The Annals of Statistics
2018, Vol. 46, No. 6A, 3099–3129
https://doi.org/10.1214/17-AOS1651
© Institute of Mathematical Statistics, 2018

OVERCOMING THE LIMITATIONS OF PHASE TRANSITION BY HIGHER ORDER ANALYSIS OF REGULARIZATION TECHNIQUES[1]

BY HAOLEI WENG, ARIAN MALEKI AND LE ZHENG

Columbia University

We study the problem of estimating a sparse vector $\beta \in \mathbb{R}^p$ from the response variables $y = X\beta + w$, where $w \sim N(0, \sigma_w^2 I_{n\times n})$, under the following high-dimensional asymptotic regime: given a fixed number δ, $p \to \infty$, while $n/p \to \delta$. We consider the popular class of ℓ_q-regularized least squares (LQLS), a.k.a. bridge estimators, given by the optimization problem

$$\hat{\beta}(\lambda, q) \in \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q,$$

and characterize the almost sure limit of $\frac{1}{p}\|\hat{\beta}(\lambda, q) - \beta\|_2^2$, and call it asymptotic mean square error (AMSE). The expression we derive for this limit does not have explicit forms, and hence is not useful in comparing LQLS for different values of q, or providing information in evaluating the effect of δ or sparsity level of β. To simplify the expression, researchers have considered the ideal "error-free" regime, that is, $w = 0$, and have characterized the values of δ for which AMSE is zero. This is known as the phase transition analysis.

In this paper, we first perform the phase transition analysis of LQLS. Our results reveal some of the limitations and misleading features of the phase transition analysis. To overcome these limitations, we propose the small error analysis of LQLS. Our new analysis framework not only sheds light on the results of the phase transition analysis, but also describes when phase transition analysis is reliable, and presents a more accurate comparison among different regularizers.

1. Introduction.

1.1. *Objective.* Consider the linear regression problem where the goal is to estimate the parameter vector $\beta \in \mathbb{R}^p$ from a set of n response variables $y \in \mathbb{R}^n$, under the model $y = X\beta + w$. This problem has been studied extensively in the last two centuries since Gauss and Legendre developed the least squares estimate of β. The instability or high variance of the least squares estimates led to the development of the regularized least squares. One of the most popular regularization

Received March 2017; revised September 2017.

[1] Supported by NSF Grant CCF-1420328.

MSC2010 subject classifications. Primary 62J05, 62J07.

Key words and phrases. Bridge regression, phase transition, comparison of estimators, small error regime, second-order term, asymptotic mean square error, optimal tuning.

classes is the ℓ_q-regularized least squares (LQLS), a.k.a. bridge regression [24, 25], given by the following optimization problem:

$$\hat{\beta}(\lambda, q) \in \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q, \tag{1.1}$$

where $\|\beta\|_q^q = \sum_{i=1}^p |\beta_i|^q$ and $1 \le q \le 2$.[2] LQLS has been extensively studied in the literature. In particular, one can prove the consistency of $\hat{\beta}(\lambda, q)$ under the classical asymptotic analysis (p fixed while $n \to \infty$) [30]. However, this asymptotic regime becomes irrelevant for high-dimensional problems in which n is *not* much larger than p. Under this high-dimensional setting, if β does not have any specific "structure," we do not expect any estimator to perform well. One of the structures that has attracted attention in the last twenty years is the sparsity, that assumes only k of the elements of β are nonzero and the rest are zero. To understand the behavior of the estimators under structured linear model in high dimension, a new asymptotic framework has been proposed in which it is assumed that $X_{ij} \overset{\text{i.i.d.}}{\sim} N(0, 1/n)$, $k, n, p \to \infty$, while $n/p \to \delta$ and $k/p \to \varepsilon$, where δ and ε are fixed numbers [1, 6, 16, 19, 22].

One of the main notions that has been widely studied in this asymptotic framework is the phase transition [1, 16, 19, 39]. Intuitively speaking, phase transition analysis assumes the error w equals zero and characterizes the value of δ above which an estimator converges to the true β (in certain sense that will be clarified later). While there is always an error in the response variables, it is believed that phase transition analysis provides reliable information when the errors are small. In this paper, we start by studying the phase transition diagrams of LQLS for $1 \le q \le 2$. Our analysis reveals several limitations of the phase transition analysis. We will clarify these limitations in the next section. We then propose a higher order analysis of LQLS in the small-error regime. As we will explain in the next section, our new framework sheds light on the peculiar behavior of the phase transition diagrams, and explains when we can rely on the results of phase transition analysis in practice.

1.2. *Limitations of the phase transition and our solution.*

In this section, we intuitively describe the results of phase transition analysis, its limitations and our new framework. Consider the class of LQLS estimators and suppose that we would like to compare the performance of these estimators through the phase transition diagrams. For the purpose of this section, we assume that the vector β has only k nonzero elements, where $k/p \to \varepsilon$ with $\varepsilon \in (0, 1)$. Since phase transition analysis is concerned with $w = 0$ setting, it considers $\lim_{\lambda \to 0} \hat{\beta}(\lambda, q)$ which is equivalent

[2] Bridge regression is a name used for LQLS with any $q \ge 0$. In this paper, we focus on $1 \le q \le 2$. To analyze the case $0 \le q < 1$, Zheng et al. [46] has used the replica method from statistical physics.

to the following estimator:

$$\arg\min_{\beta} \|\beta\|_q^q, \qquad \text{subject to } y = X\beta. \tag{1.2}$$

Below we informally state the results of the phase transition analysis. We will formalize the statement and describe in detail the conditions under which this result holds in Section 3.

INFORMAL RESULT 1. For a given $\varepsilon > 0$ and $q \in [1, 2]$, there exists a number $M_q(\varepsilon)$ such that as $p \to \infty$, if $\delta \geq M_q(\varepsilon) + \gamma$ ($\gamma > 0$ is an arbitrary number), then (1.2) succeeds in recovering β, while if $\delta \leq M_q(\varepsilon) - \gamma$, (1.2) fails.[3]

The curve $\delta = M_q(\varepsilon)$ is called the phase transition curve of (1.2). We will show that $M_q(\varepsilon)$ is given by the following formula:

$$M_q(\varepsilon) = \begin{cases} 1 & \text{if } 2 \geq q > 1, \\ \inf\limits_{\chi \geq 0} (1-\varepsilon)\mathbb{E}\eta_1^2(Z;\chi) + \varepsilon\big(1+\chi^2\big) & \text{if } q = 1, \end{cases} \tag{1.3}$$

where $\eta_1(u;\chi) = (|u| - \chi)_+ \operatorname{sign}(u)$ denotes the soft thresholding function and $Z \sim N(0, 1)$. While the above phase transition curves can be obtained with different techniques, such as the statistical dimension framework proposed in [1] and Gordon's lemma applied in [36, 42], we will derive them as a simple byproduct of our main results in Section 3 under message passing framework. Also the phase transition analysis of the regularized least squares has already been performed in the literature [16, 20, 35]. Hence, we should emphasize that the presentation of the phase transition results for bridge regression is not our main contribution here. We rather use it to motivate our second-order analysis of the asymptotic risk, which will appear later in this section. Before we proceed further, let us mention some of the properties of $M_1(\varepsilon)$ that will be useful in our later discussions.

LEMMA 1. *$M_1(\varepsilon)$ satisfies the following properties*:

(i) *$M_1(\varepsilon)$ is an increasing function of ε.*
(ii) $\lim_{\varepsilon\to 0} M_1(\varepsilon) = 0$.
(iii) $\lim_{\varepsilon\to 1} M_1(\varepsilon) = 1$.
(iv) *$M_1(\varepsilon) > \varepsilon$, for $\varepsilon \in (0, 1)$.*

PROOF. Define $F(\chi,\varepsilon) \triangleq (1-\varepsilon)\mathbb{E}\eta_1^2(Z;\chi) + \varepsilon(1+\chi^2)$. It is straightforward to verify that $F(\chi,\varepsilon)$, as a function of χ over $[0,\infty)$, is strongly convex and has a

[3] Different notions of success have been studied in the phase transition analysis. We will mention one notion later in our paper.

unique minimizer. Let $\chi^*(\varepsilon)$ be the minimizer. We write it as $\chi^*(\varepsilon)$ to emphasize its dependence on ε. By employing the chain rule, we have

$$\begin{aligned}\frac{dM_1(\varepsilon)}{d\varepsilon} &= \frac{\partial F(\chi^*(\varepsilon), \varepsilon)}{\partial \varepsilon} + \frac{\partial F(\chi^*(\varepsilon), \varepsilon)}{\partial \chi} \cdot \frac{d\chi^*(\varepsilon)}{d\varepsilon} = \frac{\partial F(\chi^*(\varepsilon), \varepsilon)}{\partial \varepsilon} \\ &= 1 + (\chi^*(\varepsilon))^2 - \mathbb{E}\eta_1^2(Z; \chi^*(\varepsilon)) > 1 + (\chi^*(\varepsilon))^2 - \mathbb{E}|Z|^2 \\ &= (\chi^*(\varepsilon))^2 > 0,\end{aligned}$$

which completes the proof of part (i). To prove (ii), note that

$$\begin{aligned}0 &\le \lim_{\varepsilon\to 0} \min_{\chi \ge 0} (1-\varepsilon)\mathbb{E}\eta_1^2(Z; \chi) + \varepsilon(1+\chi^2) \\ &\le \lim_{\varepsilon\to 0} (1-\varepsilon)\mathbb{E}\eta_1^2(Z; \log(1/\varepsilon)) + \varepsilon(1+\log^2(1/\varepsilon)) \\ &= \lim_{\varepsilon\to 0} 2(1-\varepsilon) \int_{\log(1/\varepsilon)}^{\infty} (z - \log(1/\varepsilon))^2 \phi(z)\,dz \\ &= \lim_{\varepsilon\to 0} 2(1-\varepsilon) \int_0^{\infty} z^2 \phi(z + \log(1/\varepsilon))\,dz \\ &\le \lim_{\varepsilon\to 0} 2(1-\varepsilon)\mathrm{e}^{-\frac{\log^2(1/\varepsilon)}{2}} \int_0^{\infty} z^2\phi(z)\,dz = 0,\end{aligned}$$

where $\phi(\cdot)$ is the density function of standard normal. Regarding the proof of part (iii), first note that as $\varepsilon \to 1$, $\chi^*(\varepsilon) \to 0$. Otherwise, suppose $\chi^*(\varepsilon) \to \chi_0 > 0$ (taking a convergent subsequence if necessary). Since $\mathbb{E}\eta_1^2(Z; \chi^*(\varepsilon)) \le \mathbb{E}|Z|^2 = 1$, we obtain

$$\lim_{\varepsilon\to 1} F(\chi^*(\varepsilon), \varepsilon) = 1 + \chi_0^2 > 1.$$

On the other hand, it is clear that

$$\lim_{\varepsilon\to 1} F(\chi^*(\varepsilon), \varepsilon) \le \lim_{\varepsilon\to 1} F(0, \varepsilon) = 1.$$

A contradiction arises. Hence, the fact $\chi^*(\varepsilon) \to 0$ as $\varepsilon \to 1$ leads directly to

$$\lim_{\varepsilon\to 1} M_1(\varepsilon) = \lim_{\varepsilon\to 1} F(\chi^*(\varepsilon), \varepsilon) = 1.$$

Part (iv) is clear from the definition of $M_1(\varepsilon)$. □

Figure 1 shows $M_q(\varepsilon)$ for different values of q. We observe several peculiar features: (i) As is clear from both Lemma 1 and Figure 1, $q = 1$ requires much fewer observations than all the other values of $q > 1$ for successful recovery of β. (ii) The values of the nonzero elements of β do not have any effect on the phase transition curves. In fact, even the sparsity level does not have any effect on the

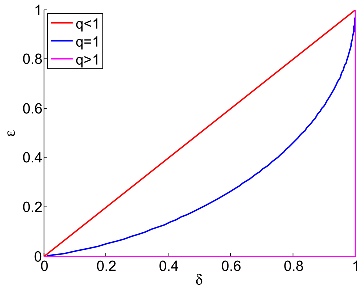


FIG. 1. *Phase transition curves of LQLS for* (i) $q < 1$: *these results are derived in* [46] *by the nonrigorous replica method from statistical physics. We have just included them for comparison purposes. In this paper, we have focused on* $q \geq 1$. (ii) $q = 1$: *the blue curve exhibits the phase transition of LASSO. Below this curve LASSO can "successfully" recover* β. (iii) $q > 1$: *The magenta curve represents the phase transition of LQLS for any* $q > 1$. *This figure is based on Informal Result* 1 *and will be carefully defined and derived in Section* 3.

phase transition for $q > 1$. (iii) For every $q > 1$, the phase transition of (1.2) happens at exactly the same value.

These features raise the following question: how much and to what extent are these phase transition results useful in applications, where at least small amount of error is present in the response variables? For instance, intuitively speaking, we do not expect to see much difference between the performance of LQLS for $q = 1.01$ and $q = 1$. However, according to the phase transition analysis, $q = 1$ outperforms $q = 1.01$ by a wide margin. In fact, the performance of LQLS for $q = 1.01$ seems to be closer to that of $q = 2$ than $q = 1$. Also, in contrast to the phase transition implication, we may not expect LQLS to perform the same for β with different values of nonzero elements. The main goal of this paper is to present a new analysis that will shed light on the misleading features of the phase transition analysis. It will also clarify when and under what conditions the phase transition analysis is reliable for practical guidance.

In our new framework, the variance σ_w^2 of the error w is assumed to be small. We consider (1.1) with the optimal value of λ for which the asymptotic mean square error, that is, $\lim_{p\to\infty} \frac{\|\hat{\beta}(\lambda,q)-\beta\|_2^2}{p}$, is minimized. We first obtain the formula for the asymptotic mean square error (AMSE) characterized through a series of nonlinear equations. Since σ_w is assumed small, we then derive the asymptotic expansions for AMSE as $\sigma_w \to 0$. As we will describe later, the phase transition of LQLS for different values of q can be obtained from the first dominant term in the expansion. More importantly, we will show that the second dominant term is capable of evaluating the importance of the phase transition analysis for practical situations and also provides a much more accurate analysis of different bridge estimators. Here is one of the results of our paper, presented informally to clarify our claims. All the technical conditions will be determined in Sections 2 and 3.

INFORMAL RESULT 2. If λ_* denotes the optimal value of λ, then for any $q \in (1, 2)$, $\delta > 1$, and $\varepsilon < 1$

$$\lim_{p\to\infty} \frac{1}{p}\|\hat{\beta}(\lambda_*, q) - \beta\|_2^2 = \frac{\sigma_w^2}{1 - 1/\delta} - \sigma_w^{2q} \frac{\delta^{q+1}(1-\varepsilon)^2(\mathbb{E}|Z|^q)^2}{(\delta-1)^{q+1}\varepsilon\mathbb{E}|G|^{2q-2}} + o(\sigma_w^{2q}),$$

where $Z \sim N(0, 1)$ and G is a random variable whose distribution is specified by the nonzero elements of β. We will clarify this in the next section. Finally, the limit notation we have used above is the almost sure limit.

As we will discuss in Section 3, the first term $\frac{\sigma_w^2}{1-1/\delta}$ determines the phase transition. Moreover, we have further derived the second dominant term in the expansion of the asymptotic mean square error. This term enables us to clarify some of the confusing features of the phase transitions. Here are some important features of this term: (i) It is negative. Hence, the AMSE that is predicted by the first term (and phase transition analysis) is overestimated specially when q is close to 1. (ii) Fixing q, the magnitude of the second dominant term grows as ε decreases. Hence, for small values of σ_w all values of $1 < q < 2$ benefit from the sparsity of β. Also, smaller values of q seem to benefit more. (iii) Fixing ε and δ, the power of σ_w decreases as q decreases. This makes the absolute value of the second dominant term bigger. As q decreases to one, the order of the second dominant term gets closer to that of the first dominant term, and thus the predictions of phase transition analysis become less accurate. We will present a more detailed discussion of the second order term in Section 3. To show some more interesting features of our approach, we also informally state a result we prove for LASSO.

INFORMAL RESULT 3. Suppose that the nonzero elements of β are all larger than a fixed number μ. If λ_* denotes the value of λ that leads to the smallest AMSE, and if $\delta > M_1(\varepsilon)$, then

$$\lim_{p\to\infty} \frac{1}{p}\|\hat{\beta}(\lambda_*, 1) - \beta\|_2^2 - \frac{\delta M_1(\varepsilon)\sigma_w^2}{\delta - M_1(\varepsilon)} = O(\exp(-\tilde{\mu}/\sigma_w^2)), \tag{1.4}$$

where $\tilde{\mu}$ is a constant that depends on μ.

As can be seen here, compared to the other values of q, $q = 1$ has smaller first order term (according to Lemma 1), but much smaller (in magnitude) second-order term. The first implication of this result is that the first dominant term provides an accurate approximation of AMSE. Hence, phase transition analysis in this case is reliable even if small amount of noise is present; that is one of the main reasons why the theoretically derived phase transition curve matches the empirical one for LASSO. Furthermore, note that in order to obtain Informal result 3, we have made certain assumption about the nonzero components of β. As will be shown in Section 3, any violation of this assumption has major impact on the second dominant term.

In the rest of the paper, we first state all the assumptions required for our analysis. We then present the formal statements of aforementioned and related results and provide a more comprehensive discussion.

1.3. *Organization of the paper.* The rest of the paper is organized as follows: Section 2 presents the asymptotic framework of our analysis. Section 3 discusses the main contributions of our paper. Section 4 compares our results with the related work. Section 5 shows some simulation results and discusses some open problems that require further research. Section 6 and the Supplementary Material [45] are devoted to the proofs of all the results.

2. The asymptotic framework. The main goal of this section is to formally introduce the asymptotic setting under which we study LQLS. In the current and next sections only, we may write vectors and matrices as $\beta(p)$, $X(p)$, $w(p)$ to emphasize the dependence on the dimension of β. Similarly, we may use $\hat{\beta}(\lambda, q, p)$ as a substitute for $\hat{\beta}(\lambda, q)$. Note that since we assume $n/p \to \delta$, we do not include n in our notation. Now we define a specific type of a sequence known as a converging sequence. Our definition is borrowed from other papers [2, 3, 18] with some minor modifications.

DEFINITION 1. A sequence of instances $\{\beta(p), X(p), w(p)\}$ is called a converging sequence if the following conditions hold:

- The empirical distribution[4] of $\beta(p) \in \mathbb{R}^p$ converges weakly to a probability measure p_β with bounded second moment. Further, $\frac{1}{p}\|\beta(p)\|_2^2$ converges to the second moment of p_β.
- The empirical distribution of $w(p) \in \mathbb{R}^n$ converges weakly to a zero mean distribution with variance σ_w^2. And, $\frac{1}{n}\|w(p)\|_2^2 \to \sigma_w^2$.
- The elements of $X(p)$ are i.i.d. with distribution $N(0, 1/n)$.

For each of the problem instances in a converging sequence, we solve the LQLS problem (1.1) and obtain $\hat{\beta}(\lambda, q, p)$ as the estimator. The interest is to evaluate the accuracy of this estimator. Below we define the asymptotic mean square error.

DEFINITION 2. Let $\hat{\beta}(\lambda, q, p)$ be the sequence of solutions of LQLS for the converging sequence of instances $\{\beta(p), X(p), w(p)\}$. The asymptotic mean square error is defined as the almost sure limit of

$$\mathrm{AMSE}(\lambda, q, \sigma_w) \triangleq \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} |\hat{\beta}_i(\lambda, q, p) - \beta_i(p)|^2,$$

where the subscript i is used to denote the ith component of a vector.

[4] It is the distribution that puts a point mass $1/p$ at each of the p elements of the vector.

Note that we have suppressed δ and p_β in the notation of AMSE for simplicity, despite the fact that the asymptotic mean square error depends on them as well. In the above definition, we have assumed that the almost sure limit exists. Under the current asymptotic setting, the existence of AMSE can be proved. In fact, we are able to derive the asymptotic limit for general loss functions as presented in the following theorem.

THEOREM 2.1. *Consider a converging sequence* $\{\beta(p), X(p), w(p)\}$. *For any given* $q \in [1, 2]$, *suppose that* $\hat{\beta}(\lambda, q, p)$ *is the solution of LQLS defined in* (1.1). *Then for any pseudo-Lipschitz function*[5] $\psi : \mathbb{R}^2 \to \mathbb{R}$, *almost surely*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi\big(\hat{\beta}_i(\lambda, q, p), \beta_i(p)\big) = \mathbb{E}_{B,Z}\big[\psi\big(\eta_q\big(B + \bar{\sigma} Z; \bar{\chi}\bar{\sigma}^{2-q}\big), B\big)\big], \tag{2.1}$$

where B and Z are two independent random variables with distributions p_β and $N(0, 1)$, respectively; the expectation $\mathbb{E}_{B,Z}(\cdot)$ is taken with respect to both B and Z; $\eta_q(\cdot; \cdot)$ is the proximal operator for the function $\|\cdot\|_q^q$;[6] *and $\bar{\sigma}$ and $\bar{\chi}$ satisfy the following equations*:

$$\bar{\sigma}^2 = \sigma_\omega^2 + \frac{1}{\delta}\mathbb{E}_{B,Z}\big[\big(\eta_q\big(B + \bar{\sigma} Z; \bar{\chi}\bar{\sigma}^{2-q}\big) - B\big)^2\big], \tag{2.2}$$

$$\lambda = \bar{\chi}\bar{\sigma}^{2-q}\Big(1 - \frac{1}{\delta}\mathbb{E}_{B,Z}\big[\eta_q'\big(B + \bar{\sigma} Z; \bar{\chi}\bar{\sigma}^{2-q}\big)\big]\Big), \tag{2.3}$$

where $\eta_q'(\cdot; \cdot)$ denotes the derivative of η_q with respect to its first argument.

The result for $q = 1$ has been proved in [3]. The key ideas of the proof for generalizing to $q \in (1, 2]$ are similar to those of [3]. We describe the main proof steps in Section J of the Supplementary Material [45]. According to Theorem 2.1, in order to calculate the asymptotic mean square error (or any other loss) of $\hat{\beta}(\lambda, q, p)$, we have to solve (2.2) and (2.3) for $(\bar{\sigma}, \bar{\chi})$. The following lemma shows these two nonlinear equations have a unique solution.

LEMMA 2. *For any positive values of $\lambda, \delta, \sigma_w > 0$, any random variable B with finite second moment, and any $q \in [1, 2]$, there exists a unique pair $(\bar{\sigma}, \bar{\chi})$ that satisfies both* (2.2) *and* (2.3).

[5] A function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is pseudo-Lipschitz of order k if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^2$, we have $|\psi(x) - \psi(y)| \le L(1 + \|x\|_2^{k-1} + \|y\|_2^{k-1})\|x - y\|_2$. We consider pseudo-Lipschitz functions with order 2 in this paper.

[6] Proximal operator of $\|\cdot\|_q^q$ is defined as $\eta_q(u; \chi) \triangleq \arg\min_z \frac{1}{2}(u - z)^2 + \chi|z|^q$. For further information on these functions, please refer to the Supplementary Material [45].

The proof of this lemma can be found in Section D of the Supplementary Material [45]. Theorem 2.1 provides the first step in our analysis of LQLS. We first calculate $\bar{\sigma}$ and $\bar{\chi}$ from (2.2) and (2.3). Then, incorporating $\bar{\sigma}$ and $\bar{\chi}$ in (2.1), gives the following expression for the asymptotic mean square error:

$$\text{AMSE}(\lambda, q, \sigma_w) = \mathbb{E}_{B,Z}\big(\eta_q(B + \bar{\sigma} Z; \bar{\chi}\bar{\sigma}^{2-q}) - B\big)^2. \tag{2.4}$$

Given the distribution of B (the sparsity level ε included), the variance of the error σ_w^2, the number of response variables (normalized by the number of predictors) δ and the regularization parameter λ, it is straightforward to write a computer program to find the solution of (2.2) and (2.3) and then compute the value of AMSE. However, it is needless to say that this approach does not shed much light on the performance of bridge regression estimates, since there are many factors involved in the computation and each affects the result in a nontrivial fashion. In this paper, we would like to perform an analytical study on the solution of (2.2) and (2.3) and obtain an explicit characterization of AMSE in the small-error regime.

3. Our main contributions.

3.1. *Optimal tuning of* λ. The performance of LQLS, as defined in (1.1), depends on the tuning parameter λ. In this paper, we consider the value of λ that gives the minimum AMSE. Let $\lambda_{*,q}$ denote the value of λ that minimizes AMSE given in (2.4). Then LQLS is solved with this specific value of λ, that is,

$$\hat{\beta}(\lambda_{*,q}, q, p) \in \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda_{*,q}\|\beta\|_q^q. \tag{3.1}$$

Note that this is the best performance that LQLS can achieve in terms of the AMSE. Theorem 2.1 enables us to evaluate this optimal AMSE of LQLS for every $q \in [1, 2]$. The key step is to compute the solution of (2.2) and (2.3) with $\lambda = \lambda_{*,q}$. Since $\lambda_{*,q}$ has to be chosen optimally, it seemingly causes an extra complication for our analysis. However, as we show in the following corollary, the study of equations (2.2) and (2.3) can be simplified to some extent.

COROLLARY 1. *Consider a converging sequence* $\{\beta(p), X(p), w(p)\}$. *Suppose that* $\hat{\beta}(\lambda_{*,q}, q, p)$ *is the solution of LQLS defined in* (3.1). *Then for any* $q \in [1, 2]$

$$\text{AMSE}(\lambda_{*,q}, q, \sigma_w) = \min_{\chi \ge 0} \mathbb{E}_{B,Z}\big(\eta_q(B + \bar{\sigma} Z; \chi) - B\big)^2, \tag{3.2}$$

where B *and* Z *are two independent random variables with distributions* p_β *and* $N(0, 1)$, *respectively*; *and* $\bar{\sigma}$ *is the unique solution of the following equation*:

$$\bar{\sigma}^2 = \sigma_\omega^2 + \frac{1}{\delta}\min_{\chi \ge 0} \mathbb{E}_{B,Z}\big[\big(\eta_q(B + \bar{\sigma} Z; \chi) - B\big)^2\big]. \tag{3.3}$$

The proof of Corollary 1 is shown in Section E of the Supplementary Material [45]. Corollary 1 enables us to focus the analysis on a single equation (3.3), rather than two equations (2.2) and (2.3). The results we will present in the next section are mainly based on investigating the solution of (3.3).

3.2. *Analysis of AMSE.* In this paper, since we are focused on the sparsity structure of β, from now on we assume that the distribution, to which the empirical distribution of $\beta \in \mathbb{R}^p$ converges, has the form

$$p_\beta(b) = (1-\varepsilon)\delta_0(b) + \varepsilon g(b),$$

where $\delta_0(\cdot)$ denotes a point mass at zero, and $g(\cdot)$ is a generic distribution that does not have any point mass at 0. Here, the mixture proportion $\varepsilon \in (0,1)$ is a fixed number that represents the sparsity level of β. The smaller ε is, the sparser β will be. The distribution $g(b)$ specifies the values of nonzero components of β. We will use G to denote a random variable having such a distribution. Since our results and proof techniques look very different for the case $q > 1$ and $q = 1$, we study these cases separately.

3.2.1. *Results for $q > 1$.* Our first result is concerned with the optimal AMSE of LQLS for $1 < q \leq 2$, when the number of response variables is larger than the number of predictors p, that is, $\delta > 1$.

THEOREM 3.1. *Suppose $\mathbb{P}(|G| \leq t) = O(t)$ (as $t \to 0$) and $\mathbb{E}|G|^2 < \infty$, then for $1 < q < 2$, $\delta > 1$ and $\varepsilon \in (0,1)$, we have*

$$\mathrm{AMSE}(\lambda_{*,q}, q, \sigma_w) = \frac{\sigma_w^2}{1-1/\delta} - \frac{\delta^{q+1}(1-\varepsilon)^2(\mathbb{E}|Z|^q)^2}{(\delta-1)^{q+1}\varepsilon\mathbb{E}|G|^{2q-2}}\sigma_w^{2q} + o(\sigma_w^{2q}). \tag{3.4}$$

For $q = 2$, $\delta > 1$ and $\varepsilon \in (0,1)$, if $\mathbb{E}|G|^2 < \infty$, we have

$$\mathrm{AMSE}(\lambda_{*,q}, q, \sigma_w) = \frac{\sigma_w^2}{1-1/\delta} - \frac{\delta^3\sigma_w^4}{(\delta-1)^3\varepsilon\mathbb{E}|G|^2} + o(\sigma_w^4).$$

Note that $Z \sim N(0,1)$ and $G \sim g(\cdot)$ are independent.

The proof of the result is presented in Section F of the Supplementary Material [45]. There are several interesting features of this result that we would like to discuss: (i) The second dominant term of AMSE is negative. This means that the actual AMSE is smaller than the one predicted by the first-order term, especially for smaller values of q. (ii) Neither the sparsity level nor the distribution of the nonzero components of β appear in the first dominant term, that is, $\frac{\sigma_w^2}{1-1/\delta}$. As we will discuss later in this section, the first dominant term is the one that specifies the phase transition curve. Hence, these calculations show a peculiar feature of phase transition analysis we discussed in Section 1.2, that the phase transition of

$q \in (1,2]$ is neither affected by nonzero components of β or the sparsity level. However, we see that both factors come into play in the second dominant term. (iii) For the fully dense vector, that is, $\varepsilon = 1$, (3.4) may imply that for $1 < q < 2$,

$$\mathrm{AMSE}(\lambda_{*,q}, q, \sigma_w) = \frac{\sigma_w^2}{1 - 1/\delta} + o(\sigma_w^{2q}).$$

Hence, we require a different analysis to obtain the second dominant term (with different orders). We refer the interested readers to [44] for further information about this case. (iv) For $\varepsilon < 1$, the choice of $q \in (1,2]$ does not affect the first dominant term. That is the reason why all the values of $q \in (1,2]$ share the same phase transition curve. However, the value of q has a major impact on the second dominant term. In particular, as q approaches 1, the order of the second dominant term in terms of σ_w gets closer to that of the first dominant term. This means that in any practical setting, phase transition analysis may lead to misleading conclusions. Specifically, in contrast to the conclusion from phase transition analysis that $q \in (1,2]$ have the same performance, the second-order expansion enables us to conclude that the closer to 1 the value of q is, the better its performance will be. Our next theorem discusses the AMSE when $\delta < 1$.

THEOREM 3.2. *Suppose* $\mathbb{E}|G|^2 < \infty$, *then for* $1 < q \leq 2$ *and* $\delta < 1$,

$$\lim_{\sigma_w \to 0} \mathrm{AMSE}(\lambda_{*,q}, q, \sigma_w) > 0. \tag{3.5}$$

The proof of this theorem is presented in Section G of the Supplementary Material [45]. Theorems 3.1 and 3.2 together show a notion of phase transition. For $\delta > 1$, as $\sigma_w \to 0$, $\mathrm{AMSE} = O(\sigma_w^2)$, and hence it will go to zero, while $\mathrm{AMSE} \nrightarrow 0$ for $\delta < 1$. In fact, the phase transition curve $\delta = 1$ can be derived from the first dominant term in the expansion of AMSE. If $\delta = 1$, the first dominant term is infinity and there will be no successful recovery, while it becomes zero when $\sigma_w = 0$ if $\delta > 1$. A more rigorous justification can be found in the proof of Theorems 3.1 and 3.2. Therefore, we may conclude that the first-order term contains the phase transition information. Moreover, the derived second-order term offers us additional important information regarding the accuracy of the phase transition analysis. To provide a comprehensive understanding of these two terms, in Section 5 we will evaluate the accuracy of first- and second-order approximations to AMSE through numerical studies.

3.2.2. *Results for* $q = 1$. So far we have studied the case $1 < q \leq 2$. In this section, we study $q = 1$, a.k.a. LASSO. In Theorems 3.1 and 3.2, we have characterized the behavior of LQLS with $q \in (1,2]$ for a general class of G. It turns out that the distribution of G has a more serious impact on the second dominant term of AMSE for LASSO. We thus analyze it in two different settings. Our first theorem considers the distributions that do not have any mass around zero.

THEOREM 3.3. *Suppose* $\mathbb{P}(|G| > \mu) = 1$ *with* μ *being a positive constant and* $\mathbb{E}|G|^2 < \infty$, *then for* $\delta > M_1(\varepsilon)$,[7]

$$\text{AMSE}(\lambda_{*,1}, 1, \sigma_w) = \frac{\delta M_1(\varepsilon)}{\delta - M_1(\varepsilon)}\sigma_w^2 + o\left(\phi\left(\sqrt{\frac{\delta - M_1(\varepsilon)}{\delta}}\frac{\tilde{\mu}}{\sigma_w}\right)\right), \tag{3.6}$$

where $\tilde{\mu}$ *is any positive constant smaller than* μ *and* $\phi(\cdot)$ *is the density function of standard normal.*

The proof of Theorem 3.3 is given in Section 6. Different from LQLS with $q \in (1, 2]$, we have not derived the exact analytical expression of second dominant term for LASSO. However, since it is exponentially small, the first-order term (or phase transition analysis) is sufficient for evaluating the performance of LASSO in the small-error regime. This will be further confirmed by the numerical studies in Section 5. Below is our result for the distributions of G that have more mass around zero.

THEOREM 3.4. *Suppose that* $\mathbb{P}(|G| \le t) = \Theta(t^\ell)$ *(as* $t \to 0$*) with* $\ell > 0$ *and* $\mathbb{E}|G|^2 < \infty$, *then for* $\delta > M_1(\varepsilon)$,

$$\begin{aligned} -\Theta(\sigma_w^{\ell+2}) &\gtrsim \text{AMSE}(\lambda_{*,1}, 1, \sigma_w) - \frac{\delta M_1(\varepsilon)}{\delta - M_1(\varepsilon)}\sigma_w^2 \\ &\gtrsim -\Theta(\sigma_w^{\ell+2}) \cdot \Bigl(\underbrace{\log\log\cdots\log}_{m\ times}\Bigl(\frac{1}{\sigma_w}\Bigr)\Bigr)^{\ell/2}, \end{aligned}$$

where m *is an arbitrary but finite natural number, and* $a \gtrsim b$ *means* $a \ge b$ *holds for sufficiently small* σ_w.

The proof of this theorem can be found in Section H of the Supplementary Material [45]. It is important to notice the difference between Theorems 3.3 and 3.4. The first point we would like to emphasize is that the first dominant terms are the same in both cases. The second dominant terms are different though. As we will show in Section 6 and Section H from the Supplementary Material [45], similar to LQLS for $1 < q \le 2$, the second dominant terms are in fact negative. Hence, the actual AMSE will be smaller than the one predicted by the first dominant term. Furthermore, note that the magnitude of the second dominant term in Theorem 3.4 is much larger than that in Theorem 3.3. This seems intuitive. LASSO tends to shrink the parameter coefficients toward zero, and hence, if the true β has more mass around zero, the AMSE will be smaller. The more mass the distribution of G has around zero, the better the second-order term will be. Our next theorem discusses what happens if $\delta < M_1(\varepsilon)$.

[7] Recall $M_1(\varepsilon) = \inf_{\chi \ge 0}(1 - \varepsilon)\mathbb{E}\eta_1^2(Z; \chi) + \varepsilon(1 + \chi^2)$ with $Z \sim N(0, 1)$.

THEOREM 3.5. *Suppose that* $\mathbb{E}|G|^2 < \infty$. *Then for* $\delta < M_1(\varepsilon)$,

$$\lim_{\sigma_w \to 0} \mathrm{AMSE}(\lambda_{*,1}, 1, \sigma_w) > 0. \tag{3.7}$$

The proof is presented in Section I of the Supplementary Material [45]. Similarly, as we discussed in Section 3.2.1, Theorems 3.3, 3.4 and 3.5 imply the phase transition curve of LASSO. Such information can be obtained from the first dominant term in the expansion of AMSE as well.

4. Related work.

4.1. *Other phase transition analyses and $n/p \to \delta$ asymptotic results.* The asymptotic framework that we considered in this paper evolved in a series of papers by Donoho and Tanner [13, 14, 19, 20]. This framework was used before on similar problems in engineering and physics [10, 26, 41]. Donoho and Tanner characterized the phase transition curve for LASSO and some of its variants. Inspired by this framework, many researchers started exploring the performance of different algorithms or estimates under these asymptotic settings [1–3, 6, 11, 12, 15, 17, 21, 22, 32, 37, 39, 42, 46].

Our paper performs the analysis of LQLS under such asymptotic framework. Also, we adopt the message passing analysis that was developed in a series of papers [2, 3, 16, 18, 33]. The notion of phase transition we consider is similar to the one introduced in [18]. However, there are three major differences: (i) The analysis of [18] is performed for LASSO, while we have generalized the analysis to any LQLS with $1 < q \leq 2$. (ii) The analysis of [18] is performed on the least favorable distribution for LASSO, while here we characterize the effect of the distribution of G on the AMSE as well. (iii) Finally, [18] is only concerned with the first dominant term in AMSE of LASSO, while we derive the second dominant term whose importance has been discussed in the last few sections.

Another line of research that has connections with our analysis is presented in a series of papers [35, 36, 42]. In [42], the authors have derived a minimax formulation that (if it has a unique solution and is solved) can give an accurate characterization of the asymptotic mean square error. Compared with Theorem 2.1 in our paper, that result works for more general penalized M-estimators, while Theorem 2.1 holds for general pseudo-Lipschitz loss functions. When applying the minimax formulation in [42] to bridge regression, the AMSE formula in (2.4) can be recovered. However, the derivation of phase transition curves for bridge regression under optimal tuning is not found in [42]. Furthermore, [35, 36] proposed a geometric approach to characterize the risk of penalized least square estimates with general convex penalties. In particular, both papers obtained phase transition results based on a key convex geometry quantity called “Gaussian squared-distances.” However, [36] only rigorously proved the negative results (equivalent to Theorems 3.2 and 3.5) and left the positive part as a conjecture. The phase transition results in [35] and are concerned with the prediction errors $\|y - X\hat{\beta}\|_2^2$ and $\|X\beta - X\hat{\beta}\|_2^2$, rather than the estimation error $\|\hat{\beta} - \beta\|_2^2$. Also, neither of the two papers went beyond the first-order or phase transition analysis of the risk.

Several researchers have also worked on the analysis of LQLS for $q < 1$ [29, 37, 46]. These analyses are based on nonrigorous, but widely accepted replica method from statistical physics. The current paper extends the analysis of [46] to $q \geq 1$ case, makes the analysis rigorous by using the message passing framework rather than the replica method, and finally provides a higher order analysis.

4.2. *Other analysis frameworks.* One of the first papers that compared the performance of penalization techniques is [27] which showed that there exists a value of λ with which Ridge regression, that is, LQLS with $q = 2$, outperforms the vanilla least squares estimator. Since then, many more regularizers have been introduced to the literature each with a certain purpose. For instance, we can mention LASSO [43], elastic net [47], SCAD [23], bridge regression [24] and more recently SLOPE [5]. There has been a large body of work on studying all these regularization techniques. We partition all the work into the following categories and explain what in each category has been done about the bridge regression:

(i) Simulation results: One of the main motivations for our work comes from the nice simulation study of the bridge regression presented in [25]. This paper finds the optimal values of λ and q by generalized cross validation and compares the performance of the resulting estimator with both LASSO and ridge. The main conclusion is that the bridge regression can outperform both LASSO and ridge. Given our results we see that if sparsity is present in β, then smaller values of q perform better than ridge (in their second dominant term).

(ii) Asymptotic study: Knight and Fu [30] studied the asymptotic properties of bridge regression under the setting where $n \to \infty$, while p is fixed. They established the consistency and asymptotic normality of the estimates under quite general conditions. Huang et al. [28] studied LQLS for $q < 1$ under a high-dimensional asymptotic setting in which p grows with n but is still assumed to be less than n. They not only derived the asymptotic distribution of the estimators, but also proved LQLS has oracle properties in the sense of Fan and Li [23]. They have also considered the case $p > n$, and have shown that under partial orthogonality assumption on X, bridge regression distinguishes correctly between covariates with zero and nonzero coefficients. Note that under the asymptotic regime of our paper, both LASSO and the other bridge estimators have false discoveries [40] and possibly nonzero AMSE. Hence, they may not provide consistent estimates. We should also mention that the analysis of bridge regression with $q \in [0, 1)$ under the asymptotic regime $n/p \to \delta$ is presented in [46]. Finally, the performance of LASSO under a variety of conditions has been studied extensively. We refer the reader to [7] for the review of those results.

(iii) Nonasymptotic bounds: One of the successful approaches that has been employed for studying the performance of regularization techniques such as LASSO is the minimax analysis [4, 38]. We refer the reader to [7] for a complete list of references on this direction. In this minimax approach, a lower bound for the prediction error or mean square error of any estimation technique is first derived. Then a specific estimate, like the one returned by LASSO, is considered and an upper bound is derived assuming the design matrices satisfy certain conditions such as restrictive eigenvalue assumption [4, 31], restricted isometry condition [9] or coherence conditions [8]. These conditions can be confirmed for matrices with i.i.d. sub-Gaussian elements. Based on these evaluations, if the order of the upper bound for the estimate under study matches the order of the lower bound, we can claim that the estimate (e.g., LASSO) is minimax rate-optimal. This approach has some advantages and disadvantages compared to our asymptotic approach: (i) It works under more general conditions. (ii) It provides information for any sample size. The price paid in the minimax analysis is that the constants derived in the results are usually not sharp, and hence many schemes have similar guarantees and cannot be compared to each other. Our asymptotic framework loses the generality and in return gives sharp constants that can then be used in evaluating and comparing different schemes as we do in this paper. Along similar directions, [31] has studied the penalized empirical risk minimization with ℓ_q penalties for the values of $q \in [1, 1 + \frac{1}{\log p}]$ and has found upper bounds on the excess risk of these estimators (oracle inequalities). Similar to minimax analysis, although the results of this analysis enjoy generality, they suffer from loose constants that impede an accurate comparisons of different bridge estimators.

5. Numerical results and discussions.

5.1. *Summary.* The analysis of AMSE we presented in Section 3.2 is performed as $\sigma_w \to 0$. For such asymptotic analysis, it would be interesting to check the approximation accuracy of the first- and second-order expansions of AMSE over a reasonable range of σ_w. Toward this goal, this section performs several numerical studies to (i) evaluate the accuracy of the first- and second-order expansions discussed in Section 3.2, (ii) discover situations in which the first-order approximation is not accurate (for reasonably small noise levels) while the second-order expansion is, and (iii) identify situations where both first- and second-orders are inaccurate and propose methods for improving the approximations. Sections 5.2 and 5.3 study the performance of LASSO and other bridge regression estimators with $q > 1$, respectively. Finally, we should also mention that all the results presented in this paper are concerned with the asymptotic setting $n, p \to \infty$ and $n/p \to \delta$. To evaluate the accuracy of these results for finite sample sizes, we have performed additional simulations whose results are presented in Section B of the Supplementary Material [45].

5.2. *LASSO*. One of the conclusions from Theorem 3.3 is that the first dominant term provides a good approximation of AMSE for the LASSO problem when the distribution of G does not have a large mass around 0. To test this claim, we conduct the following numerical experiment. We set the parameters of our problem instances in the following way:

1. δ can take any value in $\{1.1, 1.5, 2\}$.
2. ε can take values in $\{0.25, 0.7\}$.
3. σ_w ranges within the interval $[0, 0.25]$.
4. the distribution of G is specified as $g(b) = 0.5\delta_1(b) + 0.5\delta_{-1}(b)$, where $\delta_a(\cdot)$ denotes a point mass at point a.

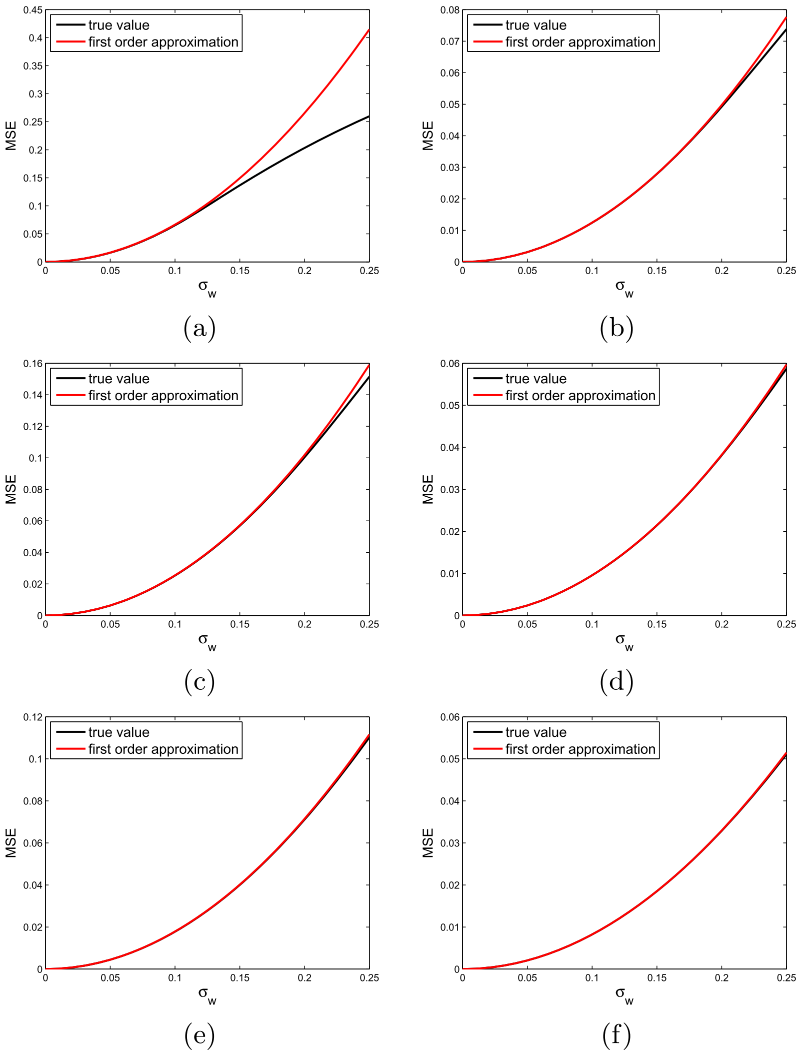
We then use the formula in Corollary 1 to calculate $\mathrm{AMSE}(\lambda_{*,1}, 1, \sigma_w)$. Finally, we compare $\mathrm{AMSE}(\lambda_{*,1}, 1, \sigma_w)$, computed numerically from (3.2) and (3.3), with its first-order approximation provided in Theorem 3.3. The results of this experiment are summarized in Figure 2. As is clear in this figure, the first-order expansion gives a very good approximation for AMSE over a large range of σ_w.

5.3. *Bridge regression estimators with $q > 1$*. In this numerical experiment, we would like to vary σ_w and see under what conditions our first-order or second-order expansions can lead to accurate approximation of AMSE for a wide range of σ_w. Throughout this section, we set the distribution of G to $g(b) = 0.5\delta_1(b) + 0.5\delta_{-1}(b)$, as we did in Section 5.2. We then investigate different conditions by specifying various values of other parameters in our problem instances. The expansion of AMSE for $q > 1$ is presented in Theorem 3.1. For $q \in (1, 2)$, recall the two terms in the expansion below:

$$\mathrm{AMSE}(\lambda_{*,q}, q, \sigma_w) = \frac{\sigma_w^2}{1 - 1/\delta} - \frac{\delta^{q+1}(1-\varepsilon)^2(\mathbb{E}|Z|^q)^2}{(\delta-1)^{q+1}\varepsilon\mathbb{E}|G|^{2q-2}}\sigma_w^{2q} + o(\sigma_w^{2q}). \tag{5.1}$$

We expect the first-order term to present a good approximation over a reasonably large range of σ_w, when the second-order term is sufficiently small. According to the analytical form of the second-order term in (5.1), it is small if the following three conditions hold simultaneously: (i) δ is not close to 1, (ii) ε is not small and (iii) q is not close to 1. Our first numerical result shown in Figure 3 is in agreement with this claim. In this simulation, we have set three different cases for δ, ε and q so that they satisfy the above three conditions. The nonzero elements of β are independently drawn from $0.5\delta_1(b) + 0.5\delta_{-1}(b)$. As demonstrated in this figure, the first- order term approximates AMSE accurately. Another interesting finding is that the second-order expansion provides an even better approximation.

To understand the limitation of the first-order approximation, we consider the cases in which the second-order term is large and suggests that at least the first-order approximation is not necessarily good. This happens when either δ decreases to 1, ε decreases to 0 or q decreases to 1. The settings of our experiments and the results are summarized below.



(a) (b) (c) (d) (e) (f)

FIG. 2. *Plots of actual AMSE and its approximations for* (a) $\delta = 1.1$ *and* $\varepsilon = 0.7$, (b) $\delta = 1.1$ *and* $\varepsilon = 0.25$, (c) $\delta = 1.5$ *and* $\varepsilon = 0.7$, (d) $\delta = 1.5$ *and* $\varepsilon = 0.25$, (e) $\delta = 2$ *and* $\varepsilon = 0.7$, (f) $\delta = 2$ *and* $\varepsilon = 0.25$.

1. We keep $q = 1.5$ and $\varepsilon = 0.7$ fixed and study different values of $\delta \in \{5, 2, 1.5, 1.1\}$. Figure 4 summarizes the results of this simulation. As is clear in this figure (and is consistent with the message of the second dominant term), as we decrease δ, the first-order approximation becomes less accurate. The second-order approximation in these cases is more accurate than the first-order approximation. However interestingly, the second-order approximation becomes less accurate as δ decreases also. These observations suggest that to have a good approximation for

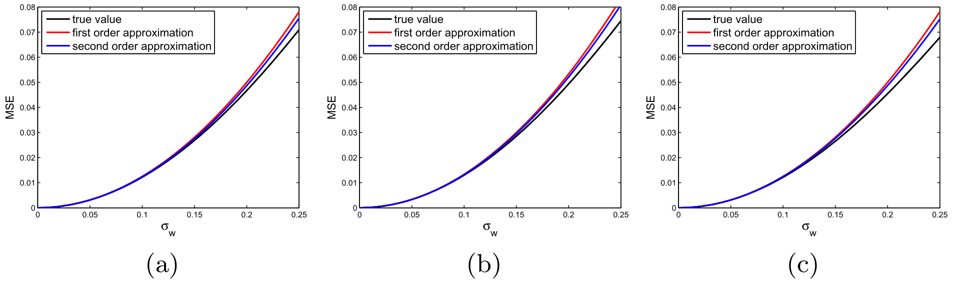


FIG. 3. *Plots of actual AMSE and its approximations for* (a) $\delta = 5$, $\varepsilon = 0.7$,$q = 1.5$, (b) $\delta = 4$, $\varepsilon = 0.7$, $q = 1.6$, (c) $\delta = 5$, $\varepsilon = 0.6$, $q = 1.8$.

the values of δ that are very close to 1, although the second-order approximation outperforms the first-order, it may not be sufficient and higher order terms are required. Such terms can be derived with strategies similar to the ones we used in the proof of Theorem 3.1. Note that the insufficiency of the second-order expansion partially results from the wide range of $\sigma_w \in [0, 0.25]$. If we evaluate the approx-

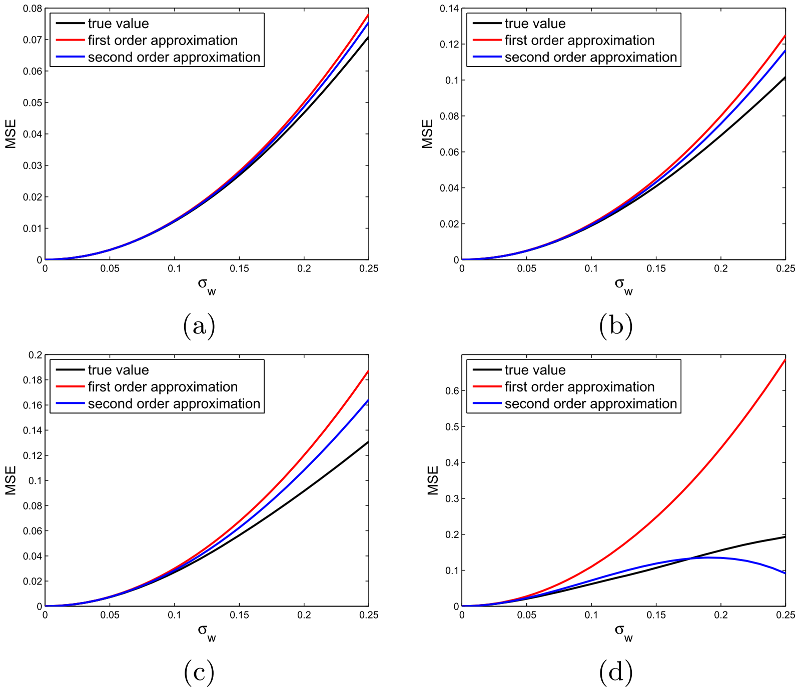
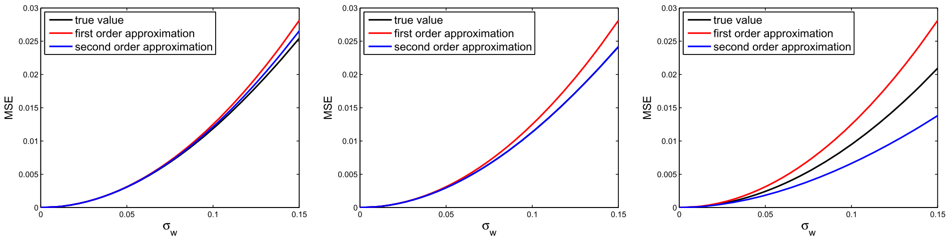


FIG. 4. *Plots of actual AMSE and its approximations for* $q = 1.5$, $\varepsilon = 0.7$ *with* (a) $\delta = 5$, (b) $\delta = 2$, (c) $\delta = 1.5$ *and* (d) $\delta = 1.1$.



(a) (b) (c)

FIG. 5. *Plots of actual AMSE and its approximations for* $\delta = 5$, $\varepsilon = 0.4$ *with* (a) $q = 1.8$, (b) $q = 1.5$ *and* (c) $q = 1.1$.

imation when σ_w is small enough, we will expect the success of the second-order expansion.

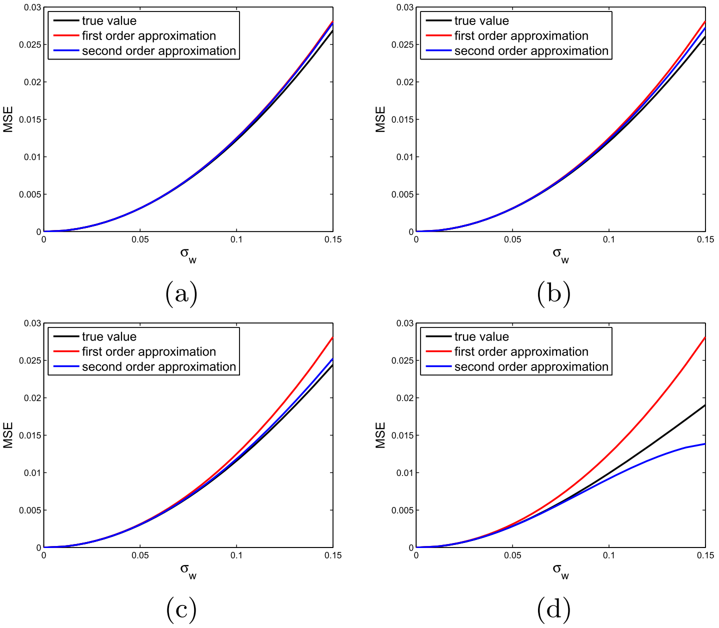
2. In our second simulation, we fix $\delta = 5$, $\varepsilon = 0.4$ and let $q \in \{1.8, 1.5, 1.1\}$. All the simulation results are summarized in Figure 5. As we expected, the first-order approximation becomes less accurate when q decreases. Furthermore, we notice that when q is very close to 1 (check $q = 1.1$ in the figure), even the second-order approximation is not necessarily good. This again calls for higher order approximation of the AMSE.

3. For the last simulation, we fix $\delta = 5$, $q = 1.8$, and let $\varepsilon \in \{0.7, 0.5, 0.3, 0.1\}$. Our simulation results are presented in Figure 6. We see that as ε decreases the first-order approximation becomes less accurate. The second-order approximation is always better than the first one. Moreover, we observe that when ε is very close to 0 (check $\varepsilon = 0.1$ in the figure), even the second-order approximation is not necessarily sufficient. As we discussed in the previous two simulations, we might need higher order approximation of the AMSE in such cases.

5.4. *Discussion.* First, our numerical studies confirm that the first-order term gives good approximations of AMSE for LASSO in the case where the distribution of nonzero elements of β is bounded away from zero. Second, as the numerical results for $q > 1$ demonstrate, while the second-order approximation always improves over the first-order term and works well in many cases, in the following situations it may not provide very accurate evaluation of AMSE: (i) when δ is close to 1, (ii) ε is close to zero and (iii) q is close to 1. In such cases, the value of the second-order term becomes large, and hence the approximation is only accurate for very small value of σ_w. The remedy that one can propose is to derive higher order expansions. Such terms can be calculated with the same strategy that we used to obtain the second dominant term.

6. Proof of Theorem 3.3.

Due to the limited space in the main text, we only present the proof of Theorem 3.3, one of our main results for LASSO, in this section. The proofs of all the other results are deferred to the Supplementary Material



(a) (b)

(c) (d)

FIG. 6. *Plots of actual AMSE and its approximations for* $\delta = 5$ *and* $q = 1.8$ *with* (a) $\epsilon = 0.7$, (b) $\epsilon = 0.5$, (c) $\epsilon = 0.3$ *and* (d) $\epsilon = 0.1$.

[45]. Although some techniques used in the proofs are quite different for LASSO and LQLS with $q \in (1, 2]$, the roadmap remains the same. Hence, we suggest readers to first read the proof in this section. Once this relatively simple proof is clear, the other more complicated proofs will be easier to read.

6.1. *Roadmap of the proof.* Since the proof of this result has several steps and is long, we lay out the roadmap of the proof here to help readers navigate through the details. According to Corollary 1 (let us accept Corollary 1 for the moment; its proof will be fully presented in Section E of the Supplementary Material [45]), in order to evaluate $\mathrm{AMSE}(\lambda_{*,1}, 1, \sigma_w)$ as $\sigma_w \to 0$, the crucial step is to characterize $\bar{\sigma}$ from the following equation:

$$\bar{\sigma}^2 = \sigma_\omega^2 + \frac{1}{\delta} \min_{\chi \geq 0} \mathbb{E}_{B,Z}\big[\big(\eta_1(B + \bar{\sigma} Z; \chi) - B\big)^2\big]. \tag{6.1}$$

To study (6.1), the key part is to analyze the term $\min_{\chi \geq 0} \mathbb{E}_{B,Z}[(\eta_1(B + \bar{\sigma} Z; \chi) - B)^2]$. A useful fact that we will prove in Section 6.4 can simplify the analysis of (6.1): The condition $\delta > M_1(\epsilon)$ implies that $\bar{\sigma} \to 0$, as $\sigma_w \to 0$. Hence, one of the main steps of this proof is to derive the convergence rate of $\min_{\chi \geq 0} \mathbb{E}_{B,Z}[(\eta_1(B + \sigma Z; \chi) - B)^2]$, as $\sigma \to 0$. Once we obtain that rate, we then characterize the convergence rate for $\bar{\sigma}$ as $\sigma_w \to 0$ from (6.1). Finally, we connect $\bar{\sigma}$ to AMSE$(\lambda_{*,1}, 1, \sigma_w)$ based on Corollary 1, and derive the expansion for AMSE$(\lambda_{*,1}, 1, \sigma_w)$ as $\sigma_w \to 0$. We introduce the following notation:

$$R(\chi, \sigma) = \mathbb{E}_{B,Z}\big[\big(\eta_1(B/\sigma + Z; \chi) - B/\sigma\big)^2\big], \qquad \chi^*(\sigma) = \arg\min_{\chi \geq 0} R(\chi, \sigma),$$

where we have suppressed the subscript B, Z in $\mathbb{E}$ for notational simplicity. According to [34], $R(\chi, \sigma)$ is a quasi-convex function of χ and has a unique global minimizer. Hence, $\chi^*(\sigma)$ is well defined. It is straightforward to confirm

$$\min_{\chi \geq 0} \mathbb{E}_{B,Z}\big[\big(\eta_1(B + \sigma Z; \chi) - B\big)^2\big] = \sigma^2 R\big(\chi^*(\sigma), \sigma\big).$$

Throughout the proof, we may write χ^* for $\chi^*(\sigma)$ when no confusion is caused, and we use $F(g)$ to denote the distribution function of $|G|$. The rest of the proof of Theorem 3.3 is organized in the following way:

1. We first prove $R(\chi^*(\sigma), \sigma) \to M_1(\varepsilon)$, as $\sigma \to 0$ in Section 6.2.
2. We further bound the convergence rate of $R(\chi^*(\sigma), \sigma)$ in Section 6.3.
3. We finally utilize the convergence rate bound derived in Section 6.3 to characterize the convergence rate of $\bar{\sigma}$ and then derive the expansion for AMSE$(\lambda_{*,1}, 1, \sigma_w)$ in Section 6.4.

6.2. *Proof of $R(\chi^*(\sigma), \sigma) \to M_1(\varepsilon)$, as $\sigma \to 0$.* Our goal in this section is to prove the following lemma.

LEMMA 3. *Suppose $\mathbb{E}|G|^2 < \infty$, then $\lim_{\sigma \to 0} \chi^*(\sigma) = \chi^{**}$ and*

$$\lim_{\sigma \to 0} R\big(\chi^*(\sigma), \sigma\big) = (1 - \varepsilon)\mathbb{E}\big(\eta_1\big(Z; \chi^{**}\big)\big)^2 + \varepsilon\big(1 + \big(\chi^{**}\big)^2\big),$$

where χ^{**} is the unique minimizer of $(1-\varepsilon)\mathbb{E}(\eta_1(Z; \chi))^2 + \varepsilon(1+\chi^2)$ over $[0, \infty)$, and $Z \sim N(0, 1)$.

PROOF. By taking derivatives, it is straightforward to verify that $(1 - \varepsilon)\mathbb{E}(\eta_1(Z; \chi))^2 + \varepsilon(1 + \chi^2)$, as a function of χ over $[0, \infty)$, is strongly convex and has a unique minimizer. Hence, χ^{**} is well defined.

We first claim that $\chi^*(\sigma_n)$ is bounded for any given sequence $\sigma_n \to 0$. Otherwise there exists an unbounded subsequence $\chi^*(\sigma_{n_k}) \to +\infty$ with $\sigma_{n_k} \to 0$. Since the distribution of G does not have point mass at zero and

$$\eta_1\big(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})\big) = \operatorname{sign}(G/\sigma_{n_k} + Z)\big(|G/\sigma_{n_k} + Z| - \chi^*(\sigma_{n_k})\big)_+,$$

it is not hard to conclude that

$$\big|\eta_1\big(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})\big) - G/\sigma_{n_k}\big| \to +\infty \qquad \text{a.s.}$$

By Fatou's lemma, we then have

$$R\big(\chi^*(\sigma_{n_k}), \sigma_{n_k}\big) \geq \varepsilon\mathbb{E}\big(\eta_1\big(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})\big) - G/\sigma_{n_k}\big)^2 \to +\infty. \tag{6.2}$$

On the other hand, the optimality of $\chi^*(\sigma_{n_k})$ implies

$$R(\chi^*(\sigma_{n_k}), \sigma_{n_k}) \le R(0, \sigma_{n_k}) = 1,$$

contradicting the unboundedness in (6.2).

We next show the sequence $\chi^*(\sigma_n)$ converges to a finite constant, for any $\sigma_n \to 0$. Taking a convergent subsequence $\chi^*(\sigma_{n_k})$, due to the boundedness of $\chi^*(\sigma_n)$, the limit of the subsequence is finite. Call it $\tilde{\chi}$. Note that

$$\begin{aligned}
&\mathbb{E}(\eta_1(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})) - G/\sigma_{n_k})^2 \\
&\quad = 1 + \mathbb{E}(\eta_1(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})) - G/\sigma_{n_k} - Z)^2 \\
&\qquad + 2\mathbb{E}Z(\eta_1(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})) - G/\sigma_{n_k} - Z).
\end{aligned}$$

Since $\eta_1(u; \chi) = \text{sign}(u)(|u| - \chi)_+$, we have the following three inequalities:

$$\begin{aligned}
|\eta_1(Z; \chi^*(\sigma_{n_k}))|^2 &\le |Z|^2, \\
(\eta_1(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})) - G/\sigma_{n_k} - Z)^2 &\le (\chi^*(\sigma_{n_k}))^2, \\
|Z(\eta_1(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})) - G/\sigma_{n_k} - Z)| &\le |Z|\chi^*(\sigma_{n_k}).
\end{aligned}$$

Furthermore, all the terms on the right-hand side of the above inequalities are integrable. Therefore, we can apply the Dominated Convergence Theorem (DCT) to obtain

$$\begin{aligned}
&\lim_{n_k \to \infty} R(\chi^*(\sigma_{n_k}), \sigma_{n_k}) \\
&\quad = \lim_{n_k \to \infty} (1 - \varepsilon)\mathbb{E}(\eta_1(Z; \chi^*(\sigma_{n_k})))^2 \\
&\qquad + \varepsilon\mathbb{E}(\eta_1(G/\sigma_{n_k} + Z; \chi^*(\sigma_{n_k})) - G/\sigma_{n_k})^2 \\
&\quad = (1 - \varepsilon)\mathbb{E}(\eta_1(Z; \tilde{\chi}))^2 + \varepsilon(1 + \tilde{\chi}^2).
\end{aligned}$$

Moreover, since $\chi^*(\sigma_{n_k})$ is the optimal threshold value for $R(\chi, \sigma_{n_k})$,

$$\begin{aligned}
\lim_{n_k \to \infty} R(\chi^*(\sigma_{n_k}), \sigma_{n_k}) &\le \lim_{n_k \to \infty} R(\chi^{**}, \sigma_{n_k}) \\
&= (1 - \varepsilon)\mathbb{E}(\eta_1(Z; \chi^{**}))^2 + \varepsilon(1 + (\chi^{**})^2).
\end{aligned}$$

Combining the last two limiting results, we can conclude $\tilde{\chi} = \chi^{**}$. Since $\chi^*(\sigma_{n_k})$ is an arbitrary convergent subsequence, this implies that the sequence $\chi^*(\sigma_n)$ converges to χ^{**} as well. This is true for any $\sigma_n \to 0$; hence, $\chi^*(\sigma) \to \chi^{**}$, as $\sigma \to 0$. $\lim_{\sigma \to 0} R(\chi^*(\sigma), \sigma)$ can then be directly derived. □

6.3. *Bounding the convergence rate of* $R(\chi^*(\sigma),\sigma)$. In Section 6.2, we have shown $R(\chi^*(\sigma),\sigma)\to M_1(\varepsilon)$ as $\sigma\to 0$. Our goal in this section is to bound the difference $R(\chi^*(\sigma),\sigma)-M_1(\varepsilon)$. For that purpose, we first bound the convergence rate of $\chi^*(\sigma)$.

LEMMA 4. *Suppose* $\mathbb{P}(|G|\geq\mu)=1$ *with* μ *being a positive constant and* $\mathbb{E}|G|^2<\infty$, *then as* $\sigma\to 0$:

$$|\chi^*(\sigma)-\chi^{**}|=O(\phi(-\mu/\sigma+\chi^{**})),$$

where $\phi(\cdot)$ *is the density function of the standard normal.*

PROOF. Since $\chi^*(\sigma)$ minimizes $R(\chi,\sigma)$, we have $\frac{\partial R(\chi^*(\sigma),\sigma)}{\partial\chi}=0$, which gives the following expression for $\chi^*(\sigma)$:

$$\chi^*(\sigma)=\frac{2(1-\varepsilon)\phi(\chi^*)+\varepsilon\mathbb{E}\phi(\chi^*-G/\sigma)+\varepsilon\mathbb{E}\phi(\chi^*+G/\sigma)}{2(1-\varepsilon)\int_{\chi^*}^{\infty}\phi(z)\,dz+\varepsilon\mathbb{E}\int_{\chi^*-G/\sigma}^{\infty}\phi(z)\,dz+\varepsilon\mathbb{E}\int_{-\infty}^{-\chi^*-G/\sigma}\phi(z)\,dz}.$$

Letting σ go to zero on both sides in the above equation, we then obtain

$$\chi^{**}=\frac{2(1-\varepsilon)\phi(\chi^{**})}{2(1-\varepsilon)\int_{\chi^{**}}^{\infty}\phi(z)\,dz+\varepsilon},$$

where we have applied Dominated Convergence Theorem (DCT). To bound $|\chi^*(\sigma)-\chi^{**}|$, we first bound the convergence rate of the terms in the expression of $\chi^*(\sigma)$. A direct application of the mean value theorem leads to

$$\phi(\chi^*)-\phi(\chi^{**})=(\chi^{**}-\chi^*)\tilde{\chi}\phi(\tilde{\chi}),\tag{6.3}$$

$$\int_{\chi^*}^{\infty}\phi(z)\,dz-\int_{\chi^{**}}^{\infty}\phi(z)\,dz=(\chi^{**}-\chi^*)\phi(\tilde{\tilde{\chi}}),\tag{6.4}$$

with $\tilde{\chi}$, $\tilde{\tilde{\chi}}$ being two numbers between χ^* and χ^{**}. We now consider the other four terms. By the condition $\mathbb{P}(|G|\geq\mu)=1$, we can conclude that for sufficiently small σ,

$$\mathbb{E}\phi(\chi^*-G/\sigma)\leq\mathbb{E}\phi(\chi^*-|G|/\sigma)\leq\phi(\mu/\sigma-\chi^*),\tag{6.5}$$

$$\mathbb{E}\phi(\chi^*+G/\sigma)\leq\mathbb{E}\phi(\chi^*-|G|/\sigma)\leq\phi(\mu/\sigma-\chi^*).\tag{6.6}$$

Moreover, it is not hard to derive

$$\begin{aligned}&1-\mathbb{E}\int_{\chi^*-G/\sigma}^{\infty}\phi(z)\,dz-\mathbb{E}\int_{-\infty}^{-\chi^*-G/\sigma}\phi(z)\,dz\\&\quad=\int_0^{\infty}\int_{-\chi^*-g/\sigma}^{\chi^*-g/\sigma}\phi(z)\,dz\,dF(g)\\&\quad\leq\int_{-\chi^*-\mu/\sigma}^{\chi^*-\mu/\sigma}\phi(z)\,dz\leq 2\chi^*\phi(\mu/\sigma-\chi^*),\end{aligned}\tag{6.7}$$

where to obtain the last two inequalities we have used the condition $\mathbb{P}(|G| \geq \mu) = 1$ and the fact $\chi^* - \mu/\sigma < 0$ for σ small enough. We are now in the position to bound $|\chi^*(\sigma) - \chi^{**}|$. Define the following notation:

$$\begin{aligned}
&e_1 \triangleq \varepsilon \mathbb{E} \int_{\chi^* - G/\sigma}^{\infty} \phi(z)\,dz + \varepsilon \mathbb{E} \int_{-\infty}^{-\chi^* - G/\sigma} \phi(z)\,dz - \varepsilon, \\
&e_2 \triangleq \varepsilon \mathbb{E}\phi(\chi^* - G/\sigma) + \varepsilon \mathbb{E}\phi(\chi^* + G/\sigma), \\
&S \triangleq 2(1-\varepsilon)\phi(\chi^{**}), \qquad T \triangleq 2(1-\varepsilon) \int_{\chi^{**}}^{\infty} \phi(z)\,dz + \varepsilon.
\end{aligned}$$

Using the new notation and equations (6.3) and (6.4), we obtain

$$\chi^*(\sigma) = \frac{S + 2(1-\varepsilon)(\chi^{**} - \chi^*)\tilde{\chi}\phi(\tilde{\chi}) + e_2}{T + 2(1-\varepsilon)(\chi^{**} - \chi^*)\phi(\tilde{\tilde{\chi}}) + e_1}, \qquad \chi^{**} = \frac{S}{T}.$$

Hence, we can do the following calculations:

$$\begin{aligned}
\chi^*(\sigma) - \chi^{**} &= \frac{S + 2(1-\varepsilon)(\chi^{**} - \chi^*)\tilde{\chi}\phi(\tilde{\chi}) + e_2}{T + 2(1-\varepsilon)(\chi^{**} - \chi^*)\phi(\tilde{\tilde{\chi}}) + e_1} - \frac{S}{T} \\
&= \frac{2(1-\varepsilon)(\chi^{**} - \chi^*)\tilde{\chi}\phi(\tilde{\chi}) + e_2}{T + 2(1-\varepsilon)(\chi^{**} - \chi^*)\phi(\tilde{\tilde{\chi}}) + e_1} \\
&\quad - \frac{S(2(1-\varepsilon)(\chi^{**} - \chi^*)\phi(\tilde{\tilde{\chi}}) + e_1)}{T(T + 2(1-\varepsilon)(\chi^{**} - \chi^*)\phi(\tilde{\tilde{\chi}}) + e_1)} \\
&= \frac{2(1-\varepsilon)(\chi^{**} - \chi^*)(\tilde{\chi}\phi(\tilde{\chi}) - \chi^{**}\phi(\tilde{\tilde{\chi}}))}{T + 2(1-\varepsilon)(\chi^{**} - \chi^*)\phi(\tilde{\tilde{\chi}}) + e_1} \\
&\quad + \frac{e_2 - \chi^{**}e_1}{T + 2(1-\varepsilon)(\chi^{**} - \chi^*)\phi(\tilde{\tilde{\chi}}) + e_1}.
\end{aligned} \tag{6.8}$$

From (6.8), we obtain

$$\begin{aligned}
&(\chi^*(\sigma) - \chi^{**})\biggl(1 + \frac{2(1-\varepsilon)(\tilde{\chi}\phi(\tilde{\chi}) - \chi^{**}\phi(\tilde{\tilde{\chi}}))}{T + 2(1-\varepsilon)(\chi^{**} - \chi^*(\sigma))\phi(\tilde{\tilde{\chi}}) + e_1}\biggr) \\
&\quad = \frac{e_2 - \chi^{**}e_1}{T + 2(1-\varepsilon)(\chi^{**} - \chi^*(\sigma))\phi(\tilde{\tilde{\chi}}) + e_1}.
\end{aligned} \tag{6.9}$$

Note that in the above expression we have $\tilde{\chi} \to \chi^{**}$ and $\tilde{\tilde{\chi}} \to \chi^{**}$ since $\chi^*(\sigma) \to \chi^{**}$. Therefore, we conclude that $\tilde{\chi}\phi(\tilde{\chi}) - \chi^{**}\phi(\tilde{\tilde{\chi}}) \to 0$ and $(\chi^{**} - \chi^*(\sigma))\phi(\tilde{\tilde{\chi}}) \to 0$. Moreover, since (6.5), (6.6) and (6.7) together show both e_1 and e_2 go to 0 exponentially fast, we conclude from (6.9) that $(\chi^*(\sigma) - \chi^{**})/\sigma \to 0$.

This enables us to proceed:

$$\lim_{\sigma\to 0}\frac{|\chi^*(\sigma)-\chi^{**}|}{\phi(\mu/\sigma-\chi^{**})}=\lim_{\sigma\to 0}\frac{|\chi^*(\sigma)-\chi^{**}|}{\phi(\mu/\sigma-\chi^{*})}\overset{\text{(a)}}{=}\lim_{\sigma\to 0}\frac{|e_2-\chi^{**}e_1|}{T\phi(\mu/\sigma-\chi^{*})}$$
$$\overset{\text{(b)}}{\le}\lim_{\sigma\to 0}\frac{2\varepsilon(1+\chi^*(\sigma)\chi^{**})\phi(\mu/\sigma-\chi^*)}{T\phi(\mu/\sigma-\chi^*)}=\frac{2\varepsilon(1+(\chi^{**})^2)}{T}.$$

We have used (6.9) to obtain (a). We derived (b) by the following steps:

1. According to (6.7), $|e_1|\le 2\varepsilon\chi^*\phi(\mu/\sigma-\chi^*)$.
2. According to (6.5) and (6.6), $|e_2|\le 2\varepsilon\phi(\mu/\sigma-\chi^*)$.

This completes the proof of Lemma 4. □

The next step is to bound the convergence rate of $R(\chi^*(\sigma),\sigma)$ based on the convergence rate of $\chi^*(\sigma)$ we have derived in Lemma 4.

LEMMA 5. *Suppose* $\mathbb{P}(|G|\ge\mu)=1$ *with* μ *being a positive constant and* $\mathbb{E}|G|^2<\infty$, *then as* $\sigma\to 0$,

$$|R(\chi^*(\sigma),\sigma)-M_1(\varepsilon)|=O(\phi(\mu/\sigma-\chi^{**})),$$

where $\phi(\cdot)$ *is the density function of the standard normal.*

PROOF. We recall the two quantities

$$M_1(\varepsilon)=(1-\varepsilon)\mathbb{E}(\eta_1(Z;\chi^{**}))^2+\varepsilon(1+(\chi^{**})^2), \tag{6.10}$$

$$\begin{aligned}R(\chi^*(\sigma),\sigma)&=(1-\varepsilon)\mathbb{E}(\eta_1(Z;\chi^*))^2\\&\quad+\varepsilon[1+\mathbb{E}(\eta_1(G/\sigma+Z;\chi^*)-G/\sigma-Z)^2]\\&\quad+2\varepsilon\mathbb{E}Z(\eta_1(G/\sigma+Z;\chi^*)-G/\sigma-Z).\end{aligned} \tag{6.11}$$

We bound $|R(\chi^*(\sigma),\sigma)-M_1(\varepsilon)|$ by bounding the difference between the corresponding terms in (6.11) and (6.10). From the proof of Lemma 4, we know $e_1<0$ and $e_2>0$. Hence, (6.9) implies $\chi^*(\sigma)>\chi^{**}$ for small enough σ. We start with

$$\begin{aligned}&|\mathbb{E}(\eta_1(Z;\chi^*))^2-\mathbb{E}(\eta_1(Z;\chi^{**}))^2|\\&\quad=|\mathbb{E}(\eta_1(Z;\chi^*)-\eta_1(Z;\chi^{**}))(\eta_1(Z;\chi^*)+\eta_1(Z;\chi^{**}))|\\&\quad\le\mathbb{E}[|\chi^*-\chi^{**}+\chi^*\mathbb{I}(|Z|\in(\chi^{**},\chi^*))|\cdot|\eta_1(Z;\chi^*)+\eta_1(Z;\chi^{**})|]\\&\quad\overset{\text{(a)}}{\le}2(\chi^*-\chi^{**})\cdot\mathbb{E}|Z|+2\chi^*\mathbb{E}[\mathbb{I}(|Z|\in(\chi^{**},\chi^*))|Z|]\\&\quad\le 2(\chi^*-\chi^{**})\cdot\mathbb{E}|Z|+4\chi^*(\chi^*-\chi^{**})\tilde{\chi}\phi(\tilde{\chi})=O(\phi(\mu/\sigma-\chi^{**})),\end{aligned} \tag{6.12}$$

where we have used the fact $|\eta_1(u;\chi)| \le |u|$ to obtain (a); $\tilde{\chi}$ is a number between $\chi^*(\sigma)$ and χ^{**}; and the last equality is due to Lemma 4. We next bound the difference between $\mathbb{E}(\eta_1(G/\sigma+Z;\chi^*)-G/\sigma-Z)^2$ and $(\chi^{**})^2$:

$$\begin{aligned}&|(\chi^{**})^2-\mathbb{E}(\eta_1(G/\sigma+Z;\chi^*)-G/\sigma-Z)^2|\\ &\quad\le|(\chi^*)^2-\mathbb{E}(\eta_1(G/\sigma+Z;\chi^*)-G/\sigma-Z)^2|+|(\chi^{**})^2-(\chi^*)^2|.\end{aligned}\tag{6.13}$$

To bound the two terms on the right-hand side of (6.13), first note that

$$\begin{aligned}0&\le(\chi^*)^2-\mathbb{E}(\eta_1(G/\sigma+Z;\chi^*)-G/\sigma-Z)^2\\ &=\mathbb{E}[\mathbb{I}(|G/\sigma+Z|\le\chi^*)\cdot((\chi^*)^2-(G/\sigma+Z)^2)]\\ &\le(\chi^*)^2\int_0^\infty\int_{-g/\sigma-\chi^*}^{-g/\sigma+\chi^*}\phi(z)\,dz\,dF(g)\\ &\overset{(b)}{\le}(\chi^*)^2\int_{-\mu/\sigma-\chi^*}^{-\mu/\sigma+\chi^*}\phi(z)\,dz\le2(\chi^*)^3\phi(\mu/\sigma-\chi^*)\\ &=O(\phi(\mu/\sigma-\chi^{**})),\end{aligned}\tag{6.14}$$

where (b) is due to the condition $\mathbb{P}(|G|\ge\mu)=1$, and the last equality holds since $(\chi^*-\chi^{**})/\sigma\to0$ implied by Lemma 4. Furthermore, Lemma 4 yields

$$(\chi^*)^2-(\chi^{**})^2=O(\phi(\mu/\sigma-\chi^{**})).\tag{6.15}$$

Combining (6.13), (6.14) and (6.15), we obtain

$$|(\chi^{**})^2-\mathbb{E}(\eta_1(G/\sigma+Z;\chi^*)-G/\sigma-Z)^2|=O(\phi(\mu/\sigma-\chi^{**})).\tag{6.16}$$

Regarding the remaining term in $R(\chi^*(\sigma),\sigma)$, we can derive

$$\begin{aligned}0&\le\mathbb{E}Z(G/\sigma+Z-\eta_1(G/\sigma+Z;\chi^*))\\ &\overset{(c)}{=}\mathbb{E}(1-\partial_1\eta_1(G/\sigma+Z;\chi^*))\\ &=\mathbb{P}(|G/\sigma+Z|\le\chi^*)\overset{(d)}{=}O(\phi(\mu/\sigma-\chi^{**})).\end{aligned}\tag{6.17}$$

We have employed Stein's lemma (see Lemma E in the Supplementary Material) to obtain (c). Equality (d) holds due to (6.7). Putting the results (6.12), (6.16) and (6.17) together completes the proof. □

6.4. *Deriving the expansion of* AMSE$(\lambda_{*,1},1,\sigma_w)$. In this section, we utilize the convergence rate result of $R(\chi^*(\sigma),\sigma)$ from Section 6.3 to derive the expansion of AMSE$(\lambda_{*,1},1,\sigma_w)$ in (3.6), and thus complete the proof of Theorem 3.3. Toward that goal, we first prove a useful lemma.

LEMMA 6. *Let $\bar{\sigma}$ be the solution to the following equation*:

$$\bar{\sigma}^2 = \sigma_\omega^2 + \frac{1}{\delta} \min_{\chi \geq 0} \mathbb{E}_{B,Z}\big[\big(\eta_1(B + \bar{\sigma} Z; \chi) - B\big)^2\big]. \tag{6.18}$$

Suppose $\delta > M_1(\varepsilon)$, then

$$\lim_{\sigma_w \to 0} \frac{\sigma_w^2}{\bar{\sigma}^2} = \frac{\delta - M_1(\varepsilon)}{\delta}.$$

PROOF. We first claim that $\mathbb{E}(\eta_1(\alpha + Z; \chi) - \alpha)^2$ is an increasing function of α, because

$$\frac{d}{d\alpha} \mathbb{E}\big(\eta_1(\alpha + Z; \chi) - \alpha\big)^2 = 2\mathbb{E}\big(\alpha \mathbb{I}\big(|\alpha + Z| \leq \chi\big)\big) \geq 0.$$

Hence, we obtain

$$\mathbb{E}\big(\eta_1(\alpha + Z; \chi) - \alpha\big)^2 \leq \lim_{\alpha \to \infty} \mathbb{E}\big(\eta_1(\alpha + Z; \chi) - \alpha\big)^2 = 1 + \chi^2. \tag{6.19}$$

Inequality (6.19) then yields

$$\begin{aligned} R(\chi, \bar{\sigma}) &= (1 - \varepsilon)\mathbb{E}\big(\eta_1(Z; \chi)\big)^2 + \varepsilon \mathbb{E}\big(\eta_1(G/\bar{\sigma} + Z; \chi) - G/\bar{\sigma}\big)^2 \\ &\leq (1 - \varepsilon)\mathbb{E}\big(\eta_1(Z; \chi)\big)^2 + \varepsilon\big(1 + \chi^2\big). \end{aligned}$$

Taking minimum over χ on both sides above gives us

$$R\big(\chi^*(\bar{\sigma}), \bar{\sigma}\big) \leq M_1(\varepsilon). \tag{6.20}$$

Moreover, since $\bar{\sigma}$ is the solution of (6.18), it satisfies

$$\bar{\sigma}^2 = \sigma_w^2 + \frac{\bar{\sigma}^2}{\delta} R\big(\chi^*(\bar{\sigma}), \bar{\sigma}\big). \tag{6.21}$$

Combining (6.20) and (6.21) with the condition $\delta > M_1(\varepsilon)$, we have

$$\bar{\sigma}^2 \leq \frac{\sigma_w^2}{1 - M_1(\varepsilon)/\delta},$$

which leads to $\bar{\sigma} \to 0$, as $\sigma_w \to 0$. Then applying Lemma 3 shows

$$\lim_{\sigma_w \to 0} R\big(\chi^*(\bar{\sigma}), \bar{\sigma}\big) = \lim_{\bar{\sigma} \to 0} R\big(\chi^*(\bar{\sigma}), \bar{\sigma}\big) = M_1(\varepsilon).$$

Diving both sides of (6.21) by $\bar{\sigma}^2$ and letting $\sigma_w \to 0$ completes the proof. □

To complete the proof of Theorem 3.3, first note that Corollary 1 tells us

$$\text{AMSE}(\lambda_{*,1}, 1, \sigma_w) = \bar{\sigma}^2 R\big(\chi^*(\bar{\sigma}), \bar{\sigma}\big), \qquad \sigma_w^2 = \bar{\sigma}^2 - \frac{\bar{\sigma}^2}{\delta} R\big(\chi^*(\bar{\sigma}), \bar{\sigma}\big).$$

We then have

$$
\begin{aligned}
&\mathrm{AMSE}(\lambda_{*,1}, 1, \sigma_w) - \frac{\delta M_1(\varepsilon)}{\delta - M_1(\varepsilon)}\sigma_w^2 \\
&\quad = \bar{\sigma}^2 R(\chi^*(\bar{\sigma}), \bar{\sigma}) - \frac{\delta M_1(\varepsilon)}{\delta - M_1(\varepsilon)} \cdot \left[\bar{\sigma}^2 - \frac{\bar{\sigma}^2}{\delta} R(\chi^*(\bar{\sigma}), \bar{\sigma})\right] \\
&\quad = \frac{\delta(R(\chi^*(\bar{\sigma}), \bar{\sigma}) - M_1(\varepsilon))}{\delta - M_1(\varepsilon)}\bar{\sigma}^2 \overset{\text{(a)}}{=} O(\bar{\sigma}^2 \phi(\mu/\bar{\sigma} - \chi^{**})),
\end{aligned}
\tag{6.22}
$$

where (a) is due to Lemma 5. Finally, since $\lim_{\sigma_w \to 0} \frac{\sigma_w^2}{\bar{\sigma}^2} = \frac{\delta - M_1(\varepsilon)}{\delta}$ according to Lemma 6, it is not hard to see

$$
O(\bar{\sigma}^2 \phi(\mu/\bar{\sigma} - \chi^{**})) = o(\phi(\bar{\mu}/\bar{\sigma})) = o\left(\phi\left(\sqrt{\frac{\delta - M_1(\varepsilon)}{\delta}} \frac{\tilde{\mu}}{\sigma_w}\right)\right),
\tag{6.23}
$$

where $\bar{\mu}$ and $\tilde{\mu}$ are any constants satisfying $0 \le \tilde{\mu} < \bar{\mu} < \mu$. Results (6.22) and (6.23) together close the proof of Theorem 3.3.

REMARK. (6.20) and (6.22) together imply that the second dominant term of $\mathrm{AMSE}(\lambda_{*,1}, 1, \sigma_w)$ is in fact negative.

Acknowledgments. We thank the Editor, Associate Editor and two anonymous referees for their helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Overcoming the limitations of phase transition by higher order analysis of regularization techniques" (DOI: 10.1214/17-AOS1651SUPP; .pdf). Due to space constraints, additional simulations and technical proofs are relegated a supplementary document in [45], which contains Sections A–J.

REFERENCES

[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453

[2] 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

[3] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017. MR2951312

[4] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

[6] BRADIC, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electron. J. Stat.* **10** 3894–3944. MR3581957

[7] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

[8] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[9] CANDÈS, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris* **346** 589–592. MR2412803

[10] COOLEN, A. C. C. (2005). *The Mathematical Theory of Minority Games*: *Statistical Mechanics of Interacting Agents*. Oxford Univ. Press, Oxford. MR2127290

[11] DONOHO, D. and MONTANARI, A. (2015). Variance breakdown of Huber (M)-estimators. $n/p \to m$. Preprint.

[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

[13] DONOHO, D. L. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.* **35** 617–652. MR2225676

[14] DONOHO, D. L. (2006). For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59** 907–934. MR2222440

[15] DONOHO, D. L., GAVISH, M. and MONTANARI, A. (2013). The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proc. Natl. Acad. Sci. USA* **110** 8405–8410. MR3082268

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

[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

[18] 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

[19] DONOHO, D. L. and TANNER, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA* **102** 9446–9451. MR2168715

[20] DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** 9452–9457. MR2168716

[21] EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. Preprint. Available at arXiv:1311.2445.

[22] EL KAROUI, N., BEAN, D., BICKEL, P., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.

[23] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[24] FRANK, L. and FRIEDMAN, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

[25] FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

[26] GUO, D. and VERDÚ, S. (2005). Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Trans. Inform. Theory* **51** 1983–2010. MR2235278

[27] HOERL, A. and KENNARD, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[28] HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

[29] KABASHIMA, Y., WADAYAMA, T. and TANAKA, T. (2009). A typical reconstruction limit for compressed sensing based on L_p-norm minimization. *J. Stat. Mech. Theory Exp.* **2009** L09003.

[30] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[31] KOLTCHINSKII, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.* **45** 7–57. MR2500227

[32] KRZAKALA, F., MÉZARD, M., SAUSSET, F., SUN, Y. and ZDEBOROVÁ, L. (2012). Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X* **2** 021005.

[33] MALEKI, A. (2010). Approximate message passing algorithms for compressed sensing. Ph.D. thesis, Stanford Univ., Stanford, CA.

[34] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2017). Consistent parameter estimation for LASSO and approximate message passing. *Ann. Statist.* **45** 2427–2454. MR3737897

[35] OYMAK, S. and HASSIBI, B. (2016). Sharp MSE bounds for proximal denoising. *Found. Comput. Math.* **16** 965–1029. MR3529131

[36] OYMAK, S., THRAMPOULIDIS, C. and HASSIBI, B. (2013). The squared-error of generalized lasso: A precise analysis. In 51*st Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 1002–1009. IEEE, New York.

[37] RANGAN, S., GOYAL, V. and FLETCHER, A. (2009). Asymptotic analysis of map estimation via the replica method and compressed sensing. In *Advances in Neural Information Processing Systems* 1545–1553.

[38] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

[39] STOJNIC, M. (2009). Various thresholds for ℓ_1-optimization in compressed sensing. Preprint. Available at arXiv:0907.3666.

[40] SU, W., BOGDAN, M. and CANDES, E. (2015). False discoveries occur early on the lasso path. Preprint. Available at arXiv:1511.01957.

[41] TANAKA, T. (2002). A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Trans. Inform. Theory* **48** 2888–2910. MR1945581

[42] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2016). Precise error analysis of regularized M-estimators in high-dimensions. Preprint. Available at arXiv:1601.06233.

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

[44] WENG, H. and MALEKI, A. (2017). Low noise sensitivity analysis of ℓ_q-minimization in oversampled systems. Preprint. Available at arXiv:1705.03533.

[45] WENG, H., MALEKI, A. and ZHENG, L. (2018). Supplement to "Overcoming the limitations of phase transition by higher order analysis of regularization techniques." DOI:10.1214/17-AOS1651SUPP.

[46] ZHENG, L., MALEKI, A., WENG, H., WANG, X. and LONG, T. (2016). Does ℓ_p-minimization outperform ℓ_1-minimization? Preprint. Available at arXiv:1501.03704v2.

[47] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

H. WENG
A. MALEKI
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
1255 AMSTERDAM AVENUE
NEW YORK, NEW YORK, 10027
USA
E-MAIL: hw2375@columbia.edu
arian@stat.columbia.edu

L. ZHENG
DEPARTMENT OF ELECTRICAL ENGINEERING
COLUMBIA UNIVERSITY
1300 S. W. MUDD BUILDING, MC 4712
500 W. 120TH STREET
NEW YORK, NEW YORK 10027
USA
E-MAIL: le.zheng.cn@gmail.com