# Adaptive robust estimation in sparse vector model

Olivier Collier, Laëtitia Comminges, A. Tsybakov

Submitted

# Some effects in adaptive robust estimation under sparsity

**O. Collier, L. Comminges, A. B. Tsybakov**

Modal'X, UPL, Univ. Paris Nanterre, F92000 Nanterre France
CEREMADE, Université Paris-Dauphine, PSL Research University, 75016 Paris, France
CREST, ENSAE, 5 avenue Henry Le Chatelier, 91764 Palaiseau, France

*Abstract.* Adaptive estimation in the sparse mean model and in sparse regression exhibits some interesting effects. This paper considers estimation of a sparse target vector, of its $\ell_2$-norm and of the noise variance in the sparse linear model. We establish the optimal rates of adaptive estimation when adaptation is considered with respect to the triplet "noise level – noise distribution – sparsity". These rates turn out to be different from the minimax non-adaptive rates when the triplet is known. A crucial issue is the ignorance of the noise level. Moreover, knowing or not knowing the noise distribution can also influence the rate. For example, the rates of estimation of the noise level can differ depending on whether the noise is Gaussian or sub-Gaussian without a precise knowledge of the distribution. Estimation of noise level in our setting can be viewed as an adaptive variant of robust estimation of scale in the contamination model, where instead of fixing the "nominal" distribution in advance we assume that it belongs to some class of distributions. We also show that in the problem of estimation of a sparse vector under the $\ell_2$-risk when the variance of the noise in unknown, the optimal rate depends dramatically on the design. In particular, for noise distributions with polynomial tails, the rate can range from sub-Gaussian to polynomial depending on the properties of the design.
*Key words:* variance estimation, functional estimation, sparsity, robust estimation, adaptivity, sub-Gaussian noise.

## 1. INTRODUCTION

This paper considers estimation of the unknown sparse vector, of its $\ell_2$-norm and of the noise level in the sparse mean model and in the sparse linear regression model. We focus on construction of estimators that are optimally adaptive in a minimax sense with respect to the noise level, to the form of the noise distribution, and to the sparsity.

We consider first the sparse mean model defined as follows. Let the signal $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ be observed with some noise of unknown magnitude $\sigma > 0$:

$$Y_i = \theta_i + \sigma\xi_i, \quad i = 1, \dots, d.$$

The noise random variables $\xi_1, \dots, \xi_d$ are assumed to be i.i.d. and we denote by $P_\xi$ the unknown distribution of $\xi_1$. We assume throughout that the noise is zero-mean, $\mathbf{E}(\xi_1) = 0$, and that $\mathbf{E}(\xi_1^2) = 1$, which ensures identifiability of $\sigma$. We denote by $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}$ the distribution of $(Y_1, \dots, Y_d)$ when the signal is $\boldsymbol{\theta}$, the noise level is $\sigma$ and the distribution of the noise variables is $P_\xi$. We also denote by $\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}$ the expectation with respect to $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}$.

We assume that the signal $\boldsymbol{\theta}$ is $s$-sparse, *i.e.,*

$$\|\boldsymbol{\theta}\|_0 = \sum_{i=1}^{d} \mathbb{1}_{\theta_i \neq 0} \leq s,$$

where $s \in \{1, \dots, d\}$ is an integer. Set $\Theta_s = \{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_0 \le s\}$. Our aim is to estimate $\sigma^2$ and the $\ell_2$-norm

$$\|\boldsymbol{\theta}\|_2 = \Big(\sum_{i=1}^d \theta_i^2\Big)^{1/2}.$$

In the case of standard Gaussian noise ($P_\xi = \mathcal{N}(0,1)$) and known noise level $\sigma$, a rate optimal estimator of the $\ell_2$-norm of $\boldsymbol{\theta}$ on the class $\Theta_s$ was found in Collier, Comminges and Tsybakov [4]. In particular, it was proved that the estimator

$$\hat{N}_{\sigma,s} = \Big(\sum_{i=1}^d (Y_i^2 - \alpha_s \sigma^2)\mathbb{1}_{Y_i^2 > 2\sigma^2 \log(1+d/s^2)}\Big)^{\frac{1}{2}}, \quad \text{with } \alpha_s = \mathbf{E}\big(\xi_1^2 | \xi_1^2 > 2\log(1+d/s^2)\big),$$

satisfies the bound

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\big(\hat{N}_{\sigma,s} - \|\boldsymbol{\theta}\|_2\big)^2}{\sigma^2} \le c_0 s \log(1+d/s^2), \qquad \forall s \le \sqrt{d}, \tag{1}$$

for some positive constant $c_0$, and the rate in (1) cannot be improved for $s \le \sqrt{d}$. For $s > \sqrt{d}$, the simple estimator

$$\big(\sum_{i=1}^d (Y_i^2 - \sigma^2)\big)^{\frac{1}{2}}$$

is rate-optimal. Thus, in the whole range $1 \le s \le d$, if we define the minimax risk by the infimum of the quantities in the left-hand side of (1) over all estimators, the minimax rate has the form

$$\phi_{\mathcal{N}(0,1),\mathsf{norm}} = \min\{\sqrt{s\log(1+d/s^2)}, d^{1/4}\}.$$

The estimator $\hat{N}_{\sigma,s}$ depends on $\sigma$, $s$, and crucially uses the fact that the variables $\xi_i$ are standard Gaussian. A natural question is whether one can construct an adaptive estimator independent of these parameters/assumptions that attains the same rate. We show that this is not possible. Furthermore, we find the best rate that can be achieved by adaptive estimators. The deterioration of the rate due to adaptation turns out to depend on the assumptions on the noise distribution $P_\xi$. We consider two types of assumptions: either the noise belongs to a class of sub-Gaussian distributions $\mathcal{G}_\tau$, *i.e.,* for some $\tau > 0$,

$$P_\xi \in \mathcal{G}_\tau \quad \text{iff} \quad \mathbf{E}(\xi_1) = 0,\ \mathbf{E}(\xi_1^2) = 1 \text{ and } \quad \forall t \ge 2,\ \mathbf{P}\big(|\xi_1| > t\big) \le 2e^{-(t/\tau)^2}, \tag{2}$$

or to a class of distributions with polynomially decaying tails $\mathcal{P}_{a,\tau}$, *i.e.,* for some $\tau > 0$ and $a \ge 2$,

$$P_\xi \in \mathcal{P}_{a,\tau} \quad \text{iff} \quad \mathbf{E}(\xi_1) = 0,\ \mathbf{E}(\xi_1^2) = 1 \text{ and } \quad \forall t \ge 2,\ \mathbf{P}\big(|\xi_1| > t\big) \le \Big(\frac{\tau}{t}\Big)^a. \tag{3}$$

In particular, the rate of estimation of the $\ell_2$-norm functional deteriorates dramatically if $P_\xi$ belongs to $\mathcal{P}_{a,\tau}$ as compared to the sub-Gaussian case $P_\xi \in \mathcal{G}_\tau$.

The problem of estimation of the variance $\sigma^2$ exhibits similar effects. When the variables $\xi_i$ are standard Gaussian, a suitably rescaled median of the squared observations achieves the rate $\phi_{\mathcal{N}(0,1)}(s,d) = \max\Big(\frac{1}{\sqrt{d}}, \frac{s}{d}\Big)$ as shown in Section 2.1 below. However, such an estimator fails to work on the class of sub-Gaussian distributions $\mathcal{G}_\tau$, and we show that adaptive estimation

of $\sigma^2$ on this class is only possible with slower rate. We obtain similar conclusions for the class $\mathcal{P}_{a,\tau}$ where the best rate of adaptive estimation is even slower and crucially depends on $a$.

Finally, we consider extensions of these results to the high-dimensional linear regression model. We propose estimators of $\sigma^2$ and $\|\boldsymbol{\theta}\|_2$ based on the Square-Root Slope estimator of $\boldsymbol{\theta}$. Using the methods of [1] and [6] we show that these estimators are simultaneously adaptive to $\sigma$, $s$, and to $P_\xi$ in the class $\mathcal{G}_\tau$.

We conclude this section by a discussion of related work. Chen, Gao and Ren [3] explore the problem of robust estimation of variance and of covariance matrix under Hubers's contamination model. As explained in Section 2.1 below, this problem is very similar to estimation of noise level in our setting. The main difference is that instead of fixing in advance the Gaussian nominal distribution of the contamination model we assume that it belongs to a class of distributions, such as (2) or (3). Therefore, one can view the part of the current paper dealing with noise level estimation as a variation on robust estimation of scale where, in contrast to the classical setting, we are interested in adaptation to the unknown nominal law. Another aspect of robust estimation of scale is analyzed by Minsker and Wei [11]. They consider classes of distributions similar to $\mathcal{P}_{a,\tau}$ rather than the contamination model. Their main aim is to construct estimators having sub-Gaussian deviations under weak moment assumptions. Our setting is different in that we consider the sparsity class $\Theta_s$ of vectors $\boldsymbol{\theta}$ and the rates that we obtain depend on $s$. We also mention the recent paper by Golubev and Krymova [8] that deals with estimation of variance in linear regression in a framework that does not involve sparsity.

The part of this paper dealing with estimation of the $\ell_2$-norm studies the same questions as in Collier, Comminges, Tsybakov and Verzelen [5] where the problem of estimation of linear functional $L(\boldsymbol{\theta}) = \sum_{i=1}^d \theta_i$ was considered. In that case, adaptation is less costly than for the $\ell_2$-norm. A minimax rate-optimal estimator of $L(\boldsymbol{\theta})$ with known $s$ and $\sigma$ and standard Gaussian $\xi_i$ can be found in [4] and has the form:

$$\hat{L}_{\sigma,s} = \sum_{i=1}^d Y_i \mathbb{1}_{Y_i^2 > 2\sigma^2 \log(1+d/s^2)}. \tag{4}$$

As shown in [5], adaptive estimation of $L(\boldsymbol{\theta})$ can be achieved with the rate, which is almost the same as the minimax non-adaptive rate. For this, it is enough to replace the unknown $s$ and $\sigma$ in (4) by some data-dependent quantities. In particular, for $\sigma$ it can be a rough overestimator. Furthermore, it is straightforward to extend the method of [5] to sub-Gaussian noise distributions. However, using such an approach for adaptive estimation of the $\ell_2$-norm meets difficulties. Indeed, replacing the unknown parameters in $\hat{N}_{\sigma,s}$ by some statistics is more problematic since along with an estimator of $\sigma$ one needs a very accurate estimate of the coefficient $\alpha_s$. Therefore, we will proceed in a different way.

## 2. ESTIMATING THE VARIANCE OF THE NOISE

### 2.1 Sample median estimator

In the sparse setting when $\|\boldsymbol{\theta}\|_0$ is small, estimation of the noise level can be viewed as a problem of robust estimation of scale. Indeed, our aim is to recover the second moment of $\sigma\xi_1$ but the sample second moment cannot be used as an estimator because of the presence of a small number of outliers $\theta_i \neq 0$. The models in robustness and sparsity problems are essentially equivalent but the questions of interest are different. When robust estimation of $\sigma$ is considered, the object of interest is the pure noise component of the sparsity model while the non-zero $\theta_i$ that are of major interest in the sparsity model play a role of nuisance.

In the context of robustness, it is known that the estimator based on sample median can be successfully applied. This is a special case of $M$-estimator of scale (*cf.* [9]) defined as

$$\hat{\sigma}^2_{\mathsf{med}} = \frac{\hat{\gamma}}{\beta} \tag{5}$$

where $\hat{\gamma}$ is the sample median of $(Y_1^2, \dots, Y_d^2)$, that is

$$\hat{\gamma} \in \arg\min_{x>0} \big|F_d(x) - 1/2\big|,$$

and $\beta$ is the median of the distribution of $\xi_1^2$. Here, $F_d$ denotes the empirical c.d.f of $(Y_1^2, \dots, Y_d^2)$. It is easy to see that if $\xi_1$ has a symmetric distribution, then

$$\beta = \big(F^{-1}(3/4)\big)^2 \tag{6}$$

where $F$ denotes the c.d.f. of $\xi_1$.

The following proposition specifies the rate of convergence of the estimator $\hat{\sigma}^2_{\mathsf{med}}$.

PROPOSITION 1. *Let $\xi_1$ have a symmetric c.d.f. $F$ with positive density, and let $\beta$ be given by* (6)*. There exist constants $c_* > 0$ and $c'_* > 0$ depending only on $F$ such that for $t > 0$ and integers $s, d$ satisfying $\sqrt{\frac{t}{d}} + \frac{s}{d} \le c'_*$ we have*

$$\sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},F,\sigma} \left( \Big| \frac{\hat{\sigma}^2_{\mathsf{med}}}{\sigma^2} - 1 \Big| \ge c_* \left( \sqrt{\frac{t}{d}} + \frac{s}{d} \right) \right) \le 2e^{-t}.$$

When the noise is Gaussian, Chen, Gao and Ren [3] study a generalization of the estimator (5) based on Tukey depth. A model in [3] related to our setting is the contamination model where the observations are Gaussian $\mathcal{N}(0, \sigma^2)$ with probability $1 - \varepsilon$ and arbitrary with probability $\varepsilon \in (0, 1)$. If $\varepsilon = s/d$, this can be compared with our model. But in contrast to [3], in our case the number of outliers $s$ is fixed rather than random. Also, the minimax risk in [3] is different from ours, since the loss is not rescaled by $\sigma^2$, and $\sigma^2$ is assumed uniformly bounded. Modulo these differences, the case $\xi_1 \sim \mathcal{N}(0, 1)$ of Proposition 1 can be viewed as an analog of Theorem 3.1 in [3].

In particular, when $\xi_1 \sim \mathcal{N}(0, 1)$, Proposition 1 shows that the rate of convergence of $\hat{\sigma}^2_{\mathsf{med}}$ in probability is

$$\phi_{\mathcal{N}(0,1)}(s, d) := \max\left( \frac{1}{\sqrt{d}}, \frac{s}{d} \right). \tag{7}$$

Note also that, akin to the results in [3], Proposition 1 does not allow to derive rates of convergence in expectation because $t$ is assumed to be bounded from above.

The main drawback of the estimator $\hat{\sigma}^2_{\mathsf{med}}$ is the dependence on the parameter $\beta$. It reflects the fact that the estimator is tailored for a given and known distribution $F$ of noise, for example, the standard Gaussian distribution. Furthermore, as shown below, the rate (7) cannot be achieved when it is only known that $F$ belongs to the class of sub-Gaussian distributions.

### 2.2 Distribution-free variance estimator

Instead of using one particular quantile, like the median in the previous section, we propose to estimate $\sigma^2$ by an integral over all quantiles, which allows us to avoid considering distribution-dependent quantities like (6).

Indeed, with the notation $q_\alpha = G^{-1}(1-\alpha)$ where $G$ is the c.d.f. of $(\sigma\xi_1)^2$ and $0 < \alpha < 1$, the variance of the noise can be expressed as

$$\sigma^2 = \mathbf{E}(\sigma\xi_1)^2 = \int_0^1 q_\alpha \, d\alpha.$$

Discarding the higher order quantiles that are dubious in the presence of outliers and replacing $q_\alpha$ by the empirical quantile $\hat{q}_\alpha$ of level $\alpha$ we obtain the following estimator

$$\hat{\sigma}^2 = \int_0^{1-s/d} \hat{q}_\alpha \, d\alpha = \frac{1}{d}\sum_{k=1}^{d-s} Y_{(k)}^2, \tag{8}$$

where $Y_{(1)}^2 \le \ldots \le Y_{(d)}^2$ are the ordered values of the squared observations $Y_1^2, \ldots, Y_d^2$. Note that $\hat{\sigma}^2$ is an $L$-estimator, *cf.* [9]. Also, up to a constant factor, $\hat{\sigma}^2$ coincides with the statistic used in Collier, Comminges and Tsybakov [4] .

The following theorem provides an upper bound on the risk of the estimator $\hat{\sigma}^2$ under the assumption that the noise is sub-Gaussian. Set

$$\phi_{\mathsf{sg}}(s,d) = \max\left(\frac{1}{\sqrt{d}}, \frac{s}{d}\log\left(\frac{ed}{s}\right)\right).$$

THEOREM 1. *Let $\tau$ be a positive real number, and let $s, d$ be integers satisfying $1 \le s < d/2$. Then, the estimator $\hat{\sigma}^2$ defined in (8) satisfies*

$$\sup_{P_\xi \in \mathcal{G}_\tau} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\hat{\sigma}^2 - \sigma^2\right)^2}{\sigma^4} \le c_1 \phi_{\mathsf{sg}}^2(s,d)$$

*where $c_1 > 0$ is a constant depending only on $\tau$.*

Note that an assumption of the type $s < d/2$ is natural in the context of variance estimation. Indeed, we need $s < cd$ for some $0 < c < 1$ since $\sigma$ is not identifiable if $s = d$.

The next theorem establishes the performance of the noise level estimator in the case of distributions with polynomially decaying tails. Set

$$\phi_{\mathsf{pol}}(s,d) = \max\left(\frac{1}{\sqrt{d}}, \left(\frac{s}{d}\right)^{1-\frac{2}{a}}\right).$$

THEOREM 2. *Let $\tau > 0, a > 4$, and let $s, d$ be integers satisfying $1 \le s < d/2$. Then, the estimator $\hat{\sigma}^2$ defined in (8) satisfies*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\hat{\sigma}^2 - \sigma^2\right)^2}{\sigma^4} \le c_2 \phi_{\mathsf{pol}}^2(s,d),$$

*where $c_2 > 0$ is a constant depending only on $\tau$ and $a$.*

We assume here that the noise distribution has a moment of order greater than 4, which is close to the minimum requirement since we deal with the squared error of a quadratic function of the observations.

We now state the lower bounds matching the results of Theorems 1 and 2. These lower bounds are obtained in more generality than the upper bounds since they cover a large class of loss functions rather than only the squared loss.

We denote by $\mathcal{L}$ the set of all monotone non-decreasing functions $\ell : [0, \infty) \to [0, \infty)$ such that $\ell(0) = 0$ and $\ell \not\equiv 0$.

THEOREM 3. *Let $\tau$ be a positive real number, and let $s, d$ be integers satisfying $1 \le s \le d$. Let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{G}_\tau} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\, \ell\Big(c_3(\phi_{\mathsf{sg}}(s,d))^{-1} \Big| \frac{\hat{T}}{\sigma^2} - 1 \Big| \Big) \ge c_3',$$

*where $c_3 > 0$ is a constant depending only on $\ell(\cdot)$, $c_3' > 0$ is a constant depending only on $\ell(\cdot)$ and $\tau$, and $\inf_{\hat{T}}$ denotes the infimum over all estimators.*

Theorems 1 and 3 imply that the estimator $\hat{\sigma}^2$ is rate optimal in a minimax sense when the noise is sub-Gaussian. Interestingly, an extra logarithmic factor appears in the optimal rate when passing from the pure Gaussian distribution of $\xi_i$'s (*cf.* Proposition 1) to the class of all sub-Gaussian distributions. This factor can be seen as a price to pay for the lack of information regarding the exact form of the distribution.

Under polynomial tail assumption on the noise, we have the following minimax lower bound.

THEOREM 4. *Let $\tau > 0$, $a \ge 2$, and let $s, d$ be integers satisfying $1 \le s \le d$. Let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\, \ell\Big(c_4(\phi_{\mathsf{pol}}(s,d))^{-1} \Big| \frac{\hat{T}}{\sigma^2} - 1 \Big| \Big) \ge c_4'$$

*where $c_4 > 0$ is a constant depending only on $\ell(\cdot)$, $\tau$ and $a$, $c_4' > 0$ is a constant depending only on $\ell(\cdot)$ and $\inf_{\hat{T}}$ denotes the infimum over all estimators.*

This theorem shows that the rate obtained in Theorem 2 cannot be improved in a minimax sense.

## 3. ESTIMATION OF THE $\ell_2$-NORM

In this section, we consider the problem of estimation of the $\ell_2$-norm of a sparse vector. We first state a lower bound on the performance of any estimators of the $\ell_2$-norm when the noise level $\sigma$ is unknown and the unknown noise distribution $P_\xi$ has either sub-Gaussian or polynomially decreasing tails.

THEOREM 5. *Let $\tau > 0$, $a \ge 2$, and let $s, d$ be integers satisfying $1 \le s \le d$. Let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Set*

$$\phi_{\mathsf{sg,norm}}(s,d) = \sqrt{s \log(ed/s)}, \qquad \phi_{\mathsf{pol,norm}}(s,d) = \sqrt{s}(d/s)^{1/a}.$$

*Then*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{G}_\tau} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\, \ell\Big(C_1(\phi_{\mathsf{sg,norm}}(s,d))^{-1} \Big| \frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma} \Big| \Big) \ge C_1', \tag{9}$$

*and*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\Big(C_2(\phi_{\mathsf{pol},\mathsf{norm}}(s,d))^{-1}\Big|\frac{\hat{T}-\|\boldsymbol{\theta}\|_2}{\sigma}\Big|\Big) \ge C_2' \tag{10}$$

*where $C_1, C_2 > 0$ are constants depending only on $\ell(\cdot)$, $\tau$ and $a$, $C_1', C_2' > 0$ are constants depending only on $\ell(\cdot)$ and $\inf_{\hat{T}}$ denotes the infimum over all estimators.*

The lower bound (10) implies that the rate of estimation of the $\ell_2$-norm of a sparse vector deteriorates dramatically if the bounded moment assumption is imposed on the noise instead of the sub-Gaussian assumption.

The lower bound (9) for the indicator loss $\ell(t) = \mathbb{1}_{t\ge 1}$ is tight and it is achieved by an estimator independent of $s$ or $\sigma$, which is stated in the following theorem.

THEOREM 6. *There exist absolute constants $c \in (0,1)$, $\bar{c} > 0, c' > 0$, and an estimator $\hat{N}$ independent of $s$ and $\sigma$ and such that, for all $1 \le s \le cd$, we have*

$$\sup_{P_\xi \in \mathcal{G}_1} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\left|\frac{\hat{N}-\|\boldsymbol{\theta}\|_2}{\sigma}\right| \ge c'\phi_{\mathsf{sg},\mathsf{norm}}(s,d)\right) \le \frac{\bar{c}}{d}.$$

This theorem is a special case of Theorem 9 in Section 4 where we set $\mathbb{X}$ to be the identity matrix. The estimator $\hat{N}$ is the corresponding special case of the Square-Root Slope estimator defined in Section 4.

We now compare these results with findings in Collier, Comminges and Tsybakov [4] regarding the (nonadaptive) estimation of $\|\boldsymbol{\theta}\|_2$ when $\xi_i$ are standard Gaussian and $\sigma$ is known. It is shown in Collier, Comminges and Tsybakov [4] that in this case the optimal rate of estimation of $\|\boldsymbol{\theta}\|_2$ has the form

$$\phi_{\mathcal{N}(0,1),\mathsf{norm}} = \min\{\sqrt{s\log(1+d/s^2)}, d^{1/4}\}.$$

In particular, the following lower bound holds (*cf.* [4]):

PROPOSITION 2. *For any $\sigma > 0$ and any integers $s, d$ satisfying $1 \le s \le d$, we have*

$$\inf_{\hat{T}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\big(\hat{T}-\|\boldsymbol{\theta}\|_2\big)^2 \ge c_5\sigma^2 \min\Big\{s\log(1+d/s^2), \sqrt{d}\Big\},$$

*where $c_5 > 0$ is an absolute constant and $\inf_{\hat{T}}$ denotes the infimum over all estimators.*

We see that, in contrast to these results, in the case of unknown $\sigma$ the optimal rate $\phi_{\mathsf{sg},\mathsf{norm}}(s,d)$ does not exhibit an elbow at $s = \sqrt{d}$ between the "sparse" and "dense" regimes. Another conclusion is that, in the "dense" zone $s > \sqrt{d}$, adaptation to $\sigma$ is only possible with a significant deterioration of the rate. On the other hand, in the "sparse" zone $s \le \sqrt{d}$ the nonadaptive rate $\sqrt{s\log(1+d/s^2)}$ differs only slightly from the adaptive rate $\sqrt{s\log(1+d/s)}$ and the difference vanishes outside a small vicinity of $s = \sqrt{d}$.

We now turn to the class of distributions with polynomial tails $\mathcal{P}_{a,\tau}$, for which we have the lower bound (10) with the rate $\phi_{\mathsf{pol},\mathsf{norm}}$. Our aim now is to show that this rate is achievable when both $P_\xi \in \mathcal{P}_{a,\tau}$ and $\sigma$ are unknown. We will do it by proving a more general fact. Namely, we will show that the same rate $\phi_{\mathsf{pol},\mathsf{norm}}$ is minimax optimal not only for the estimation of the $\ell_2$-norm of $\boldsymbol{\theta}$ but also for the estimation of the whole vector $\boldsymbol{\theta}$ under the $\ell_2$-norm. To this

end, we first note that (10) obviously implies a lower bound with the same rate $\phi_{\mathsf{pol,norm}}$ for the estimation of the $s$-sparse vector $\boldsymbol{\theta}$ under the $\ell_2$-norm. We will show in the rest of this section that this lower bound is tight; the rate $\phi_{\mathsf{pol,norm}}$ appearing in (10) is achieved by an adaptive-to-$\sigma$ version of the Slope estimator $\hat{\boldsymbol{\theta}}$ when the loss is measured by $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2/\sigma$. This immediately implies the achievability of the bound (10) by the estimator $\hat{N} = \|\hat{\boldsymbol{\theta}}\|_2$ of the $\ell_2$-norm.

First, we define a preliminary estimator $\tilde{\sigma}^2$ of $\sigma^2$ that will be used in the definition of $\hat{\boldsymbol{\theta}}$. Let $\gamma \in (0, 1/2]$ be a constant that will be chosen small enough and depending only on $a$ and $\tau$. Divide $\{1, \dots, d\}$ into $m = \lfloor \gamma d \rfloor$ disjoint subsets $B_1, \dots, B_m$, each of size $\mathrm{Card}(B_i) \geq k := \lfloor d/m \rfloor \geq \frac{1}{\gamma} - 1$. Set

$$\bar{\sigma}_i^2 = \frac{1}{\mathrm{Card}(B_i)} \sum_{j \in B_i} Y_j^2, \quad i = 1, \dots, m.$$

Finally, define $\tilde{\sigma}^2$ as a median of $(\bar{\sigma}_1^2, \dots, \bar{\sigma}_m^2)$. The next proposition shows that with high probability the estimator $\tilde{\sigma}^2$ is close $\sigma^2$ to within a constant factor.

PROPOSITION 3. *Let $\tau > 0, a > 2$. There exist constants $\gamma \in (0, 1/2]$ and $C > 0$ depending only on $a$ and $\tau$ such that for any integers $s$ and $d$ satisfying $1 \leq s < \lfloor \gamma d \rfloor/4$ we have*

$$\inf_{P_\xi \in \mathcal{P}_{a,\tau}} \inf_{\sigma > 0} \inf_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\Big(1/2 \leq \frac{\tilde{\sigma}^2}{\sigma^2} \leq 3/2\Big) \geq 1 - \exp(-Cd).$$

Consider now the estimator $\hat{\boldsymbol{\theta}}$ defined as follows:

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \Big( \sum_{i=1}^d (Y_i - \theta_i)^2 + \tilde{\sigma} \|\boldsymbol{\theta}\|_* \Big). \tag{11}$$

Here, $\|\cdot\|_*$ denotes the sorted $\ell_1$-norm:

$$\|\boldsymbol{\theta}\|_* = \sum_{i=1}^d \lambda_i |\theta|_{(d-i+1)}, \tag{12}$$

where $|\theta|_{(1)} \leq \cdots \leq |\theta|_{(d)}$ are the order statistics of $|\theta_1|, \dots, |\theta_d|$, and $\lambda_1 \geq \cdots \geq \lambda_p > 0$ are tuning parameters.

THEOREM 7. *Let $\tau > 0, a > 2$. There exists a constant $\gamma \in (0, 1/2]$ such that for any integers $s, d$ satisfying $1 \leq s < \lfloor \gamma d \rfloor/4$ the following holds. Let $\hat{\boldsymbol{\theta}}$ be the estimator defined by (11) and (12) with*

$$\lambda_i = t\Big(\frac{d}{i}\Big)^{1/a}, \quad i = 1, \dots, d,$$

*and $t > 4\big((2e)^{1/a}\tau \vee 2\big)$. Then, there exist constants $c > 0$ and $c' > 0$ depending only on $a$ and $\tau$ such that*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\left( \frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2}{\sigma} \geq ct\phi_{\mathsf{pol,norm}}(s, d) \right) \leq \frac{c'}{t^a} + \exp(-d/c').$$

Note that the estimator $\hat{\boldsymbol{\theta}}$ in Theorem 7 does not require the knowledge of $\sigma$ or $s$.

The results of this section show that the optimal rate $\phi_{\mathsf{pol,norm}}$ under polynomially decaying noise is very different from the optimal rate $\phi_{\mathsf{sg,norm}}$ for sub-Gaussian noise. This phenomenon does not appear in the regression model when the design is "well spread". Indeed, Gautier and Tsybakov [7] consider sparse linear regression with unknown noise level $\sigma$ and show that the Self-Tuned Dantzig estimator achieves a sub-Gaussian rate (which differs from the optimal rate only in a log-factor) under the assumption that the noise is symmetric and has only a bounded moment of order $a > 2$. Belloni, Chernozhukov and Wang [2] show for the same model that a square-root Lasso estimator achieves analogous behavior under the assumption that the noise has a bounded moment of order $a > 2$. A crucial condition in [2] is that the design is "well spread", that is all components of the design vectors are random with positive variance. The same type of condition is needed in Gautier and Tsybakov [7] to obtain a sub-Gaussian rate. This condition is not satisfied in the sparse mean model considered here. In this model, the design is deterministic with only one non-zero component. Such a degenerate design turns out to be the worst from the point of view of the convergence rate, while the "well spread" design is the best one. An interesting general conclusion is that the optimal rate of convergence of estimators under sparsity when the noise level is unknown depends dramatically on the properties of the design. There remains a whole spectrum of possibilities between the degenerate and "well spread" designs where a variety of new rates can arise depending on the properties of the design.

## 4. EXTENSION TO LINEAR REGRESSION SETTING

A major drawback of the estimator $\hat{\sigma}^2$ proposed in Section 2.2 is its dependence on $s$. Nevertheless, it is a very simple estimator. It only requires to sort the observations, which can be done in nearly linear time.

In this section, we propose rate optimal estimators of $\sigma^2$ and $\|\boldsymbol{\theta}\|_2$ that do not depend on $s$ and do not require the exact knowledge of the noise distribution. We only assume that the noise is sub-Gaussian. Furthermore, we study the performance of the estimators in the more general context of linear regression.

Assume that we observe a vector $\boldsymbol{Y} \in \mathbb{R}^n$ satisfying

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\theta} + \sigma\boldsymbol{\xi},$$

where $\mathbb{X}$ is a known $n \times p$ non-random design matrix, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the unknown parameter with $\|\boldsymbol{\theta}\|_0 \le s$ and $\boldsymbol{\xi}$ is a noise vector with i.i.d. components $\xi_i$ such that their distribution satisfies (2). For simplicity, we assume that condition (2) holds with $\tau = 1$. We also assume that

$$\max_{j=1,\dots,p} \|\mathbb{X}_j\|_2^2 \le 1 \tag{13}$$

where $\mathbb{X}_j$ denotes the $j$-th column of $\mathbb{X}$. In this section, we denote by $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}$ the distribution of $\boldsymbol{Y}$ when the signal is $\boldsymbol{\theta}$, the noise level is $\sigma$ and the distribution of the noise variables $\xi_i$ is $P_\xi$.

We consider the Square-Root Slope estimator defined as

$$\hat{\boldsymbol{\theta}}_{\mathsf{srs}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \Big\{ \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\theta}\|_2 + \sqrt{n}\|\boldsymbol{\theta}\|_* \Big\}, \tag{14}$$

where $\|\cdot\|_*$ is the sorted $\ell_1$-norm given by (12).

Minimax optimal bounds on the risk of this estimator in our adaptive setting can be obtained by combining ideas from Bellec, Lecué and Tsybakov [1] and Derumigny [6]. We say that the design matrix satisfies a $WRE(s,c)$ condition if

$$\kappa := \inf_{\delta \in C_{WRE}(s,c)} \frac{\|\mathbb{X}\delta\|_2}{\sqrt{n}\|\delta\|_2} > 0,$$

where

$$C_{WRE}(s,c) = \Big\{\delta \in \mathbb{R}^p \,|\, \delta \neq 0, \|\delta\|_* \leq (1+c)\|\delta\|_2 \big(\sum_{i=1}^{s} \lambda_i^2\big)^{1/2}\Big\}.$$

The following proposition holds.

PROPOSITION 4. *There exist absolute positive constants $c_6, c_7$ and $c_8$ such that the following holds. Let the tuning parameters $\lambda_j$ in the definition of the norm $\|\cdot\|_*$ be chosen as*

$$\lambda_j = c_6\sqrt{\frac{\log(2p/j)}{n}}, \qquad j = 1, \dots, p.$$

*Assume that $\mathbb{X}$ satisfies the $WRE(s,20)$ condition and (13), and that*

$$\frac{s}{n}\log\Big(\frac{2ep}{s}\Big) \leq c_7\kappa^2.$$

*Then, for all $P_\xi \in \mathcal{G}_1$, $\sigma > 0$ and $\|\boldsymbol{\theta}\|_0 \leq s$, we have with $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}$-probability at least $1 - c_8(s/p)^s - c_8 e^{-n/c_8}$,*

$$\|\mathbb{X}(\hat{\boldsymbol{\theta}}_{\mathsf{srs}} - \boldsymbol{\theta})\|_2^2 \leq c_9\sigma^2 s\log(ep/s),$$
$$\|\hat{\boldsymbol{\theta}}_{\mathsf{srs}} - \boldsymbol{\theta}\|_2^2 \leq c_9\frac{\sigma^2 s}{n}\log(ep/s),$$
$$\|\hat{\boldsymbol{\theta}}_{\mathsf{srs}} - \boldsymbol{\theta}\|_* \leq c_9\frac{\sigma s}{n}\log(ep/s),$$

*where $c_9 > 0$ is a positive constant depending only on $\kappa$.*

PROOF. We follow the proof of Theorem 6.1 in Derumigny [6] but we have now sub-Gaussian noise and not the Gaussian noise as in [6]. The two minor modifications needed to make the proof work for sub-Gaussian case consist in using Theorem 9.1 from [1] instead of Lemma 7.7 from [6], and in noticing that Lemma 7.6 from [6] remains valid, with possibly different constants, when the noise is sub-Gaussian. □

We now use $\hat{\boldsymbol{\theta}}_{\mathsf{srs}}$ to define an estimator of $\sigma^2$, which is adaptive to $s$ and to the distribution of the noise. Set

$$\hat{\sigma}^2_{\mathsf{srs}} = \frac{1}{n}\|Y - \mathbb{X}\hat{\boldsymbol{\theta}}_{\mathsf{srs}}\|_2^2.$$

The following theorem establishes the rate of convergence of this estimator.

THEOREM 8. *Let the assumptions of Proposition 4 be satisfied. There exists an absolute constant $c_{10} > 0$ such that for any $t > 0$ satisfying $t/n \leq c_{10}$ we have*

$$\sup_{P_\xi \in \mathcal{G}_1}\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(\Big|\frac{\hat{\sigma}^2_{\mathsf{srs}}}{\sigma^2} - 1\Big| \geq c_{11}\Big(\sqrt{\frac{t}{n}} + \frac{s}{n}\log(ep/s)\Big)\Big)$$
$$\leq c_8((s/p)^s + e^{-n/c_8}) + (5/2)e^{-t},$$

*where $c_{11} > 0$ depends only on $\kappa$, and $c_8 > 0$ is a constant from Proposition 4.*

Note that $(s/p)^s \le c/p, s = 1, \dots, p/2$, for an absolute constant $c > 0$.

Finally, we consider the problem of estimation of the $\ell_2$-norm in the regression model. The following theorem follows immediately from Proposition 4 and the triangle inequality.

THEOREM 9. *Let the assumptions of Proposition 4 be satisfied. Let $\hat{\boldsymbol{\theta}}_{\mathsf{srs}}$ be the estimator defined in (14) and set*

$$\hat{N}_{\mathsf{srs}} = \|\hat{\boldsymbol{\theta}}_{\mathsf{srs}}\|_2.$$

*Then*

$$\sup_{P_\xi \in \mathcal{G}_1} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\left( \left| \frac{\hat{N}_{\mathsf{srs}} - \|\boldsymbol{\theta}\|_2}{\sigma} \right| \ge c_9 \sqrt{\frac{s}{n} \log(ep/s)} \right) \le c_8((s/p)^s + e^{-n/c_8}),$$

*where $c_8 > 0$ and $c_9 > 0$ are the constants from Proposition 4.*

Note that the estimator $\hat{N}_{\mathsf{srs}}$ is adaptive to $\sigma$, $s$ and to the distribution of noise in the class $\mathcal{G}_1$. By Theorem 9, this estimator achieves the rate $\sqrt{\frac{s}{n}\log(ep/s)}$, which is shown to be optimal in the case of identity matrix $\mathbb{X}$ and $n = p = d$ in Section 3.

## 5. PROOFS

### 5.1 Proof of Proposition 1

Denote by $G$ the cdf of $(\sigma\xi_1)^2$ and by $G_d$ the empirical cdf of $((\sigma\xi_i)^2 : i \notin S)$, where $S$ is the support of $\boldsymbol{\theta}$. Let $\gamma$ be the median of $G$, that is $G(\gamma) = 1/2$. By the definition of $\hat{\gamma}$,

$$|F_d(\hat{\gamma}) - 1/2| \le |F_d(\gamma) - 1/2|.$$

It is easy to check that $|F_d(x) - G_d(x)| \le s/d$ for all $x > 0$. Therefore,

$$|G_d(\hat{\gamma}) - 1/2| \le |G_d(\gamma) - 1/2| + 2s/d.$$

The DKW inequality [16, page 99], yields that $\mathbf{P}(\sup_{x\in\mathbb{R}} |G_d(x) - G(x)| \ge u) \le 2e^{-2u^2(d-s)}$ for all $u > 0$. Fix $t > 0$ satisfying the assumption of the proposition with $c_*'$ chosen smaller that $1/8$, and consider the event

$$\mathcal{A} := \left\{ \sup_{x\in\mathbb{R}} |G_d(x) - G(x)| \le \sqrt{\frac{t}{2(d-s)}} \right\}.$$

Then, $\mathbf{P}(\mathcal{A}) \ge 1 - 2e^{-t}$. On the event $\mathcal{A}$, we have

$$|G(\hat{\gamma}) - 1/2| \le |G(\gamma) - 1/2| + 2\left( \sqrt{\frac{t}{2(d-s)}} + \frac{s}{d} \right) \le 2\left( \sqrt{\frac{t}{d}} + \frac{s}{d} \right) \le \frac{1}{4}, \tag{15}$$

where the last two inequalities are due to the fact that $G(\gamma) = 1/2$ and to the assumption of the proposition with $c_*' < 1/8$. Notice that

$$|G(\hat{\gamma}) - 1/2| = |G(\hat{\gamma}) - G(\gamma)| = 2\big|F(\sqrt{\hat{\gamma}}/\sigma) - F(\sqrt{\gamma}/\sigma)\big|. \tag{16}$$

Using (15), (16) and the fact that $\gamma = \sigma^2(F^{-1}(3/4))^2$ we obtain that, on the event $\mathcal{A}$,

$$F^{-1}(5/8) \le \sqrt{\hat{\gamma}}/\sigma \le F^{-1}(7/8).$$

This and (16) imply

$$|G(\hat{\gamma}) - 1/2| \geq c_{**}\big|\sqrt{\hat{\gamma}}/\sigma - \sqrt{\gamma}/\sigma\big| = c_{**}\sqrt{\beta}\,\big|\hat{\sigma}_{\mathsf{med}}/\sigma - 1\big|.$$

where $c_{**} = 2\min_{x\in[F^{-1}(5/8),F^{-1}(7/8)]} F'(x) > 0$, and $\beta = (F^{-1}(3/4))^2$. Combining the last inequality with (15) we get that, on the event $\mathcal{A}$,

$$\big|\hat{\sigma}_{\mathsf{med}}/\sigma - 1\big| \leq 4c_{**}^{-1}\beta^{-1/2}\left(\frac{s}{d} + \sqrt{\frac{t}{d}}\right).$$

Using this inequality when $c'_*$ in the assumption of the proposition is chosen small enough we obtain the result.

### 5.2 Proof of Theorems 1 and 2

Let $\|\boldsymbol{\theta}\|_0 \leq s$ and denote by $S$ the support of $\boldsymbol{\theta}$. Note first that, by the definition of $\hat{\sigma}^2$,

$$\frac{\sigma^2}{d}\sum_{i=1}^{d-2s}\xi_{(i)}^2 \leq \hat{\sigma}^2 \leq \frac{\sigma^2}{d}\sum_{i\in S^c}\xi_i^2, \tag{17}$$

where $\xi_{(1)}^2 \leq \cdots \leq \xi_{(d)}^2$ are the ordered values of $\xi_1^2,\dots,\xi_d^2$. Indeed, the right hand inequality in (17) follows from the relations

$$\sum_{k=1}^{d-s} Y_{(k)}^2 = \min_{J:|J|=d-s}\sum_{i\in J} Y_{(i)}^2 \leq \sum_{i\in S^c} Y_{(i)}^2 = \sum_{i\in S^c}\sigma^2\xi_i^2.$$

To show the left hand inequality in (17), notice that at least $d-2s$ among the $d-s$ order statistics $Y_{(1)}^2,\dots,Y_{(d-s)}^2$ correspond to observations $Y_k$ of pure noise, *i.e.,* $Y_k = \sigma\xi_k$. The sum of squares of such observations is bounded from below by the sum of the smallest $d-2s$ values $\sigma^2\xi_{(1)}^2,\dots,\sigma^2\xi_{(d-2s)}^2$ among $\sigma^2\xi_1^2,\dots,\sigma^2\xi_d^2$.

Using (17) we get

$$\Big(\hat{\sigma}^2 - \frac{\sigma^2}{d}\sum_{i=1}^{d}\xi_i^2\Big)^2 \leq \frac{\sigma^4}{d^2}\Big(\sum_{i=d-2s+1}^{d}\xi_{(i)}^2\Big)^2,$$

so that

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(\hat{\sigma}^2 - \frac{\sigma^2}{d}\sum_{i=1}^{d}\xi_i^2\Big)^2 \leq \frac{\sigma^4}{d^2}\Big(\sum_{i=1}^{2s}\sqrt{\mathbf{E}\xi_{(d-i+1)}^4}\Big)^2.$$

Then

$$\begin{aligned}\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}(\hat{\sigma}^2-\sigma^2)^2 &\leq 2\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(\hat{\sigma}^2 - \frac{\sigma^2}{d}\sum_{i=1}^{d}\xi_i^2\Big)^2 + 2\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(\frac{\sigma^2}{d}\sum_{i=1}^{d}\xi_i^2 - \sigma^2\Big)^2\\ &\leq \frac{2\sigma^4}{d^2}\Big(\sum_{i=1}^{2s}\sqrt{\mathbf{E}\xi_{(d-i+1)}^4}\Big)^2 + \frac{2\sigma^4\mathbf{E}(\xi_1^4)}{d}.\end{aligned}$$

Now, to prove Theorem 1 it suffices to note that under the sub-Gaussian assumption (2) we have $\mathbf{E}(\xi_1^4) < \infty$ and Lemma 1 yields

$$\begin{aligned}\sum_{i=1}^{2s}\sqrt{\mathbf{E}\xi_{(d-i+1)}^4} &\leq 2s\sqrt{C}\sum_{i=1}^{2s}\log\big(ed/i\big) = 2s\sqrt{C}\log\Big(\frac{(ed)^{2s}}{(2s)!}\Big)\\ &\leq 2s\sqrt{C}\log\big(e^2d/s\big).\end{aligned}$$

To prove Theorem 2 we act analogously by using Lemma 2 and the fact that $\mathbf{E}(\xi_1^4) < \infty$ under assumption (3) with $a > 4$.

### 5.3 Proof of Theorems 3 and 4

Since we have $\ell(t) \geq \ell(A)\mathbb{1}_{t>A}$ for any $A > 0$, it is enough to prove the theorems for the indicator loss $\ell(t) = \mathbb{1}_{t>A}$.

(i) We first prove the lower bounds with the rate $1/\sqrt{d}$ in Theorems 3 and 4. Let $f_0 : \mathbb{R} \to [0, \infty)$ be a probability density with the following properties: $f_0$ is continuously differentiable, symmetric about 0, supported on $[-3/2, 3/2]$, with variance 1 and finite Fisher information $I_{f_0} = \int (f_0'(x))^2 (f_0(x))^{-1} dx$. The existence of such $f_0$ is shown in Lemma 6. Denote by $F_0$ the probability distribution corresponding to $f_0$. Since $F_0$ is zero-mean, with variance 1 and supported on $[-3/2, 3/2]$ it belongs to $\mathcal{G}_\tau$ with any $\tau > 0$ and to $\mathcal{P}_{a,\tau}$ with any $\tau > 0$ and $a > 0$. Define $\mathbf{P}_0 = \mathbf{P}_{0,F_0,1}$, $\mathbf{P}_1 = \mathbf{P}_{0,F_0,\sigma_1}$ where $\sigma_1^2 = 1 + c_0/\sqrt{d}$ and $c_0 > 0$ is a small constant to be fixed later. Denote by $H(\mathbf{P}_1, \mathbf{P}_0)$ the Hellinger distance between $\mathbf{P}_1$ and $\mathbf{P}_0$. We have

$$H^2(\mathbf{P}_1, \mathbf{P}_0) = 2\big(1 - (1 - h^2/2)^d\big) \tag{18}$$

where $h^2 = \int (\sqrt{f_0(x)} - \sqrt{f_0(x/\sigma_1)/\sigma_1})^2 dx$. By Theorem 7.6. in Ibragimov and Hasminskii [10],

$$h^2 \leq \frac{(1-\sigma_1)^2}{4} \sup_{t \in [1, \sigma_1]} I(t)$$

where $I(t)$ is the Fisher information corresponding to the density $f_0(x/t)/t$, that is $I(t) = t^{-2} I_{f_0}$. It follows that $h^2 \leq \bar{c} c_0^2/d$ where $\bar{c} > 0$ is a constant. This and (18) imply that for $c_0$ small enough we have $H(\mathbf{P}_1, \mathbf{P}_0) \leq 1/2$. Finally, choosing such a small $c_0$ and using Theorem 2.2(ii) in Tsybakov [13] we obtain

$$\begin{aligned}
&\inf_{\hat{T}} \max \Big\{ \mathbf{P}_0\Big(\Big|\hat{T} - 1\Big| > \frac{c_0}{2(1+c_0)\sqrt{d}}\Big), \mathbf{P}_1\Big(\Big|\frac{\hat{T}}{\sigma_1^2} - 1\Big| > \frac{c_0}{2(1+c_0)\sqrt{d}}\Big) \Big\} \\
&\geq \inf_{\hat{T}} \max \Big\{ \mathbf{P}_0\Big(|\hat{T} - 1| > \frac{c_0}{2\sqrt{d}}\Big), \mathbf{P}_1\Big(|\hat{T} - \sigma_1^2| > \frac{c_0}{2\sqrt{d}}\Big) \Big\} \geq \frac{1 - H(\mathbf{P}_1, \mathbf{P}_0)}{2} \geq \frac{1}{4}.
\end{aligned}$$

(ii) We now prove the lower bound with the rate $\frac{s}{d}\log(ed/s)$ in Theorem 3. It is enough to conduct the proof for $s \geq s_0$ where $s_0 > 0$ is an arbitrary absolute constant. Indeed, for $s \leq s_0$ we have $\frac{s}{d}\log(ed/s) \leq C/\sqrt{d}$ where $C > 0$ is an absolute constant and thus Theorem 3 follows already from the lower bound with the rate $1/\sqrt{d}$ proved in item (i). Therefore, in the rest of this proof we assume without loss of generality that $s \geq 32$.

We take $P_\xi = U$ where $U$ is the uniform distribution on $\{-1, 1\}$. Clearly, $U \in \mathcal{G}_\tau$. Consider i.i.d. Bernoulli variables $\delta_i$:

$$\delta_1, \dots, \delta_d \overset{iid}{\sim} \mathcal{B}\Big(\frac{s}{2d}\Big),$$

and i.i.d. Rademacher variables $\epsilon_1, \dots, \epsilon_d$ that are independent of $(\delta_1, \dots, \delta_d)$. Denote by $\mu$ the distribution of $(\alpha\delta_1\epsilon_1, \dots, \alpha\delta_d\epsilon_d)$ where $\alpha = (\tau/2)\sqrt{\log(ed/s)}$. Note that $\mu$ is not necessarily supported on $\Theta_s = \{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_0 \leq s\}$ as the number of nonzero coefficients of a vector drawn from $\mu$ can be larger than $s$. Therefore, we consider a restricted to $\Theta_s$ version of $\mu$ defined by

$$\bar{\mu}(A) = \frac{\mu(A \cap \Theta_s)}{\mu(\Theta_s)} \tag{19}$$

for all Borel subsets $A$ of $\mathbb{R}^d$. Finally, we introduce two mixture probability measures

$$\mathbb{P}_\mu = \int \mathbf{P}_{\boldsymbol{\theta},U,1}\, d\mu(\boldsymbol{\theta}) \quad \text{and} \quad \mathbb{P}_{\bar\mu} = \int \mathbf{P}_{\boldsymbol{\theta},U,1}\, d\bar\mu(\boldsymbol{\theta}). \tag{20}$$

Notice that there exists a probability measure $\tilde P \in \mathcal{G}_\tau$ such that

$$\mathbb{P}_\mu = \mathbf{P}_{0,\tilde P,\sigma_0} \tag{21}$$

where $\sigma_0 > 0$ is defined by

$$\sigma_0^2 = 1 + \frac{\tau^2 s}{8d}\log(ed/s) \le 1 + \frac{\tau^2}{8}. \tag{22}$$

Indeed, $\sigma_0^2 = 1 + \frac{\alpha^2 s}{2d}$ is the variance of zero-mean random variable $\alpha\delta\epsilon + \xi$, where $\xi \sim U$, $\epsilon \sim U$, $\delta \sim \mathcal{B}\big(\frac{s}{2d}\big)$ and $\epsilon, \xi, \delta$ are jointly independent. Thus, to prove (21) it is enough to show that, for all $t \ge 2$,

$$\mathbf{P}\big((\tau/2)\sqrt{\log(ed/s)}\,\delta\epsilon + \xi > t\sigma_0\big) \le e^{-t^2/\tau^2}. \tag{23}$$

But this inequality immediately follows from the fact that for $t \ge 2$ the probability in (23) is smaller than

$$\mathbf{P}(\epsilon = 1, \delta = 1)\,\mathbb{1}_{(\tau/2)\sqrt{\log(ed/s)} > t-1} \le \frac{s}{4d}\mathbb{1}_{\tau\sqrt{\log(ed/s)} > t} \le e^{-t^2/\tau^2}.$$

Now, for any estimator $\hat T$ and any $u > 0$ we have

$$\begin{aligned}
&\sup_{P_\xi \in \mathcal{G}_\tau}\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(\Big|\frac{\hat T}{\sigma^2} - 1\Big| \ge u\Big) \\
&\quad \ge \max\Big\{\mathbf{P}_{0,\tilde P,\sigma_0}(|\hat T - \sigma_0^2| \ge \sigma_0^2 u), \int \mathbf{P}_{\boldsymbol{\theta},U,1}(|\hat T - 1| \ge u)\bar\mu(d\boldsymbol{\theta})\Big\} \\
&\quad \ge \max\Big\{\mathbb{P}_\mu(|\hat T - \sigma_0^2| \ge \sigma_0^2 u), \mathbb{P}_{\bar\mu}(|\hat T - 1| \ge \sigma_0^2 u)\Big\}
\end{aligned} \tag{24}$$

where the last inequality uses (21). Write $\sigma_0^2 = 1 + 2\phi$ where $\phi = \frac{\tau^2 s}{16d}\log(ed/s)$ and choose $u = \phi/\sigma_0^2 \ge \phi/(1 + \tau^2/8)$. Then, the expression in (24) is bounded from below by the probability of error in the problem of distinguishing between two simple hypotheses $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar\mu}$, for which Theorem 2.2 in Tsybakov [13] yields

$$\max\Big\{\mathbb{P}_\mu(|\hat T - \sigma_0^2| \ge \phi), \mathbb{P}_{\bar\mu}(|\hat T - 1| \ge \phi)\Big\} \ge \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu})}{2} \tag{25}$$

where $V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu})$ is the total variation distance between $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar\mu}$. The desired lower bound follows from (25) and Lemma 4 for any $s \ge 32$.

(iii) Finally, we prove the lower bound with the rate $\tau^2(s/d)^{1-2/a}$ in Theorem 4. Again, we do not consider the case $s \ge 32$ since in this case the rate $1/\sqrt{d}$ is dominating and Theorem 4 follows from item (i) above. For $s \ge 32$, the proof uses the same argument as in item (ii) above but we choose $\alpha = (\tau/2)(d/s)^{1/a}$. Then the variance of $\alpha\delta\epsilon + \xi$ is equal to

$$\sigma_0^2 = 1 + \frac{\tau^2(s/d)^{1-2/a}}{8}.$$

Furthermore, with this definition of $\sigma_0^2$ there exists $\tilde P \in \mathcal{P}_{a,\tau}$ such that (21) holds. Indeed, analogously to (23) we now have, for all $t \ge 2$,

$$\mathbf{P}\big(\alpha\,\delta\epsilon + \xi > t\sigma_0\big) \le \mathbf{P}(\epsilon = 1, \delta = 1)\,\mathbb{1}_{(\tau/2)(d/s)^{1/a} > t-1} \le \frac{s}{4d}\mathbb{1}_{\tau(d/s)^{1/a} > t} \le (t/\tau)^a.$$

To finish the proof, it remains to repeat the argument of (24) and (25) with $\phi = \frac{\tau^2(s/d)^{1-2/a}}{16}$.

### 5.4 Proof of Theorem 5

As in Section 5.3, we consider only the case $s \geq 32$. Indeed, for $s < 32$ it is enough to use the lower bound of Proposition 2 since in this case $s\log(ed/s) \leq 32\log(ed) \leq Cs\log(1+d/s^2)$ where $C > 0$ is an absolute constant.

The proof for $s \geq 32$ is very close to the argument in Section 5.3. Set $\alpha = (\tau/2)\sqrt{\log(ed/s)}$ when proving the bound on the class $\mathcal{G}_\tau$, and $\alpha = (\tau/2)(d/s)^{1/a}$ when proving the bound on $\mathcal{P}_{\alpha,\tau}$. In what follows, we only deal with the class $\mathcal{G}_\tau$ since the proof for $\mathcal{P}_{\alpha,\tau}$ is analogous. Consider the measures $\mu$ $\bar{\mu}$, $\mathbb{P}_\mu$, $\mathbb{P}_{\bar{\mu}}$ and $\tilde{P}$ defined in Section 5.3. Similarly to (24), for any estimator $\hat{T}$ and any $u > 0$ we have

$$
\begin{aligned}
&\sup_{P_\xi \in \mathcal{G}_\tau} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat{T} - \|\boldsymbol{\theta}\|_2| \geq \sigma u\big) \\
&\quad \geq \max\Big\{\mathbf{P}_{0,\tilde{P},\sigma_0}(|\hat{T}| \geq \sigma_0 u), \int \mathbf{P}_{\boldsymbol{\theta},U,1}(|\hat{T} - \|\boldsymbol{\theta}\|_2| \geq u)\bar{\mu}(d\boldsymbol{\theta})\Big\} \\
&\quad \geq \max\Big\{\mathbb{P}_\mu(|\hat{T}| \geq \sigma_0 u), \mathbb{P}_{\bar{\mu}}(|\hat{T} - \|\boldsymbol{\theta}\|_2| \geq \sigma_0 u)\Big\} \\
&\quad \geq \max\Big\{\mathbb{P}_\mu(|\hat{T}| \geq \sigma_0 u), \mathbb{P}_{\bar{\mu}}(|\hat{T}| < \sigma_0 u, \|\boldsymbol{\theta}\|_2 \geq 2\sigma_0 u)\Big\} \\
&\quad \geq \min_B \max\big\{\mathbb{P}_\mu(B), \mathbb{P}_{\bar{\mu}}(B^c)\big\} - \bar{\mu}(\|\boldsymbol{\theta}\|_2 \leq 2\sigma_0 u), \qquad (26)
\end{aligned}
$$

where $\sigma_0$ is defined in (22), $U$ denotes the Rademacher law and $\min_B$ is the minimum over all Borel sets. The third line in the last display is due to (21) and to the inequality $\sigma_0 \geq 1$.

Set $u = \frac{\alpha\sqrt{s}}{4\sigma_0}$. Then Lemma 5 implies that

$$
\bar{\mu}(\|\boldsymbol{\theta}\|_2 \leq 2\sigma_0 u) \leq 2e^{-\frac{s}{16}}, \qquad (27)
$$

while Theorem 2.2 in [13] yields

$$
\min_B \max\big\{\mathbb{P}_\mu(B), \mathbb{P}_{\bar{\mu}}(B^c)\big\} \;\geq \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}})}{2}, \qquad (28)
$$

where $V(\mathbb{P}_{\bar{\mu}}, \mathbb{P}_\mu)$ is the total variation distance between $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar{\mu}}$. It follows from (26) – (28) and Lemma 4 that

$$
\sup_{P_\xi \in \mathcal{G}_\tau} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat{T} - \|\boldsymbol{\theta}\|_2|/\sigma \geq \alpha\sqrt{s}/(4\sigma_0)\big) \geq \frac{1 - 5e^{-\frac{s}{16}}}{2}.
$$

This proves Theorem 5 for $s \geq 32$ and thus completes the proof.

### 5.5 Proof of Proposition 3

In this proof, we denote by $C_j(a,\tau)$ positive constants depending only on $a$ and $\tau$. Fix $\boldsymbol{\theta} \in \Theta_s$ and let $S$ be the support of $\boldsymbol{\theta}$. We will call outliers the observations $Y_i$ with $i \in S$. There are at least $m - s$ blocks $B_i$ that do not contain outliers. Without loss of generality, we assume that the blocks $B_1, \ldots, B_{m-s}$ contain no outliers.

As $a > 2$, there exist constants $L = L(a,\tau)$ and $r = r(a,\tau)$ such that $\mathbf{E}|\xi_1^2 - 1|^r \leq L$ and $1 < r \leq 2$. Using von Bahr-Esseen inequality (*cf.* [15]) and the fact that $\mathrm{Card}(B_i) \geq k$ we get

$$
\mathbf{P}\Big(\Big|\frac{1}{\mathrm{Card}(B_i)}\sum_{j\in B_i}\xi_j^2 - 1\Big| > 1/2\Big) \leq \frac{2^{r+1}L}{k^{r-1}}, \quad i = 1, \ldots, m.
$$

Hence, there exists a constant $C_1 = C_1(a, \tau)$ such that if $k \geq C_1$ (i.e., if $\gamma$ is small enough depending on $a$ and $\tau$), then

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\bar{\sigma}_i^2 \notin I) \leq \frac{1}{4}, \quad i = 1, \dots, m, \tag{29}$$

where $I = [\frac{\sigma^2}{2}, \frac{3\sigma^2}{2}]$. Next, by the definition of the median, for any interval $I \subseteq \mathbb{R}$ we have

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\tilde{\sigma}^2 \notin I) \leq \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\Big(\sum_{i=1}^{m} \mathbb{1}_{\bar{\sigma}_i^2 \notin I} \geq \frac{m}{2}\Big) \leq \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\Big(\sum_{i=1}^{m-s} \mathbb{1}_{\bar{\sigma}_i^2 \notin I} \geq \frac{m}{2} - s\Big).$$

Now, $s \leq \frac{\lfloor \gamma d \rfloor}{4} = \frac{m}{4}$, so that $\frac{m}{2} - s \geq \frac{m-s}{3}$. Set $\eta_i = \mathbb{1}_{\bar{\sigma}_i^2 \notin I}$, $i = 1, \dots, m - s$. Due to (29) we have $\mathbf{E}(\eta_i) \leq 1/4$, and $\eta_1, \dots, \eta_{m-s}$ are independent. Using these remarks and Hoeffding's inequality we find

$$\mathbf{P}\Big(\sum_{i=1}^{m-s} \eta_i \geq \frac{m}{2} - s\Big) \leq \mathbf{P}\Big(\sum_{i=1}^{m-s} (\eta_i - \mathbf{E}(\eta_i)) \geq \frac{m-s}{12}\Big) \leq \exp(-C_2(a, \tau)(m - s)).$$

Here, $m - s \geq 3m/4 = 3\lfloor \gamma d \rfloor / 4$. Thus, if $\gamma$ is chosen small enough depending only on $a$ and $\tau$ then

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\tilde{\sigma}^2 \notin I) \leq \exp(-C_3(a, \tau)d).$$

### 5.6 Proof of Theorem 7

Set $\boldsymbol{u} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$. It follows from Lemma A.2 in [1] that

$$2\|\boldsymbol{u}\|_2^2 \leq 2\sigma \sum_{i=1}^{d} \xi_i u_i + \tilde{\sigma}\|\boldsymbol{\theta}\|_* - \tilde{\sigma}\|\hat{\boldsymbol{\theta}}\|_*,$$

where $u_i$ are the components of $\boldsymbol{u}$. Next, Lemma A.1 in [1] yields

$$\|\boldsymbol{\theta}\|_* - \|\hat{\boldsymbol{\theta}}\|_* \leq \Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} \|\boldsymbol{u}\|_2 - \sum_{j=s+1}^{d} \lambda_j |u|_{(d-j+1)}$$

where $|u|_{(k)}$ is the $k$th order statistic of $|u_1|, \dots, |u_d|$. Combining these two inequalities we get

$$2\|\boldsymbol{u}\|_2^2 \leq 2\sigma \sum_{i=1}^{d} \xi_i u_i + \tilde{\sigma}\Big\{\Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} \|\boldsymbol{u}\|_2 - \sum_{j=s+1}^{d} \lambda_j |u|_{(d-j+1)}\Big\}.$$

Thus, on the event $\mathcal{B} = \Big\{|\xi|_{(d-j+1)} \leq \lambda_j/4, \ \forall j = 1, \dots, d\Big\} \cap \Big\{1/2 \leq \tilde{\sigma}^2/\sigma^2 \leq 3/2\Big\}$ we have

$$\begin{aligned} 2\|\boldsymbol{u}\|_2^2 &\leq \frac{\sigma}{2} \sum_{i=1}^{d} \lambda_i |u|_{(d-i+1)} + \frac{3}{2}\sigma\Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} \|\boldsymbol{u}\|_2 - \frac{\sigma}{2} \sum_{j=s+1}^{d} \lambda_j |u|_{(d-j+1)} \\ &\leq 2\sigma\Big(\sum_{j=1}^{s} \lambda_j^2\Big)^{1/2} \|\boldsymbol{u}\|_2. \end{aligned}$$

This inequality and the definition of $\lambda_j$ imply

$$\|\boldsymbol{u}\|_2 \le \sigma t \sqrt{\sum_{j=1}^{s} \Big(\frac{d}{j}\Big)^{2/a}} \le c\sigma t\, \phi_{\mathsf{pol},\mathsf{norm}}(s,d),$$

where $c>0$ is a constant depending only on $a$. Finally, combining Lemma 2 and Proposition 3 we get

$$\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}(\mathcal{B}) \ge 1 - \frac{c'}{t^a} - \exp(-t/c')$$

for a constant $c'>0$ depending only on $a$ and $\tau$.

### 5.7 Proof of Theorem 8

Using the definition of $\hat\sigma^2_{\mathsf{srs}}$ in (14) we get

$$|\hat\sigma^2_{\mathsf{srs}} - \sigma^2| \le \frac{1}{n}\|\mathbb{X}(\hat{\boldsymbol{\theta}}_{\mathsf{srs}} - \boldsymbol{\theta})\|_2^2 + \frac{\sigma^2}{n}\Big|\|\boldsymbol{\xi}\|_2^2 - n\Big| + \frac{2\sigma}{n}\Big|\boldsymbol{\xi}^T\mathbb{X}(\hat{\boldsymbol{\theta}}_{\mathsf{srs}} - \boldsymbol{\theta})\Big|. \tag{30}$$

To control the second term on the right-hand side of (30), we apply Bernstein's inequality (*cf.* e.g. Corollary 5.17 in [14]): If $P_\xi \in \mathcal{G}_1$, then

$$\mathbf{P}\Big(\big|\|\boldsymbol{\xi}\|_2^2 - n\big| > u\Big) \le 2e^{-c\left(\frac{u^2}{n}\wedge u\right)}, \quad \forall u>0,$$

where $c>0$ is an absolute constant. This yields that for any $t>0$ such that $\sqrt{t/n}\le c$,

$$\big|\|\boldsymbol{\xi}\|_2^2 - n\big| \le \sqrt{nt/c} \tag{31}$$

with probability at least $1-2e^{-t}$.

Next, to bound the first and the third terms on the right-hand side of (30) we place ourselves on the event of probability at least $1 - c_8(s/p)^s - c_8 e^{-n/c_8}$ where the result of Proposition 4 holds. We denote this event by $\mathcal{C}$. By Proposition 4, on the event $\mathcal{C}$ the first term on the right-hand side of (30) satisfies

$$\frac{1}{n}\|\mathbb{X}(\hat{\boldsymbol{\theta}}_{\mathsf{srs}} - \boldsymbol{\theta})\|_2^2 \le c_9 \frac{\sigma^2 s}{n}\log(ep/s). \tag{32}$$

Finally, to bound the third term on the right-hand side of (30), we use Proposition 9.1 in Bellec, Lecué and Tsybakov [1] that we state here in the following form.

PROPOSITION 5. *Let $t>0$ and let $\mathbb{X}$ be a design matrix satisfying (13). Assume that $P_\xi \in \mathcal{G}_1$. Then, there is a constant $K_1$ such that for all $\boldsymbol{u}\in\mathbb{R}^p$,*

$$\frac{1}{n}|\boldsymbol{\xi}^T\mathbb{X}\boldsymbol{u}| \le K_1\big(G(\boldsymbol{u}) \vee \|\boldsymbol{u}\|_*\big)$$

*with probability at least $1-e^{-t}/2$, where*

$$G(\boldsymbol{u}) = \frac{\sqrt{t}}{n}\ \|\mathbb{X}\boldsymbol{u}\|_2.$$

In what follows, we set $\boldsymbol{u} = \hat{\boldsymbol{\theta}}_{\mathsf{srs}} - \boldsymbol{\theta}$, and we denote by $\mathcal{C}'$ the intersection of $\mathcal{C}$ with the event of probability at least $1 - e^{-t}/2$, on which Proposition 5 holds. Clearly, $\mathbf{P}(\mathcal{C}') \geq 1 - c_8(s/p)^s - c_8 e^{-n/c_8} - e^{-t}/2$. On the event $\mathcal{C}$, Proposition 4 yields

$$\|\boldsymbol{u}\|_* \leq c_9 \frac{\sigma s}{n} \log(ep/s),$$
$$G(\boldsymbol{u}) \leq \frac{\sigma t}{n} + \frac{\|\mathbb{X}\boldsymbol{u}\|_2^2}{\sigma n} \leq \frac{\sigma t}{n} + c_9 \frac{\sigma s}{n} \log(ep/s).$$

This and Proposition 5 imply that, on the event $\mathcal{C}'$,

$$\frac{1}{n}|\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u}| \leq K_1 \sigma \Big( \frac{t}{n} + c_9 \frac{s}{n} \log(ep/s) \Big). \tag{33}$$

Plugging (31), (32), and (33) in (30) we obtain that if $\sqrt{t/n} \leq c$, then

$$|\hat{\sigma}^2_{\mathsf{srs}} - \sigma^2| \leq c_{11} \sigma^2 \Big( \sqrt{t/n} + \frac{s}{n} \log(ep/s) \Big)$$

with probability at least $1 - c_8(s/p)^s - c_8 e^{-n/c_8} - 5e^{-t}/2$, where $c_{11} > 0$ is a constant depending only on $\kappa$.

## 6. LEMMAS

LEMMA 1. *Let $z_1, \dots, z_d \overset{iid}{\sim} P$ with $P \in \mathcal{G}_\tau$ for some $\tau > 0$ and let $z_{(1)} \leq \dots \leq z_{(d)}$ be the order statistics of $|z_1|, \dots, |z_d|$. Then*

$$\mathbf{E}\big(z^4_{(d-j+1)}\big) \leq C \log^2\big(ed/j\big), \qquad j = 1, \dots, d,$$

*where $C > 0$ is a constant depending only on $\tau$.*

PROOF. Using the definition of $\mathcal{G}_\tau$ we get that, for any $t \geq 2$,

$$\mathbf{P}\big(z_{(d-j+1)} \geq t\big) \leq \binom{d}{j} \mathbf{P}^j(|z_1| \geq t) \leq 2\Big(\frac{ed}{j}\Big)^j e^{-j(t/\tau)^2}.$$

Then, for $v = \tau\sqrt{2\log(ed/j)} \vee 2$ we get

$$\begin{aligned} \mathbf{E}\big(z^4_{(d-j+1)}\big) &= 4\int_0^{+\infty} t^3\, \mathbf{P}\big(z_{(d-j+1)} > t\big)\, dt, \\ &\leq v^4 + 8\int_0^{+\infty} t^3\, e^{-j(t/\tau)^2/2}\, dt = v^4 + \frac{16\tau^4}{j^2}. \end{aligned} \tag{34}$$

The result follows. □

LEMMA 2. *Let $z_{(1)} \leq \dots \leq z_{(d)}$ be as in Lemma 1 with $P \in \mathcal{P}_{a,\tau}$ for some $\tau > 0$ and $a > 0$. Then if $u > (2e)^{1/a}\tau \vee 2$, we have*

$$\mathbf{P}\Big(z_{(d-j+1)} \leq u\Big(\frac{d}{j}\Big)^{1/a}, \forall\, j = 1, \dots, d\Big) \geq 1 - \frac{2e\tau^a}{u^a} \tag{35}$$

*and if $a > 4$,*

$$\mathbf{E}\big(z^4_{(d-j+1)}\big) \leq C\Big(\frac{de}{j}\Big)^{4/a}, \qquad j = 1, \dots, d, \tag{36}$$

*where $C > 0$ is a constant depending only on $\tau$ and $a$.*

PROOF. Similarly to the proof of Lemma 1 we have, for all $t \geq 2$,

$$\mathbf{P}\big(z_{(d-j+1)} \geq t\big) \leq \Big(\frac{ed}{j}\Big)^j \Big(\frac{\tau}{t}\Big)^{ja}.$$

Set $t_j = u\Big(\frac{d}{j}\Big)^{1/a}$ and $q = e(\tau/u)^a$. Using the assumption that $q < 1/2$ we get, for $u \geq 2$,

$$\mathbf{P}\Big(\exists\, j \in \{1,\dots,d\} : z_{(d-j+1)} \geq u\Big(\frac{d}{j}\Big)^{1/a}\Big) \leq \sum_{j=1}^{d} \Big(\frac{ed}{j}\Big)^j \Big(\frac{\tau}{t_j}\Big)^{ja} = \sum_{j=1}^{d} q^j \leq 2q.$$

This proves (35). The proof of (36) is obtained analogously to (34). □

LEMMA 3 (Deviations of the binomial distribution). *Let $\mathcal{B}(d,p)$ denote the binomial random variable with parameters $d$ and $p \in (0,1)$. Then, for any $\lambda > 0$,*

$$\mathbf{P}\big(\mathcal{B}(d,p) \geq \lambda\sqrt{d} + dp\big) \leq \exp\bigg(-\frac{\lambda^2}{2p(1-p)\big(1+\frac{\lambda}{3p\sqrt{d}}\big)}\bigg), \tag{37}$$

$$\mathbf{P}\big(\mathcal{B}(d,p) \leq -\lambda\sqrt{d} + dp\big) \leq \exp\bigg(-\frac{\lambda^2}{2p(1-p)}\bigg). \tag{38}$$

Inequality (37) is a combination of formulas (3) and (10) on pages 440–441 in [12]. Inequality (38) is formula (6) on page 440 in [12].

LEMMA 4. *Let $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar\mu}$ be the probability measures defined in (20). Then,*

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) \leq e^{-\frac{3s}{16}},$$

*where $V(\cdot,\cdot)$ denotes the total variation distance.*

PROOF. We have

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) = \sup_B \bigg|\int \mathbf{P}_{\boldsymbol\theta,U,1}(B)d\mu(\boldsymbol\theta) - \int \mathbf{P}_{\boldsymbol\theta,U,1}(B)d\bar\mu(\boldsymbol\theta)\bigg| \leq \sup_{|f|\leq 1}\bigg|\int f d\mu - \int f d\bar\mu\bigg| = V(\mu,\bar\mu).$$

Furthermore, $V(\mu,\bar\mu) \leq \mu(\Theta_s^c)$ since for any Borel subset $B$ of $\mathbb{R}^d$ we have $\big|\mu(B) - \bar\mu(B)\big| \leq \mu(B \cap \Theta_s^c)$. Indeed,

$$\mu(B) - \bar\mu(B) \leq \mu(B) - \mu(B\cap\Theta) = \mu(B\cap\Theta^c)$$

and

$$\bar\mu(B) - \mu(B) = \frac{\mu(B\cap\Theta)}{\mu(\Theta)} - \mu(B\cap\Theta) - \mu(B\cap\Theta^c) \geq -\mu(B\cap\Theta^c).$$

Thus,

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) \leq \mu(\Theta_s^c) = \mathbf{P}\Big(\mathcal{B}\Big(d, \frac{s}{2d}\Big) > s\Big).$$

The lemma now follows from (37). □

LEMMA 5. *Let $\bar\mu$ be defined in (19) with some $\alpha > 0$. Then*

$$\bar\mu\Big(\|\boldsymbol\theta\|_2 \leq \frac{\alpha}{2}\sqrt{s}\Big) \leq 2e^{-\frac{s}{16}}.$$

PROOF. First, note that

$$\mu\Big(\|\boldsymbol{\theta}\|_2 \le \frac{\alpha}{2}\sqrt{s}\Big) = \mathbf{P}\Big(\mathcal{B}(d, \frac{s}{2d}) \le \frac{s}{4}\Big) \le e^{-\frac{s}{16}} \tag{39}$$

where the last inequality follows from (38). Next, inspection of the proof of Lemma 4 yields that $\bar{\mu}(B) \le \mu(B) + e^{-\frac{3s}{16}}$ for any Borel set $B$. Taking here $B = \{\|\boldsymbol{\theta}\|_2 \le \alpha\sqrt{s}/2\}$ and using (39) proves the lemma. □

LEMMA 6. *There exists a probability density $f_0 : \mathbb{R} \to [0, \infty)$ with the following properties: $f_0$ is continuously differentiable, symmetric about 0, supported on $[-3/2, 3/2]$, with variance 1 and finite Fisher information $I_{f_0} = \int (f_0'(x))^2 (f_0(x))^{-1} dx$.*

PROOF. Let $K : \mathbb{R} \to [0, \infty)$ be any probability density, which is continuously differentiable, symmetric about 0, supported on $[-1, 1]$, and has finite Fisher information $I_K$, for example, the density $K(x) = \cos^2(\pi x/2)\mathbb{1}_{|x|\le 1}$. Define $f_0(x) = [K_h(x + (1-\varepsilon)) + K_h(x - (1-\varepsilon))]/2$ where $h > 0$ and $\varepsilon \in (0, 1)$ are constants to be chosen, and $K_h(u) = K(u/h)/h$. Clearly, we have $I_{f_0} < \infty$ since $I_K < \infty$. It is straightforward to check that the variance of $f_0$ satisfies $\int x^2 f_0(x) dx = (1-\varepsilon)^2 + h^2\sigma_K^2$ where $\sigma_K^2 = \int u^2 K(u) du$. Choosing $h = \sqrt{2\varepsilon - \varepsilon^2}/\sigma_K$ and $\varepsilon \le \sigma_K^2/8$ guarantees that $\int x^2 f_0(x) dx = 1$ and the support of $f_0$ is contained in $[-3/2, 3/2]$. □

## 7. ACKNOWLEDGEMENTS

The work of O. Collier has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The work of A.B. Tsybakov was 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).

## REFERENCES

[1] P. BELLEC, G. LECUÉ AND A.B. TSYBAKOV. Slope meets Lasso: improved oracle bounds and optimality. arXiv preprint, arXiv:1605.08651v3. To appear in the *Annals of Statistics*.

[2] A. BELLONI, V. CHERNOZHUKOV AND L. WANG. Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics* **42** 757–788, 2014.

[3] M. CHEN, C. GAO AND Z. REN. Robust covariance and scatter matrix estimation under Huber's contamination model. arXiv preprint, arXiv:1506.00691. To appear in the *Annals of Statistics*.

[4] O. COLLIER, L. COMMINGES AND A.B. TSYBAKOV. Minimax estimation of linear and quadratic functionals under sparsity constraints. *Annals of Statistics*, **45**, 923–958, 2017.

[5] O. COLLIER, L. COMMINGES, A.B. TSYBAKOV AND N. VERZELEN. Optimal adaptive estimation of linear functionals under sparsity. arXiv preprint, arxiv:1611.09744. To appear in the *Annals of Statistics*.

[6] A. DERUMIGNY. Improved oracle bounds for Square-Root Lasso and Square-Root Slope. arXiv preprint, arXiv:1703.02907v2.

[7] E. GAUTIER AND A.B. TSYBAKOV. Pivotal estimation in high-dimensional regression via linear programming. In: Empirical Inference. Festschrift in Honor of Vladimir N. Vapnik, B.Schölkopf, Z. Luo, V. Vovk eds., 195 - 204. Springer, New York e.a., 2013.

[8] Y. GOLUBEV AND E. KRYMOVA. On estimation of the noise variance in high-dimensional linear models. arXiv preprint, arXiv:1711.09208.

[9] P.J. HUBER. *Robust Statistics,* J. Wiley, 1981.

[10] I.A.IBRAGIMOV AND R.Z.HASMINSKII. *Statistical Estimation. Asymptotic Theory,* Springer, New York, 1981.

[11] S. MINSKER AND X. WEI. Estimation of the covariance structure of heavy-tailed distributions. arXiv preprint, arXiv:1708.00502.

[12] G. Shorack and J. Wellner. *Empirical Processes with Applications to Statistics,* John Wiley, New York, 1986.
[13] A.B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, New York, 2009.
[14] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In: *Compressed sensing*, 210–268, Cambridge Univ. Press, Cambridge, 2012.
[15] B. von Bahr and C.-G. Esseen. Inequalities for the $r$-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.* **36**, 299–303, 1965.
[16] L. Wasserman. *All of Statistics.* Springer, New York, 2005.