Journal of Machine Learning Research 17 (2016) 1-20 Submitted 11/15; Revised 9/16; Published 12/16

# A Practical Scheme and Fast Algorithm to Tune the Lasso With Optimality Guarantees

**Michaël Chichignoud** MICHAEL.CHICHIGNOUD@GMAIL.COM
*Seminar for Statistics*
*ETH Zürich*

**Johannes Lederer*** LEDERERJ@UW.EDU
*Department of Statistics*
*and Department of Biostatistics*
*University of Washington*

**Martin J. Wainwright** WAINWRIG@BERKELEY.EDU
*Department of Statistics*
*and Department of Electrical Engineering and Computer Sciences*
*University of California at Berkeley*

**Editor:** Francis Bach

## Abstract

We introduce a novel scheme for choosing the regularization parameter in high-dimensional linear regression with Lasso. This scheme, inspired by Lepski's method for bandwidth selection in non-parametric regression, is equipped with both optimal finite-sample guarantees and a fast algorithm. In particular, for any design matrix such that the Lasso has low sup-norm error under an "oracle choice" of the regularization parameter, we show that our method matches the oracle performance up to a small constant factor, and show that it can be implemented by performing simple tests along a single Lasso path. By applying the Lasso to simulated and real data, we find that our novel scheme can be faster and more accurate than standard schemes such as Cross-Validation.

**Keywords:** Lasso, regularization parameter, tuning parameter, high-dimensional regression, oracle inequalities

## 1. Introduction

Regularized estimators—among them the Lasso (Tibshirani, 1996), the Square-Root and the Scaled Lasso (Antoniadis, 2010; Belloni et al., 2011; Städler et al., 2010; Sun and Zhang, 2012), as well as estimators based on nonconvex penalties such as MCP (Zhang, 2010) and SCAD (Fan and Li, 2001)—all hinge on finding a "suitable" choice of tuning parameters. There are many possible methods for solving this so-called calibration problem, but for high-dimensional regression problems, there is a not a single method that is computationally tractable and for which the non-asymptotic theory is well understood.

The focus of this paper is the calibration of the Lasso for sparse linear regression, where the tuning parameter needs to be adjusted to both the noise distribution and the

*. Corresponding author. Postal address: Department of Statistics at University of Washington, Box 354322, Seattle, WA 98195

©2016 Michaël Chichignoud, Johannes Lederer, Martin J. Wainwright.

design matrix (van de Geer and Lederer, 2013; Hebiri and Lederer, 2013; Dalalyan et al., 2014). Calibration schemes for this setting are typically based on Cross-Validation (CV) or BIC-type criteria. However, CV-based procedures can be computationally intensive and are currently lacking in non-asymptotic theory for high-dimensional problems. BIC-type criteria, on the other hand, are computationally simpler but also lacking in non-asymptotic guarantees. Another approach is to replace the Lasso with Square-Root Lasso or TREX (Lederer and Müller, 2015); however, Square-Root Lasso still contains a tuning parameter that needs to be calibrated to certain aspects of the model, and the theory for TREX is currently fragmentary. For these reasons and given the extensive use of the Lasso in practice, understanding the calibration of Lasso is important.

In this paper, we introduce a new scheme for calibrating the Lasso in the supremum norm ($\ell_\infty$)-loss, which we refer to as *Adaptive Validation for* $\ell_\infty$ ($\mathrm{AV}_\infty$). This method is based on tests that are inspired by Lepski's method for non-parametric regression (Lepski, 1990; Lepski et al., 1997), see also Chichignoud and Lederer (2014). In contrast to current schemes for the Lasso, our method is equipped with both optimal theoretical guarantees and a fast computational routine.

The remainder of this paper is organized as follows. In Section 2, we introduce the $\mathrm{AV}_\infty$ method. Our main theoretical results show that this method enjoys finite sample guarantees for the calibration of Lasso with respect to sup-norm loss (Theorem 3) and variable selection (Remark 4). In addition, we provide a simple and fast algorithm (Algorithm 1). In Section 3, we illustrate these features with applications to simulated data and to biological data. We conclude with a discussion in Section 4.

*Notation:* The indicator of events is denoted by $\mathbb{1}\{\cdot\} \in \{0,1\}$, the cardinality of sets by $|\cdot|$, the sup-norm or maximum norm of vectors in $\mathbb{R}^p$ vectors $\|\cdot\|_\infty$, the number of non-zero entries by $\|\cdot\|_0$, the $\ell_1$- and $\ell_2$-norms by $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, and $[p] := \{1, \ldots, p\}$. For given vector $\beta \in \mathbb{R}^p$ and subset $A$ of $[p]$, $\beta_A \in \mathbb{R}^{|A|}$ and $\beta_{A^c} \in \mathbb{R}^{|A^c|}$ denote the components in $A$ and in its complement $A^c$, respectively.

## 2. Background and Methodology

In this section, we introduce some background and then move onto a description of the $\mathrm{AV}_\infty$ method.

### 2.1 Framework

We study the calibration of the Lasso tuning parameter in high-dimensional linear regression models that can contain many predictors and allow for the possibility of correlated and heavy-tailed noise. More specifically, we assume that the data $(Y, X)$ with outcome $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n \times p}$ is distributed according to a linear regression model

$$Y = X\beta^* + \varepsilon, \tag{Model}$$

where $\beta^* \in \mathbb{R}^p$ is the regression vector and $\varepsilon \in \mathbb{R}^n$ is a random noise vector. Our framework allows for $p$ larger than $n$ and requires that the noise variables $\varepsilon$ satisfy only the second moment condition

$$\max_{i \in \{1,\ldots,n\}} \mathbb{E}[\varepsilon_i^2] < \infty. \tag{1}$$

A standard approach for estimating $\beta^*$ in such a model is by computing the $\ell_1$-regularized least-squares estimate, known as the Lasso, and given by

$$\widehat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right\}. \tag{Lasso}$$

Note that this equation actually defines a family of estimators indexed by the tuning parameter $\lambda > 0$, which determines the level of regularization.

Intuitively, the optimal choice of $\lambda$ is dictated by a trade-off between bias and some form of variance control. Bias is induced by the shrinkage effect of the $\ell_1$-regularizer, which acts even on non-zero coordinates of the regression vector. Thus, the bias grows as $\lambda$ is increased. On the other hand, $\ell_1$-regularization is useful in canceling out fluctuations in the score function, which for the linear regression model is given by $X^\top \varepsilon / n$. Thus, an optimal choice of $\lambda$ is the smallest one that is large enough to control these fluctuations.

A large body of theoretical work (e.g., van de Geer and Bühlmann (2009); Bickel et al. (2009); Bühlmann and van de Geer (2011); Negahban et al. (2012)) has shown that an appropriate formalization of this intuition is based on the event

$$\mathcal{T}_\lambda := \left\{ \frac{\|X^\top \varepsilon\|_\infty}{n} \leq \frac{\lambda}{4} \right\}. \tag{2}$$

When this event holds, then as long as the design matrix $X$ is "well-behaved", it is possible to obtain bounds on the sup-norm error of the Lasso estimate. There are various ways of characterizing well-behaved design matrices; of most relevance for sup-norm error control are mutual incoherence conditions (Bunea, 2008; Lounici, 2008) as well as $\ell_\infty$-restricted eigenvalues (Ye and Zhang, 2010). See van de Geer and Bühlmann (2009) and Section 2.3 for further discussion of these design conditions.

In order to bring sharp focus to the calibration problem, rather than focusing on any particular design condition, it is useful to instead work under the generic assumption that the Lasso sup-norm error is controlled under the event $\mathcal{T}_\lambda$ defined in equation (2). More formally, we state:

**Assumption 1** ($\ell_\infty(C)$) *There is a numerical constant $C$ such that conditioned on $\mathcal{T}_\lambda$, the Lasso $\ell_\infty$-error is upper bounded as $\|\widehat{\beta}_\lambda - \beta^*\|_\infty \leq C\lambda$.*

As mentioned above, there are many conditions on the design matrix $X$ under which Assumption $\ell_\infty(C)$ is valid, and we consider a number of them in the sequel.

With this set-up in place, we can now focus specifically on how to choose the regularization parameter. Since we can handle only finitely many tuning parameters in practice, we restrict ourselves to the selection of a tuning parameter among a finite but arbitrarily large number of choices. It is easy to see that $\lambda_{\max} := 2\|X^\top Y\|_\infty / n$ is the smallest tuning parameter for which $\widehat{\beta}_\lambda$ equals zero. Accordingly, for a given positive integer $N \in \mathbb{N}$, let us form the grid

$$0 < \lambda_1 < \cdots < \lambda_N = \lambda_{\max},$$

denoted by $\Lambda := \{\lambda_1, \ldots, \lambda_N\}$ for short. Assumption $\ell_\infty(C)$ guarantees that the sup-norm error is proportional to $\lambda$ whenever the event $\mathcal{T}_\lambda$ holds; consequently, for a given probability

of error $\delta \in (0,1)$, it is natural to choose the smallest $\lambda$ for which event $\mathcal{T}_\lambda$ holds with probability at least $1-\delta$, assuming that it is finite. This criterion can be formalized as follows:

**Definition 1 (Oracle tuning parameter)** *For any constant $\delta \in (0,1)$, the oracle tuning parameter is given by*

$$\lambda_\delta^* := \arg\min_{\lambda \in \Lambda} \left\{ \mathbb{P}\left(\mathcal{T}_\lambda\right) \geq 1-\delta \right\}. \tag{3}$$

Note that by construction, if we solve the Lasso using the oracle choice $\lambda_\delta^*$, and if the design matrix $X$ fulfills Assumption $\ell_\infty(C)$, then the resulting estimate satisfies the bound $\|\widehat{\beta}_{\lambda_\delta^*} - \beta^*\|_\infty \leq C\lambda_\delta^*$ with probability at least $1-\delta$. Unfortunately, the oracle choice is inaccessible to us, since we cannot compute the probability of the event $\mathcal{T}_\lambda$ based on the observed data. However, as we now describe, we can mimic this performance, up to a factor of three, using a simple data-dependent procedure.

### 2.2 Adaptive Calibration Scheme

Let us now describe a data-dependent scheme for choosing the regularization parameter, referred to as Adaptive Calibration for $\ell_\infty$ ($\mathrm{AV}_\infty$):

**Definition 2 ($\mathbf{AV}_\infty$)** *Under Assumption $\ell_\infty(C)$ and for a given constant $\overline{C} \geq C$, Adaptive Calibration for $\ell_\infty$ ($AV_\infty$) selects the tuning parameter*

$$\hat{\lambda} := \min \left\{ \lambda \in \Lambda \,\Big|\, \max_{\substack{\lambda', \lambda'' \in \Lambda \\ \lambda', \lambda'' \geq \lambda}} \left[ \frac{\|\widehat{\beta}_{\lambda'} - \widehat{\beta}_{\lambda''}\|_\infty}{\lambda' + \lambda''} - \overline{C} \right] \leq 0 \right\}. \tag{4}$$

The definition is based on tests for sup-norm differences of Lasso estimates with different tuning parameters. We stress that Definition 2 requires neither prior knowledge about the regression vector nor about the noise.

The tests in Definition 2 can be formulated in terms of the binary random variables

$$\widehat{t}_{\lambda_j} := \prod_{k=j}^{N} \mathbb{1}\left\{ \frac{\|\widehat{\beta}_{\lambda_j} - \widehat{\beta}_{\lambda_k}\|_\infty}{\lambda_j + \lambda_k} - \overline{C} \leq 0 \right\} \qquad \text{for } j \in [N],$$

from the $\mathrm{AV}_\infty$ tuning parameter $\hat{\lambda}$ can be computed as follows:

**Data:** $\widehat{\beta}_{\lambda_1}, \ldots, \widehat{\beta}_{\lambda_N}, \overline{C}$
**Result:** $\hat{\lambda} \in \Lambda$

Set initial index: $j \leftarrow N$

**while** $\widehat{t}_{\lambda_{j-1}} \neq 0$ *and* $j > 1$ **do**
  Update index: $j \leftarrow j-1$
**end**
Set output: $\hat{\lambda} \leftarrow \lambda_j$

**Algorithm 1:** Algorithm for $\mathrm{AV}_\infty$ in Definition 2.

This algorithm can be readily implemented and only requires the computation of one Lasso solution path. In strong contrast, $k$-fold Cross-Validation requires the computation of $k$ solution paths. Consequently, the Lasso with $\mathrm{AV}_\infty$ can be computed about $k$ times faster than Lasso with $k$-fold Cross-Validation.

The following result guarantees that the Lasso with $\mathrm{AV}_\infty$ method achieves the sup-norm error up to a constant pre-factor:

**Theorem 3 (Optimality of $\mathbf{AV}_\infty$ )** *Suppose that condition $\ell_\infty(C)$ holds and the $AV_\infty$ method is implemented with parameter $\overline{C} \geq C$. Then for any $\delta \in (0,1)$, the $AV_\infty$ output pair $(\hat{\lambda}, \widehat{\beta}_{\hat{\lambda}})$ given by the rule* (4) *satisfies the bounds*

$$\hat{\lambda} \leq \lambda^*_\delta \quad \textit{and} \quad \|\widehat{\beta}_{\hat{\lambda}} - \beta^*\|_\infty \leq 3\overline{C}\lambda^*_\delta \tag{5}$$

*with probability at least $1-\delta$.*

**Remark 4 (Relevance for estimation and variable selection)** *The $\ell_\infty$-bound from equation* (5) *directly implies that the $AV_\infty$ scheme is adaptively optimal for the estimation of the regression vector $\beta^*$ in $\ell_\infty$-loss. As another important feature, Theorem 3 entails strong variable selection guarantees. First, the $\ell_\infty$-bound implies that $AV_\infty$ recovers all non-zero entries of the regression vector $\beta^*$ that are larger than $3\overline{C}\lambda^*_\delta$ in absolute value. Additionally, by virtue of the bound $\hat{\lambda} \leq \lambda^*_\delta$, thresholding $\widehat{\beta}_{\hat{\lambda}}$ by $3\overline{C}\hat{\lambda}$ leads to exact support recovery if all non-zero entries of $\beta^*$ are larger than $6\overline{C}\lambda^*_\delta$ in absolute value. In strong contrast, standard calibration schemes are not equipped with comparable variable selection guarantees, and there is no theoretically sound guidance for how to threshold standard schemes.*

We prove Theorem 3 in Appendix A; here let us make a few remarks about its consequences. First, if we knew the oracle value $\lambda^*_\delta$ defined in equation (3), then under Assumption $\ell_\infty(C)$, the Lasso estimate $\widehat{\beta}$ would satisfy the $\ell_\infty$-bound $\|\widehat{\beta} - \beta^*\|_\infty \leq C\lambda^*_\delta$. Consequently, when the $\mathrm{AV}_\infty$ method is implemented with parameter $C$, then its sup-norm error is optimal up to a factor of three. For standard calibration schemes, among them Cross-Validation, no comparable guarantees are available in the literature. In fact, we are not aware of *any* finite sample guarantees for standard calibration schemes.

We point out that Theorem 3—in contrast to asymptotic results or results with unspecified constants—provides explicit guarantees for arbitrary sample sizes. Moreover, Theorem 3 does not presume prior knowledge about the regression vector or the noise distribution and allows, in particular, for correlated, heavy-tailed noise. From the perspective of theoretical sharpness, the best choice for $\overline{C}$ is $\overline{C} = C$. However, Theorem 3 shows that it also suffices to know an upper bound for $C$. We provide more details on choices of $C$ and $\overline{C}$ below.

We finally observe that the specific choice of the grid enters Theorem 3 only via the oracle. Indeed, for any choice of the grid, Theorem 3 ensures that $\hat{\lambda}$ performs as well as the oracle tuning parameter $\lambda^*_\delta$, which is the "best" tuning parameter on the grid.

### 2.3 Conditions on the Design Matrix for $\ell_\infty$-guarantees

Let us now describe some conditions on the design matrix $X$ that are sufficient for Assumption $\ell_\infty(C)$. We stress that these are conditions to ensure that the *Lasso* satisfies

$\ell_\infty$-bounds; importantly, our method itself does not impose any additional restrictions. We defer all proofs of the results stated here to Appendix B and, for simplicity, we assume in the following that the sample covariance $\widehat{\Sigma} := X^\top X/n$ has been normalized such that $\widehat{\Sigma}_{jj} = 1$ for all $j \in [p]$.

The significance of the event $\mathcal{T}_\lambda$ lies in the following implication: when $\mathcal{T}_\lambda$ holds, then it can be shown (e.g.,Bickel et al. (2009); Bühlmann and van de Geer (2011); Negahban et al. (2012)) that the Lasso error $\widehat{\Delta} := \widehat{\beta}_\lambda - \beta^*$ must belong to the cone

$$\mathbb{C}(S) := \big\{\Delta \in \mathbb{R}^p \mid \|\Delta_{S^c}\|_1 \leq 2\|\Delta_S\|_1\big\}, \tag{6}$$

where $S$ denotes the support of $\beta^*$, and $S^c$ its complement. Accordingly, all known conditions involve controlling the behavior of the sample covariance matrix $\widehat{\Sigma}$ for vectors lying within this cone.

The most directly stated sufficient condition is based on lower bounding the $\ell_\infty$*-restricted eigenvalue:* there exists some $\gamma > 0$ such that

$$\|\widehat{\Sigma}\Delta\|_\infty \geq \gamma\|\Delta\|_\infty \qquad \text{for all } \Delta \in \mathbb{C}(S). \tag{7}$$

See van de Geer and Bühlmann (2009) for an overview of various conditions for the Lasso, and their relations. Based on (7), we prove in Appendix B.1 the following result:

**Lemma 5 ($\ell_\infty$-restricted eigenvalue)** *Suppose that $\widehat{\Sigma}$ satisfies the $\gamma$-RE condition (7) and that $\mathcal{T}_\lambda$ holds. Then Assumption $\ell_\infty(C)$ is valid with $C = \frac{5}{4\gamma}$.*

Although this result is cleanly stated, the RE condition cannot be verified in practice, since it involves the unknown support set $S$. Accordingly, let us now state some sufficient and verifiable conditions for obtaining bounds on the restricted eigenvalues, and hence for verifying Assumption $\ell_\infty(C)$.

For a given integer $\tilde{s} \in [2, p]$ and scalar $\nu > 0$, let us say that the sample covariance $\widehat{\Sigma}$ is diagonally dominant with parameters $(\tilde{s}, \nu)$ if

$$\max_{\substack{|T|=\tilde{s} \\ T \subset [p]\setminus\{j\}}} \sum_{k \in T} |\widehat{\Sigma}_{jk}| < \nu \qquad \text{for all } j \in [p]. \tag{8}$$

In the context of this definition, the reader should recall that we have assumed that $\widehat{\Sigma}_{jj} = 1$ for all $j \in [p]$. Note that this condition can be verified in polynomial-time, since the subset $T$ achieving the maximum in row $j$ can be obtained simply by sorting the entries $\{|\widehat{\Sigma}_{jk}|, k \in [p]\setminus j\}$. The significance of this condition lies in the following result:

**Lemma 6 (Diagonal dominance of order $\tilde{s}$)** *Suppose that $\tilde{s} \geq 9|S|$ and $\widehat{\Sigma}$ is $\tilde{s}$-order diagonally dominant with parameter $\nu \in [0, 1)$. Then under the event $\mathcal{T}_\lambda$, Assumption $\ell_\infty(C)$ is valid with $C = \frac{5}{4(1-\nu)}$.*

See Appendix B.2 for the proof.

It is worth noting that the diagonal dominance condition is weaker than the pairwise incoherence conditions that have been used in past work on sup-norm error (Lounici, 2008). The pairwise incoherence of the sample covariance is given by $\rho(\widehat{\Sigma}) = \max_{j \neq k} |\widehat{\Sigma}_{jk}|$. If the pairwise incoherence satisfies the bound $\rho(\widehat{\Sigma}) \leq \nu/\tilde{s}$, then it follows that $\widehat{\Sigma}$ is diagonally dominant with parameters $(\tilde{s}, \nu)$.

By combining Lemma 6 with Theorem 3, we obtain the following corollary:

**Corollary 7** *Suppose that $\tilde{s} \geq 9|S|$ and $\widehat{\Sigma}$ is $\tilde{s}$-order diagonally dominant with parameter $\nu \in [0,1)$. Then for any $\delta \in (0,1)$, the $AV_\infty$ method with $\overline{C} = \frac{5}{4(1-\nu)}$ returns an estimate $\widehat{\beta}_{\hat{\lambda}}$ such that*

$$\|\widehat{\beta}_{\hat{\lambda}} - \beta^*\|_\infty \leq \frac{15}{4(1-\nu)}\lambda_\delta^* \tag{9}$$

*with probability at least $1-\delta$.*

Another sufficient condition for the sup-norm optimality of $\mathrm{AV}_\infty$is a design compatibility condition due to van de Geer (2007). For each index $j \in [p]$, suppose that we define the deterministic vector

$$\eta^j \in \arg\min_{\substack{\beta \in \mathbb{R}^p \\ \beta_j = -1}} \left\{ \frac{\|X\beta\|_2^2}{n} + \sqrt{\frac{\log(p)}{n}}\|\beta\|_1 \right\}.$$

Note that this optimization problem defining the vector regression of the $j$th column of the design matrix on the set of all other columns, where we have imposed an $\ell_1$-penalty with weight $\sqrt{\frac{\log(p)}{n}}$. We can then derive the following sup-norm bound for the Lasso.

**Lemma 8 (Lasso bound under compatibility)** *Assume that $X$ fulfills the compatibility condition*

$$\min_{\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1} \left\{ \frac{\sqrt{|S|}\|X\beta\|_2}{\sqrt{n}\|\beta_S\|_1} \right\} \geq t \tag{Compatibility}$$

*for a constant $t > 0$. Additionally, assume that*

$$\sup_{j \in [p]} \frac{|S|}{t^2\|\eta^j\|_1} \leq \frac{1}{\log n}\sqrt{\frac{n}{\log p}}.$$

*Then under the event $\mathcal{T}_\lambda$, Assumption $\ell_\infty(C)$ is valid with*

$$C := \left(\frac{3}{4} + \frac{1}{\log(n)}\right) \max_{j \in [p]} \frac{\|\eta^j\|_1}{\|X\eta^j\|_2^2/n + \sqrt{\log(p)/n}\|\eta^j\|_{-j}/2}.$$

This bound is a consequence of results in (van de Geer, 2014); the proof is deferred to Section B.3. We are now ready to state the optimality of $\mathrm{AV}_\infty$ with respect to this bound.

**Corollary 9 (Optimality of $\mathbf{AV}_\infty$)** *Assume that the assumptions in Lemma 8 are met. Then for any constant $\delta > 0$, the following bound for Lasso $AV_\infty$ with $\overline{C} = C$, and $C$ as above, holds with probability at least $1-\delta$:*

$$\|\widehat{\beta}_{\hat{\lambda}} - \beta^*\|_\infty \leq 3C\lambda_\delta^*. \tag{10}$$

This result demonstrates the optimality of $\mathrm{AV}_\infty$ for sup-norm loss under the compatibility condition.

**Remark 10 (Constant $\overline{C}$ in practice)** *The optimal choice is $\overline{C} = C$ in view of our theoretical results. The constant $C$ (or an upper bound of it) can be readily computed, because it depends only on $X$ (cf. Lemma 8) or on $X$ and an upper bound on $s$ (cf. Lemma 6). However, we propose the universal choice $C = 0.75$ for all practical purposes. Note that accurate support recovery and $\ell_\infty$-estimation is possible only if the design is near orthogonal. A direct computation yields the bound $\|\widehat{\beta}_\lambda - \beta^*\|_\infty \leq C\lambda$ with $C = 0.75$ for orthogonal design. Letting $\alpha \to \infty$ in Theorem 1 due to Lounici (2008) yields the same bound with $C \approx 0.75$ for near orthogonal designs. This provides strong theoretical support for the choice $\overline{C} = 0.75$. The empirical evidence in Section 3 indicates that a further calibration is indeed not necessary.*

## 3. Simulations

In this section, we perform experiments on both simulated and real data to demonstrate the practical performance of $\mathrm{AV}_\infty$.

### 3.1 Simulated Data

We simulate data from linear regression models as in equation (Model) with $n = 200$ observations and $p \in \{300, 900\}$ parameters. More specifically, we sample each row of the design matrix $X \in \mathbb{R}^{n\times p}$ from a $p$-dimensional normal distribution with mean 0 and covariance matrix $(1-\kappa)\,\mathrm{I} + \kappa \mathbb{1}$, where I is the identity matrix, $\mathbb{1} := (1,\ldots,1)^\top(1,\ldots,1)$ is the matrix of ones, and $\kappa \in \{0, 0.2, 0.4\}$ is the magnitude of the mutual correlations. For the entries of the noise $\varepsilon \in \mathbb{R}^n$, we take the one-dimensional normal distribution with mean 0 and variance 1. The entries of $\beta^*$ are first set to 0 except for 6 uniformly at random chosen entries that are each set to 1 or $-1$ with equal probability. The entire vector $\beta^*$ is then rescaled such that the signal-to-noise ratio $\|X\beta^*\|_2^2/n$ is equal to 5. We finally consider a grid of 100 tuning parameters $\Lambda := \{\lambda_{\max}/1.3^0, \lambda_{\max}/1.3^1, \ldots, \lambda_{\max}/1.3^{99}\}$ with $\lambda_{\max} := 2\|X^\top Y\|_\infty/n$. We run 100 experiments for each set of parameters and report the corresponding means (thick, colored bars) and standard deviations (thin, black lines). All computations are conducted with the software R (R Core Team, 2013) and the glmnet package (Friedman et al., 2010). While we restrict the presentation to the parameter settings described, we found similar results over a wide range of settings.

We compare the sup-norm and variable selection performance of the following three procedures:

- Oracle: Lasso with the tuning parameter that minimizes the $\ell_\infty$ loss (*this tuning parameter is unknown in practice*);
- $\mathrm{AV}_\infty$: Lasso with $\mathrm{AV}_\infty$ and $\overline{C} = 0.75$;
- Cross-Validation: Lasso with 10-fold Cross-Validation.

Our choice $\overline{C} = 0.75$ is motivated by a theorem due to Lounici (2008) in the regime $\alpha \to \infty$; see Remark 10 for details.

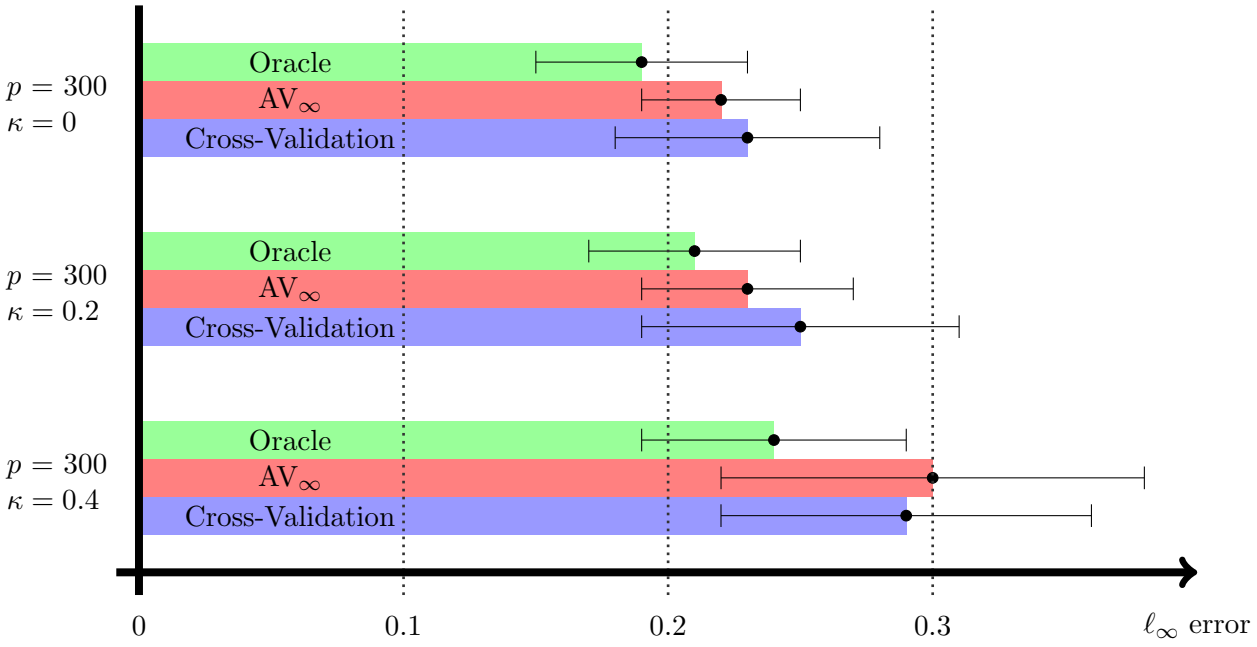


Figure 1: Sup-norm error $\|\widehat{\beta}_\lambda - \beta^*\|_\infty$ of the Lasso with three different calibration schemes for the tuning parameter $\lambda$. Depicted are the results for three simulation settings that differ in the correlation level $\kappa$. The simulation settings and the calibration schemes are specified in the body of the text.

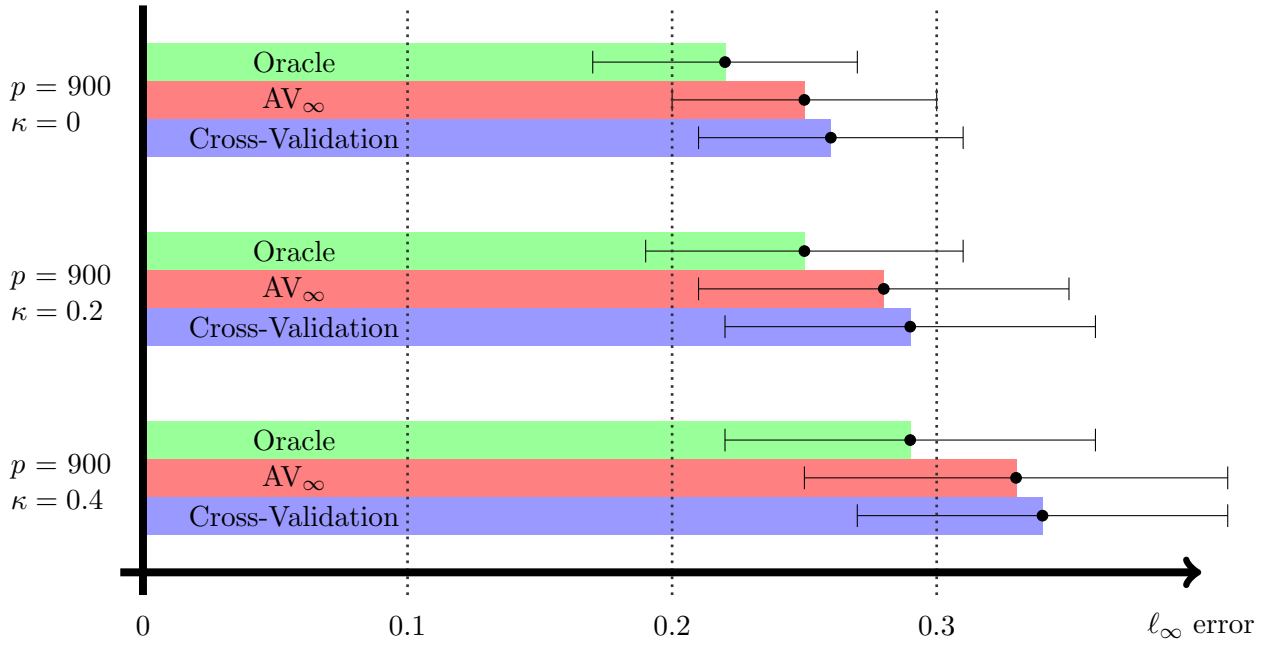


Figure 2: Sup-norm error $\|\widehat{\beta}_\lambda - \beta^*\|_\infty$ of the Lasso with three different calibration schemes for the tuning parameter $\lambda$. Depicted are the results for three simulation settings that differ in the correlation level $\kappa$. The simulation settings and the calibration schemes are specified in the body of the text.

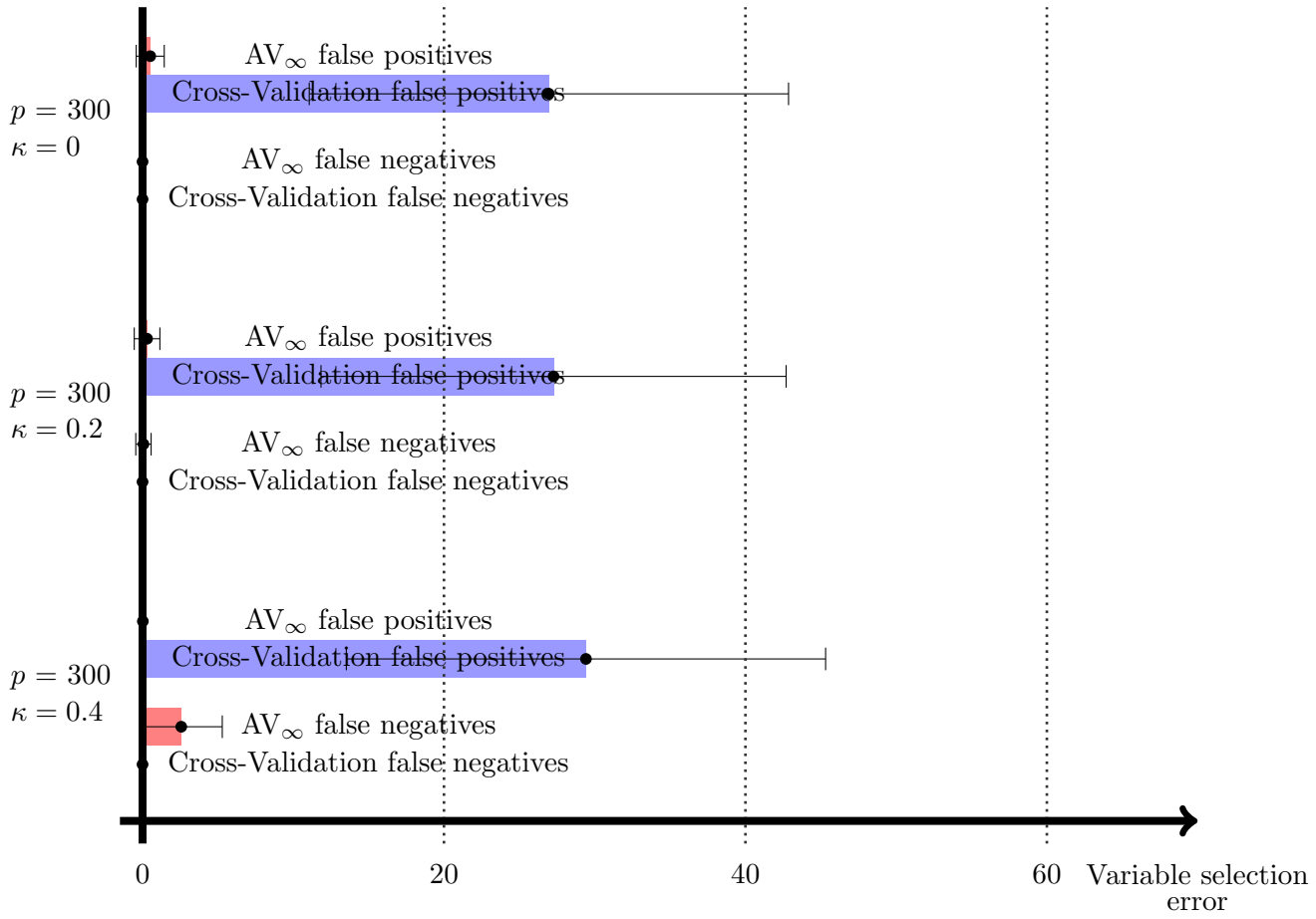


Figure 3: Number of false positives $|\{j : \beta_j^* = 0, (\widehat{\beta}_\lambda)_j \neq 0\}|$ and false negatives $|\{j : \beta_j^* \neq 0, (\widehat{\beta}_\lambda)_j = 0\}|$ of the Lasso with $\mathrm{AV}_\infty$ and Cross-Validation as calibration schemes for the tuning parameter $\lambda$. For $\mathrm{AV}_\infty$, the safe threshold described after Theorem 3 is applied. The simulations settings correspond to those in Figure 1.

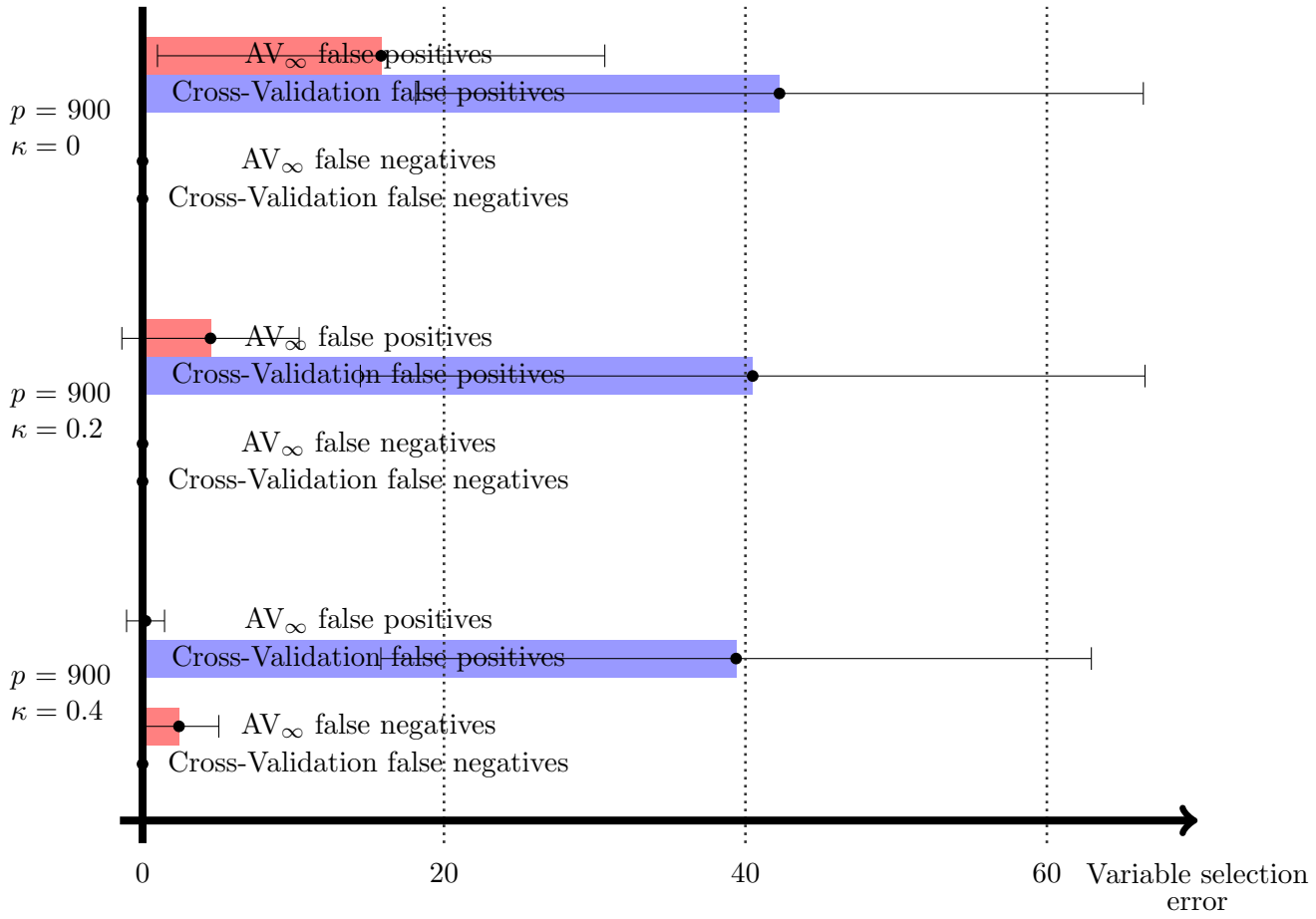


Figure 4: Number of false positives $|\{j : \beta_j^* = 0, (\widehat{\beta}_\lambda)_j \neq 0\}|$ and false negatives $|\{j : \beta_j^* \neq 0, (\widehat{\beta}_\lambda)_j = 0\}|$ of the Lasso with $\mathrm{AV}_\infty$ and Cross-Validation as calibration schemes for the tuning parameter $\lambda$. For $\mathrm{AV}_\infty$, the safe threshold described after Theorem 3 is applied. The simulations settings correspond to those in Figure 2.

*Sup-norm error:* In Figures 1 and 2, we compare the $\ell_\infty$ error of the four procedures. We observe that $\mathrm{AV}_\infty$ outperforms Cross-Validation for most settings under consideration. We also mention that the same conclusions can be drawn if the normal distribution for the noise is replaced by other, possibly heavy-tailed distributions (for conciseness, we do not show the outputs).

*Variable selection:* In Figures 3 and 4, we compare the variable selection performance of $\mathrm{AV}_\infty$ and Cross-Validation. More specifically, we compare the number of false positives $|\{j : \beta_j^* = 0, (\widehat{\beta}_\lambda)_j \neq 0\}|$ and the number of false negatives $|\{j : \beta_j^* \neq 0, (\widehat{\beta}_\lambda)_j = 0\}|$. In contrast to Cross-Validation, $\mathrm{AV}_\infty$ allows for a safe threshold of size $3\overline{C}\hat{\lambda}$ (recall the discussion after Theorem 3). Therefore, we report the results of Lasso with $\mathrm{AV}_\infty$ and an additional threshold of size $3\overline{C}\hat{\lambda}$ applied to each component (that is, we consider the vector with entries $(\widehat{\beta}_\lambda)_j \mathbb{1}\{|(\widehat{\beta}_\lambda)_j| \geq 3\overline{C}\hat{\lambda}\}$ ), and we report the results of Lasso with Cross-Validation (without threshold). We observe that, as compared to Cross-Validation, $\mathrm{AV}_\infty$ with subsequent thresholding can lead to a considerably smaller number of false positives, while keeping the number of false negatives on a low level. Note that one could perform a similar thresholding of the Cross-Validation solution, but unlike for $\mathrm{AV}_\infty$, there is no theory to guide the choice of the threshold. This problem also applies to other standard calibration schemes.

*Computational complexity:* Cross-Validation with 10 folds requires the computation 10 Lasso paths, while $\mathrm{AV}_\infty$ requires the computation of only one Lasso path - or even less. $\mathrm{AV}_\infty$ is therefore about 10 times more efficient than 10-fold Cross-Validation.

Let us conclude with remarks on the scope of the simulations. First, many methods have been proposed for tuning the regularization parameter in the Lasso, including Cross-Validation, BIC and AIC-type criteria, Stability Selection (Meinshausen and Bühlmann, 2010), LinSelect (Baraud et al., 2014; Giraud et al., 2012), permutation approaches (Sabourin et al., 2015), and many more. On top of that, there are many modifications and extensions of the Lasso itself, including BoLasso (Bach, 2008), Square-Root/Scaled Lasso (Antoniadis, 2010; Belloni et al., 2011; Städler et al., 2010; Sun and Zhang, 2012), SCAD (Fan and Li, 2001), MCP (Zhang, 2010), and others. Detailed comparisons among the selection schemes and the methods can be found in the cited papers. We also refer to Leeb and Pötscher (2008) for theoretical insights about limitations of the methods.

In our simulations, we instead focus on the Lasso and, since we are not aware of guarantees similar to ours for any selection scheme, we compare to the most popular and most extensively studied selection scheme, Cross-Validation. This comparison shows that, beyond its theoretical properties and the easy and efficient implementation, $\mathrm{AV}_\infty$ is also a competitor in numerical experiments.

### 3.2 Riboflavin Production in *B. subtilis*

We now consider variable selection for a data set that describes the production of riboflavin (vitamin $B_2$) in *B. subtilis* (*Bacillus subtilis*), see (Bühlmann et al., 2014). The data set comprises the expressions of $p = 4088$ genes and the corresponding riboflavin production rates for $n = 71$ strains of *B. subtilis*. We apply $\mathrm{AV}_\infty$ and then impose the threshold $3\overline{C}\hat{\lambda}$.

The resulting genes and the corresponding parameter values are given in the first column of Table 1. We see that these results commensurate with the results from previous

| $AV_\infty$ | Stability Selection | B-TREX |
|---|---|---|
| YXLD_at -0.405 | YXLD_at | YXLD_at |
| YOAB_at -0.420 | YOAB_at | YOAB_at |
| YEBC_at -0.146 | LYSC_at | YXLE_at |
| ARGF_at -0.313 | | |
| XHLB_at 0.278 | | |

Table 1: Variable selection results for the riboflavin data set. The first column depicts the genes and the corresponding parameter values yielded by $AV_\infty$. The second and third column depict the genes returned by approaches based on Stability Selection and TREX.

approaches based on Stability Selection (Bühlmann et al., 2014) and TREX (Lederer and Müller, 2015), which are given in the third and fourth column.

## 4. Conclusions

We have introduced a novel method for sup-norm calibration, known as $AV_\infty$, that is equipped with finite sample guarantees for estimation in $\ell_\infty$-loss and for variable selection. Moreover, we have shown that $AV_\infty$ allows for simple and fast implementations. These properties make $AV_\infty$ a competitive algorithm, as standard methods such as Cross-Validation are computationally more demanding and lack non-asymptotic guarantees.

In order to bring sharp focus to the issue, we have focused this paper exclusively on the calibration of the Lasso. However, we suspect that the methods and techniques developed here could be more generally applicable, for instance to problems with nonconvex penalties (e.g., SCAD, MCP). In particular, the paper (Loh and Wainwright, 2014) provides guarantees for $\ell_\infty$-recovery using such nonconvex methods, which could be combined with our results. Another interesting direction for future work is the use of our methods for more general decomposable penalty functions (Negahban et al., 2012), including the nuclear norm that is often used in matrix estimation.

We also stress that our goals are $\ell_\infty$-estimation and variable selection, which are feasible only under strict conditions on the design matrix. Other objectives, including prediction and $\ell_2$-estimation, can typically be achieved under less stringent conditions. However, the corresponding oracle inequalities contain quantities (such as the sparsity level) that are typically unknown in practice. Adaptations of our method to objectives beyond the ones considered here thus need further investigation. We refer to (Chételat et al., 2014) for ideas in this direction. However, there might be no approach that is uniformly optimal for all objectives, see also the papers (Yang, 2005; Zhao and Yu, 2006).

Finally, as pointed out by one of the reviewers, another field for further study is model misspecification. It would be interesting to see how robust the Lasso with the $AV_\infty$ scheme is with respect to, for example, non-linearities in the model.

## Acknowledgments

We thank Sara van de Geer and Sébastien Loustau for the inspiring discussions. We also thank the reviewers for the careful reading of the manuscript and the insightful comments. This work was partially supported by NSF Grant DMS-1107000, and Air Force Office of Scientific Research AFOSR-FA9550-14-1-0016 to MJW.

## Appendix A. Proof of Theorem 1

Define the event $\mathcal{T}_\delta^* := \left\{ \frac{\|X^\top \varepsilon\|_\infty}{n} \leq \frac{\lambda_\delta^*}{4} \right\}$ and note that $\mathbb{P}[\mathcal{T}_\delta^*] \geq 1 - \delta$ by our definition of the oracle tuning parameter in (3). Thus, it suffices to show that the two bounds hold conditioned on the event $\mathcal{T}_\delta^*$.

**Bound on $\hat{\lambda}$:** To show that $\hat{\lambda} \leq \lambda_\delta^*$, we proceed by proof by contradiction. If $\hat{\lambda} > \lambda_\delta^*$, then the definition of the $\mathrm{AV}_\infty$ method implies that there must exist two tuning parameters $\lambda', \lambda'' \geq \lambda_\delta^*$ such that

$$\|\widehat{\beta}_{\lambda'} - \widehat{\beta}_{\lambda''}\|_\infty > \overline{C}\,(\lambda' + \lambda''). \tag{11}$$

However, since $\mathcal{T}_{\lambda'}$ and $\mathcal{T}_{\lambda''}$ are both subsets of $\mathcal{T}_\delta^*$, Assumption $\ell_\infty(C)$ implies that we must have the simultaneous inequalities $\|\widehat{\beta}_{\lambda'} - \beta^*\|_\infty \leq C\lambda'$ and $\|\widehat{\beta}_{\lambda''} - \beta^*\|_\infty \leq C\lambda''$. Combined with the triangle inequality, we find that

$$\|\widehat{\beta}_{\lambda'} - \widehat{\beta}_{\lambda''}\|_\infty \leq \|\widehat{\beta}_{\lambda'} - \beta^*\|_\infty + \|\beta^* - \widehat{\beta}_{\lambda''}\|_\infty \;\leq\; C\,(\lambda' + \lambda'').$$

Since $\overline{C} \geq C$, this upper bound contradicts our earlier conclusion (11) and, therefore, yields the desired claim.

**Bound on the sup-norm error:** On the event $\mathcal{T}_\delta^*$, we have $\hat{\lambda} \leq \lambda_\delta^*$, and so the $\mathrm{AV}_\infty$ definition implies that

$$\|\widehat{\beta}_{\hat{\lambda}} - \widehat{\beta}_{\lambda_\delta^*}\|_\infty \leq \overline{C}\left(\hat{\lambda} + \lambda_\delta^*\right) \;\leq\; 2\overline{C}\lambda_\delta^*.$$

Combined with the triangle inequality, we find that

$$\|\widehat{\beta}_{\hat{\lambda}} - \beta^*\|_\infty \leq \|\widehat{\beta}_{\hat{\lambda}} - \widehat{\beta}_{\lambda_\delta^*}\|_\infty + \|\widehat{\beta}_{\lambda_\delta^*} - \beta^*\|_\infty \leq 2\overline{C}\lambda_\delta^* + \|\widehat{\beta}_{\lambda_\delta^*} - \beta^*\|_\infty.$$

Finally, under $\mathcal{T}_\delta^*$ and $\overline{C} \geq C$, Assumption $\ell_\infty(C)$ implies that $\|\widehat{\beta}_{\lambda_\delta^*} - \beta^*\|_\infty \leq C\lambda_\delta^* \leq \overline{C}\lambda_\delta^*$, and combining the pieces completes the proof. ■

## Appendix B. Remaining Proofs for Section 2

In this appendix, we provide the proofs of Lemmas 5, 6, and 8.

### B.1 Proof of Lemma 5

By the first-order stationarity conditions for an optimum, the Lasso solution $\widehat{\beta}_\lambda$ must satisfy the stationary condition $\frac{1}{n}X^\top\big(X\widehat{\beta}_\lambda - Y\big) + \lambda\widehat{z} = 0$, where $\widehat{z} \in \mathbb{R}^p$ belongs to the subdifferential of the $\ell_1$-norm at $\widehat{\beta}_\lambda$. Since $Y = X\beta^* + \varepsilon$, we find that

$$\widehat{\Sigma}\big(\widehat{\beta}_\lambda - \beta^*\big) = -\lambda\widehat{z} + \frac{X^\top\varepsilon}{n}.$$

Taking the $\ell_\infty$-norm of both sides and applying the triangle inequality yields

$$\|\widehat{\Sigma}\big(\widehat{\beta}_\lambda - \beta^*\big)\|_\infty \leq \lambda\|\widehat{z}\|_\infty + \left\|\frac{X^\top\varepsilon}{n}\right\|_\infty \leq \lambda + \frac{\lambda}{4} \;=\; \frac{5}{4}\lambda,$$

using the bound from event $\mathcal{T}_\lambda$, and the fact that $\|\widehat{z}\|_\infty \leq 1$, by definition of the $\ell_1$-subdifferential. As noted previously, under the event $\mathcal{T}_\lambda$, the error vector $\widehat{\Delta} = \widehat{\beta}_\lambda - \beta^*$ belongs to the cone $\mathbb{C}(S)$ in (6), so that the $\gamma$-RE condition can be applied so as to obtain the lower bou $\|\widehat{\Sigma}(\widehat{\beta}_\lambda - \beta^*)\|_\infty \geq \gamma\|\widehat{\beta}_\lambda - \beta^*\|_\infty$. Combining the pieces concludes the proof. ■

### B.2 Proof of Lemma 6

Since $\Delta \in \mathbb{C}(S)$, we have

$$\|\Delta\|_1^2 \leq 9\|\Delta_S\|_1^2 \leq 9|S|\|\Delta_S\|_2^2 \leq 9|S|\|\Delta\|_2^2 \leq 9|S|\|\Delta\|_1\|\Delta\|_\infty,$$

which implies $\|\Delta\|_1 \leq 9|S|\|\Delta\|_\infty$. In view of Lemma 5, it thus suffices to prove the lower bound

$$\|\widehat{\Sigma}\Delta\|_\infty \geq (1-\nu)\|\Delta\|_\infty \qquad \text{for all } \Delta \in A := \mathbb{B}_1(9|S|) \cap \mathbb{B}_\infty(1), \tag{12}$$

where we set $\mathbb{B}_d(r) := \{\beta \in \mathbb{R}^p : \|\beta\|_d \leq r\}$ for $d \in [0,\infty]$ and $r \geq 0$. We claim that

$$\underbrace{\mathbb{B}_1(9|S|) \cap \mathbb{B}_\infty(1)}_{A} \subseteq \underbrace{2\,\mathrm{cl\,conv}\,\big\{\mathbb{B}_0(9|S|) \cap \mathbb{B}_\infty(1)\big\}}_{B}, \tag{13}$$

where cl conv denotes the closed convex hull. Taking this as given for the moment, let us use it to prove the desired claim. We have

$$\max_{\Delta\in A}\frac{\|(\widehat{\Sigma}-\mathrm{I})\Delta\|_\infty}{\|\Delta\|_\infty} = \max_{\Delta\in A/2}\frac{\|(\widehat{\Sigma}-\mathrm{I})\Delta\|_\infty}{\|\Delta\|_\infty} \leq \max_{\Delta\in B}\frac{\|(\widehat{\Sigma}-\mathrm{I})\Delta\|_\infty}{\|\Delta\|_\infty} \;\leq \max_{j\in[p]}\max_{\substack{|T|=9|S|\\ T\subset[p]\setminus j}}\sum_{k\in T}|\widehat{\Sigma}_{jk}| \;\leq\; \nu \tag{14}$$

using the diagonal dominance (8). Combined with the triangle inequality, the lower bound (12) follows.

It remains to prove the inclusion (13). Since both $A$ and $B$ are closed and convex, it suffices to prove that $\phi_A(\theta) \leq \phi_B(\theta)$ for all $\theta \in \mathbb{R}^p$, where $\phi_A(\theta) := \sup_{z\in A}\langle z, \theta\rangle$ and $\phi_B(\theta) := \sup_{z\in B}\langle z, \theta\rangle$ are the support functions. For a given vector $\theta \in \mathbb{R}^p$, let $T$ be the

subset indexing its top $9|S|$ values in absolute value. By construction, we are guaranteed to have the bound $9|S|\|\theta_{T^c}\|_\infty \leq \|\theta_T\|_1$, and consequently

$$\begin{aligned}\sup_{z\in A}\left(\langle z_T,\, \theta_T\rangle + \langle z_{T^C},\, \theta_{T^C}\rangle\right)\phi_A(\theta) &\leq \sup_{z\in A}\left(\|z_T\|_\infty\|\theta_T\|_1 + \|z_{T^C}\|_1\|\theta_{T^C}\|_\infty\right)\\ &\leq \|\theta_T\|_1 + 9|S|\,\|\theta_{T^c}\|_\infty\\ &\leq\ 2\|\theta_T\|_1.\end{aligned}$$

On the other hand, for this same subset $T$, we have $\phi_B(\theta) \geq \sup_{z\in B}\langle z_T,\, \theta_T\rangle\ =\ 2\|\theta_T\|_1$, which completes the proof. ■

### B.3 Proof of Lemma 8

In order to prove Lemma 8, we use a somewhat simplified version of a recent result due to van de Geer (2014). So as to simplify notation, we first define the norms $\|a\|_j := |a_j|$ and $\|a\|_{-j} := \sum_{i\neq j}|a_i|$ for any vector $a$. We then have:

**Lemma 11 (van de Geer (2014), Lemma 2.1)** *Given any tuning parameter $\lambda > 0$, it holds that*

$$\|\widehat{\beta}_\lambda - \beta^*\|_j \leq D_j\left(\frac{\|X^\top\varepsilon\|_\infty}{n} + \frac{\sqrt{\log(p)}\|\widehat{\beta}_\lambda - \beta^*\|_{-j}}{2\sqrt{n}\|\eta^j\|_1} + \frac{\lambda}{2}\right) \qquad \textit{for all } j = 1,\ldots,p,$$

*where for each $j \in [p]$,*

$$D_j := \frac{\|\eta^j\|_1}{\|X\eta^j\|_2^2/n + \sqrt{\log(p)/n}\|\eta^j\|_{-j}/2}.$$

This result provides a specific bound for each coordinate of Lasso. Lemma 8 can then readily be proven using this result together with Theorem 6.1 from Bühlmann and van de Geer (2011). ■

## Appendix C. Strong Correlations

In this paper, we assume that the correlations in design matrix are small, which is needed for precise $\ell_\infty$-estimation and variable selection. In the interest of completeness, however, we add here two simulations where the correlations are large. Overall, we use the same settings as described in the main part of the paper, but we set $\kappa = 0.9$. The results are summarized in Figure 5 (note that the x-scale in the upper part of the figure is different from the scales of the corresponding plots in the main part of the paper). We find that $\text{AV}_\infty$ misses about half of the pertinent variables but has almost no false positives. Cross-Validation, on the other hand, has less false negatives but selects many irrelevant variables. As expected, none of the methods, including the oracle, provide accurate $\ell_\infty$-estimation.

## References

A. Antoniadis. Comments on: $\ell_1$-penalization for mixture regression models. *Test*, 19(2): 257–258, 2010.

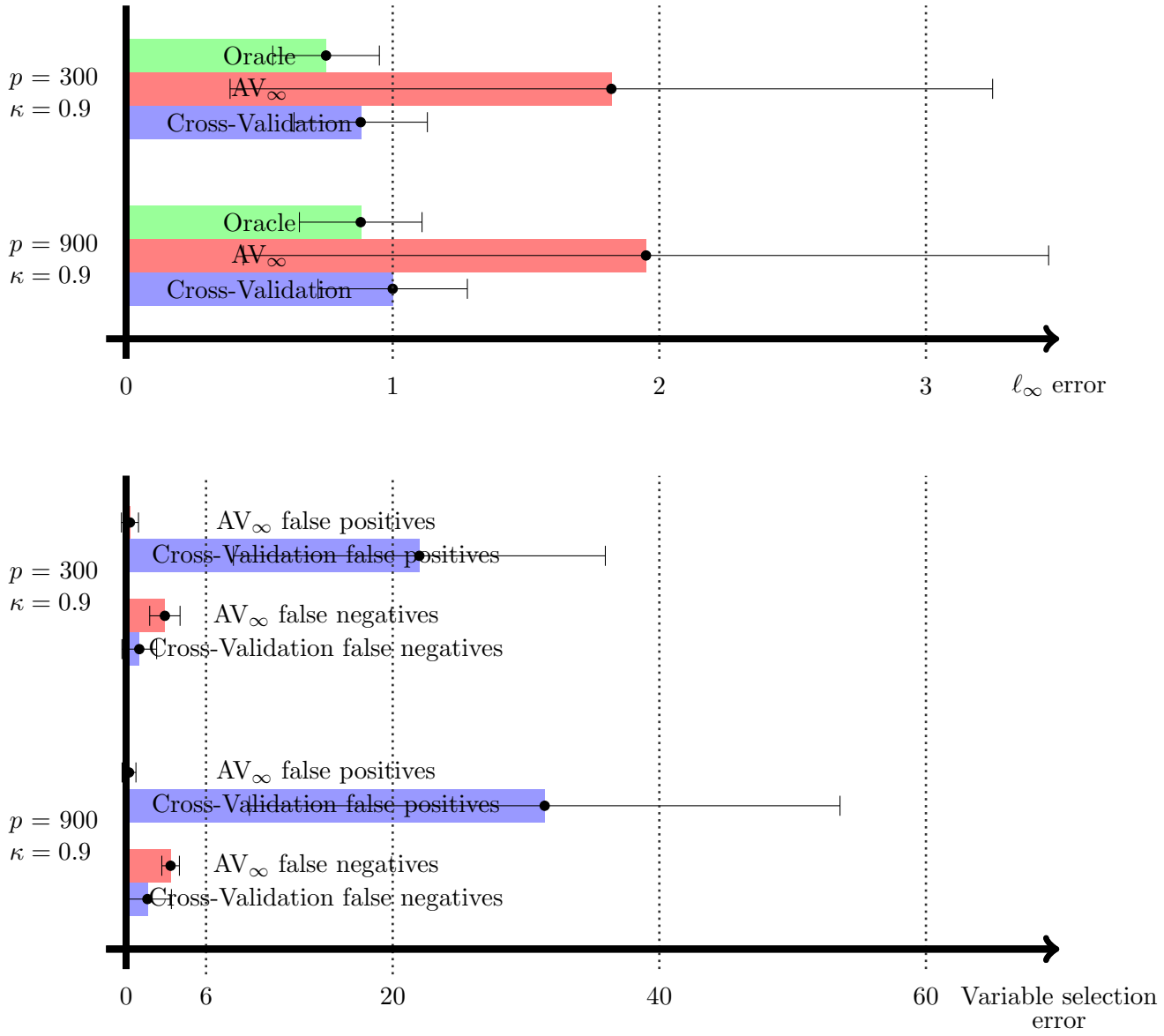


Figure 5: Sup-norm and variable selection errors of the Lasso with three/two different calibration schemes for the tuning parameter $\lambda$. Depicted are the results for two simulation settings that differ in the number of parameters $p$. The simulation settings and the calibration schemes are specified in the main part of the paper.

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, pages 33–40, 2008.

Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the gaussian setting. In *Ann. Inst. H. Poincaré Probab. Statist.*, volume 50, pages 1092–1119, 2014.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of the Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: Methods, theory and applications*. Springer Series in Statistics. Springer, 2011.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014.

F. Bunea. Honest variable selection in linear and logistic regression models via $\ell_1$ and $\ell_1+\ell_2$ penalization. *Electron. J. Stat.*, 2:1153–1194, 2008.

D. Chételat, J. Lederer, and J. Salmon. Optimal two-step prediction in regression. *arXiv:1410.5014*, 2014.

M. Chichignoud and J. Lederer. A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression. *Bernoulli*, 20(3):1560–1599, 2014.

A. Dalalyan, M. Hebiri, and J. Lederer. On the Prediction Performance of the Lasso. *Bernoulli, in press*, 2014.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *Statistical Science*, 27(4):500–518, 2012.

M. Hebiri and J. Lederer. How correlations influence Lasso prediction. *IEEE Trans. Inform. Theory*, 59(3):1846–1854, 2013.

J. Lederer and C. Müller. Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the trex. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

H. Leeb and B. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211, 2008.

O. Lepski. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990. ISSN 0040-361X.

O. Lepski, E. Mammen, and V. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997.

P.-L. Loh and M. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *arXiv:1412.5632*, 2014.

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 72(4):417–473, 2010.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/.

J. Sabourin, W. Valdar, and A. Nobel. A permutation approach for selecting the penalty parameter in penalized model selection. *Biometrics*, 71(4):1185–1194, 2015.

N. Städler, P. Bühlmann, and S. van de Geer. $\ell_1$-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

S. van de Geer. The deterministic Lasso. *2007 Proc. Amer. Math. Soc. [CD-ROM], see also www.stat.math.ethz.ch/˜geer/lasso.pdf*, 2007.

S. van de Geer. Worst possible sub-directions in high-dimensional models. *J. Multivariate Anal., in press*, 2014.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

S. van de Geer and J. Lederer. The Lasso, correlated design, and improved oracle inequalities. *IMS Collections*, 9:303–316, 2013.

Y. Yang. Can the strengths of aic and bic be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

F. Ye and C.-H. Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *J. Mach. Learn. Res.*, 11:3519–3540, 2010.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, pages 894–942, 2010.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006.