Journal of Machine Learning Research 19 (2018) 1-48 Submitted 1/17; Revised 10/18; Published 10/18

On Tight Bounds for the Lasso

Sara van de Geer GEER@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zürich
8092 Zürich, Switzerland

Editor: Francis Bach

Abstract

We present upper and lower bounds for the prediction error of the Lasso. For the case of random Gaussian design, we show that under mild conditions the prediction error of the Lasso is up to smaller order terms dominated by the prediction error of its noiseless counterpart. We then provide exact expressions for the prediction error of the latter, in terms of compatibility constants. Here, we assume the active components of the underlying regression function satisfy some "betamin" condition. For the case of fixed design, we provide upper and lower bounds, again in terms of compatibility constants. As an example, we give an up to a logarithmic term tight bound for the least squares estimator with total variation penalty.

Keywords: Compatibility, Lasso, Linear Model, Lower Bound

1. Introduction

Let $X \in \mathbb{R}^{n\times p}$ be an input matrix and $\beta^0 \in \mathbb{R}^p$ a vector of unknown coefficients. Consider an n-vector of noisy observations

$$Y = X\beta^0 + \epsilon$$

where the noise $\epsilon \in \mathbb{R}^n$ is a vector of i.i.d. standard Gaussians independent of X. The Lasso estimator $\hat\beta$ is

$$\hat\beta \in \arg\min_{b\in\mathbb{R}^p}\left\{ \|Y - Xb\|_2^2 + 2\lambda\|b\|_1 \right\} \tag{1}$$

with $\lambda > 0$ a regularization parameter (Tibshirani (1996)). Its prediction error is $\|X(\hat\beta - \beta^0)\|_2^2$. Main aim of this paper is to provide lower bounds for this prediction error, bounds which show that compatibility constants necessarily enter into the picture.

The results of this paper can be summarized as follows. Firstly, suppose the design is random and that $\Sigma_0 := \mathbb{E}X^TX/n$ exists. Let β^* be the noiseless Lasso for random design

$$\beta^* \in \arg\min_{b\in\mathbb{R}^p}\left\{ n\|\Sigma_0^{1/2}(b-\beta^0)\|_2^2 + 2\lambda\|b\|_1 \right\}. \tag{2}$$

For the case where the rows of X are i.i.d $\mathcal{N}(0, \Sigma_0)$, we compare $\|X(\hat\beta - \beta^0)\|_2$ with $\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2$ in Theorem 11. We assume here some mild condition on the growth of the compatibility constants as n increases. The theorem has as an important corollary that $\|X(\hat{\beta}-\beta^0)\|_2$ is up to lower order terms equal to $\sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2$ whenever (after normalizing the co-variance matrix Σ_0 to having bounded entries) the largest eigenvalue $\Lambda_{\max}^2$ of Σ_0 is of small order $\log n$, see Corollary 12. Secondly, we provide in Theorem 14 exact expressions for the prediction error of the noiseless Lasso in terms of compatibility constants. We require here "betamin" conditions, which roughly say that the non-zero coefficients of β^0 should have the appropriate signs and remain above the noise level in absolute value. Thirdly, for the case of fixed design, we present upper and lower bounds for the prediction error $\|X(\hat{\beta}-\beta^0)\|_2^2$ in terms of weighted compatibility constants. Theorem 17 states the lower bounds, assuming again certain betamin conditions. The upper bounds we present are similar to those obtained the literature and presented for completeness. They are stated as a consequence of Theorem 18 in Corollary 19. Another application of Theorem 18 is given in Corollary 20. It presents an upper bound for $\|X(\hat{\beta}-\beta^*)\|_2$ where β^* is now the counterpart of (2) for the fixed design case. As an illustration we consider least squares estimation with a (one-dimensional) total variation penalty. For this case we arrive in Corollary 22 at lower and upper bounds that are the same up to a logarithmic term.

There are general upper bounds in the literature, in particular *sharp oracle bounds* as in Koltchinskii et al. (2011) (see also Giraud, 2014, Theorem 4.1 or van de Geer, 2016, Theorem 2.2). The oracle bounds involve a compatibility constant, and an improved version of this constant has been developed in Sun and Zhang (2012), Belloni and Wang (2014) and Dalalyan et al. (2017).

Main theme of this paper is to gain further insight into the role of the compatibility constant when applying the Lasso and to see how it occurs in lower bounds. In Zhang et al. (2014) it is shown that for a given sparsity level, there is a design and a lower bound for the mean prediction error in the noisy case, that holds for any polynomial time algorithm. This lower bound is close to the known upper bounds and in particular shows that compatibility conditions or restricted eigenvalue conditions cannot be avoided. This has also been shown by Bellec (2017), where a choice of the particular vector of regression coefficients β^0 leads to a lower bound matching the upper bound. We further elaborate on this issue, and provide lower bounds that hold for a large class of vectors β^0.

To get an idea of the flavour of the type of bounds we are after, we present in Theorem 1 the case of random design. Details of its proof can be found in Subsection 11.9. We provide more explicit statements in Theorem 11.

Throughout the paper, the active set of β^0 is denoted by $S_0 := \{j : \ \beta_j^0 \neq 0\}$. Its size is denoted by $s_0 := |S_0|$. Our betamin condition is as follows (its meaning should become more clear after looking at Section 3 where compatibility constants are defined).

Betamin condition *Let*

$$b^* \in \arg\min \biggl\{ \|\Sigma_0^{1/2} b\|_2 : \ \sum_{j\in S_0} |b_j| - \sum_{j\notin S_0} |b_j| = 1 \biggr\}$$

and for $j \in S_0$ let z_j^* be the sign of b_j^*. We say that β^0 satisfies the betamin condition for the noiseless case with random design if

$$z_j^* \beta_j^0 > \frac{z_j^* b_j^*}{\|\Sigma_0^{1/2} b^*\|_2^2} \frac{\lambda}{n}, \ \forall j \in S_0. \tag{3}$$

We will make asymptotic statements with the sample size n tending to infinity and apply (stochastic) order symbols. All quantities in the paper are allowed to depend on n unless otherwise stated.

Theorem 1 *Let the rows of X be i.i.d. $\mathcal{N}(0, \Sigma_0)$, let $\|\Sigma_0\|_\infty$ be the maximal entry in the co-variance matrix Σ_0 and $\Lambda_{\max}^2$ be its largest eigenvalue. For $S \subset \{1, \ldots, p\}$, let $\kappa^2(S)$ be the compatibility constant defined in Definition 2. Suppose that*

$$\Lambda_{\max}^2 / \|\Sigma_0\|_\infty = o(\log(2p)),$$

and

$$\max\left\{ \left(\frac{\|\Sigma_0\|_\infty}{\kappa^2(S)} \right) \frac{\log(2p)|S|}{n} : \ S \subset \{1, \ldots, p\}, \ |S| \le \left(\frac{\Lambda_{\max}^2}{\kappa^2(S_0)} \right) 4 s_0 \right\} = o(1).$$

For some $t > 0$, take the tuning parameter λ to satisfy

$$3\|\Sigma_0\|_\infty^{1/2} \left(\sqrt{2n(\log(2p) + t)} + 2(\log(2p) + t) \right) \le \lambda = \mathcal{O}\left(\sqrt{\|\Sigma_0\|_\infty^{1/2} \log(2p)} \right).$$

Then, under condition (3) (the betamin condition for the noiseless case with random design), we have

$$\|X(\hat{\beta} - \beta^0)\|_2^2 = \frac{\lambda^2 / n}{\|\Sigma_0^{1/2} b^*\|_2^2} (1 + o_{\mathbf{P}}(1)) + \mathcal{O}_{\mathbf{P}}(1)$$

(where in fact $s_0 \|\Sigma_0^{1/2} b^*\|_2^2 = \kappa^2(S_0)$).

2. Organization of the Paper

In Section 3 the definition of compatibility constants is given and also some of their properties are discussed. Section 4 shows that for the case of random design the squared "bias" of the Lasso dominates its "variance". Section 5 then gives expressions for this "bias", i.e. for the noiseless Lasso. Here, we examine fixed design but the results carry over immediately to random design. In Section 6 the result of Section 5 is illustrated with the total variation penalty (in one dimension). Section 7 presents lower bounds for the noisy case with fixed design, and Section 8 presents some upper bounds. Corollary 19 is essentially as in the papers Sun and Zhang (2012), Belloni and Wang (2014) and Dalalyan et al. (2017), albeit that do not consider the approximately sparse case to avoid digressions. Section 9 has upper and lower bounds for the least squares estimator with total variation penalty in the noisy case. Section 10 concludes. Section 11 contains the proofs.

3. Compatibility Constants

We introduce some notation in order to be able to define the compatibility constants. This notation will also be helpful at other places. For $S \subset \{1, \ldots, p\}$ and a vector $b \in \mathbb{R}^p$ let $b_S \in \mathbb{R}^p$ be the vector with entries $b_{j,S} := b_j \text{l}\{j \in S\}$, $j = 1, \ldots, p$. We apply the same notation for the $|S|$-dimensional vector $\{b_j\}_{j \in S}$. We moreover write $b_{-S} := b_{S^c}$ where S^c is the the complement of the set S.

3.1. Theoretical Compatibility Constants

The population version of the compatibility constant will be used for the case of random design X. We call the population version the theoretical compatibility constant.

Definition 2 *Let $\Sigma_0 := \mathbb{E}X^T X/n$ (assumed to exist). Let $S \subset \{1, \ldots, p\}$ be a set of indices and $u \geq 0$ be a constant. The theoretical compatibility constant is*

$$\kappa^2(u, S) := \min\left\{ |S| \|\Sigma_0^{1/2} b\|_2^2 : \ \|b_S\|_1 - u\|b_{-S}\|_1 = 1 \right\}.$$

For $u = 1$ we write $\kappa(1, S) =: \kappa(S)$.

3.2. Empirical Compatibility Constants

For a vector w we let $W := \text{diag}(w)$ be the diagonal matrix with w on the diagonal.

Definition 3 *(Belloni and Wang, 2014, Dalalyan et al., 2017) Let $S \subset \{1, \ldots, p\}$ be a set of indices and $w \in \mathbb{R}^{p-|S|}$ be a vector of non-negative weights. The (empirical) compatibility constant is is*

$$\hat{\kappa}^2(w, S) := \min\left\{ |S| \|Xb\|_2^2/n : \ \|b_S\|_1 - \|W b_{-S}\|_1 = 1 \right\}.$$

For the case where $w = \mathbf{1}$ where $\mathbf{1}$ denotes a vector with all entries equal to one, put $\hat{\kappa}^2(S) := \hat{\kappa}^2(\mathbf{1}, S)$.

3.3. Some Properties of Compatibility Constants

One readily sees that the theoretical and empirical compatibility constants differ only in terms of the matrix used in the quadratic form (which is Σ_0 in the theoretical case and the Gram matrix $\hat{\Sigma} := X^T X/n$ in the empirical case). Thus, when discussing their basic properties it suffices to deal with only one of the two. In this section, we therefore restrict attention to the empirical version $\hat{\kappa}(w, S)$. Note that we have generalized the empirical version as compared to the theoretical one, by considering general weight vectors, not just constant vectors. With some abuse of notation, we write $\hat{\kappa}(u, S) = \hat{\kappa}(u\mathbf{1}, S)$ when the weights are the constant vector $u\mathbf{1}$ (it should be clear from the context what is meant).

The empirical compatibility constant as given in Definition 3 is from Belloni and Wang (2014) or Dalalyan et al. (2017). Another version, from for instance van de Geer (2007) or van de Geer (2016) and its references, is presented in the next definition.

Definition 4 *Let $S \subset \{1, \ldots, p\}$ be a set of indices and $u > 0$ be a constant. The (older) compatibility constant is*

$$\hat{\phi}^2(u, S) := \min\left\{ |S| \|Xb\|_2^2/n : \ \|b_S\|_1 = 1, \ \|b_{-S}\|_1 \le 1/u \right\}.$$

Let $\hat{\phi}^2(S) := \hat{\phi}^2(1, S)$ be the compatibility constant for the case $u = 1$.

The constant $\hat{\phi}(u, S)$ compares, for b's satisfying a "cone condition" $\|b_{-S}\|_1 \le \|b_S\|_1/u$, the ℓ_2-norm $\|Xb\|_2$ with the ℓ_1-norm $\|b_S\|_1$. The constant $\hat{\kappa}(u, S)$ is similar, but takes in the comparison more advantage of a "cone condition" $\|b_S\|_1 - u\|b_{-S}\|_1 > 0$. When $\hat{\kappa}^2(S) > 0$ the null space property holds (Donoho and Tanner, 2005). We will need throughout that the compatibility constant is strictly positive at S_0 (if it is zero our results cease to be of any interest). This means that we implicitly require throughout

Invertibility condition

$$\text{The matrix } X_{S_0}^T X_{S_0} \text{ is invertible.} \tag{4}$$

Here, for any $S \subset \{1, \ldots, p\}$ the matrix $X_S = \{X_j\}_{j \in S}$ is the $n \times |S|$ matrix consisting of the columns of X corresponding to the set S.

The newer version $\hat{\kappa}(u, S)$ is an improvement over $\hat{\phi}(u, S)$ in the sense that $\hat{\kappa}(u, S)$ is the larger of the two.

Lemma 5 *For all $u > 0$ it is true that*

$$\hat{\kappa}^2(u, S) \ge \hat{\phi}^2(u, S).$$

Let now for some $v > 0$

$$b^* \in \arg\min\left\{ \|Xb\|_2^2/n : \ \|b_S\|_1 - v\|b_{-S}\|_1 = 1 \right\}.$$

Then by definition

$$\hat{\kappa}^2(v, S) = |S| \|Xb^*\|_2^2/n.$$

The restriction $\|b_S\|_1 - v\|b_{-S}\|_1 = 1$ does not put any bound on the ℓ_1-norm of b_S^*. However, if there is a little room to spare, its ℓ_1-norm is bounded. This will be useful to understand the betamin conditions (conditions (3) and (8)). For simplicity we examine only the value $v = 1$.

Lemma 6 *Let*

$$b^* \in \arg\min\left\{ \|Xb\|_2^2/n : \ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \right\}.$$

Then for $0 \le u < 1$

$$\|b_S^*\|_1 \le \frac{\hat{\kappa}(S) - u\hat{\kappa}(u, S)}{(1 - u)\hat{\kappa}(u, S)}.$$

3.4. Comparing Empirical and Theoretical and Compatibility

Having random quadratic forms in mind, the fact that $\|b_S\|_1 - \|b_{-S}\|_1 = 1$ gives no bound on the ℓ_1-norm can be a problem. Again, if there is a little room to spare in the value of u in the compatibility constant, one *does* get a bound on the ℓ_1-norm. We show this in Lemma 7, and with this tool in hand we lower bound the empirical compatibility constant in terms of the theoretical one in Lemma 8.

Lemma 7 *Let $v > u > 0$. Then*

$$\hat{\kappa}^2(v,S) \geq \min\biggl\{ |S| \|Xb\|_2^2/n : \ \|b_S\|_1 - u\|b_{-S}\|_1 = 1, \ \|b\|_1 \leq 1 + (1+u)/(v-u) \biggr\}.$$

The following lemma will be applied when bounding the prediction error of $\hat{\beta}$ in terms of that of the noiseless Lasso β^*. The lemma may also be of interest in itself with applications elsewhere.

Lemma 8 *Suppose the rows of X are i.i.d. $\mathcal{N}(0,\Sigma_0)$. Let $\|\Sigma_0\|_\infty$ be the largest entry in the matrix Σ_0. For $v > u$, $(1+u)/(v-u) = \mathcal{O}(1)$ and*

$$\left(\frac{\|\Sigma_0\|_\infty}{\kappa^2(u,S)} \right) \frac{s \log(2p)}{n} = o(1),$$

it is true with probability tending to one that

$$\hat{\kappa}^2(v,S) \geq (1-\eta)^2 \kappa^2(u,S).$$

where $\eta = o(1)$.

4. Comparison With the Noiseless Lasso When the Design is Random

In this section we assume that the rows of X are i.i.d. copies of a Gaussian row vector with mean zero and co-variance matrix Σ_0. We denote the largest eigenvalue of Σ_0 by $\Lambda_{\max}^2$ and let $\|\Sigma_0\|_\infty$ be its largest entry. We define a noiseless version β^* of the Lasso where also the random design is replaced by its population counterpart:

$$\beta^* \in \arg\min_{b \in \mathbb{R}^p} \biggl\{ n \|\Sigma_0^{1/2}(b - \beta^0)\|_2^2 + 2\lambda \|b\|_1 \biggr\}.$$

The normalization with n is to put things on the scale of the empirical version, as $\mathbb{E}X^T X = n\Sigma_0$. One may think of $\|X(\beta^* - \beta^0)\|_2$ as "bias" and $\|X(\hat{\beta} - \beta^*)\|_2^2$ as "variance". We first investigate in some detail the "variance" part in Theorems 9 and 10. Then we apply the triangle inequality as a way to establish that the squared "bias" dominates the "variance", see Theorem 11.

Theorem 9 *Suppose that*

$$\rho^2 := \max\biggl\{ \left(\frac{\|\Sigma_0\|_\infty}{\kappa^2(S)} \right) \frac{\log(2p)|S|}{n} : \ S \subset \{1,\ldots,p\}, \ |S| \leq \left(\frac{\Lambda_{\max}^2}{\kappa^2(S_0)} \right) 4 s_0 \biggr\} = o(1).$$

Take for some $t > 0$

$$\lambda \ge 3\|\Sigma_0\|_\infty^{1/2}\left(\sqrt{2n(\log(2p)+t)} + 2(\log(2p)+t)\right)$$

and define

$$\gamma := (2\Lambda_{\max})\sqrt{n}/\lambda + (2/\|\Sigma_0\|_\infty^{1/2})\rho\lambda/\sqrt{n\log(2p)}.$$

Then we have for all $x > 0$ *with probability at least* $1 - 4\exp[-t] - \exp[-x] - o(1)$ *that*

$$\|X(\hat\beta - \beta^*)\|_2 \le \gamma\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2 + \sqrt{2x}.$$

Using concentration of measure, one can remove the dependency of the confidence level on the value of t. This value appears in the choice of the tuning parameter λ. We make some rather arbitrary choices for the constants.

Theorem 10 *With the conditions and notations of Theorem 9, and assuming in addition that* $4\exp[-t] < 1/8$ *(say), for* n *large enough and for all* $x > 0$*, with probability at least* $1 - 2\exp[-x]$,

$$\|X(\hat\beta - \beta^*)\|_2 \le \gamma\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2 + 4\sqrt{\log 2} + \sqrt{2x}.$$

We can now make a type of bias-variance decomposition. The triangle inequality tells us that

$$\left|\|X(\hat\beta - \beta^0)\|_2 - \|X(\beta^* - \beta^0)\|_2\right| \le \|X(\hat\beta - \beta^*)\|_2.$$

We then approximate the empirical "bias" $\|X(\beta^* - \beta^0)\|_2$ by the theoretical "bias" $\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta_0)\|_2$ (which is easy as β^* and β^0 are non-random vectors), and use Theorem 9 or 10 to bound the "variance" $\|X(\hat\beta - \beta^*)\|_2^2$.

Theorem 11 *With the conditions and notations of Theorem 10, we have for* n *sufficiently large, for all* $x > 0$ *with probability at least* $1 - 2\exp[-x]$

$$\left|\|X(\hat\beta - \beta^0)\|_2 - \sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2\right|$$

$$\le (\gamma + o(1))\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2 + 4\sqrt{\log 2} + \sqrt{2x}.$$

Corollary 12 *Recall that we defined* γ *as*

$$\gamma := (2\Lambda_{\max})\sqrt{n}/\lambda + (2/\|\Sigma_0\|_\infty^{1/2})\rho\lambda/\sqrt{n\log(2p)}.$$

Therefore, with the conditions and notations of Theorem 11, and assuming in addition
- $\Lambda_{\max}^2/\|\Sigma_0\|_\infty = o(\log(2p))$,
and
- $\lambda = o(\sqrt{\|\Sigma_0\|_\infty n\log(2p)})/\rho$,
we get with probability at least $1 - 2\exp[-x]$

$$\left|\|X(\hat\beta - \beta^0)\|_2 - \sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2\right| = o(\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2) + 4\sqrt{\log 2} + \sqrt{2x}.$$

In words: the squared "bias" dominates the "variance".

Remark 13 *With the help of Lemma 45, one may also prove bounds for* $\sqrt{n}\|\Sigma_0(\hat\beta-\beta^0)\|_2$ *to complete those for of* $\|X(\hat\beta-\beta^0)\|_2$*. We refrain from doing this here to avoid digressions.*

5. The Noiseless Case with Fixed Design

In this section we study fixed design X and the noiseless Lasso

$$\beta^* \in \arg\min_{b\in\mathbb{R}^p}\biggl\{ \|X(b-\beta^0)\|_2^2 + 2\lambda^*\|b\|_1 \biggr\}. \tag{5}$$

In principle the noiseless Lasso considered here differs from (2), although one can say that for fixed design $\hat\Sigma = \mathbb{E}\hat\Sigma =: \Sigma_0$, with $\hat\Sigma := X^TX/n$ being the Gram matrix. In what follows in this section, we do not use any specific properties of $\hat\Sigma$ and the theory goes through for any positive semi-definite matrix, Σ say. In the upcoming illustration on functions of bounded variation, the fixed design setup is the natural one.

Note that we supplied the tuning parameter λ^* with a supscript $*$. This is because in Theorem 18 we consider a case with different tuning parameters for the noisy and the noiseless case, say λ and λ^*.

The Karush-Kuhn-Tucker (KKT) conditions for the noiseless Lasso read

$$X^TX(\beta^*-\beta^0) + \lambda^*\zeta^* = 0,\ \zeta^* \in \partial\|\beta^*\|_1, \tag{6}$$

where $\partial\|b\|_1$ denotes the sub-differential of $b \mapsto \|b\|_1$:

$$\partial\|b\|_1 = \biggl\{ z \in \mathbb{R}^p :\ z^Tb = \|b\|_1,\ \|z\|_\infty \le 1 \biggr\}.$$

Recall that

$$\hat\kappa^2(S) = |S|\|Xb^*\|_2^2/n$$

where

$$b^* \in \arg\min_{b\in\mathbb{R}^p}\biggl\{ \|Xb\|_2 :\ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \biggr\}. \tag{7}$$

Note that b^* given in (7) is not unique, for example we can flip the signs of b^* (i.e., replace b^* by $-b^*$).

In Theorem 14 below we give a tight result for the noiseless case under the condition that the active coefficients in β^0 are sufficiently large in absolute value: Condition 8. Here sufficiently large depends on the magnitude of the entries of a solution b^* of (7) with $S = S_0$. Therefore, it is of interest to know how large b^* is. Lemma 6 considers its ℓ_1-norm, and in view of this lemma we conclude that if there is a little room to spare, the ℓ_1-norm of $\|b_S^*\|_1$ is bounded, or - in other words - $\{b_j^*|S|\}_{j\in S}$ is bounded "on average".

For the next condition it is useful to know that we show in Lemma 28 that for b^* given in (7), each coefficient b_j^* with $j \in S$ is nonzero (provided $\hat\kappa^2(S) > 0$).

Betamin condition *Suppose $\hat{\kappa}^2(S_0) > 0$. Let b^* satisfy (7) with $S = S_0$. Denote, for $j \in S_0$, the sign of b_j^* as z_j^*. We say that β^0 satisfies the betamin condition for the noiseless case with fixed design if*

$$z_j^* \beta_j^0 > \frac{z_j^* b_j^* s_0}{\hat{\kappa}^2(S_0)} \frac{\lambda^*}{n} \ \forall \ j \in S_0. \tag{8}$$

Here is the main theorem for the noiseless case.

Theorem 14 *Suppose $\hat{\kappa}^2(S_0) > 0$. Let b^* satisfy (7) with $S = S_0$. If β^0 satisfies condition (8) (the betamin condition for the noiseless case with fixed design), then there exists a solution β^* of the KKT conditions (6) such that*

$$\|X(\beta^* - \beta^0)\|_2^2 = \frac{s_0}{\hat{\kappa}^2(S_0)} \frac{\lambda^{*2}}{n}.$$

6. The Total Variation Penalty in the Noiseless Case

In this section Theorem 14 is illustrated with the total variation penalty. For a vector $f \in \mathbb{R}^n$, its total variation is defined as

$$\text{TV}(f) := \sum_{i=2}^n |f_i - f_{i-1}|.$$

Fix a vector $f^0 \in \mathbb{R}^n$ and let $f^* \in \mathbb{R}^n$ is the least squares approximation of f^0 with total variation penalty:

$$f^* \in \arg\min_{f \in \mathbb{R}^n} \left\{ \|f - f^0\|_2^2 + 2\lambda^* \text{TV}(f) \right\}. \tag{9}$$

Theorem 15 presents an explicit expression for the compatibility constant $\hat{\kappa}^2(S_0)$ where S_0 is the set consisting of the locations of the jumps of f^0. Invoking Theorem 14 one then arrives at an explicit expression for $\|f^* - f^0\|_2^2$ provided the jumps of f^0 are sufficiently large, see Corollary 16.

First, we need to rewrite problem (9) as a (noiseless) Lasso problem. Indeed, for $j = 1 \ldots, n$,

$$f_j = \sum_{i=1}^n (f_i - f_{i-1}) \mathrm{l}\{j \geq i\} =: (Xb)_j,$$

where $X_{j,i} = \mathrm{l}\{j \geq i\}$ and $b_i = f_i - f_{i-1}$, with $f_0 := 0$. Hence we can say that $f^0 = X\beta^0$ and $f^* = X\beta^*$ with

$$\beta^* := \arg\min_{b \in \mathbb{R}^n} \left\{ \|X(b - \beta^0)\|_2^2 + 2\lambda^* \sum_{i=2}^n |b_i| \right\}.$$

Note that the first coefficient b_1 is not penalized. It is therefore typically active, and we consider the active set as the location of the jumps augmented with the index $\{1\}$. We slightly adjust the definition of the compatibility constant to deal with the a coefficient without penalty: we set for $S \subset \{2, \ldots, n\}$

$$\kappa^2(S) := \min\biggl\{ |S \cup \{1\}| \|Xb\|_2^2 : \|b_S\|_1 - \|b_{-(S\cup\{1\})}\|_1 = 1 \biggr\}. \tag{10}$$

Let now $S := \{d_1 + 1, d_1 + d_2 + 1, \ldots, d_1 + \cdots + d_s + 1\}$ for some $\{d_j\}_{j=1}^s \subset \{2, \ldots, n\}$ satisfying $\sum_{j=1}^s d_j + 2 < n$. The set S represents locations of jumps, d_1 is the location of the first jump and $\{d_j\}_{j=2}^s$ are the distances between jumps. Let $d_{s+1} := n - \sum_{j=1}^s d_j$ the distance between the last jump and the end point. For simplicity we assume that d_j is even for all $j \in \{2, \ldots, s\}$.

Theorem 15 *The compatibility constant $\hat{\kappa}^2(S)$ is, up the constant 4 and the scaling by $1/n$, the harmonic mean of of the distances between jumps, including the distance between starting point and first jump and last jump and endpoint:*

$$\hat{\kappa}^2(S) = \frac{s+1}{\frac{n}{d_1} + \sum_{j=2}^s \frac{4n}{d_j} + \frac{n}{d_{s+1}}}.$$

In fact

$$\hat{\kappa}^2(S) = (s+1)\|Xb^*\|_2^2/n$$

where $b_j^* = 0$ for all $j \notin S$ and $b^* = \tilde{b}/\|\tilde{b}\|_1$ with

$$\begin{aligned}
\tilde{b}_{d_1+1} &= \frac{n}{d_1} + \frac{2n}{d_2}, \\
\tilde{b}_{d_2+1} &= -\biggl(\frac{2n}{d_2} + \frac{2n}{d_3}\biggr), \\
&\vdots \\
\tilde{b}_{d_s} &= (-1)^{s+1}\biggl(\frac{2n}{d_s} + \frac{n}{d_{s+1}}\biggr).
\end{aligned}$$

Corollary 16 *Suppose f^0 jumps at $S_0 := S = \{d_1 + 1, d_1 + d_2 + 1, \ldots, d_1 + \cdots + d_s + 1\}$, with $s = s_0$. Assume f^0 alternates between jumps up and jumps down. Suppose moreover that*

$$\begin{aligned}
|f^0_{d_1+1} - f^0_{d_1}| &\geq \biggl(\frac{n}{d_1} + \frac{2n}{d_2}\biggr)\frac{\lambda^*}{n}, \\
|f^0_{d_2+1} - f^0_{d_2}| &\geq \biggl(\frac{2n}{d_2} + \frac{2n}{d_3}\biggr)\frac{\lambda^*}{n}, \\
&\vdots \\
|f^0_{d_{s_0}+1} - f^0_{d_{s_0}}| &\geq \biggl(\frac{2n}{d_{s_0}} + \frac{n}{d_{s_0+1}}\biggr)\frac{\lambda^*}{n}.
\end{aligned}$$

Then by Theorem 14 combined with Theorem 15

$$\|f^* - f^0\|_2^2 = \biggl(\frac{n}{d_1} + \sum_{j=2}^{s_0} \frac{4n}{d_j} + \frac{n}{d_{s_0+1}}\biggr)\frac{\lambda^{*2}}{n}.$$

At this point it may be helpful to look how this normalizes. Say we choose $\lambda^* = \sqrt{n \log n}$. *Suppose* $\max_{1 \le j \le s_0+1} n/d_j = \mathcal{O}(s_0 + 1)$. *Then the jumps of* f^0 *are required to be of order at least* $(s_0 + 1)\sqrt{\log n/n}$. *We then obtain*

$$\|f^* - f^0\|_2^2 = \mathcal{O}\biggl((s_0 + 1)^2 \log n\biggr).$$

7. A Lower Bound in the Noisy Case with Fixed Design

We now turn to the Lasso $\hat{\beta}$ in the noisy case, given by

$$\hat{\beta} \in \arg\min_{b \in \mathbb{R}^p} \Bigl\{ \|Y - Xb\|_2^2 + 2\lambda \|b\|_1 \Bigr\}$$

where

$$Y = X\beta^0 + \epsilon.$$

We investigate the case of fixed design X. Recall that we assume throughout i.i.d. standard Gaussian noise.

7.1. Towards Betamin Conditions

Consider some vector $\bar{v} \in \mathbb{R}^{p-s_0}$ with $0 < \bar{v}_j < 1$ for all j. This vector represents the "noise" that is to be overruled by the penalty. Define the collection of weights

$$\mathcal{W}(\bar{v}) := \Bigl\{ w \in \mathbb{R}^{p-s_0} : \ \ 1 - \bar{v}_j \le w_j \le 1 + \bar{v}_j \ \forall \ j \Bigr\}.$$

Let for $\bar{W} := \mathrm{diag}(1 + \bar{v})$

$$b^*(\bar{v}) \in \arg\min \Bigl\{ \|Xb\|_2^2 : \ \|b_{S_0}\|_1 - \|\bar{W} b_{-S_0}\|_1 = 1 \Bigr\}, \ z_j^*(\bar{v}) := \mathrm{sign}(b_j^*(\bar{v})), \ j \in S_0.$$

Then by definition $\hat{\kappa}^2(1 + \bar{v}, S_0) = s_0 \|Xb^*(\bar{v})\|_2^2/n$. We remark here that by a slight adjustment of Lemma 28, the assumption $\hat{\kappa}(1 + \bar{v}, S_0) > 0$ ensures that $b_j^*(\bar{v}) \neq 0$ for all $j \in S_0$.

For $w \in \mathcal{W}(\bar{v})$ we define the convex problem with linear and convex constraints

$$b(w) \in \arg\min \Bigl\{ \|Xb\|_2^2 : \ z_{S_0}^{*T}(\bar{v}) b_{S_0} - \|W b_{-S_0}\|_1 \ge 1 \Bigr\}.$$

Finally, define

$$\mathbf{b}_j(\bar{v}) := \max_{w \in \mathcal{W}(\bar{v})} |b_j(w)|/\|Xb(w)\|_2^2, \ j \in S_0.$$

7.2. Projections

We denote the projection of X_{-S_0} on the space spanned by the columns of X_{S_0} by $X_{-S_0}\mathrm{P}X_{S_0}$. The projection is always defined but as it is implicitly assumed that $X_{S_0}^T X_{S_0}$ is invertible (condition (4)), we can clarify what we mean by projection by writing

$$X_{-S_0}\mathrm{P}X_{S_0} := X_{S_0}(X_{S_0}^T X_{S_0})^{-1} X_{S_0}^T X_{-S_0}.$$

The anti-projection is denoted by

$$X_{-S_0}\mathrm{A}X_{S_0} = X_{-S_0} - X_{-S_0}\mathrm{P}X_{S_0}.$$

We define the matrix

$$\begin{aligned} V_{-S_0,-S_0} &:= \Big(X_{-S_0}\mathrm{A}X_{S_0}\Big)^T \Big(X_{-S_0}\mathrm{A}X_{S_0}\Big) \\ &= X_{-S_0}^T \Big(I - X_{S_0}(X_{S_0}^T X_{S_0})^{-1} X_{S_0}^T\Big) X_{-S_0}, \end{aligned}$$

and let $\{v_j^2\}_{j \notin S_0}$ be the diagonal elements of this matrix.

7.3. A Lower Bound

The main result for the noisy case is presented in the next theorem. Here, we use the notations and definitions of the previous two subsections.

Theorem 17 *Take for some* $t > 0$,

$$\lambda > \|v_{-S_0}\|_\infty \sqrt{2(\log(2p) + t)}. \tag{11}$$

Define

$$\bar{v}_j := v_j \sqrt{2(\log(2p) + t)}/\lambda, \ j \notin S_0$$

and

$$\bar{u}_j := u_j \sqrt{2(\log(2p) + t)}/\lambda, \ j \in S_0.$$

where $\{u_j\}_{j \in S_0}$ *are the diagonal elements of the matrix* $(X_{S_0}^T X_{S_0})^{-1}$. *Assume that* $\hat{\kappa}(1 + \bar{v}, S_0) > 0$ *and that the following betamin condition holds:*

$$|\beta_j^0| > \lambda(\mathbf{b}_j(\bar{v}) + \bar{u}_j), \ \mathrm{sign}(\beta_j^0) = z_j^*(\bar{v}) \ \forall j \in S_0.$$

Then for all $x > 0$ *with probability at least* $1 - \exp[-t] - \exp[-x]$ *there is a solution* $\hat{\beta}$ *of the KKT conditions such that*

$$\|X(\hat{\beta} - \beta^0)\|_2 \geq \sqrt{\frac{s_0}{\hat{\kappa}^2(1 + \bar{v}, S_0)}} \sqrt{\frac{\lambda^2}{n}} - \sqrt{s_0} - \sqrt{2x}. \tag{12}$$

Note that for $j \in S_0$, the quantity u_j is the variance of the ordinary least squares estimator of β_j^0 for the case S_0 is known. Thus the betamin condition of Theorem 17 needs that the magnitude of the active coefficients should exceed the noise level of the ordinary least squares estimator for known S_0.

8. Comparison with the Noiseless Lasso when the Design is Fixed

This section studies the case of fixed design and compares the noisy Lasso

$$\hat{\beta} := \arg\min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|_2^2 + 2\lambda \|b\|_1 \right\}$$

with the noiseless Lasso

$$\beta^* := \arg\min_{b \in \mathbb{R}^p} \left\{ \|X(b - \beta^0)\|_2^2 + 2\lambda^* \|b\|_1 \right\}$$

where $\lambda^* \leq \lambda$. We let S_* be active set of β^* and its cardinality $s_* := |S_*|$. We investigate the error $\|X(\hat{\beta} - \beta^*)\|_2$ in Theorem 18. For $\lambda^* = 0$ we see that $\beta^* = \beta^0$ and then Theorem 18 gives a bound for $\|X(\hat{\beta} - \beta^0)\|_2$. This is elaborated upon in Corollary 19. The case $\lambda^* = \lambda$ is detailed in Corollary 20. The error $\|X(\hat{\beta} - \beta^*)\|_2^2$ can then seen as "variance" and $\|X(\beta^* - \beta^0)\|_2$ as "bias".

8.1. Projections

We now introduce some notations and definitions similar to the ones in Subsections 7.2, now for general S instead of just $S = S_0$. The projection of X_{-S} on the space spanned by the columns of X_S is denoted by $X_{-S}\mathrm{P}X_S$. Recall that such projections are defined, also if X_S does not have full column rank. The anti-projection is

$$X_{-S}\mathrm{A}X_S := X_{-S} - X_{-S}\mathrm{P}X_S.$$

Define the matrix

$$V_{-S,-S}^S := \Big(X_{-S}\mathrm{A}X_S \Big)^T \Big(X_{-S}\mathrm{A}X_S \Big)$$

and let $\{(v_j^S)^2\}_{j \notin S}$ be the diagonal elements of this matrix.

8.2. Upper Bound

Recall the KKT conditions for β^* as given in (6), involving the vector ζ^* in the sub-differential $\partial \|\beta^*\|_1$.

Theorem 18 *Fix a set S with cardinality $|S| = s$. Assume that that for some $t > 0$*

$$\lambda > \|v_{-S}^S\|_\infty \sqrt{2(\log(2p) + t)} \tag{13}$$

and write

$$\bar{v}_j^S := v_j^S \sqrt{2(\log(2p) + t)}/\lambda, \ j \notin S. \tag{14}$$

Suppose that

$$\lambda^* |\zeta_j^*|/\lambda < 1 - \bar{v}_j^S \ \forall \ j \notin S.$$

Define

$$\bar{w}_j^S := \frac{1 - \bar{v}_j^S - \lambda^* |\zeta_j^*|/\lambda}{1 - \lambda^*/\lambda}, \ j \notin S.$$

We have for all x with probability at least $1 - \exp[-t] - \exp[-x]$

$$\|X(\hat{\beta} - \beta^*)\|_2 \le \sqrt{\frac{s}{\hat{\kappa}^2(\bar{w}^S, S)}} \sqrt{\frac{(\lambda - \lambda^*)^2}{n}} + \sqrt{s} + \sqrt{2x}. \tag{15}$$

Corollary 19 *If we take the tuning parameter λ^* of the noiseless Lasso equal to zero, Theorem 18 gives the following: with probability at least $1 - \exp[-t] - \exp[-x]$*

$$\|X(\hat{\beta} - \beta^0)\|_2 \le \sqrt{s_0/\hat{\kappa}^2(1 - \bar{v}, S_0)} \sqrt{\lambda^2/n} + \sqrt{s_0} + \sqrt{2x}.$$

This result is comparable to results in Sun and Zhang (2012), Belloni and Wang (2014) and Dalalyan et al. (2017), albeit that we do not deal with the extension to the approximately sparse case. One may check that the the combined conclusions of this corollary with that of Theorem 17 also hold with probability at least $1 - \exp[-t] - \exp[-x]$.

Corollary 20 *We can also take $\lambda^* = \lambda$ in Theorem 18. We then formally put $\bar{w}_j^S = \infty$ for all $j \notin S$ and we put $\hat{\kappa}(\bar{w}) = \infty$ as well. Let S with $|S| = s$. Assume that*

$$|\zeta_j^*| < 1 - \bar{v}_j^S \ \forall \ j \notin S \tag{16}$$

(this implies $S \supset S_*$). We have with probability at least $1 - \exp[-t] - \exp[-x]$

$$\|X(\hat{\beta} - \beta^*)\|_2 \le \sqrt{s} + \sqrt{2x}.$$

This result is as in van de Geer (2016), Problem 2.4.

Corollary 20 is of interest only when $\sqrt{s}$ is small enough This is the case if $\hat{\Sigma} := X^T X/n$ has a well behaved maximal eigenvalue $\hat{\Lambda}_{\max}^2$. Indeed, one can show in the same way as in Lemma 24 (where $\hat{\Sigma}$ is replaced by Σ_0) that

$$s \le \left(\frac{\hat{\Lambda}_{\max}^2}{(1 - \|\bar{v}^S\|_\infty)^2} \right) \frac{n}{\lambda^2} \|X(\beta^* - \beta^0)\|_2^2.$$

Thus if $\hat{\Lambda}_{\max}^2 / (\|\hat{\Sigma}\|_\infty (1 - \|\bar{v}^S\|_\infty)^2) = o(\log(2p))$, then $s = o(\|X(\beta^* - \beta^0)\|_2^2)$. However, for the case of fixed design, one might not want to impose such eigenvalue conditions. Alternatively, one may want to resort to irrepresentable conditions. To this end, fix a set $S \supset S_0$. Let for $j \notin S$, the projection of the j^{th} column X_j on X_S be denoted by

$$X_j \mathrm{P} X_S := X_S \gamma_{S,j}.$$

Then it is not difficult to see that for $j \notin S$ $|\zeta_j^*| \le \|\gamma_{S,j}\|_1$. In other words, a sufficient condition for (16) to hold is the irrepresentable condition

$$\|\gamma_{S,j}\|_1 \le 1 - \bar{v}_j^S, \ \forall j \notin S.$$

We conclude that under irrepresentable conditions the squared "bias" $\|X(\beta^* - \beta^0)\|_2^2$ dominates the "variance" $\|X(\hat{\beta} - \beta^*)\|_2^2$.

9. The Total Variation Penalty in the Noisy Case

We continue with the total variation penalty of Section 6, but now in a noisy setting:

$$Y = f^0 + \epsilon,$$

where $f^0 \in \mathbb{R}^n$ is an unknown vector. The least squares estimator with total variation penalty is

$$\hat{f} \in \arg\min_{f \in \mathbb{R}^n} \left\{ \|Y - f\|_2^2 + 2\lambda \text{TV}(f) \right\}. \tag{17}$$

As has become clear from the previous sections, to assess the prediction error in the noisy case one needs to evaluate the compatibility constant $\hat{\kappa}(w, S)$ with weights $w_j \neq 1$ for $j \notin S$. For the upper bound on the prediction error, we need lower bounds on $\hat{\kappa}(w, S)$. These are derived in Dalalyan et al. (2017), Proposition 2. We re-derive (and slightly improve) their result using a different proof (the proof in Dalalyan et al., 2017 applies a probabilistic argument).

Suppose as in Section 6 that the locations of the jumps are $S := \{d_1 + 1, d_1 + d_2 + 1, \ldots, d_1 + \cdots + d_s + 1\}$ for some $\{d_j\}_{j=1}^s \subset \{2, \ldots, n\}$ satisfying $\sum_{j=1}^s d_j + 2 < n$. Let $d_{s+1} := n - \sum_{j=1}^s d_j$. Assume again for simplicity that d_j is even for all $j \in \{2, \ldots, s\}$.

Lemma 21 *Let $w_1, \ldots, w_n$ be non-negative weights. We have*

$$\frac{\sqrt{s+1}}{\hat{\kappa}(w, S)} \leq \|w\|_\infty \frac{\sqrt{s+1}}{\hat{\kappa}(S)} + \sqrt{n \sum_{i=2}^n (w_i - w_{i-1})^2},$$

where as in Theorem 15

$$\frac{s+1}{\hat{\kappa}^2(S)} = \frac{n}{d_1} + \sum_{j=2}^s \frac{4n}{d_j} + \frac{n}{d_{s+1}}.$$

Corollary 22 *Using the notation of Section 8 suppose that λ satisfies (13) with and let $\bar{v} = \bar{v}^{S_0}$ be given in (14), both with $S := S_0$. Define $\bar{v}_i = 0$ for all $i \in S_0$. We then have with $w_i := 1 - \bar{v}_i$, $j \notin S_0 \cup \{1\}$, $w_1 = w_2$ and $w_i = 1$, $i \in S_0$ that*

$$|w_i - w_{i-1}| \leq |v_i - v_{i-1}| / \|v\|_\infty, \quad i = \{2, \ldots, n\}.$$

In Dalalyan et al. (2017) it is shown in their Proposition 3 that

$$\sum_{i=2}^n (v_i - v_{i-1})^2 / \|v\|_\infty^2 \leq (s_0 + 1) \log n / n.$$

Hence one obtains from Lemma 21 with $S = S_0$, combined with Corollary 19,

$$\frac{\sqrt{s_0+1}}{\hat{\kappa}(1 - \bar{v}, S_0)} \leq \sqrt{\frac{s_0+1}{\hat{\kappa}(S_0)}} + \sqrt{(s_0+1)\log n}$$

where as before

$$\frac{s_0+1}{\hat{\kappa}^2(S_0)} = \frac{n}{d_1} + \sum_{j=2}^{s_0} \frac{4n}{d_j} + \frac{n}{d_{s_0+1}}.$$

Thus, with probability at least $1-\exp[-t]-\exp[-x]$

$$\|\hat{f}-f^0\|_2 \le \lambda\biggl(\sqrt{\frac{(s_0+1)}{n\hat{\kappa}^2(S_0)}} + \sqrt{\frac{(s_0+1)\log n}{n}}\biggr) + \sqrt{s_0} + \sqrt{2x}.$$

Theorem 15 implies that

$$\hat{\kappa}(1+\bar{v}, S_0) \le \hat{\kappa}(S_0).$$

Recall that for the combined conclusion of Theorem 17 and Corollary 19 we do not have to change the confidence level (which is $1-\exp[-t]-\exp[-x]$*). We therefore obtain that if the jumps of* f^0 *are sufficiently large in absolute value, as given in Theorem 17, then with probability at least* $1-\exp[-t]-\exp[-x]$

$$\lambda\sqrt{\frac{s_0+1}{n\hat{\kappa}^2(S_0)}} - \sqrt{s_0} - \sqrt{2x} \;\le\; \|\hat{f}-f^0\|_2 \le \lambda\sqrt{\frac{s_0+1}{n\hat{\kappa}^2(S_0)}} + \sqrt{s_0} + \sqrt{2x} + \lambda\sqrt{\frac{(s_0+1)\log n}{n}}.$$

10. Conclusion

This paper establishes that in a sense the squared "bias" of the Lasso dominates the "variance". Moreover, lower bounds for the prediction error are given. These lower bounds often match up to constants or logarithmic factors the upper bounds, or are in fact tight up to smaller order terms. The bounds show that compatibility constants necessarily enter into the picture. The lower bounds require "betamin" conditions, and - for the case of random design - also certain sparsity conditions. It is as yet unclear what can be said when betamin conditions fail to hold. In combination with this, it would also be of great interest to know what happens when the regression coefficients are not (approximately) sparse. The question to what extent the Lasso will have large prediction error when sparseness assumptions are violated (i.e. when the Lasso is used in a scenario not meant for it) still has some open ends.

11. Proofs

11.1. Proofs of the Lemmas in Section 3

Proof of Lemma 5. We have to show that $\hat{\kappa}^2(u,S) \ge \hat{\phi}^2(u,S)$. Write

$$A := \Bigl\{ b : \ \|b_{-S}\|_1 \le \|b_S\|_1/u, \ \|b_S\|_1 > 0 \Bigr\}$$

and

$$B := \left\{ b : \ \|b_S\|_1 - u\|b_{-S}\|_1 > 0 \right\}.$$

Then

$$B \subset A.$$

Thus

$$\begin{aligned} \hat{\phi}^2(u,S) &= \min\left\{ \frac{|S|\|Xb\|_2^2/n}{\|b_S\|_1^2} : \ b \in A \right\} \\ &\leq \min\left\{ \frac{|S|\|Xb\|_2^2/n}{\|b_S\|_1^2} : \ b \in B \right\} \\ &= \hat{\kappa}^2(u,S). \end{aligned}$$

□

Proof of Lemma 6. This lemma bounds the ℓ_1-norm of the minimizer b^* if there is a little room to spare. We have

$$\begin{aligned} \|b_S^*\|_1 - u\|b_{-S}^*\|_1 &\leq \sqrt{|S|/n}\|Xb^*\|_2/\hat{\kappa}(u,S) \\ &= \hat{\kappa}(S)/\hat{\kappa}(u,S). \end{aligned}$$

On the other hand

$$\begin{aligned} \|b_S^*\|_1 - u\|b_{-S}^*\|_1 &= \|b_S^*\|_1 - \|b_{-S}^*\|_1 + (1-u)\|b_{-S}^*\|_1 \\ &= 1 + (1-u)\|b_{-S}^*\|_1. \end{aligned}$$

Thus

$$\|b_{-S}^*\|_1 \leq \frac{\hat{\kappa}(S) - \hat{\kappa}(u,S)}{(1-u)\hat{\kappa}(u,S)},$$

yielding

$$\|b_S^*\|_1 = 1 + \|b_{-S}^*\|_1 \leq \frac{\hat{\kappa}(S) - u\hat{\kappa}(u,S)}{(1-u)\hat{\kappa}(u,S)}.$$

□

Proof of Lemma 7. This lemma shows that one has a bound for the ℓ_1-norm in the "cone condition" if there is a little room to spare. Consider a vector $b \in \mathbb{R}^p$ satisfying

$$\|b_S\|_1 - v\|b_{-S}\|_1 = 1.$$

Since

$$\|b_S\|_1 - v\|b_{-S}\|_1 = \|b_S\|_1 - u\|b_{-S}\|_1 - (v-u)\|b_{-S}\|_1$$

we obtain

$$(v-u)\|b_{-S}\|_1 = \|b_S\|_1 - u\|b_{-S}\|_1 - 1 \leq \|b_S\|_1 - u\|b_{-S}\|_1.$$

Moreover, clearly

$$\|b_S\|_1 - u\|b_{-S}\|_1 = (v-u)\|b_{-S}\|_1 + 1 \geq 1.$$

It follows that

$$\min\Bigg\{ \|Xb\|_2 : \ \|b_S\|_1 - v\|b_{-S}\|_1 = 1 \Bigg\}$$

$$\ge \min\Bigg\{ \|Xb\|_2 : \ (v-u)\|b_{-S}\|_1 \le \|b_S\|_1 - u\|b_{-S}\|_1 , \ \|b_S\|_1 - u\|b_{-S}\|_1 \ge 1 \Bigg\}.$$

Suppose now that for some $c > 1$

$$(v-u)\|b_{-S}\|_1 \le \|b_S\|_1 - u\|b_{-S}\|_1, \ \|b_S\|_1 - u\|b_{-S}\|_1 = c.$$

Define

$$\tilde{b} := b/c.$$

Then

$$(v-u)\|\tilde{b}_{-S}\|_1 \le 1, \ \|\tilde{b}_S\|_1 - u\|\tilde{b}_{-S}\|_1 = 1.$$

Moreover

$$\|Xb\|_2 = c\|X\tilde{b}\|_2 > \|X\tilde{b}\|_2.$$

Therefore

$$\begin{aligned} & \min\Bigg\{ \|Xb\|_2 : \ (v-u)\|b_{-S}\|_1 \le \|b_S\|_1 - u\|b_{-S}\|_1 , \ \|b_S\|_1 - u\|b_{-S}\|_1 \ge 1 \Bigg\} \\ = \ & \min\Bigg\{ \|Xb\|_2 : \ (v-u)\|b_{-S}\|_1 \le 1 , \ \|b_S\|_1 - u\|b_{-S}\|_1 = 1 \Bigg\}. \end{aligned}$$

But if $(v-u)\|b_{-S}\|_1 \le 1$ and $\|b_S\|_1 - u\|b_{-S}\|_1 = 1$ we see that

$$\begin{aligned} \|b\|_1 & \le \|b_S\|_1 + \|b_{-S}\|_1 = 1 + (1+u)\|b_{-S}\|_1 \\ & \le 1 + (1+u)/(v-u). \end{aligned}$$

□

Proof of Lemma 8. This lemma lower bounds the empirical compatibility constant by the theoretical one. Here is a proof. If $\|b_S\|_1 - u\|b_{-S}\|_1 = 1$ we know that

$$1 \le \|\Sigma_0^{1/2} b\|_2 \sqrt{s}/\kappa(u,S).$$

It therefore follows from Lemma 7 that

$$\hat{\kappa}^2(v,S) \ge \Bigg\{ |S|\|Xb\|_2^2/n : \ \|b_S\|_1 - u\|b_{-S}\|_1 = 1, \ \|b\|_1 \le M(u,v)\|\Sigma_0^{1/2} b\|_2 \Bigg\}$$

where

$$M(u,v) := (1 + (1+u)/(v-u))\sqrt{s}/\kappa(u,S) = o(\sqrt{n/(\|\Sigma_0\|_\infty \log(2p))}).$$

In view of Lemma 45 we know that when $M = o(\sqrt{n/(\|\Sigma_0\|_\infty \log(2p))})$, then with probability tending to one

$$\inf_{\|b\|_1 \le M \|\Sigma_0^{1/2} b\|_2} \frac{\|Xb\|_2^2/n}{\|\Sigma_0^{1/2} b\|_2^2} \ge (1-\eta_M)^2$$

for suitable $\eta_M = o(1)$. Hence with probability tending to one

$$\min\left\{ \|Xb\|_2^2/n : \ \|b_S\|_1 - u\|b_{-S}\|_1 = 1, \ \|b\|_1 \le M(u,v)\|\Sigma_0^{1/2} b\|_2 \right\}$$

$$\ge (1-\eta_{M(u,v)})^2 \min\left\{ \|\Sigma_0^{1/2} b\|_2^2 : \ \|b_S\|_1 - u\|b_{-S}\|_1 = 1 \right\} = (1-\eta_{M(u,v)})^2 \kappa^2(u,S).$$

□

11.2. Proof of Theorem 9

The proof is organized as follows. We first present a bound for $\|\Sigma_0(\beta^* - \beta_0)\|_2$ in Lemma 23. This will be used to bound later the number of active variables s_* of β^*, or rather some extended version of it involving sub-differential calculus, see Lemma 24. We then establish in Lemma 25 a deterministic bound assuming we are on some subset of the underlying probability space. Then in Lemma 26 we show that this subset has large probability.

The noiseless Lasso β^* given in (2) satisfies the KKT conditions

$$n\Sigma_0(\beta^* - \beta^0) + \lambda\zeta^* = 0, \ \zeta^* \in \partial\|\beta^*\|_1, \tag{18}$$

where $\partial\|b\|_1$ is the sub-differential of $b \mapsto \|b\|_1$:

$$\partial\|b\|_1 := \left\{ z : \ \|z\|_\infty \le 1, \ z^T b = \|b\|_1 \right\}.$$

This will be used in Lemma 24 and again in Lemma 25. In the latter we also invoke the KKT conditions for $\hat\beta$

$$X^T X(\hat\beta - \beta^0) + \lambda\hat\zeta = X^T\epsilon, \ \hat\zeta \in \partial\|\hat\beta\|_1. \tag{19}$$

11.2.1. A Bound for the Number of Active Variables of β^*

First we bound the prediction error of β^*.

Lemma 23 *Suppose $\kappa^2(S_0) > 0$. Then*

$$n\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2 \le \frac{s_0}{\kappa^2(S_0)} \frac{\lambda^2}{n}.$$

Proof of Lemma 23. This follows from results in the literature and also from a slight adjustment of Theorem 18 in this paper. Let us present a self-contained proof as well. By the KKT conditions (18)

$$-(\beta^* - \beta^0)^T\zeta^* \le \|\beta^0\|_1 - \|\beta^*\|_1 \le \|\beta^*_{S_0} - \beta^0\|_1 - \|\beta^*_{-S_0}\|_1.$$

So if $\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2 > 0$ we obtain by the definition of the compatibility constant $\kappa^2(S_0)$ that

$$n\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2 \le \lambda\sqrt{s_0}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2/\kappa(S_0).$$

This yields the result of the lemma. □

Consider the set $S_* := \{\beta_j^* \neq 0\}$ of active coefficients of β^*. We bound the size of this set. In fact we look at bound for the size of a potentially larger set, namely the set $S_*(\nu) := \{j : |\zeta_j^*| \ge 1-\nu\}$ where $0 \le \nu < 1$ is arbitrary. Note that indeed $S_* \subset S_*(\nu)$. We pin down the value of ν to $\nu = 1/2$ but the argument goes through for other values if one adjusts the constants accordingly. We still keep the symbol ν at places to facilitate tracking the constants.

Lemma 24 *We have that*

$$|S_*(\nu)| \le \frac{\Lambda_{\max}^2}{(1-\nu)^2} \frac{n^2}{\lambda^2} \|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2 \le \frac{\Lambda_{\max}^2}{(1-\nu)^2} \frac{s_0}{\kappa^2(S_0)}.$$

Proof of Lemma 24. Since

$$\|\zeta^*\|_2^2 \ge \|\zeta^*_{S_*(\nu)}\|_2^2 \ge (1-\nu)^2 |S_*(\nu)|$$

it follows from the KKT conditions (18) that

$$(1-\nu)^2 |S_*(\nu)| \le \|\Sigma_0(\beta^* - \beta^0)\|_2^2 \frac{n^2}{\lambda^2} \le \Lambda_{\max}^2 \|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2 \frac{n^2}{\lambda^2}.$$

The proof is completed by applying the upper bound of Lemma 23

$$\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2 \le \frac{s_0}{\kappa^2(S_0)} \frac{\lambda^2}{n^2}.$$

□

11.2.2. Projections

Let $S := S_*(\nu)$, $s := |S|$ (where $\nu = 1/2$). Set

$$\mathbf{U}(S) := \|\epsilon \mathrm{P} X_S\|_2$$

where $\epsilon \mathrm{P} X_S$ is the projection of ϵ on the space spanned by the columns of X_S. Denote the anti-projection of X_{-S} on this space by

$$X_{-S} \mathrm{A} X_S := X_{-S} - X_{-S} \mathrm{P} X_S.$$

11.2.3. Choice of λ

Recall we take for some $t > 0$

$$\lambda \ge 3 \|\Sigma_0\|_\infty^{1/2} \left(\sqrt{2(\log(2p) + t)} + 2(\log(2p) + t) \right).$$

11.2.4. The Sets $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$

Write

$$v_0 := \|\Sigma_0\|_\infty^{1/2}\left(\sqrt{2n(\log(2p)+t)} + 2(\log(2p)+t)\right)/\lambda.$$

We now define a suitable subset of the underlying probability space, on which we can derive the searched for inequality. This subset will be the intersection of the following sets:

$$\begin{aligned} \mathcal{T}_1 &:= \left\{\|(X_{-S}\mathrm{A}X_S)^T\epsilon\|_\infty \le \lambda v_0,\ \mathbf{U}(S) \le \sqrt{s} + \sqrt{2x}\right\}, \\ \mathcal{T}_2 &:= \left\{\ \|(X^TX - n\Sigma_0)(\beta^* - \beta^0)\|_\infty \le \lambda\delta\right\}, \\ \mathcal{T}_3 &:= \left\{\hat\kappa^2((v - v_0 - \delta)/\delta, S) \ge (1-\eta)^2\kappa^2(S)\right\}, \end{aligned}$$

where $x > 0$ is arbitrary, $\delta := \|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2$, and where $\eta \in (0,1)$ is arbitrary. We pin down η to $\eta = 1/2$ like we did with ν. We require that $\nu - v_0 - 2\delta > 0$. Since $\nu = 1/2$ and $v_0 \le 1/3$ this is the case for $\delta \le 1/(12)$. In view of Lemma 23, Theorem 9 is about the case $\delta = o(1)$, so $\delta \le 1/(12)$ will be true for n sufficiently large.

11.2.5. Deterministic Part

Lemma 25 *On $\mathcal{T}_1 \cap \mathcal{T}_2 \cap \mathcal{T}_3$ it holds that*

$$\|X(\hat\beta - \beta^*)\|_2 \le \left(\frac{\Lambda_{\max}}{(1-\nu)}\frac{\sqrt{n}}{\lambda} + \sqrt{\frac{s}{\kappa^2(S)}\frac{\lambda}{(1-\eta)n}}\right)\sqrt{n}\delta + \sqrt{2x}.$$

Proof of Lemma 25. The KKT conditions (18) and (19), for β^* and $\hat\beta$ respectively, are

$$X^TX(\beta^* - \beta^0) + \lambda\zeta^* = Z,$$

with $Z := (X^TX - n\Sigma_0)(\beta^* - \beta^0)$, and

$$X^TX(\hat\beta - \beta^0) + \lambda\hat\zeta = X^T\epsilon.$$

So subtracting the first from the second

$$X^TX(\hat\beta - \beta^*) + \lambda\hat\zeta - \lambda\zeta^* = X^T\epsilon - Z.$$

Multiplying with $\hat\beta - \beta^*$ yields

$$\|X(\hat\beta - \beta^*)\|_2^2 + \lambda(\hat\beta - \beta^*)^T(\hat\zeta - \zeta^*) = (\hat\beta - \beta^*)^T(X^T\epsilon - Z). \tag{20}$$

We write (as in the proof of Theorem 18 ahead) with $S := S_*(\nu)$, $s := |S|$,

$$X_S\hat b_S := X_S(\hat\beta_S - \beta_S^*) + (X_{-S}\mathrm{P}X_S)\hat\beta_{-S}.$$

Since $|\zeta_j^*| \le 1-\nu < 1$ for all $j \notin S$, it must be true that $\beta^*_{-S} = 0$. Therefore

$$X(\hat\beta - \beta^*) = X_S \hat b_S + (X_{-S} \mathrm{A} X_S)\hat\beta_{-S}.$$

So

$$(\hat\beta - \beta^*)^T X^T \epsilon = \hat b_S^T X_S^T \epsilon + \hat\beta_{-S}^T (X_{-S}\mathrm{A}X_S)^T \epsilon.$$

We use that (on $\mathcal{T}_1$)

$$\begin{aligned} \hat b_S^T X_S^T \epsilon &\le \mathbf{U}(S) \|X_S \hat b_S\|_2 \\ &\le \mathbf{U}(S) \|X(\hat\beta - \beta^*)\|_2 \\ &\le (\sqrt{s} + \sqrt{2x}) \|X(\hat\beta - \beta^*)\|_2 \end{aligned}$$

and

$$\hat\beta_{-S}^T (X_{-S}\mathrm{A}X_S)^T \epsilon \le \|\hat\beta_{-S}\|_1 \|(X_{-S}\mathrm{A}X_S)^T\epsilon\|_\infty \le \lambda v_0 \|\hat\beta_{-S}\|_1.$$

Moreover (on $\mathcal{T}_2$)

$$-(\hat\beta - \beta^*)^T Z \le \|\hat\beta - \beta^*\|_1 \|Z\|_\infty \le \lambda\delta \|\hat\beta - \beta^*\|_1.$$

Then

$$\begin{aligned} (\hat\beta - \beta^*)^T(\zeta^* - \hat\zeta) &= \hat\beta^T \zeta^* - \beta^{*T}\zeta^* + \beta^{*T}\hat\zeta - \hat\beta^T\hat\zeta \\ &= \hat\beta^T\zeta^* - \|\beta^*\|_1 + \beta^{*T}\hat\zeta - \|\hat\beta\|_1 \\ &\le \|\hat\beta_S\|_1 - \|\beta_S^*\|_1 + \|\beta_S^*\|_1 - \|\hat\beta_S\|_1 \\ &+ \hat\beta_{-S}^T \zeta_{-S}^* - \|\hat\beta_S\|_1 \\ &= \hat\beta_{-S}^T \zeta_{-S}^* - \|\hat\beta_S\|_1 \\ &\le (1-\nu)\|\hat\beta_{-S}\|_1 - \|\hat\beta_{-S}\|_1 \\ &= -\nu \|\hat\beta_{-S}\|_1. \end{aligned}$$

Inserting these bounds in (20) gives

$$\|X(\hat\beta - \beta^*)\|_2^2 + \lambda(\nu - v_0 - \delta)\|\hat\beta_{-S}\|_1 \le (\sqrt{s} + \sqrt{2x})\|X(\hat\beta - \beta^*)\|_2 + \lambda\delta\|\hat\beta_S - \beta_S^*\|_1.$$

If

$$\|X(\hat\beta - \beta^*)\|_2 \le (\sqrt{s} + \sqrt{2x})$$

we are done as by Lemma 24, $\sqrt{s} \le \Lambda_{\max}\delta n/((1-\nu)\lambda)$. If

$$\|X(\hat\beta - \beta^*)\|_2 > (\sqrt{s} + \sqrt{2x})$$

we get

$$(\nu - v_0 - \delta)\|\hat\beta_{-S}\|_1 < \delta\|\hat\beta_S - \beta_S^*\|_1$$

or

$$\|\hat\beta_S - \beta^*\|_1 - ((\nu - v_0 - \delta)/\delta)\|\hat\beta_{-S}\|_1 > 0.$$

But (on $\mathcal{T}_3$)

$$\begin{aligned}
& \|\hat{\beta}_S - \beta^*_S\|_1 - ((\nu - v_0 - \delta)/\delta)\|\hat{\beta}_{-S}\|_1 \\
\leq & \frac{\sqrt{s}\|X(\hat{\beta} - \beta^*)\|_2}{\sqrt{n}\hat{\kappa}((\nu - v_0 - \delta)/\delta), S)} \\
\leq & \frac{\sqrt{s}\|X(\hat{\beta} - \beta^*)\|_2}{\sqrt{n}\kappa(S)(1-\eta)}.
\end{aligned}$$

This gives

$$\|X(\hat{\beta} - \beta^*)\|_2 \leq \sqrt{s} + \sqrt{2x} + \lambda\delta\sqrt{s}/(\sqrt{n}\kappa(S)(1-\eta)).$$

Again, by Lemma 24, $\sqrt{s} \leq \Lambda_{\max}\delta n/((1-\nu)\lambda)$. We see that

$$\|X(\hat{\beta} - \beta^*)\|_2 \leq \left(\frac{\Lambda_{\max}}{(1-\nu)} \frac{\sqrt{n}}{\lambda} + \frac{\sqrt{s}}{\kappa(S)(1-\eta)} \frac{\lambda}{(1-\eta)n} \right) \sqrt{n}\delta + \sqrt{2x}.$$

□

11.2.6. Random Part

We apply the tools of Section 12.

Lemma 26 *It holds that*

$$\mathbb{P}\left(\mathcal{T}_1 \cap \mathcal{T}_2 \cap \mathcal{T}_3 \right) \geq 1 - 4\exp[-t] - \exp[-x] - o(1).$$

Proof of Lemma 26 . We first show that $\mathbb{P}(\mathcal{T}_1) \geq 1 - 2\exp[-t] - \exp[-x]$. One component of this is to show that with probability at least $1 - 2\exp[-t]$

$$\|(X_{-S}\mathrm{A}X_S)^T\epsilon\|_\infty \leq \lambda v_0.$$

For a square matrix B, let $\mathrm{diag}(B)$ be its diagonal. By Lemma 41 we know that with probability at least $1 - \exp[-t]$

$$\|(X_{-S}\mathrm{A}X_S)^T\epsilon\|_\infty \leq \|\mathrm{diag}((X_{-S}\mathrm{A}X_S)^T(X_{-S}\mathrm{A}X_S))\|_\infty^{1/2}\sqrt{2(\log(2p)+t)}.$$

But

$$\|\mathrm{diag}((X_{-S}\mathrm{A}X_S)^T(X_{-S}\mathrm{A}X_S))\|_\infty \leq \|\mathrm{diag}(X^TX)\|_\infty.$$

Moreover in view of Lemma 42, and using the union bound, with probability at least $1 - \exp[-t]$

$$\left| \|\mathrm{diag}(X^TX)\|_\infty^{1/2} - \sqrt{n}\|\mathrm{diag}(\Sigma_0)\|_\infty^{1/2} \right| \leq \|\Sigma_0\|_\infty^{1/2}\sqrt{2(\log(2p)+t)}.$$

So with probability at least $1 - 2\exp[-t]$

$$\|(X_{-S}\mathrm{A}X_S)^T\epsilon\|_\infty \leq \|\Sigma_0\|_\infty^{1/2}\left(\sqrt{2n(\log(2p)+t)} + 2(\log(2p)+t) \right) \leq \lambda v_0.$$

The second component is to show that

$$\mathbb{P}(\mathbf{U}(S) \le \sqrt{s} + \sqrt{2x}) \le \exp[-x],$$

but this follows immediately from Lemma 42.

Next we show that $\mathbb{P}(\mathcal{T}_2) \le 2\exp[-t]$. Set $Z := (X^T X - n\Sigma_0)(\beta^* - \beta^0)$. Clearly $X(\beta^* - \beta^0)$ is a Gaussian vector with i.i.d. entries with mean zero and variance $\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2$. Hence, applying Lemma 43 with $\sigma_u^2 \le \|\Sigma_0\|_\infty$, $\sigma_v^2 = \|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2^2$ and using the union bound, we obtain that with probability at least $1 - 2\exp[-t]$

$$\|Z\|_\infty \le 3\|\Sigma_0\|_\infty^{1/2}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2\biggl(\sqrt{2n(\log(2p) + t} + \log(2p) + t\biggr).$$

Finally, the result $\mathbb{P}(\mathcal{T}_3) = 1 - o(1)$ follows from Lemma 8. □

11.2.7. Collecting the pieces

Combining Lemma 25 with Lemma 26 completes the proof of Theorem 9.

11.3. Proof of Theorems 10 and 11

We use concentration of measure, Lemma 44.

Proof of Theorem 10. Let $m^* := \mathbb{E}(\|X(\hat\beta - \beta^*)\|_2 | X)$. Then we have (by Lemma 44) that with probability at least $1 - 1/8 - 3/4 - o(1)$

$$\|X(\hat\beta - \beta^*)\|_2 \ge m^* - 2\sqrt{\log 2}$$

as well as (by Theorem 9),

$$\|X(\hat\beta - \beta^*)\|_2 \le \gamma\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2 + 2\sqrt{\log 2}.$$

Thus

$$m^* \le \gamma\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2 + 4\sqrt{\log 2}.$$

Applying again Lemma 44 we see that

$$\mathbb{P}\biggl(\|X(\hat\beta - \beta^*)\| \ge \gamma\sqrt{n}\|\Sigma_0^{1/2}(\beta^* - \beta^0)\|_2 + 4\sqrt{\log 2} + \sqrt{2x}\biggr)$$
$$\le \mathbb{P}\biggl(\|X(\hat\beta - \beta^*)\| \ge m^* + \sqrt{2x}\biggr) \le 2\exp[-x].$$

□

Proof of Theorem 11. By the triangle inequality

$$\biggl|\|X(\hat\beta - \beta^0)\|_2 - \|X(\beta^* - \beta^0)\|_2\biggr| \le \|X(\hat\beta - \beta^*)\|_2.$$

By Lemma 42, with with probability at least $1-2/n$

$$\left| \|X(\beta^*-\beta^0)\|_2 - \sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2 \right| \le (\sqrt{2\log n})\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2.$$

So, invoking Theoerem 9, with probability at least $1-4\exp[-t]-\exp[-x]-o(1)-2/n$ (subtracting the term $2/n$ to follow the argument, as of course it can be included in the $o(1)$ term)

$$\left| \|X(\hat\beta-\beta^0)\|_2 - \sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2 \right| \le (\gamma+\sqrt{2\log n/n})\sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2+\sqrt{2x}.$$

Let $m^0 := \mathbb{E}(\|X(\hat\beta-\beta^0)\|_2|X)$. Using the same arguments as in Theorem 10, we arrive at

$$m^0-2\sqrt{\log 2} \le (1+\gamma+\sqrt{2\log n/n})\sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2+2\sqrt{\log 2}$$

and

$$(1-\gamma-\sqrt{2\log n/n})\sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2-2\sqrt{\log 2} \le m^0+2\sqrt{\log 2},$$

or

$$\left| m^0 - \sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2 \right| \le \left(\gamma+\sqrt{2\log n/n}\right)\sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2+4\sqrt{\log 2}.$$

Thus, inserting the triangle inequality,

$$\begin{aligned} &\left| \|X(\hat\beta-\beta^0)\|_2 - \sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2 \right| \\ \le\ & \left| \|X(\hat\beta-\beta^0)\|_2 - m^0 \right| + \left| m^0 - \sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2 \right| \\ \le\ & \left| \|X(\hat\beta-\beta^0)\|_2 - m^0 \right| + (\gamma+\sqrt{2\log n/n})\sqrt{n}\|\Sigma_0^{1/2}(\beta^*-\beta^0)\|_2+4\sqrt{\log 2}. \end{aligned}$$

Apply Lemma 44 again to finalize the result. □

11.4. Proof of Theorem 14

To establish Theorem 14, we first need to study the minimizer b^* in (7). The minimization

$$\min\left\{ \|Xb\|_2^2 :\ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \right\}$$

has non-convex constraints. If we fix the signs within S of a possible solution b, one can reformulate it as a convex problem with convex constraints. This is done in Lemma 27. We then show that $b_j^* \ne 0$ for all $j \in S$ in Lemma 28. This is important because given the signs within S of a potential solution b, we want the restrictions on these signs to be non-active so that the Lagrangian formulation is of a similar form as the KKT conditions (6) for the noiseless Lasso. This Lagrangian form is then given in Lemma 31 with Lemma 30 serving as a preparation. The Lagrangian form of Lemma 31 with $S=S_0$ in a sense resembles the KKT conditions (6) when the active coefficients in the vector β_S^0 have appropriate signs and $|\beta_j^0|$ is for $j\in S_0$ large enough. This allows one to find a solution β^* of the KKT conditions (6) with the prescribed prediction error.

11.4.1. Non-Sparseness within S

Our first step is to ascertain that a solution

$$b^* \in \arg\min_{b \in \mathbb{R}^p} \left\{ \|Xb\|_2 : \ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \right\}$$

can be found by searching over (at most) $2^{|S|}$ convex problems with convex constraints. This is done in the next lemma, where we also show that the equality constraint $\|b_S\|_1 - \|b_{-S}\|_1 = 1$ can be replaced by an inequality constraint $\|b_S\|_1 - \|b_{-S}\|_1 \geq 1$.

Lemma 27 *We have*

$$\begin{aligned} & \min\left\{ \|Xb\|_2^2 : \ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \right\} \\ = \ & \min\left\{ \|Xb\|_2^2 : \ \|b_S\|_1 - \|b_{-S}\|_1 \geq 1 \right\} \\ = \ & \min_{z_S \in \{\pm 1\}^{|S|}} \min_b \left\{ \|Xb\|_2^2 : \ z_S^T b_S - \|b_{-S}\|_1 \geq 1, \ z_j b_j \geq 0 \ \forall \ j \in S \right\}. \end{aligned}$$

Proof of Lemma 27. To show that the equality constraint can be turned into an inequality constraint let us consider some $b \in \mathbb{R}^p$ for which it holds that $\|b_S\|_1 - \|b_{-S}\|_1 = c$, where c is a constant bigger than 1. Let $\tilde{b} := b/c$. Then

$$\|\tilde{b}_S\|_1 - \|\tilde{b}_{-S}\|_1 = \left(\|b_S\|_1 - \|b_{-S}\|_1 \right) / c = 1.$$

Moreover

$$\|X\tilde{b}\|_2 = \|Xb\|_2 / c < \|Xb\|_2.$$

Thus the first equality of the lemma must be true.

We now show the second equality of the lemma. If for some $z_S \in \{\pm 1\}$ it holds that $z_j b_j \geq 0$ for all $j \in S$, we have $z_S^T b_S = \|b_S\|_1$. Conversely, if we define for $j \in S$ with $b_j \neq 0$, $z_j := b_j / |b_j|$ as the sign of b_j, and define $z_j \in \{\pm 1\}$ arbitrarily for $j \in S$ with $b_j = 0$, then we have $z_j b_j \geq 0$ for all $j \in S$. Thus

$$\left\{ b : \ \|b_S\|_1 - \|b_{-S}\|_1 \geq 1 \right\} = \cup_{z_S \in \{\pm 1\}^{|S|}} \left\{ b : \ z_S^T b_S - \|b_{-S}\|_1 \geq 1, \ z_j b_j \geq 0 \right\}.$$

□

We establish in the next lemma that sign constraints on b_S^* are not active: b_S^* is so to speak maximally non-sparse. We assume that $\hat{\kappa}^2(S) > 0$, so for $S = S_0$ we implicitly assume Condition 4.

Lemma 28 *Suppose that $\hat{\kappa}(S) \neq 0$. Then for any minimizer b^* of the problem*

$$\min\left\{ \|Xb\|_2 : \ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \right\}$$

it holds that $b_j^* \neq 0$ for all $j \in S$.

Remark 29 *A (very) special case of Lemma 28 is the minimization problem*

$$\mathrm{b}_S^* \in \arg\min\left\{ \|b_S\|_2^2 : \ \|b_S\|_1 = 1 \right\}.$$

Clearly the solution has $|\mathrm{b}_j^*| = 1/|S| \neq 0$ *for all* $j \in S$*. More generally, for the case without "*b_{-S}*-part" one can apply a geometric argument to show that whenever* $X_S^T X_S$ *is non-singular*

$$\mathrm{b}_S^* \in \arg\min\{ \|X b_S\|_2 : \ \|b_S\|_1 = 1\}$$

must have all its components in S *nonzero.*

Proof of Lemma 28. We use the representation of Lemma 27. Let $z_S^* \in \{\pm 1\}^{|S|}$ satisfy $z_S^{*T} b_S^* = \|b_S^*\|_1$ and $z_j^* b_j^* \geq 0$ for all $j \in S$. Then b^* is a solution of the convex minimization problem with (linear and) convex constraints

$$\min\left\{ \|Xb\|_2^2 : \ z_S^{*T} b_S - \|b_{-S}\|_1 \geq 1, z_j^* b_j \geq 0, \ \forall \ j \in S \right\}.$$

Note that in the minimization, one may replace the inequality constraint $z_S^{*T} b_S - \|b_{-S}\|_1 \geq 1$ by an inequality constraint $z_S^{*T} b_S - \|b_{-S}\|_1 = 1$. This follows from the same arguments as used in the proof of Lemma 27. A reason to replace the equality constraint by an inequality constraint is that the restrictions become convex.

The solution of the convex problem with convex constraints can be found using Lagrange multipliers $\tilde{\lambda}$ and μ_S, where $\tilde{\lambda} \geq 0$ and where μ_S is an $|S|$-vector with non-negative entries. The Lagrangian formulation is

$$\min\left\{ \|Xb\|_2^2 + 2\tilde{\lambda}\left(\|b_{-S}\|_1 - z_S^{*T} b_S - 1 \right) - 2 \sum_{j \in S} \mu_{j,S} z_j^* b_j \right\}.$$

Because the inequality constraint can be replaced by an equality constraint, we know that in fact $\tilde{\lambda} > 0$. The Lagrangian formulation has has KKT conditions

$$X^T X b^* = \tilde{\lambda} z^* + \mathrm{diag}(\mu_S) z_S^*,$$

where z_{-S}^* is an element of the sub-differential

$$-\partial \|b_{-S}^*\|_1 = \left\{ z_{-S} : \ \|z_{-S}\|_1 \leq 1, \ z_{-S}^T b_{-s}^* = -\|b_{-S}^*\|_1 \right\}.$$

It follows that for $j \in S$

$$b_j^* \neq 0 \ \Rightarrow \mu_{j,S} = 0.$$

Let $\mathcal{N} := \{ j \in S : \ b_j^* = 0\}$. Then we have by the above argument

$$(X^T X b^*)_{-\mathcal{N}} = \tilde{\lambda} z_{-\mathcal{N}}^*$$

$$(X^T X b^*)_{\mathcal{N}} = \tilde{\lambda} z_{\mathcal{N}}^* + \mathrm{diag}(\mu_{\mathcal{N}}) z_{\mathcal{N}}^*.$$

The tangent plane of $\{b: \ \|Xb\|_2 = \|Xb^*\|_2\}$ at b^* is

$$\mathcal{U} := \{u = b^* + v: \ v^T X^T X b^* = 0\}.$$

The idea of the proof is now to take an element $u = b^* + tv$ in this tangent plane with $t > 0$ and with $v_j \neq 0$ for at least one $j \in \mathcal{N}$ and such that $v_j \neq 0$ has the same sign as b_j^* for all $j \in S \backslash \mathcal{N}$. For $j \notin S$ we take $v_j = 0$. Then $\tilde{b} := b^* + tv$ has $\|\tilde{b}_S\|_1 - \|\tilde{b}_{-S}\|_1 > 1$ and this leads for a suitable scale t to

$$\frac{\|X\tilde{b}\|_2}{\|\tilde{b}_S\|_1 - \|\tilde{b}_{-S}\|_1} < \|Xb^*\|_2.$$

Let us now work out this idea. It cannot be true that $b_j^* = 0$ for all $j \in S$ as $\|b_S^*\|_1 \geq 1$. Hence $S \backslash \mathcal{N} \neq \emptyset$. Take (for example) $v_j = z_j^*$ for all $j \in S \backslash \mathcal{N}$. Then

$$v_{S\backslash\mathcal{N}}^T z_{S\backslash\mathcal{N}}^* = z_{S\backslash\mathcal{N}}^{*T} z_{S\backslash\mathcal{N}}^* = |S\backslash\mathcal{N}|.$$

Now $\tilde{\lambda} > 0$ and the entries of $\mu_{\mathcal{N}}$ are all positive as well (since $\mu_j = 0$ for some $j \in \mathcal{N}$ would imply $b_j^* = 0$ for this j, which is not possible by the definition of $\mathcal{N}$). Therefore we can choose

$$v_{\mathcal{N}}^T (\tilde{\lambda} z_{\mathcal{N}}^* + \text{diag}(\mu_{\mathcal{N}}) z_{\mathcal{N}}^*) = -\tilde{\lambda} |S\backslash\mathcal{N}|.$$

Then at least one entry of $v_{\mathcal{N}}$ has to be non-zero and moreover

$$\begin{aligned} v^T X^T X b^* &= \tilde{\lambda} v_{S\backslash\mathcal{N}}^T z_{S\backslash\mathcal{N}}^* + v_{\mathcal{N}}^T (\tilde{\lambda} z_{\mathcal{N}}^* + \text{diag}(\mu_{\mathcal{N}}) z_{\mathcal{N}}^*) \\ &= \tilde{\lambda} |S\backslash\mathcal{N}| - \tilde{\lambda} |S\backslash\mathcal{N}| \\ &= 0. \end{aligned}$$

We thus have for all $t > 0$

$$\|X(b^* + tv)\|_2^2 = \|Xb^*\|_2^2 + t^2 \|Xv\|_2^2.$$

Moreover

$$\begin{aligned} \|b_S^* + tv_S\|_1 &= \|b_{S\backslash\mathcal{N}}^*\|_1 + t\|v_{S\backslash\mathcal{N}}\|_1 + t\|v_{\mathcal{N}}\|_1 \\ &= \|b_S^*\|_1 + t\|v\|_1. \end{aligned}$$

Therefore

$$\begin{aligned} \|b_S^* + tv_S\|_1 - \|b_{-S}^*\|_1 &= \|b_S^*\|_1 - \|b_{-S}^*\|_1 + t\|v\|_1 \\ &= 1 + t\|v\|_1. \end{aligned}$$

It follows that

$$\begin{aligned} &\frac{\|X(b^* + tv)\|_2^2}{(\|b_S^* + tv_S\|_1 - \|b_{-S}^*\|_1)^2} \\ &= \frac{\|Xb^*\|_2^2 + t^2 \|Xv\|_2^2}{(1 + t\|v\|_1)^2}. \end{aligned}$$

Define

$$\begin{aligned} \mathrm{A} &:= \|Xb^*\|_2^2 + t^2\|Xv\|_2^2 - \|Xb^*\|_2^2(1+t\|v\|_1)^2 \\ &= t^2\|Xv\|_2^2 - 2t\|Xb^*\|_2^2\|v\|_1 - t^2\|Xb^*\|_2^2\|v\|_1^2 \\ &= t^2(\|Xv\|_2^2 - \|Xb^*\|_2^2\|v\|_1^2) - 2t\|Xb^*\|_2^2\|v\|_1^2. \end{aligned}$$

We will show that for suitable $t > 0$ the constant A is strictly negative. This means

$$\|X(b^* + tv)\|_2^2 < \|Xb^*\|_2^2(\|b_S^* + tv_S\|_1 - \|b_{-S}^*\|_1)^2$$

and so we arrive at a contradiction. To show $\mathrm{A} < 0$ we distinguish two cases. If

$$\|Xv\|_2^2 \le \|Xb^*\|_2^2\|v\|_1^2$$

then $\mathrm{A} < 0$ for all $t > 0$. If

$$\|Xv\|_2^2 > \|Xb^*\|_2^2\|v\|_1^2$$

then $\mathrm{A} < 0$ for all t satisfying

$$0 < t < \frac{2\|Xb^*\|_2^2\|v\|_1^2}{\|Xv\|_2^2 - \|Xb^*\|_2^2\|v\|_1^2}.$$

Here we used the assumption that $\|Xb^*\|_2^2 > 0$ so that the above right hand side is indeed strictly positive. □

11.4.2. Lagrangian Form

We now present the Lagrangian form given the signs within the set S and given that within the set S the solution has non-zero entries. Let for each $z_S \in \{\pm 1\}^{|S|}$

$$b^*(z_S) \in \arg\min\left\{ \|Xb\|_2^2 : \ z_S^T b_S - \|b_{-S}\|_1 \ge 1, z_j b_j \ge 0, \ \forall\ j \in S \right\}.$$

Define

$$\mathcal{Z}_S := \left\{ z_S \in \{-1, 1\}^{|S|} : \ z_j b_j^*(z_S) > 0 \ \forall\ j \in S \right\}.$$

Lemma 30 *We have for all* $z_S \in \mathcal{Z}_S$

$$X^T X b^*(z_S) = z^*(z_S)\|Xb^*(z_S)\|_2^2$$

where $z_S^*(z_S) = z_S$ *and* $z_{-S}^*(z_S) \in -\partial\|b_{-S}^*(z_S)\|_1$.

Proof of Lemma 30. To prove this result it is useful to repeat some arguments of the proof of Lemma 28. The convex minimization problem with (linear and) convex constraints

$$\min\left\{ \|Xb\|_2^2 : \ z_S^T b_S - \|b_{-S}\|_1 \ge 1, z_j b_j \ge 0, \ \forall\ j \in S \right\}$$

can be solved using Lagrange multipliers $\tilde{\lambda}$ and μ_S, where $\tilde{\lambda} > 0$ and μ_S is an $|S|$-vector with non-negative entries. The Lagrangian formulation is

$$\min\left\{ \|Xb\|_2^2 + 2\tilde{\lambda}\left(\|b_{-S}\|_1 - z_S^T b_S - 1 \right) - 2\sum_{j\in S} \mu_{j,S} z_j b_j \right\}.$$

This has KKT conditions

$$X^T X b^*(z_S) = \tilde{\lambda} z^* + \mathrm{diag}(\mu_S) z_S,$$

where $z_S^* = z_S$ and $z_{-S}^* = z_{-S}^*(z_S)$ depends on z_S and is an element of the sub-differential

$$-\partial \|b_{-S}^*(z_S)\|_1 = \left\{ z_{-S} :\ \|z_{-S}\|_\infty \le 1,\ z_{-S}^T b_{-S}^*(z_S) = -\|b_{-S}^*\|_1 \right\}.$$

It follows that for $j \in S$

$$b_j^*(z_S) \ne 0 \Rightarrow \mu_{j,S} = 0.$$

The assumption that $z_S \in \mathcal{Z}_S$ thus gives $\mu_S = 0$. The KKT conditions then read

$$X^T X b^*(z_S) = \tilde{\lambda} z^*.$$

One sees that

$$1 = z^{*T} b^*(z_S) = b^{*T}(z_S) X^T X b^*(z_S)/\tilde{\lambda} = \|Xb^*(z_S)\|_2^2/\tilde{\lambda}.$$

This gives

$$\tilde{\lambda} = \|Xb^*(z_S)\|_2^2 \ .$$

□

We apply the above lemma with $z_S := \partial \|b_S^*\|_1$. This gives the following result.

Lemma 31 *Suppose* $\hat{\kappa}(S) \ne 0$. *Let*

$$b^* \in \arg\min\left\{ \|Xb\|_2^2 :\ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \right\}$$

Then

$$X^T X b^* = z^* \|Xb^*\|_2^2.$$

where $z_S^* = \partial \|b_S^*\|_1$ *and* $z_{-S}^* \in -\partial \|b_{-S}^*\|_1$.

Proof of Lemma 31. By Lemma 28, for each

$$b^* \in \arg\min\left\{ \|Xb\|_2^2 :\ \|b_S\|_1 - \|b_{-S}\|_1 = 1 \right\}$$

it holds that $b_j^* \ne 0$ for all $j \in S$. We can therefore define $z_j^* := b_j^*/|b_j^*|$ for all $j \in S$ and then $z_S^* = \partial \|b_S^*\|_1 \in \mathcal{Z}_S$. The result now follows from Lemma 30. □

11.4.3. Finalizing the Proof of Theorem 14

With the help of Lemma 31 we are now in the position to prove Theorem 14.

Proof of Theorem 14. Let b^* and z^* be as in Lemma 31, with $S = S_0$. Define

$$\beta' = \beta^0 - \frac{b^* s_0}{\hat{\kappa}^2(S_0)} \frac{\lambda^*}{n}.$$

Then

$$\begin{aligned} X^T X(\beta' - \beta^0) &= -\frac{\lambda^* X^T X b^* s_0}{n\hat{\kappa}^2(S_0)} \\ &= -\frac{\lambda^* X^T X b^*}{\|Xb^*\|_2^2} \\ &= -\lambda^* z^*. \end{aligned}$$

Let $S_* := \{j : \ b_j^* \neq 0\}$. Then by Lemma 28, $S_0 \subset S_*$. Furthermore

$$z_j^* \beta_j' = \begin{cases} z_j^* \beta_j^0 - \lambda z_j^* b_j^* / \|Xb^*\|_2^2 > 0 & j \in S_0 \\ -\lambda^* z_j^* b_j^* / \|Xb^*\|_2^2 > 0 & j \in S_* \backslash S_0 \\ 0 & j \notin S_* \end{cases}.$$

It follows that $z^* \in \partial \|\beta'\|_1$. Thus, $\beta' =: \beta^*$ is a solution of the KKT conditions (6) with $\zeta^* = z^*$. It holds moreover that

$$\|X(\beta^* - \beta^0)\|_2^2 = \frac{\lambda^{*2} \|Xb^*\|_2^2}{\|Xb^*\|_2^4} = \frac{\lambda^{*2} s_0}{n\hat{\kappa}^2(S_0)}.$$

□

11.5. Proof of Theorem 15

The proof of Theorem 15 consists of several steps. First we note that, given the sizes of its jumps, the total variation of a function is the smallest when this function is decreasing or increasing. This is stated in Lemma 32 as a trivial fact. As a consequence, if one subtracts from an arbitrary function value - or minus this value - the total variation, the result will be at most the average of the absolute values. This is shown in Lemma 33. Lemma 33 is then applied at each jump separately, as $\|b_S\|_1 - \|b_{-S\cup\{1\}}\|_1$ in this example amounts to subtracting at each jump some total variation to the left or to the right of this jump. Lemma 34 shows how this works for one jump. Then Theorem 15 is in part proved by applying this lemma to each jump. This leads to a lower bound for $\hat{\kappa}^2(S)$. The proof is completed by showing that this lower bound is achieved by the vector b^* as given in Theorem 15.

For $f \in \mathbb{R}^n$ we define the ordered vector

$$f_{(1)} \leq \cdots \leq f_{(n)},$$

with arbitrary ordering within ties.

Lemma 32 *It holds that*

$$TV(f) \geq f_{(n)} - f_{(1)}$$

with equality if f is increasing or decreasing.

Proof of Lemma 32. Trivial. □

Lemma 33 *It holds for any $j \in \{1, \ldots, n\}$ that*

$$f_j - \text{TV}(f) \leq f_{(1)} \leq \frac{1}{n} \sum_{i=1}^n |f_i|,$$

and

$$-f_j - \text{TV}(f) \leq -f_{(n)} \leq \frac{1}{n} \sum_{i=1}^n |f_i|.$$

Proof of Lemma 33. We have from Lemma 32 that $\text{TV}(f) \geq f_{(n)} - f_{(1)}$. Moreover, $f_j \leq f_{(n)}$. Thus

$$\begin{aligned} f_j - \text{TV}(f) &\leq f_j - (f_{(n)} - f_{(1)}) \\ &\leq f_{(n)} - (f_{(n)} - f_{(1)}) \\ &= f_{(1)}. \end{aligned}$$

Case 1: if $f_{(1)} < 0$ obviously $f_{(1)} < \frac{1}{n} \sum_{i=1}^n |f_i|$.
Case 2: if $f_{(1)} \geq 0$ then $f_i \geq 0$ for all i and then

$$f_{(1)} \leq \sum_{i=1}^n f_i/n = \sum_{i=1}^n |f_i|/n.$$

In the same way

$$\begin{aligned} -f_j - \text{TV}(f) &\leq -f_j - (f_{(n)} - f_{(1)}) \\ &\leq -f_{(1)} - (f_{(n)} - f_{(1)}) \\ &= -f_{(n)}. \end{aligned}$$

Case 1: if $f_{(n)} > 0$ then $-f_{(n)} < \frac{1}{n} \sum_{i=1}^n |f_i|$.
Case 2: if $f_{(n)} \leq 0$ then $f_i \leq 0$ for all i and then

$$-f_{(n)} \leq -\sum_{i=1}^n f_i/n = \sum_{i=1}^n |f_i|/n.$$

□

Lemma 34 *Let $f \in \mathbb{R}^n$ with total variation* $\text{TV}(f) = \sum_{i=2}^n |f_i - f_{i-1}|$ *and $g \in \mathbb{R}^m$ with total variation* $\text{TV}(g) = \sum_{i=2}^m |g_i - g_{i-1}|$. *Then for any $j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$*

$$|f_j - g_k| - \text{TV}(f) - \text{TV}(g) \leq \frac{1}{n} \sum_{i=1}^n |f_i| + \frac{1}{m} \sum_{i=1}^m |g_i|.$$

Proof of Lemma 34. Suppose without loss of generality that $f_j \ge g_k$. Then by Lemma 33

$$\begin{aligned}|f_j - g_k| - \text{TV}(f) - \text{TV}(g) &= \underbrace{(f_j - \text{TV}(f))}_{\le \sum_{i=1}^n |f_i|/n} + \underbrace{(-g_k - \text{TV}(g))}_{\le \sum_{i=1}^m |g_i|/m} \\ &\le \frac{1}{n}\sum_{i=1}^n |f_i| + \frac{1}{m}\sum_{i=1}^m |g_i|.\end{aligned}$$

□

Proof of Theorem 15. Let for $j = 2, \ldots, s$, $u_j \in \mathbb{N}$ satisfy $1 \le u_j \le d_j - 1$. We may write for $f = Xb$,

$$\begin{aligned}
&\|b_S\|_1 - \|b_{-(S\cup\{1\})}\|_1 \\
= \ & |f_{d_1+1} - f_{d_1}| - \sum_{i=2}^{d_1} |f_i - f_{i-1}| - \sum_{i=d_1+2}^{d_1+u_2} |f_i - f_{i-1}| \\
+ \ & |f_{d_1+d_2+1} - f_{d_1+d_2}| - \sum_{i=d_1+u_2+1}^{d_1+d_2} |f_i - f_{i-1}| - \sum_{i=d_1+d_2+2}^{d_1+d_2+u_3} |f_i - f_{i-1}| \\
& \cdots \\
+ \ & |f_{d_1+\cdots+d_{s-1}+1} - f_{d_1+\cdots+d_{s-1}}| \\
- \ & \sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+1}^{d_1+\cdots+d_{s-1}} |f_i - f_{i-1}| - \sum_{i=d_1+\cdots+d_{s-1}+2}^{d_1+\cdots+d_{s-1}+u_s} |f_i - f_{i-1}| \\
+ \ & |f_{d_1+\cdots+d_s+1} - f_{d_1+\cdots+d_s}| \\
- \ & \sum_{i=d_1+\cdots+d_{s-1}+u_s+1}^{d_1+\cdots+d_s} |f_i - f_{i-1}| - \sum_{i=d_1+\cdots+d_s+2}^{n} |f_i - f_{i-1}|
\end{aligned}$$

$$\begin{aligned}
\le \ & \frac{1}{d_1}\sum_{i=1}^{d_1} |f_i| + \frac{1}{u_2}\sum_{i=d_1+1}^{d_1+u_2} |f_i| \\
+ \ & \frac{1}{d_2 - u_2}\sum_{i=d_1+u_2+1}^{d_1+d_2} |f_i| + \frac{1}{u_3}\sum_{i=d_1+d_2+1}^{d_1+d_2+u_3} |f_i| \\
& \cdots \\
+ \ & \frac{1}{d_{s-1} - u_{s-1}}\sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+1}^{d_1+\cdots+d_{s-1}} |f_i| + \frac{1}{u_s}\sum_{i=d_1+\cdots+d_{s-1}+1}^{d_1+\cdots+d_{s-1}+u_s} |f_i| \\
+ \ & \frac{1}{d_s - u_s}\sum_{i=d_1+\cdots+d_{s-1}+u_s+1}^{d_1+\cdots+d_s} |f_i| + \frac{1}{d_{s+1}}\sum_{i=d_1+\cdots+d_s+1}^{n} |f_i|
\end{aligned}$$

$$\leq \sqrt{\frac{1}{d_1} + \frac{1}{u_2} + \frac{1}{d_2 - u_2} + \cdots + \frac{1}{d_{s-1} - u_{s-1}} + \frac{1}{u_s} + \frac{1}{d_s - u_s} + \frac{1}{d_{s+1}}}$$
$$\times \sqrt{\sum_{i=1}^{n} |f_i|^2},$$

where in the first inequality we applied Lemma 34 and the second one follows from the Cauchy-Schwarz inequality. The assumption that for all $j \in \{2, \ldots, s\}$ d_j is even allows us to take $u_j = d_j/2$ to arrive at

$$\kappa^2(S) \geq \frac{s+1}{\frac{n}{d_1} + \sum_{j=2}^{s} \frac{4n}{d_j} + \frac{n}{d_{s+1}}}.$$

Now for the reverse inequality, let $\tilde{b}$ be given as in the theorem and and $\tilde{f} := X\tilde{b}$. Then $\tilde{f}$ is equal to

$$\tilde{f}_i = \begin{cases} -\frac{n}{d_1} & i = 1, \ldots, d_1 \\ \frac{2n}{d_2} & i = d_1 + 1, \ldots, d_1 + d_2 \\ \vdots & \\ (-1)^s \frac{2n}{d_s} & i = \sum_{j=1}^{s-1} d_j + 1, \ldots, \sum_{j=1}^{s} d_j \\ (-1)^{s+1} \frac{n}{d_{s+1}} & i = \sum_{j=1}^{s} d_j + 1, \ldots, n \end{cases}.$$

By the definition of $\tilde{f} = X\tilde{b}$,

$$\begin{aligned} \|\tilde{b}_S\|_1 = \sum_{j=1}^{s} |\tilde{f}_{d_j+1} - \tilde{f}_{d_j}| &= \frac{n}{d_1} + \frac{2n}{d_2} \\ &+ \frac{2n}{d_2} + \frac{2n}{d_3} \\ &\vdots \\ &+ \frac{2n}{d_{s-1}} + \frac{2n}{d_s} \\ &+ \frac{2n}{d_s} + \frac{n}{d_{s+1}} \\ &= \frac{n}{d_1} + \sum_{j=2}^{s} \frac{4n}{d_j} + \frac{n}{d_{s+1}}, \end{aligned}$$

and also

$$\begin{aligned} \sum_{i=1}^{n} \tilde{f}_i^2 &= d_1 \tilde{f}_{t_1}^2 + \cdots + d_{s+1} \tilde{f}_{d_{s+1}}^2 \\ &= \frac{n^2}{d_1} + 4 \sum_{j=2}^{s} \frac{n^2}{d_j} + \frac{n^2}{d_{s+1}}. \end{aligned}$$

Note also that

$$\begin{aligned} & \|\tilde{b}_{-(S\cup\{1\})}\|_1 \\ = & \sum_{i=2}^{d_1} |\tilde{f}_i - \tilde{f}_{i-1}| + \sum_{i=d_1+2}^{d_2} |\tilde{f}_i - \tilde{f}_{i-1}| + \cdots + \sum_{i=d_1+\cdots+d_s+2}^{n} |\tilde{f}_i - \tilde{f}_{i-1}| \\ = & \ 0 \end{aligned}$$

It follows that

$$\begin{aligned} \frac{(s+1)\|X\tilde{b}\|_2^2/n}{(\|\tilde{b}_S\|_1 - \|\tilde{b}_{-(S\cup\{1\})}\|_1)^2} &= \frac{\sum_{i=1}^n \tilde{f}_i^2/n}{\left(\sum_{j=1}^s |\tilde{f}_{d_j+1} - \tilde{f}_{d_j}|\right)^2} \\ &= \frac{s+1}{\frac{n}{d_1} + \sum_{j=2}^s \frac{4n}{d_j} + \frac{n}{d_{s+1}}}. \end{aligned}$$

□

11.6. Proof of Theorem 17

To prove Theorem 17, we first establish the Lagrangian form of the minimization problem where we have the convex constraint $z_{S_0}^{*T}(\bar{v})b_{S_0} - \|Wb_{-S_0}\|_1 \geq 1$. Then we recall the projections and we introduce a subset $\mathcal{T}$ of the underlying probability space where the lower bound of Theorem 17 holds. The latter is shown in Lemma 36. Finally, we show that the subset $\mathcal{T}$ has large probability.

11.6.1. Lagrangian Form

Recall for $w \in \mathcal{W}(\bar{v})$ the convex problem with linear and convex constraints

$$b(w) \in \arg\min \left\{ \|Xb\|_2^2 : \ z_{S_0}^{*T}(\bar{v})b_{S_0} - \|Wb_{-S_0}\|_1 \geq 1 \right\}.$$

Note that here we do not require the positivity constraint $z_j^{*T}(\bar{v})b_j \geq 0$ for all $j \in S_0$. The next lemma gives its Lagrangian form. This form plays in the proof of Theorem 17 the same role as in the proof of Theorem 14 for the noiseless version. We also show that for $w \in \mathcal{W}(\bar{v})$ the minimum $\|Xb(w)\|_2^2$ is not larger than $\|Xb^*(\bar{v})\|_2^2$ (recall that by definition $\hat{\kappa}^2(1+\bar{v}, S_0) = s_0\|Xb^*(\bar{v})\|_2^2/n$).

Lemma 35 *We have*

$$X^T Xb(w) = \|Xb(w)\|_2^2 Wz(w),$$

with

$$z_{S_0}(w) = z_{S_0}^*(\bar{v}), \ z_{-S_0}(w) \in -\partial\|b_{-S_0}(w)\|_1.$$

Moreover, for $w \in \mathcal{W}(\bar{v})$

$$s_0\|Xb(w)\|_2^2/n \leq \hat{\kappa}^2(1+\bar{v}, S_0).$$

Proof of Lemma 35. The problem

$$\min\left\{ \|Xb\|_2^2 : \ z_{S_0}^{*T}(\bar v) b_{S_0} - \|W b_{-S_0}\|_1 \ge 1 \right\}$$

has Lagrangian

$$X^T X b(w) = \tilde\lambda W z(w)$$

with $z_{S_0}(w) = z_{S_0}^*(\bar v)$ and $z_{-S_0}(w) \in -\partial \|b_{-S_0}(w)\|_1$. Moreover

$$\|Xb(w)\|_2^2 = \tilde\lambda b(w)^T W z(w) = z_{S_0}^{*T}(\bar v) b_{S_0} - \|W b_{-S_0}\|_1 = 1$$

because the minimum is reached at the boundary. So

$$\tilde\lambda = \|Xb(w)\|_2^2 .$$

To obtain the second statement of the lemma, we use similar arguments as in the proof of Lemma 5. We have

$$\|Xb(w)\|_2 = \min_{b \in \mathbb{R}^p} \left\{ \frac{\|Xb\|_2}{z_{S_0}^{*T}(\bar v) b_{S_0} - \|W_{-S_0} b_{-S_0}\|_1} : \ z_{S_0}^{*T}(\bar v) b_{S_0} - \|W_{-S_0} b_{-S_0}\|_1 > 0 \right\}$$

But for $w \in \mathcal{W}$ and $\bar w := 1 + \bar v$, we know

$$\|W b_{-S_0}\|_1 \le \|\bar W b_{-S_0}\|_1$$

and so

$$z_{S_0}^{*T}(\bar v) b_{S_0} - \|W b_{-S_0}\|_1 > z_{S_0}^{*T}(\bar v) b_{S_0} - \|\bar W b_{-S_0}\|_1 .$$

Let

$$A := \left\{ b : \ z_{S_0}^{*T}(\bar v) b_{S_0} - \|W b_{-S_0}\|_1 > 0 \right\}$$

and

$$B := \left\{ b : \ z_{S_0}^{*T}(\bar v) b_{S_0} - \|\bar W b_{-S_0}\|_1 > 0 \right\}.$$

Then $B \subset A$. Hence

$$\begin{aligned}
\|Xb(w)\|_2 &= \min_{b \in A} \frac{\|Xb\|_2}{z_{S_0}^{*T}(\bar v) b_{S_0} - \|W b_{-S_0}\|_1} \\
&\le \min_{b \in B} \frac{\|Xb\|_2}{z_{S_0}^{*T}(\bar v) b_{S_0} - \|W b_{-S_0}\|_1} \\
&\le \min_{b \in B} \frac{\|Xb\|_2}{z_{S_0}^{*T}(\bar v) b_{S_0} - \|\bar W b_{-S_0}\|_1} \\
&= \|Xb^*(\bar v)\|_2 \\
&= \frac{\sqrt{n}\hat\kappa(1+\bar v, S_0)}{\sqrt{s_0}} .
\end{aligned}$$

□

11.6.2. Projections

Recall the notation of Subsection 7.2 and that moreover the diagonal elements of the matrix $(X_{S_0}^T X_{S_0})^{-1}$ are denoted by $\{u_j^2\}_{j\in S_0}$. We write

$$\hat{u}_{S_0} := (X_{S_0}^T X_{S_0})^{-1} X_{S_0}^T \epsilon.$$

We denote the projection of ϵ on the space spanned by the columns of X_{S_0} by

$$\epsilon \mathrm{P} X_{S_0} := X_{S_0}(X_{S_0}^T X_{S_0})^{-1} X_{S_0}^T \epsilon = X_{S_0}\hat{u}_{S_0}$$

and write

$$\mathbf{U}(S_0) := \|\epsilon \mathrm{P} X_{S_0}\|_2.$$

11.6.3. Choice of λ

Recall that we require that for some $t > 0$

$$\lambda > \|v_{-S_0}\|_\infty \sqrt{2(\log(2p) + t)}.$$

11.6.4. The Set $\mathcal{T}$

Recall

$$\bar{u}_j := u_j \sqrt{2(\log(2p)+t)}/\lambda,\ j \in S_0,\ \bar{v}_j := v_j\sqrt{2(\log(2p)+t)}/\lambda,\ j \notin S_0. \tag{21}$$

Let $\mathcal{T}$ be the set

$$\begin{aligned} \mathcal{T} \ &:= \ \Big\{ |\hat{u}_j| \le \lambda \bar{u}_j \ \forall j \in S_0 \Big\} \\ &\cap \ \Big\{ |\hat{v}_j| \le \lambda \bar{v}_j \ \forall j \notin S_0 \Big\} \cap \Big\{ \mathbf{U}(S_0) \le \sqrt{s_0} + \sqrt{2x} \Big\}. \end{aligned}$$

We show in Subsection 11.6.6 that $\mathbb{P}(\mathcal{T}) \ge 1 - \exp[-t] - \exp[-x]$.

11.6.5. Deterministic Part

The idea is now to incorporate the noisy part of the KKT conditions for the noisy Lasso into a weighted sub-differential, creating in that way KKT conditions of the same for as the noiseless KKT conditions (see (22) in the proof). To do so, we first put part of the noise in the vector β^0 without adding additional non-zeros. This makes it possible not to change the sub-differential at S_0. The rewriting of the KKT conditions make them resemble the Lagrangian form of Lemma 35.

We will use the KKT conditions (19) for $\hat{\beta}$:

$$-X^T(Y - X\hat{\beta}) = -\lambda\hat{\zeta},\ \hat{\zeta} \in \partial\|\hat{\beta}\|_1.$$

Lemma 36 *Suppose we are on the set $\mathcal{T}$ defined in Subsection 11.6.4. Then under the conditions of Theorem 17*

$$\|X(\hat\beta - \beta^0)\|_n \ge \frac{\lambda\sqrt{s_0}}{\sqrt{n}\hat\kappa(1+\bar v, S)} + \sqrt{2x}$$

Proof of Lemma 36. Set

$$\hat\beta^0_{S_0} := \beta^0 + \hat u_{S_0},\ \hat\beta^0_{-S_0} := 0.$$

Then

$$\begin{aligned} Y &= X\beta^0 + \epsilon \\ &= X_{S_0}\beta^0_{S_0} + X_{S_0}\hat u_{S_0} + \epsilon \mathrm{A} X_{S_0} \\ &= X\hat\beta^0 + \epsilon \mathrm{A} X_{S_0}. \end{aligned}$$

The KKT conditions (19) are

$$-X^T(Y - X\hat\beta) = -\lambda\hat\zeta.$$

We have

$$Y - X\hat\beta = -X(\hat\beta - \hat\beta^0) - \epsilon\mathrm{A}X_{S_0}.$$

Therefore

$$-X^T(Y - X\hat\beta) = X^TX(\hat\beta - \hat\beta^0) - X^T(\epsilon\mathrm{A}X_{S_0}).$$

But

$$X^T_{S_0}(\epsilon\mathrm{A}X_{S_0}) = 0,$$

and

$$\begin{aligned} X^T_{-S_0}(\epsilon\mathrm{A}X_{S_0}) &= X^T_{-S_0} - X^T_{-S_0}X_{S_0}(X^T_{S_0}X_{S_0})^{-1}X^T_{S_0}\epsilon \\ &= (X_{-S_0}\mathrm{A}X_{S_0})^T\epsilon. \end{aligned}$$

Hence the KKT conditions read

$$X^TX(\hat\beta - \hat\beta^0) = -\lambda\hat\zeta + \hat v,$$

where

$$\hat v_{S_0} = 0,\ \hat v_{-S_0} = (X_{-S_0}\mathrm{A}X_{S_0})^T\epsilon.$$

Set $\hat S := \{j :\ \hat\beta_j \ne 0\}$ and define for all $j \in \hat S \backslash S_0$

$$\hat w_j := 1 + \hat v_j/(\lambda\hat\zeta_j).$$

By assumption (since we are on $\mathcal{T}$) $|\hat v_j| < \lambda\bar v_j$. so $\hat w_j \ge 1 - \bar v_j$ for all $j \in \hat S\backslash S_0$. For $j \notin \hat S \cup S_0$ we define

$$\hat w_j := \max\{|1 + \hat v_j/\lambda|, 1 - \bar v_j\}.$$

Then for $j \notin \hat{S} \cup S_0$

$$\begin{aligned}\lambda\hat{\zeta}_j + \hat{v}_j &= \lambda|\hat{\zeta}_j + \hat{v}_j/\lambda|\text{sign}(\hat{\zeta}_j + \hat{v}_j/\lambda)\\ &= \begin{cases}\hat{w}_j\text{sign}(\hat{\zeta}_j + \hat{v}_j/\lambda), & |\hat{\zeta}_j + \hat{v}_j/\lambda| \geq 1 - \bar{v}_j\\ \hat{w}_j\frac{|\hat{\zeta}_j + \hat{v}_j/\lambda|}{1-\bar{v}_j}\text{sign}(\hat{\zeta}_j + \hat{v}_j/\lambda) & |\hat{\zeta}_j + \hat{v}_j/\lambda| \leq 1 - \bar{v}_j\end{cases}\\ &= \hat{w}_j\tilde{\zeta}_j,\end{aligned}$$

where

$$\tilde{\zeta}_j := \begin{cases}\text{sign}(\hat{\zeta}_j + \hat{v}_j/\lambda), & |\hat{\zeta}_j + \hat{v}_j/\lambda| \geq 1 - \bar{v}_j\\ \frac{|\hat{\zeta}_j + \hat{v}_j/\lambda|}{1-\bar{v}_j}\text{sign}(\hat{\zeta}_j + \hat{v}_j/\lambda) & |\hat{\zeta}_j + \hat{v}_j/\lambda| \leq 1 - \bar{v}_j\end{cases}.$$

One readily verifies that (on $\mathcal{T}$) $\hat{w}_j \leq 1 + \bar{v}_j$ for all $j \notin S_0$. Taking $\tilde{\zeta}_j = \hat{\zeta}_j$ for $j \in S \cup S_0$ we arrive at the KKT conditions

$$X^T X(\hat{\beta} - \hat{\beta}^0) = -\lambda\hat{W}\tilde{\zeta},\ \tilde{\zeta} \in \partial\|\hat{\beta}\|_1 \tag{22}$$

and where $\hat{W} = \text{diag}(\hat{w})$ with $\hat{w} \in \mathcal{W}(\bar{v})$. Let now $S_0^+ := \{j \in S_0 :\ z_j^*(\bar{v})b_j(\hat{w}) > 0\}$ and $S_0^- := \{j \in S_0 :\ z_j^*(\bar{v})b_j(\hat{w}) \leq 0\}$. Take

$$\beta' = \hat{\beta}^0 - \lambda b_j(\hat{w})/\|Xb(\hat{w})\|_2^2.$$

Case 1 Let $j \in S_0$. By the condition on β^0 we know that $|\beta_j^0| > \lambda|b_j(\hat{w})|/\|Xb(\hat{w})\|_2^2 + |\hat{u}_{S_0}|$, so $|\hat{\beta}_j^0| \geq |\beta_j^0| - |\hat{u}_{S_0}| > \lambda|b_j(\hat{w})|/\|Xb(\hat{w})\|_2^2$. If $z_j^*(\bar{v}) = 1$ and $b_j(\hat{w}) > 0$, then $\hat{\beta}_j^0 > 0$ and

$$\beta_j' = |\hat{\beta}_j^0| - \lambda|b_j(\hat{w})|/\|Xb(\hat{w})\|_2^2 > 0.$$

If $z_j^*(\bar{v}) = 1$ and $b_j(\hat{w}) \leq 0$, then $\hat{\beta}_j^0 > 0$ and we have

$$\beta_j' = |\hat{\beta}_j^0| + \lambda|b_j(\hat{w})|/\|Xb(\hat{w})\|_2^2 > 0.$$

If $z_j^*(\bar{v}) = -1$ and $b_j(\hat{w}) < 0$, then $\hat{\beta}_j^0 < 0$ and

$$\beta_j' = -|\hat{\beta}_j^0| + \lambda|b_j(\hat{w})|/\|Xb(\hat{w})\|_2^2 < 0.$$

If $z_j^*(\bar{v}) = -1$ and $b_j(\hat{w}) \geq 0$, then $\hat{\beta}_j^0 < 0$ and

$$\beta_j' = -|\hat{\beta}_j^0| - \lambda|b_j(\hat{w})|/\|Xb(\hat{w})\|_2^2 < 0.$$

Case 2 Let now $j \notin S_0$. Then

$$\beta_j' = -\lambda b_j(\hat{w})/\|Xb(\hat{w})\|_2^2,$$

so

$$z_j(\hat{w})\beta_j' = -\lambda z_j(\hat{w})b_j(\hat{w})/\|Xb(\hat{w})\|_2^2 > 0.$$

Thus

$$z(\hat{w}) \in \partial \|\beta'\|_1.$$

Furthermore, by the first part of Lemma 35,

$$X^T X(\beta' - \hat{\beta}^0) = -\lambda X^T X b(\hat{w}) / \|X b(\hat{w})\|_2 = -\lambda \hat{W} z(\hat{w}).$$

So $\beta' =: \hat{\beta}$ satisfies the KKT conditions with $\tilde{\zeta} = z(\hat{w})$. We further have

$$\begin{aligned} \|X(\hat{\beta} - \hat{\beta}^0)\|_2^2 &= \lambda^2 b^T(\hat{w}) \hat{W} z(\hat{w}) / \|X b(\hat{w})\|_2^2 \\ &= \lambda^2 / \|X b(\hat{w})\|^2 \\ &\geq \lambda^2 s_0 / (n \hat{\kappa}^2 (1 + \bar{v}, S_0)) \end{aligned}$$

where in the last step we used the second part of Lemma 35. Finally, by the triangle inequality

$$\begin{aligned} \|X(\hat{\beta} - \beta^0)\|_2 &\geq \|X(\hat{\beta} - \hat{\beta}^0)\|_2 - \mathbf{U}(S_0) \\ &\geq \frac{\lambda \sqrt{s_0}}{\sqrt{n} \hat{\kappa}(1 + \bar{v}, S_0)} - \mathbf{U}(S_0) \\ &\geq \frac{\lambda \sqrt{s_0}}{\sqrt{n} \hat{\kappa}(1 + \bar{v}, S_0)} - \sqrt{s_0} - \sqrt{2x}. \end{aligned}$$

□

11.6.6. Random Part

In Lemma 36, we showed that the conclusion (12) of Theorem 17 holds on the set $\mathcal{T}$. This subsection obtains that $\mathbb{P}(\mathcal{T}) \geq 1 - \exp[-t] + \exp[-x]$.

Lemma 37 *It holds that*

$$\mathbb{P}(\mathcal{T}) \geq 1 - \exp[-t] - \exp[-x].$$

Proof of Lemma 37. Apply Lemma 41 with $Z_j = \hat{u}_j / u_j$ for $j \in S_0$ and $Z_j = \hat{v}_j / v_j$ for $j \notin S_0$ to find that with probability at least $1 - \exp[-t]$

$$|\hat{u}_j| \leq \lambda \bar{u}_j \ \forall j \in S_0, \ |\hat{v}_j| \leq \lambda \bar{v}_j \ \forall j \notin S_0.$$

Furthermore, the random variable $\mathbf{U}^2(S_0)$ has a chi-squared distribution with s_0 degrees of freedom. Lemma 42 gives that with probability at least $1 - \exp[-x]$,

$$\mathbf{U}(S_0) \leq \sqrt{s_0} + \sqrt{2x}.$$

□

11.6.7. Collecting the Pieces

Combining Lemma 36 with Lemma 37 completes the proof of Theorem 17.

11.7. Proof of Theorem 18

The proof is along the lines of Theorem 9.

11.7.1. Comparing the KKT Conditions

We compare the KKT conditions for the noisy Lasso with those for the noiseless Lasso.

Lemma 38 *It holds that*

$$\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda\|\hat{\beta}\|_1 - \lambda^*\hat{\beta}^T z^* \le (\hat{\beta}-\beta^*)^T X^T \epsilon + (\lambda-\lambda^*)\|\beta^*\|_1.$$

Proof of Lemma 38. The KKT conditions (19) for $\hat{\beta}$ can be written as

$$X^T X(\hat{\beta}-\beta^0) + \lambda\hat{\zeta} = X^T\epsilon.$$

where $\hat{\zeta} \in \partial\|\hat{\beta}\|_1$. By the KKT conditions (6) for β^*

$$X^T X(\beta^*-\beta^0) + \lambda^*\zeta^* = 0.$$

Hence, taking the difference

$$X^T X(\hat{\beta}-\beta^*) + \lambda\hat{\zeta} - \lambda^*\zeta^* = X^T\epsilon.$$

Multiply by $(\hat{\beta}-\beta^*)^T$ to find

$$\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda(\hat{\beta}-\beta^*)^T\hat{\zeta} - \lambda^*(\hat{\beta}-\beta^*)^T\zeta^* = (\hat{\beta}-\beta^*)^T X^T\epsilon.$$

But

$$\begin{aligned} & \lambda(\hat{\beta}-\beta^*)^T\hat{\zeta} - \lambda^*(\hat{\beta}-\beta^*)^T\zeta^* \\ = \; & \lambda\|\hat{\beta}\|_1 - \lambda^*\hat{\beta}^T\zeta^* + \lambda^*\|\beta^*\|_1 - \lambda\beta^{*T}\hat{\zeta} \\ = \; & \lambda\|\hat{\beta}\|_1 - \lambda^*\hat{\beta}^T\zeta^* + \lambda\|\beta^*\|_1 - \lambda\beta^{*T}\hat{\zeta} - (\lambda-\lambda^*)\|\beta^*\|_1 \\ \ge \; & \lambda\|\hat{\beta}\|_1 - \lambda^*\hat{\beta}^T\zeta^* - (\lambda-\lambda^*)\|\beta^*\|_1 \end{aligned}$$

where we used that

$$\|\beta^*\|_1 - \beta^{*T}\hat{\zeta} \ge 0.$$

Therefore

$$\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda\|\hat{\beta}\|_1 - \lambda^*\hat{\beta}^T z^* \le (\hat{\beta}-\beta^*)^T X^T \epsilon + (\lambda-\lambda^*)\|\beta^*\|_1.$$

□

11.7.2. Projections

Recall the notation of Subsection 8.1. We let moreover $\hat{v}_{-S_0}$ be the vector

$$\hat{v}_{-S}^S := (X_{-S} \mathrm{A} X_S)^T\epsilon.$$

As before, we denote the projection of ϵ on the space spanned by the columns of X_S by $\epsilon \mathrm{P} X_S$ and write

$$\mathbf{U}(S) := \|\epsilon \mathrm{P} X_S\|_2.$$

11.7.3. Choice of λ

Recall that we require that for some $t > 0$

$$\lambda > \|v_{-S}^S\|_\infty \sqrt{2(\log(2p) + t)}.$$

11.7.4. The Set $\mathcal{T}^S$

Recall

$$\bar{v}^S := v_j^S \sqrt{2(\log(2p) + t)}/\lambda, \ j \notin S.$$

Let

$$\mathcal{T}^S := \{|\hat{v}_j| \le \lambda \bar{v}_j \ \forall \ j \notin S\} \cap \{\mathbf{U}(S) \le \sqrt{s} + \sqrt{2x}\}.$$

11.7.5. Deterministic Part

Lemma 39 *On the set $\mathcal{T}^S$ it holds that*

$$\|X(\hat{\beta} - \beta^*)\|_2 \le \sqrt{s} + \sqrt{2x} + (\lambda - \lambda^*)\sqrt{s/n}/\hat{\kappa}(\bar{w}^S, S).$$

Proof of Lemma 39. Since $S_* \subset S$

$$X(\hat{\beta} - \beta^*) = X_S \hat{b}_S + X_{-S} \mathrm{A} X_S \hat{\beta}_{-S}$$

where

$$X_S \hat{b}_S = X_S(\hat{\beta}_S - \beta_S^*) + (X_{-S} \mathrm{P} X_S)\hat{\beta}_{-S}.$$

In view of Lemma 38,

$$\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda \|\hat{\beta}\|_1 - \lambda^* \beta^T z^*$$

$$\le \hat{b}_S^T X_S^T \epsilon + \left[X_{-S} \mathrm{A} X_S \hat{\beta}_S \right]^T \epsilon + (\lambda - \lambda^*) \|\beta^*\|_1$$

By the Cauchy-Schwarz inequality and since we are on $\mathcal{T}^S$

$$\hat{b}_S^T X_S^T \epsilon \le \mathbf{U}(S) \|X \hat{b}_S\|_2 \le (\sqrt{s} + \sqrt{2x}) \|X \hat{b}_S\|_2 \le (\sqrt{s} + \sqrt{2x}) \|X(\hat{\beta} - \beta^*)\|_2$$

where in the last inequality we used Pythagoras rule. Moreover, by the definition of $\hat{v}_{-S}^S$ and since we are on the set $\mathcal{T}^S$

$$\left[X_{-S} \mathrm{A} X_S \hat{\beta}_{-S} \right]^T \epsilon = \hat{\beta}_{-S}^T \hat{v}_{-S}^S \le \lambda \sum_{j \notin S} \bar{v}_{-S}^S |\hat{\beta}_j|.$$

On the other hand,

$$\lambda \|\hat{\beta}_{-S}\|_1 - \lambda^* \zeta_{-S}^{*T} \hat{\beta}_{-S} \ge \lambda \sum_{j \notin S} (1 - \lambda^* |\zeta_j^*|/\lambda) |\hat{\beta}_j|$$

and

$$(\lambda - \lambda^*)\|\beta^*\|_1 - \lambda\|\hat{\beta}_S\|_1 + \lambda^* z^{*T}\hat{\beta}_S \leq (\lambda - \lambda^*)\|\hat{\beta}_S - \beta^*_S\|_1.$$

If $\|X(\hat{\beta}-\beta^*)\|_2 \leq \sqrt{s}+\sqrt{2x}$ we are done. Suppose therefore that $\|X(\hat{\beta}-\beta^*\|_2 > \sqrt{s}+\sqrt{2x}$. Then we see that

$$\begin{aligned} & \|X(\hat{\beta} - \beta^*)\|_2^2 - (\sqrt{s} + \sqrt{2x})\|X(\hat{\beta} - \beta^*)\|_2 \\ = \ & \|X(\hat{\beta} - \beta^*)\|_2 \biggl(\|X(\hat{\beta} - \beta^*)\|_2 - \sqrt{s} - \sqrt{2x} \biggr) \\ > \ & 0. \end{aligned}$$

But then

$$\lambda \sum_{j \notin S} (1 - \bar{v}_j^S - \lambda^* |\zeta_j^*|/\lambda)|\hat{\beta}_j| < (\lambda - \lambda^*)\|\hat{\beta}_S - \beta^*\|_1.$$

or

$$\|\hat{\beta}_S - \beta^*_S\|_1 - \|\bar{W}^S \hat{\beta}_{-S}\|_1 > 0.$$

Then

$$\|\hat{\beta}_S - \beta^*_S\|_1 - \|\bar{W}^S \hat{\beta}_{-S}\|_1 \leq (\sqrt{s/n})\|X(\hat{\beta} - \beta^*)\|_2 / \hat{\kappa}(\bar{w}^S, S).$$

We thus arrive at

$$\begin{aligned} & \|X(\hat{\beta} - \beta^*)\|_2^2 \\ \leq \ & \biggl(\sqrt{s} + \sqrt{2x} + (\lambda - \lambda^*)\sqrt{s/n}/\hat{\kappa}(\bar{w}^S, S) \biggr) \|X(\hat{\beta} - \beta^*)\|_2 \end{aligned}$$

or

$$\|X(\hat{\beta} - \beta^*)\|_2 \leq \sqrt{s} + \sqrt{2x} + (\lambda - \lambda^*)\sqrt{s/n}/\hat{\kappa}(\bar{w}^S, S).$$

□

11.7.6. Random Part

Lemma 40 *We have*

$$\mathbb{P}(\mathcal{T}^S) \geq 1 - \exp[-t] - \exp[-x].$$

Proof of Lemma 40. This follows from Lemma 41 and Lemma 42. □

11.7.7. Finalizing the Proof of Theorem 18

Combine Lemma 39 with Lemma 40.

11.8. Proof of the Lemma in Section 9

Proof of Lemma 21. Write $g_i := w_i f_i$, $i = 1, \ldots, n$ and $u_j := d_j/2$, $j = 2, \ldots, s$. Then we have

$$\sum_{j=1}^{s} |g_{d_j+1} - g_{d_j}| - \sum_{i=2}^{d_1} |g_i - g_{i-1}| - \sum_{j=2}^{s-1} \sum_{i=d_j+1}^{d_{j+1}} |g_i - g_{i-1}| - \sum_{i=d_s+1}^{n} |g_i - g_{i-1}|$$

$$\begin{aligned}
\leq\ & \frac{1}{d_1} \sum_{i=1}^{d_1} |g_i| + \frac{1}{u_2} \sum_{i=d_1+1}^{d_1+u_2} |g_i| \\
+\ & \frac{1}{d_2 - u_2} \sum_{i=d_1+u_2+1}^{d_1+d_2} |g_i| + \frac{1}{u_3} \sum_{i=d_1+d_2+1}^{d_1+d_2+u_3} |g_i| \\
& \cdots \\
+\ & \frac{1}{d_{s-1} - u_{s-1}} \sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+1}^{d_1+\cdots+d_{s-1}} |g_i| + \frac{1}{u_s} \sum_{i=d_1+\cdots+d_{s-1}+1}^{d_1+\cdots+d_{s-1}+u_s} |g_i| \\
+\ & \frac{1}{d_s - u_s} \sum_{i=d_1+\cdots+d_{s-1}+u_s+1}^{d_1+\cdots+d_s} |g_i| + \frac{1}{d_{s+1}} \sum_{i=d_1+\cdots+d_s+1}^{n} |g_i|
\end{aligned}$$

$$\begin{aligned}
\leq\ & \Bigg(\frac{1}{d_1^2} \sum_{i=1}^{d_1} w_i^2 + \frac{1}{u_2^2} \sum_{i=d_1+1}^{d_1+u_2} w_i^2 \\
+\ & \frac{1}{(d_2 - u_2)^2} \sum_{i=d_1+u_2+1}^{d_1+d_2} w_i^2 + \frac{1}{u_3^2} \sum_{i=d_1+d_2+1}^{d_1+d_2+u_3} w_i^2 \\
& \cdots \\
+\ & \frac{1}{(d_{s-1} - u_{s-1})^2} \sum_{i=d_1+\cdots+d_{s-2}+u_{s-1}+1}^{d_1+\cdots+d_{s-1}} w_i^2 + \frac{1}{u_s^2} \sum_{i=d_1+\cdots+d_{s-1}+1}^{d_1+\cdots+d_{s-1}+u_s} w_i^2 \\
+\ & \frac{1}{(d_s - u_s)^2} \sum_{i=d_1+\cdots+d_{s-1}+u_s+1}^{d_1+\cdots+d_s} w_i^2 + \frac{1}{d_{s+1}^2} \sum_{i=d_1+\cdots+d_s+1}^{n} w_i^2 \Bigg)^{1/2} \\
& \times \left(\sum_{i=1}^{n} f_i^2 \right)^{1/2}
\end{aligned}$$

$$\begin{aligned}
\leq\ & \sqrt{\frac{n}{d_1} + \frac{n}{u_2} + \frac{n}{d_2 - u_2} + \cdots + \frac{n}{d_{s-1} - u_{s-1}} + \frac{n}{u_s} + \frac{n}{d_s - u_s} + \frac{n}{d_{s+1}}} \\
& \times \sqrt{\sum_{i=1}^{n} |f_i|^2/n} \\
& \times \|w\|_\infty .
\end{aligned}$$

Moreover

$$\begin{aligned}
& \sum_{j=1}^{s} w_{d_j+1}|f_{d_j+1} - f_{d_j}| - \sum_{i=2}^{d_1} w_i|f_i - f_{i-1}| \\
- & \sum_{j=2}^{s-1} \sum_{i=d_j+1}^{d_{j+1}} w_i|f_i - f_{i-1}| - \sum_{i=d_s+1}^{n} w_i|f_i - f_{i-1}| \\
\le & \sum_{j=1}^{s} |g_{d_j+1} - g_{d_j}| - \sum_{i=2}^{d_1} |g_i - g_{i-1}| - \sum_{j=2}^{s-1} \sum_{i=d_j+1}^{d_{j+1}} |g_i - g_{i-1}| - \sum_{i=d_s+1}^{n} |g_i - g_{i-1}| \\
+ & \sum_{i=2}^{n} |w_i - w_{i-1}||f_{i-1}|,
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i=2}^{n} |w_i - w_{i-1}||f_{i-1}| &\le \sqrt{\sum_{i=2}^{n} (w_i - w_{i-1})^2} \sqrt{\sum_{i=2}^{n} f_{i-1}^2} \\
&\le \sqrt{\sum_{i=2}^{n} (w_i - w_{i-1})^2} \sqrt{\sum_{i=1}^{n} f_{i}^2}
\end{aligned}$$

Thus we conclude

$$\begin{aligned}
& \sum_{j=1}^{s} w_{d_j+1}|f_{d_j+1} - f_{d_j}| \\
- & \sum_{i=2}^{d_1} w_i|f_i - f_{i-1}| - \sum_{j=2}^{s-1} \sum_{i=d_j+1}^{d_{j+1}} w_i|f_i - f_{i-1}| - \sum_{i=d_s+1}^{n} w_i|f_i - f_{i-1}| \\
\le & \left(\|w\|_\infty \sqrt{\frac{n}{d_1} + \sum_{j=2}^{s} \frac{4n}{d_j} + \frac{n}{d_{s+1}}} + \sqrt{n \sum_{i=2}^{n} (w_i - w_{i-1})^2} \right) \sqrt{\sum_{i=1}^{n} f_i^2/n}.
\end{aligned}$$

□

11.9. Proof of Theorem 1

This follows from Corollary 12 combined with Theorem 14, where in the latter we replace $\hat{\Sigma} := X^T X/n$ by the population version Σ_0. This works because we replaced condition (8) by its population counterpart condition (3).

12. Tools from Probability Theory

We first present three standard lemmas for Gaussian random variables, Lemmas 41, 42 and 43. These three lemmas are followed by a concentration of measure result and a result for Gaussian quadratic forms.

Lemma 41 *Let $Z_1, \ldots , Z_p$ be standard normal random variables. Then it holds for all $t > 0$ that*

$$\mathbb{P}\left(\max_{1 \le j \le p} |Z_j| \ge \sqrt{2(\log(2p) + t)} \right) \le \exp[-t].$$

Proof of Lemma 41. For each $t > 0$

$$\mathbb{P}(|Z_1| \ge \sqrt{2t}) \le 2 \exp[-t].$$

So by the union bound, for any $t > 0$,

$$\begin{aligned} \mathbb{P}\left(\max_{1 \le j \le p} |Z_j| > \sqrt{2(\log(2p) + t)} \right) &\le p\mathbb{P}(|Z_1| \ge \sqrt{2(\log(2p) + t)}) \\ &\le 2p \exp[-(\log(2p + t)] = \exp[-t]. \end{aligned}$$

□

Lemma 42 *Let $Z := (Z_1, \ldots , Z_T)^T$ be a vector with i.i.d. standard Gaussian entries. Then it holds for all $x > 0$ that*

$$\mathbb{P}\left(\|Z\|_2 \ge \sqrt{T} + \sqrt{2x} \right) \le \exp[-x]$$

and

$$\mathbb{P}\left(|\|Z\|_2 - \sqrt{T}| \ge \sqrt{2x} \right) \le 2 \exp[-x].$$

Proof of Lemma 42. This follows from concentration of measure (Borell, 1975, Giné and Nickl, 2015, Theorem 2.5.7) because the map $Z \mapsto \|Z\|_2$ is Lipschitz. Alternatively, one may apply Lemma 1 in Laurent and Massart (2000). □

Lemma 43 *Let $(U, V) \in \mathbb{R}^{n \times 2}$ have i.i.d Gaussian rows with mean zero and covariance matrix*

$$\begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}.$$

Then for all $t > 0$, with probability at least $1 - 4 \exp[-t]$

$$|U^T V - n\sigma_{uv}| \le 3\sigma_u\sigma_v \left(\sqrt{2nt} + t \right).$$

Proof of Lemma 43. By standard arguments (see van de Geer (2017) for tracking down some constants) one can derive that with probability at least $1 - 4 \exp[-t]$

$$|U^T V - n\sigma_{uv}| \le (\sigma_u\sigma_v + 2|\sigma_{u,v}|)\sqrt{2nt} + (\sigma_u\sigma_v + 2|\sigma_{u,v}|)t.$$

We simplify this to: with probability at least $1 - 4\exp[-t]$

$$|U^T V - n\sigma_{uv}| \le 3\sigma_u\sigma_v\biggl(\sqrt{2nt} + t\biggr).$$

□

This is the concentration of measure lemma that we use in Section 4.

Lemma 44 *For any $b \in \mathbb{R}^p$ and all $x > 0$, we have*

$$\mathbb{P}\biggl(\|X(\hat\beta - b)\|_2 \ge m_b + \sqrt{2x}\biggr) \le \exp[-x]$$

and

$$\mathbb{P}\biggl(\biggl|\|X(\hat\beta - b)\|_2 - m_b\biggr| \ge \sqrt{2x}\biggr) \le 2\exp[-x]$$

where $m_b := \mathbb{E}(\|X(\hat\beta - b)\|_2 | X)$.

Proof of Lemma 44. This follows from concentration of measure see e.g. Borell (1975), or Giné and Nickl (2015), Theorem 2.5.7, as the map $\epsilon \mapsto \|X(\hat\beta - b)\|^2$ is Lipschitz, see also van de Geer and Wainwright (2017). □

Finally, we give a result for Gaussian quadratic forms.

Lemma 45 *Let X have i.i.d. $\mathcal{N}(0, \Sigma_0)$-distributed rows and let M be a (sequence of) constant(s) such that*

$$M^2 = o\biggl(n/(\|\Sigma_0\|_\infty \log(2p))\biggr).$$

Then, for a suitable sequence $\eta_M = o(1)$, with probability tending to one

$$\inf_{\|b\|_1 \le M \|\Sigma_0^{1/2} b\|_2} \frac{\|Xb\|_2^2/n}{\|\Sigma_0^{1/2} b\|_2^2} \ge (1 - \eta_M)^2.$$

Proof of Lemma 45. See for example Chapter 16 in van de Geer (2016) and its references, or van de Geer and Muro (2014). □

Acknowledgments

We thank Rico Zenklusen from the Institute of Operations Research, ETH Zürich, and Hamza Fawzi from the Department of Applied Mathematics and Theoretical Physics at the University of Cambridge, for very helpful discussions. Research supported by Isaac Newton Institute for Mathematical Sciences, program *Statistical Scalability*, EPSRC Grant Number LNAG/036 RG91310.

References

P. C. Bellec. Optimistic lower bounds for convex regularized least-squares. *arXiv preprint arXiv:1703.01332*, 2017.

V. Belloni, A.and Chernozhukov and L. Wang. Pivotal estimation via square-root Lasso in nonparametric regression. *Annals of Statistics*, 42(2):757–788, 2014.

C. Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae*, 30 (2):207–216, 1975.

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

D.L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Models*. Cambridge University Press, 2015.

C. Giraud. *Introduction to High-Dimensional Statistics*, volume 138. CRC Press, 2014.

V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

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

R. Tibshirani. Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

S. van de Geer. *Estimation and Testing Under Sparsity: École d'Eté de Probabilités de Saint Flour XLV-2016*. Springer Science & Business Media, 2016.

S. van de Geer. On the efficiency of the de-biased Lasso, 2017. arXiv:1708.07986.

S. van de Geer and A. Muro. On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electronic Journal of Statistics*, 8:3031–3061, 2014.

S. van de Geer and M. Wainwright. On concentration for (regularized) empircal risk minimization. *Sankhyā*, 79-A:159–200, 2017.

S.A. van de Geer. The deterministic Lasso. In *JSM proceedings, 2007, 140*. American Statistical Association, 2007.

Y. Zhang, M. Wainwright, and M. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *COLT*, pages 921–948, 2014.