arXiv:1605.02214v6 [math.ST] 6 Feb 2020

ON CROSS-VALIDATED LASSO IN HIGH DIMENSIONS[*]

By Denis Chetverikov[†] Zhipeng Liao[‡] Victor Chernozhukov[§]

UCLA[†], UCLA[‡], MIT[§]

In this paper, we derive non-asymptotic error bounds for the Lasso estimator when the penalty parameter for the estimator is chosen using K-fold cross-validation. Our bounds imply that the cross-validated Lasso estimator has nearly optimal rates of convergence in the prediction, L^2, and L^1 norms. For example, we show that in the model with the Gaussian noise and under fairly general assumptions on the candidate set of values of the penalty parameter, the estimation error of the cross-validated Lasso estimator converges to zero in the prediction norm with the $\sqrt{s \log p/n} \times \sqrt{\log(pn)}$ rate, 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 in the prediction norm up to a small logarithmic factor $\sqrt{\log(pn)}$, and similar conclusions apply for the convergence rate both in L^2 and in L^1 norms. Importantly, our results cover the case when p is (potentially much) larger than n and also allow for the case of non-Gaussian noise. Our paper therefore serves as a justification for the widely spread practice of using cross-validation as a method to choose the penalty parameter for the Lasso estimator.

1. Introduction. Since its invention by Tibshirani in [41], the Lasso estimator has become increasingly important in many fields, and a 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 proposed and theoretically justified; see [49], [13], [9], [38], and [4] among other papers. Nonetheless, in practice researchers often rely upon cross-validation to choose λ, see [19], and in fact, based on simulation evidence, using cross-validation to choose λ remains a leading recommendation in the theoretical literature (see textbook-level discussions

in [15], [26], and [25]). However, to the best of our knowledge, there exist very few results about properties of the Lasso estimator when λ is chosen using cross-validation; see a review below. The purpose of this paper is to fill this gap and to derive non-asymptotic error bounds for the cross-validated Lasso estimator in different norms.

We consider the regression model

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

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 and can be larger or even much larger than n. For simplicity of notation, however, we keep this dependence implicit.

We impose a standard assumption 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 relatively small. Under this assumption, the effective way to estimate β was proposed by Tibshirani in [41], who introduced the Lasso estimator,

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

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 time the version (2) has become more popular, probably for computational reasons). In principle, the optimization problem in (2) may have multiple solutions, but to simplify presentation and to avoid unnecessary technicalities, we assume throughout the paper, without further notice, that the distribution of X is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^p$, in which case the optimization problem in (2) has the unique solution with probability one; see Lemma 4 in [43]. Without this assumption, our results would apply to the sparsest solution.

To perform the Lasso estimator $\widehat{\beta}(\lambda)$, one has to choose the penalty parameter λ. If λ is chosen appropriately, the Lasso estimator attains the optimal rate of convergence under fairly general conditions; see, for example, [13], [8], and [34]. On the other hand, if λ is not chosen appropriately, the Lasso estimator may not be consistent or may have a slower rate of convergence; see [17]. Therefore, it is important to choose λ appropriately. In this

paper, we show that K-fold cross-validation indeed provides an appropriate way to choose λ. More specifically, we derive non-asymptotic error bounds for the Lasso estimator $\widehat{\beta}(\lambda)$ with $\lambda = \widehat{\lambda}$ being chosen by K-fold cross-validation in the prediction, L^2, and L^1 norms. Our bounds reveal that the cross-validated Lasso estimator attains the optimal rate of convergence up to certain logarithmic factors in all of these norms. For example, when the conditional distribution of the noise ε given X is Gaussian, the L^2 norm bound in Theorem 4.3 implies that

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2 = O_P\left(\sqrt{\frac{s\log p}{n}} \times \sqrt{\log(pn)}\right),$$

where for $b = (b_1, \ldots, b_p)' \in \mathbb{R}^p$, $\|b\|_2 = (\sum_{j=1}^p b_j^2)^{1/2}$ denotes the L^2 norm of b. Here, $\sqrt{s\log p/n}$ represents the optimal rate of convergence, and the cross-validated Lasso estimator attains this rate up to a small $\sqrt{\log(pn)}$ factor. Throughout the paper, we assume that K is fixed, i.e., independent of n. Our results therefore do not cover leave-one-out cross-validation.

Given that cross-validation is often used to choose the penalty parameter λ and given how popular the Lasso estimator is, understanding the rate of convergence of the cross-validated Lasso estimator seems to be an important research question. Yet, to the best of our knowledge, the only results in the literature about the cross-validated Lasso estimator are due to Homrighausen and McDonald [27, 28, 29] and Miolane and Montanari [32] but all these papers imposed extremely strong conditions and made substantial use of these conditions meaning that it is not clear how to relax them. In particular, [28] assumed that p is much smaller than n, and only showed consistency of the (leave-one-out) cross-validated Lasso estimator. [29], which strictly improves upon [27], assumed that the smallest value of λ in the candidate set, over which cross-validation search is performed, is so large that all considered Lasso estimators are guaranteed to be sparse, but, as we explain below, it is exactly the low values of λ that make the analysis of the cross-validated Lasso estimator difficult. (In addition, and equally important, the smallest value of λ in [29] exceeds the Bickel-Ritov-Tsybakov $\lambda = \lambda^*$, and we find via simulations that the cross-validated $\lambda = \widehat{\lambda}$ is smaller than λ^*, at least with high probability, whenever the candidate set is large enough, see Remarks 4.1 and 4.2 for further details; this suggests that the cross-validated λ based on the Homrighausen-McDonald candidate set will be with high probability equal to the smallest value in the candidate set, which makes the cross-validation search less interesting.) [32] assumed that p is proportional to n and that the vector X consists of i.i.d. Gaussian random variables, and their estimation error bounds do not converge to zero

whenever K is fixed (independent of n). In contrast to these papers, we allow p to be much larger than n and X to be non-Gaussian, with possibly correlated components, and we also allow for very large candidate sets.

Other papers that have been concerned with cross-validation in the context of the Lasso estimator include Chatterjee and Jafarov [19] and Lecué and Mitchell [30]. [19] 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, which is the target of study in our paper. [30] 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 (3) in the next section. Although the authors studied properties of the 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 also emphasize that our paper is not related to Abadie and Kasy [1] because they do consider the cross-validated Lasso estimator but in a very different setting, and, moreover, their results are in the spirit of those in [30]. (The results of [1] can be applied in the regression setting (1) but the application would require p to be smaller than n and their estimators in this case would differ from the cross-validated Lasso estimator studied here.)

Finally, we emphasize that deriving a rate of convergence of the cross-validated Lasso estimator is a non-standard problem. From the Lasso literature perspective, a fundamental problem is that most existing results require that λ is chosen so that $\lambda > 2\|n^{-1} \sum_{i=1}^{n} X_i \varepsilon_i\|_{\infty}$, at least with high probability, but, according to simulation evidence, this inequality typically does not hold if λ is chosen by cross-validation, meaning that existing results can not be used to analyze the cross-validated Lasso estimator; see Section 4 for more details and [25], page 105, for additional complications. Also, classical techniques to derive properties of cross-validated estimators developed, for example, in [31] do not apply to the Lasso estimator as those techniques are based on the linearity of the estimators in the vector of values $(Y_1, \ldots, Y_n)'$ of the dependent variable, which does not hold in the case of the Lasso estimator. More recent techniques, developed, for example, in [47], help to analyze sub-sample Lasso estimators like those studied in [30] but are not sufficient for the analysis of the full-sample Lasso estimator considered here. See [3] 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 novel sparsity bounds, which constitute one of the main building blocks in our analysis of the cross-validated Lasso estimator. In Section 6, we conduct a small Monte Carlo simulation study demonstrating that performance of the Lasso estimator based on the penalty parameter selected by cross-validation is comparable and often better than that of the Lasso estimator based on various plug-in rules. In Section 7, we provide proofs of the main results on the estimation error bounds. In Section 8, we provide proofs of our sparsity bounds. In Section 9, we collect 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 $b = (b_1, \ldots, b_p)' \in \mathbb{R}^p$, 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, $\|b\|_2 = (\sum_{j=1}^p b_j^2)^{1/2}$ to denote its L^2 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. Also, for any random variable Z, we use $\|Z\|_{\psi_1}$ and $\|Z\|_{\psi_2}$ to denote its ψ_1- and ψ_2- Orlicz norms. In addition, we denote $X_1^n = (X_1, \ldots, X_n)$. Moreover, we use $\mathcal{S}^p$ to denote the unit sphere in $\mathbb{R}^p$, that is, $\mathcal{S}^p = \{\delta \in \mathbb{R}^p \colon \|\delta\|_2 = 1\}$, and for any $\ell > 0$, we use $\mathcal{S}^p(\ell)$ to denote the ℓ-sparse subset of $\mathcal{S}^p$, that is, $\mathcal{S}^p(\ell) = \{\delta \in \mathcal{S}^p \colon \|\delta\|_0 \leq \ell\}$. We introduce more notation in the beginning of Section 7, as required for the proofs of the main results.

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 $\{1, \ldots, n\}$; that is, for each $k \in \{1, \ldots, K\}$, I_k is a subset of $\{1, \ldots, n\}$, for each $k, k' \in \{1, \ldots, K\}$ with $k \neq k'$, the sets I_k and $I_{k'}$ have empty intersection, and $\cup_{k=1}^K I_k = \{1, \ldots, n\}$. 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 $k = 1, \ldots, K$ and $\lambda \in \Lambda_n$, let

$$(3) \qquad \widehat{\beta}_{-k}(\lambda) = \underset{b \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{n - n_k} \sum_{i \notin I_k} (Y_i - X_i'b)^2 + \lambda\|b\|_1 \right)$$

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), the optimization problem in (3) has the unique solution with probability one under our maintained assumption that the distribution of X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$. Then the cross-validation choice of λ is

$$(4) \qquad \widehat{\lambda} = \arg\min_{\lambda \in \Lambda_n} \sum_{k=1}^{K} \sum_{i \in I_k} (Y_i - X_i' \widehat{\beta}_{-k}(\lambda))^2.$$

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 $\widehat{\beta}(\widehat{\lambda})$.

We emphasize one more time that although the properties of the estimators $\widehat{\beta}_{-k}(\widehat{\lambda})$ have been studied in great generality in [30], there are very few results in the literature regarding the properties of $\widehat{\beta}(\widehat{\lambda})$, which is the estimator used in practice.

3. Regularity Conditions. Recall that we consider the model in (1), the Lasso estimator $\widehat{\beta}(\lambda)$ in (2), and the cross-validation choice of λ in (4). Let c_1, C_1, a and q be some strictly positive numbers where $a < 1$ and $q > 4$. Also, let $r \geq 0$ be an integer. In addition, denote

$$(5) \qquad M_n = (\mathrm{E}[\|X\|_\infty^q])^{1/q}.$$

Throughout the paper, we assume that $s \geq 1$. Otherwise, one has to replace s by $s \vee 1$. To derive our results, we will impose the following regularity conditions.

ASSUMPTION 1 (Covariates). *The random vector $X = (X_1, \ldots, X_p)'$ is such that: (a) for all $\delta \in \mathcal{S}^p(n+s)$, we have $\mathrm{P}(|X'\delta| \geq c_1) \geq c_1$ and (b) for all $\delta \in \mathcal{S}^p(n^{2/q+c_1} M_n^2 s \log^3(pn))$, we have $(\mathrm{E}[|X'\delta|^2])^{1/2} \leq C_1$.*

Part (a) of this assumption can be interpreted as a probability version of the "no multicollinearity condition." It is slightly stronger than a more widely used expectation version of the same condition, namely $\mathrm{E}[(X'\delta)^2] \geq c_1$ for all $\delta \in \mathcal{S}^p(n+s)$ (with a possibly different value of the constant c_1), meaning that all $(n+s)$-sparse eigenvalues of the population Gram matrix $\mathrm{E}[XX']$ are bounded away from zero. Part (b) requires that sufficiently sparse eigenvalues of the matrix $\mathrm{E}[XX']$ are bounded from above uniformly over n. Note that neither part (a) nor part (b) of Assumption 1 imposes bounds on the eigenvalues of the empirical Gram matrix $n^{-1} \sum_{i=1}^n X_i X_i'$ (of course, if $p > n$, the smallest eigenvalue of this matrix is necessarily zero and the largest one can grow with n, potentially fast).

ASSUMPTION 2 (Growth condition). *The following growth condition is satisfied: $n^{4/q} M_n^4 s \log^4(pn) \leq C_1 n^{1-c_1}$.*

Assumption 2 is a mild growth condition restricting some moments of $\|X\|_\infty$, the number of non-zero coefficients in the model s and the number of parameters in the model p. When all components of the vector X are bounded by a constant almost surely, this assumption reduces to

$$s \log^4 p \leq C_1 n^{1-c_1}.$$

Thus, Assumptions 1 and 2 do allow for the high-dimensional case, with p being much larger than n. However, we note that these assumptions are stronger than those used with more conservative choices of λ; see [13, 8] for example.

ASSUMPTION 3 (Noise). *There exists a standard Gaussian random variable e that is independent of X and a function $Q \colon \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ that is thrice continuously differentiable with respect to the second argument such that $\varepsilon = Q(X, e)$ and for all $x \in \mathbb{R}^p$, (i) $c_1 \leq Q_2(x, e) \leq C_1(1 + |e|^r)$, (ii) $\|Q_{22}(x, e)\|_{\psi_2} \leq C_1$, and (iii) $\|Q_{222}(x, e)\|_{\psi_1} \leq C_1$, where we use index 2 to denote the derivatives with respect to the second argument, so that $Q_{222}(X, e) = \partial^3 Q(X, e)/\partial e^3$, for example.*

Letting Φ and $F_{\varepsilon|X}$ denote the cdf of the $N(0, 1)$ distribution and the conditional cdf of ε given X, respectively, it follows that whenever $F_{\varepsilon|X}$ is continuous almost surely, the random variable $e = \Phi^{-1}(F_{\varepsilon|X}(\varepsilon))$ has the $N(0, 1)$ distribution and is independent of X. In this case, we can guarantee that $\varepsilon = Q(X, e)$ by setting $Q(X, e) = Q_{\varepsilon|X}(\Phi(e))$, where $Q_{\varepsilon|X} = F_{\varepsilon|X}^{-1}$ is the conditional quantile function of ε given X. In addition, Assumption 3 imposes certain smoothness conditions. In particular, it requires that the transformation function $e \mapsto Q(X, e)$, which generates the noise variable ε from the $N(0, 1)$ variable e, is smooth in the sense that it satisfies certain derivative bounds.

Assumption 3 is rather non-standard. It appears in our analysis because, as explained in Remark 4.2 below, we rely upon the degrees of freedom formula for the Lasso estimator to establish some sparsity bounds. In turn, this formula, being a consequence of the Stein identity characterizing the standard Gaussian distribution, has a simple form whenever $\varepsilon \sim N(0, \sigma^2)$; see [49] and [42]. We extend this formula to the non-Gaussian case under the condition that the noise variable ε is a smooth transformation of $e \sim N(0, 1)$ as required by Assumption 3. Note that Assumption 3 requires the noise
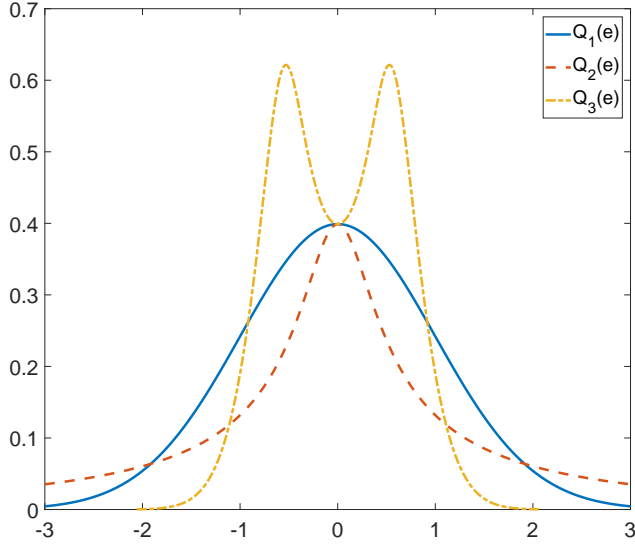
FIG 1. *The figure plots probability density functions of $\varepsilon = Q_j(e)$, $j = 1, 2, 3$, where $e \sim N(0, 1)$ and $Q_1(e) = e$, $Q_2(e) = e + e^3$, and $Q_3(e) = (e + e^3)/(1 + 2e^2)$. All three probability density functions are allowed by Assumption 3.*

variable ε to be neither sub-Gaussian nor sub-exponential. It does require, however, that the support of ε is $\mathbb{R}$. Note also that whenever ε is independent of X, we can choose the function $Q(X, e)$ to be independent of X, i.e. $Q(X, e) = Q(e)$. One simple example of a distribution that satisfies Assumption 3 is that of $\varepsilon = e + e^3$ with $e \sim N(0, 1)$. A more complicated example is $\varepsilon = (e + \gamma_1 e^3)/(1 + \gamma_2 e^2)$, where $\gamma_1, \gamma_2 > 0$ are such that $9\gamma_1^2 + \gamma_2^2 < 10\gamma_1\gamma_2$. Figure 1 presents plots of three probability density functions satisfying Assumption 3. Interestingly, the third one is bi-modal, which emphasizes the fact that Assumption 3 allows for a wide variety of distributions. Finally, note that Assumption 3 holds with $r = 0$ if the conditional distribution of ε given X is Gaussian.

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 [13] that the optimal rate of convergence of the Lasso estimator is achieved when λ is of order $(\log p/n)^{1/2}$. Since under Assumption 2, 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_n

that would yield the Lasso estimator with the optimal rate of convergence in the prediction norm. Note also that Assumption 4 allows for a rather large 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 (10) in Lemma 7.2 with probability $1 - Cn^{-c}$ and that $|\Lambda_n| \leq C \log n$, where c and C are some constants. Thus, we could for example set $\Lambda_n = \{a^l \colon l = \ldots, -2, -1, 0, 1, 2, \ldots; \ a^{-l} \leq n^{C_1}, \ a^l \leq n^{C_1}\}$.

ASSUMPTION 5 (Dataset partition). *The dataset partition $\{I_k\}_{k=1}^K$ is such that for all $k = 1, \ldots, K$, we have $n_k/n \geq c_1$, where $n_k = |I_k|$.*

Assumption 5 is mild and is typically imposed in the literature on K-fold cross-validation. This assumption ensures that the subsamples I_k are balanced in the sample size.

4. Main Results. Our first main result in this paper gives a non-asymptotic estimation error bound for the cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ in the prediction norm.

THEOREM 4.1 (Prediction Norm Bound). *Suppose that Assumptions 1 – 5 hold. Then for any $\alpha \in (0, 1)$,*

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n} \leq \sqrt{\frac{Cs \log(p/\alpha)}{n}} \times \sqrt{\log(pn) + s^{-1} \log^{r+1} n}$$

with probability at least $1 - \alpha - Cn^{-c}$, where $c, C > 0$ are constants depending only on c_1, C_1, K, a, q, and r.

REMARK 4.1 (Near-rate-optimality of cross-validated Lasso estimator in prediction norm). The results in [13] imply that under the assumptions of Theorem 4.1, setting $\lambda = \lambda^* = (C \log p/n)^{1/2}$ for sufficiently large constant C, which depends on the distribution of ε, 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 [34] that this is the optimal rate of convergence (in the minimax sense) for the estimators of β in the model (1). Therefore, Theorem 4.1 implies that the cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ has the fastest possible rate of convergence in the prediction norm up to the small $(\log(pn) + s^{-1} \log^{r+1} n)^{1/2}$ factor. Note, however, that implementing the cross-validated Lasso estimator does not require knowledge of the distribution of ε, which makes this

estimator attractive in practice. In addition, simulation evidence suggests that $\widehat{\beta}(\widehat{\lambda})$ often outperforms $\widehat{\beta}(\lambda^*)$, which is one of the main reasons why cross-validation is typically recommended as a method to choose λ. The rate of convergence following from Theorem 4.1 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 having non-zero coefficients; see, for example, [11].   ■

REMARK 4.2 (On the proof of Theorem 4.1). One of the main steps in [13] is to show that outside of the event

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

where $c > 2$ is some constant, the Lasso estimator $\widehat{\beta}(\lambda)$ satisfies the bound $\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \leq C\lambda\sqrt{s}$, where C is a constant. Thus, to obtain the Lasso estimator with a fast rate of convergence, it suffices to choose λ such that it is small enough but the event (6) holds with at most small probability. The choice $\lambda = \lambda^*$ described in Remark 4.1 satisfies these two conditions. The difficulty with cross-validation, however, is that, as we demonstrate in Section 6 via simulations, it typically yields a rather small value of λ, so that the event (6) with $\lambda = \widehat{\lambda}$ holds with non-trivial (in fact, large) 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 $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_2$ for the subsample Lasso estimators $\widehat{\beta}_{-k}(\widehat{\lambda})$ defined in (3). Second, we use the degrees of freedom formula of [49] and [42] to show that these estimators are sparse and to derive bounds on $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_1$ and $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n}$. Third, we use the two point inequality stating that for all $b \in \mathbb{R}^p$ and $\lambda > 0$,

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

which can be found in [44], with $\lambda = \widehat{\lambda}$ and $b = (K-1)^{-1} \sum_{k=1}^{K}(n - n_k)\widehat{\beta}_{-k}(\widehat{\lambda})/n$, 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 $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_2$ and $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_1$. Finally, we use the

triangle inequality to obtain a bound on $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n}$ from the bounds on $\|\widehat{\beta}(\widehat{\lambda}) - b\|_{2,n}$ and $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n}$. The details of the proof can be found in Section 7. ∎

Next, in order to obtain bounds on $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1$ and $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2$, we derive a sparsity bound for $\widehat{\beta}(\widehat{\lambda})$, that is, we show that the estimator $\widehat{\beta}(\widehat{\lambda})$ has relatively few non-zero components, at least with high-probability. Even though our sparsity bound is not immediately useful in applications itself, it will help us to translate the result in the prediction norm in Theorem 4.1 into the result in L^1 and L^2 norms in Theorem 4.3.

THEOREM 4.2 (Sparsity Bound). *Suppose that Assumptions 1 – 5 hold. Then for any $\alpha \in (0, 1)$,*

$$\|\widehat{\beta}(\widehat{\lambda})\|_0 \leq Cs \times \frac{(\log^2 p)(\log n)(\log(pn) + s^{-1}\log^{r+1} n)}{\alpha}$$

with probability at least $1 - \alpha - Cn^{-c}$, where $c, C > 0$ are constants depending only on c_1, C_1, K, a, q, and r.

REMARK 4.3 (On the sparsity bound). [9] showed that outside of the event (6), the Lasso estimator $\widehat{\beta}(\lambda)$ satisfies the bound $\|\widehat{\beta}(\lambda)\|_0 \leq Cs$, for some constant C, 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 4.2, however, cross-validation typically yields a rather small value of λ, so that the event (6) with $\lambda = \widehat{\lambda}$ holds with non-trivial (in fact, large) 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. We therefore should not necessarily expect that the inequality $\|\widehat{\beta}(\widehat{\lambda})\|_0 \leq Cs$ holds with large probability. In fact, it is well-known (from simulations) in the literature that the cross-validated Lasso estimator typically satisfies $\|\widehat{\beta}(\widehat{\lambda})\|_0 \gg s$. Our theorem, however, shows that even though the event (6) with $\lambda = \widehat{\lambda}$ may hold with large probability, the number of non-zero components in the cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ may exceed Cs only by the relatively small $(\log^2 p)(\log n)(\log(pn) + s^{-1}\log^{r+1})$ factor. ∎

With the help of Theorems 4.1 and 4.2, we immediately obtain the following bounds on the L^1 and L^2 norms of the estimation error of the cross-validated Lasso estimator, which is our second main result in this paper.

THEOREM 4.3 (L^1 and L^2 Norm Bounds). *Suppose that Assumptions 1 – 5 hold. Then for any $\alpha \in (0,1)$,*

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2 \leq \sqrt{\frac{Cs\log(p/\alpha)}{n}} \times \sqrt{\log(pn) + s^{-1}\log^{r+1} n}$$

and

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1 \leq \sqrt{\frac{Cs^2\log(p/\alpha)}{n}} \times \sqrt{\frac{(\log^2 p)(\log n)(\log(pn) + s^{-1}\log^{r+1} n)^2}{\alpha}}$$

with probability at least $1-\alpha-Cn^{-c}$, where $c, C > 0$ are constants depending only on c_1, C_1, K, a, q, and r.

REMARK 4.4 (Near-rate-optimality of cross-validated Lasso estimator in L^1 and L^2 norms). Like in Remark 4.1, the results in [13] imply that under the assumptions of Theorem 4.3, setting $\lambda = \lambda^* = (C\log p/n)^{1/2}$ for sufficiently large constant C gives the Lasso estimator $\widehat{\beta}(\lambda^*)$ satisfying $\|\widehat{\beta}(\lambda^*) - \beta\|_2 = O_P((s\log p/n)^{1/2})$ and $\|\widehat{\beta}(\lambda^*) - \beta\|_1 = O_P((s^2\log p/n)^{1/2})$, and one can use the methods from [34] to show that these rates are optimal. Therefore, the cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ has the fastest possible rate of convergence both in L^1 and in L^2 norms, up to small logarithmic factors. ∎

REMARK 4.5 (On the case with Gaussian noise). Recall that whenever the conditional distribution of ε given X is Gaussian, we can take $r = 0$ in Assumption 3. Thus, it follows from Theorems 4.1 and 4.3 that, in this case, we have

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n} \vee \|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2 \leq \sqrt{\frac{Cs\log(p/\alpha)}{n}} \times \sqrt{\log(pn)}$$

with probability at least $1-\alpha-Cn^{-c}$ for any $\alpha \in (0,1)$ and some constants $c, C > 0$. Theorems 4.2 and 4.3 can also be used to obtain the sparsity and L^1 norm bounds in this case as well. However, the sparsity and L^1 norm bounds here can be improved using results in [6]. In particular, assuming that the conditional distribution of ε given X is $N(0, \sigma^2)$ for some constant $\sigma > 0$, it follows from Theorem 4.3 in [6] that for any $\lambda > 0$,

$$\mathrm{Var}(\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n) \leq \mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n]\left(3 + 4\log\left(\frac{ep}{\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n]}\right)\right).$$

Combining this result and the same arguments as those in the proofs of Theorems 4.2 and 4.3, with Chebyshev's inequality replacing Markov's inequality in the proof of Theorem 4.2, we have

$$\|\widehat{\beta}(\widehat{\lambda})\|_0 \le Cs \times \frac{(\log^2 p)\log(pn)}{\sqrt{\alpha}}$$

and

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1 \le \sqrt{\frac{Cs^2 \log(p/\alpha)}{n}} \times \frac{(\log p)\log(pn)}{\alpha^{1/4}}$$

with probability at least $1 - \alpha - Cn^{-c}$. ∎

The near-rate-optimality of the cross-validated Lasso estimator in Theorem 4.1 may be viewed as an in-sample prediction property since the prediction norm

$$\|\widehat{\beta} - \beta\|_{2,n} = \left(\frac{1}{n}\sum_{i=1}^n (X_i'\widehat{\beta} - X_i'\beta)^2\right)^{1/2}$$

evaluates estimation errors with respect to the observed data $X_1, \ldots, X_n$. In addition, we can define an out-of-sample prediction norm

$$\|\widehat{\beta} - \beta\|_{p,2,n} = \left(\mathrm{E}\left[(X'\widehat{\beta} - X'\beta)^2 \mid (X_i, Y_i)_{i=1}^n\right]\right)^{1/2},$$

where X is independent of $X_1, \ldots, X_n$. Using Theorems 4.2 and 4.3, we immediately obtain the following corollary on the estimation error of the cross-validated Lasso estimator in the out-of-sample prediction norm:

COROLLARY 4.1 (Out-of-Sample Prediction Norm Bounds). *Suppose that Assumptions 1 – 5 hold. Then for any $\alpha \in (0,1)$,*

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{p,2,n} \le \sqrt{\frac{Cs \log(p/\alpha)}{n}} \times \sqrt{\log(pn) + s^{-1}\log^{r+1} n}$$

with probability at least $1 - \alpha - Cn^{-c}$, where $c, C > 0$ are constants depending only on $c_1, C_1, K, a, q,$ and r.

5. General Sparsity Bounds. As we mentioned in Remark 4.2, our analysis of the (full-sample) cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ requires understanding sparsity of the sub-sample cross-validated Lasso estimators $\widehat{\beta}_{-k}(\widehat{\lambda})$, that is, we need a sparsity bound showing that $\|\widehat{\beta}_{-k}(\widehat{\lambda})\|_0$, $k = 1, \ldots, K$, are sufficiently small, at least with high probability. Unfortunately, existing sparsity bounds are not good enough for our purposes because, as

we discussed in Remark 4.3, they only apply outside of the event (6) and this event holds with non-trivial (in fact, large) probability if we set $\lambda = \widehat{\lambda}$. We therefore develop here two novel sparsity bounds. The crucial feature of our bounds is that they apply for all values of λ, both large and small, independently of whether (6) holds or not. Roughly speaking, the first bound shows, under mild conditions, that the Lasso estimator $\widehat{\beta}(\lambda)$ has to be sparse, at least with large probability, whenever it has small estimation error in the L^2 norm. The second bound shows, under somewhat stronger conditions, that the Lasso estimator $\widehat{\beta}(\lambda)$ has to be sparse, at least on average, whenever it has small estimation error in the prediction norm. Both bounds turn out useful in our analysis.

THEOREM 5.1 (Sparsity Bound via Estimation Error in L^2 Norm).   *Suppose that Assumption 3 holds and let $\bar{C} > 0$ be some constant. Then for all $\lambda > 0$ and $t \geq 1$,*

$$
\mathrm{P}\left(\|\widehat{\beta}(\lambda)\|_0 \leq Cts\log^2(pn)\left(\log^r n + \frac{n(\log^2 n)\|\widehat{\beta}(\lambda) - \beta\|_2^2}{s\log(pn)} \right) \mid X_1^n \right)
$$
$$
\leq 1 - \frac{2}{ts\log(pn)} - \frac{2}{n},
$$

on the event $\sup_{\delta \in \mathcal{S}^p(s)} \|\delta\|_{2,n} \leq \bar{C}$, where $C > 0$ is a constant depending only on c_1, C_1, $\bar{C}$, and r.

THEOREM 5.2 (Sparsity Bound via Estimation Error in Prediction Norm). *Suppose that Assumption 3 holds and let $\bar{c}, \bar{C} > 0$ be some constants. Then for all $\lambda > 0$,*

$$
\mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n] \leq s + C(\log p)(nR_n(\lambda)^2 + \log^r n)
$$

on the event

$$
(7) \qquad \bar{c} \leq \inf_{\delta \in \mathcal{S}^p(J_n(\lambda))} \|\delta\|_{2,n} \quad and \quad \max_{1 \leq j \leq p} \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2} \leq \bar{C},
$$

where

$$
J_n(\lambda) = n^{1/2 + c_1/8}(\sqrt{n}R_n(\lambda) + 1), \;\; R_n(\lambda) = \mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n],
$$

and $C > 0$ is a constant depending only on c_1, C_1, $\bar{c}$, $\bar{C}$, and r.

6. 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 the 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 $X = (X_1, \ldots, X_p)'$ from the Gaussian distribution with mean zero and variance-covariance matrix given by $E[X_j X_k] = \rho^{|j-k|}$ for all $j, k = 1, \ldots, p$ with $\rho = 0.5$ and 0.75. 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. For 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 estimators with λ chosen according to [38] and [4], i.e.,

$$\lambda = n^{-1/2} \sigma \sqrt{2 \log p} \text{ and } \lambda = n^{-1/2} \sigma \sqrt{2 \log(p/s)}$$

respectively. These Lasso estimators achieve the optimal convergence rate under the prediction norm (see, e.g., [38] and [4]). The noise level σ and the true sparsity s typically have to be estimated from the data but for simplicity we assume that both σ and s are known, so we set $\sigma = 1$ and $s = 4$ in DGP1, and $\sigma = \sqrt{3}$ and $s = 4$ in DGP2. In what follows, these Lasso estimators are denoted as SZ-Lasso and B-Lasso estimators respectively, and the cross-validated Lasso estimator is denoted as CV-Lasso.

Figure 2 contains simulation results for DGP1 with $n = 100$, $p = 40$ and $\rho = 0.75$. The first four (that is, the top-left, top-right, middle-left and middle-right) panels of Figure 2 present the mean of the estimation error of the Lasso estimators in the prediction, L^2, L^1, and out-of-sample prediction norms, respectively. The out-of-sample prediction norm is defined as $\|b\|_{p,2,n} = (\mathrm{E}[(X'b)^2])^{1/2}$ for all $b \in \mathbb{R}^p$. In these panels, 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 25 as these give the most meaningful comparisons). This estimator is denoted as λ-Lasso. The solid, dotted and dashed-dotted horizontal lines represent the mean of the estimation error of CV-Lasso, SZ-Lasso, and B-Lasso, respectively.

From the top four panels of Figure 2, we see that 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, SZ-Lasso and B-Lasso tend to have larger estimation error in all four norms.

The bottom-left and bottom-right panels of Figure 2 depict the histograms for the numbers of non-zero coefficients of the CV-Lasso estimator and B-Lasso estimator respectively. Overall, these panels suggest that the CV-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. The B-Lasso estimator is more sparse than the CV-Lasso estimator: it selects around 5 to 15 covariates with large probability.

Figure 3 includes the simulation results for DGP1 when $n = 100$, $p = 400$ and $\rho = 0.75$. The estimation errors of the Lasso estimators are inflated when p is much bigger than the sample size. The estimation error of CV-Lasso under the prediction norm is increased from 0.4481 to 0.7616 when p is increased from 40 to 400, although it remains the best compared with SZ-Lasso and B-Lasso estimators. Similar phenomena are observed for the estimation error under the L^2 norm and the out-of-sample prediction norm. On the other hand, the estimation error of the CV-Lasso is slightly larger than the SZ-Lasso and B-Lasso under the L^1 norm. For the sparsity of the Lasso estimators, the CV-Lasso is much less sparse than the B-Lasso: it selects around 5 to 50 covariates with large probability while the B-Lasso only selects 8 to 22 covariates with large probability.

For all other experiments, the simulation results on the mean of estimation error of the Lasso estimators can be found in Table 1. For simplicity, we only report the minimum over $\lambda \in \Lambda_n$ of mean of the estimation error of λ-Lasso and the mean of the estimation error of B-Lasso in Table 1. The results in Table 1 confirm findings in Figure 2 and Figure 3: the mean of the estimation error of CV-Lasso is close to the minimum mean of the estimation errors of the λ-Lasso estimators under both DGPs for all combinations of n, p and ρ considered in all three norms. Their difference becomes smaller when the

sample size n increases. The mean of the estimation error of B-Lasso is larger than that of CV-Lasso in cases when p is relatively small or the regressors X have strong correlation, while the B-Lasso has smaller estimation error when p is much larger than n and the regressors X are weakly correlated. When the correlations of the regressors X become stronger and the largest eigenvalue of $\mathrm{E}[XX']$ becomes bigger, the mean of the estimation error of the CV-Lasso estimator is slightly enlarged and is much less effected compared with the B-Lasso estimator. For example, in DGP1 with $n = 100$ and $p = 40$, the mean of estimation error of CV-Lasso estimator increases 5.39% when ρ is changed from 0.5 to 0.75 (and the largest eigenvalue of $\mathrm{E}[XX']$ increases from 2.97 to 6.64), while the B-Lasso estimator has a 28% increase.

Table 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 2 confirm findings in Figure 2 and Figure 3: 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 3 provides information on the finite-sample distribution of the ratio of the maximum score $\max_{1 \le j \le p} |n^{-1} \sum_{i=1}^{n} X_{ij} \varepsilon_i|$ over $\widehat{\lambda}$, the cross-validation choice of λ. More precisely, the table shows probabilities for this ratio hitting different brackets. From Table 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 with large probability, meaning that deriving the rate of convergence of the cross-validated Lasso estimator requires new arguments since existing arguments only work for the case when (6) does not hold; see discussion in Remark 4.2 above.

7. Proofs for Section 4. In this section, we prove Theorems 4.1, 4.2, and 4.3 and Corollary 4.1. Since the proofs are long, we start with a sequence of preliminary lemmas in Subsection 7.1 and give the actual proofs of the theorems and the corollary in Subsections 7.2, 7.3, 7.4, and 7.5, respectively.

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 strictly positive constants that can change from place to place but that can be chosen to depend only on c_1, C_1, K, a, q, and r. We use the notation $a_n \lesssim b_n$ if $a_n \le Cb_n$. 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.

7.1. *Preliminary Lemmas.* Here, we collect preliminary lemmas that help to prove Theorems 4.1–4.3.

LEMMA 7.1. *Suppose that Assumptions 1 and 2 are satisfied and denote* $\ell_n = \sqrt{sn^{1+c_1/2}} \log(pn)$. *Then*

$$
(8) \qquad \sup_{\theta \in \mathcal{S}^p(\ell_n)} \left| \frac{1}{n} \sum_{i=1}^n (X_i'\theta)^2 - \mathrm{E}[(X'\theta)^2] \right| \le Cn^{-c}
$$

with probability at least $1 - Cn^{-c}$, *where* c, $C > 0$ *are some constants depending only on* c_1, C_1, *and* q.

PROOF. In this proof, c and C are strictly positive constants that depend only on c_1, C_1, and q but their values can change from place to place. By Jensen's inequality and the definition of M_n in (5),

$$
K_n := \left(\mathrm{E} \left[\max_{1 \le i \le n} \max_{1 \le j \le p} |X_{ij}|^2 \right] \right)^{1/2}
$$

$$
(9) \qquad \qquad \le \left(\mathrm{E} \left[\max_{1 \le i \le n} \max_{1 \le j \le p} |X_{ij}|^q \right] \right)^{1/q} \le n^{1/q} M_n.
$$

Therefore, given that $\ell_n \le Cn^{1-c}$ by Assumptions 1(a) and 2, which implies $\log \ell_n \le C \log n$, it follows that

$$
\delta_n := K_n \sqrt{\ell_n \log p / n} \left(1 + (\log \ell_n)(\log n)^{1/2} \right) \le Cn^{-c}
$$

by Assumption 2; here, Assumption 1(a) is used only to verify that $M_n \ge c$. Also, denoting $\ell_{n,0} = n^{2/q+c_1} M_n^2 s \log^3(pn)$,

$$
\sup_{\theta \in \mathcal{S}^p(\ell_n)} \mathrm{E}[(X'\theta)^2] \le \frac{2(\ell_n + \ell_{n,0})}{\ell_{n,0}} \sup_{\theta \in \mathcal{S}^p(\ell_{n,0})} \mathrm{E}[(X'\theta)^2] \le \frac{C(\ell_n + \ell_{n,0})}{\ell_{n,0}}
$$

by Lemma 9 in [8] and Assumption 1(b). Thus,

$$
\delta_n \sup_{\theta \in \mathcal{S}^p(\ell_n)} \left(\mathrm{E}[(X'\theta)^2] \right)^{1/2} \le Cn^{-c}
$$

by Assumption 2. Therefore, it follows from Lemma 9.3 that

$$
\mathrm{E} \left[\sup_{\theta \in \mathcal{S}^p(\ell_n)} \left| \frac{1}{n} \sum_{i=1}^n (X_i'\theta)^2 - \mathrm{E}[(X'\theta)^2] \right| \right] \le Cn^{-c}.
$$

The asserted claim follows from combining this bound and Markov's inequality. ∎

LEMMA 7.2. *Under Assumptions 1–5, there exists $\bar\lambda_0 = \bar\lambda_{n,0} \in \Lambda_n$, possibly depending on n, such that for all $k = 1, \ldots, K$, we have $\|\widehat\beta_{-k}(\bar\lambda_0)\|_0 \lesssim s$ and, in addition,*

$$(10) \qquad \|\widehat\beta_{-k}(\bar\lambda_0) - \beta\|_2^2 \lesssim \frac{s \log(pn)}{n} \quad and \quad \|\widehat\beta_{-k}(\bar\lambda_0) - \beta\|_1^2 \lesssim \frac{s^2 \log(pn)}{n}$$

with probability at least $1 - Cn^{-c}$.

REMARK 7.1. The result in this lemma is essentially well-known but we provide a short proof here for completeness.

PROOF. Let $T = \{j \in \{1, \ldots, p\}\colon \beta_j \neq 0\}$ and $T^c = \{1, \ldots, p\}\backslash T$. Fix $k = 1, \ldots, K$ and denote

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

and

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

To prove the asserted claims, we will apply Theorem 1 in [8] that shows that for any $\lambda \in \Lambda_n$, on the event $\lambda \geq 4\|Z_k\|_\infty$, we have

$$(11) \qquad \|\widehat\beta_{-k}(\lambda) - \beta\|_{2,n,-k} \leq \frac{3\lambda\sqrt{s}}{2\kappa_k}.$$

To use this bound, we show that there exist $c > 0$, $C > 0$, and $\bar\lambda_0 = \bar\lambda_{n,0} \in \Lambda_n$, possibly depending on n, such that

$$(12) \quad P(\kappa_k < c) \leq Cn^{-c}, \ P\left(\bar\lambda_0 < 4\|Z_k\|_\infty\right) \leq Cn^{-c}, \ \bar\lambda_0 \lesssim \left(\frac{\log(pn)}{n}\right)^{1/2}.$$

To prove the first claim in (12), note that

$$(13) \qquad\qquad 1 \lesssim \|\delta\|_{2,n,-k} \lesssim 1$$

with probability at least $1 - Cn^{-c}$ uniformly over all $\delta \in \mathbb{R}^p$ such that $\|\delta\|_2 = 1$ and $\|\delta_{T^c}\|_0 \leq s \log n$ by Lemma 7.1 and Assumptions 1, 2 and 5. Hence, the first claim in (12) follows from Lemma 10 in [8] applied with m there equal to $s \log n$ here.

To prove the second and the third claims in (12), note that we have $\max_{1 \leq j \leq p} \sum_{i \notin I_k} \mathrm{E}[|X_{ij}\varepsilon_i|^2] \lesssim n$ by Assumptions 1(b) and 3. Also,

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

Thus, by Lemma 9.1 and Assumption 2,

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

Hence, applying Lemma 9.2 with $t = (n \log n)^{1/2}$ and Z there replaced by $(n - n_k)\|Z_k\|_\infty$ here and noting that $nM_n^q/(n \log n)^{q/2} \leq Cn^{-c}$ by Assumption 2 implies that

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

with probability at least $1 - Cn^{-c}$. Hence, noting that $\log^4(pn) \leq Cn$ by Assumptions 1(a) and 2, it follows from Assumption 4 that there exists $\bar{\lambda}_0 \in \Lambda_n$ such that the second and the third claims in (12) hold. By (11), this $\bar{\lambda}_0$ satisfies the following bound:

$$(14) \qquad P\left(\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|^2_{2,n,-k} > \frac{Cs\log(pn)}{n}\right) \leq Cn^{-c}.$$

Now, to prove the asserted claims, note that using (12) and (13) and applying Theorem 2 in [8] with $m = s \log n$ there shows that $\|\widehat{\beta}_{-k}(\bar{\lambda}_0)\|_0 \lesssim s$ with probability at least $1 - Cn^{-c}$. Hence,

$$\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|^2_1 \lesssim s\|\widehat{\beta}_{-k}(\bar{\lambda}_0) - \beta\|^2_2$$

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

again with probability at least $1 - Cn^{-c}$, where the second inequality follows from (13), and the third one from (14). This gives all asserted claims and completes the proof of the lemma. ■

LEMMA 7.3. *Under Assumptions 1–5, we have for all $k = 1, \ldots, K$ that*

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

with probability $1 - Cn^{-c}$ for $\bar{\lambda}_0$ defined in Lemma 7.2.

REMARK 7.2. We thank one of the anonymous referees for suggesting the proof below. The suggestion relaxes the condition $s^2/n = o(1)$ in our early proof to $s/n = o(1)$, up to some log factors. ∎

PROOF. Fix $k = 1, \ldots, K$ and denote $\widehat{\beta} = \widehat{\beta}_{-k}(\bar{\lambda}_0)$. By Lemma 7.2, $\|\widehat{\beta}\|_0 \lesssim s$ with probability at least $1 - Cn^{-c}$. Hence,

$$\|\widehat{\beta} - \beta\|_{2,n,k}^2 \lesssim (\widehat{\beta} - \beta)' \mathrm{E}[XX'](\widehat{\beta} - \beta) + \|\widehat{\beta} - \beta\|_2^2 \lesssim \|\widehat{\beta} - \beta\|_2^2 \lesssim \frac{s \log(pn)}{n}$$

with probability at least $1 - Cn^{-c}$, where the first inequality follows from Lemma 7.1 and Assumption 5, the second from Assumption 1(b), and the third from Lemma 7.2. The asserted claim follows. ∎

LEMMA 7.4. *Under Assumptions 1–5, we have for all $k = 1, \ldots, K$ that*

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2 \lesssim \frac{s \log(pn)}{n} + \frac{\log^{r+1} n}{n}$$

with probability at least $1 - Cn^{-c}$.

PROOF. By the definition of $\widehat{\lambda}$ in (4),

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

for $\bar{\lambda}_0$ defined in Lemma 7.2. 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, for all $k = 1, \ldots, K$, denote $D_k = \{(X_i, Y_i)_{i \notin I_k}, (X_i)_{i \in I_k}\}$ and

$$Z_k = \max_{\lambda \in \Lambda_n} \left| \frac{\sum_{i \in I_k} \varepsilon_i X_i' (\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0))}{\sqrt{n_k} \|\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}} \right|.$$

Then by Lemma 9.1 and Assumptions 3 and 4, we have that $\mathrm{E}[Z_k \mid D_k] \lesssim$

$\sqrt{\log \log n} + M_{n,k} \log \log n$, where

$$
\begin{aligned}
M_{n,k} &= \left(\mathrm{E} \left[\max_{\lambda \in \Lambda_n} \max_{i \in I_k} \frac{(\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0)))^2}{n_k \|\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}^2} \mid D_k \right] \right)^{1/2} \\
&\leq \left(\mathrm{E} \left[\max_{\lambda \in \Lambda_n} \max_{i \in I_k} \frac{|\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0))|^u}{n_k^{u/2} \|\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}^u} \mid D_k \right] \right)^{1/u} \\
&\leq \left(\mathrm{E} \left[\sum_{\lambda \in \Lambda_n} \sum_{i \in I_k} \frac{|\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0))|^u}{n_k^{u/2} \|\widehat{\beta}_{-k}(\lambda) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}^u} \mid D_k \right] \right)^{1/u}
\end{aligned}
$$

for any $u \geq 2$. In turn, the last expression is bounded from above by

$$
C \left((u(1+r))^{u(1+r)/2} \log n \right)^{1/u}
$$

since (i) $\mathrm{E}[|\varepsilon_i|^u \mid D_k] \leq C^u \mathrm{E}[(1 + |e| + |e|^{r+1})^u] \leq C^u (u(1+r))^{u(1+r)/2}$ under Assumption 3 and (ii) for any sequence $(a_i)_{i \in I_k}$ in $\mathbb{R}$, we have $\sum_{i \in I_k} |a_i|^u \leq (\sum_{i \in I_k} |a_i|^2)^{u/2}$. Using this bound with $u = 3$ (for example) gives $\mathrm{E}[Z_k \mid D_k] \lesssim \sqrt{\log n}$. In addition, using this bound with $u = \log n$, it follows from Lemma 9.2 that

$$
\mathrm{P}\left(Z_k > 2\mathrm{E}[Z_k \mid D_k] + C\sqrt{\log^{r+1} n} \mid D_k \right) \leq Cn^{-c}.
$$

Hence, $Z_k \lesssim \sqrt{\log^{r+1} n}$ with probability at least $1 - Cn^{-c}$, and so, with the same probability,

$$
\left| \sum_{i \in I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)) \right| \lesssim \sqrt{n \log^{r+1} n} \|\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}.
$$

Therefore, since $n_k/n \geq c_1$ by Assumption 5, we have with the same probability that

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

and thus, by the triangle inequality,

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

where $\widehat{k}$ is a value of $k = 1, \ldots, K$ that maximizes $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}$. Therefore, by Lemma 7.3,

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

and thus, for all $k = 1, \ldots, K$,

$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \widehat{\beta}_{-k}(\bar{\lambda}_0)\|_{2,n,k}^2 \le \|\widehat{\beta}_{-\widehat{k}}(\widehat{\lambda}) - \widehat{\beta}_{-\widehat{k}}(\bar{\lambda}_0)\|_{2,n,\widehat{k}}^2 \lesssim \frac{s \log(pn)}{n} + \frac{\log^{r+1} n}{n},$$

again with probability at least $1 - Cn^{-c}$. The asserted claim now follows from combining this bound with the triangle inequality and Lemma 7.3. This completes the proof of the lemma. ∎

LEMMA 7.5.   *Under Assumptions 1 – 5, we have for all $k = 1, \ldots, K$ that*
$$\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_2^2 \lesssim \frac{s \log(pn)}{n} + \frac{\log^{r+1} n}{n}$$
with probability at least $1 - Cn^{-c}$.

PROOF. Fix $k = 1, \ldots, K$. For $\lambda \in \Lambda_n$, let $\delta_\lambda = (\widehat{\beta}_{-k}(\lambda) - \beta)/\|\widehat{\beta}_{-k}(\lambda) - \beta\|_2$ and observe that conditional on $(X_i, Y_i)_{i \notin I_k}$, $(\delta_\lambda)_{\lambda \in \Lambda_n}$ is non-stochastic. Hence, by Lemma 8.1, Assumptions 1(a) and 5 and Chebyshev's inequality applied conditional on $(X_i, Y_i)_{i \notin I_k}$, for any $\lambda \in \Lambda_n$, $(n_k)^{-1} \sum_{i \in I_k} (X_i' \delta_\lambda)^2 \ge c$ with probability at least $1 - Cn^{-c}$, and so $\|\widehat{\beta}_{-k}(\lambda) - \beta\|_2^2 \le C\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,k}^2$ with the same probability. Therefore, by Assumption 4 and the union bound, $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_2^2 \le C\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,k}^2$ with probability at least $1 - Cn^{-c}$. The asserted claim follows from combining this inequality and Lemma 7.4. ∎

LEMMA 7.6.   *Fix $k = 1, \ldots, K$ and denote*

$$(15) \quad \Lambda_{n,k}(X_1^n, T) = \left\{ \lambda \in \Lambda_n \colon \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n] \le T \right\}, \quad T > 0.$$

Then under Assumptions 1 – 5, we have that $\widehat{\lambda} \in \Lambda_{n,k}(X_1^n, T_n)$ with probability at least $1 - Cn^{-c}$, where

$$(16) \qquad T_n = C \left(\frac{s \log(pn)}{n} + \frac{\log^{r+1} n}{n} \right)^{1/2}.$$

PROOF. Fix $k = 1, \ldots, K$ and note that by Assumption 1(b) and Lemma 7.1, $\sup_{\delta \in \mathcal{S}^p(s)} \|\delta\|_{2,n} \leq C$ with probability at least $1 - Cn^{-c}$. Hence, by Lemma 7.5, Theorem 5.1, Assumption 4, and the union bound, $\|\widehat{\beta}_{-k}(\widehat{\lambda})\|_0 \leq n^{c_1/4} s \log^2(pn)$ with probability at least $1 - Cn^{-c}$. Further, by Assumption 2, $n^{c_1/4} s \log^2(pn) \leq \sqrt{sn^{1+c_1/2}} \log(pn)$ for all $n \geq n_0$ with n_0 depending only on c_1 and C_1, and so, by Assumption 1(b) and Lemma 7.1, $\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 \lesssim \|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_2^2$ with probability at least $1 - Cn^{-c}$. Combining this bound with Lemma 7.5 now gives

$$(17) \qquad \mathrm{P}\left(\|\widehat{\beta}_{-k}(\widehat{\lambda}) - \beta\|_{2,n,-k}^2 > C\left(\frac{s \log(pn)}{n} + \frac{\log^{r+1} n}{n} \right) \right) \leq Cn^{-c}.$$

Now,

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

where the first and the second terms on the right-hand side are at most Cn^{-c} by (17) and Lemma 8.2 applied with $\kappa = \log n$, respectively, as long the constant C in the definition of T_n is large enough. The asserted claim follows. ∎

LEMMA 7.7. *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 Section 2.4 in [44], where the proof is also provided. ∎

7.2. *Proof of Theorem 4.1.* Throughout the proof, we can assume that $\alpha \in [1/n, e^{-2}]$ since the results for $\alpha > e^{-2}$ and $\alpha < 1/n$ follow from the cases $\alpha = e^{-2}$ and $\alpha = 1/n$, respectively, with suitably increased constant C. We proceed in three steps. In the first step, for any given $\lambda > 0$, we use Lemma

7.7 to provide an upper bound on the conditional median of $n\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^2$ given X_1^n via some functionals of subsample estimators $\widehat{\beta}_{-k}(\lambda)$. In the second step, we derive bounds on these functionals for relevant values of λ with the help of Theorem 5.2. In the third step, we use Lemma 7.6 to show that $\widehat{\lambda}$ belongs to the relevant set with high probability and Lemma 8.2 to replace conditional medians by conditional expectations and complete the proof.

Step 1. For any random variable Z and any number α, let $Q_\alpha(Z \mid X_1^n)$ denote the αth quantile of the conditional distribution of Z given X_1^n. In this step, we show that for any $\lambda > 0$,

$$
Q_{1/2}(n\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^2 \mid X_1^n)
$$

$$
\lesssim \sum_{k=1}^{K} Q_{1-1/(16K)} \left(\sum_{i \notin I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 \mid X_1^n \right)
$$

$$
+ \sum_{k=1}^{K} Q_{1-1/(16K)} \left(\sum_{i \in I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 \mid X_1^n \right)
$$

$$
+ \sum_{k=1}^{K} Q_{1-1/(16K)} \left(\left| \sum_{i \notin I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta) \right| \mid X_1^n \right)
$$

$$
\text{(18)} \quad + \sum_{k=1}^{K} Q_{1-1/(16K)} \left(\left| \sum_{i \in I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta) \right| \mid X_1^n \right).
$$

To do so, fix any $\lambda > 0$ and denote

$$
\text{(19)} \quad b(\lambda) = \frac{1}{K-1} \sum_{k=1}^{K} \frac{n - n_k}{n} \widehat{\beta}_{-k}(\lambda).
$$

Then

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

$$
= \frac{1}{K-1} \sum_{k=1}^{K} \left(\sum_{i \notin I_k} (Y_i - X_i'\widehat{\beta}(\lambda))^2 + (n - n_k)\lambda\|\widehat{\beta}(\lambda)\|_1 \right)
$$

$$
\geq \frac{1}{K-1} \sum_{k=1}^{K} \left(\sum_{i \notin I_k} (Y_i - X_i'\widehat{\beta}_{-k}(\lambda))^2 + (n - n_k)\lambda\|\widehat{\beta}_{-k}(\lambda)\|_1 \right)
$$

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

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

$$\frac{1}{K-1} \sum_{k=1}^{K} \sum_{i \notin I_k} (Y_i - X_i' \widehat{\beta}_{-k}(\lambda))^2$$

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

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

Thus, by Lemma 7.7,

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

Substituting here $Y_i = X_i'\beta + \varepsilon_i$, $i = 1, \ldots, n$, and the definition of $b(\lambda)$ in (19) and using the triangle inequality gives

$$n\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^2 \lesssim n\|\widehat{\beta}(\lambda) - b(\lambda)\|_{2,n}^2 + n\|b(\lambda) - \beta\|_{2,n}^2$$

$$\lesssim \sum_{k=1}^{K} \sum_{i \notin I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 + \sum_{k=1}^{K} \sum_{i \in I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2$$

(20)
$$+ \sum_{k=1}^{K} \left| \sum_{i \notin I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta) \right| + \sum_{k=1}^{K} \left| \sum_{i \in I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta) \right|.$$

The claim of this step, inequality (18), follows from (20) and Lemma 9.6.

Step 2. Denote

(21) $$\Lambda_n(X_1^n, T) = \cap_{k=1}^{K} \Lambda_{n,k}(X_1^n, T), \quad T > 0,$$

for $\Lambda_{n,k}(X_1^n, T)$ defined in (15) of Lemma 7.6. In this step, we show that

$$(22) \quad \mathrm{P}\Big(\max_{\lambda \in \Lambda_n(X_1^n, T_n)} Q_{1/2}(n\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^2 \mid X_1^n)$$

$$> C(\log p)(s \log(pn) + \log^{r+1} n)\Big) \leq C n^{-c}$$

for T_n defined in (16) of Lemma 7.6.

To do so, we apply the result in Step 1 and bound all terms on the right-hand side of (18) in turn. To start, fix $k = 1, \ldots, K$. Then for any $\lambda \in \Lambda_n(X_1^n, T_n)$,

$$Q_{1-1/(16K)}\left(\sum_{i \notin I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 \mid X_1^n \right)$$

$$\leq n Q_{1-1/(16K)}\left(\|\widehat{\beta}_{-k}(\lambda) - \lambda\|_{2,n,-k} \mid X_1^n \right)^2$$

$$(23) \qquad \leq 16Kn\left(\mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n] \right)^2 \lesssim s\log(pn) + \log^{r+1} n,$$

where the third line follows from Markov's inequality and the definition of T_n.

Next, since Assumption 1(a) implies that $\mathrm{E}[(X'\delta)^2] \geq c$ for all $\delta \in \mathcal{S}^p(n + s)$, it follows from Lemma 7.1 and Assumptions 1 and 2 that

$$c \leq \inf_{\delta \in \mathcal{S}^p(n^{1/2+c_1/8}(\sqrt{n}T_n+1))} \|\delta\|_{2,n,-k} \quad \text{and} \quad \max_{1 \leq j \leq p} \sqrt{\frac{1}{n - n_k} \sum_{i \notin I_k} X_{ij}^2} \leq C$$

with probability at least $1 - Cn^{-c}$. Thus, by Theorem 5.2 and the union bound,

$$\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda)\|_0 \mid X_1^n] \lesssim s + (\log p)(nT_n^2 + \log^r n)$$

$$(24) \qquad\qquad\qquad\qquad \lesssim (\log p)(s\log(pn) + \log^{r+1} n)$$

with probability at least $1 - Cn^{-c}$. Thus, by Markov's inequality, Lemma 7.1, and Assumptions 1, 2, and 5,

$$\max_{\lambda \in \Lambda_n(X_1^n, T_n)} Q_{1-1/(16K)}\left(\sum_{i \in I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 \mid X_1^n \right)$$

$$\lesssim \max_{\lambda \in \Lambda_n(X_1^n, T_n)} Q_{1-1/(32K)}\left(n\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k}^2 \mid X_1^n \right)$$

$$(25) \qquad \lesssim s\log(pn) + \log^{r+1} n,$$

with probability at least $1 - Cn^{-c}$, where the last inequality follows from the same argument as that in (23).

Next, by Markov's inequality and the definition of M_n in (5), $\sum_{i \notin I_k} \|X_i\|_\infty^q \leq n^{1+c_1} M_n^q$ with probability at least $1 - Cn^{-c}$. Hence, by Lemma 9.1 and Assumptions 2 and 3,

$$\mathrm{E}\left[\left\| \sum_{i \notin I_k} \varepsilon_i X_i \right\|_\infty \mid X_1^n \right] \lesssim \sqrt{n\log p} + n^{(1+c_1)/q} M_n \log p \lesssim \sqrt{n\log p}$$

with probability at least $1 - Cn^{-c}$. Therefore, by Proposition A.1.6 in [45],

$$\mathrm{E}\left[\left\|\sum_{i \notin I_k} \varepsilon_i X_i\right\|_\infty^4 \mid X_1^n\right] \lesssim \left(\mathrm{E}\left[\left\|\sum_{i \notin I_k} \varepsilon_i X_i\right\|_\infty \mid X_1^n\right]\right)^4 + \mathrm{E}\left[\max_{i \notin I_k} \|\varepsilon_i X_i\|_\infty^4 \mid X_1^n\right]$$

$$\lesssim \left(\sqrt{n \log p} + n^{(1+c_1)/q} M_n\right)^4 \lesssim (n \log p)^2$$

with probability at least $1 - Cn^{-c}$. Thus, proceeding as in the proof of Theorem 5.2, getting from (43) to (48), with ψ_i's replaced by ε_i's, we obtain

$$\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \mathrm{E}\left[\left|\sum_{i \notin I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\right| \mid X_1^n\right]$$

$$\lesssim \sqrt{n \log p}\left(T_n + \sqrt{\frac{\log^r n}{n}}\right) \max_{\lambda \in \Lambda_n(X_1^n, T_n)} (\mathrm{E}[\|\widehat{\beta}_{-k}(\lambda)\|_0 + s \mid X_1^n])^{1/2}$$

$$+ \sqrt{n} T_n + \sqrt{\log^r n} + \max_{\lambda \in \Lambda_n(X_1^n, T_n)} \mathrm{E}[\|\widehat{\beta}_{-k}(\lambda)\|_0 + s \mid X_1^n]$$

$$\lesssim s + (\log p)(n T_n^2 + \log^r n) \lesssim (\log p)(s \log(pn) + \log^{r+1} n)$$

with probability at least $1 - Cn^{-c}$, where the second inequality follows from (24). Hence, by Markov's inequality,

$$\max_{\lambda \in \Lambda_n(X_1^n, T_n)} Q_{1-1/(16K)}\left(\left|\sum_{i \notin I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\right| \mid X_1^n\right)$$

$$\leq 16K \max_{\lambda \in \Lambda_n(X_1^n, T_n)} \mathrm{E}\left[\left|\sum_{i \notin I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\right| \mid X_1^n\right]$$

$$\lesssim (\log p)(s \log(pn) + \log^{r+1} n)$$

with probability at least $1 - Cn^{-c}$.

Finally, by Markov's inequality, for any $A_1, A_2, \lambda > 0$,

$$\mathrm{P}\left(\left|\sum_{i \in I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\right| > \sqrt{A_1 A_2 (s \log(pn) + \log^{r+1} n)} \mid X_1^n\right)$$

$$\leq \mathrm{P}\left(\sum_{i \in I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 > A_2 (s \log(pn) + \log^{r+1} n) \mid X_1^n\right)$$

$$+ \mathrm{E}\left[1\left\{\sum_{i \in I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 \leq A_2 (s \log(pn) + \log^{r+1} n)\right\}\right.$$

$$\left. \times \frac{\mathrm{E}[|\sum_{i \in I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)|^2 \mid X_1^n, (Y_i)_{i \notin I_k}]}{A_1 A_2 (s \log(pn) + \log^{r+1} n)} \mid X_1^n\right]$$

$$\lesssim \mathrm{P}\left(\sum_{i \in I_k} (X_i'(\widehat{\beta}_{-k}(\lambda) - \beta))^2 > A_2 (s \log(pn) + \log^{r+1} n) \mid X_1^n\right) + 1/A_1.$$

Choosing both A_1 and A_2 here large enough and using the same argument as that in (25) shows that

$$\max_{\lambda \in \Lambda_n(X_1^n, T_n)} Q_{1-1/(16K)}\left(\Big| \sum_{i \in I_k} \varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\Big| \mid X_1^n\right)$$

$$\lesssim \sqrt{s \log(pn) + \log^{r+1} n} \lesssim s \log(pn) + \log^{r+1} n$$

with probability at least $1 - Cn^{-c}$. Combining all inequalities presented above together and using Step 1 gives (22), which is the asserted claim of this step.

Step 3. Here we complete the proof. To do so, note that by Lemma 8.2 applied with $\kappa = 2$, for any $\lambda > 0$,

$$\left| \|\widehat{\beta}(\lambda) - \beta\|_{2,n} - \mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n] \right| \lesssim \sqrt{\frac{\log^r n}{n}}$$

with probability at least $3/4$, which implies that

$$\left| Q_{1/2}(\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n) - \mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n] \right| \lesssim \sqrt{\frac{\log^r n}{n}}.$$

Combining this inequality with (22) in Step 2 shows that

$$(26) \quad \mathrm{P}\Bigg(\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n]$$

$$> \sqrt{\frac{Cs \log p}{n}} \times \sqrt{\log(pn) + s^{-1} \log^{r+1} n} \Bigg) \leq Cn^{-c}.$$

Also, applying Lemma 8.2 with $\kappa = \log(1/\alpha) \leq \log n$ and

$$t = \left(\frac{\widetilde{C} \log(1/\alpha) \log^{r+1} n}{n} \right)^{1/2}$$

with sufficiently large $\widetilde{C}$, which can be chosen to depend only on C_1 and r, it follows that for any $\lambda > 0$,

$$\mathrm{P}\left(\left| \|\widehat{\beta}(\lambda) - \beta\|_{2,n} - \mathrm{E}[\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \mid X_1^n] \right| > t \right) \leq \left(\frac{C}{\widetilde{C} \log n} \right)^{\frac{\log(1/\alpha)}{2}} \leq \frac{\alpha}{|\Lambda_n|}$$

since $\alpha \leq e^{-2}$. Combining these inequalities and using the union bound, we obtain

$$(27) \quad \mathrm{P}\left(\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \|\widehat{\beta}(\lambda) - \beta\|_{2,n} \right.$$
$$\left. > \sqrt{\frac{Cs \log(p/\alpha)}{n}} \times \sqrt{\log(pn) + s^{-1} \log^{r+1} n}\right) \leq \alpha + Cn^{-c}.$$

Finally, by Lemma 7.6 and the union bound,

$$(28) \qquad\qquad \mathrm{P}(\widehat{\lambda} \in \Lambda_n(X_1^n, T_n)) \geq 1 - Cn^{-c}.$$

Combining the last two inequalities gives the asserted claim and completes the proof of the theorem. ∎

7.3. *Proof of Theorem 4.2.* Define $\Lambda_n(X_1^n, T_n)$ as in Step 2 of the proof of Theorem 4.1. Then by Assumptions 1 and 2, Lemma 7.1, Theorem 5.2, and (26) in the proof of Theorem 4.1,

$$\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n] \lesssim s(\log^2 p)(\log(pn) + s^{-1} \log^{r+1} n)$$

with probability at least $1 - Cn^{-c}$. Thus, by Markov's inequality, the union bound, and Assumption 4, for any $\bar{s} > 0$,

$$\mathrm{P}\left(\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \|\widehat{\beta}(\lambda)\|_0 > \bar{s} \mid X_1^n\right)$$
$$\lesssim s(\log^2 p)(\log n)(\log(pn) + s^{-1} \log^{r+1} n)/\bar{s}$$

with probability at least $1 - Cn^{-c}$. The asserted claim of the theorem follows from combining this bound with (28) in the proof of Theorem 4.1 and substituting

$$\bar{s} = Cs(\log^2 p)(\log n)(\log(pn) + s^{-1} \log^{r+1} n)/\alpha$$

with a sufficiently large constant $C > 0$. This completes the proof of the theorem. ∎

7.4. *Proof of Theorem 4.3.* Applying Theorem 4.2 with $\alpha = n^{-c_1/4}$ shows that

$$\|\widehat{\beta}(\widehat{\lambda})\|_0 \lesssim sn^{c_1/4}(\log^2 p)(\log n)(\log(pn) + s^{-1} \log^{r+1} n)$$

with probability at least $1 - Cn^{-c}$. Thus, by Lemma 7.1 and Assumptions 1(a) and 2, $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2 \lesssim \|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{2,n}$ with probability at least $1 - Cn^{-c}$. The asserted claim regarding $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2$ follows from this bound and Theorem 4.1.

Also, by the Cauchy-Schwarz and triangle inequalities,

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1 \leq \sqrt{\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_0}\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2 \leq \sqrt{\|\widehat{\beta}(\widehat{\lambda})\|_0 + s}\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2.$$

The asserted claim regarding $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_1$ follows from this bound, Theorem 4.2, and the asserted claim regarding $\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2$. This completes the proof of the theorem. ∎

7.5. *Proof of Corollary 4.1.* By Assumptions 1(a) and 2 and Theorem 4.2,
$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{p,2,n} \leq C\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_2$$
with probability at least $1 - Cn^{-c}$. Hence, by Theorem 4.3,

$$\|\widehat{\beta}(\widehat{\lambda}) - \beta\|_{p,2,n} \leq \sqrt{\frac{Cs\log(p/\alpha)}{n}} \times \sqrt{\log(pn) + s^{-1}\log^{r+1} n},$$

with probability at least $1 - \alpha - Cn^{-c}$. The asserted claim follows. ∎

8. Proofs for Section 5. In this section, we prove Theorems 5.1 and 5.2. Since the proofs are long, we start with a sequence of preliminary lemmas.

8.1. *Preliminary Lemmas.*

LEMMA 8.1. *For all $\lambda > 0$, 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 property: 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 any fixed value of $X_1^n = (X_1, \ldots, X_n)$ is well-defined and is Lipschitz-continuous with Lipschitz constant one with respect to Euclidean norm. Moreover, there always exists a Lasso estimator $\widehat{\beta}(\lambda)$ such that $\|\widehat{\beta}(\lambda)\|_0 \leq n$ almost surely. Finally, $\widehat{\beta}(\lambda)$ is unique almost surely whenever the distribution of X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$.*

PROOF. All the asserted claims in this lemma can be found in the literature. Here we give specific references for completeness. The fact that the function $(\varepsilon_i)_{i=1}^n \mapsto (X_i'\widehat{\beta}(\lambda))_{i=1}^n$ is well-defined follows from Lemma 1 in [43], which shows that even if the solution $\widehat{\beta}(\lambda)$ of the optimization problem (2)

is not unique, $(X_i'\widehat{\beta}(\lambda))_{i=1}^n$ is the same across all solutions. The Lipschitz property then follows from Proposition 2 in [5]. Moreover, by discussion in Section 2.1 in [43], there always exists a Lasso solution, say $\widehat{\beta}(\lambda)$, taking the form in (10) of [43], and such a solution satisfies $\|\widehat{\beta}(\lambda)\|_0 \leq n$. Finally, the last claim follows from Lemma 4 in [43]. ∎

LEMMA 8.2. *Suppose that Assumption 3 holds. Then for all $\kappa \geq 1$, $n \geq e^\kappa$, and $\lambda > 0$, we have*

$$(29) \quad P\Big(\Big|\|\widehat{\beta}(\lambda) - \beta\|_{2,n} - E[\|\widehat{\beta} - \beta\|_{2,n} \mid X_1^n]\Big| > t\Big) \leq \Big(\frac{C\kappa \log^r n}{t^2 n}\Big)^{\kappa/2}$$

for some constant $C > 0$ depending only on C_1 and r.

PROOF. Fix $\kappa \geq 1$, $n \geq e^\kappa$, and $\lambda > 0$. Also, let ξ be a $N(0,1)$ random variable that is independent of the data and let C be a positive constant that depends only on C_1 and r but whose value can change from place to place. Then by Lemma 8.1, 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

$$(\varepsilon_i)_{i=1}^n \mapsto \Big(\sum_{i=1}^n (X_i'(\widehat{\beta}(\lambda) - \beta))^2\Big)^{1/2} = \sqrt{n}\|\widehat{\beta} - \beta\|_{2,n}.$$

Therefore, applying Lemma 9.5 with $u(x) = (x \vee 0)^\kappa$ and using Markov's inequality and Assumption 3 shows that for any $t > 0$,

$$P\Big(\|\widehat{\beta}(\lambda) - \beta\|_{2,n} - E[\|\widehat{\beta} - \beta\|_{2,n} \mid X_1^n] > t \mid X_1^n\Big)$$
$$\leq \Big(\frac{C_1\pi}{2t\sqrt{n}}\Big)^\kappa E\Big[\max_{1 \leq i \leq n}(1 + |e_i|^r)^\kappa\Big]E[|\xi|^\kappa] \leq \Big(\frac{C}{t\sqrt{n}}\Big)^\kappa E\Big[\max_{1 \leq i \leq n}|e_i|^{r\kappa}\Big]E[|\xi|^\kappa]$$
$$\leq \Big(\frac{C}{t\sqrt{n}}\Big)^\kappa \Big(E\Big[\max_{1 \leq i \leq n}|e_i|^{r\log n}\Big]\Big)^{\kappa/\log n}E[|\xi|^\kappa]$$
$$\leq \Big(\frac{C(r\log n)^{r/2}\sqrt{\kappa}}{t\sqrt{n}}\Big)^\kappa = \Big(\frac{C\sqrt{\kappa\log^r n}}{t\sqrt{n}}\Big)^\kappa = \Big(\frac{C\kappa\log^r n}{t^2 n}\Big)^{\kappa/2}.$$

This gives one side of the bound (29). Since the other side follows similarly, the proof is complete. ∎

LEMMA 8.3. *Suppose that Assumption 3 holds and let $Q^{-1}\colon \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ be the inverse of $Q\colon \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ with respect to the second argument. Then for all $\lambda > 0$,*

$$(30) \quad E[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n] = \sum_{i=1}^n E[\psi_i X_i'(\widehat{\beta}(\lambda) - \beta) \mid X_1^n],$$

where

$$\psi_i = \frac{e_i}{Q_2(X_i, e_i)} + \frac{Q_{22}(X_i, e_i)}{Q_2(X_i, e_i)^2} \ \ and \ e_i = Q^{-1}(X_i, \varepsilon_i)$$

for all $i = 1, \ldots, n$. In addition,

$$\mathrm{E}\left[\left(\|\widehat{\beta}(\lambda)\|_0 - \sum_{i=1}^{n} \psi_i X_i'(\widehat{\beta} - \beta) \right)^2 \mid X_1^n \right]$$

$$= \sum_{i=1}^{n} \mathrm{E}\left[\gamma_i (X_i'(\widehat{\beta}(\lambda) - \beta))^2 \mid X_1^n \right] + \mathrm{E}[\|\widehat{\beta}(\lambda)\|_0 \mid X_1^n],$$

where

$$\gamma_i = \frac{1}{Q_2(X_i, e_i)^2} - \frac{e_i Q_{22}(X_i, e_i)}{Q_2(X_i, e_i)^3} + \frac{Q_{222}(X_i, e_i)}{Q_2(X_i, e_i)^3} - \frac{2Q_{22}(X_i, e_i)^2}{Q_2(X_i, e_i)^4}$$

for all $i = 1, \ldots, n$. Moreover,

$$\mathrm{Var}\left(\sum_{i=1}^{n} \psi_i X_i'(\widehat{\beta}(\lambda) - \beta) \mid X_1^n \right) \leq 2 \sum_{i=1}^{n} \mathrm{E}\left[\left(\gamma_i Q_2(X_i, e_i) X_i'(\widehat{\beta} - \beta) \right)^2 \mid X_1^n \right]$$

$$+ C\mathrm{E}\left[\|\widehat{\beta}(\lambda)\|_0 + 1 \mid X_1^n \right] (\log p)(\log^r n),$$

where $C > 0$ is a constant depending only on c_1, C_1, and r.

REMARK 8.1. Here, the inverse Q^{-1} exists because by Assumption 3, Q is strictly increasing and continuous with respect to its second argument. ∎

PROOF. This lemma extends some of the results in [42] and [6] to the non-Gaussian case. All arguments in the proof are conditional on $X_1, \ldots, X_n$ but we drop the conditioning sign for brevity of notation. Also, we use C to denote a positive constant that depends only on c_1, C_1 and r but whose value can change from place to place.

Fix $\lambda > 0$ and denote $\widehat{\beta} = \widehat{\beta}(\lambda)$ and $\widehat{T} = \{j \in \{1, \ldots, p\} \colon \widehat{\beta}_j \neq 0\}$. For all $i = 1, \ldots, n$ we will use $X_{i\widehat{T}}$ to denote the sub-vector of X_i in $\mathbb{R}^{|\widehat{T}|}$ corresponding to indices in $\widehat{T}$. By results in [42], we then have

$$(31) \qquad \frac{\partial X_i'(\widehat{\beta} - \beta)}{\partial \varepsilon_j} = X_{i\widehat{T}}' \left(\sum_{l=1}^{n} X_{l\widehat{T}} X_{l\widehat{T}}' \right)^{-1} X_{j\widehat{T}}, \quad i, j = 1, \ldots, n;$$

see, in particular, the proof of Theorem 1 there. Taking the sum over $i = j = 1, \ldots, n$ and applying the trace operator on the right-hand side of this identity gives

$$(32) \qquad \|\widehat{\beta}\|_0 = |\widehat{T}| = \sum_{i=1}^{n} \frac{\partial(X_i'(\widehat{\beta} - \beta))}{\partial \varepsilon_i}.$$

Also, for all $i = 1, \ldots, n$, under Assumption 3 (and conditional on X_1^n), the random variable ε_i is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$ with continuously differentiable pdf χ_i defined by

$$\chi_i(Q(X_i, e)) = \frac{\phi(e)}{Q_2(X_i, e)}, \quad e \in \mathbb{R},$$

where ϕ is the pdf of the $N(0,1)$ distribution. Taking the derivative over e here gives

$$\chi_i'(Q(X_i, e))Q_2(X_i, e) = -\frac{e\phi(e)}{Q_2(X_i, e)} - \frac{\phi(e)Q_{22}(X_i, e)}{Q_2(X_i, e)^2}, \quad e \in \mathbb{R},$$

and so

$$\frac{\chi_i'(\varepsilon_i)}{\chi_i(\varepsilon_i)} = \frac{\chi_i'(Q(X_i, e_i))}{\chi_i(Q(X_i, e_i))} = -\frac{e_i}{Q_2(X_i, e_i)} - \frac{Q_{22}(X_i, e_i)}{Q_2(X_i, e_i)^2} = -\psi_i.$$

Therefore, by Lemma 9.4, whose application is justified by Assumption 3 and Lemma 8.1,

$$(33) \qquad \mathrm{E}\left[\frac{\partial(X_i'(\widehat{\beta} - \beta))}{\partial \varepsilon_i}\right] = \mathrm{E}[\psi_i X_i'(\widehat{\beta} - \beta)], \quad i = 1, \ldots, n.$$

Combining (32) and (33) gives the first asserted claim.

To prove the second asserted claim, we proceed along the lines in the proof of Theorem 1.1 in [6]. Specifically, let $f_1, \ldots, f_n$ be twice continuously differentiable functions mapping $\mathbb{R}^n$ to $\mathbb{R}$ with bounded first and second derivatives. Also, let $\bar{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. Then, it follows from Lemma 9.4 that

$$\mathrm{E}\left[\psi_i f_i(\bar{\varepsilon})\left(\sum_{j=1}^{n} \psi_j f_j(\bar{\varepsilon}) - \sum_{l=1}^{n} \frac{\partial f_l(\bar{\varepsilon})}{\partial \varepsilon_l}\right)\right]$$

$$= \mathrm{E}\left[\frac{\partial f_i(\bar{\varepsilon})}{\partial \varepsilon_i}\left(\sum_{j=1}^{n} \psi_j f_j(\bar{\varepsilon}) - \sum_{l=1}^{n} \frac{\partial f_l(\bar{\varepsilon})}{\partial \varepsilon_l}\right)\right]$$

$$+ \mathrm{E}\left[f_i(\bar{\varepsilon})\left(\gamma_i f_i(\bar{\varepsilon}) + \sum_{j=1}^{n} \psi_j \frac{\partial f_j(\bar{\varepsilon})}{\partial \varepsilon_i} - \sum_{l=1}^{n} \frac{\partial^2 f_l(\bar{\varepsilon})}{\partial \varepsilon_l \partial \varepsilon_i}\right)\right]$$

for all $i = 1, \ldots, n$ and, in addition,

$$\mathrm{E}\left[\psi_j f_i(\bar{\varepsilon})\frac{\partial f_j(\bar{\varepsilon})}{\partial \varepsilon_i}\right] = \mathrm{E}\left[\frac{\partial f_i(\bar{\varepsilon})}{\partial \varepsilon_j}\frac{\partial f_j(\bar{\varepsilon})}{\partial \varepsilon_i} + f_i(\bar{\varepsilon})\frac{\partial^2 f_j(\bar{\varepsilon})}{\partial \varepsilon_i \partial \varepsilon_j}\right]$$

for all $j = 1, \ldots, n$. Combining these results, rearranging the terms, and taking the sum over $i = 1, \ldots, n$, we obtain

$$\mathrm{E}\left[\left(\sum_{i=1}^{n}\psi_i f_i(\bar{\varepsilon}) - \sum_{i=1}^{n}\frac{\partial f_i(\bar{\varepsilon})}{\partial \varepsilon_i}\right)^2\right] = \sum_{i=1}^{n}\mathrm{E}[\gamma_i f_i(\bar{\varepsilon})^2] + \sum_{i,j=1}^{n}\mathrm{E}\left[\frac{\partial f_i(\bar{\varepsilon})}{\partial \varepsilon_j}\frac{\partial f_j(\bar{\varepsilon})}{\partial \varepsilon_i}\right],$$

and since all second-order derivatives cancell out, it follows from a convolution argument that the same identity holds for any Lipschitz functions $f_1, \ldots, f_n$; see Appendix A of [6] for details. We now substitute $f_i(\bar{\varepsilon}) = X_i'(\widehat{\beta} - \beta)$ for all $i = 1, \ldots, n$ in this identity and note that

$$\sum_{i,j=1}^{n}\frac{\partial f_i(\bar{\varepsilon})}{\partial \varepsilon_j}\frac{\partial f_j(\bar{\varepsilon})}{\partial \varepsilon_i} = \|\widehat{\beta}\|_0$$

by (31) in this case. This gives the second asserted claim.

To prove the third asserted claim, we have by the Gaussian Poincare inequality, Theorem 3.20 in [14] that

$$\mathrm{Var}\left(\sum_{i=1}^{n}\psi_i X_i'(\widehat{\beta} - \beta)\right) \leq \sum_{j=1}^{n}\mathrm{E}\left[\left(\frac{\partial}{\partial e_j}\sum_{i=1}^{n}\psi_i X_i'(\widehat{\beta} - \beta)\right)^2\right]$$

$$\leq 2\sum_{j=1}^{n}\mathrm{E}\left[\left(\frac{\partial \psi_j}{\partial e_j}X_j'(\widehat{\beta} - \beta)\right)^2\right] + 2\sum_{j=1}^{n}\mathrm{E}\left[\left(\sum_{i=1}^{n}\psi_i\frac{\partial X_i'(\widehat{\beta} - \beta)}{\partial e_j}\right)^2\right].$$

Here, the first term on the right-hand side is equal to

$$2\sum_{j=1}^{n}\mathrm{E}\left[\left(\gamma_j Q_2(X_j, e_j)X_j'(\widehat{\beta} - \beta)\right)^2\right].$$

Also, by (31), the second term is equal to

$$
2\sum_{j=1}^{n} \mathrm{E}\left[\left(\sum_{i=1}^{n}\psi_i Q_2(X_j,e_j)X'_{i\widehat{T}}\left(\sum_{l=1}^{n}X_{l\widehat{T}}X'_{l\widehat{T}}\right)^{-1}X_{j\widehat{T}}\right)^2\right]
$$

$$
= 2\mathrm{E}\left[\max_{1\le j\le p}Q_2(X_j,e_j)^2\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\psi_i X'_{i\widehat{T}}\left(\sum_{l=1}^{n}X_{l\widehat{T}}X'_{l\widehat{T}}\right)^{-1}X_{j\widehat{T}}\right)^2\right]
$$

$$
= 2\mathrm{E}\left[\max_{1\le j\le p}Q_2(X_j,e_j)^2\sum_{i=1}^{n}\psi_i X'_{i\widehat{T}}\left(\sum_{l=1}^{n}X_{l\widehat{T}}X'_{l\widehat{T}}\right)^{-1}\sum_{j=1}^{n}X_{j\widehat{T}}\psi_j\right].
$$

Next, observe that

$$
\sum_{i=1}^{n}\psi_i X'_{i\widehat{T}}\left(\sum_{l=1}^{n}X_{l\widehat{T}}X'_{l\widehat{T}}\right)^{-1}\sum_{j=1}^{n}X_{j\widehat{T}}\psi_j
$$

is equal to $\|P_{\widehat{T}}\bar{\psi}\|_2^2$, where $\bar{\psi}=(\psi_1,\ldots,\psi_n)'$ and $P_{\widehat{T}}$ is the matrix projecting on $(X_{1\widehat{T}},\ldots,X_{n\widehat{T}})'$. In turn, we can bound $\mathrm{E}[\|P_{\widehat{T}}\bar{\psi}\|_2^2]$ using arguments from the proof of Theorem 4.3 in [6]. In particular, for any $M\subset\{1,\ldots,p\}$, letting P_M denote the matrix projecting on $(X_{1M},\ldots,X_{nM})'$, we have $\mathrm{E}[\|P_M\bar{\psi}\|_2^2]\le C|M|$, and so, by Assumption 3 and the Hanson-Wright inequality, Theorem 1.1 in [36],

$$
P\left(\|P_M\bar{\psi}\|_2^2 > C(|M|+x)\right)\le e^{-x}
$$

for all $x>0$. Thus, applying the union bound twice,

$$
\mathrm{P}\left(\max_{M\subset\{1,\ldots,p\}}\left(\|P_M\bar{\psi}\|_2^2 - C\left(|M|+\log\binom{p}{|M|}+\log p+x\right)\right)>0\right)\le e^{-x},
$$

and so

$$
\mathrm{P}\left(\max_{M\subset\{1,\ldots,p\}}\left(\|P_M\bar{\psi}\|_2^2 - C\left((|M|+1)\log p+x\right)\right)>0\right)\le e^{-x}.
$$

By Fubini's theorem and simple calculations, we then have

$$
(34)\qquad \mathrm{E}\left[\|P_{\widehat{T}}\bar{\psi}\|_2^2\right]\le C\mathrm{E}\left[\|\widehat{\beta}\|_0+1\right]\log p.
$$

Also,

$$
(35)\qquad \mathrm{E}\left[\|P_{\widehat{T}}\bar{\psi}\|_2^4\right]\le \mathrm{E}\left[\|\bar{\psi}\|_2^4\right]\le Cn^2.
$$

Hence, for a sufficiently large constant A that can be chosen to depend on c_1, C_1 and r only,

$$\mathrm{E}\left[\max_{1\le j\le p} Q_2(X_j,e_j)^2 \sum_{i=1}^n \psi_i X'_{i\widehat{T}} \left(\sum_{l=1}^n X_{l\widehat{T}} X'_{l\widehat{T}}\right)^{-1} \sum_{j=1}^n X_{j\widehat{T}}\psi_j\right]$$

$$= \mathrm{E}\left[\max_{1\le j\le p} Q_2(X_j,e_j)^2 \|P_{\widehat{T}}\bar{\psi}\|_2^2 1\left\{\max_{1\le j\le p} Q_2(X_j,e_j)^2 \le A\log^r n\right\}\right]$$

$$+ \mathrm{E}\left[\max_{1\le j\le p} Q_2(X_j,e_j)^2 \|P_{\widehat{T}}\bar{\psi}\|_2^2 1\left\{\max_{1\le j\le p} Q_2(X_j,e_j)^2 > A\log^r n\right\}\right].$$

Here, by (34), the first term on the right-hand side is bounded from above by $C\mathrm{E}[\|\widehat{\beta}\|_0+1](\log p)(\log^r n)$ and by (35), Assumption 3, and Hölder's inequality, the second term is bounded from above by C since A is large enough. Combining all presented inequalities together gives the third asserted claim and completes the proof of the lemma. ∎

8.2. *Proof of Theorem 5.1.* All arguments in this proof are conditional on $X_1,\ldots,X_n$ but we drop the conditioning sign for brevity of notation. Throughout the proof, we will assume that

$$(36) \qquad\qquad \sup_{\delta\in\mathcal{S}^p(s)} \|\delta\|_{2,n} \le \bar{C}.$$

Also, we use C to denote a positive constant that depends only on c_1, C_1, $\bar{C}$, and r but whose value can change from place to place.

Fix $\lambda > 0$ and denote $\widehat{\beta} = \widehat{\beta}(\lambda)$, $\widehat{s} = \|\widehat{\beta}\|_0$, and $R_n = \mathrm{E}[\|\widehat{\beta} - \beta\|_{2,n}]$. We start with some preliminary inequalities. First, by Hölder's inequality and Assumption 3,

$$\sum_{i=1}^n \mathrm{E}\left[\gamma_i (X'_i(\widehat{\beta} - \beta))^2\right]$$

$$(37) \qquad \le n\mathrm{E}\left[\|\widehat{\beta} - \beta\|_{2,n}^2 \times \max_{1\le i\le n} |\gamma_i|\right] \le C(n\log n)\sqrt{\mathrm{E}[\|\widehat{\beta} - \beta\|_{2,n}^4]}$$

and, similarly,

$$(38) \quad \sum_{i=1}^n \mathrm{E}\left[\left(\gamma_i Q_2(X_i,e_i) X'_i(\widehat{\beta} - \beta)\right)^2\right] \le C(n\log^2 n)\sqrt{\mathrm{E}[\|\widehat{\beta} - \beta\|_{2,n}^4]}.$$

Second, by the triangle inequality and Fubini's theorem,

$$
\mathrm{E}[\|\widehat{\beta} - \beta\|_{2,n}^4] \leq C\left(R_n^4 + \mathrm{E}\left[(\|\widehat{\beta} - \beta\|_{2,n} - R_n)^4\right]\right)
$$

$$
= C\left(R_n^4 + \int_0^\infty \mathrm{P}\left(|\|\widehat{\beta} - \beta\|_{2,n} - R_n| > t^{1/4}\right) dt\right)
$$

$$
(39) \qquad \leq C\left(R_n^4 + \left(\frac{\log^r n}{n}\right)^2\right),
$$

where the last line follows from Lemma 8.2 applied with $\kappa = 5$ (for example). Third,

$$
(40) \qquad \mathrm{P}\left(R_n > \|\widehat{\beta} - \beta\|_{2,n} + \sqrt{\frac{C\log^{r+1} n}{n}}\right) \leq \frac{1}{n}
$$

by Lemma 8.2 applied with $\kappa = \log n$. Fourth, by Lemma 9 of [8] and (36),

$$
\|\widehat{\beta} - \beta\|_{2,n}^2 \leq \|\widehat{\beta} - \beta\|_2^2 \times \sup_{\delta \in \mathcal{S}^p(\widehat{s}+s)} \|\delta\|_{2,n}^2
$$

$$
(41) \qquad \leq \|\widehat{\beta} - \beta\|_2^2 \times \frac{2(\widehat{s}+s)}{s} \sup_{\delta \in \mathcal{S}^p(s)} \|\delta\|_{2,n}^2 \leq \frac{C(\widehat{s}+s)}{s}\|\widehat{\beta} - \beta\|_2^2.
$$

We now prove the theorem with the help of these bounds. Denote

$$
V_1 = \mathrm{Var}\left(\widehat{s} - \sum_{i=1}^n \psi_i X_i'(\widehat{\beta} - \beta)\right) \text{ and } V_2 = \mathrm{Var}\left(\sum_{i=1}^n \psi_i X_i'(\widehat{\beta} - \beta)\right).
$$

Then for any $\bar{t} > 0$, with probability at least $1 - 2/\bar{t}^2$, by Chebyshev's inequality and Lemma 8.3,

$$
\widehat{s} \leq \sum_{i=1}^n \psi_i X_i'(\widehat{\beta} - \beta) + \bar{t}\sqrt{V_1}
$$

$$
= \sum_{i=1}^n \psi_i X_i'(\widehat{\beta} - \beta) + \mathrm{E}\left[1 + \sum_{i=1}^n \psi_i X_i'(\widehat{\beta} - \beta)\right] - \mathrm{E}[1 + \widehat{s}] + \bar{t}\sqrt{V_1}
$$

$$
(42) \qquad \leq 1 + 2\sum_{i=1}^n \psi_i X_i'(\widehat{\beta} - \beta) - \mathrm{E}[1 + \widehat{s}] + \bar{t}(\sqrt{V_1} + \sqrt{V_2}).
$$

Here, $\bar{t}(\sqrt{V_1} + \sqrt{V_2})$ is bounded from above by

$$C\bar{t}\Big(\mathrm{E}[1 + \widehat{s}](\log p)(\log^r n)\Big)^{1/2} + C\bar{t}\sqrt{n}\log n\left(R_n + \sqrt{\frac{\log^r n}{n}}\right)$$

$$\leq \mathrm{E}[1 + \widehat{s}] + C\bar{t}^2(\log p)(\log^r n) + C\bar{t}\sqrt{n}\log n\left(R_n + \sqrt{\frac{\log^r n}{n}}\right)$$

by Lemma 8.3 and inequalities (37), (38), and (39). Also, with probability at least $1 - 1/n$,

$$R_n \leq C\left(\sqrt{\frac{\widehat{s} + s}{s}}\|\widehat{\beta} - \beta\|_2 + \sqrt{\frac{\log^{r+1} n}{n}}\right)$$

by (40) and (41). In addition,

$$\sum_{i=1}^{n} \psi_i X_i'(\widehat{\beta} - \beta) \leq \left\|\sum_{i=1}^{n} \psi_i X_i\right\|_{\infty} \times \|\widehat{\beta} - \beta\|_1,$$

where $\|\widehat{\beta} - \beta\|_1 \leq (\widehat{s} + s)^{1/2} \times \|\widehat{\beta} - \beta\|_2$ and with probability at least $1 - 1/n$,

$$\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_{\infty} \leq C\sqrt{\log(pn)} \times \max_{1 \leq j \leq p}\left(\sum_{i=1}^{n} X_{ij}^2\right)^{1/2} \leq C\sqrt{n\log(pn)},$$

with the first inequality following from Assumption 3 and the union bound and the second from (36). Substituting all these bounds into (42) and using $\bar{t} = (ts\log(pn))^{1/2}$ with $t \geq 1$ gives

$$\widehat{s} \leq Cts(\log^r n)\log^2(pn) + C\sqrt{t(\widehat{s} + s)n(\log^2 n)\log(pn)}\|\widehat{\beta} - \beta\|_2$$

with probability at least $1 - 2/(ts\log(pn)) - 2/n$. Solving this inequality for $\widehat{s}$ gives the asserted claim and completes the proof of the theorem. ∎

8.3. *Proof of Theorem 5.2.* All arguments in this proof are conditional on $X_1, \ldots, X_n$ but we drop the conditioning sign for brevity of notation. Throughout the proof, we will assume that (7) holds. Also, we use C to denote a positive constant that depends only on c_1, C_1, $\bar{c}$, $\bar{C}$, and r but whose value can change from place to place.

Fix $\lambda > 0$ and denote $\widehat{\beta} = \widehat{\beta}(\lambda)$, $\widehat{s} = \|\widehat{\beta}\|_0$, $J_n = J_n(\lambda)$, and $R_n = R_n(\lambda)$. Then by Lemma 8.3,

$$(43) \qquad \mathrm{E}[\widehat{s}] = \sum_{i=1}^{n} \mathrm{E}[\psi_i X_i'(\widehat{\beta} - \beta)] = \mathcal{I}_1 + \mathcal{I}_2,$$

where

$$
(44) \qquad \mathcal{I}_1 = \sum_{i=1}^{n} \mathrm{E}\left[\psi_i X_i'(\widehat{\beta} - \beta) 1\left\{\bar{c}\|\widehat{\beta} - \beta\|_2 \le \|\widehat{\beta} - \beta\|_{2,n}\right\}\right],
$$

$$
(45) \qquad \mathcal{I}_2 = \sum_{i=1}^{n} \mathrm{E}\left[\psi_i X_i'(\widehat{\beta} - \beta) 1\left\{\bar{c}\|\widehat{\beta} - \beta\|_2 > \|\widehat{\beta} - \beta\|_{2,n}\right\}\right].
$$

We bound $\mathcal{I}_1$ and $\mathcal{I}_2$ in turn. To bound $\mathcal{I}_1$, note that as in (39) of the proof of Theorem 5.1,

$$
(46) \qquad \mathrm{E}[\|\widehat{\beta} - \beta\|_{2,n}^4] \le C\left(R_n^4 + \left(\frac{\log^r n}{n}\right)^2\right).
$$

Also, by Assumption 3 and (7),

$$
\mathrm{E}\left[\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_\infty^4\right] \le C n^2 \log^2 p.
$$

Therefore,

$$
\begin{aligned}
\mathcal{I}_1 &\le \mathrm{E}\left[\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_\infty \|\widehat{\beta} - \beta\|_1 1\left\{\bar{c}\|\widehat{\beta} - \beta\|_2 \le \|\widehat{\beta} - \beta\|_{2,n}\right\}\right] \\
&\le \mathrm{E}\left[\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_\infty \|\widehat{\beta} - \beta\|_2 (\widehat{s} + s)^{1/2} 1\left\{\bar{c}\|\widehat{\beta} - \beta\|_2 \le \|\widehat{\beta} - \beta\|_{2,n}\right\}\right] \\
&\le C \mathrm{E}\left[\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_\infty \|\widehat{\beta} - \beta\|_{2,n} (\widehat{s} + s)^{1/2} 1\left\{\bar{c}\|\widehat{\beta} - \beta\|_2 \le \|\widehat{\beta} - \beta\|_{2,n}\right\}\right] \\
&\le C \left(\mathrm{E}\left[\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_\infty^2 \|\widehat{\beta} - \beta\|_{2,n}^2\right] \mathrm{E}\left[\widehat{s} + s\right]\right)^{1/2},
\end{aligned}
$$

where the last line follows from Hölder's inequality. In turn,

$$
\begin{aligned}
\left(\mathrm{E}\left[\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_\infty^2 \|\widehat{\beta} - \beta\|_{2,n}^2\right]\right)^{1/2} & \\
&\le \left(\mathrm{E}\left[\left\|\sum_{i=1}^{n} \psi_i X_i\right\|_\infty^4\right] \mathrm{E}\left[\|\widehat{\beta} - \beta\|_{2,n}^4\right]\right)^{1/4} \\
&\le C\sqrt{n \log p}\left(\mathrm{E}\left[\|\widehat{\beta} - \beta\|_{2,n}^4\right]\right)^{1/4} \le C\sqrt{n \log p}\left(R_n + \sqrt{\frac{\log^r n}{n}}\right).
\end{aligned}
$$

Thus,

$$
(47) \qquad \mathcal{I}_1 \le C\sqrt{n \log p}\left(R_n + \sqrt{\frac{\log^r n}{n}}\right)(\mathrm{E}[\widehat{s} + s])^{1/2}.
$$

To bound $\mathcal{I}_2$, denote

$$A_1 = \sqrt{\sum_{i=1}^{n} \psi_i^2} \quad \text{and} \quad A_2 = \sqrt{\sum_{i=1}^{n}(X_i'(\widehat{\beta}-\beta))^2} = \sqrt{n}\|\widehat{\beta}-\beta\|_{2,n}$$

and observe that by Hölder's inequality,

$$\mathcal{I}_2 \le \mathrm{E}\left[A_1 A_2 1\left\{\bar{c}\|\widehat{\beta}-\beta\|_2 > \|\widehat{\beta}-\beta\|_{2,n}\right\}\right] \le \mathcal{I}_{2,1} + \mathcal{I}_{2,2},$$

where

$$\mathcal{I}_{2,1} = \mathrm{E}\left[A_1 A_2 1\left\{A_1 A_2 > \mathcal{C}\left(nR_n + \sqrt{n\log^{r+1}n}\right)\right\}\right],$$

$$\mathcal{I}_{2,2} = \mathcal{C}\left(nR_n + \sqrt{n\log^{r+1}n}\right)\mathrm{P}\left(\bar{c}\|\widehat{\beta}-\beta\|_2 > \|\widehat{\beta}-\beta\|_{2,n}\right),$$

for some constant $\mathcal{C}$ to be chosen later. To bound $\mathcal{I}_{2,1}$, note that

$$\mathrm{P}(A_1 > \sqrt{\mathcal{C}n}) \le 1/n$$

by Chebyshev's inequality and Assumption 3 if $\mathcal{C}$ is large enough. Also, by Lemma 8.2 applied with $\kappa = \log n$,

$$\mathrm{P}\left(A_2/\sqrt{n} > R_n + \sqrt{\mathcal{C}\log^{r+1}n/n}\right) \le 1/n$$

if $\mathcal{C}$ is large enough. Hence, if we set $\mathcal{C}$ in the definition of $\mathcal{I}_{2,1}$ and $\mathcal{I}_{2,2}$ large enough (note that $\mathcal{C}$ can be chosen to depend only on c_1, C_1, and r), it follows that

$$\mathrm{P}\left(A_1 A_2 > \mathcal{C}\left(nR_n + \sqrt{n\log^{r+1}n}\right)\right)$$

$$\le \mathrm{P}(A_1 > \sqrt{\mathcal{C}n}) + \mathrm{P}\left(A_2/\sqrt{n} > R_n + \sqrt{\mathcal{C}\log^{r+1}n/n}\right) \le 2/n,$$

and so $\mathcal{I}_{2,1}$ is bounded from above by

$$\left(\mathrm{E}[A_1^2 A_2^2]\right)^{1/2}\left(\mathrm{P}\left(A_1 A_2 > \mathcal{C}\left(nR_n + \sqrt{n\log^{r+1}n}\right)\right)\right)^{1/2}$$

$$\le Cn\left(R_n + \sqrt{\log^r n/n}\right)/\sqrt{n} \le C(\sqrt{n}R_n + \sqrt{\log^r n}),$$

where the first inequality follows from Hölder's inequality, Assumption 3, and (46). Also, by (7) and Markov's inequality,

$$\mathrm{P}\left(\bar{c}\|\widehat{\beta} - \beta\|_2 > \|\widehat{\beta} - \beta\|_{2,n}\right) \leq \mathrm{P}\left(\widehat{s} + s > J_n\right) \leq \frac{\mathrm{E}[\widehat{s} + s]}{J_n},$$

so that

$$\mathcal{I}_{2,2} \leq \frac{C(nR_n + \sqrt{n\log^{r+1} n})}{J_n}\mathrm{E}[\widehat{s} + s],$$

and so

$$\mathcal{I}_{2,2} \leq 3^{-1}\mathrm{E}[\widehat{s} + s]$$

for all $n \geq n_0$ depending only on c_1, C_1, $\bar{c}$, $\bar{C}$, and r by the definition of J_n.

Combining all inequalities, it follows that for all $n \geq n_0$,

$$\mathrm{E}[\widehat{s}] \leq C\sqrt{n\log p}\left(R_n + \sqrt{\frac{\log^r n}{n}}\right)(\mathrm{E}[\widehat{s} + s])^{1/2}$$

(48)
$$+ C\left(\sqrt{n}R_n + \sqrt{\log^r n}\right) + 3^{-1}\mathrm{E}[\widehat{s} + s],$$

and so

$$\mathrm{E}[\|\widehat{\beta}\|_0] = \mathrm{E}[\widehat{s}] \leq s + C(\log p)(nR_n^2 + \log^r n).$$

This gives the asserted claim for all $n \geq n_0$ and since the asserted claim for $n < n_0$ is trivial, the proof is complete. ■

9. Technical Lemmas.

LEMMA 9.1.   *Let $X_1, \ldots, X_n$ be independent centered random vectors in $\mathbb{R}^p$ with $p \geq 2$. Define $Z = \|\sum_{i=1}^n X_i\|_\infty$, $M = \max_{1 \leq i \leq n}\|X_i\|_\infty$, and $\sigma^2 = \max_{1 \leq j \leq p}\sum_{i=1}^n \mathrm{E}[X_{ij}^2]$. Then*

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

where $K > 0$ is a universal constant.

PROOF.   See Lemma E.1 in [23]. ■

LEMMA 9.2.   *Consider the setting of Lemma 9.1. For every $\eta > 0$, $t > 0$, and $q \geq 1$, we have*

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

where the constant $K > 0$ depends only on η and q.

PROOF. See Lemma E.2 in [23]. ∎

REMARK 9.1. In Lemmas 9.1 and 9.2, if, in addition, we assume that $X_1, \ldots, X_n$ are Gaussian, then $\mathrm{E}[Z] \leq \sigma\sqrt{2\log p}$ by Lemma A.3.1 in [39] and for every $t > 0$, $\mathrm{P}(Z > \mathrm{E}[Z] + t) \leq \exp(-t^2/(2\sigma^2))$ by Theorem 2.1.1 in [2]. ∎

LEMMA 9.3. 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 $\ell \geq 1$, let

$$\delta_n = \frac{K\sqrt{\ell \log p}}{\sqrt{n}} \left(1 + (\log \ell)(\log^{1/2} n)\right).$$

Moreover, let $\mathcal{S}^p(\ell) = \{\theta \in \mathbb{R}^p \colon \|\theta\| = 1 \text{ and } \|\theta\|_0 \leq \ell\}$. Then

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

where $C > 0$ is a universal constant.

PROOF. See Lemma B.1 in [12]. See also [35] for the original result. ∎

REMARK 9.2. If $X_1, \ldots, X_n$ are centered Gaussian random vectors in $\mathbb{R}^p$ with $p \geq 2$, then for any $\epsilon_1, \epsilon_2, \ell > 0$ such that $\epsilon_1 + \epsilon_2 < 1$ and $\ell \leq \min(p, \epsilon_1^2 n)$,

$$\sup_{\theta \in \mathcal{S}^p(\ell)} \left|\frac{1}{n}\sum_{i=1}^n (X_i'\theta)^2 - \mathrm{E}[(X_1'\theta)^2]\right| \leq 3(\epsilon_1 + \epsilon_2) \sup_{\theta \in \mathcal{S}^p(\ell)} \mathrm{E}[(X_1'\theta)^2]$$

with probability at least $1 - 2p^m e^{-n\epsilon_2^2/2}$ by the proof of Proposition 2 in [48]. ∎

LEMMA 9.4. Let ε be a random variable that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$ with continuously differentiable pdf χ and suppose that $f \colon \mathbb{R} \to \mathbb{R}$ is either Lipschitz-continuous or continuously differentiable with finite $\mathrm{E}[|f'(\varepsilon)|]$. Suppose also that both $\mathrm{E}[|f(\varepsilon)|]$ and $\mathrm{E}[|f(\varepsilon)\chi'(\varepsilon)|/\chi(\varepsilon)]$ are finite. Then

(49) $\mathrm{E}[f'(\varepsilon)] = -\mathrm{E}[f(\varepsilon)\chi'(\varepsilon)/\chi(\varepsilon)].$

REMARK 9.3. When ε has a $N(0, \sigma^2)$ distribution, the formula (49) reduces to the well-known Stein identity, $\mathrm{E}[f'(\varepsilon)] = \mathrm{E}[\varepsilon f(\varepsilon)]/\sigma^2$. ∎

PROOF. The proof follows immediately from integration by parts and the Lebesgue dominated convergence theorem; for example, see Section 13.1.1 in [20] for similar results. ∎

LEMMA 9.5. *Let $e = (e_1, \ldots, e_n)$ be a standard Gaussian random vector and let $Q_i \colon \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$ be some strictly increasing continuously differentiable functions. Denote $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ where $\varepsilon_i = Q_i(e_i)$, $i = 1, \ldots, n$, and let $f \colon \mathbb{R}^n \to \mathbb{R}$ be Lipschitz-continuous with Lipschitz constant $L > 0$. Then for any convex $u \colon \mathbb{R} \to \mathbb{R}_+$, the random variable*

$$V = f(\varepsilon) = f(\varepsilon_1, \ldots, \varepsilon_n)$$

satisfies the following inequality:

$$\mathrm{E}[u(V - \mathrm{E}[V])] \leq \mathrm{E}\left[u\left(\frac{\pi L}{2} \max_{1 \leq i \leq n} Q_i'(e_i)\xi \right) \right],$$

where ξ is a standard Gaussian random variable that is independent of e.

REMARK 9.4. The proof of this lemma given below mimics the well-known interpolation proof of the Gaussian concentration inequality for Lipschitz functions; see Theorem 2.1.12 in [40] for example. ∎

PROOF. To prove the asserted claim, let $\widetilde{e} = (\widetilde{e}_1, \ldots, \widetilde{e}_n)$ be another standard Gaussian random vector that is independent of e. Also, define

$$F(x) = F(x_1, \ldots, x_n) = f(Q_1(x_1), \ldots, Q_n(x_n)), \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Then

$$\begin{aligned}
\mathbb{E}[u(V - \mathbb{E}V)] &= \mathbb{E}\Big[u(F(e) - \mathbb{E}F(e))\Big] = \mathbb{E}\Big[u(F(e) - \mathbb{E}F(\widetilde{e}))\Big] \\
&= \mathbb{E}\Big[u(\mathbb{E}[F(e) - F(\widetilde{e}) \mid e])\Big] \leq \mathbb{E}\Big[\mathbb{E}\Big[u(F(e) - F(\widetilde{e})) \mid e\Big]\Big] \\
&= \mathbb{E}\Big[u(F(e) - F(\widetilde{e}))\Big].
\end{aligned}$$

Further, define

$$h(\theta) = F\Big(\widetilde{e}\cos(\pi\theta/2) + e\sin(\pi\theta/2)\Big), \quad \theta \in [0, 1],$$

so that $h(1) = F(e)$, $h(0) = F(\widetilde{e})$, and for all $\theta \in (0, 1)$,

$$\begin{aligned}
h'(\theta) &= \frac{\pi}{2} \sum_{i=1}^{n} F_i(\widetilde{e}\cos(\pi\theta/2) + e\sin(\pi\theta/2))(e_j\cos(\pi\theta/2) - \widetilde{e}_j\sin(\pi\theta/2)) \\
&= \frac{\pi}{2}(\nabla F(\widetilde{W}_\theta), W_\theta),
\end{aligned}$$

where we denoted

$$W_\theta = e \cos(\pi\theta/2) - \widetilde{e} \sin(\pi\theta/2) \text{ and } \widetilde{W}_\theta = \widetilde{e} \cos(\pi\theta/2) + e \sin(\pi\theta/2).$$

Note that for each $\theta \in (0,1)$, the random vectors W_θ and $\widetilde{W}_\theta$ are independent standard Gaussian. Hence,

$$
\begin{aligned}
\mathrm{E}\Big[u(F(e) - F(\widetilde{e}))\Big] = \mathbb{E}\Big[u(h(1) - h(0))\Big] &= \mathbb{E}\Big[u\Big(\int_0^1 h'(\theta)d\theta\Big)\Big] \\
&\leq \mathbb{E}\Big[\int_0^1 u(h'(\theta))d\theta\Big] = \mathbb{E}\Big[\int_0^1 u\Big(\frac{\pi}{2}(\nabla F(\widetilde{W}_\theta), W_\theta)\Big)d\theta\Big] \\
&= \int_0^1 \mathbb{E}\Big[u\Big(\frac{\pi}{2}(\nabla F(\widetilde{W}_\theta), W_\theta)\Big)\Big]d\theta \\
&= \int_0^1 \mathbb{E}\Big[u\Big(\frac{\pi}{2}(\nabla F(e), \widetilde{e})\Big)\Big]d\theta = \mathbb{E}\Big[u\Big(\frac{\pi}{2}(\nabla F(e), \widetilde{e})\Big)\Big].
\end{aligned}
$$

Next, note that since e and $\widetilde{e}$ are independent standard Gaussian random vectors, conditional on e, the random variable $(\nabla F(e), \widetilde{e})$ is zero-mean Gaussian with variance

$$
\begin{aligned}
\sum_{i=1}^n \Big(\frac{\partial F}{\partial e_i}(e)\Big)^2 = \sum_{i=1}^n \Big(\frac{\partial f}{\partial \varepsilon_i}(\varepsilon)\Big)^2 (Q_i'(e_i))^2 \\
\leq \max_{1 \leq i \leq n} (Q_i'(e_i))^2 \sum_{i=1}^n \Big(\frac{\partial f}{\partial \varepsilon_i}(\varepsilon)\Big)^2 \leq L^2 \max_{1 \leq i \leq n}(Q_i'(e_i))^2.
\end{aligned}
$$

Therefore, using the fact that u is convex, we conclude that

$$
\begin{aligned}
\mathbb{E}\Big[u\Big(\frac{\pi}{2}(\nabla F(e), \widetilde{e})\Big)\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[u\Big(\frac{\pi}{2}(\nabla F(e), \widetilde{e})\Big) \mid e\Big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[u\Big(\frac{\pi}{2}\Big(\sum_{i=1}^n \Big(\frac{\partial F}{\partial e_i}(e)\Big)^2\Big)^{1/2}\xi\Big) \mid e\Big]\Big] \\
&\leq \mathbb{E}\Big[\mathbb{E}\Big[u\Big(\frac{\pi L}{2} \max_{1 \leq i \leq n} Q_i'(e_i)\xi\Big) \mid e\Big]\Big] \\
&= \mathbb{E}\Big[u\Big(\frac{\pi L}{2} \max_{1 \leq i \leq n} Q_i'(e_i)\xi\Big)\Big],
\end{aligned}
$$

where ξ is a standard Gaussian random variable that is independent of the vector e. Combining presented inequalities gives the asserted claim. ∎

LEMMA 9.6. *Let* $X_1, \ldots, X_m$ *be random variables (not necessarily independent). Then for all* $\alpha \in (0,1)$,

$$Q_{1-\alpha}(X_1 + \cdots + X_m) \leq Q_{1-\alpha/(2m)}(X_1) + \cdots + Q_{1-\alpha/(2m)}(X_m),$$

where for any random variable Z and any number $\alpha \in (0,1)$, $Q_\alpha(Z)$ denotes the αth quantile of the distribution of Z, i.e. $Q_\alpha(Z) = \inf\{z \in \mathbb{R} : \alpha \leq \mathrm{P}(Z \leq z)\}$.

PROOF. To prove the asserted claim, suppose to the contrary that

$$Q_{1-\alpha}(X_1 + \cdots + X_m) > Q_{1-\alpha/(2m)}(X_1) + \cdots + Q_{1-\alpha/(2m)}(X_m).$$

Then by the union bound,

$$\begin{aligned}
\alpha &\leq \mathrm{P}(X_1 + \cdots + X_m \geq Q_{1-\alpha}(X_1 + \cdots + X_m)) \\
&\leq \mathrm{P}(X_1 + \cdots + X_m > Q_{1-\alpha/(2m)}(X_1) + \cdots + Q_{1-\alpha/(2m)}(X_m)) \\
&\leq \mathrm{P}(X_1 > Q_{1-\alpha/(2m)}(X_1)) + \cdots + \mathrm{P}(X_m > Q_{1-\alpha/(2m)}(X_m)) \\
&\leq \alpha/(2m) + \cdots + \alpha/(2m) = \alpha/2,
\end{aligned}$$

which is a contradiction. Thus, the asserted claim follows. ∎

FIG 2. *DGP1, $n = 100$, $p = 40$, and $\rho = 0.75$. The top-left, top-right, middle-left, and middle-right panels show the mean of estimation error of Lasso estimators in the prediction, L^2, L^1, and out-of-sample prediction 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 25 as these give the most meaningful comparisons). The solid, dotted, and dashed-dotted horizontal lines represent the mean of the estimation error of the CV-Lasso, SZ-Lasso, and B-Lasso estimators, respectively.*
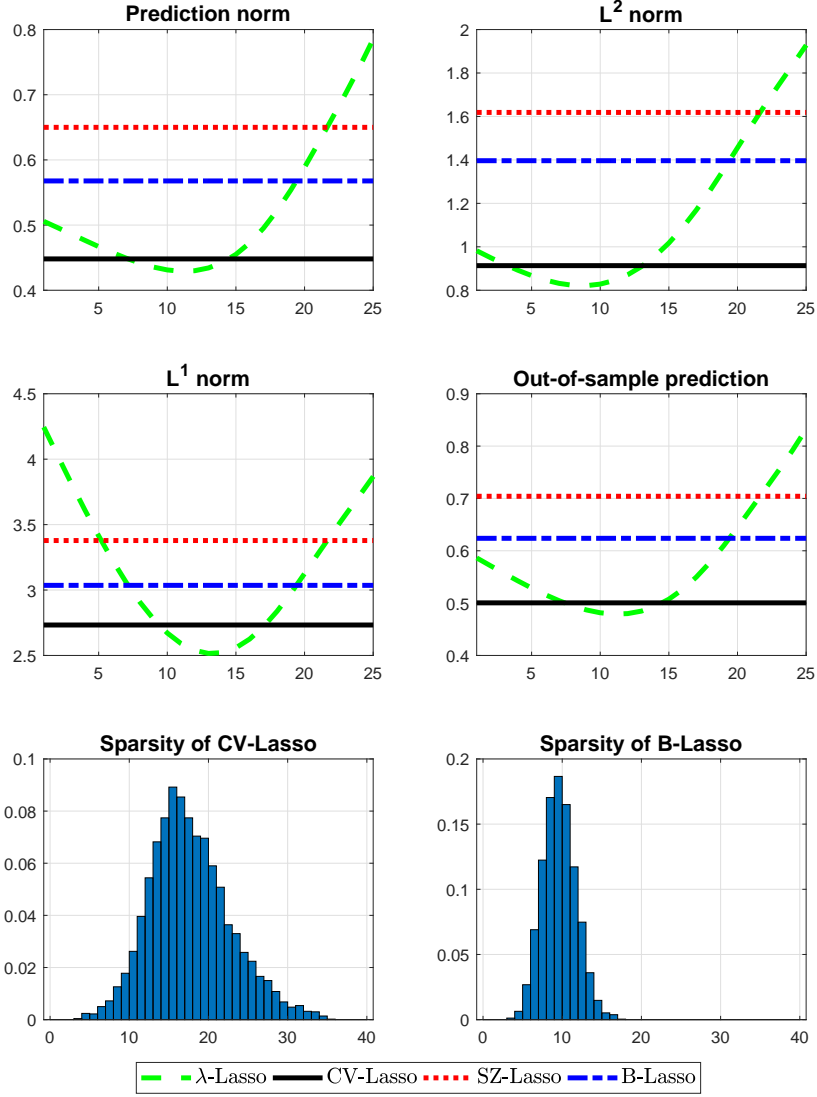
FIG 3. *DGP1, $n = 100$, $p = 400$, and $\rho = 0.75$. The top-left, top-right, middle-left, and middle-right panels show the mean of estimation error of Lasso estimators in the prediction, L^2, L^1, and out-of-sample prediction norms. The dashed line represents the mean of estimation error of the Lasso 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 25 as these give the most meaningful comparisons). The solid, dotted, and dashed-dotted horizontal lines represent the mean of the estimation error of the CV-Lasso, SZ-Lasso, and B-Lasso estimators, respectively.*
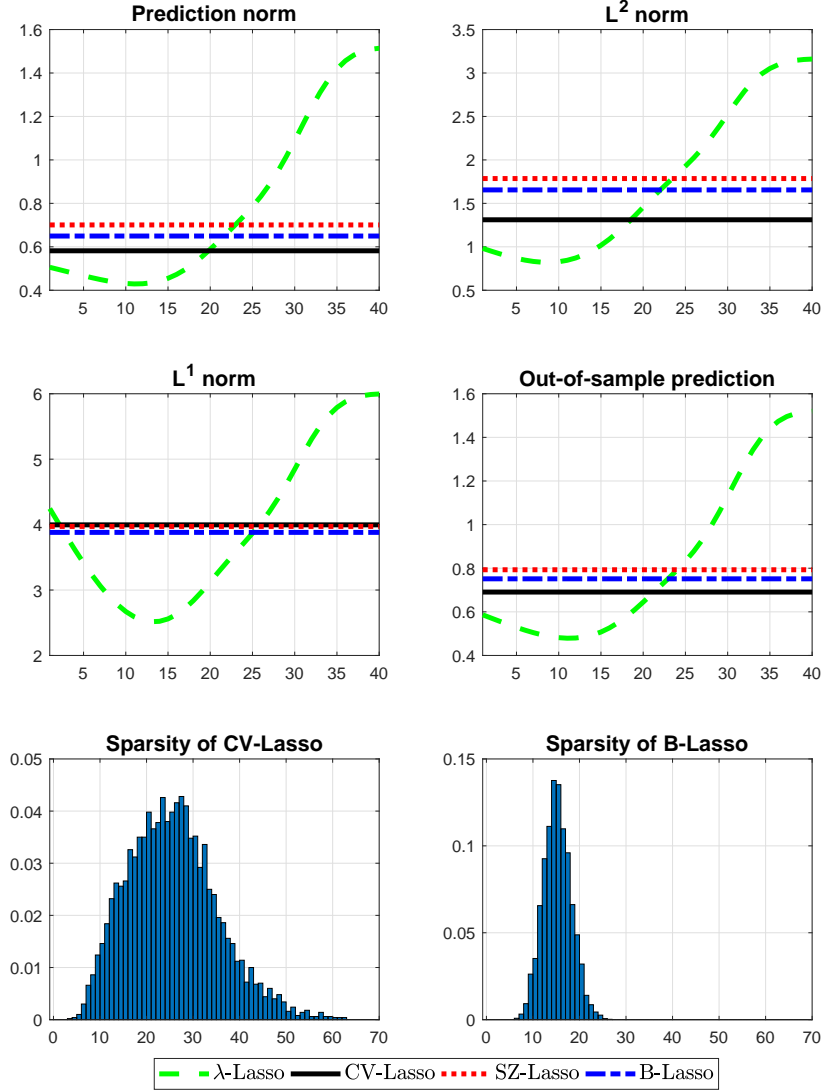
TABLE 1
The mean of estimation error of Lasso estimators

	DGP1 ($\rho = 0.5$)								
	Prediction norm			L^2 norm			Out-of-Sample prediction norm		
	CV-Lasso	λ-Lasso	B-Lasso	CV-Lasso	λ-Lasso	B-Lasso	CV-Lasso	λ-Lasso	B-Lasso
(n, p)=(100, 40)	0.4252	0.4097	0.4435	0.6164	0.5700	0.7013	0.4701	0.4530	0.4883
(n, p)=(100, 100)	0.5243	0.5040	0.5303	0.8206	0.7598	0.8897	0.6091	0.5885	0.6139
(n, p)=(100, 400)	0.7023	0.6448	0.6595	1.2629	1.1624	1.2548	0.8852	0.8474	0.8565
(n, p)=(400, 40)	0.2116	0.2047	0.2174	0.2875	0.2634	0.3186	0.2164	0.2095	0.2224
(n, p)=(400, 100)	0.2581	0.2501	0.2561	0.3674	0.3301	0.3790	0.2667	0.2588	0.2648
(n, p)=(400, 400)	0.3300	0.3206	0.3206	0.5018	0.4546	0.4807	0.3473	0.3391	0.3391
	DGP2 ($\rho = 0.5$)								
	Prediction norm			L^2 norm			Out-of-Sample prediction norm		
(n, p)=(100, 40)	0.7532	0.7123	0.7672	1.1041	0.9907	1.2107	0.8293	0.7857	0.8419
(n, p)=(100, 100)	0.9237	0.8641	0.8917	1.4644	1.3044	1.4792	1.0551	1.0048	1.0264
(n, p)=(100, 400)	1.1497	1.0465	1.0493	1.9868	1.8541	1.8962	1.3631	1.3103	1.3118
(n, p)=(400, 40)	0.3647	0.3521	0.3746	0.4961	0.4529	0.5485	0.3731	0.3603	0.3831
(n, p)=(400, 100)	0.4470	0.4325	0.4431	0.6351	0.5717	0.6550	0.4616	0.4473	0.4577
(n, p)=(400, 400)	0.5739	0.5564	0.5561	0.8714	0.7882	0.8333	0.6037	0.5885	0.5882
	DGP1 ($\rho = 0.75$)								
	Prediction norm			L^2 norm			Out-of-Sample prediction norm		
	CV-Lasso	λ-Lasso	B-Lasso	CV-Lasso	λ-Lasso	B-Lasso	CV-Lasso	λ-Lasso	B-Lasso
(n, p)=(100, 40)	0.4481	0.4292	0.5677	0.9133	0.8213	1.3963	0.5005	0.4791	0.6238
(n, p)=(100, 100)	0.5817	0.5486	0.6496	1.3110	1.1144	1.6547	0.6907	0.6611	0.7514
(n, p)=(100, 400)	0.7616	0.6957	0.7288	2.0360	1.8350	2.0207	0.9836	0.9525	0.9543
(n, p)=(400, 40)	0.2206	0.2141	0.2829	0.4143	0.3745	0.6556	0.2263	0.2196	0.2894
(n, p)=(400, 100)	0.2782	0.2717	0.3322	0.5381	0.4688	0.7766	0.2897	0.2830	0.3436
(n, p)=(400, 400)	0.3847	0.3771	0.4112	0.8217	0.6751	0.9774	0.4151	0.4081	0.4402
	DGP2 ($\rho = 0.75$)								
	Prediction norm			L^2 norm			Out-of-Sample prediction norm		
(n, p)=(100, 40)	0.7730	0.7285	0.8393	1.6151	1.3895	1.9690	0.8520	0.8072	0.9105
(n, p)=(100, 100)	0.9619	0.8843	0.9407	2.1316	1.8093	2.2295	1.0938	1.0293	1.0631
(n, p)=(100, 400)	1.2454	1.0586	1.0740	2.8271	2.4914	2.6602	1.3966	1.3298	1.3298
(n, p)=(400, 40)	0.3811	0.3696	0.4876	0.7141	0.6427	1.1292	0.3907	0.3788	0.4984
(n, p)=(400, 100)	0.4859	0.4719	0.5710	0.9443	0.8132	1.3320	0.5061	0.4920	0.5910
(n, p)=(400, 400)	0.6790	0.6499	0.6834	1.5102	1.1683	1.6067	0.7229	0.7028	0.7291

TABLE 2

Probabilities for the number of non-zero coefficients of the CV-Lasso estimator hitting different brackets

	[0, 5]	[6, 10]	[11, 15]	[16, 20]	[21, 25]	[26, 30]	[31, 35]	[36, p]
DGP1 ($\rho = 0.5$)								
(n, p)=(100, 40)	0.0008	0.0766	0.3598	0.3548	0.1582	0.0390	0.0088	0.0020
(n, p)=(100, 100)	0.0006	0.0120	0.0822	0.2146	0.2606	0.1994	0.1186	0.1120
(n, p)=(100, 400)	0.0010	0.0190	0.0480	0.0760	0.0978	0.1196	0.1288	0.5098
(n, p)=(400, 40)	0.0006	0.0964	0.3926	0.3460	0.1292	0.0316	0.0034	0.0002
(n, p)=(400, 100)	0.0006	0.0176	0.1404	0.2624	0.2596	0.1780	0.0828	0.0586
(n, p)=(400, 400)	0.0000	0.0016	0.0212	0.0728	0.1372	0.1618	0.1664	0.4390

	[0, 5]	[6, 10]	[11, 15]	[16, 20]	[21, 25]	[26, p]	[31, p]	[36, p]
DGP2 ($\rho = 0.5$)								
(n, p)=(100, 40)	0.0142	0.1436	0.3418	0.3070	0.1402	0.0432	0.0094	0.0006
(n, p)=(100, 100)	0.0158	0.1096	0.1866	0.2186	0.1828	0.1338	0.0754	0.0774
(n, p)=(100, 400)	0.0310	0.0988	0.1586	0.1752	0.1446	0.1042	0.0830	0.2046
(n, p)=(400, 40)	0.0008	0.1030	0.4032	0.3334	0.1258	0.0268	0.0060	0.0010
(n, p)=(400, 100)	0.0002	0.0202	0.1358	0.2530	0.2684	0.1704	0.0814	0.0706
(n, p)=(400, 400)	0.0002	0.0020	0.0274	0.0798	0.1280	0.1590	0.1592	0.4444

	[0, 5]	[6, 10]	[11, 15]	[16, 20]	[21, 25]	[26, p]	[31, p]	[36, p]
DGP1 ($\rho = 0.75$)								
(n, p)=(100, 40)	0.0028	0.0448	0.2658	0.3920	0.2050	0.0716	0.0178	0.0002
(n, p)=(100, 100)	0.0006	0.0316	0.1080	0.1604	0.1948	0.2000	0.1470	0.1576
(n, p)=(100, 400)	0.0206	0.0194	0.0506	0.1110	0.1534	0.1660	0.1398	0.3392
(n, p)=(400, 40)	0.0000	0.0278	0.2926	0.4222	0.1966	0.0510	0.0090	0.0008
(n, p)=(400, 100)	0.0000	0.0002	0.0136	0.1156	0.2480	0.2920	0.1836	0.1470
(n, p)=(400, 400)	0.0000	0.0000	0.0002	0.0004	0.0060	0.0192	0.0530	0.9212

	[0, 5]	[6, 10]	[11, 15]	[16, 20]	[21, 25]	[26, p]	[31, p]	[36, p]
DGP2 ($\rho = 0.75$)								
(n, p)=(100, 40)	0.0254	0.2152	0.3326	0.2546	0.1206	0.0392	0.0116	0.0008
(n, p)=(100, 100)	0.0904	0.1024	0.2192	0.2262	0.1606	0.0958	0.0502	0.0552
(n, p)=(100, 400)	0.3916	0.1022	0.0988	0.0906	0.0826	0.0650	0.0558	0.1134
(n, p)=(400, 40)	0.0002	0.0290	0.2976	0.4314	0.1862	0.0468	0.0082	0.0006
(n, p)=(400, 100)	0.0000	0.0050	0.0282	0.1264	0.2370	0.2820	0.1804	0.1410
(n, p)=(400, 400)	0.0002	0.0134	0.0582	0.0974	0.1156	0.1020	0.0860	0.5272

TABLE 3
Probabilities for $\max_{1\le j\le p} n^{-1}|\sum_{i=1}^{n} X_{ij}\varepsilon_i|/\lambda$ *hitting different brackets*

	[0, 0.5)	[0.6, 1)	[1, 1.5)	[1.5, 2)	[2, 2.5)	[2.5, 3)	[3, ∞)
DGP1 ($\rho = 0.50$)							
(n, p)=(100, 40)	0.0002	0.0910	0.3458	0.2842	0.1460	0.0670	0.0648
(n, p)=(100, 100)	0.0000	0.1560	0.4376	0.2470	0.0910	0.0322	0.0338
(n, p)=(100, 400)	0.0116	0.3262	0.3374	0.1396	0.0592	0.0282	0.0566
(n, p)=(400, 40)	0.0000	0.1118	0.4292	0.3042	0.0988	0.0364	0.0182
(n, p)=(400, 100)	0.0000	0.2648	0.5784	0.1362	0.0158	0.0032	0.0004
(n, p)=(400, 400)	0.0000	0.5828	0.3972	0.0162	0.0004	0.0000	0.0000
DGP2 ($\rho = 0.50$)							
(n, p)=(100, 40)	0.0020	0.1522	0.3296	0.2502	0.1322	0.0596	0.0674
(n, p)=(100, 100)	0.0084	0.3096	0.3772	0.1650	0.0624	0.0254	0.0208
(n, p)=(100, 400)	0.0394	0.5254	0.2252	0.0616	0.0210	0.0090	0.0252
(n, p)=(400, 40)	0.0002	0.1170	0.4452	0.2860	0.1006	0.0302	0.0204
(n, p)=(400, 100)	0.0000	0.2676	0.5656	0.1422	0.0198	0.0022	0.0014
(n, p)=(400, 400)	0.0000	0.5908	0.3894	0.0156	0.0008	0.0002	0.0000
DGP1 ($\rho = 0.75$)							
(n, p)=(100, 40)	0.0000	0.0224	0.1220	0.2250	0.2012	0.1488	0.2796
(n, p)=(100, 100)	0.0008	0.1144	0.2546	0.2306	0.1698	0.0944	0.1312
(n, p)=(100, 400)	0.0316	0.4068	0.3408	0.1072	0.0346	0.0164	0.0284
(n, p)=(400, 40)	0.0000	0.0098	0.1384	0.2800	0.2620	0.1526	0.1572
(n, p)=(400, 100)	0.0000	0.0144	0.2918	0.4250	0.1868	0.0592	0.0228
(n, p)=(400, 400)	0.0000	0.0684	0.6724	0.2304	0.0242	0.0040	0.0006
DGP2 ($\rho = 0.75$)							
(n, p)=(100, 40)	0.0062	0.1090	0.2424	0.2142	0.1508	0.1040	0.1674
(n, p)=(100, 100)	0.0686	0.2298	0.3256	0.1842	0.0798	0.0382	0.0518
(n, p)=(100, 400)	0.3616	0.3000	0.1594	0.0508	0.0186	0.0080	0.0118
(n, p)=(400, 40)	0.0000	0.0102	0.1306	0.2918	0.2750	0.1482	0.1442
(n, p)=(400, 100)	0.0000	0.0292	0.2984	0.4072	0.1864	0.0560	0.0226
(n, p)=(400, 400)	0.0004	0.3798	0.4626	0.1344	0.0134	0.0016	0.0002

References.

[1] Abadie, A. and Kasy, M. (2018). Choosing among regularized estimators in empirical economics. *Review of Economics and Statistics, forthcoming*.

[2] Adler, R. and Taylor, J. (2007). Random fields and geometry. *Springer*.

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

[4] Pierre C. Bellec (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. *arXiv preprint*, arXiv:1804.01230.

[5] Bellec, P. and Tsybakov, A. (2017). Bounds on the prediction error of penalized least squares estimators with convex penalty. *Modern Problems of Stochastic Analysis and Statistics, Selected Contributions in Honor of Valentin Konakov*, 315-333.

[6] Bellec, P. and Zhang, C.-H. (2018). Second order Stein: SURE for SURE and other applications in high-dimensional inference. *arXiv:1811.04121*.

[7] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80**, 2369-2429.

[8] 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.

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

[10] Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and regularized GMM. *Arxiv: 1806.01888*.

[11] 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.

[12] 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*.

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

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

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

[16] Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., and Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review Papers and Proceedings*, **106**, 124-127.

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

[18] Chatterjee, S. (2015). High dimensional regression and matrix estimation without tuning parameters. *arXiv:1510.07294*.

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

[20] Chen, L., Goldstein, L., and Shao, Q.-M. (2011). Normal approximation by Stein's method. *Springer: Probability and Its Applications*.

[21] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, **21**, 1-68.

[22] 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.

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

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

[25] Giraud, C. (2015). Introduction to high-dimensional statistics. *CRC Press.*

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

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

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

[29] Homrighausen, D. and McDonald, D. (2017). Risk consistency of cross-validation with Lasso-type procedures. *Statistica Sinica*, **27**, 1017-1036.

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

[31] 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.

[32] Miolane, L. and Montanari, A. (2018). The distribution of the Lasso: uniform control over sparse balls and adaptive parameter tuning. *arXiv: 1811.01212.*

[33] Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, **31**, 87-106.

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

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

[36] Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, **82**, 1-9.

[37] Sala-I-Martin, X. (1997). I just run two million regressions. *American Economic Review Papers and Proceedings*, **87**, 178-183.

[38] Sun, T. and C.-H. Zhang (2013). Sparse matrix inversion with scaled Lasso. *Journal of Machine Learning Research*, **14**, 3385-3418.

[39] Talagrand, T. (2011). Mean field models for spin glasses. *Springer.*

[40] Tao, T. (2012). Topics in random matrix theory. *American Mathematical Society.*

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

[42] Tibshirani, R. and Taylor, J. (2012). Degrees of freedom in Lasso problems. *The Annals of Statistics*, **40**, 1198-1232.

[43] Tibshirani, R. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics*, **7**, 1456-1490.

[44] van de Geer, S. (2016). Estimation and testing under sparsity. *Springer: Lecture Notes in Mathematics.*

[45] van der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes. *Springer Series in Statistics.*

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

Press.

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

[48] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**, 1567-1594.

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

Department of Economics, UCLA
Bunche Hall, 8369
315 Portola Plaza
Los Angeles, CA 90095, USA.
E-mail: chetverikov@econ.ucla.edu

Department of Economics, UCLA
Bunche Hall, 8379
315 Portola Plaza
Los Angeles, CA 90095, USA.
E-mail: zhipeng.liao@econ.ucla.edu

Department of Economics and
Operations Research Center, MIT
50 Memorial Drive
Cambridge, MA 02142, USA.
E-mail: vchern@mit.edu