*The Annals of Statistics*
2020, Vol. 48, No. 5, 3090–3111
https://doi.org/10.1214/19-AOS1923


# ASYMPTOTIC RISK AND PHASE TRANSITION OF $l_1$-PENALIZED ROBUST ESTIMATOR

BY HANWEN HUANG

*Department of Epidemiology and Biostatistics, University of Georgia, huanghw@uga.edu*

Mean square error (MSE) of the estimator can be used to evaluate the performance of a regression model. In this paper, we derive the asymptotic MSE of $l_1$-penalized robust estimators in the limit of both sample size $n$ and dimension $p$ going to infinity with fixed ratio $n/p \to \delta$. We focus on the $l_1$-penalized least absolute deviation and $l_1$-penalized Huber's regressions. Our analytic study shows the appearance of a sharp phase transition in the two-dimensional sparsity-undersampling phase space. We derive the explicit formula of the phase boundary. Remarkably, the phase boundary is identical to the phase transition curve of LASSO which is also identical to the previously known Donoho–Tanner phase transition for sparse recovery. Our derivation is based on the asymptotic analysis of the generalized approximation passing (GAMP) algorithm. We establish the asymptotic MSE of the $l_1$-penalized robust estimator by connecting it to the asymptotic MSE of the corresponding GAMP estimator. Our results provide some theoretical insight into the high-dimensional regression methods. Extensive computational experiments have been conducted to validate the correctness of our analytic results. We obtain fairly good agreement between theoretical prediction and numerical simulations on finite-size systems.

## 1. Introduction.

1.1. *Motivation*. Consider the problem of reconstructing $\boldsymbol{\beta}_0 \in \mathbf{R}^p$ from the measurements

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix and $\boldsymbol{\varepsilon}$ denotes random noise which has zero-mean components $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ i.i.d. with distribution $p_\varepsilon$. The $l_1$-penalized least square regression, also called LASSO [29], is one of the successful methods for estimating $\boldsymbol{\beta}_0$. The performance of LASSO has been studied in the literature by evaluating the upper bound of its mean square error (MSE). For instance, [11] prove that the MSE of LASSO estimator is bounded by the size of the error multiplying by a constant. These types of results are very robust but suffer from loose constants and cannot provide quantitative recommendations in practice.

Inspired by the seminal work of [16], researchers have started performing asymptotic analyses of LASSO under the setting $n, p \to \infty$ with fixed ratio $n/p \to \delta$. These type of analyses can provide sharp quantitative guidelines because they allow to derive exact high-dimensional limit for the LASSO risk [4]. One interesting result in this direction is the phase transition of the LASSO minimax risk which is defined as the minimum of the worst-case MSE of LASSO estimator over the regularization parameter. Let $k = \|\boldsymbol{\beta}_0\|_0$ denote the number of nonzero elements of $\boldsymbol{\beta}_0$ and $\epsilon = k/p$ denote the sparsity rate. It was shown in [17] that the LASSO minimax risk exhibits a phase transition in the two-dimensional phase space $(\delta, \epsilon) \in [0, 1]^2$. More specifically, a curve $\delta = \delta_c(\epsilon)$ was explicitly computed to divide the phase space into

Received December 2018; revised October 2019.
*MSC2020 subject classifications.* Primary 62J05, 62J07; secondary 62H12.
*Key words and phrases.* Mean square error, minimax, penalized, phase transition, robust.

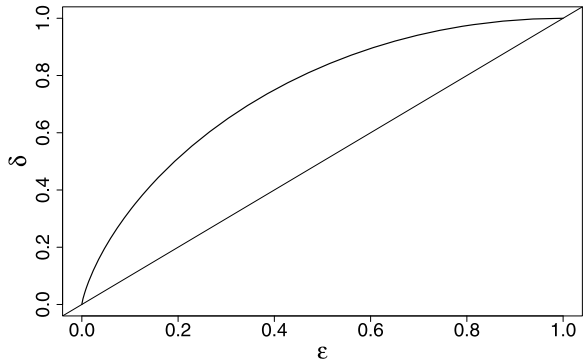


FIG. 1. *LASSO minimax risk phase transition in the plane* $(\epsilon, \delta)$. *The solid curve represents the phase transition boundary* $\delta = \delta_c(\epsilon)$.

two components as shown in Figure 1. The LASSO minimax risk is bounded in the region $\delta > \delta_c(\epsilon)$ and unbounded in the region $\delta < \delta_c(\epsilon)$. Remarkably, the phase boundary $\delta = \delta_c(\epsilon)$ is identical to the previously known phase transition curve in the problem of reconstructing the underdetermined linear systems in compressed sensing in the $k$-sparse noiseless case [13].

The least square loss is efficient for normal distributed errors and homogeneous data. However, data subject to heavy-tailed errors or outliers are commonly encountered in applications. In this case, the least square estimation is inefficient and can be biased. To overcome this problem, robust estimators such as those based on the least absolute deviation (LAD) or Huber-type losses can be useful. Toward this end, penalized robust regression methods have been proposed in the literature to handle the robustness and sparsity simultaneously. Examples include [22, 32] among many others. Substantial efforts in this field have been devoted to developing efficient algorithms for solving the optimization problem and characterizing the performances of the estimators in low-dimensional setting. However, the characterization of the penalized robust estimators in the high-dimensional setting has not been explored much.

The objective of this paper is to derive the asymptotic MSE of penalized robust estimators. We focus on $l_1$-penalized LAD and $l_1$-penalized Huber's estimators. Using the results of asymptotic MSE, we study the phase diagram and associated transitions which describe the undersampling sparsity trade-off for the reconstruction of signal using penalized robust estimators. We will show that the phase boundary is identical to the phase transition curve of LASSO in Figure 1. Our study can provide insight in the theory of regression models.

Our analysis is based on the application of the generalized approximate message passing (GAMP) algorithm to the problem of penalized robust regression. GAMP is a recently developed iterative algorithm by [24] which is a generalization of the approximate message passing (AMP) algorithm and can handle not only least square loss but also more general convex loss functions. The advantage of the GAMP framework is that its asymptotic expression can be explicitly described by the state evolution equations at each iteration. By showing that the GAMP estimators converge to the corresponding penalized robust estimators in the large system limit, we derive the asymptotic MSE of the penalized robust estimator by using state evolution of the corresponding GAMP estimators. All analytical results are confirmed by extensive numerical experiments on finite-size systems and our formulas are clarified to work well even for moderate-size systems.

1.2. *Related work.* This phase-transition curve shown in Figure 1 was originally derived in [13] by methods in combinatorial geometry. Donoho et al. [16] rederived this boundary by applying the AMP algorithm to the LASSO problem in the noiseless case. General analysis of phase transition for AMP was presented in [14] for both the noiseless and noisy cases.

In the high-dimensional regime of $n, p \to \infty$ with $n/p = \delta > 1$, [5, 12, 18] examine the exact stochastic representation for the distribution of nonpenalized robust estimators.

Bradic [8] proposed a sparse approximate message passing (RAMP) algorithm and explored the effects of model selection on the estimation of asymptotic MSE for the $l_1$-penalized robust estimators in the setting of both $p < n$ and $p > n$. However, the convergence of the RAMP estimators to the solution of the penalized robust estimation problem has not been proved completely. Moreover, they did not investigate the noise sensitivity phase transition features of the penalized robust estimators. The prediction of phase transitions for different denoisors besides the soft thresholding one has been studied in [14], but the results are still based on least square loss.

## 2. Asymptotic behavior of $l_1$-penalized robust estimator.

2.1. *$l_1$-Penalized robust estimator.* For problem (1.1), we are interested in estimating $\boldsymbol{\beta}_0$ using the following $l_1$-penalized robust estimators:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.1}$$

where $\lambda > 0$ is the tuning parameter for penalty. Here, the loss function $\rho : \mathbb{R} \to \mathbb{R}_+$ is a nonnegative convex function. Examples of the loss function include:

- Least square loss: $\rho(u) = \frac{1}{2}u^2$.
- Least absolute deviation (LAD) loss: $\rho(u) = |u|$.
- Huber's loss:

$$\rho(u) = \begin{cases} \dfrac{u^2}{2}, & |u| \le \gamma, \\ \gamma|u| - \dfrac{\gamma^2}{2}, & |u| > \gamma, \end{cases}$$

where $\gamma > 0$ is a fixed positive constant.

The regression model based on least square loss is also called LASSO and has been well studied. Here, we consider the robust regression model based on LAD loss and Huber's loss. The robust regression is an alternative to least square regression when the data subject to heavy-tailed errors or are contaminated with outliers. It can also be used for the purpose of detecting influential observations. Given a vector $\mathbf{v} \in \mathbb{R}^p$ and a scalar function $f : \mathbb{R} \to \mathbb{R}$, we write $f(\mathbf{v})$ for the vector obtained by applying $f$ componentwise. Further, $\langle \mathbf{v} \rangle = p^{-1} \sum_{i=1}^{p} v_i$ is the average of the vector $\mathbf{v}$, and $\mathbf{X}^T$ is the transpose of matrix $\mathbf{X}$.

2.2. *Generalized approximate message passing algorithm.* Many algorithms have been developed in the literature to solve the optimization problem (2.1). Here, we use the generalized approximate message passing algorithm (GAMP). Our goal is to study the asymptotic behavior of the regularized robust estimators (2.1) in the limit of $n, p \to \infty$ with fixed ratio $n/p = \delta$. We start from the asymptotic behavior of the corresponding GAMP estimators in the large system limit which can be well characterized by a simple set of state evolution equations. Then we show that the regularized robust estimators asymptotically converge to the corresponding GAMP estimators.

Approximate message passing algorithm (AMP) is a recently developed optimization method for solving the LASSO minimization problem [16]. The advantage of the AMP framework is that it provides an exact expression for the asymptotic MSE of the LASSO estimator instead of an upper bound. AMP has been extended to GAMP in [24] for solving

general convex optimization problem. In order to apply GAMP to the problem (2.1), we first need to define the following two functions:

$$\eta(a,b)=\operatorname*{argmin}_{\beta}\left\{|\beta|+\frac{1}{2b}(\beta-a)^2\right\}=\big(|a|-b\big)_+\operatorname{sign}(a), \tag{2.2}$$

$$G_0(a,b)=b\partial\rho\big(\hat{u}(a,b)\big), \tag{2.3}$$

where $\partial\rho(\cdot)$ represents the subgradient of $\rho(\cdot)$ function and

$$\hat{u}(a,b)=\operatorname*{argmin}_{u}\left\{\rho(u)+\frac{1}{2b}(u-a)^2\right\}. \tag{2.4}$$

Based on the definition of $G_0(a,b)$, it can be easily shown that

$$G_0(a,b)=a-\hat{u}(a,b). \tag{2.5}$$

For the LAD loss $\rho(u)=|u|$, (2.4) gives

$$\hat{u}(a,b)=\operatorname{sign}(a)\big(|a|-b\big)_+,$$

which leads to

$$G_0(a,b)=\begin{cases}a & |a|\le b,\\ b & a>b,\\ -b & a<-b.\end{cases} \tag{2.6}$$

For Huber's loss, (2.4) gives

$$\hat{u}(a,b)=\begin{cases}\dfrac{a}{1+b} & |a|\le(1+b)\gamma,\\ a-\gamma b & a>(1+b)\gamma,\\ a+\gamma b & a<-(1+b)\gamma\end{cases}$$

which leads to

$$G_0(a,b)=\begin{cases}\dfrac{b}{1+b}a & |a|\le(1+b)\gamma,\\ \gamma b & a>(1+b)\gamma,\\ -\gamma b & a<-(1+b)\gamma.\end{cases} \tag{2.7}$$

Let $\{\theta_t,a_t,\pi_t,\omega_t\}_{t\ge0}$ denote four sequences of nonnegative parameters. Starting with initial conditions $\boldsymbol{\beta}^0=0\in\mathbb{R}^p$, $a_0=1$ and $G_0(\mathbf{z}^{-1},a_{-1})=0\in\mathbb{R}^n$, the general form of GAMP algorithm for (2.1) constructs a sequence of estimates $\boldsymbol{\beta}^t\in\mathbb{R}^p$ and residuals $\mathbf{z}^t\in\mathbb{R}^n$ according to the iteration

$$\begin{cases}\mathbf{z}^t=\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^t+\dfrac{1}{\omega_{t-1}}G_0\big(\mathbf{z}^{t-1},a_{t-1}\big)\pi_{t-1},\\ \boldsymbol{\beta}^{t+1}=\eta\left(\boldsymbol{\beta}^t+\dfrac{1}{\omega_t}\mathbf{X}^TG_0\big(\mathbf{z}^t,a_t\big),\theta_t\right),\end{cases} \tag{2.8}$$

with the parameters $\theta_t,a_t,\pi_t,\omega_t$ updated through

$$\begin{cases}\pi_t=\dfrac{1}{\delta}\left\langle\partial_1\eta\left(\boldsymbol{\beta}^{t-1}+\dfrac{1}{\omega_{t-1}}\mathbf{X}^TG_0\big(\mathbf{z}^{t-1},a_{t-1}\big),\theta_{t-1}\right)\right\rangle,\\ a_t=\dfrac{a_{t-1}\pi_{t-1}}{\omega_{t-1}},\\ \omega_t=\big\langle\partial_1G_0\big(\mathbf{z}^t,a_t\big)\big\rangle,\\ \theta_t=\dfrac{\lambda a_t}{\omega_t},\end{cases} \tag{2.9}$$

where $\partial_1$ represents the derivative over the first argument of the function. A detailed derivation of (2.8) and (2.9) is provided in the Supplementary Material [19]. The connection of (2.8) and (2.9) to the $l_1$-penalized robust estimator (2.1) can be formalized by the proposition below.

PROPOSITION 2.1. *Let $(\boldsymbol{\beta}^\star, \mathbf{z}^\star)$ be a fixed point of the iteration (2.8) for $\theta_t = \theta_\star$, $a_t = a_\star$, $\pi_t = \omega_t = \pi_\star$ fixed. Then $\boldsymbol{\beta}^\star$ is a minimum of the cost function (2.1) for*

$$\lambda = \frac{\theta_\star \pi_\star}{a_\star}. \tag{2.10}$$

As a consequence of this proposition, if the estimates $\boldsymbol{\beta}^t$ based on (2.8) and (2.9) converge, then we are guaranteed that the limit is a $l_1$-penalized robust estimator.

Although GAMP has been successfully applied to many problems, its convergence is still not fully understood [25, 31]. Various modification procedures have been proposed to improve the convergence of GAMP; see, for example, [26, 28]. To facilitate the convergent study, here we fix certain parameters throughout the iteration to its fixed-point values. For this purpose, we take $\omega_t = \pi_\star$, the fixed point of $\pi_t$, and choose $\theta_t$ in a way that will be discussed in Section 2.4. Let

$$G(a, b) = \frac{1}{\pi_\star} G_0(a, b) \tag{2.11}$$

denote the rescaled min regularized effective score function, we eventually take the following form of GAMP algorithm:

$$\begin{cases} \mathbf{z}^t = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^t + \dfrac{1}{\delta} G(\mathbf{z}^t, a_t) \langle \partial_1 \eta(\boldsymbol{\beta}^t + \mathbf{X}^T G(\mathbf{z}^t, a_t), \theta_t) \rangle, \\ \boldsymbol{\beta}^t = \eta(\boldsymbol{\beta}^{t-1} + \mathbf{X}^T G(\mathbf{z}^{t-1}, a_{t-1}), \theta_{t-1}), \end{cases} \tag{2.12}$$

where the sequence $\{a_t\}_{t \geq 0}$ is determined by

$$\langle \partial_1 G(\mathbf{z}^t, a_t) \rangle = 1. \tag{2.13}$$

The RAMP algorithm proposed in [8] chooses $\omega_t = \|\boldsymbol{\beta}_0\|_0/(p\delta)$, where $\|\boldsymbol{\beta}_0\|_0$ denotes the number of nonzero elements in the true coefficient vector. The advantage of the choice $\omega_t = \pi_\star$ over other choices is that it is more relevant for mathematical analysis. Particularly, as it will be discussed below, it allows us to establish the convergence of GAMP in a more convenient way.

2.3. *State evolution of GAMP.* It has been shown in [3] that AMP algorithm has several unique advantages. Particularly, the asymptotic limit of the AMP estimates as $n, p \to \infty$ for any fixed $t$ can be described by the state evolution (SE). The SE not only predicts the evolution of numerical statistical properties of $\boldsymbol{\beta}^t$ with the iteration number $t$, it also correctly predicts the success/failure to converge to the correct result.

We will show that the GAMP algorithm enjoys the same properties. We will consider sequences of instances of increasing sizes which are completely determined by the measurement matrix $\mathbf{X}$, the signal $\boldsymbol{\beta}_0$, and the error vector $\boldsymbol{\varepsilon}$. We assume the following conditions in order for SE to hold.

ASSUMPTION 1. $n/p \to \delta \in (0, \infty)$.

ASSUMPTION 2. The empirical distribution of the entries of $\boldsymbol{\beta}_0$ converges weakly to a probability measure $p_{\beta_0}$ on $\mathbb{R}$ with bounded second moment. Further $\frac{1}{p}\sum_{i=1}^p \beta_{0,i}^2 \to E_{\beta_0}(\beta_0^2)$.

ASSUMPTION 3. The empirical distribution of the entries of $\boldsymbol{\varepsilon}$ converges weakly to a probability measure $p_\varepsilon$ on $\mathbb{R}$ with bounded second moment. Further $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 \to E_\varepsilon(\varepsilon^2)$.

ASSUMPTION 4. The entries of $\mathbf{X}$ are i.i.d. normal with mean 0 and variance $1/n$.

Note that the hypothesis of Gaussian measurement matrix $\mathbf{X}$ (Assumption 4) is necessary for the proof technique to be applicable. Extensive numerical simulations carried out in [17] showed that, for LASSO, the result is universal over a broader class of i.i.d. matrices. Recently, [6] generalizes the SE result for AMP to standard Gaussian design with nonseparable denoisers. This work enables the applicability of AMP to Gaussian design with nontrivial covariance $\boldsymbol{\Sigma}$ and a separable denoiser by a change of variable $\tilde{\mathbf{X}} \to \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$, $\tilde{\boldsymbol{\beta}}_0 \to \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0$. However, the challenge of this extension is that the convergence of AMP under general non-i.i.d. matrices has not been fully understood. In fact, recent works in [25, 31] have shown that, even for the simplest least square loss, GAMP can diverge under mildly ill-conditioned $\mathbf{X}$.

We say a function $\psi : \mathbb{R}^k \to \mathbb{R}$ is pseudo-Lipschitz if there exists a constant $L > 0$ such that for all $\mathbf{x}, \mathbf{y} \in R^k$ : $|\psi(\mathbf{x}) - \psi(\mathbf{y})| \le L(1 + \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)\|\mathbf{x} - \mathbf{y}\|_2$. The following proposition is a simple application of the existing results in [8]. It shows that the GAMP iterations (2.12) and (2.13) admit a high-dimensional limit as $n, p \to \infty$ with fixed $n/p = \delta$. The proof is still included in the Appendix for the completeness of the paper.

PROPOSITION 2.2. *Let $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a pseudo-Lipschitz function. Then, under Assumptions 1–4, almost surely*

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \psi\big(\beta_i^{t+1}, \beta_{0,i}\big) = E\big\{\psi\big(\eta(\beta_0 + \tau_t Z, \theta_t), \beta_0\big)\big\},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\big(z_i^t, \varepsilon_i\big) = E\big\{\psi(\varepsilon + \sigma_t Z, \varepsilon)\big\},$$

*where $Z \sim N(0,1)$ is independent of $\beta_0 \sim p_{\beta_0}$ and $\varepsilon \sim p_\varepsilon$. The state evolution sequences $\{\tau_t^2, \sigma_t^2\}_{t\ge 0}$ are obtained by the following iterative equations*:

$$\tau_t^2 = E\big\{G(\varepsilon + \sigma_t Z, a_t)^2\big\}, \tag{2.14}$$

$$\sigma_t^2 = \frac{1}{\delta} E\big\{\big(\eta(\beta_0 + \tau_{t-1} Z, \theta_{t-1}) - \beta_0\big)^2\big\}, \tag{2.15}$$

*with the parameters $a_t$ determined by*

$$E\big\{\partial_1 G(\varepsilon + \sigma_t Z, a_t)\big\} = 1. \tag{2.16}$$

According to the standard weak convergence arguments, Proposition 2.2 indicates that the empirical distribution of the entries of the GAMP estimator $\boldsymbol{\beta}^t$ converges weakly to the distribution of the random variable $\eta(\beta_0 + \tau_{t-1} Z, \theta_{t-1})$ with $Z \sim N(0,1)$ independent of $\beta_0$. Similarly, the empirical distribution of the entries of the residuals $\mathbf{z}^t$ converges weakly to the distribution of the random variable $\varepsilon + \sigma_t Z$ with $Z \sim N(0,1)$ independent of $\varepsilon$.

2.4. *Fixed-point equations and convergence.* Define $\tau_\star$, $\sigma_\star$, $a_\star$ and $\theta_\star$ the solutions of the SE fixed-point equations

$$\tau_\star^2 = E\big\{G(\varepsilon + \sigma_\star Z, a_\star)^2\big\}, \tag{2.17}$$

$$\sigma_\star^2 = \frac{1}{\delta} E\big\{\big(\eta(\beta_0 + \tau_\star Z, \theta_\star) - \beta_0\big)^2\big\}, \tag{2.18}$$

$$1 = E\big\{\partial_1 G(\varepsilon + \sigma_\star Z, a_\star)\big\}. \tag{2.19}$$

Clearly, the solutions depend on $\delta$ as well as the distributions $p_{\beta_0}$ and $p_\varepsilon$. Then the quantity $\pi_\star$ can be obtained as

$$\pi_\star = \frac{1}{\delta} E\{\partial_1 \eta(\beta_0 + \tau_\star Z, \theta_\star)\}. \tag{2.20}$$

Using the explicit forms of (2.6) and (2.7), we obtain the following proposition.

PROPOSITION 2.3. *Define $\tilde{Z} = \varepsilon + \sigma_\star Z$. Then the SE fixed-point equations of $l_1$-LAD-GAMP are*

$$\tau_\star^2 = \frac{1}{\pi_\star^2}\big[E\{\tilde{Z}^2 I(|\tilde{Z}| \le a_\star)\} + E\{a_\star^2 I(|\tilde{Z}| \ge a_\star)\}\big], \tag{2.21}$$

$$\pi_\star = p(|\tilde{Z}| \le a_\star) = \frac{1}{\delta} E\{\partial_1 \eta(\beta_0 + \tau_\star Z, \theta_\star)\}. \tag{2.22}$$

*The SE fixed-point equations of $l_1$-Huber-GAMP are*

$$\begin{aligned}\tau_\star^2 = \frac{1}{\pi_\star^2}\bigg[&E\bigg\{\frac{a_\star^2}{(1+a_\star)^2}\tilde{Z}^2 I(|\tilde{Z}| \le (1+a_\star)\gamma)\bigg\} \\ &+ E\{a_\star^2 \gamma^2 I(|\tilde{Z}| \ge (1+a_\star)\gamma)\}\bigg],\end{aligned} \tag{2.23}$$

$$\pi_\star = \frac{a_\star}{1+a_\star} p(|\tilde{Z}| \le (1+a_\star)\gamma) = \frac{1}{\delta} E\{\partial_1 \eta(\beta_0 + \tau_\star Z, \theta_\star)\}. \tag{2.24}$$

Combining (2.14) and (2.15), we obtain the one-dimensional update for $\tau_t^2$ as

$$\tau_{t+1}^2 = V(\tau_t^2, \theta_t),$$

where

$$V(\tau^2, \theta) = E\{G(\varepsilon + \sigma(\tau, \theta)Z, a(\tau, \theta))^2\}, \tag{2.25}$$

$$\sigma(\tau, \theta)^2 = \frac{1}{\delta} E\{(\eta(\beta_0 + \tau Z; \theta) - \beta_0)^2\} \tag{2.26}$$

with $a(\tau, \theta)$ implied by

$$E\{\partial_1 G(\varepsilon + \sigma(\tau, \theta)Z, a(\tau, \theta))\} = 1.$$

Now we discuss the choice of the sequence of thresholds $\theta_t$. We take $\theta_t = \alpha\tau_t$ with $\alpha$ fixed throughout the iterations. As discussed in [4], the main advantage of such choice is that the convergence of the corresponding recursion $\tau_{t+1}^2 = V(\tau_t^2, \alpha\tau_t)$ can be well established. Moreover, it is a natural choice from an intuitive point of view. At each step, we apply the soft thresholding denoiser $\eta(\cdot, \theta_t)$ to an effective observation $\beta_0 + \tau_t Z$ which can be regarded as the signal $\beta_0$ corrupted by Gaussian noise $\tau_t Z$. Therefore, this suggests to choose $\theta_t$ proportional to the standard deviation of the noise $\tau_t$. More discussion about the choice of $\theta_t$ was given in [23].

Let $\alpha_l = \alpha_l(\delta)$ be the unique nonnegative solution of the equation

$$2\Phi(-\alpha) = \delta, \tag{2.27}$$

where $\Phi(z) = \int_{-\infty}^{z} \phi(x)\,dx$ and $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the standard Gaussian density function. The following proposition indicates the convergence of the SE equations (2.14) and (2.15).

PROPOSITION 2.4. *For any $\sigma^2 > 0$ and $\alpha > \alpha_l$, the fixed-point equation*

$$\tau^2 = V(\tau^2, \alpha\tau), \tag{2.28}$$

*admits at least one solution. Denoting by $\tau_\star^2 = \tau_\star^2(\alpha)$ the largest solution, we have $\lim_{t\to\infty} \tau_t^2 = \tau_\star^2(\alpha)$ for large enough initial condition $\tau_{t=0}^2$.*

2.5. *Connection of GAMP to regularized robust estimator.* Before stating our main results, we have to describe a calibration mapping between $\alpha$ and $\lambda$ which will depend on $p_{\beta_0}$. For $p_{\beta_0}$, we consider the sparse class

$$\mathcal{F}_\epsilon \equiv \{\nu : \nu \text{ is a probability measure with } \nu(\{0\}) \geq 1 - \epsilon\} \tag{2.29}$$

which put mass at least $1 - \epsilon$ on 0. Let $\alpha_u = \alpha_u(\delta)$ be the unique nonnegative solution of the equation

$$\epsilon + 2(1 - \epsilon)\Phi(-\alpha) = \delta.$$

Clearly, $\alpha_l < \alpha_u$ for $\epsilon > 0$. We then define the function $\alpha \to \lambda(\alpha)$ on $(0, \infty)$ by

$$\lambda(\alpha) = \frac{\alpha\tau_\star(\alpha)\pi_\star(\alpha)}{a_\star(\alpha)}. \tag{2.30}$$

This function defines a correspondence between the threshold $\alpha\tau_\star$ and the regularization parameter $\lambda$. We need to invert this function and define $\alpha : (0, \infty) \to (0, \infty)$ in such a way that

$$\alpha(\lambda) = \{a \in (0, \infty) : \lambda(a) = \lambda\}. \tag{2.31}$$

The next result implies that the set on the right-hand side is nonempty and therefore the function $\lambda \to \alpha(\lambda)$ is well defined.

PROPOSITION 2.5. *For any $\sigma^2 > 0$, there exist an $\alpha_{\min} \in [\alpha_l, \alpha_u]$ which depends on $p_{\beta_0}$ such for any $\alpha > \alpha_{\min}$, the function $\alpha \to \lambda(\alpha)$ is continuous on the interval $(\alpha_{\min}, \infty)$ with $\lambda(\alpha_{\min}+) = 0$ and $\lim_{\alpha\to\infty} \lambda(\alpha) = \infty$. Therefore, the function $\lambda \to \alpha(\lambda)$ satisfying (2.31) exists.*

State evolution provides the limit of GAMP estimation in the high-dimensional setting. By showing that the GAMP estimator converges to the regularized robust estimator, one can obtain the distributional limit for the latter as well. The following theorem shows that the empirical distribution of the entries of the regularized robust estimator $\hat{\boldsymbol{\beta}}$ from (2.1) converges weakly to the distribution of the random variable $\eta(\beta_0 + \tau_\star Z; \theta_\star)$ with $Z \sim N(0, 1)$ independent of $\beta_0$.

THEOREM 2.1. *Under Assumptions 1–4, denote by $f(x)$ the density function for the distribution of error term $\varepsilon$. Further assume that for any $\omega > 0$, there exists a large enough $C > 0$ such that $f(x) > 0$ for all $x \in [-C, C]$ and the probability $p(x \in [-C, C]) \geq 1 - \omega$. Let $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a pseudo-Lipschitz function. Then, almost surely*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi(\hat{\beta}(\lambda)_i, \beta_{0,i}) = E\{\psi(\eta(\beta_0 + \tau_\star Z, \alpha(\lambda)\tau_\star), \beta_0)\},$$

*where $Z \sim N(0, 1)$ is independent of $\beta_0 \sim p_{\beta_0}$, $\tau_\star = \tau_\star(\alpha(\lambda))$ is the solution of the fixed-point equation (2.28) with $\alpha = \alpha(\lambda)$.*

Theorem 2.1 allows us to theoretically study the MSE of the regularized robust estimator. Using function $\psi(a,b)=(a-b)^2$, we obtain

$$\lim_{p\to\infty}\frac{1}{p}\|\hat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}_0\|^2=E\{[\eta(\beta_0+\tau_\star Z,\alpha(\lambda)\tau_\star)-\beta_0]^2\},$$

which depends on $\delta$, $\lambda$, $p_{\beta_0}$ and $p_\varepsilon$. Therefore, the asymptotic risk of the regularized robust estimator can be determined for any specific distributions $p_{\beta_0}$ and $p_\varepsilon$ by solving the fixed-point equation (2.28). Note that Theorem 2.1 allows us to predict MSE for any fixed $\lambda$. This is different from the traditional large $n$ fixed $p$ situations where the usual value for $\lambda$ is chosen growing to 0. We prove Theorem 2.1 by proving the following result.

THEOREM 2.2. *Under the same assumptions used for Theorem 2.1, we have*

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\|\boldsymbol{\beta}^t-\hat{\boldsymbol{\beta}}\|_2^2=0,$$

*almost surely.*

After establishing the convergence of the state evolution, the proof technique of Theorem 2.2 is very similar to the existing proof techniques used in LASSO paper [4]. For completeness, we outline the main proof in the Appendix and move the technical lemmas into the Supplementary Material.

**3. Phase transition and minimax risk.** In this section, we study the minimax risk and phase transition properties of the $l_1$-penalized robust estimator based on the asymptotic results derived in Section 2. Here, the minimax risk refers to minimizing the worst-case MSE over $\lambda$ for estimators based on a specific $l_1$-penalized robust method. It is not a minimax over all possible estimators. We start from the fixed-point equations (2.17)–(2.20) and focus on two special robust regression methods: $l_1$-LAD and $l_1$-Huber's regression. The results depend on distributions $p_\varepsilon$ and $p_{\beta_0}$. We consider the sparse class $\mathcal{F}_\epsilon$ defined in (2.29) for $p_{\beta_0}$ and have a phase space $0\le\epsilon,\delta\le 1$ expressing different combinations of under-sampling $\delta$ and sparsity $\epsilon$.

3.1. *Noiseless $l_1$-LAD.* Let us first consider the noiseless case, that is, $\sigma_0^2=0$. We study under which condition the original signal $\boldsymbol{\beta}_0$ can be correctly reconstructed from the measurement $\mathbf{y}$ using *$l_1$-LAD* regression after appropriately tuning the $\lambda$.

It is well known that in the problem of reconstructing the underdetermined linear system, exact reconstruction takes place subject to a trade-off between under-sampling $\delta$ and sparsity $\epsilon$ [13]. There is a function $\delta_c(\epsilon)$ whose graph partitions the domain $(\epsilon,\delta)\in[0,1]^2$ into two regions, a "success" region, where exact reconstruction occurs, and a "failure" region where exact reconstruction fails. In the lower region, where $\delta<\delta_c(\epsilon)$, the probability of exact reconstruction tends to zero as $k,n,p\to\infty$ with $k/p\to\epsilon$ and $n/p\to\delta$. In the upper region, where $\delta>\delta_c(\epsilon)$, the corresponding probability of exact reconstruction tends to one. Hence the curve $\delta=\delta_c(\epsilon)$ for $0<\delta<1$ indicates the precise trade-off between under-sampling and sparsity.

Note that, for LASSO, $\delta_c(\epsilon)$ is independent of the actual signal distribution $p_{\beta_0}$. This is different from the $l_p$-penalized regressions with $0\le p<1$ studied in [34] in which $p_{\beta_0}$ has a substantial effect on the phase transition curve. As shown in [16], the curve $\delta_c(\epsilon)$ admits the following simple form:

$$\delta_c(\epsilon)=\frac{2\phi(\alpha_c)}{\alpha_c+2(\phi(\alpha_c)-\alpha_c\Phi(-\alpha_c))},\tag{3.1}$$

where $\alpha_c$ is determined by

$$\epsilon = \frac{2(\phi(\alpha_c) - \alpha_c \Phi(-\alpha_c))}{\alpha_c + 2(\phi(\alpha_c) - \alpha_c \Phi(-\alpha_c))}, \tag{3.2}$$

where $\alpha_c \in [0, \infty)$ is the parameter.

The following theorem shows that a phase transition also occurs in the $l_1$-LAD regression for noiseless case. The domain has two phases: a "success" phase, where the $l_1$-LAD regression succeeds to recover $\boldsymbol{\beta}_0$, and a "failure" phase where it fails to reconstruct $\boldsymbol{\beta}_0$. Moreover, this phase transition boundary is exactly $\delta = \delta_c(\epsilon)$.

THEOREM 3.1. *Under Assumptions 1, 2 and 4, denote $\hat{\boldsymbol{\beta}}$ the estimator from (2.1) based on the LAD loss in the noiseless case. For any $\delta > \delta_c(\epsilon)$, we can tune the parameter $\lambda$ and have $\lim_{p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = 0$ almost surely. Thus we can make consistent estimation for the original signal $\boldsymbol{\beta}_0$ in this region. For any $\delta < \delta_c(\epsilon)$, we have $\lim_{p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 > 0$ almost surely for any tuning parameter $\lambda$. Thus the consistent estimation in this region fails.*

Theorem 3.1 shows that, in the "success" region, we can tune $\alpha$ such that the equation (2.28) admits a unique solution $\tau_\star = 0$ which corresponds to $\lambda = 0$ for the original problem (2.1) according to the calibration mapping (2.30).

3.2. *Noisy $l_1$-LAD.* Next, we study the MSE of $\hat{\boldsymbol{\beta}}$ in the noisy case, that is, $\sigma_0^2 \neq 0$. In this case, the probability of exact reconstruction tends to zero and the MSE result depends on $p_{\beta_0}$. In the more realistic situation, we do not know $p_{\beta_0}$. However, if we consider the sparsity class $\mathcal{F}_\epsilon$ and take the worst case MSE over this class and then minimize over $\lambda$, we get a result that is independent of $p_{\beta_0}$. Toward this end, we define the minimax risk as

$$M(\delta, \epsilon) = \min_{\lambda} \sup_{p_{\beta_0} \in \mathcal{F}_\epsilon} \lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \{|\hat{\beta}_i(\lambda) - \beta_{0,i}|^2\}. \tag{3.3}$$

That is, the regularization parameter $\lambda$ is optimally chosen such that the maximal MSE based on a specific $l_1$-penalized robust method for the class $\mathcal{F}_\epsilon$ is minimized.

The following theorem shows that in the presence of noise, the phase space $0 \leq \delta, \epsilon \leq 1$ is partitioned by the curve $\delta = \delta_c(\epsilon)$ into two regions. The minimax risk of $l_1$-LAD is bounded throughout the "success" region and unbounded throughout the "failure" region.

THEOREM 3.2. *Under Assumptions 1–4 with the condition of bounded second moment for $p_{\beta_0}$ in Assumption 2 removed, recall that $M(\delta, \epsilon)$ defined in (3.3) denotes the minimax risk of $l_1$-LAD. Then, for any $\delta > \delta_c(\epsilon)$, $M(\delta, \epsilon)$ is bounded; for any $\delta < \delta_c(\epsilon)$, $M(\delta, \epsilon)$ is unbounded.*

Note that the optimal choice of regularization parameter for minimizing the maximum risk is related to $\alpha_c$ provided by (3.1). The upper bounds for the minimax risk of the sparse regression estimators for (1.1) have been studied in the literature; see, for example, [7, 9, 27, 30]. There is a lower bound result in [30] which shows that the lower bound of estimating $\boldsymbol{\beta}$ in high dimension is $C_1 \frac{\epsilon}{\delta}\{1 + \log(1/\epsilon)\}$ for $1 \leq k \leq (n-1)/4$ and $\boldsymbol{\Sigma} = \mathbf{I}$ (see (6.7) in Proposition 6.4 in [30]). Our results complement this type of "rough and robust" bounds by providing asymptotic formal expression for the MSE of $\hat{\boldsymbol{\beta}}$.

Based on Theorem 3.2, in the upper region $\delta > \delta_c(\epsilon)$, we can estimate the actual minimax risk of $l_1$-LAD. Denote

$$M^\star(\epsilon) = \delta_c(\epsilon) \tag{3.4}$$

which is defined in (3.1). It was shown in [17] that the LASSO minimax risk is simply given by

$$M(\delta,\epsilon)=\frac{\sigma_0^2 M^\star(\epsilon)}{1-M^\star(\epsilon)/\delta} \tag{3.5}$$

which does not depend on the distribution $p_\varepsilon$. In contrast, the explicit form of minimax risk of $l_1$-LAD depends on the distribution $p_\varepsilon$. We consider three different distributions here.

PROPOSITION 3.1. *Normal random error. Assume that the noise term $\varepsilon$ follows a normal distribution $\varepsilon\sim N(0,\sigma_0^2)$. Then for $\delta>\delta_c(\epsilon)$, the minimax MSE of $l_1$-LAD is*

$$M(\delta,\epsilon)=\frac{\sigma_0^2 F(c)M^\star(\epsilon)}{1-F(c)M^\star(\epsilon)/\delta}, \tag{3.6}$$

*where*

$$F(c)=\frac{2\Phi(c)-1+2c(c\Phi(-c)-\phi(c))}{(2\Phi(c)-1)^2}, \tag{3.7}$$

*where $c$ depends on $(\epsilon,\delta)$ through $2\Phi(c)-1=\frac{M^\star(\epsilon)}{\delta}$.*

Clearly, $l_1$-LAD risk (3.6) is larger than the corresponding LASSO risk (3.5) because $F(c)\geq 1$. It is interesting to check that

$$F(c)M^\star(\epsilon)/\delta\leq 1.$$

To show this, notice that $\frac{M^\star(\epsilon)}{\delta}=2\Phi(c)-1$ which leads to

$$F(c)M^\star(\epsilon)/\delta=\frac{2\Phi(c)-1+2c(c\Phi(-c)-\phi(c))}{2\Phi(c)-1}\leq 1$$

because $c\Phi(-c)\leq\phi(c)$ for any $c\geq 0$. Therefore, for normal distributed $\varepsilon$, $l_1$-LAD minimax risk is larger than LASSO minimax risk. This is consistent with the classical statistical analysis, that is, least square loss is optimal for Gaussian errors.

PROPOSITION 3.2. *Laplace random error. Assume that the noise term $\varepsilon$ follows a Laplace distribution*

$$\varepsilon\sim\frac{1}{2b_0}\exp\left(-\frac{|\varepsilon|}{b_0}\right)$$

*which has mean 0 and variance $2b_0^2$. Denote $c=\frac{a_\star}{\sqrt{\sigma_0^2+\sigma_\star^2}}$ and $b=\frac{b_0}{\sigma_\star}$. Then the minimax MSE of $l_1$-LAD is $\tau_\star^2 M^\star(\epsilon)$, where $\tau_\star^2$ (together with $c$) is determined by the equations*

$$\begin{cases}\tau_\star^2=\left(2b_0^2+\dfrac{\tau_\star^2 M^\star(\epsilon)}{\delta}\right)B(b,c),\\ \dfrac{\tau_\star M^\star(\epsilon)}{\delta}=D(b,c),\end{cases} \tag{3.8}$$

*where*

$$\begin{aligned}B(b,c)&=\frac{1}{D(b,c)}+\frac{N(b,c)+c^2(1-D(b,c))}{D(b,c)^2(1+2b^2)},\\ D(b,c)&=\Phi(c)-\Phi(-c)+f_1(b,c)-f_2(b,c),\\ N(b,c)&=\big(c^2-1\big)\big(f_1(b,c)-f_2(b,c)\big)-2bc\big(f_1(b,c)+f_2(b,c)\big)-2c\Phi(c),\end{aligned}$$

$$f_1(b,c) = \exp\left\{\frac{1}{2}\left(\frac{1}{b^2} + \frac{2c}{b}\right)\right\}\Phi\left(-c - \frac{1}{b}\right),$$
$$f_2(b,c) = \exp\left\{\frac{1}{2}\left(\frac{1}{b^2} - \frac{2c}{b}\right)\right\}\Phi\left(c - \frac{1}{b}\right).$$

In the classical statistical theory for $p$ fixed and $n \to \infty$, it is well known that MLE based LAD loss is optimal for Laplace errors. Actually, from (2.21) and (2.22), we obtain that as $\delta \to \infty$, $l_1$-LAD $M(\delta, \epsilon) \to \frac{\sigma_0^2 M^\star(\epsilon)}{4\varphi(0)}$, where $\varphi(x)$ is the density function of $\varepsilon/\sigma_0$. Comparing it to (3.5), we conclude that, in contrast to LASSO, the minimax MSE of $l_1$-LAD is larger for normal error but smaller for Laplace distribution error. It is interesting to see that this conclusion is independent of the choice of the denoiser. Because from the fixed-point equations (2.17) and (2.18), different denoisers lead to different $\eta(\cdot)$ function, and thus different $M^\star(\epsilon)$ but the result of $\tau_\star$ will not be affected if $\delta$ is large enough.

But the above conclusion is no longer true in the regime when $p$ is large and comparable to $n$. Specifically, our numerical studies in Section 4 show that, in this regime, the results depend on $\delta$. For very small $\delta$, LASSO is better than $l_1$-LAD; but as $\delta$ increases, $l_1$-LAD eventually outperforms LASSO and yields smaller MSE. This observation is due to the extra Gaussian noise $\sigma_\star Z$ which is dominant over $\varepsilon$ at very small $\delta$ and can be negligible comparing to $\varepsilon$ when $\delta$ is large enough.

PROPOSITION 3.3. *Gaussian mixture random error. Assume that the noise term $\varepsilon$ follows a mixture of two component Gaussian distribution $\varepsilon \sim \epsilon_1 N(0, \sigma_1^2) + \epsilon_2 N(0, \sigma_2^2)$ with $\epsilon_1 + \epsilon_2 = 1$. Denote $c_1 = a_\star/\sigma_1$ and $c_2 = a_\star/\sigma_2$. Then the minimax MSE of $l_1$-LAD is $\tau_\star^2 M^\star(\epsilon)$, where $\tau_\star^2$ (together with $c_1$ and $c_2$) is determined by the equations*

$$\begin{cases} \tau_\star^2 = \dfrac{\epsilon_1(\sigma_\star^2 + \sigma_1^2) f_b(c_1) + \epsilon_2(\sigma_\star^2 + \sigma_2^2) f_b(c_2)}{\{\epsilon_1(2\Phi(c_1) - 1) + \epsilon_2(2\Phi(c_2) - 1)\}^2}, \\ \dfrac{\tau_\star M^\star(\epsilon)}{\delta} = \epsilon_1\big(2\Phi(c_1) - 1\big) + \epsilon_2\big(2\Phi(c_2) - 1\big), \end{cases} \tag{3.9}$$

*where*

$$f_b(c) = 2\Phi(c) - 1 + 2c\big\{c\Phi(-c) - \phi(c)\big\}.$$

There is no closed form solution for $\tau_\star^2$ and we have to use numerical methods. In Section 4, we have also performed extensive numerical studies to compare minimax MSE of $l_1$-LAD with LASSO for Gaussian mixture distributed $\varepsilon$.

3.3. *Penalized Huber's regression*. The phase transition and minimax risk of $l_1$-penalized Huber's regression can be studied using the same procedure as we did for $l_1$-LAD. The following theorem shows that the same phase transition also occurs in $l_1$-Huber regression for both noiseless and noisy cases. This phase transition boundary is exactly $\delta = \delta_c(\epsilon)$.

THEOREM 3.3. *Under Assumptions 1, 2 and 4, denote $\hat{\boldsymbol{\beta}}$ the estimator from (2.1) based on the Huber loss. For any $\delta > \delta_c(\epsilon)$, by tuning the parameter $\lambda$, we can have $\lim_{p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = 0$ almost surely. Thus we can make consistent estimation for the original signal $\boldsymbol{\beta}_0$ in this region. For any $\delta < \delta_c(\epsilon)$, we have $\lim_{p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 > 0$ almost surely for any tuning parameter $\lambda$. Thus the consistent estimation in this region fails.*

We have the following theorem in the presence of measurement noise, that is, $\sigma_0^2 \neq 0$.

THEOREM 3.4. *Under Assumptions 1–4 with the condition of bounded second moment for $p_{\beta_0}$ in Assumption 2 removed, recall that $M(\delta,\epsilon)$ denotes the minimax MSE of $l_1$-Huber. Then, for any $\delta > \delta_c(\epsilon)$, $M(\delta,\epsilon)$ is bounded; for any $\delta < \delta_c(\epsilon)$, $M(\delta,\epsilon)$ is unbounded.*

For a given $\delta$, $\epsilon$ satisfying $\delta > \delta_c(\epsilon)$, we can estimate the corresponding minimax risk of $l_1$-Huber. The results depend on $\delta$ as well as the form of the error distribution $p_\varepsilon$. We have derived the explicit formulas under three different error distributions: normal, Laplace and Gaussian mixture. The following proposition shows the result for normal error.

PROPOSITION 3.4. *Denote $c_0 = \frac{(1+a_\star)\gamma}{\sqrt{\sigma_0^2+\sigma_\star^2}}$. Assume that the noise term $\varepsilon$ follows a normal distribution $\varepsilon \sim N(0,\sigma_0^2)$. Then for $\delta > \delta_c(\epsilon)$, the minimax MSE of $l_1$-Huber is*

$$M(\delta,\epsilon) = \frac{\sigma_0^2 F(c_0) M^\star(\epsilon)}{1 - F(c_0)M^\star(\epsilon)/\delta}, \tag{3.10}$$

*where $F(c_0)$ is defined in (3.7) and $\frac{a_\star}{1+a_\star}(2\Phi(c_0)-1) = \frac{M^\star(\epsilon)}{\delta}$.*

Comparing (3.10) with (3.6) and (3.5), it can be shown that, for normal random error, the minimax of MSE $l_1$-Huber is less than $l_1$-LAD but larger than LASSO due to the fact that $F(c)$ is decreasing with $c$ and $F(c) \geq 1$. The minimax MSE formulas of $l_1$-Huber for Laplace and Gaussian mixture errors are complicated but the derivation is very straightforward and similar to the ones for $l_1$-LAD. To save space, we will not show the details here.

3.4. *Technical novelties.* Since the least squares loss is strongly convex and the LAD and Huber losses are not, our main results in Section 2 cannot be seen as a straightforward extension of the results in [4] for LASSO research. For example, the proof of Theorem 2.1 is much more sophisticated than the proof of the corresponding LASSO result in [4], and requires nontrivial extensions. The main reason is that, according to (2.13), the least square loss can lead to a constant $a_t$, and thus a strictly concave function $\tau^2 \mapsto V(\tau^2, \alpha\tau)$. This substantially simplify the convergence analysis of the SE. On contrary, for LAD and Huber losses, $a_t$ is not constant and the form of $\tau^2 \mapsto V(\tau^2,\alpha\tau)$ is quite complicated and not concave. As a consequence, the proof of the SE convergence for GAMP is much more difficult than for AMP. To overcome this difficulty, as suggested by one reviewer, we first construct a modified GAMP procedure by appropriately fixing certain parameters throughout the iteration to the final values they will converge to. Then, instead of using concavity, we prove the convergence of $\tau_t$ by exploring the large $\tau^2$ behavior of $V(\tau^2,\alpha\tau)$. An addition level of complexity due to nonstrongly convex loss comes from the proof of Lemma S2. For example, (S1) can be obtained immediately for strongly convex loss $\rho(\cdot)$ as shown in [12]. However, for nonstrongly convex loss, we have to develop new techniques in Lemma S1 to prove it. Moreover, in Section 3, we study the phase transition phenomenon of the penalized robust estimators. Some techniques used in the proofs of Theorems 3.1–3.3 are outside the scope of current AMP results. To the best of our knowledge, all previous rigorous analyses of phase transition were for the case of least square loss.

## 4. Numerical results.

4.1. *Comparison between theoretical prediction and simulation on finite-size systems.* In this section, we conduct Monte Carlo simulations to test the validity of our analytical estimation and to determine the finite-size effect. We first confirm that our theoretical results presented in Sections 2 and 3 are reliable. For this purpose, we focus on the comparison of

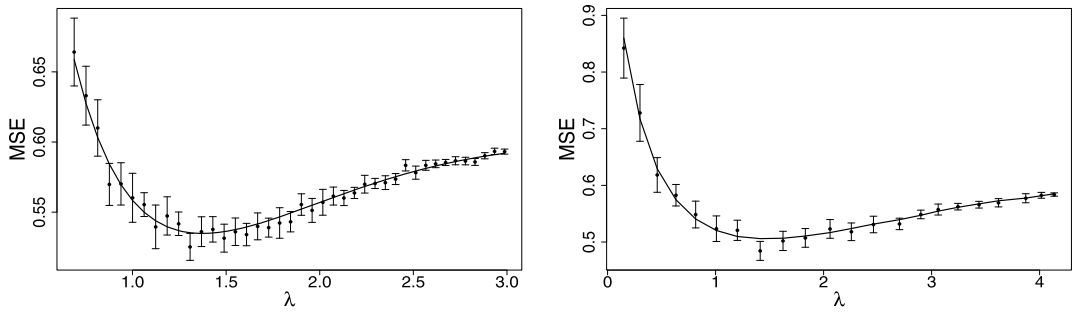


FIG. 2. *Comparison between theoretical estimation and simulation study for the change of MSE against tuning parameter* $\lambda$. *Here,* $\mu = 2$, $\delta = 0.5$, $\epsilon = 0.15$. *The error term follows a Laplace distribution with mean* 0 *and variance* 1. *The simulation is based on* $p = 1000$ *and each setting was repeated* 100 *times. The solid curve represents the theoretical estimation and the error bars represent the mean and* 95% *confidence interval summarized over* 100 *simulated data. Left panel:* $l_1$*-LAD. Right panel:* $l_1$*-Huber with* $\gamma = 1$.

the estimated MSE from theory and the MSE computed from numerical algorithms for finite system. We consider two methods: $l_1$-LAD and $l_1$-Huber. The errors follow a Laplace distribution. The Supplementary Material contains more simulation results with other types of error distributions.

For each setting, we first found the fixed point of the state evolution for $\tau_\star^2$ by numerically solving (2.28) with the corresponding $p_{\beta_0}$. Then using Theorem 2.1 and (2.18), we obtain that $\mathrm{MSE} = \delta\sigma_\star^2$. The comparisons between theoretical estimation and Monte Carlo simulation for MSE of $l_1$-LAD are shown in the left panel of Figure 2. We fix the undersampling and sparsity parameters as $\delta = 0.5$ and $\epsilon = 0.15$. The signal is assumed to follow a three-point distribution $p_{\beta_0} \sim (1-\epsilon)\delta_0 + \frac{\epsilon}{2}\delta_\mu + \frac{\epsilon}{2}\delta_{-\mu}$ with $\mu = 2$. The change of MSE as a function of tuning parameter $\lambda$ is plotted. The dimension of the simulated data $p = 1000$ and we repeat simulation 100 times for each parameter setting. The mean and standard errors over 100 replications are presented. We use R package *quantreg* to fit the $l_1$-LAD estimators. Our analytical curves (solid lines) show a fairly good agreement with the direct computations from numerical algorithms (error bar) for simulated data. Thus our analytical formulas provide reliable estimates for moderate system sizes.

The comparisons between theoretical estimation and Monte Carlo simulation for MSE of $l_1$-Huber with $\gamma = 1$ are shown in the right panel of Figure 2. We use the same parameter settings as we did for $l_1$-LAD. We use R package *hqreg* to solve the $l_1$-Huber optimization problem. Similar to $l_1$-LAD, we obtain fairly good agreement between analytic estimation and simulation study for $l_1$-Huber as well.

4.2. *Phase transition.* For the noiseless case, we compare the theoretical phase transition with the empirical one estimated by applying the $l_1$-LAD algorithm to simulated data. We first fix a grid of 30 $\delta$ values between 0.05 and 1. For each $\delta$, we consider a series of $\epsilon$ values between $\epsilon_c(\delta) - 0.1$ and $\epsilon_c(\delta) + 0.1$, where $\epsilon_c(\delta) = \{\epsilon : \delta_c(\epsilon) = \delta\}$. We then have a grid of $\delta$, $\epsilon$ values in parameter space $[0,1]^2$. At each $\delta$, $\epsilon$, we generate 20 problem instances $(\mathbf{X}, \boldsymbol{\beta}_0)$ with size $p = 1000$. Then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0$. For the $i$th problem instance, we obtain an output $\hat{\boldsymbol{\beta}}_i$ by using the $l_1$-LAD regression method to the $i$th simulated data with $\lambda$ chosen to minimize the MSE. We set the success indicator variable $S_i = 1$ if $\frac{\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0\|_2}{\|\boldsymbol{\beta}_0\|_2} \le 10^{-4}$ and $S_i = 0$ otherwise. Then at each $(\delta, \epsilon)$ combination, we have $S = \sum_{i=1}^{20} S_i$.

We analyze the simulated dataset to estimate the phase transition. At each fixed value of $\delta$ in our grid, we model the dependence of $S$ on $\epsilon$ using logistic regression. We assume that $S$ follows a binomial $B(\pi, 20)$ distribution with $\mathrm{logit}(\pi) = a + b\epsilon$. We define the phase

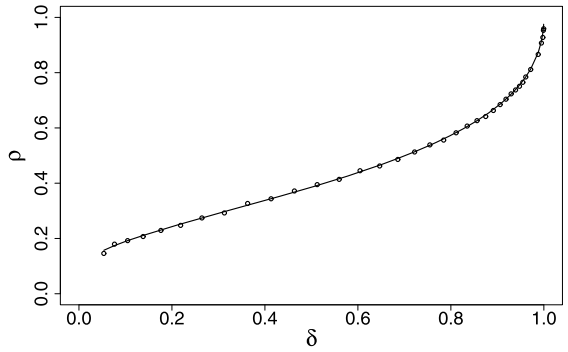


FIG. 3. *Observed phase transition for $l_1$-LAD in the plane $(\delta, \rho)$. Solid curve represents the theoretical result and circle points represent estimation from simulated data on finite-size systems with $p = 1000$.*

transition as the value of $\epsilon$ at which the success probability $\pi = 0.5$. In terms of the fitted parameters $\hat{a}$, $\hat{b}$, we have the estimated phase transition $\hat{\epsilon}(\delta) = -\hat{a}/\hat{b}$. Figure 3 shows that the agreement between the estimated phase transition curve based on the simulated finite-size systems and the analytical curve based on asymptotic theorem is remarkably good. Note that here we follow Donoho–Tanner notation and plot $\rho = \epsilon/\delta$, the number of nonzero elements in the signal per measurement, as a function of $\delta$.

Figure 4 displays the average MSE of $l_1$-LAD over 20 replications as a function of $\epsilon$ at 3 different $\delta$ values. It is apparent that MSE closes to zero for $\epsilon$ below the critical value $\epsilon_c(\delta)$ and is nonzero for $\epsilon$ above the critical value $\epsilon_c(\delta)$. Therefore, we can get exact reconstruction for below but not for above.

Figure 5 shows the location of the noise sensitivity boundary $\rho_c(\delta) = \epsilon_c(\delta)/\delta$ as well as the level lines of $M(\delta, \epsilon)$ for $\rho < \rho_c(\delta)$. The different contour lines show positions in the $\delta$, $\rho$ plane where a given minimax MSE is achieved. It is apparent that the MSE increases dramatically as one approaches the phase boundary. Above $\rho_c(\delta)$, the $l_1$-LAD MSE is not uniformly bounded.

4.3. *Minimax risk.* In the noisy case, for fixed $(\delta, \epsilon)$ with $\delta > \delta_c(\epsilon)$, we compare the estimated minimax risk among different regression methods. Figure 6 displays the estimated minimax MSE based on three different regression methods for errors following standard norm, Laplace and mixture of two component Gaussian distributions. The change of minimax MSE as a function of $\delta$ for fixed $\epsilon$ is plotted.

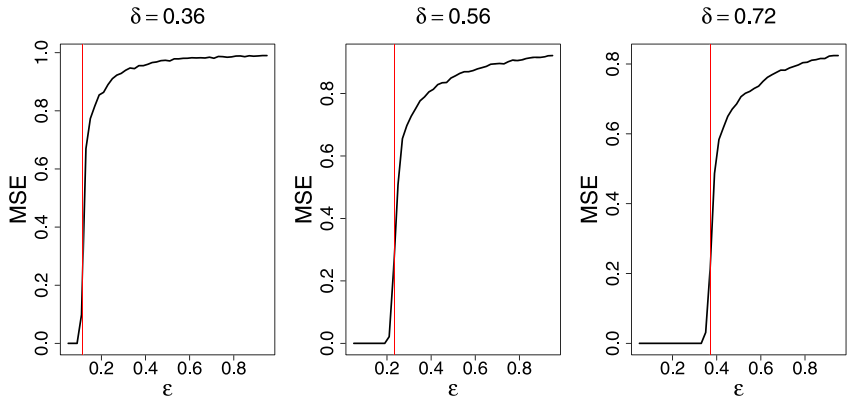


FIG. 4. *Empirical average MSE over 20 replications as a function of $\epsilon$ at fixed $\delta$ for $l_1$-LAD regression on simulated noiseless data. The red vertical lines represent the critical $\epsilon$ values at the corresponding $\delta$ which is equal to $\epsilon_c(\delta)$.*

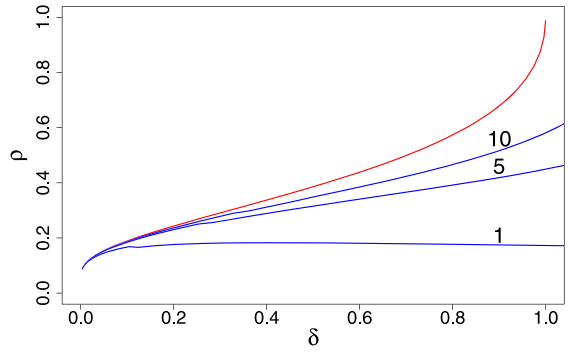


FIG. 5. *Phase diagram for the $l_1$-LAD regression method in the plane $(\delta, \rho)$ for noisy case. Red line*: *The phase transition boundary $\rho = \rho_c(\delta)$. Blue lines*: *Level curves for the $l_1$-LAD minimax risk $M(\delta, \epsilon)$. Notice that $M(\delta, \epsilon) \to \infty$ as $\rho \to \rho_c(\delta)$.*

The left panel of Figure 6 shows that for normally distributed random error, LASSO (blue curve) gives the smallest minimax MSE at all $\delta$ values. But for Laplace or Gaussian mixture distributed errors, LASSO is not the best. As illustrated in the middle panel of Figure 6 for Laplace distributed error, LASSO gives smaller minimax MSE than $l_1$-LAD (red curve) at very small $\delta$ value. When $\delta$ increases, $l_1$-LAD eventually exceeds LASSO and yields smaller minimax MSE. Therefore, the optimal loss is not always the negative log likelihood function in high-dimensional regime. Similar situations happen to Gaussian mixture distributed random error as illustrated in the right panel of Figure 6. The performance of $l_1$-penalized Huber's regression (black curve) depends on the parameter $\gamma$. Here, for $\gamma = 1$, it gives the smallest minimax MSE at all $\delta$ values for Gaussian mixture distributed error. For other two error distributions, its performance is in between. Actually $l_1$-penalized Huber's regression can always achieve the best performance if we tune the parameter $\gamma$ optimally. It is because that Huber's regression is indeed a hybrid of quantile regression and least square regression.

The numerical results shown in Figure 6 are consistent with our theoretical findings in Section 2. Theorem 2.2 indicates that the observed errors follow a mixture of two component distribution including the original error component $\varepsilon$ and an extra Gaussian component $\sigma_\star Z$. Therefore, classical MLE estimation constructed based on the original error distribution cannot always achieve the best performance especially in situations when $\delta \ll 1$. For example, the middle panel of Figure 6 illustrates that LASSO gives a smaller MSE than MLE based $l_1$-LAD when $\delta < 0.2$ even if the error actually follows a Laplace distribution. This phenomenon cannot be explained by classical concepts. But if $\delta$ is large enough, $l_1$-LAD is eventually optimal which is consistent with the classical large $n$ fixed $p$ asymptotic theory.

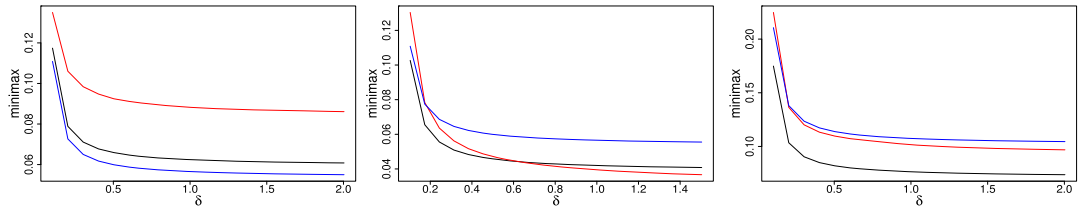


FIG. 6. *Minimax MSE as a function of $\delta$ at fixed $\epsilon = 0.01$. The blue, red and black lines represent the estimation from LASSO, $l_1$-LAD and $l_1$-Huber ($\gamma = 1$), respectively. Left panel*: *standard normal distributed error. Middle panel*: *Laplace distributed error with mean* 0 *and variance* 1. *Right panel*: *Gaussian mixture distributed error with $\varepsilon \sim 0.9N(0, 1) + 0.1N(0, 10)$.*

Our numerical results provide some guidelines for statisticians to decide among different loss functions given the noise density. Although appropriately tuning Huber's regression is the best in all situations, it is often computationally expensive to find the optimal $\gamma$ in practice. Therefore, based on our simulation studies, for normal noise, we recommend LASSO; for heavy tailed errors such as exponential ones, we recommend $l_1$-LAD for large $\delta$ and LASSO for small $\delta$; for mixture of Gaussian errors, we recommend $l_1$-Huber.

**5. Discussion.** In this paper, we study the asymptotic MSE of $l_1$-penalized robust estimators in the framework of GAMP algorithm 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. Our analysis shows the existence of a sharp phase transition in the two-dimensional $(\delta, \epsilon)$ plane which is consistent with the phase transition in LASSO. The analytical calculations are compared with numerical simulations on finite-size systems and the agreement between the data analysis and theoretical prediction is fairly good, and thus, our formulas are validated. Our numerical studies show that when $\delta$ is small enough, least squares loss becomes preferable to LAD loss for Laplace errors. Therefore, the most optimal loss function is no longer the negative log likelihood function. This is because of the extra Gaussian component caused by high-dimensional asymptotics. This new phenomenon was first discovered in [18] and later confirmed in [12] for robust regression. We show here that this phenomenon can be characterized rigorously using GAMP techniques for penalized robust regression as well. Our results can be applied to the case of $k\log(p)/n \to 0$ in practice by letting $\delta \to \infty$ or $\epsilon \to 0$. As shown by Figure S6 in the Supplementary Material, we can obtain fairly good agreement between analytic estimation and simulation study in the case of small $k\log(p)/n$ with $n < p$.

The focus of this paper is on mathematical analysis of the $l_1$-penalized robust regression methods. The analytical predictions are derived based on given signal distribution $p_{\beta_0}$ and error distribution $p_\varepsilon$. However, in practice, these quantities are usually unknown. As shown in [10], even for LASSO, the loss of estimator is hard to estimate based on the data set. There has been some recent work for the case of LASSO that uses Stein's Unbiased Risk Estimator (SURE) to propose unbiased predictions of the risk without using signal distribution $p_{\beta_0}$; see, for example, [1, 23]. Interesting areas for the future include following that direction to provide analytical predictions in real data analysis based on robust estimators.

In additional to the $l_1$-penalization, [14] studied the phase transitions of regularized least square estimators under a wide range of penalizations. Krzakala [21] proposed a probabilistic approach to reconstruct the signal by assuming that the signals follow a parametric Gauss–Bernoulli distribution. Then they add an expectation maximization based procedure to learn the unknown distribution parameters. The phase diagram of this model has also been analyzed and compared with the result based on $l_1$-penalization. One of our future research topics is to study the phase transition of regularized robust regression under other types of penalizations.

An important assumption made for deriving the asymptotic result in GAMP is that the matrix $\mathbf{X}$ has i.i.d. Gaussian entries. The reason is that the rigorous proof for SE of GAMP can only be given for this class of matrices although simulation studies have shown that the results hold for a much broader class of matrices. For least square loss, the SE is proved in [2] under i.i.d. sub-Gaussian matrices. [20] has derived the asymptotic result of LASSO estimators under general Gaussian matrices using nonrigorous replica method. Our another future direction is to rigorously prove this result using the method developed in [6]. For this purpose, the main challenge is to find conditions for $\mathbf{X}$ under which the GAMP can converge to those fixed points.

## APPENDIX A: PROOF OF THEOREM 3.1

PROOF. For $\sigma_0^2 = 0$, denote $c = \frac{a_\star}{\sigma_\star}$ and $\hat{Z} = \frac{\tilde{Z}}{\sigma_\star}$. Then the fixed-point equation (2.28) becomes

$$\tau_\star^2 = V(\tau_\star^2, \alpha\tau_\star) = \sigma_\star^2 F(c), \tag{A.1}$$

where $F(c)$ is defined in (3.7) and $c$ is determined through

$$P(|\hat{Z}| \le c) = \pi_\star = 2\Phi(c) - 1. \tag{A.2}$$

It can be shown that $F(c)$ is a decreasing function with $F(0) = \pi/2$ and $F(\infty) = 1$.

As shown in the proof of Proposition 2.3, $V(\tau^2, \alpha\tau) < \tau^2$ as $\tau \to \infty$. We argue that if the only fixed point satisfies $\tau_\star^2 = 0$, it must be true that the derivative of $V(\tau^2, \alpha\tau)$ at $\tau^2 = 0$ is smaller than or equal to 1. That is,

$$\left.\frac{dV(\tau^2, \alpha\tau)}{d\tau^2}\right|_{\tau^2=0} \le 1$$

for appropriately chosen $\alpha$. Starting from (A.1), we obtain

$$\left.\frac{dV(\tau^2, \alpha\tau)}{d\tau^2}\right|_{\tau^2=0} = \left.\frac{d\sigma^2}{d\tau^2}\right|_{\tau^2=0} F\big(c(\tau)\big) + \sigma_\star^2 F'\big(c(\tau)\big)\left.\frac{dc(\tau)}{d\tau^2}\right|_{\tau^2=0}.$$

Note that

$$\pi_\star = \frac{1}{\delta} E\big\{\Phi(-\alpha + w_\star) + \Phi(-\alpha - w_\star)\big\}, \tag{A.3}$$

where $w_\star = \frac{\beta_0}{\tau}$. Taking derivative on both side of (A.2), we have

$$2\phi(c)\frac{dc}{d\tau^2} = \frac{1}{\delta} E\left\{-\frac{w_\star}{2\tau^2\delta}\big[\phi(-\alpha + w_\star) - \phi(-\alpha - w_\star)\big]\right\} \xrightarrow{\tau^2 \to 0} 0.$$

Therefore, $\frac{dc(\tau)}{d\tau^2}|_{\tau_\star^2=0} = 0$ and for $p_{\beta_0} \in \mathcal{F}$,

$$\begin{aligned}&\left.\frac{dV(\tau^2, \alpha\tau)}{d\tau^2}\right|_{\tau^2=0} \\ &\quad = \big\{\epsilon\big(1+\alpha^2\big) + (1-\epsilon)\big[2\big(1+\alpha^2\big)\Phi(-\alpha) - 2\alpha\phi(\alpha)\big]\big\}\frac{F(c)}{\delta},\end{aligned} \tag{A.4}$$

where we have used (S2). Denote $\delta_{\min}(\epsilon)$ the smallest $\delta$ for a fixed $\epsilon$ such that

$$\min_{\alpha \ge 0}\left\{\left.\frac{dV(\tau^2, \alpha\tau)}{d\tau^2}\right|_{\tau^2=0}\right\} \le 1. \tag{A.5}$$

The terms inside $\{\cdot\}$ on the right-hand side of (A.4) are minimized when we choose $\alpha = \alpha_c$ such that

$$\epsilon = 1 - \frac{\alpha_c}{\alpha_c - 2\alpha_c\Phi(-\alpha_c) + 2\phi(\alpha_c)}. \tag{A.6}$$

Then if we take

$$\delta_{\min}(\epsilon) = \frac{2\phi(\alpha_c)}{\alpha_c - 2\alpha_c\Phi(-\alpha_c) + 2\phi(\alpha_c)}, \tag{A.7}$$

we can show that the second term $F(c)$ on the right-hand side of (A.4) is also minimized at this value. Toward this end, substituting (A.6) and (A.7) into (A.3), we get

$$\pi_\star \xrightarrow{\tau^2 \to 0} \frac{\epsilon + 2(1-\epsilon)\Phi(-\alpha_c)}{\delta_{\min}(\epsilon)} = 1.$$

Therefore, from (A.2), we have $c=\infty$, and thus $F(c)=1$ which is also the smallest value for $F(c)$. Therefore, the smallest $\delta$ that satisfies (A.5) is just the $\delta_{\min}(\epsilon)$ defined in (A.7) which is exactly equal to the $\delta_c(\epsilon)$ defined in (3.1).

Next, we prove that if $\delta>\delta_c(\epsilon)$, we have $V(\tau^2,\alpha_c\tau)<\tau^2$ for any $\tau^2>0$. Thus the unique fixed-point solution is $\tau^2=0$. For $p_{\beta_0}\in\mathcal{F}_\epsilon$, (S2) can be written as

$$\begin{aligned}\sigma^2=\frac{\tau^2}{\delta}\big\{&(1-\epsilon)\big\{2\big(1+\alpha^2\big)\Phi(-\alpha)-2\alpha\phi(\alpha)\big\}+\epsilon\big(1+\alpha^2\big)\\&+\epsilon E\big\{\big(1+\alpha^2-w_\star^2\big)\big[\Phi(-\alpha-w_\star)-\Phi(\alpha-w_\star)\big]\\&-(\alpha+w_\star)\phi(\alpha-w_\star)-(\alpha-w_\star)\phi(\alpha+w_\star)\big\}\big\}.\end{aligned}$$

Plug in (A.6), we obtain

$$\sigma^2=\frac{\tau^2}{\delta}\big\{\delta_c(\epsilon)+\epsilon E\big\{\Phi(-\alpha_c-w_\star)-\Phi(\alpha_c-w_\star)\big\}+\epsilon g(\alpha_c)\big\},\tag{A.8}$$

where

$$\begin{aligned}g(\alpha)=&\big(\alpha^2-w_\star^2\big)E\big\{\Phi(-\alpha-w_\star)-\Phi(\alpha-w_\star)\big\}\\&-(\alpha+w_\star)E\big\{\phi(\alpha-w_\star)\big\}-(\alpha-w_\star)E\big\{\phi(\alpha+w_\star)\big\}.\end{aligned}$$

Since $g(0)=0$ and

$$g'(\alpha)=2\alpha E\big\{\Phi(-\alpha-w_\star)-\Phi(\alpha-w_\star)\big\}-E\big\{\phi(\alpha-w_\star)\big\}-E\big\{\phi(\alpha+w_\star)\big\}$$

which is less than 0 for any $\alpha>0$, therefore, $g(\alpha)<0$ for any $\alpha>0$. On the other hand, for $\alpha=\alpha_c$, from (A.2) and (A.3), we have

$$2\Phi(c)-1=\frac{1}{\delta}\big[\delta_c(\epsilon)+\epsilon E\big\{\Phi(-\alpha_c-w_\star)-\Phi(\alpha_c-w_\star)\big\}\big].$$

From (3.7) and using $c\Phi(-c)<\phi(c)$ for all $c>0$, we get

$$F(c)<\frac{1}{2\Phi(c)-1}=\frac{\delta}{\delta_c(\epsilon)+\epsilon E\{\Phi(-\alpha_c-w_\star)-\Phi(\alpha_c-w_\star)\}}.\tag{A.9}$$

Combining it to (A.1), (A.8) and the fact that $g(\alpha)<0$, we obtain

$$V\big(\tau^2,\alpha_c\tau\big)=\sigma^2F(c)<\tau^2\tag{A.10}$$

for any $\tau^2>0$. □

## APPENDIX B: PROOF OF THEOREM 3.2

PROOF. From Theorem 2.1, we have

$$M(\delta,\epsilon)=\min_{\alpha}\sup_{p_{\beta_0}\in\mathcal{F}_\epsilon}E\big\{\big\|\eta(\beta_0+\tau_\star Z;\alpha\tau_\star)-\beta_0\big\|^2\big\}.$$

Since the class $\mathcal{F}_\epsilon$ is invariant by rescaling, the worst case MSE must be proportional to the only scale in the problem, that is, $\tau_\star^2$. We get

$$M(\delta,\epsilon)=\tau_\star^2M^\star(\epsilon),$$

where

$$M^\star(\epsilon)=\min_{\alpha}\sup_{p_{w_\star}\in\mathcal{F}_\epsilon}E\big[\eta(w_\star+Z,\alpha)-w_\star\big]^2,\tag{B.1}$$

where $w_\star = \beta_0/\tau_\star$. Minimax MSE of soft thresholding was studied in [15, 17, 33] where one can find a considerable amount of information about the behavior of the optimal threshold $\alpha$ and the least favorable distribution $p_{\beta_0} \in \mathcal{F}_\epsilon$. Particularly, the supremum is achieved only by a three-point mixture on the centered real line $\mathbb{R} \cup \{-\infty, \infty\}$:

$$p_{\beta_0} = \frac{\epsilon}{2}\delta_{+\infty} + (1-\epsilon)\delta_0 + \frac{\epsilon}{2}\delta_{-\infty}. \tag{B.2}$$

Then the explicit formula of $M^\star(\epsilon)$ takes the form

$$\begin{aligned} M^\star(\epsilon) &= \min_\alpha \{\epsilon(1+\alpha^2) + (1-\epsilon)[2(1+\alpha^2)\Phi(-\alpha) - 2\alpha\phi(\alpha)]\} \\ &= \frac{2\phi(\alpha_c)}{\alpha_c + 2(\phi(\alpha_c) - \alpha_c\Phi(-\alpha_c))} = \delta_c(\epsilon), \end{aligned} \tag{B.3}$$

where $\alpha_c$ is defined in (3.2). Therefore, the finiteness of $M(\delta, \epsilon)$ depends on $\tau_\star$ which is the solution of fix-point equation $\tau^2 = V(\tau^2, \alpha_c\tau)$.

For $\sigma_0^2 \neq 0$, it is apparent that $V(\tau^2, \alpha_c\tau)|_{\tau^2=0} > 0$. Thus, we can get finite solution for equation $\tau^2 = V(\tau^2, \alpha_c\tau)$ if $V(\tau^2, \alpha_c\tau) < \tau^2$ for large enough $\tau^2$. Let us first determine the phase transition curve by finding the smallest $\delta$ such that $V(\tau^2, \alpha_c\tau) < \tau^2$ as $\tau^2 \to \infty$ for fixed $\epsilon$.

From (S2), we have that $\tau^2 \to \infty$ leads to $\sigma^2 \to \infty$ and $\hat{Z} \to N(0,1)$. In this situation, $V(\tau^2, \alpha_c\tau)$ takes the form

$$V(\tau^2, \alpha_c\tau) \to \frac{\tau^2 M^\star(\epsilon)}{\delta} F(c) \leq \frac{\tau^2 M^\star(\epsilon)}{\delta(2\Phi(c)-1)}, \tag{B.4}$$

where the second step is obtained using (A.9) and the equality only holds for $c = \infty$. From (S3), we get

$$2\Phi(c) - 1 = \frac{1}{\delta}\{\epsilon + 2(1-\epsilon)\Phi(-\alpha)\} \tag{B.5}$$

as $\tau_\star \to \infty$. If we plug in the optimal $\alpha$ from (3.2), we have

$$2\Phi(c) - 1 = \frac{1}{\delta}\frac{2\phi(\alpha_c)}{\alpha_c + 2(\phi(\alpha_c) - \alpha_c\Phi(-\alpha_c))} = \frac{M^\star(\epsilon)}{\delta}. \tag{B.6}$$

For fixed $\epsilon$, plugging it into (B.4), we have $V(\tau^2, \alpha_c\tau) \leq \tau^2$ for large enough $\tau^2$ if $\delta \leq \delta_c(\epsilon) = M^\star(\epsilon)$. If $\delta > \delta_c(\epsilon)$, we can strictly have $V(\tau^2, \alpha_c\tau) < \tau^2$ as $\tau^2 \to \infty$ and thus at least one finite solution $\tau_\star^2$ for the fix-point equation which leads to a finite minimax MSE.

On the other hand, if $\delta < \delta_c(\epsilon)$, we can prove that $V(\tau^2, \alpha_c\tau) > \tau^2$ for any $\tau \geq 0$, thus the only solution is $\tau_\star^2 = \infty$ which leads to an infinite minimax MSE. Toward this end, we start from

$$V(\tau^2, \alpha_c\tau) = E\{G(\tilde{Z}, a)^2\},$$

where $\tilde{Z} = \varepsilon + \sigma Z$ which can be described as $\tilde{Z} = F_\varepsilon \star N(0, \sigma^2)$ (here $\star$ denotes convolution). Denote $\xi_{\tilde{Z}}$ the score function for location of $\tilde{Z}$, then using (2.19) we have $1 = |E_{\tilde{Z}} G'| = |E_{\tilde{Z}} G\xi_{\tilde{Z}}|$. Meanwhile, by Cauchy–Schwarz, $|E_{\tilde{Z}} G\xi_{\tilde{Z}}| \leq \sqrt{E_{\tilde{Z}} G^2}\sqrt{E_{\tilde{Z}}\xi_{\tilde{Z}}^2}$. We conclude that

$$V(\tau^2, \alpha_c\tau) \geq \frac{|E_{\tilde{Z}} G\xi_{\tilde{Z}}|^2}{E_{\tilde{Z}}\xi_{\tilde{Z}}^2} = \frac{|E_{\tilde{Z}} G'|^2}{E_{\tilde{Z}}\xi_{\tilde{Z}}^2} = \frac{1}{I(\tilde{Z})},$$

where $I(\tilde{Z})$ is the Fisher information of $F_{\tilde{Z}}$. From convexity and translation-invariance of Fisher information $I(\tilde{Z}) = I(F_\varepsilon \star N(0, \sigma^2)) < I(N(0, \sigma^2)) = 1/\sigma^2$ if $\operatorname{var}(\varepsilon) = \sigma_0^2 > 0$. Therefore,

$$V(\tau^2, \alpha_c\tau) > \sigma^2(\alpha_c) \overset{(a)}{=} \frac{\tau^2 M^\star(\epsilon)}{\delta} = \frac{\tau^2\delta_c(\epsilon)}{\delta} \tag{B.7}$$

which is larger than $\tau^2$ for $\delta < \delta_c(\epsilon)$. In step (a) of (B.7), we have used (2.17), (B.1) and (B.3). We conclude that the phase transition curve is determined by (B.3), which is exactly the same as the transition curve (A.6) and (A.7) derived in noiseless case. □

**Acknowledgments.** The author thanks the Editor, Associate Editor and four referees for many insightful comments and suggestions which have led to to great improvement of this article. This research is supported in part by Division of Mathematical Sciences (National Science Foundation) Grant DMS-1916411.

## SUPPLEMENTARY MATERIAL

**Supplement: More simulations and proofs** (DOI: 10.1214/19-AOS1923SUPP; .pdf). The supplement provides more simulation results and proofs of Theorems 2.2, 3.3, 3.4 as well as all propositions and lemmas.

## REFERENCES

[1] BAYATI, M., ERDOGDU, M. A. and MONTANARI, A. (2013). Estimating LASSO risk and noise level. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 944–952. Curran Associates.

[2] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[3] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[4] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612

[5] BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568. https://doi.org/10.1073/pnas.1307845110

[6] BERTHIER, R., MONTANARI, A. and NGUYEN, P. (2017). State evolution for approximate message passing with non-separable functions. CoRR abs/1708.03950.

[7] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[8] BRADIC, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electron. J. Stat.* **10** 3894–3944. MR3581957 https://doi.org/10.1214/16-EJS1212

[9] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[10] CAI, T. T. and GUO, Z. (2018). Accuracy assessment for high-dimensional linear regression. *Ann. Statist.* **46** 1807–1836. MR3819118 https://doi.org/10.1214/17-AOS1604

[11] CANDÈS, E. J., ROMBERG, J. K. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846 https://doi.org/10.1002/cpa.20124

[12] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z

[13] DONOHO, D. and TANNER, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4273–4293. MR2546388 https://doi.org/10.1098/rsta.2009.0152

[14] DONOHO, D. L., JOHNSTONE, I. and MONTANARI, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inform. Theory* **59** 3396–3433. MR3061255 https://doi.org/10.1109/TIT.2013.2239356

[15] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

[16] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919. https://doi.org/10.1073/pnas.0909892106

[17] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inform. Theory* **57** 6920–6941. MR2882271 https://doi.org/10.1109/TIT.2011.2165823

[18] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562. https://doi.org/10.1073/pnas.1307842110

[19] HUANG, H. (2020). Supplement to "Asymptotic risk and phase transition of $l_1$-penalized robust estimator." https://doi.org/10.1214/19-AOS1923SUPP.

[20] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629

[21] KRZAKALA, F., MÉZARD, M., SAUSSET, F., SUN, Y. F. and ZDEBOROVÁ, L. (2012). Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X* **2** 021005. https://doi.org/10.1103/PhysRevX.2.021005

[22] LAMBERT-LACROIX, S. and ZWALD, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Stat.* **5** 1015–1053. MR2836768 https://doi.org/10.1214/11-EJS635

[23] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2017). Consistent parameter estimation for LASSO and approximate message passing. *Ann. Statist.* **45** 2427–2454. MR3737897 https://doi.org/10.1214/16-AOS1529

[24] RANGAN, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings* (*ISIT*), 2011 *IEEE International Symposium on* 2168–2172. https://doi.org/10.1109/ISIT.2011.6033942

[25] RANGAN, S., SCHNITER, P. and FLETCHER, A. (2014). On the convergence of approximate message passing with arbitrary matrices. In 2014 *IEEE International Symposium on Information Theory* 236–240. https://doi.org/10.1109/ISIT.2014.6874830

[26] RANGAN, S., SCHNITER, P. and FLETCHER, A. K. (2017). Vector approximate message passing. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 1588–1592.

[27] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

[28] SCHNITER, P., RANGAN, S. and FLETCHER, A. K. (2016). Vector approximate message passing for the generalized linear model. In 2016 50*th Asilomar Conference on Signals, Systems and Computers* 1525–1529. https://doi.org/10.1109/ACSSC.2016.7869633

[29] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[30] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90. MR2879672 https://doi.org/10.1214/12-EJS666

[31] VILA, J., SCHNITER, P., RANGAN, S., KRZAKALA, F. and ZDEBOROVÁ, L. (2015). Adaptive damping and mean removal for the generalized approximate message passing algorithm. In 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) 2021–2025. https://doi.org/10.1109/ICASSP.2015.7178325

[32] WANG, L. (2013). The $L_1$ penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.* **120** 135–151. MR3072722 https://doi.org/10.1016/j.jmva.2013.04.001

[33] ZHANG, C.-H. (2012). Minimax $\ell_q$ risk in $\ell_p$ balls. In *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman. Inst. Math. Stat.* (*IMS*) *Collect.* **8** 78–89. IMS, Beachwood, OH. MR3202504 https://doi.org/10.1214/11-IMSCOLL806

[34] ZHENG, L., MALEKI, A., WENG, H., WANG, X. and LONG, T. (2017). Does $\ell_p$-minimization outperform $\ell_1$-minimization? *IEEE Trans. Inform. Theory* **63** 6896–6935. MR3724407 https://doi.org/10.1109/TIT.2017.2717585