Mach Learn (2014) 97:65–78
DOI 10.1007/s10994-014-5438-z

Leave-one-out cross-validation is risk consistent for lasso

Darren Homrighausen · Daniel J. McDonald

Received: 3 March 2013 / Accepted: 6 March 2014 / Published online: 28 March 2014

Abstract The lasso procedure pervades the statistical and signal processing literature, and as such, is the target of substantial theoretical and applied research. While much of this research focuses on the desirable properties that lasso possesses—predictive risk consistency, sign consistency, correct model selection—these results assume that the tuning parameter is chosen in an oracle fashion. Yet, this is impossible in practice. Instead, data analysts must use the data twice, once to choose the tuning parameter and again to estimate the model. But only heuristics have ever justified such a procedure. To this end, we give the first definitive answer about the risk consistency of lasso when the smoothing parameter is chosen via cross-validation. We show that under some restrictions on the design matrix, the lasso estimator is still risk consistent with an empirically chosen tuning parameter.

Keywords Stochastic equicontinuity · Uniform convergence · Persistence

1 Introduction

Since its introduction in the statistical (Tibshirani 1996) and signal processing (Chen et al. 1998) communities, the lasso has become a fixture as both a data analysis tool (see for example Lee et al. 2010; Shi et al. 2008) and as an object for deep theoretical investigations (Fu and Knight 2000; Greenshtein and Ritov 2004; Meinshausen and Bühlmann 2006). To fix ideas, suppose that the observational model is of the form

Editors: Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný.

D. Homrighausen
Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA
e-mail: darrenho@stat.colostate.edu

D. J. McDonald (✉)
Department of Statistics, Indiana University, Bloomington, IN 47408, USA
e-mail: dajmcdon@indiana.edu

$$Y = \mathbb{X}\theta + \sigma W. \tag{1.1}$$

where $Y = (Y_1, \ldots, Y_n)^\top$ is the vector of responses and $\mathbb{X} \in \mathbb{R}^{n \times p}$ is the feature matrix, with rows $(X_i^\top)_{i=1}^n$, W is a noise vector, and σ is the signal-to-noise ratio. Under Eq. (1.1), the lasso estimator, $\widehat{\theta}(\lambda)$, is defined to be the minimizer of the following functional:

$$\widehat{\theta}(\lambda) := \operatorname*{argmin}_{\theta} \frac{1}{2n} \|Y - \mathbb{X}\theta\|^2 + \lambda \|\theta\|_1. \tag{1.2}$$

Here, $\lambda \geq 0$ is a tuning parameter controlling the trade-off between fidelity to the data (small λ) and sparsity (large λ). We tacitly assume that $\mathbb{X}$ has full column rank, and thus, $\widehat{\theta}(\lambda)$ is the unique minimum.

Throughout the paper, we will reserve $||\cdot||$ for the ℓ^2- or Euclidean norm when applied to vectors or the spectral norm when applied to matrices. We indicate other norms with an appropriate subscript: $||\cdot||_F$ is the Frobenius norm of a matrix, $||\cdot||_1$ is the ℓ^1 norm of a vector, and $||x||_A^2 = x^\top A x$ is the ℓ^2-norm of x weighted by a matrix A.

Under conditions on the matrix $\mathbb{X}$, noise vector W, and the parameter θ, the optimal choice of λ leads to predictive risk consistency (Greenshtein and Ritov 2004). However, arguably the most crucial aspect of any procedure's performance is the selection of the tuning parameters. Typically, theory advocating the lasso's empirical properties specifies only the rates. That is, this theory claims "if $\lambda = \lambda_n$ goes to zero at the correct rate, then $\widehat{\theta}(\lambda_n)$ will be consistent in some sense." For the regularized problem in Eq. (1.2), taking $\lambda_n = o((\log(n)/n)^{1/2})$ gives risk consistency under very general conditions. However, this type of theoretical guidance says nothing about the properties of the lasso when the tuning parameter is chosen using the data.

There are several proposed techniques for choosing λ, such as minimizing the empirical risk plus a penalty term based on the degrees of freedom (Zou et al. 2007; Tibshirani and Taylor 2012) or using an adapted Bayesian information criterion (Wang and Leng 2007). In many papers, (Efron et al. 2004; van de Geer and Lederer 2011; Greenshtein and Ritov 2004; Hastie et al. 2009; Tibshirani 1996, 2011; Zou et al. 2007, for example), the recommended technique for selecting λ is to choose $\lambda = \widehat{\lambda}_n$ such that $\widehat{\lambda}_n$ minimizes a cross-validation estimator of the risk.

Some results supporting the use of cross-validation for statistical algorithms other than lasso are known. For instance, kernel regression (Györfi et al. 2002, [Theorem 8.1]), k-nearest neighbors (Györfi et al. 2002, [Theorem 8.1]), and various classification algorithms (Schaffer 1993) all behave well with tuning parameters selected using the data. Additionally, suppose we form the adaptive ridge regression estimator (Grandvalet 1998)

$$\operatorname*{argmin}_{\theta, (\lambda_j)} ||Y - \mathbb{X}\theta||^2 + \sum_{j=1}^{p} \lambda_j \theta_j^2 \tag{1.3}$$

subject to the contraint $\lambda \sum_{j=1}^p 1/\lambda_j = p$. Then the solution to Eq. (1.3) is equivalent, under a reparameterization of λ (Grandvalet, 1998, §3), to the solution to Eq. (1.2). As ridge regression has been shown to have good asymptotic properties under (generalized) cross-validation, there is reason to believe these properties may carry over to lasso and cross-validation using this equivalence. However, rigorous results for the lasso have yet to be developed.

The supporting theory for other methods indicates that there should be corresponding theory for the lasso. However, other results are not so encouraging. In particular, (Shao 1993) shows that cross-validation is inconsistent for model selection. As lasso implicitly does model

selection, and shares many connections with forward stagewise regression (Efron et al. 2004), this raises a concerning possibility that lasso might also be predictive risk inconsistent under cross-validation. Likewise, (Leng et al. 2006) shows that using prediction accuracy (which is what cross-validation estimates) as a criterion for choosing the tuning parameter fails to recover the sparsity pattern consistently in an orthogonal design setting. Furthermore, (Xu et al. 2008) shows that sparsity inducing algorithms like lasso are not (uniformly) algorithmically stable. In other words, leave-one-out versions of the lasso estimator are not uniformly close to each other. As shown in Bousquet and Elisseeff (2002), algorithmic stability is a sufficient, but not necessary, condition for predictive risk consistency.

These results taken as a whole leave the lasso in an unsatisfactory position, with some theoretical results and generally accepted practices advocating the use of cross-validation while others suggest that it may not work. Our result partially resolves this antagonism by showing that, in some cases, the lasso with cross-validated tuning parameter is indeed risk consistent.

In this paper we provide a first result about the predictive risk consistency of lasso with the tuning parameter selected by cross-validation under some assumptions about $\mathbb{X}$ and W. In Sect. 2 we introduce our notation and state our main theorem. In Sect. 3 we state some results necessary for our proof methods and in Sect. 4 we provide the proof. Lastly, in Sect. 5 we mention some implications of our main theorem and some directions for future research.

2 Notation, assumptions, and main results

The main assumptions we make for this paper ensure that the sequence $(X_i)_{i=1}^n$ is sufficiently regular. These are

Assumption A

$$C_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \to C, \tag{2.1}$$

where C is a positive definite matrix with minimum eigenvalue $\text{eigen}_{\min}(C) = c_{\min} > 0$.

Assumption B There exists a constant $C_X < \infty$ independent of n such that

$$||X_i|| \leq C_X. \tag{2.2}$$

Note that Assumption A appears repeatedly in the literature in various contexts (Tibshirani 1996; Fu and Knight 2000; Osborne et al. 2000; Leng et al. 2006, for example). Additionally, Assumption B is similar to the standard requirement $\max_i\{||X_i||, 1 \leq i \leq n\} = O(1)$ as $n \to \infty$. See Chatterjee and Lahiri (2011) for example.

It is important to note that $\widehat{\theta}(\lambda)$ as defined in Eq. (1.2) is random whether or not λ is chosen stochastically. While in this paper we investigate the behavior of $\widehat{\theta}(\lambda)$ when λ is a function of the data, and hence itself stochastic, very nice results describe the asymptotic behavior of $\widehat{\theta}(\lambda)$ for deterministic sequences $\lambda = \lambda_n$. Under Assumptions A and B, (Fu and Knight 2000) shows that for a deterministic sequence such that $\lambda/\sqrt{n} \to 0$, $\sqrt{n}(\widehat{\theta}(\lambda)-\theta) \to N(0, \sigma^2 C^{-1})$ in distribution. Alternatively, if $\lambda/\sqrt{n} \to \lambda_0 > 0$, then $\sqrt{n}(\widehat{\theta}(\lambda) - \theta) \to \text{argmin}(V)$, where

$$V(u) = -2u^\top W + u^\top C u + \lambda_0 \sum_{j=1}^{p} u_j \text{sgn}(\theta_j)\mathbf{1}(\theta_j \neq 0) + |\theta_j|\mathbf{1}(\theta_j = 0).$$

Here $\mathbf{1}$ is a binary indicator function and $\text{sgn}(x) = \mathbf{1}(x \geq 0) - \mathbf{1}(x < 0)$. From this we see that if λ grows slowly relative to n, then $\widehat{\theta}(\lambda)$ has the same asymptotic behavior as $\widehat{\theta}(0)$ while if λ grows more quickly, it will have an asymptotic bias. This is further evidence for the need to investigate the asymptotic behavior of the lasso solution with cross-validated tuning parameter.

We define the predictive risk and the leave-one-out cross-validation estimator of risk to be

$$R_n(\lambda) := \frac{1}{n}\mathbb{E}\|\mathbb{X}(\widehat{\theta}(\lambda) - \theta)\|^2 + \sigma^2 = \mathbb{E}\|\widehat{\theta}(\lambda) - \theta)\|^2_{C_n} + \sigma^2 \tag{2.3}$$

and

$$\widehat{R}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^\top \widehat{\theta}^{(i)}(\lambda))^2, \tag{2.4}$$

respectively. Throughout this paper, we assume that $W_i \sim P_i$ are independently distributed. We define $\mathbb{P} = \prod_i P_i$ to be the n-fold product distribution of the W_i's and use $\mathbb{E}$ to denote the expected value with respect to this product measure. In Eq. (2.4) we are using $\widehat{\theta}^{(i)}(\lambda)$ to indicate the lasso estimator $\widehat{\theta}(\lambda)$ computed using all but the i^{th} observation.

Lastly, let Λ be a large, compact subset of $[0, \infty)$ the specifics of which are unimportant. In practical situations, any $\lambda \in [\max_j |\widehat{\theta}_j(0)|, \infty)$ will result in the same solution, namely $\widehat{\theta}_j(\lambda) = 0$ for all j, so any large finite upper bound is sufficient. Then define

$$\widehat{\lambda}_n := \underset{\lambda\in\Lambda}{\text{argmin}}\, \widehat{R}_n(\lambda), \qquad \text{and} \qquad \lambda_n := \underset{\lambda\in\mathbb{R}^+}{\text{argmin}}\, R_n(\lambda),$$

where $\mathbb{R}^+ = [0, \infty)$. For $\widehat{\theta}(\lambda)$ to be consistent, it must hold that $\lambda \to 0$ as $n \to \infty$. Hence, for some N, $n \geq N$ implies $\lambda_n \in \Lambda \subset \mathbb{R}^+$. Therefore, without loss of generality, we assume that $\lambda_n \in \Lambda$ for all n.

We spend the balance of this paper discussing and proving the following result:

Theorem 2.1 (Main Theorem) *Suppose that Assumptions A and B hold and that there exists a $C_\theta < \infty$ such that $||\theta||_1 \leq C_\theta$. Let W_i be independently distributed such that $\mathbb{E}[W_i] = 0$, $\mathbb{E}[W_i]^2 = 1$, and $\mathbb{E}[W_i^4] \leq C_W$, for some constant $C_W < \infty$ independent of i. Then*

$$R_n(\widehat{\lambda}_n) - R_n(\lambda_n) \to 0. \tag{2.5}$$

Essentially, this result states that under some conditions on the design matrix $\mathbb{X}$ and the noise vector W, the predictive risk of the lasso estimator with tuning parameter chosen via cross-validation converges to the predictive risk of the lasso estimator with the oracle tuning parameter. In other words, the typical procedure for a data analyst is asymptotically equivalent (in terms of predictive risk) to the optimal procedure.

To prove this theorem, we show that $\sup_{\lambda\in\Lambda} |\widehat{R}_n(\lambda) - R_n(\lambda)| \to 0$ in probability. Then Eq. (2.5) follows as

$$\begin{aligned} R_n(\widehat{\lambda}_n) - R_n(\lambda_n) &= \left(R_n(\widehat{\lambda}_n) - \widehat{R}_n(\widehat{\lambda}_n)\right) + \left(\widehat{R}_n(\widehat{\lambda}_n) - R_n(\lambda_n)\right) \\ &\leq \left(R_n(\widehat{\lambda}_n) - \widehat{R}_n(\widehat{\lambda}_n)\right) + \left(\widehat{R}_n(\lambda_n) - R_n(\lambda_n)\right) \\ &\leq 2\sup_{\lambda\in\Lambda}\left(R_n(\lambda) - \widehat{R}_n(\lambda)\right) \\ &= o_{\mathbb{P}}(1). \end{aligned}$$

In fact, the term $R_n(\widehat{\lambda}_n) - R_n(\lambda_n)$ is non-stochastic (the expectation in the risk integrates out the randomness in the data) and therefore convergence in probability implies sequential convergence and hence $o_{\mathbb{P}}(1) = o(1)$.

We can write

$$
\begin{aligned}
&|R_n(\lambda) - \widehat{R}_n(\lambda)| \\
&\quad = \left| \frac{1}{n}\mathbb{E}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2 + \frac{1}{n}||\mathbb{X}\theta||^2 - \frac{1}{n}2\mathbb{E}(\mathbb{X}\widehat{\theta}(\lambda))^\top \mathbb{X}\theta + \sigma^2 \right. \\
&\qquad \left. - \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^2 + (X_i^\top \widehat{\theta}^{(i)}(\lambda))^2 - 2Y_i X_i^\top \widehat{\theta}^{(i)}(\lambda)\right)\right| \\
&\leq \underbrace{\left|\frac{1}{n}\mathbb{E}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2 - \frac{1}{n}\sum_{i=1}^{n}(X_i^\top \widehat{\theta}^{(i)}(\lambda))^2\right|}_{(a)} + 2\underbrace{\left|\frac{1}{n}\mathbb{E}(\mathbb{X}\widehat{\theta}(\lambda))^\top \mathbb{X}\theta - \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^\top \widehat{\theta}^{(i)}(\lambda)\right|}_{(b)} \\
&\qquad + \underbrace{\left|\frac{1}{n}||\mathbb{X}\theta||^2 + \sigma^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i^2\right|}_{(c)}. \qquad (2.6)
\end{aligned}
$$

Our proof follows by addressing (a), (b), and (c) in lexicographic order in Sect. 4. To show that each term converges in probability to zero uniformly in λ, we will need a few preliminary results.

3 Preliminary material

In this section, we present some definitions and lemmas which are useful for proving risk consistency of the lasso with cross-validated tuning parameter. First, we give some results regarding the uniform convergence of measurable functions. Then, we use these results to show that the leave-one-out lasso estimator converges uniformly to the full-sample lasso estimator.

3.1 Equicontinuity

Our proof of Theorem 2.1 uses a number of results relating uniform convergence with convergence in probability. The essential message is that particular measurable functions behave nicely over compact sets. Mathematically, such collections of functions are called *stochastically equicontinuous*.

To fix ideas, we first present the definition of stochastic equicontinuity in the context of statistical estimation. Suppose that we are interested in estimating some functional of a parameter β, $\overline{Q}_n(\beta)$, using $\widehat{Q}_n(\beta)$ where $\beta \in \mathcal{B}$.

Definition 3.1 (*Stochastic equicontinuity*) If for every $\varepsilon, \eta > 0$ there exists a random variable $\Delta_n(\varepsilon, \eta)$ and constant $n_0(\varepsilon, \eta)$ such that for $n \geq n_0(\varepsilon, \eta)$, $\mathbb{P}(|\Delta_n(\varepsilon, \eta)| > \varepsilon) < \eta$ and for each $\beta \in \mathcal{B}$ there is an open set $\mathcal{N}(\beta, \varepsilon, \eta)$ containing β such that for $n \geq n_0(\varepsilon, \eta)$,

$$\sup_{\beta' \in \mathcal{N}(\beta,\varepsilon,\eta)} \left|\widehat{Q}_n(\beta') - \widehat{Q}_n(\beta)\right| \leq \Delta_n(\varepsilon, \eta),$$

then we call $\{\widehat{Q}_n\}$ *stochastically equicontinuous over* $\mathcal{B}$.

An alternative formulation of stochastic equicontinuity which is often more useful can be found via a Lipschitz-type condition.

Theorem 3.2 (Theorem 21.10 in Davidson (1994)) *Suppose there exists a random variable B_n and a function h such that $B_n = O_{\mathbb{P}}(1)$ and for all $\beta', \beta \in \mathcal{B}$, $|\widehat{Q}_n(\beta') - \widehat{Q}_n(\beta)| \leq B_n h(d(\beta', \beta))$, where $h(x) \downarrow 0$ as $x \downarrow 0$ and d is a metric on $\mathcal{B}$ and the downward facing arrow indicates convergence from above. Then $\{\widehat{Q}_n\}$ is stochastically equicontinuous.*

The importance of stochastic equicontinuity is in showing uniform convergence, as is expressed in the following two results.

Theorem 3.3 (Theorem 2.1 in Newey (1991)) *If $\mathcal{B}$ is compact, $|\widehat{Q}_n(\beta) - \overline{Q}_n(\beta)| = o_{\mathbb{P}}(1)$ for each $\beta \in \mathcal{B}$, $\{\widehat{Q}_n\}$ is stochastically equicontinuous over $\mathcal{B}$, and $\{\overline{Q}_n\}$ is equicontinuous, then* $\sup_{\beta\in\mathcal{B}} |\widehat{Q}_n(\beta) - \overline{Q}_n(\beta)| = o_{\mathbb{P}}(1)$.

This theorem allows us to show uniform convergence of estimators $\widehat{Q}_n(\beta)$ of statistical functionals to $\overline{Q}_n(\beta)$ over compact sets $\mathcal{B}$. However, we may also be interested in the uniform convergence of random quantities to each other. While one could use the above theorem to show such a result, the following theorem of Davidson (1994) is often simpler.

Theorem 3.4 (Davidson (1994)) *If $\mathcal{B}$ is compact, then* $\sup_{\beta\in\mathcal{B}} G_n(\beta) = o_{\mathbb{P}}(1)$ *if and only if $G_n(\beta) = o_{\mathbb{P}}(1)$ for each β in a dense subset of $\mathcal{B}$ and $\{G_n(\beta)\}$ is stochastically equicontinuous.*

3.2 Uniform convergence of lasso estimators

Using stochastic equicontinuity, we prove two lemmas about lasso estimators which, while intuitive, are nonetheless novel. The first shows that the lasso estimator converges uniformly over Λ to its expectation. The second shows that the lasso estimator computed using the full sample converges in probability uniformly over Λ to the lasso estimator computed with all but one observation.

Before stating our lemmas, we include without proof some standard results about uniform convergence of functions. A function $f : [a, b] \to \mathbb{R}$ has the Luzin N property if, for all $N \subset [a, b]$ that has Lebesgue measure zero, $f(N)$ has Lebesgue measure zero as well. Also, a function f is of bounded variation if and only if it can be written as $f = f_1 - f_2$ for non-decreasing functions f_1 and f_2.

Theorem 3.5 *A function f is absolutely continuous if and only if it is of bounded variation, continuous, and has the Luzin N property.*

Theorem 3.6 *If a function $f : [a, b] \to \mathbb{R}$ is absolutely continuous, and hence differentiable almost everywhere, and satisfies $|f'(x)| \leq C_L$ for almost all $x \in [a, b]$ with respect to Lebesgue measure, then it is Lipschitz continuous with constant C_L.*

Throughout this paper, we use C_L as generic notation for a Lipschitz constant; its actual value changes from line to line. The following result is useful for showing the uniform convergence of $\widehat{\theta}(\lambda)$.

Proposition 3.7 *The random function $\widehat{\theta}(\lambda)$ is Lipschitz continuous over Λ. That is, there exists $C_L < \infty$ such that for any $\lambda, \lambda' \in \Lambda$,*

$$\left\|\widehat{\theta}(\lambda) - \widehat{\theta}(\lambda')\right\| \leq C_L|\lambda - \lambda'|. \tag{3.1}$$

Additionally, $C_L = O(1)$ as $n \to \infty$.

Proof The solution path of the lasso is piecewise linear over λ with a finite number of 'kinks.' Using the notation developed in (Tibshirani, 2013, Sect. 3.1), over each such interval, the nonzero entries in $\widehat{\theta}(\lambda)$ behave as a linear function with slope $n(\mathbb{X}_{\mathcal{E}}^{\top}\mathbb{X}_{\mathcal{E}})^{-1}s_{\mathcal{E}}$, where $\mathcal{E} \subset \{1, \ldots, p\}$ is the set of the indices of the active variables, $s_{\mathcal{E}}$ is the vector of signs, and $\mathbb{X}_{\mathcal{E}}$ is the feature matrix with columns restricted to the indices in $\mathcal{E}$.

Therefore, as $\left|\left|n(\mathbb{X}_{\mathcal{E}}^{\top}\mathbb{X}_{\mathcal{E}})^{-1}s_{\mathcal{E}}\right|\right| \leq \left|\left|n(\mathbb{X}_{\mathcal{E}}^{\top}\mathbb{X}_{\mathcal{E}})^{-1}\right|\right|$, $\widehat{\theta}(\lambda)$ is Lipschitz continuous with

$$C_L = \max_{\mathcal{E}\subset\{1,\ldots,p\}} \left|\left|n(\mathbb{X}_{\mathcal{E}}^{\top}\mathbb{X}_{\mathcal{E}})^{-1}\right|\right|$$

By Assumption A, for any $\mathcal{E}$, $\frac{1}{n}\mathbb{X}_{\mathcal{E}}^{\top}\mathbb{X}_{\mathcal{E}} \to C_{\mathcal{E}}$. Also, $\text{eigen}_{\min}(C_{\mathcal{E}}) \geq c_{\min}$ for any $\mathcal{E}$. Fix $\epsilon = c_{\min}/2$. Then, there exists an N such that for all $n \geq N$ and any $\mathcal{E}$,

$$\frac{1}{n}\text{eigen}_{\min}(\mathbb{X}_{\mathcal{E}}^{\top}\mathbb{X}_{\mathcal{E}}) \geq \epsilon. \tag{3.2}$$

Therefore, for n large enough, $C_L \leq \frac{1}{\epsilon} < \infty$, which is independent of n. □

Lemma 3.8 *For any i,*

$$\sup_{\lambda\in\Lambda} \left|\left|\widehat{\theta}(\lambda) - \widehat{\theta}^{(i)}(\lambda)\right|\right| \xrightarrow{\mathbb{P}} 0.$$

Proof The pointwise convergence of $\|\widehat{\theta}(\lambda) - \widehat{\theta}^{(i)}(\lambda)\|_2$ to zero follows by (Fu and Knight, 2000, Theorem 1). Hence, we invoke the consequent of Theorem 3.4 as long as $\|\widehat{\theta}(\lambda) - \widehat{\theta}^{(i)}(\lambda)\|_2$ is stochastically equicontinuous. For this, it is sufficient to show that $\widehat{\theta}(\lambda)$ and $\widehat{\theta}^{(i)}(\lambda)$ are Lipschitz in the sense of Theorem 3.2. This follows for both estimators by Proposition 3.7. □

Lemma 3.9 *For all $1 \leq j \leq p$, $\{\widehat{\theta}_j(\lambda)\}$ is stochastically equicontinuous, $\{\mathbb{E}[\widehat{\theta}_j(\lambda)]\}$ is equicontinuous, and $|\widehat{\theta}_j(\lambda) - \mathbb{E}\widehat{\theta}_j(\lambda)| = o_{\mathbb{P}}(1)$. Thus,*

$$\sup_{\lambda\in\Lambda} |\widehat{\theta}_j(\lambda) - \mathbb{E}\widehat{\theta}_j(\lambda)| = o_{\mathbb{P}}(1).$$

Furthermore,

$$\sup_{\lambda\in\Lambda} \left|\left|\widehat{\theta}(\lambda) - \mathbb{E}\widehat{\theta}(\lambda)\right|\right|_{C_n}^2 = o_{\mathbb{P}}(1),$$

where this notation is introduced in Eq. (2.3).

Proof To show this claim, we use Theorem 3.3. For pointwise convergence, note that $\widehat{\theta}(\lambda)$ converges in probability to an non-stochastic limit (Fu and Knight, 2000, Theorem 1), call it $\theta(\lambda)$. Also, $|\widehat{\theta}_j(\lambda)| \leq \left|\left|\widehat{\theta}(0)\right|\right|_1$, which is integrable. By the Skorohod representation theorem, there exists random variables $\widehat{\theta}_j(\lambda)'$ such that $\widehat{\theta}_j(\lambda)' \to \theta(\lambda)$ almost surely and $\widehat{\theta}_j(\lambda)'$ has the same distribution as $\widehat{\theta}_j(\lambda)$ for each n. By the dominated convergence theorem,

$$\lim \mathbb{E}\widehat{\theta}_j(\lambda) = \lim \mathbb{E}\widehat{\theta}_j(\lambda)' = \mathbb{E}\theta(\lambda) = \theta(\lambda).$$

Therefore, $|\widehat{\theta}_j(\lambda) - \mathbb{E}\widehat{\theta}_j(\lambda)| \to 0$ in probability.

Stochastic equicontinuity follows by Proposition 3.7 and Theorem 3.2. Hence, Theorem 3.3 is satisfied as long as $\{\mathbb{E}\widehat{\theta}_j(\lambda)\}$ is equicontinuous. Observe that the expectation and differentiation operations commute for $\widehat{\theta}(\lambda)$. Therefore, the result follows by Proposition 3.7.

Finally, we have

$$\begin{aligned}\left|\left|\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda)\right|\right|_{C_n}^2 &= (\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda))^\top C_n(\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda))\\ &\le \left|\left|\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda)\right|\right|\left|\left|C_n(\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda))\right|\right|\\ &\le \left|\left|\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda)\right|\right|^2||C_n||\\ &= ||C_n||\sum_{j=1}^{p}|\widehat{\theta}_j(\lambda)-\mathbb{E}\widehat{\theta}_j(\lambda)|^2,\end{aligned}$$

which goes to zero uniformly, as $||C_n|| \to ||C|| < \infty$ □

4 Proofs

In this section, we address each component of the decomposition in (2.6). Parts (a) and (b) follow from uniform convergence of the lasso estimator to its expectation (Lemma 3.9) and asymptotic equivalence of the leave-one-out lasso estimator and the full-sample lasso estimator (Lemma 3.8) while part (c) uses Markov's inequality.

Proposition 4.1 (Part (a))

$$\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\mathbb{E}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\frac{1}{n}\sum_{i=1}^{n}\left(X_i^\top\widehat{\theta}^{(i)}(\lambda)\right)^2\right| = o_{\mathbb{P}}(1).$$

Proof Observe

$$\begin{aligned}&\left|\frac{1}{n}\mathbb{E}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\widehat{\theta}^{(i)}(\lambda))^2\right|\\ &\le \underbrace{\left|\frac{1}{n}\mathbb{E}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\frac{1}{n}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2\right|}_{(ai)}+\underbrace{\left|\frac{1}{n}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\widehat{\theta}^{(i)}(\lambda))^2\right|}_{(aii)}\end{aligned}$$

For (ai), note that $\mathbb{E}\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2 = \text{trace}(\mathbb{X}^\top\mathbb{X}\mathbb{V}\widehat{\theta}(\lambda)) + \left|\left|\mathbb{X}\mathbb{E}\widehat{\theta}(\lambda)\right|\right|^2$, where $\mathbb{V}\widehat{\theta}(\lambda) = \mathbb{E}\left[(\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda))(\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda))^\top\right]$ is the variance matrix of $\widehat{\theta}(\lambda)$. Hence,

$$\begin{aligned}(ai) &\le \left|\text{trace}(C_n\mathbb{V}\widehat{\theta}(\lambda))\right|+\frac{1}{n}\left|\left|\mathbb{E}\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2\\ &\le ||C_n||_F\left|\left|\mathbb{V}\widehat{\theta}(\lambda)\right|\right|_F+\frac{1}{n}\left|\left|\mathbb{E}\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2\\ &\le ||C_n||_F\left|\left|\mathbb{V}\widehat{\theta}(0)\right|\right|_F+\frac{1}{n}\left|\left|\mathbb{E}\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2\\ &= \sigma^2||C_n||_F\left|\left|(\mathbb{X}^\top\mathbb{X})^{-1}\right|\right|_F+\frac{1}{n}\left|\left|\mathbb{E}\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2-\left|\left|\mathbb{X}\widehat{\theta}(\lambda)\right|\right|^2\\ &\le \frac{\sigma^2}{n}||C_n||_F\left|\left|C_n^{-1}\right|\right|_F+\left|\left|\widehat{\theta}(\lambda)+\mathbb{E}\widehat{\theta}(\lambda)\right|\right|_{C_n}\left|\left|\widehat{\theta}(\lambda)-\mathbb{E}\widehat{\theta}(\lambda)\right|\right|_{C_n}.\end{aligned}$$

The third inequality follows from (Osborne et al. 2000, [Eq. 4.1]) which bounds the variance of the lasso estimator with the variance of the least-squares estimator. The last line goes to zero uniformly by Lemma 3.9.

For (aii), note that

$$\begin{aligned}\frac{1}{n}\left|\|\mathbb{X}\widehat{\theta}(\lambda)\|^2-\sum_{i=1}^{n}(X_i^\top\widehat{\theta}^{(i)}(\lambda))^2\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left((X_i^\top\widehat{\theta}(\lambda))^2-(X_i^\top\widehat{\theta}^{(i)}(\lambda))^2\right)\right| \\ &\le \frac{1}{n}\sum_{i=1}^{n}\left|(X_i^\top\widehat{\theta}(\lambda))^2-(X_i^\top\widehat{\theta}^{(i)}(\lambda))^2\right| \\ &= \frac{1}{n}\sum_{i=1}^{n}\left|X_i^\top\widehat{\theta}(\lambda)\widehat{\theta}(\lambda)^\top X_i - X_i^\top\widehat{\theta}^{(i)}(\lambda)\widehat{\theta}^{(i)}(\lambda)^\top X_i\right| \\ &= \frac{1}{n}\sum_{i=1}^{n}\left|X_i^\top\left(\widehat{\theta}(\lambda)\widehat{\theta}(\lambda)^\top-\widehat{\theta}^{(i)}(\lambda)\widehat{\theta}^{(i)}(\lambda)^\top\right)X_i\right| \\ &\le \frac{1}{n}\sum_{i=1}^{n}||X_i||^2\left|\left|\widehat{\theta}(\lambda)\widehat{\theta}(\lambda)^\top-\widehat{\theta}^{(i)}(\lambda)\widehat{\theta}^{(i)}(\lambda)^\top\right|\right|_F.\end{aligned}$$

Furthermore,

$$\left|\left|\widehat{\theta}(\lambda)\widehat{\theta}(\lambda)^\top-\widehat{\theta}^{(i)}(\lambda)\widehat{\theta}^{(i)}(\lambda)^\top\right|\right|_F \le \sum_{j=1}^{p}\left|\left|\widehat{\theta}_j(\lambda)\widehat{\theta}(\lambda)-\widehat{\theta}_j^{(i)}(\lambda)\widehat{\theta}^{(i)}(\lambda)\right|\right|.$$

Finally,

$$\begin{aligned}&\left|\left|\left(\widehat{\theta}_j(\lambda)\widehat{\theta}(\lambda)-\widehat{\theta}_j^{(i)}(\lambda)\widehat{\theta}^{(i)}(\lambda)\right)-\left(\widehat{\theta}_j(\lambda')\widehat{\theta}(\lambda')-\widehat{\theta}_j^{(i)}(\lambda')\widehat{\theta}^{(i)}(\lambda')\right)\right|\right| \\ &\le |\widehat{\theta}_j(\lambda)|\left|\left|\widehat{\theta}(\lambda)-\widehat{\theta}(\lambda')\right|\right|+\left|\widehat{\theta}_j(\lambda)-\widehat{\theta}_j(\lambda')\right|\left|\left|\widehat{\theta}(\lambda')\right|\right|+ \\ &\quad\left|\widehat{\theta}_j^{(i)}(\lambda)\right|\left|\left|\widehat{\theta}^{(i)}(\lambda')-\widehat{\theta}^{(i)}(\lambda)\right|\right|+\left|\widehat{\theta}_j^{(i)}(\lambda')-\widehat{\theta}_j^{(i)}(\lambda)\right|\left|\left|\widehat{\theta}^{(i)}(\lambda')\right|\right| \\ &\le |\widehat{\theta}_j(0)|\left|\left|\widehat{\theta}(\lambda)-\widehat{\theta}(\lambda')\right|\right|+\left|\widehat{\theta}_j(\lambda)-\widehat{\theta}_j(\lambda')\right|\left|\left|\widehat{\theta}(0)\right|\right|_1+ \\ &\quad\left|\widehat{\theta}_j^{(i)}(0)\right|\left|\left|\widehat{\theta}^{(i)}(\lambda')-\widehat{\theta}^{(i)}(\lambda)\right|\right|+\left|\widehat{\theta}_j^{(i)}(\lambda')-\widehat{\theta}_j^{(i)}(\lambda)\right|\left|\left|\widehat{\theta}^{(i)}(0)\right|\right|_1.\end{aligned}$$

The least squares solution is $O_p(1)$, so Proposition 3.7 and Theorem 3.2 imply that $\left|\left|\widehat{\theta}(\lambda)\widehat{\theta}(\lambda)^\top-\widehat{\theta}^{(i)}(\lambda)\widehat{\theta}^{(i)}(\lambda)^\top\right|\right|_F$ is stochastically equicontinuous and, since $||X_i||^2 \le C_X^2$ by Assumption B, Theorem 3.3 implies that (aii) goes to zero uniformly in probability over Λ. □

Proposition 4.2 (Part (b))

$$\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\mathbb{E}(\mathbb{X}\widehat{\theta}(\lambda))^\top\mathbb{X}\theta-\frac{1}{n}\sum_{i=1}^{n}Y_iX_i^\top\widehat{\theta}^{(i)}(\lambda)\right| = o_{\mathbb{P}}(1).$$

Proof Observe,

$$\sum_{i=1}^{n}Y_iX_i^\top\widehat{\theta}^{(i)}(\lambda) = \sum_{i=1}^{n}(X_i^\top\theta+\sigma^2W_i)(X_i^\top\widehat{\theta}^{(i)}(\lambda)) \tag{4.1}$$

$$= \sum_{i=1}^{n}X_i^\top\theta X_i^\top\widehat{\theta}^{(i)}(\lambda)+\sum_{i=1}^{n}\sigma^2W_iX_i^\top\widehat{\theta}^{(i)}(\lambda). \tag{4.2}$$

So,

$$\begin{aligned}
&\left|\frac{1}{n}\mathbb{E}(\mathbb{X}\widehat{\theta}(\lambda))^\top\mathbb{X}\theta - \frac{1}{n}\sum_{i=1}^n Y_i X_i^\top\widehat{\theta}^{(i)}(\lambda)\right| \\
&\leq \left|\mathbb{E}\widehat{\theta}(\lambda)^\top C_n\theta - \widehat{\theta}(\lambda)^\top C_n\theta\right| + \left|\widehat{\theta}(\lambda)^\top C_n\theta - \frac{1}{n}\sum_{i=1}^n Y_i X_i^\top\widehat{\theta}^{(i)}(\lambda)\right| \\
&= \left|(\mathbb{E}\widehat{\theta}(\lambda) - \widehat{\theta}(\lambda))^\top C_n\theta\right| + \left|\widehat{\theta}(\lambda)^\top C_n\theta - \frac{1}{n}\sum_{i=1}^n Y_i X_i^\top\widehat{\theta}^{(i)}(\lambda)\right| \\
&\leq \underbrace{\left|\left|\mathbb{E}\widehat{\theta}(\lambda) - \widehat{\theta}(\lambda)\right|\right|_{C_n}||\theta||_{C_n}}_{(bi)} + \underbrace{\left|\frac{1}{n}\widehat{\theta}(\lambda)^\top\mathbb{X}^\top\mathbb{X}\theta - \frac{1}{n}\sum_{i=1}^n X_i^\top\theta X_i^\top\widehat{\theta}^{(i)}(\lambda)\right|}_{(bii)} + \\
&\quad + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \sigma^2 W_i X_i^\top\widehat{\theta}^{(i)}(\lambda)\right|}_{(biii)}.
\end{aligned}$$

By Lemma 3.9, (bi) goes to zero uniformly. For (bii),

$$\begin{aligned}
\frac{1}{n}\left|\widehat{\theta}(\lambda)^\top\mathbb{X}^\top\mathbb{X}\theta - \sum_{i=1}^n X_i^\top\theta X_i^\top\widehat{\theta}^{(i)}(\lambda)\right| &= \frac{1}{n}\left|\sum_{i=1}^n \theta^\top X_i X_i^\top\left(\widehat{\theta}(\lambda) - \widehat{\theta}^{(i)}(\lambda)\right)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n\left(||\theta||||X_i||^2\left|\left|\widehat{\theta}(\lambda) - \widehat{\theta}^{(i)}(\lambda)\right|\right|\right) \\
&\leq C_\theta C_X^2\frac{1}{n}\sum_{i=1}^n\left|\left|\widehat{\theta}(\lambda) - \widehat{\theta}^{(i)}(\lambda)\right|\right|.
\end{aligned}$$

This goes to zero uniformly by Lemma 3.8.

For $(biii)$, $\|\widehat{\theta}^{(i)}(\lambda)\|_1 \leq \|\widehat{\theta}^{(i)}(0)\|_1$ for any λ, i. So:

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n \sigma^2 W_i X_i^\top\widehat{\theta}^{(i)}(\lambda)\right| &= \frac{\sigma^2}{n}\left|\sum_{i=1}^n W_i X_i^\top\widehat{\theta}^{(i)}(\lambda)\right| \\
&\leq \frac{\sigma^2}{n}\left|\sum_{i=1}^n W_i||X_i||_\infty\left|\left|\widehat{\theta}^{(i)}(\lambda)\right|\right|_1\right| \\
&\leq \frac{\sigma^2 C_X}{n}\left|\sum_{i=1}^n W_i\left|\left|\widehat{\theta}^{(i)}(0)\right|\right|_1\right| \xrightarrow{ae} 0.
\end{aligned}$$

The proof of almost-everywhere convergence is given in the appendix. This completes the proof of Proposition 4.2. □

Proposition 4.3 (Part (c))

$$\left|\frac{1}{n}||\mathbb{X}\theta||^2 + \sigma^2 - \frac{1}{n}\sum_{i=1}^n Y_i^2\right| = o_{\mathbb{P}}(1).$$

Proof Using Markov's inequality and the fact that for any random variable Z, $\mathbb{E}|Z| \leq \sqrt{\mathbb{E}Z^2}$, and using the notation $\mathbb{V}Z = \mathbb{E}[(Z - \mathbb{E}Z)^2]$ for the variance, we see that

$$\begin{aligned}
&\delta^2 P\left(\left|\frac{1}{n}||\mathbb{X}\theta||^2 + \sigma^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i^2\right| > \delta\right)^2 \\
&\leq \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\left[\left((X_i^\top\theta)^2 + \sigma^2 W_i^2 + 2X_i^\top\theta\sigma W_i\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\sigma^4\mathbb{V}[W_i^2] + 4(X_i^\top\theta)^2\sigma^2\mathbb{V}[W_i] + 4\sigma^3 X_i^\top\theta\mathbb{E}W_i^3\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\sigma^4(\mathbb{E}W_i^4 - 1) + 4(X_i^\top\theta)^2\sigma^2 + 4\sigma^3 X_i^\top\theta\mathbb{E}W_i^3\right) \\
&\leq \frac{1}{n}\sigma^4(C_W - 1) + \frac{4\sigma^2}{n^2}\|\mathbb{X}\theta\|^2 + \frac{4\sigma^3 C_W}{n^2}\sum_{i=1}^{n} X_i^\top\theta \\
&\frac{1}{n}\sigma^4(C_W - 1) + \frac{4\sigma^2}{n}C_X^2 C_\theta^2 + \frac{4\sigma^3 C_W C_X C_\theta}{n} \\
&= O(1/n).
\end{aligned}$$

□

5 Discussion and future work

A common practice in data analysis is to estimate the coefficients of a linear model with the lasso and choose the regularization parameter by cross-validation. Unfortunately, no definitive theoretical results existed as to the effect of choosing the tuning parameter in this data-dependent way. In this paper, we provide a solution to the problem by demonstrating, under particular assumptions on the design matrix, that the lasso is risk consistent even when the tuning parameter is selected via leave-one-out cross-validation.

However, a number of important open questions remain. The first is to generalize to other forms of cross-validation, especially K-fold. In fact, this generalization should be possible using the methods developed herein. Lemma 3.8 holds when more than one training example is held out, provided that the size of the datasets used to form the estimators still increases to infinity with n. Furthermore, with careful accounting of the held out sets, Proposition 4.2 should hold as well.

A second question is to determine whether cross-validation holds in the high-dimensional setting where $p > n$. However, our methods are not easily extensible to this setting. We rely heavily on Assumption A which says that $n^{-1}\mathbb{X}^\top\mathbb{X}$ has a positive definite limit as well as the related results of Fu and Knight (2000) which are not available in high dimensions or with random design. Additionally, an interesting relaxation of our results would be to assume that the matrices C_n are all non-singular, but tend to a singular limit. This would provide a more realistic scenario where regularization is more definitively useful.

Finally, one of the main benefits of lasso is its ability to induce sparsity and hence perform variable selection. While selecting the correct model is far more relevant in high dimensions, it may well be desirable in other settings as well. As mentioned in the introduction, various authors have shown that cross-validation and model selection are in some sense incompatible. In particular, CV is inconsistent for model selection. Secondly, using prediction accuracy (which is what $\widehat{R}_n(\lambda)$ is estimating) as the method for choosing λ fails to recover the sparsity pattern even under orthogonal design. Thus, while we show that the predictions of the model are asymptotically equivalent to those with the optimal tuning parameter, we should not expect to have the correct model even if θ were sparse. In particular, $\widehat{\theta}(\lambda)$ does not necessarily converge to the OLS estimator, and may not converge to θ. We do show (Lemma 3.9) that $\widehat{\theta}(\lambda)$ converges to its expectation uniformly for all λ. While this expectation may be sparse, it may not be. But we are unable to show that with cross-validated tuning parameter, the lasso will select the *correct* model. While this is not surprising in light of previous research, neither is it comforting. The question of whether lasso with cross-validated tuning parameter can recover an unknown sparsity pattern remains open. Empirically, our experience is that cross-validated tuning parameters lead to over-parameterized estimated models, but this has yet to be validated theoretically.

6 Supplementary results

We state here the proof of Proposition 4.2.

Proof (Almost everywhere convergence of $(biii)$)

$$\begin{aligned}\frac{\sigma^2 C_X}{n}\left|\sum_{i=1}^n W_i\left|\left|\widehat{\theta}^{(i)}(0)\right|\right|_1\right| &\leq \frac{\sigma^2 C_X}{n}\left|\sum_{i=1}^n W_i(\left|\left|\widehat{\theta}^{(i)}(0)-\theta\right|\right|_1+||\theta||_1)\right| \\ &\leq \frac{\sigma^2 C_X}{n}\left(\left|\sum_{i=1}^n W_i\left|\left|\widehat{\theta}^{(i)}(0)-\theta\right|\right|_1\right|+\left|C_\theta\sum_{i=1}^n W_i\right|\right)\end{aligned}$$

The second term goes to zero in probability by the strong law of large numbers. For the first term, define $C_{i,n}=(n-||X_i||_2^2)^{-1}$, then

$$\begin{aligned}\widehat{\theta}^{(i)}(0) &= (\mathbb{X}_{(i)}^\top\mathbb{X}_{(i)})^{-1}\mathbb{X}_{(i)}^\top Y_{(i)} \\ &= \widehat{\theta}(0)-C_{i,n}X_iY_i+C_{i,n}X_iX_i^\top\widehat{\theta}(0) \\ &= \frac{1}{n}(C_{i,n}X_iX_i^\top+I)\mathbb{X}^\top\mathbb{X}\theta+\frac{1}{n}(C_{i,n}X_iX_i^\top+I)\mathbb{X}^\top W \\ &\quad -C_{i,n}X_iX_i^\top\theta-C_{i,n}X_iW_i \\ &= \theta+\frac{1}{n}(C_{i,n}X_iX_i^\top+I)\mathbb{X}^\top W-C_{i,n}X_iW_i.\end{aligned}$$

So,

$$\begin{aligned}&\left|\left|\widehat{\theta}^{(i)}(0)-\theta\right|\right|_1 \\ &=\left|\left|\frac{1}{n}(C_{i,n}X_iX_i^\top+I)\mathbb{X}^\top W-C_{i,n}X_iW_i\right|\right|_1 \\ &=\left|\left|\frac{1}{n}C_{i,n}X_iX_i^\top\mathbb{X}_{(i)}^\top W_{(i)}+\frac{1}{n}C_{i,n}X_iX_i^\top X_iW_i+\frac{1}{n}\mathbb{X}^\top W-C_{i,n}X_iW_i\right|\right|_1\end{aligned}$$

$$= \frac{1}{n}\left|\left|C_{i,n}X_i X_i^\top \mathbb{X}_{(i)}^\top W_{(i)} - X_i W_i + \mathbb{X}^\top W\right|\right|_1$$
$$= \frac{1}{n}\left|\left|C_{i,n}X_i X_i^\top \mathbb{X}_{(i)}^\top W_{(i)} + \mathbb{X}_{(i)}^\top W_{(i)}\right|\right|_1.$$

Therefore,

$$\frac{\sigma^2 C_X}{n}\left|\sum_{i=1}^n W_i\left|\left|\widehat{\theta}^{(i)}(0)-\theta\right|\right|_1\right|$$
$$= \frac{\sigma^2 C_X}{n^2}\left|\sum_{i=1}^n W_i\left|\left|C_{i,n}X_i X_i^\top \mathbb{X}_{(i)}^\top W_{(i)} + \mathbb{X}_{(i)}^\top W_{(i)}\right|\right|_1\right|$$
$$= \frac{\sigma^2 C_X}{n^2}\left|\sum_{i=1}^n W_i \sum_{k=1}^p \left|C_{i,n}\mathbb{X}_{ik}\sum_{\ell=1}^p \mathbb{X}_{i\ell}\sum_{j\neq i}\mathbb{X}_{j\ell}W_j + \sum_{j\neq i}\mathbb{X}_{jk}W_j\right|\right|$$
$$= \frac{\sigma^2 C_X}{n^2}\left|\sum_{i=1}^n W_i \sum_{k=1}^p \left|\sum_{j\neq i} W_j\left(C_{i,n}\mathbb{X}_{ik}\sum_{\ell=1}^p \mathbb{X}_{i\ell}\mathbb{X}_{j\ell} + \mathbb{X}_{jk}\right)\right|\right|$$
$$\leq \frac{\sigma^2 C_X^3}{n^2}\left|\sum_{i=1}^n W_i \sum_{k=1}^p \left|\sum_{j\neq i} W_j C_{i,n}\mathbb{X}_{ik}\right|\right| + \frac{\sigma^2 C_X}{n^2}\left|\sum_{i=1}^n W_i \sum_{k=1}^p \left|\sum_{j\neq i} W_j \mathbb{X}_{jk}\right|\right|$$
$$\leq \frac{\sigma^2 C_X^4 C_n^*}{n^2}\left|\sum_{i=1}^n W_i\left|\sum_{j\neq i} W_j\right|\right| + \frac{\sigma^2 C_X^2}{n^2}\left|\sum_{i=1}^n W_i\left|\sum_{j\neq i} W_j\right|\right|,$$

where $C_n^* := (n - \max_i ||X_i||_2^2)^{-1} = \max_i C_{i,n}$. To bound $\frac{1}{n^2}\left|\sum_{i=1}^n W_i \sum_{j\neq i} W_j\right|$, observe

$$\frac{1}{n^2}\left|\sum_{i=1}^n W_i\left|\sum_{j=1}^n W_j - W_i\right|\right| \leq \frac{1}{n^2}\left|\sum_{i=1}^n W_i\left(\left|\sum_{j=1}^n W_j\right| + |W_i|\right)\right|$$
$$\leq \frac{1}{n^2}\left|\sum_{i=1}^n W_i|W_i|\right| + \frac{1}{n^2}\left|\sum_{i=1}^n W_i\left|\sum_{j=1}^n W_j\right|\right|$$
$$\leq \frac{1}{n^2}\left|\sum_{i=1}^n W_i|W_i|\right| + \frac{1}{n^2}\sum_{i=1}^n |W_i|\left|\sum_{j=1}^n W_j\right|$$
$$= \frac{1}{n}\left|\frac{1}{n}\sum_{i=1}^n W_i|W_i|\right| + \frac{1}{n}\sum_{i=1}^n |W_i|\left|\frac{1}{n}\sum_{j=1}^n W_j\right| \xrightarrow{ae} 0.$$

□

References

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, *37*(4), 1705–1732.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, *2*, 499–526.

Bunea, F., Tsybakov, A., & Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, *1*, 169–194.
Chatterjee, A., & Lahiri, S. (2011). Strong consistency of lasso estimators. *Sankhya A-Mathematical Statistics and Probability*, *73*(1), 55–78.
Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*(1), 33–61.
Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford: Oxford university press.
Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.
Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378.
van de Geer, S., & Lederer, J. (2013). The Lasso, correlated design, and improved oracle inequalities. (2011). http://arxiv.org/abs/1107.0189
Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In ICANN 98 (pp. 201-206). London: Springer
Greenshtein, E., & Ritov, Y. A. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, *10*(6), 971–988.
Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression, Verlag: Springer.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Verlag: Springer.
Lee, S., Zhu, J., & Xing, E. P. (2010). Adaptive multi-task Lasso: With application to eQTL detection. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.) Advances in neural information processing systems, vol. 23 (pp. 1306–1314).
Leng, C., Lin, Y., & Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, *16*(4), 1273–1284.
Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.
Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, *59*(4), 1161–1167.
Osborne, M., Presnell, B., & Turlach, B. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics*, *9*(2), 319–337.
Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine Learning*, *13*, 135–143.
Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*, 486–494.
Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R., & Klein, B. (2008). LASSO-patternsearch algorithm with application to ophthalmology and genomic data. *Statistics and its Interface*, *1*(1), 137.
Stromberg, K. (1994). *Probability for analysts*. London: Chapman & Hall.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288.
Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282.
Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, *7*, 1456–1490.
Tibshirani, R. J., & Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, *40*, 1198–1232.
Wang, H., & Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, *102*(479), 1039–1048.
Xu, H., Mannor, S., & Caramanis, C. (2008). Sparse algorithms are not stable: A no-free-lunch theorem. In: Proceedings of the IEEE 46th Annual Allerton Conference on Communication, Control, and Computing, (pp. 1299–1303).
Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, *35*(5), 2173–2192.