cemmap
centre for microdata methods and practice

On cross-validated Lasso

Denis Chetverikov
Zhipeng Liao

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP47/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

On Cross-Validated Lasso*

Denis Chetverikov† Zhipeng Liao‡

Abstract

In this paper, we derive a rate of convergence of the Lasso estimator when the penalty parameter λ for the estimator is chosen using K-fold cross-validation; in particular, we show that in the model with Gaussian noise and under fairly general assumptions on the candidate set of values of λ, the prediction norm of the estimation error of the cross-validated Lasso estimator is with high probability bounded from above up-to a constant by $(s \log p/n)^{1/2} \cdot (\log^{7/8} n)$ as long as $p \log n/n = o(1)$ and some other mild regularity conditions are satisfied, where n is the sample size of available data, p is the number of covariates, and s is the number of non-zero coefficients in the model. Thus, the cross-validated Lasso estimator achieves the fastest possible rate of convergence up-to the logarithmic factor $\log^{7/8} n$. In addition, we derive a sparsity bound for the cross-validated Lasso estimator; in particular, we show that under the same conditions as above, the number of non-zero coefficients of the estimator is with high probability bounded from above up-to a constant by $s \log^5 n$. Finally, we show that our proof technique generates non-trivial bounds on the prediction norm of the estimation error of the cross-validated Lasso estimator even if p is much larger than n and the assumption of Gaussian noise fails; in particular, the prediction norm of the estimation error is with high-probability bounded from above up-to a constant by $(s \log^2(pn)/n)^{1/4}$ under mild regularity conditions.

1 Introduction

Machine learning techniques are gradually making their way into economics; see NBER Summer Institute Lectures Chernozhukov et al. (2013) and Athey and Imbens (2015). Using these techniques, for example, Cesarini et al. (2009) analyzed genetic factors of social preferences, Belloni and Chernozhukov (2011) found country characteristics associated with long-run growth in the cross-county growth study, Saiz and Simonsohn (2013) constructed corruption measures by country and by US state. Belloni et al. (2013) and Wager and Athey (2015) developed machine-learning-type techniques for estimating heterogeneous treatment effects.

*This version: August 24, 2016. We are extremely thankful to Victor Chernozhukov for posing the research question and for many helpful discussions. We also thank Moshe Buchinsky, Matias Cattaneo, and Rosa Matzkin for useful comments.

†Department of Economics, UCLA, Bunche Hall, 8283, 315 Portola Plaza, Los Angeles, CA 90095, USA; E-Mail address: chetverikov@econ.ucla.edu

‡Department of Economics, UCLA, Bunche Hall, 8283, 315 Portola Plaza, Los Angeles, CA 90095, USA; E-Mail address: zhipeng.liao@econ.ucla.edu

The most popular machine learning technique in econometrics is certainly the Lasso estimator. Since its invention by Tibshirani (1996), large number of papers have studied its properties. Many of these papers have been concerned with the choice of the penalty parameter λ required for the implementation of the Lasso estimator. As a result, several methods to choose λ have been developed and theoretically justified; see, for example, Zou et al. (2007), Bickel et al. (2009), and Belloni and Chernozhukov (2013). However, in practice researchers often rely upon cross-validation to choose λ (see Bülmann and van de Geer (2011), Hastie, Tibshirani, and Wainwright (2015), and Chatterjee and Jafarov (2015) for examples), and to the best of our knowledge, there exist few results in the literature about properties of the Lasso estimator when λ is chosen using cross-validation; see a review of existing results below. The purpose of this paper is to fill this gap and to derive a rate of convergence of the cross-validated Lasso estimator.

We consider the regression model

$$Y = X'\beta + \varepsilon, \quad \mathrm{E}[\varepsilon \mid X] = 0, \tag{1}$$

where Y is a dependent variable, $X = (X_1, \ldots, X_p)'$ a p-vector of covariates, ε unobserved scalar noise, and $\beta = (\beta_1, \ldots, \beta_p)'$ a p-vector of coefficients. Assuming that a random sample of size n, $(X_i, Y_i)_{i=1}^n$, from the distribution of the pair (X, Y) is available, we are interested in estimating the vector of coefficients β. We consider triangular array asymptotics, so that the distribution of the pair (X, Y), and in particular the dimension p of the vector X, is allowed to depend on n. For simplicity of notation, however, we keep this dependence implicit.

We assume that the vector of coefficients β is sparse in the sense that $s = s_n = \|\beta\|_0 = \sum_{j=1}^p 1\{\beta_j \neq 0\}$ is (potentially much) smaller than p. Under this assumption, the effective way to estimate β was introduced by Tibshirani (1996) who suggested the Lasso estimator:

$$\widehat{\beta}(\lambda) \in \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left(\frac{1}{n} \sum_{i=1}^n (Y_i - X_i'b)^2 + \lambda \|b\|_1 \right), \tag{2}$$

where for $b = (b_1, \ldots, b_p)' \in \mathbb{R}^p$, $\|b\|_1 = \sum_{j=1}^p |b_j|$ denotes the L^1 norm of b, and λ is some penalty parameter (the estimator suggested in Tibshirani's paper takes a slightly different form but over the time the version (2) has become more popular, probably for computational reasons). Whenever the solution of the optimization problem in (2) is not unique, we assume for concreteness that one solution is chosen according to some pre-specified rule; in particular, we assume that a solution with the smallest number of non-zero components is selected.

To perform the Lasso estimator $\widehat{\beta}(\lambda)$, one has to choose the penalty parameter λ. If λ is chosen appropriately, the Lasso estimator is consistent with $(s \log p/n)^{1/2}$ rate of convergence in the prediction norm under fairly general conditions; see, for example, Bickel et al. (2009) or Belloni and Chernozhukov (2011). On the other hand, if λ is not chosen appropriately, the Lasso estimator may not be consistent or may have slower rate of convergence; see Chatterjee (2014). Therefore, it is important to select λ appropriately. In practice, it is often recommended to choose λ using cross-validation as described in the next section. In this paper, we analyze properties of the Lasso estimator $\widehat{\beta}(\lambda)$ when $\lambda = \widehat{\lambda}$ is chosen using (K-fold) cross-validation and

in particular, we demonstrate that under certain mild regularity conditions, if the conditional distribution of ε given X is Gaussian and $p \log n/n = o(1)$, then

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n} \lesssim \left(\frac{s \log p}{n}\right)^{1/2} \cdot (\log^{7/8} n) \tag{3}$$

with probability $1 - o(1)$ up-to some constant C, where for $b = (b_1, \ldots, b_p)' \in \mathbb{R}^p$, $\|b\|_{2,n} = (n^{-1} \sum_{i=1}^n (X_i' b)^2)^{1/2}$ denotes the prediction norm of b. Thus, under our conditions, the cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ achieves the fastest possible rate of convergence in the prediction norm up-to the logarithmic factor $\log^{7/8} n$. We do not know whether this logarithmic factor can or can not be dropped.

Under the same conditions as above, we also derive a sparsity bound for the cross-validated Lasso estimator; in particular, we show that

$$\|\widehat{\beta}(\widehat{\lambda})\|_0 \lesssim s \log^5 n$$

with probability $1 - o(1)$ up-to some constant C. Moreover, we demonstrate that our proof technique generates a non-trivial rate of convergence in the prediction norm for the cross-validated Lasso estimator even if p is (potentially much) larger than n (high-dimensional case) and the Gaussian assumption fails. Because some steps used to derive (3) do not apply, however, the rate turns out to be sub-optimal, and our bound is probably not sharp in this case. Nonetheless, we are hopeful that our proof technique will help to derive the sharp bound for the non-Gaussian high-dimensional case in the future.

Given that cross-validation is often used to choose the penalty parameter λ for the Lasso estimator and given how popular the Lasso estimator is, deriving a rate of convergence of the cross-validated Lasso estimator is an important question in the literature; see, for example, Chatterjee and Jafarov (2015), where further motivation for the topic is provided. Yet, to the best of our knowledge, the only results in the literature about cross-validated Lasso estimator are due to Homrighausen and McDonald (2013a,b, 2014). Homrighausen and McDonald (2013a) showed that if the penalty parameter is chosen using K-fold cross-validation from a range of values determined by their techniques, the Lasso estimator is risk consistent, which under our conditions is equivalent to consistency in the L^2 norm. Homrighausen and McDonald (2014) derived a similar result for leave-one-out cross-validation. Homrighausen and McDonald (2013b) derived a rate of convergence of the cross-validated Lasso estimator that depends on n via $n^{-1/4}$ but they substantially restricted the range of values over which cross-validation search is performed. These are useful results but we emphasize that in practice the cross-validation search is often conducted over a fairly large set of values of the penalty parameter, which could potentially be much larger than required in their results. In contrast, we derive a rate of convergence that depends on n via $n^{-1/2}$, and we impose only minor conditions on the range of values of λ used by cross-validation.

Other papers that have been concerned with cross-validation in the context of the Lasso estimator include Chatterjee and Jafarov (2015) and Lecué and Mitchell (2012). Chatterjee

and Jafarov (2015) developed a novel cross-validation-type procedure to choose λ and showed that the Lasso estimator based on their choice of λ has a rate of convergence depending on n via $n^{-1/4}$. Their procedure to choose λ, however, is related to but different from the classical cross-validation procedure used in practice. Lecué and Mitchell (2012) studied classical cross-validation but focused on estimators that differ from the Lasso estimator in important ways. For example, one of the estimators they considered is the average of subsample Lasso estimators, $K^{-1}\sum_{k=1}^{K}\widehat{\beta}_{-k}(\lambda)$, for $\widehat{\beta}_{-k}(\lambda)$ defined in (4) in the next section. Although the authors studied properties of cross-validated version of such estimators in great generality, it is not immediately clear how to apply their results to obtain bounds for the cross-validated Lasso estimator itself.

We emphasize that deriving a rate of convergence of the cross-validated Lasso estimator is a non-standard problem. In particular, classical techniques to derive properties of cross-validated estimators developed for example in Li (1987) do not apply to the Lasso estimator as those techniques are based on the linearity of the estimators in the vector of dependent variables $(Y_1,\dots,Y_n)'$, which does not hold in the case of the Lasso estimator. More recent techniques, developed for example in Wegkamp (2003), help to analyze sub-sample Lasso estimators like those studied in Lecué and Mitchell (2012) but are not sufficient for the analysis of the full-sample Lasso estimator. See Arlot and Celisse (2010) for an extensive review of results on cross-validation available in the literature.

The rest of the paper is organized as follows. In the next section, we describe the cross-validation procedure. In Section 3, we state our regularity conditions. In Section 4, we present our main results. In Section 5, we describe results of our simulation experiments. In Section 6, we provide proofs of the main results. In Section 7, we give some technical lemmas that are useful for the proofs of the main results.

Notation. Throughout the paper, we use the following notation. For any vector $b=(b_1,\dots,b_p)'\in\mathbb{R}^p$, we use $\|b\|_0=\sum_{j=1}^{p}1\{b_j\neq 0\}$ to denote the number of non-zero components of b, $\|b\|_1=\sum_{j=1}^{p}|b_j|$ to denote its L^1 norm, $\|b\|=(\sum_{j=1}^{p}b_j^2)^{1/2}$ to denote its L^2 norm (the Euclidean norm), $\|b\|_\infty=\max_{1\leq j\leq p}|b_j|$ to denote its L^∞ norm, and $\|b\|_{2,n}=(n^{-1}\sum_{i=1}^{n}(X_i'b)^2)^{1/2}$ to denote its prediction norm. In addition, we use the notation $a_n\lesssim b_n$ if $a_n\leq Cb_n$ for some constant C that is independent of n. Moreover, we use $\mathcal{S}^p$ to denote the unit sphere in $\mathbb{R}^p$, that is, $\mathcal{S}^p=\{\delta\in\mathbb{R}^p\colon\|\delta\|=1\}$. Further, for any matrix $A\in\mathbb{R}^{p\times p}$, we use $\|A\|=\sup_{x\in\mathcal{S}^p}\|Ax\|$ to denote its spectral norm. Also, with some abuse of notation, we use X_j to denote the jth component of the vector $X=(X_1,\dots,X_p)'$ and we use X_i to denote the ith realization of the vector X in the random sample $(X_i,Y_i)_{i=1}^n$ from the distribution of the pair (X,Y). Finally, for any finite set S, we use $|S|$ to denote the number of elements in S. We introduce more notation in the beginning of Section 6, as required for the proofs in the paper.

2 Cross-Validation

As explained in the Introduction, to choose the penalty parameter λ for the Lasso estimator $\widehat{\beta}(\lambda)$, it is common practice to use cross-validation. In this section, we describe the procedure in details. Let K be some strictly positive (typically small) integer, and let $(I_k)_{k=1}^K$ be a partition of the set $\{1,\dots,n\}$; that is, for each $k \in \{1,\dots,K\}$, I_k is a subset of $\{1,\dots,K\}$, for each $k, k' \in \{1,\dots,K\}$ with $k \neq k'$, the sets I_k and $I_{k'}$ have empty intersection, and $\cup_{k=1}^K I_k = \{1,\dots,n\}$. For our asymptotic analysis, we will assume that K is a constant that does not depend on n. Further, let Λ_n be a set of candidate values of λ. Now, for $k = 1,\dots,K$ and $\lambda \in \Lambda_n$, let

$$\widehat{\beta}_{-k}(\lambda) \in \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left(\frac{1}{n - n_k} \sum_{i \notin I_k} (Y_i - X_i' b)^2 + \lambda \|b\|_1 \right) \tag{4}$$

be the Lasso estimator corresponding to all observations excluding those in I_k where $n_k = |I_k|$ is the size of the subsample I_k. As in the case with the full-sample Lasso estimator $\widehat{\beta}(\lambda)$ in (2), whenever the optimization problem in (4) has multiple solutions, we choose one with the smallest number of non-zero components. Then the cross-validation choice of λ is

$$\widehat{\lambda} = \operatorname*{arg\,min}_{\lambda \in \Lambda_n} \sum_{k=1}^{K} \sum_{i \in I_k} (Y_i - X_i' \widehat{\beta}_{-k}(\lambda))^2. \tag{5}$$

The cross-validated Lasso estimator in turn is $\widehat{\beta}(\widehat{\lambda})$. In the literature, the procedure described here is also often referred to as K-fold cross-validation. For brevity, however, we simply refer to it as cross-validation. Below we will study properties of $\widehat{\beta}(\widehat{\lambda})$.

3 Regularity Conditions

Recall that we consider the model given in (1), the Lasso estimator $\widehat{\beta}(\lambda)$ given in (2), and the cross-validation choice of λ given in (5). Let c_1, C_1, a, and q be some strictly positive numbers where $a < 1$ and $q > 4$. Also, let $(\xi_n)_{n \geq 1}$, $(\gamma_n)_{n \geq 1}$, and $(\Gamma_n)_{n \geq 1}$ be sequences of positive numbers, possibly growing to infinity. To derive our results, we will impose the following regularity conditions.

Assumption 1 (Covariates)**.** *The random vector $X = (X_1,\dots,X_p)'$ is such that we have $c_1 \leq (\mathrm{E}[|X'\delta|^2])^{1/2} \leq C_1$ and $(\mathrm{E}[|X'\delta|^4])^{1/4} \leq \Gamma_n$ for all $\delta \in \mathcal{S}^p$. In addition, $\max_{1 \leq j \leq p} (\mathrm{E}[|X_j|^4])^{1/4} \leq \gamma_n$ and $nP(\|X\| > \xi_n) = o(1)$.*

The first part of Assumption 1 means that all eigenvalues of the matrix $\mathrm{E}[XX']$ are bounded from above and below from zero. The second part of this assumption, that is, the condition that $(\mathrm{E}[|X'\delta|^4])^{1/4} \leq \Gamma_n$ for all $\delta \in \mathcal{S}^p$, is often assumed in the literature with $\Gamma_n \lesssim 1$; see Mammen (1993) for an example. To develop some intuition about this and other parts of Assumption 1, we consider three examples.

Example 1 (Gaussian independent covariates)**.** Suppose that the vector X consists of independent standard Gaussian random variables. Then for all $\delta \in \mathcal{S}^p$, the random variable $X'\delta$ is standard Gaussian as well, and so the condition that $(\mathrm{E}[|X'\delta|^4])^{1/4} \leq \Gamma_n$ for all $\delta \in \mathcal{S}^p$ is satisfied with $\Gamma_n = 3^{1/4}$. Similarly, the condition that $\max_{1\leq j\leq p}(\mathrm{E}[|X_j|^4])^{1/4} \leq \gamma_n$ holds with $\gamma_n = 3^{1/4}$. In addition, $\|X\|^2$ is a chi-square random variable with p degrees of freedom in this case, and so for all $t > 0$, we have $P(\|X\|^2 > p + 2\sqrt{pt} + 2t) \leq e^{-t}$; see, for example, Section 2.4 and Example 2.7 in Boucheron, Lugosi, and Massart (2013). Setting $t = 2\log n$ in this inequality shows that the condition that $nP(\|X\| > \xi_n) = o(1)$ is satisfied with $\xi_n = (2p + 6\log n)^{1/2}$. ∎

Example 2 (Bounded independent covariates)**.** Suppose that the vector X consists of independent zero-mean bounded random variables. In particular, suppose for simplicity that $\max_{1\leq j\leq p}|X_j| \leq 1$ almost surely. Then for all $t > 0$ and $\delta \in \mathcal{S}^p$, we have $P(|X'\delta| > t) \leq 2\exp(-t^2/2)$ by Hoeffding's inequality. Therefore, the condition that $(\mathrm{E}[|X'\delta|^4])^{1/4} \leq \Gamma_n$ for all $\delta \in \mathcal{S}^p$ is satisfied with $\Gamma_n = 2$ by the standard calculations. Also, the condition that $\max_{1\leq j\leq p}(\mathrm{E}[|X_j|^4])^{1/4} \leq \gamma_n$ is satisfied with $\gamma_n = 1$, and the condition that $nP(\|X\| > \xi_n) = o(1)$ is satisfied with $\xi_n = p^{1/2}$. ∎

Example 3 (Bounded non-independent covariates)**.** Suppose that the vector X consists of not necessarily independent bounded random variables. In particular, suppose for simplicity that $\max_{1\leq j\leq p}|X_j| \leq 1$ almost surely. Then the condition that $(\mathrm{E}[|X'\delta|^4])^{1/4} \leq \Gamma_n$ for all $\delta \in \mathcal{S}^p$ is satisfied with $\Gamma_n = C_1^{1/2}p^{1/4}$ since $\mathrm{E}[(X'\delta)^4] \leq \mathrm{E}[(X'\delta)^2\|X\|^2\|\delta\|^2] \leq p\mathrm{E}[(X'\delta)^2] \leq C_1^2 p$. Also, like in Example 2, the conditions that $\max_{1\leq j\leq p}(\mathrm{E}[|X_j|^4])^{1/4} \leq \gamma_n$ and that $nP(\|X\| > \xi_n) = o(1)$ are satisfied with $\gamma_n = 1$ and $\xi_n = p^{1/2}$. ∎

Assumption 2 (Noise)**.** *We have* $c_1 \leq \mathrm{E}[\varepsilon^2 \mid X] \leq C_1$ *almost surely.*

This assumption means that the variance of the conditional distribution of ε given X is bounded from above and below from zero. The lower bound is needed to avoid potential super-efficiency of the Lasso estimator. Such bounds are typically imposed in the literature.

Assumption 3 (Growth conditions)**.** *We have* $M_n^2 s(\log^4 n)(\log p)/n^{1-2/q} = o(1)$ *where* $M_n = (\mathrm{E}[\|X\|_\infty^q])^{1/q}$*. In addition,* $\gamma_n^4 s^2 \log p/n = o(1)$ *and* $\Gamma_n^4(\log n)(\log\log n)^2/n = o(1)$*.*

Assumption 3 is a mild growth condition restricting some moments of X and also the number of non-zero coefficients in the model, s. In the remark below, we discuss conditions of this assumption in three examples given above.

Remark 1 (Growth conditions in Examples 1, 2, and 3)**.** In Example 1 above, this assumption reduces to the following conditions: (i) $s(\log n)^4(\log p)^2/n^{1-\epsilon} = o(1)$ for some constant $\epsilon > 0$ and (ii) $s^2\log p/n = o(1)$ since in this case, $M_n \leq C_q(\log p)^{1/2}$ for all $q > 4$ and some constant C_q that depends only on q. In Example 2, Assumption 3 reduces to the following conditions: (i) $s(\log n)^4(\log p)/n^{1-\epsilon} = o(1)$ for some constant $\epsilon > 0$ and (ii) $s^2\log p/n = o(1)$ since in this case, $M_n \leq 1$ for all $q > 4$. In Example 3, Assumption 3 reduces to the following conditions: (i)

$s^2 \log p/n = o(1)$ and (ii) $p(\log n)(\log\log n)/n = o(1)$. Indeed, under assumptions of Example 3, we have $M_n \leq 1$ for all $q > 4$, and so the condition that $M_n^2 s(\log^4 n)(\log p)/n^{1-2/q} = o(1)$ follows from the condition that $s(\log^4 n)(\log p)/n^{1-2/q} = o(1)$ but for q large enough, this condition follows from $s^2 \log p/n = o(1)$ and $p(\log n)(\log\log n)/n = o(1)$. Note that our conditions in Examples 1 and 2 allow for the high-dimensional case, where p is (potentially much) larger than n but conditions in Example 3 hold only in the moderate-dimensional case, where p is asymptotically smaller than n. ∎

Assumption 4 (Candidate set)**.** *The candidate set Λ_n takes the following form: $\Lambda_n = \{C_1 a^l : l = 0, 1, 2, \ldots;\ a^l \geq c_1/n\}$.*

It is known from Bickel et al. (2009) that the optimal rate of convergence of the Lasso estimator in the prediction norm is achieved when λ is of order $(\log p/n)^{1/2}$. Since under Assumption 3, we have $\log p = o(n)$, it follows that our choice of the candidate set Λ_n in Assumption 4 makes sure that there are some λ's in the candidate set Λ that would yield the Lasso estimator with the optimal rate of convergence in the prediction norm. Note also that Assumption 4 gives a rather flexible choice of the candidate set Λ_n of values of λ; in particular, the largest value, C_1, can be set arbitrarily large and the smallest value, c_1/n, converges to zero rather fast. In fact, the only two conditions that we need from Assumption 4 is that Λ_n contains a "good" value of λ, say $\bar{\lambda}_0$, such that the subsample Lasso estimators $\widehat{\beta}_{-k}(\bar{\lambda}_0)$ satisfy the bound (9) in Lemma 1 with probability $1 - o(1)$, and than $|\Lambda_n| \lesssim \log n$ up-to a constant that depend only on c_1 and C_1. Thus, we could for example set $\Lambda_n = \{a^l : l = \ldots, -2, -1, 0, 1, 2, \ldots;\ a^{-l} \leq n^{C_1},\ a^l \leq n^{C_1}\}$.

Assumption 5 (Dataset partition)**.** *For all $k = 1, \ldots, K$, we have $n_k/n \geq c_1$.*

Assumption 5 is mild and is typically imposed in the literature on K-fold cross-validation. This assumption ensures that all subsamples I_k are balanced and their sizes are of the same order.

4 Main Results

Recall that for $b \in \mathbb{R}^p$, we use $\|b\|_{2,n} = (n^{-1}\sum_{i=1}^n (X_i' b)^2)^{1/2}$ to denote the prediction norm of b. Our first main result in this paper derives a rate of convergence of the cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ in the prediction norm for the Gaussian case where $\xi_n^2 \log n/n = o(1)$. As explained in Remark 4 below, the last condition implies that this is a moderate-dimensional case, where p is asymptotically smaller than n.

Theorem 1 (Gaussian moderate-dimensional case)**.** *Suppose that Assumptions 1 – 5 hold. In addition, suppose that $\xi_n^2 \log n/n = o(1)$. Finally, suppose that the conditional distribution of ε given X is Gaussian. Then*

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n} \lesssim \left(\frac{s \log p}{n}\right)^{1/2} \cdot (\log^{7/8} n)$$

with probability $1 - o(1)$ up-to a constant depending only on c_1, C_1, K, a, and q.

Remark 2 (Near-optimality of cross-validated Lasso estimator)**.** Let σ be a constant such that $\mathrm{E}[\varepsilon^2 \mid X] \leq \sigma^2$ almost surely. The results in Bickel et al. (2009) imply that under assumptions of Theorem 1, setting $\lambda = \lambda^* = C\sigma(\log p/n)^{1/2}$ for sufficiently large constant C gives the Lasso estimator $\widehat{\beta}(\lambda^*)$ satisfying $\|\widehat{\beta}(\lambda^*) - \beta\|_{2,n} = O_P((s \log p/n)^{1/2})$, and it follows from Rigollet and Tsybakov (2011) that this is the optimal rate of convergence (in the minimax sense) for the estimators of β in the model (1). Therefore, Theorem 1 shows that the cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ has the fastest possible rate of convergence in the prediction norm up-to the logarithmic factor $\log^{7/8} n$. Note, however, that implementing the cross-validated Lasso estimator does not require knowledge of σ, which makes this estimator attractive in practice. The rate of convergence established in Theorem 1 is also very close to the oracle rate of convergence, $(s/n)^{1/2}$, that could be achieved by the OLS estimator if we knew the set of covariates X_j having non-zero coefficient β_j; see, for example, Belloni et al. (2015a). ∎

Remark 3 (On the proof of Theorem 1)**.** One of the ideas in Bickel et al. (2009) is to show that outside of the event

$$\lambda < c \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \varepsilon_i \right|, \tag{6}$$

where $c > 2$ is some constant, the Lasso estimator $\widehat{\beta}(\lambda)$ satisfies the bound $\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \lesssim \lambda\sqrt{s}$. Thus, to obtain the Lasso estimator with fast rate of convergence, it suffices to choose λ such that λ is small enough but the event (6) holds at most with probability $o(1)$. The choice $\lambda = \lambda^*$ described in Remark 2 satisfies these two conditions. The difficulty with cross-validation, however, is that, as we demonstrate in Section 5 via simulations, it typically yields a rather small value of λ, so that the event (6) with $\lambda = \widehat{\lambda}$ holds with non-trivial probability even in large samples, and little is known about properties of the Lasso estimator $\widehat{\beta}(\lambda)$ when the event (6) does not hold, which is perhaps one of the main reasons why there are only few results on the cross-validated Lasso estimator in the literature. We therefore take a different approach. First, we use the fact that $\widehat{\lambda}$ is the cross-validation choice of λ to derive bounds on $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|$ and $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n}$ for the subsample Lasso estimators $\widehat{\beta}_{-k}(\widehat{\lambda})$ defined in (4). Second, we use the "degrees of freedom estimate" of Zou et al. (2007) to derive a sparsity bound for these estimators, and so to bound $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_1$. Third, we use the two point inequality

$$\|\widehat{\beta}(\lambda) - b\|_{2,n}^2 \leq \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \|b\|_1 - \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' \widehat{\beta}(\lambda))^2 - \lambda \|\widehat{\beta}(\lambda)\|_1, \quad \text{for all } b \in \mathbb{R}^p,$$

which can be found in van de Geer (2016), with $b = (K-1)^{-1} \sum_{k=1}^{K} (n - n_k) \widehat{\beta}_{-k}(\widehat{\lambda})/n$, a convex combination of the subsample Lasso estimators $\widehat{\beta}_{-k}(\widehat{\lambda})$, and derive a bound for its right-hand side using the definition of estimators $\widehat{\beta}_{-k}(\widehat{\lambda})$ and bounds on $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|$ and $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_1$. Finally, we use the triangle inequality to obtain a bound on $\|\widehat{\beta}(\lambda) - \beta\|_{2,n}$ from the bounds on $\|\widehat{\beta}(\lambda) - b\|_{2,n}$ and $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n}$. The details of the proof, including a short proof of the two point inequality, can be found in Section 6. ∎

Remark 4 (On the condition $\xi_n^2 \log n/n = o(1)$)**.** Note that in Examples 1, 2, and 3 above, the condition that $\xi_n^2 \log n/n = o(1)$ reduces to $p \log n/n = o(1)$, which we used in the abstract and in the Introduction. In fact, Lemma 17 in Section 7 shows that under Assumptions 1 and 3, we have $\sqrt{p} \lesssim \xi_n$, so that p is necessarily asymptotically smaller than n under the condition $\xi_n^2 \log n/n = o(1)$. This is why we refer to the case where $\xi_n^2 \log n/n = o(1)$ as the moderate-dimensional case. ∎

In addition to the bound on the prediction norm of the estimation error of the cross-validated Lasso estimator given in Theorem 1, we derive in the next theorem a bound on the sparsity of the estimator.

Theorem 2 (Sparsity bound for Gaussian moderate-dimensional case)**.** *Suppose that all conditions of Theorem 1 are satisfied. Then*

$$\|\widehat{\beta}(\widehat{\lambda})\|_0 \lesssim s \log^5 n \tag{7}$$

with probability $1 - o(1)$ *up-to a constant depending only on* c_1*,* C_1*,* K*,* a*, and* q*.*

Remark 5 (On the sparsity bound)**.** Belloni and Chernozhukov (2013) showed that if λ is chosen so that the event (6) holds at most with probability $o(1)$, then the Lasso estimator $\widehat{\beta}(\lambda)$ satisfies the bound $\|\widehat{\beta}(\lambda)\|_0 \lesssim s$ with probability $1 - o(1)$, so that the number of covariates that have been mistakenly selected by the Lasso estimator is at most of the same order as the number of non-zero coefficients in the original model (1). As explained in Remark 3, however, cross-validation typically yields a rather small value of λ, so that the event (6) with $\lambda = \widehat{\lambda}$ holds with non-trivial probability even in large samples, and it is typically the case that smaller values of λ lead to the Lasso estimators $\widehat{\beta}(\lambda)$ with a larger number of non-zero coefficients. However, using the result in Theorem 1 and the "degrees of freedom estimate" of Zou et al. (2007), we are still able to show that the cross-validated Lasso estimator is typically rather sparse, and in particular satisfies the bound (7) with probability $1 - o(1)$. ∎

With the help of Theorems 1 and 2, we immediately arrive at the following corollary for the bounds on L^2 and L^1 norms of the estimation error of the cross-validated Lasso estimator:

Corollary 1 (Other bounds for Gaussian moderate-dimensional case)**.** *Suppose that all conditions of Theorem 1 are satisfied. Then*

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\| \lesssim \left(\frac{s \log p}{n}\right)^{1/2} \cdot (\log^{7/8} n) \quad \textit{and} \quad \|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1 \lesssim \left(\frac{s^2 \log p}{n}\right)^{1/2} \cdot (\log^{27/8} n)$$

with probability $1 - o(1)$ *up-to a constant depending only on* c_1*,* C_1*,* K*,* a*, and* q*.*

To conclude this section, we consider the non-Gaussian case. One of the main complications in our derivations for this case is that without the assumption of Gaussian noise, we can not apply the "degrees of freedom estimate" derived in Zou et al. (2007) that provides a bound on the number of non-zero coefficients of the Lasso estimator, $\|\widehat{\beta}(\lambda)\|_0$, as a function of the

prediction norm of the estimation error of the estimator, $\|\widehat{\beta}(\lambda) - \beta\|_{2,n}$; see Lemmas 6 and 9 in the next section. Nonetheless, we can still derive an interesting bound on $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n}$ in this case even if p is much larger than n (high-dimensional case):

Theorem 3 (Sub-Gaussian high-dimensional case)**.** *Suppose that Assumptions 1 – 5 hold. In addition, suppose that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \leq C_1 t^2$. Finally, suppose that $M_n^4 s(\log^8 n)(\log^2 p)/n^{1-4/q} \lesssim 1$. Then*

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n} \lesssim \left(\frac{s \log^2(pn)}{n}\right)^{1/4} \tag{8}$$

with probability $1 - o(1)$ up-to a constant depending only on c_1, C_1, K, a, and q.

Remark 6 (On conditions of Theorem 3)**.** This theorem does not require the noise ε to be Gaussian conditional on X. Instead, it imposes a weaker condition that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \leq C_1 t^2$, which means that the conditional distribution of ε given X is sub-Gaussian; see, for example, Vershynin (2012). Also, we want to emphasize that the condition that $M_n^4 s(\log^8 n)(\log^2 p)/n^{1-4/q} \lesssim 1$ is not necessary to derive a non-trivial bound on $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n}$ but it does simplify the bound (8). Inspecting the proof of Theorem 3 reveals that without this condition, the bound (8) would take the form:

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n} \lesssim \left(\frac{s \log^2(pn)}{n}\right)^{1/4} + (n^{1/q} M_n \log^2 n \log^{1/2} p) \cdot \left(\frac{s \log(pn)}{n}\right)^{1/2}$$

with probability $1 - o(1)$ up-to a constant depending only on c_1, C_1, K, a, and q. ∎

5 Simulations

In this section, we present results of our simulation experiments. The purpose of the experiments is to investigate finite-sample properties of the cross-validated Lasso estimator. In particular, we are interested in (i) comparing estimation error of the cross-validated Lasso estimator in different norms to the Lasso estimator based on other choices of λ; (ii) studying sparsity properties of the cross-validated Lasso estimator; and (iii) estimating probability of the event (6) for $\lambda = \widehat{\lambda}$, the cross-validation choice of λ.

We consider two data generating processes (DGPs). In both DGPs, we simulate the vector of covariates X from the Gaussian distribution with mean zero and variance-covariance matrix given by $E[X_j X_k] = 0.5^{|j-k|}$ for all $j, k = 1, \ldots, p$. Also, we set $\beta = (1, -1, 2, -2, 0_{1\times(p-4)})'$. We simulate ε from the standard Gaussian distribution in DGP1 and from the uniform distribution on $[-3, 3]$ in DGP2. In both DGPs, we take ε to be independent of X. Further, for each DGP, we consider samples of size $n = 100$ and 400. For each DGP and each sample size, we consider $p = 40$, 100, and 400. To construct the candidate set Λ_n of values of the penalty parameter λ, we use Assumption 4 with $a = 0.9$, $c_1 = 0.005$ and $C_1 = 500$. Thus, the set Λ_n contains values of λ ranging from 0.0309 to 500 when $n = 100$ and from 0.0071 to 500 when $n = 400$, that is, the

set Λ_n is rather large in both cases. In all experiments, we use 5-fold cross-validation ($K = 5$). We repeat each experiment 5000 times.

As a comparison to the cross-validated Lasso estimator, we consider the Lasso estimator with λ chosen according to the Bickel-Ritov-Tsybakov rule:

$$\lambda = 2c\sigma n^{-1/2}\Phi^{-1}(1 - \alpha/(2p)),$$

where $c > 1$ and $\alpha \in (0, 1)$ are some constants, σ is the standard deviation of ε, and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard Gaussian distribution; see Bickel et al. (2009). Following Belloni and Chernozhukov (2011), we choose $c = 1.1$ and $\alpha = 0.1$. The noise level σ is typically have to be estimated from the data but for simplicity we assume that σ is known, so we set $\sigma = 1$ in DGP1 and $\sigma = \sqrt{3}$ in DGP2. In what follows, this Lasso estimator is denoted as P-Lasso and the cross-validated Lasso estimator is denoted as CV-Lasso.

Figure 5.1 contains simulation results for DGP1 with $n = 100$ and $p = 40$. The first three (that is, the top-left, top-right, and bottom-left) panels of Figure 5.1 present the mean of the estimation error of the Lasso estimators in the prediction, L^2, and L^1 norms, respectively. In addition to the solid and dotted horizontal lines representing the mean of the estimation error of CV-Lasso and P-Lasso, respectively, these panels also contain the curved dashed line representing the mean of the estimation error of the Lasso estimator as a function of λ in the corresponding norm (we perform the Lasso estimator for each value of λ in the candidate set Λ_n; we sort the values in Λ_n from the smallest to the largest, and put the order of λ on the horizontal axis; we only show the results for values of λ up to order 32 as these give the most meaningful comparisons). This estimator is denoted as λ-Lasso.

From these three panels of Figure 5.1, we see that the estimation error of CV-Lasso is only slightly above the minimum of the estimation error over all possible values of λ not only in the prediction and L^2 norms but also in the L^1 norm. In comparison, P-Lasso tends to have much larger estimation error in all three norms.

The bottom-right panel of Figure 5.1 depicts the histogram for the the number of non-zero coefficients of the cross-validated Lasso estimator. Overall, this panel suggests that the cross-validated Lasso estimator tends to select too many covariates: the number of selected covariates with large probability varies between 5 and 30 even though there are only 4 non-zero coefficients in the true model. Thus, we conjecture that even if it might be possible to decrease the power of the logarithm in the inequality $\|\widehat{\beta}(\widehat{\lambda})\|_0 \lesssim s\log^5 n$ obtained in Theorem 2, it is probably not possible to avoid the logarithm itself.

For all other experiments, the simulation results on the mean of the estimation error of the Lasso estimators can be found in Table 5.1. For simplicity, we only report the minimum over $\lambda \in \Lambda_n$ of mean of the estimation error of λ-Lasso in Table 5.1. The results in Table 5.1 confirm findings in Figure 5.1: the mean of the estimation error of CV-Lasso is very close to the minimum mean of the estimation errors of the λ-Lasso estimators under both DGPs for all combinations of n and p considered in all three norms. Their difference becomes smaller when the sample size n increases. The mean of the estimation error of P-Lasso is much larger than that

of CV-Lasso in most cases and is smaller than that of CV-Lasso only in L^1-norm when $n = 100$ and $p = 400$. Thus, given that the estimation error, for example, in the prediction norm of the Lasso estimator $\widehat{\beta}(\lambda)$ satisfies the bound $\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \lesssim (s \log p/n)^{1/2}$ with probability $1 - o(1)$ when λ is chosen using the Bickel-Ritov-Tsybakov rule, we conjecture that it might be possible to avoid the additional $\log^{7/8} n$ factor in the inequality $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n} \lesssim (s \log p/n)^{1/2} \cdot (\log^{7/8} n)$ obtained in Theorem 1.

Table 5.2 reports model selection results for the cross-validated Lasso estimator. More precisely, the table shows probabilities for the number of non-zero coefficients of the cross-validated Lasso estimator hitting different brackets. Overall, the results in Table 5.2 confirm findings in Figure 5.1: the cross-validated Lasso estimator tends to select too many covariates. The probability of selecting larger models tends to increase with p but decreases with n.

Table 5.3 provides information on the finite-sample distribution of the ratio of the maximum score $\max_{1 \leq j \leq p} |n^{-1} \sum_{i=1}^n X_{ij} \varepsilon_i|$ over $\widehat{\lambda}$, the cross-validation choice of λ. Specifically, the table shows probabilities for this ratio hitting different brackets. From Table 5.3, we see that this ratio is above 0.5 with large probability in all cases and in particular this probability exceeds 99% in most cases. Hence, (6) with $\lambda = \widehat{\lambda}$ holds not only with non-trivial but actually with large probability, meaning that existing arguments used to derive the rate of convergence of the Lasso estimator based, for example, on the Bickel-Ritov-Tsybakov choice of λ do not apply to the cross-validated Lasso estimator (see Remark 3 above) and justifying novel analysis developed in this paper.

6 Proofs

In this section, we prove Theorems 1, 2, 3, and Corollary 1. Since the proofs are long, we start with a sequence of preliminary lemmas. For convenience, we use the following additional notation. For $k = 1, \ldots, K$, we denote

$$\|\delta\|_{2,n,k} = \left(\frac{1}{n_k} \sum_{i \in I_k} (X_i' \delta)^2 \right)^{1/2} \quad \text{and} \quad \|\delta\|_{2,n,-k} = \left(\frac{1}{n - n_k} \sum_{i \notin I_k} (X_i' \delta)^2 \right)^{1/2}$$

for all $\delta \in \mathbb{R}^p$. We use c and C to denote constants that can change from place to place but that can be chosen to depend only on c_1, C_1, K, a, and q. We use the notation $a_n \lesssim b_n$ if $a_n \leq C b_n$. In addition, we denote $X_1^n = (X_1, \ldots, X_n)$. Moreover, for $\delta \in \mathbb{R}^p$ and $M \subset \{1, \ldots, p\}$, we use δ_M to denote the vector in $\mathbb{R}^{|M|}$ consisting of all elements of δ corresponding to indices in M (with order of indices preserved). Finally, for $\delta = (\delta_1, \ldots, \delta_p)' \in \mathbb{R}^p$, we denote $\text{supp}(\delta) = \{j \in \{1, \ldots, p\} \colon \delta_j \neq 0\}$.

In Lemmas 1 – 5, we will impose the condition that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \leq C_1 t^2$. Note that under Assumption 2, this condition is satisfied if the conditional distribution of ε given X is Gaussian.

Lemma 1. *Suppose that Assumptions 1 – 5 hold. In addition, suppose that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \le C_1 t^2$. Then there exists $\bar{\lambda}_0 = \bar{\lambda}_{n,0} \in \Lambda_n$, possibly depending on n, such that for all $k = 1, \dots, K$, we have*

$$\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_{2,n,-k}^2 \lesssim \frac{s(\log p + \log\log n)}{n} \quad \text{and} \quad \|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_1^2 \lesssim \frac{s^2(\log p + \log\log n)}{n} \tag{9}$$

with probability $1 - o(1)$.

Proof. Let $T = \mathrm{supp}(\beta)$ and $T^c = \{1, \dots, p\} \backslash T$. Also, for $k = 1, \dots, K$, denote

$$Z_k = \frac{1}{n - n_k} \sum_{i \notin I_k} X_i \varepsilon_i$$

and

$$\kappa_k = \inf \left\{ \frac{\sqrt{s}\|\delta\|_{2,n,-k}}{\|\delta_T\|_1} : \ \delta \in \mathbb{R}^p, \ \|\delta_{T^c}\|_1 < 3\|\delta_T\|_1 \right\}.$$

To prove the first asserted claim, we will apply Theorem 1 in Belloni and Chernozhukov (2011) that shows that for any $k = 1, \dots, K$ and $\lambda \in \Lambda_n$, on the event $\lambda \ge 4\|Z_k\|_\infty$, we have

$$\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \le \frac{3\lambda\sqrt{s}}{2\kappa_k}.$$

Thus, it suffices to show that there exists $c > 0$ such that

$$P(\kappa_k < c) = o(1), \tag{10}$$

for all $k = 1, \dots, K$, and that there exist $\bar{\lambda}_0 = \bar{\lambda}_{n,0} \in \Lambda_n$, possibly depending on n, such that

$$P\left(\bar{\lambda}_0 < 4\|Z_k\|_\infty\right) = o(1) \tag{11}$$

for all $k = 1, \dots, K$ and

$$\bar{\lambda}_0 \lesssim \left(\frac{\log p + \log\log n}{n} \right)^{1/2}. \tag{12}$$

To prove (10), note that by Jensen's inequality,

$$\begin{aligned} L_n &= \Big(\mathrm{E}\Big[\max_{1 \le i \le n} \max_{1 \le j \le p} |X_{ij}|^2 \Big] \Big)^{1/2} \le \Big(\mathrm{E}\Big[\max_{1 \le i \le n} \max_{1 \le j \le p} |X_{ij}|^q \Big] \Big)^{1/q} \\ &\le \Big(\textstyle\sum_{i=1}^n \mathrm{E}\Big[\max_{1 \le j \le p} |X_{ij}|^q \Big] \Big)^{1/q} \le n^{1/q} M_n. \end{aligned}$$

Thus, for $l_n = s \log n$,

$$\begin{aligned} \gamma_n &= \frac{L_n \sqrt{l_n}}{\sqrt{n}} \cdot \Big(\log^{1/2} p + (\log l_n) \cdot (\log^{1/2} p) \cdot (\log^{1/2} n) \Big) \\ &\lesssim \frac{L_n \sqrt{s}}{\sqrt{n}} \cdot (\log^2 n) \cdot (\log^{1/2} p) \le \frac{M_n \sqrt{s}}{\sqrt{n^{1-2/q}}} \cdot (\log^2 n) \cdot (\log^{1/2} p) = o(1) \end{aligned}$$

by Assumption 3. Hence, noting that (i) all eigenvalues of the matrix $\mathrm{E}[XX']$ are bounded from above and below from zero by Assumption 1 and that (ii) $(n-n_k)^{-1} \lesssim n^{-1}$ by Assumption 5 and applying Lemma 15 with k, K, and δ_n there replaced by l_n, L_n, and γ_n here shows that

$$1 \lesssim \|\delta\|_{2,n,-k} \lesssim 1 \tag{13}$$

with probability $1-o(1)$ uniformly over all $\delta \in \mathbb{R}^p$ such that $\|\delta\| = 1$ and $\|\delta_{T^c}\|_0 \leq s \log n$ and all $k = 1, \ldots, K$. Hence, (10) follows from Lemma 10 in Belloni and Chernozhukov (2011) applied with m there equal to $s \log n$ here.

To prove (11) and (12) fix $k = 1, \ldots, K$ and note that

$$\max_{1 \leq j \leq p} \sum_{i \notin I_k} \mathrm{E}[|X_{ij}\varepsilon_i|^2] \lesssim n$$

by Assumptions 1 and 2. Also,

$$\begin{aligned}\Big(\mathrm{E}\Big[\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |X_{ij}\varepsilon_i|^2\Big]\Big)^{1/2} &\leq \Big(\mathrm{E}\Big[\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |X_{ij}\varepsilon_i|^q\Big]\Big)^{1/q} \\ &\leq \Big(\textstyle\sum_{i=1}^n \mathrm{E}\Big[\max_{1 \leq j \leq p} |X_{ij}\varepsilon_i|^q\Big]\Big)^{1/q} \lesssim n^{1/q} M_n\end{aligned}$$

by Jensen's inequality, the definition of M_n, and the assumption on the moment generating function of the conditional distribution of ε given X. Thus, by Lemma 13 and Assumption 3,

$$\mathrm{E}\Big[(n-n_k)\|Z_k\|_\infty\Big] \lesssim \sqrt{n \log p} + n^{1/q} M_n \log p \lesssim \sqrt{n \log p}.$$

Hence, applying Lemma 14 with $t = (n \log\log n)^{1/2}$ and Z there replaced by $(n-n_k)\|Z_k\|_\infty$ here and noting that $nM_n^q/(n \log\log n)^{q/2} = o(1)$ by Assumption 3 implies that

$$\|Z_k\|_\infty \lesssim \left(\frac{\log p + \log\log n}{n}\right)^{1/2}$$

with probability $1-o(1)$. Hence, noting that $\log p + \log\log n = o(n)$ by Assumption 3, it follows from Assumption 4 that there exists $\bar{\lambda}_0 \in \Lambda_n$ such that (11) and (12) hold.

Further, to prove the second asserted claim, note that using (10) and (13) and applying Theorem 2 in Belloni and Chernozhukov (2011) with $m = s \log n$ there shows that $\|\widehat{\beta}_{-k}(\bar{\lambda}_0)\|_0 \lesssim s$ with probability $1-o(1)$ for all $k = 1, \ldots, K$. Hence,

$$\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_1^2 \lesssim s\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|^2 \lesssim s\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_{2,n,-k}^2 \lesssim \frac{s^2(\log p + \log\log n)}{n}$$

with probability $1-o(1)$ for all $k = 1, \ldots, K$, where the second inequality follows from (13), and the third one from the first asserted claim. This completes the proof of the lemma. ∎

Lemma 2. *Suppose that Assumptions 1 – 5 hold. In addition, suppose that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \leq C_1 t^2$. Then we have for all $k = 1, \ldots, K$ that*

$$\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_{2,n,k}^2 \lesssim \frac{s(\log p + \log\log n)}{n}$$

with probability $1-o(1)$ for $\bar{\lambda}_0$ defined in Lemma 1.

Proof. Fix $k = 1, \ldots, K$ and denote $\widehat{\beta} = \widehat{\beta}_{-k}(\bar{\lambda}_0)$. We have

$$\begin{aligned}
\Big| \|\widehat{\beta} - \beta\|_{2,n,-k}^2 - \|\widehat{\beta} - \beta\|_{2,n,k}^2 \Big| &= \Big| (\widehat{\beta} - \beta)' \Big(\frac{1}{n - n_k} \sum_{i \notin I_k} X_i X_i' - \frac{1}{n_k} \sum_{i \in I_k} X_i X_i' \Big) (\widehat{\beta} - \beta) \Big| \\
&\leq \Big| (\widehat{\beta} - \beta)' \Big(\frac{1}{n - n_k} \sum_{i \notin I_k} X_i X_i' - \mathrm{E}[XX'] \Big) (\widehat{\beta} - \beta) \Big| \\
&\quad + \Big| (\widehat{\beta} - \beta)' \Big(\frac{1}{n_k} \sum_{i \in I_k} X_i X_i' - \mathrm{E}[XX'] \Big) (\widehat{\beta} - \beta) \Big| \\
&\leq \|\widehat{\beta} - \beta\|_1^2 \max_{1 \leq j,l \leq p} \Big| \frac{1}{n - n_k} \sum_{i \notin I_k} X_{ij} X_{il} - \mathrm{E}[X_j X_l] \Big| \\
&\quad + \|\widehat{\beta} - \beta\|_1^2 \max_{1 \leq j,l \leq p} \Big| \frac{1}{n_k} \sum_{i \in I_k} X_{ij} X_{il} - \mathrm{E}[X_j X_l] \Big|
\end{aligned}$$

by the triangle inequality. Further, by Lemma 1, $\|\widehat{\beta} - \beta\|_1^2 \lesssim s^2(\log p + \log\log n)/n$ with probability $1 - o(1)$ and by Lemma 13,

$$\begin{aligned}
\mathrm{E}\Big[\max_{1 \leq j,l \leq p} \Big| \frac{1}{n - n_k} \textstyle\sum_{i \notin I_k} X_{ij} X_{il} - \mathrm{E}[X_j X_l] \Big| \Big] &\lesssim \Big(\frac{\gamma_n^4 \log p}{n} \Big)^{1/2} + \frac{M_n^2 \log p}{n^{1 - 2/q}}, \\
\mathrm{E}\Big[\max_{1 \leq j,l \leq p} \Big| \frac{1}{n_k} \textstyle\sum_{i \in I_k} X_{ij} X_{il} - \mathrm{E}[X_j X_l] \Big| \Big] &\lesssim \Big(\frac{\gamma_n^4 \log p}{n} \Big)^{1/2} + \frac{M_n^2 \log p}{n^{1 - 2/q}},
\end{aligned}$$

since $1/n_k \lesssim 1/n$ and $1/(n - n_k) \lesssim 1/n$ by Assumption 5 and

$$\max_{1 \leq j,l \leq p} \mathrm{E}[X_{ij}^2 X_{il}^2] \leq \max_{1 \leq j \leq p} \mathrm{E}[X_{ij}^4] \leq \gamma_n^4$$

by Hölder's inequality and Assumption 1. Noting that

$$\gamma_n^4 s^2 \log p / n = o(1) \quad \text{and} \quad M_n^2 s \log p / n^{1 - 2/q} = o(1),$$

which hold by Assumption 3, and combining presented inequalities implies that

$$\Big| \|\widehat{\beta} - \beta\|_{2,n,-k}^2 - \|\widehat{\beta} - \beta\|_{2,n,k}^2 \Big| \lesssim \frac{s(\log p + \log\log n)}{n} \cdot o(1)$$

with probability $1 - o(1)$. In addition, by Lemma 1, $\|\widehat{\beta} - \beta\|_{2,n,-k}^2 \lesssim s(\log p + \log\log n)/n$ with probability $1 - o(1)$. Therefore, it follows that

$$\|\widehat{\beta} - \beta\|_{2,n,k}^2 \lesssim \frac{s(\log p + \log\log n)}{n}$$

with probability $1 - o(1)$. This completes the proof. ∎

Lemma 3. *Suppose that Assumptions 1 – 5 hold. In addition, suppose that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \leq C_1 t^2$. Then we have for all $k = 1, \ldots, K$ that*

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2 \lesssim \frac{s(\log p + \log\log n)}{n} + \frac{(\log\log n)^2}{n}$$

with probability $1 - o(1)$.

Proof. We have

$$\sum_{k=1}^{K}\sum_{i\in I_k}(Y_i - X_i'\widehat{\beta}_{-k}(\widehat{\lambda}))^2 \leq \sum_{k=1}^{K}\sum_{i\in I_k}(Y_i - X_i'\widehat{\beta}_{-k}(\bar{\lambda}_0))^2$$

for $\bar{\lambda}_0$ defined in Lemma 1. Therefore,

$$\sum_{k=1}^{K} n_k\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2 \leq \sum_{k=1}^{K} n_k\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_{2,n,k}^2 + 2\sum_{k=1}^{K}\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)).$$

Further, by assumptions of the lemma, for $\lambda \in \Lambda_n$, $k = 1,\ldots,K$, and $D_k = \{(X_i,Y_i)_{i\notin I_k}; (X_i)_{i\in I_k}\}$, we have for all $t \in \mathbb{R}$ that

$$\log \mathrm{E}\left[\exp\left(t\textstyle\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0))\right) \mid D_k\right] \lesssim t^2 n_k\|\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}^2.$$

Therefore, since $|\Lambda_n| \lesssim \log n$ by Assumption 4, we have with probability $1 - o(1)$ that for all $k = 1,\ldots,K$ and $\lambda \in \Lambda_n$,

$$\left|\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0))\right| \lesssim \sqrt{n_k}\cdot(\log\log n)\cdot\|\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}$$

by the union bound and Markov's inequality; in particular, since $\widehat{\lambda} \in \Lambda_n$, we have with probability $1 - o(1)$ that for all $k = 1,\ldots,K$,

$$\left|\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0))\right| \lesssim \sqrt{n_k}\cdot(\log\log n)\cdot\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}.$$

Hence, since $n_k/n \geq c_1$ by Assumption 5, we have with probability $1 - o(1)$ that

$$\sum_{k=1}^{K}\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2 \lesssim \sum_{k=1}^{K}\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_{2,n,k}^2 + \frac{\log\log n}{\sqrt{n}}\sum_{k=1}^{K}\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}.$$

Let $\widehat{k}$ be a $k = 1,\ldots,K$ that maximizes $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}$. Then with probability $1-o(1)$,

$$\|\widehat{\beta}_{-\widehat{k}}(\widehat{\lambda}) - \beta\|_{2,n,\widehat{k}}^2 \lesssim \sum_{k=1}^{K}\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|_{2,n,k}^2 + \frac{\log\log n}{\sqrt{n}}\|\widehat{\beta}_{-\widehat{k}}(\widehat{\lambda}) - \widehat{\beta}_{-\widehat{k}}(\bar{\lambda}_0)\|_{2,n,\widehat{k}},$$

and so, by Lemma 2 and the triangle inequality, with probability $1 - o(1)$,

$$\begin{aligned}\|\widehat{\beta}_{-\widehat{k}}(\widehat{\lambda}) - \beta\|_{2,n,\widehat{k}}^2 \lesssim{}& \frac{s(\log p + \log\log n)}{n} \\ &+ \frac{\log\log n}{\sqrt{n}}\sqrt{\frac{s(\log p + \log\log n)}{n}} + \frac{\log\log n}{\sqrt{n}}\|\widehat{\beta}_{-\widehat{k}}(\widehat{\lambda}) - \beta\|_{2,n,\widehat{k}}.\end{aligned}$$

Conclude that for all $k = 1,\ldots,K$, with probability $1 - o(1)$,

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2 \lesssim \|\widehat{\beta}_{-\widehat{k}}(\widehat{\lambda}) - \beta\|_{2,n,\widehat{k}}^2 \lesssim \frac{s(\log p + \log\log n)}{n} + \frac{(\log\log n)^2}{n}.$$

This completes the proof. ∎

Lemma 4. *Suppose that Assumptions 1 – 5 hold. In addition, suppose that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \leq C_1 t^2$. Then we have for all $k = 1, \ldots, K$ that*

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|^2 \lesssim \frac{s(\log p + \log\log n)}{n} + \frac{(\log\log n)^2}{n}$$

with probability $1 - o(1)$.

Proof. Fix $k = 1, \ldots, K$. For $\lambda \in \Lambda_n$, let $\delta_\lambda = (\widehat{\beta}_{-k}(\lambda) - \beta)/\|\widehat{\beta}_{-k}(\lambda) - \beta\|$. Observe that conditional on $\bar{D}_k = (X_i, Y_i)_{i \notin I_k}$, $(\delta_\lambda)_{\lambda \in \Lambda_n}$ is non-stochastic. Therefore, $\max_{\lambda \in \Lambda_n} \sum_{i \in I_k} \mathrm{E}[(X_i'\delta_\lambda)^4 \mid \bar{D}_k] \lesssim \Gamma_n^4 n$ by Assumption 1 since $\|\delta_\lambda\| = 1$ for all $\lambda \in \Lambda_n$. In addition,

$$\Big(\mathrm{E}\Big[\max_{i \in I_k} \max_{\lambda \in \Lambda_n} (X_i'\delta_\lambda)^4 \mid \bar{D}_k\Big]\Big)^{1/2} \leq \Gamma_n^2 \cdot (n|\Lambda_n|)^{1/2}.$$

So, by Lemma 13,

$$R = \max_{\lambda \in \Lambda_n} \left| \frac{1}{n_k} \sum_{i \in I_k} \Big((X_i'\delta_\lambda)^2 - \mathrm{E}[(X_i'\delta_\lambda)^2 \mid \bar{D}_k] \Big) \right|$$

satisfies

$$\mathrm{E}[R] \lesssim \sqrt{\frac{\Gamma_n^4 \log|\Lambda_n|}{n}} + \frac{\Gamma_n^2 \cdot (n|\Lambda_n|)^{1/2} \log|\Lambda_n|}{n} = o(1)$$

by Assumption 3 since $|\Lambda_n| \lesssim \log n$ by Assumption 4. Moreover, by Assumption 1, for any $\lambda \in \Lambda_n$,

$$\begin{aligned}
\|\widehat{\beta}_{-k}(\lambda) - \beta\|^2 &\lesssim \frac{1}{n_k} \sum_{i \in I_k} \mathrm{E}[(X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 \mid \bar{D}_k] \\
&\leq \frac{1}{n_k} \sum_{i \in I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 + R\|\widehat{\beta}_{-k}(\lambda) - \beta\|^2 \\
&= \|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,k}^2 + R\|\widehat{\beta}_{-k}(\lambda) - \beta\|^2.
\end{aligned}$$

Therefore, with probability $1 - o(1)$,

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|^2 \lesssim \|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2 \lesssim \frac{s(\log p + \log\log n)}{n} + \frac{(\log\log n)^2}{n},$$

where the second inequality follows from Lemma 3. The asserted claim follows. ∎

Lemma 5. *Suppose that Assumptions 1 – 5 hold. In addition, suppose that for all $t \in \mathbb{R}$, we have $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \leq C_1 t^2$. Finally, suppose that $\xi_n^2 \log n / n = o(1)$. Then we have for all $k = 1, \ldots, K$ that*

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 \lesssim \frac{s(\log p + \log\log n)}{n} + \frac{(\log\log n)^2}{n}$$

with probability $1 - o(1)$.

Proof. Since $nP(\|X\| > \xi_n) = o(1)$ and for all $\delta \in \mathcal{S}^p$, we have $(\mathrm{E}[|X'\delta|^2])^{1/2} \leq C_1$ and $(\mathrm{E}[|X'\delta|^4])^{1/4} \leq \Gamma_n$ by Assumption 1, applying Lemma 16 shows that with probability $1 - o(1)$,

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 \lesssim \|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|^2 \cdot \left(1 + \Gamma_n^2 \cdot (P(\|X\| > \xi_n))^{1/2} + \sqrt{\frac{\xi_n^2 \log(pn)}{n}} + \frac{\xi_n^2 \log(pn)}{n}\right).$$

In addition, $\Gamma_n^2 \cdot (P(\|X\| > \xi_n)^{1/2} = o(1)$ by Assumptions 1 and 3. Also, $\xi_n^2 \log(pn)/n = o(1)$ since we have $\xi_n^2 \log n/n = o(1)$ and it follows from Lemma 17 that $\sqrt{p} \lesssim \xi_n$. Thus, the asserted claim follows from Lemma 17. ∎

Lemma 6. *For all $\lambda \in \Lambda_n$, the Lasso estimator $\widehat{\beta}(\lambda)$ given in (2) based on the data $(X_i, Y_i)_{i=1}^n = (X_i, X_i'\beta + \varepsilon_i)_{i=1}^n$ has the following properties: (i) the function $(\varepsilon_i)_{i=1}^n \mapsto (X_i\widehat{\beta}(\lambda))_{i=1}^n$ mapping $\mathbb{R}^n$ to $\mathbb{R}^n$ for a fixed value of $X_1^n = (X_1, \ldots, X_n)$ is Lipschitz-continuous with Lipschitz constant one whenever the matrix $(X_1, \ldots, X_n)'$ has full column rank; (ii) if for all $i = 1, \ldots, n$, the conditional distribution of ε_i given X_i is $N(0, \sigma_i^2)$ and the pairs (X_i, ε_i) are independent across i, then*

$$\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n] = \sum_{i=1}^n \sigma_i^{-2} \mathrm{E}[\varepsilon_i X_i'(\widehat{\beta}(\lambda) - \beta) \mid X_1^n] \tag{14}$$

on the event that the matrix $(X_1, \ldots, X_n)'$ has full column rank.

Proof. This lemma is an extension of the main result in Zou et al. (2007) to the heteroscedastic case (we allow σ_i^2's to vary over i).

Fix $\lambda \in \Lambda_n$ and $X_1^n = (X_1, \ldots, X_n)$ such that the matrix $(X_1, \ldots, X_n)'$ has full column rank. Denote $\widehat{\beta} = \widehat{\beta}(\lambda)$ and $\widehat{T} = \mathrm{supp}(\widehat{\beta})$. Note that $\widehat{\beta}$ is well-defined because the solution of the optimization problem (2) is unique since the optimized function is strictly convex under the condition that the matrix $(X_1, \ldots, X_n)'$ has full column rank. Also, let $\mathcal{D}$ denote the set of all vectors in $\mathbb{R}^p$ whose elements are either -1 or 1. Moreover, let $\mathcal{N}_{X_1^n}$ be a subset of $\mathbb{R}^n$ consisting of all values of $e = (e_1, \ldots, e_n)' \in \mathbb{R}^n$ such that for some $M \subset \{1, \ldots, p\}$ with $M \neq \{1, \ldots, p\}$, $j \in \{1, \ldots, p\} \setminus M$, and $d \in \mathcal{D}$, we have

$$\left|\frac{2}{n}\sum_{i=1}^n (e_i + X_i'(\beta - \widehat{b}))X_{ij}\right| = \lambda$$

where $\widehat{b} = (\widehat{b}_1, \ldots, \widehat{b}_p)'$ is a vector in $\mathbb{R}^p$ such that

$$(\widehat{b})_M = \left(\frac{1}{n}\sum_{i=1}^n (X_i)_M (X_i)_M'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n (X_i)_M'(e_i + X_i'\beta) - \lambda d_M/2\right)$$

and $\widehat{b}_j = 0$ for all $j \notin M$. Note that $\widehat{b}$ is well-defined because the matrix $((X_1)_M, \ldots, (X_n)_M)'$ has full column rank under the condition that the matrix $(X_1, \ldots, X_n)'$ has full column rank. It follows that $\mathcal{N}_{X_1^n}$ is contained in a finite set of hyperplanes in $\mathbb{R}^n$.

Next, by the Kuhn-Tucker conditions, for all $j \in \widehat{T}$, we have

$$\frac{2}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})X_{ij} = \lambda \cdot \mathrm{sign}(\widehat{\beta}_j),$$

and for all $j \notin \widehat{T}$, we have

$$\left| \frac{2}{n} \sum_{i=1}^{n} (Y_i - X_i' \widehat{\beta}) X_{ij} \right| \leq \lambda.$$

Thus, since the matrix $((X_1)_{\widehat{T}}, \ldots, (X_n)_{\widehat{T}})'$ has full column rank under the condition that the matrix $(X_1, \ldots, X_n)'$ has full column rank, we have for all $l = 1, \ldots, n$ that

$$X_l' \widehat{\beta} = (X_l)_{\widehat{T}}' \left(\frac{1}{n} \sum_{i=1}^{n} (X_i)_{\widehat{T}} (X_i)_{\widehat{T}}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} (X_i)_{\widehat{T}} Y_i - \lambda \cdot \operatorname{sign}(\widehat{\beta}_{\widehat{T}})/2 \right). \tag{15}$$

Moreover, since $\widehat{\beta}$ is the unique solution of the optimization problem (2), it follows that the functions $(\varepsilon_i)_{i=1}^n \mapsto \widehat{T}$ and $(\varepsilon_i)_{i=1}^n \mapsto \operatorname{sign}(\widehat{\beta}_{\widehat{T}})$ are well-defined and are locally constant for all values of $(\varepsilon_i)_{i=1}^n$ satisfying $(\varepsilon_1, \ldots, \varepsilon_n)' \notin \mathcal{N}_{X_1^n}$.

Now we are ready to show that the function $(\varepsilon_i)_{i=1}^n \mapsto (X_i' \widehat{\beta})_{i=1}^n$ is Lipschitz-continuous with Lipschitz constant one, which is the first asserted claim. To this end, consider $\widehat{\beta}$ as a function of $(\varepsilon_i)_{i=1}^n$. Let ε^1 and ε^2 be two values of $(\varepsilon_i)_{i=1}^n$, and let $\widehat{\beta}(\varepsilon^1)$ and $\widehat{\beta}(\varepsilon^2)$ be corresponding values of $\widehat{\beta}$. Suppose first that the line segment $\mathcal{P} = \{t\varepsilon^2 + (1-t)\varepsilon^1 \colon t \in [0,1]\}$ does not intersect $\mathcal{N}_{X_1^n}$. Then $\widehat{T}$ and $\operatorname{sign}(\widehat{\beta}_{\widehat{T}})$ are constant on $\mathcal{P}$, and so (15) implies that

$$\sum_{i=1}^{n} \left(X_i' \widehat{\beta}(\varepsilon^2) - X_i' \widehat{\beta}(\varepsilon^1) \right)^2 \leq \|\varepsilon^2 - \varepsilon^1\|^2. \tag{16}$$

Second, suppose that $\mathcal{P}$ has a non-empty intersection with $\mathcal{N}_{X_1^n}$. Recall that the set $\mathcal{N}_{X_1^n}$ is contained in a finite collection of hyperplanes, and so we can find $0 = t_0 < t_1 < \cdots < t_k = 1$ such that $\widehat{T}$ remains constant on each line segment $\{t\varepsilon^2 + (1-t)\varepsilon^1 \colon t \in (t_{j-1}, t_j)\}$, $j = 1, \ldots, k$, of $\mathcal{P}$. In addition, note that the function $(\varepsilon_i)_{i=1}^n \mapsto (X_i' \widehat{\beta})_{i=1}^n$ is continuous (otherwise we could use, for example, the fact that the optimized function in (2) is strictly convex to arrive at a contradiction). Hence, (16) holds in this case as well by the triangle inequality. This gives the first asserted claim.

Next, we prove (14), which is the second asserted claim. Note that since for all values of $(\varepsilon_i)_{i=1}^n$ satisfying $(\varepsilon_1, \ldots, \varepsilon_n)' \notin \mathcal{N}_{X_1^n}$, the functions $(\varepsilon_i)_{i=1}^n \mapsto \widehat{T}$ and $(\varepsilon_i)_{i=1}^n \mapsto \operatorname{sign}(\widehat{\beta}_{\widehat{T}})$ are locally constant, it follows from (15) that for the same values of $(\varepsilon_i)_{i=1}^n$, the functions $(\varepsilon_i)_{i=1}^n \mapsto X_l' \widehat{\beta}$ are differentiable. Moreover,

$$\frac{\partial (X_l' \widehat{\beta})}{\partial \varepsilon_l} = \frac{1}{n} (X_l)_{\widehat{T}}' \left(\frac{1}{n} \sum_{i=1}^{n} (X_i)_{\widehat{T}} (X_i)_{\widehat{T}}' \right)^{-1} (X_l)_{\widehat{T}},$$

and so

$$\begin{aligned} \sum_{l=1}^{n} \frac{\partial (X_l' \widehat{\beta})}{\partial \varepsilon_l} &= \frac{1}{n} \sum_{l=1}^{n} (X_l)_{\widehat{T}}' \Big(\frac{1}{n} \sum_{i=1}^{n} (X_i)_{\widehat{T}} (X_i)_{\widehat{T}}' \Big)^{-1} (X_l)_{\widehat{T}} \\ &= \frac{1}{n} \sum_{l=1}^{n} \operatorname{tr} \Big((X_l)_{\widehat{T}}' \Big(\frac{1}{n} \sum_{i=1}^{n} (X_i)_{\widehat{T}} (X_i)_{\widehat{T}}' \Big)^{-1} (X_l)_{\widehat{T}} \Big) \end{aligned}$$

$$= \frac{1}{n}\sum_{l=1}^{n} \mathrm{tr}\Big(\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i)_{\widehat{T}}(X_i)_{\widehat{T}}'\Big)^{-1}(X_l)_{\widehat{T}}(X_l)_{\widehat{T}}'\Big)$$
$$= \mathrm{tr}\Big(\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i)_{\widehat{T}}(X_i)_{\widehat{T}}'\Big)^{-1}\frac{1}{n}\sum_{l=1}^{n}(X_l)_{\widehat{T}}(X_l)_{\widehat{T}}'\Big) = |\widehat{T}|$$

whenever $(\varepsilon_1, \ldots, \varepsilon_n)' \notin \mathcal{N}_{X_1^n}$. Since $\mathrm{P}((\varepsilon_1, \ldots, \varepsilon_n)' \in \mathcal{N}_{X_1^n} \mid X_1^n) = 0$, it follows that

$$\sum_{l=1}^{n} \mathrm{E}\left[\frac{\partial(X_l'\widehat{\beta})}{\partial \varepsilon_l} \mid X_1^n\right] = \mathrm{E}[|\widehat{T}| \mid X_1^n].$$

In addition, the first asserted claim implies that the functions $(\varepsilon_i)_{i=1}^n \mapsto X_l'\widehat{\beta}$ are absolutely continuous, and so applying Stein's lemma (see, for example, Lemma 2.1 in Chen, Goldstein, and Shao, 2011) conditional on X_1^n and using the fact that pairs (X_i, ε_i) are independent across i shows that

$$\mathrm{E}[|\widehat{T}| \mid X_1^n] = \sum_{l=1}^{n} \mathrm{E}\left[\frac{\partial(X_l'\widehat{\beta})}{\partial \varepsilon_l} \mid X_1^n\right] = \sum_{l=1}^{n} \sigma_l^{-2}\mathrm{E}[\varepsilon_l X_l'\widehat{\beta} \mid X_1^n] = \sum_{l=1}^{n} \sigma_l^{-2}\mathrm{E}[\varepsilon_l X_l'(\widehat{\beta} - \beta) \mid X_1^n],$$

which gives (14), the second asserted claim, since $|\widehat{T}| = \|\widehat{\beta}(\lambda)\|_0$. This completes the proof of the lemma. ∎

Lemma 7. *Suppose that Assumptions 2 and 5 hold. In addition, suppose that the conditional distribution of ε given X is Gaussian. Then for all $\lambda \in \Lambda_n$ and $t > 0$, we have*

$$\mathrm{P}\Big(\Big|\|\widehat{\beta}(\lambda) - \beta\|_{2,n} - \mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n]\Big| > t \mid X_1^n\Big) \leq Ce^{-cnt^2},$$

and for all $k = 1, \ldots, K$, $\lambda \in \Lambda_n$, and $t > 0$, we have

$$\mathrm{P}\Big(\Big|\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} - \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n]\Big| > t \mid X_1^n\Big) \leq Ce^{-cnt^2}$$

on the event that the matrix $(X_1, \ldots, X_n)'$ has the full column rank, where $c > 0$ and $C > 0$ are some constants that depend only on c_1 and C_1.

Proof. Fix $\lambda \in \Lambda_n$ and $X_1^n = (X_1, \ldots, X_n)$ such that the matrix $(X_1, \ldots, X_n)'$ has full column rank. By Lemma 6, the function $(\varepsilon_i)_{i=1}^n \mapsto (X_i'\widehat{\beta}(\lambda))_{i=1}^n$ is Lipschitz-continuous with Lipschitz constant one, and so is $(\varepsilon_i)_{i=1}^n \mapsto (\sum_{i=1}^n (X_i'(\widehat{\beta}(\lambda) - \beta))^2)^{1/2}$. In turn, $(\sum_{i=1}^n (X_i'(\widehat{\beta}(\lambda) - \beta))^2)^{1/2} = \sqrt{n}\|\widehat{\beta}(\lambda) - \beta\|_{2,n}$. Thus, by the Gaussian concentration inequality (see, for example, Theorem 2.1.12 in Tao, 2012),

$$\mathrm{P}\Big(\Big|\sqrt{n}\|\widehat{\beta}(\lambda) - \beta\|_{2,n} - \mathrm{E}[\sqrt{n}\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n]\Big| > t \mid X_1^n\Big) \leq Ce^{-ct^2},$$

for some constants $c > 0$ and $C > 0$ that depend only on c_1 and C_1. Replacing t by $\sqrt{n}t$ in this inequality gives the first asserted claim. The second asserted claim follows similarly. This completes the proof of the theorem. ∎

Lemma 8. *For some sufficiently large constant C, let*

$$T_n = C\left(\left(\frac{s(\log p + \log\log n)}{n}\right)^{1/2} + \frac{\log\log n}{\sqrt{n}}\right),$$

and for $k = 1, \ldots, K$, let

$$\Lambda_{n,k}(X_1^n) = \left\{\lambda \in \Lambda_n \colon \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n] \le T_n\right\}.$$

Suppose that Assumptions 1 – 5 hold. In addition, suppose that the conditional distribution of ε given X is Gaussian. Finally, suppose that $\xi_n^2 \log n/n = o(1)$. Then $\widehat{\lambda} \in \Lambda_{n,k}(X_1^n)$ for all $k = 1, \ldots, K$ with probability $1 - o(1)$.

Proof. Fix $k = 1, \ldots, K$. We have

$$\begin{aligned}\mathrm{P}\Big(\widehat{\lambda} \notin \Lambda_{n,k}(X_1^n)\Big) &\le \mathrm{P}\Big(\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 > T_n^2/4\Big) \\ &\quad + \mathrm{P}\Big(\max_{\lambda\in\Lambda_n}\Big|\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} - \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n]\Big|^2 > T_n^2/4\Big).\end{aligned}$$

The first term on the right-hand side of this inequality is $o(1)$ by Lemma 5 (recall that the fact that the conditional distribution of ε given X is Gaussian combined with Assumption 2 implies that $\log \mathrm{E}[\exp(t\varepsilon) \mid X] \le C_1 t^2$ for all $t > 0$ if C_1 in this inequality is large enough).

Further, since $nP(\|X\| > \xi_n) = o(1)$ and for all $\delta \in \mathcal{S}^p$, we have $c_1 \le (\mathrm{E}[|X'\delta|^2])^{1/2} \le C_1$ and $(\mathrm{E}[|X'\delta|^4])^{1/4} \le \Gamma_n$ by Assumption 1, Lemma 16 implies that all eigenvalues of the matrix $(n - n_k)^{-1}\sum_{i\notin I_k} X_i X_i'$ are bounded below from zero with probability $1 - o(1)$, like in Lemma 5. In addition, on the event that the matrix $(n - n_k)^{-1}\sum_{i\notin I_k} X_i X_i'$ is non-singular, we have by Lemma 7 and the union bound that the expression

$$\mathrm{P}\left(\max_{\lambda\in\Lambda_n}\Big|\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} - \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n]\Big|^2 > T_n^2/4 \mid X_1^n\right)$$

is bounded from above by $C|\Lambda_n|\exp(-C\log n)$ for arbitrarily large constant C as long as the constant C in the statement of the lemma is large enough. Since $|\Lambda_n| \lesssim \log n$ by Assumption 4, it follows that

$$\mathrm{P}\left(\max_{\lambda\in\Lambda_n}\Big|\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} - \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n]\Big|^2 > T_n^2/4\right) = o(1).$$

Hence, the asserted claim follows. ∎

Lemma 9. *Suppose that Assumptions 1 – 5 hold. In addition, suppose that the conditional distribution of ε given X is Gaussian. Finally, suppose that $\xi_n^2 \log n/n = o(1)$. Then for all $k = 1, \ldots, K$,*

$$\|\widehat{\beta}_{-k}(\widehat{\lambda})\|_0 \lesssim s \cdot (\log^{3/2} p) \cdot (\log^2 n)$$

with probability $1 - o(1)$.

Proof. Fix $k = 1, \dots, K$. Similarly to the proof of Lemma 5, we have by Lemma 16 that the smallest eigenvalue of the matrix $(n - n_k)^{-1} \sum_{i \notin I_k} X_i X_i'$ is bounded from below by $c_1^2/2$ with probability $1 - o(1)$ since the smallest eigenvalue of the matrix $\mathrm{E}[XX']$ is bounded from below by c_1^2 by Assumption 1. Fix $X_1^n = (X_1, \dots, X_n)$ such that the smallest eigenvalue of the matrix $(n - n_k)^{-1} \sum_{i \notin I_k} X_i X_i'$ is bounded from below by $c_1^2/2$. Also, fix $\lambda \in \Lambda_{n,k}(X_1^n)$ for $\Lambda_{n,k}(X_1^n)$ defined in the statement of Lemma 8. Then $\mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n] \leq T_n$ for T_n defined in the statement of Lemma 8. Hence, by Fubini's theorem and Lemma 7, we have

$$\begin{aligned}
\mathrm{E}\Big[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k}^4 \mid X_1^n\Big] &= \int_0^\infty \mathrm{P}\Big(\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k}^4 > t \mid X_1^n\Big) dt \\
&\leq T_n^4 + \int_{T_n^4}^\infty \mathrm{P}\Big(\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} > t^{1/4} \mid X_1^n\Big) dt \\
&\lesssim T_n^4 + \int_{T_n^4}^\infty \exp\Big(- cn(t^{1/4} - T_n)^2\Big) dt \\
&\lesssim T_n^4 + \frac{1}{\sqrt{n}} \int_0^\infty \Big(t/\sqrt{n} + T_n\Big)^3 \exp(-ct^2) dt \lesssim T_n^4.
\end{aligned}$$

Thus,

$$\Big(\mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k}^4 \mid X_1^n]\Big)^{1/4} \lesssim T_n. \tag{17}$$

Then by Assumption 2 and Lemma 6 applied to the data $(X_i, Y_i)_{i \notin I_k}$ and the Lasso estimator $\widehat{\beta}_{-k}(\lambda)$,

$$\begin{aligned}
\mathrm{E}[\|\widehat{\beta}_{-k}(\lambda)\|_0 \mid X_1^n] &\lesssim \textstyle\sum_{i \notin I_k} \mathrm{E}[\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta) \mid X_1^n] \\
&\lesssim \mathrm{E}\Big[\Big\|\textstyle\sum_{i \notin I_k} \varepsilon_i X_i\Big\|_\infty \cdot \|\widehat{\beta}_{-k}(\lambda) - \beta\|_1 \mid X_1^n\Big] \\
&\leq \mathrm{E}\Big[\Big\|\textstyle\sum_{i \notin I_k} \varepsilon_i X_i\Big\|_\infty \cdot \|\widehat{\beta}_{-k}(\lambda) - \beta\| \cdot \sqrt{\|\widehat{\beta}_{-k}(\lambda)\|_0 + s} \mid X_1^n\Big] \\
&\leq \Big(\mathrm{E}\Big[\Big\|\textstyle\sum_{i \notin I_k} \varepsilon_i X_i\Big\|_\infty^2 \cdot \|\widehat{\beta}_{-k}(\lambda) - \beta\|^2 \mid X_1^n\Big] \cdot \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda)\|_0 + s \mid X_1^n]\Big)^{1/2},
\end{aligned}$$

where the last line follows from Hölder's inequality. In turn, with probability $1 - o(1)$,

$$\begin{aligned}
&\Big(\mathrm{E}\Big[\Big\|\textstyle\sum_{i \notin I_k} \varepsilon_i X_i\Big\|_\infty^2 \cdot \|\widehat{\beta}_{-k}(\lambda) - \beta\|^2 \mid X_1^n\Big]\Big)^{1/2} \\
&\quad \leq \Big(\mathrm{E}\Big[\Big\|\textstyle\sum_{i \notin I_k} \varepsilon_i X_i\Big\|_\infty^4 \mid X_1^n\Big] \cdot \mathrm{E}\Big[\|\widehat{\beta}_{-k}(\lambda) - \beta\|^4 \mid X_1^n\Big]\Big)^{1/4} \\
&\quad \lesssim \sqrt{n \log p}\Big(\mathrm{E}\Big[\|\widehat{\beta}_{-k}(\lambda) - \beta\|^4 \mid X_1^n\Big]\Big)^{1/4} \lesssim \sqrt{n \log p}\Big(\mathrm{E}\Big[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k}^4 \mid X_1^n\Big]\Big)^{1/4}
\end{aligned}$$

and the last expression is bounded from above up-to a constant C by

$$s^{1/2} \cdot (\log^{1/2} p) \cdot (\log p + \log\log n)^{1/2} + (\log^{1/2} p) \cdot (\log\log n)$$

by (17). Hence, with probability $1 - o(1)$,

$$\mathrm{E}[\|\widehat{\beta}_{-k}(\lambda)\|_0 \mid X_1^n] \lesssim s \cdot (\log p) \cdot (\log p + \log\log n) + (\log p) \cdot (\log\log n)^2.$$

So, by Markov's inequality and the union bound,

$$\begin{aligned}\|\widehat{\beta}_{-k}(\lambda)\|_0 &\lesssim (\log^{3/2} n) \cdot \Big(s \cdot (\log p) \cdot (\log p + \log\log n) + (\log p) \cdot (\log\log n)^2\Big) \\ &\lesssim s \cdot (\log^{3/2} p) \cdot (\log^2 n)\end{aligned}$$

for all $\lambda \in \Lambda_{n,k}(X_1^n)$ with probability $1 - o(1)$ since $|\Lambda_{n,k}(X_1^n)| \leq |\Lambda_n| \lesssim \log n$ by Assumption 4 and since $\log p \lesssim \log n$ by the assumption that $\xi_n^2 \log n/n = o(1)$ (recall that by Lemma 17, $\sqrt{p} \lesssim \xi_n$). The asserted claim follows since by Lemma 8, $\widehat{\lambda} \in \Lambda_{n,k}(X_1^n)$ for all $k = 1, \ldots, K$ with probability $1 - o(1)$. ∎

Lemma 10. *For all $\lambda \in \Lambda_n$ and $b \in \mathbb{R}^p$, we have*

$$\|\widehat{\beta}(\lambda) - b\|_{2,n}^2 \leq \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'b)^2 + \lambda\|b\|_1 - \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta}(\lambda))^2 - \lambda\|\widehat{\beta}(\lambda)\|_1.$$

Proof. The result in this lemma is sometimes referred to as the two point inequality; see van de Geer (2016). Here we give a short proof of this inequality using an argument similar to that of Lemma 5.1 in Chatterjee (2015). Fix $\lambda \in \Lambda_n$ and denote $\widehat{\beta} = \widehat{\beta}(\lambda)$. Take any $t \in (0, 1)$. We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})^2 + \lambda\|\widehat{\beta}\|_1 &\leq \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'(tb + (1-t)\widehat{\beta}))^2 + \lambda\|tb + (1-t)\widehat{\beta}\|_1 \\ &\leq \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta} + tX_i'(\widehat{\beta} - b))^2 + t\lambda\|b\|_1 + (1-t)\lambda\|\widehat{\beta}\|_1.\end{aligned}$$

Hence,

$$t\lambda(\|\widehat{\beta}\|_1 - \|b\|_1) \leq t^2\|\widehat{\beta} - b\|_{2,n}^2 + \frac{2t}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})(X_i'\widehat{\beta} - X_i'b),$$

and so

$$\lambda(\|\widehat{\beta}\|_1 - \|b\|_1) \leq t\|\widehat{\beta} - b\|_{2,n}^2 + \frac{2}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})(X_i'\widehat{\beta} - X_i'b).$$

Since $t \in (0, 1)$ is arbitrary, we obtain

$$\lambda(\|\widehat{\beta}\|_1 - \|b\|_1) \leq \frac{2}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})(X_i'\widehat{\beta} - X_i'b).$$

Thus,

$$\begin{aligned}\|\widehat{\beta} - b\|_{2,n}^2 &= \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'b - (Y_i - X_i'\widehat{\beta}))^2 \\ &= \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'b)^2 + \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})^2 - \frac{2}{n}\sum_{i=1}^n (Y_i - X_i'b)(Y_i - X_i'\widehat{\beta}) \\ &= \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'b)^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})^2 - \frac{2}{n}\sum_{i=1}^n (X_i'\widehat{\beta} - X_i'b)(Y_i - X_i'\widehat{\beta}) \\ &\leq \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'b)^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})^2 - \lambda(\|\widehat{\beta}\|_1 - \|b\|_1).\end{aligned}$$

The asserted claim follows. ∎

Proof of Theorem 1

Let

$$b = \frac{1}{K-1}\sum_{k=1}^{K}\frac{n-n_k}{n}\widehat{\beta}_{-k}(\widehat{\lambda}), \tag{18}$$

so that b is a convex combination of $\{\widehat{\beta}_{-1}(\widehat{\lambda}),\ldots,\widehat{\beta}_{-K}(\widehat{\lambda})\}$. We have

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - X_i'\widehat{\beta}(\widehat{\lambda}))^2 + n\widehat{\lambda}\|\widehat{\beta}(\widehat{\lambda})\|_1 &= \frac{1}{K-1}\sum_{k=1}^{K}\Big(\sum_{i\notin I_k}(Y_i - X_i'\widehat{\beta}(\widehat{\lambda}))^2 + (n-n_k)\widehat{\lambda}\|\widehat{\beta}(\widehat{\lambda})\|_1\Big)\\
&\geq \frac{1}{K-1}\sum_{k=1}^{K}\Big(\sum_{i\notin I_k}(Y_i - X_i'\widehat{\beta}_{-k}(\widehat{\lambda}))^2 + (n-n_k)\widehat{\lambda}\|\widehat{\beta}_{-k}(\widehat{\lambda})\|_1\Big)\\
&\geq \frac{1}{K-1}\sum_{k=1}^{K}\sum_{i\notin I_k}(Y_i - X_i'\widehat{\beta}_{-k}(\widehat{\lambda}))^2 + n\widehat{\lambda}\|b\|_1
\end{aligned}$$

where the second line follows from the definition of $\widehat{\beta}_{-k}(\widehat{\lambda})$'s and the third from the triangle inequality. Also,

$$\begin{aligned}
\frac{1}{K-1}\sum_{k=1}^{K}\sum_{i\notin I_k}(Y_i - X_i'\widehat{\beta}_{-k}(\widehat{\lambda}))^2 &\geq \frac{1}{K-1}\sum_{k=1}^{K}\sum_{i\notin I_k}\Big((Y_i - X_i'b)^2 + 2(Y_i - X_i'b)(X_i'b - X_i'\widehat{\beta}_{-k}(\widehat{\lambda}))\Big)\\
&= \sum_{i=1}^{n}(Y_i - X_i'b)^2 + \frac{2}{K-1}\sum_{k=1}^{K}\sum_{i\notin I_k}(Y_i - X_i'b)(X_i'b - X_i'\widehat{\beta}_{-k}(\widehat{\lambda})).
\end{aligned}$$

Thus, by Lemma 10,

$$n\|\widehat{\beta}(\widehat{\lambda}) - b\|_{2,n}^2 \leq \frac{2}{K-1}\sum_{k=1}^{K}\left|\sum_{i\notin I_k}(Y_i - X_i'b)(X_i'b - X_i'\widehat{\beta}_{-k}(\widehat{\lambda}))\right| \leq \frac{2}{K-1}\sum_{k=1}^{K}(\mathcal{I}_{1,k} + \mathcal{I}_{2,k})$$

where

$$\mathcal{I}_{1,k} = \left|\sum_{i\notin I_k}\varepsilon_i X_i'(b - \widehat{\beta}_{-k}(\widehat{\lambda}))\right|, \qquad \mathcal{I}_{2,k} = \left|\sum_{i\notin I_k}(X_i'\beta - X_i'b)\cdot(X_i'b - X_i'\widehat{\beta}_{-k}(\widehat{\lambda}))\right|. \tag{19}$$

Next, for all $k = 1,\ldots,K$, we have

$$\mathcal{I}_{1,k} \leq \max_{1\leq j\leq p}\left|\textstyle\sum_{i\notin I_k}\varepsilon_i X_{ij}\right|\cdot\|b - \widehat{\beta}_{-k}(\widehat{\lambda})\|_1.$$

Now, $\max_{1\le j\le p}|\sum_{i\notin I_k}\varepsilon_i X_{ij}|\lesssim\sqrt{n\log n}$ with probability $1-o(1)$. In addition, with probability $1-o(1)$, for all $k=1,\dots,K$,

$$\begin{aligned}\|b-\widehat{\beta}_{-k}(\widehat{\lambda})\|_1&\le\|b-\beta\|_1+\|\widehat{\beta}_{-k}(\widehat{\lambda})-\beta\|_1\lesssim\sum_{l=1}^K\|\widehat{\beta}_{-l}(\widehat{\lambda})-\beta\|_1\\&\lesssim\Big(s\cdot(\log^2 n)\cdot(\log^{3/2}p)\Big)^{1/2}\sum_{l=1}^K\|\widehat{\beta}_{-l}(\widehat{\lambda})-\beta\|\\&\lesssim\Big(s\cdot(\log^2 n)\cdot(\log^{3/2}p)\Big)^{1/2}\frac{s^{1/2}\cdot(\log^{1/4}p)\cdot(\log^{1/4}n)}{n^{1/2}}=\frac{s\cdot(\log p)\cdot(\log^{5/4}n)}{n^{1/2}}\end{aligned}$$

where the first line follows from the triangle inequality and (18), the second from Lemma 9, and the third from Lemma 4 (again, recall that the fact that the conditional distribution of ε given X is Gaussian combined with Assumption 2 implies that $\log\mathrm{E}[\exp(t\varepsilon)\mid X]\le C_1t^2$ for all $t>0$ if C_1 in this inequality is large enough) and the observation that $\log p\lesssim\log n$, which follows from $\xi_n^2\log n/n=o(1)$ and Lemma 17. Thus, with probability $1-o(1)$, for all $k=1,\dots,K$,

$$\mathcal{I}_{1,k}\lesssim s\cdot(\log p)\cdot(\log^{7/4}n).$$

Also, with probability $1-o(1)$, for all $k=1,\dots,K$,

$$\begin{aligned}\mathcal{I}_{2,k}&\le(n-n_k)\|\beta-b\|_{2,n,-k}\cdot\|b-\widehat{\beta}_{-k}(\widehat{\lambda})\|_{2,n,-k}\\&\lesssim(n-n_k)\Big(\textstyle\sum_{l=1}^K\|\widehat{\beta}_{-l}(\widehat{\lambda})-\beta\|_{2,n,-k}\Big)^2\lesssim s\log n\end{aligned}$$

where the first line follows from Hölder's inequality and the second from (18) and Lemmas 3 and 5 since $\log p\lesssim\log n$. Combining presented inequalities shows that with probability $1-o(1)$,

$$\|\widehat{\beta}(\widehat{\lambda})-b\|_{2,n}^2\lesssim\frac{s\log p}{n}\cdot(\log^{7/4}n).$$

Finally, with probability $1-o(1)$,

$$\begin{aligned}\|b-\beta\|_{2,n}^2&\lesssim\sum_{k=1}^K\|\widehat{\beta}_{-k}(\widehat{\lambda})-\beta\|_{2,n}^2\\&\lesssim\sum_{k=1}^K\Big(\|\widehat{\beta}_{-k}(\widehat{\lambda})-\beta\|_{2,n,k}^2+\|\widehat{\beta}_{-k}(\widehat{\lambda})-\beta\|_{2,n,-k}^2\Big)\lesssim\frac{s\log n}{n}\end{aligned}$$

by Lemmas 3 and 5. Thus, by the triangle inequality,

$$\|\widehat{\beta}(\widehat{\lambda})-\beta\|_{2,n}^2\lesssim\frac{s\log p}{n}\cdot(\log^{7/4}n)$$

with probability $1-o(1)$. This completes the proof of the theorem. ∎

Proof of Theorem 2

Let

$$\Lambda_n(X_1^n) = \left\{\lambda \in \Lambda_n \colon \mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n] \le C \cdot \left(\frac{s \log p}{n}\right)^{1/2} \cdot (\log^{7/8} n)\right\}$$

for some sufficiently large constant C. Since by Theorem 1,

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n}^2 \le \frac{C^2}{4} \cdot \frac{s \log p}{n} \cdot (\log^{7/4} n)$$

with probability $1 - o(1)$ if C is large enough, it follows by the same argument as that used in the proof of Lemma 8 that $\widehat{\lambda} \in \Lambda_n(X_1^n)$ with probability $1 - o(1)$.

Further, as in the proof of Lemma 9, fix $X_1^n = (X_1, \ldots, X_n)$ such that the smallest eigenvalue of the matrix $n^{-1} \sum_{i=1}^n X_i X_i'$ is bounded from below by $c_1^2/2$, which happens with probability $1 - o(1)$, and fix $\lambda \in \Lambda_n(X_1^n)$. Then

$$\left(\mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^4 \mid X_1^n]\right)^{1/4} \lesssim \left(\frac{s \log p}{n}\right)^{1/2} \cdot (\log^{7/8} n),$$

and so

$$\begin{aligned}
\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n] &\lesssim \sum_{i=1}^n \mathrm{E}[\varepsilon_i X_i'(\widehat{\beta}(\lambda) - \beta) \mid X_1^n] \\
&\lesssim \left(\mathrm{E}\left[\left\|\textstyle\sum_{i=1}^n \varepsilon_i X_i\right\|_\infty^4 \mid X_1^n\right]\right)^{1/4} \cdot \left(\mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^4 \mid X_1^n]\right)^{1/4} \cdot \left(\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 + s \mid X_1^n]\right)^{1/2} \\
&\lesssim s^{1/2} \cdot (\log p) \cdot (\log^{7/8} n) \cdot \left(\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 + s \mid X_1^n]\right)^{1/2} \\
&\lesssim s^{1/2} \cdot (\log^{15/8} n) \cdot \left(\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 + s \mid X_1^n]\right)^{1/2}
\end{aligned}$$

by the same argument as that used in the proof of Lemma 9. Therefore,

$$\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n] \lesssim s \log^{15/4} n.$$

Hence, given that $|\Lambda_n(X_1^n)| \le |\Lambda_n| \lesssim \log n$ by Assumption 4, it follows from Markov's inequality and the union bound that

$$\|\widehat{\beta}(\widehat{\lambda})\|_0 \lesssim s \log^5 n$$

with probability $1 - o(1)$. This completes the proof of the theorem. ∎

Proof of Corollary 1

Like in the proof of Lemma 5, it follows from Lemma 16 that

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|^2 \lesssim \|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n}^2 \Big/ \left(1 - \Gamma_n^2 \cdot (P(\|X\| > \xi_n)^{1/2} - \sqrt{\frac{\xi_n^2 \log(pn)}{n}} - \frac{\xi_n^2 \log(pn)}{n}\right)$$

with probability $1 - o(1)$, and so the first asserted claim follows from Theorem 1 and the assumption that $\xi_n^2 \log n = o(n)$. The second asserted claim follows from the first claim combined with the observation that

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1^2 \le (\|\widehat{\beta}(\widehat{\lambda})\|_0 + s) \cdot \|\widehat{\beta}(\widehat{\lambda}) - \beta\|^2$$

and Theorem 2. This completes the proof of the corollary. ∎

Lemma 11. *For all $\lambda \in \Lambda_n$ and $k = 1, \ldots, K$, we have $\|\widehat{\beta}_{-k}(\lambda)\|_0 \leq n$.*

Proof. Fix $\lambda \in \Lambda_n$ and $k = 1, \ldots, K$. Denote $\widehat{\beta} = \widehat{\beta}_{-k}(\lambda)$ and $\widehat{T} = \mathrm{supp}(\widehat{\beta})$. Suppose to the contrary that $\|\widehat{\beta}\|_0 > n$. Then the matrix $((X_1)_{\widehat{T}}, \ldots, (X_n)_{\widehat{T}})'$ does not have full column rank, and there exists $\gamma \in \mathbb{R}^p$ with $\|\gamma\| \neq 0$ and $\mathrm{supp}(\gamma) \subset \widehat{T}$ such that $X_i'\gamma = 0$ for all $i = 1, \ldots, n$. Also, the function $\alpha \mapsto \|\widehat{\beta} + \alpha\gamma\|_1$ mapping $\mathbb{R}$ into $\mathbb{R}$ is constant in some neighborhood around zero, and so as α increases in this neighborhood, some of the components of the vector $\widehat{\beta} + \alpha\gamma$ increase and others decrease. Thus, as we move α, we can always find a vector $\widehat{\beta} + \alpha\gamma$ that is a solution of the optimization problem (4) but is also such that $\|\widehat{\beta} + \alpha\gamma\|_0 < \|\widehat{\beta}\|_0$, which contradicts to our assumption that whenever the lasso optimization problem in (4) has multiple solutions, we choose one with the smallest number of non-zero components. This completes the proof of the lemma. ∎

Lemma 12. *Suppose that Assumptions 1 – 5 hold. Then we have for all $k = 1, \ldots, K$ that*

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 \lesssim (n^{2/q} M_n^2 \log^4 n \log p) \cdot \frac{s \log(pn)}{n}$$

with probability $1 - o(1)$.

Proof. Note that $\|\widehat{\beta}_{-k}(\widehat{\lambda})\|_0 \leq n$ by Lemma 11. Also, recall that $\|\beta\|_0 = s$. In addition,

$$\left(\mathrm{E}\left[\max_{i \notin I_k} \max_{1 \leq j \leq p} |X_{ij}|^2 \right] \right)^{1/2} \leq n^{1/q} M_n$$

for M_n defined in Assumption 3. Therefore, since all eigenvalues of the matrix $\mathrm{E}[XX']$ are bounded from above by Assumption 1, applying Lemma 15 with $k = n + s \lesssim n$ shows that with probability $1 - o(1)$,

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 \lesssim \|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|^2 \cdot n^{2/q} M_n^2 \log^4 n \log p.$$

Combining this inequality with Lemma 4 and noting that

$$\frac{s(\log p + \log\log n)}{n} + \frac{(\log\log n)^2}{n} \lesssim \frac{s \log(pn)}{n}$$

gives the asserted claim. ∎

Proof of Theorem 3

Define b as in (18). Also, for $k = 1, \ldots, K$, define $\mathcal{I}_{1,k}$ and $\mathcal{I}_{2,k}$ as in (19). Then it follows as in the proof of Theorem 1 that

$$n\|\widehat{\beta}(\widehat{\lambda}) - b\|_{2,n}^2 \leq \frac{2}{K-1} \sum_{k=1}^{K} (\mathcal{I}_{1,k} + \mathcal{I}_{2,k}).$$

Fix $k = 1, \dots, K$. To bound $\mathcal{I}_{1,k}$, we have by the triangle inequality and Lemmas 4 and 11 that

$$\|b - \widehat{\beta}_{-k}(\widehat{\lambda})\|_1 \leq \|b - \beta\|_1 + \|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1 \lesssim \sum_{l=1}^{K} \|\widehat{\beta}_{-l}(\widehat{\lambda}) - \beta\|_1$$
$$\leq \sqrt{n} \sum_{l=1}^{K} \|\widehat{\beta}_{-l}(\widehat{\lambda}) - \beta\| \lesssim (s \log(pn))^{1/2}$$

with probability $1 - o(1)$. Thus,

$$\mathcal{I}_{1,k} \leq \max_{1 \leq j \leq p} \left| \sum_{i \notin I_k} X_{ij} \varepsilon_i \right| \cdot \|b - \widehat{\beta}_{-k}(\widehat{\lambda})\|_1 \lesssim (n \log(pn))^{1/2} \cdot (s \log(pn))^{1/2} = (sn)^{1/2} \log(pn)$$

with probability $1 - o(1)$. Further, to bound $\mathcal{I}_{2,k}$, we have by Hölder's inequality and Lemmas 3 and 12 that

$$\mathcal{I}_{2,k} \leq (n - n_k) \|\beta - b\|_{2,n,-k} \cdot \|b - \widehat{\beta}_{-k}(\widehat{\lambda})\|_{2,n,-k}$$
$$\lesssim (n - n_k) \Big(\textstyle\sum_{l=1}^{K} \|\widehat{\beta}_{-l}(\widehat{\lambda}) - \beta\|_{2,n,-k} \Big)^2 \lesssim (n^{2/q} M_n^2 \log^4 n \log p) \cdot (s \log(pn))$$

with probability $1 - o(1)$. Hence,

$$\|\widehat{\beta}(\widehat{\lambda}) - b\|_{2,n}^2 \lesssim \left(\frac{s \log^2(pn)}{n} \right)^{1/2} + (n^{2/q} M_n^2 \log^4 n \log p) \cdot \frac{s \log(pn)}{n} \lesssim \left(\frac{s \log^2(pn)}{n} \right)^{1/2}$$

with probability $1 - o(1)$ where the second inequality holds by the assumption that

$$\frac{M_n^4 s (\log^8 n)(\log^2 p)}{n^{1-4/q}} \lesssim 1.$$

Finally, with probability $1 - o(1)$,

$$\|b - \beta\|_{2,n}^2 \lesssim \sum_{k=1}^{K} \|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n}^2$$
$$\lesssim \sum_{k=1}^{K} \Big(\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2 + \|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 \Big) \lesssim \Big(\frac{s \log^2(pn)}{n} \Big)^{1/2}$$

by Lemmas 3 and 12. Combining these inequalities and using the triangle inequality shows that

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n}^2 \lesssim \left(\frac{s \log^2(pn)}{n} \right)^{1/2}$$

with probability $1 - o(1)$. This completes the proof of the theorem. ∎

7 Technical Lemmas

Lemma 13. *Let $X_1, \dots, X_n$ be independent centered random vectors in $\mathbb{R}^p$ with $p \geq 2$. Define $Z = \max_{1 \leq j \leq p} |\sum_{i=1}^n X_{ij}|$, $M = \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |X_{ij}|$, and $\sigma^2 = \max_{1 \leq j \leq p} \sum_{i=1}^n \mathrm{E}[X_{ij}^2]$. Then*

$$\mathrm{E}[Z] \leq K \Big(\sigma \sqrt{\log p} + \sqrt{\mathrm{E}[M^2]} \log p \Big)$$

where K is a universal constant.

Proof. See Lemma E.1 in Chernozhukov, Chetverikov, and Kato (2014). ∎

Lemma 14. *Consider the setting of Lemma 13. For every $\eta > 0$, $t > 0$, and $q \geq 1$, we have*

$$P\Big(Z \geq (1+\eta)\mathrm{E}[Z] + t\Big) \leq \exp(-t^2/(3\sigma^2)) + K\mathrm{E}[M^q]/t^q$$

where the constant K depends only on η and q.

Proof. See Lemma E.2 in Chernozhukov, Chetverikov, and Kato (2014). ∎

Lemma 15. *Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^p$ with $p \geq 2$. Also, let $K = (\mathrm{E}[\max_{1\leq i\leq n} \max_{1\leq j\leq p} |X_{ij}^2|])^{1/2}$ and for $k \geq 1$, let*

$$\delta_n = \frac{K\sqrt{k}}{\sqrt{n}}\Big(\log^{1/2} p + (\log k)\cdot(\log^{1/2} p)\cdot(\log^{1/2} n)\Big).$$

Moreover, let $\mathcal{S}^p = \{\theta \in \mathbb{R}^p \colon \|\theta\| = 1\}$. Then

$$\mathrm{E}\left[\sup_{\theta\in\mathcal{S}^p\colon \|\theta\|_0\leq k} \Big|\frac{1}{n}\sum_{i=1}^n (X_i'\theta)^2 - \mathrm{E}[(X_1'\theta)^2]\Big|\right] \lesssim \delta_n^2 + \delta_n \sup_{\theta\in\mathcal{S}^p\colon \|\theta\|_0\leq k} \Big(\mathrm{E}[(X_1'\theta)^2]\Big)^{1/2}$$

up-to an absolute constant.

Proof. See Lemma B.1 in Belloni et al. (2015b). See also Rudelson and Vershynin (2008) for the original result. ∎

Lemma 16. *Let $X_1, \ldots, X_n$ be a random sample from the distribution of a p-dimensional random vector X such that for all $\delta \in \mathcal{S}^p$, we have $(\mathrm{E}[|X'\delta|^2)^{1/2} \leq C$ and $(\mathrm{E}[|X'\delta|^4)^{1/4} \leq \Gamma$ for some constants C and Γ. Then for any constants $\xi > 0$ and $t > 0$ such that $t > 0$, we have*

$$\begin{aligned} P\Big(\Big\|\tfrac{1}{n}\textstyle\sum_{i=1}^n X_iX_i' - E[XX']\Big\| > t + \Gamma^2\cdot(P(\|X\| > \xi))^{1/2}\Big) \\ \leq \exp\Big(\log(2p) - \frac{Ant^2}{\xi^2(1+t)}\Big) + nP(\|X\| > \xi) \end{aligned}$$

where A is a constant depending only on C.

Proof. An application of Corollary 6.2.1 in Tropp (2012) shows that

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n X_iX_i'1\{\|X_i\| \leq \xi\} - E\Big[XX'1\{\|X\| \leq \xi\}\Big]\right\| > t\right) \leq \exp\left(\log(2p) - \frac{Ant^2}{\xi^2(1+t)}\right) \quad (20)$$

where A depends only on C; see, for example, Lemma 10 in Chetverikov and Wilhelm (2015).

Now, let $\mathcal{D}$ be the event that $\|X_i\| \leq \xi$ for all $i = 1, \ldots, n$ and let $\mathcal{D}^c$ be its complement. Then $P(\mathcal{D}^c) \leq nP(\|X\| > \xi)$ by the union bound. Also,

$$\begin{aligned} \Big\|\mathrm{E}\Big[XX'1\{\|X\| > \xi\}\Big]\Big\| &= \sup_{\delta\in\mathbb{R}^p\colon \|\delta\|=1} \mathrm{E}\Big[|X'\delta|^2 1\{\|X\| > \xi\}\Big] \\ &\leq \sup_{\delta\in\mathbb{R}^p\colon \|\delta\|=1} \Big(\mathrm{E}[\|X'\delta|^4]\Big)^{1/2}\Big(P(\|X\| > \xi)\Big)^{1/2} \leq \Gamma^2\cdot\Big(P(\|X\| > \xi)\Big)^{1/2} \end{aligned}$$

by Hölder's inequality. Hence,

$$\begin{aligned}
&P\Big(\Big\|\frac{1}{n}\textstyle\sum_{i=1}^n X_iX_i' - \mathrm{E}[XX']\Big\| > t + \Gamma^2 \cdot (P(\|X\| > \xi))^{1/2}\Big) \\
&\quad \le P\Big(\Big\{\Big\|\frac{1}{n}\textstyle\sum_{i=1}^n X_iX_i' - \mathrm{E}[XX']\Big\| > t + \Gamma^2 \cdot (P(\|X\| > \xi))^{1/2}\Big\} \cap \mathcal{D}\Big) + P(\mathcal{D}^c) \\
&\quad \le P\Big(\Big\|\frac{1}{n}\textstyle\sum_{i=1}^n X_iX_i'1\{\|X_i\| \le \xi\} - \mathrm{E}[XX'1\{\|X\| \le \xi\}]\Big\| > t\Big) + nP(\|X\| > \xi)
\end{aligned}$$

by the triangle inequality. The asserted claim follows by combining this inequality with (20). ∎

Lemma 17. *Let X be a random vector in $\mathbb{R}^p$ implicitly indexed by n, where the dimension $p = p_n$ is also allowed to depend on n. Suppose that $(\mathrm{E}[|X'\delta|^2])^{1/2} \ge c_1$ for all $\delta \in \mathcal{S}^p$, $n \ge 1$, and some constant $c_1 > 0$. Also, suppose that $nP(\|X\| > \xi_n) = o(1)$ as $n \to \infty$ for some sequence of constants $(\xi_n)_{n\ge1}$. In addition, suppose that $\Gamma_n = \sup_{\delta\in\mathcal{S}^p}(\mathrm{E}[|X'\delta|^4])^{1/4}$ satisfies $\Gamma_n^4/n = o(1)$ as $n \to \infty$. Then $\xi_n^2 > p \cdot \min(1, c_1^2/2)$ for all sufficiently large n.*

Proof. Note that if $c_1^2/2 > 1$, the assumption that $(\mathrm{E}[|X'\delta|^2])^{1/2} \ge c_1$ for all $\delta \in \mathcal{S}^p$ implies that $(\mathrm{E}[|X'\delta|^2])^{1/2} \ge \sqrt{2}$ for all $\delta \in \mathcal{S}^p$. Hence, it suffices to consider the case where $c_1^2/2 \le 1$ and to show that $\xi_n^2 > c_1^2p/2$ for all sufficiently large n in this case.

To this end, suppose to the contrary that $\xi_n^2 \le c_1^2p/2$ for all $n = n_k$, $k \ge 1$, where $(n_k)_{k\ge1}$ is some increasing sequence of integers. Then

$$\mathrm{E}\Big[\|X\|^2 1\{\|X\| \le \xi_n\}\Big] \le \xi_n^2 \le c_1^2p/2$$

for all $n = n_k$. On the other hand,

$$\mathrm{E}[\|X\|^2] = \mathrm{E}[X'X] = \mathrm{tr}(\mathrm{E}[X'X]) = \mathrm{E}[\mathrm{tr}(X'X)] = \mathrm{E}[\mathrm{tr}(XX')] = \mathrm{tr}(\mathrm{E}[XX']) \ge c_1^2p$$

for all $n \ge 1$ by the assumption that $(\mathrm{E}[|X'\delta|^2])^{1/2} \ge c_1$ for all $\delta \in \mathcal{S}^p$, where for any $p \times p$ matrix A, we use $\mathrm{tr}(A)$ to denote the sum of its diagonal terms, that is, $\mathrm{tr}(A) = \sum_{j=1}^p A_{jj}$. Hence,

$$\mathrm{E}\Big[\|X\|^2 1\{\|X\| > \xi_n\}\Big] \ge c_1^2p/2 \tag{21}$$

for all $n = n_k$. In addition, the assumption that $nP(\|X\| > \xi_n) = o(1)$ implies that

$$\mathrm{E}\Big[\|X\|^2 1\{\xi_n < \|X\| \le \sqrt{p}\}\Big] = o(p/n)$$

as $n \to \infty$, and so it follows from (21) that

$$\mathrm{E}\Big[\|X\|^2 1\{\|X\| > \sqrt{p}\}\Big] \ge c_1^2p/4 \tag{22}$$

for all $n = n_k$ with sufficiently large k.

Let us now use (22) to bound $\Gamma_n = \sup_{\delta\in\mathcal{S}^p}(\mathrm{E}[|X'\delta|^4])^{1/4}$ from below to obtain a contradiction with the assumption that $\Gamma_n^4/n = o(1)$. Let $N = (N_1, \ldots, N_p)'$ be a standard Gaussian random vector in $\mathbb{R}^p$ that is independent of X. Then $U = N/\|N\|$ is distributed uniformly on S^p and is independent of $\|N\|$. Further, let $Z = X/\|X\|$. Then for all $\delta \in S^p$,

$$\mathrm{E}[|X'\delta|^4] \ge \mathrm{E}\Big[|X'\delta|^4 1\{\|X\| > \sqrt{p}\}\Big] = \mathrm{E}\Big[\|X\|^4 |Z'\delta|^4 1\{\|X\| > \sqrt{p}\}\Big],$$

and so

$$\sup_{\delta \in S^p} \mathrm{E}[|X'\delta|^4] \geq \mathrm{E}\Big[\|X\|^4 |Z'U|^4 1\{\|X\| > \sqrt{p}\}\Big] = \mathrm{E}\Big[\|X\|^4 1\{\|X\| > \sqrt{p}\} \mathrm{E}[|Z'U|^4 \mid X]\Big]. \quad (23)$$

On the other hand, for any non-stochastic $z \in S^p$,

$$3 = \mathrm{E}[|z'N|^4] = \mathrm{E}[\|N\|^4 \cdot |z'U|^4] = \mathrm{E}[\|N\|^4] \cdot \mathrm{E}[|z'U|^4] \leq 3p^2 \mathrm{E}[|z'U|^4],$$

so that $\mathrm{E}[|z'U|^4] \geq p^{-2}$. Hence, it follows from (23) that

$$\sup_{\delta \in S^p} \mathrm{E}[|X'\delta|^4] \geq p^{-2} \mathrm{E}\Big[\|X\|^4 1\{\|X\| > \sqrt{p}\}\Big]. \quad (24)$$

However, (22) implies by Hölder's inequality that

$$\begin{aligned} c_1^2 p/4 &\leq \mathrm{E}\Big[\|X\|^2 1\{\|X\| > \sqrt{p}\}\Big] = \mathrm{E}\Big[\|X\|^2 1\{\|X\| > \sqrt{p}\} \cdot 1\{\|X\| > \sqrt{p}\}\Big] \\ &\leq \Big(\mathrm{E}\Big[\|X\|^4 1\{\|X\| > \sqrt{p}\}\Big] \cdot P(\|X\| > \sqrt{p})\Big)^{1/2}, \end{aligned}$$

so that

$$\mathrm{E}\Big[\|X\|^4 1\{\|X\| > \sqrt{p}\}\Big] \geq \frac{c_1^4 p^2}{16 P(\|X\| > \sqrt{p})}$$

for all $n = n_k$ with sufficiently large k. Therefore, it follows from (24) that

$$\Gamma_n^4 = \sup_{\delta \in S^p} \mathrm{E}[|X'\delta|^4] \geq \frac{c_1^4}{16 P(\|X\| > \sqrt{p})}$$

for all $n = n_k$ with sufficiently large k. This contradicts to the assumptions that $nP(\|X\| > \xi_n) = o(1)$ and $\Gamma_n^4/n = o(1)$ since under our condition that $c_1^2/2 \leq 1$, we have $\xi_n^2 \leq c_1^2 p/2 \leq p$ and $P(\|X\| > \xi_n) \geq P(\|X\| > \sqrt{p})$ for all $n = n_k$. This completes the proof of the lemma. ∎

References

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, 40-79.

Athey, S. and Imbens, G. (2015). Lectures on machine learning. *NBER Summer Institute Lectures*.

Belloni, A. and Chernozhukov, V. (2011). High dimensional sparse econometric models: an introduction. *Chapter 3 in Inverse Problems and High-Dimensional Estimation*, **203**, 121-156.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521-547.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015a). Some new asymptotic theory for least squares series: pointwise and uniform results. *Journal of Econometrics*, **186**, 345-366.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2015b). Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework. *Arxiv:1512.07619*.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2013). Program evaluation with high-dimensional data. *Arxiv:1311.2645*.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, **37**, 1705-1732.

Boucheron, S., Lugosi, G., and Massart, P. (2013). Concentration inequalities: a nonasymptotic theory of independent. *Oxford University Press*.

Bülmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data: methods, theory, and applications. *Springer Series in Statistics*.

Cesarini, D., Dawes, C., Johannesson, M., Lichtenstein, P., and Wallace, B. (2009). Genetic variation in preferences for giving and risk taking. *Quarterly Journal of Economics*, **124**, 809-842.

Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *The Annals of Statistics*, **42**, 2340-2381.

Chatterjee, S. (2015). High dimensional regression and matrix estimation without tuning parameters. *Arxiv:1510.07294*.

Chatterjee, S. and Jafarov, J. (2015). Prediction error of cross-validated lasso. *Arxiv:1502.06292*.

Chen, L., Goldstein, L., and Shao, Q.-M. (2011). *Normal approximation by Stein's method*. Springer.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, **41**, 2786-2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Central limit theorems and bootstrap in high dimensions. *Arxiv:1412.3661*.

Chernozhukov, V., Gentzkow, M., Hansen, C., Shapiro, J., and Taddy, M. (2013). Econometric methods for high-dimensional data. *NBER Summer Institute Lectures*.

Chetverikov, D. and Wilhelm, D. (2015). Nonparametric instrumental variable estimation under monotonicity. *Arxiv:1507.05270*.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity: The Lasso and generalizations. *CRC Press*.

Homrighausen, D. and McDonald, D. (2013a). The lasso, persistence, and cross-validation. *Proceedings of the 30th International Conference on Machine Learning*, **28**.

Homrighausen, D. and McDonald, D. (2013b). Risk consistency of cross-validation with Lasso-type procedures. *Arxiv:1308.0810*.

Homrighausen, D. and McDonald, D. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Mach. Learn.*, **97**, 65-78.

Lecué, G. and Mitchell, C. (2012). Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics*, **6**, 1803-1837.

Li, K. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, **15**, 958-975.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, **21**, 255-285.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, **39**, 731-771.

Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, **61**, 1025-1045.

Saiz, A. and Simonsohn, U. (2013). Proxying for unobservable variables with internet document-frequency. *Journal of the European Economic Association*, **11**, 137-165.

Tao, T. (2012). *Topics in random matrix theory*. Graduate Studies in Mathematics, **132**.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

Tropp, J. (2012). *User-friendly tools for random matrices: an introduction*.

van de Geer, S. (2016). *Estimation and testing under sparsity*, Springer.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *Chapter 5 in Compressed Sensing: Theory and Applications*, Cambridge University Press.

Wager, S. and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *Arxiv:1510.04342*.

Wegkamp, M. (2003). Model selection in nonparametric regression. *The Annals of Statistics*, **31**, 252-273.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, **35**, 2173-2192.

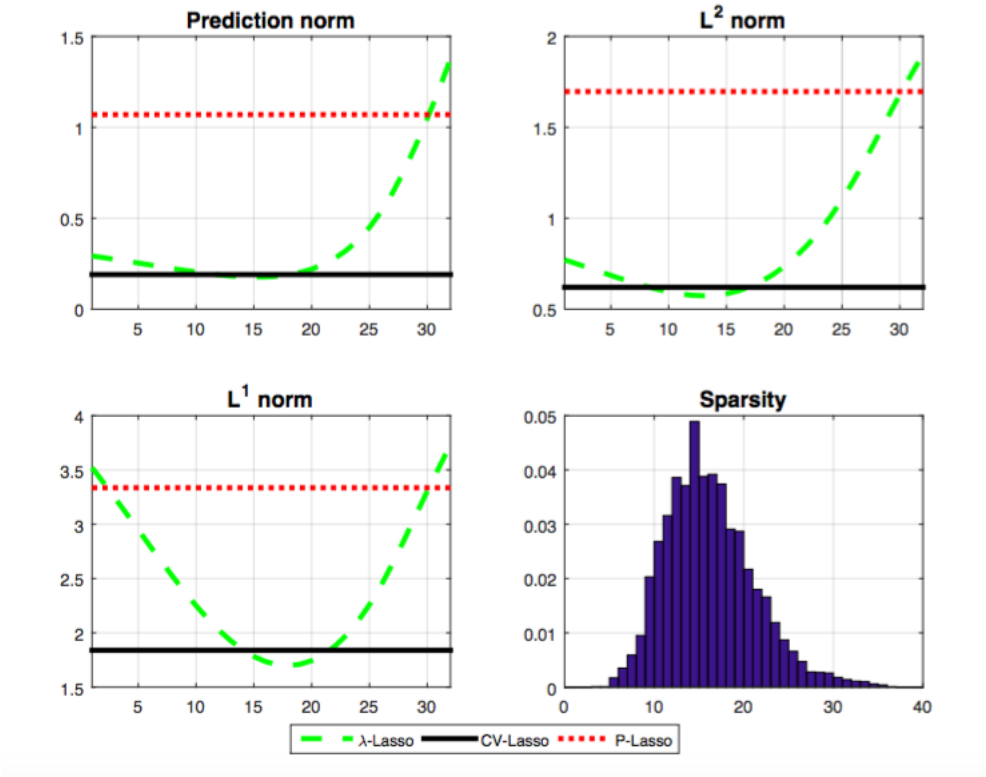

Figure 5.1: DGP1, $n = 100$, and $p = 40$. The top-left, top-right, and bottom-left panels show the mean of estimation error of Lasso estimators in the prediction, L^2, and L^1 norms. The dashed line represents the mean of estimation error of the Lasso estimator as a function of λ (we perform the Lasso estimator for each value of λ in the candidate set Λ_n; we sort the values in Λ_n from the smallest to the largest, and put the order of λ on the horizontal axis; we only show the results for values of λ up to order 32 as these give the most meaningful comparisons). The solid and dotted horizontal lines represent the mean of the estimation error of the cross-validated Lasso estimator and the Lasso estimator with λ chosen according to the Bickel-Ritov-Tsybakov rule, respectively.

Table 5.1: The mean of estimation error of Lasso estimators

	DGP1								
	Prediction norm			L^2 norm			L^1 norm		
	CV-Lasso	λ-Lasso	P-Lasso	CV-Lasso	λ-Lasso	P-Lasso	CV-Lasso	λ-Lasso	P-Lasso
(n, p)=(100, 40)	0.1902	0.1765	1.0697	0.6203	0.5738	1.6965	1.8389	1.6986	3.3357
(n, p)=(100, 100)	0.2834	0.2606	1.1944	0.8169	0.7569	1.7980	2.6853	2.3876	3.5653
(n, p)=(100, 400)	0.5137	0.4238	1.3545	1.2701	1.1651	1.9302	4.2179	3.5655	3.8927
(n, p)=(400, 40)	0.0467	0.0435	0.2903	0.2884	0.2633	0.8649	0.8330	0.7780	1.6872
(n, p)=(400, 100)	0.0688	0.0646	0.3422	0.3677	0.3307	0.9410	1.1309	1.0312	1.8374
(n, p)=(400, 400)	0.1108	0.1044	0.4208	0.5013	0.4537	1.0461	1.6110	1.3765	2.0449
	DGP2								
	Prediction norm			L^2 norm			L^1 norm		
	CV-Lasso	λ-Lasso	P-Lasso	CV-Lasso	λ-Lasso	P-Lasso	CV-Lasso	λ-Lasso	P-Lasso
(n, p)=(100, 40)	0.5912	0.5255	2.1759	1.1044	0.9915	2.3235	3.1826	2.9263	4.5308
(n, p)=(100, 100)	0.8850	0.7690	2.4533	1.4693	1.3092	2.4566	4.3838	3.9265	4.7881
(n, p)=(100, 400)	1.3417	1.1076	2.7829	1.9803	1.8530	2.6102	5.7023	4.7242	5.0990
(n, p)=(400, 40)	0.1381	0.1290	0.8518	0.4952	0.4530	1.4815	1.4282	1.3375	2.8924
(n, p)=(400, 100)	0.2051	0.1923	0.9827	0.6355	0.5718	1.5926	1.9483	1.7781	3.1136
(n, p)=(400, 400)	0.3340	0.3138	1.1466	0.8701	0.7868	1.7205	2.7962	2.3803	3.3735

Table 5.2: Probabilities for the number of non-zero coefficients of the cross-validated Lasso estimator hitting different brackets

	DGP1							
	[0, 5]	[6, 10]	[11, 15]	[16, 20]	[21, 25]	[26, 30]	[31, 35]	[36, p]
(n, p)=(100, 40)	0.0004	0.0818	0.3660	0.3464	0.1538	0.0388	0.0118	0.0010
(n, p)=(100, 100)	0.0000	0.0102	0.0814	0.2164	0.2596	0.1968	0.1210	0.1146
(n, p)=(100, 400)	0.0018	0.0230	0.0552	0.0720	0.0896	0.1234	0.1312	0.5038
(n, p)=(400, 40)	0.0010	0.1028	0.3906	0.3392	0.1288	0.0312	0.0058	0.0006
(n, p)=(400, 100)	0.0002	0.0192	0.1304	0.2658	0.2636	0.1672	0.0952	0.0584
(n, p)=(400, 400)	0.0000	0.0028	0.0236	0.0680	0.1338	0.1680	0.1670	0.4368
	DGP2							
	[0, 5]	[6, 10]	[11, 15]	[16, 20]	[21, 25]	[26, p]	[31, p]	[36, p]
(n, p)=(100, 40)	0.0164	0.1394	0.3430	0.3134	0.1334	0.0426	0.0096	0.0022
(n, p)=(100, 100)	0.0142	0.1116	0.1952	0.2002	0.1818	0.1400	0.0774	0.0796
(n, p)=(100, 400)	0.0300	0.0934	0.1646	0.1728	0.1426	0.1160	0.0778	0.2028
(n, p)=(400, 40)	0.0012	0.0988	0.4022	0.3322	0.1304	0.0308	0.0044	0.0000
(n, p)=(400, 100)	0.0002	0.0210	0.1360	0.2620	0.2560	0.1802	0.0872	0.0574
(n, p)=(400, 400)	0.0000	0.0024	0.0238	0.0766	0.1348	0.1664	0.1592	0.4368

Table 5.3: Probabilities for $\max_{1\leq j\leq p} n^{-1}|\sum_{i=1}^{n} X_{ij}\varepsilon_i|/\widehat{\lambda}$ hitting different brackets

	DGP1						
	[0, 0.5)	[0.6, 1)	[1, 1.5)	[1.5, 2)	[2, 2.5)	[2.5, 3)	[3, ∞)
(n, p)=(100, 40)	0.0000	0.0902	0.3478	0.2852	0.1404	0.0668	0.0696
(n, p)=(100, 100)	0.0000	0.1578	0.4446	0.2310	0.0976	0.0360	0.0330
(n, p)=(100, 400)	0.0124	0.3708	0.3426	0.1326	0.0570	0.0284	0.0562
(n, p)=(400, 40)	0.0002	0.1164	0.4352	0.2900	0.1072	0.0318	0.0192
(n, p)=(400, 100)	0.0000	0.2672	0.5664	0.1456	0.0186	0.0018	0.0004
(n, p)=(400, 400)	0.0000	0.5886	0.3956	0.0148	0.0008	0.0002	0.0000
	DGP2						
	[0, 0.5)	[0.6, 1)	[1, 1.5)	[1.5, 2)	[2, 2.5)	[2.5, 3)	[3, ∞)
(n, p)=(100, 40)	0.0018	0.1522	0.3474	0.2402	0.1308	0.0600	0.0676
(n, p)=(100, 100)	0.0066	0.3444	0.3732	0.1542	0.0710	0.0260	0.0246
(n, p)=(100, 400)	0.0380	0.6188	0.2250	0.0610	0.0242	0.0122	0.0208
(n, p)=(400, 40)	0.0000	0.1188	0.4450	0.2880	0.0996	0.0306	0.0180
(n, p)=(400, 100)	0.0000	0.2698	0.5764	0.1320	0.0196	0.0018	0.0004
(n, p)=(400, 400)	0.0000	0.5792	0.4028	0.0174	0.0006	0.0000	0.0000