ON CROSS-VALIDATED LASSO IN HIGH DIMENSIONS

Author(s): Denis Chetverikov, Zhipeng Liao and Victor Chernozhukov

Source: *The Annals of Statistics*, June 2021, Vol. 49, No. 3 (June 2021), pp. 1300-1317

Published by: Institute of Mathematical Statistics

Stable URL: https://www.jstor.org/stable/10.2307/27169421

The Annals of Statistics
2021, Vol. 49, No. 3, 1300–1317
https://doi.org/10.1214/20-AOS2000
© Institute of Mathematical Statistics, 2021

ON CROSS-VALIDATED LASSO IN HIGH DIMENSIONS

BY DENIS CHETVERIKOV[1,*], ZHIPENG LIAO[1,†] AND VICTOR CHERNOZHUKOV[2]

[1]*Department of Economics, UCLA,* [*]*chetverikov@econ.ucla.edu;* [†]*zhipeng.liao@econ.ucla.edu*
[2]*Department of Economics and Operations Research Center, MIT, vchern@mit.edu*

In this paper, we derive nonasymptotic 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 nonzero 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 [24], 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 [7, 9, 23, 32] and [5] among other papers. Nonetheless, in practice researchers often rely upon cross-validation to choose λ (see [11]), 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 [10, 14] and [13]). 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 nonasymptotic error bounds for the cross-validated Lasso estimator in different norms.

We consider the regression model

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

where Y is a dependent variable, $X = (X_1, \dots, X_p)'$ a p-vector of covariates, ε unobserved scalar noise and $\beta = (\beta_1, \dots, \beta_p)'$ 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.

Received February 2019; revised July 2020.
MSC2020 subject classifications. 62J07.
Key words and phrases. Cross-validation, Lasso, high-dimensional models, sparsity, nonasymptotic bounds.

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 [24], who introduced the Lasso estimator,

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

where for $b = (b_1, \ldots, b_p)' \in \mathbb{R}^p$, $\|b\|_1 = \sum_{j=1}^p |b_j|$ denotes the L^1 norm of b, and λ is some penalty parameter (the estimator suggested in Tibshirani's paper takes a slightly different form but over 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 [25].

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, [6, 9], and [22]. In this paper, we show that K-fold cross-validation indeed provides an appropriate way to choose λ. More specifically, we derive nonasymptotic 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, that is, 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 [15–17] and Miolane and Montanari [21] but all these papers imposed strong conditions and made substantial use of these conditions meaning that it is not clear how to relax them. In particular, [16] assumed that p is much smaller than n, and only showed consistency of the (leave-one-out) cross-validated Lasso estimator. [17], which strictly improves upon [15], 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. Miolane and Montanari [21] 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 [11] and Lecué and Mitchell [19]. Chatterjee and Jafarov [11] 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. Lecué and Mitchell [19] 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 [19]. (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 nonstandard problem. From the Lasso literature perspective, a fundamental problem is that existing results require that λ satisfies certain restrictions, for example, $\lambda > 2\|n^{-1}\sum_{i=1}^{n}X_i\varepsilon_i\|_\infty$ with high probability, but it is not clear how to verify whether cross-validated choice of λ satisfies these restrictions; see Section 4 for more details and [13], page 105, for additional complications. Also, classical techniques to derive properties of cross-validated estimators developed, for example, in [20] do not apply to the Lasso estimator as those techniques are based on the linearity of the estimators in the vector of values $(Y_1,\dots,Y_n)'$ of the dependent variable, which does not hold in the case of the Lasso estimator. More recent techniques, developed, for example, in [29], help to analyze subsample Lasso estimators like those studied in [19] but are not sufficient for the analysis of the full-sample Lasso estimator considered here. See [2] 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 provide proofs of the main results on the estimation error bounds. In Section 7, 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 8, we provide proofs of our sparsity bounds. In Section 9, we prove a few lemmas stated in Section 6. In Section 10, we collect some technical lemmas that are useful for the proofs of the main results. Sections 7, 8, 9 and 10 are contained in the Supplementary Material [12].

Notation. Throughout the paper, we use the following notation. For any vector $b = (b_1,\dots,b_p)' \in \mathbb{R}^p$, we use $\|b\|_0 = \sum_{j=1}^{p}1\{b_j \neq 0\}$ to denote the number of nonzero 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, $\|b\|_\infty = \max_{1\le j\le p}|b_j|$ 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 $X_1^n = (X_1,\dots,X_n)$. Moreover, we use $\mathcal{S}^p$ to denote the unit sphere in $\mathbb{R}^p$, that is, $\mathcal{S}^p = \{\delta \in \mathbb{R}^p : \|\delta\|_2 = 1\}$, and for any $\ell > 0$, we use $\mathcal{S}^p(\ell)$ to denote the ℓ-sparse subset of $\mathcal{S}^p$, that is, $\mathcal{S}^p(\ell) = \{\delta \in \mathcal{S}^p : \|\delta\|_0 \le \ell\}$.

We introduce more notation in the beginning of Section 6, 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 $\{1,\dots,n\}$; that is, for each $k \in \{1,\dots,K\}$, I_k is a subset of $\{1,\dots,n\}$, for each $k, k' \in \{1,\dots,K\}$ with $k \neq k'$, the sets I_k and $I_{k'}$ have empty intersection, and $\bigcup_{k=1}^K I_k = \{1,\dots,n\}$. For our asymptotic analysis, we will assume that K is a constant that does not depend on n. Further, let Λ_n be a set of candidate values of λ. Now, for $k = 1,\dots,K$ and $\lambda \in \Lambda_n$, let

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

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

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

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 [19], 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). Throughout the paper, let c_1, C_1, a and q be some fixed strictly positive numbers where $a < 1$ and $q > 4$. Also, let $r \geq 0$ be an integer. In addition, denote

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

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 $X = (X_1,\dots,X_p)'$ 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: $n^{4/q}M_n^4 s\log^4(pn) \le C_1 n^{1-c_1}$.

Assumption 2 is a mild growth condition restricting some moments of $\|X\|_\infty$, the number of nonzero 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

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

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 [6, 9], 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) $c_1 \le Q_2(x, e) \le C_1(1 + |e|^r)$, (ii) $\|Q_{22}(x, e)\|_{\psi_2} \le C_1$ and (iii) $\|Q_{222}(x, e)\|_{\psi_1} \le C_1$, 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 nonstandard. It appears in our analysis because, as explained in Remark 4.3 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 [32] and [26]. 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 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, that is, $Q(X, e) = Q(e)$. 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 bimodal, 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: $\Lambda_n = \{C_1 a^l : l = 0, 1, 2, \dots; a^l \ge c_1/n\}$.

It is known from [9] 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

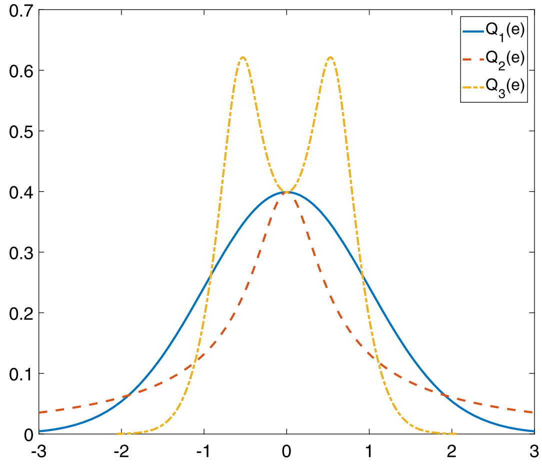

FIG. 1. *The figure plots probability density functions of* $\varepsilon = Q_j(e)$, $j = 1, 2, 3$, *where* $e \sim N(0, 1)$ *and* $Q_1(e) = e$, $Q_2(e) = e + e^3$, *and* $Q_3(e) = (e + e^3)/(1 + 2e^2)$. *All three probability density functions are allowed by Assumption* 3.

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 (9) in Lemma 6.2 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 $\Lambda_n = \{a^l : l = \cdots, -2, -1, 0, 1, 2, \ldots; a^{-l} \leq n^{C_1}, a^l \leq n^{C_1}\}$.

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 nonasymptotic 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 [9] 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 [22] 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 nonzero coefficients; see, for example, [8].

REMARK 4.2 (Varying number of folds K). Theorem 4.1 requires that the number of folds K in the cross-validation procedure is a fixed integer. Therefore, it does not explain how properties of the cross-validated Lasso estimator change with K. In contrast, [21] has recently derived an upper bound on the estimation error of the cross-validated Lasso estimator that is decreasing in K. Their upper bound does not converge to zero, however, whenever p is proportional to n and K is independent of n. Better understanding how properties of the cross-validated Lasso estimator depend on the number of folds K therefore remains an open question.

REMARK 4.3 (On the proof of Theorem 4.1). One of the main steps in [9] is to show that outside of the event

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

where $c > 2$ is some constant, the Lasso estimator $\widehat{\beta}(\lambda)$ satisfies the bound $\|\widehat{\beta}(\lambda) - \beta\|_{2,n} \le C\lambda\sqrt{s}$, 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 it typically yields a rather small value of λ, so that the event (6) with $\lambda = \widehat{\lambda}$ holds with nontrivial (in fact, large) probability even in large samples. Other papers, for example, [3, 4, 18, 30, 31], impose different/weaker conditions on λ but it is still unclear whether these conditions are satisfied by the cross-validated choice of $\lambda = \widehat{\lambda}$. 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 [32] and [26] 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 $\lambda > 0$,

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

which can be found in [27], 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 6.

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 nonzero 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 4.4 (On the sparsity bound). In the previous literature, for example, [7, 30, 31], it has been shown that the Lasso estimator $\widehat{\beta}(\lambda)$ satisfies the bound $\|\widehat{\beta}(\lambda)\|_0 \leq Cs$, for some constant C, outside of the event (6) or under certain different but related conditions. As explained in Remark 4.3, however, cross-validation typically yields a random and rather small value of λ making these conditions difficult to verify/satisfy with $\lambda = \widehat{\lambda}$, and it is typically the case that smaller values of λ lead to the Lasso estimators $\widehat{\beta}(\lambda)$ with a larger number of nonzero 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 that the cross-validated Lasso estimator typically selects too many variables. Our theorem provides a bound on "too": it shows that the number of nonzero 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 4.5 (Near-rate-optimality of cross-validated Lasso estimator in L^1 and L^2 norms). Like in Remark 4.1, the results in [9] 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 [22] 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 4.6 (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 [3]. 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 [3] that for any $\lambda > 0$,

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

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

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

and

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

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 $X_1, \dots, X_n$. 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 $X_1, \dots, X_n$. 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 $\alpha \in (0, 1)$,*

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

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.

5. General sparsity bounds.

As we mentioned in Remark 4.3, our analysis of the (full-sample) cross-validated Lasso estimator $\widehat{\beta}(\widehat{\lambda})$ requires understanding sparsity of the subsample cross-validated Lasso estimators $\widehat{\beta}_{-k}(\widehat{\lambda})$, that is, we need to show that the random variables $\|\widehat{\beta}_{-k}(\widehat{\lambda})\|_0$, $k = 1, \dots, K$, 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 4.4, they depend on difficult to verify/satisfy conditions. 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. 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 $t \geq 1$,*

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

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 $\lambda > 0$,*

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

on the event

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

where

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

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

6. 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 Section 6.1 and give the actual proofs of the theorems and the corollary in Sections 6.2, 6.3, 6.4 and 6.5, respectively.

For convenience, we use the following additional notation. For $k = 1, \dots, K$, we denote

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

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 $a_n \leq Cb_n$. Moreover, for $\delta \in \mathbb{R}^p$ and $M \subset \{1, \dots, p\}$, we use δ_M to denote the vector in $\mathbb{R}^{|M|}$ consisting of all elements of δ corresponding to indices in M.

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

LEMMA 6.1. *Suppose that Assumptions 1 and 2 are satisfied and denote* $\ell_n = \sqrt{sn^{1+c_1/2}}\log(pn)$. *Then*

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

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

LEMMA 6.2. *Under Assumptions 1–5, there exists* $\bar\lambda_0 = \bar\lambda_{n,0} \in \Lambda_n$, *possibly depending on* n, *such that for all* $k = 1,\dots,K$, *we have* $\|\widehat\beta_{-k}(\bar\lambda_0)\|_0 \lesssim s$ *and, in addition,*

$$\|\widehat\beta_{-k}(\bar\lambda_0) - \beta\|_2^2 \lesssim \frac{s\log(pn)}{n} \quad \textit{and} \quad \|\widehat\beta_{-k}(\bar\lambda_0) - \beta\|_1^2 \lesssim \frac{s^2\log(pn)}{n} \tag{9}$$

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

REMARK 6.1. The result in this lemma is essentially well known but we provide a short proof here for completeness.

LEMMA 6.3. *Under Assumptions 1–5, we have for all* $k = 1,\dots,K$ *that*

$$\|\widehat\beta_{-k}(\bar\lambda_0) - \beta\|_{2,n,k}^2 \lesssim \frac{s\log(pn)}{n}$$

with probability $1 - Cn^{-c}$ *for* $\bar\lambda_0$ *defined in Lemma 6.2.*

REMARK 6.2. 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.

LEMMA 6.4. *Under Assumptions 1–5, we have for all* $k = 1,\dots,K$ *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 $1 - Cn^{-c}$.

LEMMA 6.5. *Under Assumptions 1–5, we have for all* $k = 1,\dots,K$ *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 $1 - Cn^{-c}$.

LEMMA 6.6. *Fix* $k = 1,\dots,K$ *and denote*

$$\Lambda_{n,k}(X_1^n, T) = \{\lambda \in \Lambda_n : \mathrm{E}[\|\widehat\beta_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n] \le T\}, \quad T > 0. \tag{10}$$

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*

$$T_n = C\left(\frac{s\log(pn)}{n} + \frac{\log^{r+1} n}{n}\right)^{1/2}. \tag{11}$$

LEMMA 6.7. *For all $\lambda \in \Lambda_n$ and $b \in \mathbb{R}^p$, we have*

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

6.2. *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 6.7 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 6.6 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 $\lambda > 0$,

$$\begin{aligned}
&Q_{1/2}\big(n\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^2 \mid X_1^n\big)\\
&\quad\lesssim \sum_{k=1}^K Q_{1-1/(16K)}\Big(\sum_{i\notin I_k}\big(X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\big)^2 \,\Big|\, X_1^n\Big)\\
&\qquad + \sum_{k=1}^K Q_{1-1/(16K)}\Big(\sum_{i\in I_k}\big(X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\big)^2 \,\Big|\, X_1^n\Big)\\
&\qquad + \sum_{k=1}^K Q_{1-1/(16K)}\Big(\Big|\sum_{i\notin I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\Big| \,\Big|\, X_1^n\Big)\\
&\qquad + \sum_{k=1}^K Q_{1-1/(16K)}\Big(\Big|\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda) - \beta)\Big| \,\Big|\, X_1^n\Big).
\end{aligned} \tag{12}$$

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

$$b(\lambda) = \frac{1}{K-1}\sum_{k=1}^K \frac{n-n_k}{n}\widehat{\beta}_{-k}(\lambda). \tag{13}$$

Then

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

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

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

Thus, by Lemma 6.7,

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

Substituting here $Y_i = X_i'\beta + \varepsilon_i$, $i = 1, \dots, n$, and the definition of $b(\lambda)$ in (13) and using the triangle inequality gives

$$\begin{aligned}
n\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^2 &\lesssim n\|\widehat{\beta}(\lambda) - b(\lambda)\|_{2,n}^2 + n\|b(\lambda) - \beta\|_{2,n}^2 \\
&\lesssim \sum_{k=1}^{K}\sum_{i\notin I_k}\big(X_i'\big(\widehat{\beta}_{-k}(\lambda) - \beta\big)\big)^2 + \sum_{k=1}^{K}\sum_{i\in I_k}\big(X_i'\big(\widehat{\beta}_{-k}(\lambda) - \beta\big)\big)^2 \\
&\quad + \sum_{k=1}^{K}\bigg|\sum_{i\notin I_k}\varepsilon_i X_i'\big(\widehat{\beta}_{-k}(\lambda) - \beta\big)\bigg| + \sum_{k=1}^{K}\bigg|\sum_{i\in I_k}\varepsilon_i X_i'\big(\widehat{\beta}_{-k}(\lambda) - \beta\big)\bigg|.
\end{aligned} \tag{14}$$

The claim of this step, inequality (12), follows from (14) and Lemma 10.6.

Step 2. Denote

$$\Lambda_n\big(X_1^n, T\big) = \bigcap_{k=1}^{K}\Lambda_{n,k}\big(X_1^n, T\big), \quad T > 0, \tag{15}$$

for $\Lambda_{n,k}(X_1^n, T)$ defined in (10) of Lemma 6.6. In this step, we show that

$$\begin{aligned}
&\mathrm{P}\Big(\max_{\lambda\in\Lambda_n(X_1^n, T_n)} Q_{1/2}\big(n\|\widehat{\beta}(\lambda) - \beta\|_{2,n}^2 \mid X_1^n\big) \\
&\quad > C(\log p)\big(s\log(pn) + \log^{r+1} n\big)\Big) \leq Cn^{-c}
\end{aligned} \tag{16}$$

for T_n defined in (11) of Lemma 6.6.

To do so, we apply the result in Step 1 and bound all terms on the right-hand side of (12) in turn. To start, fix $k = 1, \dots, K$. Then, for any $\lambda \in \Lambda_n(X_1^n, T_n)$,

$$\begin{aligned}
&Q_{1-1/(16K)}\bigg(\sum_{i\notin I_k}\big(X_i'\big(\widehat{\beta}_{-k}(\lambda) - \beta\big)\big)^2 \;\bigg|\; X_1^n\bigg) \\
&\quad\leq nQ_{1-1/(16K)}\big(\|\widehat{\beta}_{-k}(\lambda) - \lambda\|_{2,n,-k} \mid X_1^n\big)^2 \\
&\quad\leq 16Kn\big(\mathrm{E}[\|\widehat{\beta}_{-k}(\lambda) - \beta\|_{2,n,-k} \mid X_1^n]\big)^2 \lesssim s\log(pn) + \log^{r+1} n,
\end{aligned} \tag{17}$$

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 6.1 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,

$$\begin{aligned}\max_{\lambda\in\Lambda_n(X_1^n,T_n)} \mathrm{E}\big[\|\widehat{\beta}_{-k}(\lambda)\|_0 \mid X_1^n\big] &\lesssim s + (\log p)\big(nT_n^2 + \log^r n\big)\\ &\lesssim (\log p)\big(s\log(pn) + \log^{r+1} n\big)\end{aligned} \tag{18}$$

with probability at least $1 - Cn^{-c}$. Thus, by Markov's inequality, Lemma 6.1 and Assumptions 1, 2 and 5,

$$\begin{aligned}&\max_{\lambda\in\Lambda_n(X_1^n,T_n)} Q_{1-1/(16K)}\Big(\sum_{i\in I_k}\big(X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\big)^2 \,\Big|\, X_1^n\Big)\\ &\quad\lesssim \max_{\lambda\in\Lambda_n(X_1^n,T_n)} Q_{1-1/(32K)}\big(n\|\widehat{\beta}_{-k}(\lambda)-\beta)\|_{2,n,-k}^2 \mid X_1^n\big)\\ &\quad\lesssim s\log(pn) + \log^{r+1} n,\end{aligned} \tag{19}$$

with probability at least $1 - Cn^{-c}$, where the last inequality follows from the same argument as that in (17).

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 10.1 and Assumptions 2 and 3,

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

with probability at least $1 - Cn^{-c}$. Therefore, by Proposition A.1.6 in [28],

$$\begin{aligned}\mathrm{E}\Big[\Big\|\sum_{i\notin I_k}\varepsilon_i X_i\Big\|_\infty^4 \,\Big|\, X_1^n\Big] &\lesssim \Big(\mathrm{E}\Big[\Big\|\sum_{i\notin I_k}\varepsilon_i X_i\Big\|_\infty \,\Big|\, X_1^n\Big]\Big)^4 + \mathrm{E}\Big[\max_{i\notin I_k}\|\varepsilon_i X_i\|_\infty^4 \mid X_1^n\Big]\\ &\lesssim \big(\sqrt{n\log p} + n^{(1+c_1)/q}M_n\big)^4 \lesssim (n\log p)^2\end{aligned}$$

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

$$\begin{aligned}&\max_{\lambda\in\Lambda_n(X_1^n,T_n)} \mathrm{E}\Big[\Big|\sum_{i\notin I_k}\varepsilon_i X_i'\big(\widehat{\beta}_{-k}(\lambda)-\beta\big)\Big| \,\Big|\, X_1^n\Big]\\ &\quad\lesssim \sqrt{n\log p}\bigg(T_n + \sqrt{\frac{\log^r n}{n}}\bigg)\max_{\lambda\in\Lambda_n(X_1^n,T_n)}\big(\mathrm{E}\big[\|\widehat{\beta}_{-k}(\lambda)\|_0 + s \mid X_1^n\big]\big)^{1/2}\\ &\qquad + \sqrt{n}T_n + \sqrt{\log^r n} + \max_{\lambda\in\Lambda_n(X_1^n,T_n)} \mathrm{E}\big[\|\widehat{\beta}_{-k}(\lambda)\|_0 + s \mid X_1^n\big]\\ &\quad\lesssim s + (\log p)\big(nT_n^2 + \log^r n\big) \lesssim (\log p)\big(s\log(pn) + \log^{r+1} n\big)\end{aligned}$$

with probability at least $1 - Cn^{-c}$, where the second inequality follows from (18). Hence, by Markov's inequality,

$$\max_{\lambda\in\Lambda_n(X_1^n,T_n)} Q_{1-1/(16K)}\biggl(\biggl|\sum_{i\notin I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\biggr| \biggm| X_1^n\biggr)$$
$$\le 16K \max_{\lambda\in\Lambda_n(X_1^n,T_n)} \mathrm{E}\biggl[\biggl|\sum_{i\notin I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\biggr| \biggm| X_1^n\biggr]$$
$$\lesssim (\log p)\bigl(s\log(pn) + \log^{r+1} n\bigr)$$

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

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

$$\mathrm{P}\biggl(\biggl|\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\biggr| > \sqrt{A_1A_2\bigl(s\log(pn)+\log^{r+1} n\bigr)} \biggm| X_1^n\biggr)$$
$$\le \mathrm{P}\biggl(\sum_{i\in I_k}\bigl(X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\bigr)^2 > A_2\bigl(s\log(pn)+\log^{r+1} n\bigr) \biggm| X_1^n\biggr)$$
$$+ \mathrm{E}\biggl[1\biggl\{\sum_{i\in I_k}\bigl(X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\bigr)^2 \le A_2\bigl(s\log(pn)+\log^{r+1} n\bigr)\biggr\}$$
$$\times \frac{\mathrm{E}[|\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)|^2 \mid X_1^n, (Y_i)_{i\notin I_k}]}{A_1A_2(s\log(pn)+\log^{r+1} n)} \biggm| X_1^n\biggr]$$
$$\lesssim \mathrm{P}\biggl(\sum_{i\in I_k}\bigl(X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\bigr)^2 > A_2\bigl(s\log(pn)+\log^{r+1} n\bigr) \biggm| X_1^n\biggr) + 1/A_1.$$

Choosing both A_1 and A_2 here large enough and using the same argument as that in (19) shows that

$$\max_{\lambda\in\Lambda_n(X_1^n,T_n)} Q_{1-1/(16K)}\biggl(\biggl|\sum_{i\in I_k}\varepsilon_i X_i'(\widehat{\beta}_{-k}(\lambda)-\beta)\biggr| \biggm| X_1^n\biggr)$$
$$\lesssim \sqrt{s\log(pn)+\log^{r+1} n} \lesssim s\log(pn)+\log^{r+1} n$$

with probability at least $1 - Cn^{-c}$. Combining all inequalities presented above together and using Step 1 gives (16), 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 $\lambda > 0$,

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

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

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

Combining this inequality with (16) in Step 2 shows that

$$\mathrm{P}\biggl(\max_{\lambda\in\Lambda_n(X_1^n,T_n)} \mathrm{E}[\|\widehat{\beta}(\lambda)-\beta\|_{2,n} \mid X_1^n]$$
$$> \sqrt{\frac{Cs\log p}{n}} \times \sqrt{\log(pn)+s^{-1}\log^{r+1} n}\biggr) \le Cn^{-c}. \tag{20}$$

Also, applying Lemma 8.2 with $\kappa = \log(1/\alpha) \le \log n$ 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 $\lambda > 0$,

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

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

$$\begin{aligned} &\mathrm{P}\Big(\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \|\widehat{\beta}(\lambda) - \beta\|_{2,n} \\ &\quad > \sqrt{\frac{Cs \log(p/\alpha)}{n}} \times \sqrt{\log(pn) + s^{-1} \log^{r+1} n}\Big) \le \alpha + Cn^{-c}. \end{aligned} \tag{21}$$

Finally, by Lemma 6.6 and the union bound,

$$\mathrm{P}\big(\widehat{\lambda} \in \Lambda_n(X_1^n, T_n)\big) \ge 1 - Cn^{-c}. \tag{22}$$

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

6.3. *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 6.1, Theorem 5.2 and (20) in the proof of Theorem 4.1,

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

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

$$\begin{aligned} &\mathrm{P}\Big(\max_{\lambda \in \Lambda_n(X_1^n, T_n)} \|\widehat{\beta}(\lambda)\|_0 > \bar{s} \mid X_1^n\Big) \\ &\quad \lesssim s(\log^2 p)(\log n)\big(\log(pn) + s^{-1} \log^{r+1} n\big)/\bar{s} \end{aligned}$$

with probability at least $1 - Cn^{-c}$. The asserted claim of the theorem follows from combining this bound with (22) in the proof of Theorem 4.1 and substituting

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

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

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

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

with probability at least $1 - Cn^{-c}$. Thus, by Lemma 6.1 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,

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

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. □

6.5. *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. □

Acknowledgments. We thank Mehmet Caner, Matias Cattaneo, Yanqin Fan, Sara van de Geer, Jerry Hausman, James Heckman, Roger Koenker, Andzhey Koziuk, Miles Lopes, Jinchi Lv, Rosa Matzkin, Anna Mikusheva, Whitney Newey, Jesper Sorensen, Vladimir Spokoiny, Larry Wasserman and seminar participants in many places for helpful comments.

Funding. Chetverikov's work was partially funded by NSF Grant SES–1628889. Liao's work was partially funded by NSF Grant SES–1628889.

SUPPLEMENTARY MATERIAL

Supplement to "On cross-validated Lasso in high dimensions" (DOI: 10.1214/20-AOS2000SUPP; .pdf). This supplemental file contains results of a small-scale Monte Carlo simulation study as well as proofs for Section 5, proofs of lemmas in Section 6 and several technical lemmas.

REFERENCES

[1] ABADIE, A. and KASY, M. (2018). Choosing among regularized estimators in empirical economics. *Rev. Econ. Stat.* **100** 743–762.

[2] ARLOT, S. and CELISSE, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* **4** 40–79. MR2602303 https://doi.org/10.1214/09-SS054

[3] BELLEC, P. and ZHANG, C.-H. (2018). Second order Stein: SURE for SURE and other applications in high-dimensional inference. Preprint. Available at arXiv:1811.04121.

[4] BELLEC, P. and ZHANG, C.-H. (2020). De-biasing the Lasso with degrees-of-freedom adjustment. Preprint. Available at arXiv:1902.08885v2.

[5] BELLEC, P. C. (2018). The noise barrier and the large signal bias of the Lasso and other convex estimators. Preprint. Available at arXiv:1804.01230.

[6] BELLONI, A. and CHERNOZHUKOV, V. (2011). High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation*. *Lect. Notes Stat. Proc.* **203** 121–156. Springer, Heidelberg. MR2868201 https://doi.org/10.1007/978-3-642-19989-9_3

[7] BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163 https://doi.org/10.3150/11-BEJ410

[8] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

[9] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[10] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[11] CHATTERJEE, S. and JAFAROV, J. (2015). Prediction error of cross-validated Lasso. Preprint. Available at arXiv:1502.06292.

[12] CHETVERIKOV, D., LIAO, Z. and CHERNOZHUKOV, V. (2021). Supplement to "On cross-validated Lasso in high dimensions." https://doi.org/10.1214/20-AOS2000SUPP

[13] GIRAUD, C. (2015). *Introduction to High-Dimensional Statistics*. *Monographs on Statistics and Applied Probability* **139**. CRC Press, Boca Raton, FL. MR3307991

[14] HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

[15] HOMRIGHAUSEN, D. and MCDONALD, D. (2013). The Lasso, persistence, and cross-validation. In *Proceedings of the* 30*th International Conference on Machine Learning* **28**.

[16] HOMRIGHAUSEN, D. and MCDONALD, D. J. (2014). Leave-one-out cross-validation is risk consistent for Lasso. *Mach*. *Learn*. **97** 65–78. MR3252827 https://doi.org/10.1007/s10994-014-5438-z

[17] HOMRIGHAUSEN, D. and MCDONALD, D. J. (2017). Risk consistency of cross-validation with Lasso-type procedures. *Statist*. *Sinica* **27** 1017–1036. MR3699692

[18] LECUÉ, G. and MENDELSON, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann*. *Statist*. **46** 611–641. MR3782379 https://doi.org/10.1214/17-AOS1562

[19] LECUÉ, G. and MITCHELL, C. (2012). Oracle inequalities for cross-validation type procedures. *Electron*. *J*. *Stat*. **6** 1803–1837. MR2988465 https://doi.org/10.1214/12-EJS730

[20] LI, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann*. *Statist*. **15** 958–975. MR0902239 https://doi.org/10.1214/aos/1176350486

[21] MIOLANE, L. and MONTANARI, A. (2018). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. Preprint. Available at arXiv:1811.01212.

[22] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann*. *Statist*. **39** 731–771. MR2816337 https://doi.org/10.1214/10-AOS854

[23] SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled Lasso. *J*. *Mach*. *Learn*. *Res*. **14** 3385–3418. MR3144466

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

[25] TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron*. *J*. *Stat*. **7** 1456–1490. MR3066375 https://doi.org/10.1214/13-EJS815

[26] TIBSHIRANI, R. J. and TAYLOR, J. (2012). Degrees of freedom in Lasso problems. *Ann*. *Statist*. **40** 1198–1232. MR2985948 https://doi.org/10.1214/12-AOS1003

[27] VAN DE GEER, S. (2016). *Estimation and Testing Under Sparsity*. *Lecture Notes in Math*. **2159**. Springer, Cham. MR3526202 https://doi.org/10.1007/978-3-319-32774-7

[28] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[29] WEGKAMP, M. (2003). Model selection in nonparametric regression. *Ann*. *Statist*. **31** 252–273. MR1962506 https://doi.org/10.1214/aos/1046294464

[30] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann*. *Statist*. **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

[31] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann*. *Statist*. **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

[32] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the “degrees of freedom” of the Lasso. *Ann*. *Statist*. **35** 2173–2192. MR2363967 https://doi.org/10.1214/009053607000000127