On the efficiency of the de-biased Lasso

Sara van de Geer
Seminar for Statistics, ETH Zürich
August 26, 2017

Abstract. We consider the high-dimensional linear regression model
\[ Y = X \beta^0 + \epsilon \]
with Gaussian noise \( \epsilon \) and Gaussian random design \( X \). We assume
that \( \Sigma_0 := \mathbb{E} X^T X / n \) is non-singular and write its inverse as \( \Theta^0 := \Sigma_0^{-1} \). The
parameter of interest is the first component \( \beta_1^0 \) of \( \beta^0 \). We show that the asymp-
totic variance of a de-biased Lasso estimator can be smaller than \( \Theta_{1,1}^0 \), under
the conditions: \( \beta^0 \) is sparse in the sense that it has \( s_0 = o(\sqrt{n}/\log p) \) non-zero
entries and the first column \( \Theta_{1,1}^0 \) of \( \Theta^0 \) is not sparse. As by-product, we obtain
some results for the Lasso estimator of \( \beta^0 \) for cases where \( \beta^0 \) is not sparse.

1 Introduction

We consider the regression model
\[ Y = X \beta^0 + \epsilon, \]
where \( Y \in \mathbb{R}^n \) is a response vector, \( X \in \mathbb{R}^{n \times p} \) is a design matrix, \( \beta^0 \in \mathbb{R}^p \) is an
unknown vector and \( \epsilon \in \mathbb{R}^n \) is unobserved noise. The number of parameters \( p \)
is allowed to be much larger than \( n \). Our parameter of interest is \( \beta_1^0 \), the first
component of the vector \( \beta^0 \). The topic of this paper is the study of asymptotic
bounds on the performance of an estimator of \( \beta_1^0 \).

We assume throughout that \( \epsilon \) is a vector of i.i.d. Gaussian random variables
with mean zero and variance \( \sigma_\epsilon^2 \) which for simplicity is assumed to be known,
say \( \sigma_\epsilon^2 = 1 \). Furthermore, we assume random design, with \( X \) independent of \( \epsilon \).
The rows of \( X \) are assumed to be i.i.d. Gaussian random vectors with mean zero
and non-singular covariance matrix \( \Sigma_0 \) with \( \text{diag}(\Sigma_0) = I \), the \( p \times p \) identity
matrix. The smallest eigenvalue of \( \Sigma_0 \) is denoted by \( \Lambda_{\min}^2 \) and the precision
matrix is \( \Theta^0 := \Sigma_0^{-1} \). For the asymptotics, we assume the standardizations
\( 1/\Lambda_{\min} = \mathcal{O}(1) \) and \( \beta_0^T \Sigma_0 \beta^0 = \mathcal{O}(1) \).

The main question we address is: is \( \Theta_{1,1}^0 \) the efficient asymptotic variance for
estimating \( \beta_1^0 \)? Below in (2) the definition of a de-biased Lasso is given. In
Theorems 2.1 and 2.2 we show that a de-biased Lasso can have an asymptotic
variance that is orders of magnitude smaller than \( \Theta_{1,1}^0 \) (see also Corollary 2.1
where the two theorems are combined). The answer thus appears to be “no”: \( \Theta_{1,1}^0 \)
is not always the efficient asymptotic variance! Nevertheless, there remains
room for discussion as the result of course relies on some premises.

In the high-dimensional context considered here, a common assumption on
\( \beta^0 \) is sparsity. Sparsity of \( \beta^0 \) typically means that the number of non-zero
components \( s_0 \) of \( \beta^0 \) is small. One may also impose weakly sparse variants
of this assumption, involving for some \( 0 < q < 1 \) the \( \ell_q \)-“norm”, defined for
\( b \in \mathbb{R}^p \) as \( \|b\|_q := \left( \sum_{j=1}^p |b_j|^q \right)^{1/q} \). We adopt the by now commonplace notation \( \|b\|_0 := \{|j: \ b_j \neq 0\| \} \) (although it would be more consistent to write this as \( \|b\|_0^0 \)). Thus, \( s_0 := \|\beta^0\|_0 \). Two prevalent sparsity assumptions are:

(i) \( s_0 = o(n/\log p) \)

(ii) \( s_0 = o(\sqrt{n}/\log p) \)

(or their weakly sparse variants).

Before commenting on these assumptions, we introduce the notation applied throughout this paper. We set

\[ \|v\|^2_n := \sum_{i=1}^n v_i^2/n, \ v \in \mathbb{R}^n. \]

Moreover, we put for every non-random \( b \in \mathbb{R}^p \)

\[ \|Y - Xb\|^2 = \mathbf{E}\|Y - Xb\|^2_n. \]

If \( \hat{b} \) is a random vector, we let \( \|Y - X\hat{b}\|^2 \) be the function \( b \mapsto \|Y - Xb\|^2 \) evaluated at \( \hat{b} \) (so \( \|Y - X\hat{b}\|^2 \) is random). The quantities \( \|Xb\| \) and \( \|X\hat{b}\| \) are defined similarly. We let \( X_1 \) be the first column of \( X \) and \( X_{-1} \) be the matrix consisting of the remaining columns. The normalized Gram matrix is written as \( \hat{\Sigma} := X^TX/n \). Thus for any \( b \in \mathbb{R}^p \), \( \|Xb\|^2_n = b^T\hat{\Sigma}b \) and \( \|Xb\|^2 = b^T\Sigma_0b \).

Let us return to the sparsity assumptions. Clearly variant (i) is weaker than variant (ii). It is for example invoked to establish \( \ell_2 \)-consistency of the Lasso \( \hat{\beta} \) (see Bickel et al. [2009] or the monographs Koltchinskii [2011], Bühlmann and van de Geer [2011] and Giraud [2014], and their references). The Lasso is defined as (Tibshirani [1996])

\[ \hat{\beta} := \arg \min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|^2_n + 2\lambda\|b\|_1 \right\}. \tag{1} \]

Here \( \lambda > 0 \) is a tuning parameter. (We will take of order \( \sqrt{\log p/n} \) but not too small.)

Variant (ii) is applied for establishing asymptotic normality of the de-biased Lasso given by

\[ \hat{b}_1 = \hat{\beta}_1 + \hat{\Theta}_1^T X^T(Y - X\hat{\beta})/n. \tag{2} \]

The \( p \)-dimensional vector \( \hat{\Theta}_1 \) is some estimate of the first column \( \Theta^0_1 \) of the precision matrix \( \Theta^0 \). The de-biased Lasso was introduced in Zhang and Zhang [2014] and further developed in Javanmard and Montanari [2014a] and van de Geer et al. [2014] for example. Related work is Belloni et al. [2015] and Belloni et al. [2016]. In Javanmard and Montanari [2014b] and Cai and Guo [2017] one can find refined results on the conditions on \( s_0 \) for estimating \( \beta^0_1 \).

We will show that under sparsity variant (ii) there are cases where the asymptotic variance of a de-biased Lasso is orders of magnitude smaller than \( \Theta^0_1 \). This follows from Theorems 2.1 and 2.2 in the next section. We argue in Section 3 that this smaller asymptotic variance is not in contradiction with the asymptotic Cramér-Rao lower bound. We omit there the details (these can be found
The message is that if one assumes sparsity, this gives one extra information which can be used to improve estimation performance.

Under sparsity variant (i) we were unable to establish an estimator with essentially smaller asymptotic variance. We comment on this issue briefly in Section 4, where we discuss the case $\Sigma_0$ known.

The question on efficiency leads to the study of Lasso estimators in non-sparse situations. It concerns the Lasso for the regression of $X_1$ on $X_{-1}$. For ease of notation, we replace in Section 5 $X_1$ by $Y$ and $X_{-1}$ by $X$. We address the question: what happens to the Lasso $\hat{\beta}$ if $\beta_0$ is not sparse?

With the findings of Section 5 the proof of the main results in Theorems 2.1 and 2.2 is relatively straightforward. This proof is presented in Section 6.

Section 7 contains some basic probability inequalities for products of Gaussians, which are applied in Section 5.

2 A potential improvement over $\Theta_{1,1}^0$ when $\Theta_1^0$ is not sparse

As pointed out in the previous section, the de-biased Lasso is based on some estimate $\hat{\Theta}_1$ of the first column $\Theta_1^0$ of the precision matrix. Following Zhang and Zhang [2014], we use in this section a Lasso estimator, based on the following representation of $\Theta_1^0$. The vector $\{-\Theta_{1,j}^0\}_{j=2}^p$ is proportional to the coefficients $\gamma_0$ of the projection of $X_1$ on the remaining columns $X_{-1}$. More precisely if we write $\gamma_0 := \arg \min_{c \in \mathbb{R}^p} \|X_1 - X_{-1}c\|^2$

then $\Theta_1^0 = (1, -\gamma^0)/\|X_1 - X_{-1}\gamma^0\|^2$. This can be mimicked by replacing the theoretical norm $\|\cdot\|$ by its empirical version $\|\cdot\|_n$. To deal with the possible high-dimensionality of the problem a Lasso-penalty is added. Let $\bar{\lambda} > 0$ be a tuning parameter and set

$$\hat{\gamma} := \arg \min_{c \in \mathbb{R}^{p-1}} \left\{ \|X_1 - X_{-1}c\|^2_n + 2\bar{\lambda}\|c\|_1 \right\}.$$ 

Then take $\hat{\Theta}_1 := (1, -\hat{\gamma})^T/(\|X_1 - X_{-1}\hat{\gamma}\|^2_n + \bar{\lambda}\|\hat{\gamma}\|_1)$.

We show below in Theorem 2.1 that $\hat{\gamma}$ is close to the penalized version of the theoretical regression of $X_1$ on $X_{-1}$, which is

$$\gamma^* := \arg \min_{c \in \mathbb{R}^{p-1}} \left\{ \|X_1 - X_{-1}c\|^2 + 2\bar{\lambda}_*\|c\|_1 \right\}.$$ 

Here $\bar{\lambda}_*$ is another “tuning parameter” which we will take strictly smaller than $\bar{\lambda}$. We let $\Theta_1^* = (1, -\gamma^*)/\|X_1 - X_{-1}\gamma^*\|^2$ and $\bar{s}_* := \|\gamma^*\|_0$. 

3
Theorem 2.1 Let $\hat{b}_1$ be the de-biased Lasso with $\hat{\Theta}_1 = \hat{\Theta}_1$. Assume
- $\lambda = \mathcal{O}(\sqrt{\log p/n})$ is sufficiently large and that $s_0 = o(\sqrt{n}/\log p)$,
- $\hat{\lambda}_s = \mathcal{O}(\sqrt{\log p/n})$ and $\hat{s}_s = o(\sqrt{n}/\log p)$,
- $\bar{\lambda} = \mathcal{O}(\sqrt{\log p/n})$ is sufficiently large (in particular $\lambda > \lambda_s$).

Then
$$\sqrt{n}(\hat{b}_1 - \beta_1^0) = \Theta_1^* X^T \epsilon / \sqrt{n} + \text{remainder}$$
where remainder $= o_P(1)$ and
$$\text{var}(\Theta_1^* X^T \epsilon / \sqrt{n}) = \Theta_{1,1}^* = \frac{\Theta_{1,1}^0}{1 + \Theta_{1,1}^0 \|X_{-1}(\gamma^* - \gamma^0)\|^2}.$$ 

In Theorem 2.1 the penalized version $\gamma^*$ can be quite different from the unpenalized version $\gamma^0$, in particular when $\gamma^0$ is not sparse. If $\gamma^0$ is sparse, in the sense that $\|\gamma^0\|_0 = 1 = o(n/\log p)$, then one can show that $\|X_{-1}(\gamma^* - \gamma^0)\| = o(1)$ (this follows from Corollary 6.1). Therefore an improvement over $\Theta_{1,1}^0$ can only occur when $\gamma^0$ (or equivalently $\Theta_1^0$) is not sparse in this sense.

Note that we assume sparsity variant (ii) on $s_0$ and also the corresponding variant (ii) on $s_s$.

An accompanying result for Theorem 2.1 says that one can construct a large class of covariance matrices $\Sigma_0$ such that the conditions of Theorem 2.1 are met and such that $\|X_{-1}(\gamma^* - \gamma^0)\|$ is not vanishing. Thus, in these settings the improvement over $\Theta_{1,1}^0$ is not asymptotically negligible.

We define $\Sigma_{-1,-1} = \mathbb{E} X^T X_{-1}/n$ and let $\lambda_{-1,-1}$ be the smallest eigenvalue of $\Sigma_{-1,-1}$. Its largest eigenvalue is denoted by $\Lambda_{-1,-1}$.

Theorem 2.2 Fix $\bar{\lambda}_s = \mathcal{O}(\sqrt{\log p/n})$ and $\Sigma_{-1,-1}$. Suppose $1/\Lambda_{-1,-1} = \mathcal{O}(1)$ and $\Lambda_{-1,-1} = \mathcal{O}(1)$. There exists a $(p-1)$-vector $\gamma^0$ such that for
$$\Sigma_0 = \begin{pmatrix} 1 & \gamma^{0T} \Sigma_{-1,-1} \\ \Sigma_{-1,-1} \gamma^{0} & \Sigma_{0,-1,-1} \end{pmatrix}$$
it holds that $1/\Lambda_{\min}^2 = \mathcal{O}(1)$ and
$$\Theta_{1,1}^0 = \frac{\bar{\lambda}_{-1,-1}}{(\Lambda_{-1,-1} - \bar{\lambda}_{-1,-1})}.$$ 

Moreover $\gamma^* \equiv 0$ and (hence) $\Theta_{1,1}^* = 1$.

In the above theorem the covariance matrix $\Sigma_0$ is constructed in such a way that it satisfies the conditions of Section 4, namely is has 1's on the diagonal and $1/\Lambda_{\min}^2 = \mathcal{O}(1)$. The construction works for any sub-matrix $\Sigma_{-1,-1}$ with well-behaved eigenvalues.

Corollary 2.1 Let $\Sigma_0$ be as in Theorem 2.2. Suppose $\lambda = \mathcal{O}(\sqrt{\log p/n})$ is sufficiently large and $s_0 = o(\sqrt{n}/\log p)$. Then in view of Theorem 2.1 the
asymptotic variance of $\hat{b}_1$ with $\hat{\Theta}_1 = \Theta_1$ is reduced from

$$\Theta^{0}_{1,1} = \frac{\bar{\Lambda}_{-1,-1}}{(\bar{\Lambda}_{-1,-1} - \frac{1}{\bar{\Lambda}_{-1,-1}})}$$

to 1.

## 3 A quick look at the Cramér-Rao lower bound

We briefly discuss the Cramér-Rao lower bound (CRLB) without presenting details. The purpose of this section is merely to show that the improved asymptotic variance is not in contradiction with efficiency theory.

Suppose $\beta^0$ is known to lie in some subset $B$ of $\mathbb{R}^p$. We then define $\vartheta_{\text{CRLB}}$ as

$$\vartheta_{\text{CRLB}} := \max_{h \in \mathcal{H}} \frac{h^T e_1}{h^T \Sigma_0 h},$$

where $e_1$ is the first unit vector in $\mathbb{R}^p$ and

$$\mathcal{H} := \{ h \in \mathbb{R}^p : \| h \|_2 = 1, \ \beta^0 + th \in B \ \forall \ |t| \text{ sufficiently small} \}.$$ 

Note that $\Sigma_0$ is the Fisher-information matrix and $h^T e_1$ is the derivative of the function $g(\beta) = \beta_1$ in the direction $h$. For the rationale why $\vartheta_{\text{CRLB}}$ represents a lower bound for the asymptotic variance of an asymptotically unbiased estimator, we refer to [Janková and van de Geer 2016].

One may write the inverse Cramér-Rao lower bound as

$$\frac{1}{\vartheta_{\text{CRLB}}} = \min_{c \in \mathbb{R}^{p-1}} \| X_1 - X_{-1} c \|^2,$$

Let as in Section 2

$$\| X_1 - X_{-1} \gamma^0 \|^2 := \min_{c \in \mathbb{R}^{p-1}} \| X_1 - X_{-1} c \|^2.$$ 

Then $\| X_1 - X_{-1} \gamma^0 \|^2 = 1/\Theta^{0}_{1,1}$. Thus, if $\mathcal{H}$ is sufficiently rich we have the equality

$$\vartheta_{\text{CRLB}} = \Theta^{0}_{1,1}.$$ 

Otherwise $\vartheta_{\text{CRLB}}$ might be orders of magnitude smaller.

Therefore, a main point is to decide what the model $B$ for $\beta^0$ is. In the sparsity context, a possible model is

$$B = \{ b : \| b \|_0 \leq M \}$$

(or weakly sparse variants). For the constant $M$ one may for example require variant (i): $M = o(n/\log p)$ or variant (ii): $M = o(\sqrt{n}/\log p)$. The class $\mathcal{H}$ of possible sub-directions is $\mathcal{H} = \{ h : \| h \|_2 = 1, \| h_{-S_0} \|_0 \leq M - s_0 \}$, with $h_{-S_0} = \{ h_j \}_{j \not\in S_0}$ and $S_0 = \{ j : \beta^0_j \neq 0 \}$ the support set of $\beta^0$. Hence, if $\gamma^0$ (or equivalently $\Theta^{0}_{1,1}$) is not sparse in the sense $\gamma^0 \not\in \mathcal{H}$, then $\vartheta_{\text{CRLB}}$ is smaller than $\Theta^{0}_{1,1}$.
4 Some remarks on sparsity variant (i) and the case where $\Sigma_0$ known

Recall that de-biased Lasso is defined as

$$\hat{b}_1 = \hat{\beta}_1 + \hat{\Theta}_1^T X^T (Y - X\hat{\beta}) / n.$$  

After some algebra, one sees that

$$\hat{b}_1 - \beta_0 = \hat{\Theta}_1^T X^T \epsilon / n - (\hat{\Theta}_1^T \hat{\Sigma} - e_1^T)(\hat{\beta} - \beta_0),$$

where

$$\text{remainder}(b) := (\hat{\Theta}_1^T \hat{\Sigma} - e_1^T)(b - \beta_0), \quad b \in \mathbb{R}^p.$$  

We remark that $\text{remainder}(\hat{\beta})$ is different from the one in Theorem 2.1.

In the following two subsections we will use consistency results for $\hat{\beta}$ under sparsity variants (i) and (ii). These results are quite standard by now, we refer to [Bühlmann and van de Geer 2011] and its references. One may also deduce these consistency results from Lemma 5.1 in this paper (apply it with the vector $\beta^*$ introduced there equal to $\beta^0$).

4.1 The case $\Sigma_0$ known

If $\Sigma_0$ is known, we may proceed as follows. Let $\hat{\Theta}_1 := \Theta_0^T$ be the first column of $\Theta^0$. Since $e_1 = \Sigma_0 \Theta_0^T$ we then have for any fixed $b$

$$\text{remainder}(b) = (\hat{\Theta}_1^T \hat{\Sigma} - e_1^T)(b - \beta_0) = \Theta_0^T (\hat{\Sigma} - \Sigma_0)(b - \beta_0).$$

This is the average of zero mean i.i.d. random variables with variance of order $\|b - \beta_0\|^2$. Thus

$$\text{remainder}(b) = \mathcal{O}(b - \beta_0) / \sqrt{n}). (3)$$

Under sparsity assumption variant (i): $s_0 = \alpha_0(n/\log p)$ the $\ell_2$-consistency of $\hat{\beta}$ is guaranteed: $\|(\hat{\beta} - \beta_0\|^2 = \alpha_0(1)$. However (3) tells us nothing about $\text{remainder}(\hat{\beta})$ because $\hat{\beta}$ is a random vector that depends on $\hat{\Sigma}$. By sample splitting one can overcome this problem. One can invoke a version of (3) with the Lasso estimator of $\beta_0$ based on one half of the sample and the empirical covariance matrix based on the other half of the sample. By reversing the roles of the two half-samples one obtains two de-biased estimators of $\beta_1^0$, and by averaging them one obtains an asymptotically unbiased asymptotically normal estimator with asymptotic variance $\Theta_{1,1}^0$. We omit the details. The conclusion is that in case $\Sigma_0$ is known and assuming only sparsity variant (i) one can construct an estimator with asymptotic variance $\Theta_{1,1}^0$.  

6
Can we improve this, again under sparsity variant (i)? Let us study the de-biased Lasso for general $\hat{\Theta}_1$. The remainder can be written as

$$\text{remainder}(\hat{\beta}) = (\Theta_1^T \hat{\Sigma} - e_1^T)(\hat{\beta} - \beta^0) + (\hat{\Theta}_1 - \Theta_1^0)^T \Sigma_0 (\hat{\beta} - \beta^0).$$

The first term $I$ can be dealt with as before. Let us investigate the second term $II$. Recall that under sparsity variant (i) we have $\ell_2$-consistency. This follows in fact from $1/\Lambda_{\text{min}} = O(1)$ and $\|X(\hat{\beta} - \beta^0)\| = o_P(1)$. The dual norm of $\ell_2$ is $\ell_\infty$. We invoke this fact, that is, we apply Cauchy-Schwarz to obtain for the second term $II$

$$|(\hat{\Theta}_1 - \Theta_1^0)^T \Sigma_0 (\hat{\beta} - \beta^0)| \leq \|X(\hat{\Theta}_1 - \Theta_1^0)\| \times \|X(\hat{\beta} - \beta^0)\|.$$

That is $II = o_P(1/\sqrt{n})$ as soon as $\|X(\hat{\Theta}_1 - \Theta_1^0)\| = O(1/\sqrt{n})$. Moreover, for non-random $\hat{\Theta}_1$ it holds that

$$\text{var}(\hat{\Theta}_1 X^T \epsilon/\sqrt{n}) = \|X\hat{\Theta}_1\|^2.$$

Hence if $\|X(\hat{\Theta}_1 - \Theta_1^0)\| = O(1/\sqrt{n}) (= o(1))$ then

$$\text{var}(\hat{\Theta}_1 X^T \epsilon/\sqrt{n}) = \Theta_{1,1}^0 + o(1).$$

So under sparsity assumption variant (i) we do not seem to be able to establish a decrease in the asymptotic variance.

### 4.2 The case $\Sigma_0$ unknown

If $\Sigma_0$ is unknown, the more stringent variant (ii): $s_0 = o(\sqrt{n}/\log p)$ is often imposed when using the de-biased Lasso. Under sparsity variant (ii) we have consistency in $\ell_1$ as well, with the rate $1/\sqrt{\log p}$: $\|\hat{\beta} - \beta^0\|_1 = o_P(1/\sqrt{\log p})$. The dual norm of $\ell_1$ is $\ell_\infty$, and the $\ell_\infty$-norm can be orders of magnitude smaller than the $\ell_2$-norm. Therefore consistency in $\ell_1$ with rate $1/\sqrt{\log p}$ supplies us a powerful tool for dealing with the term remainder($\hat{\beta}$):

$$|\text{remainder}(\hat{\beta})| = |(\hat{\Theta}_1^T \hat{\Sigma} - e_1^T)(\hat{\beta} - \beta^0)| \leq \|\hat{\Sigma} \hat{\Theta}_1 - e_1\|_\infty \times \|\hat{\beta} - \beta^0\|_1.$$

Thus remainder($\hat{\beta}$) = $o_P(1/\sqrt{n})$ as soon as

$$\|\hat{\Sigma} \hat{\Theta}_1 - e_1\|_\infty = O(\sqrt{\log p/n}).$$

Moreover, for possibly random $\hat{\Theta}_1$ converging sufficiently fast to some nonrandom limit $\Theta_1^*$ in the sense $\|\hat{\Theta}_1 - \Theta_1^*\|_1 = o_P(1/\sqrt{\log p})$ we see that

$$\hat{\Theta}_1^T X^T \epsilon/n = \Theta_1^* X^T \epsilon/n + (\hat{\Theta}_1^T - \Theta_1^* T) X^T \epsilon/n$$

where

$$|III| \leq \|\hat{\Theta}_1 - \Theta_1^*\|_1 \times \|X^T \epsilon\|_\infty/n = o_P(n^{-1/2})$$
since $\|X^T \epsilon\|_\infty/n = O_P(\sqrt{\log p/n})$. The latter follows from Lemma 7.2 and the union bound. We thus end up with asymptotic variance

$$\text{var}(\Theta_1^T X^T \epsilon / \sqrt{n}) = \|X\Theta_1\|^2$$

which is potentially much smaller than $\Theta_0^T X_1^T \epsilon / \sqrt{n}$. If we invoke the Lasso $\hat{\Theta}_1$ for $\tilde{\Theta}_1$, then indeed, as we see from Theorems 2.1 and 2.2 depending on $\Sigma$, improvements can be made which are not asymptotically negligible.

A final remark in this section is about the assumption in this paper that the unknown $\Sigma_0$ has 1’s on the diagonal. In practice, one generally applies the Lasso after standardizing the variables. This is equivalent with replacing the penalty $\lambda \sum_{j=1}^p \|X_j\| b_j$ where $X_j$ is the $j$th column of $X$ ($j = 1, \ldots, p$).

The theory is easily adjusted to such a weighted penalty. Similarly, invoking a constant term and dropping the assumption that $X$ has mean zero leads to a few extra technicalities which are easily dealt with.

5 What happens if $\beta^0$ is not sparse?

As we have described in Section 3 an issue is whether the first column $\Theta_0^T X_1^T \epsilon$ of $\Theta_0$ is sparse or not. If it is not sparse, then $\vartheta_{\text{CRLB}}$ is possibly much smaller than $\Theta_0^T X_1^T \epsilon / \sqrt{n}$. The quantity $\vartheta_{\text{CRLB}}$ is the inverse of the residual after “projection” on a subset of linear space where the sparsity regime holds. Such a sparse approximation of a non-sparse vector can have large error, resulting in $\vartheta_{\text{CRLB}}$ being much smaller than $\Theta_0^T X_1^T \epsilon / \sqrt{n}$. The question is then: are there estimators that reach $\vartheta_{\text{CRLB}}$? We will not answer this question, but Theorem 2.1 and 2.2 show that when $\Theta_0^T X_1^T \epsilon / \sqrt{n}$ is not sparse indeed improvements over $\Theta_0^T X_1^T \epsilon / \sqrt{n}$ are possible. This result is based on Lemmas 5.1 and 5.2 below, where we study for ease of notation the regression of $Y$ on $X$ instead of the regression of $X_1$ on $X_{-1}$.

We consider the Lasso estimator

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

in the situation where $\beta^0$ is possibly not sparse. We compare $\hat{\beta}$ with a theoretical counterpart

$$\beta^* := \arg \min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|^2_n + 2\lambda_\ast \|b\|_1 \right\}.$$ 

We take $\lambda_\ast = O(\sqrt{\log p/n})$ (possibly much smaller). Set $s_\ast = \|\beta^*\|_0$.

Lemma 5.1 below actually does not rely on the Lasso per sé, but rather on the KKT (Karush-Kuhn-Tucker) conditions. Let for $b \in \mathbb{R}^p$, the sub-differential of $b \mapsto \|b\|_1$ be

$$\partial \|b\|_1 = \{ z \in \mathbb{R}^p : \|z\|_\infty \leq 1, \ z^T b = \|b\|_1 \}.$$ 

For the Lasso $\hat{\beta}$ the KKT conditions read

$$X^T(Y - X\hat{\beta})/n = \lambda \hat{z}, \ \hat{z} \in \partial \|\hat{\beta}\|_1.$$ (4)
For $\beta^*$ they read
\[ \Sigma_0(\beta^* - \beta^0) = \lambda_* z^*, \quad z^* \in \partial \|\beta^*\|_1. \]

It follows that
\[ \|\Sigma_0(\beta^* - \beta^0)\|_\infty \leq \lambda_* \] (5)

Let $0 < \alpha < 1$ be a given error level. It need not be fixed: $\alpha = o(1)$ is allowed.

Let $M = o(\sqrt{n/\log p})$ be a constant such that for some $0 < \eta_M = o(1)$ it holds that
\[ P \left( \inf_{b : \|b\|_1 \leq 4M} \|Xb\|_2^2 \leq 1 - \eta_M \right) \leq \frac{1}{2} \alpha. \] (6)

Given $\alpha$, the existence of such constants $M$ and $\eta_M$ follow from general theory on quadratic forms (see for example Chapter 16 in van de Geer [2016] and its references).

Let us denote the support of $\beta^*$ as $S_* := \{j : \beta^*_j \neq 0\}$. For a vector $b \in \mathbb{R}^p$ we let $b_{S_*} \in \mathbb{R}^p$ be the vector with entries $b_{j,S_*} := \begin{cases} b_j & j \in S_* \\ 0 & j \notin S_* \end{cases}$, $j = 1, \ldots, p$

and we set $b_{-S_*} := b - b_{S_*}$.

**Lemma 5.1** Let $\lambda_* = O(\sqrt{\log p/n})$ and let $\beta^*$ be a vector for which (5) holds, and which satisfies $s_* \leq M^2$ with $M$ as in (6). Define $\sigma_* := \|Y - X\beta^*\|$ and
\[ \lambda_0 := \lambda_* + (\sqrt{2\sigma_* + 2\lambda_*})\tilde{\lambda} + (\sigma_* + 2\lambda_*)\tilde{\lambda}^2, \]
with $\tilde{\lambda} = \sqrt{\log(6p/\alpha)/n}$. Let $\hat{\beta}$ be a solution of the KKT conditions (4). Then for $\lambda \geq 2\lambda_0$, with probability at least $1 - \alpha$,
\[ \|\hat{\beta} - \beta^*\|_1 \leq 4\|\hat{\beta}_{S_*} - \beta^*\|_1, \]
\[ \|X(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{(\lambda + \lambda_0)^2 s_*}{\Lambda_{\min}^2(1 - \eta_M)}, \]
\[ \|X(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{(\lambda + \lambda_0)^2 s_*}{\Lambda_{\min}^2(1 - \eta_M)^2} \]
and
\[ \|\hat{\beta} - \beta^*\|_1 \leq \frac{4(\lambda + \lambda_0)s_*}{\Lambda_{\min}^2(1 - \eta_M)}. \]

**Proof.**
We first examine the random part of the problem. Let $\epsilon^* := \epsilon + X(\beta^0 - \beta^*)$.

Invoke condition (5), Lemma 7.3 and the union bound, to obtain
\[ P \left( \|X^T\epsilon^*\|_\infty / n \geq \lambda_0 \right) \leq 3p \exp[-\log(6p/\alpha)] = \frac{1}{2} \alpha. \]
We have now shown that the event
\[
\left\{ \|X^T\epsilon^*\|_\infty/n \leq \lambda_0 \right\} \cap \inf_{b: \|b\|_1 \leq 4M, \|Xb\|_1 = 1} \{ \|X_b\|^2_n \geq 1 - \eta_M \}
\]
has probability at least $1 - \alpha$. In the remainder of the proof, we assume this event to hold.

By the KKT conditions [14]
\[
\hat{\Sigma}(\hat{\beta} - \beta^0) = X^T\epsilon/n - \lambda\hat{\beta}.
\]
Thus
\[
\hat{\Sigma}(\hat{\beta} - \beta^*)/n = X^T\epsilon^*/n - \lambda\hat{\beta}
\]
and therefore
\[
\|X(\hat{\beta} - \beta^*)\|^2_n = (\hat{\beta} - \beta^*)^T X^T\epsilon^*/n - \lambda\|\hat{\beta}\|\|\hat{\beta}^*\|\|\hat{\beta}^*\|\|
\leq \|\hat{\beta} - \beta^*\|_1\|X^T\epsilon^*/\|\hat{\beta}\|_1 + \lambda\|\beta^*\|_1
\leq \lambda\|\hat{\beta} - \beta^*\|_1 - \lambda\|\hat{\beta}\|_1 + \lambda\|\beta^*\|_1
\leq \lambda\|\hat{\beta} - \beta^*\|_1 - (\lambda - \lambda_0)\|\beta_{-S^*}\|_1.
\]
It follows that
\[
\|X(\hat{\beta} - \beta^*)\|^2_n + (\lambda - \lambda_0)\|\hat{\beta}_{-S^*}\|_1 \leq (\lambda + \lambda_0)\|\hat{\beta}_{-S^*} - \beta^*\|_1 \quad (7)
\]
and so
\[
\|\hat{\beta} - \beta^*\|_1 \leq \left( 1 + \frac{\lambda + \lambda_0}{\lambda - \lambda_0} \right) \|\hat{\beta}_{-S^*} - \beta^*\|_1
\leq 4\|\hat{\beta}_{-S^*} - \beta^*\|_1 \leq 4\sqrt{s^*}\|\hat{\beta} - \beta^*\|_2
\leq 4\sqrt{s^*}\|X(\hat{\beta} - \beta^*)\|/\Lambda_{\min}.
\]
and hence
\[
\|X(\hat{\beta} - \beta^*)\|_n \geq \sqrt{1 - \eta_M}\|X(\hat{\beta} - \beta^*)\|.
\]
Continuing with (7) we find
\[
\|X(\hat{\beta} - \beta^*)\|^2_n + (\lambda - \lambda_0)\|\beta_{-S^*}\|_1 \leq (\lambda + \lambda_0)\|\hat{\beta}_{S^*} - \beta^*\|_1
\leq (\lambda + \lambda_0)\sqrt{s^*}\|X(\hat{\beta} - \beta^*)\|/\Lambda_{\min}
\leq \frac{(\lambda + \lambda_0)\sqrt{s^*}\|X(\hat{\beta} - \beta^*)\|}{\Lambda_{\min}\sqrt{1 - \eta_M}}
\leq \frac{1}{2} \frac{(\lambda + \lambda_0)^2 s^*}{\Lambda_{\min}^2 (1 - \eta_M)} + \frac{1}{2} \|X(\hat{\beta} - \beta^*)\|_n^2.
\]
We conclude that
\[
\|X(\hat{\beta} - \beta^*)\|^2_n \leq \frac{(\lambda + \lambda_0)^2 s^*}{\Lambda_{\min}^2 (1 - \eta_M)},
\]
\[
\|X(\hat{\beta} - \beta^*)\|^2 \leq \frac{(\lambda + \lambda_0)^2 s^*}{\Lambda_{\min}^2 (1 - \eta_M)^2}.
\]
Lemma 5.1 says that potentially, the residual $\|Y - X\hat{\beta}\|^2$ of the Lasso is much larger than $\sigma_\epsilon^2$. (Recall that we assume throughout that $\sigma_\epsilon = 1$ but it may be helpful to write $\sigma_\epsilon$ at places so that it is clear where certain constants come from.) It can only happen if $\|X(\beta^* - \beta^0)\|$ is not vanishing, i.e., if the $\ell_1$-penalty forces $\beta^*$ to be a bad approximation of $\beta^0$. Hence, it can only happen if $\beta^0$ is not sparse (in a generic sense). (One may also derive a variant of Lemma 5.1 under weak sparsity conditions on $\beta^*$, of the form $\|\beta^*_q\|^q = o((n/\log n)^{2/q})$.)

**Remark 5.1** In Lemma 5.1 the tuning parameter $\lambda$ has to be chosen large depending on the unknown $\lambda_*$ and $\sigma_*$. One may for example use $\|Y\|_n$ as high probability bound for $\sigma_*$. Moreover if $\lambda_* = o(\log p/n)$ then $\lambda$ will asymptotically meet the requirement of Lemma 5.2 if we take it at least $2\sqrt{2}\|Y\|_n\lambda$ (the factor 2 in front of the $\sqrt{2}$ here is chosen for explicitness, it can be replaced by any fixed constant strictly larger than 1). We see in Lemma 5.2 below that if $p \gg n/\log p$, one can indeed take $\lambda_* = o(\sqrt{\log p/n})$ in the construction applied there.

We now construct values of $\beta^0$ such that $\|X(\beta^* - \beta^0)\|$ is not vanishing.

**Lemma 5.2** Suppose $p \geq n/\log p$. Let $\beta^*$ be a vector with $\|X\beta^*\| = O(1)$ and $s_* := \|\beta^*\|_0 = o(n/\log p)$. Consider a vector $z \in \mathbb{R}^p$ with entries $z_j \in \{0, 1\}$ for all $j$. Let $s_z = \|z\|_0$ be its number of entries equal to 1. Take $s_z$ of the form

$$2s_z = \frac{\Lambda^2_{\text{min}}}{\lambda_*^2},$$

where $\lambda_* = O(\sqrt{\log p/n})$. Finally, take

$$\beta^0 := \beta^* + \lambda_* \Theta_0 z.$$

Then

$$\|\Sigma_0(\beta^* - \beta^0)\|_\infty \leq \lambda_*$$

and

$$\frac{1}{2}\frac{\Lambda^2_{\text{min}}}{\Lambda^2_{\text{max}}} \leq \|X(\beta^* - \beta^0)\|^2 \leq \frac{1}{2}$$

where $\Lambda^2_{\text{max}}$ is the maximal eigenvalue of $\Sigma_0$. Moreover if $\Lambda^2_{\text{max}}$ remains bounded: $\Lambda^2_{\text{max}} = O(1)$, then $\beta^0$ is not sparse, in the sense that $\sqrt{n/\log p} = O(\|\beta^0\|_1)$.

**Proof.** Since $\Sigma_0(\beta^0 - \beta^*) = \lambda_* z$ it follows immediately that $\|\Sigma_0(\beta^* - \beta^0)\|_\infty \leq \lambda_*$. Moreover,

$$\|X(\beta^* - \beta^0)\|^2 = (\beta^* - \beta^0)^T \Sigma_0 (\beta^* - \beta^0) = \lambda_*^2 z^T \Theta_0 z \leq \|z\|^2 \frac{\lambda_*^2}{\Lambda^2_{\text{min}}} = s_z \lambda_*^2 / \Lambda^2_{\text{min}} = \frac{1}{2},$$

and

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{4\sqrt{s_\epsilon} \|X(\hat{\beta} - \beta^*)\|}{\Lambda_{\text{min}}} \leq \frac{4(\lambda + \lambda_0)s^*}{\Lambda^2_{\text{min}}(1 - \eta_M)}.$$
and
\[ \|X(\beta^* - \beta^0)\|^2 = \lambda_*^2 z^T \Theta_0^T z \geq \|z\|^2 \lambda_*^2 / \Lambda_{\max}^2 \]
\[ = s \lambda_*^2 / \Lambda_{\max}^2 = \frac{1}{2} \Lambda_{\min}^2 / \Lambda_{\max}^2. \]

Thus
\[ \frac{1}{2} \Lambda_{\min}^2 / \Lambda_{\max}^2 \leq \|X(\beta^* - \beta^0)\|^2 = \lambda_* (\beta^* - \beta^0)^T z \leq \lambda_* \|\beta^* - \beta^0\|_1, \]
so
\[ \|\beta^* - \beta^0\|_1 \geq \frac{1}{2} \lambda_*^{-1} \Lambda_{\min}^2 / \Lambda_{\max}^2. \]

Since
\[ \|\beta^*\|_1 \leq \sqrt{s_*} \|\beta^*\|_2 \leq \sqrt{s_*} \|X \beta^*\| / \Lambda_{\min} = o(\sqrt{n / \log p}), \]
we see that \( \sqrt{n / \log p} = O(\|\beta^0\|_1) \). \( \Box \)

**Example 5.1** Suppose that \( \Sigma_0 = I \) and that \( \beta^0 = \lambda_* z \) where \( z \) is a \( p \)-vector consisting of only 1’s. We assume that \( p \geq n / \log p \) and \( \lambda_* \) satisfies \( \|\beta^0\|_2^2 = \lambda_*^2 p = \frac{1}{4} \). (Note that if \( p \gg n / \log p \), then \( \lambda_* \) is very small in the sense \( \lambda_* = o(\sqrt{\log p / n}) \).) Take \( \beta^* = 0 \). Then
\[ \Sigma_0 \beta^0 = \beta^0 = \lambda_* z \]
and
\[ \|X \beta^0\|^2 = \|\beta^0\|_2^2 = \frac{1}{2}. \]

By Lemma 5.1 we know that also \( \hat{\beta} \equiv 0 \) with large probability. The residual sum of squares satisfies
\[ \|Y - X \hat{\beta}\|^2_n = \|Y\|^2_n + o_p(1) = \|Y\|^2 + o_p(1) = \sigma^2 + \frac{1}{2} + o_p(1). \]

## 6 Proof of Theorems 2.1 and 2.2

### 6.1 Translation of the results of Section 5

Lemmas 5.1 and 5.2 can be straightforwardly translated to the regression of \( X_1 \) on \( X_{-1} \). This gives the following corollaries.

**Corollary 6.1** Let \( \bar{\lambda}_* = O(\sqrt{\log p / n}) \) and
\[ \|\Sigma_{-1,-1}(\gamma^* - \gamma^0)\|_\infty \leq \bar{\lambda}_*. \]

Suppose \( \gamma^* \) is sparse: \( \|\gamma^*\|_0 = o(n / \log p) \). Consider the Lasso
\[ \hat{\gamma} := \arg \min_{c \in \mathbb{R}^{p-1}} \left\{ ||X_1 - X_{-1}c||_n^2 + 2\lambda ||c||_1 \right\} \]
or in fact any solution $\hat{\gamma}$ of the KKT conditions of the above minimization problem. For $\lambda \sim O(\sqrt{\log p/n})$ sufficiently large we have

$$
\|X_1(\hat{\gamma} - \gamma^*)\|^2_n = \mathcal{O}_P(s_* \log p/n), \quad \|X_1(\hat{\gamma} - \gamma^*)\|^2 = \mathcal{O}_P(s_* \log p/n)
$$

and $\|\hat{\gamma} - \gamma^*\|_1 = \mathcal{O}_P(s_* \sqrt{\log p/n})$.

In the translation of Lemma 5.2 we for simplicity take $\gamma^* = 0$.

**Corollary 6.2** Fix $\Sigma_{-1,-1}$ with smallest eigenvalue $\Lambda^2_{-1,-1}$ and largest eigenvalue $\bar{\Lambda}^2_{-1,-1}$. Consider a vector $\bar{z} \in \mathbb{R}^{p-1}$ with entries $\bar{z}_j \in \{0,1\}$ for all $j$. Let $\|\bar{z}\|_0$ be its number of entries equal to 1. Take

$$
2\|\bar{z}\|_0 = \frac{\Lambda^{2}_{-1,-1}}{\bar{\Lambda}}
$$

where $\bar{\lambda} = O(\sqrt{\log p/n})$, and

$$
\gamma^0 := \bar{\lambda}_* \Sigma_{-1,-1}^T \bar{z}.
$$

Then

$$
\|\Sigma_{-1,-1} \gamma^0\|_\infty \leq \bar{\lambda}_*
$$

and

$$
\frac{1}{2} \Lambda_{-1,-1}^2/\bar{\Lambda}_{-1,-1}^2 \leq \|X_{-1} \gamma^0\|^2 \leq \frac{1}{2}.
$$

**6.2 Proof of Theorem 2.1**

We use the dual norm inequality as in Subsection 4.2. First note that by construction and the KKT conditions

$$
e^T_1 \hat{\Sigma} \hat{\Theta}^0_1 = 1
$$

and for all $j \neq 1$, and with $e_j$ the $j$th unit vector

$$
|e_j^T \hat{\Sigma} \hat{\Theta}^0_1| \leq \bar{\lambda}/\|X_1 - X_{-1} \hat{\gamma}\|^2_n.
$$

Thus

$$
\|\hat{\Sigma} \Theta_1 - e_1\|_\infty \leq \bar{\lambda}/\|X_1 - X_{-1} \hat{\gamma}\|^2_n.
$$

But by Corollary 6.1

$$
\|X_1 - X_{-1} \hat{\gamma}\|_n \geq \|X_1 - X_{-1} \gamma^*\|_n - \|X_{-1}(\hat{\gamma} - \gamma^*)\|_n
$$

$$
= \|X_1 - X_{-1} \gamma^*\|_n + o_P(1)
$$

$$
= \|X_1 - X_{-1} \gamma^*\| + o_P(1)
$$

$$
\geq \|X_1 - X_{-1} \gamma^0\| + o_P(1)
$$

$$
\geq 1/\Lambda_{\min} + o_P(1)
$$

so

$$
1/\|X_1 - X_{-1} \hat{\gamma}\|^2_n = O_P(1)
$$
and therefore
\[ \|\hat{\Sigma}_1 - e_1\|_\infty = O_p(\bar{\lambda}) = O_p(\sqrt{\log p/n}). \]
Since \( \|\hat{\beta} - \beta^0\|_1 = o_p(1/\sqrt{\log n}) \) we see that
\[ \hat{b}_1 - \beta^0_1 = \hat{\Theta}^T_1 X^T \epsilon/n + o_p(1/\sqrt{n}) \]
and that
\[ \text{var}(\hat{\Theta}^T_1 X^T \epsilon/\sqrt{n}|X) = \Theta^*_T 1_X + o(1). \]
We now use the condition \( s_* = o(\sqrt{n}/\log p) \) to get an unconditional variance. Namely, we have
\[ \hat{\Theta}^*_1 X^T \epsilon/n = \Theta^*_1 X^T \epsilon/n + (\hat{\Theta}^*_1 - \Theta^*_1) X^T \epsilon/n, \]
and
\[ |(\hat{\Theta}_1 - \Theta^*_1)^T X^T \epsilon/n| \leq \|\hat{\Theta}_1 - \Theta^*_1\|_1 \times \|X^T \epsilon\|_\infty/n = O_p(s_* \sqrt{n/\log p}) \times O_p(\sqrt{\log p/n}) = o_p(1/\sqrt{n}) \]
since by Lemma 7.2 and the union bound \( \|X^T \epsilon\|_\infty/n = O_p(\sqrt{\log p/n}) \). This completes the proof of Theorem 2.1. \( \square \)

6.3 Proof of Theorem 2.2

Theorem 2.2 follows immediately from Corollary 6.2 except that we still need to show that \( 1/\Lambda_{\min} = O(1) \).

Let
\[ \Sigma_0 := \begin{pmatrix} 1 & \gamma^0_T \Sigma_{-1,-1} \\ \Sigma_{-1,-1} \gamma^0 & \Sigma^0_{-1,-1} \end{pmatrix} \]
where
\[ \|X_{-1} \gamma^0\|^2 \leq \frac{1}{2} < 1. \]
For \( a \in \mathbb{R} \) and \( c \in \mathbb{R}^{p-1} \) satisfying \( a^2 + \|X_{-1} c\|^2 = 1 \),
\[ \begin{pmatrix} a \\ c \end{pmatrix}^T \Sigma_0 \begin{pmatrix} a \\ c \end{pmatrix} = a^2 + 2a \gamma^0_T \Sigma_{-1,-1} c + c^T \Sigma_{-1,-1} c \]
\[ = 1 + 2a \gamma^0_T \Sigma_{-1,-1} c \]
\[ \geq 1 - 2a \|X_{-1} \gamma^0\| \|X_{-1} c\| \]
\[ = 1 - 2a \sqrt{1 - a^2} \|X_{-1} \gamma^0\| \]
\[ \geq 1 - \|X_{-1} \gamma^0\|. \]
Hence
\[ \Lambda^2_{\min} \geq (1 - \|X_{-1} \gamma^0\|) \Lambda^2_{-1,-1}. \]
\( \square \)
7 Probability inequalities

**Lemma 7.1** Let $U$ and $W$ be two independent $\mathcal{N}(0, 1)$-distributed random variables. Then for all $L > 1$
\[
\mathbb{E} \exp \left[ \frac{UW}{L} \right] \leq \exp \left[ \frac{1}{2L^2 - 2L} \right].
\]

**Proof.** We have for $L > 1$
\[
\mathbb{E} \exp \left[ \frac{UW}{L} \right] \leq \mathbb{E} \exp \left( \frac{(U + W)^2 - (U - W)^2}{4L} \right)
= \mathbb{E} \exp \left( \frac{(U + W)^2 - 2}{4L} \right) \exp \left( \frac{2 - (U - W)^2}{4L} \right)
= \mathbb{E} \exp \left( \frac{(U + W)^2 - 2}{4L} \right) \mathbb{E} \exp \left( \frac{2 - (U - W)^2}{4L} \right)
\leq \exp \left[ \frac{1}{4L^2 - 4L} \right] \mathbb{E} \exp \left[ \frac{2 - (U - W)^2}{4L} \right].
\]

(see Lemma 1 and its proof in Laurent and Massart [2000]) or Section 8.4 in Van de Geer [2016]). But
\[
\mathbb{E} \exp \left[ \frac{2 - (U - W)^2}{4L} \right] = \frac{1}{\sqrt{2\pi}} \int \exp \left[ \frac{1 - v^2}{2L} \right] \exp \left[ -\frac{v^2}{2} \right] dv
= \exp \left[ \frac{1}{2L} \right] \frac{1}{\sqrt{2\pi}} \int \exp \left[ -\frac{1}{2} \left( 1 + \frac{1}{L} \right) v^2 \right] dv
= \exp \left[ \frac{1}{2L} \right] \left( 1 + \frac{1}{L} \right)^{-1/2}
= \exp \left[ \frac{1}{2L} - \frac{1}{2} \log(1 + 1/L) \right].
\]

Since
\[
\log(1 + 1/L) \geq 1/L - \frac{1}{2}(1/L^2)
\]
we obtain
\[
\mathbb{E} \exp \left[ \frac{2 - (U + W)^2}{4L} \right] \leq \exp \left[ \frac{1}{4L^2} \right]
\leq \exp \left[ \frac{1}{4L^2 - 4L} \right].
\]

It follows that
\[
\mathbb{E} \exp \left[ \frac{UW}{L} \right] \leq \exp \left[ \frac{2}{4L^2 - 4L} \right] = \exp \left[ \frac{1}{2L^2 - 2L} \right].
\]
\[\square\]
Lemma 7.2 Let $U = (U_1, \ldots, U_n)^T$ and $W = (W_1, \ldots, W_n)^T$ be two independent standard Gaussian $n$-dimensional random vectors. Then for all $t > 0$

$$
P\left(U^T W/n \geq \sqrt{2t/n + t/n}\right) \leq \exp[-t].$$

**Proof.** By Lemma 7.1 and using the independence

$$
\mathbb{E} \exp\left[\frac{U^T W}{L}\right] \leq \exp\left[\frac{n}{2L^2 - 2L}\right].
$$

This gives for all $t > 0$

$$
P\left(U^T W \geq \sqrt{2nt} + t\right) \leq \exp[-t]
$$
(see e.g. Lemma 8.3 in van de Geer [2016]).

Lemma 7.3 Let $\{(U_i, V_i)\}_{i=1}^n$ be i.i.d. two-dimensional Gaussians with mean zero. Suppose $\text{var}(U_1) = 1$. Define $\lambda_* := \mathbb{E}U_1V_1$ and $\sigma_*^2 = \mathbb{E}V_1^2$. Then for all $t > 0$

$$
P\left(U^T V/n \geq \lambda_* + (\sqrt{2}\sigma_* + 2\lambda_*)\sqrt{t/n + (\sigma_* + 2\lambda_*)t/n}\right) \leq 2\exp[-t].$$

**Proof.** For all $i$ the projection of $V_i$ on $U_i$ is $\mathbb{E}U_iV_i/\text{var}(U_i)U_i = \lambda_* U_i$. Hence we may write for all $i$

$$
V_i = \lambda_* U_i + W_i,
$$
where $W_i$ is a zero-mean Gaussian random variable independent of $U_i$. It follows that

$$
U^T V/n = \lambda_* \|U\|_n^2 + U^T W/n.
$$

Since $\text{var}(W_i) \leq \sigma_*^2$ for all $i$ we see from Lemma 7.2 that

$$
P\left(U^T W/n \geq \sqrt{2}\sigma_* \sqrt{t/n + \sigma_*^2 t/n}\right) \leq \exp[-t].
$$

Moreover (see Lemma in Laurent and Massart [2000], also given in van de Geer [2016] as Lemma 8.6)

$$
P\left(\|U\|_n^2 - 1 \geq 2\sqrt{t/n + 2t/n}\right) \leq \exp[-t].
$$

Thus

$$
P\left(U^T V/n \geq \lambda_* + (\sqrt{2}\sigma_* + 2\lambda_*)\sqrt{t/n + (\sigma_* + 2\lambda_*)t/n}\right) \leq 2\exp[-t].$$

$\square$
References

A Belloni, V Chernozhukov, and K Kato. Uniform postselection inference for
lad regression models. *Biometrika*, 102:77–94, 2015.

A. Belloni, V. Chernozhukov, and Y. Wei. Post-selection inference for gener-
alized linear models with many controls. *Journal of Business & Economic
Statistics*, 34(4):606–619, 2016.

P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and
Dantzig selector. *Annals of Statistics*, pages 1705–1732, 2009.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Meth-
ods, Theory and Applications*. Springer, 2011.

T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression:
Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

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

J. Janková and S. van de Geer. Semi-parametric efficiency bounds for high-
dimensional models, 2016. arXiv:1601.00815, To appear in Ann. Statist.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing
for high-dimensional regression. *Journal of Machine Learning Research*, 15
(1):2869–2909, 2014a.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional re-
gression under the gaussian random design model: Asymptotic theory. *IEEE
Transactions on Information Theory*, 60(10):6522–6554, 2014b.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse
Recovery Problems: Ecole d’Eté de Probabilités de Saint-Flour XXXVIII-
2008*, volume 38. Springer Science & Business Media, 2011.

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

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

S. van de Geer. *Estimation and Testing Under Sparsity: Ecole d’Eté de Proba-
bilités de Saint-Flour XLV-2016*. Springer Science & Business Media, 2016.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically
optimal confidence regions and tests for high-dimensional models. *Annals of
Statistics*, 42:1166–1202, 2014.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional pa-
rameters in high dimensional linear models. *Journal of the Royal Statistical
Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.