Robust and Tuning-Free Sparse Linear Regression via Square-Root Slope

Stanislav Minsker*, Mohamed Ndaoud †, and Lang Wang*

Abstract. We consider the high-dimensional linear regression model and assume that a fraction of the measurements are altered by an adversary with complete knowledge of the data and the underlying distribution. We are interested in a scenario where dense additive noise is heavy-tailed while the measurement vectors follow a sub-Gaussian distribution. Within this framework, we establish minimax lower bounds for the performance of an arbitrary estimator that depend on the the fraction of corrupted observations as well as the tail behavior of the additive noise. Moreover, we design a modification of the so-called Square-Root Slope estimator with several desirable features: (a) it is provably robust to adversarial contamination, and satisfies performance guarantees in the form of sub-Gaussian deviation inequalities that match the lower error bounds, up to logarithmic factors; (b) it is fully adaptive with respect to the unknown sparsity level and the variance of the additive noise, and (c) it is computationally tractable as a solution of a convex optimization problem. To analyze performance of the proposed estimator, we prove several properties of matrices with sub-Gaussian rows that may be of independent interest.

Key words. robust inference, sub-Gaussian deviation, Slope, sparse linear regression

MSC codes. 62F35, 62J07

1. Introduction.

Robust statistics, broadly speaking, is an arsenal of estimation and inference techniques that are resistant to model perturbations. Data generated from a perturbed model will often contain atypical observations, commonly referred to as outliers. This paper is devoted to robust estimation in the context of high-dimensional sparse linear regression. Assume that a sequence of random pairs $(X_1,y_1),\dots,(X_n,y_n)$ is generated according to the model

$$y_i = X_i^T\beta^* + \sqrt{n}\theta_i^* + \sigma\xi_i, \quad i=1,\dots,n.$$

Here, each $y_i \in \mathbb{R}$ is a linear measurements of an unknown vector $\beta^* \in \mathbb{R}^p$ that has s non-zero coordinates. The measurement vectors $X_i \in \mathbb{R}^p$, $j=1,\dots,n$ are independently sampled from a distribution with unknown covariance matrix Σ. We assume that the measurements y_i are contaminated by the noise $\sigma\xi_i$ where $\sigma>0$ and $\xi_1,\dots,\xi_n$ are i.i.d. random variables with unit variance, independent from $X_1,\dots,X_n$. Finally, the adversarial noise is modeled by the additive term $\sqrt{n}\theta_i^*$[1], where the sequence $\theta_1^*,\dots,\theta_n^*$ has $o<n$ non-zero elements and is generated by an adversary who has access to $\{(y_i,X_i,\xi_i)\}_{i=1}^n,\beta^*$, σ, as well as the joint distribution of all random variables involved. We are interested in the situation when (a) when p is possibly much larger than n but s is smaller than n, and (b) the random variables $\{\xi_i\}_{i=1}^n$ are possibly heavy-tailed, distributed according to a law with polynomially decaying tails.

The well-known Lasso estimator [30], as well as its sibling, the Dantzig selector [5], provably achieve strong performance guarantees in the sparse prediction and estimation tasks. For example, the

*Department of Mathematics, University of Southern California (minsker@usc.edu,langwang@usc.edu).

†Department of Information Systems, Decision Sciences and Statistics, ESSEC Business School (ndaoud@essec.edu).

[1]See remark 2.1 related to the $\sqrt{n}$ factor.

Lasso estimator is the solution to the following optimization problem:

$$\widetilde{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\left[\frac{1}{n}\sum_{j=1}^{n}(y_j - X_j^T\beta)^2 + \lambda\|\beta\|_1\right]$$

where $\lambda > 0$ is the regularization parameter and $\|\cdot\|_p$ stands for the ℓ_p norm of a vector, $p \geq 1$. It is known that theoretically optimal value of the parameter λ is proportional to $\sigma\sqrt{\frac{\log(ep/s)}{n}}$ [1, 22], in particular, it depends on the unknown variance of the noise as well as the unknown sparsity level s. In addition, the Lasso estimator $\widetilde{\beta}$ is not robust to the presence of gross outliers, i.e. the norm $\|\beta^* - \widetilde{\beta}\|_2$ can be arbitrarily large if θ^* has just 1 non-zero element that can take arbitrary values. In this paper, we propose robust version of the pivotal Slope ("Sorted ℓ-One Penalized Estimation") algorithm [10] that is provably robust to the heavy-tailed additive noise and the adversarial corruption; moreover, it is tuning-free. The proposed estimator combines the ideas behind the original Slope algorithm [4] that eliminates dependence of the optimal choice of λ on the sparsity level s, the square-root Lasso [2] that allows to set λ independently of the noise variance σ^2, and moreover takes advantage of the robustness stemming from connections between the Huber's loss and the ℓ_1 - penalized squared loss [25, 14]. Specifically, we prove that the estimator $\widehat{\beta}$ produced by robust pivotal Slope and formally defined via (2) below admits the following performance guarantees under suitable assumptions on the covariance matrix Σ of the design vectors:

(a) If both the design vectors X_j and the noise variables ξ_j, $j = 1,\dots,n$ have sub-Gaussian distributions, then

$$\|\widehat{\beta} - \beta^*\|_\Sigma^2 \lesssim \sigma^2\left(\frac{s\log(ep/s)}{n} + \left(\frac{o\log(n/o)}{n}\right)^2 + \frac{\log(1/\delta)}{n}\right)$$

with probability at least $1-\delta$, where $\lesssim$ denotes the inequality up to an absolute constant and $\|x\|_\Sigma^2 := \langle \Sigma x, x\rangle$. In particular, the upper bound is minimax optimal with respect to sparsity level s and the number of outliers o.

(b) If X_j's are sub-Gaussian but ξ_j's are heavy tailed, meaning that $\mathbf{E}(|\xi|^\tau) < \infty$ for some $\tau \geq 4$, and in the absence of adversarial contamination (i.e. $\theta^* \equiv 0$),

$$\|\widehat{\beta} - \beta^*\|_\Sigma^2 \lesssim \sigma^2\left(\frac{s\log(ep/s)}{n} + \frac{\log(1/\delta)}{n}\right)$$

with probability at least $1-\delta$. In other words, $\|\widehat{\beta} - \beta^*\|_\Sigma$ admits sub-Gaussian deviation guarantees despite the fact that the noise is allowed to be heavy-tailed.

(c) Finally, if $\mathbf{E}(|\xi|^\tau) < \infty$ for some $\tau \geq 2$ and $s\log(p/s) + \log(1/\delta) + o \lesssim n$, a version of the proposed estimator satisfies

$$\|\widehat{\beta} - \beta^*\|_\Sigma^2 \lesssim \sigma^2\left(\frac{s\log(ep/s)}{n} + \left(\frac{o}{n}\right)^{2-2/\tau}\log\left(\frac{n}{o}\right)\left(1 + \left(\frac{o}{\log(1/\delta)}\right)^{2/\tau}\right) + \frac{\log(1/\delta)}{n}\right)$$

with probability at least $1-\delta$. It implies that whenever $\log(1/\delta) \lesssim o$, the upper bound depends optimally (up to a logarithmic factor) on the number of adversarial outliers, as well as on the sparsity level s. Note that $\|\widehat{\beta}-\beta^*\|_\Sigma$ admits sub-Gaussian deviation guarantees in this case as well.

1.1. Structure of the paper. The rest of the exposition is organized as follows: notation and key definitions are summarized in section 1.2. In section 2, we explain the main ideas leading to the definition of the pivotal Slope estimator and state the theoretical guarantees related to its performance, along with the information-theoretic lower bounds. This is followed by a discussion and comparison to existing results in section 3. Finally, the proofs of the main results are presented in the appendix.

1.2. Notation. Absolute constants that do not depend on any parameters of the problem are going to be denoted by C, C', C_1, etc as well as c, c', c_1, with the convention that capital C stands for "a sufficiently large absolute constant" while the lower case c is a synonym of "a sufficiently small absolute constant". It is assumed that C and c can denote different absolute constants in different parts of the expression. For $a, b \in \mathbb{R}$, let $a \vee b := \max\{a,b\}$ and $a \wedge b := \min\{a,b\}$.

Given a vector $v \in \mathbb{R}^p$, we denote its ℓ_1 and ℓ_2- norms via $\|v\|_1 := \sum_{i=1}^p |v_i|$ and $\|v\|_2 := \sqrt{\sum_{i=1}^p |v_i|^2}$ respectively. If $\Sigma \in \mathbb{R}^{p\times p}$ is a symmetric positive-definite matrix, we define $\|v\|_\Sigma := \langle \Sigma v, v\rangle^{1/2}$. Given two vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$, let $[u;v] \in \mathbb{R}^n \times \mathbb{R}^p$ be the $(p+n)$-dimensional vector created by the vertical concatenation of u and v. Let $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_p \geq 0$ be a non-increasing sequence. The corresponding sorted ℓ_1 norm is defined as

$$\|v\|_\gamma := \sum_{i=1}^p \gamma_i |v|_{(i)},$$

where $|v|_{(i)}$ is the i-th largest coordinate of the vector $(|v|_1, \ldots, |v|_p)$; the fact that this is indeed a norm is established in [4, Proposition 1.2].

Capital S and O will be reserved for the supports of vectors β^* and θ^*, the subsets of $\{1,\ldots,p\}$ and $\{1,\ldots,n\}$ respectively that contain the indices of non-zero coordinates of these vectors. We will also set $s = |S| := \mathrm{Card}(S)$ and $o = |O| := \mathrm{Card}(O)$.

2. Main results. Recall that we observe n random pairs of predictor-response values $(X_1, y_1), \ldots, (X_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$ that are assumed to be generated according to the model

$$y_i = X_i^T \beta^* + \sqrt{n}\theta_i^* + \sigma\xi_i, \qquad i = 1,\ldots,n.$$

Alternatively,

$$Y = X\beta^* + \sqrt{n}\theta^* + \sigma\xi,$$

where $X = [X_1, \ldots, X_n]^T$ is the $n \times p$ design matrix, $Y = (y_1, \ldots, y_n)^T$ is the response vector, $\xi = (\xi_1, \ldots, \xi_n)^T$ is the additive noise vector and $\theta^* = (\theta_1^*, \ldots, \theta_n^*)^T$ is the vector of adversarial outliers.

Remark 2.1. *The $\sqrt{n}$ factor in front of θ^* is introduced for technical convenience: with this scaling, the columns of the augmented design matrix $[X \,|\, \sqrt{n} I_n]$, where I_n is $n \times n$ identity matrix, are of similar length.*

Let us now define the robust pivotal Slope estimator. To this end, set

$$Q(\beta,\theta) := \frac{1}{2n}\sum_{j=1}^{n}(y_j - X_j^T\beta - \sqrt{n}\theta_j)^2 = \frac{1}{2n}\left\|Y - X\beta - \sqrt{n}\theta\right\|_2^2,$$

and let

$$L(\beta,\theta) = Q(\beta,\theta)^{\frac{1}{2}} + \|\beta\|_\lambda + \|\theta\|_\mu$$

for some positive non-increasing sequences $\{\lambda_j\}_{j=1}^p$, $\{\mu\}_{j=1}^p$. The pivotal Slope estimator $\widehat{\beta}$ of β^* is then defined via the solution of a convex minimization problem

$$(\widehat{\beta},\widehat{\theta}) = \operatorname*{argmin}_{\beta\in\mathbb{R}^p,\theta\in\mathbb{R}^n} L(\beta,\theta).$$

The estimator in (2) can be seen as a generalization of the square-root Slope estimator (see [28, 10]). The idea of introducing the square root of the quadratic term $\sqrt{Q(\beta,\theta)}$, as opposed to $Q(\beta,\theta)$ itself, was originally developed in [2] for the Lasso estimator with the goal of removing the dependence of the regularization parameters on unknown σ, while retaining the convexity of the loss function. Note that estimator (2) is equivalent to the argmin over $\beta\in\mathbb{R}^p$, $\theta\in\mathbb{R}^n$ and $\sigma>0$ of the loss function

$$\widetilde{L}(\beta,\theta,\sigma) = \frac{Q(\beta,\theta)}{\sigma} + \sigma + \|\beta\|_\lambda + \|\theta\|_\mu,$$

provided that the minimum is attained at a positive value of σ (see [31, Chapter 3]). Indeed, the term $Q(\beta,\theta)^{\frac{1}{2}}$ in (2) appears when one performs minimization of $\widetilde{L}(\beta,\theta,\sigma)$ with respect to $\sigma>0$ first, and the optimal value of σ can be viewed as an estimator of the unknown standard deviation of the noise. The couple $(\widehat{\beta},\widehat{\theta})$ can in turn be viewed as the usual square root Slope estimator for the vector $[\beta^*;\theta^*]$ of unknown regression coefficients corresponding the augmented design matrix $[X;I_n]$. This is a natural approach since both β^* and θ^* are sparse vectors.

In the following subsections, we will show that the estimator defined in (2) achieves the optimal error bound under suitable choices of the sequences $(\lambda)_p$ and $(\mu)_n$. To this end, we will need the following assumptions:

Assumption 1. *ξ_i's are i.i.d. random variables with distribution P_ξ in $\mathscr{P}_{\tau,a}$ for some $\tau\geq 2$ with $\mathbf{E}(\xi_i)=0$ and $Var(\xi_i)=1$, where*

$$\mathscr{P}_{\tau,a} = \{P_\xi \text{ such that } \mathbf{E}(|\xi|^\tau)\leq a^\tau\}.$$

When $\tau=\infty$, we instead impose the condition $\Pr(|\xi|\geq t)\leq 2\exp(-(t/a)^2)$. We will also assume that τ or its lower bound is known and that a is bounded by a sufficiently large numerical constant.

Assumption 2. *Assume that ξ satisfies a "small ball-type" condition, namely*

$$\mathbf{E}(\xi^2\mathbf{1}\{|\xi|\leq 1/2\})\geq 1/4.$$

The "small ball" property and related conditions essentially state that the distribution of ξ assigns sufficient probability to the neighborhood of 0. It turns out to be a convenient and rather mild assumption

that allows one to control lower tails of sums of independent random variables. The specific choice of the constants $1/2$ and $1/4$ is not of essence, as any other choice of absolute constants would yield similar results. It is well known [for example, see section 4 in 20] that ξ satisfies this type of bounds for some positive constants in place if $1/2$ and $1/4$ if for some $\tau > 2$ and $c(\tau)$,

$$(\mathbb{E}|\xi|^\tau)^{1/\tau} \le c(\tau)\left(\mathbb{E}\xi^2\right)^{1/2}.$$

More specifically, to deduce the "small ball-type" bound, it suffices to write that $\mathbb{E}\xi^2 = \mathbf{E}(\xi^2\mathbf{1}\{|\xi| \le C\}) + \mathbf{E}(\xi^2\mathbf{1}\{|\xi| > C\})$, followed by the application of Hölder's inequality and the relation (2).

Assumption 3. *For $i = 1,\dots,n$ $X_i = \Sigma^{1/2}Y_i$ for some (unknown) matrix Σ satisfying $\Sigma_{ii} \le 1$ and $Y_1,\dots,Y_n$ are i.i.d. centered 1-sub-Gaussian random vectors such that $\mathbf{E}(Y_1Y_1^\top) = I_n$. Here, "1-sub-Gaussian" means that $\mathbb{E}e^{\lambda\langle Y_1,v\rangle} \le e^{\lambda^2\frac{\|v\|_2^2}{2}}$ for any $v \in \mathbb{R}^p$.*

We will also need to introduce the following objects:

- Define the cone

$$\mathscr{C}(s,c_0) = \left\{u \in \mathbb{R}^p : \|u\|_\lambda \le c_0\sqrt{\sum_{i=1}^s \lambda_i^2}\,\|u\|_2\right\}.$$

  This is a set of vectors with s largest coordinates that "dominate" the remaining ones, in a sense made precise above. We will assume that the covariance matrix Σ satisfies the following version of the *restricted eigenvalue condition*: for any $u \in \mathscr{C}(s,4)$,

$$\|u\|_\Sigma^2 \ge \kappa(s)\|u\|_2^2.$$

  In particular, if Σ is non-degenerate, then it is always true that

$$\kappa(s) \ge \lambda_{\min}(\Sigma) > 0.$$

- We will also be interested in a similar cone in the augmented space $\mathbb{R}^p \times \mathbb{R}^n$ that is defined via

$$\mathscr{C}(s,c_0,o,\delta,\Sigma) = \Bigg\{(u,v) \in \mathbb{R}^p \times \mathbb{R}^n : \|u\|_\lambda + \|v\|_\mu \le c_0\left(\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)} + \frac{\log(1/\delta)}{n}}\,\|u\|_\Sigma + \sqrt{\sum_{i=1}^o \mu_i^2}\,\|v\|_2\right)\Bigg\}.$$

- Finally, define the sequence

$$\lambda_i = C\sqrt{\frac{\log(ep/i)}{n}},\ i = 1,\dots,p,$$

  where C is a sufficiently large absolute constant. We will use the ordered $\|\cdot\|_\lambda$ norm corresponding to this sequence.

2.1. Upper error bounds. Depending on the value of the parameter τ controlling the tails of the additive noise ξ, we will need to set the sequence $(\mu_n)_n$ differently (recall that τ, or its lower bound, is assumed to be known). Specifically, let

$$\mu_i = \frac{C}{\sqrt{n}}\left(\frac{n}{i}\right)^{1/\tau},\ i=1,\dots,n$$

and

$$\mu_i = C\sqrt{\frac{\log(en/i)}{n}},\ i=1,\dots,n$$

for $\tau=\infty$. Solution of the minimization problem (2) corresponding to this choice of penalization will be denoted via $\hat{\beta}_{\text{sorted}}$. Similarly, given $0<\delta<1$, we denote by $\hat{\beta}_{\text{fixed}}$ the solution of (2) corresponding to

$$\mu_i = \frac{C}{\sqrt{n}}\left(\frac{n}{\log(1/\delta)}\right)^{1/\tau},\ i=1,\dots,n.$$

Observe that both estimators $\hat{\beta}_{\text{sorted}}$ and $\hat{\beta}_{\text{fixed}}$ are fully adaptive, the only requirement being the prior knowledge of τ. In addition, notice that $\hat{\beta}_{\text{fixed}}$ requires the desired confidence level δ as an input while $\hat{\beta}_{\text{sorted}}$ does not.

Theorem 2.1. *Assume that $\tau\geq 2$ and that assumptions 1, 2 and 3 hold. There exist absolute positive constants c,C' with the following properties: let $0<\delta<1$ be fixed, and assume that $s\log(p/s)/\kappa(s)+\log(1/\delta)+o\leq cn$. Then with probability at least $1-\delta$,*

$$\|\widehat{\beta}_{\text{fixed}}-\beta^*\|_\Sigma^2 \leq C'\sigma^2\left(\frac{s\log(ep/s)}{\kappa(s)n}+\left(\frac{o}{n}\right)^{2-2/\tau}\log(n/o)\left(1+\left(\frac{o}{\log(1/\delta)}\right)^{2/\tau}\right)+\frac{\log(1/\delta)}{n}\right).$$

It follows from Theorem 2.3 stated below that the bound (2.1) is minimax optimal with respect to the contamination proportion $\frac{o}{n}$, up to the logarithmic factors, as long as $\log(1/\delta)\geq o$. Note that similar types of conditions have appeared in the context of robust regression for methods based on the median of means estimator [17]. In general, the condition $\log(1/\delta)\geq o$ is also required for robust mean estimation for instance using the trimmed mean [19] or self-normalized sums [21]. Finally, observe that (2.1) is meaningful, although sub-optimal, even when $o\gg\log(1/\delta)$.

Theorem 2.2. *Assume that $\tau>2$ and that assumptions 1,2 and 3 hold. There exist absolute positive constants c,C' with the following properties: for any δ such that $s\log(p/s)/\kappa(s)+\log(1/\delta)+o\leq cn$, the inequality*

$$\|\widehat{\beta}_{\text{sorted}}-\beta^*\|_\Sigma^2 \leq C'\sigma^2\left(\frac{s\log(ep/s)}{\kappa(s)n}+\left(\frac{o}{n}\right)^{2-4/\tau}+\left(\frac{\log(1/\delta)}{n}\right)^{2-4/\tau}+\frac{\log(1/\delta)}{n}\right)$$

holds with probability at least $1-\delta$. In particular, when the noise ξ has sub-Gaussian distribution,

$$\|\widehat{\beta}_{\text{sorted}}-\beta^*\|_\Sigma^2 \leq C'\sigma^2\left(\frac{s\log(ep/s)}{\kappa(s)n}+\left(\frac{o\log(n/o)}{n}\right)^{2}+\frac{\log(1/\delta)}{n}\right).$$

For $\tau = \infty$, the bound implied by the inequality (2.2) is minimax optimal up to the logarithmic factors. However, it is sub-optimal for $\tau < \infty$. Interestingly, the bound holds uniformly over the range confidence levels $e^{-cn} < \delta < 1$ for the fixed choice of the regularization sequences $\{\lambda_i\}$ and $\{\mu_i\}$. In the special case where $o = 0$ (no adversarial corruption) and $\tau \geq 4$, inequality (2.2) yields a sub-Gaussian deviation bound with optimal dependence on the sparsity level s, despite the fact that the noise can be heavy-tailed.

2.2. Lower error bounds. It is well known [e.g. see 1] that in the absence of adversarial contamination and with $\Sigma = I_p$,

$$\inf_{\hat{\beta}} \sup_{|\beta|_0 \leq s} \Pr\left(\|X(\hat{\beta} - \beta)\|_2^2/n \geq C'\sigma^2 \frac{s\log(ep/s)}{n} \right) \geq c$$

for some positive constants c, C'. In the Huber's contamination framework coupled with the assumption that the additive noise ξ is Gaussian, results in [7] yield that no estimator can achieve the error smaller than $C'\sigma^2 \left(\frac{s\log(ep/s)}{n} + \left(\frac{o}{n}\right)^2 \right)$; of course, this lower bound is also valid for the adversarial contamination model. However, we could not find readily available lower bounds for the noise distributions beyond Gaussian. The following result gives an answer in this case.

Theorem 2.3. *Assume that at least one of the columns of X belongs to $\{-1,1\}^n$. Then*

$$\inf_{\hat{\beta}} \sup_{|\beta|_0 \leq s} \sup_{|\theta|_0 \leq o} \sup_{\sigma > 0} \sup_{P_\xi \in \mathscr{P}_{\tau,1}} \Pr_{(\beta,\theta,\sigma,P_\xi)} \left(\|X(\hat{\beta} - \beta)\|^2/n \geq C\sigma^2 \left(\frac{o}{n}\right)^{2-2/\tau} \right) \geq c,$$

for some $C, c > 0$ where the infimum is taken over all measurable estimators. For sub-Gaussian noise the inequality takes the form

$$\inf_{\hat{\beta}} \sup_{|\beta|_0 \leq s} \sup_{|\theta|_0 \leq o} \sup_{\sigma > 0} \sup_{P_\xi \in \mathscr{P}_{\infty,1}} \Pr_{(\beta,\theta,\sigma,P_\xi)} \left(\|X(\hat{\beta} - \beta)\|^2/n \geq C\sigma^2 \left(\frac{o}{n}\right)^2 \log(n/o) \right) \geq c.$$

The assumption on the design is very mild: indeed, it suffices that $\mathbf{1_n}$ is a column of X, which is equivalent to including the intercept term in the regression. Another special case is the Rademacher design, implying that the lower bound holds for the class of sub-Gaussian design matrices.

2.3. Main ideas of the proofs. In this section, we give a brief summary of the key ideas used in the proofs of Theorems 2.1 and 2.2. We start by discussing several useful properties of sub-Gaussian design vectors.

- We show that the design matrix X acts as a near-isometry on approximately sparse vectors: indeed, conditions of these type are crucial to guarantee success of sparse recovery. A detailed overview of similar assumptions that appear in the literature can be found in [32]. The specific form of the inequality that we prove is the following: with high probability, for all $u \in \mathbb{R}^p$ simultaneously

$$\frac{\|Xu\|_2^2}{n} \geq \frac{1}{2}\|u\|_\Sigma^2 - \|u\|_\lambda^2/4.$$

- The columns of the design matrix X need to be "nearly uncorrelated" with the columns of the identity matrix I_n. Indeed, consider a particular case where $n = p$ and $X = \sqrt{n} I_n$. In such a case, the model (2) becomes
$$y_i = \sqrt{n}(\beta^* + \theta^*) + \sigma\xi_i,$$
whence the only identifiable vector is $\beta^* + \theta^*$, making it impossible to consistently estimate β^* itself. Assumptions of this type are commonly referred to as the *incoherence conditions*. The incoherence employed in our proof takes the following form: with probability at least $1-\delta$, for all $u, v \in \mathbb{R}^p$ simultaneously and some absolute constant C',
$$\frac{1}{\sqrt{n}}|v^\top X u| \le \|u\|_\lambda \|v\|_2/10 + \|v\|_\lambda \|u\|_\Sigma/10 + C'\sqrt{\frac{\log(1/\delta)+1}{n}}\|u\|_\Sigma \|v\|_2.$$
Similarly, for any fixed $v \in \mathbb{R}^n$, the following inequality holds with probability at least $1-\delta$ uniformly over all $u \in \mathbb{R}^p$:
$$\frac{1}{\sqrt{n}}|v^\top X u| \le \|u\|_\lambda \|v\|_2/10 + C'\sqrt{\frac{\log(1/\delta)+1}{n}}\|u\|_\Sigma \|v\|_2.$$
In [9], authors establish very similar conditions for Gaussian design matrices.

We summarize the important properties of sub-Gaussian design matrices in the following result.

Theorem 2.4. *Assume that $X = Y\Sigma^{1/2}$, where Y has independent 1-sub-Gaussian rows. Then, with probability at least $1-e^{-cn}$, X satisfies (2.3), and with probability at least $1-\delta$, X satisfies (2.3) and (2.3).*

We note that properties (2.3), (2.3) and (2.3) are the only conditions we require from the design. Next, we explain the way we deal with heavy-tailed noise. Fix an integer $o' \le n$: we can then treat the largest, in absolute value, o' coordinates of the noise vector ξ as "outliers" that we merge with the vector θ^*, while the remaining coordinates of ξ are sufficiently well-behaved and "light-tailed." Therefore, we can replace $\xi_{(i)}$ by $\xi_{(i)}\mathbf{1}\{i \ge o'\}$ and $\sqrt{n}\theta_{(i)}$ by $\sqrt{n}\theta_{(i)} + \xi_{(i)}\mathbf{1}\{i \le o'\}$, where $\xi_{(i)}$ denotes the i-th largest, in absolute value, element of the vector ξ. Note that this new noise vector is no longer centered, and that $o + o'$ becomes a new upper bound of the number of outliers. We then define the "good" event $\mathscr{E}$ via
$$\mathscr{E} = \left\{ n/10 \le \sum_{j \ge o'} |\xi|^2_{(j)} \le 2n \text{ and } \forall j \ge o', |\xi|_{(j)} \le \sqrt{n}\mu_j/20 \right\},$$
and show that $\mathscr{E}$ holds with high probability (see Lemmas A.2 and A.1). The following inequality is our main result which in turn implies the bounds of Theorems 2.1 and 2.2 under various assumptions on ξ.

Theorem 2.5. *Fix any $o' \ge o$. There exist absolute positive constants c, C' with the following properties: assume that $\sum_{i=1}^s \lambda_i^2/\kappa(s) + \log(1/\delta)/n + \sum_{i=1}^{o'} \mu_i^2 \le c$ and that event $\mathscr{E}$ occurs. If moreover properties (2.3), (2.3) and (2.3) of the design matrix hold, the following bound is valid whenever $\lambda_i \le \mu_i$:*
$$\|\widehat{\beta} - \beta^*\|_\Sigma^2 \le C'\sigma^2 \left(\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)} + \max_{j \ge o'}(\lambda_j^2/\mu_j^2)\left(\sum_{j=1}^{o'} \mu_j^2\right)^2 + \frac{\log(1/\delta)}{n} \right).$$

3. Discussion and comparison to existing results. Below, we give a brief overview of existing literature and results that are most closely related to the problem considered in this work. For an extended overview of the classical and modern approaches to robust regression, we refer the reader to the excellent discussions in [27, section 2] and [9, section 4].

The idea of taking advantage of sparsity of the sequence of outliers and applying Lasso or Dantzig selector-type algorithms has been previously suggested in [6, 16, 27, 23, 12], among other works. In particular, in [12] authors note that the solution $\widetilde{\beta}$ of the convex problem

$$\left(\widetilde{\beta}, \widetilde{\theta}\right) = \underset{\beta\in\mathbb{R}^p, \theta\in\mathbb{R}^n}{\operatorname{argmin}} \left[\frac{1}{2n}\sum_{j=1}^{n}(y_j - X_j^T\beta - \theta)^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\theta\|_1\right]$$

can be equivalently written, after carrying out minimization over θ explicitly, as

$$\widetilde{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \left[\lambda_2^2\sum_{j=1}^{n} H\left(\frac{y_j - X_j^T\beta}{\lambda_2\sqrt{n}}\right) + \lambda_1\|\beta\|_1\right],$$

where

$$H(x) = \begin{cases} x^2/2, & |x| \le 1, \\ |x| - 1/2, & |x| > 1 \end{cases}$$

is the Huber's loss function; similar connection has been used in several earlier works, including [25, 14]. More recently, in [9, 29] authors improved the bounds proven in [23] and showed that for the Gaussian design and Gaussian additive noise, $\|\widetilde{\beta} - \beta^*\|_\Sigma^2 = O_P\left(\frac{s\log(p)}{n} + \left(\frac{o\log(n)}{n}\right)^2\right)$ which is nearly minimax optimal in the ratio $\frac{o}{n}$ (note that the additional $\log(n)$ factor makes the bound suboptimal [7]). At the same time, estimators that achieve minimax optimality, such as the methods based on regression depth [15], are not computationally feasible, Our work has been partially motivated by the question raised by the authors of [9], namely, whether the penalized ERM-type methods can also handle the case of heavy-tailed additive noise variables $\{\xi_i\}_{i=1}^n$ and yield optimal or near-optimal rates. Results of the present paper give a generally affirmative answer and make an extra step by proving that it is possible to be computationally efficient and minimax optimal with respect to the sparsity level and contamination level, while being completely adaptive and achieve strong concentration of the resulting estimators simultaneously. Related results in the literature, such as the work [13], establish strong theoretical guarantees for the estimators that are not efficiently computable.

Very recently, a model with adversarially contaminated design and response was considered in [26], however, the resulting bounds are only valid for very sparse signals such that $s \lesssim \sqrt{n}$. A similar setup was also considered in [11] and [24] without the sparsity assumptions. For example, in [24] the authors used a black-box "filtering" algorithm to eliminate outliers from the design matrix provided that the covariance matrix Σ is known. Our goal was to show that similar results hold for a simple procedure and without additional knowledge about the parameters of the problem. Finally, let us remark that in the low-dimensional case $p < n$, there exist estimators capable of approximating β^* regardless of the number of outliers o as long the following conditions hold: (i) $o < cn$, (ii) the contamination is oblivious and (iii) the design matrix X is sufficiently nice (e.g., has normally distributed rows); this fact was proven in [3].

Acknowledgements. Stanislav Minsker and Lang Wang acknowledge support by the National Science Foundation grants CIF-1908905 and DMS CAREER-2045068.
The work of Mohamed Ndaoud was supported by a Chair of Excellence in Data Science granted by the CY Initiative.

References.

[1] Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets Lasso: improved oracle bounds and optimality. *Annals of Statistics*, 46(6B):3603–3642, 2018.
[2] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.
[3] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.
[4] Malgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. SLOPE—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103, 2015.
[5] Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of statistics*, 35(6):2313–2351, 2007.
[6] Emmanuel J Candes and Paige A Randall. Highly robust error correction byconvex programming. *IEEE Transactions on Information Theory*, 54(7):2829–2840, 2008.
[7] Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for Huber's ε-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.
[8] Laëtitia Comminges, Olivier Collier, Mohamed Ndaoud, and Alexandre B Tsybakov. Adaptive robust estimation in sparse vector model. *The Annals of Statistics*, 49(3):1347–1377, 2021.
[9] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. *Advances in neural information processing systems*, 32, 2019.
[10] Alexis Derumigny. Improved bounds for square-root LASSO and square-root SLOPE. *Electronic Journal of Statistics*, 12(1):741–766, 2018.
[11] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.
[12] David Donoho and Andrea Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969, 2016.
[13] Gianluca Finocchio, Alexis Derumigny, and Katharina Proksch. Robust-to-outliers square-root Lasso, simultaneous inference with a MOM approach. *arXiv preprint arXiv:2103.10420*, 2021.
[14] Irène Gannaz. Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing*, 17(4):293–310, 2007.
[15] Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.
[16] Jason N Laska, Mark A Davenport, and Richard G Baraniuk. Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. In *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pages 1556–1560. IEEE, 2009.
[17] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020.

[18] Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis*, pages 277–299. Springer, 2017.
[19] Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *ArXiv preprint 1907.11391*, 2019.
[20] Shahar Mendelson. Learning without Concentration. Technical report arXiv:1401.0304, 2014.
[21] Stanislav Minsker and Mohamed Ndaoud. Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15(2):6036–6070, 2021.
[22] Mohamed Ndaoud. Scaled minimax optimality in high-dimensional linear regression: A non-convex algorithmic regularization approach. *10.48550/ARXIV.2008.12236*, 2020.
[23] Nam H Nguyen and Trac D Tran. Robust Lasso with missing and grossly corrupted observations. *IEEE transactions on information theory*, 59(4):2036–2058, 2012.
[24] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.
[25] Sylvain Sardy, Paul Tseng, and Andrew Bruce. Robust wavelet denoising. *IEEE Transactions on Signal Processing*, 49(6):1146–1152, 2001.
[26] Takeyuki Sasai and Hironori Fujisawa. Outlier robust and sparse estimation of linear regression coefficients. *arXiv preprint arXiv:2208.11592*, 2022.
[27] Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.
[28] Benjamin Stucky and Sara Van De Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18(67):1–29, 2017.
[29] Philip Thompson. Outlier-robust sparse/low-rank least-squares regression and robust matrix completion. *arXiv preprint arXiv:2012.06750*, 2020.
[30] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
[31] Sara A Van de Geer. *Estimation and testing under sparsity*. Springer, 2016.
[32] Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Appendix A. Technical results.

Recall that $\mathbf{E}(|\xi_i|^\tau) \le a^\tau$ for some $\tau \ge 2$ and $a \ge 1$, implying that $\Pr(|\xi| \ge t) \le (a/t)^\tau$ for all t. Without loss of generality, we will assume that $a = 1$, otherwise we can simply replace ξ_i by ξ_i/a and σ by $\sigma \cdot a$.

Lemma A.1. *For any $1 \le i \le n$, set $\mu_i = \frac{C}{\sqrt{n}} \left(\frac{n}{i}\right)^{1/\tau}$ for $\tau \ge 2$ and $C \ge 80$. Then for any $k \ge 1$, the following inequality holds:*

$$\Pr\left(\max_{i \ge k} \frac{|\xi|_{(i)}}{\sqrt{n}\mu_i} \ge 1/20\right) \le 2e^{-k}.$$

Moreover, for all i such that $\log(n) \le i \le n$,

$$\mathbf{E}\left(\frac{|\xi|_{(i)}}{\sqrt{n}\mu_i}\right) \le 1.$$

The result holds for sub-Gaussian noise as well with the choice $\mu_i = \lambda_i$.

Proof. For any fixed $i \geq k$,

$$\Pr\left(\frac{|\xi|_{(i)}}{\sqrt{n}\mu_i} \geq 1/20\right) = \Pr\left(\exists |I| = i, \forall j \in I \quad \frac{|\xi_j|}{\sqrt{n}\mu_i} \geq 1/20\right).$$

Therefore, applying the inequality $\binom{n}{i} \leq e^{i\log(en/i)}$ and the assumption $C \geq 80$, we deduce that

$$\Pr\left(\frac{|\xi|_{(i)}}{\sqrt{n}\mu_i} \geq 1/20\right) \leq e^{i\log(en/i)}\Pr\left(\frac{|\xi_1|}{\sqrt{n}\mu_i} \geq 1/20\right)^i \leq e^{-i}.$$

We conclude using the union bound over $i \geq k$ and the fact that $\sum_{i=k}^n e^{-i} \leq 2e^{-k}$.
To get the result in expectation, let us denote $\gamma := \frac{|\xi|_{(i)}}{\sqrt{n}\mu_i}$. Observe that

$$\mathbf{E}(\gamma) \leq \mathbf{E}(\gamma \mathbf{1}\{\gamma \leq 1/20\}) + \mathbf{E}(\gamma \mathbf{1}\{\gamma \geq 1/20\}).$$

Using Cauchy-Schwarz inequality, we get that

$$\mathbf{E}\left(\frac{|\xi|_{(i)}}{\sqrt{n}\mu_i}\right) \leq 1/20 + \sqrt{\mathbf{E}\left(\frac{|\xi|^2_{(i)}}{n\mu_i^2}\right)\Pr\left(\frac{|\xi|_{(i)}}{\sqrt{n}\mu_i} \geq 1/20\right)}.$$

Since $n\mu_i^2 \geq 8$ and $\mathbf{E}(|\xi|^2_{(i)}) \leq \mathbf{E}\left(\sum_{i=1}^n \xi_i^2\right) \leq n$ we conclude using (1) that

$$\mathbf{E}\left(\frac{|\xi|_{(i)}}{\sqrt{n}\mu_i}\right) \leq 1/20 + \frac{\sqrt{n}}{2}e^{-i/2} \leq 1,$$

as long as $i \geq \log(n)$. ∎

Lemma A.2. *Assume that* $\mathbf{E}(\xi_i^2 \mathbf{1}\{|\xi_i| \leq 1/2\}) \geq 1/4$ *and that* $o \leq n/1000$. *Then*

$$\Pr\left(n/10 \leq \sum_{i=o}^n |\xi|^2_{(i)} \leq 2n\right) \geq 1 - 3e^{-co},$$

for an absolute constant $c > 0$.

Proof. For the upper bound, we only need to control the random variables bounded by $C\sqrt{n/o}$, in view of Lemma A.1 applied with $k = o$. Set

$$R = C\left(\frac{n}{o}\right)^{1/2}.$$

Observe that, as long as $|\xi|_{(o)} \leq R$, we have

$$\sum_{i=o}^n |\xi|^2_{(i)} \leq \sum_{i=1}^n \xi_i^2 \mathbf{1}\{|\xi_i| \leq R\}.$$

Since $\xi_i^2 \mathbf{1}\{|\xi_i| \leq R\} \leq R^2$ and $\mathbf{E}(\xi_i^2 \mathbf{1}\{|\xi_i| \leq R\}) \leq 1$, Hoeffding's inequality yields that

$$\Pr\left(\left\{\sum_{i=o}^n |\xi|^2_{(i)} \geq 2n\right\} \cap \left\{|\xi|_{(o)} \leq R\right\}\right) \leq \exp(-n/R^2).$$

Noticing that $n/R^2 = o/C^2$, the upper bound follows for $c = 1/C^2$ from the inequality (1), since

$$\Pr\left(\sum_{i=o}^{n} |\xi|_{(i)}^2 \geq 2n\right) \leq \Pr\left(\left\{\sum_{i=o}^{n} |\xi|_{(i)}^2 \geq 2n\right\} \cap \left\{|\xi|_{(o)} \leq R\right\}\right) + \Pr\left(|\xi|_{(o)} \geq R\right) \leq 2\exp(-co).$$

For the lower bound, observe that

$$\Pr\left(n/10 \geq \sum_{i=o}^{n} |\xi|_{(i)}^2\right) = \Pr\left(n/10 \geq \min_{|I|=n-o+1} \sum_{i\in I} |\xi_i|^2\right) \leq e^{o\log(\frac{en}{o-1})} \Pr\left(n/10 \geq \sum_{i=1}^{n-o+1} |\xi_i|^2\right),$$

where we use the union bound together with the relations $\binom{n}{n-o+1} = \binom{n}{o-1} \leq e^{o\log(\frac{en}{o-1})}$. Since

$$\mathbf{E}(|\xi_i|^2 \mathbf{1}\{|\xi_i| \leq 1/2\}) \geq 1/4,$$

$$\Pr\left(n/10 \geq \sum_{i=o}^{n} |\xi|_{(i)}^2\right) \leq e^{o\log(en/(o-1))} \Pr\left(-n/10 \geq \sum_{i=1}^{n-o+1} |\xi_i|^2 \mathbf{1}\{|\xi_i| \leq 1/2\} - \mathbf{E}(|\xi_i|^2 \mathbf{1}\{|\xi_i| \leq 1/2\})\right).$$

We conclude, using Hoeffding's inequality, that

$$\Pr\left(n/10 \geq \sum_{i=o}^{n} |\xi|_{(i)}^2\right) \leq e^{o\log(en/(o-1)) - n/100} \leq \exp(-cn),$$

for c small enough. ∎

Lemma A.3. *Assume that ξ is a centered Gaussian vector such that $\mathbf{E}(\xi_i^2) \leq 1$ for all $1 \leq i \leq n$. Then*

$$\mathbf{E}\left(\max_{i=1,\dots,n} \frac{|\xi|_{(i)}}{\sqrt{\log(en/i)}}\right) \leq 20.$$

Proof. Set $\lambda_i^2 = 4\log(en/i)$. Let $\hat{\imath}$ be an index such that $\max_{i=1,\dots,n} \frac{|\xi|_{(i)}}{\lambda_i} = \frac{|\xi|_{(\hat{\imath})}}{\lambda_{\hat{\imath}}}$. Then

$$\begin{aligned}
\mathbf{E}\left(\max_{i=1,\dots,n} \frac{|\xi|_{(i)}}{\lambda_i}\right) &\leq 1 + \mathbf{E}\left(\frac{|\xi|_{(\hat{\imath})}}{\lambda_{\hat{\imath}}} \mathbf{1}\left(\frac{|\xi|_{(\hat{\imath})}}{\lambda_{\hat{\imath}}} \geq 1\right)\right) \\
&\leq 1 + \int_1^\infty \Pr\left(|\xi|_{(\hat{\imath})} \geq t\lambda_{\hat{\imath}}\right) \mathrm{d}t \\
&\leq 1 + \int_1^\infty \Pr\left(\hat{\imath}\exp\left(\xi_{(\hat{\imath})}^2/4\right) \geq \hat{\imath}\exp\left(t^2\lambda_{\hat{\imath}}^2/4\right)\right) \mathrm{d}t.
\end{aligned}$$

On the one hand, we have that

$$\hat{\imath}\exp\left(\xi_{(\hat{\imath})}^2/4\right) \leq \sum_{j=1}^{i} \exp\left(\xi_{(j)}^2/4\right) \leq \sum_{j=1}^{n} \exp\left(\xi_j^2/4\right).$$

On the other hand, for $t^2 \geq 1$

$$\hat{i}\exp\left(t^2\lambda_{\hat{i}}^2/4\right) = \hat{i}\left(en/\hat{i}\right)^{t^2} \geq ne^{t^2}.$$

Therefore,

$$\mathbf{E}\left(\max_{i=1,\dots,n}\frac{|\xi|_{(i)}}{\lambda_i}\right) \leq 1 + \int_1^\infty \Pr\left(\sum_{j=1}^n \exp\left(\xi_j^2/4\right) \geq ne^{t^2}\right) \mathrm{d}t.$$

Since $\mathbf{E}\left(\exp\left(\xi_j^2/4\right)\right) \leq 5$, we conclude using Markov's inequality that

$$\mathbf{E}\left(\max_{i=1,\dots,n}\frac{|\xi|_{(i)}}{\lambda_i}\right) \leq 1 + 5\int_1^\infty e^{-t^2}\mathrm{d}t \leq 10.$$

Re-scaling λ_i by 2 yields the result. ∎

Appendix B. Proofs of the main results.

The proof of the lower bound in inspired by results in [8] where the goal was to estimate the nuisance parameter θ rather than the signal β itself.

B.1. Proof of Theorem 2.3. Assume that the first column v of X is such that $v \in \{\pm 1\}^n$. Let us choose β proportional to the canonical basis vector e_1 (recall that β is sparse) such that $X\beta = \frac{\|X\beta\|_2}{\sqrt{n}}v$. Moreover, let ξ be a vector of i.i.d. Rademacher random variables. Clearly, $P_\xi \in \mathscr{P}_{\tau,1}$ for all values of τ. The vector θ will be chosen to be random with i.i.d entries such that $\theta_i = \sigma\left(\frac{o}{n}\right)^{-1/\tau}\alpha_i v_i$, where α_i are i.i.d. Bernoulli random variables with parameter o/n. Therefore,

$$\mathbf{E}(\theta_i) = \sigma\left(\frac{o}{n}\right)^{1-1/\tau} v_i \text{ and } \mathrm{Var}(\theta_i) = \sigma^2\left(\frac{o}{n}\right)^{1-2/\tau}\left(1-\left(\frac{o}{n}\right)^{1-2/\tau}\right) \leq \sigma^2\left(\frac{o}{n}\right)^{1-2/\tau}.$$

Notice that θ is not exactly of sparsity less than o but we will deal with this technicality exactly as in [8]. Finally, set

$$\frac{\|X\beta\|_2}{\sqrt{n}} = \sigma\left(\frac{o}{n}\right)^{1-1/\tau}.$$

Notice that

$$Y_i = (X\beta - \theta + \sigma\xi)_i = -(\theta_i - \mathbf{E}(\theta)_i)v_i + \sigma\xi_i.$$

Hence, the distributions of Y defined by the model corresponding to $(\beta, -\theta, \sigma, P_\xi)$ and $(0,0,\tilde{\sigma},\tilde{P}_\xi)$ are identical. Here, $\tilde{\sigma}^2 = \sigma^2(1+\left(\frac{o}{n}\right)^{1-2/\tau}(1-\left(\frac{o}{n}\right)^{1-2/\tau})) \sim \sigma^2$ and $\tilde{P}_\xi$ is the distribution of $\zeta = \sigma\left(\frac{o}{n}\right)^{-1/\tau}((\alpha_i - \left(\frac{o}{n}\right))v_i + \left(\frac{o}{n}\right)^\tau \xi_i)/\tilde{\sigma}$. Notice that $|\zeta| \geq 2$ only if $\alpha_i = 1$, hence for all $2 \leq t \leq (o/n)^{-1/\tau}$ we have that

$$\Pr(|\zeta| \geq t) = o/n \leq \left(\frac{1}{t}\right)^\tau,$$

and for $t > \left(\frac{o}{n}\right)^{-1/\tau}$,

$$\Pr(|\zeta| \geq t) = 0 \leq \left(\frac{1}{t}\right)^\tau.$$

Therefore, $\tilde{P}_\xi \in \mathscr{P}_{\tau,1}$. Let us denote

$$\mathscr{R}^* = \inf_{\hat{\beta}} \sup_{|\beta|_0 \le s} \sup_{|\theta|_0 \le o} \sup_{\sigma>0} \sup_{P_\xi \in \mathscr{P}_{\tau,1}} \Pr_{(\beta,\theta,\sigma,P_\xi)} \left(\|X(\hat{\beta}-\beta)\|^2/n \ge \sigma^2/16 \left(\frac{o}{n}\right)^{2-2/\tau} \right).$$

It is easy to notice that

$$\mathscr{R}^* \ge \inf_{\hat{T}} \left(\Pr_{(0,0,\tilde{\sigma},\tilde{P}_\xi)} \left(|\hat{T}| \ge \tilde{\sigma}/4 \left(\frac{o}{n}\right)^{1-1/\tau} \right) \bigvee \Pr_{(\beta,-\theta,\sigma,P_\xi)} \left(\left| \hat{T} - \frac{\|X\beta\|_2}{\sqrt{n}} \right| \ge \sigma/4 \left(\frac{o}{n}\right)^{1-1/\tau} \right) \right),$$

where $\hat{T}$ is an estimator of $\frac{\|X\beta\|_2}{\sqrt{n}}$. Since $\tilde{\sigma} \ge \sigma$ and the distributions $\Pr_{(0,0,\tilde{\sigma},\tilde{P}_\xi)}$, $\Pr_{(\beta,-\theta,\sigma,P_\xi)}$ are equal, we deduce that

$$\mathscr{R}^* \ge \inf_{\hat{T}} \left(\Pr\left(|\hat{T}| \ge \tilde{\sigma}/4 \left(\frac{o}{n}\right)^{1-1/\tau} \right) \vee \Pr\left(|\hat{T}| \le \tilde{\sigma}/4 \left(\frac{o}{n}\right)^{1-1/\tau} \right) \right),$$

as long as $\frac{\|X\beta\|}{\sqrt{n}} \ge \frac{\tilde{\sigma}\left(\frac{o}{n}\right)^{1-1/\tau}}{2}$. The last condition is satisfied since $\tilde{\sigma} \le 2\sigma$. We conclude that

$$\mathscr{R}^* \ge 1/2.$$

In the case of sub-Gaussian noise, we choose $\theta_i = \sigma\sqrt{\log\left(\frac{o}{n}\right)}\alpha_i v_i$ and follow the same argument.

B.2. Proof of Theorem 2.4. We start with the property given by the inequality (2.3). We will show first that for all vectors u,

$$\frac{\|Xu\|_2^2}{n} \ge \frac{1}{2}\|u\|_\Sigma^2 - \|u\|_\lambda^2/4,$$

where $\|u\|_\Sigma^2 = u^\top \Sigma u$. Define $\tilde{X}$ such that $X = \tilde{X}\Sigma^{1/2}$, let A be the set

$$A = \left\{ u : \ \|u\|_\Sigma^2 \ge \|u\|_\lambda^2/2 \right\},$$

and $K = \{\Sigma^{1/2}u, \ u \in A\}$. Moreover, we will denote the sphere of radius 1 in $\mathbb{R}^p$ via S^{p-1}. Since $\tilde{X}$ is isotropic, Corollary 1.5 in [18] implies that for any $h \ge 0$ and for all $v \in K \cap S^{p-1}$

$$\frac{\|\tilde{X}v\|_2}{\sqrt{n}} \ge 1 - C'\left(\frac{\omega(K\cap S^{p-1}) + h}{\sqrt{n}}\right),$$

with probability at least $1 - e^{-ch^2}$. Here, C' is a positive absolute constant and $\omega(T)$ corresponds to the Gaussian mean width of $T \subseteq \mathbb{R}^p$ defined via

$$\omega(T) = \mathbb{E} \sup_{u,v\in T} \langle \xi, u - v\rangle$$

where ξ has standard normal law. It is clear that

$$\omega(K\cap S^{p-1}) = \mathbf{E}\left(\sup_{v\in K\cap S^{p-1}} \xi^\top v \right) = \mathbf{E}\left(\sup_{u\in A} \xi_\Sigma^\top u/\|u\|_\Sigma \right),$$

where ξ_Σ is a centered Gaussian random vector with covariance Σ. Therefore,

$$\omega(K\cap S^{p-1}) \le \mathbf{E}\left(\max_i \frac{|\xi_\Sigma|_{(i)}}{\lambda_i}\right)\sup_{u\in A}\|u\|_\lambda/\|u\|_\Sigma.$$

Since diagonal elements of Σ do not exceed 1 and $\lambda_i = C\sqrt{\frac{\log(ep/i)}{n}}$, $i=1,\dots,p$, we deduce from the bound of Lemma A.3 that whenever C in the definition of λ is large enough,

$$\omega(K\cap S^{p-1}) \le \frac{1}{10C'}.$$

Hence, for all $u\in A$

$$\frac{\|Xu\|_2}{\sqrt{n}} \ge \|u\|_\Sigma/\sqrt{2},$$

with probability at least $1-e^{-cn}$. This concludes the first part of the proof, since the inequality is always true for $u\notin A$. We will now prove inequality (2.3). Fix $v\in\mathbb{R}^p$. We want to show that for all $u\in\mathbb{R}^p$,

$$\frac{1}{\sqrt{n}}|v^\top Xu| \le \|u\|_\lambda\|v\|_2/10 + C\sqrt{\frac{1+\log(1/\delta)}{n}}\|u\|_\Sigma\|v\|_2.$$

This is equivalent to establishing that for all $u\in\mathbb{R}^p$,

$$\frac{1}{\sqrt{n}}\xi^\top\Sigma^{1/2}u \le \|u\|_\lambda/10 + C\sqrt{\frac{1+\log(1/\delta)}{n}}\|u\|_\Sigma,$$

where ξ is an isotropic sub-Gaussian vector. For $\alpha>0$, let A_α be the set

$$A_\alpha = \{u:\ \|u\|_\Sigma = 1,\quad \|u\|_\lambda \le \alpha\},$$

and let K_α be the set $K_\alpha = \{\Sigma^{1/2}u,\ u\in A_\alpha\}$. Notice that $K_\alpha\subset S^{p-1}$ and that

$$\sup_{u\in A_\alpha}\frac{1}{\sqrt{n}}\xi^\top\Sigma^{1/2}u = \sup_{v\in K_\alpha}\frac{1}{\sqrt{n}}\xi^\top v.$$

Hence, applying Theorem 4.1 in [18] on K_α, we get that

$$\sup_{u\in A_\alpha}\frac{1}{\sqrt{n}}\xi^\top\Sigma^{1/2}u \le C'\left(\mathbf{E}\left(\sup_{u\in A_\alpha}\frac{1}{\sqrt{n}}\xi^\top\Sigma^{1/2}u\right)+\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

for some $C'>0$ with probability $1-\delta$, where ξ is a standard Gaussian random vector. Using the bound (B.2) we deduce that the inequality

$$\sup_{u\in A_\alpha}\frac{1}{\sqrt{n}}\xi^\top\Sigma^{1/2}u \le \alpha/20 + C'\sqrt{\frac{\log(1/\delta)}{n}}$$

holds with probability at least $1-\delta$. We can now conclude, using the peeling argument as in [9, Lemma 5] that

$$\sup_{\|u\|_\Sigma=1}\frac{1}{\sqrt{n}}\xi^\top\Sigma^{1/2}u \le \|u\|_\lambda/10 + C'\sqrt{\frac{1+\log(1/\delta)}{n}},$$

again with probability at least $1-\delta$. The proof is complete by homogeneity of the norm. For the remaining part of the proof, we need to show that for all u, v

$$\frac{1}{\sqrt{n}}|v^\top \tilde{X}\Sigma^{1/2}u| \le \|u\|_\lambda \|v\|_2/10 + \|v\|_\lambda \|u\|_\Sigma/10 + C\sqrt{\frac{1+\log(1/\delta)}{n}}\|u\|_\Sigma \|v\|_2,$$

with probability at least $1-\delta$. For $\alpha, \beta > 0$, let A_α and B_β be the sets

$$A_\alpha = \{u : \ \|u\|_\Sigma = 1, \ \|u\|_\lambda \le \alpha\},$$

and

$$B_\beta = \{v : \ \|v\|_2 = 1, \ \|v\|_\lambda \le \beta\},$$

and let K_α be the set $K_\alpha = \{\Sigma^{1/2}u, \ u \in A_\alpha\}$. Notice that $K_\alpha \subset S^{p-1}$ and that

$$\sup_{(u,v)\in A_\alpha \times B_\beta} \frac{1}{\sqrt{n}} v^\top \tilde{X}\Sigma^{1/2}u = \sup_{(b,v)\in K_\alpha \times B_\beta} \frac{1}{\sqrt{n}} v^\top \tilde{X} b.$$

Let us denote by $Z_{v,b}$ the sub-Gaussian process $v^\top \tilde{X} b$ where v, b are both of norm 1. We see that

$$\mathbf{E}\left(Z_{v,b} - Z_{v',b'}\right)^2 = 2(1 - \langle v, v'\rangle\langle b, b'\rangle) \le 4 - 2(\langle v, v'\rangle + \langle b, b'\rangle) \le \|v - v'\|^2 + \|b - b'\|^2.$$

Therefore,

$$\mathbf{E}\left(Z_{v,b} - Z_{v',b'}\right)^2 \le \mathbf{E}\left(\xi^\top (v - v')\right)^2 + \mathbf{E}\left(\tilde{\xi}^\top (b - b')\right)^2,$$

where ξ and $\tilde{\xi}$ are both standard Gaussian vectors. Applying Theorem 4.1 in [18] on $K_\alpha \times B_\beta$, we deduce that

$$\sup_{(u,v)\in A_\alpha \times B_\beta} \frac{1}{\sqrt{n}} v^\top \tilde{X}\Sigma^{1/2}u \le C'\left(\mathbf{E}\left(\sup_{u\in A_\alpha} \frac{1}{\sqrt{n}} \tilde{\xi}^\top \Sigma^{1/2}u\right) + \mathbf{E}\left(\sup_{v\in B_\alpha} \frac{1}{\sqrt{n}} \tilde{\xi}^\top v\right) + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

for some $C' > 0$ with probability at least $1-\delta$. Using the inequality (B.2), we get that

$$\sup_{(u,v)\in A_\alpha \times B_\beta} \frac{1}{\sqrt{n}} v^\top \tilde{X}\Sigma^{1/2}u \le \alpha/20 + \beta/20 + C'\sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1-\delta$. We can now conclude, using the double-peeling argument as in [9, Lemma 6] that

$$\sup_{\|u\|_\Sigma = 1, \|v\|_2 = 1} \frac{1}{\sqrt{n}} v^\top \tilde{X}\Sigma^{1/2}u \le \|u\|_\lambda/10 + \|v\|_\lambda/10 + C'\sqrt{\frac{1+\log(1/\delta)}{n}},$$

with probability at least $1-\delta$. The desired result now follows by homogeneity of the norm.

Proposition B.1. *Let X be a matrix with sub-Gaussian rows. Then for all vectors u, v*

$$\|Xu/\sqrt{n} + v\|_2^2 \ge \frac{1}{8}\|u\|_\Sigma^2 + \frac{1}{8}\|v\|_2^2 - \|u\|_\lambda^2/2 - \|v\|_\mu^2/2,$$

with probability at least $1 - e^{-cn}$.

Proof. We have

$$\|Xu/\sqrt{n}+v\|_2^2 \geq \|Xu/\sqrt{n}\|_2^2 + \|v\|_2^2 - \frac{2}{\sqrt{n}}|v^\top Xu|.$$

We conclude using the inequalities (2.3) and (2.3) with $\delta = e^{-cn}$ for c small enough. ∎

B.3. Proof of Theorem 2.5. Throughout this section, we set $\Delta^\beta = \widehat{\beta} - \beta^*$, $\Delta^\theta = \widehat{\theta} - \theta^*$ and let $\Delta = [\Delta^\beta; \Delta^\theta] \in \mathbb{R}^{p+n}$ be the augmented error vector. We also introduce the following additional notation with the goal of simplifying the expressions; recall that

$$Q(\beta,\theta) := \frac{1}{2n}\left\|Y - X\beta - \sqrt{n}\theta\right\|_2^2,$$

and let

- $\widehat{Q} := Q(\widehat{\beta},\widehat{\theta})$ and $Q^* := Q(\beta^*,\theta^*)$;
- $A^{(n)} := \frac{1}{\sqrt{n}}A$, whenever A is a scalar, vector or a matrix;
- $\widehat{\xi} := Y - X\widehat{\beta} - \sqrt{n}\widehat{\theta}$.

Moreover, note that $\widehat{Q} = \frac{1}{2n}\left\|\widehat{\xi}\right\|_2^2$ and $Q^* = \frac{1}{2n}\|\xi\|_2^2$. In the rest of the proof we assume that the event

$$\mathscr{E} = \left\{ n/10 \leq \sum_{j\geq o'} \xi_{(j)}^2 \leq 2n \text{ and } \forall j \geq o', |\xi|_{(j)} \leq \sqrt{n}\mu_{(j)}/20 \right\}$$

occurs and that Properties 1, 2 and 3 of the design matrix hold as well.

Given $o' > o$, let us replace the dense noise vector ξ by the new vector obtained from ξ by replacing the largest o' coordinates by 0. We will treat the largest o' coordinates removed from ξ as a subset of adversarial outliers, and will replace the corruption parameter o by $2o'$. From now on, we can assume that the entries of the "new" noise vector ξ are bounded by $|\xi|_{(o')}$.

Let us note that several steps of the proof below follow the argument in [9]. First, recall that $S = \{j : \beta_j^* \neq 0\}$, $O = \{j : \theta_j^* \neq 0\}$ and $s = \mathrm{Card}(S)$, $\mathrm{Card}(O) \leq 2o'$. The following lemma ensures that the augmented error vector Δ belongs to the cone defined in (2), with $c_0 = 4$.

Lemma B.2. *The following inequality holds:*

$$\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_\mu \leq 4\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)} + \frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma + 4\sqrt{\sum_{i=1}^{o'} \mu_i^2}\|\Delta^\theta\|_2.$$

Equivalently, $\Delta \in \mathscr{C}(4, s, o', \delta, \Sigma)$.

Proof of Lemma B.2. By the definition of $\widehat{\beta}, \widehat{\theta}$, we have that

$$\widehat{Q}^{\frac{1}{2}} - {Q^*}^{\frac{1}{2}} \leq \left(\|\beta^*\|_\lambda - \left\|\widehat{\beta}\right\|_\lambda\right) + \left(\|\theta^*\|_\mu - \left\|\widehat{\theta}\right\|_\mu\right).$$

Using Lemma A.1 in [1], we get that

$$\left(\|\beta^*\|_\lambda - \left\|\widehat{\beta}\right\|_\lambda\right) \leq 2\sqrt{\sum_{i=1}^s \lambda_i^2}\|\Delta^\beta\|_2 - \|\Delta^\beta\|_\lambda.$$

If $\Delta^\beta \in \mathscr{C}(s,4)$, then

$$\left(\|\beta^*\|_\lambda - \left\|\widehat{\beta}\right\|_\lambda\right) \leq 2\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)}}\|\Delta^\beta\|_\Sigma - \|\Delta^\beta\|_\lambda.$$

Otherwise,

$$\left(\|\beta^*\|_\lambda - \left\|\widehat{\beta}\right\|_\lambda\right) \leq -\|\Delta^\beta\|_\lambda/2.$$

In both cases we have that

$$\left(\|\beta^*\|_\lambda - \left\|\widehat{\beta}\right\|_\lambda\right) \leq 2\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)}}\|\Delta^\beta\|_\Sigma - \|\Delta^\beta\|_\lambda/2.$$

Similarly, we observe that

$$\left(\|\theta^*\|_\mu - \left\|\widehat{\theta}\right\|_\mu\right) \leq 2\sqrt{\sum_{i=1}^{o'} \mu_i^2}\|\Delta^\theta\|_2 - \|\Delta^\theta\|_\mu.$$

Combining the inequalities above with (5), we deduce that

$$\widehat{Q}^{\frac{1}{2}} - Q^{*\frac{1}{2}} \leq 2\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)}}\|\Delta^\beta\|_\Sigma + 2\sqrt{\sum_{i=1}^{o'} \mu_i^2}\|\Delta^\theta\|_2 - (\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_\mu)/2.$$

Recall that the property (2.3) of sub-Gaussian designs together with the inequality $\|\xi\| \leq 2\sqrt{n}$ yields the bound

$$\frac{1}{n}|\xi^\top X\Delta^\beta| \leq \|\Delta^\beta\|_\lambda/10 + 2\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma.$$

On the other hand, convexity of $Q(\beta,\theta)^{\frac{1}{2}}$ implies that

$$\widehat{Q}^{\frac{1}{2}} - Q^{*\frac{1}{2}} \geq \left\langle \partial_\beta\left(Q^{\frac{1}{2}}\right)(\beta^*,\theta^*), \widehat{\beta} - \beta^* \right\rangle + \left\langle \partial_\theta\left(Q^{\frac{1}{2}}\right)(\beta^*,\theta^*), \widehat{\theta} - \theta^* \right\rangle,$$

where $\partial_\beta\left(Q^{\frac{1}{2}}\right)(\beta^*,\theta^*)$ $\left(\text{ or } \partial_\theta\left(Q^{\frac{1}{2}}\right)(\beta^*,\theta^*)\right)$ represents the subgradient of $Q^{\frac{1}{2}}$ with respect to β (or θ), evaluated at the point (β^*,θ^*). If $Q^* \neq 0$, we have that

$$\partial_\beta\left(Q^{\frac{1}{2}}\right)(\beta^*,\theta^*) = -\frac{1}{2}Q^{*-\frac{1}{2}} \cdot \frac{1}{n}X^T\left(Y - X\beta^* - \sqrt{n}\theta^*\right)$$

and

$$\partial_\theta\left(Q^{\frac{1}{2}}\right)(\beta^*,\theta^*) = -\frac{1}{2}Q^{*-\frac{1}{2}} \cdot \frac{1}{\sqrt{n}}\left(Y - X\beta^* - \sqrt{n}\theta^*\right).$$

Therefore,

$$\widehat{Q}^{\frac{1}{2}} - Q^{*\frac{1}{2}} \geq -\frac{\frac{1}{n}\sum_{j=1}^{n}(y_j - X_j^T\beta^* - \sqrt{n}\theta_j^*)X_j^T(\widehat{\beta}-\beta^*)}{2Q^{*\frac{1}{2}}}$$
$$- \frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(y_j - X_j^T\beta^* - \sqrt{n}\theta_j^*)(\widehat{\theta}_j-\theta_j^*)}{2Q^{*\frac{1}{2}}}$$
$$\geq -\sigma\frac{|\xi^\top X\Delta^\beta|/n}{2Q(\beta^*,\theta^*)^{\frac{1}{2}}} - \sigma\frac{\frac{1}{\sqrt{n}}\max_{j\geq o'}(|\xi|_{(j)}/\mu_j)\left\|\widehat{\theta}-\theta^*\right\|_\mu}{2Q^{*\frac{1}{2}}}$$
$$\geq -\sigma\frac{(\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_\mu)/10 + 2\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma}{2Q^{*\frac{1}{2}}},$$

where the last inequality follows from (5) combined with that fact that

$$\max_{j\geq o'}(|\xi|_{(j)}/(\sqrt{n}\mu_j)) \leq 1/10.$$

We conclude that

$$2Q^{*\frac{1}{2}}\left(\widehat{Q}^{\frac{1}{2}} - Q^{*\frac{1}{2}}\right)/\sigma \geq -(\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_\mu)/10 - 2\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma.$$

Recall that $\|\xi\|_2^2 \geq n/10$ on event $\mathscr{E}$. Hence, in the view of the inequality (5), we see that

$$\left(\widehat{Q}^{\frac{1}{2}} - Q^{*\frac{1}{2}}\right) \geq -1/8\|\Delta^\beta\|_\lambda - 1/8\|\Delta^\theta\|_\mu - 2\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma.$$

Combining the upper and the lower bounds above, we deduce the following result:

$$\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_\mu \leq 4\sqrt{\frac{\sum_{i=1}^{s}\lambda_i^2}{\kappa(s)} + \frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma + 4\sqrt{\sum_{i=1}^{o'}\mu_i^2}\|\Delta^\theta\|_2,$$

as claimed. ∎

We are ready to proceed with the proof of the main result. Since the loss function $L(\beta,\theta)$ defined in (2) is convex, the first-order optimality conditions ensure the existence of $\widehat{v}\in\partial\left\|\widehat{\beta}\right\|_1$, $\widehat{u}\in\partial\left\|\widehat{\theta}\right\|_1$ such that $\widehat{v}^T\widehat{\beta} = \left\|\widehat{\beta}\right\|_\lambda$, $\widehat{u}^T\widehat{\theta} = \left\|\widehat{\theta}\right\|_\mu$, and

$$0 = -\frac{\frac{1}{n}X^T(Y - X\widehat{\beta} - \sqrt{n}\widehat{\theta})}{2\widehat{Q}^{\frac{1}{2}}} + \widehat{v},$$

$$0 = -\frac{\frac{1}{\sqrt{n}}(Y - X\widehat{\beta} - \sqrt{n}\widehat{\theta})}{2\widehat{Q}^{\frac{1}{2}}} + \widehat{u}$$

under the assumption that $\widehat{Q}\neq 0$. The two equations above are equivalent to

$$[X^{(n)}, I_n]^T(Y^{(n)} - X^{(n)}\widehat{\beta} - \widehat{\theta}) = 2\widehat{Q}^{\frac{1}{2}}\left[(\widehat{v})^T, (\widehat{u})^T\right]^T.$$

Note that when $\widehat{Q}=0$, we have that $\widehat{\xi}=0$ and hence (B.3) is still valid. Next, recall that $Y^{(n)}=X^{(n)}\widehat{\beta}+\widehat{\theta}+\sigma\xi^{(n)}$, so by (B.3),

$$[X^{(n)},I_n]^T[X^{(n)},I_n]\Delta=\sigma[X^{(n)},I_n]^T\xi^{(n)}-2\widehat{Q}^{\frac{1}{2}}\left[(\widehat{v})^T,(\widehat{u})^T\right]^T.$$

Multiplying both sides of this equation by Δ^T from the left, we get

$$\left\|X^{(n)}\Delta^\beta+\Delta^\theta\right\|_2^2 \;=\; \sigma(\Delta^\beta)^T(X^{(n)})^T\xi^{(n)} \;+\; \sigma(\Delta^\theta)^T\xi^{(n)} \;-\; 2\widehat{Q}^{\frac{1}{2}}(\Delta^\beta)^T\widehat{v} \;-\; 2\widehat{Q}^{\frac{1}{2}}(\Delta^\theta)^T\widehat{u}.$$

Note that $\max_i(|\widehat{v}|_{(i)}/\lambda_i)\le 1$ and $\widehat{v}^T\widehat{\beta}=\left\|\widehat{\beta}\right\|_\lambda$, hence

$$-(\Delta^\beta)^T\widehat{v}=(\beta^*-\widehat{\beta})^T\widehat{v}=(\beta^*)^T\widehat{v}-\left\|\widehat{\beta}\right\|_1\le\|\beta^*\|_\lambda-\left\|\widehat{\beta}\right\|_\lambda.$$

Similarly, we can show that

$$-(\Delta^\theta)^T\widehat{u}\le\|\theta^*\|_\mu-\left\|\widehat{\theta}\right\|_\mu.$$

Combining these bounds with equation (B.3) and noticing that $(Q^*)^{\frac{1}{2}}=\frac{\sigma}{\sqrt{2n}}\|\xi\|_2$, we deduce that on event $\mathscr{E}$,

$$\left\|X^{(n)}\Delta^\beta+\Delta^\theta\right\|_2^2\le\sigma(\|\Delta^\beta\|_\lambda+\|\Delta^\theta\|_\mu)/2+\sigma\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma \\ +2\widehat{Q}^{\frac{1}{2}}\left(\|\beta^*\|_\lambda-\left\|\widehat{\beta}\right\|_\lambda\right)+2\widehat{Q}^{\frac{1}{2}}\left(\|\theta^*\|_\mu-\left\|\widehat{\theta}\right\|_\mu\right).$$

For brevity, set $x=\left\|X^{(n)}\Delta^\beta+\Delta^\theta\right\|_2$. Note that

$$\hat{Q}^{1/2}=\frac{1}{\sqrt{2n}}\|Y-X\hat{\beta}-\sqrt{n}\hat{\theta}\|_2=\frac{1}{\sqrt{2}}\|\xi^*/\sqrt{n}-X^{(n)}\Delta^\beta-\Delta^\theta\|_2.$$

Since $(Q^*)^{1/2}=\frac{1}{\sqrt{2n}}\|\xi^*\|_2$, we deduce that

$$\hat{Q}^{1/2}\le(Q^*)^{1/2}+x/\sqrt{2}.$$

Therefore, we derive using the inequality (B.3) that

$$x^2\le(5\sigma+x)(\|\Delta^\beta\|_\lambda+\|\Delta^\theta\|_\mu)+\sigma\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma,$$

where in the last step we used the bound $\|u\|_\lambda-\|v\|_\lambda\le\|u-v\|_\lambda$ that holds for any vectors u and v. Lemma B.2 implies that

$$\|\Delta^\beta\|_\lambda+\|\Delta^\theta\|_\mu\le 4\sqrt{\frac{\sum_{i=1}^s\lambda_i^2}{\kappa(s)}+\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma+4\sqrt{\sum_{i=1}^{o'}\mu_i^2}\|\Delta^\theta\|_2.$$

Using Proposition B.1 and the condition on the sample size n, we have that

$$x^2 \geq (\|\Delta^\beta\|_\Sigma^2 + \|\Delta^\theta\|_2^2)/16,$$

and that

$$\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_\mu \leq x/2.$$

Combining these bounds with (B.3), we get

$$(\|\Delta^\beta\|_\Sigma^2 + \|\Delta^\theta\|_2^2)/16 \leq x^2 \leq 10\sigma\left(\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)} + \frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma + \sqrt{\sum_{i=1}^{o'} \mu_i^2}\|\Delta^\theta\|_2\right).$$

Therefore,

$$\left\|\Delta^\beta\right\|_\Sigma^2 + \|\Delta^\theta\|_2^2 \leq 100\sigma^2\left(\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)} + \frac{\log(1/\delta)}{n} + \sum_{i=1}^{o'} \mu_i^2\right).$$

Moreover,

$$\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_\mu \leq 100\sigma\left(\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)} + \frac{\log(1/\delta)}{n} + \sum_{i=1}^{o'} \mu_i^2\right)$$

which concludes the proof.

Remark B.1. *Direct computation shows that*

$$\left|(Q^*)^{\frac{1}{2}} - (\widehat{Q})^{\frac{1}{2}}\right| = \frac{1}{\sqrt{2}}\left|\left\|Y - X^{(n)}\beta^* - \theta^*\right\|_2 - \left\|Y - X^{(n)}\widehat{\beta} - \widehat{\theta}\right\|_2\right| \leq \frac{1}{\sqrt{2}}\left\|X^{(n)}\Delta^\beta + \Delta^\theta\right\|_2,$$

from which we derive that

$$\left|(Q^*)^{\frac{1}{2}} - (\widehat{Q})^{\frac{1}{2}}\right| \leq (Q^*)^{\frac{1}{2}}/20.$$

This shows that under the assumptions of the lemma, $\sqrt{\widehat{Q}}$ can be close to $\sqrt{Q^*}$. This fact will be important in the next proof.

Finally, we present the proof of the error bound for the estimator $\widehat{\beta}$ only, as opposed to the vector $(\widehat{\beta}, \widehat{\theta})$. The main idea of the proof, first used in [9], is to treat θ^* as a "nuisance parameter" and repeat parts of the previous argument. Recall the key notation:

- $\widehat{Q} := Q(\widehat{\beta}, \widehat{\theta})$ and $Q^* := Q(\beta^*, \theta^*)$;
- $A^{(n)} := \frac{1}{\sqrt{n}}A$, whenever A is a number, vector or matrix;
- $\widehat{\xi} := Y - X\widehat{\beta} - \sqrt{n}\widehat{\theta}$;
- Also, we note that $\widehat{Q} = \frac{1}{2n}\left\|\widehat{\xi}\right\|_2^2$ and $Q^* = \frac{1}{2n}\|\xi\|_2^2$.

Since

$$\widehat{\beta} \in \operatorname*{argmin}_\beta\left\{\frac{1}{\sqrt{2n}}\left\|Y - X\beta - \sqrt{n}\widehat{\theta}\right\|_2 + \|\beta\|_\lambda\right\},$$

there exists $\widehat{v} \in \partial \left\|\widehat{\beta}\right\|_1$ with $\widehat{v}^T\widehat{\beta} = \left\|\widehat{\beta}\right\|_\lambda$ such that

$$-\frac{1}{n}X^T(Y - X\widehat{\beta} - \sqrt{n}\widehat{\theta}) + 2(\widehat{Q})^{\frac{1}{2}}\widehat{v} = 0.$$

It implies, together with the identity $Y^{(n)} = X^{(n)}\widehat{\beta} + \widehat{\theta} + \sigma\xi^{(n)}$, that

$$(X^{(n)})^T\left(X^{(n)}\Delta^\beta + \Delta^\theta - \sigma\xi^{(n)}\right) + 2(\widehat{Q})^{\frac{1}{2}}\lambda_s\widehat{v} = 0.$$

Multiplying both sides from the left by $\left(\Delta^\beta\right)^T$, we get that

$$\left\|X^{(n)}\Delta^\beta\right\|_2^2 = -\left\langle X^{(n)}\Delta^\beta, \Delta^\theta\right\rangle + \left\langle X^{(n)}\Delta^\beta, \sigma\xi^{(n)}\right\rangle - 2(\widehat{Q})^{\frac{1}{2}}\left\langle \Delta^\beta, \widehat{v}\right\rangle.$$

Recall that

$$-\left\langle \Delta^\beta, \widehat{v}\right\rangle \le \|\beta^*\|_\lambda - \left\|\widehat{\beta}\right\|_\lambda \le 2\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)}}\|\Delta^\beta\|_\Sigma - \|\Delta^\beta\|_\lambda/2,$$

which together with remark B.1 implies that

$$\frac{1}{2}\sigma \le (\widehat{Q})^{\frac{1}{2}} \le \frac{3}{2}\sigma.$$

Moreover, from the inequality (2.3) we see that

$$\left\langle X^{(n)}\Delta^\beta, \xi^{(n)}\right\rangle \le \|\Delta^\beta\|_\lambda/20 + \sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma.$$

Combining these results with the relation (B.3), we deduce that on the event $\mathscr{E}$,

$$\left\|X^{(n)}\Delta^\beta\right\|_2^2 \le -\left\langle X^{(n)}\Delta^\beta, \Delta^\theta\right\rangle + \sigma\left(\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma + 2\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)}}\|\Delta^\beta\|_\Sigma - \|\Delta^\beta\|_\lambda/2\right).$$

Inequality (2.3) yields that

$$-\left\langle X^{(n)}\Delta^\beta, \Delta^\theta\right\rangle \le \|\Delta^\beta\|_\lambda\|\Delta^\theta\|_2/10 + \|\Delta^\theta\|_\lambda\|\Delta^\beta\|_\Sigma/10 + C\sqrt{\frac{\log(1/\delta)}{n}}\|\Delta^\beta\|_\Sigma\|\Delta^\theta\|_2.$$

Applying (2.3), we deduce the inequality

$$\left\|X^{(n)}\Delta^\beta\right\|_2^2 \ge \frac{1}{2}\|\Delta^\beta\|_\Sigma^2 - \|\Delta^\beta\|_\lambda^2/4.$$

Since $\|\Delta^\beta\|_\lambda + \|\Delta^\theta\|_2 \ll 1$,

$$\frac{1}{2}\|\Delta^\beta\|_\Sigma^2 \le \sigma\left(\sqrt{\frac{\log(1/\delta)}{n}} + 2\sqrt{\frac{\sum_{i=1}^s \lambda_i^2}{\kappa(s)}} + \|\Delta^\theta\|_\lambda/10\right)\|\Delta^\beta\|_\Sigma.$$

Therefore,

$$\|\Delta^\beta\|_\Sigma^2 \le 100\sigma^2\left(\frac{\log(1/\delta)}{n}+\frac{\sum_{i=1}^s\lambda_i^2}{\kappa(s)}+\|\Delta^\theta\|_\lambda^2/2\right).$$

Observe that

$$\begin{aligned}\|\Delta^\theta\|_\lambda^2/2 &\le \left(\sum_{j\ge o'}\lambda_j|\Delta^\theta|_{(j)}\right)^2+\|\Delta^\theta\|_2^2\sum_{j=1}^{o'}\lambda_j^2\\ &\le \left(\sum_{j\ge o'}(\lambda_j/\mu_j)\mu_j|\Delta^\theta|_{(j)}\right)^2+10\sum_{j=1}^{o'}\lambda_j^2(1+\sum_{j=1}^{o'}\mu_j^2)\\ &\le \max_{j\ge o'}(\lambda_j^2/\mu_j^2)(\|\Delta^\theta\|_\mu^2+20(\sum_{j=1}^{o'}\mu_j^2)^2)+10\sum_{j=1}^{o'}\lambda_j^2,\end{aligned}$$

where we used the inequality $\sum_{j=1}^{o'}\lambda_j^2\le 2o'\lambda_{o'}^2\le 2(\lambda_{o'}^2/\mu_{o'}^2)\sum_{j=1}^{o'}\mu_j^2$. Since $\lambda_j\le\mu_j$, we conclude that

$$\|\Delta^\beta\|_\Sigma^2 \le 100\sigma^2\left(\frac{\log(1/\delta)}{n}+\frac{\sum_{i=1}^s\lambda_i^2}{\kappa(s)}+\max_{j\ge o'}(\lambda_j^2/\mu_j^2)(\sum_{j=1}^{o'}\mu_j^2)^2\right).$$

B.4. Proof of Theorems 2.1 and 2.2. Recall that event $\mathscr{E}$ and the properties of sub-Gaussian designs expressed via the inequalities (2.3), (2.3), (2.3) hold for a given δ with probability at least $1-3\delta-5\exp(-co')$. Observe that $\sum_{i=1}^s\lambda_i^2\le Cs\log(ep/s)/n$. For the case of fixed thresholds $\mu_i=\frac{C}{\sqrt{n}}\left(\frac{n}{m}\right)^{1/\tau}$, where $m=\log(1/\delta)$, we get that

$$\max_{j\ge o'}\left(\lambda_j^2/\mu_j^2\right)\left(\sum_{j=1}^{o'}\mu_j^2\right)^2 = C\left(\frac{o'}{n}\right)^2\log(n/o')\left(\frac{n}{m}\right)^{2/\tau}=C\left(\frac{o'}{n}\right)^{2-2/\tau}\log(n/o')\left(\frac{o'}{m}\right)^{2/\tau}.$$

Choosing $o'=o+m$ with $m=\log(1/\delta)$, we deduce that

$$\max_{j\ge o'}(\lambda_j^2/\mu_j^2)(\sum_{j=1}^{o'}\mu_j^2)^2\le C\left(\left(\frac{o}{n}\right)^{2-2/\tau}\log(n/o)(1+(o/m)^{2/\tau})+\frac{\log(1/\delta)}{n}\right).$$

Theorem 2.5 immediately yields that with probability at least $1-8\delta$,

$$\|\Delta^\beta\|_\Sigma^2\le C\sigma^2\left(\frac{\log(1/\delta)}{n}+\frac{s\log(ep/s)/n}{\kappa(s)}+\left(\frac{o}{n}\right)^{2-2/\tau}\log(n/o)(1+(o/\log(1/\delta))^{2/\tau})\right).$$

This completes the proof of Theorem 2.1.

For the case of the adaptive threshold $\mu_i=\frac{C}{\sqrt{n}}\left(\frac{n}{i}\right)^{1/\tau}$, for any δ we can choose $o'=o+m$ with $m=\log(1/\delta)$ so that

$$\max_{j\ge o'}\left(\lambda_j^2/\mu_j^2\right)\left(\sum_{j=1}^{o'}\mu_j^2\right)^2\le C\left(\frac{\sum_{i=1}^{o'}(n/i)^{2/\tau}}{n}\right)^2\le C\frac{(\sum_{i=1}^{o'}(1/i)^{2/\tau})^2}{n^{2-4/\tau}}.$$

Hence, for $\tau > 2$ we have

$$\max_{j \geq o'} (\lambda_j^2/\mu_j^2)(\sum_{j=1}^{o'} \mu_j^2)^2 \leq C(o'/n)^{2-4/\tau}.$$

In view of Theorem 2.5,

$$\|\Delta^\beta\|_\Sigma^2 \leq C\sigma^2 \left(\frac{\log(1/\delta)}{n} + \frac{s\log(ep/s)/n}{\kappa(s)} + (o + \log(1/\delta)/n)^{2-4/\tau} \right)$$

with probability at least $1 - 8\delta$. Finally, if the noise is sub-Gaussian, then

$$\max_{j \geq o'} (\lambda_j^2/\mu_j^2)(\sum_{j=1}^{o'} \mu_j^2)^2 \leq C \left(\frac{o'\log(n/o')}{n} \right)^2,$$

and we get that

$$\|\Delta^\beta\|_\Sigma^2 \leq C\sigma^2 \left(\frac{\log(1/\delta)}{n} + \frac{s\log(ep/s)/n}{\kappa(s)} + \left(\frac{o\log(n/o)}{n} \right)^2 \right),$$

again with probability at least $1 - 8\delta$. The last inequality holds since

$$\frac{\log(1/\delta)}{n} \geq \left(\frac{\log(1/\delta)\log(n/\log(1/\delta))}{n} \right)^2$$

whenever $\log(1/\delta)$ is smaller than n. This completes the proof of Theorem 2.2.