The Annals of Statistics
2018, Vol. 46, No. 6B, 3603–3642
https://doi.org/10.1214/17-AOS1670

SLOPE MEETS LASSO: IMPROVED ORACLE BOUNDS AND OPTIMALITY[1]

BY PIERRE C. BELLEC[*,†,§], GUILLAUME LECUÉ[*,†,‡] AND ALEXANDRE B. TSYBAKOV[*,†]

ENSAE[*], CREST (UMR CNRS 9194)[†], CNRS[‡] and Rutgers University[§]

We show that two polynomial time methods, a Lasso estimator with adaptively chosen tuning parameter and a Slope estimator, adaptively achieve the minimax prediction and ℓ_2 estimation rate $(s/n)\log(p/s)$ in high-dimensional linear regression on the class of s-sparse vectors in $\mathbb{R}^p$. This is done under the Restricted Eigenvalue (RE) condition for the Lasso and under a slightly more constraining assumption on the design for the Slope. The main results have the form of sharp oracle inequalities accounting for the model misspecification error. The minimax optimal bounds are also obtained for the ℓ_q estimation errors with $1 \le q \le 2$ when the model is well specified. The results are nonasymptotic, and hold both in probability and in expectation. The assumptions that we impose on the design are satisfied with high probability for a large class of random matrices with independent and possibly anisotropically distributed rows. We give a comparative analysis of conditions, under which oracle bounds for the Lasso and Slope estimators can be obtained. In particular, we show that several known conditions, such as the RE condition and the sparse eigenvalue condition are equivalent if the ℓ_2-norms of regressors are uniformly bounded.

1. Introduction. One of the important issues in high-dimensional statistics is to construct methods that are both computable in polynomial time, and have optimal statistical performance in the sense that they attain the optimal convergence rates on suitable classes of underlying objects (vectors, matrices) such as, for example, the classes of s-sparse vectors. It has been recently shown that, in some testing problems, this task cannot be achieved, and there is a gap between the optimal rates in a minimax sense and the best rate achievable by polynomial time algorithms [3]. However, the question about the existence of such a gap remains open for the most famous problem, namely, that of estimation and prediction in

Received May 2016; revised May 2017.

[1]Supported by GENES and by the French National Research Agency (ANR) under the Grants IPANEMA (ANR-13-BSH1-0004-02) and Labex Ecodec (ANR-11-LABEX-0047). It was also supported by the "Chaire Economie et Gestion des Nouvelles Données," under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

MSC2010 subject classifications. Primary 60K35, 62G08; secondary 62C20, 62G05, 62G20.

Key words and phrases. Sparse linear regression, minimax rates, high-dimensional statistics, Slope, Lasso.

high-dimensional linear regression on the classes of s-sparse parameters in $\mathbb{R}^p$. The known polynomial time methods such as the Lasso, the Dantzig selector and several other were shown to attain the prediction or ℓ_2-estimation rate $(s/n)\log(p)$ [4, 8] while the minimax rate for the problem is $(s/n)\log(p/s)$ (cf. [1, 7, 19, 23, 24, 27, 29] and Section 7 below). The recent papers [16, 25] inspire hope that computationally feasible methods can achieve the minimax rate $(s/n)\log(p/s)$. Specifically, [25] shows that for a particular random design (i.i.d. standard normal regressors) the rate $(s/n)\log(p/s)$ is asymptotically achieved by a Slope estimator, which is computable in polynomial time. An extension of [25] to sub-Gaussian designs is given in [16] that provides a nonasymptotic bound with the same rate. However, akin to [25], a key assumption in [16] is that the design is isotropic, so that its covariance matrix is proportional to the identity matrix.

The Slope estimator suggested in [5] is defined as a solution of the convex minimization problem given in (2.4) below. This estimator requires p tuning parameters $\lambda_1, \dots, \lambda_p$ not all equal to 0 and such that $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. These $\lambda_1, \dots, \lambda_p$ are the weights of the sorted ℓ_1 norm; cf. (2.2) below.

In this paper, we show that under a Restricted Eigenvalue (RE)-type condition on the design, the Slope estimator with suitably chosen tuning parameters achieves the optimal rate $(s/n)\log(p/s)$ for both the prediction and the ℓ_2 estimation risks, and both in probability and in expectation. The recommended tuning parameters are given in (2.5) below. Furthermore, we show that a large class of random design matrices with independent and possibly anisotropically distributed rows satisfies this RE-type condition with high probability. In other words, our conditions on the design for the Slope estimator are very close to those usually assumed for the Lasso estimator while the rate improves from $(s/n)\log(p)$ (previously known for the Lasso) to the optimal rate $(s/n)\log(p/s)$. Next, with the same method of proof, we show that the Lasso estimator also achieves this improved (and optimal) rate when the sparsity s is known. If s is unknown, we propose to replace s by an estimator $\hat{s}$ such that the bound $\hat{s} \leq s$ holds with high probability. We show that the suggested $\hat{s}$ is such that the Lasso estimator with tuning parameter of order $\sqrt{\log(p/\hat{s})/n}$ achieves the optimal rate $(s/n)\log(p/s)$.

The main results are obtained in the form of sharp oracle inequalities accounting for the model misspecification error. The minimax optimal bounds are also established for the ℓ_q-estimation errors with $1 \leq q \leq 2$ when the model is well specified. All our results are nonasymptotic.

As a by-product, we cover some other related issues of independent interest:

- We give a comparative analysis of conditions, under which oracle bounds for the Lasso and Slope estimators can be obtained showing, in particular, that several known conditions are equivalent.
- Due to the new techniques, we obtain bounds in probability with fast rate $(s/n)\log(p/s)$ at any level of confidence while using the same tuning parameter. As opposed to the previous work on the Lasso, the level of confidence is not

linked to the tuning parameter of the method. As a corollary, this implies rate optimal bounds on any moments of the estimation and prediction errors.

2. Statement of the problem and organization of the paper. Assume that we observe the vector

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\xi},$$

where $\mathbf{f} \in \mathbb{R}^n$ is an unknown deterministic mean and $\boldsymbol{\xi} \in \mathbb{R}^n$ is a noise vector. Let $\sigma > 0$. Everywhere except for Section 9 we assume that $\boldsymbol{\xi}$ is normal $\mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$, where $I_{n\times n}$ denotes the $n \times n$ identity matrix.

For all $\boldsymbol{u} = (u_1, \dots, u_n) \in \mathbb{R}^n$, define the empirical norm of $\boldsymbol{u}$ by

$$\|\boldsymbol{u}\|_n^2 = \frac{1}{n}\sum_{i=1}^n u_i^2.$$

Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ be a given matrix that we will call the design matrix. If $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y})$ is an estimator valued in $\mathbb{R}^p$, the value $\mathbb{X}\hat{\boldsymbol{\beta}}$ is used as a prediction for $\mathbf{f}$. The prediction error of an estimator $\hat{\boldsymbol{\beta}}$ is given by $\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2$. If the model is well specified, that is, $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$, then $\hat{\boldsymbol{\beta}}$ is used as an estimator of $\boldsymbol{\beta}^*$. The estimation error of $\hat{\boldsymbol{\beta}}$ is given by $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_q^q$ for some $q \in [1, 2]$, where $|\cdot|_q$ denotes the ℓ_q-norm in $\mathbb{R}^p$.

Two estimators will be studied in this paper: the Lasso estimator and the Slope estimator. The Lasso estimator $\hat{\boldsymbol{\beta}}$ is a solution of the minimization problem

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\big(\|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2 + 2\lambda|\boldsymbol{\beta}|_1\big), \tag{2.1}$$

where $\lambda > 0$ is a tuning parameter. Section 4 studies the prediction and estimation performance of the Lasso estimator with tuning parameter of order $\sigma\sqrt{\log(p/s)/n}$, where $s \in \{1, \dots, p\}$ is a sparsity parameter which is supposed to be known. In Section 5, we propose an adaptive choice of this parameter. Section 5 defines an estimator $\hat{s}$ valued in $\{1, \dots, p\}$ and studies the performance of the Lasso estimator with a data-driven tuning parameter of order $\sigma\sqrt{\log(p/\hat{s})/n}$.

Section 6 studies the Slope estimator [5], which is defined as follows. Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$ be a vector of tuning parameters not all equal to 0 such that $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$. For any $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$, let $(\beta_1^\sharp, \dots, \beta_p^\sharp)$ be a nonincreasing rearrangement of $|\beta_1|, \dots, |\beta_p|$. Set

$$|\boldsymbol{\beta}|_* = \sum_{j=1}^p \lambda_j \beta_j^\sharp, \qquad \boldsymbol{\beta} \in \mathbb{R}^p, \tag{2.2}$$

which defines a norm on $\mathbb{R}^p$; cf. [5], Proposition 1.2. Equivalently, we can write

$$|\boldsymbol{\beta}|_* = \max_\phi \sum_{j=1}^p \lambda_j |\beta_{\phi(j)}|, \tag{2.3}$$

where the maximum is taken over all permutations $\phi = (\phi(1), \dots, \phi(p))$ of $\{1, \dots, p\}$. The Slope estimator $\hat{\boldsymbol{\beta}}$ is defined as a solution of the minimization problem:

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \big(\|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2 + 2|\boldsymbol{\beta}|_* \big). \tag{2.4}$$

Section 6 establishes oracle inequalities and estimation error bounds for the Slope estimator with tuning parameters

$$\lambda_j = A\sigma \sqrt{\frac{\log(2p/j)}{n}}, \qquad j = 1, \dots, p, \tag{2.5}$$

for any constant $A > 4 + \sqrt{2}$.

Section 7 gives nonasymptotic minimax lower bounds showing that the upper bounds of Sections 4–6 cannot be improved. In Section 8, we provide a comparison of the conditions on the design matrix $\mathbb{X}$, under which the results are obtained. In particular, we prove that the oracle inequalities for the Slope estimator in Section 6 hold for design matrices with independent and possibly anisotropically distributed sub-Gaussian rows. Section 9 explains that, up to changes in numerical constants, all results of the paper remain valid if the components of the noise vector $\boldsymbol{\xi}$ are independent sub-Gaussian random variables. The proofs are given in the Appendix.

Notation and preliminaries. We will assume that the diagonal elements of the Gram matrix $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ are at most 1, that is, $\max_{j=1,\dots,p} \|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$ where $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_p)$ is the canonical basis in $\mathbb{R}^p$. Let $\boldsymbol{g} = (g_1, \dots, g_p)$ be the random vector with components

$$g_j = \frac{1}{\sqrt{n}} \boldsymbol{x}_j^T \boldsymbol{\xi} \qquad \text{where } \boldsymbol{x}_j = \mathbb{X}\boldsymbol{e}_j,\ j = 1, \dots, p. \tag{2.6}$$

If $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ it follows from the inequality $\|\boldsymbol{x}_j\|_n \le 1$ that the random variables g_j are zero mean Gaussian with variance at most σ^2. We denote by $\boldsymbol{g}^\sharp = (g_1^\sharp, \dots, g_p^\sharp)$ a nonincreasing rearrangement of $(|g_1|, \dots, |g_p|)$. We also use the notation

$$|\boldsymbol{\beta}|_0 = \sum_{j=1}^p I(\beta_j \neq 0), \qquad |\boldsymbol{\beta}|_\infty = \max_{j=1,\dots,p} |\beta_j| \quad \text{and} \quad |\boldsymbol{\beta}|_q = \left(\sum_{j=1}^p |\beta_j|^q \right)^{1/q}$$

for $0 < q < \infty$. Here, $I(\cdot)$ is the indicator function. For any set $J \subset \{1, \dots, p\}$, denote by J^c its complement, by $|J|$ its cardinality, and for any $\boldsymbol{u} = (u_1, \dots, u_p) \in \mathbb{R}^p$, let $\boldsymbol{u}_J \in \mathbb{R}^p$ be the vector such that its jth component is equal to u_j if $j \in J$ and equal to 0 otherwise. For two real numbers a, b, we will use the notation $a \vee b = \max(a, b)$.

We denote by $\operatorname{Med}[Z]$ a median of a real valued random variable Z, that is, any real number such that $\mathbb{P}(Z \ge \operatorname{Med}[Z]) \ge 1/2$ and $\mathbb{P}(Z \le \operatorname{Med}[Z]) \ge 1/2$.

The following bounds on the sum $\sum_{j=1}^{s}\log(2p/j)$ will be useful. From Stirling's formula, we easily deduce that $s\log(s/e)\le\log(s!)\le s\log(s)$, and thus

$$s\log(2p/s)\le\sum_{j=1}^{s}\log(2p/j)=s\log(2p)-\log(s!)\le s\log(2ep/s). \tag{2.7}$$

Finally, for a given $\delta_0\in(0,1)$ and for any $\boldsymbol{u}=(u_1,\dots,u_p)\in\mathbb{R}^p$ we set

$$\begin{aligned} H(\boldsymbol{u})&\triangleq(4+\sqrt{2})\sum_{j=1}^{p}u_j^{\sharp}\sigma\sqrt{\frac{\log(2p/j)}{n}},\\ G(\boldsymbol{u})&\triangleq(4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}\|\mathbb{X}\boldsymbol{u}\|_n, \end{aligned} \tag{2.8}$$

where $(u_1^{\sharp},\dots,u_p^{\sharp})$ is a nonincreasing rearrangement of $(|u_1|,\dots,|u_p|)$.

3. The tuning parameter of the Lasso need not be tied to a confidence level. In this section, we denote by $\hat{\boldsymbol{\beta}}$ the Lasso estimator defined by (2.1) and provide an improved probability estimate for the performance of the Lasso estimator with a tuning parameter of order $\sigma\sqrt{2\log p}$. First, we state a version of the restricted eigenvalue condition that we will refer to in the sequel. Let $s\in\{1,\dots,p\}$, and let $c_0>0$ be a constant.

RE(s,c_0) CONDITION. *The design matrix* $\mathbb{X}$ *satisfies* $\|\mathbb{X}\boldsymbol{e}_j\|_n\le 1$ *for all* $j=1,\dots,p$, *and*

$$\kappa(s,c_0)\triangleq\inf_{\boldsymbol{\delta}\in\mathcal{C}_{\mathrm{RE}}(s,c_0):\boldsymbol{\delta}\neq\boldsymbol{0}}\frac{\|\mathbb{X}\boldsymbol{\delta}\|_n}{|\boldsymbol{\delta}|_2}>0, \tag{3.1}$$

where $\mathcal{C}_{\mathrm{RE}}(s,c_0)=\{\boldsymbol{\delta}\in\mathbb{R}^p:|\boldsymbol{\delta}|_1\le(1+c_0)\sum_{j=1}^{s}\delta_j^{\sharp}\}$ *and* $\delta_1^{\sharp}\ge\cdots\ge\delta_p^{\sharp}$ *denotes a nonincreasing rearrangement of* $|\delta_1|,\dots,|\delta_p|$.

Though stated in somewhat different form, inequality (3.1) is equivalent to the original RE condition of [4]. Indeed, let $\boldsymbol{\delta}\in\mathbb{R}^p$, and let $J_*=J_*(\boldsymbol{\delta})\subseteq\{1,\dots,p\}$ be the set of indices of the s largest in absolute value components of $\boldsymbol{\delta}$. Then $\sum_{j=1}^{s}\delta_j^{\sharp}=|\boldsymbol{\delta}_{J_*}|_1$. Therefore, the condition $|\boldsymbol{\delta}|_1\le(1+c_0)\sum_{j=1}^{s}\delta_j^{\sharp}$ can be written as $|\boldsymbol{\delta}_{J_*^c}|_1\le c_0|\boldsymbol{\delta}_{J_*}|_1$. Thus, an equivalent form of (3.1) is obtained by replacing the cone $\mathcal{C}_{\mathrm{RE}}(s,c_0)$ with

$$\mathcal{C}'_{\mathrm{RE}}(s,c_0)=\bigcup_{J\subseteq\{1,\dots,p\}:|J|\le s}\{\boldsymbol{\delta}\in\mathbb{R}^p:|\boldsymbol{\delta}_{J^c}|_1\le c_0|\boldsymbol{\delta}_J|_1\}, \tag{3.2}$$

which is the standard cone of the RE condition as introduced in [4]. One minor difference from [4] is that in (3.1) we have $|\boldsymbol{\delta}|_2$ rather than $|\boldsymbol{\delta}_{J_*}|_2$ in the denominator. This only modifies the constant $\kappa(s,c_0)$ by factor $\sqrt{1+c_0}$. Indeed, for any

$\boldsymbol{\delta} \in \mathcal{C}'_{\mathrm{RE}}(s, c_0)$,

$$|\boldsymbol{\delta}_{J_*^c}|_2^2 = \sum_{j \in J_*^c} \delta_j^2 \le \sum_{j \in J_*^c} |\delta_j| \Big(\frac{1}{|J_*|} \sum_{k \in J_*} |\delta_k| \Big) \le \frac{|\boldsymbol{\delta}_{J_*}|_1 |\boldsymbol{\delta}_{J_*^c}|_1}{|J_*|} \le \frac{c_0 |\boldsymbol{\delta}_{J_*}|_1^2}{|J_*|} \le c_0 |\boldsymbol{\delta}_{J_*}|_2^2,$$

and thus $|\boldsymbol{\delta}|_2 = (|\boldsymbol{\delta}_{J_*}|_2^2 + |\boldsymbol{\delta}_{J_*^c}|_2^2)^{1/2} \le \sqrt{1 + c_0} |\boldsymbol{\delta}_{J_*}|_2$. Another difference from [4] is that we include the assumption $\|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$ in the statement of the RE condition. This is a mild assumption that is omnipresent in the literature on the Lasso. Often it is stated with equality and is interpreted as a normalization.

Consider the following result based on [4, 10, 14], which is representative of the nonasymptotic bounds obtained so far for the prediction performance of the Lasso.

PROPOSITION 3.1 ([4, 10, 14]). *Let $p \ge 4$, $s \in \{1, \dots, p\}$, $\varepsilon > 0$ and set $c_0 = 1 + 1/\varepsilon$. For any $\delta \in (0, 1/2)$ and $\varepsilon \in (0, 1)$, the Lasso estimator (2.1) with tuning parameter*

$$\lambda = (1 + \varepsilon) \sigma \sqrt{2 \log(p/\delta)/n} \tag{3.3}$$

satisfies with probability $1 - 2\delta$ the oracle inequality

$$\begin{aligned} \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le{}& \inf_{\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\beta}|_0 \le s} \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \\ &+ \frac{\sigma^2}{n} \Big(\frac{(1 + \varepsilon)\sqrt{2s \log(p/\delta)}}{\kappa(s, c_0)} + \sqrt{s} + \sqrt{2 \log(1/\delta)} \Big)^2. \end{aligned} \tag{3.4}$$

An oracle inequality of the same kind as (3.4) was first obtained in [14], Theorem 6.1, and in a slightly less general form, with some factor $C > 1$ in front of $\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2$ in [4]. The numerical constants in Proposition 3.1 are taken from the proof of Theorem 3 in [10] (cf. the inequality preceding (25) in [10]).

A notable feature of Proposition 3.1 and of other nonasymptotic bounds for the Lasso available in the literature is that the confidence level $1 - 2\delta$ is tied to the tuning parameter λ; cf. (3.3). If a confidence level closer to one is desired (i.e., smaller δ), the previous results suggest that the tuning parameter should be increased according to the relationship (3.3) between λ and δ. We claim that it is not needed. Indeed, the following proposition holds.

PROPOSITION 3.2. *Let $p \ge 2$, $s \in \{1, \dots, p\}$, $\varepsilon > 0$ and set $c_0 = 1 + 1/\varepsilon$. Let $\hat{\boldsymbol{\beta}}$ be the Lasso estimator (2.1) with tuning parameter*

$$\lambda = (1 + \varepsilon) \sigma \sqrt{2 \log(p)/n}. \tag{3.5}$$

Then for any $\delta \in (0, 1)$,

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le \inf_{\boldsymbol{\beta}\in\mathbb{R}^p : |\boldsymbol{\beta}|_0 \le s} \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\left(\frac{(1+\varepsilon)\sqrt{2s\log p}}{\kappa(s, c_0)} + \sqrt{s} + \sqrt{2\log(1/\delta)} + 2.8\right)^2 \tag{3.6}$$

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

The proof of Proposition 3.2 is given in Appendix I. Let us highlight some features of Proposition 3.2 that are new.

First, the tuning parameter (3.5) does not depend on the confidence constant δ. The Lasso estimator with tuning parameter (3.5) enjoys the oracle inequality (3.6) for any $\delta \in (0, 1)$. This contrasts with Proposition 3.1 where the oracle inequality is for a single δ tied to the tuning parameter (3.3). As a consequence, (3.6) immediately implies bounds for all moments of $\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2$, which cannot be obtained from Proposition 3.1.

Second, the probability that an oracle inequality with error term of order $s\log(p)/n$ holds is substantially closer to 1 than it was commonly understood before. For example, take $\delta = p^{-s}$, which balances the remainder term in (3.6). Then Proposition 3.2 yields that the Lasso estimator with tuning parameter (3.5) converges with the rate smaller than $s\log(p)/n$ up to a multiplicative constant. On the other hand, the choice $\delta = p^{-s}$ in Proposition 3.1 yields only a suboptimal rate of order $s^2\log(p)/n$.

The constant term 2.8 in (3.6) is negligible. Thus, in asymptotic regimes where $p \to \infty$ or $\delta \to 0$, the constant term 2.8 is dominated by $\sqrt{\log p}$ or $\sqrt{\log(1/\delta)}$. If we ignore this constant term, the bound (3.6) strictly improves upon (3.4).

In spite of these improvements, the result of Proposition 3.2 is not completely satisfying since only the rate $s\log(p)/n$ and not the optimal rate $(s/n)\log(p/s)$ is proved. In the next section, we show that the optimal rate can be achieved by the Lasso estimator with tuning parameter of the order $\sigma\sqrt{\log(p/s)/n}$.

4. Optimal rates for the Lasso estimator. In this section, we denote by $\hat{\boldsymbol{\beta}}$ the Lasso estimator defined by (2.1), and we derive upper bounds for its prediction and estimation errors. As usual in the Lasso context, the argument contains two main ingredients. First, all randomness is removed from the problem by reducing the consideration to a suitably chosen random event of high probability. Second, the error bounds are derived on this event by a purely deterministic argument. In our case, such a deterministic argument is given in Theorem 4.2 below, while the "randomness removing tool" is provided by the next theorem. As we will see in Section 6, this theorem is common to the study of both the Lasso and the Slope estimators.

THEOREM 4.1. *Let $\delta_0 \in (0,1)$ and let $\mathbb{X} \in \mathbb{R}^{n\times p}$ be a matrix such that $\max_{j=1,\dots,p} \|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$. Let $H(\cdot)$ and $G(\cdot)$ be defined in (2.8). If $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$, then the random event*

$$\left\{\frac{1}{n}\boldsymbol{\xi}^T\mathbb{X}\boldsymbol{u} \le \max\big(H(\boldsymbol{u}), G(\boldsymbol{u})\big), \forall \boldsymbol{u} \in \mathbb{R}^p\right\} \tag{4.1}$$

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

The proof of Theorem 4.1 is given in Appendix E. We now discuss consequences of Theorem 4.1 leading to tighter bounds. We will need the following condition on the design matrix $\mathbb{X}$ that will be called the *Strong restricted Eigenvalue* condition or shortly the SRE condition. Let $c_0 > 0$ and $s \in \{1,\dots,p\}$ be fixed.

SRE(s,c_0) CONDITION. *The design matrix $\mathbb{X}$ satisfies $\|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$ for all $j = 1,\dots,p$, and*

$$\theta(s,c_0) \triangleq \min_{\boldsymbol{\delta}\in\mathcal{C}_{\mathrm{SRE}}(s,c_0):\boldsymbol{\delta}\neq\mathbf{0}} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_n}{|\boldsymbol{\delta}|_2} > 0, \tag{4.2}$$

where $\mathcal{C}_{\mathrm{SRE}}(s,c_0) \triangleq \{\boldsymbol{\delta}\in\mathbb{R}^p : |\boldsymbol{\delta}|_1 \le (1+c_0)\sqrt{s}|\boldsymbol{\delta}|_2\}$ is a cone in $\mathbb{R}^p$.

Inequality (4.2) differs from its analog (3.1) in the RE(s,c_0) condition only in the definition of the cone $\mathcal{C}_{\mathrm{SRE}}(s,c_0)$, and in general (4.2) is more constraining. Indeed, the cone $\mathcal{C}_{\mathrm{RE}}(s,c_0)$ of the RE(s,c_0) condition is the set of all $\boldsymbol{\delta}\in\mathbb{R}^p$ such that $|\boldsymbol{\delta}|_1 \le (1+c_0)\sum_{j=1}^s \delta_j^\sharp$. By the Cauchy–Schwarz inequality, $\sum_{j=1}^s \delta_j^\sharp \le \sqrt{s}(\sum_{j=1}^s (\delta_j^\sharp)^2)^{1/2} \le \sqrt{s}|\boldsymbol{\delta}|_2$ so that

$$\mathcal{C}_{\mathrm{RE}}(s,c_0) \subseteq \mathcal{C}_{\mathrm{SRE}}(s,c_0). \tag{4.3}$$

Note that we have included the requirement that $\|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$ for all j in the RE and the SRE conditions. It can be replaced by $\|\mathbb{X}\boldsymbol{e}_j\|_n \le \theta_0$ for some $\theta_0 > 0$ but for brevity and w.l.o.g. we take here $\theta_0 = 1$. Interestingly, due to the inclusion of the assumption $\|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$, the RE condition becomes equivalent to the SRE condition up to absolute constants. Moreover, the equivalence further extends to the s-sparse eigenvalue condition. This is detailed in Section 8 below.

Under the SRE condition, we now establish a deterministic result, which is central in our argument. We first introduce some notation. Let $\gamma \in (0,1)$ be a constant. For any tuning parameter $\lambda > 0$, set

$$\begin{aligned} \delta(\lambda) &\triangleq \exp\left(-\left(\frac{\gamma\lambda\sqrt{n}}{(4+\sqrt{2})\sigma}\right)^2\right) \\ &\text{so that } \lambda = \frac{(4+\sqrt{2})\sigma}{\gamma}\sqrt{\frac{\log(1/\delta(\lambda))}{n}}. \end{aligned} \tag{4.4}$$

For given $s \in \{1, \dots, p\}$, the following theorem holds under the condition:

$$\lambda \geq \frac{(4+\sqrt{2})\sigma}{\gamma}\sqrt{\frac{\log(2ep/s)}{n}} \quad \text{or equivalently} \quad \delta(\lambda) \leq s/(2ep). \tag{4.5}$$

THEOREM 4.2. *Let $s \in \{1, \dots, p\}$, $\gamma \in (0,1)$ and $\tau \in [0, 1-\gamma)$. Assume that the* SRE(s, c_0) *condition holds with $c_0 = c_0(\gamma, \tau) = \frac{1+\gamma+\tau}{1-\gamma-\tau}$. Let λ be a tuning parameter such that (4.5) holds. Let $\delta_0 \in (0,1)$. Then, on the event (4.1), the Lasso estimator $\hat{\boldsymbol{\beta}}$ with tuning parameter λ satisfies*

$$2\tau\lambda|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \leq \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + C_{\gamma,\tau}(s, \lambda, \delta_0)\lambda^2 s, \tag{4.6}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\boldsymbol{\beta}|_0 \leq s$, and all $\mathbf{f} \in \mathbb{R}^n$, where

$$C_{\gamma,\tau}(s, \lambda, \delta_0) \triangleq (1+\gamma+\tau)^2\Big(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \vee \frac{1}{\theta^2(s, c_0(\gamma,\tau))}\Big).$$

Furthermore, if $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \leq s$ then on the event (4.1), we have for any $1 \leq q \leq 2$,

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_q \leq \Big(\frac{C_{\gamma,\tau}}{2\tau}\Big)^{2/q-1}\Big(\frac{C_{\gamma,0}}{1+\gamma}\Big)^{2-2/q}\lambda s^{1/q}. \tag{4.7}$$

Theorem 4.2 is proved in Appendix B. Before the statement of its corollaries, a few comments are in order.

The conclusions of Theorem 4.2 hold on the event (4.1), which is independent of γ and τ. Thus, on the event (4.1), for all choices of τ, γ and λ such that (4.5) holds, the oracle inequality (4.6) and the estimation bound (4.7) are satisfied.

The constants γ and τ are such that $\gamma + \tau < 1$. For the ease of presentation, the particular choice $\gamma = 1/2$ and $\tau = 1/4$ will be used below to derive two corollaries of Theorem 4.2. If $\gamma = 1/2$ and $\tau = 1/4$, then the constants in Theorem 4.2 have the form

$$c_0 = 7, \qquad (1+\gamma+\tau)^2 = \frac{49}{16}, \qquad \frac{1+\gamma}{1-\gamma} = 3, \qquad 1+\gamma = 3/2, \tag{4.8}$$

while inequality (4.7) can be transformed into

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_q \leq \frac{49}{8}\Big(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \vee \frac{1}{\theta^2(s,7)}\Big)\lambda s^{1/q}, \tag{4.9}$$

where we have used the inequalities $\theta^2(s, c_0) \leq \theta^2(s, \frac{1+\gamma}{1-\gamma})$ and $(2\tau)^{1-2/q} \leq 2$.

We now take a closer look at the constant $C_{\gamma,\tau}(s, \lambda, \delta_0)$. This constant is always greater than or equal to $(1+\gamma+\tau)^2/\theta^2(s, c_0)$. Furthermore, the value

$$\delta_0^* = \big(\delta(\lambda)\big)^{\frac{s}{\theta^2(s,c_0)}} \tag{4.10}$$

is the smallest $\delta_0 \in (0,1)$ such that $C_{\gamma,\tau}(s,\lambda,\delta_0) = (1+\gamma+\tau)^2/\theta^2(s,c_0)$. If λ satisfies (4.5), then

$$\delta_0^* \le \left(\frac{s}{2ep}\right)^{\frac{s}{\theta^2(s,c_0)}}.$$

Using these remarks, we obtain the following corollary of Theorems 4.1 and 4.2 with the choice $\gamma = 1/2$, $\tau = 1/4$, and $\delta_0 = \delta_0^*$.

COROLLARY 4.3. *Let $s \in \{1,\dots,p\}$. Assume that the* SRE$(s,7)$ *condition holds. Let $\hat{\boldsymbol{\beta}}$ be the Lasso estimator with tuning parameter λ satisfying* (4.5) *for $\gamma = 1/2$. Then, with probability at least $1 - \frac{1}{2}(\frac{s}{2ep})^{\frac{s}{\theta^2(s,7)}}$, we have*

$$\frac{\lambda}{2}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{49\lambda^2 s}{16\theta^2(s,7)} \tag{4.11}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\boldsymbol{\beta}|_0 \le s$, and all $\mathbf{f} \in \mathbb{R}^n$. Furthermore, if $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \le s$ then, for any $1 \le q \le 2$,

$$\mathbb{P}\left(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_q \le \frac{49\lambda s^{1/q}}{8\theta^2(s,7)}\right) \ge 1 - \frac{1}{2}\left(\frac{s}{2ep}\right)^{\frac{s}{\theta^2(s,7)}}. \tag{4.12}$$

Since $\theta^2(s,7) \le 1$, the probability in Corollary 4.3 is greater than $1 - \frac{1}{2}(\frac{s}{2ep})^s$. If the tuning parameter is chosen such that (4.5) holds with equality, then $\lambda^2 s$ is equal to

$$\frac{\sigma^2 s \log(2ep/s)}{n}$$

up to a multiplicative constant. This is the minimax rate with respect to the prediction error over the class of all s-sparse vectors $B_0(s) = \{\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\beta}|_0 \le s\}$. The rate $\lambda s^{1/q}$ in (4.12) is minimax optimal for the ℓ_q estimation problem. A more detailed discussion of the minimax rates is given in Section 7.

Finally, the conclusions of Theorem 4.2 hold for all $\delta_0 \le \delta_0^*$. This allows us to integrate the oracle inequality (4.6) and the estimation bound (4.7) to obtain the following results in expectation.

COROLLARY 4.4. *Let $s \in \{1,\dots,p\}$. Assume that the* SRE$(s,7)$ *condition holds. Let $\hat{\boldsymbol{\beta}}$ be the Lasso estimator with tuning parameter λ satisfying* (4.5) *for $\gamma = 1/2$. Then*

$$\begin{aligned}\mathbb{E}\left[\frac{\lambda}{2}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2\right] \\ \le \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{49\lambda^2 s}{16}\left(\frac{1}{\theta^2(s,7)} + \frac{1}{2\log(2ep)}\right)\end{aligned} \tag{4.13}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$ *such that* $|\boldsymbol{\beta}|_0 \le s$, *and all* $\mathbf{f} \in \mathbb{R}^n$. *Furthermore, if* $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ *for some* $\boldsymbol{\beta}^* \in \mathbb{R}^p$ *with* $|\boldsymbol{\beta}^*|_0 \le s$ *then, for any* $1 \le q \le 2$,

$$\mathbb{E}\big[|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}|_q^q\big] \le \frac{(49)^q \lambda^q s}{8^q}\bigg(\frac{1}{\theta^{2q}(s,7)} + \frac{1}{(\log(2ep))^q}\bigg). \tag{4.14}$$

Corollary 4.4 is proved in Section B.

REMARK 1. The smallest values of λ, for which Theorem 4.2 and Corollaries 4.3 and 4.4 hold is given in (4.5); in particular, it depends on s. For this choice of λ, the prediction risk and the ℓ_2 risk of the Lasso estimator attain the nonasymptotic minimax optimal rate $(s/n)\log(p/s)$. However, the knowledge of the sparsity index s is needed to achieve this, which raises a problem of adaptation to sparsity s. In Section 5, we propose a data-driven Lasso estimator, independent of s, solving this adaptation problem, for which we prove essentially the same results as above. The argument there uses Corollary 4.3 as a building block.

REMARK 2. Since the assumptions on tuning parameter λ in Theorem 4.2 and Corollaries 4.3 and 4.4 are given by inequalities, the case of λ defined with $\log(2ep)$ instead of $\log(2ep/s)$ is also covered. With such a choice of λ, the estimators do not depend on s and the results take the same form as in the standard Lasso framework (cf. [4]), in which the prediction risk and the ℓ_2 risk achieve the suboptimal rate $(s\log p)/n$. However, even in this case, Theorem 4.2 brings in some novelty. Indeed, to our knowledge, bounds in expectation (cf. Corollary 4.4), or in probability with arbitrary $\delta_0 \in (0,1)$ (cf. Theorem 4.2), were not available. The previous work provided only bounds in probability for fixed δ_0 proportional to $1/p^c$ for an absolute constant $c > 0$, in the spirit of (3.4). Such bounds do not allow for control of the moments of the estimation and prediction errors without imposing extra assumptions. To our understanding, there was no way to fix this problem within the old proof techniques. On the contrary, bounds for the moments of any order can be readily derived from Theorem 4.2.

REMARK 3. In this section, the variance σ was supposed to be known. The case of unknown σ can be treated in a standard way as described, for example, in [12]. Namely, we replace σ in (4.5) by a suitable statistic $\hat{\sigma}$. For example, it can be shown that under the RE condition, the scaled Lasso estimator $\hat{\sigma}^S$ is such that $\sigma/2 \le \hat{\sigma}^S \le 2\sigma$ with high probability provided that $s \le cn$ for some constant $c > 0$, cf. [12], Sections 5.4 and 5.6.2. Then, replacing σ by $\hat{\sigma} \triangleq 2\hat{\sigma}^S$ in the expression for λ [cf. (4.5)], we obtain that under the same mild conditions, Corollary 4.3 remains valid with this choice of λ independent of σ, up to a change in numerical constants. This remark also applies to upper bounds in probability obtained in the next sections.

5. Aggregated Lasso estimator and adaptation to sparsity. In this section, we assume that $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$, so that we have a linear regression model:

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}. \tag{5.1}$$

We also assume that $\boldsymbol{\beta}^* \in B_0(s) = \{\boldsymbol{\beta}^* \in \mathbb{R}^p : |\boldsymbol{\beta}^*|_0 \le s\}$, where $s \le s_*$. Here, $s_* \in [1, p/2]$ is a given integer. Our aim is to construct an adaptive to s estimator $\tilde{\boldsymbol{\beta}}$ whose prediction risk attains the optimal rate $(s/n)\log(p/s)$ simultaneously on the classes $B_0(s)$, $1 \le s \le s_*$. This will be done by aggregating a small number of Lasso estimators using a Lepski-type procedure. Given a lower bound for a restricted eigenvalue at $s = s_*$, the resulting adaptive estimator $\tilde{\boldsymbol{\beta}}$ is computed in polynomial time and its computational complexity exceeds that of the Lasso only by a $\log_2 p$ factor. Furthermore, we propose an estimator $\hat{s}$ such that the bound $\hat{s} \le s$ holds with high probability without the beta-min condition and without any strong assumptions on the matrix $\mathbb{X}$ such as the irrepresentability condition.

We denote by $\hat{\boldsymbol{\beta}}_s$ the Lasso estimator with tuning parameter

$$\lambda(s) = 2(4+\sqrt{2})\sigma\sqrt{\frac{\log(2ep/s)}{n}},$$

and we set for brevity $\theta_* = \theta(2s_*, 7)$. We will assume that $\theta_* > 0$. Then, $\theta^2(s,7) \ge \theta_* > 0$, $s = 1, \dots, 2s_*$. It follows from Corollary 4.3 that for any $s = 1, \dots, 2s_*$

$$\begin{aligned}\sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{P}_{\boldsymbol{\beta}^*}\left(\|\mathbb{X}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*)\|_n \ge C_0\sigma\sqrt{\frac{s\log(2ep/s)}{n}}\right) \\ \le \frac{1}{2}\left(\frac{s}{2ep}\right)^{\frac{s}{\theta^2(s,7)}} \le \left(\frac{s}{p}\right)^s,\end{aligned} \tag{5.2}$$

where

$$C_0 = 7(4+\sqrt{2})/(2\theta_*), \tag{5.3}$$

and $\mathbb{P}_{\boldsymbol{\beta}^*}$ is the probability measure associated to the model (5.1). Furthermore, since $\lambda(s) \ge \lambda(2s)$ we also have for any $s = 1, \dots, s_*$

$$\sup_{\boldsymbol{\beta}^* \in B_0(2s)} \mathbb{P}_{\boldsymbol{\beta}^*}\left(\|\mathbb{X}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*)\|_n \ge \sqrt{2}C_0\sigma\sqrt{\frac{s\log(2ep/s)}{n}}\right) \le \left(\frac{2s}{p}\right)^{2s}. \tag{5.4}$$

Set $b_j = 2^{j-1}$, $j \in \mathbb{N}$. In what follows, we assume w.l.o.g. that $s_* \ge 2$ since the problem of adaptation does not arise for $s_* = 1$. Then the integer $M \triangleq \max\{m \in \mathbb{N} : b_m \le s_*\}$ satisfies $M \ge 2$. Note also that $M \le \log_2(p)$. We now construct a data-driven selector from the set of estimators $\{\hat{\boldsymbol{\beta}}_{b_m}, m = 2, \dots, M\}$. We select the index m as follows:

$$\hat{m} \triangleq \max\{m \in \{2, \dots, M\} : d(\hat{\boldsymbol{\beta}}_{b_m}, \hat{\boldsymbol{\beta}}_{b_{m-1}}) > 2w(b_m)\} \tag{5.5}$$

with the convention that $\hat{m} = 2$ if the set in the above display is empty. Here, $d(\boldsymbol{\beta}, \boldsymbol{\beta}') = \|\mathbb{X}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_n$ for all $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p$, and $w(b) = C_0\sigma\sqrt{\frac{b\log(2ep/b)}{n}}$, $\forall b \in [1, p]$. Next, we define adaptive estimators of s and $\boldsymbol{\beta}^*$ as follows:

$$\hat{s} = b_{\hat{m}-1} = 2^{\hat{m}-1} \quad \text{and} \quad \tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{2\hat{s}}. \tag{5.6}$$

THEOREM 5.1. *Let $s_* \in \{2, \dots, p\}$ be such that $\theta_* > 0$ and $s_* \le p/(2e)$. Then there exists an absolute constant $C_1 > 0$ such that, for C_0 given in (5.3) and all $s = 1, \dots, s_*$,*

$$\begin{aligned}\sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{P}_{\boldsymbol{\beta}^*}\left(\|\mathbb{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_n \ge C_1C_0\sigma\sqrt{\frac{s\log(2ep/s)}{n}}\right) \\ \le (2\log_2(p) + 1)\left(\frac{2s}{p}\right)^{2s}\end{aligned} \tag{5.7}$$

and

$$\sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{P}_{\boldsymbol{\beta}^*}(\hat{s} \le s) \ge 1 - 2\log_2(p)\left(\frac{2s}{p}\right)^{2s}. \tag{5.8}$$

The proof of this theorem is given in Appendix C. It uses only the properties (5.2) and (5.4) of the family of estimators $\{\hat{\boldsymbol{\beta}}_s\}$. Theorem 5.1 shows that the prediction error of $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{2\hat{s}}$ achieves the optimal rate of order $(s/n)\log(p/s)$. Without (5.7), the fact that estimator $\hat{s}$ satisfies (5.8) is not interesting. Indeed, the dummy estimator $\tilde{s} = 0$ satisfies $\mathbb{P}(\tilde{s} \le s) = 1$. The estimator $\hat{s}$ is of interest because of the conjunction of (5.7) and (5.8); not only it satisfies $\hat{s} \le s$ with high probability [cf. (5.8)] but also the Lasso estimator with the tuning parameter $\lambda = 2(4 + \sqrt{2})\sigma\sqrt{\log(ep/\hat{s})/n}$ achieves the optimal rate [cf. (5.7)]. Theorem 5.1 shows that this choice of the tuning parameter improves upon the prediction bounds for the Lasso estimator [4] with the universal tuning parameter of order $\sigma\sqrt{(\log p)/n}$.

A procedure of the same type is adaptive to the sparsity when measuring the accuracy by the ℓ_q estimation error for $1 \le q \le 2$. In this case, the risk bound is proved in the same way as in Theorem 5.1 due to the following observations. First, it follows from Corollary 4.3 that for all $s = 1, \dots, 2s_*$, $1 \le q \le 2$,

$$\sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{P}_{\boldsymbol{\beta}^*}\left(|\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*|_q \ge C_0'\sigma s^{1/q}\sqrt{\frac{\log(2ep/s)}{n}}\right) \le \left(\frac{s}{p}\right)^s, \tag{5.9}$$

where $C_0' = 49(4 + \sqrt{2})/(4\theta_*)$. Second, analogously to (5.4), we have for all $s = 1, \dots, s_*$, $1 \le q \le 2$,

$$\sup_{\boldsymbol{\beta}^* \in B_0(2s)} \mathbb{P}_{\boldsymbol{\beta}^*}\left(|\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*|_q \ge 2C_0'\sigma s^{1/q}\sqrt{\frac{\log(2ep/s)}{n}}\right) \le \left(\frac{2s}{p}\right)^{2s}. \tag{5.10}$$

THEOREM 5.2. *Let $s_* \in \{2, \dots, p\}$ be such that $\theta_* > 0$ and $s_* \le p/(2e)$. Let $1 \le q \le 2$ and let $\hat{m}$ be defined by (5.5) with $d(\boldsymbol{\beta}, \boldsymbol{\beta}') = |\boldsymbol{\beta} - \boldsymbol{\beta}'|_q$ for all $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p$ and $w(b) = C_0 \sigma b^{1/q} \sqrt{\frac{\log(2ep/b)}{n}}$, $\forall b \in [1, p]$, where $C_0 = 49(4 + \sqrt{2})/(4\theta_*)$. Let $\tilde{\boldsymbol{\beta}}$ be defined as in (5.6). Then there exists an absolute constant $C_1 > 0$ such that, for all $s = 1, \dots, s_*$,*

$$\sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{P}_{\boldsymbol{\beta}^*}\left(|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_q \ge C_1 C_0 \sigma s^{1/q} \sqrt{\frac{\log(2ep/s)}{n}} \right) \le (2\log_2(p) + 1)\left(\frac{2s}{p}\right)^{2s}.$$

The proof of this theorem is given in Appendix C. Due to (5.9) and (5.10), it is quite analogous to the proof of Theorem 5.1.

6. Optimal rates for the Slope estimator. In this section, we study the Slope estimator with weights λ_j given in (2.5). We will use the following assumption on the design matrix $\mathbb{X}$ that we call the *Weighted Restricted Eigenvalue* condition, or shortly the WRE condition. Let $c_0 > 0$, $s \in \{1, \dots, p\}$ be constants, and let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$ be a vector of weights not all equal to 0 such that $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$.

WRE(s, c_0) CONDITION. *The design matrix $\mathbb{X}$ satisfies $\|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$ for all $j = 1, \dots, p$ and*

$$\vartheta(s, c_0) \triangleq \min_{\boldsymbol{\delta} \in \mathcal{C}_{\mathrm{WRE}}(s, c_0): \boldsymbol{\delta} \ne \mathbf{0}} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_n}{|\boldsymbol{\delta}|_2} > 0,$$

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

This condition is stated for any weights $\lambda_1 \ge \cdots \ge \lambda_p \ge 0$ but we will use it only for λ_j given in (2.5) and in that case the cone is equivalently defined as

$$\mathcal{C}_{\mathrm{WRE}}(s, c_0) = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : \sum_{j=1}^p \delta_j^\sharp \sqrt{\log(2p/j)} \le (1 + c_0)|\boldsymbol{\delta}|_2 \sqrt{\sum_{j=1}^s \log(2p/j)} \right\}.$$

Let us compare the WRE condition with the SRE condition. Assume that $\boldsymbol{\delta}$ belongs to the cone $\mathcal{C}_{\mathrm{SRE}}(s, c_0)$, that is, $|\boldsymbol{\delta}|_1 \le (1 + c_0)\sqrt{s}|\boldsymbol{\delta}|_2$. Then also $\sum_{j=s+1}^p \delta_j^\sharp \le (1 + c_0)\sqrt{s}|\boldsymbol{\delta}|_2$, and we have

$$\begin{aligned} \sum_{j=s+1}^p \delta_j^\sharp \sqrt{\log(2p/j)} &\le \sqrt{\log(2p/s)} \sum_{j=s+1}^p \delta_j^\sharp \le (1 + c_0)|\boldsymbol{\delta}|_2 \sqrt{s \log(2p/s)} \\ &\le (1 + c_0)|\boldsymbol{\delta}|_2 \sqrt{\sum_{j=1}^s \log(2p/j)}, \end{aligned}$$

where the last inequality follows from (2.7). For the first s components, the Cauchy–Schwarz inequality yields

$$\sum_{j=1}^{s} \delta_j^\sharp \sqrt{\log(2p/j)} \le |\boldsymbol{\delta}|_2 \sqrt{\sum_{j=1}^{s} \log(2p/j)}.$$

Combining the last two displays, we find that $\boldsymbol{\delta} \in \mathcal{C}_{\mathrm{WRE}}(s, 1+c_0)$. Thus, $\mathcal{C}_{\mathrm{SRE}}(s, c_0) \subseteq \mathcal{C}_{\mathrm{WRE}}(s, 1+c_0)$, so that the $\mathrm{WRE}(s, 1+c_0)$ condition implies the $\mathrm{SRE}(s, c_0)$ condition. A more detailed comparison between these two conditions as well as examples of random matrices, for which both conditions hold are given in Section 8. We are now ready to state our main result on the Slope estimator.

THEOREM 6.1. *Let $s \in \{1, \dots, p\}$, $\gamma \in (0,1)$ and $\tau \in [0, 1-\gamma)$. Set $c_0 = c_0(\gamma, \tau) = \frac{1+\gamma+\tau}{1-\gamma-\tau}$. Let the tuning parameters λ_j be defined by (2.5) with constant*

$$A \ge (4+\sqrt{2})/\gamma. \tag{6.1}$$

Let $\delta_0 \in (0,1)$. Then, on the event (4.1), the Slope estimator $\hat{\boldsymbol{\beta}}$ that minimizes (2.4) with the weights $\lambda_1, \dots, \lambda_p$ satisfies

$$2\tau|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_* + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + C'_{\gamma,\tau}(s, \delta_0) \sum_{j=1}^{s} \lambda_j^2 \tag{6.2}$$

simultaneously for all $\mathbf{f} \in \mathbb{R}^n$, all $s = 1, \dots, p$, and all $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\boldsymbol{\beta}|_0 = s$, where we set

$$C'_{\gamma,\tau}(s, \delta_0) \triangleq (1+\gamma+\tau)^2 \Big(\frac{\log(1/\delta_0)}{s \log(2p/s)} \vee \frac{1}{\vartheta^2(s, c_0(\gamma,\tau))} \Big), \qquad s = 1, \dots, p,$$

if $\mathrm{WRE}(s, c_0)$ holds, and $C'_{\gamma,\tau}(s, \delta_0) = \infty$ otherwise. Furthermore, if $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \le s$, then on the event (4.1) we have

$$2\tau|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_* \le C'_{\gamma,\tau}(s, \delta_0) \sum_{j=1}^{s} \lambda_j^2, \tag{6.3}$$

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \le \frac{C'_{\gamma,0}(s, \delta_0)}{1+\gamma} \Big(\sum_{j=1}^{s} \lambda_j^2 \Big)^{1/2}. \tag{6.4}$$

The proof of Theorem 6.1 is given in Section D. It follows the same route as the proof of Theorem 4.2. Since $\lambda_1, \dots, \lambda_p$ satisfy (2.5) then by (2.7), for all $s = 1, \dots, p$ we have

$$\frac{A^2\sigma^2 s \log(2p/s)}{n} \le \sum_{j=1}^{s} \lambda_j^2 \le \frac{A^2\sigma^2 s \log(2ep/s)}{n} \tag{6.5}$$

so that the Slope estimator achieves the optimal rate for the prediction error and the ℓ_2-estimation error. The presentation of Theorem 6.1 is similar to that of Theorem 4.2 for the Lasso, although there are some differences that will be highlighted after the following corollaries. Corollary 6.2 below is an immediate consequence of Theorems 4.1 and 6.1 with $\gamma = 1/2$, $\tau = 1/4$ and $\delta_0 = (\frac{s}{2p})^{\frac{s}{\vartheta^2(s,7)}}$ or $\tau = 0$ and $\delta_0 = (\frac{s}{2p})^{\frac{s}{\vartheta^2(s,3)}}$.

COROLLARY 6.2. *Let $s \in \{1, \dots, p\}$. Assume that the* WRE$(s, 7)$ *condition holds. Let $\hat{\boldsymbol{\beta}}$ be the Slope estimator with tuning parameters $\lambda_1, \dots, \lambda_p$ satisfying* (2.5) *for $A \geq 2(4 + \sqrt{2})$. Then, with probability at least $1 - \frac{1}{2}(\frac{s}{2p})^{\frac{s}{\vartheta^2(s,7)}}$, we have*

$$\frac{1}{2}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_* + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \leq \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{49\sum_{j=1}^s \lambda_j^2}{16\vartheta^2(s,7)}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\boldsymbol{\beta}|_0 \leq s$, and all $\mathbf{f} \in \mathbb{R}^n$. Furthermore, if $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \leq s$ then

$$\mathbb{P}\biggl(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2^2 \leq \frac{9\sum_{j=1}^s \lambda_j^2}{4\vartheta^4(s,3)}\biggr) \geq 1 - \frac{1}{2}\biggl(\frac{s}{2p}\biggr)^{\frac{s}{\vartheta^2(s,3)}}.$$

The fact that Theorems 4.1 and 6.1 hold for any $\delta_0 \in (0, 1)$ allows us to integrate the bounds (6.2) and (6.4) to obtain the following oracle inequalities and bounds on the estimation error in expectation.

COROLLARY 6.3. *Let $s \in \{1, \dots, p\}$. Assume that the* WRE$(s, 7)$ *condition holds. Let $\hat{\boldsymbol{\beta}}$ be the Slope estimator with tuning parameters $\lambda_1, \dots, \lambda_p$ satisfying* (2.5) *for $A \geq 2(4 + \sqrt{2})$. Then*

$$\mathbb{E}\biggl[\frac{1}{2}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_* + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2\biggr] \leq \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{49\sum_{j=1}^s \lambda_j^2}{16}\biggl(\frac{1}{\vartheta^2(s,7)} + \frac{1}{2\log(2p)}\biggr)$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\boldsymbol{\beta}|_0 \leq s$, and all $\mathbf{f} \in \mathbb{R}^n$. Furthermore, if $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \leq s$, then

$$\mathbb{E}\bigl[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2^2\bigr] \leq \frac{9\sum_{j=1}^s \lambda_j^2}{4}\biggl(\frac{1}{\vartheta^4(s,3)} + \frac{1}{(\log(2p))^2}\biggr).$$

Since $\hat{\boldsymbol{\beta}}$ does not depend on s, the first inequality in Corollary 6.3 and (6.5) imply a "balanced" oracle inequality:

$$\begin{aligned}(6.6)\qquad &\mathbb{E}\bigl[\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2\bigr] \\ &\quad\leq \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\biggl[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + C(|\boldsymbol{\beta}|_0)\frac{|\boldsymbol{\beta}|_0}{n}\log\biggl(\frac{2ep}{|\boldsymbol{\beta}|_0 \vee 1}\biggr)\biggr], \qquad \forall \mathbf{f} \in \mathbb{R}^n,\end{aligned}$$

where

$$C(|\boldsymbol{\beta}|_0) = \frac{49A^2\sigma^2}{16}\left(\frac{1}{\vartheta^2(|\boldsymbol{\beta}|_0, 7)} + \frac{1}{2\log(2p)}\right)$$

if $\vartheta^2(|\boldsymbol{\beta}|_0, 7) \neq 0$ and $C(|\boldsymbol{\beta}|_0) = \infty$ otherwise. This formulation might be of interest in the context of aggregation as explained, for example, in [24].

Corollaries 6.2 and 6.3 are the analogs of Corollaries 4.3 and 4.4 for the Lasso. The proof of Corollary 6.3 is omitted. It is deduced from Theorem 6.1 exactly in the same way as Corollary 4.4 is deduced from Theorem 4.2 in Section B.

The results in Section 4 and in the present section show that both the Lasso estimator with tuning parameter of order $\sigma\sqrt{\log(p/s)/n}$ and the Slope estimator with weights (2.5) achieve the optimal rate $(s/n)\log(p/s)$ for the ℓ_2-estimation and the prediction error. We now highlight some differences between these results on Slope and Lasso.

The first difference, in favor of Slope, is that Slope achieves the optimal rate adaptively to the unknown sparsity s. This was previously established in [25] for random design matrices $\mathbb{X}$ with i.i.d. $\mathcal{N}(0, 1)$ entries and in [16] for random $\mathbb{X}$ with independent sub-Gaussian isotropically distributed rows. The results of the present section show that, in reasonable generality, Slope achieves rate optimality for deterministic design matrices. Namely, it is enough to check a rather general condition WRE, which is only slightly more constraining than the RE condition commonly used in the context of Lasso. It is also shown in Section 8 that the WRE condition holds with high probability for a large class of random design matrices. This includes design matrices with i.i.d. anisotropically distributed rows, for example, matrices with i.i.d. rows distributed as $\mathcal{N}(\mathbf{0}, \Sigma)$ where $\Sigma \in \mathbb{R}^{p\times p}$ is not invertible.

The second difference is that our results for the Lasso are obtained in greater generality than for the Slope. Indeed, the SRE condition required in Section 4 for the Lasso is weaker that the WRE condition required here for the Slope. We refer to Section 8 for a more detailed comparison of these conditions. Furthermore, for $1 \leq q < 2$, in Section 4 we obtain rate optimal bounds on the ℓ_q-errors of the Lasso estimator, while for its Slope counterpart we can only control the rate in the $|\cdot|_*$ and ℓ_2 norms; cf. (6.3) and (6.4) (and all the interpolation norms in between but those are not classical to measure statistical performances). Of course, for the weights (2.5), the trivial relation $|\boldsymbol{\beta}|_* \geq C\sigma|\boldsymbol{\beta}|_1/\sqrt{n}$ holds, where $C > 0$ is a constant. This and (6.3) lead to a bound on the ℓ_1-error of the Slope estimator, which is however suboptimal. The same problem arises with the ℓ_q-norms with $1 < q < 2$ if the bounds are obtained by interpolation between such a suboptimal bound for the ℓ_1-error and the ℓ_2 bound (6.4).

Finally, note that the aggregation scheme proposed in Section 5 requires the knowledge of a lower bound θ_* on the SRE constants $\theta(\cdot, 7)$ in order to adaptively achieve the optimal rate $s\log(p/s)/n$. If $n > cs_*\log(p/s_*)$ for some numerical

constant $c > 0$, and the regressors are random and satisfy suitable assumptions (see Theorem 8.3), one can compute the numerical constant θ_* that gives the required bound with high probability. It can be also computed when the matrix $\mathbb{X}$ is deterministic and satisfies the mutual coherence condition, but in the general case, it is hard to compute such θ_*. On the other hand, the Slope estimator adaptively achieves the optimal rate $s\log(p/s)/n$ without the knowledge of any RE-type constants.

7. Minimax lower bounds. In this section, we provide the minimax lower bounds for the prediction risk and ℓ_q-estimation risk on the class $B_0(s)$. Several papers have addressed this issue for the prediction risk [1, 19, 23, 24, 27], for the ℓ_2-estimation risk [7, 23, 27] and for the ℓ_q-estimation risk with general q [19, 23, 29]. We are interested here in nonasymptotic bounds and, therefore, the results in [23, 29] obtained in some asymptotics do not fit in our context. Another issue is that the papers cited above, except for [19], deal with lower bounds for the expected squared risk or power risk [23, 29], and thus cannot be used to match the upper bounds in probability that are in the focus of our study in this paper. The only result that can be applied in this context is Theorem 6.1 in [19]. It gives a nonasymptotic lower bound for general loss functions under the condition that the ratio of minimal and maximal $2s$-sparse eigenvalue of the Gram matrix $\mathbb{X}^T\mathbb{X}/n$ is bounded from below by a constant. It matches our upper bounds both for the prediction risk and for the ℓ_q-estimation risk. Note that Theorem 6.1 in [19] deals with group sparsity and is therefore more general than in our setting. Thus, we only refer to the case $T = 1$ of Theorem 6.1 in [19] corresponding to ordinary sparsity. Here, we provide an improvement on it, in the sense that for the lower bound in ℓ_q, we drop the ratio of sparse eigenvalues condition. Thus, in the next theorem the lower bound for the ℓ_q-estimation risk holds for any design matrix $\mathbb{X}$. For the prediction risk, the lower bound that we state below is borrowed from [19], Theorem 6.1, and it is meaningful only if the minimal sparse eigenvalue is positive. For any matrix $\mathbb{X} \in \mathbb{R}^{n\times p}$ and any $s \in [1, p]$, define the minimal and maximal s-sparse eigenvalues as follows:

$$\bar{\theta}_{\min}(\mathbb{X}, s) \triangleq \min_{\boldsymbol{\delta}\in B_0(s)\setminus\{0\}} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_n}{|\boldsymbol{\delta}|_2}, \qquad \bar{\theta}_{\max}(\mathbb{X}, s) \triangleq \max_{\boldsymbol{\delta}\in B_0(s)\setminus\{0\}} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_n}{|\boldsymbol{\delta}|_2}. \tag{7.1}$$

In particular, $\bar{\theta}_{\max}(\mathbb{X}, 1) = \max_{j=1,\dots,p} \|\mathbb{X}\boldsymbol{e}_j\|_n$.

Define

$$\psi_{n,q} = \sigma s^{1/q}\sqrt{\frac{\log(ep/s)}{n}}, \qquad 1 \le q \le \infty,$$

where we set $s^{1/\infty} \triangleq 1$. Let $\mathbb{E}_{\boldsymbol{\beta}}$ denote the expectation with respect to the measure $\mathbb{P}_{\boldsymbol{\beta}}$.

THEOREM 7.1. *Let $p \ge 2$, $s \in [1, p/2]$, $n \ge 1$ be integers, and let $1 \le q \le \infty$. Let $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ be a nondecreasing function such that $\ell(0) = 0$ and $\ell \not\equiv 0$. Assume that $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$, $\sigma > 0$. Then the following holds*:

(i) *There exist positive constants $\bar{b}, \bar{c}$ depending only on $\ell(\cdot)$ and q such that*

$$\inf_{\hat{\boldsymbol{\tau}}} \inf_{\mathbb{X}} \sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{E}_{\boldsymbol{\beta}^*} \ell\big(\bar{b}\psi_{n,q}^{-1} \bar{\theta}_{\max}(\mathbb{X}, 1) |\hat{\boldsymbol{\tau}} - \boldsymbol{\beta}^*|_q\big) \ge \bar{c}, \tag{7.2}$$

where $\inf_{\hat{\boldsymbol{\tau}}}$ denotes the infimum over all estimators $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\beta}^*$, and $\inf_{\mathbb{X}}$ denotes the infimum over all matrices $\mathbb{X} \in \mathbb{R}^{n\times p}$.

(ii) *There exist positive constants $\bar{b}, \bar{c}$ depending only on $\ell(\cdot)$ such that*

$$\inf_{\hat{\boldsymbol{\tau}}} \inf_{\mathbb{X}} \sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{E}_{\boldsymbol{\beta}^*} \ell\bigg(\bar{b}\psi_{n,2}^{-1} \frac{\bar{\theta}_{\max}(\mathbb{X}, 1)}{\bar{\theta}_{\min}(\mathbb{X}, 2s)} \|\mathbb{X}(\hat{\boldsymbol{\tau}} - \boldsymbol{\beta}^*)\|_n\bigg) \ge \bar{c}, \tag{7.3}$$

where, by definition, the expression under the expectation is $+\infty$ for matrices $\mathbb{X}$ such that $\bar{\theta}_{\min}(\mathbb{X}, 2s) = 0$.

PROOF. We first prove part (i). We use Lemma F.1 stated in the Appendix. Let $\mathcal{B} = \{\boldsymbol{\beta} = a\boldsymbol{\omega} : \boldsymbol{\omega} \in \Omega\}$, where Ω is a subset of $\{1, 0, -1\}^p$ described in Lemma F.1, $a = \alpha\bar{\theta}_{\max}^{-1}(\mathbb{X}, 1)\psi_{n,q} s^{-1/q}$ and $0 < \alpha < \tilde{c}^{1/2}/4$, where $\tilde{c}$ is a constant appearing in Lemma F.1. It follows from Lemma F.1 that $\mathcal{B} \subset B_0(s)$, and

$$|\boldsymbol{\beta} - \boldsymbol{\beta}'|_q \ge 4^{-1/q} \alpha \bar{\theta}_{\max}^{-1}(\mathbb{X}, 1)\psi_{n,q} \tag{7.4}$$

for any two distinct elements $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ of $\mathcal{B}$. Again from Lemma F.1, for all $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ in $\mathcal{B}$, the Kullback–Leibler divergence $\mathcal{K}(\mathbb{P}_{\boldsymbol{\beta}}, \mathbb{P}_{\boldsymbol{\beta}'})$ between the probability measures $\mathbb{P}_{\boldsymbol{\beta}}$ and $\mathbb{P}_{\boldsymbol{\beta}'}$ satisfies

$$\begin{aligned} \mathcal{K}(\mathbb{P}_{\boldsymbol{\beta}}, \mathbb{P}_{\boldsymbol{\beta}'}) &= \frac{n}{2\sigma^2} \|\mathbb{X}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_n^2 \le \frac{\alpha^2 n}{\sigma^2} \psi_{n,q}^2 s^{1-2/q} \\ &= \frac{\alpha^2 s}{n} \log\bigg(\frac{ep}{s}\bigg) \le \frac{\tilde{c}s}{16n} \log\bigg(\frac{ep}{s}\bigg) \le \frac{1}{16} \log(\operatorname{Card} \mathcal{B}). \end{aligned} \tag{7.5}$$

The bound (7.2) now follows from (7.4) and (7.5) in view of [26], Theorem 2.7.

To prove part (ii), set $q = 2$ and let $\mathcal{B}$, Ω and a be as above. Then (7.4) implies that for any $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathcal{B}$ we have

$$\|\mathbb{X}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_n^2 \ge \bar{\theta}_{\min}(\mathbb{X}, 2s) |\boldsymbol{\beta} - \boldsymbol{\beta}'|_2 \ge 4^{-1/2} \alpha \frac{\bar{\theta}_{\min}(\mathbb{X}, 2s)}{\bar{\theta}_{\max}(\mathbb{X}, 1)} \psi_{n,2}. \tag{7.6}$$

The bound (7.3) follows from (7.6) and (7.5) in view of [26], Theorem 2.7. □

As a consequence of Theorem 7.1, we get, for example, lower bounds for the squared loss $\ell(u) = u^2$ and for the indicator loss $\ell(u) = I\{u \ge 1\}$. The indicator loss is relevant for comparison with the upper bounds in probability obtained in the

previous sections. For example, Theorem 7.1 with this loss and $1 \le q \le 2$ implies that for any estimator $\hat{\boldsymbol{\tau}}$ there exists $\boldsymbol{\beta}^* \in B_0(s)$ such that, with $\mathbb{P}_{\boldsymbol{\beta}^*}$-probability at least $\bar{c} > 0$,

$$\|\mathbb{X}(\hat{\boldsymbol{\tau}} - \boldsymbol{\beta}^*)\|_n^2 \ge C\frac{\sigma^2 s}{n}\log\left(\frac{ep}{s}\right)$$

and

$$|\hat{\boldsymbol{\tau}} - \boldsymbol{\beta}^*|_q \ge C\frac{\sigma s^{1/q}}{\sqrt{n}}\log\left(\frac{ep}{s}\right)^{1/2}, \qquad 1 \le q \le 2,$$

where $\bar{c} > 0$ is a numerical constant and $C > 0$ is some constant depending only on $\mathbb{X}$. The rates on the right-hand side of these inequalities have the same form as in the corresponding upper bounds for the Lasso and Slope estimators obtained in Corollaries 4.3 and 6.2. The fact that the constants C here depend on the design implies that the optimality is not guaranteed for all configurations of n, s, p. Thus, we get the rate optimality under the assumption that $s\log(ep/s) < cn$ for the ℓ_2-risk, and under the assumption $s\log(ep/s) < cR$ for the prediction risk, where $c > 0$ is a constant. Here, R denotes the rank of matrix $\mathbb{X}$. Concerning the prediction risk, this remark is based on the following fact.

COROLLARY 7.2. *Let $p \ge 2$, $s \in [1, p/2]$ and $n \ge 1$ be integers. If for some matrix $\mathbb{X} \in \mathbb{R}^{n\times p}$ with rank R and some $b > 0$, we have $\frac{\bar{\theta}_{\max}(\mathbb{X},1)}{\bar{\theta}_{\min}(\mathbb{X},2s)} \le b$, then there exists $c = c(b) > 0$ such that $s\log(ep/s) < cR$.*

This corollary follows immediately from (7.3) with $\ell(u) = u^2$ and the fact that the minimax expected squared risk is bounded from above by $\sigma^2 R/n$ (cf., e.g., [24]).

In view of Corollary 7.2, the bound (7.3) is nontrivial only when $s\log(ep/s) < cR$. The bound (7.2) does not have such a restriction and remains nontrivial for all n, s, p. However, it is known that for $q = 2$ and $s\log(ep/s) \gg n$ this bound is not optimal [27]. Anyway, (7.2) shows that if $s\log(ep/s) \gg n$, the ℓ_2-risk diverges, so this case is of minor interest. Also note that the upper bounds of Theorems 4.2, 6.1 and their corollaries rely on RE type conditions, and those conditions imply that $s\log(ep/s) < cn$. This follows from Proposition 2.2.18 in [9] and the fact that the RE condition implies the exact reconstruction property by ℓ_1 minimization as defined in [9].

8. Assumptions on the design matrix.

8.1. *Equivalence between* RE, SRE *and s-sparse eigenvalue conditions.* Along with the RE and SRE conditions defined in Section 4, we consider here the s-sparse eigenvalue condition defined as follows, for any $s \in \{1, \dots, p\}$.

s-SPARSE EIGENVALUE CONDITION. *The design matrix* $\mathbb{X}$ *satisfies* $\|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$ *for all* $j = 1, \dots, p$, *and* $\bar{\theta}_{\min}(\mathbb{X}, s) > 0$.

In this section, we will write for brevity $\bar{\theta}_{\min}(s) = \bar{\theta}_{\min}(\mathbb{X}, s)$. The next proposition establishes the equivalence between the three conditions mentioned above.

PROPOSITION 8.1. *Let* $c_0 > 0$ *and* $s \in \{1, \dots, p\}$. *We have the following implications*:

(i) *If condition* $\mathrm{SRE}(s, c_0)$ *holds*, *then condition* $\mathrm{RE}(s, c_0)$ *holds and* $\kappa(s, c_0) \ge \theta(s, c_0)$.

(ii) *If condition* $\mathrm{RE}(s, c_0)$ *holds*, *then the* s*-sparse eigenvalue condition holds and* $\bar{\theta}_{\min}(s) \ge \kappa(s, c_0)$.

(iii) *Let* $\theta_1 > 0$. *If the* s*-sparse eigenvalue condition holds with* $\bar{\theta}_{\min}(s) \ge \theta_1$, *then the* $\mathrm{SRE}(s_1, c_0)$ *condition holds and* $\theta(s_1, c_0) \ge \theta_1/\sqrt{2}$ *for* $s_1 \le (s-1)\theta_1^2/(2c_0^2)$.

PROOF. Part (i) follows from (4.3). Next, if $\boldsymbol{\delta} \in B_0(s)$ then obviously $|\boldsymbol{\delta}|_1 \le (1+c_0)\sum_{j=1}^{s} \delta_j^\sharp$ with $c_0 = 0$. Thus, the set of all s-sparse vectors $B_0(s)$ is included in the cone $\mathcal{C}_{\mathrm{RE}}(s, c_0)$ for any $c_0 > 0$. This implies (ii). To prove (iii), we use Lemma 2.7 in [17], which implies that if $\bar{\theta}_{\min}(s) \ge \theta_1$, and $\|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$ for all $j = 1, \dots, p$, then $\|\mathbb{X}\boldsymbol{\delta}\|_n^2 \ge \theta_1^2|\boldsymbol{\delta}|_2^2 - |\boldsymbol{\delta}|_1^2/(s-1)$ for all $\boldsymbol{\delta} \in \mathbb{R}^p$. □

The message of the above proposition is that the three conditions—$\mathrm{RE}(s, c_0)$, $\mathrm{SRE}(s, c_0)$ and the s-sparse eigenvalue condition—are equivalent up to absolute constants. This equivalence has two main consequences for the results of the present paper:

- First, the results on the Lasso in Sections 4 and 5 are proved under the $\mathrm{SRE}(s, c_0)$ condition. The above equivalence shows that, for some integer s_1, which is of the same order as s, the oracle inequalities and the estimation bounds of Sections 4 and 5 are valid under the Restricted Eigenvalue condition $\mathrm{RE}(s_1, c_0)$.
- Second, the s-sparse eigenvalue condition is known to hold with high probability for rather general random matrices with i.i.d. rows. By the above equivalence, conditions $\mathrm{RE}(s, c_0)$ and $\mathrm{SRE}(s, c_0)$ are satisfied for the same random matrices. A useful sufficient condition for the s-sparse eigenvalue condition is the small ball condition [15, 20, 22]. A random vector $\boldsymbol{x}$ valued in $\mathbb{R}^p$ is said to satisfy the small ball condition over $B_0(s_1)$ if there exist positive numbers u and β such that

$$\mathbb{P}\big[|\boldsymbol{\delta}^T\boldsymbol{x}| \ge u|\boldsymbol{\delta}|_2\big] \ge \beta \qquad \forall \boldsymbol{\delta} \in B_0(s_1). \tag{8.1}$$

Let $\mathbb{X} \in \mathbb{R}^{n \times p}$ be a matrix with i.i.d. rows that have the same distribution as $\boldsymbol{x}$ satisfying (8.1). Corollary 2.5 in [17] establishes that, for such $\mathbb{X}$

we have $\bar{\theta}_{\min}(s_1) > u/\sqrt{2}$ with probability at least $1 - \exp(-Cn\beta^2)$ if $n \geq (C'/\beta^2)s\log(ep/s)$ for some absolute constants $C, C' > 0$.

Note that condition (8.1) is very mild. For instance, a vector $\boldsymbol{x}$ with independent components that have a Cauchy distribution satisfies this condition (to see this, notice that for any nonzero $\boldsymbol{\delta} \in \mathbb{R}^p$, the random variable $\boldsymbol{\delta}^T\boldsymbol{x}/|\boldsymbol{\delta}|_2$ has Cauchy distribution with parameter $|\boldsymbol{\delta}|_1/|\boldsymbol{\delta}|_2 \geq 1$, hence (8.1) holds). Thus, condition (8.1) is quite different in nature from any concentration property. On the other hand, the property $\max_{j=1,\dots,p} \|\mathbb{X}\boldsymbol{e}_j\|_n \leq 1$ assumed in the above three conditions (which is usually seen as a simple normalization) requires concentration. Indeed, this inequality can be written as

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{e}_j)^2 \leq 1 \qquad \forall j = 1,\dots,p,$$

where $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ are the i.i.d. rows of $\mathbb{X}$. We have here a sum of i.i.d. positive random variables. Satisfying $\max_{j=1,\dots,p} \|\mathbb{X}\boldsymbol{e}_j\|_n \leq 1$ with high probability requires $\mathbb{E}[(\boldsymbol{x}^T\boldsymbol{e}_j)^2] < 1$ and some concentration property of the random variables $(\boldsymbol{x}^T\boldsymbol{e}_j)^2$ for all $j = 1,\dots,p$. It is proved in [17] that this property holds with high probability when the components of $\boldsymbol{x}$ (that do not have to be independent) have moments with a polynomial growth up to the order $\log(ep)$ and that this condition may be violated with probability greater than $1/2$ if the coordinates only have $\log(ep)/\log\log(ep)$ such moments.

In conclusion, for a large class of random matrices with i.i.d. rows, condition $\mathrm{SRE}(s, c_0)$ holds with high probability if

$$s\log(ep/s) \leq cn, \tag{8.2}$$

where $c > 0$ is a constant.

8.2. *Design conditions for the Slope estimator.* Theorem 6.1 and Corollaries 6.2, 6.3 establish prediction and estimation bounds for the Slope estimator under the $\mathrm{WRE}(s, c_0)$ condition. It was explained in Section 6 that the $\mathrm{WRE}(s, c_0)$ condition implies the $\mathrm{SRE}(s, 1 + c_0)$ condition. The converse is not true; there is no equivalence between the two conditions. However, a simple observation leads to the following sufficient condition for $\mathrm{WRE}(s, c_0)$.

PROPOSITION 8.2. *Let $s \in \{1,\dots,p\}$, $c_0 > 0$, and let the weights λ_j be defined by (2.5). Set $s_2 = \lceil s\log(2ep/s)/\log 2\rceil$. If the* $\mathrm{SRE}(s_2, c_0)$ *condition holds then the* $\mathrm{WRE}(s, c_0)$ *condition holds, and $\vartheta(s, c_0) \geq \theta(s_2, c_0)$.*

PROOF. If $\boldsymbol{\delta} \in \mathcal{C}_{\mathrm{WRE}}(s, c_0)$, then

$$(1 + c_0)\left(\sum_{j=1}^{s}\lambda_j^2\right)^{1/2}|\boldsymbol{\delta}|_2 \geq |\boldsymbol{\delta}|_* = \sum_{j=1}^{p}\lambda_j\delta_j^\sharp \geq \lambda_p\sum_{j=1}^{p}\delta_j^\sharp = \lambda_p|\boldsymbol{\delta}|_1.$$

This, together with (2.5) and (2.7) imply $|\boldsymbol{\delta}|_1 \le (1+c_0)\sqrt{s\log(2ep/s)/\log 2}|\boldsymbol{\delta}|_2$. Thus, $\boldsymbol{\delta} \in \mathcal{C}_{\mathrm{SRE}}(s_2, c_0)$. $\square$

Proposition 8.2 implies that, under the same assumptions as discussed in Section 8.1, for large classes of random matrices with i.i.d. rows, condition $\mathrm{WRE}(s, c_0)$ holds with high probability whenever $s\log^2(ep/s) \le cn$ where $c > 0$ is a constant. This inequality on s, p and n differs from (8.2) only in an extra logarithmic factor. Moreover, the next theorem shows that this extra factor is not necessary if the row vectors of $\mathbb{X}$ are sub-Gaussian.

THEOREM 8.3. *There exist absolute constants $C, C' > 0$ such that the following holds. Let $c_0, \kappa > 0$ and let $s \in \{1, \dots, p\}$. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be i.i.d. copies of a mean zero random variable $\boldsymbol{x}$ valued in $\mathbb{R}^p$ with covariance matrix $\Sigma = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = (\Sigma_{ij})_{1\le i,j\le p}$. Let $L \ge 1$ and assume that $\boldsymbol{x}$ is L-sub-Gaussian in the sense that*

$$\mathbb{E}\exp(\boldsymbol{\delta}^T\boldsymbol{x}) \le \exp\left(\frac{L^2|\Sigma^{1/2}\boldsymbol{\delta}|_2^2}{2}\right) \qquad \forall \boldsymbol{\delta} \in \mathbb{R}^p. \tag{8.3}$$

Assume that the covariance matrix Σ satisfies

$$\max_{j=1,\dots,p} \Sigma_{jj} \le \frac{1}{2}, \qquad \min_{\boldsymbol{\delta}\in\mathcal{C}_{\mathrm{WRE}}(s,c_0):\boldsymbol{\delta}\ne\mathbf{0}} \frac{|\Sigma^{1/2}\boldsymbol{\delta}|_2}{|\boldsymbol{\delta}|_2} \ge \kappa. \tag{8.4}$$

Let $\mathbb{X}$ be the random matrix in $\mathbb{R}^{n\times p}$ with row vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. If

$$n \ge \frac{CL^4(1+c_0)^2}{\kappa^2} s\log(2ep/s) \tag{8.5}$$

then, with probability at least $1 - 3\exp(-C'n/L^4)$ we have

$$\max_{j=1,\dots,p} \|\mathbb{X}\boldsymbol{e}_j\|_n^2 \le 1, \qquad \inf_{\boldsymbol{\delta}\in\mathcal{C}_{\mathrm{WRE}}(s,c_0):\boldsymbol{\delta}\ne\mathbf{0}} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_n}{|\boldsymbol{\delta}|_2} \ge \frac{\kappa}{\sqrt{2}}. \tag{8.6}$$

9. Extension to sub-Gaussian noise. The goal of this section is to show that all results of the present paper extend to sub-Gaussian noise. This is due to the following analog of Theorem 4.1.

THEOREM 9.1. *Let $\delta_0 \in (0,1)$. Assume that the components of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ are independent, with zero mean and sub-Gaussian in the sense that, for some $\sigma > 0$,*

$$\mathbb{E}\exp(\xi_i^2/\sigma^2) \le e, \qquad i = 1, \dots, n. \tag{9.1}$$

Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ be any matrix such that $\max_{j=1,\dots,p} \|\mathbb{X}\boldsymbol{e}_j\|_n \le 1$. Then, with probability at least $1 - \delta_0$ we have, for all $\boldsymbol{u} \in \mathbb{R}^p$,

$$\frac{1}{n}\boldsymbol{\xi}^T\mathbb{X}\boldsymbol{u} \le 40\sigma \max\left(\sum_{j=1}^p u_j^\sharp \sqrt{\frac{\log(2p/j)}{n}}, \|\mathbb{X}\boldsymbol{u}\|_n \frac{\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)}}{\sqrt{n}}\right).$$

Theorem 9.1 implies that $\frac{1}{n}\boldsymbol{\xi}^T\mathbb{X}\boldsymbol{u} \le \max(\bar{H}(\boldsymbol{u}), \bar{G}(\boldsymbol{u}))$ with probability at least $1-\delta_0$, where $\bar{H}(\cdot)$ and $\bar{G}(\cdot)$ have the same form as $H(\cdot)$ and $G(\cdot)$ up to numerical constants.Thus, under the sub-Gaussian assumption of Theorem 9.1, all results of Sections 4–6 remain valid up to differences in the numerical constants.

The proof of Theorem 9.1 relies on the following deviation inequality, which is proved in Appendix H using symmetrization and contraction arguments.

PROPOSITION 9.2. *Assume that the components of $\boldsymbol{\xi}$ are independent, with zero mean, and satisfy (9.1). Let $U \subseteq \{\boldsymbol{u} \in \mathbb{R}^n : |\boldsymbol{u}|_2 \le 1\}$ be a subset of the unit ball. For any $x > 0$, with probability at least $1-\exp(-x)$ we have*

$$\begin{aligned}\sup_{\boldsymbol{u}\in U} \boldsymbol{\xi}^T\boldsymbol{u} &\le 8\sigma\mathbb{E}\Big[\sup_{\boldsymbol{u}\in U} \boldsymbol{z}^T\boldsymbol{u}\Big] + 8\sigma\sqrt{2x} \\ &\le 8\sigma\,\mathrm{Med}\Big[\sup_{\boldsymbol{u}\in U} \boldsymbol{z}^T\boldsymbol{u}\Big] + 8\sigma(\sqrt{\pi/2}+\sqrt{2x}),\end{aligned} \tag{9.2}$$

where $\boldsymbol{z}$ is a standard normal $\mathcal{N}(\mathbf{0}, I_{n\times n})$ random vector.

APPENDIX A: PRELIMINARIES FOR THE PROOFS

LEMMA A.1. *Let $s \in \{1,\dots,p\}$ and $\tau \in [0,1]$. For any two $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that $|\boldsymbol{\beta}|_0 \le s$ we have*

$$\tau|\boldsymbol{u}|_* + |\boldsymbol{\beta}|_* - |\hat{\boldsymbol{\beta}}|_* \le (1+\tau)\Bigg(\sum_{j=1}^s \lambda_j^2\Bigg)^{1/2}|\boldsymbol{u}|_2 - (1-\tau)\sum_{j=s+1}^p \lambda_j u_j^\sharp, \tag{A.1}$$

where $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (u_1,\dots,u_p)$ and $(u_1^\sharp,\dots,u_p^\sharp)$ is a nonincreasing rearrangement of $(|u_1|,\dots,|u_p|)$. If $\lambda_1 = \cdots = \lambda_p = \lambda$ for some $\lambda > 0$, then $|\cdot|_* = \lambda|\cdot|_1$ and (A.1) yields

$$\tau\lambda|\boldsymbol{u}|_1 + \lambda|\boldsymbol{\beta}|_1 - \lambda|\hat{\boldsymbol{\beta}}|_1 \le (1+\tau)\lambda\sqrt{s}|\boldsymbol{u}|_2 - (1-\tau)\lambda\sum_{j=s+1}^p u_j^\sharp. \tag{A.2}$$

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

$$|\boldsymbol{\beta}|_* = \sum_{j=1}^s \lambda_j|\beta_{\phi(j)}| \quad\text{and}\quad |u_{\phi(s+1)}| \ge |u_{\phi(s+2)}| \ge \cdots \ge |u_{\phi(p)}|. \tag{A.3}$$

By (2.3) applied to $|\hat{\boldsymbol{\beta}}|_*$, we have

$$\begin{aligned}|\boldsymbol{\beta}|_* - |\hat{\boldsymbol{\beta}}|_* &\le \sum_{j=1}^s \lambda_j\big(|\beta_{\phi(j)}| - |\hat{\beta}_{\phi(j)}|\big) - \sum_{j=s+1}^p \lambda_j|\hat{\beta}_{\phi(j)}| \\ &\le \sum_{j=1}^s \lambda_j|u_{\phi(j)}| - \sum_{j=s+1}^p \lambda_j|\hat{\beta}_{\phi(j)}| = \sum_{j=1}^s \lambda_j|u_{\phi(j)}| - \sum_{j=s+1}^p \lambda_j|u_{\phi(j)}|,\end{aligned}$$

since $u_{\phi(j)} = \hat{\beta}_{\phi(j)} - \beta_{\phi(j)}$ for $j = 1, \dots, s$ and $u_{\phi(j)} = \hat{\beta}_{\phi(j)}$ for all $j > s$. Since the sequence λ_j is nonincreasing, we have $\sum_{j=1}^{s} \lambda_j |u_{\phi(j)}| \le \sum_{j=1}^{s} \lambda_j u_j^\sharp$. Next, the fact that permutation ϕ satisfies (A.3) implies $\sum_{j=s+1}^{p} \lambda_j u_j^\sharp \le \sum_{j=s+1}^{p} \lambda_j |u_{\phi(j)}|$. Finally, $\sum_{j=1}^{s} \lambda_j u_j^\sharp \le (\sum_{j=1}^{s} \lambda_j^2)^{1/2} |\boldsymbol{u}|_2$ by the Cauchy–Schwarz inequality. □

LEMMA A.2. *Let $h : \mathbb{R}^p \to \mathbb{R}$ be a convex function, let $\mathbf{f}, \boldsymbol{\xi} \in \mathbb{R}^n$, $\mathbf{y} = \mathbf{f} + \boldsymbol{\xi}$ and let $\mathbb{X}$ be any $n \times p$ matrix. If $\hat{\boldsymbol{\beta}}$ is a solution of the minimization problem $\min_{\boldsymbol{\beta} \in \mathbb{R}^p} (\|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2 + h(\boldsymbol{\beta}))$, then $\hat{\boldsymbol{\beta}}$ satisfies for all $\boldsymbol{\beta} \in \mathbb{R}^p$*

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \le \frac{2}{n} \boldsymbol{\xi}^T \mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + h(\boldsymbol{\beta}) - h(\hat{\boldsymbol{\beta}}) - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2. \tag{A.4}$$

PROOF. Define the functions f and g by the relations $g(\boldsymbol{\beta}) = \|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2$, and $f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Since f is convex and $\hat{\boldsymbol{\beta}}$ is a minimizer of f, it follows that $\mathbf{0}$ belongs to the subdifferential of f at $\hat{\boldsymbol{\beta}}$. By the Moreau–Rockafellar theorem, there exists $\boldsymbol{v}$ in the subdifferential of h at $\hat{\boldsymbol{\beta}}$ such that $\mathbf{0} = \nabla g(\hat{\boldsymbol{\beta}}) + \boldsymbol{v}$. Here, $\nabla g(\hat{\boldsymbol{\beta}}) = \frac{2}{n} \mathbb{X}^T (\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{y})$. Using these remarks and some algebra, we obtain

$$\begin{aligned}
&\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \\
&\quad = \frac{2}{n} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbb{X}^T (\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}) - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2 \\
&\quad = \frac{2}{n} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbb{X}^T (\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}) - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \big(\nabla g(\hat{\boldsymbol{\beta}}) + \boldsymbol{v}\big) \\
&\quad = \frac{2}{n} \boldsymbol{\xi}^T \mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{v}.
\end{aligned}$$

To complete the proof, notice that by definition of the subdifferential of h at $\hat{\boldsymbol{\beta}}$, we have $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{v} \le h(\boldsymbol{\beta}) - h(\hat{\boldsymbol{\beta}})$. □

LEMMA A.3. *Let $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, I_{p \times p})$, and let $f : \mathbb{R}^p \to \mathbb{R}$ be a 1-Lipschitz function, that is, $|f(\boldsymbol{u}) - f(\boldsymbol{u}')| \le |\boldsymbol{u} - \boldsymbol{u}'|_2$ for any $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}^p$. Then $|\operatorname{Med}[f(\boldsymbol{z})] - \mathbb{E}[f(\boldsymbol{z})]| \le \sqrt{\pi/2}$.*

This lemma is proved in the discussion after equation (1.6) in [18], p. 21.

APPENDIX B: PROOFS FOR THE LASSO ESTIMATOR

PROOF OF THEOREM 4.2. Using inequality (A.4) with $h(\cdot) = 2\lambda |\cdot|_1$ we get that, almost surely, for all $\boldsymbol{\beta} \in \mathbb{R}^p$ and all $\mathbf{f} \in \mathbb{R}^n$,

$$2\tau\lambda |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2 + \triangle^*, \tag{B.1}$$

where

$$\triangle^* \triangleq 2\tau\lambda|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_1+\frac{2}{n}\boldsymbol{\xi}^T\mathbb{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+2\lambda|\boldsymbol{\beta}|_1-2\lambda|\hat{\boldsymbol{\beta}}|_1.$$

Let $\boldsymbol{u}=\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ and assume that $|\boldsymbol{\beta}|_0\le s$. Define

$$\tilde{H}(\boldsymbol{u}) \triangleq \frac{(4+\sqrt{2})\sigma}{\sqrt{n}}\left(|\boldsymbol{u}|_2\left(\sum_{j=1}^{s}\log(2p/j)\right)^{1/2}+\sum_{j=s+1}^{p}u_j^\sharp\sqrt{\log(2p/j)}\right). \tag{B.2}$$

Using the Cauchy–Schwarz inequality, it is easy to see that

$$H(\boldsymbol{u})\le\tilde{H}(\boldsymbol{u})\le\gamma\lambda\left(\sqrt{s}|\boldsymbol{u}|_2+\sum_{j=s+1}^{p}u_j^\sharp\right)\triangleq F(\boldsymbol{u})\qquad\forall\boldsymbol{u}\in\mathbb{R}^p, \tag{B.3}$$

where $H(\cdot)$ is defined in (2.8), and the last inequality follows from (2.7) and (4.5).

On the event (4.1), using (B.3) and Lemma A.1 we obtain

$$\begin{aligned}\triangle^* &\le 2\lambda\bigl(\tau|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_1+|\boldsymbol{\beta}|_1-|\hat{\boldsymbol{\beta}}|_1\bigr)+2\max\bigl(H(\boldsymbol{u}),G(\boldsymbol{u})\bigr)\\ &\le 2\lambda\left((1+\tau)\sqrt{s}|\boldsymbol{u}|_2-(1-\tau)\sum_{j=s+1}^{p}u_j^\sharp\right)+2\max\bigl(F(\boldsymbol{u}),G(\boldsymbol{u})\bigr).\end{aligned}$$

By definition of $\delta(\lambda)$, we have

$$G(\boldsymbol{u})=\lambda\sqrt{s}\gamma\sqrt{\log(1/\delta_0)/\bigl(s\log\bigl(1/\delta(\lambda)\bigr)\bigr)}\|\mathbb{X}\boldsymbol{u}\|_n.$$

We now consider the following two cases:

(i) *Case* $G(\boldsymbol{u})>F(\boldsymbol{u})$. Then

$$|\boldsymbol{u}|_2\le\sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}}\|\mathbb{X}\boldsymbol{u}\|_n. \tag{B.4}$$

Thus,

$$\begin{aligned}\triangle^* &\le 2\lambda(1+\tau)\sqrt{s}|\boldsymbol{u}|_2+2G(\boldsymbol{u})\\ &\le 2\lambda\sqrt{s}(1+\tau+\gamma)\sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}}\|\mathbb{X}\boldsymbol{u}\|_n\\ &\le \lambda^2 s(1+\tau+\gamma)^2\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}+\|\mathbb{X}\boldsymbol{u}\|_n^2.\end{aligned} \tag{B.5}$$

(ii) *Case* $G(\boldsymbol{u})\le F(\boldsymbol{u})$. In this case, we get

$$\triangle^*\le 2\lambda\left((1+\gamma+\tau)\sqrt{s}|\boldsymbol{u}|_2-(1-\gamma-\tau)\sum_{j=s+1}^{p}u_j^\sharp\right)\triangleq\triangle.$$

If $\triangle > 0$, then $\boldsymbol{u}$ belongs to the cone $\mathcal{C}_{\mathrm{SRE}}(s, c_0)$ and we can use the $\mathrm{SRE}(s, c_0)$ condition, which yields $|\boldsymbol{u}|_2 \le \|\mathbb{X}\boldsymbol{u}\|_n/\theta(s, c_0)$. Therefore,

$$\triangle^* \le \triangle \le \frac{2(1+\gamma+\tau)\lambda\sqrt{s}}{\theta(s, c_0)}\|\mathbb{X}\boldsymbol{u}\|_n \le \left(\frac{(1+\gamma+\tau)\lambda\sqrt{s}}{\theta(s, c_0)}\right)^2 + \|\mathbb{X}\boldsymbol{u}\|_n^2. \tag{B.6}$$

If $\triangle \le 0$, then (B.6) holds trivially.

Combining (B.5) and (B.6) with (B.1) completes the proof of (4.6).

Let now $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \le s$. Then (4.6) with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ implies

$$2\tau|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \le C_{\gamma,\tau}(s, \lambda, \delta_0)\lambda s. \tag{B.7}$$

Next, we show that

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \le \frac{C_{\gamma,0}(s, \lambda, \delta_0)}{1+\gamma}\frac{\lambda\sqrt{s}}{\theta^2(s, \frac{1+\gamma}{1-\gamma})}. \tag{B.8}$$

To prove (B.8), we take $\tau = 0$, and consider the cases (i) and (ii) as above with $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$.

- If $G(\boldsymbol{u}) > F(\boldsymbol{u})$, then from (B.1) and (B.5) with $\tau = 0$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ we get

$$\|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_n^2 \le \lambda^2(1+\gamma)^2\frac{\log(1/\delta_0)}{\log(1/\delta(\lambda))}.$$

  This and (B.4) imply

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \le (1+\gamma)\lambda\sqrt{s}\left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}\right). \tag{B.9}$$

- If $G(\boldsymbol{u}) \le F(\boldsymbol{u})$, then it follows from (B.1) with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ that $\triangle^* \ge 0$ almost surely, and thus $\triangle \ge \triangle^* \ge 0$. Hence, $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathcal{C}_{\mathrm{SRE}}(s, \frac{1+\gamma}{1-\gamma})$. Thus, we can apply the $\mathrm{SRE}(s, \frac{1+\gamma}{1-\gamma})$ condition, which yields

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \le \frac{\|\mathbb{X}\boldsymbol{u}\|_n}{\theta(s, \frac{1+\gamma}{1-\gamma})} \le \frac{(1+\gamma)\lambda\sqrt{s}}{\theta^2(s, \frac{1+\gamma}{1-\gamma})}, \tag{B.10}$$

  where the second inequality is due to the combination of (B.1) and (B.6) with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\tau = 0$.

Putting together (B.9) and (B.10) proves (B.8). To conclude, it is enough to notice that $\theta^2(s, \frac{1+\gamma}{1-\gamma}) \ge \theta^2(s, c_0)$ and then to bound $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_q$ from above using (B.7), (B.8) and the norm interpolation inequality $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_q \le |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1^{2/q-1}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2^{2-2/q}$. □

PROOF OF COROLLARY 4.4. Let $\gamma = 1/2$, $\tau = 1/4$, and let δ_0^* be defined in (4.10). Set

$$Z \triangleq \frac{16\log(2ep/s)}{49\lambda^2} \sup_{\boldsymbol{\beta}\in\mathbb{R}^p:|\boldsymbol{\beta}|_0\le s} \left(\frac{\lambda}{2}|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_1 + \|\mathbb{X}\hat{\boldsymbol{\beta}}-\mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta}-\mathbf{f}\|_n^2\right).$$

For all $\delta_0 \in (0, \delta_0^*]$, by Theorem 4.2 we have $Z \le \log(1/\delta_0)$ with probability at least $1-\delta_0/2$. That is, $\mathbb{P}(Z > t) \le \frac{e^{-t}}{2}$ for all $t \ge T \triangleq \log(1/\delta_0^*) = \frac{s\log(2ep/s)}{\theta^2(s,7)}$. By integration,

$$\mathbb{E}[Z] \le \int_0^\infty \mathbb{P}(Z>t)\,dt \le T + \int_T^\infty \frac{e^{-t}}{2}\,dt \le T + \frac{1}{2},$$

which yields

$$\begin{aligned}\mathbb{E} \sup_{\boldsymbol{\beta}\in\mathbb{R}^p:|\boldsymbol{\beta}|_0\le s} &\left(\frac{\lambda}{2}|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_1 + \|\mathbb{X}\hat{\boldsymbol{\beta}}-\mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta}-\mathbf{f}\|_n^2\right)\\ &\le \frac{49}{16}\left(\frac{\lambda^2 s}{\theta^2(s,7)} + \frac{\lambda^2 s}{2s\log(2ep/s)}\right) \le \frac{49}{16}\left(\frac{\lambda^2 s}{\theta^2(s,7)} + \frac{\lambda^2 s}{2\log(2ep)}\right).\end{aligned}$$

This completes the proof of (4.13).

Let now $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \le s$. For all $t \ge T = \log(2/\delta_0^*)$, by Theorem 4.2 we have

$$Z_q \triangleq \frac{8s\log(2ep/s)}{49\lambda s^{1/q}}|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_q \le t$$

with probability at least $1-\frac{e^{-t}}{2}$. To prove (4.14), it remains to note that

$$\begin{aligned}\mathbb{E}[Z_q^q] &= \int_0^\infty q t^{q-1}\mathbb{P}(Z_q > t)\,dt\\ &\le \int_0^T q t^{q-1}\,dt + \int_T^\infty \frac{q t^{q-1}e^{-t}}{2}\,dt \le T^q + \frac{q}{2}\Gamma(q) \le T^q + 1. \qquad \square\end{aligned}$$

APPENDIX C: LASSO WITH ADAPTIVE CHOICE OF λ

PROOF OF THEOREM 5.1. In this proof, we set for brevity $\mathbb{P} = \mathbb{P}_{\boldsymbol{\beta}^*}$. Fix $s \in [1, s_*]$ and assume that $\boldsymbol{\beta}^* \in B_0(s)$. If $s < b_M$, let m_0 be the index such that b_{m_0} is the minimal element greater than s in the collection $\{b_m, m = 2, \dots, M\}$. Then $b_{m_0-1} \le s < b_{m_0}$. If $s \in [b_M, s_*]$, set $m_0 = M$. For any $a > 0$, we have

$$\mathbb{P}\big(d(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \ge a\big) \le \mathbb{P}\big(d(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \ge a, \hat{m} \le m_0\big) + \mathbb{P}(\hat{m} \ge m_0 + 1). \tag{C.1}$$

On the event $\{\hat{m} \le m_0\}$, we have

$$d(\hat{\boldsymbol{\beta}}_{b_{\hat{m}}}, \hat{\boldsymbol{\beta}}_{b_{m_0}}) \le \sum_{k=\hat{m}+1}^{m_0} d(\hat{\boldsymbol{\beta}}_{b_k}, \hat{\boldsymbol{\beta}}_{b_{k-1}}) \le 2C_0\sigma \sum_{k=\hat{m}+1}^{m_0} \sqrt{\frac{b_k \log(2ep/b_k)}{n}} \tag{C.2}$$

$$\le 2C_0\sigma \sum_{k=2}^{m_0} \sqrt{\frac{b_k \log(2ep/b_k)}{n}} \le c' C_0 \sigma \sqrt{\frac{b_{m_0} \log(2ep/b_{m_0})}{n}}, \tag{C.3}$$

where $c' > 2$ is an absolute constant. We deduce that, on the event $\{\hat{m} \le m_0\}$,

$$d(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le d(\hat{\boldsymbol{\beta}}_{b_{\hat{m}}}, \hat{\boldsymbol{\beta}}_{b_{m_0}}) + d(\hat{\boldsymbol{\beta}}_{m_0}, \boldsymbol{\beta}^*) \le c' w(b_{m_0}) + d(\hat{\boldsymbol{\beta}}_{m_0}, \boldsymbol{\beta}^*). \tag{C.4}$$

The following two cases are possible:

(i) *Case* $s < b_M$. Then, by definition of m_0 we have $b_{m_0}/2 = b_{m_0-1} \le s < b_{m_0}$. Since the function $w(\cdot)$ is increasing on $[1, p]$, we easily deduce that

$$w(b_{m_0})/\sqrt{2} \le w(s) \le w(b_{m_0}).$$

(ii) *Case* $s \in [b_M, s_*]$. Then $b_{m_0} = b_M \le s$, while $s \le 2b_M$. Therefore, in this case $b_{m_0} \le s < 2b_{m_0}$, which implies

$$w(b_{m_0}) \le w(s) \le \sqrt{2} w(b_{m_0}).$$

In both cases (i) and (ii), we have $w(s) \ge w(b_{m_0})/\sqrt{2}$. This remark and the fact that (C.4) holds on the event $\{\hat{m} \le m_0\}$ imply

$$\begin{aligned} &\mathbb{P}\big(d(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \ge \sqrt{2}(\sqrt{2} + c')w(s), \hat{m} \le m_0\big) \\ &\quad \le \mathbb{P}\big(d(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \ge (\sqrt{2} + c')w(b_{m_0}), \hat{m} \le m_0\big) \\ &\quad \le \mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{m_0}, \boldsymbol{\beta}^*) \ge \sqrt{2} w(b_{m_0})\big). \end{aligned} \tag{C.5}$$

Next, in both cases (i) and (ii), we have $s \le 2b_{m_0}$, which implies that $\boldsymbol{\beta}^* \in B_0(2b_{m_0})$. Using this fact together with (C.5) and (5.4), we obtain

$$\begin{aligned} &\sup_{\boldsymbol{\beta}^* \in B_0(s)} \mathbb{P}\big(d(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \ge \sqrt{2}(\sqrt{2} + c')w(s), \hat{m} \le m_0\big) \\ &\quad \le \sup_{\boldsymbol{\beta}^* \in B_0(2b_{m_0})} \mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{m_0}, \boldsymbol{\beta}^*) \ge \sqrt{2} w(b_{m_0})\big) \le \left(\frac{2b_{m_0}}{p}\right)^{2b_{m_0}} \le \left(\frac{2s}{p}\right)^{2s}, \end{aligned} \tag{C.6}$$

where we have used that the function $b \mapsto (b/p)^b$ is decreasing on the interval $[1, p/e]$ and, in both cases (i) and (ii), $2b_{m_0} \le 2s_* \le p/e$.

We now estimate the probability $\mathbb{P}(\hat{m} \ge m_0 + 1)$. We have

$$\mathbb{P}(\hat{m} \ge m_0 + 1) = \sum_{m=m_0+1}^{M} \mathbb{P}(\hat{m} = m).$$

Now, from the definition of $\hat{m}$ we obtain

$$\begin{aligned}\mathbb{P}(\hat{m}=m) &\le \mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{b_m},\hat{\boldsymbol{\beta}}_{b_{m-1}})>2w(b_m)\big)\\ &\le \mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{b_m},\boldsymbol{\beta}^*)>w(b_m)\big)+\mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{b_{m-1}},\boldsymbol{\beta}^*)>w(b_m)\big)\\ &\le \mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{b_m},\boldsymbol{\beta}^*)>w(b_m)\big)+\mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{b_{m-1}},\boldsymbol{\beta}^*)>w(b_{m-1})\big)\\ &\le 2\max_{k=m_0,\dots,M}\mathbb{P}\big(d(\hat{\boldsymbol{\beta}}_{b_k},\boldsymbol{\beta}^*)>w(b_k)\big)\triangleq 2a_{m_0},\end{aligned}$$

where we have used the triangle inequality, that $b_m > b_{m-1}$ and the monotonicity of $w(\cdot)$. Thus,

$$\mathbb{P}(\hat{m}\ge m_0+1)\le 2Ma_{m_0}. \tag{C.7}$$

The probability in (C.7) is nonzero only if $m_0 < M$. This implies that $s < b_{m_0}$, and hence $\boldsymbol{\beta}^* \in B_0(b_{m_0}) \subset B_0(b_k)$ for all $k \ge m_0$. Therefore, using (5.2) we obtain

$$\sup_{\boldsymbol{\beta}^*\in B_0(s)}\mathbb{P}(\hat{m}\ge m_0+1)\le 2M\max_{k\ge m_0}\left(\frac{b_k}{p}\right)^{b_k}. \tag{C.8}$$

Recall that $M \le \log_2(p)$. Note also that the function $b \mapsto (\frac{b}{p})^b$ is decreasing on the interval $[1, p/e]$, while $b_j \le s_*$ for all j and $s_* \le p/e$ by assumption. Finally, in both cases (i) and (ii), $b_{m_0} \le 2s$. Using these remarks, we get

$$\sup_{\boldsymbol{\beta}^*\in B_0(s)}\mathbb{P}(\hat{m}\ge m_0+1)\le 2M\left(\frac{b_{m_0}}{p}\right)^{b_{m_0}}\le 2\log_2(p)\left(\frac{2s}{p}\right)^{2s}. \tag{C.9}$$

Combining this bound with (C.7) and (C.1) where we set $a=\sqrt{2}(\sqrt{2}+c')w(s)$ proves (5.7). Finally, inequality (5.8) follows from (C.9) and the relations $\{\hat{s}\le s\}=\{b_{\hat{m}-1}\le s\}=\{b_{\hat{m}}/2\le s\}\supseteq\{b_{\hat{m}}\le b_{m_0}\}=\{\hat{m}\le m_0\}$. □

PROOF OF THEOREM 5.2. The constant C_0 in Theorem 5.2 is chosen to satisfy $C_0=\sqrt{2}C_0'$. Thus, inequalities (5.9) and (5.10) imply the inequalities of the same form as (5.2) and (5.4). Note that $w(b)=C_0\sigma b^{1/q}\sqrt{\frac{\log(2ep/b)}{n}}$ is increasing in b for $b\in[1,p]$. Note also that (C.2) remains valid if we replace there the expressions of the form $\sqrt{b\log(2ep/b)}$ by $b^{1/q}\sqrt{\log(2ep/b)}$. Using these remarks, we obtain the result of Theorem 5.2 by repeating, with minor modifications, the proof of Theorem 5.1 if we set $d(\boldsymbol{\beta},\boldsymbol{\beta}')=|\boldsymbol{\beta}-\boldsymbol{\beta}'|_q, \forall \boldsymbol{\beta},\boldsymbol{\beta}'\in\mathbb{R}^p$, and replace the references (5.2) and (5.4) with (5.9) and (5.10), respectively. We omit further details. □

APPENDIX D: PROOFS FOR THE SLOPE ESTIMATOR

PROOF OF THEOREM 6.1. By (A.4) with $h(\cdot) = 2|\cdot|_*$, we have that almost surely, for all $\boldsymbol{\beta} \in \mathbb{R}^p$ and all $\mathbf{f} \in \mathbb{R}^n$,

$$2\tau|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_* + \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2 + \triangle^*, \tag{D.1}$$

where $\triangle^* \triangleq 2\tau|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_* + \frac{2}{n}\boldsymbol{\xi}^T\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2|\boldsymbol{\beta}|_* - 2|\hat{\boldsymbol{\beta}}|_*$. Let $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Now consider the event (4.1) and the quantities $H(\boldsymbol{u})$ and $G(\boldsymbol{u})$ defined in (2.8). On the event (4.1), using (B.3) and Lemma A.1 we obtain

$$\begin{aligned}\triangle^* &\le 2\big(\tau|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_* + |\boldsymbol{\beta}|_* - |\hat{\boldsymbol{\beta}}|_*\big) + 2\max\big(H(\boldsymbol{u}), G(\boldsymbol{u})\big)\\ &\le 2\left((1+\tau)|\boldsymbol{u}|_2\Lambda(s) - (1-\tau)\sum_{j=s+1}^{p}\lambda_j u_j^\sharp\right) + 2\max\big(\tilde{H}(\boldsymbol{u}), G(\boldsymbol{u})\big),\end{aligned}$$

where $\tilde{H}(\cdot)$ is defined in (B.2) and $\Lambda(s) = (\sum_{j=1}^s \lambda_j^2)^{1/2}$. We now consider the following two cases:

(i) *Case* $\tilde{H}(\boldsymbol{u}) \le G(\boldsymbol{u})$. In this case, the definition of $\tilde{H}$ [cf. (B.2)], implies

$$|\boldsymbol{u}|_2 \le \frac{G(\boldsymbol{u})}{(4+\sqrt{2})(\sum_{j=1}^s \log(2p/j))^{1/2}} \le \|\mathbb{X}\boldsymbol{u}\|_n\sqrt{\frac{\log(1/\delta_0)}{s\log(2p/s)}}, \tag{D.2}$$

where we used (2.7) for the second inequality. As the weights (2.5) and the constant A satisfies (6.1), by (2.7) we also have $G(\boldsymbol{u}) \le \gamma\Lambda(s) \times \sqrt{\log(1/\delta_0)/(s\log(2p/s))}\|\mathbb{X}\boldsymbol{u}\|_n$. Using (D.2), we obtain

$$\begin{aligned}\triangle^* &\le 2(1+\tau)\Lambda(s)|\boldsymbol{u}|_2 + 2G(\boldsymbol{u})\\ &\le 2(1+\gamma+\tau)\Lambda(s)\sqrt{\log(1/\delta_0)/\big(s\log(2p/s)\big)}\|\mathbb{X}\boldsymbol{u}\|_n\\ &\le (1+\gamma+\tau)^2\Lambda^2(s)\left(\frac{\log(1/\delta_0)}{s\log(2p/s)}\right) + \|\mathbb{X}\boldsymbol{u}\|_n^2.\end{aligned} \tag{D.3}$$

(ii) *Case* $\tilde{H}(\boldsymbol{u}) > G(\boldsymbol{u})$. In this case,

$$\triangle^* \le 2(1+\gamma+\tau)|\boldsymbol{u}|_2\Lambda(s) - 2(1-\gamma-\tau)\sum_{j=s+1}^{p}\lambda_j u_j^\sharp \triangleq \triangle.$$

If $\triangle \le 0$, then (6.2) holds trivially in view of (D.1). If $\triangle > 0$, then $\boldsymbol{u}$ belongs to the cone $\mathcal{C}_{\mathrm{WRE}}(s, c_0)$, and we can use the $\mathrm{WRE}(s, c_0)$ condition, which yields

$$\begin{aligned}\triangle^* \le \triangle &\le 2(1+\gamma+\tau)\Lambda(s)|\boldsymbol{u}|_2\\ &\le \frac{2(1+\gamma+\tau)\Lambda(s)\|\mathbb{X}\boldsymbol{u}\|_n}{\vartheta(s,c_0)}\\ &\le \frac{(1+\gamma+\tau)^2\Lambda^2(s)}{\vartheta^2(s,c_0)} + \|\mathbb{X}\boldsymbol{u}\|_n^2.\end{aligned} \tag{D.4}$$

Combining the last inequality with (D.3) and (D.1) completes the proof of (6.2).

Let now $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with $|\boldsymbol{\beta}^*|_0 \le s$. Then (D.1) implies that $\triangle^* \ge 0$ almost surely. The estimation bound (6.3) is a direct consequence of (6.2). To prove (6.4), we set in what follows $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $\tau = 0$ so that $c_0 = \frac{1+\gamma}{1-\gamma}$ and consider the same two cases as above.

- If $\tilde{H}(\boldsymbol{u}) \le G(\boldsymbol{u})$, then (D.2) holds. Combining this bound with (D.1)–(D.3) we conclude that (6.4) is satisfied in this case.
- If $\tilde{H}(\boldsymbol{u}) > G(\boldsymbol{u})$, then $\triangle \ge \triangle^*$. It follows from (D.1) with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ that $\triangle^* \ge 2\|\mathbb{X}\boldsymbol{u}\|_n^2$. Thus, $\triangle \ge 0$. Therefore, $\boldsymbol{u}$ belongs to the cone $\mathcal{C}_{\mathrm{WRE}}(s, \frac{1+\gamma}{1-\gamma})$, and we can apply the WRE$(s, \frac{1+\gamma}{1-\gamma})$ condition, which yields $|\boldsymbol{u}|_2 \le \|\mathbb{X}\boldsymbol{u}\|_n / \vartheta(s, \frac{1+\gamma}{1-\gamma})$. Combining this bound with the inequality $2\|\mathbb{X}\boldsymbol{u}\|_n^2 \le \triangle$ and (D.4) we obtain that (6.4) is satisfied. □

APPENDIX E: BOUND ON THE STOCHASTIC ERROR

Here, we prove Theorem 4.1. The proof is based on a sequence of propositions.

PROPOSITION E.1. *Let $g_1, \dots, g_p$ be zero-mean Gaussian random variables with variance at most σ^2. Denote by $(g_1^\sharp, \dots, g_p^\sharp)$ be a nonincreasing rearrangement of $(|g_1|, \dots, |g_p|)$. Then*

$$\mathbb{P}\left(\frac{1}{s\sigma^2}\sum_{j=1}^{s}(g_j^\sharp)^2 > t\log\left(\frac{2p}{s}\right)\right) \le \left(\frac{2p}{s}\right)^{1-\frac{3t}{8}}$$

for all $t > 0$ and $s \in \{1, \dots, p\}$.

PROOF. By Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\exp\left(\frac{3}{8s\sigma^2}\sum_{j=1}^{s}(g_j^\sharp)^2\right) &\le \frac{1}{s}\sum_{j=1}^{s}\mathbb{E}\exp\left(\frac{3(g_j^\sharp)^2}{8\sigma^2}\right)\\ &\le \frac{1}{s}\sum_{j=1}^{p}\mathbb{E}\exp\left(\frac{3g_j^2}{8\sigma^2}\right) \le \frac{2p}{s},\end{aligned} \tag{E.1}$$

where we have used the fact that $\mathbb{E}[\exp(3\eta^2/8)] = 2$ when $\eta \sim \mathcal{N}(0,1)$. The Chernoff bound completes the proof. □

PROPOSITION E.2. *Under the assumptions of Proposition E.1,*

$$\mathbb{P}\left(\max_{j=1,\dots,p}\frac{g_j^\sharp}{\sigma\sqrt{\log(2p/j)}} \le 4\right) \ge \frac{1}{2}. \tag{E.2}$$

PROOF. Proposition E.1 with $t = 16/3$, and the inequality $(g_j^\sharp)^2 \le \frac{1}{j}\sum_{k=1}^j (g_k^\sharp)^2$ imply

$$\text{(E.3)} \qquad \mathbb{P}\Big((g_j^\sharp)^2 \le \frac{16\sigma^2}{3}\log(2p/j)\Big) \ge 1 - \frac{j}{2p}, \qquad j = 1, \dots, p.$$

Let $q \ge 0$ be the integer such that $2^q \le p < 2^{q+1}$. Applying (E.3) to $j = 2^l$ for $l = 0, \dots, q-1$, and using the union bound, we obtain that the event

$$\Omega_0 \triangleq \Big\{ \max_{l=0,\dots,q-1} \frac{g_{2^l}^\sharp \sqrt{3}}{4\sigma\sqrt{\log(2p/2^l)}} \le 1 \Big\}$$

satisfies $\mathbb{P}(\Omega_0) \ge 1 - \sum_{l=0}^{q-1} \frac{2^l}{2p} = 1 - \frac{2^q - 1}{2p} \ge 1/2$. For any $j < 2^q$, there exists $l \in \{0, \dots, q-1\}$ such that $2^l \le j < 2^{l+1}$, and thus, on the event Ω_0,

$$g_j^\sharp \le g_{2^l}^\sharp \le \frac{4\sigma}{\sqrt{3}}\sqrt{\log(2p/2^l)} \le \frac{4\sigma}{\sqrt{3}}\sqrt{\log(4p/j)} \le 4\sigma\sqrt{\log(2p/j)} \qquad \forall j < 2^q.$$

Next, for $2^q \le j \le p$ we have

$$g_j^\sharp \le g_{2^{q-1}}^\sharp \le \frac{4\sigma}{\sqrt{3}}\sqrt{\log(2p/2^{q-1})} < \frac{4\sigma}{\sqrt{3}}\sqrt{\log(8p/j)} \le 4\sigma\sqrt{\log(2p/j)}.$$

Thus, on the event Ω_0 we have $g_j^\sharp \le 4\sigma\sqrt{\log(2p/j)}$ for all $j = 1, \dots, p$. □

A function $N : \mathbb{R}^p \to [0, \infty)$ will be called positive homogeneous if $N(a\boldsymbol{u}) = aN(\boldsymbol{u})$ for all $a \ge 0$, $\boldsymbol{u} \in \mathbb{R}^p$ and $N(\boldsymbol{u}) > 0$ for $\boldsymbol{u} \ne \boldsymbol{0}$.

PROPOSITION E.3. *Let $\delta_0 \in (0, 1)$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_{n\times n})$. Let $N : \mathbb{R}^p \to [0, +\infty)$ be a positive homogeneous function. Assume that the event*

$$\Omega_4 \triangleq \Big\{ \sup_{\boldsymbol{v}\in\mathbb{R}^p : N(\boldsymbol{v})\le 1} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{v} \le 4 \Big\}$$

satisfies $\mathbb{P}(\Omega_4) \ge 1/2$. Then for all $\delta_0 \in (0, 1)$ we have

$$\mathbb{P}\Big(\forall \boldsymbol{u} \in \mathbb{R}^p : \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u} \le (4+\sqrt{2}) \max\Big(N(\boldsymbol{u}), \|\mathbb{X}\boldsymbol{u}\|_n \sigma \sqrt{\frac{\log(1/\delta_0)}{n}}\Big)\Big) \ge 1 - \delta_0/2.$$

PROOF. By homogeneity, it is enough to consider only $\boldsymbol{u} \in \mathbb{R}^p$ such that $\max(N(\boldsymbol{u}), \|\mathbb{X}\boldsymbol{u}\|_n/L) = 1$ where $L \triangleq (n/(\sigma^2 \log(1/\delta_0)))^{1/2}$. Define $T \subset \mathbb{R}^p$ and $f : \mathbb{R}^n \to \mathbb{R}$ by

$$\text{(E.4)} \quad T \triangleq \Big\{ \boldsymbol{u} \in \mathbb{R}^p : \max\Big(N(\boldsymbol{u}), \frac{1}{L}\|\mathbb{X}\boldsymbol{u}\|_n\Big) \le 1 \Big\}, \qquad f(\boldsymbol{v}) \triangleq \sup_{\boldsymbol{u}\in T} \frac{1}{n}(\sigma\boldsymbol{v})^T \mathbb{X}\boldsymbol{u}$$

for all $\boldsymbol{v} \in \mathbb{R}^n$. Then f is a Lipschitz function with Lipschitz constant $\sigma L/\sqrt{n}$. Thus, by [18], Inequality (1.4), we have with probability at least $1 - \delta_0/2$,

$$\begin{aligned}\sup_{\boldsymbol{u}\in T} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u} &\le \operatorname{Med}\Big[\sup_{\boldsymbol{u}\in T} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u}\Big] + \sigma L\sqrt{\frac{2\log(1/\delta_0)}{n}} \\ &\le \operatorname{Med}\Big[\sup_{\boldsymbol{u}\in\mathbb{R}^p : N(\boldsymbol{v})\le 1} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u}\Big] + \sigma L\sqrt{\frac{2\log(1/\delta_0)}{n}} \\ &\le 4 + \sigma L\sqrt{\frac{2\log(1/\delta_0)}{n}} = 4 + \sqrt{2},\end{aligned}$$

where we used the fact that $\mathbb{P}(\Omega_4) \ge 1/2$ to bound from above the median. □

PROOF OF THEOREM 4.1. Set

$$N(\boldsymbol{u}) = \sum_{j=1}^{p} u_j^\sharp \sigma\sqrt{\frac{\log(2p/j)}{n}}. \tag{E.5}$$

Then, using that $g_j = \frac{1}{\sqrt{n}}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{e}_j$ for all $j = 1, \dots, p$ we have

$$\begin{aligned}\sup_{\boldsymbol{u}\in\mathbb{R}^p : N(\boldsymbol{u})\le 1} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u} &\le \sup_{\boldsymbol{u}\in\mathbb{R}^p : N(\boldsymbol{u})\le 1} \sum_{j=1}^{p} u_j^\sharp \sigma\sqrt{\frac{\log(2p/j)}{n}} \frac{g_j^\sharp}{\sigma\sqrt{\log(2p/j)}} \\ &\le \max_{j=1,\dots,p} \frac{g_j^\sharp}{\sigma\sqrt{\log(2p/j)}}.\end{aligned} \tag{E.6}$$

By Proposition E.2, we have $\mathbb{P}(\Omega_4) \ge 1/2$, where Ω_4 is the event introduced in Proposition E.3. Therefore, Theorem 4.1 follows immediately from Proposition E.3. □

APPENDIX F: TOOLS FOR LOWER BOUNDS

LEMMA F.1. *For any integers $p \ge 2$, $n \ge 1$, $s \in [1, p/2]$, and any matrix $\mathbb{X} \in \mathbb{R}^{n\times p}$, there exists a subset Ω of the set $\{1, 0, -1\}^p$ with the following properties*:

$$|\boldsymbol{\omega}|_0 = s \quad \text{and} \quad \|\mathbb{X}\boldsymbol{\omega}\|_n^2 \le \bar{\theta}_{\max}^2(\mathbb{X}, 1)s \qquad \forall \boldsymbol{\omega} \in \Omega,$$

$$\log(|\Omega|) \ge \tilde{c}s\log\Big(\frac{ep}{s}\Big),$$

where $\tilde{c} > 0$ is an absolute constant, and for any two distinct elements $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ of Ω,

$$|\boldsymbol{\omega} - \boldsymbol{\omega}'|_q \ge (s/4)^{1/q} \qquad \forall 1 \le q \le \infty.$$

The proof of this lemma is omitted since it closely follows the argument in [27], pp. 79–80.

APPENDIX G: RANDOM DESIGN MATRICES

PROOF OF THEOREM 8.3. In this proof, we denote by C_i absolute positive constants. Using (3.2), the second inequality in (8.4), and (2.7) we get the inclusion $\mathcal{C}_{\mathrm{WRE}}(s, c_0) \cap \{\boldsymbol{v} \in \mathbb{R}^p : |\Sigma^{1/2}\boldsymbol{v}|_2 = 1\} \subset T$ where

$$T \triangleq \left\{ \boldsymbol{v} \in \mathbb{R}^p : \sum_{j=1}^{p} v_j^\sharp \sqrt{\log(2p/j)} \le r, |\Sigma^{1/2}\boldsymbol{v}|_2 = 1 \right\} \quad \text{and} \tag{G.1}$$
$$r = \frac{1+c_0}{\kappa}\sqrt{s\log(2ep/s)}.$$

It follows from [11, 21, 28] (cf., for instance, Theorem 1.12 in [21]) that for all $u > 0$, with probability at least $1 - 2\exp(-C_2 \min(u^2, u\sqrt{n}))$,

$$\sup_{\boldsymbol{v}\in T} \left| \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{v}^T \boldsymbol{x}_i)^2 - \mathbb{E}[(\boldsymbol{v}^T \boldsymbol{x})^2] \right| \le C_1 \left(\frac{L\gamma}{\sqrt{n}} + \frac{\gamma^2}{n} + \frac{uL^2}{\sqrt{n}} \right), \tag{G.2}$$

where $\gamma = \mathbb{E}[\sup_{\boldsymbol{v}\in T} G_{\boldsymbol{v}}]$ and $(G_{\boldsymbol{v}})_{\boldsymbol{v}\in T}$ is a centered Gaussian process indexed by T with covariance structure given by $\mathbb{E}[G_{\boldsymbol{v}} G_{\boldsymbol{u}}] = \boldsymbol{v}^T \Sigma \boldsymbol{u}$ for all $\boldsymbol{u}, \boldsymbol{v} \in T$. For instance, one can take $G_{\boldsymbol{v}} = \boldsymbol{v}^\top \Sigma^{1/2} \boldsymbol{z}$ for $\boldsymbol{v} \in T$, where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I_{p\times p})$. Then $\gamma = \mathbb{E}\sup_{\boldsymbol{v}\in T} \boldsymbol{v}^\top \Sigma^{1/2} \boldsymbol{z}$.

By (G.2), if we take $u = \sqrt{n}/(4C_1L^2)$ and if the number of observations n satisfies $n \ge 64(C_1 \vee C_1^2)L^2\gamma^2$, then with probability at least $1 - 2\exp(-C_3 n/L^4)$,

$$\frac{1}{2} \le \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{v}^T \boldsymbol{x}_i)^2 \le \frac{3}{2} \qquad \forall \boldsymbol{v} \in T.$$

Next, we evaluate the Gaussian mean width $\gamma = \mathbb{E}\sup_{\boldsymbol{v}\in T} \boldsymbol{v}^\top \Sigma^{1/2} \boldsymbol{z}$.

LEMMA G.1. *Let T be as in (G.1), with arbitrary $r > 0$. Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I_{p\times p})$ and let $\Sigma \in \mathbb{R}^{p\times p}$ be a positive semidefinite matrix with $\max_{j=1,\dots,p} \Sigma_{jj} \le 1$. Then*

$$\mathbb{E}\sup_{\boldsymbol{v}\in T} \boldsymbol{v}^\top \Sigma^{1/2} \boldsymbol{z} \le 4r + \sqrt{\pi/2}.$$

PROOF. In this proof, we set $g_j = \boldsymbol{z}^T \Sigma^{1/2} \boldsymbol{e}_j$, $j = 1, \dots, p$. As $\Sigma_{jj} \le 1$, the variance of g_j is at most 1. By Proposition E.2, the event

$$\Omega_0 \triangleq \left\{ \max_{j=1,\dots,p} \frac{g_j^\sharp}{\sqrt{\log(2p/j)}} \le 4 \right\}$$

has probability at least $1/2$. On the event Ω_0, for all $\boldsymbol{v} \in T$ we have

$$\boldsymbol{v}^T \Sigma^{1/2} \boldsymbol{z} = \sum_{j=1}^{p} g_j v_j \le \sum_{j=1}^{p} g_j^\sharp v_j^\sharp \le 4 \sum_{j=1}^{p} v_j^\sharp \sqrt{\log(2p/j)} \le 4r.$$

Set $f(z) \triangleq \sup_{v \in T}(v^T \Sigma^{1/2} z)$. We have $f(z) \le 4r$ with probability at least $1/2$, and thus $\mathrm{Med}[f(z)] \le 4r$. Furthermore, since the constraint $|\Sigma^{1/2} v|_2 = 1$ is satisfied for any $v \in T$, the function $f(\cdot)$ is 1-Lipschitz. Therefore, by Lemma A.3, $|\mathrm{Med}[f(z)] - \mathbb{E}[f(z)]| \le \sqrt{\pi/2}$. $\square$

It follows from Lemma G.1 and (G.2) that if $n \ge C_4 L^2 r^2$ then with probability at least $1 - 2\exp(-C_5 n/L^4)$, we have $(1/2) \le \|\mathbb{X} v\|_n^2 \le 3/2$ for any $v \in T$. By rescaling, for any $v \in \mathcal{C}_{\mathrm{WRE}}(s, c_0)$ we obtain $\|\mathbb{X} v\|_n^2 \ge (1/2)|\Sigma^{1/2} v|_2^2 \ge (\kappa^2/2)|v|_2^2$. This proves that the second inequality in (8.6) is satisfied if $n \ge C_4 L^2 r^2$.

We now prove the first inequality in (8.6). Let $j \in \{1, \dots, p\}$ and note that $\mathbb{X} e_j$ is a vector in $\mathbb{R}^n$ with i.i.d. sub-Gaussian coordinates since for all $i = 1, \dots, n$ and all $t \ge 0$, $\mathbb{E}\exp(t x_i^T e_j) \le \exp(t^2 L^2 \Sigma_{jj}/2)$. Hence, $\|\mathbb{X} e_j\|_n^2 - \Sigma_{jj}$ is a sum of independent zero-mean subexponential variables. It follows from Bernstein's inequality that for $u = 1/L^2$, with probability at least $1 - \exp(-C_6 n u^2)$, $\|\mathbb{X} e_j\|_n^2 \le \Sigma_{jj} + u L^2 \Sigma_{jj} \le 1$ when $\Sigma_{jj} \le 1/2$. By the union bound, the condition $\max_{j=1,\dots,p} \|\mathbb{X} e_j\|_n \le 1$ holds with probability at least $1 - p e^{-C_6 n/L^4} \ge 1 - e^{-C_6 n/(2L^4)}$ if $n \ge (2L^4/C_6)\log p$.

In conclusion, both inequalities in (8.6) are satisfied if

$$n \ge \frac{C_9(1+c_0)^2 L^2}{\kappa^2} s \log\left(\frac{2ep}{s}\right) + \frac{2L^4 \log p}{C_6}.$$

Since $\kappa^2 \le 1$, $L \ge 1$, and $s\log(2ep/s) \ge \log p$ the inequality in the last display is satisfied if (8.5) holds for some large enough absolute constant $C > 0$. $\square$

APPENDIX H: SUB-GAUSSIAN NOISE

To prove Proposition 9.2, we need the following lemma.

LEMMA H.1. *Let $\sigma > 0$, $z \sim \mathcal{N}(0,1)$, and let ξ_i be a random variable satisfying (9.1). Then, for all $t \ge 0$, $\mathbb{P}(|\xi_i| > t) \le 4\mathbb{P}(\sigma|z| > t)$.*

PROOF. By homogeneity, it is enough to consider $\sigma = 1$. From a standard lower bound on the Gaussian tail probability (cf. [2], Formula 7.1.13), we get

$$\mathbb{P}(|z| > t) \ge \frac{4\exp(-t^2/2)}{\sqrt{2\pi}(t + \sqrt{4+t^2})} \ge \frac{e^{1-t^2}}{4} \qquad \forall t \ge 0,$$

while if ξ_i satisfies (9.1) with $\sigma = 1$, a Chernoff bound yields that $\mathbb{P}(|\xi_i| > t) \le \exp(1 - t^2)$. $\square$

PROOF OF PROPOSITION 9.2. Let $\eta > 0$. Let $(\varepsilon_1, \dots, \varepsilon_n)$ be a vector of i.i.d. Rademacher variables independent of $\boldsymbol{\xi}$. The symmetrization inequality (cf., e.g., [13], Theorem 2.1) yields

$$\mathbb{E}\exp\left(\eta \sup_{\boldsymbol{u}\in U} \sum_{i=1}^{n} \xi_i u_i\right) \le \mathbb{E}\exp\left(\eta \sup_{\boldsymbol{u}\in U} \sum_{i=1}^{n} 2\varepsilon_i \xi_i u_i\right).$$

By Lemma H.1, we have $\mathbb{P}(|\varepsilon_i \xi_i| > t) \le K\mathbb{P}(\sigma|z_i| > t)$ for $K = 4$ and for $i = 1, \dots, n$. It follows from the contraction inequality as stated in [18], Lemma 4.6, that

$$\mathbb{E}\exp\left(\eta \sup_{\boldsymbol{u}\in U} \sum_{i=1}^{n} 2\varepsilon_i \xi_i u_i\right) \le \mathbb{E}\exp\Big(2\eta K\sigma \sup_{\boldsymbol{u}\in U} \boldsymbol{z}^T \boldsymbol{u}\Big).$$

Since U is a subset of the unit sphere, the function $f : \boldsymbol{z} \to \sup_{\boldsymbol{u}\in U} \boldsymbol{z}^T \boldsymbol{u}$ is 1-Lipschitz. Thus, by [6], Theorem 5.5, the right-hand side of the previous display is bounded from above by

$$\exp\Big(2\eta K\sigma \mathbb{E}\Big[\sup_{\boldsymbol{u}\in U} \boldsymbol{z}^T \boldsymbol{u}\Big] + 2\eta^2 K^2 \sigma^2\Big).$$

Furthermore, by Lemma A.3, $|\operatorname{Med}[f(\boldsymbol{z})] - \mathbb{E}f(\boldsymbol{z})| \le \sqrt{\pi/2}$. A Chernoff argument completes the proof. □

PROOF OF THEOREM 9.1. Let $N(\cdot)$ be defined in (E.5) and let $\boldsymbol{z}$ be a standard normal $\mathcal{N}(\mathbf{0}, I_{n\times n})$ random vector. It follows from (E.6) and Proposition E.2 that

$$\operatorname{Med}\Big[\sup_{\boldsymbol{u}\in\mathbb{R}^p : N(\boldsymbol{u})\le 1} \frac{1}{n}(\sigma \boldsymbol{z})^T \mathbb{X}\boldsymbol{u}\Big] \le 4. \tag{H.1}$$

Let $T \subset \mathbb{R}^p$ be defined in (E.4) with $L = \sqrt{n}/(\sigma(\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)}))$. Using Proposition 9.2 and then (H.1), we obtain that, with probability at least $1 - \delta_0$,

$$\begin{aligned}\sup_{\boldsymbol{u}\in T} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u} &\le 8\sigma \operatorname{Med}\Big[\sup_{\boldsymbol{u}\in T} \frac{1}{n} z^T \mathbb{X}\boldsymbol{u}\Big] + \frac{8\sigma L}{\sqrt{n}}\big(\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)}\big) \\ &\le 32 + \frac{8\sigma L}{\sqrt{n}}\big(\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)}\big) = 40.\end{aligned}$$

□

APPENDIX I: LASSO WITH UNIVERSAL TUNING PARAMETER

PROOF OF PROPOSITION 3.2. Let $\boldsymbol{\beta} \in \mathbb{R}^p$ be a minimizer of the right-hand side of (3.6) and let T be the support of $\boldsymbol{\beta}$, so that $|T| \le s$. Using inequality (A.4) with $h(\cdot) = 2\lambda|\cdot|_1$, we get that, almost surely,

$$\begin{aligned}&\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \\ &\quad\le (2/n)\boldsymbol{\xi}^T \mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2\lambda|\boldsymbol{\beta}|_1 - 2\lambda|\hat{\boldsymbol{\beta}}|_1 - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2.\end{aligned} \tag{I.1}$$

Let $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and define the function f as follows:

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{v}\in\mathbb{R}^p:\|\mathbb{X}\boldsymbol{v}\|_n=1} \big((1/\sqrt{n})\boldsymbol{x}^T\mathbb{X}\boldsymbol{v} + \sqrt{n}\lambda\big(|\boldsymbol{v}_T|_1 - |\boldsymbol{v}_{T^c}|_1\big)\big), \qquad \boldsymbol{x}\in\mathbb{R}^n.$$

By simple algebra, $2\lambda|\boldsymbol{\beta}|_1 - 2\lambda|\hat{\boldsymbol{\beta}}|_1 \le 2\lambda(|\boldsymbol{u}_T|_1 - |\boldsymbol{u}_{T^c}|_1) = \|\mathbb{X}\boldsymbol{u}\|_n 2\lambda(|\boldsymbol{w}_T|_1 - |\boldsymbol{w}_{T^c}|_1)$ where $\boldsymbol{w} = (1/\|\mathbb{X}\boldsymbol{u}\|_n)\boldsymbol{u}$. Hence, the right-hand side of (I.1) is bounded from above by

$$2\|\mathbb{X}\boldsymbol{u}\|_n f(\boldsymbol{\xi})/\sqrt{n} - \|\mathbb{X}\boldsymbol{u}\|_n^2 \le f^2(\boldsymbol{\xi})/n.$$

Since the function f is 1-Lipschitz, by the Gaussian concentration bound [18], inequality (1.4), we have, for all $\delta\in(0,1)$,

$$\mathbb{P}\big(f(\boldsymbol{\xi}) \le \operatorname{Med}\big[f(\boldsymbol{\xi})\big] + \sigma\sqrt{2\log(1/\delta)}\big) \ge 1-\delta.$$

To complete the proof, it remains to show that

$$\operatorname{Med}\big[f(\boldsymbol{\xi})\big] \le \sigma\bigg(\frac{1+\varepsilon}{\kappa(s,c_0)}\sqrt{2s\log p} + \sqrt{s} + 2.8\bigg). \tag{I.2}$$

Let $\Pi_T\in\mathbb{R}^{n\times n}$ be the orthogonal projection onto the linear span of $\{\boldsymbol{x}_j, j\in T\}$, where $\boldsymbol{x}_j = \mathbb{X}e_j$. Then almost surely,

$$\begin{aligned} f(\boldsymbol{\xi}) &= \sup_{\boldsymbol{v}\in\mathbb{R}^p:\|\mathbb{X}\boldsymbol{v}\|_n=1} \big[(1/\sqrt{n})\boldsymbol{\xi}^T\Pi_T\mathbb{X}\boldsymbol{v} + (1/\sqrt{n})\boldsymbol{\xi}^T(I_{n\times n}-\Pi_T)\mathbb{X}\boldsymbol{v}_{T^c} \\ &\qquad + \sqrt{n}\lambda\big(|\boldsymbol{v}_T|_1 - |\boldsymbol{v}_{T^c}|_1\big)\big] \\ &\le |\Pi_T\boldsymbol{\xi}|_2 + \sup_{\boldsymbol{v}\in\mathbb{R}^p:\|\mathbb{X}\boldsymbol{v}\|_n=1} \big[(1/\sqrt{n})\boldsymbol{\xi}^T(I_{n\times n}-\Pi_T)\mathbb{X}\boldsymbol{v}_{T^c} \\ &\qquad + \sqrt{n}\lambda\big(|\boldsymbol{v}_T|_1 - |\boldsymbol{v}_{T^c}|_1\big)\big]. \end{aligned}$$

The random variable $|\Pi_T\boldsymbol{\xi}|_2^2/\sigma^2$ has a chi-square distribution with at most s degrees of freedom. Standard bounds on the tails of the chi-square distribution yield that the event $\Omega_1 = \{|\Pi_T\boldsymbol{\xi}|_2 \le \sigma(\sqrt{s} + \sqrt{2\log(50)})\}$ satisfies $\mathbb{P}(\Omega_1)\ge 1-1/50 = 0.98$. Let $Z = \max_{j=1,\dots,p}|\boldsymbol{\xi}^T(I_{n\times n}-\Pi_T)\boldsymbol{x}_j|$ and define $\Omega_2 = \{Z\le\sigma\sqrt{2\log p}\}$. The random variable Z is the maximum of $2p$ centered Gaussian random variables with variance at most σ^2, and thus $\mathbb{P}(Z>\sigma x)\le 2pe^{-x^2/2}/(x\sqrt{2\pi})$ for all $x>0$. The choice $x=\sqrt{2\log p}$ yields that $\mathbb{P}(\Omega_2)\ge 1-1/\sqrt{\pi\log p}\ge 0.5208$ for all $p\ge 4$. Direct calculation shows that the same bound on $\mathbb{P}(\Omega_2)$ is true for $p\in\{2,3\}$. Combining these remarks, we find that, for all $p\ge 2$,

$$\mathbb{P}(\Omega_1\cap\Omega_2) \ge 1 - \mathbb{P}\big(\Omega_1^c\big) - \mathbb{P}\big(\Omega_2^c\big) \ge 1 - 0.02 - 0.4792 > 1/2.$$

Thus, by definition of the median, an upper bound on $\operatorname{Med}[f(\boldsymbol{\xi})]$ is given by an upper bound on $f(\boldsymbol{\xi})$ on the event $\Omega_1\cap\Omega_2$:

$$\begin{aligned} \operatorname{Med}\big[f(\boldsymbol{\xi})\big] \le \sigma(\sqrt{s}+2.8) + \sup_{\boldsymbol{v}\in\mathbb{R}^p:\|\mathbb{X}\boldsymbol{v}\|_n=1} \big[\sigma\sqrt{2\log p}|\boldsymbol{v}_{T^c}|_1 \\ + \sqrt{n}\lambda\big(|\boldsymbol{v}_T|_1 - |\boldsymbol{v}_{T^c}|_1\big)\big], \end{aligned} \tag{I.3}$$

where we have used that $\sqrt{2\log(50)} \le 2.8$. Recall that $\sqrt{n}\lambda = (1+\varepsilon)\sigma\sqrt{2\log p}$. Thus, if $|\boldsymbol{v}_{T^c}|_1 > (1+1/\varepsilon)|\boldsymbol{v}_T|_1$, the supremum in (I.3) is negative and (I.2) follows. Finally, if $|\boldsymbol{v}_{T^c}|_1 \le (1+1/\varepsilon)|\boldsymbol{v}_T|_1$ then, by the definition of the RE constant $\kappa(s, c_0)$, with $c_0 = 1 + 1/\varepsilon$ we obtain $|\boldsymbol{v}_T|_1 \le \sqrt{s}|\boldsymbol{v}_T|_2 \le \sqrt{s}\|\mathbb{X}\boldsymbol{v}\|_n/\kappa(s, c_0)$. Using this remark to bound the supremum in (I.3) proves (I.2). □

REFERENCES

[1] ABRAMOVICH, F. and GRINSHTEIN, V. (2010). MAP model selection in Gaussian regression. *Electron. J. Stat.* **4** 932–949. MR2721039

[2] ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. *National Bureau of Standards Applied Mathematics Series* **55**. Washington, DC. MR0167642

[3] BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815.

[4] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140.

[6] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, London.

[7] CANDÈS, E. J. and DAVENPORT, M. A. (2013). How well can we estimate a sparse vector? *Appl. Comput. Harmon. Anal.* **34** 317–323. MR3008569

[8] CANDES, E. J. and TAO, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory* **52** 5406–5425.

[9] CHAFAÏ, D., GUÉDON, O., LECUÉ, G. and PAJOR, A. (2012). *Interactions Between Compressed Sensing Random Matrices and High Dimensional Geometry*. *Panoramas et Synthèses* **37**. Société Mathématique de France, Paris. MR3113826

[10] DALALYAN, A. S., HEBIRI, M., LEDERER, J. et al. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581.

[11] DIRKSEN, S. (2015). Tail bounds via generic chaining. *Electron. J. Probab.* **20** no. 53, 29. MR3354613

[12] GIRAUD, C. (2015). *Introduction to High-Dimensional Statistics*. *Monographs on Statistics and Applied Probability.* **139**. CRC Press, Boca Raton, FL.

[13] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Ecole D'Ete de Probabilites de Saint-Flour XXXVIII*-2008. Springer, New York.

[14] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

[15] KOLTCHINSKII, V. and MENDELSON, S. (2015). Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN* **23** 12991–13008. MR3431642

[16] LECUÉ, G. and MENDELSON, S. (2015). Regularization and the small-ball method I: Sparse recovery Technical report, CNRS, ENSAE and Technion, I.I.T.

[17] LECUÉ, G. and MENDELSON, S. (2017). Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.* (*JEMS*) **19** 881–904. MR3612870

[18] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) **23**. Springer, Berlin. MR1102015

[19] LOUNICI, K., PONTIL, M., TSYBAKOV, A. B. and VAN DE GEER, S. A. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204.

[20] MENDELSON, S. (2014). Learning without concentration. In *Proceedings of the* 27*th Annual Conference on Learning Theory COLT*14 25–39.

[21] MENDELSON, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.* **126** 3652–3680.

[22] MENDELSON, S. (2015). Learning without concentration. *J. ACM* **62** Art. 21, 25.

[23] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

[24] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337

[25] SU, W. and CANDÈS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953

[26] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[27] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90.

[28] WITOLD, B. (2013). Concentration via chaining method and its applications. Technical report, Univ. Warsaw. Available at arXiv:1405.0676.

[29] YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540.

P. C. BELLEC
DEPARTMENT OF STATISTICS AND BIOSTATISTICS
BUSCH CAMPUS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: pierre.bellec@rutgers.edu

G. LECUÉ
A. B. TSYBAKOV
ENSAE
3 AVENUE PIERRE LAROUSSE
92240 MALAKOFF
FRANCE
E-MAIL: guillaume.lecue@ensae.fr
alexandre.tsybakov@ensae.fr