Phase transition and higher order analysis of L_q regularization under dependence

Hanwen Huang

Department of Epidemiology and Biostatistics
University of Georgia, Athens, GA 30602
huanghw@uga.edu

Peng Zeng

Department of Mathematics & Statistics
Auburn University, Auburn, AL 36849
zengpen@auburn.edu

Qinglong Yang

School of Statistics and Mathematics
Zhongnan University of Economics and Law
Wuhan, Hubei 430073, P. R. China
yangqinglong@zuel.edu.cn

Abstract

We study the problem of estimating a k-sparse signal $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ from a set of noisy observations $\mathbf{y} \in \mathbb{R}^n$ under the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + w$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the measurement matrix the row of which is drawn from distribution $N(0, \boldsymbol{\Sigma})$. We consider the class of L_q-regularized least squares (LQLS) given by the formulation $\hat{\boldsymbol{\beta}}(\lambda, q) = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q^q$, where $\|\cdot\|_q$ ($0 \le q \le 2$) denotes the L_q-norm. In the setting $p, n, k \to \infty$ with fixed $k/p = \epsilon$ and $n/p = \delta$, we derive the asymptotic risk of $\hat{\boldsymbol{\beta}}(\lambda, q)$ for arbitrary covariance matrix $\boldsymbol{\Sigma}$ which generalizes the existing results for standard Gaussian design, i.e. $X_{ij} \overset{i.i.d}{\sim} N(0, 1)$. We perform a higher-order analysis for LQLS in the small-error regime in which the first dominant term can be used to determine the phase transition behavior of LQLS. Our results show that the first dominant term does not depend on the covariance structure of $\boldsymbol{\Sigma}$ in the cases $0 \le q < 1$ and $1 < q \le 2$ which indicates that the correlations among predictors only affect the phase transition curve in the case $q = 1$ a.k.a. LASSO. To study the influence of the covariance structure of $\boldsymbol{\Sigma}$ on the performance of LQLS in the cases $0 \le q < 1$ and $1 < q \le 2$, we derive the explicit formulas for the second dominant term in the expansion of the asymptotic risk in terms of small error. Extensive

computational experiments confirm that our analytical predictions are consistent with numerical results.

Keywords: L_q-regularization, least squares, phase transition, higher order analysis

1   Introduction

The goal of linear regression is to estimate the parameter vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ from a set of n response variables $\mathbf{y} \in \mathbb{R}^n$ under the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{w}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a known measurement matrix and $\mathbf{w} \in \mathbb{R}^n$ is a noise vector. We consider the popular class of L_q-regularized least square methods (LQLS), given by the optimization problem

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q^q, \tag{2}$$

where $\|\cdot\|_q$ $(0 \le q \le 2)$ denotes the L_q-norm and $\lambda \ge 0$ is a fixed number.

In many modern statistical applications with large p, the true $\boldsymbol{\beta}_0$ vector is often sparse, i.e. only k of the elements of $\boldsymbol{\beta}_0$ are non-zero and the rest are zero. We would like to identify the useful predictors and also obtain good estimates of their coefficients. Among all LQLSs, the case $q = 1$, or in other words L_1 optimization a.k.a. LASSO, is the most popular and best studied scheme for identifying a set of relevant features from a given list. LASSO enjoys attractive statistical properties (Fu and Knight, 2000; Donoho, 2006; Meinshausen and Bühlmann, 2006) and has been effectively used for simultaneously producing accurate and parsimonious models.

However, as a convex relaxation of L_0-norm, the L_1-norm may be sub-optimal for recovering a real sparse signal because it tends to result in biased estimation by shrinking all the entries toward to zero simultaneously, and sometimes leads to over-penalization. Many researchers have shown that using the L_q-norm $(0<q<1)$ to approximate the L_0-norm is in fact a better choice than using the L_1-norm because the former provides a tighter optimization relaxation to the original L_0 sparsity finding formulation. For example, in the noiseless settings $(w = 0)$, Chartrand and Staneva (2008); Saab et al. (2008); Saab and Özgür Yılmaz (2010) have shown that the global minima of (2) for $q<1$ outperforms the solution of LASSO. They have derived the Restricted Isometry Property conditions that can guarantee the accurate recovers of the sparse vector from (2).

To understand the behavior of these algorithms in high dimension more reliably, a new asymptotic has been proposed recently which assumes that $p, n, k \to \infty$ with fixed $k/p = \epsilon$ and $n/p = \delta$. One of the main studies undertaken in this asymptotic framework is to find the relationship between the sparsity of the vector $\boldsymbol{\beta}_0$ and the successful recovery of it using

L_q-minimization algorithm (2) under the noiseless setting. In other word, for different values of δ, we need to know the boundary of sparsity ϵ that differentiates the success and failure of recovering by L_q-minimization. This boundary is known as the phase transition curve.

Stojnic (2013) characterizes the phase transition curve for the case of $q = 0$. Wang et al. (2011) analyzes the phase transition curve for $\delta \to 1$. Kabashima et al. (2009) and Rangan et al. (2009) derives the exact value of phase transition curve for any value of $0 \leq q \leq 1$ and any value of δ using the non-rigorous replica method. In addition to the sharp characterization of the phase transition in noiseless case, Weng et al. (2016); Zheng et al. (2017) also present accurate calculation of the mean squared error (MSE) in the presence of noise and compare the accuracy of estimator (2) for different values of q under the optimal tuning of the parameter λ. They observe that if the measurement noise w is zero or small, then the global minima of (2) for $q<1$ (when λ is optimally picked) outperforms the solution of LASSO with optimal λ. Furthermore, all values of $q<1$ have the same performance when $w = 0$. When w is small, LQLS with the value of q closer to 0 has a better performance. This coincides with the intuition that the performance of L_q- minimization is improved when q decreases since it is closer to L_0-minimization. The analysis of Zheng et al. (2017) is based on approximate message passing (AMP) algorithm and replica method.

The phase transition of LQLS in the case $1<q \leq 2$, also known as bridge regression, has been thoroughly studied in last several decades with different techniques, such as combinatorial geometry (Donoho and Tanner, 2005), statistical dimension framework (Amelunxen et al., 2013), Gordon's lemma (Thrampoulidis et al., 2018), and AMP (Donoho et al., 2011; Weng et al., 2018; Wang et al., 2020; Ma et al., 2019). It was summarized in Weng et al. (2018); Weng and Maleki (2019); Wang et al. (2021) that the phase transition curve is $\delta_q(\epsilon) = 1$ if $2 \geq q>1$ and $\delta_q(\epsilon) = M_1(\epsilon)$ if $q = 1$, where $M_1(\epsilon)$ is an increasing function with $M_1(0) = 0$ and $M_1(1) = 1$. Similar to Zheng et al. (2017), they also propose a higher-order analysis for LQLS in the small-error regime in which the first dominant term is the one that specifies the phase transition curve and the second dominant term can be used to compare the performance of different values of q. They observed that the actual MSE is smaller than the one predicted by the first-order term.

All the above results are based on the i.i.d. Gaussian design assumption i.e. $X_{i,j} \sim N(0, 1/n)$. Our aim in this paper is to study the phase transition under arbitrary covariance dependence, i.e. $\mathbf{X}$ consists of i.i.d. Gaussian rows $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} \succ 0$ and $\boldsymbol{\Sigma} \neq \mathbf{I}_p$. The phase transition curve in the case $q = 1$, i.e. LASSO, under arbitrary covariance dependence has been studied in Huang (2021). It was found that the LASSO phase transition curve changes with the correlation coefficients when the signed patterns of the nonzero components of $\boldsymbol{\beta}_0$ are not symmetric. In current paper, we focus on $0 \leq q<1$ and $1<q \leq 2$. Toward this goal, we first derive the limiting prediction risk of LQLS estimator (2) under asymptotic setting $p, n \to \infty$ with fixed $n/p = \delta$ using the non-rigorous, but widely accepted replica method from statistical physics. Then we apply the higher order analysis to the

asymptotic prediction risk in terms of small noise error for a simple block diagonal covariance matrix structure.

Here is the summary of our results: 1. The first dominant term is not influenced by the actual covariance matrix $\boldsymbol{\Sigma}$ in both cases of $0 \le q<1$ and $1<q \le 2$. This is in sharp contrast to the case $q = 1$ and indicates that the phase transition curve is $\delta = \epsilon$ if $0 \le q<1$ and $\delta = 1$ if $1<q \le 2$ for any covariance structure; 2. The second dominant term depends on the correlation coefficient ρ which is positive in the case $0<q<1$ and negative in the case $1<q \le 2$. To the best of our knowledge, this is the first result to illustrate the phase transition and higher analysis for LQLS estimators under design matrices $\mathbf{X}$ that have non-independent entries.

The rest of this paper is organized as follows: In Section 2, we present the asymptotic framework of our analysis from which the distributional limit of estimator (2) is derived. In Section 3, we study the phase transition and higher analysis for the case $0 \le q<1$. In Section 4, we study the phase transition and higher analysis for the case $1<q \le 2$. In Section 5, we show some simulation results to verify that our analytic derivations are indeed correct. The last section is devoted to the conclusion and the proofs of our main propositions are placed in Appendix.

2 Distributional limit of LQLS estimator

The main goal of this section is to formally introduce the asymptotic setting under which we study the prediction risk of LQLS estimator (2). We write a sequence of instances $\{\boldsymbol{\beta}_0(p), \mathbf{w}(p), \boldsymbol{\Sigma}(p), \mathbf{X}(p)\}$ indexed by p which is called a converging sequence if the following conditions hold:

1. $p, n \to \infty$ with $n/p \to \delta$ for some positive constant δ.

2. The empirical distribution of the entries of $\boldsymbol{\beta}_0(p)$ converges weakly to a probability measure p_{β_0} on $\mathbb{R}$ with $\sum_{i=1}^{p} \beta_{0,i}(p)^2/p \to E_{p_{\beta_0}}\{\beta_0^2\}<\infty$.

3. The empirical distribution of the entries of $\mathbf{w}(p)$ converges weakly to a probability measure p_w on $\mathbb{R}$ with $\sum_{i=1}^{n} w_i(p)^2/n^2 \to \sigma_w^2<\infty$.

4. Denote $\lambda_{min}(\boldsymbol{\Sigma}(p))$ and $\lambda_{max}(\boldsymbol{\Sigma}(p))$ the smallest and largest eigenvalues of $\boldsymbol{\Sigma}(p)$. Then $1/\lambda_{min}(\boldsymbol{\Sigma}(p)) = O(1)$ and $\lambda_{max}(\boldsymbol{\Sigma}(p)) = O(1)$.

5. The rows of $\mathbf{X}(p)$ are drawn independently from distribution $N(0, \boldsymbol{\Sigma}(p)/n)$.

6. The sequence of functions

$$\mathcal{E}^{(p)}(a,b) \;\equiv\; \frac{1}{p}E \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(p) - \sqrt{a}\boldsymbol{\Sigma}(p)^{-1/2}\mathbf{z}\|_{\boldsymbol{\Sigma}(p)}^2 + b\|\boldsymbol{\beta}\|_q^q \right\} \qquad (3)$$

4

admits a differentiable limit $\mathcal{E}(a,b)$ on $\mathbb{R}_+ \times \mathbb{R}_+$ with $\operatorname{div}\mathcal{E}^{(p)}(a,b) \to \operatorname{div}\mathcal{E}(a,b)$, where $\|\mathbf{v}\|_\Sigma^2 = \mathbf{v}^T\Sigma\mathbf{v}$ and the expected value is with respect to two independent random vectors $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ and $\boldsymbol{\beta}_0(p)$.

Assume we observe training data $(\mathbf{X}(p), \mathbf{y}(p))$, where $\mathbf{y}(p) = \mathbf{X}(p)\boldsymbol{\beta}_0(p) + \mathbf{w}(p)$. We insist on the fact that $\boldsymbol{\beta}_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p), \mathbf{y}(p)$ depend on p. However, we will drop this dependence most of the time to ease the reading. In order to present our main result, for any $\mathbf{v} \in \mathbb{R}^p$ and $\alpha > 0$, we need to introduce the generalized thresholding operation $\boldsymbol{\eta}_q : \mathbb{R}^p \to \mathbb{R}^p$ which is defined as

$$\boldsymbol{\eta}_q(\mathbf{v}, \alpha) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{\beta} - \mathbf{v}\|_\Sigma^2 + \alpha\|\boldsymbol{\beta}\|_q^q \right\}. \tag{4}$$

It can be easily verified that $\boldsymbol{\eta}_q(\mathbf{v}, \alpha) \xrightarrow{\alpha\to 0} \mathbf{v}$ and $\boldsymbol{\eta}_q(\mathbf{v}, \alpha) \xrightarrow{\alpha\to\infty} 0$. For $\Sigma = \mathbf{I}_{p\times p}$, each component of $\boldsymbol{\eta}_q(\mathbf{v}, \alpha)$ can be solved independently using the corresponding scalar thresholding operator $\eta_q(u, \alpha) = \operatorname{argmin}_{\beta\in\mathbb{R}} \left\{ \frac{1}{2}(\beta - u)^2 + \alpha|\beta|^q \right\}$ whose solution is

$$\eta_q(u,\alpha) = \begin{cases} uI(|u|>\sqrt{2\alpha}) & if \quad q=0 \\ \operatorname{sign}(u)\tilde{\beta}I[|u|>C_q\alpha^{1/(2-q)}] & if \quad 0<q<1 \\ \operatorname{sign}(u)(|u| - \alpha)I(|u|>\alpha) & if \quad q=1 \\ \operatorname{sign}(u)\bar{\beta} & if \quad 1<q<2 \\ u/(1+2\alpha) & if \quad q=2 \end{cases}$$

where $C_q = [2(1-q)]^{1/(2-q)} + q[2(1-q)]^{(q-1)/(2-q)}$, $\tilde{\beta}$ is the largest solution of $\beta + \frac{q\alpha}{\beta^{1-q}} = |u|$, and $\bar{\beta}$ is the solution of $\beta + q\alpha\beta^{q-1} = |u|$. Figure 1 exhibits η_q for different values of q.

For a converging sequence of instances, we can define the function

$$\psi_q(\tau^2, \lambda) = \sigma_w^2 + \lim_{p\to\infty}\frac{1}{p\delta}E\left(\|\boldsymbol{\eta}_q(\boldsymbol{\beta}_0 + \tau\Sigma^{-1/2}\mathbf{z}, \lambda) - \boldsymbol{\beta}_0\|_\Sigma^2\right), \tag{5}$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ is independent of $\boldsymbol{\beta}_0$. Notice that the function $\psi_q(\cdot, \cdot)$ depends implicitly on the law $p_{\boldsymbol{\beta}_0}$.

Denote $\hat{\boldsymbol{\beta}}(\lambda, p)$ the LQLS estimator for instance $\{\boldsymbol{\beta}_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}$ with $\lambda > 0$ based on (2). Then the following proposition establishes the distributional limit of $(\hat{\boldsymbol{\beta}}(\lambda, p), \boldsymbol{\beta}_0)$.

Proposition 1 $(\hat{\boldsymbol{\beta}}(\lambda, p), \boldsymbol{\beta}_0)$ *converges in distribution to the random vector* $(\boldsymbol{\eta}_q(\boldsymbol{\beta}_0 + \tau_\star\Sigma^{-1/2}\mathbf{z}, \alpha), \boldsymbol{\beta}_0)$ *as* $p \to \infty$, *where* $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ *is independent of* $\boldsymbol{\beta}_0 \sim p_{\boldsymbol{\beta}_0}$, *and* $\alpha, \tau_\star^2$ *satisfy the following equations*

$$\tau_\star^2 = \psi_q(\tau_\star^2, \alpha), \tag{6}$$

$$\lambda = \alpha\left(1 - \frac{1}{\delta}E\{\langle\boldsymbol{\eta}_q'(\boldsymbol{\beta}_0 + \tau_\star\Sigma^{-1/2}\mathbf{z}, \alpha)\rangle\}\right), \tag{7}$$

where the expectation is with respect to $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ *and* $\boldsymbol{\beta}_0$, $\boldsymbol{\eta}_q'(\cdot, \cdot)$ *is the derivative of the generalized thresholding function over its first argument, and* $\langle\mathbf{v}\rangle \equiv \sum_{j=1}^p v_j/p$ *is the average of the vector* $\mathbf{v} \in \mathbb{R}^p$.
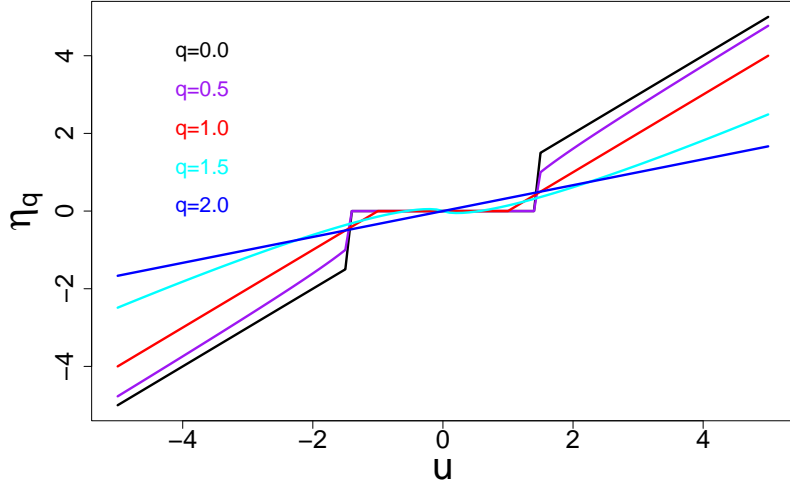
5

Figure 1: $\eta_q(u, \alpha)$ for 5 different values of q. α is set to 1.

The derivation of this proposition is presented in the supplementary material using the replica method which is a non-rigorous technique invented in statistical physics to study the behavior of large magnetic and disordered systems. This method has been used to analyze the accuracy of $\hat{\boldsymbol{\beta}}(\lambda, p)$ for i.i.d. Gaussian $\mathbf{X}$, i.e. $\boldsymbol{\Sigma} = \mathbf{I}_p$ in Rangan et al. (2009). Here we generalize it to arbitrary $\boldsymbol{\Sigma}$.

The performance of LQLS estimator (2) depends on the choice of the threshold parameters λ. Any fair comparison between LQLS for different values of q must take this fact into account. Consider a test point $\mathbf{x}_0 \sim N(0, \boldsymbol{\Sigma})$ independent of the train data. For an estimator $\hat{\boldsymbol{\beta}}$ (a function of the training data $\mathbf{X}, \mathbf{y}$), we define its prediction risk (or simply, risk) as

$$R_q(\hat{\boldsymbol{\beta}}(\lambda)) = \frac{1}{p} E(\mathbf{x}_0^T \hat{\boldsymbol{\beta}}(\lambda) - \mathbf{x}_0^T \boldsymbol{\beta}_0) = \frac{1}{p} E \|\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0\|_{\boldsymbol{\Sigma}}^2.$$

For $\boldsymbol{\Sigma} = \mathbf{I}_p$, $R_q(\hat{\boldsymbol{\beta}}(\lambda))$ is equivalent to the MSE defined in Weng et al. (2016). Hence, it is important to pick λ optimally by minimizing the prediction risk, i.e. $\lambda_\star = \operatorname{argmin}_\lambda R_q(\hat{\boldsymbol{\beta}}(\lambda))$. This is equivalent to consider the following optimal thresholding policy

$$\alpha_\star(\tau) = \operatorname{argmin}_\alpha \lim_{p\to\infty} \frac{1}{p} E \left(\|\boldsymbol{\eta}_q(\boldsymbol{\beta}_0 + \tau \boldsymbol{\Sigma}^{-1/2} \mathbf{z}, \alpha) - \boldsymbol{\beta}_0\|_{\boldsymbol{\Sigma}}^2 \right) \tag{8}$$

since $R_q(\hat{\boldsymbol{\beta}}(\lambda_\star)) = \delta[\psi(\tau_\star^2, \alpha_\star(\tau_\star)) - \sigma_w^2]$. This enables us to focus the analysis on a single equation rather than two equations (6) and (7). The results we will present in the next two sections are mainly based on investigating the solution of the following fixed point equation

$$\tau_\star^2 = \psi_q(\tau_\star^2, \alpha_\star(\tau_\star)). \tag{9}$$

We consider the cases $0 \leq q < 1$ and $1 < q \leq 2$ separately.

6

3   Asymptotic performance for $0 \leq q{<}1$

When $0 \leq q{<}1$, the L_q regularization has the nice property of creating coefficient sparsity because the generalized thresholding operation $\boldsymbol{\eta}_q(\mathbf{v}, \lambda)$ defined in (4) maps small values of $\mathbf{v}$ to zero. Also the smaller values of q correspond to a preference for increasingly sparse solutions. However, the penalty term is not convex for $q{<}1$ which makes solving the optimization problem much more difficult because the local minima may not be unique.

3.1   Noiseless Settings

This section discusses our main results in the noiseless setting $\sigma_w^2 = 0$. Since there is no measurement noise, $\psi_q(\tau^2, \alpha_\star(\tau)) \to 0$ as $\tau^2 \to 0$, thus $\tau^2 = 0$ is a fixed point of (9). If $\tau^2 = 0$ is also a stable fixed point of (9), i.e. there exists $\tau_i{>}0$ such that for every $0{<}\tau{<}\tau_i$, $\psi_q(\tau^2, \alpha_\star(\tau)){<}\tau^2$, then we say LQLS has successfully recovered the sparse solution of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0$. Otherwise, we say LQLS has failed.

The shape of $\psi_q(\tau^2, \alpha_\star(\tau))$ and its fixed points depend on the distribution p_{β_0}. In this work we focus on $p_{\beta_0} \in \mathcal{F}_\epsilon$, where $\mathcal{F}_\epsilon$ denotes the set of distributions whose mass at zero is greater than or equal to $1 - \epsilon$. In other words, $\beta_0 \sim p_{\beta_0}$ implies that $P(\beta_0 \neq 0) \leq \epsilon$. This class of distributions is considered as a good model for exactly sparse signals and has been studied in many other papers. The following proposition identifies the conditions under which zero is a stable fixed point under noiseless setting.

Proposition 2 *Let p_{β_0} be an arbitrary distribution in $\mathcal{F}_\epsilon$. For any $0 \leq q{<}1$, zero is the stable fixed point of $\psi_q(\tau^2, \alpha_\star(\tau))$ under noiseless settings if and only if $\delta{>}\epsilon$. Thus the phase transition curve is $\delta = \epsilon$.*

The proof of this result can be found in Section 7.2. This result extends the conclusion of Theorem 4 in Zheng et al. (2017) from $\boldsymbol{\Sigma} = \mathbf{I}_{p \times p}$ to general $\boldsymbol{\Sigma}$. There are three main features of this proposition that we would like to emphasize. Firstly, the actual distribution that is picked from $\mathcal{F}_\epsilon$ does not have any impact on the behavior of the fixed point at zero. Secondly, this proposition is universal in the sense that it does not depend on the actual structure of covariance matrix $\boldsymbol{\Sigma}$. Thirdly, the number of the measurements δ that is required for the stability of this fixed point is the same as the sparsity level ϵ. As long as $\delta{>}\epsilon$, zero is a stable fixed point and LQLS recovers $\boldsymbol{\beta}_0$ accurately for every $0 \leq q{<}1$. If we are concerned with the noiseless settings, all L_q-minimization algorithms are the same.

Note that since L_q is not convex for $0 \leq q{<}1$, $\psi_q(\tau^2, \alpha_\star(\tau))$ may have additional stable fixed points than $\tau^2 = 0$. The conditions under which this case happens depend on q, p_{β_0}, and the actual structure of $\boldsymbol{\Sigma}$. We will provide a numeric study on this type of phase transition in Section 5.1 for $q = 0$ and a block diagonal $\boldsymbol{\Sigma}$.

As pointed out in Zheng et al. (2017), this result seems to be counter-intuitive. For instance, we expect to see that the performance difference between $q = 0$ and $q = 0.9$ is bigger than the performance difference between $q = 0$ and $q = 0.1$. We also expect to see the impact of covariance structure of Σ on the performance of LQLS estimators. However, according to the phase transition analysis of Proposition 2, the performance of all LQLS are the same under the noiseless settings. This naturally leads to the following question: to what extent are the phase transition analysis applicable if the response variables include small amount of errors in practice, i.e. the variance σ_w^2 of the error $\mathbf{w}$ is assumed to be small. In the next section, we will clarify this surprising phenomenon by performing high order analysis in terms of σ_w.

3.2   Noisy Settings

In this section, we assume that $\sigma_w^2 > 0$. This implies that zero is no longer a fixed point of $\psi_q(\tau^2, \alpha_\star(\tau))$ defined in (9) and the reconstruction error of LQLS is greater than zero for all values of q. Denote τ_l the lowest fixed point of $\psi_q(\tau^2, \alpha_\star(\tau))$. Our first result is concerned with the first dominant term for small amount of noise.

Proposition 3 *If $\epsilon < \delta$, then there exists τ_s^2 such that for every $\sigma_w^2 < \tau_s^2$, $\lim_{\sigma_w^2 \to 0} \frac{\tau_l^2}{\sigma_w^2} = \frac{\delta}{\delta - \epsilon}$.*

As discussed in Section 3.1, the first dominant term determines the phase transition which is the same for all q and Σ. In order to see the discrepancy between different values of q and different structures of Σ, we have to do high-order analysis and explore how $\tau_l^2 - \frac{\delta \sigma_w^2}{\delta - \epsilon}$ behaves for small values of σ_w^2. We consider the simplest block diagonal matrix Σ with two-dimensional block $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where $0 \le \rho < 1$. For this choice of the covariance matrix, it can be easily verified that the function (3) defined in Condition 5 has a differentiable limit.

Consider $p_{\boldsymbol{\beta}_0} \in \mathcal{F}_\epsilon$. Let $p_{\beta_0} = (1 - \epsilon)\delta_0 + \epsilon G$ and U is a random variable with distribution G, where δ_0 denotes a point mass at zero and G denotes the distribution of nonzero elements of $\boldsymbol{\beta}_0$. Let us discuss the result for $q = 0$ and $0 < q < 1$ separately in the following two propositions.

Proposition 4 *Suppose $E|U|^2 < \infty$ and $P(|U| > \mu) = 1$, where $\mu = \sup_v \{v : P(|U| > v) = 1\} > 0$. Then for $q = 0$ and $\epsilon < \delta$,*

$$\lim_{\sigma_w \to 0} \tau_l^2 = \frac{\delta}{\delta - \epsilon} \sigma_w^2 + o\left(\phi\left(\tilde{\mu} \sqrt{\frac{\delta - \epsilon}{\delta}} \sigma_w^{-1} \right) \right), \tag{10}$$

where ϕ denotes the standard normal density function and $\tilde{\mu}$ is any constant that is smaller than $(2 - \rho - \sqrt{1 - \rho^2})\sqrt{1 - \rho^2}\mu/2$.

Proposition 5 *Suppose $E|U|^2 < \infty$ and $P(|U| > \mu) = 1$, where $\mu = \sup_v \{v : P(|U| > v) = 1\} > 0$. Then for $0 < q < 1$ and $\epsilon < \delta$,*

$$\lim_{\sigma_w \to 0} \frac{\tau_l^2 - \frac{\delta}{\delta - \epsilon} \sigma_w^2}{\sigma_w^{4 - 2q} (\log \frac{1}{\sigma_w})^{2 - q}} = \frac{\epsilon C_q^{4 - 2q} q^2 (\epsilon D_0 + (1 - \epsilon)C_0)\delta^{2 - q}}{(4 - 4q)^{2 - q}(\delta - \epsilon)^{3 - q}}, \tag{11}$$

8

where $C_q = [2(1-q)]^{1/(2-q)} + q[2(1-q)]^{(q-1)/(2-q)}$, $D_0 = [E\{|U|^{2q-2}\} - \rho(E\{|U|^{q-1}sign(U)\})^2]/(1-\rho^2)$, *and* $C_0 = E\{|U|^{2q-2}\}$.

It can be easily verified that when $\rho = 0$, results from Proposition 4 and Proposition 5 are the same as the results from Theorem 8 and Theorem 9 in Zheng et al. (2017). Comparing (10) and (11), we conclude that the second dominant term for $q = 0$ decays exponentially faster than the polynomial rate for $0 < q < 1$ in low noise regime. Since the second dominant term in (11) is positive, optimally tuned L_0 regularization will outperform optimally tuned L_q regularization for $0 < q < 1$ in this regime. Another interesting feature of Proposition 5 is that the second dominant term is proportional to σ_w^{4-2q}, thus if $q_1 < q_2$, $LQLS$ for q_1 outperforms $LQLS$ for q_2 for small enough σ_w^2. Moreover, the second dominant term increases with ϵ and decreases with δ. Examining the impact of ρ based on (11), we conclude that if the distribution of U is symmetric about zero, the second dominant term increases with ρ since D_0 is proportional to $1/(1-\rho^2)$ and $E\{|U|^{q-1}sign(U)\} = 0$. All these observations are consistent with the numerical studies shown in Section 5.

4 Asymptotic Performance for $1 < q \le 2$

The difference between the regularization for $1 < q \le 2$ and the regularization for $0 \le q < 1$ is that the former is convex and its optimization problem has a unique global minimum. However, since its penalty is differentiable everywhere, it never leads to a coefficient been zero rather only minimizes it when $\lambda \ne 0$. Our first result is concerned with the first dominant term of the optimally tuned asymptotic prediction risk defined as $R_q(\tau_\star, \sigma_w) = \min_\lambda \frac{1}{p}\|\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0\|_{\boldsymbol{\Sigma}}^2$ which is related to the phase transition curve for successful recovery.

Proposition 6 *For $\boldsymbol{\beta}_0 \in \mathcal{F}_\epsilon$, suppose $E|U|^2 \le \infty$ and $P(|U| > \mu) = 1$, where $\mu = \sup_v \{v : P(|U| > v) = 1\} > 0$. Then for $1 < q \le 2$, we have*

$$\lim_{\sigma_w^2 \to 0} R_q(\tau_\star, \sigma_w) \begin{cases} > 0 & if \quad \delta < 1 \\ \to 0 & if \quad \delta > 1 \end{cases}.$$

Therefore, assume the error σ_w^2 equals zero, for $1 < q \le 2$ and any $\gamma > 0$, if $\delta > 1 + \gamma$, then (2) succeeds in recovering $\boldsymbol{\beta}$, while if $\delta < 1 - \gamma$, (2) fails. The phase transition curve $\delta = 1$ is independent of ϵ, q, and the covariance structure of $\boldsymbol{\Sigma}$. As pointed out by Zheng et al. (2017), phase transition analysis based on the first dominant term may lead to misleading conclusions in any practical setting where there is always error in the response variables. To study the impact of them, we need to conduct higher-order analysis of LQLS in the small-error regime in the expansion of prediction risk $R_q(\tau_\star, \sigma_w)$.

Our second result is concerned with the second dominant term of the optimally tuned MSE of LQLS when the number of response variables is larger than the number of predictors p, i.e. $\delta > 1$. The key is to characterize the convergence rate for $R_q(\tau_\star, \sigma_w)$ as $\sigma_w \to 0$.

Proposition 7 *Denote $Z \sim N(0,1)$ and U a random variable whose distribution is specified by the nonzero elements of β_0. Suppose $E|U|^2 < \infty$ and $P(|U| > \mu) = 1$, where $\mu = \sup_v \{v : P(|U| > v) = 1\} > 0$. Then for $1 < q \le 2$ and $\delta > 1$, we have*

$$
\begin{aligned}
&R_q(\tau_\star, \sigma_w) \\
&= \frac{\delta \sigma_w^2}{\delta - 1} - \frac{\delta^{q+1}}{(\delta-1)^{q+1}} \frac{(1-\epsilon)^2}{\epsilon} \frac{[E(|Z|^q)]^2 \sigma_w^{2q}}{(1-\rho^2)^{q-1} \{ E(|U|^{2q-2}) - \epsilon \rho [E(|U|^{q-1} sign(U))]^2 \}}.
\end{aligned}
\tag{12}
$$

The first term $\frac{\delta \sigma_w^2}{\delta - 1}$ shows a notion of phase transition. For as $\sigma_w \to 0$, $R_q(\tau_\star, \sigma_w) = O(\sigma_w^2)$, and will go to zero. For $\rho = 0$, the second dominant term is the same as the one derived in Theorem 3.1 in Weng et al. (2018). The important facts of the second dominant term include: (1) it is negative; (2) its order is σ_w^{2q}; (3) its magnitude decreases with ϵ for fixed q and ρ; (4) its magnitude decreases with q for fixed ϵ and ρ; (5) its magnitude increases with ρ for fixed q and ϵ. Facts (3) and (4) have been studied thoroughly in Weng et al. (2018). Our numerical studies in Section 5 focus on the verification of fact (5).

5 Numerical results

This section performs several numerical studies to evaluate the $R_q(\tau_\star, \sigma_w)$ of LQLS as a function of σ_w for several different values of q and ρ. Here we consider a block diagonal matrix $\boldsymbol{\Sigma}$ with two-dimensional block $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The equation can be decomposed into 2-dimensional blocks. Sections 5.1 and 5.2 study the performance of LQLS with $0 \le q < 1$ and $1 < q \le 2$ respectively.

5.1 $0 \le q < 1$

We first study the phase transition result for $q = 0$, i.e. $\sigma_w = 0$, in Figure 2 to identify conditions under which $\psi_q(\tau^2, \alpha_\star(\tau))$ has additional stable fixed points than $\tau^2 = 0$. The following quantity plays an important role in our analysis

$$
M_0(\epsilon, \rho) = \sup_{\mu_1, \mu_2} \inf_\alpha \left\{ \frac{\epsilon^2}{2} E\|\hat{\beta}_1 - \boldsymbol{\beta}_0^1\|_{\boldsymbol{\Sigma}}^2 + \epsilon(1-\epsilon) E\|\hat{\beta}_2 - \boldsymbol{\beta}_0^2\|_{\boldsymbol{\Sigma}}^2 + \frac{(1-\epsilon)^2}{2} E\|\hat{\beta}_3\|_{\boldsymbol{\Sigma}}^2 \right\},
\tag{13}
$$

where

$$
\hat{\beta}_1 = \boldsymbol{\eta}_0(\boldsymbol{\beta}_0^1 + \boldsymbol{\Sigma}^{-1/2}\mathbf{z}, \alpha), \ \ \hat{\beta}_2 = \boldsymbol{\eta}_0(\boldsymbol{\beta}_0^2 + \boldsymbol{\Sigma}^{-1/2}\mathbf{z}, \alpha), \ \ \hat{\beta}_3 = \boldsymbol{\eta}_0(\boldsymbol{\Sigma}^{-1/2}\mathbf{z}, \alpha),
$$

where

$$
\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \ \boldsymbol{\beta}_0^1 = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \boldsymbol{\beta}_0^2 = \begin{pmatrix} \mu_1 \\ 0 \end{pmatrix}.
$$

Here we consider 3 different situations according to the nonzero components of 2-dimensional vector $\boldsymbol{\beta}_0$. The first term in (13) is for situation that both components are nonzero. The

10

second term is for situation that one is nonzero and the other is zero. The third term is for situation that both are zero. From (8) we obtain that if $\delta > M_0(\epsilon, \rho)$, $\psi_0(\tau^2, \lambda_\star(\tau))$ has a unique stable fixed point at zero. On the other hand, if $\delta < M_0(\epsilon, \rho)$, there exists $P_{\beta_0} \in \mathcal{F}_\epsilon$ for which $\psi_0(\tau^2, \lambda_\star(\tau))$ has more than one stable fixed point. Figure 2 compares the phase transition curves for different values of ρ. For larger ρ, we need larger δ to achieve unique fixed point for $\psi_0(\tau^2, \lambda_\star(\tau))$. Figures 3 and 4 exhibit the lowest fixed point τ_l^2 of $\psi_0(\tau^2, \lambda_\star(\tau))$ as a function of σ_w^2 for different values of ρ. Figure 3 is for the case of $\delta > M_0(\epsilon, \rho)$ while Figure 4 is for the case of $\delta < M_0(\epsilon, \rho)$. Note that the performance of L_0 regularized method for smaller ρ is better than its performance for larger ρ. Moreover, τ_l^2 is a continuous function of σ_w^2 if $\psi_0(\tau^2, \lambda_\star(\tau))$ has a unique stable fixed point as shown in Figure 3. On the other hand, if $\psi_0(\tau^2, \lambda_\star(\tau))$ has more than one stable fixed points, the function is discontinuous as shown in Figure 4.

Figure 2: Comparison of the phase transition of the L_0 regularized method. The phase transition exhibited is the value of δ at which the number of stable fixed points changes from one to more than one for at least some prior $P_{\beta_0} \in \mathcal{F}$.

Figure 5 illustrates the dependence of the risk function on the correlation coefficient ρ for $0 < q < 1$ under different choices of q. It is shown clearly that the risk increases with ρ for fixed q which is consistent with the analytical result (11) derived in Proposition 5 based on high order analysis. Moreover, the magnitude of changes due to ρ is larger for small q than that for large q. In particular, as q close to 1, the curve for $\rho = 0$ and the curve for $\rho = 0.9$ in the right panel of Figure 5 are almost indistinguishable. This is consistent with the conclusion drawn in Theorems 3.3 of Weng et al. (2018) which states that the first order term is sufficient to describe the performance of LASSO (q=1) in the small-error regime because the second dominant term is exponentially small. Figure 6 depicts the impact of q on risk for fixed ρ. It is shown that
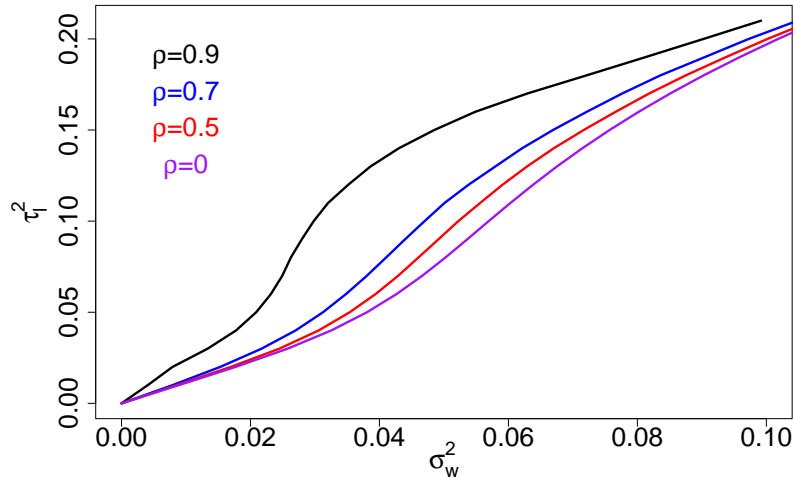
Figure 3: The curve of τ_l^2 as a function of σ_w^2 for $\rho \in [0, 0.5.0.7.0.9]$. Here $\delta = 0.9$, $\epsilon = 0.1$, and the non-zero elements of β_0 are i.i.d. ± 1 with probability 0.5.
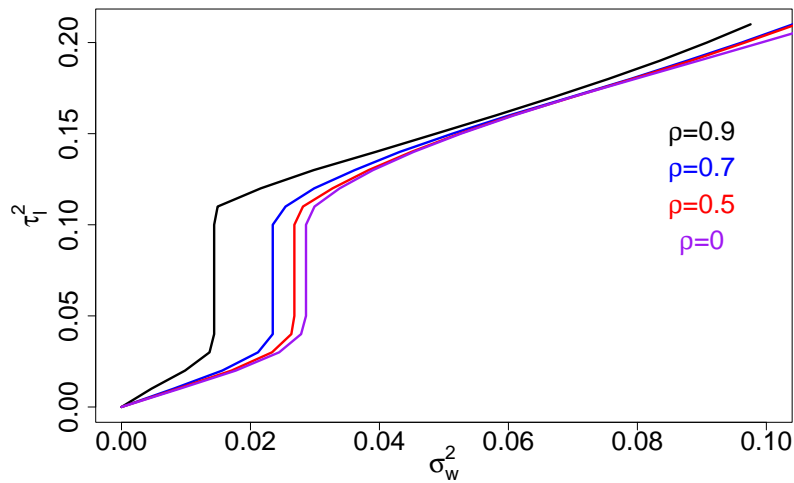

Figure 4: The curve of τ_l^2 as a function of σ_w^2 for $\rho \in [0, 0.5.0.7.0.9]$. Here $\delta = 0.1$, $\epsilon = 0.01$, and the non-zero elements of β_0 are i.i.d. ± 1 with probability 0.5.

$LQLS$ for q_1 outperforms $LQLS$ for q_2 when $q_1 < q_2$. This is consistent with the conclusion in Weng et al. (2018) and also confirms our observation in Proposition 5. Moreover, it is shown from Figure 6 that the discrepancies caused by different values of q decreases with ρ.

Figure 5: The curve of MSE as a function of σ_w^2 for $\rho \in [0, 0.9]$ under three different values of q. Here $\delta = 0.9$, $\epsilon = 0.1$, and the non-zero elements of β_0 are i.i.d. ± 3 with probability 0.5.

5.2 $1 < q \leq 2$

In this section, we check the approximation accuracy of the first and second order expansions of the risk over a reasonable range of σ_w in the case $1 < q \leq 2$. The dependence of risk on parameters δ, ϵ, and q has been studied extensively in Weng et al. (2018). Here we focus on the dependence of risk on the correlation coefficient ρ. Throughout this section, we set the distribution of U to $f(b) = 0.5\delta_1(b) + 0.5\delta_{-1}(b)$.

Figures 7 and 8 compare the true value, first order approximation, and second order approximation of risk under different choices of ρ for $q = 1.5$ and $q = 1.9$ respectively. Our numerical results in Figures 7 and 8 show that both the first order and second order expansions present a good approximation when σ_w is small enough. As we increase σ_w, both first order and second order expansions are larger than the true values and the second order approximation is more accurate than the first order approximation. The discrepancies between the approximation and the true values increase with ρ as demonstrated in these Figures. Therefore, the expansion approximation is less accurate under large correlation than under small correlation. Consistence with the analytical form of the second-order term in (12), we conclude from the numerical
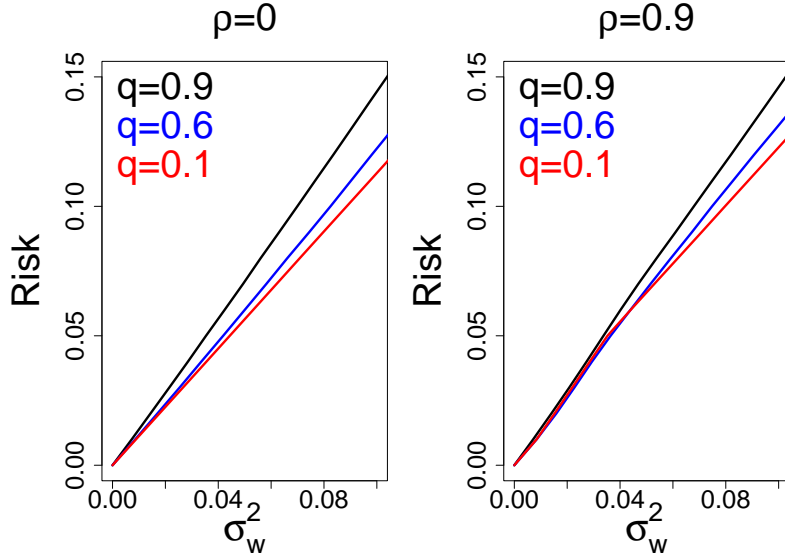
Figure 6: The curve of MSE as a function of σ_w^2 for $q \in [0.1, 0.6, 0.9]$ under two different values of ρ. Here $\delta = 0.9$, $\epsilon = 0.1$, and the non-zero elements of β_0 are i.i.d. ± 3 with probability 0.5.

studies that the second order approximation outperforms the first order, however, it may not be sufficient if the following conditions hold: (i) δ is close to 1; (ii) ϵ is small, (iii) q is close to 1, (iv) ρ is close to 1.

Figure 9 exhibits the impact of ρ on the true value of risk for different values of q. As is clear in this figure, the risk decreases with both ρ and q which is in agreement with the second order approximation in (12).

6   Discussion

L_q-regularized least square is one of the most popular schemes for recovering a high-dimensional sparse vector from low-dimensional measurements. This paper focuses on asymptotic behavior of LQLS estimators in the framework in which it is assumed that $p, n, k \to \infty$ while $n/p \to \delta$ and $k/p \to \epsilon$ are fixed. We first derive the distributional limit of LQLS estimators for nonstandard Gaussian design models where the row of design matrix $\mathbf{X}$ are drawn independently from distribution $N(0, \boldsymbol{\Sigma})$ with arbitrary $\boldsymbol{\Sigma}$. Then we obtain an explicit characterization of the asymptotic risk by deriving its higher order expansion in the small-error regime. Our analysis is performed on both the convex case $1 < q \le 2$ and non-convex case $0 \le q < 1$. We conclude that the first order term does not depend on the covariance structure of $\boldsymbol{\Sigma}$ and thus the phase transition curves are the same as the in the case of standard Gaussian design. This is different from the case $q = 1$ in which the correlation can change the phase transition boundary if the
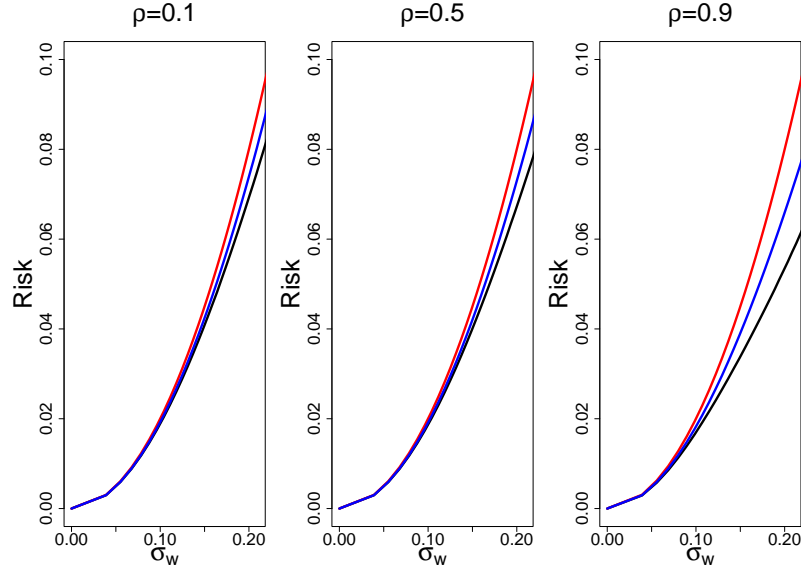
Figure 7: Plots of actual risk and its approximation for $q = 1.5, \delta = 2$, and $\epsilon = 0.7$. The true value, first order approximation, and second order approximation are denoted by black, red, and blue curves respectively.
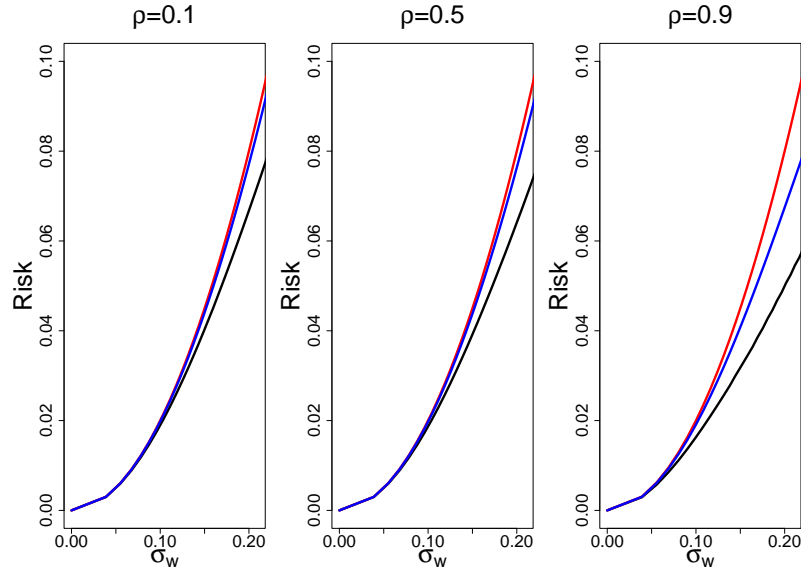

Figure 8: Plots of actual risk and its approximation for $q = 1.9, \delta = 2$, and $\epsilon = 0.7$. The true value, first order approximation, and second order approximation are denoted by black, red, and blue curves respectively.
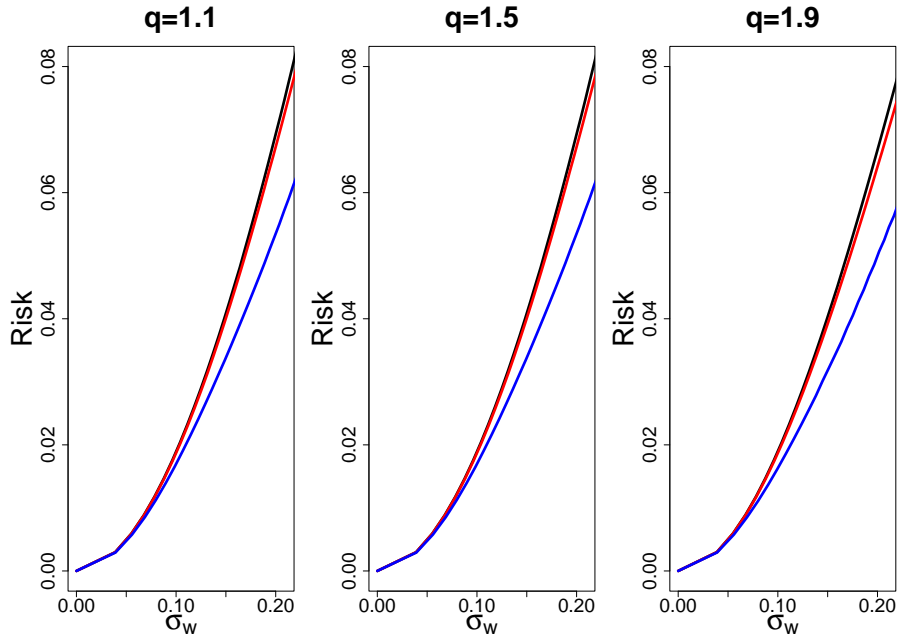
Figure 9: Plots of actual risk for $q = 1.5, \delta = 2$, and $\epsilon = 0.7$. The results for $\rho = 0.1$, $\rho = 0.5$, and $\rho = 0.9$ are denoted by black, red, and blue curves respectively.

signed pattern of signal is not symmetric. The second order term depends on the correlation and the explicit formulas are derived in both cases based on a simple two-dimensional block diagonal covariance model.

Note that part of our analysis is based on the replica method which are not fully rigorous yet. So far the rigorous work in this area mainly focuses on i.i.d. randomness. For example, the mean square error of LQLS for the case of $1 \leq q \leq 2$ is characterized using Gordon's lemma in Thrampoulidis et al. (2018) and AMP in Weng et al. (2018). For nonstandard Gaussian design models, $\Sigma \neq I_{p \times p}$ the rigorous analysis for the case $q = 1$ have been conducted recently Celentano et al. (2020) using Gordon's comparison inequality and in Huang (2021) using AMP. Our next step is to derive the rigorous results for Proposition 1 for the case $1 < q \leq 2$. The case $0 < q \leq 1$ is more challenging because the penalty term is not convex and even in the i.i.d. case its rigorous results have not been established yet.

Our conclusion in this paper is based on higher-order analysis in small error region. It may not be applicable if the noisy error is large enough. As shown in Zheng et al. (2017), for large values of noise, the LASSO outperforms L_0-minimization and if $q_1 > q_2$, then optimal L_{q_1}-minimization outperforms optimal L_{q_2}-minimization. Proposition 3 of Zheng et al. (2017) provides theoretical justification on this high-noise phenomenon. Another direction of our future research is to study the high-noise phenomenon of LQLS for nonstandard Gaussian

16

design models.

7 Appendix

This appendix outlines the proofs of the Propositions in the main text.

7.1 Proof of Proposition 1

Proof 1 *We limit ourselves to the main steps of replica calculations leading to Propositions 1. For a general introduction to the method and its motivation, we refer to Mezard et al. (1987); Mézard and Montanari (2009).*

We consider L_q-regularized least squares estimators of the form [1]

$$\hat{\boldsymbol{\theta}} = argmin_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_q^q \right\}. \tag{A1}$$

Define $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ a continuous function strictly convex in its first argument. Then its Lagrange dual $\tilde{g}(x,y) \equiv max_{\mu \in \mathbb{R}} \{\mu x - g(\mu, y)\}$ is also a continuous function convex in its first argument. We start from estimating the following moment generating function also called partition function

$$Z_p(\beta, s) = \int \exp \left\{ -\frac{\beta}{2n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 - \beta \lambda \|\boldsymbol{\theta}\|_q^q - \beta s \sum_{j=1}^p [g(\mu_j, \beta_{0,j}) - \mu_j \theta_j] \right\} d\boldsymbol{\theta} d\boldsymbol{\mu}, \tag{A2}$$

where $s>0$, $(y_i, \mathbf{x}_i)$ are i.i.d. pairs distributed as model (2) in the main text, and $\beta>0$ is the inverse temperature parameter. The free energy is the low temperature limit

$$\mathcal{F}(s) = -\lim_{p \to \infty} \lim_{\beta \to \infty} \frac{1}{p\beta} \log Z_p(\beta, s), \tag{A3}$$

which is assumed to converge to

$$\mathcal{F}(s) = -\lim_{p \to \infty} \lim_{\beta \to \infty} \frac{1}{p\beta} E \log Z_p(\beta, s), \tag{A4}$$

where the expectation is with respect to the distribution of $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$. Using Laplace method in the integral (A2), we have

$$\mathcal{F}(s) = \lim_{p \to \infty} \frac{1}{p} \min_{\boldsymbol{\theta}, \boldsymbol{\mu} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 + \lambda \|\boldsymbol{\theta}\|_q^q + s \sum_{j=1}^p [g(\mu_j, \beta_{0,j}) - \mu_j \theta_j] \right\}. \tag{A5}$$

---

[1]There is a subtle discrepancy between this definition and (2) in the main text, i.e. $\mathbf{X}$ and $\mathbf{y}$ are scaled by $1/\sqrt{n}$.

Taking the derivative of $\mathcal{F}(s)$ as $s \to 0$, we have

$$\frac{d\mathcal{F}}{ds}(s=0) = \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \min_{\mu_i \in \mathbb{R}} [g(\mu_j, \beta_{0,j}) - \mu_j \hat{\theta}_j] = -\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \tilde{g}(\hat{\theta}_j, \beta_{0,j}), \tag{A6}$$

where $\hat{\boldsymbol{\theta}}$ is the solution of (A1).

Now we compute $E \log Z_p(\beta, s)$ using the replica method. Instead of performing integration over the log-term directly, we use the following identity

$$E \log Z_p(\beta, s) = \left. \frac{\partial E\{Z_p(\beta, s)^d\}}{\partial d} \right|_{d=0}. \tag{A7}$$

We first compute the expectation for integer d, and then extrapolate it as $d \to 0$.

Define the measure $\nu(d\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \mathbb{R}^p$ as follows

$$\nu(d\boldsymbol{\theta}) = \int \exp\left\{ -\beta\lambda\|\boldsymbol{\theta}\|_q^q - \beta s \sum_{j=1}^{p} [g(\mu_j, \theta_{0,j}) - \mu_j \theta_j] \right\} d\boldsymbol{\mu} d\boldsymbol{\theta}, \tag{A8}$$

where the integration is with respect to $d\boldsymbol{\mu}$. Then for integer d, define $\nu^d(d\boldsymbol{\theta}) = \nu(d\boldsymbol{\theta}^1) \cdots \nu(d\boldsymbol{\theta}^d)$ be a measure over $(\mathbb{R}^p)^d$, with $\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^d \in \mathbb{R}^p$. With these notations and substituting (1.1), we have

$$E\{Z_p(\beta, s)^d\} = \int E \exp\left\{ -\frac{\beta}{2n} \sum_{a=1}^{d} \sum_{i=1}^{n} \left[\mathbf{x}_i^T(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a) + w_i \right]^2 \right\} \nu^d(d\boldsymbol{\theta}),$$

where the expectation is with respect to $\mathbf{w}$ and $\mathbf{x}_i$. Using identity

$$\int \delta(u_i^a - \mathbf{x}_i^T(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)) du_i^a = 1, \tag{A9}$$

and Fourier transformation

$$\delta(u_i^a - \mathbf{x}_i^T(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)) = \frac{1}{2\pi i} \int \exp\{i\hat{u}_i^a(u_i^a - \mathbf{x}_i^T(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a))\} d\hat{u}_i^a, \tag{A10}$$

where $\delta(\cdot)$ is the Dirac delta function, we obtain

$$E\{Z_p(\beta, s)^d\} = \frac{1}{(2\pi i)^{nd}} \left(\frac{\beta}{n}\right)^{\frac{nd}{2}} \int E \exp\left\{ -\frac{\beta}{2n} \sum_{a,i} (u_i^a + w_i)^2 \right.$$

$$\left. + i\sqrt{\frac{\beta}{n}} \sum_{a,i} \hat{u}_i^a(u_i^a - \mathbf{x}_i^T(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)) \right\} \nu^d(d\boldsymbol{\theta})\nu^d(d\mathbf{u})\nu^d(d\hat{\mathbf{u}}), \tag{A11}$$

where $\nu^d(d\mathbf{u}) = \nu(d\mathbf{u}^1) \cdots \nu(d\mathbf{u}^d)$ is a measure over $(\mathbb{R}^n)^d$ with $\mathbf{u}^1, \cdots, \mathbf{u}^d \in \mathbb{R}^n$ and $\nu^d(d\hat{\mathbf{u}})$ is defined similarly. For fixed $\sum_{a=1}^{d}(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)$, the product term $\mathbf{x}_i^T \sum_{a=1}^{d}(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)$ follows a

multivariate normal distribution with mean zero and covariance matrix $\sum_{a,b}(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)^T \boldsymbol{\Sigma}(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^b)$. After integration over $\mathbf{x}_1, \cdots, \mathbf{x}_n$, we have

$$
\begin{aligned}
E\{Z_p(\beta,s)^d\} &= \frac{1}{(2\pi i)^{nd}}\left(\frac{\beta}{n}\right)^{\frac{nd}{2}} \int E \exp\left\{-\frac{\beta}{2n}\sum_{a,i}(u_i^a + w_i)^2 + i\sqrt{\frac{\beta}{n}}\sum_{a,i}\hat{u}_i^a u_i^a \right. \\
&\left. -\frac{\beta}{2n}\sum_{a,b}(\hat{\mathbf{u}}^a)^T\hat{\mathbf{u}}^b(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)^T\boldsymbol{\Sigma}(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^b)\right\}\nu^d(d\boldsymbol{\theta})\nu^d(d\mathbf{u})\nu^d(d\hat{\mathbf{u}}). \quad \text{(A12)}
\end{aligned}
$$

We next follow the similar procedure as in Javanmard and Montanari (2014) and use the identity

$$
\exp(-xy) = \frac{1}{2\pi i}\int_{(-i\infty,i\infty)}\int_{(-\infty,\infty)}\exp(-\zeta\hat{\zeta} + \zeta x - \hat{\zeta}y)d\zeta d\hat{\zeta} \quad \text{(A13)}
$$

to (A12) and introduce integration variables $\mathbf{Q} = (Q_{ab})_{1\le a,b\le d}$ and $\boldsymbol{\Lambda} = (\Lambda_{ab})_{1\le a,b\le d}$. Letting $d\mathbf{Q} = \prod_{a,b}dQ_{ab}$ and $d\boldsymbol{\Lambda} = \prod_{a,b}d\Lambda_{ab}$, we have

$$
E\{Z_p(\beta,s)^d\} = \left(\frac{\beta n}{4\pi i}\right)^{d^2}\int\exp\{-p\mathcal{S}_d(\mathbf{Q},\boldsymbol{\Lambda})\}d\mathbf{Q}d\boldsymbol{\Lambda}, \quad \text{(A14)}
$$

$$
\mathcal{S}_d(\mathbf{Q},\boldsymbol{\Lambda}) = \frac{\beta\delta}{2}\sum_{a,b}\Lambda_{ab}Q_{ab} - \frac{1}{p}\log\xi(\boldsymbol{\Lambda}) - \delta\log\hat{\xi}(\mathbf{Q}), \quad \text{(A15)}
$$

$$
\xi(\boldsymbol{\Lambda}) = \int\exp\left\{\frac{\beta}{2}\sum_{a,b}\Lambda_{ab}(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^a)^T\boldsymbol{\Sigma}(\boldsymbol{\beta}_0 - \boldsymbol{\theta}^b)\right\}\nu^d(d\boldsymbol{\theta}),
$$

$$
\hat{\xi}(\mathbf{Q}) = \frac{1}{(2\pi i)^d}E\exp\left\{\frac{\beta}{2n}w^2\sum_{a,b}(\mathbf{I}_{d\times d} + (\beta\mathbf{Q})^{-1})_{ab}^{-1} - \frac{1}{2}\log det(\mathbf{I}_{d\times d} + \beta\mathbf{Q}) - \frac{\beta}{2n}dw^2\right\},
$$

where in $\hat{\xi}(\mathbf{Q})$ we have performed integrations over $d\mathbf{u}$ and $d\hat{\mathbf{u}}$ and the remaining expectation is over the noise variable w. We next use the saddle point method in (A14) to obtain

$$
-\lim_{p\to\infty}\frac{1}{p}\log E\{Z_p(\beta,s)^d\} = \mathcal{S}_d(\mathbf{Q}^\star, \boldsymbol{\Lambda}^\star), \quad \text{(A16)}
$$

where $\mathbf{Q}^\star, \boldsymbol{\Lambda}^\star$ is the saddle-point location. From replica symmetry, $\mathbf{Q}^\star, \boldsymbol{\Lambda}^\star$ are invariant under permutations of the row/column indices, which is equivalent to

$$
Q_{ab}^\star = \begin{cases} q_1 & if\ a = b, \\ q_0 & otherwise \end{cases}, \qquad \Lambda_{ab}^\star = \begin{cases} \beta\zeta_1 & if\ a = b, \\ \beta\zeta_0 & otherwise \end{cases}, \quad \text{(A17)}
$$

The next step is to substitute the above expressions for $\mathbf{Q}^\star$ and $\boldsymbol{\Lambda}^\star$ in (A15) and then taking

the limit $d \to 0$. Hence

$$\lim_{d \to 0} \frac{\beta\delta}{2d} \sum_{a,b} \Lambda_{ab}^\star Q_{ab}^\star = \frac{\beta^2\delta}{2}(\zeta_1 q_1 - \zeta_0 q_0)$$

$$\lim_{d \to 0} \frac{-\delta \log \hat{\xi}(\mathbf{Q}^\star)}{d} = \frac{\delta}{2}\log(1 + \beta(q_1 - q_0)) + \frac{\delta}{2}\frac{\beta(q_0 + \sigma_w^2)}{1 + \beta(q_1 - q_0)}$$

$$\lim_{d \to 0} \frac{-\log \xi(\mathbf{\Lambda}^\star)}{d} = E\left\{\log\left[\int \exp\left\{\frac{\beta^2}{2}(\zeta_1 - \zeta_0)\|\boldsymbol{\theta} - \boldsymbol{\beta}_0\|_{\mathbf{\Sigma}}^2\right.\right.\right.$$
$$\left.\left.\left. + \beta\sqrt{\zeta_0}\mathbf{z}^T\mathbf{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\beta}_0)\right\}\nu(d\boldsymbol{\theta})\right]\right\},$$

where expectation is with respect to $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$. In order for the exponent in above equation to be extensive in p, we introduce $q_1 - q_0 = q/\beta$ and $\zeta_1 - \zeta_0 = -\zeta/\beta$. We find in the asymptotic limit $\beta \to \infty$ the free energy becomes

$$\mathcal{F}(s) = -\lim_{\beta \to \infty}\lim_{p \to \infty}\frac{1}{p\beta}E\log Z_p(\beta, s) = \lim_{\beta \to \infty}\lim_{d \to 0}\frac{1}{d\beta}\mathcal{S}_d(\mathbf{Q}^\star, \mathbf{\Lambda}^\star)$$

$$= \frac{\delta}{2}(\zeta_0 q - \zeta q_0) + \frac{\delta}{2}\frac{q_0 + \sigma_w^2}{1 + q}$$

$$+ \lim_{p \to \infty}\frac{1}{p}E\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\|\boldsymbol{\theta} - \boldsymbol{\beta}_0\|_{\mathbf{\Sigma}}^2 - \sqrt{\zeta_0}\mathbf{z}^T\mathbf{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\beta}_0) + D(\boldsymbol{\theta}; s)\right\}, \quad \text{(A18)}$$

where

$$D(\boldsymbol{\theta}; s) = \min_{\boldsymbol{\mu}\in\mathbb{R}^p}\left\{\lambda\|\boldsymbol{\theta}\|_q^q - s\boldsymbol{\theta}^T\boldsymbol{\mu} + s\sum_{j=1}^p g(\mu_j, \theta_{0,j})\right\}. \quad \text{(A19)}$$

After eliminating q and q_0 using their saddle-point equations and renaming $\zeta_0 = \zeta^2\tau^2$, we have

$$\mathcal{F}(s) = -\frac{1}{2}(1 - \delta)\zeta\tau^2 - \frac{\delta}{2}\zeta^2\tau^2 + \frac{\delta}{2}\sigma_0^2\zeta$$

$$+ \lim_{p \to \infty}\frac{1}{p}E\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\|\boldsymbol{\theta} - \boldsymbol{\beta}_0 - \tau\mathbf{\Sigma}^{-1/2}\mathbf{z}\|_{\mathbf{\Sigma}}^2 + D(\boldsymbol{\theta}, s)\right\}. \quad \text{(A20)}$$

In the case $s = 0$, solving the saddle-point equations for ζ and τ^2, we finally get

$$\tau^2 = \sigma_w^2 + \frac{1}{\delta}\lim_{p \to \infty}\frac{1}{p\delta}E\left(\|\boldsymbol{\eta}_q(\boldsymbol{\beta}_0 + \tau\mathbf{\Sigma}^{-1/2}\mathbf{z}, \lambda/\zeta) - \boldsymbol{\beta}_0\|_{\mathbf{\Sigma}}^2\right),$$

$$\zeta = 1 - \frac{1}{\delta}E\{\langle\boldsymbol{\eta}_q'(\boldsymbol{\beta}_0 + \tau\mathbf{\Sigma}^{-1/2}\mathbf{z}, \lambda/\zeta)\rangle\}, \quad \text{(A21)}$$

where $\boldsymbol{\eta}_q(\mathbf{v}, \lambda)$ is defined in (2.2). Then taking the derivative of (A20) as $s \to 0$ and minimizing over $\boldsymbol{\mu}$, we get

$$\frac{d\mathcal{F}}{ds}(s = 0) = -\lim_{p \to \infty}\frac{1}{p}\sum_{j=1}^p E\tilde{g}(\tilde{\theta}_j, \beta_{0,j}), \quad \text{(A22)}$$

where $\tilde{\boldsymbol{\theta}} = \eta_q(\boldsymbol{\beta}_0 + \tau\mathbf{\Sigma}^{-1/2}\mathbf{z}, \lambda/\zeta)$. Comparing (A22) with (A6) for a complete set of functions $\tilde{g}$ and using the standard weak convergence arguments, we prove the claim that the distributional limit of LQLS estimator does indeed hold.

20

7.2 Proof of Proposition 2

Proof 2 *Define $\lambda = \alpha_0 \tau^{2-q}$. This will enable us to employ the scale invariance properties of the general thresholding operation (4) more efficiently. We can write (8) as*

$$\psi_q(\tau^2, \alpha_0\tau^{2-q}) \;\; = \;\; \frac{1}{p\delta} E\|\boldsymbol{\eta}_q(\boldsymbol{\beta}_0 + \tau\boldsymbol{\Sigma}^{-1/2}\mathbf{z}, \alpha_0\tau^{2-q}) - \boldsymbol{\beta}_0\|_\Sigma^2 \qquad (A23)$$

Note that $\tau^2 = 0$ is actually a fixed point of $\psi_q(\tau^2, \alpha_0\tau^{2-q})$. Furthermore, it is straightforward to see that 0 is a stable fixed point if and only if

$$\frac{d\psi_q(\tau^2, \alpha_0\tau^{2-q})}{d\tau^2}\bigg|_{\tau^2=0} = \lim_{\tau^2 \to 0} \frac{\psi_q(\tau^2, \alpha_0\tau^{2-q})}{\tau^2} < 1$$

Denote $U = \boldsymbol{\beta}_0/\tau$. Then (A23) can be written as

$$\psi_q(\tau^2, \alpha_0\tau^{2-q}) \;\; = \;\; \frac{\tau^2}{\delta} R_q(\tau^2, \alpha_0),$$

where

$$R_q(\tau^2, \alpha_0) \;\; = \;\; \frac{1}{p}E\|\hat{\boldsymbol{\beta}} - U\|_\Sigma^2, \qquad (A24)$$

where

$$\hat{\boldsymbol{\beta}} \;\; = \;\; argmin_{\boldsymbol{\beta}}\left\{\frac{1}{2}\|\boldsymbol{\beta} - \mathbf{U} - \Sigma^{-1/2}\mathbf{z}\|_\Sigma^2 + \alpha_0\|\boldsymbol{\beta}\|_q^q\right\}. \qquad (A25)$$

For fixed U, denote $S = \{j|U_j \neq 0\}$. As $\tau \to 0$, $\hat{\boldsymbol{\beta}}_S = U_s + o_P(\mu/\tau)$, we obtain

$$\boldsymbol{\Sigma}_{SS}(\hat{\boldsymbol{\beta}}_S - \mathbf{U}_S) + \boldsymbol{\Sigma}_{S\bar{S}}\hat{\boldsymbol{\beta}}_{\bar{S}} - (\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S + q\alpha_0|\hat{\boldsymbol{\beta}}_S|^{q-1}sign(\hat{\boldsymbol{\beta}}_S) = 0,$$

which can be written as

$$\hat{\boldsymbol{\beta}}_S - \mathbf{U}_S = (\boldsymbol{\Sigma}^{-1/2}\mathbf{z})_S - \boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{S\bar{S}}\hat{\boldsymbol{\beta}}_{\bar{S}} - q\alpha_0\boldsymbol{\Sigma}_{SS}^{-1}|\hat{\boldsymbol{\beta}}_S|^{q-1}sign(\hat{\boldsymbol{\beta}}_S). \qquad (A26)$$

Substituting into (A24), we have

$$
\begin{aligned}
R_q(\tau^2, \alpha_0) \;\; = \;\; & \frac{1}{p}E\{(\hat{\boldsymbol{\beta}}_S - \mathbf{U}_S)^T\boldsymbol{\Sigma}_{SS}(\hat{\boldsymbol{\beta}}_S - \mathbf{U}_S) + 2(\hat{\boldsymbol{\beta}}_S - \mathbf{U}_S)^T\boldsymbol{\Sigma}_{S\bar{S}}\hat{\boldsymbol{\beta}}_{\bar{S}} + \hat{\boldsymbol{\beta}}_{\bar{S}}^T\boldsymbol{\Sigma}_{\bar{S}\bar{S}}\hat{\boldsymbol{\beta}}_{\bar{S}}\} \\
= \;\; & \frac{1}{p}E[\{(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S - q\alpha_0|\hat{\boldsymbol{\beta}}_S|^{q-1}sign(\hat{\boldsymbol{\beta}}_S) + \boldsymbol{\Sigma}_{S\bar{S}}\hat{\boldsymbol{\beta}}_{\bar{S}}\}^T\boldsymbol{\Sigma}_{SS}^{-1} \\
& \{(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S - q\alpha_0|\hat{\boldsymbol{\beta}}_S|^{q-1}sign(\hat{\boldsymbol{\beta}}_S) - \boldsymbol{\Sigma}_{S\bar{S}}\hat{\boldsymbol{\beta}}_{\bar{S}}\} + \hat{\boldsymbol{\beta}}_{\bar{S}}^T\boldsymbol{\Sigma}_{\bar{S}\bar{S}}\hat{\boldsymbol{\beta}}_{\bar{S}}] \\
= \;\; & \frac{1}{p}E[\{(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S - q\alpha_0|\hat{\boldsymbol{\beta}}_S|^{q-1}sign(\hat{\boldsymbol{\beta}}_S)\}^T\boldsymbol{\Sigma}_{SS}^{-1}\{(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S - q\alpha_0|\hat{\boldsymbol{\beta}}_S|^{q-1}sign(\hat{\boldsymbol{\beta}}_S)\} \\
& + \hat{\boldsymbol{\beta}}_{\bar{S}}^T(\boldsymbol{\Sigma}_{\bar{S}\bar{S}} - \boldsymbol{\Sigma}_{\bar{S}S}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{S\bar{S}})\hat{\boldsymbol{\beta}}_{\bar{S}}] \qquad (A27)
\end{aligned}
$$

21

As $\tau \to 0$, $\hat{\boldsymbol{\beta}}_S \to U_S$. Thus, $|\hat{\boldsymbol{\beta}}_S|^{q-1} = O_p(\tau^{1-q})$, we have

$$R_q(\tau^2, \alpha_0) \;=\; \frac{1}{p}E[(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S^T \boldsymbol{\Sigma}_{SS}^{-1}(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S + \hat{\boldsymbol{\beta}}_{\bar{S}}^T(\boldsymbol{\Sigma}_{\bar{S}\bar{S}} - \boldsymbol{\Sigma}_{\bar{S}S}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{S\bar{S}})\hat{\boldsymbol{\beta}}_{\bar{S}}] + O(\tau^{1-q}).$$

The second term $\hat{\boldsymbol{\beta}}_{\bar{S}}^T(\boldsymbol{\Sigma}_{\bar{S}\bar{S}} - \boldsymbol{\Sigma}_{\bar{S}S}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{S\bar{S}})\hat{\boldsymbol{\beta}}_{\bar{S}} \geq 0$. Combining (A25) and (A26), we obtain the equation for $\hat{\boldsymbol{\beta}}_{\bar{S}}$ as

$$\hat{\boldsymbol{\beta}}_{\bar{S}} = argmin_{\boldsymbol{\beta}}\left\{|\hat{\boldsymbol{\beta}} - \overline{\boldsymbol{\Sigma}}_{\bar{S}\bar{S}}^{-1}\{(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S - \boldsymbol{\Sigma}_{\bar{S}S}(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_{\bar{S}}\}|_{\overline{\boldsymbol{\Sigma}}_{\bar{S}\bar{S}}} + q\alpha_0|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}})\right\}$$

which goes to 0 if α_0 is large enough. Here $\overline{\boldsymbol{\Sigma}}_{\bar{S}\bar{S}} = \boldsymbol{\Sigma}_{\bar{S}\bar{S}} - \boldsymbol{\Sigma}_{\bar{S}S}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{S\bar{S}}$. Therefore denote $\alpha_{0\star}$ the optimal α_0, we obtain

$$R_q(\tau^2, \alpha_{0\star}) \xrightarrow{\tau \to 0} \frac{1}{p}E[(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S^T \boldsymbol{\Sigma}_{SS}^{-1}(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S] = \epsilon,$$

and

$$\frac{d\psi_q(\tau^2, \alpha_{0\star}\tau^{2-q})}{d\tau^2}\bigg|_{\tau^2=0} = \frac{\epsilon}{\delta}\,.$$

7.3 Proof of Proposition 4

Proof 3 *Define* $\lambda = \alpha_0\tau^2$*, then from (4), we have for* $q = 0$

$$\begin{aligned}
\bar{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}/\tau = \boldsymbol{\eta}_0(\boldsymbol{\beta}_0/\tau + \boldsymbol{\Sigma}^{-1/2}\mathbf{z}, \alpha_0) \\
&= argmin_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0/\tau - \boldsymbol{\Sigma}^{-1/2}\mathbf{z}\|_{\boldsymbol{\Sigma}}^2 + \alpha_0\|\boldsymbol{\beta}\|_0\right\}. \quad\quad (A28)
\end{aligned}$$

For block diagonal covariance $\boldsymbol{\Sigma}$ *with block* $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$*, equation (A28) can be decomposed into 2-dimensional blocks each of which can be solved as*

$$(\bar{\beta}_1, \bar{\beta}_2) = (\hat{\beta}_1/\tau, \hat{\beta}_2/\tau) = argmin_{\beta_1,\beta_2}\mathcal{L},$$

where

$$\mathcal{L} \;=\; \frac{1}{2}\{(\beta_1 - y_1)^2 + (\beta_2 - y_2)^2 + 2\rho(\beta_1 - y_1)(\beta_2 - y_2)\} + \alpha_0(\|\beta_1\|_0 + \|\beta_2\|_0).$$

For problem (A28), $y_1 = U_1 + \xi_1$*,* $y_2 = U_2 + \xi_2$*, where* $U_1 = \beta_{0,1}/\tau$*,* $U_2 = \beta_{0,2}/\tau$*,* $\beta_{0,1}, \beta_{0,2} \overset{i.i.d.}{\sim} G$*, and*

$$\begin{aligned}
\xi_1 &= \frac{1}{2}\left(\sqrt{\frac{1}{1+\rho}} + \sqrt{\frac{1}{1-\rho}}\right)z_1 + \frac{1}{2}\left(\sqrt{\frac{1}{1+\rho}} - \sqrt{\frac{1}{1-\rho}}\right)z_2, \\
\xi_2 &= \frac{1}{2}\left(\sqrt{\frac{1}{1+\rho}} - \sqrt{\frac{1}{1-\rho}}\right)z_1 + \frac{1}{2}\left(\sqrt{\frac{1}{1+\rho}} + \sqrt{\frac{1}{1-\rho}}\right)z_2,
\end{aligned}$$

22

with $z_1, z_2 \overset{i.i.d.}{\sim} N(0,1)$.

As shown in Figure 10, the two dimensional space can be divided into nine regions. In each region, the estimator is chosen to be the one that gives the lowest loss $\mathcal{L}$. The solution can be summarized as

$$
\begin{cases}
\hat{\beta}_1 > 0 \ \& \ \hat{\beta}_2 > 0 & if \ \ \mathbf{y} \in I_1 \\
\hat{\beta}_1 > 0 \ \& \ \hat{\beta}_2 = 0 & if \ \ \mathbf{y} \in I_5 \\
\hat{\beta}_1 > 0 \ \& \ \hat{\beta}_2 < 0 & if \ \ \mathbf{y} \in I_4 \\
\hat{\beta}_1 = 0 \ \& \ \hat{\beta}_2 > 0 & if \ \ \mathbf{y} \in I_7 \\
\hat{\beta}_1 = 0 \ \& \ \hat{\beta}_2 < 0 & if \ \ \mathbf{y} \in I_8 \\
\hat{\beta}_1 < 0 \ \& \ \hat{\beta}_2 > 0 & if \ \ \mathbf{y} \in I_2 \\
\hat{\beta}_1 < 0 \ \& \ \hat{\beta}_2 = 0 & if \ \ \mathbf{y} \in I_6 \\
\hat{\beta}_1 < 0 \ \& \ \hat{\beta}_2 < 0 & if \ \ \mathbf{y} \in I_3 \\
\hat{\beta}_1 = 0 \ \& \ \hat{\beta}_2 = 0 & if \ \ \mathbf{y} \in I_9
\end{cases}
\tag{A29}
$$

where

$$
\begin{cases}
I_1 &= \ I\left(y_1 > \sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ y_2 > \sqrt{\frac{2\alpha_0}{1-\rho^2}}\right) \\
I_5 &= \ I\left(|y_2| < \sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ |y_1| > |y_2| \ \& \ y_1 + \rho y_2 > \sqrt{2\alpha_0}\right) \\
I_4 &= \ I\left(y_1 > \sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ y_2 < -\sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ y_1^2 + y_2^2 + 2\rho y_1 y_2 > \sqrt{4\alpha_0}\right) \\
I_7 &= \ I\left(|y_1| < \sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ |y_1| < |y_2| \ \& \ y_2 + \rho y_1 > \sqrt{2\alpha_0}\right) \\
I_8 &= \ I\left(|y_1| < \sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ |y_1| < |y_2| \ \& \ y_2 + \rho y_1 < -\sqrt{2\alpha_0}\right) \\
I_2 &= \ I\left(y_1 < -\sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ y_1^2 + y_2^2 + 2\rho y_1 y_2 > \sqrt{4\alpha_0} \ \& \ y_2 > \sqrt{\frac{2\alpha_0}{1-\rho^2}}\right) \\
I_6 &= \ I\left(|y_2| < \sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ |y_1| > |y_2| \ \& \ y_1 + \rho y_2 < -\sqrt{2\alpha_0}\right) \\
I_3 &= \ I\left(y_1 < -\sqrt{\frac{2\alpha_0}{1-\rho^2}} \ \& \ y_2 < -\sqrt{\frac{2\alpha_0}{1-\rho^2}}\right) \\
I_9 &= \ (\cup_{i=1}^{8} I_i)^c
\end{cases}
$$

We consider three different scenarios. The first scenario includes regions 1-4 where both $\hat{\beta}_1$ and $\hat{\beta}_2$ are nonzero and the loss is $\mathcal{L} = 2\alpha_0$. Its contribution to the prediction risk is

$$
\begin{aligned}
R_1(\tau^2, \alpha_0) &= \ E\{(\hat{\beta}_1 - \beta_{0,1})^2 + (\hat{\beta}_2 - \beta_{0,2})^2 + 2\rho(\hat{\beta}_1 - \beta_{0,1})(\hat{\beta}_2 - \beta_{0,2})\}(I_1 + I_2 + I_3 + I_4)/2 \\
&= \ \tau^2 E\{(\xi_1^2 + \xi_2^2 + 2\rho\xi_1\xi_2)\}(I_1 + I_2 + I_3 + I_4)/2.
\end{aligned}
\tag{A30}
$$

The second scenario includes regions 5-8 where $\hat{\beta}_1 \neq 0, \hat{\beta}_2 = 0$ or $\hat{\beta}_1 = 0, \hat{\beta}_2 \neq 0$ and the loss is $\mathcal{L} = \frac{1}{2}(1-\rho^2)y_2^2 + \alpha_0$ or $\mathcal{L} = \frac{1}{2}(1-\rho^2)y_1^2 + \alpha_0$. Its contribution to the prediction risk is

$$
\begin{aligned}
R_2(\tau^2, \alpha_0) &= \ \tau^2 \left[E\{(\xi_1 + \rho\xi_2)^2 + (1-\rho^2)U_2^2\}(I_5 + I_6) \right. \\
&\qquad \left. + E\{(\xi_2 + \rho\xi_1)^2 + (1-\rho^2)U_1^2\}(I_7 + I_8) \right]/2.
\end{aligned}
\tag{A31}
$$

23

The third scenario includes regions 9 where $\hat{\beta}_1 = \hat{\beta}_2 = 0$ and the loss is $\mathcal{L} = (y_1^2 + y_2^2 + 2\rho y_1 y_2)/2$. Its contribution to the prediction risk is

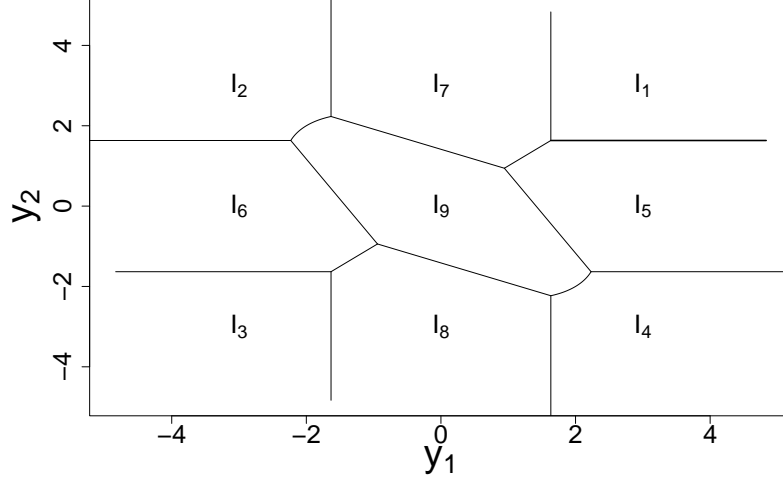$$R_3(\tau^2, \alpha_0) \quad = \quad \tau^2 E\{(U_1^2 + U_2^2 + 2\rho U_1 U_2)\} I_9/2. \tag{A32}$$

Figure 10: Illustration of the solution (A29) for equation (A28) in two dimensional space. Here $\rho = 0.5$ and $\alpha_0 = 1$.

To find the optimal α_0 for minimizing $R(\tau^2, \alpha_0) = R_1(\tau^2, \alpha_0) + R_2(\tau^2, \alpha_0) + R_3(\tau^2, \alpha_0)$, we need to take derivative over α_0. Since the explicit function $R(\tau^2, \alpha_0)$ in each region does not depend on α_0, changing α_0 only causes the change of the boundaries among different regions. According to Stokes's theorem, as in Theorem 1 of Baddeley (1977), we conclude that the dominant contributions to $\lim_{\tau \to 0} \frac{\partial R(\tau^2, \alpha_0)}{\partial \alpha_0}$ come from the boundaries between scenarios 1 and 2 as well as the boundaries between scenarios 2 and 3. For example, the contribution from the boundary between regions 1 and 5 is

$$\tau^2 \sqrt{\frac{1-\rho^2}{2\alpha_0}} E\left\{\left(\sqrt{\frac{2\alpha_0}{1-\rho^2}} - U_2\right)^2 - U_2^2\right\} I\left(U_1 + \xi_1 > \sqrt{\frac{2\alpha_0}{1-\rho^2}}\right) \delta\left(U_2 + \xi_2 = \sqrt{\frac{2\alpha_0}{1-\rho^2}}\right)$$

$$\sim \quad \epsilon_+^2 E\left\{\left(\sqrt{\frac{2\alpha_0}{1-\rho^2}} - U_2\right)^2 - U_2^2\right\} \phi(\sqrt{2\alpha_0} - \sqrt{1-\rho^2}U_2) + \epsilon_+(1-\epsilon)\frac{2\alpha_0}{1-\rho^2}\phi(\sqrt{2\alpha_0}),$$

24

where $\epsilon_+ = P(\beta_0 > 0)$. The contribution from the boundary between regions 5 and 9 is

$$\tau^2 \sqrt{\frac{1}{2\alpha_0}} E\{(\sqrt{2\alpha_0} - U_1 + \rho U_2)^2 - (U_1 + \rho U_2)^2\} I \left(\frac{\sqrt{2\alpha_0}}{1 + \rho} > U_2 + \xi_2 > -\sqrt{\frac{2\alpha_0}{1 - \rho^2}}\right)$$

$$\delta\left(U_1 + \rho U_2 + \xi_1 + \rho \xi_2 = \sqrt{2\alpha_0}\right)$$

$$\sim (1 - \epsilon)^2 E \alpha_0 \phi(\sqrt{2\alpha_0}).$$

The contributions from other boundaries can be derived similarly. The total contributions include three dominant terms: $\phi(\sqrt{2\alpha_0})$, $\phi(\sqrt{2\alpha_0} - \sqrt{1 - \rho^2}U_1)$, and $\phi(\sqrt{2\alpha_0} - \sqrt{1 - \rho^2}U_2)$. Therefore, we need to have $\lim_{\tau \to 0} \sqrt{2\alpha_{0\star}}\tau = \mu\sqrt{1 - \rho^2}/2$ to ensure the existence of the finite solutions for $\frac{\partial R(\tau^2, \alpha_0)}{\partial \alpha_0} = 0$. Substituting into (A30), (A31), and (A32), we obtain

$$\begin{aligned}
R_1(\tau^2, \alpha_{0\star}) &= \frac{\tau^2}{2} E\{(\xi_1^2 + \xi_2^2 + 2\rho\xi_1\xi_2)\} I\left(|U_1 + \xi_1| > \sqrt{\frac{2\alpha_{0\star}}{1 - \rho^2}}\right) I\left(|U_2 + \xi_2| > \sqrt{\frac{2\alpha_{0\star}}{1 - \rho^2}}\right) \\
&= \tau^2[\epsilon^2(1 + o(\sqrt{2\alpha_{0\star}}\phi(\sqrt{2\alpha_{0\star}}))) + \epsilon(1 - \epsilon)o(\sqrt{2\alpha_{0\star}}\phi(\sqrt{2\alpha_{0\star}})) + (1 - \epsilon)^2 o(\phi(\sqrt{2\alpha_{0\star}})^2)] \\
&= \tau^2[\epsilon^2 + o(\sqrt{2\alpha_{0\star}}\phi(\sqrt{2\alpha_{0\star}}))],
\end{aligned}$$

$$\begin{aligned}
R_2(\tau^2, \alpha_{0\star}) &= \frac{\tau^2}{2} E\{(\xi_1 + \rho\xi_2)^2 + (1 - \rho^2)U_2^2\} I\left(|y_1 + \rho y_2| > \sqrt{2\alpha_{0\star}}\right) I\left(|U_2 + \xi_2| < \sqrt{\frac{2\alpha_{0\star}}{1 - \rho^2}}\right) \\
&\quad + \frac{\tau^2}{2} E\{(\xi_2 + \rho\xi_1)^2 + (1 - \rho^2)U_1^2\} I\left(|y_2 + \rho y_1| > \sqrt{2\alpha_{0\star}}\right) I\left(|U_1 + \xi_1| < \sqrt{\frac{2\alpha_{0\star}}{1 - \rho^2}}\right) \\
&= \tau^2[\epsilon^2 o(\sqrt{2\alpha_{0\star}}\phi(\sqrt{2\alpha_{0\star}})) + \epsilon(1 - \epsilon)(1 + o(\sqrt{2\alpha_{0\star}}\phi(\sqrt{2\alpha_{0\star}}))) + (1 - \epsilon)^2 o(\sqrt{2\alpha_{0\star}}\phi(\sqrt{2\alpha_{0\star}}))] \\
&= \tau^2[\epsilon(1 - \epsilon) + o(\sqrt{2\alpha_{0\star}}\phi(\sqrt{2\alpha_{0\star}}))],
\end{aligned}$$

and

$$\begin{aligned}
R_3(\tau^2, \alpha_{0\star}) &= \frac{\tau^2}{2} E(U_1^2 + U_2^2 + 2\rho U_1 U_2)\left(|y_1 + \rho y_2| > \sqrt{2\alpha_{0\star}}\right)\left(|y_2 + \rho y_1| > \sqrt{2\alpha_{0\star}}\right) \\
&\leq \frac{\tau^2}{2} E(U_1^2 + U_2^2 + 2\rho U_1 U_2) I\left(|U_1 + \xi_1| < \frac{(\rho + \sqrt{1 - \rho^2})\sqrt{2\alpha}}{\sqrt{1 - \rho^2}}\right) \\
&\quad I\left(|U_2 + \xi_2| < \frac{(\rho + \sqrt{1 - \rho^2})\sqrt{2\alpha_{0\star}}}{\sqrt{1 - \rho^2}}\right) \\
&= \tau^2[\epsilon^2 o(\phi((2 - \rho - \sqrt{1 - \rho^2})\sqrt{2\alpha_{0\star}})^2) + \epsilon(1 - \epsilon)o(\sqrt{2\alpha_{0\star}}\phi((2 - \rho - \sqrt{1 - \rho^2})\sqrt{2\alpha_{0\star}}))] \\
&= \tau^2 o(\sqrt{2\alpha_{0\star}}\phi((2 - \rho - \sqrt{1 - \rho^2})\sqrt{2\alpha_{0\star}})).
\end{aligned}$$

Thus, as $\tau \to 0$, the total MSE becomes $R(\tau^2, \alpha_{0\star}) = \tau^2(\epsilon + o(\phi(\tilde{\mu}\tau^{-1})))$, where $\tilde{\mu} = (2 - \rho - \sqrt{1 - \rho^2})\sqrt{1 - \rho^2}\mu/2$. The fixed point equation for small σ_w^2 can be written as

$$\begin{aligned}
\tau_l^2 &= \sigma_w^2 + \frac{R(\tau_l^2, \alpha_{0\star})}{\delta} \\
&= \sigma_w^2 + \frac{\tau_l^2}{\delta}(\epsilon + o(\phi(\tilde{\mu}\tau_l^{-1}))).
\end{aligned}$$

25

Therefore,

$$\tau_l^2 = \frac{\delta\sigma_w^2}{\delta - \epsilon} + o\left(\phi\left(\tilde{\mu}\sqrt{\frac{\delta - \epsilon}{\delta}}\sigma_w^{-1}\right)\right).$$

7.4 Proof of Proposition 5

Proof 4 *For* $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, *denote* $\rho_1 = \frac{1}{2}(\sqrt{1+\rho} + \sqrt{1-\rho})$ *and* $\rho_2 = \frac{1}{2}(\sqrt{1+\rho} - \sqrt{1-\rho})$.
Let $\lambda = \alpha_0\tau^{2-q}$, *then the solution can be decomposed into 2-dimensional block of* $\boldsymbol{\beta} \in \mathbb{R}^2$ *as*

$$\hat{\beta}_1, \hat{\beta}_2 = argmin_{\beta_1,\beta_2}\left[\mathcal{L}\right],$$

where

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0 - \tau\Sigma^{-1/2}\mathbf{z}\|_\Sigma^2 + \alpha_0\tau^{2-q}\|\boldsymbol{\beta}\|_q^q \\
&= \frac{1}{2}\{(\beta_1 - \beta_{0,1})^2 + (\beta_2 - \beta_{0,2})^2 + 2\rho(\beta_1 - \beta_{0,1})(\beta_2 - \beta_{0,2})\} \\
&\quad - \tau\xi_1(\beta_1 - \beta_{0,1}) - \tau\xi_2(\beta_2 - \beta_{0,2}) + \frac{1}{2}\tau^2\|\mathbf{z}\|^2 + \alpha_0\tau^{2-q}(|\beta_1|^q + |\beta_2|^q), \quad (A33)
\end{aligned}
$$

with $\xi_1 = \rho_1 z_1 + \rho_2 z_2$, $\xi_2 = \rho_2 z_1 + \rho_1 z_2$. *If both* $\hat{\beta}_1$ *and* $\hat{\beta}_2$ *are nonzero, they should satisfy*

$$
\begin{cases}
\hat{\beta}_1 - \beta_{0,1} + \rho(\hat{\beta}_2 - \beta_{0,2}) - \tau\xi_1 + \alpha_0\tau^{2-q}q|\hat{\beta}_1|^{q-1}sign(\hat{\beta}_1) &= 0, \\
\rho(\hat{\beta}_1 - \beta_{0,1}) + \hat{\beta}_2 - \beta_{0,2} - \tau\xi_2 + \alpha_0\tau^{2-q}q|\hat{\beta}_2|^{q-1}sign(\hat{\beta}_2) &= 0.
\end{cases} \quad (A34)
$$

The ψ *function defined in (2.3) can now be written as*

$$
\begin{aligned}
\psi_q(\tau^2, \alpha_0\tau^{2-q}) &= \sigma_w^2 + \frac{1}{2\delta}E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\Sigma \\
&= \sigma_w^2 + \frac{1}{2\delta}E\{(\hat{\beta}_1 - \beta_{0,1})^2 + (\hat{\beta}_2 - \beta_{0,2})^2 + 2\rho(\hat{\beta}_1 - \beta_{0,1})(\hat{\beta}_2 - \beta_{0,2})\}(A35)
\end{aligned}
$$

We consider three different situations. In the first case: $\beta_{0,1} \neq 0$ *and* $\beta_{0,2} \neq 0$. *Then as* $\tau \to 0$,
both $\hat{\beta}_1$ *and* $\hat{\beta}_2$ *are nonzero and (A34) can be written as*

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \tau\Sigma^{-1/2}\mathbf{z} - \alpha_0\tau^{2-q}q\Sigma^{-1}|\boldsymbol{\beta}_0|^{q-1}sign(\boldsymbol{\beta}_0) + o_p(\tau^{2-q}).$$

Substituting into (A35), its contribution to the prediction risk is

$$
\begin{aligned}
R_1(\tau^2, \alpha_0\tau^{2-q}) &= \epsilon^2\tau^2 E\left[1 + \frac{1}{2}\alpha_0^2\tau^{2-2q}q^2\{(\boldsymbol{\beta}_0)^{q-1}sign(\boldsymbol{\beta}_0)\}^T\Sigma^{-1}(\boldsymbol{\beta}_0)^{q-1}sign(\boldsymbol{\beta}_0)\right] \\
&\quad + o(\tau^{4-2q}). \quad (A36)
\end{aligned}
$$

Taking derivative over α_0, *we obtain*

$$
\begin{aligned}
\frac{\partial R_1(\tau^2, \alpha_0\tau^{2-q})}{\tau^2\partial\alpha_0} &= \epsilon^2 E\left[\alpha_0\tau^{2-2q}q^2\{(\boldsymbol{\beta}_0)^{q-1}sign(\boldsymbol{\beta}_0)\}^T\Sigma^{-1}|\boldsymbol{\beta}_0|^{q-1}sign(\boldsymbol{\beta}_0)\right] + o(\tau^{2-2q}) \\
&= 2\epsilon^2\alpha_0\tau^{2-2q}q^2 D_0 + o(\tau^{2-2q}). \quad (A37)
\end{aligned}
$$

26

where $D_0 = E\{|\beta_0|^{2q-2}\} - \rho(E\{|\beta_0|^{q-1}sign(\beta_0)\})^2/(1-\rho^2)$.

In the second case: $\beta_{0,1} \neq 0$ and $\beta_{0,2} = 0$ (or $\beta_{0,2} \neq 0$ and $\beta_{0,1} = 0$). The first equation of (A34) becomes

$$\hat{\beta}_1 - \beta_{0,1} = \tau\xi_1 - \rho\hat{\beta}_2 - \alpha_0\tau^{2-q}q|\beta_{0,1}|^{q-1}sign(\beta_{0,1}) + o_p(\tau^{2-q}).$$

Substituting into (A33), we obtain the solution for β_2

$$\hat{\beta}_2 = argmin_\beta \left\{(1-\rho^2)[\beta - (1-\rho^2)^{-1}\tau(\xi_2 - \rho\xi_1)]^2 + \alpha_0\tau^{2-q}q|\beta|^{q-1}sign(\beta)\right\}$$

which is zero when $|\xi_2 - \rho\xi_1| < (1-\rho^2)C_q(\frac{\alpha_0}{1-\rho^2})^{1/(2-q)}$ with $C_q = [2(1-q)]^{1/(2-q)} + q[2(1-q)]^{(q-1)/(2-q)}$ and nonzero solved otherwise Denote $I_1 = I\left(|\xi_2 - \rho\xi_1| > (1-\rho^2)C_q(\frac{\alpha_0}{1-\rho^2})^{\frac{1}{2-q}}\right)$. Its contribution to the prediction risk can be written as

$$
\begin{aligned}
R_2(\tau^2, \alpha_0\tau^{2-q}) &= \epsilon(1-\epsilon)E\{(\tau\xi_1 - \alpha_0\tau^{2-q}q|\beta_{0,1}|^{q-1}sign(\beta_{0,1}))^2 + (1-\rho^2)\hat{\beta}_2^2\} \\
&= \epsilon(1-\epsilon)\tau^2\{1 + \alpha_0^2\tau^{2-2q}q^2 E|\beta_{0,1}|^{2q-2} + (1-\rho^2)E\tilde{\beta}^2 I_1\} + o(\tau^{2-2q})(A38)
\end{aligned}
$$

where $\tilde{\beta}$ is the solution of

$$(1-\rho^2)\tilde{\beta} - (\xi_2 - \rho\xi_1) + \alpha_0 q|\tilde{\beta}|^{q-1}sign(\tilde{\beta}) = 0.$$

Taking derivative over α_0, we obtain

$$
\begin{aligned}
\frac{\partial R_2(\tau^2, \alpha_0\tau^{2-q})}{\tau^2\partial\alpha_0} &= 2\epsilon(1-\epsilon)\left\{\alpha_0\tau^{2-2q}q^2 C_0 - (1-\rho^2)D_q^2\left(\frac{\alpha_0}{1-\rho^2}\right)^{\frac{2}{2-q}}C_q\left(\frac{1}{1-\rho^2}\right)^{\frac{q}{2(2-q)}}\frac{1}{2-q}\alpha_0^{\frac{q-1}{2-q}} \right. \\
&\qquad\qquad \left. \phi\left(C_q\left(\frac{1}{1-\rho^2}\right)^{\frac{q}{2(2-q)}}\alpha_0^{\frac{1}{2-q}}\right)\right\} + o(\tau^{2-q}) \qquad\qquad \text{(A39)} \\
&= 2\epsilon(1-\epsilon)\left\{\alpha_0\tau^{2-2q}q^2 C_0 - \left(\frac{1}{1-\rho^2}\right)^{\frac{3q}{2(2-q)}}\frac{C_q D_q^2}{2-q}\alpha_0^{\frac{q+1}{2-q}}\phi\left(C_q\left(\frac{1}{1-\rho^2}\right)^{\frac{q}{2(2-q)}}\alpha_0^{\frac{1}{2-q}}\right)\right\},
\end{aligned}
$$

where $C_0 = E|\beta_{0,1}|^{2q-2}$ and $D_q = [2(1-q)]^{\frac{1}{2-q}}$.

In the third case, $\beta_{0,1} = \beta_{0,2} = 0$, its contribution to MSE can be summarized as

$$R_3(\tau^2, \alpha_0\tau^{2-q}) = \frac{1}{2}E(\hat{\beta}_1^2 + \hat{\beta}_2^2 + 2\rho\hat{\beta}_1\hat{\beta}_2), \qquad\qquad \text{(A40)}$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the solution that minimizes

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{\beta} - \tau\Sigma^{-1/2}\mathbf{z}\|_\Sigma^2 + \alpha_0\tau^{2-q}\|\boldsymbol{\beta}\|_q^q. \qquad\qquad \text{(A41)}$$

As shown in Figure 11, we can divide the 2-dimensional space into nine regions. In region I_9, $\hat{\beta}_1 = 0$ and $\hat{\beta}_2 = 0$; in regions I_5 and I_6, $\hat{\beta}_2 = 0$ and $\hat{\beta}_1$ is solved by $\hat{\beta}_1 - \tau\xi_1 +$

$\alpha_0 \tau^{2-q} q |\hat{\beta}_1|^{q-1} sign(\hat{\beta}_1) = 0$; *in regions I_7 and I_8, $\hat{\beta}_1 = 0$ and $\hat{\beta}_2$ is solved by $\hat{\beta}_2 - \tau \xi_2 + \alpha_0 \tau^{2-q} q |\hat{\beta}_1|^{q-1} sign(\hat{\beta}_2) = 0$; In regions I_1, I_2, I_3, and I_4, $\hat{\beta}_1 \neq 0$ and $\hat{\beta}_2 \neq 0$ which are solved by*

$$\begin{cases} \hat{\beta}_1 + \rho \hat{\beta}_2 - \tau \xi_1 + \alpha_0 \tau^{2-q} q |\hat{\beta}_1|^{q-1} sign(\hat{\beta}_1) &= 0, \\ \rho \hat{\beta}_1 + \hat{\beta}_2 - \tau \xi_2 + \alpha_0 \tau^{2-q} q |\hat{\beta}_2|^{q-1} sign(\hat{\beta}_2) &= 0. \end{cases}$$

Then the derivative of $R_3(\tau^2, \alpha_0 \tau^{2-q})$ over α_0 involves the explicit derivative inside each region and integrals over the boundaries among different regions over 1-dimensional boundary curve. According to Stokes's theorem, as in Theorem 1 of Baddeley (1977), we conclude that the main contributions come from the four dominated boundaries which connect region 9 and four other regions 5, 6, 7, and 8 respectively. The contribution can be summarized as

$$\frac{\partial R_3(\tau^2, \alpha_0 \tau^{2-q})}{\tau^2 \partial \alpha_0} = -2(1-\epsilon)^2 \frac{C_q D_q^2}{2-q} \alpha_0^{(q+1)/(2-q)} \phi(C_q \alpha_0^{1/(2-q)}) + o(\tau^{2-q}). \tag{A42}$$

Put (A37), (A52), and (A42) together, we obtain

$$2\epsilon^2 \alpha_0 \tau^{2-2q} q^2 D_0 + 2\epsilon(1-\epsilon) \alpha_0 \tau^{2-2q} q^2 C_0$$

$$-2\epsilon(1-\epsilon) f(\rho) \frac{C_q D_q^2}{2-q} \alpha_0^{(q+1)/(2-q)} \phi\left(C_q \left(\frac{1}{1-\rho^2}\right)^{\frac{q}{2(2-q)}} \alpha_0^{1/(2-q)}\right)$$

$$-2(1-\epsilon)^2 \frac{C_q D_q^2}{2-q} \alpha_0^{(q+1)/(2-q)} \phi(C_q \alpha_0^{1/(2-q)}) = 0,$$

where $f(\rho) = (\frac{1}{1-\rho^2})^{\frac{3q}{2(2-q)}}$. Therefore, for $0<\rho<1$, we have

$$\frac{\tau^{2-2q}}{\alpha_0^{\frac{2q-1}{2-q}} \phi(C_q \alpha_0^{1/(2-q)})} = \frac{(1-\epsilon)^2 C_q D_q^2}{\epsilon q^2 (2-q)(\epsilon D_0 + (1-\epsilon)C_0)}; \tag{A43}$$

while for $\rho = 0$, we have

$$\frac{\tau^{2-2q}}{\alpha_0^{\frac{2q-1}{2-q}} \phi(C_q \alpha_0^{1/(2-q)})} = \frac{(1-\epsilon) C_q D_q^2}{\epsilon q^2 (2-q) C_0}. \tag{A44}$$

From (A43) and (A44), it is straightforward to show that

$$e^{-\frac{C_q^2 \alpha_0^{\frac{2}{2-q}}}{2}} \sim \tau^{2-2q} \quad and \quad (4-4q) \log \frac{1}{\tau} \sim C_q^2 \alpha_0^{\frac{2}{2-q}}.$$

Combining (A36), (A38), and (A40) together, we obtain the risk function

$$\begin{aligned} R(\tau^2, \alpha_0 \tau^{2-q}) &= \tau^2 \left\{ \epsilon + \epsilon \alpha_0^2 \tau^{2-q} q^2 (\epsilon D_0 + (1-\epsilon)C_0) \right\} \\ &= \tau^2 \left\{ \epsilon + \epsilon \left(\frac{4-4q}{C_q^2} \log \frac{1}{\tau}\right)^{2-q} \tau^{2-q} q^2 (\epsilon D_0 + (1-\epsilon)C_0) \right\}. \end{aligned}$$

From the high order analysis for the fixed point equation

$$\tau^2 = \sigma_w^2 + \frac{R(\tau^2, \alpha_0 \tau^{2-q})}{\delta},$$

we obtain the leading order term $\tau_0^2 = \frac{\delta}{\delta - \epsilon} \sigma_w^2$ *and the next-to-leading order term*

$$
\begin{aligned}
\tau_1^2 &= \frac{q^2 \epsilon \tau_0^{4-2q}}{\delta - \epsilon} \left(\frac{4 - 4q}{C_q^2} \log \frac{1}{\tau_0} \right)^{2-q} (\epsilon D_0 + (1 - \epsilon) C_0) \\
&= \frac{q^2 \epsilon \delta^{2-q} \sigma_w^{4-2q}}{(\delta - \epsilon)^{3-q}} \left(\frac{4 - 4q}{C_q^2} \log \frac{1}{\sigma_w} \right)^{2-q} (\epsilon D_0 + (1 - \epsilon) C_0).
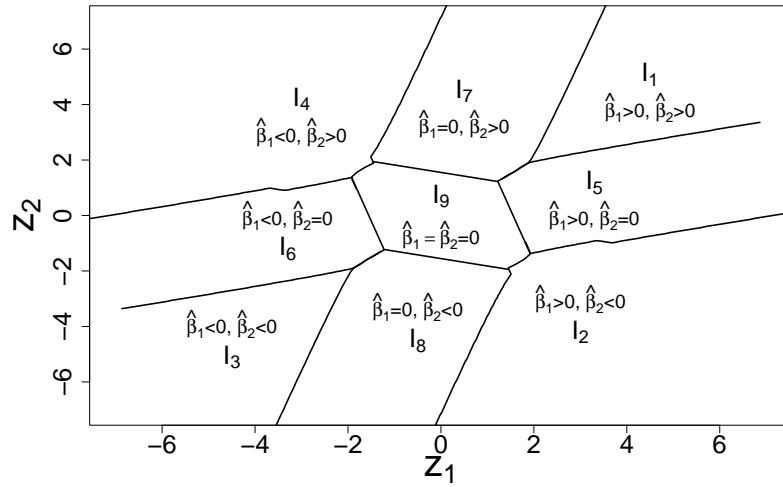\end{aligned}
$$

Hence we obtain (3.2).

Figure 11: Illustration of the solution for equation (A41) in two dimensional space. Here $\rho = 0.5$ and $\alpha = 1$.

7.5 Proof of Proposition 6

Proof 5 *Denote* $U = \boldsymbol{\beta}_0 / \tau$. *For fixed* U, *denote* $S = \{j | U_j \neq 0\}$. *From (A27), we obtain*

$$R_q(\tau^2, \alpha_0) \geq \frac{1}{p} E[\{(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S - q\alpha_0 |\hat{\boldsymbol{\beta}}_S|^{q-1} sign(\hat{\boldsymbol{\beta}}_S)\}^T \boldsymbol{\Sigma}_{SS}^{-1} \{(\boldsymbol{\Sigma}^{1/2}\mathbf{z})_S - q\alpha_0 |\hat{\boldsymbol{\beta}}_S|^{q-1} sign(\hat{\boldsymbol{\beta}}_S)\}$$

which goes to infinity as $\tau \to 0$ *for fixed* $\alpha_0 > 0$ *since* $\hat{\boldsymbol{\beta}}_S \to \frac{\boldsymbol{\beta}_{0,S}}{\tau}$ *and* $q > 1$. *Therefore,* $\alpha_{0\star}(\tau) = 0$ *and we obtain* $R_q(\tau^2, \alpha_\star(\tau)) \xrightarrow{\tau \to 0} 1$. *From (9), we have* $\tau^2 = \sigma_w^2 + \frac{\tau^2 R_q(\tau^2, \alpha_{0\star}(\tau))}{\delta}$ *which implies that* $\tau^2(\delta - R_q(\tau^2, \alpha_{0\star}(\tau))) = \delta\sigma_w^2$. *If* $\delta > 1$, *we have* $\tau \xrightarrow{\sigma_w \to 0} 0$ *since* $R_q(\tau^2, \alpha_{0\star}(\tau)) \xrightarrow{\tau \to 0} 1 < \delta$. *On the other hand, if* $\delta < 1$, *we have* $\tau > 0$ *as* $\sigma_w \to 0$ *because* $R_q(\tau^2, \alpha_{0\star}(\tau)) \xrightarrow{\tau \to 0} 1$ *and* $R_q(\tau^2, \alpha_\star(\tau)) \xrightarrow{\tau \to \infty} 0$ *which leads to* $\tau \xrightarrow{\sigma_w \to 0} \tau_\star > 0$ *such that* $R_q(\tau_\star, \alpha_{0\star}(\tau_\star)) = \delta$.

29

7.6   Proof of Proposition 7

Proof 6 *For* $\lambda = \alpha_0 \tau^{2-q}$, *the the solution of* $\boldsymbol{\eta}_q$ *in (4) in the case* $1 < q \le 2$ *is the minimizer of loss*

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0 - \tau\boldsymbol{\Sigma}^{-1/2}\mathbf{z}\|_{\boldsymbol{\Sigma}}^2 + \alpha_0\tau^{2-q}\|\boldsymbol{\beta}\|_q^q.$$

It is the solution of the following equation

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \tau\boldsymbol{\Sigma}^{1/2}\mathbf{z} + q\alpha_0\tau^{2-q}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}}) = 0 \tag{A45}$$

which can also be written as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \tau\boldsymbol{\Sigma}^{-1}\{\boldsymbol{\Sigma}^{1/2}\mathbf{z} - q\alpha_0\tau^{1-q}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}})\}, \tag{A46}$$

or

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \tau\boldsymbol{\Sigma}^{-1/2}\mathbf{z} - q\alpha_0\tau^{2-q}\boldsymbol{\Sigma}^{-1}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}}). \tag{A47}$$

Taking derivative of (A47) over $\mathbf{z}$, *we have*

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{z}} = \tau\boldsymbol{\Sigma}^{-1/2} - q(q-1)\alpha_0\tau^{2-q}\boldsymbol{\Sigma}^{-1}diag[|\hat{\boldsymbol{\beta}}|^{q-2}sign(\hat{\boldsymbol{\beta}})]\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{z}}.$$

Thus

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{z}} = \tau\left\{\mathbf{I} + q(q-1)\alpha_0\tau^{2-q}\boldsymbol{\Sigma}^{-1}diag[|\hat{\boldsymbol{\beta}}|^{q-2}sign(\hat{\boldsymbol{\beta}})]\right\}^{-1}\boldsymbol{\Sigma}^{-1/2}.$$

Taking derivative of (A47) over α_0, *we have*

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\alpha_0} = -q\tau^{2-q}\boldsymbol{\Sigma}^{-1}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}}) - q(q-1)\alpha_0\tau^{2-q}\boldsymbol{\Sigma}^{-1}|\hat{\boldsymbol{\beta}}|^{q-2}sign(\hat{\boldsymbol{\beta}})\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\alpha_0}.$$

Thus

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\alpha_0} = -\left\{\mathbf{I} + q(q-1)\alpha_0\tau^{2-q}\boldsymbol{\Sigma}^{-1}diag[|\hat{\boldsymbol{\beta}}|^{q-2}sign(\hat{\boldsymbol{\beta}})]\right\}^{-1}q\tau^{2-q}\boldsymbol{\Sigma}^{-1}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}}). \tag{A48}$$

Using (A45), we can derive the risk function as

$$\begin{aligned} R(\tau^2, \alpha_0) &= \frac{1}{p}E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_{\boldsymbol{\Sigma}}^2 \\ &= \frac{\tau^2}{p}E\left\{\|\mathbf{z}\|^2 - 2q\alpha_0\tau^{1-q}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}}) \right. \\ &\qquad \left. + q^2\alpha_0^2\tau^{2-2q}(|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}}))^T\boldsymbol{\Sigma}^{-1}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}})\right\}. \end{aligned} \tag{A49}$$

Then we can conclude that as $\tau \to 0$, the optimal $\alpha_0^\star \to 0$ because $\boldsymbol{\beta} \to \boldsymbol{\beta}_0$ and $2 - 2q < 0$. Therefore in order for the third term to be finite, we need $\alpha_0^\star \to 0$.

In order to obtain the second dominant term, we need to take the first order derivative of $R(\tau^2, \alpha_0)$ over α_0 and assume it equal to 0. For $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, the equation can be decomposed into 2-dimensional block. Denote $\rho_1 = \frac{1}{2}(\sqrt{1 + \rho} + \sqrt{1 - \rho})$ and $\rho_2 = \frac{1}{2}(\sqrt{1 + \rho} - \sqrt{1 - \rho})$, we can obtain

$$
\begin{aligned}
R(\tau^2, \alpha_0) &= \frac{1}{2} E \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_{\Sigma}^2 \\
&= \frac{1}{2} E\{(\hat{\beta}_1 - \beta_{0,1})^2 + (\hat{\beta}_2 - \beta_{0,2})^2 + 2\rho(\hat{\beta}_1 - \beta_{0,1})(\hat{\beta}_2 - \beta_{0,2})\},
\end{aligned}
$$

and the two explicit equations for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$

$$
\begin{cases}
\hat{\beta}_1 - \beta_{0,1} + \rho(\hat{\beta}_2 - \beta_{0,2}) - \tau\xi_1 + \alpha_0\tau^{2-q}q|\hat{\beta}_1|^{q-1}sign(\hat{\beta}_1) &= 0, \\
\rho(\hat{\beta}_1 - \beta_{0,1}) + \hat{\beta}_2 - \beta_{0,2} - \tau\xi_2 + \alpha_0\tau^{2-q}q|\hat{\beta}_2|^{q-1}sign(\hat{\beta}_2) &= 0,
\end{cases}
$$

where $\xi_1 = \rho_1 z_1 + \rho_2 z_2$, $\xi_2 = \rho_2 z_1 + \rho_1 z_2$.

We consider three different cases according to the nonzero components of $\beta_{0,1}$ and $\beta_{0,2}$. In the first case: $\beta_{0,1} \neq 0$ and $\beta_{0,2} \neq 0$. Using (A49), we obtain its contribution to the derivative of $R(\tau^2, \alpha_0)$ as

$$
\begin{aligned}
R_1'(\tau^2, \alpha_0) = \epsilon^2\tau^2 E \Big\{ &-2q\tau^{1-q}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}|\boldsymbol{\beta}|^{q-1}sign(\boldsymbol{\beta}) - 2q(q-1)\alpha_0\tau^{1-q}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}|\boldsymbol{\beta}|^{q-2}sign(\boldsymbol{\beta})\frac{\partial\boldsymbol{\beta}}{\partial\alpha_0} \\
&+2q^2\alpha_0\tau^{2-2q}(|\boldsymbol{\beta}|^{q-1}sign(\boldsymbol{\beta}))^T\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}|^{q-1}sign(\boldsymbol{\beta}) \\
&+2q^2(q-1)\alpha_0^2\tau^{2-2q}(|\boldsymbol{\beta}|^{q-1}sign(\boldsymbol{\beta}))^T\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}|^{q-2}sign(\boldsymbol{\beta})\frac{\partial\boldsymbol{\beta}}{\partial\alpha_0} \Big\}. \quad (A50)
\end{aligned}
$$

Since both components of $\boldsymbol{\beta}_0$ are nonzero, as $\tau \to 0$, (A47) can be written as

$$
\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \tau\boldsymbol{\Sigma}^{-1/2}\mathbf{z} - p\alpha_0\tau^{2-q}\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}_0|^{q-1}sign(\boldsymbol{\beta}_0) + o_p(\tau^q).
$$

Thus, using (A48), we obtain the two leading terms of (A51) as

$$
\begin{aligned}
R_1'(\tau^2, \alpha_0) = \frac{1}{2}\epsilon^2\tau^2 E\{ &-2q(q-1)\tau^{2-q}Tr[\boldsymbol{\Sigma}^{-1}diag\{|\boldsymbol{\beta}_0|^{q-2}sign(\boldsymbol{\beta}_0)\}] \\
&+2q^2\alpha_0\tau^{2-2q}(|\boldsymbol{\beta}_0|^{q-1}sign(\boldsymbol{\beta}_0))^T\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}_0|^{q-1}sign(\boldsymbol{\beta}_0)\}. \quad (A51)
\end{aligned}
$$

In the second case, $\beta_{0,1} \neq 0$ and $\beta_{0,2} = 0$ (or $\beta_{0,1} = 0$ and $\beta_{0,2} \neq 0$). We have

$$
\begin{aligned}
R_2(\tau^2, \alpha_0) &= \frac{1}{2}\epsilon(1 - \epsilon)E\{(\hat{\beta}_1 - \beta_{0,1})^2 + \hat{\beta}_2^2 + 2\rho(\hat{\beta}_1 - \beta_{0,1})\hat{\beta}_2\} \\
&= \frac{1}{2}\epsilon(1 - \epsilon)E\{\hat{\beta}_2^2 + [\alpha_0\tau^{2-q}q|\hat{\beta}_1|^{q-1}sign(\hat{\beta}_1) - \tau\xi_1 - \rho\hat{\beta}_2][\alpha_0\tau^{2-q}q|\hat{\beta}_1|^{q-1}sign(\hat{\beta}_1) - \tau\xi_1 + \rho\hat{\beta}_2]\} \\
&= \frac{1}{2}\epsilon(1 - \epsilon)E\{(1 - \rho^2)\hat{\beta}_2^2 + [\alpha_0\tau^{2-q}q|\hat{\beta}_1|^{q-1}sign(\hat{\beta}_1) - \tau\xi_1]^2\} \\
&= \frac{1}{2}\epsilon(1 - \epsilon)E\{(1 - \rho^2)\hat{\beta}_2^2 + \alpha_0^2\tau^{4-2q}q^2|\hat{\beta}_1|^{2q-2} + \tau^2\xi_1^2 - 2\alpha_0\tau^{3-q}q\xi_1|\hat{\beta}_1|^{q-1}sign(\hat{\beta}_1)\}. \quad (A52)
\end{aligned}
$$

31

Taking derivative over α_0, using (A48), we obtain the dominant terms as

$$
\begin{aligned}
R_2'(\tau^2, \alpha_0) &= \epsilon(1-\epsilon)E\{(1-\rho^2)\hat{\beta}_2\frac{\partial\hat{\beta}_2}{\partial\alpha_0} + q^2\alpha_0\tau^{4-2q}|\beta_{0,1}|^{2q-2}\} \\
&= -\epsilon(1-\epsilon)q\tau^{2-q}E\{|\hat{\beta}_2|^q - \rho\hat{\beta}_2|\beta_{0,1}|^{q-1}sign(\beta_{0,1}) - q\alpha_0\tau^{2-q}|\beta_{0,1}|^{2q-2}\} \\
&= -\epsilon(1-\epsilon)q\tau^{2-q}E\{|\hat{\beta}_2|^q - \frac{q\alpha_0\tau^{2-q}\rho^2|\beta_{0,1}|^{2q-2}}{1-\rho^2} - q\alpha_0\tau^{2-q}|\beta_{0,1}|^{2q-2}\} \\
&= -\epsilon(1-\epsilon)q\tau^2 E\left\{\frac{|z|^q}{(1-\rho^2)^{q/2}} - \frac{q\alpha_0\tau^{2-2q}|\beta_{0,1}|^{2q-2}}{1-\rho^2}\right\}. \quad (A53)
\end{aligned}
$$

In the third case, $\beta_{0,1} = 0$ and $\beta_{0,2} = 0$. Using (A47), we obtain

$$
\hat{\boldsymbol{\beta}} = \tau\boldsymbol{\Sigma}^{-1/2}\mathbf{z} - q\alpha_0\tau^{2-q}\boldsymbol{\Sigma}^{-1}|\hat{\boldsymbol{\beta}}|^{q-1}sign(\hat{\boldsymbol{\beta}}).
$$

Its contribution to risk function is

$$
R_3(\tau^2, \alpha_0) = \frac{1}{2}(1-\epsilon)^2 E\|\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}^2. \quad (A54)
$$

Taking derivative over α_0, we obtain the dominant term as

$$
R_3'(\tau^2, \alpha_0) = (1-\epsilon)^2 E\hat{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\alpha_0} = -(1-\epsilon)^2 q\tau^{2-q}E\langle|\hat{\boldsymbol{\beta}}|^q\rangle = -\frac{2(1-\epsilon)^2 q\tau^2}{(1-\rho^2)^{q/2}}E\{|z|^q\}. \quad (A55)
$$

Combining (A51), (A53), and (A55) together, we obtain

$$
\begin{aligned}
&R_1'(\tau^2, \alpha_0) + 2R_2'(\tau^2, \alpha_0) + R_3'(\tau^2, \alpha_0) \\
&= -\frac{2(1-\epsilon)q\tau^2}{(1-\rho^2)^{q/2}}E\{|z|^q\} + \frac{2\epsilon(1-\epsilon)q^2\tau^2\alpha_0\tau^{2-2q}E\{|\beta_0|^{2q-2}\}}{2(1-\rho^2)} \\
&\quad + \frac{2\epsilon^2 q^2\tau^2\alpha_0\tau^{2-2q}}{1-\rho^2}\left[E\{|\beta_0|^{2q-2}\} - \rho(E|\beta_0|^{q-1}sign(\beta_0))^2\right].
\end{aligned}
$$

Therefore, the optimal tuning $\alpha_{0\star}$ satisfies

$$
\alpha_{0\star}\tau^{2-2q} = \frac{(1-\epsilon)E(|Z|^q)(1-\rho^2)^{1-q/2}}{\epsilon q\{E(|\beta_0|^{2q-2}) - \epsilon\rho[E|\beta_0|^{q-1}sign(\beta_0)]^2\}}, \quad (A56)
$$

where $Z \sim N(0,1)$ is independent of β_0. Now we can obtain the dominant terms in the corresponding R functions. From (A49), we obtain

$$
R_1(\tau^2, \alpha_{0\star}) = \tau^2\epsilon^2 E\left\{1 + \frac{1}{2}q^2\alpha_{0\star}^2\tau^{2-2q}(|\boldsymbol{\beta}_0|^{q-1}sign(\boldsymbol{\beta}_0))^T\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}_0|^{q-1}sign(\boldsymbol{\beta}_0)\right\}.
$$

From (A52), we obtain

$$
R_2(\tau^2, \alpha_{0\star}) = 2\epsilon(1-\epsilon)\tau^2 E\left\{1 - q\alpha_{0\star}\tau^{-q}|\hat{\beta}_2|^q + \frac{q^2\alpha_{0\star}^2\tau^{2-2q}|\beta_0|^{2q-2}}{1-\rho^2}\right\}.
$$

32

From (A54), we obtain

$$
\begin{aligned}
R_3(\tau^2, \alpha_{0\star}) &= (1-\epsilon)^2 \tau^2 E\left\{ \frac{1}{2}\|\mathbf{z}\|^2 - q\alpha_{0\star}\tau^{1-q}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}|\hat{\boldsymbol{\beta}}|^{q-1} sign(\hat{\boldsymbol{\beta}}) \right\} \\
&= (1-\epsilon)^2 \tau^2 E\left\{ \frac{1}{2}\|\mathbf{z}\|^2 - q\alpha_{0\star}\tau^{-q}\langle|\hat{\boldsymbol{\beta}}|^q\rangle \right\} \\
&= (1-\epsilon)^2 \tau^2 \left(1 - \frac{q\alpha_{0\star}}{(1-\rho^2)^{q/2}} E\left\{\langle|\mathbf{z}|^q\rangle\right\} \right).
\end{aligned}
$$

Putting them together, we have

$$
R(\tau^2, \alpha_{0\star}) = \tau^2 \left(1 + \frac{q^2\alpha_{0\star}^2\tau^{2-2q}}{1-\rho^2}(E|\beta_0|^{2q-2} - \epsilon\rho[E|\beta_0|^{q-1}sign(\beta_0)]^2) \right). \qquad \text{(A57)}
$$

Substituting (A56) into (A57), we have

$$
R(\tau^2, \alpha_{0\star}) = \tau^2 + C\tau^{2q},
$$

where

$$
C = \frac{(1-\epsilon)^2}{\epsilon} \frac{[E(|Z|^q)]^2(1-\rho^2)^{1-q}\tau^{2q}}{\{E(|\beta_0|^{2q-2}) - \epsilon\rho[E|\beta_0|^{q-1}sign(\beta_0)]^2\}}.
$$

Therefore, the fixed point equation is

$$
\tau^2 = \sigma_w^2 + \frac{\tau^2 + C\tau^{2q}}{\delta}.
$$

The first and second dominant terms are

$$
\tau^2 = \frac{\delta}{\delta-1}\sigma_w^2 - \frac{C\tau^{2q}}{\delta-1}.
$$

We obtain the prediction risk

$$
\begin{aligned}
Risk &= \delta(\tau^2 - \sigma_w^2) = \frac{\delta\sigma_w^2}{\delta-1} - \frac{C\delta\tau^{2q}}{\delta-1} \\
&= \frac{\delta\sigma_w^2}{\delta-1} - \frac{\delta^{q+1}}{(\delta-1)^{q+1}}\frac{(1-\epsilon)^2}{\epsilon}\frac{[E(|Z|^q)]^2(1-\rho^2)^{1-q}\sigma_w^{2q}}{\{E(|\beta_0|^{2q-2}) - \epsilon\rho[E|\beta_0|^{q-1}sign(\beta_0)]^2\}}.
\end{aligned}
$$

References

Amelunxen, D., M. Lotz, M. B. McCoy, and J. A. Tropp (2013). Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA 3*, 224–294.

Baddeley, A. (1977). Integrals on a moving manifold and geometrical probability. *Advances in Applied Probability 9*(3), 588–603.

Celentano, M., A. Montanari, and Y. Wei (2020). The lasso with general gaussian designs with applications to hypothesis testing. *arXiv*, 10.48550/ARXIV.2007.13716.

Chartrand, R. and V. Staneva (2008, may). Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems 24*(3), 035020.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics 59*(6), 797–829.

Donoho, D. L., A. Maleki, and A. Montanari (2011, Oct). The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory 57*(10), 6920–6941.

Donoho, D. L. and J. Tanner (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences 102*(27), 9446–9451.

Fu, W. and K. Knight (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356 – 1378.

Huang, H. (2021). Lasso risk and phase transition under dependence. *arXiv*, 10.48550/ARXIV.2103.16035.

Javanmard, A. and A. Montanari (2014, Oct). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Kabashima, Y., T. Wadayama, and T. Tanaka (2009, sep). A typical reconstruction limit for compressed sensing based on ℓ_p-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment 2009*(09), L09003.

Ma, J., J. Xu, and A. Maleki (2019). Optimization-based amp for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Transactions on Information Theory 65*(6), 3600–3629.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics 34*(3), 1436 – 1462.

Mézard, M. and A. Montanari (2009). *Information, Physics, and Computation.* Oxford Graduate Texts. OUP Oxford.

Mezard, M., G. Parisi, and M. Virasoro (1987). *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications.* World Scientific Lecture Notes in Physics. World Scientific.

Rangan, S., V. Goyal, and A. K. Fletcher (2009). Asymptotic analysis of map estimation via the replica method and compressed sensing. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, Volume 22. Curran Associates, Inc.

Saab, R., R. Chartrand, and O. Yilmaz (2008). Stable sparse approximations via nonconvex optimization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3885–3888.

Saab, R. and Özgür Yılmaz (2010). Sparse recovery by non-convex optimization – instance optimality. *Applied and Computational Harmonic Analysis 29*(1), 30–48.

Stojnic, M. (2013). Lifting ℓ_q-optimization thresholds. *arXiv*, https://arxiv.org/abs/1306.3976.

Thrampoulidis, C., E. Abbasi, and B. Hassibi (2018). Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory 64*(8), 5592–5628.

Wang, M., W. Xu, and A. Tang (2011). On the performance of sparse recovery via ℓ_p-minimization ($0 \le p \le 1$). *IEEE Transactions on Information Theory 57*(11), 7255–7278.

Wang, S., H. Weng, and A. Maleki (2020). Which bridge estimator is the best for variable selection? *The Annals of Statistics 48*(5), 2791 – 2823.

Wang, S., H. Weng, and A. Maleki (2021). Does SLOPE outperform bridge regression? *Information and Inference: A Journal of the IMA 11*(1), 1–54.

Weng, H. and A. Maleki (2019). Low noise sensitivity analysis of -minimization in oversampled systems. *Information and Inference: A Journal of the IMA 9*(1), 113–155.

Weng, H., A. Maleki, and L. Zheng (2018). Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *The Annals of Statistics 46*(6A), 3099 – 3129.

Weng, H., L. Zheng, A. Maleki, and X. Wang (2016). Phase transition and noise sensitivity of l_p-minimization for $0 \le p \le 1$. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 675–679.

Zheng, L., A. Maleki, H. Weng, X. Wang, and T. Long (2017). Does ℓ_p-minimization outperform ℓ_1-minimization? *IEEE Transactions on Information Theory 63*(11), 6896–6935.