Electronic Journal of Statistics
Vol. 5 (2011) 688–749
ISSN: 1935-7524
DOI: 10.1214/11-EJS624

The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso)

Sara van de Geer[*] **and Peter Bühlmann**

Seminar for Statistics
ETH Zürich
e-mail: geer@stat.math.ethz.ch; buhlmann@stat.math.ethz.ch

Shuheng Zhou[*]

Department of Statistics
University of Michigan
e-mail: shuhengz@umich.edu

Abstract: We revisit the adaptive Lasso as well as the thresholded Lasso with refitting, in a high-dimensional linear model, and study prediction error, ℓ_q-error ($q \in \{1, 2\}$), and number of false positive selections. Our theoretical results for the two methods are, at a rather fine scale, comparable. The differences only show up in terms of the (minimal) restricted and sparse eigenvalues, favoring thresholding over the adaptive Lasso. As regards prediction and estimation, the difference is virtually negligible, but our bound for the number of false positives is larger for the adaptive Lasso than for thresholding. We also study the adaptive Lasso under beta-min conditions, which are conditions on the size of the coefficients. We show that for exact variable selection, the adaptive Lasso generally needs more severe beta-min conditions than thresholding. Both the two-stage methods add value to the one-stage Lasso in the sense that, under appropriate restricted and sparse eigenvalue conditions, they have similar prediction and estimation error as the one-stage Lasso but substantially less false positives. Regarding the latter, we provide a lower bound for the Lasso with respect to false positive selections.

AMS 2000 subject classifications: Primary 62J07; secondary 62G08.
Keywords and phrases: Adaptive Lasso, estimation, prediction, restricted eigenvalue, thresholding, variable selection.

Received July 2010.

Contents

[*] Research supported by SNF 20PA21-120050/1.

1. Introduction

Consider the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\beta \in \mathbb{R}^p$ is a vector of coefficients, $\mathbf{X}$ is an $(n \times p)$-design matrix, and $\mathbf{Y}$ is an n-vector of noisy observations, ϵ being the noise term. We examine the case $p \geq n$, i.e., a high-dimensional situation. The design matrix $\mathbf{X}$ is treated as fixed, and the Gram matrix is denoted by $\hat{\Sigma} := \mathbf{X}^T\mathbf{X}/n$. Throughout, we assume the normalization $\hat{\Sigma}_{j,j} = 1$ for all $j \in \{1, \ldots, p\}$.

This paper presents a theoretical comparison between the thresholded Lasso with refitting and the adaptive Lasso. Our analysis is motivated by the fact that both methods are very popular in practical applications for reducing the number of active variables. Our theoretical study shows that under suitable conditions both methods reduce the number of false positives while maintaining a good prediction error.

We emphasize here and describe later that we allow for model misspecification where the true regression function may be non-linear in the covariates. For such cases, we can consider the projection onto the linear span of the covariates. The (projected or true) linear model does not need to be sparse nor do we assume so-called beta-min conditions, requiring that the non-zero regression coefficients (from a sparse approximation) are "sufficiently large". We will show in Lemma 3.3 how beta-min conditions can be invoked to improve the result. Furthermore, we also do not require the stringent irrepresentable conditions or incoherence assumptions on the design matrix $\mathbf{X}$ but only some weaker restricted or sparse eigenvalue conditions.

Regularized estimation with the ℓ_1-norm penalty, also known as the Lasso ([25]), refers to the following convex optimization problem:

$$\hat{\beta} := \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1 \right\}, \tag{1.1}$$

where $\lambda > 0$ is a penalization parameter.

Regularization with ℓ_1-penalization in high-dimensional scenarios has become extremely popular. The methods are easy to use, due to recent progress in specifically tailored convex optimization ([21], [15]).

A two-stage version of the Lasso is the so-called adaptive Lasso

$$\hat{\beta}_{\text{adap}} := \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda_{\text{init}}\lambda_{\text{adap}} \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{j,\text{init}}|} \right\}. \tag{1.2}$$

Here, $\hat{\beta}_{\text{init}}$ is the one-stage Lasso defined in (1.1), with initial tuning parameter $\lambda = \lambda_{\text{init}}$, and $\lambda_{\text{adap}} > 0$ is the tuning parameter for the second stage. Note that when $|\hat{\beta}_{j,\mathrm{init}}| = 0$, we exclude variable j in the second stage. The adaptive Lasso was originally proposed by [39].

Another possibility is the thresholded Lasso with refitting. Define

$$\hat{S}_{\mathrm{thres}} = \{j : |\hat{\beta}_{j,\mathrm{init}}| > \lambda_{\mathrm{thres}}\}, \tag{1.3}$$

which is the set of variables having estimated coefficients larger than some given threshold λ_{thres}. The refitting is then done by ordinary least squares:

$$\hat{b}_{\mathrm{thres}} = \arg\min_{\beta_{\hat{S}_{\mathrm{thres}}}} \|\mathbf{Y} - \mathbf{X}\beta_{\hat{S}_{\mathrm{thres}}}\|_2^2/n,$$

where, for a set $S \subset \{1, \ldots, p\}$, β_S has coefficients different from zero at the components in S only.

We will present bounds for the prediction error, its ℓ_q-error ($q \in \{1, 2\}$), and the number of false positives. The bounds for the two methods are qualitatively the same. A difference is that our variable selection properties results for the adaptive Lasso depend on its prediction error, whereas for the thresholded Lasso, variable selection can be studied without reference to its prediction error. In our analysis this leads to a bound for the number of false positives of the thresholded Lasso that is smaller than the one for the adaptive Lasso, when restricted or sparse minimal eigenvalues are small and/or sparse maximal eigenvalues are large.

Of course, such comparisons depend on how the tuning parameters are chosen. Choosing these by cross validation is in our view the most appropriate, but it is beyond the scope of this paper to present a mathematically rigorous theory for the cross validation scheme for the adaptive and/or thresholded Lasso (see [1] for a recent survey on cross validation).

1.1. Related work

Consistency results for the prediction error of the Lasso can be found in [16]. The prediction error is asymptotically oracle optimal under certain conditions on the design matrix $\mathbf{X}$, see e.g. [6–8], [26], [4], [18, 19], where also estimation in terms of the ℓ_1- or ℓ_2-loss is considered. The "restricted eigenvalue condition" of [4] (see also [18, 19]) plays a key role here. Restricted eigenvalue conditions are implied by, but generally much weaker than, "incoherence" conditions, which exclude high correlations between co-variables. Also [9] allow for a major relaxation of incoherence conditions, using assumptions on the set of true coefficients.

There is however a bias problem with ℓ_1-penalization, due to the shrinking of the estimates which correspond to true signal variables. A discussion can be found in [39], and [22]. Moreover, for consistent variable selection with the Lasso, it is known that the so-called "neighborhood stability condition" ([23]) for the design matrix, which has been re-formulated in a nicer form as the "irrepresentable condition" ([36]), is sufficient and essentially necessary. The papers [30, 31] analyze the smallest sample size needed to recover a sparse signal under certain incoherence conditions. Because irrepresentable or incoherence conditions are restrictive and much stronger than restricted eigenvalue conditions (see [29] for a comparison), we conclude that the Lasso for exact variable selection only works in a rather narrow range of problems, excluding for example some cases where the design exhibits strong (empirical) correlations.

Regularization with the ℓ_q-“norm” with $q < 1$ will mitigate some of the bias problems, see [33]. Related are multi-step procedures where each of the steps involves a convex optimization only. A prime example is the adaptive Lasso which is a two-step algorithm and whose repeated application corresponds in some “loose” sense to a non-convex penalization scheme ([40]). In [39], the adaptive Lasso is analyzed in an asymptotic setup for the case where p is fixed. Further progress in the high-dimensional scenario has been achieved by [17]. Under a rather strong mutual incoherence condition between every pair of relevant and irrelevant covariables, they prove that the adaptive Lasso recovers the correct model and has an oracle property. As we will explain in Subsection 5.4, the adaptive Lasso indeed essentially needs a - still quite restrictive - weighted version of the irrepresentable condition in order to be able to correctly estimate the support of the coefficients.

The paper [24] examines the thresholding procedure, assuming all non-zero components are large enough, an assumption we will avoid. Thresholding and multistage procedures are also considered in [12, 13]. In [37, 38], it is shown that a multi-step thresholding procedure can accurately estimate a sparse vector $\beta \in \mathbb{R}^p$ under the restricted eigenvalue condition of [4]. The two-stage procedure in [35] applies “selective penalization” in the second stage. This procedure is studied assuming incoherence conditions. A more general framework for multi-stage variable selection was studied by [32]. Their approach controls the probability of false positives (type I error) but pays a price in terms of false negatives (type II error). The contribution of this paper is that we provide bounds for the adaptive Lasso that are comparable to the bounds for the Lasso followed by a thresholding procedure. Because the true regression itself, or its linear projection, is perhaps not sparse, we moreover consider a sparse approximation of the truth, somewhat in the spirit of [34].

1.2. Organization of the paper

The next section introduces the sparse oracle approximation, with which we compare the initial and adaptive Lasso. In Section 3, we present the main results. Eigenvalues and their restricted and sparse counterparts are defined in Section 4. In Section 5 we give in an example the exact solution for the initial Lasso, illustrating that it can have very many false positives. We also provide the irrepresentable conditions for the initial, adaptive and more generally, weighted Lasso. We show in Subsection 5.4 that even the adaptive Lasso needs beta-min conditions and/or strong conditions on the design for exact variable selection. This is linked to Corollary 3.2, where it is proved that the false positives of the adaptive Lasso vanish under beta-min conditions. Some conclusions are presented in Section 6.

The rest of the paper presents intermediate results and complements for establishing the main results of Section 3. In Section 7, we consider the noiseless case, i.e., the case where $\epsilon = 0$. The reason is that many of the theoretical issues involved concern the approximation properties of the two stage procedure, and not so much the fact that there is noise. By studying the noiseless case first, we separate the approximation problem from the stochastic problem.

Both initial and adaptive Lasso are special cases of a weighted Lasso. We discuss prediction error, ℓ_q-error ($q \in \{1, 2\}$) and variable selection with the weighted Lasso in Subsection 7.1. Theorem 7.1 in this section is the core of the present work as regards prediction and estimation, and Lemma 7.1 is the main result as regards variable selection. The behavior of the noiseless initial and adaptive Lasso are simple corollaries of Theorem 7.1 and Lemma 7.1. We give in Subsection 7.2 the resulting bounds for the initial Lasso and discuss in Section 7.3 its thresholded version. In Subsection 7.4 we derive results for the adaptive Lasso by comparing it with a thresholded initial Lasso.

Section 8 studies the noisy case. It is an easy extension of the results of Sections 7.1, 7.2, 7.3 and 7.4. We do however need to further specify the choice of the tuning parameters λ_{init}, λ_{thres} and λ_{adap}. After explaining the notation, we present the bounds for the prediction error, estimation error and for the number of false positives, of the weighted Lasso. This then provides us with the tools to prove the main results.

All proofs are in Section 9. There, we also present explicit constants in the bounds to highlight the non-asymptotic character of the results.

2. Model misspecification, weak variables and the oracle

Let

$$\mathbb{E}\mathbf{Y} := \mathbf{f}^0,$$

where $\mathbf{f}^0$ is the regression function. First, we note that without loss of generality, we can assume that $\mathbf{f}^0$ is linear. If $\mathbf{f}^0$ is non-linear in the covariates, we consider its projection $\mathbf{X}\beta_{\text{true}}$ onto the linear space $\{\mathbf{X}\beta : \beta \in \mathbb{R}^p\}$, i.e.,

$$\mathbf{X}\beta_{\text{true}} := \arg\min_{\mathbf{X}\beta} \|\mathbf{f}^0 - \mathbf{X}\beta\|_2.$$

It is not difficult to see that all our results still hold if $\mathbf{f}^0$ is replaced by its projection $\mathbf{X}\beta_{\text{true}}$. The statistical implication is very relevant. The mathematical argument is the orthogonality

$$\mathbf{X}^T(\mathbf{X}\beta_{\text{true}} - \mathbf{f}^0) = 0.$$

For ease of notation, we therefore assume from now on that $\mathbf{f}^0$ is indeed linear:

$$\mathbf{f^0} := \mathbf{X}\beta_{\text{true}}.$$

Nevertheless, β_{true} itself may not be sparse. Denote the active set of β_{true} by

$$S_{\text{true}} := \{j : \ \beta_{j,\text{true}} \neq 0\},$$

which has cardinality $s_{\text{true}} := |S_{\text{true}}|$. It may well be that s_{true} is quite large, but that there are many weak variables, that is, many very small non-zero coefficients in β_{true}. Therefore, the sparse object we aim to recover may not be the "true" unknown parameter $\beta_{\text{true}} \in \mathbb{R}^p$ of the linear regression, but rather a sparse approximation. We believe that an extension to the case where $\mathbf{f}^0$ is only "approximately" sparse, better reflects the true state of nature. We emphasize however that throughout the paper, it is allowed to replace the oracle approximation b^0 given below by β_{true}. This would simplify the theory. However, we have chosen not to follow this route because it generally leads to a large price to pay in the bounds.

The sparse approximation of $\mathbf{f}^0$ that we consider is defined as follows. For a set of indices $S \subset \{1, \ldots, p\}$ and for $\beta \in \mathbb{R}^p$, we let

$$\beta_{j,S} := \beta_j \mathrm{l}\{j \in S\}, \ j = 1, \ldots, p.$$

Given a set S, the best approximation of $\mathbf{f}^0$ using only variables in S is

$$\mathrm{f}_S = \mathbf{X} b^S := \arg \min_{f = \mathbf{X}\beta_S} \|f - \mathbf{f}^0\|_2.$$

Thus, f_S is the projection of $\mathbf{f}^0$ on the linear span of the variables in S. Our target is now the projection f_{S_0}, where

$$S_0 := \arg \min_{S \subset S_{\text{true}}} \left\{ \|\mathrm{f}_S - \mathbf{f}^0\|_2^2/n + 7\lambda_{\text{init}}^2 |S| / \phi^2(6, S) \right\}.$$

Here, $|S|$ denotes the size of S. Moreover, $\phi^2(6, S)$ is a "restricted eigenvalue" (see Section 4 for its definition), which depends on the Gram matrix $\hat{\Sigma}$ and on the set S. The constants are chosen in relation with the oracle result (see Corollary 9.3). In other words, f_{S_0} is the optimal ℓ_0-penalized approximation, albeit that it is discounted by the restricted eigenvalue $\phi^2(6, S_0)$. To facilitate the interpretation, we require S_0 to be a subset of S_{true}, so that the oracle is not allowed to trade irrelevant coefficients against restricted eigenvalues. With $S_0 \subset S_{\text{true}}$, any false positive selection with respect to S_{true} is also a false positive for S_0.

We refer to f_{S_0} as the "oracle". The set S_0 is called the oracle active set, and $b^0 = b^{S_0}$ are the oracle coefficients, i.e.,

$$\mathrm{f}_{S_0} = \mathbf{X} b^0.$$

We write $s_0 = |S_0|$, and assume throughout that $s_0 \geq 1$.

Inferring the sparsity pattern, i.e. variable selection, refers to the task of estimating the set of non-zero coefficients, that is, to have a limited number of false positives (type I errors) and false negatives (type II errors). It can be verified that under reasonable conditions with suitably chosen tuning parameter λ, the "ideal" estimator

$$\hat{\beta}_{\text{ideal}} := \arg \min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda^2 |\{j : \ \beta_j \neq 0\}| \right\},$$

has $O(\lambda^2 s_0)$ prediction error and $O(s_0)$ false positives (see for instance [2] and [28]). With this in mind, we generally aim at $O(s_0)$ false positives (see also [38]), yet keeping the prediction error as small as possible (see Corollary 3.1).

As regards false negative selections, we refer to Subsection 3.5, where we derive bounds based on the ℓ_q-error.

3. Main results

3.1. Conditions

The behavior of the thresholded Lasso and adaptive Lasso depends on the tuning parameters, on the design, as well as on the true $\mathbf{f}^0$, and actually on the interplay between these quantities. To keep the exposition clear, we will use order symbols. Our expressions are functions of n, p, $\mathbf{X}$, and $\mathbf{f}^0$, and also of the tuning parameters λ_{init}, λ_{thres}, and λ_{adap}. For positive functions g and h, we say that $g = O(h)$ if $\|g/h\|_\infty$ is bounded, and $g \asymp h$ if in addition $\|h/g\|_\infty$ is bounded. Moreover, we say that $g = O_{\text{suff}}(h)$ if $\|g/h\|_\infty$ is not larger than a suitably chosen sufficiently small constant, and $g \asymp_{\text{suff}} h$ if in addition $\|h/g\|_\infty$ is bounded.

Our results depend on restricted eigenvalues $\phi(L, S, N)$, minimal restricted eigenvalues $\phi_{\min}(L, S, N)$, and minimal sparse eigenvalues $\phi_{\text{sparse}}(S, N)$ (which we generally think of as being not too small), as well on maximal sparse eigenvalues $\Lambda_{\text{sparse}}(s)$ (which we generally think of being not too large). The exact definition of these constants is given in Section 4.

When using order symbols, we simplify the expressions by assuming that that

$$\|\mathrm{f}_{S_0} - \mathbf{f}^0\|_2^2/n = O(\lambda_{\text{init}}^2 s_0/\phi^2(6, S_0)) \tag{3.1}$$

(where $\phi(6, S_0) = \phi(6, S_0, s_0)$), which roughly says that the oracle "squared bias" term is not substantially larger than the oracle "variance" term. For example, in the case of orthogonal design, this condition holds if the small non-zero coefficients are small enough, or if there are not too many of them, i.e., if

$$\sum_{|\beta_{j,\text{true}}|^2 \le 7\lambda_{\text{init}}^2} |\beta_{j,\text{true}}|^2 = O(\lambda_{\text{init}}^2 s_0).$$

We stress that (3.1) is merely to write order bounds for the oracle, bounds with which we compare the ones for the various Lasso versions. If actually the "squared bias" term is the dominating term, this mathematically does not alter the theory but makes the result more difficult to interpret. We refer to Subsection 7.1 for some exact expressions, which do not rely on assumption (3.1).

We will furthermore discuss the results on the set

$$\mathcal{T} := \left\{ 4 \max_{1 \le j \le p} |\epsilon^T \mathbf{X}_j / n| \le \lambda_{\text{init}} \right\},$$

where $\mathbf{X}_j$ is the j-th column of the matrix $\mathbf{X}$. For an appropriate choice of λ_{init}, depending on the distribution of ϵ, the set $\mathcal{T}$ has large probability. Typically, λ_{init} can be taken of order

$$\sqrt{\log p/n}.$$

The next lemma serves as an example, but the results can clearly be extended to other distributions.

Lemma 3.1. *Suppose that $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Take for a given $t > 0$,*

$$\lambda_{\text{init}} = 4\sigma\sqrt{\frac{2t + 2\log p}{n}}.$$

Then

$$\mathbb{P}(\mathcal{T}) \geq 1 - 2\exp[-t].$$

The following conditions play an important role. Conditions A and AA for thresholding are similar to those in [38] (Theorems 1.2, 1.3 and 1.4).

Condition A *For the thresholded Lasso, the threshold level λ_{thres} is chosen sufficiently large, in such a way that*

$$\left[\frac{1}{\phi^2(6, S_0, 2s_0)}\right]\lambda_{\text{init}} = O_{\text{suff}}(\lambda_{\text{thres}}).$$

Condition AA *For the thresholded Lasso, the threshold level λ_{thres} is chosen sufficiently large, but such that*

$$\left[\frac{1}{\phi^2(6, S_0, 2s_0)}\right]\lambda_{\text{init}} \asymp_{\text{suff}} \lambda_{\text{thres}}.$$

Condition B *For the adaptive Lasso, the tuning parameter λ_{adap} is chosen sufficiently large, in such a way that*

$$\left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi^3_{\min}(6, S_0, 2s_0)}\right]\lambda_{\text{init}} = O_{\text{suff}}(\lambda_{\text{adap}}).$$

Condition BB *For the adaptive Lasso, the tuning parameter λ_{adap} is chosen sufficiently large, but such that*

$$\left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi^3_{\min}(6, S_0, 2s_0)}\right]\lambda_{\text{init}} \asymp_{\text{suff}} \lambda_{\text{adap}}.$$

Remark 3.1. Note that our conditions on λ_{thres} and λ_{adap} depend on the ϕ's and Λ's, which are unknown (since S_0 is unknown). Indeed, our study is of theoretical nature, revealing common features of thresholding and the adaptive Lasso. Furthermore, it is possible to remove the dependence of the ϕ's and Λ's, when one imposes stronger sparse eigenvalue conditions, along the lines of [34]. In practice, the tuning parameters are generally chosen by cross validation.

The above conditions can be considered with a zoomed-out look, neglecting the expressions in the square brackets ($[\cdots]$), and a zoomed-in look, taking into account what is inside the square brackets. One may think of λ_{init} as the noise level (see e.g. Lemma 3.1, with the $\log p$-term the price for not knowing the relevant coefficients a priori). Zooming out, Conditions A and B say that the threshold level λ_{thres} and the tuning parameter λ_{adap} are required to be at least of the same order as λ_{init}, i.e., they should not drop below the noise level. Assumption AA and BB put these parameters exactly at the noise level, i.e., at the smallest value we allow. The reason to do this is that one then can have good prediction and estimation bounds. If we zoom in, we see in the square brackets the role played by the various eigenvalues. As they are defined only later in Section 4, it is at first reading perhaps easiest to remember that the ϕ's can be small and the Λ's can be large, but one hopes they behave well, in the sense that the values in the square brackets are not too large.

3.2. The results

The next three theorems contain the main ingredients of the present work. Theorem 3.1 is not new (see e.g. [6–8], [4], [18]), albeit that we replace the perhaps non-sparse β_{true} by the sparser b^0 (see also [26]). Recall that the latter replacement is done because it yields generally an improvement of the bounds.

Theorem 3.1. *For the initial Lasso $\hat\beta_{\text{init}} = \hat\beta$ defined in (1.1), we have on $\mathcal{T}$,*

$$\|\mathbf{X}\hat\beta_{\text{init}} - \mathbf{f}^0\|_2^2/n = \left[\frac{1}{\phi^2(6, S_0)}\right] O(\lambda_{\text{init}}^2 s_0),$$

and

$$\|\hat\beta_{\text{init}} - b^0\|_1 = \left[\frac{1}{\phi^2(6, S_0)}\right] O(\lambda_{\text{init}} s_0),$$

and

$$\|\hat\beta_{\text{init}} - b^0\|_2 = \left[\frac{1}{\phi^2(6, S_0, 2s_0)}\right] O(\lambda_{\text{init}} \sqrt{s_0}).$$

We also present here a bound for the number of false positives of the initial Lasso. In this section, we confine ourselves to the following lemma. Here, $\Lambda_{\max}^2$ is the largest eigenvalue of $\hat\Sigma$, which can generally be quite large.

Lemma 3.2. *On $\mathcal{T}$,*

$$|\hat S_{\text{init}} \backslash S_0| \le \left[\frac{\Lambda_{\max}^2}{\phi^2(6, S_0)}\right] O(s_0).$$

The bound of Lemma 3.2 can be quite large. Under further conditions as given in Lemma 8.1, the bound can sometimes be improved. See Subsection 5.3 for a lower bound in a situation where these further conditions fail.

The next theorem discusses thresholding. The paper [38] contains a careful analysis of thresholding. It presents the results of Theorem 3.2, albeit under different conditions. The role of Theorem 3.2 in the present paper (with a relatively short proof) is to make a comparison possible with the adaptive Lasso, that is, it is invoked to prove similar bounds for the adaptive Lasso, as presented in Theorem 3.3.

Theorem 3.2. *Suppose Condition A holds. Then on $\mathcal{T}$,*

$$\|\mathbf{X}\hat{\beta}_{\text{thres}} - \mathbf{f}^0\|_2^2/n = \left[\Lambda^2_{\text{sparse}}(s_0)\right] \frac{\lambda^2_{\text{thres}}}{\lambda^2_{\text{init}}} O(\lambda^2_{\text{init}} s_0),$$

and

$$\|\hat{b}_{\text{thres}} - b^0\|_1 = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\text{sparse}}(S_0, 2s_0)}\right] \frac{\lambda_{\text{thres}}}{\lambda_{\text{init}}} O(\lambda_{\text{init}} s_0),$$

and

$$\|\hat{b}_{\text{thres}} - b^0\|_2 = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\text{sparse}}(S_0, 2s_0)}\right] \frac{\lambda_{\text{thres}}}{\lambda_{\text{init}}} O(\lambda_{\text{init}} \sqrt{s_0}),$$

and

$$|\hat{S}_{\text{thres}} \backslash S_0| = \left[\frac{1}{\phi^4(6, S_0, 2s_0)}\right] \frac{\lambda^2_{\text{init}}}{\lambda^2_{\text{thres}}} O(s_0).$$

Theorem 3.3. *Suppose Condition B holds. Then on $\mathcal{T}$,*

$$\|\mathbf{X}\hat{\beta}_{\text{adap}} - \mathbf{f}^0\|_2^2/n = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\text{min}}(6, S_0, 2s_0)}\right] \frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}} O(\lambda^2_{\text{init}} s_0),$$

and

$$\|\hat{\beta}_{\text{adap}} - b^0\|_1 = \left[\frac{\Lambda^{1/2}_{\text{sparse}}(s_0)}{\phi^{3/2}_{\text{min}}(6, S_0, 2s_0)}\right] \sqrt{\frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}}} O(\lambda_{\text{init}} s_0),$$

and

$$\|\hat{\beta}_{\text{adap}} - b^0\|_2 = \left[\frac{\Lambda^{1/2}_{\text{sparse}}(s_0)\phi^{1/2}_{\text{min}}(6, S_0, 2s_0)}{\phi^2_{\text{min}}(6, S_0, 3s_0)}\right] \sqrt{\frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}}} O(\lambda_{\text{init}} \sqrt{s_0}),$$

and

$$|\hat{S}_{\text{adap}} \backslash S_0| = \left[\frac{\Lambda^2_{\text{sparse}}(s_0)}{\phi^4(6, S_0, 2s_0)} \frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\text{min}}(6, S_0, 2s_0)}\right] \frac{\lambda_{\text{init}}}{\lambda_{\text{adap}}} O(s_0).$$

Theorem 3.2 and 3.3 show how the results depend on the choice of the tuning parameters λ_{thres} and λ_{adap}. The following corollary takes the choices of Conditions AA and BB, as these choices give the smallest prediction and estimation error.

Corollary 3.1. *Suppose we are on $\mathcal{T}$. Then, under Condition AA,*

$$\|\mathbf{X}\hat{b}_{\text{thres}} - \mathbf{f}^0\|_2^2/n = \left[\frac{\Lambda^2_{\text{sparse}}(s_0)}{\phi^4(6, S_0, 2s_0)}\right] O(\lambda^2_{\text{init}} s_0), \tag{3.2}$$

and

$$\|\hat{b}_{\text{thres}} - b^0\|_1 = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\text{sparse}}(S_0, 2s_0)\phi^2(6, S_0, 2s_0)}\right] O(\lambda_{\text{init}} s_0),$$

and

$$\|\hat{b}_{\text{thres}} - b^0\|_2 = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\text{sparse}}(S_0, 2s_0)\phi^2(6, S_0, 2s_0)}\right] O(\lambda_{\text{init}}\sqrt{s_0}),$$

and

$$|\hat{S}_{\text{thres}} \backslash S_0| = O(s_0). \tag{3.3}$$

Similarly, under Condition BB,

$$\|\mathbf{X}\hat{\beta}_{\text{adap}} - \mathbf{f}^0\|_2^2/n = \left[\frac{\Lambda^2_{\text{sparse}}(s_0)}{\phi^4_{\text{min}}(6, S_0, 2s_0)}\right] O(\lambda^2_{\text{init}} s_0), \tag{3.4}$$

and

$$\|\hat{\beta}_{\text{adap}} - b^0\|_1 = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi^3_{\text{min}}(6, S_0, 2s_0)}\right] O(\lambda_{\text{init}} s_0),$$

and

$$\|\hat{\beta}_{\text{adap}} - b^0\|_2 = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi^2_{\text{min}}(6, S_0, 3s_0)\phi_{\text{min}}(6, S_0, 2s_0)}\right] O(\lambda_{\text{init}}\sqrt{s_0}),$$

and

$$|\hat{S}_{\text{adap}} \backslash S_0| = \left[\frac{\Lambda^2_{\text{sparse}}(s_0)\phi^2_{\text{min}}(6, S_0, 2s_0)}{\phi^4(6, S_0, 2s_0)}\right] O(s_0). \tag{3.5}$$

3.3. Comparison with the Lasso

At the zoomed-out level, where all ϕ's and Λ's are neglected, we see that the thresholded Lasso (under Condition AA) and the adaptive Lasso (under Condition BB) achieve the same order of magnitude for the prediction error as the initial, one-stage Lasso discussed in Theorem 3.1. The same is true for their estimation errors. Zooming in on the ϕ's and the Λ's, their error bounds are generally larger than for the initial Lasso.

We will show in Subsection 5.3 that in certain examples the bound for $|\hat{S}_{\text{init}} \backslash S_0|$ of Lemma 3.2 cannot be improved, and also that the results of Theorem 3.1 for the prediction and estimation error of the initial Lasso are sharp. Therefore general message is that thresholding and the adaptive Lasso can have similar prediction and estimation error as the initial Lasso, and are often far better as regards variable selection. Of course, a careful comparison depends on the restricted eigenvalues involved. We generally think of $1/\phi_{\text{min}}$'s and Λ_{sparse}'s being $O(1)$, i.e., we stay at the zoomed-out level. For the case of very large Λ_{sparse}, we refer to Lemma 3.3. If the minimal restricted eigenvalues are very small, one actually enters a different regime, where the design is highly correlated. This has its implications on the random part of the problem, leading for example to a smaller order of magnitude for the tuning parameters. We refer to [27] for some illustrations.

In the paper [34], one can find further conditions that ensure that also for the initial Lasso, modulo ϕ's and Λ's, the number of false positives is of order s_0. These conditions are rather involved and also improve the bounds for the adaptive and thresholded Lasso. In Subsection 8.3 we briefly discuss a condition of similar nature as the one used in [34].

3.4. Comparison between adaptive and thresholded Lasso

When zooming-out, we see that the adaptive and thresholded Lasso have bounds of the same order of magnitude, for prediction, estimation and variable selection.

At the zoomed-in level, the adaptive and thresholded Lasso also have very similar bounds for the prediction error (compare (3.2) with (3.4)) in terms of the ϕ's and Λ's. A similar conclusion holds for their estimation error. We remark that our choice of Conditions AA and BB for the tuning parameters is motivated by the fact that according to our theory, these give the smallest prediction and estimation errors. It then turns out that the "optimal" errors of the two methods match at a quite detailed level. However, if we zoom-in even further and look at the definition of $\phi_{\rm sparse}$, ϕ, and $\phi_{\min}$ in Section 4, it will show up that the bounds for the adaptive Lasso prediction and estimation error are (slightly) larger.

Regarding variable selection, at zoomed-out level the results are also comparable (see (3.3) and (3.5)). Zooming-in on the the ϕ's and Λ's, the adaptive Lasso may have more false positives than the thresholded version.

A conclusion is that at the zoomed-in level, the adaptive Lasso has less favorable bounds as the refitted thresholded Lasso. However, these are still only bounds, which are based on focussing on a direct comparison between the two methods, and we may have lost the finer properties of the adaptive Lasso. Indeed, the non-explicitness of the adaptive Lasso makes its analysis a non-trivial task. The adaptive Lasso is a quite popular practical method, and we certainly do not advocate that it should always be replaced by thresholding and refitting.

3.5. Bounds for the number of false negatives

The ℓ_q-error has immediate consequences for the number of false negatives: if for some estimator $\hat\beta$, some target b^0, and some constant $\delta_q^{\rm upper}$ one has

$$\|\hat\beta - b^0\|_q \le \delta_q^{\rm upper}$$

then the number of undetected yet large coefficients cannot be very large, in the sense that

$$|\{j:\ \hat\beta_j = 0, |b_j^0| > \delta\}|^{1/q} \le \frac{\delta_q^{\rm upper}}{\delta}.$$

Therefore, on $\mathcal{T}$, for example

$$\left|\left\{ j:\ \hat\beta_{j,\rm init} = 0, \left[\frac{1}{\phi^2(6, S_0, 2s_0)}\right]\sqrt{s_0}\lambda_{\rm init} = O_{\rm suff}(|b_j^0|)\right\}\right| = 0.$$

Similar bounds hold for the thresholded and the adaptive Lasso (considering now, in terms of the ϕ's and Λ's, somewhat larger $|b_j^0|$).

One may argue that one should not aim at detecting variables that the oracle considers as irrelevant. Nevertheless, given an estimator $\hat\beta$, it is straightforward to bound $\|\hat\beta - \beta_{\rm true}\|_q$ in terms of $\|\hat\beta - b^0\|_q$: apply the triangle inequality

$$\|\hat\beta - \beta_{\rm true}\|_q \le \|\hat\beta - b^0\|_q + \|b^0 - \beta_{\rm true}\|_q.$$

Moreover, for $q = 2$, one has the inequality

$$\|b^0 - \beta_{\text{true}}\|_2^2 \le \frac{\|\mathrm{f}_{S_0} - \mathbf{f}^0\|_2^2}{n\Lambda_{\min}^2(S_{\text{true}})},$$

where $\Lambda_{\min}^2(S)$ is the smallest eigenvalue of the Gram matrix corresponding to the variables in S. One may verify that $\phi(6, S_{\text{true}}) \le \Lambda_{\min}(S_{\text{true}})$. In other words, choosing β_{true} as target instead of b^0 does in our approach not lead to an improvement in the bounds for $\|\hat{\beta} - \beta_{\text{true}}\|_2$.

3.6. Assuming beta-min conditions

Let us have a closer look at what conditions on the size of the coefficients can bring us. We call such conditions beta-min conditions. We only discuss the adaptive Lasso (for thresholding we refer to [38]).

We define

$$|b^0|_{\min} := \min_{j \in S_0} |b_j^0|.$$

Moreover, we let

$$|b^0|_{\text{harm}}^2 := \left(\frac{1}{s_0} \sum_{j \in S_0} \frac{1}{|b_j^0|^2} \right)^{-1}$$

be the harmonic mean of the squared coefficients.

In Chapter 7 of [5], sparse eigenvalues are avoided altogether and results are given under a condition which replaces the quantities $|b^0|_{\min}$ and $|b^0|_{\text{harm}}$ we consider here by a trimmed harmonic mean. This refinements require further arguments. Therefore let us here only consider a simple version.

Condition C *For the adaptive Lasso, take λ_{adap} sufficiently large, such that*

$$|b^0|_{\text{harm}} = O_{\text{suff}}(\lambda_{\text{adap}}).$$

Condition CC *For the adaptive Lasso, take λ_{adap} sufficiently large, but such that*

$$|b^0|_{\text{harm}} \asymp_{\text{suff}} \lambda_{\text{adap}}.$$

Remark 3.2. In Condition CC the choice for λ_{adap} is larger than in Condition BB. We will see in Lemma 3.3 that beta-min conditions allow for a larger tuning parameter λ_{adap} without paying a price in prediction error, but with a possible gain in variable selection properties. In some examples, one can show that the prediction optimal choice for the tuning parameter will be of the right order. In practice one may use cross validation, although we have as yet no theoretical guarantee that cross validation will mimic the theoretical values we consider.

Lemma 3.3. *Suppose that for some constant $\delta_\infty^{\text{upper}}$, on $\mathcal{T}$,*

$$\|\hat{\beta}_{\text{init}} - b^0\|_\infty \le \delta_\infty^{\text{upper}}.$$

Assume in addition that

$$|b^0|_{\min} > 2\delta_\infty^{\text{upper}}. \tag{3.6}$$

Then under Condition C,

$$\|\mathbf{X}\hat{\beta}^2_{\text{adap}} - \mathbf{f}^0\|_2^2/n = \left[\frac{1}{\phi^2(6, S_0)}\right]\frac{\lambda^2_{\text{adap}}}{|b^0|^2_{\text{harm}}}O(\lambda^2_{\text{init}}s_0),$$

and

$$\|\hat{\beta}_{\text{adap}} - b^0\|_1 = \left[\frac{1}{\phi^2(6, S_0)}\right]\frac{\lambda_{\text{adap}}}{|b^0|_{\text{harm}}}O(\lambda_{\text{init}}s_0),$$

and

$$\|\hat{\beta}_{\text{adap}} - b^0\|_2 = \left[\frac{1}{\phi^2(6, S_0, 2s_0)}\right]\frac{\lambda_{\text{adap}}}{|b^0|_{\text{harm}}}O(\lambda_{\text{init}}\sqrt{s_0}),$$

and

$$|\hat{S}_{\text{adap}}\backslash S_0| = \left(s_0 \vee \left[\frac{\Lambda^2_{\text{sparse}}(s_0)}{\phi^2(6, S_0)\phi^4(6, S_0, 2s_0)}\right]O\left(\frac{\lambda^2_{\text{init}}s_0}{|b^0|^2_{\text{harm}}}\right)\right)$$
$$\wedge \left[\frac{1}{\phi^2(6, S_0)\phi^4(6, S_0, 2s_0)}\right]O\left(\frac{\lambda^2_{\text{init}}s_0^2}{|b^0|^2_{\text{harm}}}\right).$$

It is clear that by Theorem 3.1,

$$\|\hat{\beta}_{\text{init}} - b^0\|_\infty = \left[\frac{\sqrt{s_0}}{\phi^2(6, S_0)} \wedge \frac{1}{\phi^2(6, S_0, 2s_0)}\right]O(\lambda_{\text{init}}\sqrt{s_0}).$$

This can be improved under coherence conditions on the Gram matrix. To simplify the exposition, we will not discuss such improvements in detail (see [20]).

Under Condition CC, the bound for the prediction error and estimation error is again the smallest. We moreover have the following corollary.

Corollary 3.2. *Assume the conditions of Lemma 3.3 and*

$$\left[\frac{1}{\phi^2(6, S_0)\phi^4(6, S_0, 2s_0)}\right]\lambda_{\text{init}}s_0 = O_{\text{suff}}(|b^0|_{\text{harm}}). \tag{3.7}$$

Then on $\mathcal{T}$,

$$|\hat{S}_{\text{adap}}\backslash S_0| = 0.$$

Asymptotics For the case of Gaussian errors $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, with $\sigma^2 \asymp 1$, one typically takes $\lambda_{\text{init}} \asymp \sqrt{\log p/n}$ (see Lemma 3.1). If we assume moreover that the design is such that $\phi(6, S_0, 2s_0) \asymp 1$ and all non-zero coefficients b_j^0 are in the same range, the beta-min condition we impose in (3.7) is $s_0\sqrt{\log p/n} = O(b_j^0)$ for all $j \in S_0$, that is, all non-zero coefficients should be not smaller than $s_0\sqrt{\log p/n}$ in order of magnitude.

We will see in Subsection 5.4 that in certain examples, the beta-min condition (3.7), which roughly requires a separation of order $\sqrt{s_0}$ between $|\hat{\beta}_{j,\text{init}}|$, $j \in S_0$ and $|\hat{\beta}_{j,\text{init}}|$, $j \notin S_0$ (with, according to Theorem 3.1, for $j \notin S_0$, $|\hat{\beta}_{j,\text{init}}|$ being of order $\lambda_{\text{init}}\sqrt{s_0}/\phi^2(6, S_0, 2s_0)$) is necessary for variable selection. It is clear that with thresholding, one does not need this large separation. Thus, in the beta-min context, thresholding wins from the adaptive Lasso.

4. Notation and definition of generalized eigenvalues

We reformulate the problem in $L_2(Q)$, where Q is a generic probability measure on some space $\mathcal{X}$. (This is somewhat more natural in the noiseless case, which we will consider in Section 7.) Let $\{\psi_j\}_{j=1}^p \subset L_2(Q)$ be a given dictionary. For $j = 1, \ldots, p$, the function ψ_j will play the role of the j-th co-variable. The Gram matrix is

$$\Sigma := \int \psi^T \psi dQ, \ \psi := (\psi_1, \ldots, \psi_p).$$

We assume that Σ is normalized, i.e., that $\int \psi_j^2 dQ = 1$ for all j. In our final results, we will actually take $\Sigma = \hat{\Sigma}$, the (empirical) Gram matrix corresponding to fixed design.

Write a linear function of the ψ_j with coefficients $\beta \in \mathbb{R}^p$ as

$$f_\beta := \sum_{j=1}^p \psi_j \beta_j.$$

The $L_2(Q)$-norm is denoted by $\| \cdot \|$, so that

$$\|f_\beta\|^2 = \beta^T \Sigma \beta.$$

Recall that for an arbitrary $\beta \in \mathbb{R}^p$, and an arbitrary index set S, we use the notation

$$\beta_{j,S} = \beta_j \text{l}\{j \in S\}.$$

We now present our notation for eigenvalues. We also introduce restricted eigenvalues and sparse eigenvalues.

4.1. Eigenvalues

The largest eigenvalue of Σ is denoted by $\Lambda_{\max}^2$, i.e.,

$$\Lambda_{\max}^2 := \max_{\|\beta\|_2 = 1} \beta^T \Sigma \beta.$$

We will also need the largest eigenvalue of a submatrix containing the inner products of variables in S:

$$\Lambda_{\max}^2(S) := \max_{\|\beta_S\|_2 = 1} \beta_S^T \Sigma \beta_S.$$

Its minimal eigenvalue is

$$\Lambda_{\min}^2(S) := \min_{\|\beta_S\|_2 = 1} \beta_S^T \Sigma \beta_S.$$

4.2. Restricted eigenvalues

A restricted eigenvalue is of similar nature as the minimal eigenvalue of Σ, but with the coefficients β restricted to certain subsets of $\mathbb{R}^p$. The restricted eigenvalue condition we impose corresponds to the so-called *adaptive* version as introduced in [29]. It differs from the restricted eigenvalue condition in [4] or [18, 19]. This is due to the fact that we want to mimic the oracle f_{S_0}, that is, do not choose $\mathbf{f}^0$ as target, so that we have to deal with a bias term $\|\mathrm{f}_{S_0} - \mathbf{f}^0\|$. For a given S, our restricted eigenvalue condition is stronger than the one in [4] or [18, 19]. On the other hand, we apply it to the smaller set S_0 instead of to $S_{\rm true}$.

Define for an index set $S \subset \{1, \ldots, p\}$, and for a set $\mathcal{N} \supset S$ and constant $L > 0$, the sets of restrictions

$$\mathcal{R}(L, S, \mathcal{N}) := \left\{ \beta : \ \|\beta_{\mathcal{N}^c}\|_1 \le L\sqrt{|\mathcal{N}|}\|\beta_{\mathcal{N}}\|_2, \ \max_{j \in \mathcal{N}^c} |\beta_j| \le \min_{j \notin \mathcal{N} \setminus S} |\beta_j| \right\}.$$

Definition: Restricted eigenvalue. *For $N \ge |S|$, we call*

$$\phi^2(L, S, N) := \min \left\{ \frac{\|f_\beta\|^2}{\|\beta_{\mathcal{N}}\|_2^2} : \ \mathcal{N} \supset S, \ |\mathcal{N}| \le N, \ \beta \in \mathcal{R}(L, S, \mathcal{N}) \right\}$$

the (L, S, N)-restricted eigenvalue*. The* (L, S, N)-restricted eigenvalue condition *holds if $\phi(L, S, N) > 0$.*
For the case $N = |S|$, we write $\phi(L, S) := \phi(L, S, |S|)$.
The minimal (L, S, N)-restricted eigenvalue *is*

$$\phi^2_{\min}(L, S, N) := \min_{\mathcal{N} \supset S, \ |\mathcal{N}| = N} \phi^2(L, \mathcal{N}).$$

It is easy to see that $\phi_{\min}(L, S, N) \le \phi(L, S, N) \le \phi(L, S) \le \Lambda_{\min}(S)$ for all $L > 0$. It can moreover be shown that

$$\phi^2(L, S, N) \ge \min \left\{ \|f_\beta\|^2 : \ \mathcal{N} \supset S, \ |\mathcal{N}| = N, \ \|\beta_{\mathcal{N}^c}\|_2 \le L, \ \|\beta_{\mathcal{N}}\|_2 = 1 \right\}.$$

4.3. Sparse eigenvalues

We also invoke sparse eigenvalues, in line with the sparse Riesz condition occurring in [34].

Definition: Sparse eigenvalues. *For $N \in \{1, \ldots, p\}$, the* maximal sparse eigenvalue *is*

$$\Lambda_{\rm sparse}(N) = \max_{\mathcal{N}: \ |\mathcal{N}| = N} \Lambda_{\max}(\mathcal{N}).$$

For an index set $S \subset \{1, \ldots, p\}$ with $|S| \le N$, the minimal sparse eigenvalue *is*

$$\phi_{\rm sparse}(S, N) := \min_{\mathcal{N} \supset S: \ |\mathcal{N}| = N} \Lambda_{\min}(\mathcal{N}).$$

One easily verifies that for any set $\mathcal{N}$ with $|N| = ks$, $k \in \mathbb{N}$,

$$\Lambda_{\text{max}}(\mathcal{N}) \le \sqrt{k}\Lambda_{\text{sparse}}(s).$$

Moreover, for all $L \ge 0$,

$$\phi_{\text{sparse}}(S, N) = \phi(0, S, N) \ge \phi(L, S, N).$$

See [5] for some further relations.

5. Some lower bounds for the (weighted) Lasso, and the case of random design

This section complements the upper bounds of Section 3 with some examples where the bounds (for the initial Lasso) cannot be improved. We also consider necessary conditions for exact variable selection. We first present the irrepresentable condition. Subsection 5.2 treats the case of random design. We will then consider in Subsection 5.3 an example where the irrepresentable condition does not hold, and where the Lasso has many false positives. We first consider in Subsection 5.3.1 a simple situation, where the correlation pattern follows a worst case scenario. A more realistic correlation pattern, as given in Subsection 5.3.2, requires rather refined arguments. Subsection 5.4 presents an example where the adaptive Lasso needs the beta-min conditions of Corollary 3.2 for exact variable selection, i.e., the bounds in this corollary are sharp.

5.1. The irrepresentable condition

For a symmetric $(p \times p)$ -matrix Σ, we define

$$\Sigma_{1,1}(S) := (\sigma_{j,k})_{j,k \in S},$$

$$\Sigma_{2,1}(S) := (\sigma_{j,k})_{j \notin S, k \in S}.$$

Let $0 \le \theta \le 1$. The θ-irrepresentable condition assumes

$$\|\hat{\Sigma}_{2,1}(S_{\text{true}})\hat{\Sigma}_{1,1}^{-1}(S_{\text{true}})(\tau_{\text{true}})_{S_{\text{true}}}\|_\infty \le \theta,$$

where τ_{true} is the sign-vector of β_{true}. It is known that under beta-min conditions the θ-irrepresentable condition with $\theta = 1$ is necessary for the initial Lasso to have no false positives, and that with θ sufficiently small it is also sufficient for consistent variable selection in a proper asymptotic setting (see [23] and [36]). The next lemma, from Chapter 7 in [5], presents a uniform θ-irrepresentable condition giving a non-asymptotic result. See also Lemma 6.2 in [29] for the noiseless case (where $\lambda_\epsilon = 0$).

Lemma 5.1. *Let*

$$\mathcal{T} := \mathcal{T}_\epsilon := \{\max_{1\le j\le p} 2|\epsilon^T \mathbf{X}_j| < \lambda_\epsilon\}.$$

Suppose

$$\sup_{\|\tau_{S_{\rm true}}\|_\infty \le 1} \|\hat\Sigma_{2,1}(S_{\rm true})\hat\Sigma_{1,1}^{-1}(S_{\rm true})\tau_{S_{\rm true}})\|_\infty < \frac{\lambda_{\rm init}-\lambda_\epsilon}{\lambda_{\rm init}+\lambda_\epsilon}. \tag{5.1}$$

Then on $\mathcal{T}$,

$$\hat S_{\rm init} \subset S_{\rm true}.$$

In order to keep the prediction error small, one chooses λ_ϵ small while keeping the probability of $\mathcal{T} = \mathcal{T}_\epsilon$ large, and $\lambda_{\rm init}$ of the same order as λ_ϵ. In particular, we throughout take $\lambda_{\rm init} = 2\lambda_\epsilon$. Then

$$\frac{\lambda_{\rm init}-\lambda_\epsilon}{\lambda_{\rm init}+\lambda_\epsilon} = \frac{1}{3}.$$

5.2. Random design

Let X_i be the i-th row of $\mathbf{X}$. The idea that follows is easiest understood when the covariables X_i are i.i.d. copies of some random row vector $X \in \mathbb{R}^p$ with mean zero and covariance matrix $\Sigma := \mathbb{E}X^TX$. Let $\|\hat\Sigma - \Sigma\|_\infty := \max_{j,k} |\hat\sigma_{j,k} - \sigma_{j,k}|$. Under e.g. exponential moment conditions (and also under weaker moment conditions) it holds that for $\lambda_X = O(\sqrt{\log p/n})$, one has $\|\hat\Sigma - \Sigma\|_\infty \le \lambda_X$ on a set $\mathcal{T}_X$ with large probability. Looking at the definition of restricted eigenvalues, we see that they depend on the Gram matrix under consideration. It is shown in [29] that Σ-restricted eigenvalue conditions imply $\hat\Sigma$-restricted eigenvalue conditions when $\|\hat\Sigma - \Sigma\|_\infty$ is small enough, depending on the sparsity $s = |S|$. For example, for $\|\hat\Sigma - \Sigma\|_\infty \le \lambda_X$, where $32\lambda_X s/\phi^2_\Sigma(L,S,2s) \le 1$, we have $\phi^2_{\hat\Sigma}(L,S,2s) \ge \phi^2_\Sigma(L,S,2s)/2$. Similar statements can be made for minimal restricted eigenvalues and sparse eigenvalues. Thus, we can handle sparsity s of order $s = O_{\rm suff}(\lambda_X^{-1}\phi^2_\Sigma(L,S,2s))$. With $\lambda_X \asymp \sqrt{\log p/n}$, this becomes $s = O_{\rm suff}(\sqrt{n/\log p}) \times \phi^2_\Sigma(L,S,2s)$.

Our conclusion is that $\hat\Sigma$ in a rather general setting inherits (up to constants) its generalized eigenvalues from those of Σ when the two matrices are close enough. This can be applied to establish bounds for the prediction error.

As explained above, for variable selection the irrepresentable condition plays a crucial role. Again, one may want to replace the empirical covariance matrix $\hat\Sigma$ by a population version Σ. Suppose that on the set $\mathcal{T}_\epsilon \cap \mathcal{T}_X$, we have for some $\tilde\lambda_X$,

$$\|\hat\Sigma - \Sigma\|_\infty \|\hat\beta_{\rm init} - \beta_{\rm true}\|_1 \le \tilde\lambda_X.$$

Then the condition

$$\sup_{\|\tau_{S_{\rm true}}\|_\infty \le 1} \|\Sigma_{2,1}(S_{\rm true})\Sigma_{1,1}^{-1}(S_{\rm true})\tau_{S_{\rm true}})\|_\infty < \frac{\lambda_{\rm init}-(\lambda_\epsilon+\tilde\lambda_X)}{\lambda_{\rm init}+(\lambda_\epsilon+\tilde\lambda_X)}$$

suffices for the initial Lasso to have no false positives on $\mathcal{T}_\epsilon \cap \mathcal{T}_X$ (see [5], Chapter 7). As we have seen (Theorem 3.1), on $\mathcal{T}_\epsilon$, and with $\lambda_{\rm init} = 2\lambda_\epsilon$, we have

$$\|\hat\beta_{\rm init} - b^0\|_1 = O(\lambda_{\rm init} s_0)/\phi^2(6, S_0)$$

(where $\phi^2(6, S_0) = \phi^2_{\hat\Sigma}(6, S_0)$). Suppose now that the approximation error $\|b^0 - \beta_{\rm true}\|_1$ is also of this order. Then the Σ-irrepresentable condition has slightly larger noise term $\lambda_\epsilon + \tilde\lambda_X$ (instead of λ_ϵ), but the additional $\tilde\lambda_X$ is of order $\lambda_{\rm init}\lambda_X/\phi^2(6, S_0)$. So we can use the Σ-irrepresentable condition when again the sparsity s_0 is of order $s_0 = O_{\rm suff}(\lambda_X^{-1}\phi^2(6, S_0))$.

5.3. An example illustrating that the Lasso can select too many false positives

We consider now the case of equal correlation. Subsection 5.3.1 treats an idealized setting, with "worst case" correlation pattern. Here, the Lasso selects all p variables. However, the considered correlation pattern is a-typical (e.g. for Gaussian errors it occurs with probability zero). In that sense, this idealized case mainly serves as a first step towards the more realistic setting of Subsection 5.3.2. There, we show that for Gaussian errors, the Lasso selects at least an order of magnitude $s_{\rm true}\log n$ false positives (see Theorem 5.1). To show this is quite involved and requires refined concentration inequalities. We furthermore believe that the lower bound $s_{\rm true}\log n$ is not sharp, i.e., that in fact the number of false positives can be even larger.

Of course, one can get many false positives by choosing the tuning parameter very small. We will however not do this, but allow for a value $\lambda_{\rm init}$ of the order $\max_{1\le j\le p}|\epsilon^T\psi_j|/n$, meaning that for instance for Gaussian errors the prediction error satisfies the oracle bound $s_{\rm true}\log p/n$.

5.3.1. Worst case correlation

We do not provide any proofs of the results in this section: they follow from straightforward calculations.

Let P be a probability measure on $\mathcal{X} \times \mathbb{R}$ with marginal distribution Q on $\mathcal{X}$ (possibly P is the empirical distribution $P_n = \sum_{i=1}^n \delta_{(X_i,Y_i)}/n$). We study a function $\mathbf{Y} \in L_2(P)$ satisfying $\mathbf{Y} = \mathbf{f}^0 + \epsilon$, where (as in Section 4) $\mathbf{f}^0 = \sum_{j=1}^p \beta_{\rm true}\psi_j$, and where $\psi_1, \ldots, \psi_p$ are given functions in $L_2(Q)$ (the latter again playing the role of the covariables). The Gram matrix is $\Sigma := \int \psi^T\psi dQ$, where $\psi := (\psi_1, \ldots, \psi_p)$. The $L_2(P)$ inner product is denoted by $(\cdot,\cdot)$, and $\|\cdot\|$ is the $L_2(P)$-norm. We let

$$\hat\beta_{\rm init} := \arg\min_\beta \left\{ \|\mathbf{Y} - \sum_{j=1}^p \psi_j\beta_j\|^2 + 2\lambda_{\rm init}\|\beta\|_1 \right\}.$$

Note we replaced λ_{init} by $2\lambda_{\text{init}}$. This will simplify the expressions. To make our analysis more explicit, we throughout take $\lambda_{\text{init}} = 2\lambda_\epsilon$, where

$$\lambda_\epsilon \geq \max_{1 \leq j \leq p} |(\epsilon, \psi_j)|.$$

Let ι be a p-vector of all 1's, and let

$$\Sigma := (1-\rho)I + \rho\iota\iota^T,$$

where $0 \leq \rho < 1$. This corresponds to

$$\psi_j = \sqrt{1-\rho}\tilde{\psi}_j + \sqrt{\rho}z, \; j = 1, \ldots, p,$$

where $\|z\| = 1$, $(\psi_j, z) = 0$ for all j, and $\int \tilde{\psi}^T \tilde{\psi} dQ = I$. In other words, the covariables have a variable z in common, but are otherwise uncorrelated.

The (minimal restricted) eigenvalues are as follows.

Lemma 5.2. *Let S be a set with cardinality s. We have*

$$\Lambda^2_{\max}(S) = 1 - \rho + \rho s.$$

Moreover for any L (and for s and p even),

$$\phi^2_{\min}(L, S, 2s) = \phi^2(L, S, 2s) = \Lambda^2_{\min}(S) = 1 - \rho.$$

Thus, in this example, the maximal eigenvalue $\Lambda^2_{\max}$ is equal to $1-\rho+\rho p \geq \rho p$, i.e., it can be vary large.

It is easy to see that

$$\sup_{\|\tau_{S_{\text{true}}}\|_\infty \leq 1} \|\Sigma_{2,1}(S_{\text{true}})\Sigma^{-1}_{1,1}(S_{\text{true}})(\tau)_{S_{\text{true}}})\|_\infty = \frac{\rho s_{\text{true}}}{1 - \rho + \rho s_{\text{true}}}.$$

Therefore, to be able to construct an example with false positives, we assume that

$$\Delta := \frac{\rho s_{\text{true}}}{1 - \rho + \rho s_{\text{true}}} - \frac{\lambda_{\text{init}} - \lambda_\epsilon}{\lambda_{\text{init}} + \lambda_\epsilon} > 0.$$

Note that with the choice $\lambda_{\text{init}} = 2\lambda_\epsilon$, this holds for ρ sufficiently large:

$$\frac{\rho s_{\text{true}}}{1 - \rho + \rho s_{\text{true}}} > \frac{1}{3} \text{ iff } \rho > \frac{1}{2s_{\text{true}} + 1}.$$

Now, we will actually assume that ρ stays away from 1, that is, we exclude the case

$$\rho = 1 - o(1).$$

The reason is that for ρ in a neighborhood of 1, the design is highly correlated. For such designs, a smaller order for the tuning parameter can be appropriate (see [27]).

Therefore, let us consider the range

$$\frac{2}{2s_{\text{true}}+1} \le \rho \le \frac{1}{2}.$$

We assume that

$$(\epsilon, \psi_j) = \begin{cases} -\lambda_\epsilon & j \in S_{\text{true}} \\ +\lambda_\epsilon & j \notin S_{\text{true}} \end{cases}.$$

The latter can be seen as the "worst case" correlation pattern. Some other and perhaps more typical correlation patterns (that increase the penalty on the true positives and decrease the penalty on the true negatives) will lead to similar conclusions but more involved calculations.

We further simplify the situation by assuming that

$$\beta_{j,\text{true}} = \beta_0, \ \forall \ j \in S_{\text{true}},$$

where β_0 is some positive constant. It is easy to see that $b^0 = \beta_{\text{true}}$ when β_0 is sufficiently larger than λ_{init}, i.e., when $\lambda_{\text{init}} = O_{\text{suff}}(\beta_0)$.

It is not difficult to see that the Lasso is also constant on S_{true}:

$$\hat\beta_{j,\text{init}} = \hat\beta_0, \ \forall \ j \in S_{\text{true}},$$

where $\hat\beta_0$ is some non-negative constant. Moreover,

$$\hat\beta_{j,\text{init}} = \hat\gamma, \ \forall \ j \notin S_{\text{true}},$$

where $\hat\gamma$ is some other constant.

The next lemma presents the explicit solution for the initial Lasso. We also give the order of magnitude of the terms.

Lemma 5.3. *Suppose that*

$$\frac{2}{2s_{\text{true}}+1} \le \rho \le \frac{1}{2},$$

and that

$$\beta_0 > \left(\frac{\lambda_{\text{init}} + \lambda_\epsilon}{(1-\rho+\rho s_{\text{true}})} + \frac{\rho(p-s_{\text{true}})\Delta(\lambda_{\text{init}}+\lambda_\epsilon)}{(1-\rho)(1-\rho+\rho p)} \right).$$

We have

$$\hat\beta_0 - \beta_0 = -\left(\frac{\lambda_{\text{init}} + \lambda_\epsilon}{(1-\rho+\rho s_{\text{true}})} + \frac{\rho(p-s_{\text{true}})\Delta(\lambda_{\text{init}}+\lambda_\epsilon)}{(1-\rho)(1-\rho+\rho p)} \right) \asymp -\lambda_{\text{init}},$$

and

$$\hat\gamma = \frac{\Delta(1-\rho+\rho s_{\text{true}})(\lambda_{\text{init}}+\lambda_\epsilon)}{(1-\rho)(1-\rho+\rho p)} \asymp \frac{\lambda_{\text{init}}}{1-\rho}\frac{s_{\text{true}}}{p}.$$

We also give the prediction, ℓ_1- and ℓ_2-error, and the number of false negatives, and their order of magnitude.

Lemma 5.4. *Suppose the conditions of Lemma 5.3. We have*

$$\|\hat{f}_{\text{init}} - \mathbf{f}^0\|^2 = \frac{(\lambda_{\text{init}} + \lambda_\epsilon)^2 s_{\text{true}}}{(1 - \rho + \rho s_{\text{true}})}$$
$$+\frac{\Delta^2\big[1 - \rho + \rho s_{\text{true}}\big](\lambda_{\text{init}} + \lambda_\epsilon)^2 (p - s_{\text{true}})}{(1-\rho)(1-\rho+\rho p)}$$
$$\asymp \frac{\lambda_{\text{init}}^2 s_{\text{true}}}{1-\rho}$$

and

$$\|\hat{\beta}_{\text{init}} - \beta_{\text{true}}\|_1 = \frac{(\lambda_{\text{init}} + \lambda_\epsilon) s_{\text{true}}}{(1 - \rho + \rho s_{\text{true}})} + \frac{\rho(p - s_{\text{true}}) s_{\text{true}} \Delta(\lambda_{\text{init}} + \lambda_\epsilon)}{(1-\rho)(1-\rho+\rho p)}$$
$$+\frac{\Delta(1 - \rho + \rho s_{\text{true}})(p - s_{\text{true}})(\lambda_{\text{init}} + \lambda_\epsilon)}{(1-\rho)(1-\rho+\rho p)} \asymp \frac{\lambda_{\text{init}} s_{\text{true}}}{1-\rho},$$
$$\|\hat{\beta}_{\text{init}} - \beta_{\text{true}}\|_2 = \frac{(\lambda_{\text{init}} + \lambda_\epsilon)\sqrt{s_{\text{true}}}}{(1 - \rho + \rho s_{\text{true}})} + \frac{\rho(p - s_{\text{true}})\sqrt{s_{\text{true}}}\Delta(\lambda_{\text{init}} + \lambda_\epsilon)}{2(1-\rho)(1-\rho+\rho p)}$$
$$+\frac{\Delta(1 - \rho + \rho s_{\text{true}})\sqrt{p - s_{\text{true}}}(\lambda_{\text{init}} + \lambda_\epsilon)}{(1-\rho)(1-\rho+\rho p)}$$
$$\asymp \frac{\lambda_{\text{init}}\sqrt{s_{\text{true}}}}{1-\rho},$$

and

$$|\hat{S}_{\text{init}} \backslash S_{\text{true}}| = (p - s_{\text{true}}).$$

Note that we kept $1/(1-\rho)$ in our order bounds, whereas admittedly, $1-\rho \asymp 1$ in the range considered. Nevertheless, as $1-\rho$ plays the role of the ϕ^2's, we see that the bounds we provide incorporate the ϕ's in a reasonable way.

It follows that when $\rho \asymp 1/s_{\text{true}}$, the result of Lemma 3.2 is sharp. Also for $\rho \asymp 1/s_{\text{true}}$ the maximal sparse eigenvalue is $\asymp 1$. We can then apply Theorems 3.2 and 3.3. There, the thresholds λ_{thres} and λ_{adap} appear to be too large, as $\hat{\gamma} \asymp (s_{\text{true}}/p)\lambda_{\text{init}}/(1-\rho)$. On the other hand, for $p \to s_{\text{true}}$ e.g. Theorem 3.2 gives a bound for λ_{thres} which is arbitrarily close to the bound needed.

When ρ is of larger order than $1/s_{\text{true}}$ the maximal sparse eigenvalue becomes increasing in s_{true}. Recall that Lemma 3.3 not necessarily needs sparse eigenvalue conditions (but instead needs beta-min conditions).

5.3.2. A realistic example

The proofs for this subsection are in Section 9. We consider a situation where $p = n-1$. The errors $\epsilon_1, \ldots, \epsilon_n$ are assumed to be i.i.d. $\mathcal{N}(0,1)$-distributed. The design is constructed as in the previous subsection: we let $\{\tilde{\psi}_1, \ldots, \tilde{\psi}_{n-1}, z\}$ be n orthogonal vectors in $\mathbb{R}^n$ with length $\sqrt{n}$, and

$$\psi_j := \sqrt{1-\rho}\tilde{\psi}_j + \sqrt{\rho} z, \;\; j = 1, \ldots, n-1,$$

where $0 < \rho \leq 1/2$. The Gram matrix is then

$$\hat{\Sigma} = (1-\rho)I + \rho\iota\iota^T,$$

where ι is an $(n-1)$-vector of 1's.

Again, to simplify the expressions, we replace λ_{init} by $2\lambda_{\text{init}}$, i.e., we consider the Lasso

$$\hat{\beta}_{\text{init}} := \arg\min \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + 2\lambda_{\text{init}}\|\beta\|_1 \right\}.$$

First, we present the form of the solution of this minimization problem.

Lemma 5.5. *Consider for some λ_ϵ, the set*

$$\mathcal{T} := \left\{ \max_{1\leq j\leq p} |\epsilon^T\psi_j|/n \leq \lambda_\epsilon \right\},$$

and assume $\lambda_{\text{init}} \geq C_1\lambda_\epsilon$ and $\beta_j^0 > C\lambda_{\text{init}}$ for all $j \in S_{\text{true}}$, with $C_1 > 1$ and C a sufficiently large positive constant depending only on C_1. Then on $\mathcal{T}$ the following holds. Firstly, $\hat{S}_{\text{init}} \supset S_{\text{true}}$. Moreover, all $\hat{\beta}_{j,\text{init}}$ are non-negative, and

$$\lambda_{\text{init}}|\hat{S}_{\text{init}}| - \sum_{j\in\hat{S}_{\text{init}}} \epsilon^T\psi_j/n$$

is non-negative. Finally, when $\hat{S}_{\text{init}}\backslash S_{\text{true}} \neq \emptyset$, for all $j \in \hat{S}_{\text{init}}\backslash S_{\text{true}}$,

$$\hat{\beta}_{j,\text{init}} = \epsilon^T\psi_j/n - \lambda_{\text{init}} \tag{5.2}$$
$$+\frac{\rho}{1-\rho+\rho|\hat{S}_{\text{init}}|}\left(\lambda_{\text{init}}|\hat{S}_{\text{init}}| - \sum_{j\in\hat{S}_{\text{init}}} \epsilon^T\psi_j/n\right).$$

We use concentration results for $\epsilon^T\psi_j$, $j \in S_{\text{true}}^c$, to establish the following lower bound for the number of false positives:

Theorem 5.1. *For some positive constants C_5, C_6, C_7, C_8 and C, and constant $0 < \alpha < 1$ not depending on n, with $\lambda_{\text{init}} = C_5\sqrt{\log n/n}$, $\min_{j\in S_{\text{true}}} \beta_j^0 > C\lambda_{\text{init}}$, $s_{\text{true}} > C_6$, and*

$$\frac{\rho s_{\text{true}}}{1-\rho+\rho s_{\text{true}}} > 1 - \frac{1}{C_7},$$

it holds that with probability at least α,

$$|\hat{S}_{\text{init}}\backslash S_{\text{true}}| \geq \frac{s_{\text{true}}\log n}{C_8}.$$

5.4. The weighted irrepresentable condition

This subsection will show that, even in the noiseless case, exact variable selection with the adaptive Lasso needs rather strong conditions. Let, as in Subsection

5.3, Q be a probability measure on $\mathcal{X}$, $\psi_1, \ldots, \psi_p$ be a fixed dictionary in $L_2(Q)$ and $\| \cdot \|$ be the $L_2(Q)$-norm. The Gram matrix is $\Sigma := \int \psi^T \psi dQ$. We write

$$f_\beta = \sum_{j=1}^p \beta_j \psi_j, \ \beta \in \mathbb{R}^p,$$

and denote the truth by

$$\mathbf{f}^0 = f_{\beta_{\text{true}}}.$$

The weighted Lasso is

$$\beta_{\text{weight}} := \arg\min_\beta \left\{ \|f_\beta - \mathbf{f}^0\|^2 + \lambda_{\text{init}} \lambda_{\text{weight}} \|W\beta\|_1 \right\},$$

where $W := \text{diag}(w)$ with $w := (w_1, \ldots, w_p)$ a vector of positive weights. We let $S_{\text{weight}} = \{j : \beta_{j,\text{weight}} \neq 0\}$.

As stated in Subsection 5.1, the initial Lasso essentially needs the irrepresentable condition in order to have no false positives. Similar statements can be made for the weighted Lasso. We let $W_S := \text{diag}(\{w_j\}_{j \in S})$, and $\tau_{\text{true}} := \text{sign}(\beta_{\text{true}})$. We say that the weighted θ-irrepresentable condition holds if

$$\|W^{-1}_{S^c_{\text{true}}} \Sigma_{2,1}(S_{\text{true}}) \Sigma^{-1}_{1,1}(S_{\text{true}}) W_{S_{\text{true}}} (\tau_{\text{true}})_{S_{\text{true}}} \|_\infty \leq \theta.$$

The reparametrization $\beta \mapsto \gamma := W^{-1}\beta$ leads to the following lemma, which is the weighted variant of the third part of Lemma 6.2 in [29].

Lemma 5.6. *Assume the beta-min condition*

$$|\beta_{\text{true}}|_{\min} > \lambda_{\text{weight}} \lambda_{\text{init}} \sup_{\|\tau_{S_{\text{true}}}\|_\infty \leq 1} \|\Sigma^{-1}_{1,1}(S_{\text{true}}) W_{S_{\text{true}}} \tau_{S_{\text{true}}} \|_\infty / 2.$$

Suppose $S_{\text{weight}} \subset S_{\text{true}}$. Then the weighted θ-irrepresentable condition holds with $\theta = 1$.

We now consider conditions for the weighted irrepresentable condition to hold. Let

$$w^{\min}_{S^c} := \min_{j \notin S} w_j.$$

Lemma 5.7. *Suppose that*

$$\|w_S\|_2 \leq \theta \Lambda_{\min}(S) w^{\min}_{S^c}. \tag{5.3}$$

Then

$$\sup_{\|\tau_S\|_\infty \leq 1} \|W^{-1}_{S^c} \Sigma_{2,1}(S) \Sigma^{-1}_{1,1}(S) W_S \tau_S \|_\infty \leq \theta.$$

The next example shows that the result of Lemma 5.7 cannot be improved without assuming further conditions. As a consequence, for exact variable selection, the adaptive Lasso needs a rather large lower bound on $|\beta_{\text{true}}|_{\text{harm}}$, where

$$|\beta_{\text{true}}|^2_{\text{harm}} := \left(\frac{1}{s_{\text{true}}} \sum_{j \in S_{\text{true}}} \frac{1}{\beta^2_{j,\text{true}}} \right)^{-1}$$

is the harmonic mean of the squared coefficients. The corresponding upper bound occurs in Corollary 3.2.

Example 5.1. Let $S_{\text{true}} = \{1, \ldots, s\}$, with cardinality $s := |S_{\text{true}}|$, be the active set, and write

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}.$$

We now will take a special (idealized) choice for Σ. We suppose that $\Sigma_{1,1} := I$ is the $(s \times s)$-identity matrix, and

$$\Sigma_{2,1} := \rho(c_2 c_1^T),$$

with $0 \le \rho < 1$, and with c_1 an s-vector and c_2 a $(p-s)$-vector, satisfying $\|c_1\|_2 = \|c_2\|_2 = 1$. Moreover, we suppose $\Sigma_{2,2}$ is the $((p-s) \times (p-s))$-identity matrix. Then $\Lambda_{\min}(S_{\text{true}}) = 1$ and the smallest eigenvalue of Σ is $1-\rho$. Its largest eigenvalue is $1+\rho$. Take $c_1 = w_{S_{\text{true}}}/\|w_{S_{\text{true}}}\|_2$, and $c_2 = (0, \ldots, 1, 0, \ldots)^T$, where the 1 is placed at $\arg\min_{j \in S_{\text{true}}^c} w_j$. Then, for $\tau_{S_{\text{true}}}$ being a vector of all 1's,

$$W_{S_{\text{true}}} \tau_{S_{\text{true}}} = w_{S_{\text{true}}},$$

and

$$\Sigma_{2,1}\Sigma_{1,1}^{-1} W_{S_{\text{true}}} = \rho c_2 \|w_{S_{\text{true}}}\|_2,$$

and hence the bound of Lemma 5.7 is, up to constants, sharp:

$$\|W_{S_{\text{true}}^c}^{-1} \Sigma_{2,1}\Sigma_{1,1}^{-1} W_{S_{\text{true}}} \tau_{S_{\text{true}}}\|_\infty = \rho \|w_{S_{\text{true}}}\|_2 / w_{S_{\text{true}}^c}^{\min}.$$

We now first present some heuristics concerning the consequences for the adaptive Lasso, and then provide a detailed exact example. The adaptive Lasso aims at weights

$$w_j = \frac{1}{|\beta_{j,\text{true}}|},\ j \in S_{\text{true}}.$$

Then

$$\|w_{S_{\text{true}}}\|_2 = \frac{\sqrt{s_{\text{true}}}}{|\beta_{\text{true}}|_{\text{harm}}}.$$

The θ-irrepresentable condition with $\theta = 1$ then requires

$$|\beta_{\text{true}}|_{\text{harm}} \ge \rho \sqrt{s_{\text{true}}} \delta_\infty,$$

where

$$\delta_\infty := \frac{1}{w_{S^c}^{\min}} = \|1/w_{S_{\text{true}}^c}\|_\infty.$$

Therefore for exact variable selection, one needs $|\beta_{\text{true}}|_{\text{harm}}$ to be an order of magnitude $\sqrt{s_{\text{true}}}$ larger than δ_∞. This condition also shows up in Corollary 3.2.

Let us consider in more detail the special case with $p = s+1$, $\rho = 1/2$ and $\beta_{j,\text{true}} = \beta_0$ for all $j \in S_{\text{true}}$, where β_0 is a positive constant. Then $c_1 = (1, \ldots, 1)/\sqrt{s}$, and $c_2 = 1$. Straightforward calculations show that when $\rho >$

$1/\sqrt{s}$, and $\lambda_{\rm init} = O_{\rm suff}(\beta_0)$, the initial Lasso $\beta_{\rm init}$ (for the noiseless case, see (7.2) for its definition) will select variable $(s+1)$, and

$$\beta_{s+1,{\rm init}} = \frac{\rho\sqrt{s}-1}{1-\rho^2}\lambda_{\rm init} \asymp \sqrt{s}\lambda_{\rm init}.$$

Furthermore,

$$\beta_{j,{\rm init}} = \beta_j^0 - \frac{1-\rho/\sqrt{s}}{1-\rho^2}\lambda_{\rm init},\ j \in S_{\rm true}.$$

Hence, with

$$w_j = \frac{1}{|\beta_{j,{\rm init}}|},\ \forall\ j,$$

we have

$$\delta_\infty \asymp \sqrt{s}\lambda_{\rm init},$$

and for $\beta_0 > 2\lambda_{\rm init}$,

$$\|w_{S_{\rm true}}\|_2 \asymp \sqrt{s}/\beta_0.$$

It follows from the weighted irrepresentable condition (Lemma 5.6) that under the beta-min condition $\lambda_{\rm init}\lambda_{\rm adap} = O(\beta_0^2)$, we need the lower bound

$$s\lambda_{\rm init} = O_{\rm suff}(\beta_0)$$

for exact variable selection with the adaptive Lasso. This corresponds to the upper bound of Corollary 3.2, so that the bounds are sharp.

6. Conclusions

We present some comparable bounds for the adaptive Lasso and the thresholded Lasso with refitting and we also compare them to the ordinary Lasso. The framework of our analysis allows for misspecified linear models whose best linear projection is not necessarily sparse and with possibly small non-zero regression coefficients, i.e., many weak variables. This setting is much more realistic than the usual high-dimensional framework where the model is true with only a few but strong variables.

Estimating the support S_0 of the non-zero coefficients is a hard statistical problem. The irrepresentable condition, which is essentially a necessary condition for exact recovery of the non-zero coefficients by the one-step Lasso, is much too restrictive in many cases. In this paper, our main focus is on having $O(s_0)$ false positives while achieving good prediction and estimation. This is inspired by the behavior of the "ideal" ℓ_0-penalized estimator.

We have examined thresholding the Lasso with least squares refitting and the adaptive Lasso. Our main conclusion is that both methods can have about the same prediction and estimation error as the one-stage ordinary Lasso, and that both gain over the one-stage Lasso in the sense of having less false positives. We provide additional support of this point by also proving a lower bound for the one-stage Lasso with respect to false positive selections. Moreover, according to our theory (and not exploiting the fact that the adaptive Lasso mimics thresholding and refitting using an "oracle" threshold), thresholding with least squares refitting and the adaptive Lasso perform equally well, even when considered at a rather fine scale. Our bounds for the adaptive Lasso are more sensitive to small (minimal) restricted eigenvalues or small minimal sparse eigenvalues, or large sparse maximal eigenvalues. Both thresholded and adaptive Lasso benefit from a situation with large non-zero coefficients of the oracle, i.e., from beta-min conditions. For exact variable selection, the adaptive Lasso however needs more severe beta-min conditions than thresholding.

We do not give a full account of the tightness of our bounds for both two-step methods. One can however construct (idealized) examples where the bounds are sharp in order of magnitude (as a function of $\lambda_{\rm init}$ and s_0). Moreover, the thresholded Lasso allows a rather direct analysis, and we believe there is little room for improvement of the bounds for this method. The analysis of the adaptive Lasso is more involved. Our comparison to thresholding might not do justice to the adaptive Lasso. Indeed, we have not fully exploited the finer oracle properties of the adaptive Lasso.

In practice the tuning parameters are often chosen by cross validation, which may correspond to a choice giving the smallest prediction error. It is not within the scope of this paper to prove that with cross validation, thresholding and the adaptive Lasso again have comparable theoretical performance, although we do believe this to be typically the case. As for the computational aspect, we observe the following. For the solution path for all $\lambda_{\rm adap}$, the adaptive Lasso needs $O(n|\hat{S}_{\rm init}| \min(n, |\hat{S}_{\rm init}|))$ essential operation counts. The same order of operation counts is needed when computing the thresholded Lasso for the whole solution path over all $\lambda_{\rm thres}$. Therefore, the two methods are also computationally comparable.

7. The noiseless case

Consider a fixed target $\mathbf{f}^0 = f_{\beta_{\rm true}} \in L_2(Q)$. Let $S \subset \{1, \ldots, p\}$ and let $\mathrm{f}_S := \arg\min_{f=f_{\beta_S}} \|f_{\beta_S} - \mathbf{f}^0\|$ be the projection of $\mathbf{f}^0$ on the $|S|$-dimensional linear space spanned by the variables $\{\psi_j\}_{j\in S}$. We denote the coefficients of f_S by b^S, i.e.,

$$\mathrm{f}_S = \sum_{j\in S} \psi_j b_j^S = f_{b^S}.$$

The oracle set S_0 is defined by trading off dimension against fit, namely

$$S_0 := \arg\min_{S\subset S_{\rm true}} \left\{ \|\mathrm{f}_S - \mathbf{f}^0\|^2 + \frac{3\lambda_{\rm init}^2 |S|}{\phi^2(2,S)} \right\}, \tag{7.1}$$

where the constants are now from Theorem 7.1 (or its Corollary 9.1). We call f_{S_0} the oracle, and we let $b^0 := b^{S_0}$, i.e., $\mathrm{f}_{S_0} = f_{b^0}$. Also, we let $s_0 := |S_0|$ and assume $s_0 \geq 1$.

For simplicity, we assume throughout that

$$\|\mathrm{f}_{S_0} - \mathbf{f}^0\|^2 = O(\lambda_{\text{init}}^2 s_0/\phi^2(2, S_0)),$$

which roughly says that the approximation error does not overrule the penalty term.

The initial Lasso is

$$\beta_{\text{init}} := \arg\min_\beta \biggl\{ \|f_\beta - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\|\beta\|_1 \biggr\}. \tag{7.2}$$

We assume that the tuning parameter λ_{init} is set at some fixed value. Of course, in the noiseless case, the optimal - in terms of prediction error - value for λ_{init} is $\lambda_{\text{init}} = 0$. However, in the noisy case, a strictly positive lower bound for λ_{init} is dictated by the noise level. Write

$$f_{\text{init}} := f_{\beta_{\text{init}}},\ S_{\text{init}} := \{j :\ \beta_{j,\text{init}} \ne 0\},\ \delta_{\text{init}} := \|f_{\text{init}} - \mathbf{f}^0\|. \tag{7.3}$$

Let for $\delta > 0$,

$$S_{\text{init}}^\delta := \{j :\ |\beta_{j,\text{init}}| > \delta\}.$$

Then $\mathrm{f}_{S_{\text{init}}^\delta} = f_{b^{S_{\text{init}}^\delta}}$ is the refitted Lasso after thresholding at δ. Note that we express explicitly the dependence of the thresholded estimator on the threshold level, which we now call δ (instead of λ_{thres} as we did in the introduction). The reason for this is that the analysis of the adaptive Lasso will go via the thresholded Lasso with a choice of the threshold δ that trades off prediction error against estimation error (see (9.8) in the proof of Theorem 7.4).

The adaptive Lasso is

$$\beta_{\text{adap}} := \arg\min_\beta \left\{ \|f_\beta - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{adap}} \sum_{j=1}^p \frac{|\beta_j|}{|\beta_{j,\text{init}}|} \right\}.$$

The second stage tuning parameter λ_{adap} is again assumed to be strictly positive. We denote the resulting adaptive variants of (7.3) by

$$f_{\text{adap}} := f_{\beta_{\text{adap}}},\ S_{\text{adap}} := \{j :\ \beta_{j,\text{adap}} \ne 0\},\ \delta_{\text{adap}} := \|f_{\text{adap}} - \mathbf{f}^0\|.$$

As the initial and adaptive Lasso are special cases of the weighted Lasso, many of the results in Subsections 7.2, 7.3 and 7.4 are consequences of those for the weighted Lasso as studied in Subsection 7.1. As in Subsection 5.4, the weighted Lasso is defined as

$$\beta_{\text{weight}} := \arg\min_\beta \left\{ \|f_\beta - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j=1}^p w_j|\beta_j| \right\},$$

where the $\{w_j\}_{j=1}^p$ are non-negative weights. We set $f_{\text{weight}} := f_{\beta_{\text{weight}}}$, $S_{\text{weight}} := \{j :\ \beta_{j,\text{weight}} \ne 0\}$. Moreover, we define

$$\|w_S\|_2^2 := \sum_{j\in S} w_j^2,\ w_{S^c}^{\min} := \min_{j\notin S} w_j.$$

By the reparametrization $\beta \mapsto \gamma := W\beta$, where $W = \text{diag}(w_1, \ldots, w_p)$, one sees that the weighted Lasso is a standard Lasso with Gram matrix

$$\Sigma_{\text{weight}} := W^{-1}\Sigma W^{-1}.$$

We emphasize however that Σ_{weight} is generally not normalized, i.e., generally $\text{diag}(\Sigma_{\text{weight}}) \neq I$.

7.1. The weighted Lasso

We first present a bound for the prediction and estimation error and then consider variable selection.

Theorem 7.1. *Let S be an index set with cardinality $s := |S|$, satisfying for some constants $M \geq 0$ and $L > 0$,*

$$w_{S^c}^{\min} \geq M/L, \ \ \|w_S\|_2/\sqrt{s} \leq M.$$

Then for all β, we have

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 \leq 2\|f_{\beta_S} - \mathbf{f}^0\|^2 + \frac{6\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 M^2 s}{\phi^2(2L, S)}.$$

Moreover, for all β, we have

$$\sqrt{s}\|(\beta_{\text{weight}})_S - \beta_S\|_2 + \|(\beta_{\text{weight}})_{S^c}\|_1/L \leq \frac{3\|f_{\beta_S} - \mathbf{f}^0\|^2}{\lambda_{\text{init}}\lambda_{\text{weight}} M} + \frac{3\lambda_{\text{init}}\lambda_{\text{weight}} M s}{\phi^2(2L, S)}.$$

Finally, it holds for all β, that

$$\|\beta_{\text{weight}} - \beta_S\|_2 \leq \frac{6(L \vee 1)\|f_{\beta_S} - \mathbf{f}^0\|^2}{\lambda_{\text{init}}\lambda_{\text{weight}} M \sqrt{s_0}} + \frac{6L\lambda_{\text{init}}\lambda_{\text{weight}} M(s + s_0)}{\phi^2(2L, S, s + s_0)\sqrt{s_0}}.$$

We will apply the above theorem with S the set of the smaller weights.

Corollary 7.1. *Fix some arbitrary $\delta > 0$, and let*

$$S_{\text{weight}}^{\delta} \supset \{j : \ w_j < 1/\delta\}, \ (S_{\text{weight}}^{\delta})^c \supset \{j : \ w_j > 1/\delta\}.$$

The indices j with $w_j = 1/\delta$ can be put in either $S_{\text{weight}}^{\delta}$ or in its complement. Suppose that for some $\alpha \geq 0$,

$$|S_{\text{weight}}^{\delta} \backslash S_0| \leq \alpha s_0.$$

Taking $S = S_{\text{weight}}^{\delta}$, $L = 1$ and $M = 1/\delta$ in Theorem 7.1, we get that for all β,

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 \leq 2\|f_{\beta_{S_{\text{weight}}^{\delta}}} - \mathbf{f}^0\|^2 + \frac{6\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 (1 + \alpha)s_0}{\delta^2 \phi_{\min}^2(2, S_0, (1 + \alpha)s_0)}.$$

Moreover,

$$\|\beta_{\rm weight} - \beta_{S^\delta_{\rm weight}}\|_1 \le \frac{3\delta \|f_{\beta_{S^\delta_{\rm weight}}} - \mathbf{f}^0\|^2}{\lambda_{\rm init}\lambda_{\rm weight}} + \frac{3\lambda_{\rm init}\lambda_{\rm weight}(1+\alpha)s_0}{\delta\phi^2(2, S_0, (1+\alpha)s_0)},$$

and

$$\|\beta_{\rm weight} - \beta_{S^\delta_{\rm weight}}\|_2 \le \frac{6\delta \|f_{\beta_{S^\delta_{\rm weight}}} - \mathbf{f}^0\|^2}{\sqrt{s_0}\lambda_{\rm init}\lambda_{\rm weight}} + \frac{6\lambda_{\rm init}\lambda_{\rm weight}(2+\alpha)\sqrt{s_0}}{\delta\phi^2_{\rm min}(2, S_0, (2+\alpha)s_0)}.$$

In the case $\alpha = 0$, one may replace in the last bound, $\phi^2_{\rm min}(2, S_0, (2+\alpha)s_0) = \phi_{\rm min}(2, S_0, 2s_0)$ by $\phi(2, S_0, 2s_0)$.

Our next theme is variable selection. The *Karush-Kuhn-Tucker* (*KKT*) conditions (see [3]) can be invoked to derive Lemma 7.1 below, where we use the notation

$$\|(1/w)_S\|_2^2 := \sum_{j\in S} \frac{1}{w_j^2}.$$

Lemma 7.1. *It holds that*

$$|S_{\rm weight}\backslash S_0|^2 \le 4\Lambda^2_{\rm max}(S_{\rm weight}\backslash S_0)\frac{\|f_{\rm weight} - \mathbf{f}^0\|^2}{\lambda^2_{\rm weight}}\frac{\|(1/w)_{S_{\rm weight}\backslash S_0}\|_2^2}{\lambda^2_{\rm init}}. \tag{7.4}$$

If $|S_{\rm weight}\backslash S_0| > s_0$, we have

$$|S_{\rm weight}\backslash S_0| \le 8\Lambda^2_{\rm sparse}(s_0)\frac{\|f_{\rm weight} - \mathbf{f}^0\|^2}{\lambda^2_{\rm weight}s_0}\frac{\|(1/w)_{S_{\rm weight}\backslash S_0}\|_2^2}{\lambda^2_{\rm init}}.$$

7.2. The initial Lasso

Recall that

$$\delta_{\rm init} := \|f_{\rm init} - \mathbf{f}^0\|.$$

For $q \ge 1$, we define

$$\delta_q := \|\beta_{\rm init} - b^0\|_q.$$

Theorem 7.2. *The prediction error of the initial Lasso has*

$$\delta^2_{\rm init} = \left[\frac{1}{\phi^2(2, S_0)}\right] O(\lambda^2_{\rm init}s_0),$$

and its estimation error has

$$\delta_1 = \left[\frac{1}{\phi^2(2, S_0)}\right] O(\lambda_{\rm init}s_0),\ \ \delta_2 = \left[\frac{1}{\phi^2(2, S_0, 2s_0)}\right] O(\lambda_{\rm init}\sqrt{s_0}).$$

The initial estimator has number of false positives

$$|S_{\rm init}\backslash S_0| = \left[\frac{\Lambda^2_{\rm max}(S_{\rm init}\backslash S_0)}{\phi^2(2, S_0)}\right] O(s_0).$$

Considering the variable selection result, it is clear that $\Lambda^2_{\max}(S_{\rm init}\backslash S_0) \le \Lambda^2_{\max}$. Without further conditions, this cannot be refined, and the eigenvalue $\Lambda^2_{\max}$ can be quite large (yet having the minimal eigenvalue of Σ bounded away from zero). Therefore, the result of Theorem 7.2 needs further conditions for good variable selection properties of the initial Lasso.

7.3. Thresholding the initial estimator

Variable selection results by thresholding are not difficult to obtain:

$$|S^\delta_{\rm init}\backslash S_0|^{1/q} \le \frac{\delta_q}{\delta}.$$

Hence, for $\delta \ge \delta_1/s_0 \wedge \delta_2/\sqrt{s_0}$, we get for $q \in \{1,2\}$,

$$|S^\delta_{\rm init}\backslash S_0| \le s_0. \tag{7.5}$$

If the coefficients of the oracle are sufficiently large, thresholding will improve the prediction and estimation error. Here, we do not impose such minimal size conditions. The estimation error of the thresholded Lasso is then still easy to assess. Our bound for the prediction error, however, now depends on maximal sparse eigenvalues.

At this stage, we invoke the noiseless counterparts of Conditions A and AA.

Condition a *We have* $\lambda_{\rm init}/\phi^2(2,S_0) = O_{\rm suff}(\delta)$.

Condition aa *We have* $\lambda_{\rm init}/\phi^2(2,S_0,2s_0) \asymp_{\rm suff} \delta$.

Theorem 7.3. *Assume Condition a. Then*

$$\|{\rm f}_{S^\delta_{\rm init}} - \mathbf{f}^0\|^2 = \Lambda^2_{\rm sparse}(s_0)\left[\frac{\delta^2}{\lambda^2_{\rm init}}\right] O(\lambda^2_{\rm init}s_0),$$

$$\|b^{S^\delta_{\rm init}} - b^0\|_2 = \frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\rm sparse}(S_0,2s_0)}\left[\frac{\delta}{\lambda_{\rm init}}\right] O(\lambda_{\rm init}\sqrt{s_0}),$$

and

$$|S^\delta_{\rm init}\backslash S_0| = \left[\frac{1}{\phi^4(2,S_0,2s_0)}\right]\left[\frac{\lambda^2_{\rm init}}{\delta^2}\right] O(s_0).$$

The expressions for the prediction and estimation error lead to favoring the choice $\lambda_{\rm init}/\phi^2(2,S_0,2s_0) \asymp_{\rm suff} \delta$ *of Condition aa, which yields*

$$\|{\rm f}_{S^\delta_{\rm init}} - \mathbf{f}^0\|^2 = \left[\frac{\Lambda^2_{\rm sparse}}{\phi^4(2,S_0,2s_0)}\right] O(\lambda^2_{\rm init}s_0),$$

$$\|b^{S^\delta_{\rm init}} - b^0\|_2 = \left[\frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\rm sparse}(S_0,2s_0)\phi^2(2,S_0,2s_0)}\right] O(\lambda_{\rm init}\sqrt{s_0}),$$

and

$$|S^\delta_{\rm init}\backslash S_0| = O(s_0).$$

7.4. The adaptive Lasso

Observe that the adaptive Lasso is somewhat more reluctant than thresholding and refitting: the latter ruthlessly disregards all coefficients with $|\beta_{j,\text{init}}| \le \delta$ (i.e., these coefficients get penalty ∞), and puts zero penalty on coefficients with $|\beta_{j,\text{init}}| > \delta$. The adaptive Lasso gives the coefficients with $|\beta_{j,\text{init}}| \le \delta$ a penalty of at least $\lambda_{\text{init}}(\lambda_{\text{adap}}/\delta)$ and those with $|\beta_{j,\text{init}}| > \delta$ a penalty of at most $\lambda_{\text{init}}(\lambda_{\text{adap}}/\delta)$. (Looking ahead, we will actually need to choose $\lambda_{\text{adap}} \ge \delta$ in the noisy case, see Theorem 3.3.)

Recall

$$\delta_{\text{adap}} := \|f_{\text{adap}} - \mathbf{f}^0\|.$$

The noiseless versions of Conditions B and BB are:

Condition b *We have*

$$\lambda_{\text{init}}\left[\frac{\phi_{\min}(2, S_0, 2s_0)\Lambda_{\text{sparse}}(s_0)}{\phi^4(2, S_0, 2s_0)}\right] = O_{\text{suff}}(\lambda_{\text{adap}}).$$

Condition bb *We have*

$$\lambda_{\text{init}}\left[\frac{\phi_{\min}(2, S_0, 2s_0)\Lambda_{\text{sparse}}(s_0)}{\phi^4(2, S_0, 2s_0)}\right] \asymp_{\text{suff}} \lambda_{\text{adap}}.$$

Note the slight discrepancy with the noisy versions: the noiseless versions are somewhat better. This is due to the fact that we also will need to choose λ_{adap} large enough to handle the noise.

Theorem 7.4. *Assume Condition b. Then*

$$\delta^2_{\text{adap}} = \left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\min}(2, S_0, 2s_0)}\right]\frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}}O(\lambda^2_{\text{init}}s_0),$$

and

$$\|\beta_{\text{adap}} - b^0\|_1 = \left[\frac{\Lambda^{1/2}_{\text{sparse}}(s_0)}{\phi^{3/2}_{\min}(2, S_0, 2s_0)}\right]\sqrt{\frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}}}O(\lambda_{\text{init}}s_0),$$

and

$$\|\beta_{\text{adap}} - b^0\|_2 = \left[\frac{\Lambda^{1/2}_{\text{sparse}}(s_0)\phi^{1/2}_{\min}(2, S_0, 2s_0)}{\phi^2_{\min}(2, S_0, 3s_0)}\right]\sqrt{\frac{\lambda_{\text{adap}}}{\lambda_{\text{init}}}}O(\lambda_{\text{init}}\sqrt{s_0}),$$

and

$$|S_{\text{adap}}\backslash S_0| = \frac{\Lambda^2_{\text{sparse}}(s_0)}{\phi^4(2, S_0, 2s_0)}\left[\frac{\Lambda_{\text{sparse}}(s_0)}{\phi_{\min}(2, S_0, 2s_0)}\right]\frac{\lambda_{\text{init}}}{\lambda_{\text{adap}}}O(s_0).$$

Considering the bounds for the prediction and estimation error leads to favoring the choice of Condition bb, giving

$$\delta^2_{\text{adap}} = \left[\frac{\Lambda^2_{\text{sparse}}(s_0)}{\phi^4(2, S_0, 2s_0)}\right]O(\lambda^2_{\text{init}}s_0),$$

$$\|\beta_{\rm adap} - b^0\|_1 = \left[\frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\rm min}(2, S_0, 2s_0)\phi^2(2, S_0, 2s_0)} \right] O(\lambda_{\rm init} s_0),$$

$$\|\beta_{\rm adap} - b^0\|_2 = \left[\frac{\Lambda_{\rm sparse}(s_0)\phi_{\rm min}(2, S_0, 2s_0)}{\phi^2_{\rm min}(2, S_0, 3s_0)\phi^2(2, S_0, 2s_0)} \right] O(\lambda_{\rm init}\sqrt{s_0}),$$

and

$$|S_{\rm adap} \backslash S_0| = \frac{\Lambda^2_{\rm sparse}(s_0)}{\phi^2_{\rm min}(2, S_0, 2s_0)} O(s_0).$$

8. Adding noise

After introducing the notation (Subsection 8.1), we will give the extension of the results for the weighted Lasso to the noisy case[1] (see Theorem 8.1). Once this is done, results for the initial Lasso, its thresholded version, and for the adaptive Lasso, follow in the same way as in Subsections 7.2, 7.3 and 7.4. The new point is to take care that the tuning parameters are chosen in such a way that the noisy part due to variables in S_0^c are overruled by the penalty term. In our situation, this can be done by taking $\lambda_{\rm init}$, as well as $\lambda_{\rm adap} \ge \lambda_{\rm init}$ sufficiently large.

We provide the result for the noisy weighted Lasso in Subsection 8.2. Theorems 3.1, 3.2 and 3.3 follow from this and from some further results for the noisy case (their proofs are in Subsection 9.5). In Section 8.3, we look at more restrictive sparse eigenvalue conditions in the spirit of [34].

8.1. Notation for the noisy case

Consider an n-dimensional vector of observations

$$\mathbf{Y} = \mathbf{f}^0 + \epsilon.$$

where $\mathbf{f}^0 := (\mathbf{f}^0(X_1), \ldots, \mathbf{f}^0(X_n))^T$, with $X_1, \ldots, X_n$ co-variables in some space $\mathcal{X}$. Let $\{\psi_j\}_{j=1}^p$ be a given dictionary.

The regression $\mathbf{f}^0$, the dictionary $\{\psi_j\}$, and $f_\beta := \sum \psi_j \beta_j$ are now considered as vectors in $\mathbb{R}^n$. The norm we use is the normalized Euclidean norm

$$\|f\| := \|f\|_n := \|f\|_2/\sqrt{n} : \ f \in \mathbf{R}^n,$$

induced by the inner product

$$(f, \tilde{f})_n := \frac{1}{n} \sum_{i=1}^n f_i \tilde{f}_i, \ f, \tilde{f} \in \mathbb{R}^n.$$

[1] Of separate interest is a direct comparison of the noisy initial Lasso with the noisy ℓ_0-penalized estimator. Replacing $\mathbf{f}^0$ by $\mathbf{Y}$ in Corollary 9.1 in Subsection 9.3 (and dropping the requirement $S \subset S_{\rm true}$) gives

$$\|\mathbf{Y} - \hat{f}_{\rm init}\|_n^2 \le 2 \min_S \left\{ \|\mathbf{Y} - \hat{\mathbf{f}}_S\|_n^2 + \frac{3\lambda^2_{\rm init}|S|}{\phi^2(2, S)} \right\}.$$

In other words, the probability measure Q is now $Q := Q_n = \sum_{i=1}^n \delta_{X_i}/n$, the empirical measure of the co-variables $X_1, \ldots, X_n$. With some abuse of notation, we also write

$$\|\mathbf{Y} - f\|_n^2 := \|\mathbf{Y} - f\|_2^2/n,$$

and

$$(\epsilon, f)_n := \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i).$$

The design matrix $\mathbf{X}$ is

$$\mathbf{X} = (\psi_1, \ldots, \psi_p).$$

We write the eigenvalues involved as before, e.g., $\Lambda_{\max}$ is the largest eigenvalue of the empirical Gram matrix $\hat{\Sigma} := \mathbf{X}^T\mathbf{X}/n$, and $\phi^2(L, S, N)$ is the (L, S, N)-restricted eigenvalue of $\hat{\Sigma}$. The projections in $L_2(Q_n)$ are also written as before, i.e.

$$\mathrm{f}_S := \mathbf{X}b^S := \arg\min_{f = \mathbf{X}\beta_S} \|f - \mathbf{f}^0\|_n.$$

The ℓ_0-sparse projection $\mathrm{f}_{S_0} = \sum_{j\in S_0} b_j^0$ is now defined with a larger constant (7 instead of 3) in front of the penalty term, and a larger constant ($L = 6$ instead of $L = 2$) in the restrictions of the restricted eigenvalue condition:

$$S_0 := \arg\min_{S \subset S_{\mathrm{true}}} \left\{ \|\mathrm{f}_S - \mathbf{f}^0\|_n^2 + \frac{7\lambda_{\mathrm{init}}^2 |S|}{\phi^2(6, S)} \right\}$$

(compare with formula (7.1)).

The weighted Lasso is

$$\hat{\beta}_{\mathrm{weight}} = \arg\min_{\beta} \left\{ \|\mathbf{Y} - f_\beta\|_n^2 + \lambda_{\mathrm{init}}\lambda_{\mathrm{weight}} \sum_{j=1}^p w_j|\beta_j| \right\}. \tag{8.1}$$

Let

$$\hat{f}_{\mathrm{weight}} := f_{\hat{\beta}_{\mathrm{weight}}},\ \hat{S}_{\mathrm{weight}} := \{j :\ \hat{\beta}_{j,\mathrm{weight}} \neq 0\}.$$

The initial and adaptive Lasso are defined as in Section 1. We write $\hat{f}_{\mathrm{init}} := f_{\hat{\beta}_{\mathrm{init}}}$ and $\hat{f}_{\mathrm{adap}} := f_{\hat{\beta}_{\mathrm{adap}}}$, with active sets $\hat{S}_{\mathrm{init}} := \{j :\ \hat{\beta}_{j,\mathrm{init}} \neq 0\}$ and $\hat{S}_{\mathrm{adap}} := \{j :\ \hat{\beta}_{j,\mathrm{adap}} \neq 0\}$, respectively. Let

$$\hat{\delta}_{\mathrm{init}}^2 := \|f_{\beta_{\mathrm{init}}} - \mathbf{f}^0\|_n^2,$$

be the prediction error of the initial Lasso, and and, for $q \geq 1$,

$$\hat{\delta}_q := \|\hat{\beta}_{\mathrm{init}} - b^0\|_q$$

be its ℓ_q-error. Denote the prediction error of the adaptive Lasso by

$$\hat{\delta}_{\mathrm{adap}}^2 := \|f_{\hat{\beta}_{\mathrm{adap}}} - \mathbf{f}^0\|_n^2.$$

The least squares estimator using only variables in S is also written with a “hat”:

$$\hat{\mathrm{f}}_S = f_{\hat{b}^S} := \arg \min_{f = f_{\beta_S}} \|\mathbf{Y} - f_{\beta_S}\|_n .$$

A threshold level will be denoted by δ, instead of λ_{thres} as we do in Section 1. The reason is again that we need to explicitly express dependence on the threshold level. We define, for any threshold $\delta > 0$,

$$\hat{S}_{\text{init}}^{\delta} := \{ j : \ |\hat{\beta}_{j,\text{init}}| > \delta \}.$$

The refitted version after thresholding, based on the data $\mathbf{Y}$, is $\hat{\mathrm{f}}_{\hat{S}_{\text{init}}^{\delta}}$.

To handle the (random) noise, we define the set

$$\mathcal{T} := \left\{ \max_{1 \le j \le p} 4 |(\epsilon, \psi_j)_n| \le \lambda_{\text{init}} \right\}.$$

This is the set where the (empirical) correlations between noise and design is “small”.

Here λ_{init} is chosen in such a way that

$$\mathbb{P}(\mathcal{T}) \ge 1 - \alpha$$

where $(1 - \alpha)$ is the confidence we want to achieve.

8.2. The noisy weighted Lasso

Theorem 8.1. *Suppose we are on $\mathcal{T}$. Let S be a set with cardinality $s = |S|$, which satisfies for some positive L and M*

$$\lambda_{\text{weight}} (w_{S^c}^{\min} \wedge M) \ge 1,$$

and

$$w_{S^c}^{\min} \ge M/L, \ \|w_S\|_2 / \sqrt{s} \le M.$$

Then for all β,

$$\|\hat{f}_{\text{weight}} - \mathbf{f}^0\|_n^2 \le 2 \|f_{\beta_S} - \mathbf{f}^0\|_n^2 + \frac{14 \lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 M^2 s}{\phi^2(6L, S)},$$

and

$$\sqrt{s} \|(\hat{\beta}_{\text{weight}})_S - \beta_S\|_2 + \|(\hat{\beta}_{\text{weight}})_{S^c}\|_1 / L \le \frac{5 \|f_{\beta_S} - \mathbf{f}^0\|_n^2}{\lambda_{\text{init}} \lambda_{\text{weight}} M} + \frac{7 \lambda_{\text{init}} \lambda_{\text{weight}} M s}{\phi^2(6L, S)},$$

and

$$\|\hat{\beta}_{\text{weight}} - \beta_S\|_2$$
$$\le \frac{10 L \|f_{\beta_S} - \mathbf{f}^0\|_n^2}{M \lambda_{\text{init}} \lambda_{\text{weight}} \sqrt{s_0}} + \frac{14 L \lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 M (s + s_0)}{\phi^2(6L, S, s + s_0) \lambda_{\text{init}} \lambda_{\text{weight}} \sqrt{s_0}}.$$

Moreover, under the condition $\lambda_{\rm weight} w_{S^c}^{\min} \ge 1$,

$$|(\hat{S}_{\rm weight} \cap S^c) \backslash S_0|^2$$

$$\le 16 \Lambda_{\max}^2 ((\hat{S}_{\rm weight} \cap S^c) \backslash S_0) \frac{\| \hat{f}_{\rm weight} - \mathbf{f}^0 \|_n^2}{\lambda_{\rm weight}^2} \frac{\| (1/w)_{\hat{S}_{\rm weight} \backslash S_0} \|_2^2}{\lambda_{\rm init}^2}.$$

When $|(\hat{S}_{\rm weight} \cap S^c) \backslash S_0| > s_0$, *this implies*

$$|(\hat{S}_{\rm weight} \cap S^c) \backslash S_0| \le 32 \Lambda_{\rm sparse}^2 (s_0) \frac{\| \hat{f}_{\rm weight} - \mathbf{f}^0 \|_n^2}{\lambda_{\rm weight}^2 s_0} \frac{\| (1/w)_{\hat{S}_{\rm weight} \backslash S_0} \|_2^2}{\lambda_{\rm init}^2}.$$

8.3. Another look at the number of false positives

Here, we discuss a refinement, assuming a condition corresponding to the one used in [34].

Condition D *It holds for some* $s_* \ge s_0$, *that*

$$D(s_*, s_0) := \left\{ \frac{\Lambda_{\rm sparse}^2 (s_*) s_0}{\phi^2(6, S_0) s_*} \right\} = O_{\rm suff}(1).$$

Lemma 8.1. *Suppose we are on* $\mathcal{T}$. *Then under Condition D,*

$$|\hat{S}_{\rm init} \backslash S_0| = \left[\frac{\Lambda_{\rm sparse}^2 (s_*)}{\phi^2(6, S_0)} \right] \left(1 - \frac{D(s_*, s_0)}{O_{\rm suff}(1)} \right)^{-1} O(s_0).$$

Moreover, under Condition B,

$$|\hat{S}_{\rm adap} \backslash S_0| = \Lambda_{\rm sparse}(s_*) \left[\frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\min}(6, S_0, 2s_0) \phi^4(6, S_0, 2s_0)} \right]^{1/2} \sqrt{\frac{\lambda_{\rm init}}{\lambda_{\rm adap}}} O(s_0)$$

$$+ \left[\frac{\Lambda_{\rm sparse}(s_0) \phi^2(6, S_0)}{\phi_{\min}(6, S_0, 2s_0) \phi^4(6, S_0, 2s_0)} \right] D(s, s_*) \frac{\lambda_{\rm init}}{\lambda_{\rm adap}} O(s_0).$$

Under Condition BB, this becomes

$$|\hat{S}_{\rm adap} \backslash S_0| = \left[\frac{\Lambda_{\rm sparse}(s_*)}{\phi(6, S_0)} \right] \left[\frac{\phi_{\min}^2(6, S_0, 2s_0) \phi^2(6, S_0)}{\phi^2(6, S_0, 2s_0)} \right]^{1/2} O(s_0) \tag{8.2}$$

$$+ \left[\frac{\phi_{\min}^2(6, S_0, 2s_0) \phi^2(2, S_0)}{\phi^4(6, S_0, 2s_0)} \right] D(s_*, s_0) O(s_0).$$

Under Condition D, the first term in the right hand side of (8.2) is generally the leading term. We thus see the adaptive Lasso replaces the potentially very large constant

$$\left(1 - \frac{D(s_*, s_0)}{O_{\rm suff}(1)} \right)^{-1}$$

in the bound for the number of false positives of the initial Lasso by

$$\left[\frac{\phi^2_{\min}(6, S_0, 2s_0)\phi^2(6, S_0)}{\phi^4(6, S_0, 2s_0)} \right]^{1/2},$$

a constant which is close to 1 if the ϕ's do not differ too much.

Admittedly, Condition D is difficult to interpret. On the one hand, it wants s_* to be large, but on the other hand, a large s_* also can render $\Lambda_{\text{sparse}}(s_*)$ large. We refer to [34] for examples where Condition D is met.

9. Proofs

We present five subsections, containing respectively the proofs for Subsection 5.3.2, for Subsection 5.4, Section 7, Section 8, and finally Section 3.

9.1. Proofs for Subsection 5.3.2 with the realistic example giving a lower bound for the Lasso

Proof of Lemma 5.5. By the Karush-Kuhn-Tucker (KKT) conditions (see [3]), for $\psi = (\psi_1, \ldots, \psi_p)$,

$$\hat{\Sigma}(\hat{\beta}_{\text{init}} - \beta^0) = -\lambda_{\text{init}}\hat{\tau}_{\text{init}} + \psi^T \epsilon / n,$$

where $\|\hat{\tau}_{\text{init}}\|_\infty \le 1$ and for $j \in \hat{S}_{\text{init}}$, $\hat{\tau}_{j,\text{init}} = \text{sign}(\hat{\beta}_{j,\text{init}})$. It follows that

$$(1-\rho)(\hat{\beta}_{\text{init}} - \beta^0) = (1-\rho)\hat{\Sigma}^{-1}\Big(-\lambda_{\text{init}}\hat{\tau}_{\text{init}} + \psi^T \epsilon / n \Big)$$

$$= -\lambda_{\text{init}}\hat{\tau}_{\text{init}} + \psi^T \epsilon / n + a\iota,$$

with

$$a = \frac{\rho}{1 - \rho + \rho p}\Big(\lambda_{\text{init}} \iota^T \hat{\tau}_{\text{init}} - \iota^T \psi^T \epsilon / n \Big).$$

Because β^0_j is positive and sufficiently large, we know that all $\hat{\beta}_{j,\text{init}}$ with $j \in S_{\text{true}}$ are strictly positive.

We now show that when $a \ge 0$, then all $\hat{\beta}_{j,\text{init}}$ with $j \in \hat{S}_{\text{init}} \backslash S_{\text{true}}$ are positive. Fix some $j \in \hat{S}_{\text{init}} \backslash S_{\text{true}}$. We then have $|\hat{\tau}_{j,\text{init}}| = 1$. If $\hat{\beta}_{j,\text{init}} < 0$, we must have $\hat{\tau}_{j,\text{init}} = -1$. So we get

$$0 > (1-\rho)\hat{\beta}_{j,\text{init}} = \lambda_{\text{init}} + \psi_j^T \epsilon / n + a \ge a.$$

When $a \ge 0$ this is a contradiction.

Similarly, suppose that $\hat{\tau}_{j,\text{init}} = 1$ for some $j \in S^c_{\text{true}}$. Then

$$0 < (1-\rho)\hat{\beta}_{j,\text{init}} = -\lambda_{\text{init}} + \psi^T \epsilon / n + a \le a.$$

So then $a > 0$. It follows that a can only be negative if all $\hat{\beta}_{j,\text{init}}$ with $j \notin S_\text{true}$ are negative.

Let us consider the case $a < 0$ further, and show it cannot be. Write

$$w_j = \lambda_\text{init} + \epsilon^T \psi_j / n.$$

Because $|\epsilon^T \psi_j|/n < \lambda_\text{init}$ for all j, it holds that $w_j > 0$. Furthermore,

$$a = \frac{\rho}{1 - \rho + \rho \hat{s}} \Bigg[\sum_{j \in S_\text{true}} w_j - \sum_{j \in \hat{S} \backslash S_\text{true}} w_j \Bigg].$$

For $j \in \hat{S}_\text{init} \backslash S_\text{true}$

$$0 > (1 - \rho) \hat{\beta}_{j,\text{init}} = w_j + a,$$

so that

$$0 > \sum_{j \in \hat{S} \backslash S_\text{true}} w_j + \frac{\rho(\hat{s} - s_\text{true})}{1 - \rho + \rho \hat{s}} \Bigg[\sum_{j \in S_\text{true}} w_j - \sum_{j \in \hat{S} \backslash S_\text{true}} w_j \Bigg]$$

$$= \frac{\rho}{1 - \rho + \rho \hat{s}} \Bigg[(\hat{s} - s_\text{true}) \sum_{j \in S_\text{true}} w_j + (1 - \rho + \rho s) \sum_{j \in \hat{S} \backslash S_\text{true}} w_j \Bigg] \geq 0.$$

This is a contradiction, and hence $a \geq 0$. The result now follows from writing down the KKT solution given that the $\hat{\beta}_j$'s are all strictly positive for $j \in \hat{S}_\text{init}$ (and zero outside $\hat{S}_\text{init}$). □

We now consider more precisely to what extent false positives contribute to a better fit to the data.

Lemma 9.1. *Assume the conditions of Lemma 5.5. Define for all j,*

$$w_j := \lambda_\text{init} - \epsilon^T \psi_j / n,$$

and for all index sets S,

$$\bar{w}_S := \sum_{j \in S} w_j / |S|.$$

Then on $\mathcal{T}$, $\hat{S}_\text{init} \backslash S_\text{true}$ maximizes over all $\mathcal{N} \subset S_\text{true}^c$, the expression

$$\sum_{j \in \mathcal{N}} (w_j - \bar{w}_\mathcal{N})^2$$

$$+ \frac{(1 - \rho) s_\text{true} \bar{w}_{S_\text{true}}^2 + (1 - \rho) N \bar{w}_\mathcal{N}^2 + \rho N s_\text{true} (\bar{w}_{S_\text{true}} - \bar{w}_\mathcal{N})^2}{1 - \rho + \rho(s_\text{true} + N)},$$

under the restriction that for all $j \in \mathcal{N}$,

$$\rho s_\text{true} \bar{w}_{S_\text{true}} - (1 - \rho + \rho s_\text{true}) \bar{w}_\mathcal{N} > (1 - \rho + \rho(s_\text{true} + N))(w_j - \bar{w}_\mathcal{N}). \quad (9.1)$$

Here $N := |\mathcal{N}|$.
Alternatively, on $\mathcal{T}$, $\hat{S}_{\text{init}} \backslash S_{\text{true}}$ *maximizes over all* $\mathcal{N} \subset S_{\text{true}}$ *the expression*

$$+\frac{1-\rho+\rho s_{\text{true}}}{\rho}\left[1-\frac{1-\rho+\rho s_{\text{true}}}{1-\rho+\rho(s_{\text{true}}+N)}\right](\bar{w}_{S_{\text{true}}}-\bar{w}_{\mathcal{N}})^2+\sum_{j\in\mathcal{N}}(w_j-\bar{w}_{\mathcal{N}})^2$$
$$-\frac{(1-\rho)(N-s_{\text{true}})\bar{w}^2_{S_{\text{true}}}-2(1-\rho)N\bar{w}_{S_{\text{true}}}\bar{w}_{\mathcal{N}}}{1-\rho+\rho(s_{\text{true}}+N)} \tag{9.2}$$

under the restriction (9.1) for all $j \in \mathcal{N}$.

We note that restriction (9.1) is needed to ensure that the solution (5.2) for $\hat{\beta}_{j,\text{init}}$ is indeed strictly positive.

Proof of Lemma 9.1. Let us use the short-hand notation $S := S_{\text{true}}$, $s = s_{\text{true}}$, $\hat{S} = \hat{S}_{\text{init}}$, $\hat{\mathcal{N}} := \hat{S}\backslash S$, and $\hat{N} = |\hat{\mathcal{N}}|$, and write $\lambda := \lambda_{\text{init}}$, and $\hat{\beta} = \hat{\beta}_{\text{init}}$. Moreover, let

$$\bar{w}_1 := \bar{w}_S,\ \ \bar{w}_2 := \bar{w}_{\hat{\mathcal{N}}}.$$

By definition, $\hat{\beta}$ has

$$\|\mathbf{Y}-\mathbf{X}\hat{\beta}\|_2^2/n+2\lambda\|\hat{\beta}\|_1-2\lambda\|\beta^0\|_1-\|\epsilon\|_2^2/n$$

at its minimum value. We consider this expression now in detail. First note that on $\mathcal{T}$, it is equal to

$$(\hat{\beta}_{\hat{S}}-\beta^0_{\hat{S}})^T\hat{\Sigma}_{1,1}(\hat{S})(\hat{\beta}_{\hat{S}}-\beta^0_{\hat{S}})-2\epsilon^T\mathbf{X}(\hat{\beta}-\beta^0)$$
$$+2\lambda\|\hat{\beta}\|_1-2\lambda\|\beta^0\|_1,$$

since by Lemma 5.5, $\hat{S} \supset S$. Here, we use the notation $\hat{\beta}_{\hat{S}} = \{\hat{\beta}_j\}_{j\in\hat{S}}$ and $\beta^0_{\hat{S}} = \{\beta^0_j\}_{j\in\hat{S}}$.

Also by Lemma 5.5, on $\mathcal{T}$,

$$\hat{\beta}_{\hat{S}}-\beta^0_{\hat{S}} = -\hat{\Sigma}^{-1}_{1,1}(\hat{S})w_{\hat{S}},$$

where $w_{\hat{S}} = \{w_j\}_{j\in\hat{S}}$. So

$$\hat{\Sigma}_{1,1}(\hat{S})(\hat{\beta}_{\hat{S}}-\beta^0_{\hat{S}}) = -w_{\hat{S}},$$

and therefore

$$(\hat{\beta}_{\hat{S}}-\beta^0_{\hat{S}})^T\hat{\Sigma}_{1,1}(\hat{S})(\hat{\beta}_{\hat{S}}-\beta^0_{\hat{S}}) = w^T_{\hat{S}}\hat{\Sigma}^{-1}_{1,1}(\hat{S})w_{\hat{S}}.$$

We further see that on $\mathcal{T}$,

$$2\epsilon^T\mathbf{X}(\hat{\beta}-\beta^0) = 2\sum_{j\in\hat{S}}(\epsilon^T\psi_j/n)(-w_j+a).$$

Since $\hat{\beta}_j \geq 0$,

$$\|\hat{\beta}\|_1 = \sum_{j \in \hat{S}} \hat{\beta}_j .$$

Similarly

$$\|\beta^0\|_1 = \sum_{j \in \hat{S}} \beta_j^0 ,$$

where we again used $\hat{S} \supset S$. Therefore

$$-2\epsilon^T \mathbf{X}(\hat{\beta} - \beta^0) + 2\lambda \|\hat{\beta}\|_1 - 2\lambda \|\beta^0\|_1$$

$$= -2 \sum_{j \in \hat{S}} (\epsilon^T \psi_j / n)(\hat{\beta}_j - \beta_j^0) + 2\lambda \sum_{j \in \hat{S}} (\hat{\beta}_j - \beta_j^0)$$

$$= 2 \sum_{j \in \hat{S}} w_j (\hat{\beta}_j - \beta_j^0) = 2 w_{\hat{S}}^T (\hat{\beta}_{\hat{S}} - \beta_{\hat{S}}^0) = -2 w_{\hat{S}}^T \Sigma_{1,1}^{-1}(\hat{S}) w_{\hat{S}} .$$

We thus derived that

$$\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2/n + 2\lambda \|\hat{\beta}\|_1 - 2\lambda \|\beta^0\|_1 - \|\epsilon\|_2^2/n = -w_{\hat{S}}^T \Sigma_{1,1}^{-1}(\hat{S}) w_{\hat{S}}$$

$$= -\frac{1}{1-\rho} \left[\sum_{j \in \hat{S}} w_j^2 - \frac{\rho}{1 - \rho + \rho \hat{s}} (s \bar{w}_1 + \hat{N} \bar{w}_2)^2 \right]. \tag{9.3}$$

Write

$$\sum_{j \in \hat{S}} w_j^2 = \sum_{j \in S} (w_j - \bar{w}_1)^2 + \sum_{j \in \hat{\mathcal{N}}} (w_j - \bar{w}_2)^2 + s \bar{w}_1^2 + \hat{N} \bar{w}_2^2 .$$

The expression in the square brackets in (9.3) then becomes

$$\sum_{j \in S} (w_j - \bar{w}_1)^2 + \sum_{j \in \hat{\mathcal{N}}} (w_j - \bar{w}_2)^2 + s \bar{w}_1^2 + \hat{N} \bar{w}_2^2 - \frac{\rho}{1 - \rho + \rho \hat{s}} (s \bar{w}_1 + \hat{N} \bar{w}_2)^2$$

$$= \sum_{j \in S} (w_j - \bar{w}_1)^2 + \sum_{j \in \hat{\mathcal{N}}} (w_j - \bar{w}_2)^2$$

$$+ \frac{(1 - \rho + \rho(s + \hat{N})) s \bar{w}_1^2 + (1 - \rho + \rho(s + \hat{N})) \hat{N} \bar{w}_2^2 - \rho(s^2 \bar{w}_1^2 + \hat{N}^2 \bar{w}_2^2 + 2 s \hat{N} \bar{w}_1 \bar{w}_2)}{1 - \rho + \rho(s + \hat{N})}$$

$$= \sum_{j \in S} (w_j - \bar{w}_1)^2 + \sum_{j \in \hat{\mathcal{N}}} (w_j - \bar{w}_2)^2$$

$$+ \frac{(1 - \rho) s \bar{w}_1^2 + (1 - \rho) \hat{N} \bar{w}_2^2 + \rho s \hat{N} (\bar{w}_1 - \bar{w}_2)^2}{1 - \rho + \rho \hat{s}} .$$

This proves the first result.

For the second result, we introduce

$$\bar{w}_0 := \bar{w}_1 - \bar{w}_2.$$

Then

$$\frac{(1-\rho)s\bar{w}_1^2 + (1-\rho)\hat{N}\bar{w}_2^2 + \rho s\hat{N}(\bar{w}_1 - \bar{w}_2)^2}{1-\rho+\rho\hat{s}}$$

$$= \frac{(1-\rho)\hat{s}\bar{w}_1^2 - 2(1-\rho)\hat{N}\bar{w}_1\bar{w}_0 + (1-\rho+\rho s)\hat{N}w_0^2}{1-\rho+\rho\hat{s}}$$

$$= \frac{(1-\rho)\hat{s}\bar{w}_1^2 - 2(1-\rho)\hat{N}\bar{w}_1\bar{w}_0}{1-\rho+\rho\hat{s}} + \frac{1-\rho+\rho s}{\rho}\left[1 - \frac{1-\rho+\rho s}{1-\rho+\rho\hat{s}}\right]\bar{w}_0^2.$$

□

From Lemma 5.5, in particular the restriction (9.1) for all $j \in \mathcal{N}$, it can be seen that if there are false positives j, the corresponding error terms $\epsilon^T\psi_j/n$ cannot vary much, i.e., they have to be concentrated around their average. We will consider two types of concentration.

Definition 1. Let $v_1, \ldots, v_m$ be real-valued random variables. We say that a probability concentration inequality holds for $v_1, \ldots, v_m$ if for some constant b_m, and constants u and $0 < \theta < 1$ not depending on m, one has

$$\mathbb{P}\left(\max_{1\le j\le m} v_j \in [b_m - u, b_m + u]\right) > \theta.$$

The value $u = u_m$ can also be taken dependent on m, with $u_m \to 0$ as $m \to \infty$. Then possibly $\theta = \theta_m$ will also tend to zero. This leads to the following.

Definition 2. Let $v_1, \ldots, v_m$ be real-valued random variables. We say that a density concentration inequality holds for $v_1, \ldots, v_m$ if for some constant b_m, $u_m > 0$ and θ_m, one has

$$\mathbb{P}(\max_{1\le j\le m} v_j \in [b_m - u_m, b_m + u_m]) \ge \theta_m.$$

Lemma 9.2. *For $m \le n-1$ sufficiently large, and $0 < \theta < 1$ and $t > 0$ not depending on m, the following concentration inequality holds for $v_1, \ldots, v_m$ with $v_j = \epsilon^T\tilde{\psi}_j/\sqrt{n}$,*

$$b_m = \sqrt{2\log m} - \frac{\log\log m + \log(4\pi)}{2\sqrt{2\log m}},$$

and $u_m = t/\sqrt{2\log m}$.

Proof of Lemma 9.2. We note that $v_1, \ldots, v_m$ are i.i.d. $\mathcal{N}(0,1)$-distributed. The lemma follows from a result from extreme value theory, which says that for $v_1, \ldots, v_m, \ldots$ a sequence of i.i.d. $\mathcal{N}(0,1)$-distributed random variables,

$$\sqrt{2\log m}\left(\max_{1\le j\le m} v_j - \sqrt{2\log m} + \frac{\log\log m + \log(4\pi)}{2\sqrt{2\log m}}\right)$$

converges weakly to a Gumbel distribution, see for example [14]. □

We get more refined results if in fact u_m decreases faster than $t/\sqrt{2\log m}$, when $\theta = \theta_m$ does not decrease fast. We do not elaborate on this.

By repeated application of the concentration result, one can show that with positive probability, there are many $v_j := \epsilon^T \tilde{\psi}_j/\sqrt{n}$, $j \in S_{\rm true}^c$, which are almost as large as $\max_{j\in S_{\rm true}^c} v_j$. We call this an applied concentration result. The idea is to divide the set $S_{\rm true}^c$ into $2N$ sets $S_1, \ldots, S_{2N}$ of size $\asymp n/N$. Within each set S_k, we apply Lemma 9.2 to the random variables v_j, $j \in S_k$. We then get $2N$ random variables $\max_{j\in S_k} v_j$. With positive probability, at least half of them are in the set $[b_m - u_m, b_m + u_m]$, where $m \asymp n/N$ and b_m and u_m are as in Lemma 9.2.

Applied Concentration Result *Define $n_0 := n - 1 - s_{\rm true}$. Let $N \le n_0/2$ be an integer with*

$$\log\lfloor n_0/(2N)\rfloor \ge \frac{1}{2}\log n.$$

Let $v_j := \epsilon^T \tilde{\psi}_j/\sqrt{n}$, $j = 1, \ldots, n-1$, and $v_n := \epsilon^T z/\sqrt{n}$. Define

$$c_n := \sqrt{2(1-\rho)\log\lfloor n_0/(2N)\rfloor} - \frac{(1-\rho)(\log\log\lfloor n_0/(2N)\rfloor + \log(4\pi))}{2\sqrt{2\log\lfloor n_0/(2N)\rfloor}} + \sqrt{\rho}v_n.$$

There exist constants $C_2 > 0$ and $0 < \alpha_0 < 1$ not depending on n, such that for $t_n = C_2/\sqrt{\log n}$, there is with probability at least α_0 as set $\mathcal{N} \subset S_{\rm rue}^c$ with cardinality N, such that for all $j \in \mathcal{N}$,

$$c_n - t_n \le \epsilon^T \psi_j/\sqrt{n} \le c_n + t_n. \tag{9.4}$$

We are now ready to show that there are sets $\mathcal{N}$ that satisfy the restriction (9.1) for all $j \in \mathcal{N}$.

Lemma 9.3. *Let $\lambda_\epsilon = \lambda_{\epsilon,n}$, be a positive constant depending on n, and C_1, C_3, C_4 be positive constants such that with probability at least $1 - \alpha_1$, where $\alpha_1 < \alpha_0$, with α_0 given in the Applied Concentration Result,*

$$\max_{1\le j\le p} |\epsilon^T \psi_j|/n \le \lambda_\epsilon, \quad \left|\sum_{j\in S_{\rm true}} \epsilon^T \psi_j/n\right| \le \frac{1}{2(1+C_4)}\lambda_{\rm init} s_{\rm true}.$$

Take

$$\lambda_{\rm init} \ge C_1 \lambda_\epsilon.$$

Define $n_0 := n - 1 - s_{\rm true}$. Let $N \le n_0/2$ be an integer with

$$\log\lfloor n_0/(2N)\rfloor \ge \frac{1}{2}\log n.$$

and

$$N < \frac{s_{\rm true}\sqrt{n\log n}\lambda_{\rm init}}{4(1+C_4)C_2}.$$

Assume moreover that for t_n and c_n given in the Applied Concentration Result with such a value for N, with probability at least $1 - \alpha_2$, where $\alpha_2 < \alpha_0 - \alpha_1$,

$$c_n - t_n \ge \sqrt{n}\lambda_{\rm init}/C_3,$$

and that

$$\frac{\rho s_{\rm true}}{1-\rho+\rho s_{\rm true}} > \biggl(\frac{1+C_1}{C_1}\biggr)\biggl(\frac{C_3-1}{C_3}\biggr).$$

Assume finally that $\beta_j^0 > C\lambda_{\rm init}$ for all $j \in S_{\rm true}$ with C a sufficiently large positive constant depending only on C_1. Then with probability at least $\alpha_0-\alpha_1-\alpha_2$, there is a set $\mathcal{N} \supset S_{\rm true}$ with size N satisfying the restriction (9.1) of Lemma 9.1 for all $j \in \mathcal{N}$.

Proof of Lemma 9.3. We have to show that for all $j \in \mathcal{N}$,

$$\epsilon^T\psi_j/n - \lambda_{\rm init}$$
$$+\frac{\rho}{1-\rho+\rho(s_{\rm true}+N)}\biggl(\lambda_{\rm init}(N+s_{\rm true}) - \sum_{j\in\mathcal{N}\cup S_{\rm true}}\epsilon^T\psi_j/n\biggr)$$

is strictly positive. That is, we have to show that for all such j,

$$B_j := \epsilon^T\psi_j/n - \lambda_{\rm init}$$
$$+\frac{\rho}{1-\rho+\rho(N+s_{\rm true})}\biggl(\lambda_{\rm init}s_{\rm true} - \sum_{j\in S_{\rm true}}\epsilon^T\psi_j/n + \lambda_{\rm init}N - \sum_{j\in\mathcal{N}}\epsilon^T\psi_j/n\biggr)$$

is strictly positive. Inserting the assumed bounds and the Applied Concentration Result gives

$$B_j \ge [c_n/\sqrt{n} - t_n/\sqrt{n} - \lambda_{\rm init}]$$
$$+\frac{\rho\bigl(\lambda_{\rm init}s_{\rm true} - \frac{1}{2(1+C_4)}\lambda_{\rm init}s_{\rm true} + N[\lambda_{\rm init} - c_n/\sqrt{n} + t_n/\sqrt{n}]\bigr)}{1-\rho+\rho(N+s_{\rm true})}$$
$$-\frac{2\rho N t_n/\sqrt{n}}{1-\rho+\rho(N+s_{\rm true})}.$$

This can be reorganized to

$$\Bigl(1-\rho+\rho(N+s_{\rm true})\Bigr)B_j \ge \lambda_{\rm init}\rho s_{\rm true} - (1-\rho+\rho s_{\rm true})[\lambda_{\rm init} - c_n/\sqrt{n} + t_n/\sqrt{n}]$$
$$-\frac{\rho}{2(1+C_4)}\lambda_{\rm init}s_{\rm true} - 2\rho t_n N/\sqrt{n}$$
$$:= \lambda_{\rm init}\rho s_{\rm true} - (I + II + III),$$

where

$$I = (1-\rho+\rho s_{\rm true})[\lambda_{\rm init} - c_n/\sqrt{n} + t_n/\sqrt{n}],$$

and

$$II = \frac{\rho}{2(1+C_4)}\lambda_{\rm init}s_{\rm true},$$

and

$$III = 2\rho t_n N/\sqrt{n}.$$

Clearly,

$$\lambda_{\text{init}}\rho s_{\text{true}} - (I + II + III) \geq \lambda_{\text{init}}\rho s_{\text{true}} - (1 + C_4)\max\{I/C_4, II + III\},$$

Now we verify that

$$\lambda_{\text{init}}\rho s_{\text{true}} > (1 + C_4)\max\{I/C_4, II + III\}.$$

Since $II + III < 2II$ if $II > III$ and else $II + III \leq 2III$, it is enough to show that

$$\lambda_{\text{init}}\rho s_{\text{true}} > \frac{1 + C_4}{C_4} I = \frac{1 + C_4}{C_4}(1 - \rho + \rho s_{\text{true}})[\lambda_{\text{init}} - c_n/\sqrt{n} + t_n/\sqrt{n}],$$

and

$$\lambda_{\text{init}}\rho s_{\text{true}} \geq 2(1 + C_4)II = \lambda_{\text{init}}\rho s_{\text{true}},$$

and

$$\lambda_{\text{init}}\rho s_{\text{true}} > 2(1 + C_4)III = 4(1 + C_4)\rho t_n N/\sqrt{n}.$$

The first follows from $c_n/\sqrt{n} - t_n/\sqrt{n} \geq \lambda_{\text{init}}/C_3$, and our condition

$$\frac{\rho s_{\text{true}}}{1 - \rho + \rho s_{\text{true}}} > \left(\frac{1 + C_4}{C_4}\right)\left(\frac{C_3 - 1}{C_3}\right).$$

The second is immediate. For the last one, we use

$$N < s_{\text{true}} \frac{\sqrt{n}\lambda_{\text{init}}}{4(1 + C_4)t_n}.$$

□

Finally, we have all ingredients to prove the main result of Subsection 5.3.2.

Proof of Theorem 5.1. Let $\hat{\mathcal{N}} := \hat{S}_{\text{init}} \backslash S_{\text{true}}$. We write (9.2) with general $\mathcal{N} \subset S_{\text{true}}$ with cardinality N, as $I + II + III$. So $\hat{\mathcal{N}}$ is a maximizer of $I + II + III$.

The first term is

$$I = \frac{1 - \rho + \rho s_{\text{true}}}{\rho}\left[1 - \frac{1 - \rho + \rho s_{\text{true}}}{1 - \rho + \rho(s_{\text{true}} + N)}\right](\bar{w}_{S_{\text{true}}} - \bar{w}_{\mathcal{N}})^2.$$

We argue as follows. It holds that $\bar{w}_{S_{\text{true}}} - \lambda_{\text{init}} = O(1/\sqrt{n})$. Moreover, in view of the concentration results, $\bar{w}_{\mathcal{N}} - \lambda_{\text{init}}$ can be as large as $\asymp \sqrt{(\log n - \log N)/n}$. With such a $\bar{w}_{\mathcal{N}}$, we get

$$(\bar{w}_{S_{\text{true}}} - \bar{w}_{\mathcal{N}})^2 \asymp \left(\frac{\log n - \log N}{n}\right).$$

The expression

$$\left[1 - \frac{1 - \rho + \rho s_{\text{true}}}{1 - \rho + \rho(s_{\text{true}} + N)}\right]\left(\frac{\log n - \log N}{n}\right)$$

is maximized as a function of N for $N \asymp s_{\text{true}} \log n$. The first term with this maximizing value is $\asymp s_{\text{true}} \log n/n$.

The second term $II = \sum_{j \in \mathcal{N}} (w_j - \bar{w}_{\mathcal{N}})^2$ will generally be small when $\mathcal{N}$ is small. Note that $n \sum_{j \in \mathcal{N}} (w_j - \bar{w}_{\mathcal{N}})^2/(1-\rho)$ is a χ^2-random variable with $N-1$ degrees of freedom. Therefore, it is not difficult to show that

$$\sum_{j \in \mathcal{N}} (w_j - \bar{w}_{\mathcal{N}})^2 \leq \frac{(N-1)(1-\rho)}{n} + O_{\mathbb{P}}\left(\frac{\sqrt{|N| \log p}}{n} \right),$$

uniformly over all $\mathcal{N} \subset S_{\text{true}}^c$. Hence, if N is small, it will also be small (and if N is large there is chance that it will be large). We see that with values of $N = |\mathcal{N}|$ substantially smaller than $s_{\text{true}} \log n$ the second term will be substantially smaller than $s_{\text{true}} \log n/n$. Therefore, there is no gain in the second term by choosing a smaller value than $\asymp s_{\text{true}} \log n$ for N.

As for the third term

$$III = -\frac{(1-\rho)(N - s_{\text{true}})\bar{w}_{S_{\text{true}}}^2 - 2(1-\rho)N \bar{w}_{S_{\text{true}}} \bar{w}_{\mathcal{N}}}{1 - \rho + \rho(s_{\text{true}} + N)},$$

this is of order

$$O(\bar{w}_{S_{\text{true}}}^2 + |\bar{w}_{S_{\text{true}}} \bar{w}_{\mathcal{N}}|).$$

But both $|\bar{w}_{S_{\text{true}}}|$ as well as $|\bar{w}_{\mathcal{N}}|$ are with large probability at most $\lambda_{\text{init}} \asymp \sqrt{\log n/n}$. Hence, when s_{true} is large enough, this will not be the dominant term.

Thus, one sees that the overall maximizer $\hat{\mathcal{N}}$ will choose N large, whenever feasible. As shown in Lemma 9.3, for an appropriate C_8, values

$$N = \frac{s_{\text{true}} \log n}{C_8}$$

are with positive probability indeed feasible. □

9.2. *Proofs for Subsection 5.4 on the weighted irrepresentable condition*

Proof of Lemma 5.6. This is the weighted variant of the first part of Lemma 6.2 in [29]. □

Proof of Lemma 5.7. We define, as in [29], the *adaptive restricted regression*

$$\vartheta_{\text{adaptive}}(S) := \max_{\beta \in \mathcal{R}(1, S, |S|)} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|f_{\beta_S}\|^2}.$$

Here, $(f, \tilde{f})$ denotes the inner product between f and $\tilde{f}$ as elements of $L_2(Q)$.

We will show that

$$\sup_{\|\tau_S\|_\infty \leq 1} \|W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S\|_\infty \leq \frac{\|w_S\|_2}{\sqrt{|S|} w_{S^c}^{\min}} \vartheta_{\text{adaptive}}(S). \tag{9.5}$$

It is moreover not difficult to see that $\vartheta_{\rm adaptive}(S) \le \sqrt{|S|}/\Lambda_{\min}(S)$, so then the proof of Lemma 5.7 is done.

Define

$$\beta_S := \Sigma_{1,1}^{-1}(S) W_S \tau_S .$$

Then

$$\| W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S \|_\infty = \sup_{\|\gamma_{S^c}\|_1 \le 1} |\gamma_{S^c}^T W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S|$$

$$= \sup_{\|W_{S^c}\beta_{S^c}\|_1 \le 1} |\beta_S^T \Sigma_{2,1}(S) \beta_S| = \sup_{\|W_{S^c}\beta_{S^c}\|_1 \le 1} |(f_{\beta_{S^c}}, f_{\beta_S})|$$

$$\le \sup_{\|\beta_{S^c}\|_1 \le 1/w_{S^c}^{\min}} |(f_{\beta_{S^c}}, f_{\beta_S})|$$

$$= \sup_{\|\beta_{S^c}\|_1 \le \|w_S\|_2 \|\beta_S\|_2 / w_{S^c}^{\min}} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|w_S\|_2 \|\beta_S\|_2}$$

$$= \sup_{\|\beta_{S^c}\|_1 \le \|w_S\|_2 \|\beta_S\|_2 / w_{S^c}^{\min}} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|f_{\beta_S}\|^2} \frac{\|f_{\beta_S}\|^2}{\|w_S\|_2 \|\beta_S\|_2} .$$

But

$$\frac{\|f_{\beta_S}\|^2}{\|w_S\|_2 \|\beta_S\|_2} = \frac{\tau_S^T W_S \Sigma_{1,1}^{-1}(S) W_S \tau_S}{\sqrt{\tau_S^T W_S^2 \tau_S} \sqrt{\tau_S^T W_S \Sigma_{1,1}^{-2}(S) W_S \tau_S}} \frac{\|W_S \tau_S\|_2}{\|w_S\|_2} \le 1.$$

We conclude that

$$\| W_{S^c}^{-1} \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) W_S \tau_S \|_\infty \le \sup_{\|\beta_{S^c}\|_1 \le \|w_S\|_2 \|\beta_S\|_2 / w_{S^c}^{\min}} \frac{|(f_{\beta_{S^c}}, f_{\beta_S})|}{\|f_{\beta_S}\|^2}$$

$$= \frac{\|w_S\|_2}{\sqrt{|S|} w_{S^c}^{\min}} \vartheta_{\rm adaptive}(S).$$

□

9.3. Proofs for Section 7: the noiseless case

9.3.1. Proofs for Subsection 7.1: the noiseless weighted Lasso

Proof of Theorem 7.1. Take

$$w_{S^c}^{\min} \ge M/L, \ \|w_S\|_2 / \sqrt{s} \le M.$$

We have

$$\|f_{\rm weight} - \mathbf{f}^0\|^2 + \lambda_{\rm init} \lambda_{\rm weight} \sum_{j=1}^p w_j |\beta_{j,{\rm weight}}|$$

$$\leq \|f_{\beta_S} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \in S} w_j |\beta_j|,$$

and hence

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{S^c}\|_1$$
$$\leq \|f_{\beta_S} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} \sum_{j \in S} w_j |\beta_{j,\text{weight}} - \beta_j|$$
$$\leq \|f_{\beta_S} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{s} \|(\beta_{\text{weight}})_S - \beta_S\|_2.$$

Let $\mathcal{N} \supset S$, $|\mathcal{N}| = N$. Then

$$\|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1 \leq \|(\beta_{\text{weight}})_{S^c}\|_1,$$

and

$$\|(\beta_{\text{weight}})_S - \beta_S\|_2 \leq \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2, \ \sqrt{s} \leq \sqrt{N}.$$

Therefore,

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1$$
$$\leq \|f_{\beta_S} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2.$$

Case i). If

$$\|f_{\beta_S} - \mathbf{f}^0\|^2 \leq \lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2,$$

we get

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1 \tag{9.6}$$
$$\leq 2\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2.$$

It follows that

$$\|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1 \leq 2L\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - (\beta)_S\|_2.$$

But then, by the definition of restricted eigenvalue, and invoking the triangle inequality,

$$\|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2 \leq \|f_{\text{weight}} - f_{\beta_S}\| / \phi(2L, \mathcal{N})$$
$$\leq \|f_{\text{weight}} - \mathbf{f}^0\| / \phi(2L, \mathcal{N}) + \|f_{\beta_S} - \mathbf{f}^0\| / \phi(2L, \mathcal{N}).$$

This gives

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1$$
$$\leq 2\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|f_{\text{weight}} - \mathbf{f}^0\| / \phi(2L, \mathcal{N})$$
$$+ 2\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|f_{\beta_S} - \mathbf{f}^0\| / \phi(2L, \mathcal{N})$$
$$\leq \frac{1}{2} \|f_{\text{weight}} - \mathbf{f}^0\|^2 + \|f_{\beta_S} - \mathbf{f}^0\|^2 + \frac{3\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 N M^2}{\phi^2(2L, \mathcal{N})}.$$

Hence,

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + 2\lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1$$
$$\leq 2\|f_{\beta_S} - \mathbf{f}^0\|^2 + \frac{6\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 N M^2}{\phi^2(2L, \mathcal{N})}.$$

Case ii) If

$$\|f_{\beta_S} - \mathbf{f}^0\|^2 > \lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2,$$

we get

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1 \leq 2\|f_{\beta_S} - \mathbf{f}^0\|^2.$$

The first result of the theorem now follows from taking $\mathcal{N} = S$.

For the second result, we add in Case i), $\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N}\|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2$ to the left and right hand side of (9.6):

$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + \lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2$$
$$+\lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1$$
$$\leq 3\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2.$$

The same arguments now give

$$\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1 \leq$$
$$\|f_{\text{weight}} - \mathbf{f}^0\|^2 + 3\|f_{\beta_S} - \mathbf{f}^0\|^2 + \frac{3\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 N M^2}{\phi^2(2L, \mathcal{N})}.$$

In Case ii), we have

$$\lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1 \leq 2\|f_{\beta_S} - \mathbf{f}^0\|^2,$$

and also

$$\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2 < \|f_{\beta_S} - \mathbf{f}^0\|^2.$$

So then

$$\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{N} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}} w_{S^c}^{\min} \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1$$
$$< 3\|f_{\beta_S} - \mathbf{f}^0\|^2.$$

Taking $\mathcal{N} = S$ gives the second result.

For the third result, we let $\mathcal{N}$ be the set S, complemented with the s_0 largest - in absolute value - coefficients of $(\beta_{\text{weight}})_{S^c}$. Then $\phi(2L, \mathcal{N}) \leq \phi(2L, S, s+s_0)$. Moreover, $N \geq s_0$. Thus, from the second result, we get

$$\lambda_{\text{init}}\lambda_{\text{weight}} M\sqrt{s_0} \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2 + \lambda_{\text{init}}\lambda_{\text{weight}} M \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_1 / L$$

$$\leq 3\|f_{\beta_S} - \mathbf{f}^0\|^2 + \frac{3\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 (s_0 + s) M^2}{\phi^2(2L, S, s + s_0)}.$$

Moreover, as is shown in Lemma 2.2 in [29] (with original reference [10], and [11]),

$$\|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_2 \leq \|(\beta_{\text{weight}})_{S^c}\|_1 / \sqrt{s_0}$$

$$\leq \frac{3L\|f_{\beta_S} - \mathbf{f}^0\|^2 + 3L\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 (s + s_0) M^2 / \phi^2(2L, S, s + s_0)}{\lambda_{\text{init}} \lambda_{\text{weight}} M \sqrt{s_0}}.$$

So then

$$\|\beta_{\text{weight}} - \beta_S\|_2 \leq \|(\beta_{\text{weight}})_{\mathcal{N}} - \beta_S\|_2 + \|(\beta_{\text{weight}})_{\mathcal{N}^c}\|_2$$

$$\leq \frac{6(L \vee 1)\|f_{\beta_S} - \mathbf{f}^0\|^2 + 6L\lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 (s + s_0) M^2 / \phi^2(2L, S, s + s_0)}{M\sqrt{s_0} \lambda_{\text{init}} \lambda_{\text{weight}}}.$$

□

We now turn to the proof of Lemma 7.1. An important characterization of the solution β_{weight} can be derived from the *Karush-Kuhn-Tucker* (*KKT*) conditions (see [3]).

Weighted KKT-conditions *We have*

$$2\Sigma(\beta_{\text{weight}} - \beta_{\text{true}}) = -\lambda_{\text{weight}} \lambda_{\text{init}} W \tau_{\text{weight}}.$$

Here, $\|\tau_{\text{weight}}\|_\infty \leq 1$*, and moreover*

$$\tau_{j,\text{weight}} \mathrm{l}\{\beta_{j,\text{weight}} \neq 0\} = \text{sign}(\beta_{j,\text{weight}}), \; j = 1, \ldots, p.$$

Proof of Lemma 7.1. By the weighted KKT conditions, for all j

$$2(\psi_j, f_{\text{weight}} - \mathbf{f}^0) = -\lambda_{\text{init}} \lambda_{\text{weight}} w_j \tau_{j,\text{weight}}.$$

Hence,

$$\sum_{j \in S_{\text{weight}} \setminus S_0} 4|(\psi_j, f_{\text{weight}} - \mathbf{f}^0)|^2 \geq \lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 \|w_{S_{\text{weight}} \setminus S_0}\|_2^2$$

$$\geq \lambda_{\text{init}}^2 \lambda_{\text{weight}}^2 |S_{\text{weight}} \setminus S_0|^2 / \|(1/w)_{S_{\text{weight}} \setminus S_0}\|_2^2.$$

On the other hand

$$\sum_{j \in S_{\text{weight}} \setminus S_0} |(\psi_j, f_{\text{weight}} - \mathbf{f}^0)|^2 \leq \Lambda_{\max}^2 (S_{\text{weight}} \setminus S_0) \|f_{\text{weight}} - \mathbf{f}^0\|^2.$$

Thus, we arrive at inequality (7.4):

$$|S_{\text{weight}} \setminus S_0|^2 \leq 4\Lambda_{\max}^2 (S_{\text{weight}} \setminus S_0) \frac{\|f_{\text{weight}} - \mathbf{f}^0\|^2}{\lambda_{\text{weight}}^2} \frac{\|1/w_{S_{\text{weight}} \setminus S_0}\|^2}{\lambda_{\text{init}}^2}.$$

Clearly,

$$\Lambda_{\max}^2 (S_{\text{weight}} \setminus S^0) \leq \Lambda_{\max}^2 \wedge \left(\frac{|S_{\text{weight}} \setminus S_0|}{s_0} + 1 \right) \Lambda_{\text{sparse}}^2 (s_0).$$

□

9.3.2. Proofs for Subsection 7.2: the noiseless initial Lasso

We first present the corollaries of Theorem 7.1 and Lemma 7.1 when we apply them to the case where all the weights are equal to one.

Corollary 9.1. *For the initial Lasso, $w_j = 1$ for all j, so we can apply Corollary 7.1 with $\delta = 1$ and $S^{\delta}_{\rm weight} = S_0$. Let*

$$\delta^2_{\rm oracle} := \|{\rm f}_{S_0} - \mathbf{f}^0\|^2 + \frac{3\lambda^2_{\rm init}|S_0|}{\phi^2(2, S_0)}.$$

We have

$$\delta^2_{\rm init} \le 2\|{\rm f}_{S_0} - \mathbf{f}^0\|^2 + \frac{6\lambda^2_{\rm init}|S_0|}{\phi^2(2, S_0)} = 2\delta^2_{\rm oracle}.$$

The estimation error can be bounded as follows:

$$\delta_1 \le 3\|{\rm f}_{S_0} - \mathbf{f}^0\|^2/\lambda_{\rm init} + \frac{3\lambda_{\rm init}|S_0|}{\phi^2(2, S_0)} \le 3\delta^2_{\rm oracle}/\lambda_{\rm init},$$

and

$$\delta_2 \le \left[\frac{\phi^2(2, S_0)}{\phi^2(2, S_0, 2s_0)}\right]\frac{6\delta^2_{\rm oracle}}{\lambda_{\rm init}\sqrt{s_0}}.$$

Moreover, application of Lemma 7.1 bounds the number of false positives:

$$|S_{\rm init}\backslash S_0| \le 4\Lambda^2_{\max}(S_{\rm init}\backslash S_0)\frac{\delta^2_{\rm init}}{\lambda^2_{\rm init}}.$$

Proof of Theorem 7.2. This is now a direct consequence of Corollary 9.1. □

9.3.3. Proofs for Subsection 7.3: the noiseless thresholded Lasso

We first provide some explicit bounds.

Lemma 9.4. *We have*

$$\|(\beta_{\rm init})_{S^{\delta}_{\rm init}} - b^0\|_1 \le 2\delta_1 + \delta s_0,$$

and

$$\|(\beta_{\rm init})_{S^{\delta}_{\rm init}} - b^0\|_2 \le 2\delta_2 + \delta\sqrt{s_0},$$

and

$$\|{\rm f}_{S^{\delta}_{\rm init}} - \mathbf{f}^0\| \le \|f_{(\beta_{\rm init})_{S^{\delta}_{\rm init}}} - \mathbf{f}^0\|$$

$$\le \|{\rm f}_{S_0} - \mathbf{f}^0\| + \sqrt{\left\lceil\frac{\delta_2^2}{\delta^2 s_0} + 1\right\rceil\Lambda_{\rm sparse}(s_0)(2\delta_2 + \delta\sqrt{s_0})},$$

and, for $\delta \ge \delta_2/\sqrt{s_0}$,

$$\|b^{S^{\delta}_{\rm init}} - b^0\|_2 \le \frac{\|{\rm f}_{S^{\delta}_{\rm init}} - \mathbf{f}^0\|}{\phi_{\rm sparse}(S_0, 2s_0)}.$$

Proof of Lemma 9.4. To obtain the first result, we use

$$\|(\beta_{\rm init})_{S_{\rm init}^\delta} - b^0\|_1 = \|(b^0 - \beta_{\rm init})_{S_{\rm init}^\delta}\|_1 + \|(b^0)_{S_0 \setminus S_{\rm init}^\delta}\|_1 .$$

Now,

$$\|(b^0 - \beta_{\rm init})_{S_{\rm init}^\delta}\|_1 \le \delta_1$$

Moreover

$$\begin{aligned}\|(b^0)_{S_0 \setminus S_{\rm init}^\delta}\|_1 &\le \|(b^0 - \beta_{\rm init})_{S_0 \setminus S_{\rm init}^\delta}\|_1 + \|(\beta_{\rm init})_{S_0 \setminus S_{\rm init}^\delta}\|_1 \\ &\le \|(b^0 - \beta_{\rm init})_{S_0 \setminus S_{\rm init}^\delta}\|_1 + \delta s_0 \le \delta_1 + \delta s_0 .\end{aligned}$$

Hence

$$\|(\beta_{\rm init})_{S_{\rm init}^\delta} - b^0\|_1 \le 2\delta_1 + \delta s_0 .$$

The ℓ_2-error of the second result follows by the same arguments.

The first inequality of the third result follows from the definition of $\mathrm{f}_{S_{\rm init}^\delta}$ as projection, and the second follows from the triangle inequality, where we invoke that

$$|S_{\rm init}^\delta \setminus S_0| \le \frac{\delta_2^2}{\delta^2}$$

so that

$$|S_{\rm init}^\delta| \le \frac{\delta_2^2}{\delta^2} + s_0 ,$$

and thus

$$\Lambda_{\max}^2(S_{\rm init}^\delta) \le \left[\frac{\delta_2^2}{\delta^2 s_0} + 1\right] \Lambda_{\rm sparse}^2(s_0).$$

The final result follows from

$$\Lambda_{\min}(S_{\rm init}^\delta \cup S_0) \ge \phi_{\rm sparse}(S_0, |S_{\rm init}^\delta \setminus S_0| + s_0) \ge \phi_{\rm sparse}(S_0, 2s_0).$$

□

Proof of Theorem 7.3. Inserting the bound $\delta_2 = O(\lambda_{\rm init}\sqrt{s_0}/\phi^2(2, S_0, 2s_0))$ (see Theorem 7.2), and $\|\mathrm{f}_{S_0} - \mathbf{f}^0\| = O(\lambda_{\rm init}\sqrt{s_0}/\phi^2(2, S_0))$, we get for $\lambda_{\rm init}/\phi^2(2, S_0) = O(\delta)$, $\delta \ge \delta_2/\sqrt{s_0}$,

$$\|f_{S_{\rm init}^\delta} - \mathbf{f}^0\|^2 = \Lambda_{\rm sparse}^2(s_0)\left[\frac{1}{\phi^4(2, S_0, 2s_0)} + \frac{\delta^2}{\lambda_{\rm init}^2}\right] O(\lambda_{\rm init}^2 s_0),$$

$$\|b^{S_{\rm init}^\delta} - b^0\|_2 = \frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\rm sparse}(S_0, 2s_0)} \times \left[\frac{1}{\phi^2(2, S_0, 2s_0)} + \frac{\delta}{\lambda_{\rm init}}\right] O(\lambda_{\rm init}\sqrt{s_0}),$$

and

$$|S_{\rm init}^\delta \setminus S_0| = \left[\frac{\lambda_{\rm init}^2}{\delta^2 \phi^4(2, S_0, 2s_0)}\right] O(s_0).$$

□

9.3.4. Proofs for Subsection 7.4: the noiseless adaptive Lasso

We use that when $\delta \ge \delta_2 / \sqrt{s_0}$, then $S_{\rm init}^{\delta} \backslash S_0 \le s_0$. Application of Corollary 7.1 then gives

Corollary 9.2. *We have, for all $\delta \ge \delta_2/\sqrt{s_0}$, and all β*

$$\delta_{\rm adap}^2 \le 2 \| f_{\beta_{S_{\rm init}^{\delta}}} - \mathbf{f}^0 \|^2 + \frac{12 \lambda_{\rm init}^2 \lambda_{\rm adap}^2 s_0}{\delta^2 \phi_{\rm min}^2 (2, S_0, 2s_0)},$$

and

$$\| \beta_{\rm adap} - \beta_{S_{\rm init}^{\delta}} \|_1 \le \frac{3\delta \| f_{\beta_{S_{\rm init}^{\delta}}} - \mathbf{f}^0 \|^2}{\lambda_{\rm init} \lambda_{\rm adap}} + \frac{6 \lambda_{\rm init} \lambda_{\rm adap} s_0}{\delta \phi_{\rm min}^2 (2, S_0, 2s_0)},$$

and

$$\| \beta_{\rm adap} - \beta_{S_{\rm init}^{\delta}} \|_2 \le \frac{6\delta \| f_{\beta_{S_{\rm init}^{\delta}}} - \mathbf{f}^0 \|^2}{\sqrt{s_0} \lambda_{\rm init} \lambda_{\rm adap}} + \frac{18 \lambda_{\rm init} \lambda_{\rm adap} \sqrt{s_0}}{\delta \phi_{\rm min}^2 (2, S_0, 3s_0)},$$

and, from Lemma 9.4,

$$\| f_{(\beta_{\rm init})_{S_{\rm init}^{\delta}}} - \mathbf{f}^0 \|^2 \le 2 \| \mathrm{f}_{S_0} - \mathbf{f}^0 \|^2 + 36 \Lambda_{\rm sparse}^2 (s_0) \delta^2 s_0 .$$

Furthermore, from Lemma 7.1 ,

$$|S_{\rm adap} \backslash S_0|^2 \le 4 \Lambda_{\rm max}^2 (S_{\rm adap} \backslash S_0) \frac{\delta_{\rm adap}^2}{\lambda_{\rm adap}^2} \frac{\delta_2^2}{\lambda_{\rm init}^2}.$$

If $|S_{\rm adap} \backslash S_0| > s_0$, we have

$$|S_{\rm adap} \backslash S_0| \le 8 \Lambda_{\rm sparse}^2 (s_0) \frac{\delta_{\rm adap}^2}{\lambda_{\rm adap}^2 s_0} \frac{\delta_2^2}{\lambda_{\rm init}^2} \wedge 2 \Lambda_{\rm max} \frac{\delta_{\rm adap}}{\lambda_{\rm adap}} \frac{\delta_2}{\lambda_{\rm init}}.$$

Remark 9.1. We note that in the above corollary, the use of the ℓ_2-error δ_2 is invoked for the variable selection result: with the weights $w_j = 1/|\beta_{j,\rm init}|$, we have

$$|S_{\rm adap} \backslash S_0|^2 \le \| w_{S_{\rm adap} \backslash S_0} \|_2^2 \| (1/w)_{S_{\rm adap} \backslash S_0} \|_2^2 \le \| w_{S_{\rm adap} \backslash S_0} \|_2^2 \delta_2^2 .$$

The theory can also be developed using only the ℓ_1-error δ_1, by applying an alternative version of Lemma 7.1 based on the inequality

$$|S_{\rm adap} \backslash S_0|^3 \le \| w_{S_{\rm adap} \backslash S_0} \|_2^2 \| (\beta_{\rm init})_{S_{\rm adap} \backslash S_0} \|_1^2 \le \| w_{S_{\rm adap} \backslash S_0} \|_2^2 \delta_1^2 .$$

This alternative route yields qualitatively the same results under e.g. sparse eigenvalue conditions. To avoid too many cases, we do not elaborate this.

Remark 9.2. A further observation is that the above corollary is an obstructed oracle inequality, where the oracle is restricted to choose the index set as a thresholded set of the initial Lasso. Concentrating on prediction error, it leads to defining the "oracle" threshold as

$$\delta_0 := \arg \min_{\delta \geq \delta_2/\sqrt{s_0}} \left\{ \|\mathrm{f}_{S^\delta_{\rm init}} - \mathbf{f}^0\|^2 + \frac{12\lambda_{\rm init}^2 \lambda_{\rm adap}^2 s_0}{\delta^2 \phi_{\rm min}^2(2, S_0, 2s_0)} \right\}. \tag{9.7}$$

This oracle has active set $S_{\rm init}^{\delta_0}$, with size $|S_{\rm init}^{\delta_0}| = O(s_0)$. Our following considerations however will not be based on this optimal threshold, but rather on thresholds that allow a comparison with the results for the thresholded initial Lasso. This means that we might loose here some further favorable properties of the adaptive Lasso.

Proof of Theorem 7.4. Corollary 9.2 combined with Lemma 9.4 gives that for all $\delta \geq \delta_2/\sqrt{s_0}$,

$$\delta_{\rm adap}^2 \leq 4\|\mathrm{f}_{S_0} - \mathbf{f}^0\|^2 + 72\Lambda_{\rm sparse}^2(s_0)\delta^2 s_0 + \frac{12\lambda_{\rm init}^2\lambda_{\rm adap}^2 s_0}{\delta^2 \phi_{\rm min}^2(2, S_0, 2s_0)}.$$

Using moreover that $\|\beta_{\rm adap} - b^0\|_q \leq \|\beta_{\rm adap} - \beta_{S^\delta_{\rm init}}\|_q + \|\beta_{S^\delta_{\rm init}} - b^0\|_q$ and the bound of Lemma 9.4, we get for $\delta \geq \delta_2/\sqrt{s_0}$,

$$\|\beta_{\rm adap} - b^0\|_1 \leq 3\delta s_0 + \frac{6\delta\|\mathrm{f}_{S_0} - \mathbf{f}^0\|^2}{\lambda_{\rm init}\lambda_{\rm adap}} + \frac{108\Lambda_{\rm sparse}^2(s_0)\delta^3 s_0}{\lambda_{\rm init}\lambda_{\rm adap}} + \frac{6\lambda_{\rm init}\lambda_{\rm adap}s_0}{\delta\phi_{\rm min}^2(2, S_0, 2s_0)},$$

and

$$\|\beta_{\rm adap} - b^0\|_2 \leq 3\delta\sqrt{s_0} + \frac{12\delta\|\mathrm{f}_{S_0} - \mathbf{f}^0\|^2}{\sqrt{s_0}\lambda_{\rm init}\lambda_{\rm adap}} + \frac{216\Lambda_{\rm sparse}^2(s_0)\delta^3\sqrt{s_0}}{\lambda_{\rm init}\lambda_{\rm adap}} + \frac{18\lambda_{\rm init}\lambda_{\rm adap}\sqrt{s_0}}{\delta\phi_{\rm min}^2(2, S_0, 3s_0)}.$$

Finally, again for $\delta \geq \delta_2/\sqrt{s_0}$,

$$|S_{\rm adap}\backslash S_0| \leq \frac{8\Lambda_{\rm sparse}^2(s_0)\delta_2^2}{\lambda_{\rm init}^2\lambda_{\rm adap}^2}\left(\frac{4\|\mathrm{f}_{S_0} - \mathbf{f}^0\|^2}{s_0} + 72\Lambda_{\rm sparse}^2(s_0)\delta^2 + \frac{12\lambda_{\rm init}^2\lambda_{\rm adap}^2}{\delta^2\phi_{\rm min}^2(2, S_0, 2s_0)}\right).$$

By Corollary 9.1,

$$\frac{\delta_2}{\sqrt{s_0}} = O\left(\frac{\lambda_{\rm init}}{\phi^2(2, S_0, 2s_0)}\right).$$

Taking

$$\delta^2 \asymp \frac{\lambda_{\rm init}\lambda_{\rm adap}}{\phi_{\rm min}(2, S_0, 2s_0)\Lambda_{\rm sparse}(s_0)}, \tag{9.8}$$

the requirement that $\delta \geq \delta_2/\sqrt{s_0}$ is fulfilled if take

$$\lambda_{\rm init}\left[\frac{\phi_{\rm min}(2, S_0, 2s_0)\Lambda_{\rm sparse}(s_0)}{\phi^4(2, S_0, 2s_0)}\right] = O_{\rm suff}(\lambda_{\rm adap}),$$

that is, if Condition b holds. We then obtain

$$\delta^2_{\rm adap} = \left[\frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\rm min}(2, S_0, 2s_0)} \right] O(\lambda_{\rm init}\lambda_{\rm adap} s_0),$$

$$\|\beta_{\rm adap} - b^0\|_1 = \left[\frac{\Lambda^{1/2}_{\rm sparse}(s_0)}{\phi^{3/2}_{\rm min}(2, S_0, 2s_0)} \right] O(\sqrt{\lambda_{\rm init}\lambda_{\rm adap}} s_0),$$

$$\|\beta_{\rm adap} - b^0\|_2 = \left[\frac{\Lambda^{1/2}_{\rm sparse}(s_0)\phi^{1/2}_{\rm min}(2, S_0, 2s_0)}{\phi^2_{\rm min}(2, S_0, 3s_0)} \right] O(\sqrt{\lambda_{\rm init}\lambda_{\rm adap} s_0}),$$

and

$$|S_{\rm adap} \backslash S_0| = \frac{\Lambda^2_{\rm sparse}(s_0)}{\phi^4(2, S_0, 2s_0)} \left[\frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\rm min}(2, S_0, 2s_0)} \right] \frac{\lambda_{\rm init}}{\lambda_{\rm adap}} O(s_0).$$

□

9.4. Proofs for Section 8: the noisy case

Theorem 8.1 gives bounds for prediction error, estimation error and the number of false positives of the noisy weighted Lasso.

Proof of Theorem 8.1. We can derive the prediction and estimation results in the same way as in Theorem 7.1, adding now the noise term:

$$\|\hat{f}_{\rm weight} - \mathbf{f}^0\|^2_n + \lambda_{\rm init}\lambda_{\rm weight} \sum_{j=1}^p w_j |\hat{\beta}_{j,{\rm weight}}|$$

$$\leq 2(\epsilon, \hat{f}_{\rm weight} - f_{\beta_S})_n + \|f_{\beta_S} - \mathbf{f}^0\|^2_n + \lambda_{\rm init}\lambda_{\rm weight} \sum_{j\in S} w_j |\beta_j|$$

$$\leq \lambda_{\rm init} \|\hat{\beta}_{\rm weight} - \beta_S\|_1/2 + \|f_{\beta_S} - \mathbf{f}^0\|^2_n + \lambda_{\rm init}\lambda_{\rm weight} \sum_{j\in S} w_j |\beta_j|$$

and hence, using $\lambda_{\rm weight} w^{\rm min}_{S^c} \geq 1$,

$$\|\hat{f}_{\rm weight} - \mathbf{f}^0\|^2_n + \lambda_{\rm init}\lambda_{\rm weight} w^{\rm min}_{S^c} \|\hat{\beta}_{S^c}\|_1/2$$

$$\leq \|f_{\beta_S} - \mathbf{f}^0\|^2_n + \left[\lambda_{\rm int}/2 + \lambda_{\rm init}\lambda_{\rm weight} \|w_S\|_2/\sqrt{s} \right] \sqrt{s} \|\hat{\beta}_{\rm weight} - \beta_S\|_2.$$

Now insert $w^{\rm min}_{S^c} \geq M/L$, $1 \leq \lambda_{\rm weight} M$ and $\|w_S\|_2/\sqrt{s} \leq M$:

$$\|\hat{f}_{\rm weight} - \mathbf{f}^0\|^2_n + \lambda_{\rm init}\lambda_{\rm weight} M \|\hat{\beta}_{S^c}\|_1/(2L)$$

$$\leq \|f_{\beta_S} - \mathbf{f}^0\|^2_n + \left[\lambda_{\rm init}/2 + \lambda_{\rm init}\lambda_{\rm weight} M \right] \sqrt{s} \|\hat{\beta}_{\rm weight} - \beta_S\|_2$$

$$\leq \|f_{\beta_S} - \mathbf{f}^0\|^2_n + 3\lambda_{\rm int}\lambda_{\rm weight} M \sqrt{s} \|\hat{\beta}_{\rm weight} - \beta_S\|_2/2.$$

The rest of the proof for the prediction and estimation error can therefore carried out in the same way is the proof of Theorem 7.1.

As for variable selection, we use as in Lemma 7.1 the weighted KKT conditions: for all j

$$2(\psi_j, \hat{f}_{\rm weight} - \mathbf{f}^0)_n - 2(\psi_j, \epsilon)_n = -\lambda_{\rm init}\lambda_{\rm weight} w_j \hat{\tau}_{j,{\rm weight}},$$

where $\|\hat{\tau}_{\rm weight}\|_\infty \le 1$ and $\hat{\tau}_{j,{\rm weight}} {\rm l}\{\hat{\beta}_{j,{\rm weight}} \ne 0\} = {\rm sign}(\hat{\beta}_{j,{\rm weight}})$. Invoking $\lambda_{\rm weight} w_{S^c}^{\rm min} \ge 1$, we know that for all $j \in S^c$, $\lambda_{\rm weight} w_j \ge 1$. Moreover, $2|(\epsilon, \psi_j)_n \le \lambda_{\rm init}/2$ by the definition of $\mathcal{T}$. Therefore,

$$\sum_{j \in \hat{S}_{\rm weight} \cap S^c \backslash S_0} 2|(\psi_j, \hat{f}_{\rm weight} - \mathbf{f}^0)_n|^2 \ge \lambda_{\rm init}^2 \lambda_{\rm weight}^2 \|w_{\hat{S}_{\rm weight} \cap S^c \backslash S_0}\|_2^2/4.$$

One can now proceed as in Lemma 7.1. □

9.4.1. Proof of Lemma 8.1 with the more involved conditions

To prove this lemma, we actually need some results from Section 3 and an intermediate result in their proof. One may skip the present proof at first reading and first consult the next subsection (Subsection 9.5).

The bound for the number of false positives of the initial lasso follows from the inequality

$$|\hat{S}_{\rm init} \backslash S_0| \le \frac{\Lambda_{\max}^2(\hat{S}_{\rm init} \backslash S_0)}{\phi^2(6, S_0)} O(s_0).$$

This follows from Theorem 8.1, and from inserting the bound of Theorem 3.1 for $\hat{\delta}_{\rm init}$. One can then proceed by applying the inequality

$$\Lambda_{\max}^2(\hat{S}_{\rm init} \backslash S_0) \le \left(\frac{|\hat{S}_{\rm init} \backslash S_0|}{s_*} + 1 \right) \Lambda_{\rm sparse}^2(s_*). \tag{9.9}$$

The result for the adaptive Lasso can be derived from

$$|\hat{S}_{\rm adap} \backslash S_0|^2 \le \frac{\Lambda_{\max}^2(\hat{S}_{\rm adap} \backslash S_0)}{\phi^4(6, S_0, 2s_0)} \left[\frac{\Lambda_{\rm sparse}(s_0)}{\phi_{\min}(6, S_0, 2s_0)} \right] \frac{\lambda_{\rm init}}{\lambda_{\rm adap}} O(s_0).$$

This follows from (9.11) (which can be found at the end of the proof of Theorem 3.3), invoking Condition B, and applying the bound of Theorem 3.3 for $\hat{\delta}_{\rm adap}$, and the bound of Theorem 3.1 for $\hat{\delta}_2$. Insert again (9.9) to complete the proof. □

9.5. Proofs for Section 3

9.5.1. Proof of the probability inequality of Lemma 3.1

This follows easily from the probability bound $\mathbb{P}(|Z| \ge \sqrt{2t}) \le 2\exp[-t]$ for a standard normal random variable Z. □

9.5.2. Proof of Theorem 3.1 and Lemma 3.2: the noisy initial Lasso

Theorem 3.1 and Lemma 3.2 are simplified formulation of Corollary 9.3 below. This corollary follows from Theorem 8.1 by taking $L = 1$ and $S = S_0$.

Corollary 9.3. *Let*

$$\delta^2_{\text{oracle}} := \|\mathrm{f}_{S_0} - \mathbf{f}^0\|_n^2 + \frac{7\lambda^2_{\text{init}}|S_0|}{\phi^2(6, S_0, 2s_0)}.$$

We have on $\mathcal{T}$,

$$\hat{\delta}^2_{\text{init}} \le 2\delta^2_{\text{oracle}}.$$

Moreover, on $\mathcal{T}$,

$$\hat{\delta}_1 \le 5\delta^2_{\text{oracle}}/\lambda_{\text{init}},$$

and

$$\hat{\delta}_2 \le 10\delta^2_{\text{oracle}}/(\lambda_{\text{init}}\sqrt{s_0}).$$

Also, on $\mathcal{T}$,

$$|\hat{S}_{\text{init}} \backslash S_0| \le 16\Lambda^2_{\max}(\hat{S}_{\text{init}} \backslash S_0)\frac{\hat{\delta}^2_{\text{init}}}{\lambda^2_{\text{init}}}.$$

9.5.3. Proof of Theorem 3.2: the noisy thresholded Lasso

The least squares estimator $\hat{\mathrm{f}}_{\hat{S}^\delta_{\text{init}}}$ using only variables in $\hat{S}^\delta_{\text{init}}$ (i.e., the projection of $\mathbf{Y} = \mathbf{f}^0 + \epsilon$ on the linear space spanned by $\{\psi_j\}_{j\in\hat{S}^\delta_{\text{init}}}$) has similar prediction properties as $\mathrm{f}_{\hat{S}^\delta_{\text{init}}}$ (the projection of $\mathbf{f}^0$ on the same linear space). This is because, as is shown in the next lemma, their difference is small.

Lemma 9.5. *Let $\delta \ge \hat{\delta}_2/\sqrt{s_0}$. Then on $\mathcal{T}$,*

$$\|\hat{\mathrm{f}}_{\hat{S}^\delta_{\text{init}}} - \mathrm{f}_{\hat{S}^\delta_{\text{init}}}\|_n^2 \le \frac{\lambda^2_{\text{init}}s_0}{2\phi^2_{\text{sparse}}(S_0, 2s_0)}.$$

Proof of Lemma 9.5. This follows from

$$\|\hat{\mathrm{f}}_{\hat{S}^\delta_{\text{init}}} - \mathrm{f}_{\hat{S}^\delta_{\text{init}}}\|_n^2 \le 2(\epsilon, \hat{\mathrm{f}}_{\hat{S}^\delta_{\text{init}}} - \mathrm{f}_{\hat{S}^\delta_{\text{init}}})_n,$$

and

$$2(\epsilon, \hat{\mathrm{f}}_{\hat{S}^\delta_{\text{init}}} - \mathrm{f}_{\hat{S}^\delta_{\text{init}}})_n \le \lambda_{\text{init}}\|\hat{b}^{\hat{S}^\delta_{\text{init}}} - b^{\hat{S}^\delta_{\text{init}}}\|_1/2$$

$$\le \lambda_{\text{init}}\sqrt{2s_0}\|\hat{b}^{\hat{S}^\delta_{\text{init}}} - b^{\hat{S}^\delta_{\text{init}}}\|_2/2 \le \lambda_{\text{init}}\sqrt{2s_0}\|\hat{\mathrm{f}}_{\hat{S}^\delta_{\text{init}}} - \mathrm{f}_{\hat{S}^\delta_{\text{init}}}\|_n/(2\phi_{\text{sparse}}(S_0, 2s_0)).$$

□

Proof of Theorem 3.2. The bound for $\|(\hat{\beta}_{\text{init}})_{\hat{S}^\delta_{\text{init}}} - b^0\|_2 \le 2\hat{\delta}_2 + \delta\sqrt{s_0}$ can be derived in the same way as in Lemma 9.4. The same is true for the bound

$$\|\mathrm{f}_{\hat S^\delta_{\rm init}} - \mathbf{f}^0\|_n \le \|f_{(\hat\beta_{\rm init})_{\hat S^\delta_{\rm init}}} - \mathbf{f}^0\|_n$$

$$\le \|\mathrm{f}_{S_0} - \mathbf{f}^0\|_n + \sqrt{\left\lceil \frac{\hat\delta_2^2}{\delta^2 s_0} + 1 \right\rceil} \Lambda_{\rm sparse}(s_0)(2\hat\delta_2 + \delta\sqrt{s_0}).$$

Assumption A together with Lemma 9.5 complete the proof for the bounds for prediction and estimation error, with the ℓ_1-bound being a simple consequence of the thus derived ℓ_2-bound. Also, the variable selection result follows from

$$|\hat S^\delta_{\rm init} \backslash S_0| \le \frac{\hat\delta_2^2}{\delta^2},$$

and Assumption A. □

9.5.4. Proof of Theorem 3.3: the noisy adaptive Lasso

We first apply Theorem 8.1 to the adaptive Lasso.

Corollary 9.4. *Suppose we are on $\mathcal{T}$. Take $\lambda_{\rm adap} \ge \delta \ge \hat\delta_2/\sqrt{s_0}$. Apply Theorem 8.1, with $S := \hat S^\delta_{\rm init}$, $M = \delta$ and $L = 1$, and invoke that $|\hat S^\delta_{\rm init}| \le 2s_0$ and that $\phi^2(6, \hat S^\delta_{\rm init}, |\hat S^\delta_{\rm init}|) \ge \phi^2_{\rm min}(6, S_0, 2s_0)$ and $\phi^2(6, \hat S^\delta_{\rm init}, s_0 + |\hat S^\delta_{\rm init}|) \ge \phi^2_{\rm min}(6, S_0, 3s_0)$. This then gives, for all $\delta \ge \hat\delta_2/\sqrt{s_0}$, and all β*

$$\hat\delta^2_{\rm adap} \le 2\|f_{\beta_{\hat S^\delta_{\rm init}}} - \mathbf{f}^0\|_n^2 + \frac{28\lambda^2_{\rm init}\lambda^2_{\rm adap}s_0}{\delta^2\phi^2_{\rm min}(6, S_0, 2s_0)},$$

and

$$\|\hat\beta_{\rm adap} - \beta_{\hat S^\delta_{\rm init}}\|_1 \le \frac{5\delta\|f_{\beta_{\hat S^\delta_{\rm init}}} - \mathbf{f}^0\|_n^2}{\lambda_{\rm init}\lambda_{\rm adap}} + \frac{14\lambda_{\rm init}\lambda_{\rm adap}s_0}{\delta\phi^2_{\rm min}(6, S_0, 2s_0)},$$

and

$$\|\hat\beta_{\rm adap} - \beta_{\hat S^\delta_{\rm init}}\|_2 \le \frac{10\delta\|f_{\beta_{\hat S^\delta_{\rm init}}} - \mathbf{f}^0\|_n^2}{\sqrt{s_0}\lambda_{\rm init}\lambda_{\rm adap}} + \frac{42\lambda_{\rm init}\lambda_{\rm adap}\sqrt{s_0}}{\delta\phi^2_{\rm min}(6, S_0, 3s_0)}.$$

Moreover

$$|(\hat S_{\rm adap} \cap (\hat S^\delta_{\rm init})^c)\backslash S_0| \le s_0 + 32\Lambda_{\rm sparse}(s_0)\frac{\hat\delta^2_{\rm adap}}{\lambda^2_{\rm adap}s_0}\frac{\hat\delta_2^2}{\lambda^2_{\rm init}} \wedge 4\Lambda_{\rm max}\frac{\hat\delta_{\rm adap}}{\lambda_{\rm adap}}\frac{\hat\delta_2}{\lambda_{\rm init}}.$$

Proof of Theorem 3.3. By the same arguments as used in Lemma 9.4, for $\delta \ge \hat\delta_2/\sqrt{s_0}$,

$$\|f_{(\hat\beta_{\rm init})_{\hat S^\delta_{\rm init}}} - \mathbf{f}^0\|_n \le \|\mathrm{f}_{S_0} - \mathbf{f}^0\|_n^2 + 3\sqrt{2}\Lambda^2_{\rm sparse}(s_0)\delta^2 s_0,$$

and $\|(\hat\beta_{\rm init})_{\hat S^\delta_{\rm init}} - b^0\|_2 \le 3\delta\sqrt{s_0}$. The prediction and estimation results now follow from Corollary 9.4 combined with Condition B.

We apply Corollary 9.4 with

$$\delta^2 = \frac{\lambda_{\mathrm{init}}\lambda_{\mathrm{adap}}}{\phi_{\min}(6, S_0, 2s_0)\Lambda_{\mathrm{sparse}}}. \tag{9.10}$$

Condition B requires that

$$\left[\frac{\Lambda_{\mathrm{sparse}}(s_0)}{\phi^3_{\min}(6, S_0, 2s_0)}\right]\lambda_{\mathrm{init}} = O_{\mathrm{suff}}(\lambda_{\mathrm{adap}}).$$

This ensures that $\delta \ge \hat{\delta}_2/\sqrt{s_0}$ on the set $\mathcal{T}$. Moreover, equation (9.10) gives that $\lambda_{\mathrm{adap}} \ge \delta$ as soon as

$$\lambda_{\mathrm{adap}} \ge \left[\frac{1}{\phi_{\min}(6, S_0, 2s_0)\Lambda_{\mathrm{sparse}}(s_0)}\right]\lambda_{\mathrm{init}},$$

which is also ensured by Condition B.

The variable selection result follows from: for $\delta \ge \hat{\delta}_2/\sqrt{s_0}$,

$$|\hat{S}_{\mathrm{adap}} \backslash S_0| \le |(\hat{S}_{\mathrm{adap}} \cap (\hat{S}^{\delta}_{\mathrm{init}})^c \backslash S_0| + |\hat{S}^{\delta}_{\mathrm{init}} \backslash S_0| \le |(\hat{S}_{\mathrm{adap}} \cap (\hat{S}^{\delta}_{\mathrm{init}})^c \backslash S_0| + s_0. \tag{9.11}$$

□

9.5.5. Proof of Lemma 3.3, where coefficients are assumed to be large

On $\mathcal{T}$, for $j \in S_0$, $|\hat{\beta}_{j,\mathrm{init}}| > \hat{\delta}_\infty$, and $|\hat{\beta}_{j,\mathrm{init}}| > |b^0_j|/2$, since $|b^0_j| > 2\hat{\delta}_\infty$. Moreover, for $j \in S_0^c$, $|\hat{\beta}_{j,\mathrm{init}}| \le \hat{\delta}_\infty$. Let

$$M^2 = \frac{4}{s_0}\sum_{j \in S_0}\frac{1}{|b^0_j|^2}.$$

So

$$\|w_{S_0}\|_2^2/s_0 \le M^2.$$

Note that $M \le 1/\hat{\delta}_\infty$. Since $w^{\min}_{S_0^c} \ge 1/\hat{\delta}_\infty$, the condition $\lambda_{\mathrm{adap}}M \ge 1$ implies $\lambda_{\mathrm{adap}}w^{\min}_{S_0^c} \ge 1$.

Apply Theorem 8.1 to the adaptive Lasso with $S = S_0$, and $\beta = b^0$:

$$\hat{\delta}^2_{\mathrm{adap}} \le 2\|\mathrm{f}_{S_0} - \mathbf{f}^0\|_n^2 + \frac{14\lambda^2_{\mathrm{init}}\lambda^2_{\mathrm{adap}}M^2 s_0}{\phi^2(6, S_0)} = O\left(\frac{\lambda^2_{\mathrm{init}}\lambda^2_{\mathrm{adap}}M^2 s_0}{\phi^2(6, S_0)}\right),$$

and

$$\|\hat{\beta}_{\mathrm{adap}} - b^0\|_1 \le \frac{5\|\mathrm{f}_{S_0} - \mathbf{f}^0\|_n^2}{\lambda_{\mathrm{init}}\lambda_{\mathrm{adap}}M} + \frac{7\lambda_{\mathrm{init}}\lambda_{\mathrm{adap}}Ms_0}{\phi^2(6, S_0,)} = O\left(\frac{\lambda_{\mathrm{init}}\lambda_{\mathrm{adap}}Ms_0}{\phi^2(6, S_0)}\right),$$

and

$$\|\hat{\beta}_{\mathrm{adap}} - b^0\|_2 \le \frac{10\|\mathrm{f}_{S_0} - \mathbf{f}^0\|_n^2}{M\sqrt{s_0}\lambda_{\mathrm{init}}\lambda_{\mathrm{adap}}} + \frac{28\lambda_{\mathrm{init}}\lambda_{\mathrm{adap}}M\sqrt{s_0}}{\phi^2(6, S_0, 2s_0)} = O\left(\frac{\lambda_{\mathrm{init}}\lambda_{\mathrm{adap}}M\sqrt{s_0}}{\phi^2(6, S_0, 2s_0)}\right).$$

Also, when $|\hat{S}_{\text{adap}} \backslash S_0| > s_0$, it holds that

$$
\begin{aligned}
|\hat{S}_{\text{adap}} \backslash S_0| &\le 32\Lambda^2_{\text{sparse}}(s_0) \frac{\|\hat{f}_{\text{adap}} - \mathbf{f}^0\|_n^2}{\lambda^2_{\text{adap}} s_0} \frac{\|(1/w)_{\hat{S}_{\text{adap}} \backslash S_0}\|_2^2}{\lambda^2_{\text{init}}} \\
&\le 32\Lambda^2_{\text{sparse}}(s_0) \frac{\|\hat{f}_{\text{adap}} - \mathbf{f}^0\|_n^2}{\lambda^2_{\text{adap}} s_0} \frac{\hat{\delta}_2^2}{\lambda^2_{\text{init}}} \\
&= \Lambda^2_{\text{sparse}}(s_0) O\left(\frac{\lambda^2_{\text{init}} M^2 s_0}{\phi^2(6, S_0)\phi^4(6, S_0, 2s_0)} \right).
\end{aligned}
$$

Alternatively, for any size of $\hat{S}_{\text{adap}} \backslash S_0$,

$$
\begin{aligned}
|\hat{S}_{\text{adap}} \backslash S_0|^2 &\le 16\Lambda^2_{\max}(\hat{S}_{\text{adap}} \backslash S_0) \frac{\|\hat{f}_{\text{adap}} - \mathbf{f}^0\|_n^2}{\lambda^2_{\text{adap}}} \frac{\hat{\delta}_2^2}{\lambda^2_{\text{init}}} \\
&= |\hat{S}_{\text{adap}} \backslash S_0| O\left(\frac{\lambda^2_{\text{init}} M^2 s_0^2}{\phi^2(6, S_0)\phi^4(6, S_0, 2s_0)} \right).
\end{aligned}
$$

□

Acknowledgement

We are extremely grateful to the associate editor and referee, for their careful reading and pointing out the weak spots in earlier versions of the paper.

References

[1] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010. MR2602303

[2] A. Barron, L. Birge, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999. MR1679028

[3] D. Bertsimas and J. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997.

[4] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009. MR2533469

[5] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[6] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation and sparsity via ℓ_1-penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006. Lecture Notes in Artificial Intelligence 4005*, pages 379–391, Heidelberg, 2006. Springer Verlag. MR2280619

[7] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007a. MR2351101

[8] F. Bunea, A. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007b. MR2312149

[9] E. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37:2145–2177, 2009. MR2543688

[10] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005. MR2243152

[11] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007. MR2382644

[12] E.J. Candès, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2006. MR2230846

[13] EJ Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted 11 minimization. *J. Fourier Anal. Appl*, 14:877–905, 2008. MR2461611

[14] L. De Haan and A. Ferreira. *Extreme Value theory: an Introduction*. Springer Verlag, 2006. ISBN 0387239464. MR2234156

[15] J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 2010.

[16] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004. MR2108039

[17] J. Huang, S. Ma, and C.-H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008. MR2469326

[18] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45:7–57, 2009a. MR2500227

[19] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009b. MR2555200

[20] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008. MR2386087

[21] L. Meier, S. van de Geer, and P. Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70:53–71, 2008. MR2412631

[22] N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52:374–393, 2007. MR2409990

[23] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006. MR2278363

[24] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009. MR2488351

[25] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996. MR1379242

[26] S. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645, 2008. MR2396809

[27] S. van de Geer. On non-asymptotic bounds for estimation in generalized linear models with highly correlated design. In *Asymptotics: Particles, Processes and Inverse Problems (E.A. Cator, G. Jongbloed, C. Kraaikamp, H.P. Lopuhaä, J.A. Wellner eds.)*, volume 55, pages 121–134. IMS Lecture Notes Monograph Series, 2007. MR2459935

[28] S. van de Geer. Least squares estimation with complexity penalties. *Mathematical Methods of Statistics*, pages 355–374, 2001. MR1867165

[29] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, pages 1360–1392, 2009. MR2576316

[30] M. Wainwright. Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55:5728–5741, 2007. MR2597190

[31] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009. MR2729873

[32] L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of Statistics*, 37:2178–2201, 2009. MR2543689

[33] C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010. MR2604701

[34] C.H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008. MR2435448

[35] T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Annals of Statistics*, 37:2109–2144, 2009. MR2543687

[36] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006. MR2274449

[37] S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing Systems 22*. MIT Press, 2009.

[38] S. Zhou. Thresholded lasso for high dimensional variable selection and statistical estimation, 2010. arXiv:1002.1583v2, shorter version in Advances in Neural Information Processing Systems 22(NIPS 2009).

[39] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. MR2279469

[40] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics*, 36:1509–1566, 2008. MR2435443