CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-53

Improved bounds for Square-Root Lasso and Square-Root Slope

A. DERUMIGNY[1]

[1] ENSAE; CREST. E-mail : alexis.derumigny@ensae.fr

Improved bounds for Square-Root Lasso and Square-Root Slope

Alexis Derumigny*

December 11, 2017

Abstract

Extending the results of Bellec, Lecué and Tsybakov [1] to the setting of sparse high-dimensional linear regression with unknown variance, we show that two estimators, the Square-Root Lasso and the Square-Root Slope can achieve the optimal minimax prediction rate, which is $(s/n)\log(p/s)$, up to some constant, under some mild conditions on the design matrix. Here, n is the sample size, p is the dimension and s is the sparsity parameter. We also prove optimality for the estimation error in the l_q-norm, with $q \in [1,2]$ for the Square-Root Lasso, and in the l_2 and sorted l_1 norms for the Square-Root Slope. Both estimators are adaptive to the unknown variance of the noise. The Square-Root Slope is also adaptive to the sparsity s of the true parameter. Next, we prove that any estimator depending on s which attains the minimax rate admits an adaptive to s version still attaining the same rate. We apply this result to the Square-root Lasso. Moreover, for both estimators, we obtain valid rates for a wide range of confidence levels, and improved concentration properties as in [1] where the case of known variance is treated. Our results are non-asymptotic.

MCS: Primary 62G08; secondary 62C20, 62G05.

Keywords: Sparse linear regression, Minimax rates, High-dimensional statistics, Adaptivity, Square-root Estimators.

1 Introduction

In a recent paper by Bellec, Lecué and Tsybakov [1], it is shown that there exist high-dimensional statistical methods realizable in polynomial time that achieve the minimax optimal rate $(s/n)\log(p/s)$ in the context of sparse linear regression. Here, n is the sample size, p is the dimension and s is the sparsity parameter. The result is achieved by the Lasso and Slope estimators, and the Slope estimator is adaptive to the unknown sparsity s. Bounds for more general estimators are proved by Bellec, Lecué and Tsybakov [2,3]. These articles also establish bounds in deviation that hold

*ENSAE-CREST, 5, avenue Henry Le Chatelier, TSA 96642, 91764 Palaiseau cedex, France. alexis.derumigny@ensae.fr

for any confidence level and for the risk in expectation. However, the estimators considered in [1–3] require the knowledge of the noise variance σ^2. To our knowledge, no polynomial-time methods, which would be at the same time optimal in a minimax sense and adaptive both to σ and s are available in the literature.

Estimators similar to the Lasso, but adaptive to σ are the Square-Root Lasso and the related Scaled Lasso, introduced by Sun and Zhang [13] and Belloni, Chernozhukov and Wang [4]. It has been shown to achieve the rate $(s/n)\log(p)$ in deviation with the value of the tuning parameter depending on the confidence level. A variant of this estimator is the Heteroscedastic Square-Root Lasso, which is studied in more general nonparametric and semiparametric setups by Belloni, Chernozhukov and Wang [5], but it also achieves the rate $(s/n)\log(p)$ and depends on the confidence level. We refer to the book by Giraud [8] for the link between the Lasso and the Square-Root Lasso and a short proof of oracle inequalities for the Square-root Lasso. In summary, there are two points to improve for the Square-root Lasso method:

(i) The available results on oracle inequalities are valid only for the estimators depending on the confidence level. Thus, one cannot have an oracle inequality for one given estimator at any confidence level except the one that was used to design it.

(ii) The obtained rate is $(s/n)\log(p)$ which is greater than the minimax rate $(s/n)\log(p/s)$.

The Slope, which is an acronym for Sorted L-One Penalized Estimation, is an estimator introduced by Bogdan et al. [7], that is close to the Lasso, but uses the sorted l_1 norm instead of the standard l_1 norm for penalization. Su and Candès [12] proved that, as opposed to the Lasso, the Slope estimator is asymptotically minimax, in the sense that it attains the rate $(s/n)\log(p/s)$ for two isotropic designs, that is either for $\mathbb{X}$ deterministic with $\frac{1}{n}\mathbb{X}^T\mathbb{X} = I_{p\times p}$ or when $\mathbb{X}$ is a matrix with i.i.d. standard normal entries. Moreover, their result has not only the optimal minimax rate, but also the exact optimal constant. General isotropic random designs are explored by Lecué and Mendelson [9]. For non-isotropic random designs and deterministic designs under conditions close to the Restricted Eigenvalue, the behavior of the Slope estimator is studied in [1]. The Slope estimator is adaptive only to s, and requires knowledge of σ, which is not available in practice. In order to have an estimator which is adaptive both to s and σ, we will use the Square-Root Slope, introduced by Stucky and van de Geer [11]. They give oracle inequalities for a large group of square-root estimators, including the new Square-Root Slope, but still following the scheme where (i) and (ii) cannot be avoided. The square-root estimators are also members of a more general family of penalized estimators defined by Owen [10, equations (8)-(9)] ; using their notation, these estimators correspond to the case where $\mathcal{H}_M$ is the squared loss and $\mathcal{B}_M$ is a norm (either the l_1 norm or the slope norm).

The paper is organized as follows. In Section 2, we provide the main definitions and notations. In Section 3, we show that the Square-Root Lasso is minimax optimal if s is known while being adaptive to σ under a mild condition on the design matrix (SRE). In Section 4, we show that any sequence of estimators can be made adaptive to the sparsity parameter s, while keeping

the same rate up to some constant, with a computational cost increased by a factor of $\log(s_*)$ where s_* is an upper bound on the sparsity parameter s. As an application, the Square-root Lasso modified by this procedure is still optimal while being now adaptive to s (in addition of being already adaptive to σ). In Section 5, we show how to adapt any algorithm for computing the Slope estimator to the case of the Square-root Slope estimator. In Section 6, we study the Square-Root Slope estimator, and show that it is minimax optimal and adaptive both to s and σ, under a slightly stronger condition (WRE). The (SRE) and (WRE) conditions have already been studied by Bellec, Lecué and Tsybakov [1] and hold with high probability for a large class of random matrices. Moreover, the inequalities we obtain for each estimator are valid for a wide range of confidence levels. Proofs are given in Section 7.

2 The framework

We use the notation $|\cdot|_q$ for the l_q norm, with $1 \leq q \leq \infty$, and $|\cdot|_0$ for the number of non-zero coordinates of a given vector. For any $v \in \mathbb{R}^p$, and any set of coordinates J, we denote by v_J the vector $(v_j \mathbb{1}\{i \in J\})_{i=1,\dots,p}$, where $\mathbb{1}$ is the indicator function. We also define the empirical norm of a vector $u = (u_1, \dots, u_n)$ as $||u||_n^2 := \frac{1}{n}\sum_{i=1}^n u_i^2$. For a vector $v \in \mathbb{R}^p$, we denote by $v_{(j)}$ the j-th largest component of v. As a particular case, $|v|_{(j)}$ is the j-th largest component of the vector $|v|$ whose components are the absolute values of the components of v. We use the notation $\langle \cdot, \cdot \rangle$ for the inner product with respect to the Euclidean norm and $(e_j)_{j=1,\dots,p}$ for the canonical basis in $\mathbb{R}^p$.

Let $Y \in \mathbb{R}^n$ be the vector of observations and let $\mathbb{X} \in \mathbb{R}^{n \times p}$ be the design matrix. We assume that the true model is the following

$$Y = \mathbb{X}\beta^* + \varepsilon. \tag{1}$$

Here $\beta^* \in \mathbb{R}^p$ is the unknown true parameter. We assume that ε is the random noise, with values in $\mathbb{R}^n$, distributed as $\mathcal{N}(0, \sigma^2 I_{n\times n})$, where $I_{n\times n}$ is the identity matrix. We denote by $\mathbb{P}_{\beta^*}$ the probability distribution of Y satisfying (1). In what follows, we define the set $B_0(s) := \{\beta^* \in \mathbb{R}^p : |\beta^*|_0 \leq s\}$. In the high-dimensional framework, we have typically in mind the case where s is small, p is large and possibly $p \gg n$.

We define two square-root type estimators of β^*: the Square-Root Lasso $\hat{\beta}^{SQL}$ and the Square-Root Slope $\hat{\beta}^{SQS}$ by the following relations

$$\hat{\beta}^{SQL} \in \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{\sqrt{n}} |Y - \mathbb{X}\beta|_2 + \lambda |\beta|_1 \right), \tag{2}$$

$$\hat{\beta}^{SQS} \in \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{\sqrt{n}} |Y - \mathbb{X}\beta|_2 + |\beta|_* \right), \tag{3}$$

where $\lambda > 0$ is a tuning parameter to be chosen, and the sorted l_1 norm, $|\cdot|_*$, is defined for all $u \in \mathbb{R}^p$ by $|u|_* = \sum_{i=1}^p \lambda_j |u|_{(j)}$, with tuning parameters $\lambda_1 \geq \cdots \geq \lambda_p > 0$.

3 Optimal rates for the Square-Root Lasso

In this section, we derive oracle inequalities with optimal rate for the Square-Root Lasso estimator. We will use the *Strong Restricted Eigenvalue* (SRE) condition, introduced in [1]. For $c_0 > 0$ and $s \in \{1, \dots, p\}$, it is defined as follows,

$SRE(s, c_0)$ **condition :** *The design matrix* $\mathbb{X}$ *satisfies* $\max_{j=1,\dots,p} ||\mathbb{X}e_j||_n \leq 1$ *and*

$$\kappa(s) := \min_{\delta \in C_{SRE}(s,c_0): \delta \neq 0} \frac{||\mathbb{X}\delta||_n}{|\delta|_2} > 0, \tag{4}$$

where $C_{SRE}(s, c_0) := \{\delta \in \mathbb{R}^p : |\delta|_1 \leq (1 + c_0)\sqrt{s}|\delta|_2\}$ *is a cone in* $\mathbb{R}^p$.

The condition $\max_{j=1,\dots,p} ||\mathbb{X}e_j||_n \leq 1$ is standard and corresponds to a normalization. It is shown in [1, Proposition 8.1] that the SRE condition is equivalent to the Restricted Eigenvalue (RE) condition of [6] if that is considered in conjunction with such a normalization. By the same proposition, the RE condition is also equivalent to the s-sparse eigenvalue condition, which is satisfied with high probability for a large class of random matrices. It is the case, if for instance, $n \geq Cs \log(ep/s)$ and the rows of $\mathbb{X}$ satisfies the small ball condition, which is very mild, see, e.g. [1].

Note that the minimum in (4) is the same as the minimum of the function $\delta \mapsto ||\mathbb{X}\delta||_n$ on the set $C_{SRE}(s, c_0) \cap \{\delta \in \mathbb{R}^p : |\delta|_2 = 1\}$, which is a continuous function on a compact of $\mathbb{R}^p$, therefore this minimum is attained. When there is no ambiguity over the choice of s, we will just write κ instead of $\kappa(s)$.

Theorem 3.1 *Let* $s \in \{1, \dots, p\}$ *and assume that the* $SRE(s, 5/3)$ *condition holds. Choose the following tuning parameter*

$$\lambda = \gamma \sqrt{\frac{1}{n} \log\left(\frac{2p}{s}\right)}, \tag{5}$$

and assume that

$$\gamma \geq 16 + 4\sqrt{2} \quad \text{and} \quad \frac{s}{n} \log\left(\frac{2p}{s}\right) \leq \frac{9\kappa^2}{256\gamma^2}. \tag{6}$$

Then, for every $\delta_0 \geq \exp(-n/4\gamma^2)$ *and every* $\beta^* \in \mathbb{R}^p$ *such that* $|\beta^*|_0 \leq s$, *with* $\mathbb{P}_{\beta^*}$*-probability at least* $1 - \delta_0 - (1 + e^2)e^{-n/24}$, *we have*

$$||\mathbb{X}(\hat{\beta}^{SQL} - \beta^*)||_n \leq \sigma \max\left(\frac{C_1}{\kappa^2}\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},\ C_2\sqrt{\frac{\log(1/\delta_0)}{n}}\right), \tag{7}$$

$$|\hat{\beta}^{SQL} - \beta^*|_q \leq \sigma \max\left(\frac{C_3}{\kappa^2} s^{1/q}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)},\ C_4 s^{1/q-1}\sqrt{\frac{\log^2(1/\delta_0)}{n\log(2p/s)}}\right), \tag{8}$$

where $1 \leq q \leq 2$, *and* $C_1 > 0$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$ *are constants depending only on* γ.

The values of the constants C_1, C_2, C_3 and C_4 in Theorem 3.1 can be found in the proof, in Section 7.2. Using the fact that $\kappa \leq 1$ and choosing $\delta_0 = (s/p)^s$, we get the following corollary of Theorem 3.1.

Corollary 3.2 *Under the assumptions of Theorem 3.1, with $\mathbb{P}_{\beta^*}$-probability at least $1-(s/p)^s-(1+e^2)e^{-n/24}$, we have*

$$||\mathbb{X}(\hat{\beta}^{SQL}-\beta^*)||_n \leq \frac{C_2}{\kappa^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},$$
$$|\hat{\beta}^{SQL}-\beta^*|_q \leq \frac{C_4}{\kappa^2}\sigma s^{1/q}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)},$$

where $1 \leq q \leq 2$.

Theorem 3.1 and Corollary 3.2 give bounds that hold with high probability for both the prediction error and the estimation error in the l_q norm, for every q in $[1,2]$. Note that the bounds are best when the tuning parameter is chosen as small as possible, i.e. with $\gamma = 16+4\sqrt{2}$. As shown in Section 7 of Bellec, Lecué and Tsybakov [1], the rates of estimation obtained in the latter corollary are optimal in a minimax sense on the set $B_0(s) := \{\beta^* \in \mathbb{R}^p : |\beta^*|_0 \leq s\}$. We obtain the same rate of convergence as [1] (see the paragraph after Corollary 4.3 in [1]) up to some multiplicative constant.

The rate is also the same as in Su and Candès [12], but the framework is quite different: we obtain a non-asymptotic bound in probability whereas they consider asymptotic bounds in expectation (cf. Theorem 1.1 in [12]) and in probability (Theorem 1.2) but without giving an explicit expression of the probability that their bound is valid. Our result is non-asymptotic and valid when general enough conditions on $\mathbb{X}$ are satisfied whereas the result in [12] is asymptotic as $n \to \infty$, and valid for two isotropic designs, that is either for $\mathbb{X}$ deterministic with $\frac{1}{n}\mathbb{X}^T\mathbb{X} = I_{p\times p}$ or when $\mathbb{X}$ is a matrix with i.i.d. standard normal entries.

Similarly to [1], for each tuning parameter γ, there is a wide range of levels of confidence δ_0 under which the bounds of Theorem 3.1 are valid. However, [1] allows for an arbitrary small confidence level while in our case, there is a lower bound on the size of the confidence level under which the rate is obtained. Note that this bound can be made arbitrary small by choosing a sample size n large enough.

Note that the possible values chosen for the tuning parameter λ are independent of the underlying standard deviation σ, which is unknown in practice. This gives an advantage for the Square-Root Lasso over other methods such as the ordinary Lasso. Nevertheless, this estimator is not adaptive to the sparsity s, so that we need to know that $|\beta^*|_0 \leq s$ in order to be able to apply this result. In the following section, we suggest a procedure to make the Square-root Lasso adaptive to s while keeping its optimality and adaptivity to σ.

4 Adaptation to sparsity by a Lepski-type procedure

Let s_* be an integer in $\{2,\dots,p/e\}$. We want to show that the Square-Root Lasso can also achieve the minimax optimal bound, adaptively to the sparsity s on the interval $[1,s_*]$ (in addition of being already adaptive to σ). Following [1], we will use aggregation of at most $\log_2(s_*)$ Square-Root Lasso estimators with different tuning parameters to construct an adaptive estimator $\tilde{\beta}$ of β and at the same time an estimator $\tilde{s}$ of the sparsity s.

In the following, we use the notation $\kappa_* := \kappa(2s_*)$. Note that $\kappa_* = \min_{s=1,\dots,2s_*} \kappa(s)$. Indeed, the function $\kappa(\cdot)$ is decreasing, because the minimization (4) is done on spaces that are growing with s, in the sense of the inclusion. We will assume that the condition $SRE(2s_*, 5/3)$ holds and that $(2s_*/n)\log\big(2p/(2s_*)\big) \leq 9\kappa_*^2/(256\gamma^2)$. The functions $b \mapsto (b/n)\log(2p/b)$ and $\kappa(\cdot)$ are respectively increasing (by Lemma 4.4) and decreasing, so this ensures that the second part of condition (6) is satisfied for any $s = 1, \dots, 2s_*$.

We can reformulate Corollary 3.2 as follows: for any $s = 1, \dots, 2s_*$ and any $\gamma \geq 16 + 4\sqrt{2}$

$$\sup_{\beta^* \in B_0(s)} \mathbb{P}_{\beta^*}\left(||\mathbb{X}(\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*)||_n \leq \frac{C_2(\gamma)}{\kappa_*^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)}\right) \geq 1 - \left(\frac{s}{p}\right)^s - (1+e^2)e^{-n/24}, \quad (9)$$

denoting by $\hat{\beta}^{SQL}_{(s,\gamma)}$ the estimator (2) with the tuning parameter $\lambda_{(s,\gamma)}$ given by (5). Replacing s by $2s$ in equation (9), we get that for any $s = 1, \dots, s_*$ and any $\gamma \geq 16 + 4\sqrt{2}$,

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\left(||\mathbb{X}(\hat{\beta}^{SQL}_{(2s,\gamma)} - \beta^*)||_n \leq \frac{C_2(\gamma)}{\kappa_*^2}\sigma\sqrt{\frac{2s}{n}\log\left(\frac{p}{2s}\right)}\right) \geq 1 - \left(\frac{2s}{p}\right)^{2s} - (1+e^2)e^{-n/24}. \quad (10)$$

Remark that $\lambda_{(s,\gamma)} = \gamma\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)} = \tilde{\gamma}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right) - \frac{\log(2)}{n}} = \lambda_{(2s,\tilde{\gamma})}$ for some $\tilde{\gamma} > \gamma$. As a consequence, $\hat{\beta}^{SQL}_{(s,\gamma)} = \hat{\beta}^{SQL}_{(2s,\tilde{\gamma})}$ and we can apply Equation (10), replacing γ by $\tilde{\gamma}$ and we get

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\left(||\mathbb{X}(\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*)||_n \leq \frac{C_2(\tilde{\gamma})}{\kappa_*^2}\sigma\sqrt{\frac{2s}{n}\log\left(\frac{p}{2s}\right)}\right) \geq 1 - \left(\frac{2s}{p}\right)^{2s} - (1+e^2)e^{-n/24}. \quad (11)$$

Note that equations (9) and (11) are the same as equations (5.2) and (5.4) in Bellec, Lecué and Tsybakov [1], taking $C_0 := \max\big(C_2(\gamma), C_2(\tilde{\gamma})\big)/\kappa_*^2$, except that we have a supplementary term $-(1+e^2)e^{-n/24}$. Similarly, we deduce from Corollary 3.2 that

$$\sup_{\beta^* \in B_0(s)} \mathbb{P}_{\beta^*}\left(|\mathbb{X}(\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*)|_q \leq \frac{C_4(\gamma)}{\kappa_*^2}\sigma s^{1/q}\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}\right) \geq 1 - \left(\frac{s}{p}\right)^s - (1+e^2)e^{-n/24}, \quad (12)$$

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\left(|\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*|_q \leq \frac{C_4(\tilde{\gamma})}{\kappa_*^2}\sigma s^{1/q}\sqrt{\frac{2s}{n}\log\left(\frac{2p}{2s}\right)}\right) \geq 1 - \left(\frac{2s}{p}\right)^{2s} - (1+e^2)e^{-n/24}. \quad (13)$$

We describe now an algorithm to compute this adaptive estimator. The idea is to use an estimator $\tilde{s}$ of s which can be written as $\tilde{s} := 2^{\tilde{m}}$ for some positive data-dependent integer $\tilde{m}$. We will use the notation $M := \max\{m \in \mathbb{N} : 2^m \leq s_*\}$, so that the number of estimators we consider in the aggregation is M.

The suggested procedure is detailed in Algorithm 1 below, with the distance $d(\beta, \beta') = ||\mathbb{X}(\beta - \beta')||_n$ or $d(\beta, \beta') = |\beta - \beta'|_q$ for $q \in [1, 2]$. It can be used for any family of estimators $(\hat{\beta}_{(s)})_{s=1,\dots s_*}$, and chooses the best one in terms of the distance $d(\cdot,\cdot)$, resulting in an aggregated estimator $\tilde{\beta}$. Note that the weight function $w(\cdot)$ used in the algorithm cannot depend on σ as in [1], i.e. to have the form $w(b) = C_0\sigma\sqrt{(b/n)\log(p/b)}$ (respectively

$w(b) = C_0 \sigma b^{1/q} \sqrt{(1/n)\log(p/b)}\,)$, because we are looking for a procedure adaptive to σ. Therefore, we will remove σ from w and use an estimate $\hat{\sigma}$.

Algorithm 1: Algorithm for adaptivity.

Input: a distance $d(\cdot,\cdot)$ on $\mathbb{R}^p$

Input: a function $w(\cdot) : [1, s_*] \to \mathbb{R}_+$ satisfying Assumption 4.1

Input: a family of estimators $\left(\hat{\beta}_{(s)}\right)_{s=1,\dots,s_*}$

$M \leftarrow \lfloor \log_2(s_*) \rfloor$;

for $m \leftarrow 1$ **to** $M+1$ **do**

  compute the estimator $\hat{\beta}_{(2^m)}$;

end

compute $\hat{\sigma} \leftarrow ||Y - \mathbb{X}\hat{\beta}_{(2^{M+1})}||_n$;

compute the set $S_1 \leftarrow \left\{ m \in \{1,\dots,M\} : d\Big(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\Big) \le 4\hat{\sigma} C_0 w(2^k), \text{ for all } k \ge m \right\}$;

if $S_1 \neq \emptyset$ **then** $\tilde{m} \leftarrow \min S_1$ **else** $\tilde{m} \leftarrow M$;

Output: $\tilde{s} \leftarrow 2^{\tilde{m}}$

Output: $\tilde{\beta} \leftarrow \hat{\beta}_{(\tilde{s})}$

Assumption 4.1 *The function $w(\cdot) : [1, s_*] \to \mathbb{R}_+$ satisfies the following conditions:*

1. *$w(\cdot)$ is increasing on $[1, s_*]$;*
2. *There exists a constant $C' > 0$ such that, for all $m = 1, \dots, M$, we have $\sum_{k=1}^{m} w(2^k) \le C' \cdot w(2^m)$;*
3. *There exists a constant $C'' > 0$ such that, for all $b = 1, \dots, s_*$, $w(2b) \le C'' w(b)$.*

Assumption 4.2 *The family of estimators $(\hat{\beta}_{(s)})_{s=1,\dots,s_*}$ satisfies*

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\Big(\sigma/2 \le \hat{\sigma} \le \alpha\sigma\Big) \le u_{n,p,M},$$

with a constant $\alpha > 0$, $\hat{\sigma} := ||Y - \mathbb{X}\hat{\beta}_{(2^{M+1})}||_n$, and $u_{n,p,M} > 0$.

Theorem 4.3 *Let $s_* \in \{2, \dots, p/e\}$ and let $(\hat{\beta}_{(s)})_{s=1,\dots,s_*}$ be a collection of estimators satisfying Assumption 4.2 such that, for any $s = 1, \dots, s_*$,*

$$\sup_{\beta^* \in B_0(s)} \mathbb{P}_{\beta^*}\left(d(\hat{\beta}_{(s)}\,,\,\beta^*) \le C_0 \sigma w(s)\right) \ge 1 - \left(\frac{s}{p}\right)^s - u_n, \tag{14}$$

and

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\left(d(\hat{\beta}_{(s)}\,,\,\beta^*) \le C_0 \sigma w(2s)\right) \ge 1 - \left(\frac{2s}{p}\right)^{2s} - u_n, \tag{15}$$

for a constant $C_0 > 0$, a function $w(\cdot) : [1, s_*] \to \mathbb{R}_+$ satisfying Assumption 4.1, and $u_n > 0$.

Then, there exists a constant C_5, depending on $C_0, C', C'', C_2, \kappa$ and α such that, for all $\beta^* \in B_0(s)$, the aggregated estimator $\tilde{\beta}$ satisfies:

$$\mathbb{P}_{\beta^*}\left(d(\tilde{\beta}, \beta^*) \leq C_5 \cdot \sigma w(s)\right) \geq 1 - 3(\log_2(s_*)+1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M}.$$

Furthermore,

$$\mathbb{P}_{\beta^*}\left(\tilde{s} \leq s\right) \geq 1 - 2(\log_2(s_*)+1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M}.$$

This theorem is proved in Section 7.3.1. In particular, it implies that when $\hat{\beta}_{(s)} = \hat{\beta}^{SQL}_{(s,\gamma)}$, the aggregated estimator $\tilde{\beta}$ has the same rate on $B_0(s)$ as the estimators with known s. We detail it below. The following lemmas proved in Sections 7.3.2 and 7.3.3 assure that Theorem 4.3 can be applied to the family $\hat{\beta}_{(s)} = \hat{\beta}^{SQL}_{(s,\gamma)}$.

Lemma 4.4 *Assumption 4.1 is satisfied with the choices $w(b) = \sqrt{(b/n)\log(p/b)}$ and $w(b) = b^{1/q}\sqrt{(1/n)\log(2p/b)}$, for $q \in [1,2]$.*

Lemma 4.5 *Assume that the $SRE(2s_*, 5/3)$ condition holds and*

$$\gamma \geq 16 + 4\sqrt{2} \quad \textit{and} \quad \frac{2s_*}{n}\log\left(\frac{p}{s_*}\right) \leq \min\left(\frac{9\kappa_*^2}{256\gamma^2}\,,\, \frac{\kappa_*^4}{2C_2(\gamma)^2}\left(\frac{1}{\sqrt{2}} - \frac{1}{2}\right)^2\right),$$

where $\kappa_* := \kappa(2s_*)$. Then Assumption 4.2 is satisfied with the choice $(\hat{\beta}_{(s)})_{s=1,\ldots,s_*} = (\hat{\beta}^{SQL}_{(s,\gamma)})_{s=1,\ldots,s_*}$, $\alpha = 2 + \frac{3\sqrt{2}C_2(\gamma)}{16\kappa\gamma}$ and $u_{n,p,M} = (2^{M+1}/p)^{2^{M+1}} - (1+e^2)e^{-n/24}$.

Combining equations (9), (11) with Theorem 4.3 and Lemmas 4.4 and 4.5, we obtain the following results for the case of the Square-root Lasso.

Corollary 4.6 *Under the same assumptions as in Lemma 4.5, using Algorithm 1, with $(\hat{\beta}_{(s)})_{s=1,\ldots,s_*} = (\hat{\beta}^{SQL}_{(s,\gamma)})_{s=1,\ldots,s_*}$, the distance $d(\beta,\beta') = ||\mathbb{X}(\beta - \beta')||_n$, and the weight $w(b) = \sqrt{(b/n)\log(p/b)}$, we have that, for all $\beta^* \in B_0(s)$, the aggregated estimator $\tilde{\beta}$ satisfies*

$$\mathbb{P}_{\beta^*}\left(||\mathbb{X}(\tilde{\beta} - \beta^*)||_n \leq C_5 \cdot \sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)}\right) \geq 1 - 3(\log_2(s_*)+1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

and

$$\mathbb{P}_{\beta^*}\left(\tilde{s} \leq s\right) \geq 1 - 2(\log_2(s_*)+1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

where $u_n = (1+e^2)e^{-n/24}$, $u_{n,p,M} = (2^{M+1}/p)^{2^{M+1}} - (1+e^2)e^{-n/24}$, and C_5 is a constant depending only on γ and κ_*.

Corollary 4.7 *Under the same assumptions as in Lemma 4.5, using Algorithm 1, with* $(\hat{\beta}_{(s)})_{s=1,\dots,s_*} = (\hat{\beta}^{SQL}_{(s,\gamma)})_{s=1,\dots,s_*}$*, the distance* $d(\beta,\beta') = |\beta-\beta'|_q$*, and the weight* $w(b) = b^{1/q}\sqrt{(1/n)\log(2p/b)}$*, for* $q \in [1;2]$*, we have that, for all* $\beta^* \in B_0(s)$*, the aggregated estimator* $\tilde{\beta}$ *satisfies*

$$\mathbb{P}_{\beta^*}\left(|\tilde{\beta}-\beta^*|_q \le C_5 \cdot \sigma s^{1/q}\sqrt{\frac{1}{n}\log\left(\frac{p}{s}\right)}\right) \ge 1 \,-\, 3(\log_2(s_*)+1)^2\left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

and

$$\mathbb{P}_{\beta^*}\left(\tilde{s} \le s\right) \ge 1 - 2(\log_2(s_*)+1)^2\left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

where $u_n = (1+e^2)e^{-n/24}$*,* $u_{n,p,M} = (2^{M+1}/p)^{2^{M+1}} - (1+e^2)e^{-n/24}$*, and* C_5 *is a constant depending only on* γ *and* κ_**.*

Thus, we have shown that the suggested aggregated procedure based on the Square-root Lasso is adaptive to s while still being adaptive to σ and minimax optimal. Note that the computational cost is multiplied by $O(\log(s_*))$.

5 Algorithms for computing the Square-root Slope

In this part, our goal is to provide algorithms for computing the square-root Slope estimator. A natural idea is revisiting the algorithms used for the square-root Lasso and for the Slope, then adapting or combining them.

Belloni, Chernozhukov and Wang [4, Section 4] have proposed to compute the Square-root Lasso estimator by reducing its definition to an equivalent problem, which can be solved by interior-point or first-order methods. The equivalent formulation as the Scaled Lasso, introduced by Sun and Zhang [13] allows one to view it as a joint minimization in (β,σ). Sun and Zhang [13] propose an iterative algorithm which alternates estimation of β using the ordinary Lasso and estimation of σ.

Zeng and Figueiredo [14] studied several algorithms related to estimation of the regression with the ordered weighted l_1-norm, which is the Slope penalization. Bogdan et al. [7] provide an algorithm for computing the Slope estimator using a proximal gradient.

As in the case of the Square-root Lasso, we still have for any β,

$$||Y-\mathbb{X}\beta||_n = \min_{\sigma>0}\left(\sigma + \frac{||Y-\mathbb{X}\beta||_n^2}{\sigma}\right), \tag{16}$$

where the minimum is attained for $\hat{\sigma} = ||Y-\mathbb{X}\beta||_n$. As a consequence,

$$\hat{\beta}^{SQS} \in \arg\min_{\beta\in\mathbb{R}^p}\left(||Y-\mathbb{X}\beta||_n + |\beta|_*\right)$$

is equivalent to take the estimator $\hat{\beta}$ in the joint minimization program

$$(\hat{\beta},\hat{\sigma}) \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p,\,\sigma>0}\left(\sigma + \frac{||Y-\mathbb{X}\beta||_n^2}{\sigma} + |\beta|_*\right).$$

Alternating minimization in β and in σ gives an iterative procedure for a "Scaled Slope" (see Algorithm 2).

Algorithm 2: Scaled Slope algorithm

Input: explained variable Y, design matrix $\mathbb{X}$;

Input: tuning parameters $\lambda_1 \leq \cdots \leq \lambda_p$;

choose some initialization value for $\hat{\sigma}$, for example the standard deviation of Y ;

repeat

estimate $\hat{\beta}$ by the Slope algorithm with the parameters $\hat{\sigma} \cdot \lambda_1, \ldots, \hat{\sigma} \cdot \lambda_p$;

estimate $\hat{\sigma}$ by $||Y - \mathbb{X}\hat{\beta}||_n$;

until *convergence*;

Output: a joint estimator $\left(\hat{\beta},\, \hat{\sigma}\right)$;

6 Optimal rates for the Square-Root SLOPE

In this part, we will use another condition, the *Weighted Restricted Eigenvalue* condition, introduced in [1]. For $c_0 > 0$ and $s \in \{1, \ldots, p\}$, it is defined as follows,

$WRE(s, c_0)$ **condition :** *The design matrix* $\mathbb{X}$ *satisfies* $\max_{j=1,\ldots,p} ||\mathbb{X}e_j||_n \leq 1$ *and*

$$\kappa' := \min_{\delta \in C_{WRE}(s,c_0):\delta \neq 0} \frac{||\mathbb{X}\delta||_n}{|\delta|_2} > 0, \tag{17}$$

where $C_{WRE}(s, c_0) := \{\delta \in \mathbb{R}^p : |\delta|_* \leq (1 + c_0)|\delta|_2 \sqrt{\sum_{j=1}^s \lambda_j^2}\}$ *is a cone in* $\mathbb{R}^p$.

To obtain the following result, we assume that the Weighted Restricted Eigenvalue condition holds. This condition is shown to be only slightly more constraining than the usual Restricted Eigenvalue condition of [6], but is nevertheless satisfied with high probability for a large class of random matrices, see Bellec, Lecué and Tsybakov [1] for a discussion. Note that, in a similar way as in definition (4), the minimum is attained. Indeed, κ' is equal to the minimum of the function $\delta \mapsto ||\mathbb{X}\delta||_n$ on the set $C_{WRE}(s, c_0) \cap \{\delta \in \mathbb{R}^p : |\delta|_2 = 1\}$, which is a continuous function on a compact of $\mathbb{R}^p$.

Theorem 6.1 *Let* $s \in \{1, \ldots, p\}$ *and assume that the* $WRE(s, 20)$ *condition holds. Choose the following tuning parameters*

$$\lambda_j = \gamma' \sqrt{\frac{\log(2p/j)}{n}},\ \textit{for } j = 1, \ldots, p, \tag{18}$$

and assume that

$$\gamma' \geq 16 + 4\sqrt{2} \quad \textit{and} \quad \frac{s}{n} \log\left(\frac{2ep}{s}\right) \leq \frac{\kappa'^2}{256\gamma'^2}. \tag{19}$$

Then, for every $\delta_0 \geq \exp(-n/4\gamma'^2)$ *and every* $\beta^* \in \mathbb{R}^p$ *such that* $|\beta^*|_0 \leq s$, *with* $\mathbb{P}_{\beta^*}$*-probability*

at least $1-\delta_0-(1+e^2)e^{-n/24}$*, we have*

$$||\mathbb{X}(\hat{\beta}^{SQS}-\beta^*)||_n \leq \sigma \max\left(\frac{C_1'}{\kappa'}\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)}\,,\, C_2'\sqrt{\frac{\log(1/\delta_0)}{n}}\right), \tag{20}$$

$$|\hat{\beta}^{SQS}-\beta^*|_* \leq \sigma \max\left(\frac{C_1'}{\kappa'^2}\frac{s}{n}\log\left(\frac{p}{s}\right)\,,\, C_2'\frac{\log(1/\delta_0)}{n}\right), \tag{21}$$

$$|\hat{\beta}^{SQS}-\beta^*|_2 \leq \sigma \max\left(\frac{C_1'}{\kappa'^2}\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)}\,,\, C_2'\sqrt{\frac{\log^2(1/\delta_0)}{sn\log(p/s)}}\right), \tag{22}$$

for constants $C_1'>0$ *and* $C_2'>0$ *depending only on* γ'*.*

The values of the constants C_1' and C_2' can be found in the proof, in Subsection 7.4. Note that the bounds are best when the tuning parameters is chosen as small as possible, i.e. using the choice $\gamma' = 16+4\sqrt{2}$. Using the fact that $\kappa' \leq 1$ and choosing $\delta_0 = (s/p)^s$, we get the following corollary.

Corollary 6.2 *Under the assumptions of Theorem 6.1, with* $\mathbb{P}_{\beta^*}$*-probability at least* $1-(s/p)^s-(1+e^2)e^{-n/24}$*, we have*

$$||\mathbb{X}(\hat{\beta}^{SQS}-\beta^*)||_n \leq \frac{C_1'}{\kappa'}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},$$

$$|\hat{\beta}^{SQS}-\beta^*|_* \leq \frac{C_1'}{\kappa'^2}\sigma\frac{s}{n}\log\left(\frac{p}{s}\right),$$

$$|\hat{\beta}^{SQS}-\beta^*|_2 \leq \frac{C_1'}{\kappa'^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},$$

These results show that the Square-Root Slope estimator, with a given choice of parameters, attains the optimal rate of convergence in the prediction norm $||\cdot||_n$ and in the estimation norm $|\cdot|_2$. We also provide a bound on the sorted l_1 norm $|\cdot|_*$ of the estimation error. One can note that the choice of λ_i that allows us to obtain optimal bounds does not depend on the level of confidence δ_0, but only influence the size of the range of valid δ_0. This improves upon the oracle result of Stucky and van de Geer [11], in which the parameter does depend on the level of confidence and the rate does not scale in the optimal way, i.e., as $\sqrt{(s/n)\log(p/s)}$. Moreover, we can see that our estimator is independent of the underlying standard deviation σ and of the sparsity s, even if the rates depend on them. Note that, up to some multiplicative constant, we obtain the same rates as for the Slope in Bellec, Lecué and Tsybakov [1]. In Su and Candès [12], the Slope estimator is proved to attain the sharp constant in the asymptotic framework where σ is known and for specific $\mathbb{X}$; whereas here we obtain only the minimax rates, but in a non-asymptotic framework, and under general assumptions on the design matrix $\mathbb{X}$.

For this estimator, we did not provide a bound for the l_1 norm, for the same reasons as in [1]. Indeed, the coefficients λ_j of the components of β are different in the sorted norm. As a consequence, we do not provide inequalities for l_q norms when $q<2$, that are obtained by interpolation between the l_1 and l_2 norms.

7 Proofs

7.1 Preliminary lemmas

Let $\beta^* \in \mathbb{R}^p$, $\mathcal{S} \subset \{1,\dots,p\}$ with cardinality s and denote by $\mathcal{S}^C$ the complement of $\mathcal{S}$. For $i \in \{1,\dots,p\}$, let β_i^* be the i-th component of β^* and assume that for every $i \in \mathcal{S}^C$, $\beta_i^* = 0$.

Lemma 7.1 *We have* $|(\hat{\beta}^{SQL} - \beta^*)_{\mathcal{S}^C}|_1 \leq |(\hat{\beta}^{SQL} - \beta^*)_{\mathcal{S}}|_1 + \dfrac{1}{\lambda\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon\ ,\ \hat{\beta}^{SQL} - \beta^* \right\rangle.$

The proof follows from the arguments in Giraud [8, pages 110-111], and it is therefore omitted.

Lemma 7.2 *Let* $u \in \mathbb{R}^p$ *be defined by* $u := \hat{\beta}^{SQS} - \beta^*$. *We have*

$$\sum_{j=s+1}^{p} \lambda_j |u|_{(j)} \leq \sum_{j=1}^{s} \lambda_j |u|_{(j)} + \frac{1}{\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon\ ,\ u \right\rangle.$$

Proof : We combine the arguments from Giraud [8, pages 110-111], and from the proof of Lemma A.1 in [1]. First, we remark that the sorted l_1 norm can be written as follows, for any $v \in \mathbb{R}^p$,

$$|v|_* = \max_{\phi} \sum_{j=1}^{p} \lambda_j \left|v_{\phi(j)}\right|,$$

where the maximum is taken over all permutations $\phi = (\phi(1),\dots,\phi(p))$ of $\{1,\dots,p\}$.

By definition, $\hat{\beta}^{SQS}$ is a minimizer of (3), so we have

$$|Y - \mathbb{X}\hat{\beta}^{SQS}|_2 - |Y - \mathbb{X}\beta^*|_2 \leq \sqrt{n}\left(|\beta^*|_* - |\hat{\beta}^{SQS}|_*\right).$$

Let ϕ be any permutation of $\{1,\dots,p\}$ such that

$$|\beta^*|_* = \sum_{j=1}^{s} \lambda_j |\beta^*_{\phi(j)}| \ \text{ and } \ |u_{\phi(s+1)}| \geq |u_{\phi(s+2)}| \geq \cdots \geq |u_{\phi(p)}|. \tag{23}$$

We have

$$\begin{aligned}
|\beta^*|_* - |\hat{\beta}^{SQS}|_* &\leq \sum_{j=1}^{s} \lambda_j \left(\left|\beta^*_{\phi(j)}\right| - \left|\hat{\beta}^{SQS}_{\phi(j)}\right|\right) - \sum_{j=s+1}^{p} \lambda_j \left|\hat{\beta}^{SQS}_{\phi(j)}\right| \\
&\leq \sum_{j=1}^{s} \lambda_j \left|u_{\phi(j)}\right| - \sum_{j=s+1}^{p} \lambda_j \left|\hat{\beta}^{SQS}_{\phi(j)}\right| = \sum_{j=1}^{s} \lambda_j \left|u_{\phi(j)}\right| - \sum_{j=s+1}^{p} \lambda_j \left|u_{\phi(j)}\right|.
\end{aligned}$$

Since the sequence λ_j is non-increasing, we have $\sum_{j=1}^{s} \lambda_j |u_{\phi(j)}| \leq \sum_{j=1}^{s} \lambda_j |u|_{(j)}$. The permutation ϕ satisfies (23), therefore, $\sum_{j=s+1}^{p} \lambda_j |u|_{(j)} \leq \sum_{j=s+1}^{p} \lambda_j |u_{\phi(j)}|$. From the previous inequalities, we get that

$$|Y - \mathbb{X}\hat{\beta}^{SQS}|_2 - |Y - \mathbb{X}\beta^*|_2 \leq \sqrt{n}\left(\sum_{j=1}^{s} \lambda_j |u|_{(j)} - \sum_{j=s+1}^{p} \lambda_j |u|_{(j)}\right). \tag{24}$$

By convexity of the mapping $\beta \mapsto ||Y - X\beta||_2$, we have

$$|Y - \mathbb{X}\hat{\beta}^{SQS}|_2 - |Y - \mathbb{X}\beta^*|_2 \geq -\left\langle \frac{\mathbb{X}^T\varepsilon}{|\varepsilon|_2} \ , \ \hat{\beta}^{SQS} - \beta^* \right\rangle = -\frac{1}{|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon \ , \ \hat{\beta}^{SQS} - \beta^* \right\rangle. \quad (25)$$

Combining (24) and (25), we get

$$-\frac{1}{|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon \ , \ \hat{\beta}^{SQS} - \beta^* \right\rangle \leq \sqrt{n}\left(\sum_{j=1}^{s}\lambda_j|u|_{(j)} - \sum_{j=s+1}^{p}\lambda_j|u|_{(j)}\right),$$

which concludes the proof.

□

Lemma 7.3 *We have* $|\mathbb{X}(\hat{\beta}^{SQL} - \beta^*)|_2^2 \leq \left\langle \mathbb{X}^T\varepsilon \ , \ \hat{\beta}^{SQL} - \beta^* \right\rangle + \lambda\sqrt{n}|Y - \mathbb{X}\hat{\beta}^{SQL}|_2|\hat{\beta}^{SQL} - \beta^*|_1$.

Lemma 7.4 *We have* $|\mathbb{X}(\hat{\beta}^{SQS} - \beta^*)|_2^2 \leq \left\langle \mathbb{X}^T\varepsilon \ , \ \hat{\beta}^{SQS} - \beta^* \right\rangle + \sqrt{n}|Y - \mathbb{X}\hat{\beta}|_2|\hat{\beta}^{SQS} - \beta^*|_*$.

Proof : We will give a general proof of Lemmas 7.3 and 7.4 in the case of an estimator defined by

$$\hat{\beta} := \arg\min_{\beta\in\mathbb{R}^p}\left(\frac{1}{\sqrt{n}}|Y - \mathbb{X}\beta|_2 + ||\beta||\right), \quad (26)$$

where $||\cdot||$ is a norm on $\mathbb{R}^p$. Lemmas 7.3 and 7.4 are obtained as special cases corresponding to $||\cdot|| = \lambda|\cdot|_1$ and $||\cdot|| = |\cdot|_*$. Denote by $||\cdot||_{dual}$ the norm dual to $||\cdot||$.

Since $\hat{\beta}$ is optimal, we know that $\mathbb{X}^T(Y - \mathbb{X}\hat{\beta})/(\sqrt{n}|Y - \mathbb{X}\hat{\beta}|_2)$ belongs to the subdifferential of the function $||\cdot||$ evaluated at $\hat{\beta}$. Thus, there exists $v \in \mathbb{R}^p$ such that $||v||_{dual} \leq 1$ and

$$\frac{\mathbb{X}^T(Y - \mathbb{X}\hat{\beta})}{\sqrt{n}|Y - \mathbb{X}\hat{\beta}|_2} + v = 0.$$

Thus, we have

$$|\mathbb{X}(\hat{\beta} - \beta^*)|_2^2 = \left\langle \mathbb{X}^T\varepsilon \ , \ \hat{\beta} - \beta^* \right\rangle + \sqrt{n}|Y - \mathbb{X}\hat{\beta}^*|_2\langle v \ , \ \hat{\beta} - \beta^*\rangle.$$

The conclusion results from the inequality

$$\langle v \ , \ \hat{\beta} - \beta^*\rangle \leq ||v||_{dual}||\hat{\beta} - \beta^*|| \leq ||\hat{\beta} - \beta^*||.$$

□

Lemma 7.5 *We have* $\gamma'\sqrt{(s/n)\log(2p/s)} \leq \sqrt{\sum_{j=1}^{s}\lambda_j^2} \leq \gamma'\sqrt{(s/n)\log(2ep/s)}$.

Proof : From Stirling's formula, we deduce that $s\log(s/e) \leq \log(s!) \leq s\log(s)$. Therefore

$$s\log(2p/s) \leq \sum_{j=1}^{s}\log(2p/j) = \log(2p) - \log(s!) \leq s\log(2ep/s).$$

The conclusion follows from the definition of the λ_j in (18).

□

The following simple property is proved in Giraud [8, page 112]. For convenience, it is stated here as a lemma.

Lemma 7.6 *With $\mathbb{P}_{\beta^*}$-probability at least $1-(1+e^2)e^{-n/24}$, we have*

$$\frac{\sigma}{\sqrt{2}} \leq \frac{|\varepsilon|_2}{\sqrt{n}} \leq 2\sigma.$$

We will also use the following theorem from Bellec, Lecué and Tsybakov [1, Theorem 4.1].

Lemma 7.7 *Let $0<\delta_0<1$ and let $\mathbb{X}$ in $\mathbb{R}^{n\times p}$ be a matrix such that $\max_{j=1,\dots,p}||\mathbb{X}e_j||_n \leq 1$. For any $u=(u_1,\dots u_p)$ in $\mathbb{R}^p$, we define :*

$$G(u) := (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n, \quad H(u) := (4+\sqrt{2})\sum_{j=1}^{p}|u|_{(j)}\sigma\sqrt{\frac{\log(2p/j)}{n}},$$

$$\text{and} \quad F(u) := (4+\sqrt{2})\sigma\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right).$$

If $\varepsilon \sim \mathcal{N}(0,\sigma^2 I_{n\times n})$, then the random event

$$\left\{\frac{1}{n}\varepsilon^T\mathbb{X}u \leq \max\Big(H(u),G(u)\Big), \forall u\in\mathbb{R}^p\right\},$$

is of probability at least $1-\delta_0/2$.

Moreover, by the Cauchy-Schwarz inequality, we have $H(u)\leq F(u)$, for all u in $\mathbb{R}^p$.

7.2 Proof of Theorem 3.1

Lemma 7.7 allows one to control the random variable $\varepsilon^T\mathbb{X}u$ that appears in Lemmas 7.1 and 7.3 with $u := \hat{\beta}^{SQL}-\beta^*$. Our calculations will take place on an event of probability at least $1-\delta_0-(1+e^2)e^{-n/24}$, where both Lemmas 7.6 and 7.7 can be used. Applying Lemma 7.7, we will distinguish between the two cases : $G(u)\leq F(u)$ and $F(u)<G(u)$.

First case : $G(u)\leq F(u)$.

Then we have

$$(4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n \leq (4+\sqrt{2})\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right).$$

We will show first that u is in the SRE cone, so that we can use the SRE assumption. From Lemma 7.1, we have

$$|u_{\mathcal{S}^c}|_1 \leq |u_{\mathcal{S}}|_1 + \frac{1}{\lambda\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon\ ,\ \hat{\beta}^{SQL}-\beta^*\right\rangle$$

$$\leq |u_{\mathcal{S}}|_1 + \frac{1}{\sqrt{n}\lambda|\varepsilon|_2} n\sigma(4+\sqrt{2})\sqrt{\frac{\log(2p/s)}{n}} \left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p} |u|_{(j)}\right)$$
$$\leq |u_{\mathcal{S}}|_1 + \frac{1}{4}\left(\sqrt{s}|u|_2 + |u_{\mathcal{S}^C}|_1\right),$$

where in the last inequality, we have used Lemma 7.6 and assumption (6). We deduce that

$$\frac{3}{4}|u|_1 \leq \frac{7}{4}|u_{\mathcal{S}}|_1 + \frac{1}{4}\sqrt{s}|u|_2 \leq \frac{7}{4}\sqrt{s}|u|_2 + \frac{1}{4}\sqrt{s}|u|_2 = 2\sqrt{s}|u|_2.$$

Therefore, we have

$$|u|_1 \leq \frac{8}{3}\sqrt{s}|u|_2, \tag{27}$$

and thus, the following inequality holds $|u|_1 \leq (1+c_0)\sqrt{s}|u|_2$, with $c_0 = 5/3$, allowing us to use the $SRE(s, 5/3)$ assumption.

From Lemmas 7.3 and 7.7, and using that, in view of the $SRE(s, 5/3)$ condition, $||\mathbb{X}u||_n \geq \kappa|u|_2$, we deduce that

$$||\mathbb{X}u||_n^2 \leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(2p/s)}{n}} \left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p} |u|_{(j)}\right) + \left(\frac{|\varepsilon|_2}{\sqrt{n}} + ||\mathbb{X}u||_n\right)\frac{8}{3}\lambda\sqrt{s}|u|_2$$
$$\leq (4+\sqrt{2})\frac{11}{3}\sigma\sqrt{s\frac{\log(2p/s)}{n}}\frac{||\mathbb{X}u||_n}{\kappa} + (2\sigma + ||\mathbb{X}u||_n)\frac{8}{3}\lambda\sqrt{s}\frac{||\mathbb{X}u||_n}{\kappa}.$$

Thus,

$$||\mathbb{X}u||_n \leq (4+\sqrt{2})\frac{11}{3}\sigma\sqrt{s\frac{\log(2p/s)}{n}}\frac{1}{\kappa} + (2\sigma + ||\mathbb{X}u||_n)\frac{8}{3}\lambda\sqrt{s}\frac{1}{\kappa}.$$

Under assumptions (5) and (6), we have

$$\frac{8\lambda\sqrt{s}}{3\kappa} = \frac{8\gamma}{3\kappa}\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)} \leq \frac{1}{2}.$$

Thus, we have

$$||\mathbb{X}u||_n \leq 2\left(\frac{44+11\sqrt{2}}{3\kappa}\sigma\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)} + \frac{16\sigma\lambda\sqrt{s}}{3\kappa}\right)$$
$$\leq \frac{88+22\sqrt{2}+32\gamma}{3\kappa}\sigma\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}. \tag{28}$$

We have proved in (27) that $|u|_1 \leq (1+c_0)\sqrt{s}|u|_2$, with $c_0 = 5/3$, so we get that $|u|_2 \leq ||\mathbb{X}u||_n/\kappa$. Therefore, we can deduce the following inequalities

$$|u|_2 \leq \frac{88+22\sqrt{2}+32\gamma}{3\kappa^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}, \tag{29}$$

$$|u|_1 \leq \frac{704+176\sqrt{2}+256\gamma}{9\kappa^2}\sigma s\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)}. \tag{30}$$

Second case : $F(u) \leq G(u)$.

Then we have

$$(4+\sqrt{2})\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2+\sum_{j=s+1}^{p}|u|_{(j)}\right)\leq(4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n.$$

Thus

$$|u|_1\leq\sqrt{s}|u|_2+\sum_{j=s+1}^{p}|u|_{(j)}\leq\sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n.$$

From Lemmas 7.3 and 7.7, we find

$$\begin{aligned}||\mathbb{X}u||_n^2&\leq(4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n+\lambda\left(\frac{|\varepsilon|_2}{\sqrt{n}}+||\mathbb{X}u||_n\right)|u|_1\\&\leq(4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n+\lambda\left(2\sigma+||\mathbb{X}u||_n\right)\sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n.\end{aligned}$$

Thus,

$$||\mathbb{X}u||_n\leq(4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}+\lambda\left(2\sigma+||\mathbb{X}u||_n\right)\sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}.$$

We have chosen $\lambda=\gamma\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)}$, therefore we have

$$||\mathbb{X}u||_n\leq\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}(4+\sqrt{2}+2\gamma)+||\mathbb{X}u||_n\gamma\sqrt{\frac{\log(1/\delta_0)}{n}}.$$

By assumption, $\exp(-n/4\gamma^2)\leq\delta_0$, thus we have

$$||\mathbb{X}u||_n\leq\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}(8+2\sqrt{2}+4\gamma).\tag{31}$$

As a consequence, we have

$$|u|_1\leq\sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n\leq\sigma\sqrt{\frac{\log^2(1/\delta_0)}{n\log(2p/s)}}(8+2\sqrt{2}+4\gamma).\tag{32}$$

We have also $\sqrt{s}|u|_2\leq\sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n$, thus

$$|u|_2\leq\sigma\sqrt{\frac{\log^2(1/\delta_0)}{sn\log(2p/s)}}(8+2\sqrt{2}+4\gamma).\tag{33}$$

As a conclusion, we can prove the result (7) by combining the inequalities (28) and (31). The general bound for $|u|_q$, with $1\leq q\leq 2$ is a consequence of the norm interpolation inequality $|u|_q\leq|u|_1^{2/q-1}|u|_2^{2-2/q}$ which proves (8).

□

7.3 Proofs of the adaptive procedure

7.3.1 Proof of Theorem 4.3

We choose $s\in[1,s_*]$ and assume that $\beta^*\in B_0(s)$. Define $\mathbb{P}:=\mathbb{P}_{\beta^*}$ and $m_0:=\lfloor\log_2(s)\rfloor+1$.

For any $a > 0$, we have

$$\mathbb{P}\big(d(\tilde{\beta}, \beta^*) \geq a\big) \leq \mathbb{P}\big(d(\tilde{\beta}, \beta^*) \geq a, \tilde{m} \leq m_0\big) + \mathbb{P}(\tilde{m} \geq m_0 + 1). \tag{34}$$

On the event $\{\tilde{m} \leq m_0\}$, we have the decomposition

$$d(\tilde{\beta}, \beta^*) \leq \sum_{k=\tilde{m}+1}^{m_0} d\left(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\right) + d\big(\hat{\beta}_{(2^{m_0})}, \beta^*\big). \tag{35}$$

Using Assumption 4.1, we get that,

$$\sum_{k=\tilde{m}+1}^{m_0} d\left(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\right) \leq \sum_{k=\tilde{m}+1}^{m_0} 4\hat{\sigma} C_0 w(2^k) \leq 4\hat{\sigma} C_0 C' w(2^{m_0}) \leq 4\hat{\sigma} C_0 C' C'' w(s). \tag{36}$$

We have $2^{m_0} \leq 2s$, therefore applying Assumption (15), we have with $\mathbb{P}_{\beta^*}$-probability at least $1 - (2s/p)^{2s} - u_n$,

$$d(\hat{\beta}_{(2^{m_0})}, \beta^*) \leq \frac{C_2(\tilde{\gamma})}{\kappa^2} \sigma w(2s) \leq \frac{C_2(\tilde{\gamma}) C''}{\kappa^2} \sigma w(s). \tag{37}$$

Combining equations (35), (36), (37) and Assumption 4.2, we get with $\mathbb{P}_{\beta^*}$-probability at least $1 - (2s/p)^{2s} - u_n - u_{n,p,M}$,

$$d(\tilde{\beta}, \beta^*) \leq \left(4\sigma C_0 C' C'' \alpha + \frac{C_2(\tilde{\gamma}) C''}{\kappa^2}\right) \sigma w(s). \tag{38}$$

We now bound the probability $\mathbb{P}(\tilde{m} \geq m_0 + 1)$.

$$\begin{aligned}
\mathbb{P}(\tilde{m} \geq m_0 + 1) &\leq \sum_{m=m_0+1}^{M} \mathbb{P}(\tilde{m} = m_0 + 1) \leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(d\Big(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\Big) > 4\hat{\sigma} C_0 w(2^k)\right) \\
&\leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(d\Big(\hat{\beta}_{(2^{k-1})}, \beta^*\Big) > 2\hat{\sigma} C_0 w(2^k)\right) + \mathbb{P}\left(d\Big(\hat{\beta}_{(2^k)}, \beta^*\Big) > 2\hat{\sigma} C_0 w(2^k)\right) \\
&\leq 2 \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(d\Big(\hat{\beta}_{(2^{k-1})}, \beta^*\Big) > 2\hat{\sigma} C_0 w(2^k)\right) \\
&\leq 2 \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(d\Big(\hat{\beta}_{(2^{k-1})}, \beta^*\Big) > 2\hat{\sigma} C_0 w(2^k), \hat{\sigma} \geq \frac{\sigma}{2}\right) + \mathbb{P}\left(\hat{\sigma} < \frac{\sigma}{2}\right).
\end{aligned}$$

Combining the previous equation with Assumption 4.2, and then with Assumption (15), we get

$$\begin{aligned}
\mathbb{P}(\tilde{m} \geq m_0 + 1) &\leq 2 \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(d\Big(\hat{\beta}_{(2^{k-1})}, \beta^*\Big) > \sigma C_0 w(2^k)\right) - u_{n,p,M} \\
&\leq 2M^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M} \\
&\leq 2(\log_2(s_*) + 1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M}.
\end{aligned}$$

As a consequence, we deduce the bound on $\tilde{s}$. Combining the last equation with equations (34) and (38), we finally get that

$$\begin{aligned}
\mathbb{P}\left(d(\tilde{\beta}, \beta^*) \geq \left(4\sigma C_0 C' C'' \alpha + \frac{C_2(\tilde{\gamma}) C''}{\kappa^2}\right) \sigma w(s)\right) & \\
\leq 3(\log_2(s_*) + 1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - 2u_{n,p,M}.
\end{aligned}$$

□

7.3.2 Proof of Lemma 4.4

Now, we consider the general case of the function $w(b) = b^{1/q}\sqrt{(1/n)\log(ap/b)}$, with q a fixed number of the interval $[1, 2]$. The first case will correspond to $a = 1$ and $q = 2$ and the second case will correspond to $a = 2$ with any choice of q.

We want to that the first part of Assumption 4.1 is satisfied, i.e., w is increasing on the interval $[1, s_*]$. Let $b \in [1, s_*]$. We have

$$\begin{aligned} w'(b) &= \frac{1}{q} b^{(1/q)-1} \sqrt{\frac{1}{n} \log\left(\frac{ap}{b}\right)} + b^{(1/q)} \frac{-\frac{1}{nb}}{2\sqrt{\frac{1}{n}\log\left(\frac{ap}{b}\right)}} \\ &= \frac{b^{(1/q)-1} n^{-1/2} \left((2/q)\log\left(\frac{ap}{b}\right) - 1\right)}{2\sqrt{\log\left(\frac{ap}{b}\right)}}, \end{aligned}$$

which is positive when $(2/q)\log\left(\frac{ap}{b}\right) - 1 \geq 0$, that is, when $b \leq ape^{-q/2}$.

We have $b \leq s_* \leq p/e = ape^{-q/2}$ when $a = 1$ and $q = 2$. When $a = 2$ and $q \in [1, 2]$, $p/e \leq 2pe^{-1} \leq ape^{-q/2}$. In the two cases we consider, we have proved that $w'(\cdot) \geq 0$ on the interval $[1, s_*]$, thus the function w is increasing on this interval. This proves that the first part of Assumption 4.1 is satisfied.

Let m be an integer in the interval $[1, M]$.

$$\begin{aligned} \sum_{k=1}^{m} w(2^k) &= \sum_{k=1}^{m} 2^{k/q} \sqrt{\frac{1}{n}\log\left(\frac{ap}{2^k}\right)} = \sum_{k=0}^{m-1} 2^{(m-k)/q} \sqrt{\frac{1}{n}\log\left(\frac{ap}{2^{m-k}}\right)} \\ &= \frac{2^{m/q}}{\sqrt{n}} \sum_{k=0}^{m-1} \frac{1}{2^{k/q}} \sqrt{\left(\log\left(\frac{ap}{2^m}\right) + k\log(2)\right)} \\ &\leq \frac{2^{m/q}}{\sqrt{n}} \left(\sum_{k=0}^{m-1} \frac{1}{2^{k/q}} \sqrt{\log\left(\frac{ap}{2^m}\right)} + \sum_{k=0}^{m-1} \frac{\sqrt{k}}{2^{k/q}} \sqrt{\log(2)} \right) \\ &\leq \frac{2^{m/q}}{\sqrt{n}} \left(\sqrt{\log\left(\frac{ap}{2^m}\right)} \frac{1}{1 - 2^{-1/q}} + \sum_{k=0}^{m-1} \frac{4}{2^{k/2q}} \sqrt{\log(2)} \right) \\ &\leq 2^{m/q} \sqrt{\frac{1}{n}\log\left(\frac{ap}{2^m}\right)} \left(\frac{1}{1 - 2^{-1/q}} + \frac{4\sqrt{\log(2)}}{1 - 2^{-1/(2q)}} \right), \end{aligned}$$

which proves that the second part is satisfied.

Let b be an integer of $[1, s_*]$. We have $w(2b) = (2b)^{1/q}\sqrt{(1/n)\log(2p/(2b))} \leq 2^{1/q} w(b)$, which proves that the third part is satisfied.

□

7.3.3 Proof of Lemma 4.5

We have $\beta^* \in B_0(s) \subset B_0(2^{M+1})$, therefore, we can apply Corollary 3.2 and Lemma 7.6, we have with $\mathbb{P}_{\beta^*}$-probability at least $1 - (2^{M+1}/p)^{2^{M+1}} - (1 + e^2)e^{-n/24}$

$$\begin{aligned} \hat{\sigma} &\leq ||\varepsilon||_n + \left|\left|\mathbb{X}(\hat{\beta}_{(2^{M+1})} - \beta^*)\right|\right|_n \leq 2\sigma + \frac{C_2(\gamma)}{\kappa_*^2} \sigma \sqrt{\frac{2^{M+1}}{n} \log\left(\frac{p}{2^{M+1}}\right)} \\ &\leq \sigma \left(2 + \frac{C_2(\gamma)}{\kappa_*^2} \sqrt{\frac{2s}{n}\log\left(\frac{2p}{s}\right)} \right) \leq \sigma \left(2 + \frac{3\sqrt{2}C_2(\gamma)}{16\kappa_*\gamma} \right), \end{aligned}$$

$$\hat{\sigma} \geq ||\varepsilon||_n - \left|\left|\mathbb{X}(\hat{\beta}_{(2^{M+1})} - \beta^*)\right|\right|_n \geq \frac{\sigma}{\sqrt{2}} - \frac{C_2(\gamma)}{\kappa_*^2}\sigma\sqrt{\frac{2^{M+1}}{n}\log\left(\frac{p}{2^{M+1}}\right)}$$
$$\geq \sigma\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{2}C_2(\gamma)}{\kappa_*^2}\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}\right)$$
$$\geq \sigma\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{2}C_2(\gamma)}{\kappa_*^2}\sqrt{\frac{2s_*}{n}\log\left(\frac{p}{s_*}\right)}\right)$$
$$\geq \sigma\left(\frac{1}{\sqrt{2}} - \sqrt{\left(\frac{1}{\sqrt{2}} - \frac{1}{2}\right)^2}\right) \geq \frac{\sigma}{2}.$$

□

7.4 Proof of Theorem 6.1

We act as in Section 7.2, with suitable modifications. We place ourselves in the event where both Lemmas 7.6 and 7.7 are valid, and set now $u := \hat{\beta}^{SQS} - \beta^*$. Applying Lemma 7.7, we will distinguish between the two cases : $G(u) \leq H(u) + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$ and $H(u) + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} < G(u)$.

First case : $G(u) \leq H(u) + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$.

Applying Lemma 7.2, Lemma 7.7 and then Lemma 7.6, we have

$$|u|_* = \sum_{j=1}^p \lambda_j |u|_{(j)} \leq 2\sum_{j=1}^s \lambda_j |u|_{(j)} + \frac{1}{\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon\ ,\ \hat{\beta}^{SQS} - \beta^*\right\rangle$$
$$\leq 2\sqrt{\sum_{j=1}^s \lambda_j^2}|u|_2 + \frac{n}{\sqrt{n}|\varepsilon|_2}\left((4+\sqrt{2})\frac{\sigma}{\gamma'}|u|_* + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}\right)$$
$$\leq 4\sqrt{\sum_{j=1}^s \lambda_j^2}|u|_2 + \frac{8+2\sqrt{2}}{\gamma'}|u|_*,$$

and we get

$$|u|_* \leq \frac{4|u|_2}{1 - \frac{8+2\sqrt{2}}{\gamma'}}\sqrt{\sum_{j=1}^s \lambda_j^2},$$

Using assumption (19), we have $\gamma' \geq 16 + 4\sqrt{2}$, therefore $|u|_* \leq 8|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$. As a consequence, we get $u \in C_{WRE}(s, c_0)$ with $c_0 := 8$. Invoking Lemmas 7.4, 7.5, 7.7 and using the $WRE(s, c_0)$ condition, we get

$$||\mathbb{X}u||_n^2 \leq \frac{1}{n}\left\langle \mathbb{X}^T\varepsilon\ ,\ u\right\rangle + \frac{1}{\sqrt{n}}|Y - \mathbb{X}\hat{\beta}|_2|u|_*$$
$$\leq (4+\sqrt{2})\frac{\sigma}{\gamma'}|u|_* + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} + (2\sigma + ||\mathbb{X}u||_n)|u|_*$$
$$\leq \left((32+8\sqrt{2})\frac{\sigma}{\gamma'} + 17\sigma + 8||\mathbb{X}u||_n\right)|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$$

$$\leq \left((32+8\sqrt{2})\frac{\sigma}{\gamma'}+17\sigma+8||\mathbb{X}u||_n\right)\frac{||\mathbb{X}u||_n}{\kappa'}\gamma'\sqrt{(s/n)\log(2ep/s)}.$$

Thus,

$$||\mathbb{X}u||_n \leq \frac{\sigma}{\kappa'}\sqrt{\frac{s}{n}\log\left(\frac{2ep}{s}\right)}\frac{32+8\sqrt{2}+17\gamma'}{1-\frac{8\gamma'}{\kappa'}\sqrt{\frac{s}{n}\log\left(\frac{2ep}{s}\right)}}.$$

Applying condition (19), we obtain

$$||\mathbb{X}u||_n \leq (64+16\sqrt{2}+34\gamma')\frac{\sigma}{\kappa'}\sqrt{\frac{s}{n}\log\left(\frac{2ep}{s}\right)}. \tag{39}$$

This and the WRE condition imply

$$|u|_2 \leq (64+16\sqrt{2}+34\gamma')\frac{\sigma}{\kappa'^2}\sqrt{\frac{s}{n}\log\left(\frac{2ep}{s}\right)}. \tag{40}$$

Therefore, using the inequality $|u|_* \leq 8|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$, we get from Lemma 7.5

$$|u|_* \leq 8(64+16\sqrt{2}+34\gamma')\gamma'\frac{\sigma}{\kappa'^2}\frac{s}{n}\log\left(\frac{2ep}{s}\right). \tag{41}$$

Second case : $H(u)+\sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} \leq G(u)$.

Then we have

$$(4+\sqrt{2})\frac{\sigma}{\gamma'}|u|_* + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} \leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n.$$

Therefore we have

$$|u|_* \leq \gamma'\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n, \quad \text{and} \quad |u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} \leq (4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n. \tag{42}$$

Invoking Lemmas 7.4 and 7.7, and using (42), we get

$$||\mathbb{X}u||_n^2 \leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} + (2\sigma+||\mathbb{X}u||_n)|u|_*$$

$$\leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n + \sigma(4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n + (2\sigma+||\mathbb{X}u||_n)\gamma'\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n.$$

which yields

$$||\mathbb{X}u||_n \leq (8+2\sqrt{2}+2\gamma')\sigma\sqrt{\frac{\log(1/\delta_0)}{n}} + ||\mathbb{X}u||_n\gamma'\sqrt{\frac{\log(1/\delta_0)}{n}},$$

We have chosen $\exp(-n/4\gamma'^2) \leq \delta_0$, which implies that

$$||\mathbb{X}u||_n \leq (16+4\sqrt{2}+4\gamma')\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}. \tag{43}$$

We can deduce from (42) that

$$|u|_* \leq (16 + 4\sqrt{2} + 4\gamma')\sigma\gamma' \frac{\log(1/\delta_0)}{n}, \tag{44}$$

and combining the second part of (42) with Lemma 7.5, we get

$$|u|_2 \gamma' \sqrt{\frac{s}{n} \log\left(\frac{p}{s}\right)} \leq (4 + \sqrt{2}) \sqrt{\frac{\log(1/\delta_0)}{n}} ||\mathbb{X}u||_n \leq (4 + \sqrt{2})(16 + 4\sqrt{2} + 4\gamma')\sigma \frac{\log(1/\delta_0)}{n}.$$

Finally, we get that

$$|u|_2 \leq \frac{(4 + \sqrt{2})(16 + 4\sqrt{2} + 4\gamma')}{\gamma'} \sigma \sqrt{\frac{\log^2(1/\delta_0)}{sn \log(p/s)}}. \tag{45}$$

□

Acknowledgement. This work is supported by the Labex Ecodec under the grant ANR-11-LABEX-0047 from the French Agence Nationale de la Recherche. The author thanks Professor Alexandre Tsybakov for helpful comments and discussions.

References

[1] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *ArXiv preprint, arXiv:1605.08651v3*, 2017.

[2] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Towards the study of least squares estimators with convex penalty. *Séminaires et Congrès, accepted, to appear*, 2017.

[3] P. C. Bellec and A. B. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In V. Panov, editor, *Modern Problems of Stochastic Analysis and Statistics, Selected Contributions In Honor of Valentin Konakov*. Springer, 2017.

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

[5] A. Belloni, V. Chernozhukov, L. Wang, et al. Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics*, 42(2):757–788, 2014.

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

[7] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. Slope - adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9(3):1103, 2015.

[8] C. Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

[9] G. Lecué and S. Mendelson. Regularization and the small-ball method i: sparse recovery. *Annals of Statistics, to appear*, 2017.

[10] A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

[11] B. Stucky and S. van de Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18:1–29, 2017.

[12] W. Su and E. Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. *Annals of Statistics*, 44(3):1038–1068, 2016.

[13] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, pages 1–20, 2012.

[14] X. Zeng and M. A. T. Figueiredo. The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms. *ArXiv preprint, arXiv:1409.4271*, 2014.