2-2016

Accuracy Assessment for High-Dimensional Linear Regression

Tony Cai
University of Pennsylvania

Zijian Guo
University of Pennsylvania

Accuracy Assessment for High-Dimensional Linear Regression

Abstract
This paper considers point and interval estimation of the ℓ_q loss of an estimator in high-dimensional linear regression with random design. We establish the minimax rate for estimating the ℓ_q loss and the minimax expected length of confidence intervals for the ℓ_q loss of rate-optimal estimators of the regression vector, including commonly used estimators such as Lasso, scaled Lasso, square-root Lasso and Dantzig Selector. Adaptivity of the confidence intervals for the ℓ_q loss is also studied. Both the setting of known identity design covariance matrix and known noise level and the setting of unknown design covariance matrix and unknown noise level are studied. The results reveal interesting and significant differences between estimating the ℓ_2 loss and ℓ_q loss with $1 \le q < 2$ as well as between the two settings. New technical tools are developed to establish rate sharp lower bounds for the minimax estimation error and the expected length of minimax and adaptive confidence intervals for the ℓ_q loss. A significant difference between loss estimation and the traditional parameter estimation is that for loss estimation the constraint is on the performance of the estimator of the regression vector, but the lower bounds are on the difficulty of estimating its ℓ_q loss. The technical tools developed in this paper can also be of independent interest.

Keywords
Accuracy assessment, adaptivity, confidence interval, highdimensional linear regression, loss estimation, minimax lower bound, minimaxity, sparsity

Disciplines
Physical Sciences and Mathematics

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

ACCURACY ASSESSMENT FOR HIGH-DIMENSIONAL LINEAR REGRESSION*

By T. Tony Cai, and Zijian Guo

University of Pennsylvania

This paper considers point and interval estimation of the ℓ_q loss of an estimator in high-dimensional linear regression with random design. We establish the minimax rate for estimating the ℓ_q loss and the minimax expected length of confidence intervals for the ℓ_q loss of rate-optimal estimators of the regression vector, including commonly used estimators such as Lasso, scaled Lasso, square-root Lasso and Dantzig Selector. Adaptivity of confidence intervals for the ℓ_q loss is also studied. Both the setting of known identity design covariance matrix and known noise level and the setting of unknown design covariance matrix and unknown noise level are studied. The results reveal interesting and significant differences between estimating the ℓ_2 loss and ℓ_q loss with $1 \leq q < 2$ as well as between the two settings.

New technical tools are developed to establish rate sharp lower bounds for the minimax estimation error and the expected length of minimax and adaptive confidence intervals for the ℓ_q loss. A significant difference between loss estimation and the traditional parameter estimation is that for loss estimation the constraint is on the performance of the estimator of the regression vector, but the lower bounds are on the difficulty of estimating its ℓ_q loss. The technical tools developed in this paper can also be of independent interest.

1. Introduction. In many applications, the goal of statistical inference is not only to construct a good estimator, but also to provide a measure of accuracy for this estimator. In classical statistics, when the parameter of interest is one-dimensional, this is achieved in the form of a standard error or a confidence interval. A prototypical example is the inference for a binomial proportion, where often not only an estimate of the proportion but also its margin of error are given. Accuracy measures of an estimation procedure have also been used as a tool for the empirical selection of tuning parameters. A well known example is Stein's Unbiased Risk Estimate (SURE), which has been an effective tool for the construction of data-driven adaptive estimators in normal means estimation, nonparametric signal recovery, covariance

*The research was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

MSC 2010 subject classifications: Primary 62G15; secondary 62C20, 62H35

Keywords and phrases: Accuracy assessment, adaptivity, confidence interval, high-dimensional linear regression, loss estimation, minimax lower bound, minimaxity, sparsity.

matrix estimation, and other problems. See, for instance, [25, 21, 15, 11, 32]. The commonly used cross-validation methods can also be viewed as a useful tool based on the idea of empirical assessment of accuracy.

In this paper, we consider the problem of estimating the loss of a given estimator in the setting of high-dimensional linear regression, where one observes (X, y) with $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$, and for $1 \le i \le n$,

$$y_i = X_{i\cdot}\beta + \epsilon_i.$$

Here $\beta \in \mathbb{R}^p$ is the regression vector, $X_{i\cdot} \overset{iid}{\sim} N_p(0, \Sigma)$ are the rows of X, and the errors $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ are independent of X. This high-dimensional linear model has been well studied in the literature, where the main focus has been on estimation of β. Several penalized/constrained ℓ_1 minimization methods, including Lasso [28], Dantzig selector [12], scaled Lasso [26] and square-root Lasso [3], have been proposed. These methods have been shown to work well in applications and produce interpretable estimates of β when β is assumed to be sparse. Theoretically, with a properly chosen tuning parameter, these estimators achieve the optimal rate of convergence over collections of sparse parameter spaces. See, for example, [12, 26, 3, 23, 4, 5, 30].

For a given estimator $\widehat{\beta}$, the ℓ_q loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \le q \le 2$ is commonly used as a metric of accuracy for $\widehat{\beta}$. We consider in the present paper both point and interval estimation of the ℓ_q loss $\|\widehat{\beta} - \beta\|_q^2$ for a given $\widehat{\beta}$. Note that the loss $\|\widehat{\beta} - \beta\|_q^2$ is a random quantity, depending on both the estimator $\widehat{\beta}$ and the parameter β. For such a random quantity, prediction and prediction interval are ususally used for point and interval estimation, respectively. However, we slightly abuse the terminologies in the present paper by using estimation and confidence interval to represent the point and interval estimators of the loss $\|\widehat{\beta} - \beta\|_q^2$. Since the ℓ_q loss depends on the estimator $\widehat{\beta}$, it is necessary to specify the estimator in the discussion of loss estimation. Throughout this paper, we restrict our attention to a broad collection of estimators $\widehat{\beta}$ that perform well at least at one interior point or a small subset of the parameter space. This collection of estimators includes most state-of-art estimators such as Lasso, Dantzig selector, scaled Lasso and square-root Lasso.

High-dimensional linear regression has been well studied in two settings. One is the setting with known design covariance matrix $\Sigma = \mathrm{I}$, known noise level $\sigma = \sigma_0$ and sparse β. See for example, [16, 2, 22, 30, 27, 7, 1, 19]. Another commonly considered setting is sparse β with unknown Σ and σ. We study point and interval estimation of the ℓ_q loss $\|\widehat{\beta} - \beta\|_q^2$ in both settings. Specifically, we consider the parameter space $\Theta_0(k)$ introduced in

(2.3), which consists of k-sparse signals β with known design covariance matrix $\Sigma = \mathrm{I}$ and known noise level $\sigma = \sigma_0$, and $\Theta(k)$ defined in (2.4), which consists of k-sparse signals with unknown Σ and σ.

1.1. *Our contributions.* The present paper studies the minimax and adaptive estimation of the loss $\|\widehat{\beta} - \beta\|_q^2$ for a given estimator $\widehat{\beta}$ and the minimax expected length and adaptivity of confidence intervals for the loss. A major step in our analysis is to establish rate sharp lower bounds for the minimax estimation error and the minimax expected length of confidence intervals for the ℓ_q loss over $\Theta_0(k)$ and $\Theta(k)$ for a broad class of estimators of β, which contains the subclass of rate-optimal estimators. We then focus on the estimation of the loss of rate-optimal estimators and take the Lasso and scaled Lasso estimators as generic examples. For these rate-optimal estimators, we propose procedures for point estimation as well as confidence intervals for their ℓ_q losses. It is shown that the proposed procedures achieve the corresponding lower bounds up to a constant factor. These results together establish the minimax rates for estimating the ℓ_q loss of rate-optimal estimators over $\Theta_0(k)$ and $\Theta(k)$. The analysis shows interesting and significant differences between estimating the ℓ_2 loss and ℓ_q loss with $1 \le q < 2$ as well as between the two parameter spaces $\Theta(k)$ and $\Theta_0(k)$.

- The minimax rate for estimating $\|\widehat{\beta} - \beta\|_2^2$ over $\Theta_0(k)$ is $\min\left\{\frac{1}{\sqrt{n}}, k\frac{\log p}{n}\right\}$ and over $\Theta(k)$ is $k\frac{\log p}{n}$. So loss estimation is much easier with the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$ when $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$.
- The minimax rate for estimating $\|\widehat{\beta} - \beta\|_q^2$ with $1 \le q < 2$ over both $\Theta_0(k)$ and $\Theta(k)$ is $k^{\frac{2}{q}}\frac{\log p}{n}$.

In the regime $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$, a practical loss estimator is proposed for estimating the ℓ_2 loss and shown to achieve the optimal convergence rate $\frac{1}{\sqrt{n}}$ adaptively over $\Theta_0(k)$. We say *estimation of loss is impossible* if the minimax rate can be achieved by the trivial estimator 0, which means that the estimation accuracy of the loss is at least of the same order as the loss itself. In all other considered cases, estimation of loss is shown to be impossible. These results indicate that loss estimation is difficult.

We then turn to the construction of confidence intervals for the ℓ_q loss. A confidence interval for the loss is useful even when it is "impossible" to estimate the loss, as a confidence interval can provide non-trivial upper and lower bounds for the loss. In terms of convergence rate over $\Theta_0(k)$ or $\Theta(k)$, the minimax rate of the expected length of confidence intervals for the ℓ_q loss, $\|\widehat{\beta} - \beta\|_q^2$, of any rate-optimal estimator $\widehat{\beta}$ coincides with

the minimax estimation rate. We also consider the adaptivity of confidence intervals for the ℓ_q loss of any rate-optimal estimator $\widehat{\beta}$. (The framework for adaptive confidence intervals is discussed in detail in Section 3.1.) Regarding confidence intervals for the ℓ_2 loss in the case of known $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$, a procedure is proposed and is shown to achieve the optimal length $\frac{1}{\sqrt{n}}$ adaptively over $\Theta_0(k)$ for $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$. Furthermore, it is shown that this is the only regime where adaptive confidence intervals exist, even over two given parameter spaces. For example, when $k_1 \ll \frac{\sqrt{n}}{\log p}$ and $k_1 \ll k_2$, it is impossible to construct a confidence interval for the ℓ_2 loss with guaranteed coverage probability over $\Theta_0(k_2)$ (consequently also over $\Theta_0(k_1)$) and with the expected length automatically adjusted to the sparsity. Similarly, for the ℓ_q loss with $1 \le q < 2$, construction of adaptive confidence intervals is impossible over $\Theta_0(k_1)$ and $\Theta_0(k_2)$ for $k_1 \ll k_2 \lesssim \frac{n}{\log p}$. Regarding confidence intervals for the ℓ_q loss with $1 \le q \le 2$ in the case of unknown Σ and σ, the impossibility of adaptivity also holds over $\Theta(k_1)$ and $\Theta(k_2)$ for $k_1 \ll k_2 \lesssim \frac{n}{\log p}$.

Establishing rate-optimal lower bounds requires the development of new technical tools. One main difference between loss estimation and the traditional parameter estimation is that for loss estimation the constraint is on the performance of the estimator $\widehat{\beta}$ of the regression vector β, but the lower bound is on the difficulty of estimating its loss $\|\widehat{\beta} - \beta\|_q^2$. We introduce useful new lower bound techniques for the minimax estimation error and the expected length of adaptive confidence intervals for the loss $\|\widehat{\beta} - \beta\|_q^2$. In several important cases, it is necessary to test a composite null against a composite alternative in order to establish rate sharp lower bounds. The technical tools developed in this paper can also be of independent interest.

In addition to $\Theta_0(k)$ and $\Theta(k)$, we also study an intermediate parameter space where the noise level σ is known and the design covariance matrix Σ is unknown but of certain structure. Lower bounds for the expected length of minimax and adaptive confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ over this parameter space are established for a broad collection of estimators $\widehat{\beta}$ and are shown to be rate sharp for the class of rate-optimal estimators. Furthermore, the lower bounds developed in this paper have wider implications. In particular, it is shown that they lead immediately to minimax lower bounds for estimating $\|\beta\|_q^2$ and the expected length of confidence intervals for $\|\beta\|_q^2$ with $1 \le q \le 2$.

1.2. *Comparison with other works.* Statistical inference on the loss of specific estimators of β has been considered in the recent literature. The papers [16, 2] established, in the setting $\Sigma = \mathrm{I}$ and $n/p \to \delta \in (0, \infty)$, the limit of the normalized loss $\frac{1}{p}\|\widehat{\beta}(\lambda) - \beta\|_2^2$ where $\widehat{\beta}(\lambda)$ is the Lasso estima-

tor with a pre-specified tuning parameter λ. Although [16, 2] provided an exact asymptotic expression of the normalized loss, the limit itself depends on the unknown β. In a similar setting, the paper [27] established the limit of a normalized ℓ_2 loss of the square-root Lasso estimator. These limits of the normalized losses help understand the properties of the corresponding estimators of β, but they do not lead to an estimate of the loss. Our results imply that although these normalized losses have a limit under certain regularity conditions, such losses cannot be estimated well in most settings.

A recent paper, [20], constructed a confidence interval for $\|\widehat{\beta} - \beta\|_2^2$ in the case of known $\Sigma = \mathrm{I}$, unknown noise level σ, and moderate dimension where $n/p \to \xi \in (0, 1)$ and no sparsity is assumed on β. While no sparsity assumption on β is imposed, their method requires the assumption of $\Sigma = \mathrm{I}$ and $n/p \to \xi \in (0, 1)$. In contrast, in this paper, we consider both unknown Σ and known $\Sigma = \mathrm{I}$ settings, while allowing $p \gg n$ and assuming sparse β.

Honest adaptive inference has been studied in the nonparametric function estimation literature, including [8] for adaptive confidence intervals for linear functionals, [18, 10] for adaptive confidence bands, and [9, 24] for adaptive confidence balls, and in the high-dimensional linear regression literature, including [22] for adaptive confidence set and [7] for adaptive confidence interval for linear functionals. In this paper, we develop new lower bound tools, Theorems 8 and 9, to establish the possibility of adaptive confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$. The connection between ℓ_2 loss considered in the current paper and the work [22] is discussed in more detail in Section 3.2.

1.3. *Organization.* Section 2 establishes the minimax lower bounds of estimating the loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \le q \le 2$ over both $\Theta_0(k)$ and $\Theta(k)$ and shows that these bounds are rate sharp for the Lasso and scaled Lasso estimators, respectively. We then turn to interval estimation of $\|\widehat{\beta} - \beta\|_q^2$. Sections 3 and 4 present the minimax and adaptive minimax lower bounds for the expected length of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ over $\Theta_0(k)$ and $\Theta(k)$. For Lasso and scaled Lasso estimators, we show that the lower bounds can be achieved and investigate the possibility of adaptivity. Section 5 considers the rate-optimal estimators and establishes the minimax convergence rate of estimating their ℓ_q losses. Section 6 presents new minimax lower bound techniques for estimating the loss $\|\widehat{\beta} - \beta\|_q^2$. Section 7 discusses the minimaxity and adaptivity in another setting, where the noise level σ is known and the design covariance matrix Σ is unknown but of certain structure. Section 8 applies the newly developed lower bounds to establish lower bounds for a related problem, that of estimating $\|\beta\|_q^2$. Section 9 proves the main results and additional proofs are given in the supplemental material [6].

1.4. *Notation.* For a matrix $X \in \mathbb{R}^{n\times p}$, $X_{i\cdot}$, $X_{\cdot j}$, and $X_{i,j}$ denote respectively the i-th row, j-th column, and (i,j) entry of the matrix X. For a subset $J \subset \{1,2,\cdots,p\}$, $|J|$ denotes the cardinality of J, J^c denotes the complement $\{1,2,\cdots,p\}\backslash J$, X_J denotes the submatrix of X consisting of columns $X_{\cdot j}$ with $j \in J$ and for a vector $x \in \mathbb{R}^p$, x_J is the subvector of x with indices in J. For a vector $x \in \mathbb{R}^p$, $\mathrm{supp}(x)$ denotes the support of x and the ℓ_q norm of x is defined as $\|x\|_q = (\sum_{i=1}^p |x_i|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\mathrm{supp}(x)|$ and $\|x\|_\infty = \max_{1\leq j\leq p} |x_j|$. For $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$. We use $\max \|X_{\cdot j}\|_2$ as a shorthand for $\max_{1\leq j\leq p} \|X_{\cdot j}\|_2$ and $\min \|X_{\cdot j}\|_2$ as a shorthand for $\min_{1\leq j\leq p} \|X_{\cdot j}\|_2$. For a matrix A, we define the spectral norm $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$ and the matrix ℓ_1 norm $\|A\|_{L_1} = \sup_{1\leq j\leq p} \sum_{i=1}^p |A_{ij}|$; For a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the smallest and largest eigenvalue of A. We use c and C to denote generic positive constants that may vary from place to place. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \leq Cb_n$ for all n and $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\to\infty} \frac{a_n}{b_n} = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$.

2. Minimax estimation of the ℓ_q loss.

We begin by presenting the minimax framework for estimating the ℓ_q loss, $\|\widehat{\beta} - \beta\|_q^2$, of a given estimator $\widehat{\beta}$, and then establish the minimax lower bounds for the estimation error for a broad collection of estimators $\widehat{\beta}$. We also show that such minimax lower bounds can be achieved for the Lasso and scaled Lasso estimators.

2.1. *Problem formulation.* Recall the high-dimensional linear model,

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N_n(0, \sigma^2 \mathrm{I}). \tag{2.1}$$

We focus on the random design with $X_{i\cdot} \overset{iid}{\sim} N(0, \Sigma)$ and $X_{i\cdot}$ and ϵ_i are independent. Let $Z = (X, y)$ denote the observed data and $\widehat{\beta}$ be a given estimator of β. Denoting by $\widehat{L}_q(Z)$ any estimator of the loss $\|\widehat{\beta} - \beta\|_q^2$, the minimax rate of convergence for estimating $\|\widehat{\beta} - \beta\|_q^2$ over a parameter space Θ is defined as the largest quantity $\gamma_{\widehat{\beta},\ell_q}(\Theta)$ such that

$$\inf_{\widehat{L}_q} \sup_{\theta\in\Theta} \mathbb{P}_\theta \left(|\widehat{L}_q(Z) - \|\widehat{\beta} - \beta\|_q^2| \geq \gamma_{\widehat{\beta},\ell_q}(\Theta) \right) \geq \delta, \tag{2.2}$$

for some constant $\delta > 0$ not depending on n or p. We shall write $\widehat{L}_q$ for $\widehat{L}_q(Z)$ when there is no confusion.

We denote the parameter by $\theta = (\beta, \Sigma, \sigma)$, which consists of the signal β, the design covariance matrix Σ and the noise level σ. For a given

$\theta = (\beta, \Sigma, \sigma)$, we use $\beta(\theta)$ to denote the corresponding β. Two settings are considered: The first is known design covariance matrix $\Sigma = \mathrm{I}$ and known noise level $\sigma = \sigma_0$ and the other is unknown Σ and σ. In the first setting, we consider the following parameter space that consists of k-sparse signals,

$$\Theta_0(k) = \{(\beta, \mathrm{I}, \sigma_0) : \|\beta\|_0 \le k\}, \tag{2.3}$$

and in the second setting, we consider

$$\Theta(k) = \left\{(\beta, \Sigma, \sigma) : \|\beta\|_0 \le k,\ \frac{1}{M_1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le M_1,\ 0 < \sigma \le M_2\right\}, \tag{2.4}$$

where $M_1 \ge 1$ and $M_2 > 0$ are constants. The parameter space $\Theta_0(k)$ is a subset of $\Theta(k)$, which consists of k-sparse signals with unknown Σ and σ.

The minimax rate $\gamma_{\widehat{\beta},\ell_q}(\Theta)$ for estimating $\|\widehat{\beta} - \beta\|_q^2$ also depends on the estimator $\widehat{\beta}$. Different estimators $\widehat{\beta}$ could lead to different losses $\|\widehat{\beta} - \beta\|_q^2$ and in general the difficulty of estimating the loss $\|\widehat{\beta} - \beta\|_q^2$ varies with $\widehat{\beta}$. We first recall the properties of some state-of-art estimators and then specify the collection of estimators on which we focus in this paper. As shown in [12, 4, 3, 26], Lasso, Dantzig Selector, scaled Lasso and square-root Lasso satisfy the following property if the tuning parameter is properly chosen,

$$\sup_{\theta \in \Theta(k)} \mathbb{P}_\theta\left(\|\widehat{\beta} - \beta\|_q^2 \ge C k^{\frac{2}{q}} \frac{\log p}{n}\right) \to 0, \tag{2.5}$$

where $C > 0$ is a constant. The minimax lower bounds established in [30, 23, 31] imply that $k^{\frac{2}{q}} \frac{\log p}{n}$ is the optimal rate for estimating β over the parameter space $\Theta(k)$. It should be stressed that all of these algorithms do not require knowledge of the sparsity k and are thus adaptive to the sparsity provided $k \lesssim \frac{n}{\log p}$. We consider a broad collection of estimators $\widehat{\beta}$ satisfying one of the following two assumptions.

(A1) The estimator $\widehat{\beta}$ satisfies, for some $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$,

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\beta} - \beta^*\|_q^2 \ge C^* \|\beta^*\|_0^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2\right) \le \alpha_0, \tag{2.6}$$

where $0 \le \alpha_0 < \frac{1}{4}$ and $C^* > 0$ are constants.

(A2) The estimator $\widehat{\beta}$ satisfies

$$\sup_{\{\theta = (\beta^*, \mathrm{I}, \sigma) : \sigma \le 2\sigma_0\}} \mathbb{P}_\theta\left(\|\widehat{\beta} - \beta^*\|_q^2 \ge C^* \|\beta^*\|_0^{\frac{2}{q}} \frac{\log p}{n} \sigma^2\right) \le \alpha_0, \tag{2.7}$$

where $0 \le \alpha_0 < \frac{1}{4}$ and $C^* > 0$ are constants and $\sigma_0 > 0$ is given.

In view of the minimax rate given in (2.5), Assumption (A1) requires $\widehat{\beta}$ to be a good estimator of β at at least one point $\theta_0 \in \Theta_0(k)$. Assumption (A2) is slightly stronger than (A1) and requires $\widehat{\beta}$ to estimate β well for a single β^* but over a range of noise levels $\sigma \le 2\sigma_0$ while $\Sigma = \mathrm{I}$. Of course, any estimator $\widehat{\beta}$ satisfying (2.5) satisfies both (A1) and (A2). In addition to Assumptions (A1) and (A2), we also introduce the following sparsity assumptions that will be used in various theorems.

(B1) Let c_0 be the constant defined in (9.14). The sparsity levels k and k_0 satisfy $k \le c_0 \min\{p^\gamma, \frac{n}{\log p}\}$ for some constant $0 \le \gamma < \frac{1}{2}$ and $k_0 \le c_0 \min\{k, \frac{\sqrt{n}}{\log p}\}$.

(B2) The sparsity levels k_1, k_2 and k_0 satisfy $k_1 \le k_2 \le c_0 \min\{p^\gamma, \frac{n}{\log p}\}$ for some constant $0 \le \gamma < \frac{1}{2}$ and $c_0 > 0$ and $k_0 \le c_0 \min\{k_1, \frac{\sqrt{n}}{\log p}\}$.

2.2. *Minimax estimation of the ℓ_q loss over $\Theta_0(k)$.* The following theorem establishes the minimax lower bounds for estimating the loss $\|\widehat{\beta} - \beta\|_q^2$ over the parameter space $\Theta_0(k)$.

THEOREM 1. *Suppose that the sparsity levels k and k_0 satisfy Assumption* (B1)*. For any estimator $\widehat{\beta}$ satisfying Assumption* (A1) *with $\|\beta^*\|_0 \le k_0$,*

$$\inf_{\widehat{L}_2} \sup_{\theta \in \Theta_0(k)} \mathbb{P}_\theta \left(|\widehat{L}_2 - \|\widehat{\beta} - \beta\|_2^2| \ge c \min\left\{ k\frac{\log p}{n}, \frac{1}{\sqrt{n}} \right\} \sigma_0^2 \right) \ge \delta. \tag{2.8}$$

For any estimator $\widehat{\beta}$ satisfying Assumption (A2) *with $\|\beta^*\|_0 \le k_0$,*

$$\inf_{\widehat{L}_q} \sup_{\theta \in \Theta_0(k)} \mathbb{P}_\theta \left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \ge c k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2 \right) \ge \delta, \quad \text{for } 1 \le q < 2, \tag{2.9}$$

where $\delta > 0$ and $c > 0$ are constants.

REMARK 1. Assumption (A1) restricts our focus to estimators that can perform well at at least one point $(\beta^*, \mathrm{I}, \sigma_0) \in \Theta_0(k)$. This weak condition makes the established lower bounds widely applicable as the benchmark for evaluating estimators of the ℓ_q loss of any $\widehat{\beta}$ that performs well at a proper subset, or even a single point of the whole parameter space.

In this paper, we focus on estimating the loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \le q \le 2$. Similar results can be established for the loss in the form of $\|\widehat{\beta} - \beta\|_q^q$ with $1 \le q \le 2$; Under the same assumptions as those in Theorem 1, the lower bounds for estimating the loss $\|\widehat{\beta} - \beta\|_q^q$ hold with replacing the convergence rates with their $\frac{q}{2}$ power; that is, (2.8) remains the same while the convergence

rate $k^{\frac{2}{q}}(\sqrt{\log p/n}\sigma_0)^2$ in (2.9) is replaced by $k(\sqrt{\log p/n}\sigma_0)^q$. Similarly, all the results established in the rest of the paper for $\|\widehat{\beta}-\beta\|_q^2$ hold for $\|\widehat{\beta}-\beta\|_q^q$ with corresponding convergence rates replaced by their $\frac{q}{2}$ power.

Theorem 1 establishes the minimax lower bounds for estimating the ℓ_2 loss $\|\widehat{\beta}-\beta\|_2^2$ of any estimator $\widehat{\beta}$ satisfying Assumption (A1) and the ℓ_q loss $\|\widehat{\beta}-\beta\|_q^2$ with $1 \le q < 2$ of any estimator $\widehat{\beta}$ satisfying Assumption (A2). We will take the Lasso estimator as an example and demonstrate the implications of the above theorem. We randomly split $Z=(y,X)$ into subsamples $Z^{(1)}=\left(y^{(1)},X^{(1)}\right)$ and $Z^{(2)}=\left(y^{(2)},X^{(2)}\right)$ with sample sizes n_1 and n_2, respectively. The Lasso estimator $\widehat{\beta}^L$ based on the first subsample $Z^{(1)}=\left(y^{(1)},X^{(1)}\right)$ is defined as

$$\widehat{\beta}^L = \arg\min_{\beta\in\mathbb{R}^p} \frac{\|y^{(1)}-X^{(1)}\beta\|_2^2}{n_1} + \lambda\sum_{j=1}^{p}\frac{\|X_{\cdot j}^{(1)}\|_2}{\sqrt{n_1}}|\beta_j|, \tag{2.10}$$

where $\lambda = A\sqrt{\log p/n_1}\sigma_0$ with $A > \sqrt{2}$ being a pre-specified constant. Without loss of generality, we assume $n_1 \asymp n_2$. For the case $1 \le q < 2$, (2.5) and (2.9) together imply that the estimation of the ℓ_q loss $\|\widehat{\beta}^L-\beta\|_q^2$ is impossible since the lower bound can be achieved by the trivial estimator of the loss, 0. That is, $\sup_{\theta\in\Theta_0(k)}\mathbb{P}_\theta\left(|0-\|\widehat{\beta}^L-\beta\|_q^2| \ge Ck^{\frac{2}{q}}\frac{\log p}{n}\right)\to 0$.

For the case $q=2$, in the regime $k \ll \frac{\sqrt{n}}{\log p}$, the lower bound $\frac{k\log p}{n}$ in (2.8) can be achieved by the zero estimator and hence estimation of the loss $\|\widehat{\beta}^L-\beta\|_2^2$ is impossible. However, the interesting case is when $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$, the loss estimator $\widetilde{L}_2$ proposed in (2.11) achieves the minimax lower bound $\frac{1}{\sqrt{n}}$ in (2.8), which cannot be achieved by the zero estimator. We now detail the construction of the loss estimator $\widetilde{L}_2$. Based on the second half sample $Z^{(2)}=\left(y^{(2)},X^{(2)}\right)$, we propose the following estimator,

$$\widetilde{L}_2 = \left(\frac{1}{n_2}\left\|y^{(2)}-X^{(2)}\widehat{\beta}^L\right\|_2^2 - \sigma_0^2\right)_+. \tag{2.11}$$

Note that the first subsample $Z^{(1)}=\left(y^{(1)},X^{(1)}\right)$ is used to produce the Lasso estimator $\widehat{\beta}^L$ in (2.10) and the second subsample $Z^{(2)}=\left(y^{(2)},X^{(2)}\right)$ is retained to evaluate the loss $\|\widehat{\beta}^L-\beta\|_2^2$. Such sample splitting technique is similar to cross-validation and has been used in [22] for constructing confidence sets for β and in [20] for confidence intervals for the ℓ_2 loss.

The following proposition establishes that the estimator $\widetilde{L}_2$ achieves the minimax lower bound of (2.8) over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$.

PROPOSITION 1. *Suppose that* $k \lesssim \frac{n}{\log p}$ *and* $\widehat{\beta}^L$ *is the Lasso estimator defined in* (2.10) *with* $A > \sqrt{2}$*, then the estimator of loss proposed in* (2.11) *satisfies, for any sequence* $\delta_{n,p} \to \infty$*,*

$$\limsup_{n,p\to\infty} \sup_{\theta\in\Theta_0(k)} \mathbb{P}_\theta \left(\left| \widetilde{L}_2 - \|\widehat{\beta}^L - \beta\|_2^2 \right| \geq \delta_{n,p} \frac{1}{\sqrt{n}} \right) = 0. \tag{2.12}$$

2.3. *Minimax estimation of the* ℓ_q *loss over* $\Theta(k)$. We now turn to the case of unknown Σ and σ and establish the minimax lower bound for estimating the ℓ_q loss over the parameter space $\Theta(k)$.

THEOREM 2. *Suppose that the sparsity levels* k *and* k_0 *satisfy Assumption* (B1)*. For any estimator* $\widehat{\beta}$ *satisfying Assumption* (A1) *with* $\|\beta^*\|_0 \leq k_0$*,*

$$\inf_{\widehat{L}_q} \sup_{\theta\in\Theta(k)} \mathbb{P}_\theta \left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \geq c k^{\frac{2}{q}} \frac{\log p}{n} \right) \geq \delta, \quad 1 \leq q \leq 2, \tag{2.13}$$

where $\delta > 0$ *and* $c > 0$ *are constants.*

Theorem 2 provides a minimax lower bound for estimating the ℓ_q loss of any estimator $\widehat{\beta}$ satisfying Assumption (A1), including the scaled Lasso estimator defined as

$$\{\widehat{\beta}^{SL}, \hat{\sigma}\} = \arg \min_{\beta\in\mathbb{R}^p, \sigma\in\mathbb{R}^+} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{j=1}^p \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} |\beta_j|, \tag{2.14}$$

where $\lambda_0 = A\sqrt{\log p/n}$ with $A > \sqrt{2}$. Note that for the scaled Lasso estimator, the lower bound in (2.13) can be achieved by the trivial loss estimator 0, in the sense, $\sup_{\theta\in\Theta(k)} \mathbb{P}_\theta \left(|0 - \|\widehat{\beta}^{SL} - \beta\|_q^2| \geq C k^{\frac{2}{q}} \frac{\log p}{n} \right) \to 0$, and hence estimation of loss is impossible in this case.

3. Minimaxity and adaptivity of confidence intervals over $\Theta_0(k)$. We focused in the last section on point estimation of the ℓ_q loss and showed the impossibility of loss estimation except for one regime. The results naturally lead to another question: Is it possible to construct "useful" confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ that can provide non-trivial upper and lower bounds for the loss? In this section, after introducing the framework for minimaxity and adaptivity of confidence intervals, we consider the case of known $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$ and establish the minimaxity and adaptivity lower bounds for the expected length of confidence intervals for the ℓ_q loss of a broad collection of estimators over the parameter space $\Theta_0(k)$. We also show that such

minimax lower bounds can be achieved for the Lasso estimator and then discuss the possibility of adaptivity using the Lasso estimator as an example. The case of unknown Σ and σ will be the focus of the next section.

3.1. *Framework for minimaxity and adaptivity of confidence intervals.* In this section, we introduce the following decision theoretical framework for confidence intervals of the loss $\|\widehat{\beta}-\beta\|_q^2$. Given $0<\alpha<1$ and the parameter space Θ and the loss $\|\widehat{\beta}-\beta\|_q^2$, denote by $\mathcal{I}_\alpha\left(\Theta,\widehat{\beta},\ell_q\right)$ the set of all $(1-\alpha)$ level confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ over Θ,

(3.1)
$$\mathcal{I}_\alpha\left(\Theta,\widehat{\beta},\ell_q\right)=\left\{\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)=[l\left(Z\right),u\left(Z\right)]:\inf_{\theta\in\Theta}\mathbb{P}_\theta\left(\|\widehat{\beta}-\beta(\theta)\|_q^2\in\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right)\geq 1-\alpha\right\}.$$

We will write CI_α for $\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)$ when there is no confusion. For any confidence interval $\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)=[l\left(Z\right),u\left(Z\right)]$, its length is denoted by $\mathbf{L}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right)=u\left(Z\right)-l\left(Z\right)$ and the maximum expected length over a parameter space Θ_1 is defined as

$$\mathbf{L}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right),\Theta_1\right)=\sup_{\theta\in\Theta_1}\mathbb{E}_\theta\mathbf{L}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right). \tag{3.2}$$

For two nested parameter spaces $\Theta_1\subseteq\Theta_2$, we define the benchmark $\mathbf{L}_\alpha^*\left(\Theta_1,\Theta_2,\widehat{\beta},\ell_q\right)$, measuring the degree of adaptivity over the nested spaces $\Theta_1\subset\Theta_2$,

(3.3)
$$\mathbf{L}_\alpha^*\left(\Theta_1,\Theta_2,\widehat{\beta},\ell_q\right)=\inf_{\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\in\mathcal{I}_\alpha\left(\Theta_2,\widehat{\beta},\ell_q\right)}\sup_{\theta\in\Theta_1}\mathbb{E}_\theta\mathbf{L}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right).$$

We will write $\mathbf{L}_\alpha^*\left(\Theta_1,\widehat{\beta},\ell_q\right)$ for $\mathbf{L}_\alpha^*\left(\Theta_1,\Theta_1,\widehat{\beta},\ell_q\right)$, which is the minimax expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ over Θ_1. The benchmark $\mathbf{L}_\alpha^*\left(\Theta_1,\Theta_2,\widehat{\beta},\ell_q\right)$ is the infimum of the maximum expected length over Θ_1 among all $(1-\alpha)$-level confidence intervals over Θ_2. In contrast, $\mathbf{L}_\alpha^*\left(\Theta_1,\widehat{\beta},\ell_q\right)$ is considering all $(1-\alpha)$-level confidence intervals over Θ_1. In words, if there is prior information that the parameter lies in the smaller parameter space Θ_1, $\mathbf{L}_\alpha^*\left(\Theta_1,\widehat{\beta},\ell_q\right)$ measures the benchmark length of confidence intervals over the parameter space Θ_1, which is illustrated in the left of Figure 1; however, if there is only prior information that the parameter lies in the larger parameter space Θ_2, $\mathbf{L}_\alpha^*\left(\Theta_1,\Theta_2,\widehat{\beta},\ell_q\right)$ measures the benchmark length of confidence intervals over the parameter space Θ_1, which is illustrated in the right of Figure 1.

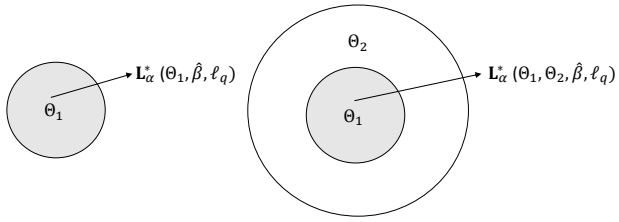

FIG 1. *The plot demonstrates definitions of* $\mathbf{L}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right)$ *and* $\mathbf{L}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right)$.

Rigorously, we define a confidence interval CI* to be simultaneously adaptive over Θ_1 and Θ_2 if $\text{CI}^* \in \mathcal{I}_\alpha\left(\Theta_2, \widehat{\beta}, \ell_q\right)$,

$$\mathbf{L}\left(\text{CI}^*, \Theta_1\right) \asymp \mathbf{L}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right), \quad \text{and} \quad \mathbf{L}\left(\text{CI}^*, \Theta_2\right) \asymp \mathbf{L}^*_\alpha\left(\Theta_2, \widehat{\beta}, \ell_q\right). \tag{3.4}$$

The condition (3.4) means that the confidence interval CI*, which has coverage over the larger parameter space Θ_2, achieves the minimax rate over both Θ_1 and Θ_2. Note that $\mathbf{L}\left(\text{CI}^*, \Theta_1\right) \geq \mathbf{L}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right)$. If $\mathbf{L}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right) \gg \mathbf{L}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right)$, then the rate-optimal adaptation (3.4) is impossible to achieve for $\Theta_1 \subset \Theta_2$. Otherwise, it is possible to construct confidence intervals simultaneously adaptive over parameter spaces Θ_1 and Θ_2. The possibility of adaptation over parameter spaces Θ_1 and Θ_2 can thus be answered by investigating the benchmark quantities $\mathbf{L}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right)$ and $\mathbf{L}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right)$. Such framework has already been introduced in [7], which studies the minimaxity and adaptivity of confidence intervals for linear functionals in high-dimensional linear regression.

We will adopt the minimax and adaptation framework discussed above and establish the minimax expected length $\mathbf{L}^*_\alpha\left(\Theta_0(k), \widehat{\beta}, \ell_q\right)$ and the adaptation benchmark $\mathbf{L}^*_\alpha\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right)$. In terms of the minimax expected length and the adaptivity behavior, there exist fundamental differences between the case $q = 2$ and $1 \leq q < 2$. We will discuss them separately in the following two subsections.

3.2. *Confidence intervals for the ℓ_2 loss over $\Theta_0(k)$.* The following theorem establishes the minimax lower bound for the expected length of confidence intervals of $\|\widehat{\beta} - \beta\|_2^2$ over the parameter space $\Theta_0(k)$.

THEOREM 3. *Suppose that $0 < \alpha < \frac{1}{4}$ and the sparsity levels k and k_0 satisfy Assumption* (B1). *For any estimator $\widehat{\beta}$ satisfying Assumption* (A1)

with $\|\beta^*\|_0 \le k_0$, *then there is some constant* $c > 0$ *such that*

$$\mathbf{L}_\alpha^* \left(\Theta_0(k), \widehat{\beta}, \ell_2 \right) \ge c \min \left\{ \frac{k \log p}{n}, \frac{1}{\sqrt{n}} \right\} \sigma_0^2. \tag{3.5}$$

In particular, if $\widehat{\beta}^L$ *is the Lasso estimator defined in* (2.10) *with* $A > \sqrt{2}$, *then the minimax expected length for* $(1-\alpha)$ *level confidence intervals of* $\|\widehat{\beta}^L - \beta\|_2^2$ *over* $\Theta_0(k)$ *is*

$$\mathbf{L}_\alpha^* \left(\Theta_0(k), \widehat{\beta}^L, \ell_2 \right) \asymp \min \left\{ \frac{k \log p}{n}, \frac{1}{\sqrt{n}} \right\} \sigma_0^2. \tag{3.6}$$

We now consider adaptivity of confidence intervals for the ℓ_2 loss. The following theorem gives the lower bound for the benchmark $\mathbf{L}_\alpha^* \left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_2 \right)$. We will then discuss Theorems 3 and 4 together.

THEOREM 4. *Suppose that* $0 < \alpha < \frac{1}{4}$ *and the sparsity levels* k_1, k_2 *and* k_0 *satisfy Assumption* (B2). *For any estimator* $\widehat{\beta}$ *satisfying Assumption* (A1) *with* $\|\beta^*\|_0 \le k_0$, *then there is some constant* $c > 0$ *such that*

$$\mathbf{L}_\alpha^* \left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_2 \right) \ge c \min \left\{ \frac{k_2 \log p}{n}, \frac{1}{\sqrt{n}} \right\} \sigma_0^2. \tag{3.7}$$

In particular, if $\widehat{\beta}^L$ *is the Lasso estimator defined in* (2.10) *with* $A > \sqrt{2}$, *the above lower bound can be achieved.*

The lower bound established in Theorem 4 implies that of Theorem 3 and both lower bounds hold for a general class of estimators satisfying Assumption (A1). There is a phase transition for the lower bound of the benchmark $\mathbf{L}_\alpha^* \left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_2 \right)$. In the regime $k_2 \ll \frac{\sqrt{n}}{\log p}$, the lower bound in (3.7) is $\frac{k_2 \log p}{n} \sigma_0^2$; when $\frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$, the lower bound in (3.7) is $\frac{1}{\sqrt{n}} \sigma_0^2$. For the Lasso estimator $\widehat{\beta}^L$ defined in (2.10), the lower bound $\frac{k \log p}{n} \sigma_0^2$ in (3.5) and $\frac{k_2 \log p}{n} \sigma_0^2$ in (3.7) can be achieved by the confidence intervals $\mathrm{CI}_\alpha^0 (Z, k, 2)$ and $\mathrm{CI}_\alpha^0 (Z, k_2, 2)$ defined in (3.15), respectively. Applying a similar idea to (2.11), we show that the minimax lower bound $\frac{1}{\sqrt{n}} \sigma_0^2$ in (3.6) and (3.7) can be achieved by the following confidence interval,

$$\mathrm{CI}_\alpha^1 (Z) = \left(\left(\frac{\psi(Z)}{\frac{1}{n_2} \chi_{1-\frac{\alpha}{2}}^2 (n_2)} - \sigma_0^2 \right)_+, \left(\frac{\psi(Z)}{\frac{1}{n_2} \chi_{\frac{\alpha}{2}}^2 (n_2)} - \sigma_0^2 \right)_+ \right), \tag{3.8}$$

where $\chi^2_{1-\frac{\alpha}{2}}(n_2)$ and $\chi^2_{\frac{\alpha}{2}}(n_2)$ are the $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ quantiles of χ^2 random variable with n_2 degrees of freedom, respectively, and

$$\psi(Z) = \min\left\{\frac{1}{n_2}\left\|y^{(2)} - X^{(2)}\widehat{\beta}^L\right\|_2^2, \sigma_0^2 \log p\right\}. \tag{3.9}$$

Note that the two-sided confidence interval (3.8) is simply based on the observed data Z, not depending on any prior knowledge of the sparsity k. Furthermore, it is a two-sided confidence interval, which tells not only just an upper bound, but also a lower bound for the loss. The coverage property and the expected length of $\mathrm{CI}^1_\alpha(Z)$ are established in the following proposition.

PROPOSITION 2. *Suppose $k \lesssim \frac{n}{\log p}$ and $\widehat{\beta}^L$ is the estimator defined in (2.10) with $A > \sqrt{2}$. Then $\mathrm{CI}^1_\alpha(Z)$ defined in (3.8) satisfies,*

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta_0(k)} \mathbb{P}\left(\|\widehat{\beta}^L - \beta\|_2^2 \in \mathrm{CI}^1_\alpha(Z)\right) \geq 1-\alpha, \tag{3.10}$$

and

$$\mathbf{L}\left(\mathrm{CI}^1_\alpha(Z), \Theta_0(k)\right) \lesssim \frac{1}{\sqrt{n}}\sigma_0^2. \tag{3.11}$$

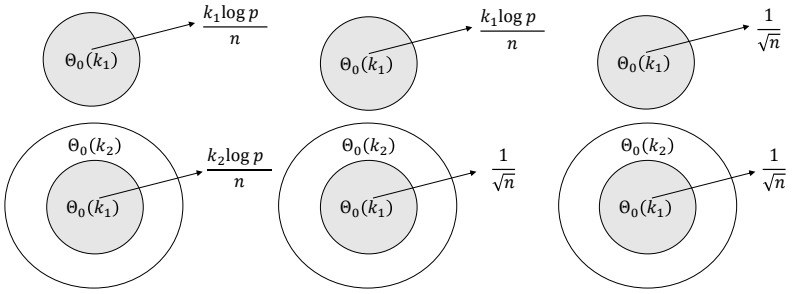

FIG 2. *Illustration of $\mathbf{L}^*_\alpha\left(\Theta_0(k_1), \widehat{\beta}^L, \ell_2\right)$ (top) and $\mathbf{L}^*_\alpha\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L, \ell_2\right)$ (bottom) over regimes $k_1 \leq k_2 \lesssim \frac{\sqrt{n}}{\log p}$ (leftmost), $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ (middle) and $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$ (rightmost).*

Regarding the Lasso estimator $\widehat{\beta}^L$ defined in (2.10), we will discuss the possibility of adaptivity of confidence intervals for $\|\widehat{\beta}^L - \beta\|_2^2$. The adaptivity behavior of confidence intervals for $\|\widehat{\beta}^L - \beta\|_2^2$ is demonstrated in Figure 2. As illustrated in the rightmost plot of Figure 2, in the regime $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$, we obtain $\mathbf{L}^*_\alpha\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L, \ell_2\right) \asymp \mathbf{L}^*_\alpha\left(\Theta_0(k_1), \widehat{\beta}^L, \ell_2\right) \asymp$

$\frac{1}{\sqrt{n}}$, which implies that adaptation is possible over this regime. As shown in Proposition 2, the confidence interval $\mathrm{CI}_\alpha^1(Z)$ defined in (3.8) is fully adaptive over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$ in the sense of (3.4).

Illustrated in the leftmost and middle plots of Figure 2, it is impossible to construct an adaptive confidence interval for $\|\widehat{\beta}^L - \beta\|_2^2$ over regimes $k_1 \leq k_2 \lesssim \frac{\sqrt{n}}{\log p}$ and $k_1 \ll \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ since $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L, \ell_2\right) \gg \mathbf{L}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}^L, \ell_2\right)$ if $k_1 \ll \frac{\sqrt{n}}{\log p}$ and $k_1 \ll k_2$. To sum up, adaptive confidence intervals for $\|\widehat{\beta}^L - \beta\|_2^2$ is only possible over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$.

Comparison with confidence balls. We should note that the problem of constructing confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$ is related to but different from that of constructing confidence sets for β itself. Confidence balls constructed in [22] are of form $\left\{\beta : \ \|\beta - \widehat{\beta}\|_2^2 \leq u_n(Z)\right\}$, where $\widehat{\beta}$ can be the Lasso estimator and $u_n(Z)$ is a data dependent squared radius. See [22] for further details. A naive application of this confidence ball leads to a one-sided confidence interval for the loss $\|\widehat{\beta} - \beta\|_2^2$,

$$\mathrm{CI}_\alpha^{\mathrm{induced}}(Z) = \left\{\|\widehat{\beta} - \beta\|_2^2 : \ \|\widehat{\beta} - \beta\|_2^2 \leq u_n(Z)\right\}. \tag{3.12}$$

Due to the reason that confidence sets for β were sought for in Theorem 1 in [22], confidence sets in the form $\left\{\beta : \ \|\beta - \widehat{\beta}\|_2^2 \leq u_n(Z)\right\}$ will suffice to achieve the optimal length. However, since our goal is to characterize $\|\widehat{\beta} - \beta\|_2^2$, we apply the unbiased risk estimation discussed in Theorem 1 of [22] and construct the two-sided confidence interval in (3.8). Such a two-sided confidence interval is more informative than the one-sided confidence interval (3.12) since the one-sided confidence interval does not contain the information whether the loss is close to zero or not. Furthermore, as shown in [22], the length of confidence interval $\mathrm{CI}_\alpha^{\mathrm{induced}}(Z)$ over the parameter space $\Theta_0(k)$ is of order $\frac{1}{\sqrt{n}} + \frac{k \log p}{n}$. The two-sided confidence interval $\mathrm{CI}_\alpha^1(Z)$ constructed in (3.8) is of expected length $\frac{1}{\sqrt{n}}$, which is much shorter than $\frac{1}{\sqrt{n}} + \frac{k \log p}{n}$ in the regime $k \gg \frac{\sqrt{n}}{\log p}$. That is, the two-sided confidence interval (3.8) provides a more accurate interval estimator of the ℓ_2 loss. This is illustrated in Figure 3.

The lower bound technique developed in the literature of adaptive confidence sets [22] can also be used to establish some of the lower bound results for the case $q = 2$ given in the present paper. However, new techniques are needed in order to establish the rate sharp lower bounds for the minimax estimation error (2.9) in the region $\frac{\sqrt{n}}{\log p} \leq k \lesssim \frac{n}{\log p}$ and for the

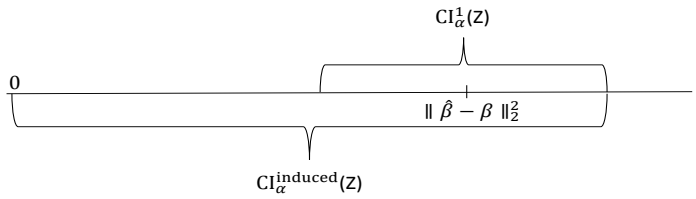

FIG 3. *Comparison of the two-sided confidence interval* $\mathrm{CI}_\alpha^1(Z)$ *with the one-sided confidence interval* $\mathrm{CI}_\alpha^{\mathrm{induced}}(Z)$.

expected length of the confidence intervals (3.18) and (7.3) in the region $\frac{\sqrt{n}}{\log p} \lesssim k_1 \le k_2 \lesssim \frac{n}{\log p}$, where it is necessary to test a composite null against a composite alternative in order to establish rate sharp lower bounds.

3.3. *Confidence intervals for the ℓ_q loss with $1 \le q < 2$ over $\Theta_0(k)$.* We now consider the case $1 \le q < 2$ and investigate the minimax expected length and adaptivity of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ over the parameter space $\Theta_0(k)$. The following theorem characterizes the minimax convergence rate for the expected length of confidence intervals.

THEOREM 5. *Suppose that $0 < \alpha < \frac{1}{4}$, $1 \le q < 2$ and the sparsity levels k and k_0 satisfy Assumption* (B1). *For any estimator $\widehat{\beta}$ satisfying Assumption* (A2) *with $\|\beta^*\|_0 \le k_0$, then there is some constant $c > 0$ such that*

$$\mathbf{L}_\alpha^* \left(\Theta_0(k), \widehat{\beta}, \ell_q \right) \ge c k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2. \tag{3.13}$$

In particular, if $\widehat{\beta}^L$ is the Lasso estimator defined in (2.10) *with $A > 4\sqrt{2}$, then the minimax expected length for $(1-\alpha)$ level confidence intervals of $\|\widehat{\beta}^L - \beta\|_q^2$ over $\Theta_0(k)$ is*

$$\mathbf{L}_\alpha^* \left(\Theta_0(k), \widehat{\beta}^L, \ell_q \right) \asymp k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2. \tag{3.14}$$

We now construct the confidence interval achieving the minimax convergence rate in (3.14),

$$\mathrm{CI}_\alpha^0(Z, k, q) = \left(0, C^*(A, k) k^{\frac{2}{q}} \frac{\log p}{n} \right), \tag{3.15}$$

where $C^*(A,k) = \max \left\{ \frac{(22A\sigma_0)^2}{\left(\frac{1}{4} - 42\sqrt{\frac{2k \log p}{n_1}}\right)^4}, \frac{\left(\frac{3\eta_0}{\eta_0+1} A\sigma_0\right)^2}{\left(\frac{1}{4} - (9+11\eta_0)\sqrt{\frac{2k \log p}{n_1}}\right)^4} \right\}$ with $\eta_0 = 1.01 \frac{\sqrt{A}+\sqrt{2}}{\sqrt{A}-\sqrt{2}}$. The following proposition establishes the coverage property and the expected length of $\mathrm{CI}_\alpha^0(Z, k, q)$.

PROPOSITION 3. *Suppose* $k \lesssim \frac{n}{\log p}$ *and* $\widehat{\beta}^L$ *is the estimator defined in* (2.10) *with* $A > 4\sqrt{2}$. *For* $1 \le q \le 2$, *the confidence interval* $\mathrm{CI}_\alpha^0(Z,k,q)$ *defined in* (3.15) *satisfies*

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta_0(k)} \mathbb{P}_\theta\left(\|\widehat{\beta}-\beta\|_q^2 \in \mathrm{CI}_\alpha^0(Z,k,q)\right) = 1, \tag{3.16}$$

and

$$\mathbf{L}\left(\mathrm{CI}_\alpha^0(Z,k,q), \Theta_0(k)\right) \lesssim k^{\frac{2}{q}} \frac{\log p}{n}\sigma_0^2. \tag{3.17}$$

In particular, for the case $q = 2$, (3.16) *and* (3.17) *also hold for the estimator* $\widehat{\beta}^L$ *defined in* (2.10) *with* $A > \sqrt{2}$.

This result shows that the confidence interval $\mathrm{CI}_\alpha^0(Z,k,q)$ achieves the minimax rate given in (3.14). In contrast to the ℓ_2 loss where the two-sided confidence interval (3.8) is significantly shorter than the one-sided interval and achieves the optimal rate over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$, for the ℓ_q loss with $1 \le q < 2$, the one-sided confidence interval achieves the optimal rate given in (3.14).

We now consider adaptivity of confidence intervals. The following theorem establishes the lower bounds for $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right)$ with $1 \le q < 2$.

THEOREM 6. *Suppose* $0 < \alpha < \frac{1}{4}$, $1 \le q < 2$ *and the sparsity levels* k_1, k_2 *and* k_0 *satisfy Assumption* (B2). *For any estimator* $\widehat{\beta}$ *satisfying Assumption* (A2) *with* $\|\beta^*\|_0 \le k_0$, *then there is some constant* $c > 0$ *such that*

$$\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right) \ge \begin{cases} ck_2^{\frac{2}{q}} \frac{\log p}{n}\sigma_0^2 & \text{if} \quad k_1 \le k_2 \lesssim \frac{\sqrt{n}}{\log p}; \\ ck_2^{\frac{2}{q}-1} \frac{1}{\sqrt{n}}\sigma_0^2 & \text{if} \quad k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}; \\ ck_2^{\frac{2}{q}-1} k_1 \frac{\log p}{n}\sigma_0^2 & \text{if} \quad \frac{\sqrt{n}}{\log p} \lesssim k_1 \le k_2 \lesssim \frac{n}{\log p}. \end{cases} \tag{3.18}$$

In particular, if $p \ge n$ *and* $\widehat{\beta}^L$ *is the Lasso estimator defined in* (2.10) *with* $A > 4\sqrt{2}$, *the above lower bounds can be achieved.*

The lower bounds of Theorem 6 imply that of Theorem 5 and both lower bounds hold for a general class of estimators satisfying Assumption (A2). However, the lower bound (3.18) in Theorem 6 has a significantly different meaning from (3.13) in Theorem 5 where (3.18) quantifies the cost of adaptation without knowing the sparsity level. For the Lasso estimator

$\widehat{\beta}^L$ defined in (2.10), by comparing Theorem 5 and Theorem 6, we obtain $\mathbf{L}^*_\alpha\left(\Theta_0(k_1),\Theta_0(k_2),\widehat{\beta}^L,\ell_q\right) \gg \mathbf{L}^*_\alpha\left(\Theta_0(k_1),\widehat{\beta}^L,\ell_q\right)$ if $k_1 \ll k_2$, which implies the impossibility of constructing adaptive confidence intervals for the case $1 \le q < 2$. There exists marked difference between the case $1 \le q < 2$ and the case $q = 2$, where it is possible to construct adaptive confidence intervals over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$.

For the Lasso estimator $\widehat{\beta}^L$ defined in (2.10), it is shown in Proposition 3 that the confidence interval $\mathrm{CI}^0_\alpha(Z,k_2,q)$ defined in (3.15) achieves the lower bound $k_2^{\frac{2}{q}}\frac{\log p}{n}\sigma_0^2$ of (3.18). The lower bounds $k_2^{\frac{2}{q}-1}k_1\frac{\log p}{n}\sigma_0^2$ and $k_2^{\frac{2}{q}-1}\frac{1}{\sqrt{n}}\sigma_0^2$ of (3.18) can be achieved by the following proposed confidence interval,

(3.19)
$$\mathrm{CI}^2_\alpha(Z,k_2,q) = \left(\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)} - \sigma_0^2\right)_+, (16k_2)^{\frac{2}{q}-1}\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)} - \sigma_0^2\right)_+\right),$$

where $\psi(Z)$ is given in (3.9). The above claim is verified in Proposition 4. Note that the confidence interval $\mathrm{CI}^1_\alpha(Z)$ defined in (3.8) is a special case of $\mathrm{CI}^2_\alpha(Z,k_2,q)$ with $q = 2$.

PROPOSITION 4. *Suppose $p \ge n$, $k_1 \le k_2 \lesssim \frac{n}{\log p}$ and $\widehat{\beta}^L$ is defined in (2.10) with $A > 4\sqrt{2}$. Then $\mathrm{CI}^2_\alpha(Z,k_2,q)$ defined in (3.19) satisfies,*

$$\liminf_{n,p\to\infty}\inf_{\theta\in\Theta_0(k_2)}\mathbb{P}_\theta\left(\|\widehat{\beta}-\beta\|_q^2 \in \mathrm{CI}^2_\alpha(Z,k_2,q)\right) \ge 1-\alpha, \tag{3.20}$$

and

$$\mathbf{L}\left(\mathrm{CI}^2_\alpha(Z,k_2,q),\Theta_0(k_1)\right) \lesssim k_2^{\frac{2}{q}-1}\left(k_1\frac{\log p}{n}+\frac{1}{\sqrt{n}}\right)\sigma_0^2. \tag{3.21}$$

4. Minimaxity and adaptivity of confidence intervals over $\Theta(k)$. In this section, we focus on the case of unknown Σ and σ and establish the minimax expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ with $1 \le q \le 2$ over $\Theta(k)$ defined in (2.4). We also study the possibility of adaptivity of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$. The following theorem establishes the lower bounds for the benchmark quantities $\mathbf{L}^*_\alpha\left(\Theta(k_i),\widehat{\beta},\ell_q\right)$ with $i = 1,2$ and $\mathbf{L}^*_\alpha\left(\Theta(k_1),\Theta(k_2),\widehat{\beta},\ell_q\right)$.

THEOREM 7. *Suppose that $0 < \alpha < \frac{1}{4}$, $1 \le q \le 2$ and the sparsity levels k_1, k_2 and k_0 satisfy Assumption* (B2). *For any estimator $\widehat{\beta}$ satisfying*

Assumption (A1) *at* $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$ *with* $\|\beta^*\|_0 \le k_0$, *there is a constant* $c > 0$ *such that*

$$\mathbf{L}_\alpha^* \left(\Theta\left(k_i\right), \widehat{\beta}, \ell_q \right) \ge c k_i^{\frac{2}{q}} \frac{\log p}{n}, \quad \textit{for} \quad i = 1, 2; \tag{4.1}$$

$$\mathbf{L}_\alpha^* \left(\{\theta_0\}, \Theta\left(k_2\right), \widehat{\beta}, \ell_q \right) \ge c k_2^{\frac{2}{q}} \frac{\log p}{n}. \tag{4.2}$$

In particular, if $\widehat{\beta}^{SL}$ *is the scaled Lasso estimator defined in* (2.14) *with* $A > 2\sqrt{2}$, *then the above lower bounds can be achieved.*

The lower bounds (4.1) and (4.2) hold for any $\widehat{\beta}$ satisfying Assumption (A1) at an interior point $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$, including the scaled Lasso estimator as a special case. We demonstrate the impossibility of adaptivity of confidence intervals for the ℓ_q loss of the scaled Lasso estimator $\widehat{\beta}^{SL}$. Since $\mathbf{L}_\alpha^* \left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q \right) \ge \mathbf{L}_\alpha^* \left(\{\theta_0\}, \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q \right)$, by (4.2), we have $\mathbf{L}_\alpha^* \left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q \right) \gg \mathbf{L}_\alpha^* \left(\Theta\left(k_1\right), \widehat{\beta}^{SL}, \ell_q \right)$ if $k_1 \ll k_2$. The comparison of $\mathbf{L}_\alpha^* \left(\Theta\left(k_1\right), \widehat{\beta}^{SL}, \ell_q \right)$ and $\mathbf{L}_\alpha^* \left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q \right)$ is illustrated in Figure 4. Referring to the adaptivity defined in (3.4), it is impossible to construct adaptive confidence intervals for $\|\widehat{\beta}^{SL} - \beta\|_q^2$.

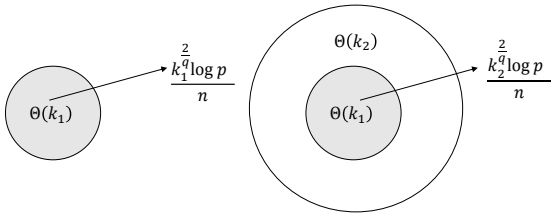

FIG 4. *Illustration of* $\mathbf{L}_\alpha^* \left(\Theta\left(k_1\right), \widehat{\beta}^{SL}, \ell_q \right)$ *(left) and* $\mathbf{L}_\alpha^* \left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q \right)$ *(right).*

Theorem 7 shows that for any confidence interval $\mathrm{CI}_\alpha \left(\widehat{\beta}, \ell_q, Z \right)$ for the loss of any given estimator $\widehat{\beta}$ satisfying Assumption (A1), under the coverage constraint that $\mathrm{CI}_\alpha \left(\widehat{\beta}, \ell_q, Z \right) \in \mathcal{I}_\alpha \left(\Theta\left(k_2\right), \widehat{\beta}, \ell_q \right)$, its expected length at any given $\theta_0 = (\beta^*, \mathrm{I}, \sigma) \in \Theta\left(k_0\right)$ must be of order $k_2^{\frac{2}{q}} \frac{\log p}{n}$. In contrast to Theorem 4 and 6, Theorem 7 demonstrates that confidence intervals must be long at a large subset of points in the parameter space, not just at a small

number of "unlucky" points. Therefore, the lack of adaptivity for confidence intervals is not due to the conservativeness of the minimax framework.

In the following, we detail the construction of confidence intervals for $\|\widehat{\beta}^{SL}-\beta\|_q^2$. The construction of confidence intervals is based on the following definition of restricted eigenvalue, which is introduced in [4],

$$\kappa(X,k,s,\alpha_0)=\min_{\substack{J_0\subset\{1,\cdots,p\},\\ |J_0|\le k}}\ \min_{\substack{\delta\neq 0,\\ \|\delta_{J_0^c}\|_1\le\alpha_0\|\delta_{J_0}\|_1}}\frac{\|X\delta\|_2}{\sqrt{n}\|\delta_{J_{01}}\|_2}, \tag{4.3}$$

where J_1 denotes the subset corresponding to the s largest in absolute value coordinates of δ outside of J_0 and $J_{01}=J_0\cup J_1$. Define the event $\mathcal{B}=\{\hat{\sigma}\le\log p\}$. The confidence interval for $\|\widehat{\beta}^{SL}-\beta\|_q^2$ is defined as

$$\mathrm{CI}_\alpha(Z,k,q)=\begin{cases}[0,\ \varphi(Z,k,q)] & \text{on } \mathcal{B}\\ \{0\} & \text{on } \mathcal{B}^c,\end{cases} \tag{4.4}$$

where

$$\varphi(Z,k,q)=\min\left\{\left(\frac{16A\max\|X_{\cdot j}\|_2^2\widehat{\sigma}}{n\kappa^2\left(X,k,k,3\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right)}\right)^2 k^{\frac{2}{q}}\frac{\log p}{n},\ \left(k^{\frac{2}{q}}\frac{\log p}{n}\log p\right)\widehat{\sigma}^2\right\}.$$

Remark 2. The restricted eigenvalue $\kappa^2\left(X,k,k,3\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right)$ is computationally infeasible. For design covariance matrix Σ of special structures, the restricted eigenvalue can be replaced by its lower bound and a computationally feasible confidence interval can be constructed. See Section 4.4 in [7] for more details.

Properties of $\mathrm{CI}_\alpha(Z,k,q)$ are established as follows.

Proposition 5. *Suppose $k\lesssim\frac{n}{\log p}$ and $\widehat{\beta}^{SL}$ is the estimator defined in (2.14) with $A>2\sqrt{2}$. For $1\le q\le 2$, then $\mathrm{CI}_\alpha(Z,k,q)$ defined in (4.4) satisfies the following properties,*

$$\liminf_{n,p\to\infty}\inf_{\theta\in\Theta(k)}\mathbb{P}_\theta\left(\|\widehat{\beta}-\beta\|_q^2\in\mathrm{CI}_\alpha(Z,k,q)\right)=1, \tag{4.5}$$

$$\mathbf{L}\left(\mathrm{CI}_\alpha(Z,k,q),\Theta(k)\right)\lesssim k^{\frac{2}{q}}\frac{\log p}{n}. \tag{4.6}$$

Proposition 5 shows that the confidence interval $\mathrm{CI}_\alpha(Z,k_i,q)$ defined in (4.4) achieves the lower bound in (4.1), for $i=1,2$, and the confidence interval $\mathrm{CI}_\alpha(Z,k_2,q)$ defined in (4.4) achieves the lower bound in (4.2).

5. Estimation of the ℓ_q loss of rate-optimal estimators. We have established minimax lower bounds for the estimation accuracy of the loss of a broad class of estimators $\widehat{\beta}$ satisfying (A1) or (A2) and also demonstrated that such minimax lower bounds are sharp for the Lasso and scaled Lasso estimators. We now show that the minimax lower bounds are sharp for the class of rate-optimal estimators satisfying the following Assumption (A).

(A) The estimator $\widehat{\beta}$ satisfies,

$$\sup_{\theta\in\Theta(k)} \mathbb{P}_\theta\left(\|\widehat{\beta}-\beta\|_q^2 \geq C^*\|\beta\|_0^{\frac{2}{q}}\frac{\log p}{n}\right) \leq Cp^{-\delta}, \tag{5.1}$$

for all $k \ll \frac{n}{\log p}$, where $\delta > 0$, $C^* > 0$ and $C > 0$ are constants not depending on k, n, or p.

We say an estimator $\widehat{\beta}$ is rate-optimal if it satisfies Assumption (A). As shown in [12, 4, 3, 26], Lasso, Dantzig Selector, scaled Lasso and square-root Lasso are rate-optimal when the tuning parameter is chosen properly. We shall stress that Assumption (A) implies Assumptions (A1) and (A2). Assumption (A) requires the estimator $\widehat{\beta}$ to perform well over the whole parameter space $\Theta(k)$ while Assumptions (A1) and (A2) only require $\widehat{\beta}$ to perform well at a single point or over a proper subset. The following proposition shows that the minimax lower bounds established in Theorem 1 to Theorem 7 can be achieved for the class of rate-optimal estimators.

PROPOSITION 6. *Let $\widehat{\beta}$ be an estimator satisfying Assumption* (A).

1. *There exist (point or interval) estimators of the loss $\|\widehat{\beta}-\beta\|_q^2$ with $1 \leq q < 2$ achieving, up to a constant factor, the minimax lower bounds* (2.9) *in Theorem 1 and* (3.13) *in Theorem 5 and estimators of loss $\|\widehat{\beta}-\beta\|_q^2$ with $1 \leq q \leq 2$ achieving, up to a constant factor, the minimax lower bounds* (2.13) *in Theorem 2 and* (4.1) *and* (4.2) *in Theorem 7.*
2. *Suppose that the estimator $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$, then there exist estimators of the loss $\|\widehat{\beta}-\beta\|_2^2$ achieving, up to a constant factor, the minimax lower bounds* (2.8) *in Theorem 1,* (3.5) *in Theorem 3 and* (3.7) *in Theorem 4.*
3. *Suppose the estimator $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ and it satisfies Assumption (A) with $\delta > 2$ and*

$$\sup_{\theta\in\Theta(k)} \mathbb{P}_\theta\left(\|(\widehat{\beta}-\beta)_{S^c}\|_1 \geq c^*\|(\widehat{\beta}-\beta)_S\|_1 \text{ where } S = \operatorname{supp}(\beta)\right) \leq Cp^{-\delta}, \tag{5.2}$$

for all $k \ll \frac{n}{\log p}$. *Then for* $p \geq n$ *there exist estimators of the loss* $\|\widehat{\beta} - \beta\|_q^2$ *with* $1 \leq q < 2$ *achieving the lower bounds given in* (3.18) *in Theorem 6.*

For reasons of space, we do not discuss the detailed construction for the point and interval estimators achieving these minimax lower bounds here and postpone the construction to the proof of Proposition 6.

REMARK 3. Sample splitting has been widely used in the literature. For example, the condition that $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ has been introduced in [22] for constructing confidence sets for β and in [20] for constructing confidence intervals for the ℓ_2 loss. Such a condition is imposed purely for technical reasons to create independence between the estimator $\widehat{\beta}$ and the subsample $Z^{(2)} = \left(y^{(2)}, X^{(2)}\right)$, which is useful to evaluate the ℓ_q loss of the estimator $\widehat{\beta}$. As shown in [4], the assumption (5.2) is satisfied for Lasso and Dantzig Selector. This technical assumption is imposed such that $\|\widehat{\beta} - \beta\|_1^2$ can be tightly controlled by $\|\widehat{\beta} - \beta\|_2^2$.

6. General tools for minimax lower bounds. A major step in our analysis is to establish rate sharp lower bounds for the estimation error and the expected length of confidence intervals for the ℓ_q loss. We introduce in this section new technical tools that are needed to establish these lower bounds.

A significant distinction of the lower bound results given in the previous sections from those for the traditional parameter estimation problems is that the constraint is on the performance of the estimator $\widehat{\beta}$ of the regression vector β, but the lower bounds are on the difficulty of estimating its loss $\|\widehat{\beta} - \beta\|_q^2$. It is necessary to develop new lower bound techniques to establish rate-optimal lower bounds for the estimation error and the expected length of confidence intervals for the loss $\|\widehat{\beta} - \beta\|_q^2$. These technical tools may also be of independent interest.

We begin with notation. Let Z denote a random variable whose distribution is indexed by some parameter $\theta \in \Theta$ and let π denote a prior on the parameter space Θ. We will use $f_\theta(z)$ to denote the density of Z given θ and $f_\pi(z)$ to denote the marginal density of Z under the prior π. Let $\mathbb{P}_\pi$ denote the distribution of Z corresponding to $f_\pi(z)$, i.e., $\mathbb{P}_\pi(\mathcal{A}) = \int 1_{z \in \mathcal{A}} f_\pi(z)\, dz$, where $1_{z \in \mathcal{A}}$ is the indicator function. For a function g, we write $\mathbb{E}_\pi(g(Z))$ for the expectation under f_π. More specifically, $f_\pi(z) = \int f_\theta(z)\, \pi(\theta)\, d\theta$ and $\mathbb{E}_\pi(g(Z)) = \int g(z)\, f_\pi(z)\, dz$. The L_1 distance between two probability distributions with densities f_0 and f_1 is given by $L_1(f_1, f_0) = \int |f_1(z) - f_0(z)|\, dz$.

The following theorem establishes the minimax lower bounds for the estimation error and the expected length of confidence intervals for the ℓ_q loss, under the constraint that $\widehat{\beta}$ is a good estimator at at least one interior point.

THEOREM 8. *Suppose* $0 < \alpha, \alpha_0 < \frac{1}{4}$, $1 \le q \le 2$, Σ_0 *is positive definite,* $\theta_0 = (\beta^*, \Sigma_0, \sigma_0) \in \Theta$, *and* $\mathcal{F} \subset \Theta$. *Define* $d = \min_{\theta \in \mathcal{F}} \|\beta(\theta) - \beta^*\|_q$. *Let* π *denote a prior over the parameter space* $\mathcal{F}$. *If an estimator* $\widehat{\beta}$ *satisfies*

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\beta} - \beta^*\|_q^2 \le \frac{1}{16} d^2\right) \ge 1 - \alpha_0, \tag{6.1}$$

then

$$\inf_{\widehat{L}_q} \sup_{\theta \in \{\theta_0\} \cup \mathcal{F}} \mathbb{P}_\theta\left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \ge \frac{1}{4} d^2\right) \ge \bar{c}_1, \tag{6.2}$$

and

$$\mathbf{L}_\alpha^*\left(\{\theta_0\}, \Theta, \widehat{\beta}, \ell_q\right) = \inf_{\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, Z\right) \in \mathcal{I}_\alpha\left(\Theta, \widehat{\beta}, \ell_q\right)} \mathbb{E}_{\theta_0} \mathbf{L}\left(\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, Z\right)\right) \ge c_2^* d^2, \tag{6.3}$$

where $\bar{c}_1 = \min\left\{\frac{1}{10}, \left(\frac{9}{10} - \alpha_0 - L_1\left(f_\pi, f_{\theta_0}\right)\right)_+\right\}$ *and* $c_2^* = \frac{1}{2}\left(1 - 2\alpha - \alpha_0 - 2L_1\left(f_\pi, f_{\theta_0}\right)\right)_+$.

REMARK 4. The minimax lower bound (6.2) for the estimation error and (6.3) for the expected length of confidence intervals hold as long as the estimator $\widehat{\beta}$ estimates β well at an interior point θ_0. Besides Condition (6.1), another key ingredient for the lower bounds (6.2) and (6.3) is to construct the least favorable space $\mathcal{F}$ with the prior π such that the marginal distributions f_π and f_{θ_0} are non-distinguishable. For the estimation lower bound (6.2), constraining that $\|\widehat{\beta} - \beta^*\|_q^2$ can be well estimated at θ_0, due to the non-distinguishability between f_π and f_{θ_0}, we can establish that the loss $\|\widehat{\beta} - \beta\|_q^2$ cannot be estimated well over $\mathcal{F}$. For the lower bound (6.3), by Condition (6.1) and the non-distinguishability between f_π and f_{θ_0}, we will show that $\|\widehat{\beta} - \beta\|_q^2$ over $\mathcal{F}$ is much larger than $\|\widehat{\beta} - \beta^*\|_q^2$ and hence the honest confidence intervals must be sufficiently long.

Theorem 8 is used to establish the minimax lower bounds for both the estimation error and the expected length of confidence intervals of the ℓ_q loss over $\Theta(k)$. By taking $\theta_0 \in \Theta(k_0)$ and $\Theta = \Theta(k)$, Theorem 2 follows from (6.2) with a properly constructed subset $\mathcal{F} \subset \Theta(k)$. By taking $\theta_0 \in \Theta(k_0)$ and $\Theta = \Theta(k_2)$, the lower bound (4.2) in Theorem 7 follows from (6.3) with

a properly constructed $\mathcal{F} \subset \Theta(k_2)$. In both cases, Assumption (A1) implies Condition (6.1).

Several minimax lower bounds over $\Theta_0(k)$ can also be implied by Theorem 8. For the estimation error, the minimax lower bounds (2.8) and (2.9) over the regime $k \lesssim \frac{\sqrt{n}}{\log p}$ in Theorem 1 follow from (6.2). For the expected length of confidence intervals, the minimax lower bounds (3.7) in Theorem 4 and (3.18) in the regions $k_1 \le k_2 \lesssim \frac{\sqrt{n}}{\log p}$ and $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ in Theorem 6 follow from (6.3). In these cases, Assumption (A1) or (A2) can guarantee that Condition (6.1) is satisfied. However, the minimax lower bounds for estimation error (2.9) in the region $\frac{\sqrt{n}}{\log p} \le k \lesssim \frac{n}{\log p}$ and for the expected length of confidence intervals (3.18) in the region $\frac{\sqrt{n}}{\log p} \lesssim k_1 \le k_2 \lesssim \frac{n}{\log p}$ cannot be established using the above theorem. The following theorem, which requires testing a composite null against a composite alternative, establishes the refined minimax lower bounds over $\Theta_0(k)$.

THEOREM 9. *Let $0 < \alpha, \alpha_0 < \frac{1}{4}$, $1 \le q \le 2$, and $\theta_0 = (\beta^*, \Sigma_0, \sigma_0)$ where Σ_0 is a positive definite matrix. Let k_1 and k_2 be two sparsity levels. Assume that for $i = 1, 2$ there exist parameter spaces $\mathcal{F}_i \subset \{(\beta, \Sigma_0, \sigma_0) : \|\beta\|_0 \le k_i\}$ such that for given dist_i and d_i*

$$\sqrt{(\beta(\theta) - \beta^*)^\intercal \Sigma_0 (\beta(\theta) - \beta^*)} = \mathrm{dist}_i \quad \text{and} \quad \|\beta(\theta) - \beta^*\|_q = d_i, \quad \text{for all } \theta \in \mathcal{F}_i.$$

Let π_i denote a prior over the parameter space $\mathcal{F}_i$ for $i = 1, 2$. Suppose that for $\theta_1 = \left(\beta^*, \Sigma_0, \sigma_0^2 + \mathrm{dist}_1^2\right)$ and $\theta_2 = \left(\beta^*, \Sigma_0, \sigma_0^2 + \mathrm{dist}_2^2\right)$, there exist constants $c_1, c_2 > 0$ such that

$$\mathbb{P}_{\theta_i}\left(\|\widehat{\beta} - \beta^*\|_q^2 \le c_i^2 d_i^2\right) \ge 1 - \alpha_0, \quad \text{for} \quad i = 1, 2. \tag{6.4}$$

Then we have

$$\inf_{\widehat{L}_q} \sup_{\theta \in \mathcal{F}_1 \cup \mathcal{F}_2} \mathbb{P}_\theta\left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \ge c_3^* d_2^2\right) \ge \bar{c}_3, \tag{6.5}$$

and

$$\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right) \ge c_4^* \left((1 - c_2)^2 d_2^2 - (1 + c_1)^2 d_1^2\right)_+, \tag{6.6}$$

where $c_3^* = \min\left\{\frac{1}{4}, \left((1 - c_2)^2 - \frac{1}{4} - (1 + c_1)^2 \frac{d_1^2}{d_2^2}\right)_+\right\}$, $c_4^* = \left(1 - 2\alpha_0 - 2\alpha - \sum_{i=1}^2 L_1(f_{\pi_i}, f_{\theta_i}) - 2L_1(f_{\pi_2}, f_{\pi_1})\right)_+$ *and* $\bar{c}_3 = \min\left\{\frac{1}{10}, \left(\frac{9}{10} - 2\alpha_0 - \sum_{i=1}^2 L_1(f_{\pi_i}, f_{\theta_i}) - 2L_1(f_{\pi_2}, f_{\pi_1})\right)_+\right\}$.

REMARK 5. As long as the estimator $\widehat{\beta}$ performs well at two points, θ_1 and θ_2, the minimax lower bounds (6.5) for the estimation error and (6.6) for the expected length of confidence intervals hold. Note that θ_i in the above theorem does not belong to the parameter space $\{(\beta, \Sigma_0, \sigma_0) : \|\beta\|_0 \le k_i\}$, for $i = 1, 2$. In contrast to Theorem 8, Theorem 9 compares composite hypotheses $\mathcal{F}_1$ and $\mathcal{F}_2$, which will lead to a sharper lower bound than comparing the simple null $\{\theta_0\}$ with the composite alternative $\mathcal{F}$. For simplicity, we construct least favorable parameter spaces $\mathcal{F}_i$ such that the points in $\mathcal{F}_i$ is of fixed (generalized) ℓ_2 distance and fixed ℓ_q distance to β^*, for $i = 1, 2$, respectively. More importantly, we construct $\mathcal{F}_1$ with the prior π_1 and $\mathcal{F}_2$ with the prior π_2 such that f_{π_1} and f_{π_2} are not distinguishable, where θ_1 and θ_2 are introduced to facilitate the comparison. By Condition (6.4) and the construction of $\mathcal{F}_1$ and $\mathcal{F}_2$, we establish that the ℓ_q loss cannot be simultaneously estimated well over $\mathcal{F}_1$ and $\mathcal{F}_2$. For the lower bound (6.6), under the same conditions, it is shown that the ℓ_q loss over $\mathcal{F}_1$ and $\mathcal{F}_2$ are far apart and any confidence interval with guaranteed coverage probability over $\mathcal{F}_1 \cup \mathcal{F}_2$ must be sufficiently long. Due to the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$, the lower bound construction over $\Theta_0(k)$ is more involved than that over $\Theta(k)$. We shall stress that the construction of $\mathcal{F}_1$ and $\mathcal{F}_2$ and the comparison between composite hypotheses are of independent interest.

The minimax lower bound (2.9) in the region $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$ follows from (6.5) and the minimax lower bound (3.18) in the region $\frac{\sqrt{n}}{\log p} \lesssim k_1 \le k_2 \lesssim \frac{n}{\log p}$ for the expected length of confidence intervals follows from (6.6). In these cases, Σ_0 is taken as I and Assumption (A2) implies Condition (6.4).

7. An intermediate setting with known $\sigma = \sigma_0$ and unknown Σ. The results given in Sections 3 and 4 show the significant difference between $\Theta_0(k)$ and $\Theta(k)$ in terms of minimaxity and adaptivity of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$. $\Theta_0(k)$ is for the simple setting with known design covariance matrix $\Sigma = \mathrm{I}$ and known noise level $\sigma = \sigma_0$, and $\Theta(k)$ is for unknown Σ and σ. In this section, we further consider minimaxity and adaptivity of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ in an intermediate setting where the noise level $\sigma = \sigma_0$ is known and Σ is unknown but of certain structure. Specifically, we consider the following parameter space,

(7.1)
$$\Theta_{\sigma_0}(k, s) = \left\{ (\beta, \Sigma, \sigma_0) : \begin{array}{ll} \|\beta\|_0 \le k, & \dfrac{1}{M_1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le M_1 \\ \|\Sigma^{-1}\|_{L_1} \le M, & \max\limits_{1 \le i \le p} \| \left(\Sigma^{-1}\right)_{i\cdot} \|_0 \le s \end{array} \right\},$$

for some constants $M_1 \geq 1$ and $M > 0$. $\Theta_{\sigma_0}(k, s)$ basically assumes known noise level σ and imposes sparsity conditions on the precision matrix of the random design. This parameter space is similar to those used in the literature of sparse linear regression with random design [29, 13, 14]. $\Theta_{\sigma_0}(k, s)$ has two sparsity parameters where k represents the sparsity of β and s represents the maximum row sparsity of the precision matrix Σ^{-1}. Note that $\Theta_0(k) \subset \Theta_{\sigma_0}(k, s) \subset \Theta(k)$ and $\Theta_0(k)$ is a special case of $\Theta_{\sigma_0}(k, s)$ with $M_1 = 1$.

Under the assumption $s \ll \sqrt{n/\log p}$, the minimaxity and adaptivity lower bounds for the expected length of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q < 2$ over $\Theta_{\sigma_0}(k, s)$ are the same as those over $\Theta_0(k)$. That is, Theorems 5 and 6 hold with $\Theta_0(k_1)$, $\Theta_0(k_2)$, and $\Theta_0(k)$ replaced by $\Theta_{\sigma_0}(k_1, s)$, $\Theta_{\sigma_0}(k_2, s)$, and $\Theta_{\sigma_0}(k, s)$, respectively. For the case $q = 2$, the following theorem establishes the minimaxity and adaptivity lower bounds for the expected length of confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$ over $\Theta_{\sigma_0}(k, s)$.

THEOREM 10. *Suppose* $0 < \alpha, \alpha_0 < 1/4$, $M_1 > 1$, $s \ll \sqrt{n/\log p}$ *and the sparsity levels* k_1, k_2 *and* k_0 *satisfy Assumption* (B2) *with the constant* c_0 *replaced by* c_0^* *defined in* (9.14). *For any estimator* $\widehat{\beta}$ *satisfying*

$$\sup_{\theta \in \Theta(k_0)} \mathbb{P}_\theta \left(\|\widehat{\beta} - \beta^*\|_q^2 \geq C^* \|\beta^*\|_0^{\frac{2}{q}} \frac{\log p}{n} \sigma^2 \right) \leq \alpha_0, \tag{7.2}$$

with a constant $C^* > 0$, *then there is some constant* $c > 0$ *such that*

$$\mathbf{L}_\alpha^* \left(\Theta_{\sigma_0}(k_1, s), \Theta_{\sigma_0}(k_2, s), \widehat{\beta}, \ell_2 \right) \geq c \min \left\{ k_2 \frac{\log p}{n}, \max \left\{ k_1 \frac{\log p}{n}, \frac{1}{\sqrt{n}} \right\} \right\} \sigma_0^2 \tag{7.3}$$

and

$$\mathbf{L}_\alpha^* \left(\Theta_{\sigma_0}(k_i, s), \widehat{\beta}, \ell_2 \right) \geq c \frac{k_i \log p}{n} \sigma_0^2 \quad \text{and} \quad i = 1, 2. \tag{7.4}$$

In particular, if $p \geq n$ *and* $\widehat{\beta}$ *is constructed based on the subsample* $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ *and satisfies Assumption (A) with* $\delta > 2$, *the above lower bounds can be attained.*

In contrast to Theorems 3 and 4, the lower bounds for the case $q = 2$ change in the absence of the prior knowledge $\Sigma = \mathrm{I}$ but the possibility of adaptivity of confidence intervals over $\Theta_{\sigma_0}(k, s)$ is similar to that over $\Theta_0(k)$. Since the Lasso estimator $\widehat{\beta}^L$ defined in (2.10) with $A > 4\sqrt{2}$ satisfies Assumption (A) with $\delta > 2$, by Theorem 10, the minimax lower bounds (7.3) and (7.4) can be attained for $\widehat{\beta}^L$. For $\widehat{\beta}^L$, only when $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim$

$\frac{n}{\log p}$, $\mathbf{L}_\alpha^*\left(\Theta_{\sigma_0}(k_1,s),\widehat{\beta}^L,\ell_2\right) \asymp \mathbf{L}_\alpha^*\left(\Theta_{\sigma_0}(k_1,s),\Theta_{\sigma_0}(k_2,s),\widehat{\beta},\ell_2\right) \asymp \frac{k_1\log p}{n}$ and adaptation between $\Theta_{\sigma_0}(k_1,s)$ and $\Theta_{\sigma_0}(k_2,s)$ is possible. In other regimes, if $k_1 \ll k_2$, then $\mathbf{L}_\alpha^*\left(\Theta_{\sigma_0}(k_1,s),\widehat{\beta}^L,\ell_2\right) \ll \mathbf{L}_\alpha^*\left(\Theta_{\sigma_0}(k_1,s),\Theta_{\sigma_0}(k_2,s),\widehat{\beta},\ell_2\right)$ and adaptation between $\Theta_{\sigma_0}(k_1,s)$ and $\Theta_{\sigma_0}(k_2,s)$ is impossible. For reasons of space, more discussion on $\Theta_{\sigma_0}(k,s)$, including the construction of adaptive confidence intervals over the regime $\frac{\sqrt{n}}{\log p} \lesssim k_1 \le k_2 \lesssim \frac{n}{\log p}$, is postponed to the supplement [6].

8. Minimax lower bounds for estimating $\|\beta\|_q^2$ with $1 \le q \le 2$. The lower bounds developed in this paper have broader implications. In particular, the established results imply the minimax lower bounds for estimating $\|\beta\|_q^2$ and the expected length of confidence intervals for $\|\beta\|_q^2$ with $1 \le q \le 2$. To build the connection, it is sufficient to note that the trivial estimator $\widehat{\beta} = 0$ satisfies Assumptions (A1) and (A2) with $\beta^* = 0$. Then we can apply the lower bounds (2.8), (2.9) and (2.13) to the estimator $\widehat{\beta} = 0$ and establish the minimax lower bounds of estimating $\|\beta\|_q^2$,

$$\inf_{\widehat{L}_2}\sup_{\theta\in\Theta_0(k)}\mathbb{P}_\theta\left(|\widehat{L}_2 - \|\beta\|_2^2| \ge c\min\left\{k\frac{\log p}{n},\frac{1}{\sqrt{n}}\right\}\sigma_0^2\right) \ge \delta; \tag{8.1}$$

$$\inf_{\widehat{L}_q}\sup_{\theta\in\Theta_0(k)}\mathbb{P}_\theta\left(|\widehat{L}_q - \|\beta\|_q^2| \ge ck^{\frac{2}{q}}\frac{\log p}{n}\sigma_0^2\right) \ge \delta, \quad \text{for } 1 \le q < 2, \tag{8.2}$$

$$\inf_{\widehat{L}_q}\sup_{\theta\in\Theta(k)}\mathbb{P}_\theta\left(|\widehat{L}_q - \|\beta\|_q^2| \ge ck^{\frac{2}{q}}\frac{\log p}{n}\right) \ge \delta, \quad \text{for } 1 \le q \le 2, \tag{8.3}$$

for some constants $\delta > 0$ and $c > 0$. Similarly, all the lower bounds for the expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ established in Theorem 3 to Theorem 7 imply corresponding lower bounds for $\|\beta\|_q^2$. The lower bound $\min\{k\frac{\log p}{n},\frac{1}{\sqrt{n}}\}\sigma_0^2$ in (8.1) is the same as the detection boundary in the sparse linear regression for the case $\Sigma = \mathrm{I}$ and $\sigma = 1$; See [19] and [1] for more details. Estimation of $\|\beta\|_2^2$ in high-dimensional linear regression has been considered in [17] under the general setting where Σ and σ are unknown and the lower bound (8.3) with $q = 2$ leads to one key component of the lower bound $ck\frac{\log p}{n}$ for estimating $\|\beta\|_2^2$.

9. Proofs. This section presents the proofs of the lower bound results. We first establish the general lower bound result, Theorem 8, in Section 9.1.

By applying Theorems 8 and 9, we prove Theorems 4 and 6 in Section 9.2. For reasons of space, the proofs of other main results, Theorems 1, 2, 3, 5, 7, 9, 10 as well as Propositions 1, 2, 3, 4, 5, 6 and the proofs of technical lemmas are postponed to the supplement [6].

We define the χ^2 distance between two density functions f_1 and f_0 by $\chi^2(f_1, f_0) = \int \frac{(f_1(z)-f_0(z))^2}{f_0(z)}dz = \int \frac{f_1^2(z)}{f_0(z)}dz - 1$, and it is well known that

$$L_1(f_1, f_0) \le \sqrt{\chi^2(f_1, f_0)}. \tag{9.1}$$

We follow the same notation used in Section 6. Let $\mathbb{P}_{Z,\theta\sim\pi}$ be the joint probability of Z and θ with the joint density function $f(\theta, z) = f_\theta(z)\,\pi(\theta)$. The following lemma, which is proved in the supplement [6], is needed in the proofs of Theorem 8 and Theorem 9.

LEMMA 1. *For any event $\mathcal{A}$, we have*

$$\mathbb{P}_\pi(Z \in \mathcal{A}) = \mathbb{P}_{Z,\theta\sim\pi}(Z \in \mathcal{A}), \tag{9.2}$$

$$|\mathbb{P}_{\pi_1}(Z \in \mathcal{A}) - \mathbb{P}_{\pi_2}(Z \in \mathcal{A})| \le L_1(f_{\pi_2}, f_{\pi_1}). \tag{9.3}$$

We will write $\mathbb{P}_\pi(\mathcal{A})$ and $\mathbb{P}_{Z,\theta\sim\pi}(\mathcal{A})$ for $\mathbb{P}_\pi(Z \in \mathcal{A})$ and $\mathbb{P}_{Z,\theta\sim\pi}(Z \in \mathcal{A})$ respectively. Recall that $\widehat{L}_q(Z)$ denotes a data-dependent loss estimator and $\beta(\theta)$ denotes the corresponding β of the parameter θ.

9.1. *Proof of Theorem 8.* We set $c_0 = \frac{1}{4}$ and $\alpha_1 = \frac{1}{10}$.

Proof of (6.2)

We assume

$$\mathbb{P}_{\theta_0}\left(\left|\widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta^*\|_q^2\right| \le \frac{1}{4}d^2\right) \ge 1 - \alpha_1. \tag{9.4}$$

Otherwise, we have

$$\mathbb{P}_{\theta_0}\left(\left|\widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta^*\|_q^2\right| \ge \frac{1}{4}d^2\right) \ge \alpha_1, \tag{9.5}$$

and hence (6.2) follows. Define the event

$$\mathcal{A}_0 = \left\{z : \|\widehat{\beta}(z) - \beta^*\|_q^2 \le c_0^2 d^2\ ,\ \left|\widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta^*\|_q^2\right| \le \frac{1}{4}d^2\right\}. \tag{9.6}$$

By (6.1) and (9.4), we have $\mathbb{P}_{\theta_0}(\mathcal{A}_0) \ge 1 - \alpha_0 - \alpha_1$. By (9.3), we obtain

$$\mathbb{P}_\pi(\mathcal{A}_0) \ge 1 - \alpha_0 - \alpha_1 - \int |f_{\theta_0}(z) - f_\pi(z)|\,dz. \tag{9.7}$$

For $z \in \mathcal{A}_0$ and $\theta \in \mathcal{F}$, by triangle inequality,

$$\|\widehat{\beta}(z) - \beta(\theta)\|_q \geq \left| \|\beta(\theta) - \beta^*\|_q - \|\widehat{\beta}(z) - \beta^*\|_q \right| \geq (1 - c_0)\, d. \tag{9.8}$$

For $z \in \mathcal{A}_0$ and $\theta \in \mathcal{F}$, then $\left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 \right| \geq \left| \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 - \|\widehat{\beta}(z) - \beta^*\|_q^2 \right| - \left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta^*\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4})d^2$, where the first inequality follows from triangle inequality and the last inequality follows from (9.6) and (9.8). Hence, for $z \in \mathcal{A}_0$, we obtain

$$\inf_{\theta \in \mathcal{F}} \left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4})d^2. \tag{9.9}$$

Note that $\sup_{\theta \in \mathcal{F}} \mathbb{P}_\theta \left(\left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4})d^2 \right) \geq \sup_{\theta \in \mathcal{F}} \mathbb{P}_\theta \left(\inf_{\theta \in \mathcal{F}} \left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4})d^2 \right)$. Since the max risk is lower bounded by the Bayesian risk, we can further lower bound the last term by $\mathbb{P}_\pi \left(\inf_{\theta \in \mathcal{F}} \left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4})d^2 \right)$. Combined with (9.9), we establish

$$\sup_{\theta \in \mathcal{F}} \mathbb{P}_\theta \left(\left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4})d^2 \right) \geq \mathbb{P}_\pi(\mathcal{A}_0). \tag{9.10}$$

Combining (9.5), (9.7) and (9.10), we establish (6.2).

Proof of (6.3)

For $\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, Z\right) \in \mathcal{I}_\alpha\left(\Theta, \widehat{\beta}, \ell_q\right)$, we have

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \left(\|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, Z\right) \right) \geq 1 - \alpha. \tag{9.11}$$

Define the event $\mathcal{A} = \left\{ z : \|\widehat{\beta}(z) - \beta^*\|_q < c_0 d, \ \ \|\widehat{\beta}(z) - \beta^*\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, L, z\right) \right\}$. By (6.1) and (9.11), we have $\mathbb{P}_{\theta_0}(\mathcal{A}) \geq 1 - \alpha - \alpha_0$. (9.2) and (9.3) imply

$$\mathbb{P}_{Z, \theta \sim \pi}(\mathcal{A}) = \mathbb{P}_\pi(\mathcal{A}) \geq 1 - \alpha - \alpha_0 - L_1(f_\pi, f_{\theta_0}). \tag{9.12}$$

Define the event $\mathcal{B}_\theta = \left\{ z : \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right) \right\}$ and $\mathcal{M} = \cup_{\theta \in \mathcal{F}} \mathcal{B}_\theta$. By (9.11), we have

$$\mathbb{P}_{Z, \theta \sim \pi}(\mathcal{M}) = \int \left(\int 1_{z \in \mathcal{M}} f_\theta(z) dz \right) \pi(\theta)\, d\theta \geq \int \left(\int 1_{z \in \mathcal{B}_\theta} f_\theta(z) dz \right) \pi(\theta)\, d\theta \geq 1 - \alpha.$$

Combined with (9.12), we have $\mathbb{P}_{Z, \theta \sim \pi}(\mathcal{A} \cap \mathcal{M}) \geq 1 - 2\alpha - \alpha_0 - L_1(f_\pi, f_{\theta_0})$. For $z \in \mathcal{M}$, there exists $\bar{\theta} \in \mathcal{F}$ such that $\|\widehat{\beta}(z) - \beta(\bar{\theta})\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right)$;

For $z \in \mathcal{A}$, we have $\|\widehat{\beta}(z) - \beta^*\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right)$ and $\|\widehat{\beta}(z) - \beta^*\|_q < c_0 d$. Hence, for $z \in \mathcal{A} \cap \mathcal{M}$, we have $\|\widehat{\beta}(z) - \beta(\bar{\theta})\|_q^2, \|\widehat{\beta}(z) - \beta^*\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right)$ and $\|\widehat{\beta}(z) - \beta(\bar{\theta})\|_q \geq \|\beta(\bar{\theta}) - \beta^*\|_q - \|\widehat{\beta}(z) - \beta^*\|_q \geq (1 - c_0)\, d$ and hence

$$\mathbf{L}\left(\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right)\right) \geq (1 - 2c_0)\, d^2. \tag{9.13}$$

Define the event $\mathcal{C} = \left\{z : \mathbf{L}\left(\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right)\right) \geq (1 - 2c_0)\, d^2\right\}$. By (9.13), we have $\mathbb{P}_\pi(\mathcal{C}) = \mathbb{P}_{Z,\theta\sim\pi}(\mathcal{C}) \geq \mathbb{P}_{Z,\theta\sim\pi}(\mathcal{A} \cap \mathcal{M}) \geq 1 - 2\alpha - \alpha_0 - L_1(f_\pi, f_{\theta_0})$. By (9.3), we establish $\mathbb{P}_{\theta_0}(\mathcal{C}) \geq 1 - 2\alpha - \alpha_0 - 2L_1(f_\pi, f_{\theta_0})$ and hence (6.3).

9.2. *Proof of Theorems 4 and 6.* We first specify some constants used in the proof. Let C^* be given in (2.6). Define $\epsilon_1 = \frac{1-2\alpha-2\alpha_0}{12}$ and

$$c_0 = \min\left\{\frac{1}{2}, 32\log\left(1+\epsilon_1^2\right), \frac{2}{3}\sqrt{\log(1+\epsilon_1^2)}, \frac{1-2\gamma}{16C^*}, \left(\frac{1-2\gamma}{16C^*}\right)^2\right\},\ c_0^* = \min\left\{c_0, \frac{\sqrt{M_1}-1}{C^*M_1 + \sqrt{M_1}-1}\right\}. \tag{9.14}$$

Theorems 4 and 6 follow from Theorem 11 below.

THEOREM 11. *Suppose $0 < \alpha < \frac{1}{4}$, $1 \leq q \leq 2$ and the sparsity levels k_1, k_2 and k_0 satisfy Assumption* (B2). *Suppose that $\widehat{\beta}$ satisfies Assumption* (A2) *with $\|\beta^*\|_0 \leq k_0$.*

1. *If $k_2 \lesssim \frac{\sqrt{n}}{\log p}$, then there is some constant $c > 0$ such that*

$$\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right) \geq c k_2^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2. \tag{9.15}$$

2. *If $\frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$, then there is some constant $c > 0$ such that*

$$\begin{aligned}&\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right)\\ &\geq c \max\left\{\left((1-c_2)^2 k_2^{\frac{2}{q}-1} k_1 \frac{\log p}{n} - (1+c_1)^2 k_1^{\frac{2}{q}} \frac{\log p}{n}\right)_+, \frac{k_2^{\frac{2}{q}-1}}{\sqrt{n}}\right\} \sigma_0^2,\end{aligned} \tag{9.16}$$

where $c_1 = \frac{C^* k_0^{\frac{1}{q}}}{(k_1-k_0)^{\frac{1}{q}}}$ and $c_2 = \frac{C^* k_0^{\frac{1}{q}}}{(k_2-k_0)^{\frac{1}{q}-\frac{1}{2}}(k_1-k_0)^{\frac{1}{2}}}$.

In particular, the minimax lower bound (9.15) *and the term $\frac{k_2^{\frac{2}{q}-1}}{\sqrt{n}}\sigma_0^2$ in* (9.16) *can be established under the weaker assumption* (A1) *with $\|\beta^*\|_0 \leq k_0$.*

By Theorem 11, we establish (3.7) in Theorem 4 and (3.18) in Theorem 6. In the regime $k_2 \lesssim \frac{\sqrt{n}}{\log p}$, the lower bound (3.7) for $q = 2$ and (3.18) for $1 \le q < 2$ follow from (9.15). For the case $q = 2$, in the regime $\frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$, the first term of the right hand side of (9.16) is 0 while the second term is $\frac{1}{\sqrt{n}}\sigma_0^2$, which leads to (3.7). For $1 \le q < 2$, let $k_1^* = \min\{k_1, \zeta_0 k_2\}$ for some constant $0 < \zeta_0 < 1$, an application of (9.16) leads to $\mathbf{L}_\alpha^*\left(\Theta_0\left(k_1^*\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right) \ge c \max\left\{k_2^{\frac{2}{q}-1} k_1^* \frac{\log p}{n}, \frac{k_2^{\frac{2}{q}-1}}{\sqrt{n}}\right\}\sigma_0^2$. By this result, if $k_1 \le \zeta_0 k_2$, then $k_1^* = k_1$ and the lower bounds (3.18) in the regions $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ and $\frac{\sqrt{n}}{\log p} \lesssim k_1 \le k_2 \lesssim \frac{n}{\log p}$ follow; if $\zeta_0 k_2 < k_1 \le k_2$, then $k_1^* = \zeta_0 k_2 \ge \zeta_0 k_1$. By the fact that $\mathbf{L}_\alpha^*\left(\Theta_0\left(k_1\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right) \ge \mathbf{L}_\alpha^*\left(\Theta_0\left(k_1^*\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right)$, the lower bounds (3.18) over the regions $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ and $\frac{\sqrt{n}}{\log p} \lesssim k_1 \le k_2 \lesssim \frac{n}{\log p}$ follow. The following lemma shows that (3.7) holds for $\widehat{\beta}^L$ defined in (2.10) with $A > \sqrt{2}$ by verifying Assumption (A1) and (3.18) holds for $\widehat{\beta}^L$ defined in (2.10) with $A > 4\sqrt{2}$ by verifying Assumption (A2). Its proof can be found in the supplement [6].

LEMMA 2. *If $A > 4\sqrt{2}$, then we have*

$$\sup_{\{\theta=(\beta^*,\mathrm{I},\sigma):\sigma\le 2\sigma_0\}} \mathbb{P}_\theta\left(\|\widehat{\beta}^L - \beta^*\|_q^2 \ge C\|\beta^*\|_0^{\frac{2}{q}} \frac{\log p}{n}\sigma^2\right) \le c\exp\left(-c'n\right) + p^{-c}.$$

In particular, the above result holds for $q = 2$ under the assumption $A > \sqrt{2}$.

References.

[1] Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

[2] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on*, 58(4):1997–2017, 2012.

[3] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[5] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[6] T Tony Cai and Zijian Guo. Supplement to "accuracy assessment for high-dimensional linear regression". 2016.

[7] T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, to appear.

[8] T Tony Cai and Mark G Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840, 2005.

[9] T Tony Cai and Mark G Low. Adaptive confidence balls. *The Annals of Statistics*, 34(1):202–228, 2006.
[10] T Tony Cai, Mark G Low, and Zongming Ma. Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association*, 109:1054–1070, 2014.
[11] T. Tony Cai and Harrison H Zhou. A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*, 37(2):569–595, 2009.
[12] Emmanuel Candès and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.
[13] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. 2015.
[14] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *arXiv preprint arXiv:1501.03430*, 2015.
[15] David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.
[16] David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941, 2011.
[17] Zijian Guo, Wanjie Wang, T Tony Cai, and Hongzhe Li. Optimal estimation of co-heritability in high-dimensional linear models. *arXiv preprint arXiv:1605.07244*, 2016.
[18] Marc Hoffmann and Richard Nickl. On adaptive inference and confidence bands. *The Annals of Statistics*, 39(5):2383–2409, 2011.
[19] Yuri I Ingster, Alexandre B Tsybakov, and Nicolas Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010.
[20] Lucas Janson, Rina Foygel Barber, and Emmanuel Candès. Eigenprism: Inference for high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1505.02097*, 2015.
[21] Ker-Chau Li. From stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, 13(4):1352–1377, 1985.
[22] Richard Nickl and Sara van de Geer. Confidence sets in sparse regression. *The Annals of Statistics*, 41(6):2852–2876, 2013.
[23] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.
[24] James Robins and Aad Van Der Vaart. Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253, 2006.
[25] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.
[26] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101(2):269–284, 2012.
[27] Christos Thrampoulidis, Ashkan Panahi, and Babak Hassibi. Asymptotically exact error analysis for the generalized ℓ_2^2-lasso. *arXiv preprint arXiv:1502.06287*, 2015.
[28] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.
[29] Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.
[30] Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phe-

nomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[31] Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

[32] Feng Yi and Hui Zou. SURE-tuned tapering estimation of large covariance matrices. *Computational Statistics & Data Analysis*, 58:339–351, 2013.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu
zijguo@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/∼tcai/