Risk consistency of cross-validation with lasso-type procedures

Darren Homrighausen  
Department of Statistics  
Colorado State University  
Fort Collins, CO 80523  
darrenho@stat.colostate.edu

Daniel J. McDonald  
Department of Statistics  
Indiana University  
Bloomington, IN 47408  
dajmcdon@indiana.edu

May 11, 2014

Abstract

The lasso and related procedures such as the group lasso, have been the target of a substantial amount of theoretical and applied research. Correspondingly, many results are known about their behavior for a fixed or optimally chosen tuning parameter specified up to unknown constants. In practice however, this oracle tuning parameter is inaccessible, so one must instead use the data to choose a tuning parameter. Common statistical practice is to use one of a few variants of cross-validation for this task. However, very little is known about the theoretical properties of the resulting linear model using these data-dependent methods. We consider the high-dimensional setting with random design wherein the number of predictors $p = n^\alpha$, $\alpha > 0$ grows with the number of observations. We show that the lasso and group lasso estimators remain risk consistent relative to their linear oracles even when the tuning parameter is chosen via cross-validation rather than optimally.

Keywords: persistence, linear oracle, high-dimensional, regularized

1 Introduction

Since its introduction in the statistical [21] and signal processing [2] communities, $\ell_1$-regularized linear regression has become a fixture as both a data analysis tool and as a subject for deep theoretical investigations. In particular, for a response vector $Y \in \mathbb{R}^n$, design matrix $X \in \mathbb{R}^{n \times p}$, and tuning parameter $t$, we consider the lasso problem of finding

$$\hat{\beta}_t := \hat{\beta}(B_t) \in \arg\min_{\beta \in B_t} \frac{1}{n} ||Y - X\beta||_2^2$$  (1)

where $B_t := \{\beta : ||\beta||_1 \leq t\}$ and $||\cdot||_2$ and $||\cdot||_1$ indicate the Euclidean and $\ell_1$-norms respectively. By convexity, for each $t$, there is always at least one solution in equation (1). While it is true that the solution is not necessarily unique if $\text{rank}(X) < p$, this detail is unimportant for our purposes and we abuse notation slightly by referring to $\hat{\beta}_t$ as ‘the’ lasso solution. We consider the constrained optimization form for the lasso instead of the more common (but equivalent) Lagrangian interpretation as it leads to a more natural formulation for our results.

The literature contains numerous results regarding the statistical properties of the lasso. While it is beyond the scope of this paper to give a complete picture of, we highlight some key results here. Early results about the asymptotic distribution of the lasso solution are shown by Fu and Knight [8] under the assumption that the sample covariance matrix has a nonnegative definite limit
and \( p \) is fixed. Many authors have investigated model selection properties of the lasso—showing that when the best predicting model is linear and sparse (i.e. the response is only a function of a handful of predictors) the lasso will tend to asymptotically recover those predictors. Donoho et al. [3], Meinshausen and Bühlmann [13], Meinshausen and Yu [14], Wainwright [26], Zhao and Yu [30] and Zou [31] have all considered this setting under sparsity and various “irrepresentability” conditions which ensure that the relevant predictors are not too highly correlated with irrelevant ones.

Theoretical results such as these and others, depend critically on the choice of tuning parameters and are typically of the form: if \( t = t_n \) is such that \( t_n = o(b_n) \) for some rate \( b_n \), such as \( b_n = (n/\log(n))^{1/4} \), then \( \hat{\beta}_{t_n} \) is consistent in some sense. However comforting results of this type are, this theoretical guidance says little about the properties of the lasso when the tuning parameter is chosen in a data-dependent way.

The criterion we focus on for this paper is risk consistency, (alternatively known as persistence). That is, we investigate the difference between the predictive risk of the lasso estimator with tuning parameter estimated by cross-validation and the risk of the best linear oracle predictor (with oracle tuning parameter). Risk consistency of lasso has previously been investigated by Bunea et al. [1], Greenshtein and Ritov [9], and van de Geer [24]. These results, in contrast to this paper, assume that the tuning parameter is selected in an oracle fashion.

There are several proposed techniques for choosing \( t \), or equivalently, the parameter in the Lagrangian formulation, commonly notated as \( \lambda \). Tibshirani and Taylor [23] and Zou et al. [32] investigate using the “degrees of freedom” of a lasso solution, which can be defined informally as the trace of the covariance between the lasso solution and the response \( Y \). Then, \( t \) can be selected by minimizing the empirical risk penalized by the degrees of freedom. Another risk estimator is the adapted Bayesian information criterion of Wang and Leng [27], which uses a plug-in estimator of the second-order Taylor’s expansion of the risk.

However, in many papers [6, 7, 9, 21, 22, 32] and in the book [25, Sections 2.4.1], the recommended technique for selecting \( t \) in the lasso problem is to choose \( t = \hat{t} \) such that \( \hat{t} \) minimizes a \( K \)-fold cross-validation estimator of the risk (see Section 2 for the precise definition), which we will refer to as CV.

The main contribution of this paper is to show that the use of cross-validation to choose the tuning parameter in lasso remains persistent relative to the theoretically optimal, but empirically unavailable, non-stochastic oracle choice. We consider the high-dimensional regime where \( p_n = n^\alpha \), for a positive \( \alpha \) that is to be discussed in Section 3.

Some results supporting the use of cross-validation for statistical methods other than lasso are known. For instance, [19, 20] outlines various conditions under which cross-validated methods can perform good predictions. More recently, [4] find finite sample bounds for using various cross-validation procedures. These results do not address the lasso nor parameter spaces with increasing dimensions. Additionally, [12] provide oracle inequalities related to using cross-validation with lasso. However, [12] treats the problem as aggregating a dictionary of lasso estimators with different tuning parameters under various regularity conditions which do not apply to the results of this paper.

The supporting theory for non-lasso methods suggests that there should be corresponding theory for the lasso. However, other results are not so encouraging. In particular, Shao [18] shows that cross-validation is inconsistent for model selection. As lasso implicitly does model selection, and shares many connections with forward stagewise regression [6], this raises a concerning possibility that lasso might similarly be inconsistent under cross-validation. Additionally, various results showing convergence of cross-validation for algorithmically stable procedures exist [11, for example]. However, Xu et al. [28] show that sparsity inducing algorithms like lasso are not algorithmically stable.
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 indicate that cross-validation may not be a sound method for selecting the tuning parameter at all. In particular, when using the CV-selected tuning parameter, one should not expect to recover the correct sparsity pattern, but maybe one can still hope to predict well.

In this paper, we show that the lasso under random design with cross-validated tuning parameter is indeed risk consistent under some conditions on the joint distribution of the design that generates X and the response Y. Additionally, we demonstrate that our framework can be used to show similar results for the group lasso.

In Section 2, we outline the mathematical setup for the lasso prediction problem and discuss some empirical concerns. Section 3 contains the main result and associated conditions. Section 5 presents some useful lemmas and provides the proof of our results, while Section 6 summarizes our contribution and presents some avenues for further research.

2 Notation and definitions

In this section, we present the setup of our problem, describing the lasso procedure and its properties.

2.1 Preliminaries

Suppose we observe pairs \( Z_i^\top = (Y_i, X_i^\top) \) of predictor variables, \( X_i \in \mathbb{R}^{p_n} \), and response variables, \( Y_i \in \mathbb{R} \), where \( Z_i \overset{i.i.d.}{\sim} F_n \) for \( i = 1, 2, \ldots, n \) and the distribution \( F_n \) is in some class \( \mathcal{F} \) to be specified later. Here, we use the notation \( p_n \) to allow the number of predictor variables to change with \( n \). Similarly, we index the distribution \( F_n \) to emphasize its dependence on \( n \). For simplicity, in what follows, we omit the subscript \( n \) when there is little risk of confusion.

We consider the problem of estimating the best linear functional

\[
f(X_1, \ldots, X_p) = \beta^\top X
\]

for predicting \( Y \), when \( Z^\top = (Y, X^\top) \sim F_n \) is a new random variable from the same distribution as the data and \( \beta = (\beta_1, \ldots, \beta_p)^\top \). We will use zero-based indexing for \( Z \) so that \( Z_0 = Y \). To measure performance, we use the \( L^2 \)-risk of the predictor \( \beta \):

\[
R(\beta) := \mathbb{E}_{F_n} [(Y - \beta^\top X)^2].
\]

Note that the (conditional) expectation is taken only over the new datum \( Z \) and not over any observables which may or may not be used to choose \( \beta \).

Using the \( n \) independent observations \( Z_1, \ldots, Z_n \), we can form the response vector \( Y := (Y_i)_{i=1}^n \) and the design matrix \( X := [X_1, \ldots, X_n]^\top \). Then, given a vector \( \beta \), we can write the squared-error empirical risk function as

\[
\hat{R}(\beta) := \frac{1}{n} ||Y - X\beta||_2^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2.
\]

Let \( B \subseteq \mathbb{R}^p \) be any regularizing set. We define the estimator generated by \( B \) to be

\[
\hat{\beta}(B) := \arg \min_{\beta \in B} \hat{R}(\beta).
\]
We focus on two estimators in particular; the lasso, which has $\mathcal{B} = B_t = \{ \beta : ||\beta||_1 \leq t \}$ as in equation (1), and the grouped lasso, which has $\mathcal{B} = B_u = \{ \beta : \sum_{g \in G} \sqrt{|g|} ||\beta_g||_2 \leq u \}$ for a partition $G$ of $\{1, \ldots, p\}$ and $|g|$ is the size of group $g$. We refer to lasso quantities with “t’s” and grouped lasso quantities with “u’s”. As the regularization sets in this paper are always indexed by a nonnegative real parameter, we refer to a regularizing set by indexing with $\omega \in \{t, u\}$. So, for example, $\hat{\beta}_t = \hat{\beta}(B_t)$ is the grouped lasso estimator and $\hat{\beta}_u = \hat{\beta}(B_u)$ is an estimator corresponding to empirical risk minimization over a generic regularization set $B_\omega$. Though this is some abuse of notation, we feel that it adds to the overall coherence of the results.

Analogously to equation (3), we write the $K$-fold cross-validation estimator of the risk with respect to the regularization set $\mathcal{B}_\omega$, which we abbreviate to CV-risk, as

$$\hat{R}_{V_n}(\omega) = \hat{R}_{V_n}(\hat{\beta}_{\omega}^{(v_1)}, \ldots, \hat{\beta}_{\omega}^{(v_K)}) := \frac{1}{K} \sum_{v \in V_n} \frac{1}{|v|} \sum_{r \in v} (Y_r - X_r^T \hat{\beta}_{\omega}^{(v)})^2 .$$

Here, $V_n = \{v_1, \ldots, v_K\}$ is a set of validation sets, $\hat{\beta}_{\omega}^{(v)}$ is the estimator in equation (4) with the observations in the validation set $v$ removed, and $|v|$ indicates the cardinality of the set $v$. Notice in particular that the cross-validation estimator of the risk is a function of $\omega$ rather than a single predictor $\beta$. Lastly, we define the CV-risk minimizing choice of tuning parameter to be

$$\hat{\omega} = \arg\min_{\omega \in \Omega} \hat{R}_{V_n}(\omega).$$

In our setting, we will take $\Omega$ is an interval subset of the nonnegative real numbers which needs to be defined by the data-analyst. The set $\Omega$ is an important part of the performance of $\hat{\beta}_\omega$ and needs to be chosen carefully.

### 2.2 Choosing the set $\Omega$

In practice, the data analyst must be able to solve the optimization problem in equation (6). If $\Omega$ is an interval, lower-bounded by 0, an upper bound must be selected for any grid-search optimization procedure. To define such an upper bound in a practical way, it should be large enough to include all possible estimators in a given class while still being finite. This implies we must choose $\Omega$ to be a function of the data. The specifics of $\Omega$ depend on the regularizing set $\mathcal{B}_\omega$. The upper bound will have a nontrivial impact on the quality of the recovery, as choosing a value too small may eliminate the best solutions. Thus, treating the upper bound as a random function of the data is more realistic from a statistical practice point of view.

In the case that $\omega = t$, $\hat{\beta}_t$ must be in the $\ell_1$-ball with radius $t$. This constraint is only binding if

$$t < \min_{\eta \in \mathcal{K}} ||\hat{\beta}_\infty + \eta||_1 =: t_0,$$

where $\hat{\beta}_\infty = \hat{\beta}(\mathbb{R}^p)$ is a least-squares solution and $\mathcal{K} := \{ a : \mathbb{X}a = 0 \}$. Observe that $\mathcal{K} = \{0\}$ if $n \geq p$ and otherwise $\mathcal{K}$ has dimension $p - n$, which would imply $\hat{\beta}_\infty$ is not unique (both of these statements assume that the columns of $\mathbb{X}$ contain a linearly independent set of size $\min\{n, p\}$). In either case, if $t \geq t_0$, then $\hat{\beta}_t$ is equal to a least-squares solution. Therefore, we define $T_n := [0, t_{\text{max}}]$, where

$$t_{\text{max}} := \left\| \hat{\beta}_\infty \right\|_1 ,$$

and $\hat{\beta}_\infty = (\mathbb{X}^T \mathbb{X})^+ \mathbb{X}^T Y$ is the least-squares solution when $(\cdot)^+$ is given by the Moore-Penrose inverse.

It is important to realize that as both $t_0$ and $t_{\text{max}}$ are functions of the data they are both random. Since $0 \in \mathcal{K}$, $t_{\text{max}} \geq t_0$. Therefore, $t_{\text{max}}$ is a bit conservative in the sense that the lasso
solution is already a least-squares solution for values of $t \in [t_0, t_{\max})$. We make this definition to eliminate the explicit dependence on the null space of $\mathbf{X}$ that appears in the definition of $t_0$. Providing further justification for this choice of $T_n$, LARS [6] and the associated R package by default uses the $\ell_1$-norm of the ‘saturated’ model to determine an appropriate value for $t_{\max}$ when confronted with a dataset. This default setting coincides with our choice of $t_{\max}$, the $\ell_1$-norm of a particular least-squares solution. Analogously to $T_n$ for lasso, we define $U_n := [0, u_{\max}]$ for group lasso, where $u_{\max} := \sum_{g \in G} \sqrt{|g|} \left\| (\beta_\infty)_{j \in g} \right\|_2$.

### 3 Main results

For any $\omega$, define the oracle estimator with respect to $\omega$ as

$$\hat{\beta}_{\omega} := \operatorname{argmin}_{\beta \in \mathcal{B}_\omega} R(\beta).$$

Suppose $\hat{\omega}$ is an estimated value of $\omega$, such as by cross-validation. Then a natural criterion for studying the performance of an estimator $\hat{\beta}_{\hat{\omega}}$ is the excess risk of $\hat{\beta}_{\hat{\omega}}$ relative to $\hat{\beta}_\omega$, which we define as

$$\mathcal{E}(\hat{\omega}, \omega) := R(\hat{\beta}_{\hat{\omega}}) - R(\hat{\beta}_\omega). \quad (7)$$

Note that $\mathcal{E}(\hat{\omega}, \omega)$ is random due to the term $R(\hat{\beta}_{\hat{\omega}})$. Here, $R(\hat{\beta}_{\hat{\omega}})$ is a function of the data, with the expectation only with respect to a new test random variable $Z$ and not with respect to the observed data used to choose either $\hat{\omega}$ or $\hat{\beta}_\omega$.

This criterion allows for meaningful theory when the oracle linear model is not risk consistent; that is, when the term $R(\hat{\beta}_\omega)$ does not necessarily converge to zero as $n$ grows. This is particularly important here, as we do not assume that the conditional expectation of $\mathcal{Y}$ given $\mathcal{X}$ is linear.

Before outlining our results, we define the following set of distributions

**Definition 3.1.** Let $Z \in \mathbb{R}^{p+1}$ be any $p + 1$-dimensional random variable with distribution $F_n$. Then define the following set of sequences of distributions

$$\mathcal{F} := \left\{ (F_n)_{n \geq 1} : \exists C < \infty \text{ s.t. } \mathbb{E}_{F_n} \max_{0 \leq j, k \leq p} (Z_j Z_k - \mathbb{E}_{F_n} Z_j Z_k)^2 \leq C \text{ for all } n \right\}.$$

While $F_n$ is a measure on $\mathbb{R}^{p+1}$ indexing with $n$ is more natural than indexing with $p$ given that our results include $p(n)$ increasing with $n$. **Definition 3.1** is a common moment condition [9, 15] for showing risk consistency of the lasso-type methods in high dimensional settings. Heuristically, $\mathcal{F}$ is the set of all triangular distributions where a universal constant exists that bounds the variance of all the $(p + 1)^2$ interaction terms. When necessary, and with some abuse of notation, we refer to this constant for a given sequence of distributions as $C_\mathcal{F}$. Note that we do not assume the design is Gaussian nor independent.

Additionally, we make the following condition.

**Condition 1.** For any system of validation sets $V_n$, there exists a constant $c_n$ such that, for all $v \in V_n$, $|v| \geq c_n$. Additionally, for any $v \neq v' \in V_n$, $v \cap v' = \emptyset$.

For example, with $K$-fold cross-validation, we can take $c_n = \lfloor n/K \rfloor$, which is the integer part of $n/K$.

We can now state our main results.
Theorem 3.2 (Lasso main result). Let \((F_n) \in \mathcal{F}\) be given such that \(\mathbb{E} \left[ \left| \beta_{\infty} \right|_{1}^{4} \right] = o(t_n^4)\) and \(p_n = n^{\alpha} \) for some \(\alpha > 0\). Also, suppose Condition 1 holds with \(c_n < n\) for all \(n\). Then, for any \(\delta > 0\),

\[
\mathbb{P}_{F_n} \left( \mathcal{E}(\hat{t},t_n) > \delta \right) = o \left( t_n^2 \sqrt{\frac{\log n}{c_n}} \right) \tag{8}
\]

Remark 1. Usually in the oracle estimation framework, \(\hat{t} = t_n\) and so the excess risk is necessarily nonnegative due to the oracle predictor, \(\beta_{tn}\), being selected as the risk minimizer over \(\mathcal{B}_{tn}\). In that case, equation (8) is the same as saying \(\mathcal{E}(\hat{t},t_n) = o \left( t_n^2 \sqrt{\frac{\log n}{c_n}} \right)\). As we are examining the case where the optimization set is estimated, \(\mathcal{E}(\hat{t},t_n)\) may be negative. However, we are only interested in bounding the case where \(\mathcal{E}(\hat{t},t_n)\) is positive, i.e. the case where the estimator is worse than the oracle.

Before discussing this result further, we present an additional theorem, which extends this result to the group lasso.

Theorem 3.3. Let \((F_n) \in \mathcal{F}\) be given such that \(\mathbb{E} \left[ \left( \sum_g \left| \left( \hat{\beta}_{\infty,j} \right)_{j \in g} \right|_{2} \right)^4 \right] = o(u_n^4)\), \(p_n = n^{\alpha} \) for some \(\alpha > 0\), and \(\max_{g \in G} |g| = a_n\). Also, suppose Condition 1 holds with \(c_n < n\) for all \(n\). Then, for any \(\delta > 0\),

\[
\mathbb{P}_{F_n} \left( \mathcal{E}(\hat{u},u_n) > \delta \right) = o \left( a_n u_n^2 \sqrt{\frac{\log n}{c_n}} \right) .
\]

Though Theorem 3.2 and Theorem 3.3 could be combined, we state them separately for clarity. Also, Theorem 3.3 has a weaker hypothesis as

\[
\mathbb{E} \left| \beta_{\infty} \right|_{1}^{4} = o(u_n^4) \Rightarrow \mathbb{E} \left[ \left( \sum_g \left| \left( \hat{\beta}_{\infty,j} \right)_{j \in g} \right|_{2} \right)^4 \right] = o(u_n^4).
\]

However, as long as the group size is bounded asymptotically, that is \(a_n = O(1)\), then the rates of convergence to the risk of the oracle predictor are the same for both methods.

4 Discussion of assumptions and examples

The quantity \(c_n\), which is the number of observations put into the smallest validation set, controls the decay of the excess risk. If \(c_n \approx n\), which is the case when using \(K\)-fold cross-validation for a fixed integer \(K\) (such as \(K = 10\), which is the default in the \texttt{R} package \texttt{glmnet} [7]), then the rate is \(o \left( t_n^2 \sqrt{\frac{\log n}{n}} \right)\).

The inclusion of \(t_n\) (or \(u_n\)) also deserves comment. Here, \(t_n\) is a sequence of constants which determines the amount of regularization imposed in the oracle. Naturally, the faster \(t_n\) goes to infinity, the less constrained the oracle is, due to \(\mathcal{B}_{tn}\) being a larger set, and the less restrictive the assumptions. We still get \(\mathbb{P}_{F_n} \left( \mathcal{E}(\hat{t},t_n) > \delta \right) = o(1)\) as long as \(t_n = o \left( \left( \frac{n}{\log n} \right)^{1/4} \right)\). As mentioned in Greenshtein and Ritov [9], if \(t_n\) grows as fast or faster than \(\left( \frac{n}{\log n} \right)^{1/4}\), then \(R \left( \hat{\beta}(\mathcal{B}_{tn}) \right) - R (\beta(\mathcal{B}_{tn}))\) does not necessarily converge to 0 in probability and hence the lasso is not necessarily risk consistent.
with this tuning parameter. A similar result for $u_\alpha$ and grouped lasso is shown in Nardi and Rinaldo [15]. This begs the question: is it feasible to have $E \left[ \| \beta_\infty \|_1^4 \right] = o(t_n^4)$ when $t_n = o \left( \left( \frac{n}{\log n} \right)^{1/4} \right)$?

Some assumptions can be used to derive sufficient conditions for the moment condition we impose. Suppose as before that $p_n = n^\alpha$ for some $\alpha > 0$. Then

$$
E \left[ \| \beta_\infty \|_1^4 \right] \leq E \left[ \| (X^T X)^{1/2} \|_2^4 \| Y \|_1^4 \right]
$$

$$
\leq p^2 E \left[ \| (X^T X)^{1/2} \|_2^4 \| Y \|_1^4 \right]
$$

$$
= p^2 E \left[ \| (X^T X)^{1/2} \|_2^4 \| Y \|_1^4 \right] E_{Y|X} \left[ ||Y||_1^4 \right]
$$

$$
= n^{2\alpha} E_{X} \left[ \sigma_{\min}^+(X)^{-4} \right] E_{Y|X} \left[ ||Y||_1^4 \right],
$$

where $\sigma_{\min}^+(A)$ is the smallest non-zero singular value of a matrix $A$, $||.||_s$ is the operator norm corresponding to the $\ell_s$ vector norm, and equation (9) uses the sub-multiplicative property of the operator norm.

Suppose that our model is the usual nonparametric regression model $Y = m(X) + e$, where $X$ and $e$ are stochastically independent random variables and $e$ has zero mean. If there exists a constant $C$ independent of $n$ such that $\text{ess sup}_x m(x) < C$ with respect to the distribution of $X$ and $E e^4 < \infty$, then there exists a $C' < \infty$, again independent of $n$, such that

$$
E [Y^4 | X] = E \left[ \sum_{l=0}^4 \binom{4}{l} m(X)^l e^{4-l} | X \right]
$$

$$
= \sum_{l=0}^4 \binom{4}{l} m(X)^l E e^{4-l} < C'.
$$

This implies that $E_{Y|X} ||Y||_1^4 = O(n)$. Combining this with equation (10) and writing $E_{X} \left[ \sigma_{\min}^+(X)^{-4} \right] = O(n^{-s})$, with $s \geq 0$, we see that

$$
E \left[ \| \beta_\infty \|_1^4 \right] = O(n^{-s+2\alpha+1}).
$$

Therefore, Condition 1 follows if $n^{-s+2\alpha+1} = o \left( \frac{n}{\log n} \right)$, which happens if, for instance, $s > 2\alpha$. So, if with high probability $\sigma_{\min}^+(X) = \Omega(n^{\alpha/2})$, i.e. is of larger order than $n^{\alpha/2}$, Condition 1 holds. Indeed, as shown by Rudelson and Vershynin [17], a random matrix composed of independent and identically distributed sub-Gaussian random variables has, with high probability,

$$
\sigma_{\min}^+(X) \geq \begin{cases} \sqrt{p_n} = n^{\alpha/2} & \text{if } \alpha > 1 \\ \sqrt{n} = n^{1/2} & \text{if } \alpha < 1. \end{cases}
$$

In either case, $\sigma_{\min}^+(X)$ is at least of order $n^{\alpha/2}$.

Alternatively, suppose that instead of random design we want to estimate the regression function nonparametrically. Let $e_i = i/n$ be equally spaced grid points and let $\phi : [0,1] \to \mathbb{R}$ be a fixed kernel function. Define $\phi_{j,h}(x) = \frac{1}{h} \phi \left( \frac{x_j - x}{h} \right)$ and let $X_{ij} = \phi_{j,h}(e_i)$ for $j = 1, \ldots, p = n$. Let $\phi$ and $h$ be such that, for all $\phi_{j,h}(e_j) - \sum_{k \neq j} \phi_{j,h}(e_k) \geq C n^{1/7}$ for some constant $0 < C < \infty$ independent
of $n$ and $j$. Note that this is always possible as long as $\phi$ is such that $h^{-1}\phi(1/h) \to 0$ as $h \to 0$—as is the case, for example, with the Gaussian kernel—and $\phi_{j,h}(e_j) \to \infty$ as $h \to 0$. Lastly, suppose for simplicity that $\phi$ is symmetric (this is not necessary, however). Then, Johnson [10] shows that

$$\sigma_{\min}^+(\mathcal{X}) \geq \max_{1 \leq i \leq n} \left\{ |X_{ii}| - \frac{1}{2} \left( \sum_{j \neq i} |X_{ij}| + \sum_{j \neq i} |X_{ij}| \right) \right\}. \tag{11}$$

Therefore,

$$\sigma_{\min}^+(\mathcal{X})^{-4} \leq \frac{1}{Cn^{-4/7}}. \tag{11}$$

In this example, $\alpha = 1$, so $s > 1/2$, as required.

Likewise, suppose we wish to estimate the regression function $m$ via some orthogonal basis $(\psi)_{j=1}^\infty$ with coefficients $m_j = \int m(\psi_j)$. Then, as the design matrix is given by $X = [\psi_j(X_i)]$, if the design is such that the Riemann sum $\frac{1}{n} \sum_{i=1}^n \psi_j(X_i) \psi_k(X_i) \approx \int \psi_j \psi_k$, the singular value condition holds. For example, suppose $(\psi_j)$ is the trigonometric basis and $X_i = i/n$. Then, if $\psi(x) = (\psi_1(x), \ldots, \psi_p(x))^\top$ and $p - 1 \leq n$, the matrix $\Psi = \frac{1}{n} \sum_{i=1}^n \psi(X_i)\psi(X_i)^\top$ is the identity.

Therefore, as $X^\top X = \sum_{i=1}^n \psi(X_i)\psi(X_i)^\top$, $\sigma_{\min}^+(\mathcal{X}) = \sqrt{\sigma_{\min}^+(n\Psi)} = \sqrt{n\sigma_{\min}^+(\Psi)} = \sqrt{n}$.

Lastly, two corollaries follow immediately from our proofs.

**Corollary 4.1.** Under the conditions of Theorem 3.2,

$$\mathbb{E}_{F_n} \left[ |\mathcal{E}\left(\hat{\ell}, t_n\right)| \right] = O \left( t_n^2 \sqrt{\frac{\log n}{c_n}} \right) + \mathbb{E}_{F_n} \left[ \mathcal{E}\left(\hat{\ell}, t_n\right) 1_{\mathcal{E}(\hat{\ell}, t_n) < 0} \right].$$

**Corollary 4.2.** Under the conditions of Theorem 3.3,

$$\mathbb{E}_{F_n} \left[ |\mathcal{E}\left(\hat{u}, u_n\right)| \right] = O \left( a_n u_n^2 \sqrt{\frac{\log n}{c_n}} \right) + \mathbb{E}_{F_n} \left[ \mathcal{E}\left(\hat{u}, u_n\right) 1_{\mathcal{E}(\hat{u}, u_n) < 0} \right].$$

5 Proof of main results

To show the main theorems of this paper, we need a few preliminary results. We will show how to rewrite the risk as a quadratic form in Section 5.1. We then present some useful lemmas in Section 5.2 before presenting our proof.

5.1 Squared-error loss and quadratic forms

We can rewrite the various formulas for the risk from Section 2 as quadratic forms. Define the parameter to be $\gamma := (-1, \beta)^\top$, with associated estimator $\hat{\gamma}_\omega := (-1, \hat{\beta}_\omega)^\top$. We can rewrite equation (2) as

$$R(\beta) = \mathbb{E}_{F_n} \left[ (Y - X\beta)^2 \right] = \gamma^\top \Sigma_n \gamma \tag{12}$$

where $\Sigma_n := \mathbb{E}_{F_n}[ZZ^\top]$. Analogously, equation (3) has the following form

$$\hat{R}(\beta) = \frac{1}{n} ||Y - X\beta||_2^2 = \gamma^\top \hat{\Sigma}_n \gamma,$$

where $\hat{\Sigma}_n = n^{-1} \sum_{i=1}^n Z_i Z_i^\top$. Lastly, we rewrite equation (5) as

$$\hat{R}_{V_n}(\omega) = \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega^{(v)})^\top \hat{\Sigma}_v \hat{\gamma}_\omega^{(v)}, \tag{13}$$
where \( \hat{\Sigma}_v = |v|^{-1} \sum_{r \in v} Z_r Z_r^\top \), \( \hat{\gamma}_v^{(v)} := (-1, \hat{\beta}_v^{(v)})^\top \), and
\[
\hat{\beta}_v^{(v)} := \text{argmin}_{\beta \in \mathcal{B}_v} \gamma^\top \hat{\Sigma}_v \gamma,
\]
with \( \hat{\Sigma}_v := (n - |v|)^{-1} \sum_{r \notin v} Z_r Z_r^\top \).

5.2 Supporting lemmas

Several times in our proof of the main results we need to bound a quadratic form given by a symmetric matrix and an estimator indexed by a tuning parameter. To this end, we state and prove the following lemma.

**Lemma 5.1.** Suppose \( a \in \mathbb{R}^p \) and \( A \in \mathbb{R}^{p \times p} \). Then
\[
a^\top Aa \leq ||a||_1 ||A||_\infty,
\]
where \( ||A||_\infty := \max_{i,j} |A_{ij}| \) is the entry-wise max norm.

**Proof of Lemma 5.1.**
\[
a^\top Aa \leq ||a||_1 ||Aa||_\infty \\
\leq ||a||_1 \max_{i,j} |A_{ij}| ||a||_1 \\
= ||a||_1^2 ||A||_\infty,
\]
where the first inequality follows by Hölder’s inequality. \( \square \)

Additionally, we include a special case of Nemirovski’s inequality for completeness. See Dümbgen et al. [5] for more general formulations.

**Lemma 5.2** (Nemirovski’s inequality). Let \( \xi_1, \ldots, \xi_n \) be independent random vectors in \( \mathbb{R}^d \), for \( d \geq 3 \) with \( E[\xi] = 0 \) and \( E ||\xi||_2^2 < \infty \). Then, for any validation set \( v \) and distribution for the \( \xi_i \)'s,
\[
E \left[ \left\| \sum_{i \in v} \xi_i \right\|_\infty^2 \right] \leq (2e \log d - e) \sum_{i \in v} E \left[ ||\xi_i||_\infty^2 \right] \\
\leq 2e \log d \sum_{i \in v} E \left[ ||\xi_i||_\infty^2 \right] \\
\lesssim \log d \sum_{i \in v} E \left[ ||\xi_i||_\infty^2 \right].
\]

Finally, we will use **Lemma 5.2** to find the rate of convergence for the sample covariance matrix to the population covariance.

**Lemma 5.3.** Let \( V_n = \{v_1, \ldots, v_K\} \) be a set of validation sets satisfying **Condition 1**. Then,
\[
E \left( \frac{1}{K} \sum_{v \in V_n} \left\| \hat{\Sigma}_v - \Sigma \right\|_\infty \right)^2 = O \left( \frac{\log n}{c_n} \right).
\]
Proof of Lemma 5.3. First, note that
\[
\mathbb{E} \left( \sum_{v \in V_n} \| \tilde{\Sigma}_v - \Sigma_n \|_\infty \right)^2 = \sum_{u \neq v \in V_n} \mathbb{E} \| \tilde{\Sigma}_v - \Sigma_n \|_\infty \mathbb{E} \| \tilde{\Sigma}_u - \Sigma_n \|_\infty + \\
+ \sum_{v \in V_n} \mathbb{E} \| \tilde{\Sigma}_v - \Sigma_n \|_\infty^2
\]
by independence and disjoint elements of \( v \). Let \( \xi_r \in \mathbb{R}^{(p+1)^2} \) be the vectorized version of the zero-mean matrix \( \frac{1}{c_n}(Z_rZ_r^T - \mathbb{E}ZZ^T) \). Then,
\[
\mathbb{E} \| \tilde{\Sigma}_v - \Sigma_n \|_\infty \leq \sqrt{\mathbb{E} \| \tilde{\Sigma}_v - \Sigma_n \|_\infty^2} = \sqrt{\mathbb{E} \left( \sum_{r \in v} \xi_r \right)^2} \tag{14}
\]
by Jensen’s inequality. Using Lemma 5.2 with \( d = (p+1)^2 \), we find
\[
\mathbb{E} \left( \sum_{r \in v} \xi_r \right)^2 \leq \log((p+1)^2) \frac{1}{c_n} \sum_{r \in v} \mathbb{E} \frac{\max_{0 \leq j,k \leq p} |Z_{rj}Z_{rk} - \mathbb{E}Z_jZ_k|^2}{\log n} \leq \log(4n^2) \frac{1}{c_n} \sum_{r \in v} C_F \leq \frac{\log n}{c_n}.
\]
Therefore,
\[
\mathbb{E} \left( \frac{1}{K} \sum_{v \in V_n} \| \tilde{\Sigma}_v - \Sigma_n \|_\infty \right)^2 \leq \frac{1}{K^2} \left( \sum_{u,v \in V_n} \frac{\log n}{c_n} \right) = \frac{\log n}{c_n}. \tag{15}
\]

5.3 Proofs
Define \( \omega_* := \min\{\omega_{\text{max}}, \omega_n\} \). Then, we can write
\[
\mathcal{E}(\tilde{\omega}, \omega_n) = R \left( \tilde{\beta}_{\tilde{\omega}} \right) - R \left( \beta_{\omega_n} \right) = R \left( \tilde{\beta}_{\tilde{\omega}} \right) - \tilde{R}_{V_n}(\tilde{\omega}) + \tilde{R}_{V_n}(\tilde{\omega}) - \tilde{R}_{V_n}(\omega_*) \tag{I}
\]
\[
+ \tilde{R}_{V_n}(\omega_*) - \tilde{R} \left( \tilde{\beta}_{\omega_*} \right) + \tilde{R} \left( \tilde{\beta}_{\omega_*} \right) - \tilde{R} \left( \tilde{\beta}_{\omega_n} \right) \tag{II}
\]
\[
+ \tilde{R} \left( \tilde{\beta}_{\omega_n} \right) - R \left( \tilde{\beta}_{\omega_n} \right) + R \left( \tilde{\beta}_{\omega_n} \right) - R \left( \beta_{\omega_n} \right) \tag{IV}
\]
where \( \tilde{\beta}_\omega = \tilde{\beta}(\mathcal{B}_\omega) \) for \( \omega \in \{ \tilde{\omega}, \omega_*, \omega_n \} \).
Now, for any \( \omega \in \Omega \),
\[
\tilde{R}_{V_n}(\tilde{\omega}) - \tilde{R}_{V_n}(\omega_*) \leq 0,
\]

10
by the definition of $\tilde{\omega}$. Therefore, as $\omega_* \in \Omega$, we have that $(II) \leq 0$. Also, by the discussion in Section 2.2,

$$\tilde{R}(\tilde{\beta}_\omega) = \begin{cases} \tilde{R}(\tilde{\beta}_\omega) & \text{if } \omega < \omega_{\text{max}} \\ \tilde{R}(\tilde{\beta}_{\omega_{\text{max}}}) & \text{if } \omega \geq \omega_{\text{max}}. \end{cases}$$

To see this, note that for any $\omega \geq \omega_{\text{max}}$, $\tilde{\beta}_\omega$ is a least-squares solution. Therefore, by the definition of $\omega_*$, $\tilde{R}(\tilde{\beta}_{\omega_*}) = \tilde{R}(\tilde{\beta}_{\omega_{\text{max}}})$ and hence $(IV) = 0$.

To bound the remaining terms, we rewrite them as quadratic forms as in Section 5.1 and use the lemmas in Section 5.2. We break the remainder of this section into parts based on the decomposition (I) - (VI).

**Final predictor and cross-validation risk (I)** Using the notation introduced in Section 5.1, note that by equations (12) and (13)

$$R(\tilde{\beta}_\omega) - R_{V_n}(\omega) = \gamma_\omega^T \Sigma_n \gamma_\omega - \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \Sigma_v \hat{\gamma}_\omega(v)$$

$$= \left[ \gamma_\omega^T \Sigma_n \gamma_\omega - \gamma_\omega^T \left( \hat{\Sigma}_n \right) \hat{\gamma}_\omega \right] + \left[ \gamma_\omega^T \left( \hat{\Sigma}_n \right) \hat{\gamma}_\omega - \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \Sigma_v \hat{\gamma}_\omega(v) \right]$$

Addressing each of the terms in order,

$$\left[ \gamma_\omega^T \Sigma_n \gamma_\omega - \gamma_\omega^T \left( \hat{\Sigma}_n \right) \hat{\gamma}_\omega \right] = \gamma_\omega^T \left( \Sigma_n - \hat{\Sigma}_n \right) \hat{\gamma}_\omega \leq \|\hat{\gamma}_\omega\|_1^2 \left\| \Sigma_n - \hat{\Sigma}_n \right\|_\infty,$$

where the inequality follows by Lemma 5.1.

Likewise,

$$\left[ \gamma_\omega^T \left( \hat{\Sigma}_n \right) \hat{\gamma}_\omega - \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \Sigma_v \hat{\gamma}_\omega(v) \right]$$

$$= \left( \gamma_\omega^T \hat{\Sigma}_n \gamma_\omega - \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \hat{\Sigma}_n \hat{\gamma}_\omega(v) \right) + \left( \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \hat{\Sigma}_n \hat{\gamma}_\omega(v) - \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \Sigma_v \hat{\gamma}_\omega(v) \right)$$

$$= \frac{1}{K} \sum_{v \in V_n} \left( \gamma_\omega^T \hat{\Sigma}_n \gamma_\omega - (\hat{\gamma}_\omega(v))^T \hat{\Sigma}_n \hat{\gamma}_\omega(v) \right) + \left( \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \left( \hat{\Sigma}_n - \Sigma_v \right) \hat{\gamma}_\omega(v) \right)$$

$$\leq \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}_\omega(v))^T \left( \hat{\Sigma}_n - \Sigma_v \right) \hat{\gamma}_\omega(v).$$

The last inequality follows as $\hat{\gamma}_\omega$ is chosen to minimize $\gamma_\omega^T \hat{\Sigma}_n \gamma_\omega$, and so for any $v \in V_n$

$$\gamma_\omega^T \hat{\Sigma}_n \gamma_\omega \leq (\hat{\gamma}_\omega(v))^T \hat{\Sigma}_n \hat{\gamma}_\omega(v).$$
Continuing and using Lemma 5.1,

\[
\frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}^{(v)}_{\omega} )^\top (\hat{\Sigma}_n - \hat{\Sigma}_v) \hat{\gamma}^{(v)}_{\omega} \\
\leq \frac{1}{K} \sum_{v \in V_n} ||\hat{\gamma}^{(v)}_{\omega}||^2_1 ||\hat{\Sigma}_n - \hat{\Sigma}_v||_{\infty} \\
\leq \frac{1}{K} \sum_{v \in V_n} ||\hat{\gamma}^{(v)}_{\omega}||^2_1 (||\hat{\Sigma}_n - \Sigma_n||_{\infty} + ||\Sigma_n - \hat{\Sigma}_v||_{\infty})
\] (17)

**Cross-validation risk and empirical risk (III)** Recall that

\[
\hat{\Sigma}_{(v)} = (n - c_n)^{-1} \sum_{v \notin V} Z_v Z_v^\top.
\]

Then,

\[
\hat{R}_{V_n}(\omega) - \tilde{R}(\hat{\beta}_{\omega}) = \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}^{(v)}_{\omega} )^\top \hat{\Sigma}_v \hat{\gamma}^{(v)}_{\omega} - \hat{\gamma}^\top_{\omega} \hat{\Sigma}_n \hat{\gamma}_{\omega} \\
= \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}^{(v)}_{\omega} )^\top \hat{\Sigma}_v \hat{\gamma}^{(v)}_{\omega} - (\hat{\gamma}^{(v)}_{\omega})^\top \hat{\Sigma}_v (\hat{\gamma}^{(v)}_{\omega}) + \\
+ \frac{1}{K} \sum_{v \in V_n} (\hat{\gamma}^{(v)}_{\omega} )^\top \hat{\Sigma}_v (\hat{\gamma}^{(v)}_{\omega}) - \hat{\gamma}^\top_{\omega} \hat{\Sigma}_n \hat{\gamma}_{\omega} \\
\leq \frac{1}{K} \sum_{v \in V_n} ||\hat{\gamma}^{(v)}_{\omega}||^2_1 ||\hat{\Sigma}_v - \hat{\Sigma}(v)||_{\infty},
\]

(18)

where the inequality follows by Lemma 5.1 and the fact that \(\hat{\gamma}^{(v)}_{\omega}\) is chosen to minimize \((\hat{\gamma}^{(v)}_{\omega})^\top \hat{\Sigma}(v) \hat{\gamma}^{(v)}_{\omega}\), which implies \((\hat{\gamma}^{(v)}_{\omega})^\top \hat{\Sigma}(v) \hat{\gamma}^{(v)}_{\omega}\) \(\leq \hat{\gamma}^\top_{\omega} \hat{\Sigma}(v) \hat{\gamma}_{\omega}\).

**Empirical risk and expected risk (V, VI)** The proof of these results is given by Greenshtein and Ritov [9]. We include a somewhat different proof for completeness. Observe the following bounds

\[
\tilde{R}(\hat{\beta}_{\omega}) - R(\hat{\beta}_{\omega}) \leq \sup_{\beta \in B_{\omega}} |R(\beta) - \tilde{R}(\beta)|
\]

and

\[
R(\hat{\beta}_{\omega}) - R(\beta_{\omega}) = R(\hat{\beta}_{\omega}) - \tilde{R}(\hat{\beta}_{\omega}) + \tilde{R}(\hat{\beta}_{\omega}) - R(\beta_{\omega}) \\
= R(\hat{\beta}_{\omega}) - \tilde{R}(\hat{\beta}_{\omega}) + \tilde{R}(\hat{\beta}_{\omega}) - R(\beta_{\omega}) + \\
+ \tilde{R}(\beta_{\omega}) - R(\beta_{\omega}).
\]

Therefore, bounds for both (V) and (VI) can be found as follows.

\[
\sup_{\beta \in B_{\omega}} |R(\beta) - \tilde{R}(\beta)| = \sup_{\beta \in B_{\omega}} \left| \gamma^\top \Sigma \gamma - \gamma^\top \hat{\Sigma}_n \gamma \right| \\
= \sup_{\beta \in B_{\omega}} \left| \gamma^\top (\Sigma_n - \hat{\Sigma}_n) \gamma \right| \\
\leq \sup_{\beta \in B_{\omega}} ||\gamma||^2_1 ||\hat{\Sigma}_n - \Sigma_n||_{\infty}
\]

(19)
Proof of Theorem 3.2. Addressing equations (16) and (17) and substituting $t$ for $\omega$, we find that

\[
(I) = R\left(\beta_t\right) - \bar{R}_{V_n}(\hat{t})
\]

\[
\leq \left\| \gamma_t \right\|^2 \left\| \Sigma_n - \tilde{\Sigma}_n \right\| + \frac{1}{K} \sum_{v \in V_n} \left(\gamma_t^{(v)}\right)^\top \left(\tilde{\Sigma}_n - \tilde{\Sigma}_v \right) \gamma_t^{(v)}
\]

\[
\leq (1 + t_{\text{max}})^2 \left(2 \left\| \Sigma_n - \tilde{\Sigma}_n \right\| + \frac{1}{K} \sum_{v \in V_n} \left\| \Sigma_n - \tilde{\Sigma}_v \right\|\right).
\]

By Lemma 5.3 with $V_n = \{1, \ldots, n\}$ and $c_n = n$,

\[
\mathbb{E} \left\| \Sigma_n - \tilde{\Sigma}_n \right\|^2 \lesssim \left(\frac{\log n}{n}\right).
\]

Additionally, by Lemma 5.3 with $V_n = \{v_1, \ldots, v_K\}$,

\[
\mathbb{E} \left(\frac{1}{K} \sum_{v \in V_n} \left\| \Sigma_n - \tilde{\Sigma}_v \right\|^2\right) \lesssim \left(\frac{\log n}{c_n}\right).
\]

Combining these two bounds together along with the Cauchy-Schwarz inequality, we get

\[
\mathbb{E} \left[\left\| \gamma_t \right\|^2 \left\| \Sigma_n - \tilde{\Sigma}_n \right\| + \frac{1}{K} \sum_{v \in V_n} \left(\gamma_t^{(v)}\right)^\top \left(\tilde{\Sigma}_n - \tilde{\Sigma}_v \right) \gamma_t^{(v)}\right]
\]

\[
\leq 2 \sqrt{\mathbb{E}[(1 + t_{\text{max}})^4] \mathbb{E} \left\| \Sigma_n - \tilde{\Sigma}_n \right\|^2 + \mathbb{E}[(1 + t_{\text{max}})^4] \mathbb{E} \left[\left(\frac{1}{K} \sum_{v \in V_n} \left\| \Sigma_n - \tilde{\Sigma}_v \right\|^2\right)^2\right]}
\]

\[
\lesssim \sqrt{\mathbb{E}[(1 + t_{\text{max}})^4] \left(\frac{\log n}{n} + \frac{\log n}{c_n}\right)}.
\]

Turning now to equation (18), we can use a straightforward adaptation of Lemma 5.3 to show that

\[
\mathbb{E} \left(\frac{1}{K} \sum_{v \in V_n} \left\| \Sigma_n - \tilde{\Sigma}_v \right\|\right)^2 \lesssim \frac{\log n}{n - c_n}.
\]

By assumption, $n > c_n$. Hence,

\[
\mathbb{E} \left(\frac{1}{K} \sum_{v \in V_n} \left\| \Sigma_n - \tilde{\Sigma}_v \right\|\right)^2 = O\left(\frac{\log n}{n}\right).
\]

Therefore, following the analogous steps that lead to equation (22), we conclude that

\[
\mathbb{E} \left[\frac{1}{K} \sum_{v \in V_n} \left|\gamma_t^{(v)}\right|^2 \left|\tilde{\Sigma}_v - \tilde{\Sigma}_v \right|\right] = O\left((1 + t_n)^2 \sqrt{\frac{\log n}{n}}\right).
\]
Lastly, we address equation (19). By Lemma 5.3,  
\[ \mathbb{E} \left\| \tilde{\Sigma}_n - \Sigma_n \right\|_\infty \leq \sqrt{\frac{\log n}{n}}, \]
which implies  
\[ \mathbb{E} \sup_{\beta \in \mathcal{B}_n} \left\| \gamma \right\|_1 \left\| \tilde{\Sigma}_n - \Sigma_n \right\|_\infty \leq (1 + t_n)^2 \mathbb{E} \left\| \tilde{\Sigma}_n - \Sigma_n \right\|_\infty \lesssim (1 + t_n)^2 \sqrt{\frac{\log n}{n}}. \]
By applying Markov’s inequality, we have that  
\[ \mathbb{P} \left( \mathcal{E}(\tilde{t}, t_n) > \delta \right) \]
\[ \leq \mathbb{P} \left( \left\| \tilde{\gamma}_v \right\|_1^2 \left\| \Sigma_n - \tilde{\Sigma}_n \right\|_\infty + \frac{1}{K} \sum_{v \in V_n} \left( \tilde{\gamma}_v \right)^{\top} \left( \tilde{\Sigma}_n - \tilde{\Sigma}_v \right) \tilde{\gamma}_v \right) + 
\frac{1}{K} \sum_{v \in V_n} \left\| \gamma_v \right\|_1^2 \left\| \tilde{\Sigma}_n - \tilde{\Sigma}_v \right\|_\infty + \sup_{\beta \in \mathcal{B}_n} \left\| \gamma \right\|_1 \left\| \tilde{\Sigma}_n - \Sigma_n \right\|_\infty > \delta \right) \]
\[ \lesssim \sqrt{\mathbb{E} \left[ (1 + t_{\max})^4 \right]} \left( \sqrt{\frac{\log n}{c_n}} + \sqrt{\frac{\log n}{n}} \right) + (1 + t_n)^2 \left( \sqrt{\frac{\log n}{n - c_n}} + \sqrt{\frac{\log n}{n}} \right) \]
Using the fact that \((1 + x)^4 \asymp x^4\), we can conclude  
\[ \mathbb{P} \left( \mathcal{E}(\tilde{t}, t_n) > \delta \right) = o \left( \sqrt{t_n \log n} \right), \]
which shows the result.

**Proof of Theorem 3.3.** Returning now to equations (16) through (19) and substituting \( u \) for \( \omega \), we can form analogous bounds to the ones in the proof of Theorem 3.2. To this end, note that  
\[ \left\| \gamma \right\|_1^2 \leq (1 + \left\| \beta \right\|_1)^2, (1 + \left\| \beta \right\|_1)^2 \asymp \left\| \beta \right\|_1^2, \]
\[ \left\| \beta \right\|_1 \leq \sum_{g \in G} \sqrt{\left\| g \right\| \left( \beta_{j,g} \right)_{j \in g}} \], and \( \max_{g \in G} \left\| g \right\| = a_n \).
Therefore, 
\[ \mathbb{E} \left\| \tilde{\gamma}_u \right\|_1^4 \leq \mathbb{E} \left( 1 + \left\| \tilde{\beta}_u \right\|_1 \right)^4 \]
\[ \times \mathbb{E} \left\| \tilde{\beta}_u \right\|_1^4 \]
\[ \leq \mathbb{E} \left( \sum_{g \in G} \sqrt{\left\| g \right\| \left( \tilde{\beta}_{u,j} \right)_{j \in g}} \right)^4 \]
\[ \leq a_n^2 \mathbb{E} \left( \sum_{g \in G} \left\| \left( \tilde{\beta}_{u,j} \right)_{j \in g} \right\|_2 \right)^4 \]
\[ \leq a_n^2 \mathbb{E} u_{\max}^4 \leq a_n^2 u_n^4. \]
The rest of the proof follows from the proof of Theorem 3.2. \( \square \)
6 Conclusion

A common practice in data analysis is to attempt to estimate the best linear predictor via lasso with
its regularization parameter chosen via cross-validation. Unfortunately, no definitive theoretical
results existed as to the effect of choosing the tuning parameter in this data-dependent way.

We have demonstrated that the lasso and group lasso with tuning parameter chosen by cross-
validation are risk consistent relative to the linear oracle predictor, provided the size of the validation
sets, $c_n$, grows with the sample size $n$. Imposing conditions on the joint distribution of the data,
we can achieve the same rate of convergence to the risk of the linear oracle predictor with a
data-dependent tuning parameter as with the optimal, yet inaccessible, oracle tuning parameter.
This shows that if the data analyst is interested in using lasso for prediction, choosing the tuning
parameter via cross-validation is still a good choice.

This work reveals three interesting open questions. First, our results do not apply for leave-
one-out cross-validation as $c_n = 1$ for all $n$ and hence the upper-bounds become trivial. Leave-
one-out cross-validation is more computationally demanding than $K$-fold cross-validation, but is
still used in practice. Results about its behavior would further contribute to our understanding.
Second, direct comparison of the behavior of the lasso estimator with tuning parameter chosen via
cross-validation versus a degrees-of-freedom-based method is of substantial interest and should be
investigated. Finally, our results depend strongly on the upper bound for $\Omega$. However, in most
cases, we will never need to use a tuning parameter this large. So it makes sense to attempt to find
more subtle theory to provide greater intuition for the behavior of lasso under cross-validation.

Acknowledgements

Thanks

References

[1] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007), “Sparsity oracle inequalities for
the lasso,” Electronic Journal of Statistics, 1, 169–194.

[2] Chen, S., Donoho, D., and Saunders, M. (1998), “Atomic decomposition by basis pur-
suit,” SIAM Journal on Scientific Computing, 20(1), 33–61.

[3] Donoho, D., Elad, M., and Temlyakov, V. (2006), “Stable recovery of sparse overcom-
plete representations in the presence of noise,” IEEE Transactions on Information Theory,
52(1), 6–18.

[4] Dudoit, S., and van der Laan, M. J. (2005), “Asymptotics of cross-validation risk
estimation in estimator selection and performance assessment,” Statistical Methodology,
131–154.

[5] Dumbgen, L., van de Geer, S., Veraar, M., and Wellner, J. (2010), “Nemirovski’s
inequalities revisited,” American Mathematical Monthly, 117(2), 138–160.

[6] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least angle regres-
sion,” The Annals of Statistics, 32(2), 407–499.

[7] Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization paths for gen-
eralized linear models via coordinate descent,” Journal of Statistical Software, 33(1), 1–22.
[8] **Fu, W., and Knight, K.** (2000), “Asymptotics for lasso-type estimators,” *Annals of Statistics, 28*(5), 1356–1378.

[9] **Greenshtein, E., and Ritov, Y.** (2004), “Persistence in high-dimensional linear predictor selection and the virtue of overparametrization,” *Bernoulli, 10*(6), 971–988.

[10] **Johnson, C.** (1989), “A Gershgorin-type lower bound for the smallest singular value,” *Linear Algebra Applications, 272*, 169–179.

[11] **Kale, S., Kumar, R., and Vassilvitskii, S.** (2011), “Cross-validation and mean-square stability,” in *Innovations in Computer Science*, pp. 487–495, Tsinghua University Press.

[12] **Lecué, G., and Mitchell, C.** (2012), “Oracle inequalities for cross-validation type procedures,” *Electronic Journal of Statistics, 6*, 1803–1834.

[13] **Meinshausen, N., and Bühlmann, P.** (2006), “High-dimensional graphs and variable selection with the lasso,” *The Annals of Statistics, 34*(3), 1436–1462.

[14] **Meinshausen, N., and Yu, B.** (2009), “Lasso-type recovery of sparse representations for high-dimensional data,” *The Annals of Statistics, 37*(1), 246–270.

[15] **Nardi, Y., and Rinaldo, A.** (2008), “On the asymptotic properties of the group lasso estimator for linear models,” *Electronic Journal of Statistics, 2*, 605–633.

[16] **Osborne, M., Presnell, B., and Turlach, B.** (2000), “On the lasso and its dual,” *Journal of Computational and Graphical statistics, 9*(2), 319–337.

[17] **Rudelson, M., and Vershynin, R.** (2009), “Smallest singular value of a random rectangular matrix,” *Communications on Pure and Applied Mathematics, 62*(12), 1707–1739.

[18] **Shao, J.** (1993), “Linear model selection by cross-validation,” *Journal of the American Statistical Association, 88*, 486–494.

[19] **Stone, M.** (1974), “Cross-validatory choice and assessment of statistical predictions,” *Journal of the Royal Statistical Society. Series B (Methodological), 36*(2), 111–147.

[20] **Stone, M.** (1977), “Asymptotics for and against cross-validation,” *Biometrika, 64*(1), 29–35.

[21] **Tibshirani, R.** (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58*(1), 267–288.

[22] **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.

[23] **Tibshirani, R., and Taylor, J.** (2012), “Degrees of freedom in lasso problems,” *Annals of Statistics, 40*, 1198–1232.

[24] **van de Geer, S.** (2008), “High-dimensional generalized linear models and the lasso,” *The Annals of Statistics, 36*(2), 614–645.

[25] **van de Geer, S., and Bühlmann, P.** (2011), *Statistics for High-Dimensional Data*, Springer Verlag.
[26] Wainwright, M. (2009), “Sharp thresholds for high-dimensional and noisy sparsity recovery using \( \ell_1 \)-constrained quadratic programming (lasso),” *IEEE Transactions on Information Theory*, 55(5), 2183–2202.

[27] Wang, H., and Leng, C. (2007), “Unified lasso estimation by least squares approximation,” *Journal of the American Statistical Association*, 102(479), 1039–1048.

[28] Xu, H., Mannor, S., and Caramanis, C. (2008), “Sparse algorithms are not stable: A no-free-lunch theorem,” in 46th Annual Allerton Conference on Communication, Control, and Computing, pp. 1299–1303, IEEE.

[29] Yuan, M., and Lin, Y. (2005), “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

[30] Zhao, P., and Yu, B. (2006), “On model selection consistency of Lasso,” *The Journal of Machine Learning Research*, 7, 2541–2563.

[31] Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101(476), 1418–1429.

[32] Zou, H., Hastie, T., and Tibshirani, R. (2007), “On the degrees of freedom of the lasso,” *The Annals of Statistics*, 35(5), 2173–2192.