H_x$  denotes a unit atom at  $x$ . For the loss function, we use the Huber's  $\rho_H(z; \lambda)$  with  $\lambda = 3$ . Starting the AMP algorithm with  $\hat{\theta}^0 = 0$ , we run 20 iterations.

Separately, we solved the M-estimation problem using CVX, obtaining  $\hat{\theta}$ .

Figure 1 (left panel) shows the progress of the AMP algorithm across iterations, presenting

$$\text{RMSE}(\hat{\theta}^t; \theta_0) \equiv \frac{1}{\sqrt{p}} \|\hat{\theta}^t - \theta_0\|_2,$$

while Figure 1 (right panel) shows the progress of AMP in approaching the M-estimate  $\hat{\theta}$ , as measured by

$$\text{RMSE}(\hat{\theta}^t; \hat{\theta}) \equiv \frac{1}{\sqrt{p}} \|\hat{\theta}^t - \hat{\theta}\|_2.$$

As is evident, the iterations converge rapidly, and they converge to the M-estimator, both in the sense of convergence of risks - measured here by  $\text{RMSE}(\hat{\theta}^t; \theta_0) \rightarrow \text{RMSE}(\hat{\theta}; \theta_0) \approx 1.6182$  - and, more directly, in convergence of the estimates themselves:  $\text{RMSE}(\hat{\theta}^t; \hat{\theta}) \rightarrow 0$ .

Figure 2 (left panel) shows the process by which the effective score parameter  $\hat{b}_t$  is obtained at iteration  $t = 3$ , while the right panel shows how  $\hat{b}_t$  behaves across iterations. In fact it converges quickly towards a limit  $b_\infty \approx 0.2710$ .

## 2.5 Contrast to iterative M-estimation

Earlier we pointed to resemblances between AMP (11) and the traditional method of scoring for obtaining M-estimators (12). In reality the two approaches are very different:

- The precise form of various terms in (9), (10) (11) is dictated by the statistical assumptions that we are making on the design  $\mathbf{X}$ . In particular the memory terms are crucial for the state evolution analysis to hold. Several papers document this point [Mon12, Sch10, SSS10, Ran11, KMZ13].
- Under classical asymptotics, where  $p$  is fixed and  $n \rightarrow \infty$ , it is sufficient to run a single step of such an algorithm [Bic75], in the high-dimensional setting it is necessary to iterate numerous times. The resulting analysis is considerably more complex because of correlations arising as the algorithm evolves.

## 3 State evolution description of AMP

State Evolution is a method for computing the operating characteristics of the AMP iterates  $\hat{\theta}^t$  and  $R^t$  for arbitrary fixed  $t$ , under the high-dimensional asymptotic limit  $n, p \rightarrow \infty$ ,  $n/p \rightarrow \delta$ .

In this section we initially describe a purely formal procedure which *assumes* that the AMP adjusted residuals  $R^t = Y - \mathbf{X}\hat{\theta}^t + \Psi(R^t; b_t)$  really behave as  $W + \tau_t Z$ , with  $W$  the error distribution and  $Z$  an independent standard normal, for  $t = 0, 1, 2, \dots$ . The variable  $\tau_t^2$  thus quantifies the extra Gaussian noise supposedly present in the adjusted residuals of AMP; we show how this ansatz allows

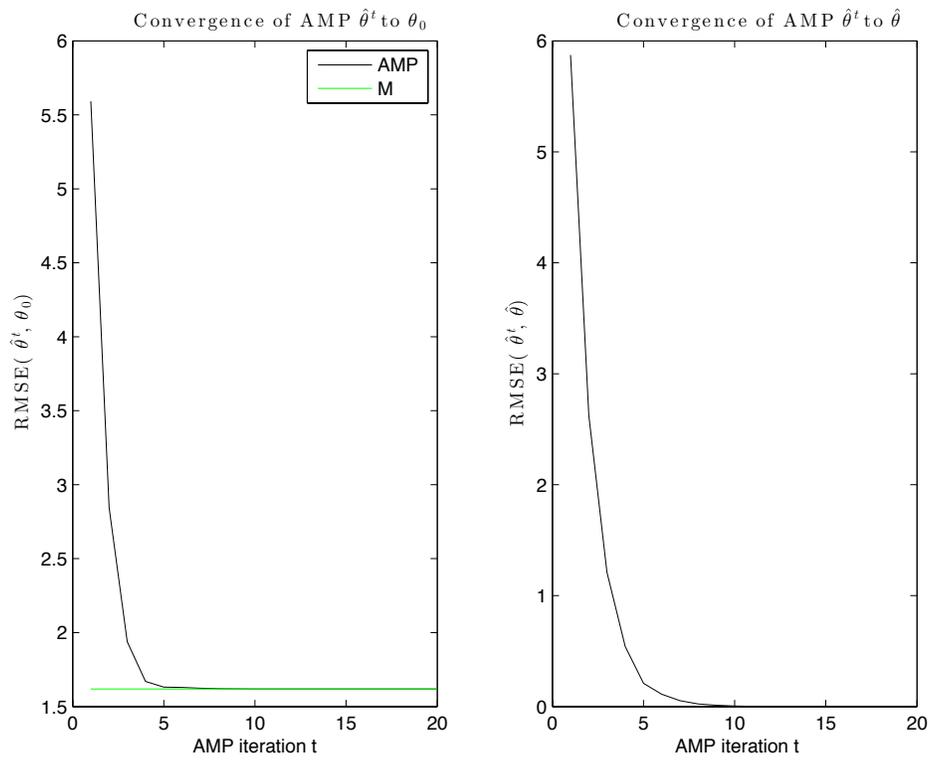


Figure 1: Left Panel: RMSE of AMP versus iteration (black curve), and its convergence to RMSE of M-estimation (constant green curve). Right Panel: Discrepancy of AMP from M-estimate, versus iteration.

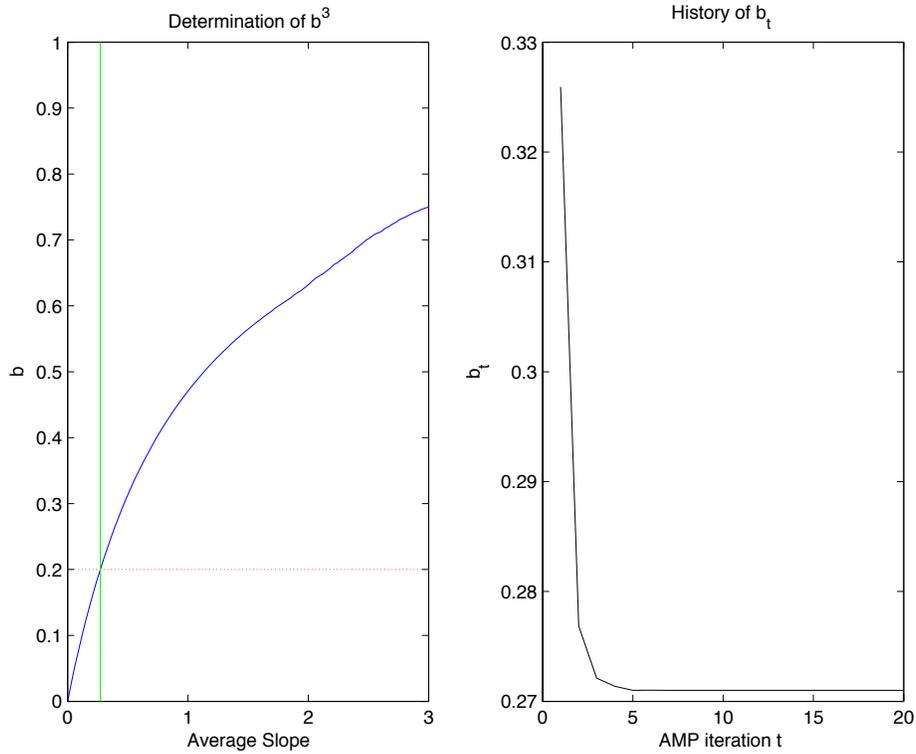


Figure 2: Left Panel: Determining the regularization parameter at iteration 3. Blue curve: Average slope (vertical) versus regularization parameter  $b$  (horizontal). The blue curve intersects desired level  $0.2 = 1/\delta$  near 0.3. Right Panel: regularization parameter  $b_t$  versus iteration; it converges rapidly to roughly 0.2710.

one to calculate  $\tau_t^2$  for each  $t = 0, 1, 2, 3, \dots$ , and to calculate the limit of  $\tau_t$  as  $t \rightarrow \infty$ . Later in the section we present a rigorous result validating the method under the following random Gaussian design assumption.

**Definition 3.1.** *We say that a sequence of random design matrices  $\{\mathbf{X}(n)\}_n$ , with  $n \rightarrow \infty$  is a Gaussian design if each  $\mathbf{X} = \mathbf{X}(n)$  has dimensions  $n \times p$ , and entries  $(X_{ij})_{i \in [n], j \in [p]}$  that are i.i.d.  $\mathbf{N}(0, 1/n)$ . Further,  $p = p(n)$  is such that  $\lim_{n \rightarrow \infty} n/p(n) = \delta \in (0, \infty)$ .*

### 3.1 Initialization of the extra variance

Under the Gaussian design assumption, suppose that  $\mathbf{u}$  is a vector in  $\mathbb{R}^p$  with norm  $\|\mathbf{u}\|_2$ . Then  $\{\mathbb{E}\|\mathbf{X}\mathbf{u}\|_2^2\} = \|\mathbf{u}\|_2^2$ . Moreover,  $\mathbf{X}\mathbf{u}$  is a Gaussian random vector with entries iid  $\mathbf{N}(0, \|\mathbf{u}\|_2^2/n)$ .

It will be convenient to introduce for any estimator  $\tilde{\theta}$  the notation

$$\text{MSE}(\tilde{\theta}, \theta_0) = \frac{1}{p} m \|\tilde{\theta} - \theta_0\|_2^2. \quad (18)$$

So initialize AMP with a deterministic estimate  $\hat{\theta}^0$ , and take  $R^{-1} = 0$ . Then the initial residual is  $R^1 = Y - \mathbf{X}\hat{\theta}^0 = W + \mathbf{X}(\theta_0 - \hat{\theta}^0)$ . The terms  $W$  and  $\mathbf{X}(\theta_0 - \hat{\theta}^0)$  are independent, and  $\mathbf{X}(\theta_0 - \hat{\theta}^0)$  is Gaussian with variance  $\tau_0^2 = \|\hat{\theta}^0 - \theta_0\|_2^2/n = \text{MSE}(\hat{\theta}^0, \theta_0)/\delta$ . Consider some fixed coordinate  $R^1(i)$  of  $R^1$ . Then

$$\text{Var}(R_i^1) = \text{Var}(W) + \text{Var}(\mathbf{X}(\theta_0 - \hat{\theta}^0)) = \text{Var}(W) + \text{MSE}(\theta^0, \theta_0)/\delta.$$

Hence, when AMP is started this way, we see that the adjusted residuals initially contain an extra Gaussian noise of variance  $\tau_0^2 = \text{MSE}(\hat{\theta}^0, \theta_0)/\delta$ .

### 3.2 Evolution of the extra Gaussian variance to its ultimate limit

Assuming the adjusted residuals continue, at later iterations, to behave as  $W + \tau_t Z$  with  $Z$  an independent standard normal, we now calculate  $\tau_t^2$  for each  $t = 1, 2, 3, \dots$ , and eventually identify the limit of  $\tau_t$  as  $t \rightarrow \infty$ .

For a given  $\tau > 0$ ,  $\delta = n/p$  and noise distribution  $F_W$ , define the *variance map*

$$\mathcal{V}(\tau^2, b; \delta, F_W) = \delta \mathbb{E}\left\{\Psi(W + \tau Z; b)^2\right\},$$

where  $W \sim F_W$ , and, independently,  $Z \sim \mathbf{N}(0, 1)$ . In this display, the reader can see that extra Gaussian noise of variance  $\tau^2$  is being added to the underlying noise  $W$ , and  $\mathcal{V}$  measures the  $\delta$ -scaled variance of the resulting output. Evidently for  $b > 0$ ,  $0 \leq \mathcal{V}(\tau^2, b) \cdot \delta \leq (\text{Var}(W) + \tau^2) \cdot \delta$ .

Under our assumptions for  $\Psi$ , for each given specification  $(\tau; \delta, F_W)$  of the ingredients besides  $b$  that go into  $\mathcal{V}$ , there is (as clarified by Lemma A.3) a well-defined value  $b = b(\tau; \delta, F_W)$  giving the smallest solution  $b \geq 0$  to

$$\frac{1}{\delta} = \mathbb{E}\left\{\Psi'(W + \tau \cdot Z; b)\right\}. \quad (19)$$

**Definition 3.2.** *State Evolution is an iterative process for computing the scalars  $\{\tau_t^2\}_{t \geq 0}$ , starting from an initial condition  $\tau_0^2 \in \mathbb{R}_{\geq 0}$  following*

$$\tau_{t+1}^2 = \mathcal{V}(\tau_t^2, b(\tau_t)) = \mathcal{V}(\tau_t, b(\tau_t; \delta, F_W); \delta, F_W). \quad (20)$$

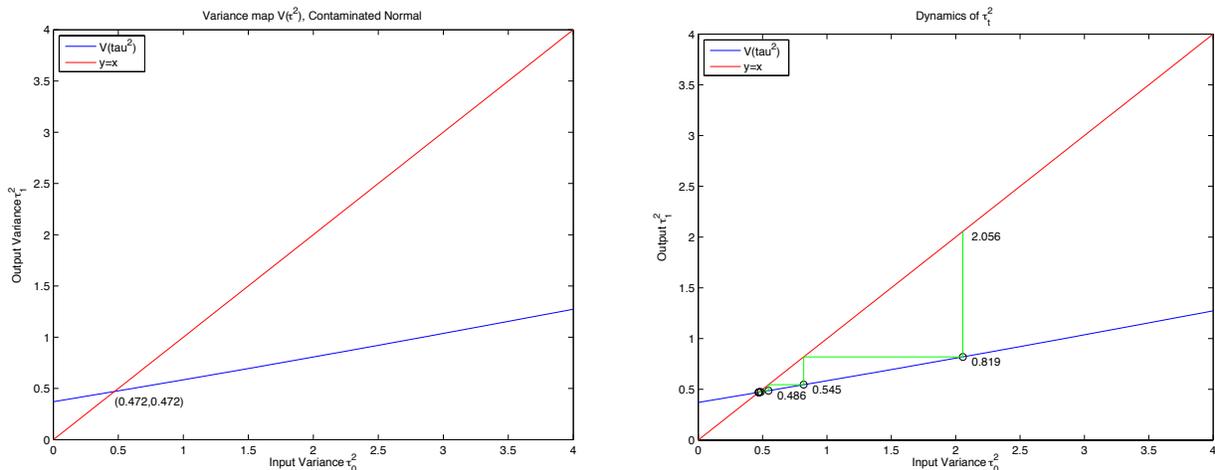


Figure 3: The State Evolution Variance Mapping. Left Panel: Blue Curve:  $\tilde{\mathcal{V}}$  versus  $\tau^2$ , Red Curve: diagonal; unique fixed point at about 0.472. Right Panel: the iteration history of state evolution, starting from  $\tau_0^2 = 2.0556$

Defining  $\tilde{\mathcal{V}}(\tau^2) = \mathcal{V}(\tau^2, b(\tau))$ , we see that the evolution of  $\tau_t$  follows the iterations of the map  $\tilde{\mathcal{V}}$ . In particular, we make these observations:

- $\tilde{\mathcal{V}}(0) > 0$ ,
- $\tilde{\mathcal{V}}(\tau^2)$  is a continuous, nondecreasing function of  $\tau$ .
- $\tilde{\mathcal{V}}(\tau^2) < \tau^2$  as  $\tau \rightarrow \infty$ .

Figure 3, left panel, considers the case where  $W$  again follows the Huber’s contaminated normal distribution  $\text{CN}(0.05, 10)$  and  $\psi$  is the standard Huber estimator with parameter  $\lambda = 3$ . The ratio  $n/p = \delta = 2$ , and the parameter vector has  $\|\theta_0\|_2^2/p = 6^2$ . It displays the function  $\tilde{\mathcal{V}}(\tau^2)$  as a function of  $\tau$ .

Evidently, there is a stable fixed point  $\tau_* = \tau_*(\delta, F_W)$ , i.e. a point obeying  $\tilde{\mathcal{V}}(\tau_*^2) = \tau_*^2$ , such that  $\tau^2 \mapsto \tilde{\mathcal{V}}(\tau^2)$  has a derivative less than 1 at  $\tau_*^2$ . We conclude that  $\tau_t$  evolves under state evolution to a nonzero limit. Figure 3, right panel, shows how  $\tau_t^2$  evolves to the fixed point near 0.472 starting from  $\tau_0^2 = 2.056$ .

### 3.3 Predicting operating characteristics from State Evolution

State Evolution offers a formal<sup>10</sup> procedure for predicting operating characteristics of the AMP iteration at any fixed iteration  $t$  or in the limit  $t \rightarrow \infty$ . Nater in this section, we will provide rigorous validation of these predictions.

---

<sup>10</sup>By *formal*, we mean a rule-based procedure which we can follow to get a prediction, without any guarantees that the prediction is correct.

Call the tuple  $S = (\tau; b, \delta, F)$  a *state*; in running the AMP algorithm we assume that the algorithm is initialized with  $\widehat{\theta}^0$  so that  $\tau_0^2 = \text{MSE}(\widehat{\theta}^0, \theta_0)/\delta$ , so that AMP starts in state  $S = (\tau_0; b^0, \delta, F)$ , and visits  $S_1 = (\tau_1; b^1, \delta, F)$ ,  $S_2 = (\tau_2; b^2, \delta, F)$ ,  $\dots$ ; eventually AMP visits states arbitrarily close to the equilibrium state  $S_* = (\tau_*; b^*, \delta, F)$ .

SE predictions of operating characteristics are provided by two rules assigning predictions to certain classes of observables, based on the state that AMP is in.

**Definition 3.3.** *The state evolution formalism assigns predictions  $\mathcal{E}$  to two types of observables under specific states.*

**Observables Involving  $\widehat{\theta} - \theta_0$ .** *Given a univariate test function  $\xi : \mathbb{R} \mapsto \mathbb{R}$ , assign the predicted value for  $p^{-1} \sum_{i \in p} \xi(\widehat{\theta}_i - \theta_{0,i})$  under state  $S$  by the rule*

$$\mathcal{E}(\xi(\widehat{\theta} - \vartheta)|S) \equiv \mathbb{E}\left\{\xi(\sqrt{\delta} \tau Z)\right\},$$

where expectation on the right hand side is with respect to  $Z \sim \mathbf{N}(0, 1)$ .

**Observables involving Residual, Error.** *Let  $R$  denote some coordinate of the adjusted residual for AMP in state  $S$  and  $W$  the same coordinate of the underlying error. Given a bivariate test function  $\xi_2 : \mathbb{R}^2 \mapsto \mathbb{R}$ , assign the prediction of  $n^{-1} \sum_{i=1}^n \xi_2(R_i, W_i)$  in state  $S$  by*

$$\mathcal{E}(\xi_2(R, W)|S) \equiv \mathbb{E}\xi_2(W + \tau Z, W)$$

where  $Z \sim \mathbf{N}(0, 1)$  and  $W \sim F_W$  is independent of  $Z$ .

The two most important predictions of operating characteristics are undoubtedly:

- **MSE at iteration  $t$ .** We let  $S_t = (\tau_t, b(\tau_t), \delta, F_W)$  denote the state of AMP at iteration  $t$ , and predict

$$\text{MSE}(\widehat{\theta}^t, \theta_0) \approx \mathcal{E}((\widehat{\vartheta} - \vartheta)^2|S_t) = \mathbb{E}\left\{(\sqrt{\delta} \tau_t Z)^2\right\} = \delta \tau_t^2.$$

- **MSE at convergence.** With  $\tau_* > 0$  the limit of  $\tau_t$ , let  $S_* = (\tau_*, b(\tau_*), \delta, F_W)$  denote the state of AMP at convergence. and predict

$$\text{MSE}(\widehat{\theta}_*, \theta_0) \approx \mathcal{E}((\widehat{\vartheta} - \vartheta)^2|S_*) = \mathbb{E}\left\{(\sqrt{\delta} \tau_* Z)^2\right\} = \delta \tau_*^2.$$

Other predictions might also be of interest. Thus, concerning the mean absolute error  $\text{MAE}(\widehat{\theta}^t, \theta_0) = \|\widehat{\theta}^t - \theta_0\|_1/p$ , state evolution predicts  $\text{MAE} \approx \sqrt{2\delta\tau_t^2/\pi}$ . Concerning functions of  $(R, W)$ , consider the ordinary residuals  $Y - X\widehat{\theta}^*$  at AMP convergence. These residuals will of course in general not have the distribution of the errors  $W$ . Setting  $\eta(z; b) = z - \Psi(z; b)$ , we have  $Y - X\widehat{\theta}^* = \eta(R; b_*)$ . State evolution predicts that the ordinary residuals will have the same distribution as  $\eta(W + \tau_* Z; b_*)$ .

### 3.4 Example of State Evolution predictions

Continuing with our running example, we again consider the case of contaminated normal data  $W \sim \text{CN}(0.05, 10)$  and Huber  $\rho$  with  $\lambda = 3$ . If we start AMP with the all-zero estimate  $\widehat{\theta}^0 = 0$ , then since  $\|\theta_0\|_2 = 6\sqrt{p}$  we start SE with  $\tau_0 = 2.056$ . Figure 4 presents predictions by state evolution for the MSE (left panel) and for the mean absolute error MAE.

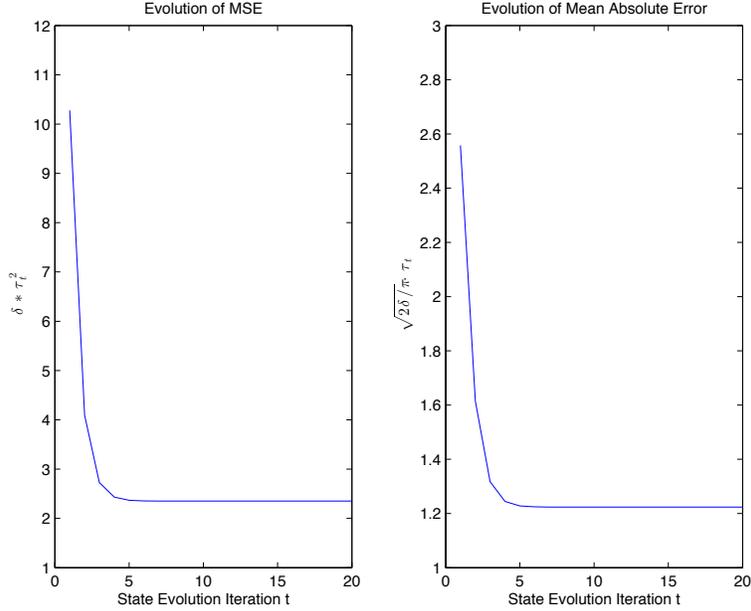


Figure 4: State Evolution predictions for  $\text{CN}(0.05, 10)$ , with Huber  $\psi$ ,  $\lambda = 3$ . Predicted evolutions of two observables of  $\hat{\theta}^t - \theta_0$ : Left: MSE, Mean Squared Error. Right: MAE, Mean Absolute Error.

Again in our running example, these predictions can be tested empirically. For illustration, we conducted a very small experiment, generating 10 independent realizations of the running model at  $n = 1000$  and  $p = 200$ , and comparing the actual evolutions of observables during AMP iterations with the predicted evolutions. Figure 5 shows that the predictions from SE are very close to the averages across realizations.

### 3.5 A lower bound on State Evolution

State Evolution cannot evolve so that  $\tau_t^2 \rightarrow 0$ ; under minimal regularity, it always exceeds a specific nonzero noise level.

**Lemma 3.4.** *Suppose that  $F_W$  has a well-defined Fisher information  $I(F_W)$ . Then for any  $t > 0$*

$$\tau_t^2 \geq \frac{1}{\delta I(F_W)}.$$

*Proof.* Let  $G = F_W \star N(0, \tau^2)$ .  $\mathbb{E}_G \Psi' = \frac{1}{\delta}$  and, if  $\xi_G$  denotes the score function for location of  $G$ , then  $|\mathbb{E}_G \Psi'| = |\mathbb{E}_G \Psi \cdot \xi_G|$ . Meanwhile, by Cauchy-Schwartz,  $|\mathbb{E}_G \Psi \cdot \xi_G| \leq \sqrt{\mathbb{E}_G \Psi^2} \sqrt{\mathbb{E}_G \xi_G^2}$ . We conclude that

$$\tilde{\mathcal{V}}(\tau) = \delta \mathbb{E}_G \Psi^2 \geq \delta \frac{|\mathbb{E}_G \Psi \cdot \xi_G|^2}{\mathbb{E}_G \xi_G^2} = \delta \frac{|\mathbb{E}_G \Psi'|^2}{I(G)} = \frac{1}{\delta I(G)}.$$

From convexity and translation-invariance of Fisher Information  $I(G) = I(F_W \star N(0, \tau^2)) < I(F_W)$ . Then  $\tau_t^2 = \tilde{\mathcal{V}}(\tau_{t-1}^2) \geq 1/(\delta I(F_W))$ .  $\square$

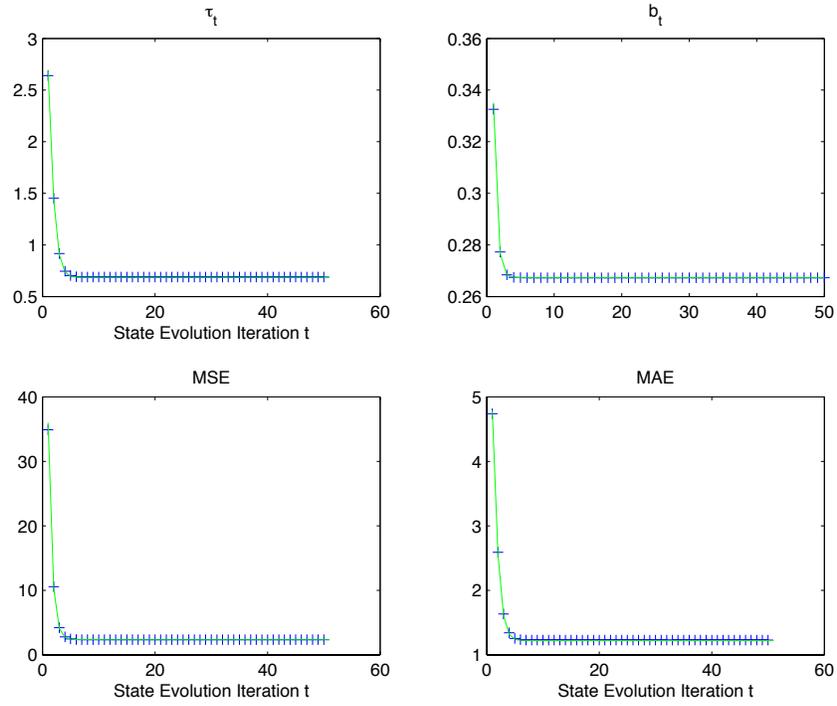


Figure 5: Experimental means from 10 simulations compared with State Evolution predictions under  $\text{CN}(0.05, 10)$ , with Huber  $\psi$ ,  $\lambda = 3$ . Upper Left:  $\hat{\tau}_t = \|\hat{\theta}^t - \theta_0\|_2 / \sqrt{n}$ . Upper Right:  $\hat{b}_t$ . Lower Left: MSE, Mean Squared Error. Lower Right: MAE, Mean Absolute Error. Blue '+' symbols: Empirical means of AMP observables. Green Curve: Theoretical predictions by SE.

We can sharpen this bound one step further. It will be convenient to write  $I(X)$  for the Fisher information of distribution  $F_X$ .

**Lemma 3.5.**

$$I(W + \tau Z) \leq \frac{I(W)}{1 + \tau^2 I(W)}.$$

*Proof.* Barron and Madiman [MB07] give the inequality  $I(W + \tau Z) \leq x^2 I(W) + (1-x)^2 I(\tau Z)$ , valid for any  $x \in (0, 1)$ . By calculus, we know that for  $a, b > 0$ ,

$$\min_{x \in (0, 1)} x^2 a + (1-x)^2 b = \frac{ab}{a+b}.$$

Setting  $a = I(W)$  and  $b = I(\tau Z) = \tau^{-2}$ , and dividing both numerator and denominator by  $b$ , we are done.  $\square$

Revisit the argument of Lemma 3.4; the inequality  $\tau_t^2 \geq 1/(\delta I(F_W))$  shows that if  $t > 0$ , then  $\tau_t^2 I(W) \geq 1/\delta$ . Using this in the previous Lemma,

$$I(W + \tau_t Z) \leq \frac{I(W)}{1 + \tau_t^2 I(W)} \leq \frac{I(W)}{1 + \frac{1}{\delta}}.$$

This yields a ‘one-step’ improvement:

**Corollary 3.6.** *Suppose that  $F_W$  has a well-defined Fisher Information  $I(F_W)$ . Then for any  $t > 1$*

$$\tau_t^2 \geq \frac{1 + \frac{1}{\delta}}{\delta I(F_W)}.$$

We can iterate this argument across many steps, obtaining that, for every  $t > k$ ,

$$\tau_t^2 \geq \frac{1 + \frac{1}{\delta} + \frac{1}{\delta^2} + \cdots + \frac{1}{\delta^k}}{\delta I(F_W)}.$$

We obtain immediately:

**Corollary 3.7.** *Suppose that  $F_W$  has a well-defined Fisher information  $I(F_W)$ . Then for every accumulation point  $\tau_*$  of State Evolution*

$$\tau_*^2 \geq \frac{1}{\delta - 1} \cdot \frac{1}{I(F_W)}.$$

### 3.6 Correctness of State Evolution predictions

The predictions of state evolution can be validated in the large-system limit  $n, p \rightarrow \infty$ , under the random Gaussian design assumption of Definition 3.1. We impose regularity conditions on the observables whose behavior we attempt to predict:

**Definition 3.8.** *A function  $\xi : \mathbb{R}^k \rightarrow \mathbb{R}$  is pseudo-Lipschitz if there exists  $L < \infty$  such that, for all  $x, y \in \mathbb{R}^k$ ,  $|\xi(x) - \xi(y)| \leq L(1 + \|x\|_2 + \|y\|_2) \|x - y\|_2$ .*

In particular,  $\xi(x) = x^2$  is pseudo-Lipschitz.

Recall also the definition of MSE in equation (18). For a sequence of estimators  $\tilde{\theta}$ , define the per-coordinate asymptotic mean squared error (AMSE) as the following large-system limit:

$$\text{AMSE}(\tilde{\theta}; \theta_0) =_{\text{a.s.}} \lim_{n, p_n \rightarrow \infty} \text{MSE}(\tilde{\theta}; \theta_0), \quad (21)$$

when the indicated limit exists.

The following result validates the predictions of State Evolution for pseudo-Lipschitz observables. Our proof is deferred to Appendix B.

**Theorem 3.9.** *Assume that the loss function  $\rho$  is convex and smooth, that the sequence of matrices  $\{\mathbf{X}(n)\}_n$  is a standard Gaussian design, and that  $\theta_0, \hat{\theta}^0$  are deterministic sequences such that  $\text{AMSE}(\theta_0, \hat{\theta}^0) = \delta\tau_0^2$ . Further assume that  $F_W$  has finite second moment and let  $\{\tau_t^2\}_{t \geq 0}$  be the state evolution sequence with initial condition  $\tau_0^2$ . Let  $\{\hat{\theta}^t, R^t\}_{t \geq 0}$  be the AMP trajectory with parameters  $b_t$  as per Eq. (19).*

Let  $\xi : \mathbb{R} \rightarrow \mathbb{R}$ ,  $\xi_2 : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be pseudo-Lipschitz functions. Then, for any  $t > 0$ , we have, for  $Z \sim \mathcal{N}(0, 1)$  independent of  $W \sim F_W$

$$\lim_{n \rightarrow \infty} \frac{1}{p} \sum_{i=1}^p \xi(\hat{\theta}_i^t - \theta_{0,i}) =_{\text{a.s.}} \mathbb{E} \left\{ \xi(\sqrt{\delta} \tau_t Z) \right\}, \quad (22)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_2(R_i^t, W_i) =_{\text{a.s.}} \mathbb{E} \left\{ \xi_2(W + \tau_t Z, W) \right\}. \quad (23)$$

In particular, we may take  $\xi(x) = x^2$  and obtain for the AMP iteration

$$\text{AMSE}(\hat{\theta}^t, \theta_0) = \delta\tau_t^2,$$

in full agreement with the predictions of state evolution in Definition 3.3.

## 4 Convergence and characterization of M-estimators

The key step for characterizing the distribution of the M-estimator  $\hat{\theta}$ , cf. Eq. (2), is to prove that the AMP iterates  $\hat{\theta}^t$  converge to  $\hat{\theta}$ . We will prove that this is indeed the case, at least in the limit  $n, p \rightarrow \infty$ , and for suitable initial conditions<sup>11</sup>.

Throughout this section, we shall assume that  $\rho$  is *strongly convex*, i.e. that  $\inf_{x \in \mathbb{R}} \rho''(x) > 0$ . This corresponds to assuming  $\inf_{x \in \mathbb{R}} \psi'(x) > 0$ , which is rather natural from the point of view of robust statistics since it ensures uniqueness of the M estimator<sup>12</sup>.

The key step is to establish the following high-dimensional convergence result.

---

<sup>11</sup>We expect convergence for arbitrary initial conditions (as long as they are independent of  $(W, \mathbf{X})$ ), but proving this claim is not needed for our main goal, and we leave it for future study. Proving this claim would require showing convergence of the state evolution recursion (20).

<sup>12</sup>The Huber estimator is not covered by the result of this section; although we expect our approach to apply in such generality. We focus here on the strongly convex case to avoid un-necessary complications.

**Theorem 4.1. (Convergence of AMP to the M-Estimator.)** *Assume the same setting as in Theorem 3.9, and further assume that  $\rho$  is strongly convex and that  $\delta > 1$ .*

*Let  $(\tau_*, b_*)$  be a solution of the two equations*

$$\tau^2 = \delta \mathbb{E} \left\{ \Psi(W + \tau Z; b)^2 \right\}, \quad (24)$$

$$\frac{1}{\delta} = \mathbb{E} \left\{ \Psi'(W + \tau Z; b) \right\}. \quad (25)$$

*and assume that  $\text{AMSE}(\hat{\theta}^0, \theta_0) = \delta \tau_*^2$ . Then*

$$\lim_{t \rightarrow \infty} \text{AMSE}(\hat{\theta}^t, \hat{\theta}) = 0. \quad (26)$$

From this and Theorem 3.9, the desired characterization of  $\hat{\theta}$  immediately follows.

To tie back to the introduction, we prove formula (4):

**Corollary 4.2. (Asymptotic Variance Formula under High-Dimensional Asymptotics.)** *Assume the setting of Theorem 3.9, and further assume that  $\rho$  is strongly convex and  $\delta > 1$ . The asymptotic variance of  $\hat{\theta}$  obeys*

$$\lim_{n, p \rightarrow \infty} \text{Ave}_{i \in [p]} \text{Var}(\hat{\theta}_i) =_{\text{a.s.}} V(\tilde{\Psi}, \tilde{F}), \quad (27)$$

*where  $\text{Ave}_{i \in [p]}$  denotes the average across indices  $i$ ,  $V(\psi, F)$  denotes the usual Huber asymptotic variance formula for M-estimates –  $V(\psi, F) = (\int \psi^2 dF) / (\int \psi' dF)^2$  – and the effective score  $\tilde{\Psi}$  is*

$$\tilde{\Psi}(\cdot) = \Psi(\cdot; b_*),$$

*while the effective noise distribution  $\tilde{F}$  is*

$$\tilde{F} = F_W \star \mathbf{N}(0, \tau_*^2).$$

*Here  $(\tau_*, b_*)$  are the unique solutions of the equations (24)-(25).*

*Proof.* By symmetry,  $\text{Ave}_{i \in [p]} \text{Var}(\hat{\theta}_i) = \mathbb{E} \text{MSE}(\hat{\theta}, \theta_0)$ . Theorem 4.1 and State Evolution show that  $\text{AMSE}(\hat{\theta}, \theta_0) = \delta \tau_*^2$ . By (24)-(25)

$$V(\tilde{\Psi}, \tilde{F}) = \frac{\mathbb{E} \Psi^2(W + \tau_* Z; b_*)}{[\mathbb{E} \Psi'(W + \tau_* Z; b_*)]^2} = \frac{\tau_*^2 / \delta}{\delta^{-2}} = \delta \tau_*^2.$$

□

Recall that the traditional information bound for M-estimators is  $V(\Psi, F) \geq \frac{1}{I(F_W)}$ , and that this is achievable under  $p$  fixed,  $n \rightarrow \infty$  asymptotics. Considering the formula for  $\tilde{F}$  we see that because  $\tau_* > 0$ , such an asymptotic variance is not achievable under high-dimensional asymptotics. We now make this effect more visible. Combining Corollary 4.2 with Corollary 3.7's lower bound on the equilibrium noise  $\tau_*$  reachable by State Evolution, we have the following.

**Corollary 4.3. Information Bound under High-Dimensional Asymptotics:**

$$V(\tilde{\Psi}, \tilde{F}) \geq \frac{1}{\delta - 1} \cdot \frac{1}{I(F_W)}.$$

In this inequality, the effect of the high-dimensional asymptotics parameter  $\delta$  is extremely clear; it shows that the classical information bound is not achievable when  $\delta = n/p < \infty$ , There is always an inflation in variance at least by  $(1 - \delta)^{-1} = \frac{n}{n-p}$ . Moreover, the inflation completely blows up as  $\delta \rightarrow 1$ .

**Corollary 4.4.** *Assume the setting of Theorem 3.9, and further assume that  $\rho$  is strongly convex and  $\delta > 1$ . Then for any pseudo-Lipschitz function  $\xi : \mathbb{R} \rightarrow \mathbb{R}$ , we have, for  $Z \sim \mathbf{N}(0, 1)$*

$$\lim_{n \rightarrow \infty} \frac{1}{p} \sum_{i=1}^p \xi(\hat{\theta}_i^t - \theta_{0,i}) =_{a.s.} \mathbb{E} \left\{ \xi(\sqrt{\delta} \tau_* Z) \right\}. \quad (28)$$

*In particular, the solution of Eqs. (24), (25) is necessarily unique.*

Among other applications, this result can be used to bound the suboptimality of AMP after a fixed number of iterations. Combining Theorems 3.9 and 4.1 gives:

**Corollary 4.5.** *Assume the same setting as in Theorem 3.9, and further assume that  $\rho$  is strongly convex and  $\delta > 1$ . Then the almost sure limits  $\text{AMSE}(\hat{\theta}^t; \theta_0)$  and  $\text{AMSE}(\hat{\theta}; \theta_0)$  exist, and obey*

$$\text{AMSE}(\hat{\theta}^t; \theta_0) - \text{AMSE}(\hat{\theta}; \theta_0) = \delta(\tau_t^2 - \tau_*^2). \quad (29)$$

Theorem 4.1 extends to cover general Gaussian matrices  $\mathbf{X}$  with i.i.d. rows.

**Definition 4.6.** *We say that a sequence of random design matrices  $\{\mathbf{X}(n)\}_n$ , with  $n \rightarrow \infty$ , is a general Gaussian design if each  $\mathbf{X} = \mathbf{X}(n)$  has dimensions  $n \times p$ , and rows  $(X_i)_{i \in [n]}$  that are i.i.d.  $\mathbf{N}(0, \Sigma/n)$ , where  $\Sigma = \Sigma(n) \in \mathbb{R}^{p \times p}$  is a strictly positive definite matrix. Further,  $p = p(n)$  is such that  $\lim_{n \rightarrow \infty} n/p(n) = \delta \in (0, \infty)$ .*

Notice that, if  $\mathbf{X}$  is a general Gaussian design, then  $\mathbf{X}\Sigma^{-1/2}$  is a standard Gaussian design. The following then follows from Corollary 4.7 together with a simple change of variables argument, cf. [EKBB13, Lemma 1].

**Corollary 4.7.** *Assume the same setting as in Theorem 3.9, but with  $\{\mathbf{X}(n)\}_{n \geq 0}$  being a general Gaussian design with covariance  $\Sigma$ , and further assume that  $\rho$  is strongly convex and  $\delta > 1$ . There is a scalar random variable  $T_n$  so that*

$$\hat{\theta} = \theta_0 + \sqrt{\delta} T_n \Sigma^{-1/2} \mathbf{Z}, \quad (30)$$

where  $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{I}_{p \times p})$  and we have the almost-sure limit  $\lim_{n \rightarrow \infty} T_n =_{a.s.} \tau_*$ , where  $\tau_*$  solves Eqs. (24), (25).

This result coincides with *Corollary 1* in [EKBB13] apart from a factor  $\sqrt{n}$  in the random part of Eq. (30) that arises because of a difference in the normalization of  $\mathbf{X}$ .

## 5 Discussion

Several generalizations of the present proof technique should be possible, and would be of interest. We list a few in order of increasing difficulty:

1. Generalize the i.i.d. Gaussian rows model for  $\mathbf{X}$  by allowing different rows to be randomly scaled copies of a common  $X \sim \mathcal{N}(0, \Sigma/n)$ . This is the setting of [EKBBL13, Result 1].
2. Remove the smoothness and strong convexity assumptions on  $\rho$ .
3. Add a regularization term to the objective function  $\mathcal{L}(\theta)$  cf. Eq. (2), of the form  $\sum_{i=1}^p J(\theta_i)$ , with  $J : \mathbb{R} \rightarrow \mathbb{R}$  a convex penalty. For  $\ell_1$  penalty and  $\ell_2$  loss, this reduces to the Lasso, studied in [BM12].
4. Generalize the present results to non-Gaussian designs. We expect –for instance– that they should hold universally across matrices  $\mathbf{X}$  with i.i.d. entries (under suitable moment conditions). A similar universality result was established in [BLM12] for compressed sensing.

Let us mention that alternative proof techniques would be worth exploring as well. In particular, Shcherbina and Tirozzi [ST03] define a statistical mechanics model with energy function that is analogous to the loss  $\mathcal{L}(\theta)$ , cf. Eq. (2), and Talagrand [Tal10, Chapter 3] proves further results on the same model. While this treatment focuses on estimating a certain partition function, in the case of strongly convex  $\rho$  it should be possible to extract properties of the minimizer from a ‘zero-temperature’ limit.

Finally, Rangan [Ran11] considers a similar regression model to the one studied here using approximate message passing algorithms, albeit from a Bayesian point of view.

## 6 Duality between robust regression and regularized least squares

The reader might have noticed many analogies between the analysis in the last pages and earlier work on estimation in the underdetermined regime  $n < p$  using the Lasso [DMM09, DMM11, DJMM11, BM12]. Most specifically, the central tool in our proof of the correctness of State Evolution is a set of lemmas and theorems about analysis of recursive systems that were developed to understand the Lasso. That the same machinery directly gives results in robust regression - see for example our proof of correctness of State Evolution in Appendix B below - might seem particularly unexpected. In this section we briefly point out that the two problems are so closely linked that phenomena which appear in one situation are bound to appear in the other.

### 6.1 Duality of optimization problems

In a very strong sense, solving an M-estimation problem with  $p < n$  is the very same thing as solving a related penalized regression problem in  $\tilde{p} > \tilde{n}$ . Given a convex function  $J : \mathbb{R} \rightarrow \mathbb{R}$ , define the  $\rho$  function

$$\rho_J(z) \equiv \min_{x \in \mathbb{R}} \left\{ \frac{1}{2}(z - x)^2 + J(x) \right\} \quad (31)$$

We then have the M-Estimation problem

$$(M_J) \quad \min_{\theta \in \mathbb{R}^p} \sum_{i=1}^n \rho_J(Y_i - \langle X_i, \theta \rangle) \quad (32)$$

This problem has  $p < n$  and is generically a determined problem. We now construct a corresponding underdetermined problem with the ‘same’ solution. Set  $\tilde{n} = n - p$ ,  $\tilde{p} = n$ . We soon will

construct a vector/matrix pair  $(\tilde{Y} \in \mathbb{R}^{\tilde{n}}, \tilde{\mathbf{X}} \in \mathbb{R}^{\tilde{n} \times \tilde{p}})$  obeying  $\tilde{n} < \tilde{p}$ , where  $\tilde{Y}$  and  $\tilde{\mathbf{X}}$  are related to  $Y$  and  $\mathbf{X}$  in a specific way. With this pair we pose the  $J$ -penalized least squares problem

$$(L_J) \quad \min_{\beta \in \mathbb{R}^{\tilde{p}}} \frac{1}{2} \|\tilde{Y} - \tilde{\mathbf{X}}\beta\|_2^2 + \sum_{i=1}^{\tilde{p}} J(\beta_i). \quad (33)$$

with solution  $\hat{\beta}(\tilde{Y}; \tilde{\mathbf{X}})$ , say.

Here is the specific pair that links  $(M_J)$  with  $(L_J)$ . We let  $\tilde{\mathbf{X}}$  be a matrix with orthonormal rows such that  $\tilde{\mathbf{X}}\mathbf{X} = 0$ , i.e.

$$\text{null}(\tilde{\mathbf{X}}) = \text{image}(\mathbf{X}), \quad (34)$$

finally, we set  $\tilde{Y} = \tilde{\mathbf{X}}Y$ .

### 6.1.1 The Lasso-Huber connection

Of special interest is the case  $J(x) = \lambda|x|$  in which case  $(L_J)$  of (33) defines the Lasso estimator. Then  $\rho_J(x) = \rho_H(x; \lambda)$  is the Huber loss and  $(M_J)$  of (32) defines the Huber M-estimate. Indeed, in that case  $(L_J)$  is more classically presented as

$$(\text{Lasso}_\lambda) \quad \min_{\beta \in \mathbb{R}^{\tilde{p}}} \frac{1}{2} \|\tilde{Y} - \tilde{\mathbf{X}}\beta\|_2^2 + \lambda \sum_{i=1}^{\tilde{p}} |\beta_i|, \quad (35)$$

while  $(M_J)$  is more classically presented as

$$(\text{Huber}_\lambda) \quad \min_{\beta \in \mathbb{R}^{\tilde{p}}} \sum_{i=1}^n \rho_H(Y_i - \langle X_i, \beta \rangle; \lambda) \quad (36)$$

In this special case, our general result from the next section implies the following:

**Proposition 6.1.** *With problem instances  $(Y, X)$  and  $(\tilde{Y}, \tilde{\mathbf{X}})$  related as above, the optimal values of the Lasso problem  $(\text{Lasso}_\lambda)$  and the Huber problem  $(\text{Huber}_\lambda)$  are identical. The solutions of the two problems are in one-one-relation. In particular, we have*

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (Y - \hat{\beta}). \quad (37)$$

In a sense the Lasso problem solution  $\hat{\beta}$  is finding the outliers in  $Y$ ; once the solution is known, the solution of the M-estimation problem is simply a least squares regression on adjusted data  $Y_{\text{adj}} \equiv (Y - \hat{\beta})$  with outliers removed.

### 6.1.2 General duality result

We will now show that the problem (32) is dual to (33) under or special choice of  $(\tilde{Y}, \tilde{\mathbf{X}})$ , via (34).

*Notation.* For  $x \in \mathbb{R}^n$ , we denote by  $\partial\rho(x)$  the subgradient of the convex function  $\sum_{i=1}^n \rho(x_i)$ , at  $x$ . Analogously, for  $z \in \mathbb{R}^{\tilde{p}}$ , we denote by  $\partial J(z)$  the subgradient of the convex function  $\sum_{i=1}^{\tilde{p}} J(z_i)$ , at  $z$ .

**Proposition 6.2.** *Assume that  $\rho(\cdot) = \rho_J(\cdot)$ , that  $\tilde{\mathbf{X}}$  has orthonormal rows with  $\text{null}(\tilde{\mathbf{X}}) = \text{image}(\mathbf{X})$ , and finally that  $\tilde{Y} = \tilde{\mathbf{X}}Y$ . Then the solutions of the regularized least squares problem (33) are in one-to-one correspondence with the solutions of the robust regression problem (2), via the mappings*

$$\hat{\beta} = Y - \mathbf{X}\hat{\theta} - u, \quad u \in \text{null}(\mathbf{X}^\top) \cap \partial\rho(y - \mathbf{X}\hat{\theta}), \quad (38)$$

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (Y - \hat{\beta}). \quad (39)$$

*Proof.* ‘Differentiating’ Eq. (31) it is easy to see that

$$u \in \partial\rho(x) \quad \text{if and only if} \quad u \in \partial J(x - u). \quad (40)$$

First assume  $\hat{\theta}$  is a minimizer of problem (32). This happens if and only if there exists  $u \in \mathbb{R}^n$  such that

$$\mathbf{X}^\top u = 0, \quad u \in \partial\rho(Y - \mathbf{X}\hat{\theta}). \quad (41)$$

We then claim that  $\hat{\beta} \equiv Y - \mathbf{X}\hat{\theta} - u$  is a minimizer of Eq. (33). Indeed

$$\tilde{\mathbf{X}}^\top (\tilde{Y} - \tilde{\mathbf{X}}\hat{\beta}) = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} (Y - \hat{\beta}) \quad (42)$$

$$= \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} (\mathbf{X}\hat{\theta} + u) = u, \quad (43)$$

where the last identity follows since, by Eq. (34),  $\text{null}(\mathbf{X}^\top) = \text{image}(\tilde{\mathbf{X}}^\top)$ , and hence  $u \in \text{image}(\tilde{\mathbf{X}}^\top)$  by Eq. (41). Using again Eqs. (41) and (40), we deduce that  $u \in \partial J(\hat{\beta})$ , i.e.

$$\tilde{\mathbf{X}}^\top (\tilde{Y} - \tilde{\mathbf{X}}\hat{\beta}) \in \partial J(\hat{\beta}), \quad (44)$$

which is the stationarity condition for the problem (33).

Viceversa a similar argument shows that, given  $\hat{\beta}$  that minimizes Eq. (33), and  $\hat{\theta} \equiv (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (Y - \hat{\beta})$  is a minimizer of the robust regression problem (32).  $\square$

## 6.2 Comparison to AMP in the $p > n$ case

The last section raises the possibility that the phenomena found in this paper for M-estimation in the  $p < n$  case are actually isomorphic to those found in our previous work on penalized regression in the  $p > n$  case; [DMM09, DMM11, DJMM11, BM12]. Here we merely content ourselves with sketching a few similarities.

To be definite, consider robust regression using the Huber loss [Hub64, HR09]  $\rho(x) = x^2/2$  for  $|x| \leq \lambda$  and  $\rho(x) = \lambda|x| - \lambda^2/2$  otherwise. In this case it is easy to see that

$$\Psi(z; b) = \begin{cases} \lambda b & \text{if } z > \lambda(1 + b), \\ bz/(1 + b) & \text{if } |z| \leq \lambda(1 + b), \\ -\lambda b & \text{if } z < -\lambda(1 + b). \end{cases} \quad (45)$$

In order to make contact with the Lasso, recall the definition of soft thresholding operator  $\eta(x; \alpha) = \text{sign}(x) (|x| - \alpha)_+$ . We have the relationship

$$\Psi(z; b) = \frac{bz}{1 + b} - \eta\left(\frac{bz}{1 + b}; \lambda b\right). \quad (46)$$

Letting  $c_t \equiv b_t/(1 + b_t)$ , the state evolution equation (20), then reads

$$\tau_{t+1}^2 = \delta c_t^2 \mathbb{E} \left\{ \left[ \eta \left( W + \tau_t Z; \lambda(1 + b_t) \right) - W - \tau_t Z \right]^2 \right\}, \quad (47)$$

This is very close to the state evolution equation in compressed sensing for reconstructing a sparse signal whose entries have distribution  $F_W$ , from an underdetermined number of linear measurements; indeed in that setting we have the state evolution recursion

$$\tau_{t+1}^2 = \delta \mathbb{E} \left\{ \left[ \eta \left( W + \tau_t Z; \lambda \tau_t \right) - W \right]^2 \right\}; \quad (48)$$

[DMM09, DMM11, DJMM11, BM12]. The connection is quite suggestive: while in compressed sensing we look for the few non-zero coefficients in the signal, in robust regression we try to identify the few outliers contaminating the linear relation. A similar duality was already pointed out in [DT09], although in a specific setting.

## Acknowledgements

This work was partially supported by the NSF CAREER award CCF-0743978, the NSF grant DMS-0806211, and the grants AFOSR/DARPA FA9550-12-1-0411 and FA9550-13-1-0036.

## A Properties of the functions $\text{Prox}$ , $\Psi$

Throughout this section  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  is convex bounded below and smooth (i.e. with bounded second derivative). Recall the definition of  $\text{Prox} : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}$  and  $\Psi : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}$ , given by

$$\text{Prox}(z; b) \equiv \arg \min_{x \in \mathbb{R}} \left\{ \rho(x) + \frac{1}{2b}(x - z)^2 \right\}, \quad (49)$$

$$\Psi(z; b) \equiv b \rho'(\text{Prox}(z; b)). \quad (50)$$

**Proposition A.1.** *The function  $\text{Prox} : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}$  is differentiable in its domain, with partial derivatives*

$$\frac{\partial \text{Prox}}{\partial z}(z; b) = \frac{1}{1 + b\rho''(x)} \Big|_{x=\text{Prox}(z; b)}, \quad \frac{\partial \text{Prox}}{\partial b}(z; b) = -\frac{\rho'(x)}{1 + b\rho''(x)} \Big|_{x=\text{Prox}(z; b)}. \quad (51)$$

*In particular, letting  $\|\rho''\|_\infty \equiv \sup_{x \in \mathbb{R}} \rho''(x)$ , and for any fixed  $b$ ,  $z \mapsto \text{Prox}(z; b)$  is strictly increasing and Lipschitz continuous, with*

$$\frac{1}{1 + b\|\rho''\|_\infty} \leq \frac{\partial \text{Prox}}{\partial z}(z; b) \leq 1 \quad (52)$$

*Proof.* Since, for  $b > 0$ ,  $x \mapsto \rho(x) + (x - z)^2/(2b)$  is differentiable and strongly convex,  $x = \text{Prox}(z; b)$  is uniquely determined by setting to zero the first derivative:

$$x + b\rho'(x) - z = 0. \quad (53)$$

The claim then follows from the Implicit Function theorem.  $\square$

**Proposition A.2.** For  $(z, b) \in \mathbb{R} \times \mathbb{R}_+$ , we have

$$\Psi(z; b) = z - \text{Prox}(z, b), \quad (54)$$

and hence  $\Psi$  is differentiable, with partial derivatives

$$\frac{\partial \Psi}{\partial z}(z; b) = \frac{b\rho''(x)}{1 + b\rho''(x)} \Big|_{x=\text{Prox}(z; b)}, \quad \frac{\partial \Psi}{\partial b}(z; b) = \frac{\rho'(x)}{1 + b\rho''(x)} \Big|_{x=\text{Prox}(z; b)}. \quad (55)$$

In particular, for any fixed  $b$ ,  $z \mapsto \Psi(z; b)$  is strictly increasing and Lipschitz continuous, with

$$\frac{b \inf_{x \in \mathbb{R}} \rho''(x)}{1 + b \inf_{x \in \mathbb{R}} \rho''(x)} \leq \frac{\partial \Psi}{\partial z}(z; b) \leq \frac{b \|\rho''\|_\infty}{1 + b \|\rho''\|_\infty}. \quad (56)$$

*Proof.* Using again the stationarity condition (53) that holds for  $x = \text{Prox}(z; b)$ , we have

$$\text{Prox}(z; b) + b\rho'(\text{Prox}(z; b)) - z = 0, \quad (57)$$

which is our first claim. The other claims immediately follow by calculus.  $\square$

Finally, we prove that Eq. (19) that defines  $b_t$  as a function of  $\tau_t$  always has at least one solution.

**Lemma A.3.** For  $\tau > 0$  fixed, let  $G : \mathbb{R}_{>0} \rightarrow \mathbb{R}$  be defined by

$$G(b) \equiv \mathbb{E} \left\{ \Psi'(W + \tau Z; b) \right\}. \quad (58)$$

Then for any  $a \in (0, 1)$ , the set of solutions

$$\mathcal{S}_a \equiv \{b \in \mathbb{R}_{>0} : G(b) = a\}, \quad (59)$$

is closed and non-empty.

*Proof.* It follows immediately from the continuity properties of  $\Psi$  that  $b \mapsto G(b)$  is continuous. The claim follows by proving that  $\lim_{b \rightarrow 0} G(b) = 0$  and  $\lim_{b \rightarrow \infty} G(b) = 1$ .

By Proposition A.2 equation (56)  $0 \leq \Psi'(z; b) \leq 1$ . The limit  $b \rightarrow 0$  follows from dominated convergence since, by the upper bound in (56)  $\lim_{b \rightarrow 0} \Psi'(z; b) = 0$  for each  $z$ .

In order to obtain the limit as  $b \rightarrow \infty$ , note that by Stein Lemma:

$$G(b) = \frac{1}{\tau} \mathbb{E} \left\{ Z \Psi(W + \tau Z; b) \right\}. \quad (60)$$

Since  $0 \leq \Psi'(z, b) \leq 1$ , the integrand is bounded in modulus by an integrable quantity. We can therefore use again dominated convergence. Now  $\lim_{b \rightarrow \infty} \text{Prox}(z; b) = \arg \min_{x \in \mathbb{R}} \rho(x) \equiv c_0$  and hence  $\lim_{b \rightarrow \infty} \Psi(z; b) = z - c_0$ . By dominated convergence we obtain

$$\lim_{b \rightarrow \infty} G(b) = \frac{1}{\tau} \mathbb{E} \left\{ Z (W + \tau Z - c_0) \right\} = 1. \quad (61)$$

$\square$

## B Proof of correctness of State Evolution (Theorem 3.9)

We will show correctness of State Evolution for the AMP algorithm using analytically defined  $b_t$ . Namely, we suppose that with  $b_t$  defined recursively as the smallest positive solution of the second equation in this system:

$$\tau_{t+1}^2 = \delta \mathbb{E} \left\{ \Psi(W + \tau_t Z; b_t)^2 \right\}, \quad (62)$$

$$\frac{1}{\delta} = \mathbb{E} \left\{ \Psi'(W + \tau_t Z; b_t) \right\}. \quad (63)$$

For analysis purposes, we consider a recursion equivalent to the AMP recursion, in which the data are recentered and the recursion is recast around recentered variables. We change the initial condition of the AMP iteration by letting  $\widehat{\theta}^{\text{cen},0} = \widehat{\theta}^0 - \theta_0$ , and change data by letting  $Y^{\text{cen}} = Y - \mathbf{X}\theta_0 \equiv W$ . Applying the AMP recursion in these new coordinates gives the new trajectory  $\widehat{\theta}^{\text{cen},t} = \widehat{\theta}^t - \theta_0$  for all  $t$ , and  $R^{\text{cen},t} = R^t$  for all  $t$ .

The new trajectory follows the recursion

$$R^{\text{cen},t} = W - \mathbf{X}\widehat{\theta}^{\text{cen},t} + \Psi(R^{\text{cen},t-1}; b_{t-1}), \quad (64)$$

$$\widehat{\theta}^{\text{cen},t+1} = \widehat{\theta}^{\text{cen},t} + \delta \mathbf{X}^\top \Psi(R^{\text{cen},t}; b_t), \quad (65)$$

In this form, the recursion can be reduced to a recursion studied in [BM11], for which State Evolution has been proven correct. The reduction is to introduce a recursion generating iterates  $\{\vartheta^t, S^t\}$  that approximates closely the iterates  $\{\widehat{\theta}^{\text{cen},t}, R^{\text{cen},t}\}$  defined by (64),(65). The new sequence is defined by letting  $\vartheta^0 = \widehat{\theta}^0 - \theta_0$  and, for all  $t \geq 0$

$$S^t = -\mathbf{X}\vartheta^t + \Psi(W + S^{t-1}; b_{t-1}), \quad (66)$$

$$\vartheta^{t+1} = \delta \mathbf{X}^\top \Psi(W + S^t; b_t) + q_t \vartheta^t, \quad (67)$$

where

$$q_t = \delta \left\{ \frac{1}{n} \sum_{i=1}^n \Psi'(W_i + S_i^t; b_t) \right\}. \quad (68)$$

The only difference between this recursion and the previous one cf. Eqs. (64), (65), lies in the new coefficient  $q_t$ , which was identically equal to 1 in the previous recursion. The benefit of this specific recursion is that we already know that State Evolution is correct.

**Lemma B.1.** *Under the assumptions of Theorem 3.9, we have, for any fixed  $t \geq 0$ ,*

$$\lim_{n \rightarrow \infty} \frac{1}{p} \sum_{i=1}^p \xi(\vartheta_i^t) =_{\text{a.s.}} \mathbb{E} \left\{ \xi(\sqrt{\delta} \tau_t Z) \right\} \quad (69)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_2(S_i^t, W_i) =_{\text{a.s.}} \mathbb{E} \left\{ \xi_2(\tau_t Z, W) \right\}. \quad (70)$$

*Proof.* This is an immediate application of Theorem 2 in [BM11]. That Theorem considers general recursions which include (66)-(67) as a special case, and corresponding state evolution equations, and

shows the correctness of state evolution, in the process establishing conclusions of the precise form shown in the conclusion of this lemma. So it is simply a matter of establishing the correspondence of variables.

In the original notation of [BM11], the generalized AMP recursions studied are

$$b^t = Aq^t - \lambda_t m^{t-1} \quad (71)$$

$$h^{t+1} = A^* m^t - \xi_t q^t \quad (72)$$

where  $b^t$ ,  $h^t$ ,  $q^t$  and  $m^t$  are vectors and  $\lambda_t$  and  $\xi_t$  scalars. In addition, the vectors  $q^t = f_t(h^t)$  and  $m^t = g^t(b^t, w)$  are produced by element wise applications of nonlinearities  $f_t$  and  $g_t$ , the latter involving the random vector  $w$ . Here  $A$  is a rectangular  $n \times N$  random matrix with iid Gaussian entries. The scalars  $\xi_t = \langle g'_t(b^t, w) \rangle$  and  $\lambda_t = \frac{1}{\delta} \langle f'_t(h^t) \rangle$ , where  $\langle \cdot \rangle$  denotes an empirical mean over the entries in a vector. and the iteration takes  $m^{-1} = 0$ . For state evolution, the Theorem 2 assumes the sequence of initial conditions  $q^0$  obeys

$$\bar{\sigma}_0^2 = \lim_{N \rightarrow \infty} \frac{1}{N\delta} = \|q^0\|_2^2.$$

and the state evolution recursion involves the pair of variables

$$\bar{\tau}_t^2 = \mathbb{E}\{g_t^2(\bar{\sigma}_t Z, W)\}, \quad \bar{\sigma}_t^2 = \mathbb{E}\{f_t^2(\bar{\tau}_{t-1} Z)\}.$$

Table 1 sets up a ‘dictionary’ of correspondences between this paper and [BM11].

|      |   |                   |                   |                             |              |                              |
|------|---|-------------------|-------------------|-----------------------------|--------------|------------------------------|
| (66) | $\vartheta^{t+1} = \delta \mathbf{X}^\top \Psi(W + S^t; b_t) + q_t \vartheta^t$ | $\vartheta^{t+1}$ | $\mathbf{X}^\top$ | $\delta \Psi(W + S^t; b_t)$ | $q_t$        | $\vartheta^t$                |
| (71) | $h^{t+1} = A^* m^t - \xi_t q^t$   | $h^{t+1}$         | $A^*$             | $m^t$                       | $\xi_t$      | $-q^t$                       |
| (67) | $S^t = -\mathbf{X} \vartheta^t + \Psi(W + S^{t-1}; b_{t-1})$                    | $S^t$             | $\mathbf{X}$      | $-\vartheta^t$              | 1            | $\Psi(W + S^{t-1}; b_{t-1})$ |
| (72) | $b^t = Aq^t - \lambda_t m^{t-1}$  | $b^t$             | $A$               | $q^t$                       | $-\lambda_t$ | $m^{t-1}$                    |

Table 1: Correspondences between terms in the recursions of this paper, (66)-(67), and the recursions (71)-(72), analyzed in [BM11].

We get exact correspondence between the two systems, provided we identify  $\delta \Psi(W + S^t; b_t)$  with  $m^t = g_t(b^t; w)$  and  $-\delta h^t$  with  $f_t(h^t)$ . One has, in particular, that  $\lambda_t = \frac{1}{\delta} \langle f'_t(h^t) \rangle = -1$ , and that  $\xi_t = \langle g'_t(b^t, w) \rangle = \langle \delta \Psi'(W + S^t; b_t) \rangle = q_t$ .

In the [BM11] general study of state evolution, there are two state variables  $\bar{\tau}_t$  and  $\bar{\sigma}_t$ . However, when applied here, the distinction vanishes. The variable called  $\bar{\tau}_t^2$  corresponds with  $\delta^2 \mathbb{E}\{\Psi(W + \bar{\sigma}_t Z)^2\}$  while the variable  $\bar{\sigma}_t^2$  corresponds with  $\mathbb{E}(\vartheta^t)^2$ . Using facts about  $\tau_t^2$  in this paper, we have the identities  $\bar{\sigma}_t^2 = \delta \tau_t^2$  and  $\bar{\tau}_t^2 = \delta \tau_t^2$ . Equations (69)-(70) now follow from Theorem 2 of [BM11].  $\square$

Theorem 3.9 now follows from the equivalence of the last two recursions – i.e. equivalence of (64)-(65) with (66)-(67).

**Lemma B.2.** *Under the assumptions of Theorem 3.9, we have, for any fixed  $t \geq 0$ ,*

$$\lim_{n \rightarrow \infty} \frac{1}{p} \|\widehat{\theta}^{\text{cen},t} - \vartheta^t\|_2^2 =_{\text{a.s.}} 0, \quad \lim_{n \rightarrow \infty} \frac{1}{n} \|R^{\text{cen},t} - S^t - W\|_2^2 =_{\text{a.s.}} 0. \quad (73)$$

## B.1 Proof of Lemma B.2 (Equivalence of recursions)

Throughout this proof, we will drop the superscript ‘cen’ from  $R^{\text{cen},t}$  and  $\widehat{\theta}^{\text{cen},t}$ . Define  $S_+^t \equiv W + S^t$ , whence

$$S_+^t = W - \mathbf{X}\vartheta^t + \Psi(S_+^{t-1}; b_{t-1}), \quad (74)$$

$$\vartheta^{t+1} = \delta \mathbf{X}^\top \Psi(S_+^t; b_t) + q_t \vartheta^t, \quad (75)$$

Comparing the first of these equations with Eq. (64), and using triangular inequality, we get

$$\|R^t - S_+^t\|_2 \leq \|\mathbf{X}\|_2 \|\widehat{\theta}^t - \vartheta^t\|_2 + \|\Psi(R^{t-1}; b_{t-1}) - \Psi(S_+^{t-1}; b_{t-1})\|_2 \quad (76)$$

$$\leq \|\mathbf{X}\|_2 \|\widehat{\theta}^t - \vartheta^t\|_2 + \|R^{t-1} - S_+^{t-1}\|_2, \quad (77)$$

where the last inequality follows since  $\Psi(\cdot; b) : \mathbb{R} \rightarrow \mathbb{R}$  is Lipschitz continuous with Lipschitz constant at most 1, cf Proposition A.2.

Comparing analogously Eq. (65) and (75), we obtain

$$\|\widehat{\theta}^{t+1} - \vartheta^{t+1}\|_2 \leq \delta \|\mathbf{X}\|_2 \|\Psi(R^t; b_t) - \Psi(S_+^t; b_t)\|_2 + \|\widehat{\theta}^t - \vartheta^t\|_2 + |q_t - 1| \|\vartheta^t\|_2 \quad (78)$$

$$\leq \delta \|\mathbf{X}\|_2 \|R^t - S_+^t\|_2 + \|\widehat{\theta}^t - \vartheta^t\|_2 + |q_t - 1| \|\vartheta^t\|_2. \quad (79)$$

Iterating the upper bounds (77), (79), and using the fact that  $\vartheta^0 = \widehat{\theta}^0$ , we conclude that there exists a constant  $A = A(\delta) < \infty$  such that

$$\|\widehat{\theta}^t - \vartheta^t\|_2 \leq (A \|\mathbf{X}\|_2)^{2t} \sum_{\ell=0}^{t-1} |q_\ell - 1| \|\vartheta^\ell\|_2 \quad (80)$$

By Lemma B.1, we have, almost surely

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{p}} \|\vartheta^\ell\|_2 = \tau_\ell < \infty, \quad (81)$$

and

$$\lim_{n \rightarrow \infty} q_t = \delta \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \Psi'(W_i + S_i^t; b_t) \quad (82)$$

$$= \delta \mathbb{E}\{\Psi'(W + \tau_t Z; b_t)\} = 1, \quad (83)$$

where the second identity follows from Lemma B.1 and, in the third, we used the definition of  $b_t$ . (Note that we are applying here Lemma B.1 to  $\xi(\cdot) = \Psi'(\cdot; b_t)$  which is bounded and non-negative but not necessarily continuous. However, since  $W + \tau_t Z$  has a density for every  $\tau_t > 0$ , the limit holds by a standard weak convergence argument, approximating  $\xi$  by simple functions. Namely, we construct a sequence of simple functions  $\xi_\ell$  such that  $\xi_\ell(t) \leq \xi(t) \leq \xi_\ell(t) + (1/\ell)$  for all  $t$ , and apply Lemma B.1 –which implies weak convergence of the empirical distribution of  $\{W_i + S_i^t\}$ – to  $\xi_\ell$ .)

Finally, it is a standard result in random matrix theory [AGZ09] that  $\lim_{n \rightarrow \infty} \|\mathbf{X}\|_2 = C(\delta) < \infty$ . Hence, by taking the limit of Eq. (80) we get, almost surely,

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{p}} \|\widehat{\theta}^t - \vartheta^t\|_2 = 0. \quad (84)$$

The norm  $\|R^t - S_+^t\|_2$  is then controlled using Eq. (77).

## C Proof that AMP converges to the M-estimator (Theorem 4.1)

Notice first of all that, by construction,  $\tau_t^2 = \tau_*^2$ ,  $b_t = b_*$  for all  $t$ .

Given  $\delta, \rho$  as in the statement of the theorem and  $\tau_*, b_*$  a solution of the fixed point equation (24), (25), we define the doubly infinite matrix  $\Gamma = (\Gamma_{t,s})_{t,s \geq 0}$  by letting, recursively for  $t, s \geq 0$

$$\Gamma_{t+1,s+1} = \delta \mathbb{E}\{\Psi(W + Z_t; b_*)\Psi(W + Z_s; b_*)\}, \quad (85)$$

where the expectation is with respect to  $(Z_t, Z_s)$  jointly Gaussian, with zero means and covariance  $\mathbb{E}\{Z_t^2\} = \Gamma_{t,t}$ ,  $\mathbb{E}\{Z_s^2\} = \Gamma_{s,s}$ ,  $\mathbb{E}\{Z_t Z_s\} = \Gamma_{t,s}$ , independent of  $W \sim F_W$ . This is supplemented with the boundary condition  $\Gamma_{0,0} = \tau_*^2$  and  $\Gamma_{0,t} = \Gamma_{t,0} = 0$  for  $t > 0$ .

Notice that, in particular,  $\Gamma_{s,t} = \Gamma_{t,s}$  for all  $s, t \geq 0$  and  $\Gamma_{t,t} = \tau_*^2$  for all  $t$ .

The significance of these quantities is clarified by the following result.

**Lemma C.1.** *Under the hypotheses of Theorem 3.9, further assume that  $\tau_*^2$  and  $\Gamma$  are defined as above. Then, for any  $t, s \geq 0$ ,*

$$\lim_{n \rightarrow \infty} \frac{1}{p} \sum_{i=1}^p \xi_2(\hat{\theta}_i^t - \theta_{0,i}, \hat{\theta}_i^s - \theta_{0,i}) =_{a.s.} \mathbb{E} \xi_2(\sqrt{\delta} Z_t, \sqrt{\delta} Z_s), \quad (86)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_2(R_i^t - W_i, R_i^s - W_i) =_{a.s.} \mathbb{E} \xi_2(Z_t, Z_s), \quad (87)$$

where the expectation is with respect to  $(Z_t, Z_s)$  jointly Gaussian, with zero means and covariance  $\mathbb{E}\{Z_t^2\} = \Gamma_{t,t}$ ,  $\mathbb{E}\{Z_s^2\} = \Gamma_{s,s}$ ,  $\mathbb{E}\{Z_t Z_s\} = \Gamma_{t,s}$ , independent of  $W \sim F_W$ .

The proof is deferred to Section C.1.

As a special case of the latter result, we have

$$\lim_{n \rightarrow \infty} \frac{1}{p} \|\hat{\theta}^t - \hat{\theta}^s\|_2^2 =_{a.s.} 2\delta (\tau_*^2 - \Gamma_{t,s}), \quad (88)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \|R^t - R^s\|_2^2 =_{a.s.} 2(\tau_*^2 - \Gamma_{t,s}). \quad (89)$$

The following lemma provides information about the asymptotic behavior of  $\Gamma_{t,s}$ . Its proof is deferred to Section C.2.

**Lemma C.2.** *Let  $\tau_*, \Gamma$  be defined as above for  $\delta > 1$ . Then*

$$\lim_{t \rightarrow \infty} \Gamma_{t,t+1} = \tau_*^2 \quad (90)$$

Applying this result to Eqs. (88) and (89) we get, for any fixed  $h \in \mathbb{N}$ ,

$$\lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{p} \|\hat{\theta}^{t+h} - \hat{\theta}^t\|_2^2 =_{a.s.} 0, \quad (91)$$

$$\lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{p} \|R^{t+h} - R^t\|_2^2 =_{a.s.} 0. \quad (92)$$

(The case  $h > 1$  follows from  $h = 1$  by the triangle inequality.)

We are now ready to prove Theorem 4.1. Recall that  $\mathcal{L}(\theta) = \mathcal{L}(\theta; Y, \mathbf{X})$  denotes the loss function defined in Eq. (2), and that its gradient and Hessian are given by

$$\nabla_{\theta} \mathcal{L}(\theta) = - \sum_{i=1}^n \rho'(Y_i - \langle X_i, \theta \rangle) X_i, \quad (93)$$

$$\nabla_{\theta}^2 \mathcal{L}(\theta) = \sum_{i=1}^n \rho''(Y_i - \langle X_i, \theta \rangle) X_i X_i^{\top}. \quad (94)$$

In particular, letting  $\sigma_{\min}(\mathbf{X})$  denote the minimum non-zero singular value of  $\mathbf{X}$ , we have

$$\lambda_{\min}(\nabla_{\theta}^2 \mathcal{L}(\theta)) \geq \inf_{x \in \mathbb{R}} \rho''(x) \cdot \sigma_{\min}(\mathbf{X})^2. \quad (95)$$

Using the hypothesis of strong convexity and standard concentration of measure for the singular values of Wishart matrices [Ver12], there exists constants  $c_0, c_1, n_0 > 0$  for  $\delta > 1$  such that for any  $n \geq n_0$ ,

$$\mathbb{P}\left(\nabla_{\theta}^2 \mathcal{L}(\theta) \succeq c_0 \mathbf{I} \quad \forall \theta \in \mathbb{R}^p\right) \geq 1 - e^{-c_1 n}. \quad (96)$$

As a consequence, with probability at least  $1 - e^{-c_1 n}$ , we have

$$\mathcal{L}(\hat{\theta}^t) \geq \mathcal{L}(\hat{\theta}) \geq \mathcal{L}(\hat{\theta}^t) + \langle \nabla_{\theta} \mathcal{L}(\hat{\theta}^t), \hat{\theta} - \hat{\theta}^t \rangle + \frac{1}{2} c_0 \|\hat{\theta} - \hat{\theta}^t\|_2^2. \quad (97)$$

Hence using Cauchy-Schwartz

$$\|\hat{\theta} - \hat{\theta}^t\|_2 \leq \frac{2}{c_0} \|\nabla_{\theta} \mathcal{L}(\hat{\theta}^t)\|_2. \quad (98)$$

The last step of the proof consists in showing that, almost surely

$$\lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{p} \|\nabla_{\theta} \mathcal{L}(\hat{\theta}^t)\|_2^2 = 0. \quad (99)$$

In order to prove this claim, reconsider Eq. (9), for time  $t + 1$ , with  $b_t = b_*$ . Using the fact that  $\Psi(z; b_*) = z - \text{Prox}(z; b_*)$ , this can be rewritten as

$$\text{Prox}(R^t; b_*) = Y - \mathbf{X} \hat{\theta}^{t+1} + R^t - R^{t+1}. \quad (100)$$

By Eq. (11), and recalling that  $\Psi(z; b) = b \rho'(\text{Prox}(z; b))$ , we have

$$\frac{1}{b_* \delta} (\hat{\theta}^{t+1} - \hat{\theta}^t) = \mathbf{X}^{\top} \rho'(\text{Prox}(R^t; b_*)) \quad (101)$$

$$= \mathbf{X}^{\top} \rho'(Y - \mathbf{X} \hat{\theta}^{t+1} + R^t - R^{t+1}), \quad (102)$$

where the last identity followed by Eq. (100). Using the triangle inequality and noting that, by the smoothness assumption  $C \equiv \sup_{z \in \mathbb{R}} \rho''(z) < \infty$ , we get

$$\|\mathbf{X}^{\top} \rho'(Y - \mathbf{X} \hat{\theta}^{t+1})\|_2 \leq \frac{1}{b_* \delta} \|\hat{\theta}^{t+1} - \hat{\theta}^t\|_2 + C \|\mathbf{X}\|_2 \|R^t - R^{t+1}\|_2. \quad (103)$$

Hence, using Eqs (91) and (91), and recalling that  $\lim_{n \rightarrow \infty} \|\mathbf{X}\|_2 < \infty$  almost surely [AGZ09], we get

$$\lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{p} \|\mathbf{X}^{\top} \rho'(Y - \mathbf{X} \hat{\theta}^{t+1})\|_2^2 = 0. \quad (104)$$

This is equivalent to the claim (99) since  $\nabla_{\theta} \mathcal{L}(\theta) = -\mathbf{X} \rho'(Y - \mathbf{X} \theta)$ .

## C.1 Proof of Lemma C.1

First of all note that, due to Lemma B.2, it is sufficient to prove that

$$\lim_{n \rightarrow \infty} \frac{1}{p} \sum_{i=1}^p \xi_2(\vartheta_i^t, \vartheta_i^s) =_{a.s.} \mathbb{E} \xi_2(\sqrt{\delta} Z_t, \sqrt{\delta} Z_s), \quad (105)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_2(s_i^t, S_i^s) =_{a.s.} \mathbb{E} \xi_2(Z_t, Z_s). \quad (106)$$

Note that a similar statement is proved in [BM12, Theorem 4.2] for characterizing the Lasso estimator. While the same argument can be followed here, we outline an alternative argument that is based on a reduction to the setting of [JM12].

We fix an even number  $q \in \mathbb{N}$ , and will prove the claim for all  $t, s \leq T \equiv (q/2) - 1$ . Let  $N \equiv n + p$ . For  $t \in \{0, \dots, T\}$ , we introduce a vector  $z^t \in (\mathbb{R}^q)^N$ , which we think of as a vector with entries in  $\mathbb{R}^q$ :  $z^t = (\mathbf{z}_1^t, \dots, \mathbf{z}_N^t)$   $\mathbf{z}_i^t \in \mathbb{R}^q$ . Its entries are defined as follows:

$$\mathbf{z}_i^t = (S_i^0, 0, S_i^1, 0, S_i^2, 0, \dots, S_i^t, 0, 0, 0, \dots, 0) \quad \text{if } 1 \leq i \leq n, \quad (107)$$

$$\mathbf{z}_i^{t+1} = (0, \vartheta_j^1, 0, \vartheta_j^2, 0, \vartheta_j^3, \dots, 0, \vartheta_i^{t+1}, 0, 0, 0, \dots, 0) \quad \text{if } n+1 \leq i = j+n \leq n+p. \quad (108)$$

Further, we let  $A \in \mathbb{R}^{N \times N}$  be a symmetric matrix with  $A_{ii} = 0$ ,  $A_{ij} = \sqrt{n/N} X_{i,j-n}$  for  $1 \leq i \leq n$  and  $n+1 \leq j \leq n+p$ , and all the other entries  $A_{ij}$   $i < j$  i.i.d.  $\mathcal{N}(0, 1/N)$ . It is then easy to see that the iteration in Eqs. (66), (67) is equivalent to the following

$$z^{t+1} = A f(z^t; t) - \mathbf{B}_t f(z^{t-1}; t-1). \quad (109)$$

Here, for each  $t$ ,  $f(\cdot; t) : (\mathbb{R}^q)^N \rightarrow (\mathbb{R}^q)^N$  is separable in the following sense

$$f(z; t) = (f_1(\mathbf{z}_1; t); f_2(\mathbf{z}_2; t); \dots; f_N(\mathbf{z}_N; t)), \quad (110)$$

with  $f_i(\cdot; t) : \mathbb{R}^q \rightarrow \mathbb{R}^q$ . These are defined as follows (letting  $\Psi_{t,i}(x) = \Psi(W_i + x; b_t)$  and  $h = \sqrt{(1+\delta)/\delta}$ )

$$f_i(\mathbf{z}_i; t) = (0, \delta h \Psi_{0,i}(\mathbf{z}_{i,1}), 0, \delta h \Psi_{1,i}(\mathbf{z}_{i,3}), 0, \dots, 0, \delta h \Psi_{t,i}(\mathbf{z}_{i,2t-1}), 0, 0, 0, \dots, 0) \quad \text{if } 1 \leq i \leq n, \quad (111)$$

$$f_i(\mathbf{z}_i; t) = (0, 0, -h \mathbf{z}_{i,2}, 0, -h \mathbf{z}_{i,4}, 0, \dots, -h \mathbf{z}_{i,2t}, 0, 0, 0, 0, \dots, 0) \quad \text{if } n+1 \leq i \leq n+p. \quad (112)$$

The matrix multiplication in Eq. (109) operates in the natural way over  $(\mathbb{R}^q)^N$ , namely we identified  $A$  with the Kronecker product  $A \otimes \mathbf{I}_{q \times q}$ . Explicitly, Eq. (109) reads

$$\mathbf{z}_i^{t+1} = \sum_{j \in [N]} A_{ij} f_j(\mathbf{z}_j^t; t) - \mathbf{B}_t f_i(\mathbf{z}_i^{t-1}; t-1). \quad (113)$$

Finally  $\mathbf{B}_t \in \mathbb{R}^{q \times q}$  is given by

$$\mathbf{B}_t = \frac{1}{N} \sum_{i=1}^N \frac{\partial f_i}{\partial \mathbf{z}}(\mathbf{z}_i^t; t). \quad (114)$$

The recursion (109) is characterized in [JM12, Theorem 1], which establishes –for instance– that, for  $\xi : \mathbb{R}^q \rightarrow \mathbb{R}$  pseudo-Lipschitz, we have, almost surely,

$$\lim_{N \rightarrow \infty} \frac{1}{p} \sum_{i=n+1}^{n+p} \xi(\mathbf{z}_i^t) = \mathbb{E}\{\xi(\mathbf{Z}_t)\}. \quad (115)$$

Here  $\mathbf{Z}_t$  is a Gaussian random vector whose covariance is fully specified in [JM12]. The proof of the lemma is finished by comparing the expressions in [JM12] for the covariance with the ones in the statement of the lemma.

## C.2 Proof of Lemma C.2

First of all we introduce the notation  $q_t \equiv \Gamma_{t,t+1}/\tau_*^2$ . We then have the recursion

$$q_{t+1} = \mathbf{H}(q_t), \quad (116)$$

$$\mathbf{H}(q) = \frac{\delta}{\tau_*^2} \mathbb{E}_q\{\Psi(W + \tau_* Z_1; b_*)\Psi(W + \tau_* Z_2; b_*)\}, \quad (117)$$

where expectation  $\mathbb{E}_q$  is with respect to the centered Gaussian vector  $(Z_1, Z_2)$  with  $\mathbb{E}_q\{Z_1^2\} = \mathbb{E}_q\{Z_2^2\} = 1$  and  $\mathbb{E}_q\{Z_1 Z_2\} = q$ , independent of  $W \sim F_W$ . We claim that:

- (i)  $\mathbf{H}(1) = 1$ ;
- (ii)  $\mathbf{H}(q)$  is increasing for  $q \in [0, 1]$ ;
- (iii)  $\mathbf{H}(q)$  is strictly convex for  $q \in [0, 1]$ .

In order to prove (i), note that, for  $q = 1$ ,  $Z_1 = Z_2 \equiv Z \sim \mathbf{N}(0, 1)$  and hence

$$\mathbf{H}(1) = \frac{\delta}{\tau_*^2} \mathbb{E}_q\{\Psi(W + \tau_* Z; b_*)^2\}, \quad (118)$$

which is equal to 1 since  $b_*, \tau_*$  satisfy Eq. (24).

In order to prove (ii), (iii), define

$$h_W(z) \equiv \Psi(W + \tau_* z; b_*), \quad (119)$$

$$\mathcal{H}(q) \equiv \mathbb{E}_q\{h_W(Z_1)h_W(Z_2)|W\}, \quad (120)$$

We will prove that  $\mathcal{H}$  is strictly increasing and convex for any  $W$ , whence claims (ii) and (iii) follow by linearity. The argument is the same as in [BM12, Lemma C.1] Let  $\{X_t\}_{t \geq 0}$  be the stationary Ornstein–Uhlenbeck process with covariance  $\mathbb{E}(X_0 X_t) = e^{-t}$ , and denote by  $\mathbf{E}$  expectation with respect to  $X$ . Then

$$\mathcal{H}(q) = \mathbf{E}\{h_W(X_0)h_W(X_t)\}\Big|_{t=\log(1/q)}, \quad (121)$$

Then we have the spectral representation (for  $t = \log(1/q)$ )

$$\mathcal{H}(q) = \sum_{\ell=0}^{\infty} c_\ell^2 e^{-\ell t} = \sum_{\ell=0}^{\infty} c_\ell^2 q^\ell, \quad (122)$$

whence the claim follows since  $c_\ell \neq 0$  for some  $\ell \geq 2$  as long as  $h_W(x)$  is non-linear.

Because of the remarks (i)-(iii) just proven, it follows that  $\lim_{t \rightarrow \infty} q_t = 1$  (and hence  $\lim_{t \rightarrow \infty} \Gamma_{t,t+1} = \tau_*^2$ ) if and only if  $H'(1) \leq 1$ . A simple calculation yields

$$H'(1) = \delta \mathbb{E}\{\Psi'(W + \tau_* Z; b_*)^2\}, \quad (123)$$

where  $Z \sim \mathcal{N}(0, 1)$ . Recalling that  $\Psi'(z; b) \in (0, 1)$ , we have  $(\Psi')^2 \leq \Psi'$  and so

$$H'(1) \leq \delta \mathbb{E}\{\Psi'(W + \tau_* Z; b_*)\} = 1, \quad (124)$$

where the last identity follows because  $(\tau_*, b_*)$  solve Eq. (25). This finishes the proof.

## References

- [AGZ09] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Cambridge University Press, 2009.
- [Bic75] Peter J Bickel, *One-step huber estimates in the linear model*, Journal of the American Statistical Association **70** (1975), no. 350, 428–434.
- [BLM12] Mohsen Bayati, Marc Lelarge, and Andrea Montanari, *Universality in polytope phase transitions and message passing algorithms*, arXiv:1207.7321 (2012).
- [BM11] 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.
- [BM12] ———, *The LASSO risk for gaussian matrices*, IEEE Trans. on Inform. Theory **58** (2012), 1997–2017.
- [CD95] S.S. Chen and D.L. Donoho, *Examples of basis pursuit*, Proceedings of Wavelet Applications in Signal and Image Processing III (San Diego, CA), 1995.
- [Cha03] C-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, Springer, 2003.
- [DJMM11] David Donoho, Iain Johnstone, Arian Maleki, and Andrea Montanari, *Compressed sensing over  $\ell_p$ -balls: Minimax mean square error*, Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on, IEEE, 2011, pp. 129–133.
- [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.
- [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.
- [DT09] David Donoho and Jared Tanner, *Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **367** (2009), no. 1906, 4273–4293.

- [EKBBL13] Nouredine El Karoui, Derek Bean, Peter J Bickel, and Bin Lim, Chingwayand Yu, *On robust regression with high-dimensional predictors*, Proceedings of the National Academy of Sciences **110** (2013), no. 36, 14557–14562.
- [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.
- [Hub73] Peter J Huber, *Robust regression: asymptotics, conjectures and monte carlo*, The Annals of Statistics **1** (1973), no. 5, 799–821.
- [JM12] A. Javanmard and A. Montanari, *State Evolution for General Approximate Message Passing Algorithms, with Applications to Spatial Coupling*, arXiv:1211.5164v1, 2012.
- [KMZ13] Florent Krzakala, Marc Mézard, and Lenka Zdeborová, *Phase diagram and approximate message passing for blind calibration and dictionary learning*, arXiv preprint arXiv:1301.5898 (2013).
- [LDSP08] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly, *Compressed sensing mri*, IEEE Signal Processing Magazine **25** (2008), 72–82.
- [MB07] Mokshay Madiman and Andrew Barron, *Generalized entropy power inequalities and monotonicity properties of information*, Information Theory, IEEE Transactions on **53** (2007), no. 7, 2317–2329.
- [MM09] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford, 2009.
- [Mon12] A. Montanari, *Graphical Models Concepts in Compressed Sensing*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012.
- [MPV87] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond*, World Scientific, 1987.
- [Ran11] S. Rangan, *Generalized Approximate Message Passing for Estimation with Random Linear Mixing*, IEEE Intl. Symp. on Inform. Theory (St. Perersbourg), August 2011.
- [Ric05] M.A. Richards, *Fundamentals of Radar Signal Processing*, McGraw-Hill, 2005.
- [Sca97] J.A. Scales, *Theory of Seismic Imaging*, Samizdat Press, 1997.
- [Sch10] P. Schniter, *Turbo Reconstruction of Structured Sparse Signals*, Proceedings of the Conference on Information Sciences and Systems (Princeton), 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 Aereospace and Electronics Conference (Dayton, OH), 2010.
- [ST03] Mariya Shcherbina and Brunello Tirozzi, *Rigorous solution of the gardner problem*, Communications in mathematical physics **234** (2003), no. 3, 383–422.

- [Tal10] M. Talagrand, *Mean field models for spin glasses: Volume i*, Springer-Verlag, Berlin, 2010.
- [Tib96] R. Tibshirani, *Regression shrinkage and selection with the Lasso*, J. Royal. Statist. Soc B **58** (1996), 267–288.
- [Ver12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012, pp. 210–268.