Adaptive robust estimation in sparse vector model

Laëtitia Comminges, Olivier Collier, M Ndaoud, A. Tsybakov

▶ **To cite this version:**

Laëtitia Comminges, Olivier Collier, M Ndaoud, A. Tsybakov. Adaptive robust estimation in sparse vector model. 2019. hal-01707612v2

HAL Id: hal-01707612

https://hal.science/hal-01707612v2

Preprint submitted on 5 Oct 2019 (v2), last revised 7 Mar 2021 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Submitted to the Annals of Statistics

ADAPTIVE ROBUST ESTIMATION IN SPARSE VECTOR MODEL

By L. Comminges[1], O. Collier[2], M. Ndaoud[3] and A.B. Tsybakov[3]

[1]*CEREMADE, Université Paris-Dauphine and CREST*
[2]*Modal'X, UPL, Université Paris Nanterre and CREST*
[3]*CREST, ENSAE*

Abstract For the sparse vector model, we consider estimation of the target vector, of its ℓ_2-norm and of the noise variance. We construct adaptive estimators and establish the optimal rates of adaptive estimation when adaptation is considered with respect to the triplet "noise level – noise distribution – sparsity". We consider classes of noise distributions with polynomially and exponentially decreasing tails as well as the case of Gaussian noise. The obtained 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 variance. Moreover, knowing or not knowing the noise distribution can also influence the rate. For example, the rates of estimation of the noise variance can differ depending on whether the noise is Gaussian or sub-Gaussian without a precise knowledge of the distribution. Estimation of noise variance 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.

1. Introduction. This paper considers estimation of the unknown sparse vector, of its ℓ_2-norm and of the noise level in the sparse sequence model. The focus is 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 the model defined as follows. Let the signal $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ be observed with noise of unknown magnitude $\sigma > 0$:

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

The noise random variables $\xi_1, \dots, \xi_d$ are assumed to be i.i.d. and we denote by P_ξ the unknown distribution of ξ_1. We assume throughout that the noise is zero-mean, $\mathbf{E}(\xi_1) = 0$, and that $\mathbf{E}(\xi_1^2) = 1$, since σ needs to be identifiable. We denote by $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}$ the distribution of $\boldsymbol{Y} = (Y_1, \dots, Y_d)$ when the signal is $\boldsymbol{\theta}$, the noise level is σ and the distribution of the noise variables is P_ξ. 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 $\Theta_s = \{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_0 \leq s\}$. We consider the problems of

Keywords and phrases: variance estimation, sparsity in linear regression, functional estimation, robust estimation, adaptive estimation, minimax rate

estimating $\boldsymbol{\theta}$ under the ℓ_2 loss, estimating the variance σ^2, and estimating the ℓ_2-norm

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

The classical Gaussian sequence model corresponds to the case where the noise ξ_i is standard Gaussian ($P_\xi = \mathcal{N}(0,1)$) and the noise level σ is known. Then, the optimal rate of estimation of $\boldsymbol{\theta}$ under the ℓ_2 loss in a minimax sense on the class Θ_s is $\sqrt{s\log(ed/s)}$ and it is attained by thresholding estimators [10]. Also, for the Gaussian sequence model with known σ, minimax optimal estimator of the norm $\|\boldsymbol{\theta}\|_2$ as well as the corresponding minimax rate are available from [8] (see Table 1).

In this paper, we study estimation of the three objects $\boldsymbol{\theta}$, $\|\boldsymbol{\theta}\|_2$, and σ^2 in the following two settings.

(a) *The distribution of ξ_i and the noise level σ are both unknown.* This is the main setting of our interest. For the unknown distribution of ξ_i, we consider two types of assumptions. Either P_ξ belongs to a class $\mathcal{G}_{a,\tau}$, *i.e.,* for some $a, \tau > 0$,

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

which includes for example sub-Gaussian distributions ($a = 2$), or to a class of distributions with polynomially decaying tails $\mathcal{P}_{a,\tau}$, *i.e.,* for some $\tau > 0$ and $a \geq 2$,

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

We propose estimators of $\boldsymbol{\theta}$, $\|\boldsymbol{\theta}\|_2$, and σ^2 that are optimal in non-asymptotic minimax sense on these classes of distributions and the sparsity class Θ_s. We establish the corresponding non-asymptotic minimax rates. They are given in the second and third columns of Table 1. We also provide the minimax optimal estimators.

(b) *Gaussian noise ξ_i and unknown σ.* The results on the non-asymptotic minimax rates are summarized in the first column of Table 1. Notice an interesting effect – the rates of estimation of σ^2 and of the norm $\|\boldsymbol{\theta}\|_2$ when the noise is Gaussian are faster than the optimal rates when the noise is sub-Gaussian. This can be seen by comparing the first column of Table 1 with the particular case $a = 2$ of the second column corresponding to sub-Gaussian noise.

Some comments about Table 1 and additional details are in order.

- The difference between the minimax rates for estimation of $\boldsymbol{\theta}$ and estimation of the ℓ_2-norm $\|\boldsymbol{\theta}\|_2$ turns out to be specific for the pure Gaussian noise model. It disappears for the classes $\mathcal{G}_{a,\tau}$ and $\mathcal{P}_{a,\tau}$. This is somewhat unexpected since $\mathcal{G}_{2,\tau}$ is the class of sub-Gaussian distributions, and it turns out that $\|\boldsymbol{\theta}\|_2$ is estimated optimally at different rates for sub-Gaussian and pure Gaussian noise. Another conclusion is that if the noise is not Gaussian and σ is unknown, the minimax rate for $\|\boldsymbol{\theta}\|_2$ does not have an elbow between the "dense" ($s > \sqrt{d}$) and the "sparse" ($s \leq \sqrt{d}$) zones.
- For the problem of estimation of variance σ^2 with *known* distribution of the noise P_ξ, we consider a more general setting than (b) mentioned above. We show that when the noise distribution is exactly known (and satisfies a rather general assumption, not necessarily Gaussian - can have polynomial tails), then the rate of estimation of σ^2 can be as fast

	Gaussian noise model	Noise in class $\mathcal{G}_{a,\tau}$,	Noise in class $\mathcal{P}_{a,\tau}$,
$\boldsymbol{\theta}$	$\sqrt{s\log(ed/s)}$ known σ [10] unknown σ [22]	$\sqrt{s}\log^{\frac{1}{a}}(ed/s)$ unknown σ	$\sqrt{s}(d/s)^{\frac{1}{a}}$ unknown σ
$\|\boldsymbol{\theta}\|_2$	$\sqrt{s\log(1+\frac{\sqrt{d}}{s})}\wedge d^{1/4}$ known σ [8] $\sqrt{s\log(1+\frac{\sqrt{d}}{s})}\vee\sqrt{\frac{s}{1+\log_+(s^2/d)}}$ unknown σ	$\sqrt{s}\log^{\frac{1}{a}}(ed/s)\wedge d^{1/4}$ known σ $\sqrt{s}\log^{\frac{1}{a}}(ed/s)$ unknown σ	$\sqrt{s}(d/s)^{\frac{1}{a}}\wedge d^{1/4}$ known σ $\sqrt{s}(d/s)^{\frac{1}{a}}$ unknown σ
σ^2	$\frac{1}{\sqrt{d}}\vee\frac{s}{d(1+\log_+(s^2/d))}$	$\frac{1}{\sqrt{d}}\vee\frac{s}{d}\log^{\frac{2}{a}}\left(\frac{ed}{s}\right)$	$\frac{1}{\sqrt{d}}\vee\left(\frac{s}{d}\right)^{1-\frac{2}{a}}$

TABLE 1
Optimal rates of convergence.

as $\max\left(\frac{1}{\sqrt{d}},\frac{s}{d}\right)$, which is faster than the optimal rate $\max\left(\frac{1}{\sqrt{d}},\frac{s}{d}\log\left(\frac{ed}{s}\right)\right)$ for the class of sub-Gaussian noise. In other words, the phenomenon of improved rate is not due to the Gaussian character of the noise but rather to the fact that the noise distribution is known.

- Our findings show that there is a dramatic difference between the behavior of optimal estimators of $\boldsymbol{\theta}$ in the sparse sequence model and in the sparse linear regression model with "well spread" regressors. It is known from [11, 2] that in sparse linear regression with "well spread" regressors (that is, having positive variance), the rates of estimating $\boldsymbol{\theta}$ are the same for the noise with sub-Gaussian and polynomial tails. We show that the situation is quite different in the sparse sequence model, where the optimal rates are much slower and depend on the polynomial index of the noise.
- The rates shown in Table 1 for the classes $\mathcal{G}_{a,\tau}$ and $\mathcal{P}_{a,\tau}$ are achieved on estimators that are adaptive to the sparsity index s. Thus, knowing or not knowing s does not influence the optimal rates of estimation when the distribution of ξ and the noise level are unknown.

We conclude this section by a discussion of related work. Chen, Gao and Ren [7] explore the problem of robust estimation of variance and of covariance matrix under Hubers's contamination model. As explained in Section 4 below, this problem has similarities with 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, the corresponding results in Section 4 can be viewed as results 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 [17] who consider classes of distributions similar to $\mathcal{P}_{a,\tau}$ rather than the contamination model. The main aim in [17] is to construct estimators having sub-Gaussian deviations under weak moment assumptions. Our setting is different in that we consider the sparsity class Θ_s of vectors $\boldsymbol{\theta}$ and the rates that we obtain depend on s. Estimation of variance in sparse linear model is discussed in [20] where some upper bounds for the rates are given. We also mention the recent paper [12] that deals with estimation of

variance in linear regression in a framework that does not involve sparsity, as well as the work on estimation of signal-to-noise ratio functionals in settings involving sparsity [23, 13] and not involving sparsity [16]. Papers [9, 6] discuss estimation of other functionals than the ℓ_2-norm $\|\boldsymbol{\theta}\|_2$ in the sparse vector model when the noise is Gaussian with unknown variance.

Notation. For $x > 0$, let $\lfloor x \rfloor$ denote the maximal integer smaller than x. For a finite set A, we denote by $|A|$ its cardinality. Let $\inf_{\hat{T}}$ denote the infimum over all estimators. The notation C, C',c, c' will be used for positive constants that can depend only a and τ and can vary from line to line.

2. Estimation of sparse vector $\boldsymbol{\theta}$. In this section, we study the problem of estimating a sparse vector $\boldsymbol{\theta}$ in ℓ_2-norm when the variance of noise σ and the distribution of ξ_i are both unknown. We only assume that the noise distribution belongs a given class, which can be either a class of distributions with polynomial tails $\mathcal{P}_{a,\tau}$, or a class $\mathcal{G}_{a,\tau}$ with exponential decay of the tails.

First, we introduce a preliminary estimator $\tilde{\sigma}^2$ of σ^2 that will be used to define an estimator of $\boldsymbol{\theta}$. Let $\gamma \in (0, 1/2]$ be a constant that will be chosen small enough and depending only on a and τ. Divide $\{1, \dots, d\}$ into $m = \lfloor \gamma d \rfloor$ disjoint subsets $B_1, \dots, B_m$, each of cardinality $|B_i| \geq k := \lfloor d/m \rfloor \geq 1/\gamma - 1$. Consider the median-of-means estimator

$$\tilde{\sigma}^2 = \mathsf{med}(\bar{\sigma}_1^2, \dots, \bar{\sigma}_m^2), \text{ where } \bar{\sigma}_i^2 = \frac{1}{|B_i|} \sum_{j \in B_i} Y_j^2, \quad i = 1, \dots, m. \tag{4}$$

Here, $\mathsf{med}(\bar{\sigma}_1^2, \dots, \bar{\sigma}_m^2)$ denotes the median of $\bar{\sigma}_1^2, \dots, \bar{\sigma}_m^2$. The next proposition shows that the estimator $\tilde{\sigma}^2$ recovers σ^2 to within a constant factor.

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

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

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left| \tilde{\sigma}^2 - \sigma^2 \right| \leq C\sigma^2,$$

and for $a > 4$,

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\tilde{\sigma}^2 - \sigma^2 \right)^2 \leq C\sigma^4.$$

Note that the result of Proposition 1 also holds for the class $\mathcal{G}_{a,\tau}$ for all $a > 0$ and $\tau > 0$. Indeed, $\mathcal{G}_{a,\tau} \subset \mathcal{P}_{a,\tau}$ for all $a > 2$ and $\tau > 0$, while for any $0 < a \leq 2$ and $\tau > 0$, there exist $a' > 4$ and $\tau' > 0$ such that $\mathcal{G}_{a,\tau} \subset \mathcal{P}_{a',\tau'}$.

We further note that assuming $s < cd$ for some $0 < c < 1$ is natural in the context of variance estimation since σ is not identifiable when $s = d$. In what follows, all upper bounds on the risks of estimators will be obtained under this assumption.

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

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \Big(\sum_{i=1}^d (Y_i - \theta_i)^2 + \tilde{\sigma} \|\boldsymbol{\theta}\|_* \Big). \tag{5}$$

Here, $\|\cdot\|_*$ denotes the sorted ℓ_1-norm:

$$\|\boldsymbol{\theta}\|_* = \sum_{i=1}^{d} \lambda_i |\theta|_{(d-i+1)}, \tag{6}$$

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

Set

$$\phi^*_{\mathsf{exp}}(s,d) = \sqrt{s}\log^{1/a}(ed/s), \qquad \phi^*_{\mathsf{pol}}(s,d) = \sqrt{s}(d/s)^{1/a}. \tag{7}$$

The next theorem shows that $\hat{\boldsymbol{\theta}}$ estimates $\boldsymbol{\theta}$ with the rates $\phi^*_{\mathsf{exp}}(s,d)$ and $\phi^*_{\mathsf{pol}}(s,d)$ when the noise distribution belongs to the class $\mathcal{G}_{a,\tau}$ and class $\mathcal{P}_{a,\tau}$, respectively.

THEOREM 1. *Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$ where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. Then for the estimator $\hat{\boldsymbol{\theta}}$ defined by (5) the following holds.*

1. *Let $\tau > 0$, $a > 0$. There exist constants $c, C > 0$ and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\lambda_j = c\log^{1/a}(ed/j), j = 1, \ldots, d$, we have*

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right) \le C\sigma^2 \left(\phi^*_{\mathsf{exp}}(s,d) \right)^2.$$

2. *Let $\tau > 0, a > 2$. There exist constants $c, C > 0$ and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\lambda_j = c(d/j)^{1/a}, j = 1, \ldots, d$, we have*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right) \le C\sigma^2 \left(\phi^*_{\mathsf{pol}}(s,d) \right)^2.$$

Furthermore, it follows from the lower bound of Theorem 2 in Section 3 that the rates $\phi^*_{\mathsf{exp}}(s,d)$ and $\phi^*_{\mathsf{pol}}(s,d)$ cannot be improved in a minimax sense. Thus, the estimator $\hat{\boldsymbol{\theta}}$ defined in (5) achieves the optimal rates in a minimax sense.

From Theorem 1, we can conclude that the optimal rate ϕ^*_{pol} under polynomially decaying noise is very different from the optimal rate ϕ^*_{exp} under exponential tails, in particular, from the rate under the sub-Gaussian noise. At first sight, this phenomenon seems to contradict some results in the literature on sparse regression model. Indeed, Gautier and Tsybakov [11] consider sparse linear regression with unknown noise level σ and show that the Self-Tuned Dantzig estimator can achieve the same rate as in the case of Gaussian noise (up to a logarithmic 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 $a > 2$. However, 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 [11] to obtain a sub-Gaussian rate. This condition of "well spreadness" is not satisfied in the sparse sequence model that we are considering here. In this model viewed as a special case of linear regression, the design is deterministic, with only one non-zero component. We see that such a degenerate design turns out to be the least favorable from the point of view of the convergence rate, while the "well spread" design

is the best one. An interesting general conclusion of comparing our findings to [11] and [2] 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 is 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. Studying them remains an open problem.

3. Estimation of the ℓ_2-norm. In this section, we consider the problem of estimation of the ℓ_2-norm of a sparse vector when the variance of the noise and the form of its distribution are both unknown. We show that the rates $\phi^*_{\mathsf{exp}}(s,d)$ and $\phi^*_{\mathsf{pol}}(s,d)$ are optimal in a minimax sense on the classes $\mathcal{G}_{a,\tau}$ and $\mathcal{P}_{a,\tau}$, respectively. We first provide a lower bound on the risks of any estimators of the ℓ_2-norm when the noise level σ is unknown and the unknown noise distribution P_ξ belongs either to $\mathcal{G}_{a,\tau}$ or $\mathcal{P}_{a,\tau}$. 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 2. *Let s, d be integers satisfying $1 \le s \le d$. Let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then, for any $a > 0, \tau > 0$,*

$$\inf_{\hat T} \sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\Big(c(\phi^*_{\mathsf{exp}}(s,d))^{-1}\Big|\frac{\hat T - \|\boldsymbol{\theta}\|_2}{\sigma}\Big|\Big) \ge c', \tag{8}$$

and, for any $a \ge 2, \tau > 0$,

$$\inf_{\hat T} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\Big(\bar c(\phi^*_{\mathsf{pol}}(s,d))^{-1}\Big|\frac{\hat T - \|\boldsymbol{\theta}\|_2}{\sigma}\Big|\Big) \ge \bar c'. \tag{9}$$

Here, $\inf_{\hat T}$ denotes the infimum over all estimators, and $c, \bar c > 0$, $c', \bar c' > 0$ are constants that can depend only on $\ell(\cdot)$, τ and a.

The lower bound (9) implies that the rate of estimation of the ℓ_2-norm of a sparse vector deteriorates dramatically if the bounded moment assumption is imposed on the noise instead, for example, of the sub-Gaussian assumption.

Note also that (8) and (9) immediately imply lower bounds with the same rates ϕ^*_{exp} and ϕ^*_{pol} for the estimation of the s-sparse vector $\boldsymbol{\theta}$ under the ℓ_2-norm.

Given the upper bounds of Theorem 1, the lower bounds (8) and (9) are tight for the quadratic loss, and are achieved by the following plug-in estimator independent of s or σ:

$$\hat N = \|\hat{\boldsymbol{\theta}}\|_2 \tag{10}$$

where $\hat{\boldsymbol{\theta}}$ is defined in (5).

In conclusion, when both P_ξ and σ are unknown the rates ϕ^*_{exp} and ϕ^*_{pol} defined in (7) are minimax optimal both for estimation of $\boldsymbol{\theta}$ and of the the norm $\|\boldsymbol{\theta}\|_2$.

We now compare these results with the findings in [8] regarding the (nonadaptive) estimation of $\|\boldsymbol{\theta}\|_2$ when ξ_i have the standard Gaussian distribution ($P_\xi = \mathcal{N}(0,1)$) and σ is known. It is shown in [8] that in this case the optimal rate of estimation of $\|\boldsymbol{\theta}\|_2$ has the form

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

Namely, the following proposition holds.

PROPOSITION 2 (Gaussian noise, known σ [8]). *For any $\sigma > 0$ and any integers s, d satisfying $1 \le s \le d$, we have*

$$c\sigma^2\phi^2_{\mathcal{N}(0,1)}(s,d) \le \inf_{\hat{T}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\big(\hat{T} - \|\boldsymbol{\theta}\|_2\big)^2 \le C\sigma^2\phi^2_{\mathcal{N}(0,1)}(s,d),$$

where $c > 0$ and $C > 0$ are absolute constants and $\inf_{\hat{T}}$ denotes the infimum over all estimators.

We have seen that, in contrast to this result, in the case of unknown P_ξ and σ the optimal rates (7) do 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 P_ξ and σ is only possible with a significant deterioration of the rate. On the other hand, for the sub-Gaussian class $\mathcal{G}_{2,\tau}$, in the "sparse" zone $s \le \sqrt{d}$ the non-adaptive rate $\sqrt{s\log(1+\sqrt{d}/s)}$ differs only slightly from the adaptive sub-Gaussian rate $\sqrt{s\log(ed/s)}$; in fact, this difference in the rate appears only in a vicinity of $s = \sqrt{d}$.

A natural question is whether such a deterioration of the rate is caused by the ignorance of σ or by the ignorance of the distribution of ξ_i within the sub-Gaussian class $\mathcal{G}_{2,\tau}$. The answer is that both are responsible. It turns out that if only one of the two ingredients (σ or the noise distribution) is unknown, then a rate faster than the adaptive sub-Gaussian rate $\phi^*_{\mathsf{exp}}(s,d) = \sqrt{s\log(ed/s)}$ can be achieved. This is detailed in the next two propositions.

Consider first the case of Gaussian noise and unknown σ. Set

$$\phi^*_{\mathcal{N}(0,1)}(s,d) = \max\left\{\sqrt{s\log(1+\sqrt{d}/s)}, \sqrt{\frac{s}{1+\log_+(s^2/d)}}\right\},$$

where $\log_+(x) = \max(0, \log(x))$ for any $x > 0$. We divide the set $\{1,\dots,d\}$ into two disjoint subsets I_1 and I_2 with $\min(|I_1|, |I_2|) \ge \lfloor d/2 \rfloor$. Let $\hat{\sigma}^2$ be the variance estimator defined by (15), cf. Section 4.1 below, and let $\hat{\sigma}^2_{\mathsf{med},1}, \hat{\sigma}^2_{\mathsf{med},2}$ be the median estimators (12) corresponding to the samples $(Y_i)_{i\in I_1}$ and $(Y_i)_{i\in I_2}$, respectively. Consider the estimator

$$\hat{N}^* = \begin{cases} \sqrt{\left|\sum_{j=1}^d (Y_j^2\, \mathbb{1}_{\{|Y_j|>\rho_j\}}) - d\alpha\hat{\sigma}^2\right|} & \text{if} \quad s \le \sqrt{d}, \\ \sqrt{\left|\sum_{j=1}^d Y_j^2 - d\hat{\sigma}^2\right|} & \text{if} \quad s > \sqrt{d}, \end{cases} \tag{11}$$

where $\rho_j = 2\hat{\sigma}_{\mathsf{med},1}\sqrt{2\log(1+d/s^2)}$ if $j \in I_2$, $\rho_j = 2\hat{\sigma}_{\mathsf{med},2}\sqrt{2\log(1+d/s^2)}$ if $j \in I_1$ and $\alpha = \mathbf{E}\left(\xi_1^2\, \mathbb{1}_{\{|\xi_1| > 2\sqrt{2\log(1+d/s^2)}\}}\right)$. Note that Y_j is independent of ρ_j for every j. Note also that the estimator $\hat{N}^*$ depends on the preliminary estimator $\tilde{\sigma}^2$ since $\hat{\sigma} > 0$ defined in (15) depends on it.

PROPOSITION 3 (Gaussian noise, unknown σ). *The following two properties hold.*

(i) Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. There exist absolute constants $C > 0$ and $\gamma \in (0, 1/2]$ such that

$$\sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\left(\hat{N}^* - \|\boldsymbol{\theta}\|_2\right)^2 \le C\sigma^2\left(\phi^*_{\mathcal{N}(0,1)}(s,d)\right)^2.$$

(ii) Let s and d be integers satisfying $1 \le s \le d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,

$$\inf_{\hat{T}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\, \ell\left(c(\phi^*_{\mathcal{N}(0,1)}(s,d))^{-1} \left| \frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma} \right| \right) \ge c',$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators, and $c > 0$, $c' > 0$ are constants that can depend only on $\ell(\cdot)$.

The proof of item (ii) of Proposition 3 (the lower bound) is given in the Supplementary material.

Proposition 3 establishes the minimax optimality of the rate $\phi^*_{\mathcal{N}(0,1)}(s,d)$. It also shows that if σ is unknown, the knowledge of the Gaussian character of the noise leads to an improvement of the rate compared to the adaptive sub-Gaussian rate $\sqrt{s\log(ed/s)}$. However, the improvement is only in a logarithmic factor.

Consider now the case of unknown noise distribution in $\mathcal{G}_{a,\tau}$ and known σ. We show in the next proposition that in this case the minimax rate is of the form

$$\phi^\circ_{\mathsf{exp}}(s,d) = \min\{\sqrt{s}\log^{\frac{1}{a}}(ed/s), d^{1/4}\}$$

and it is achieved by the estimator

$$\hat{N}^\circ_{\mathsf{exp}} = \begin{cases} \|\hat{\boldsymbol{\theta}}\|_2 & \text{if} \quad s \le \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}, \\ \left| \sum_{j=1}^d Y_j^2 - d\sigma^2 \right|^{1/2} & \text{if} \quad s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}, \end{cases}$$

where $\hat{\boldsymbol{\theta}}$ is defined in (5). Note $\phi^\circ_{\mathsf{exp}}(s,d)$ can be written equivalently (up to absolute constants) as $\min\{\sqrt{s}\log^{\frac{1}{a}}(ed), d^{1/4}\}$.

PROPOSITION 4 (Unknown noise in $\mathcal{G}_{a,\tau}$, known σ). *Let $a, \tau > 0$. The following two properties hold.*

(i) Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. There exist constants $c, C > 0$, and $\gamma \in (0, 1/2]$ depending only on (a,τ) such that if $\hat{\boldsymbol{\theta}}$ is the estimator defined in (5) with $\lambda_j = c\log^{\frac{1}{a}}(ed/j)$, $j = 1, \dots, d$, then

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma} \left(\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2 \right)^2 \le C\sigma^2 \left(\phi^\circ_{\mathsf{exp}}(s,d) \right)^2.$$

(ii) Let s and d be integers satisfying $1 \le s \le d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then, there exist constants $c > 0$, $c' > 0$ depending only on $\ell(\cdot)$, a and τ such that

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\left(c(\phi^\circ_{\mathsf{exp}}(s,d))^{-1} \left| \frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma} \right| \right) \ge c',$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators.

Proposition 4 establishes the minimax optimality of the rate $\phi^{\circ}_{\mathsf{exp}}(s,d)$. It also shows that if the noise distribution is unknown and belongs to $\mathcal{G}_{a,\tau}$, the knowledge of σ leads to an improvement of the rate compared to the case when σ is unknown. In contrast to the case of Proposition 3 (Gaussian noise), the improvement here is substantial; it results not only in a logarithmic but in a polynomial factor in the dense zone $s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$.

We end this section by considering the case of unknown polynomial noise and known σ. The next proposition shows that in this case the minimax rate, for a given $a > 4$, is of the form

$$\phi^{\circ}_{\mathsf{pol}}(s,d) = \min\{\sqrt{s}(d/s)^{\frac{1}{a}}, d^{1/4}\}$$

and it is achieved by the estimator

$$\hat{N}^{\circ}_{\mathsf{pol}} = \begin{cases} \|\hat{\boldsymbol{\theta}}\|_2 & \text{if} \quad s \le d^{\frac{1}{2}-\frac{1}{a-2}}, \\ \left|\sum_{j=1}^{d} Y_j^2 - d\sigma^2\right|^{1/2} & \text{if} \quad s > d^{\frac{1}{2}-\frac{1}{a-2}}, \end{cases}$$

where $\hat{\boldsymbol{\theta}}$ is defined in (5).

PROPOSITION 5 (Unknown noise in $\mathcal{P}_{a,\tau}$, known σ). *Let $\tau > 0, a > 4$. The following two properties hold.*

(i) Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. There exist constants $c, C > 0$, and $\gamma \in (0, 1/2]$ depending only on (a,τ) such that if $\hat{\boldsymbol{\theta}}$ is the estimator defined in (5) with $\lambda_j = c(d/j)^{\frac{1}{a}}$, $j = 1, \dots, d$, then

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\hat{N}^{\circ}_{\mathsf{pol}} - \|\boldsymbol{\theta}\|_2\right)^2 \le C\sigma^2 \left(\phi^{\circ}_{\mathsf{pol}}(s,d)\right)^2.$$

(ii) Let s and d be integers satisfying $1 \le s \le d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then, there exist constants $c > 0$, $c' > 0$ depending only on $\ell(\cdot)$, a and τ such that

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\, \ell\left(c(\phi^{\circ}_{\mathsf{pol}}(s,d))^{-1} \left|\frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma}\right|\right) \ge c',$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators.

Note that here, similarly to Proposition 4, the improvement over the case of unknown σ is in a polynomial factor in the dense zone $s > d^{\frac{1}{2}-\frac{1}{a-2}}$.

4. Estimating the variance of the noise.

4.1. *Estimating σ^2 when the distribution P_ξ is known.* 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 $\theta_i \neq 0$. Thus, the models in robustness and sparsity problems are quite similar but the questions of interest are different. When robust estimation of σ^2 is considered, the object of interest is the pure

noise component of the sparsity model while the non-zero components θ_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. Recall that, when $\boldsymbol{\theta} = 0$, the median M-estimator of scale (*cf.* [14]) is defined as

$$\hat{\sigma}^2_{\mathsf{med}} = \frac{\hat{M}}{\beta} \tag{12}$$

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

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

and β is the median of the distribution of ξ_1^2. Here, F_d denotes the empirical c.d.f. of $(Y_1^2, \dots, Y_d^2)$. When F denotes the c.d.f. of ξ_1^2, it is easy to see that

$$\beta = F^{-1}(1/2). \tag{13}$$

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

PROPOSITION 6. *Let ξ_1^2 have a c.d.f. F with positive density, and let β be given by* (13)*. There exist constants $\gamma \in (0, 1/8)$, $c > 0$, $c_* > 0$ and $C > 0$ depending only on F such that for any integers s and d satisfying $1 \le s < \gamma d$ and any $t > 0$ we have*

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

and if $\mathbf{E}|\xi_1|^{2+\epsilon} < \infty$ for some $\epsilon > 0$. Then,

$$\sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},F,\sigma} \left| \hat{\sigma}^2_{\mathsf{med}} - \sigma^2 \right|}{\sigma^2} \le C \max\left(\frac{1}{\sqrt{d}}, \frac{s}{d} \right).$$

The main message of Proposition 6 is that the rate of convergence of $\hat{\sigma}^2_{\mathsf{med}}$ in probability and in expectation is as fast as

$$\max\left(\frac{1}{\sqrt{d}}, \frac{s}{d} \right) \tag{14}$$

and it does not depend on F when F varies in a large class. The role of Proposition 6 is to contrast the subsequent results of this section dealing with unknown distribution of noise and providing slower rates. It emphasizes the fact that the knowledge of the noise distribution is crucial as it leads to an improvement of the rate of estimating the variance.

However, the rate (14) achieved by the median estimator is not necessarily optimal. As shown in the next proposition, in the case of Gaussian noise the optimal rate is even better:

$$\phi_{\mathcal{N}(0,1)}(s, d) = \max\left\{ \frac{1}{\sqrt{d}}, \frac{s}{d(1 + \log_+(s^2/d))} \right\}.$$

This rate is attained by an estimator that we are going to define now. We use the observation that, in the Gaussian case, the modulus of the empirical characteristic function $\varphi_d(t) = \frac{1}{d}\sum_{i=1}^d e^{itY_j}$ is to within a constant factor of the Gaussian characteristic function $\exp(-\frac{t^2\sigma^2}{2})$ for any t. This suggests the estimator

$$\tilde{v}^2 = -\frac{2\log(|\varphi_d(\hat{t}_1)|)}{\hat{t}_1^2},$$

with a suitable choice of $t = \hat{t}_1$ that we further set as follows:

$$\hat{t}_1 = \frac{1}{\tilde{\sigma}}\sqrt{\log\big(4(es/\sqrt{d}+1)\big)},$$

where $\tilde{\sigma}$ is the preliminary estimator (4) with some tuning parameter $\gamma \in (0, 1/2]$. The final variance estimator is defined as a truncated version of $\tilde{v}^2$:

$$\hat{\sigma}^2 = \begin{cases} \tilde{v}^2 & \text{if } |\varphi_d(\hat{t}_1)| > (es/\sqrt{d}+1)^{-1}/4, \\ \tilde{\sigma}^2 & \text{otherwise.} \end{cases} \tag{15}$$

PROPOSITION 7 (Gaussian noise). *The following two properties hold.*

(i) Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. There exist absolute constants $C > 0$ and $\gamma \in (0, 1/2]$ such that the estimator $\hat{\sigma}^2$ defined in (15) satisfies

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s} \frac{\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\left|\hat{\sigma}^2 - \sigma^2\right|}{\sigma^2} \le C\phi_{\mathcal{N}(0,1)}(s,d).$$

(ii) Let s and d be integers satisfying $1 \le s \le d$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. Then,

$$\inf_{\hat{T}}\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s} \mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\,\ell\left(c(\phi_{\mathcal{N}(0,1)}(s,d))^{-1}\left|\frac{\hat{T}}{\sigma^2} - 1\right|\right) \ge c',$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators, and $c > 0$, $c' > 0$ are constants that can depend only on $\ell(\cdot)$.

Estimators of variance or covariance matrix based on the empirical characteristic function have been studied in several papers [4, 5, 3, 6]. The setting in [4, 5, 3] is different from the ours as those papers deal with the model where the non-zero components of $\boldsymbol{\theta}$ are random with a smooth distribution density. The estimators in [4, 5] are also quite different. On the other hand, [3, 6] consider estimators close to $\tilde{v}^2$. In particular, [6] uses a similar pilot estimator for testing in the sparse vector model where it is assumed that $\sigma \in [\sigma_-, \sigma_+]$, $0 < \sigma_- < \sigma_+ < \infty$, and the estimator depends on σ_+. Although [6] does not provide explicitly stated result about the rate of this estimator, the proofs in [6] come close to it and we believe that it satisfies an upper bound as in item (i) of Proposition 7 with $\sup_{\sigma>0}$ replaced by $\sup_{\sigma\in[\sigma_-,\sigma_+]}$.

4.2. *Distribution-free variance estimators.* The main drawback of the estimator $\hat{\sigma}^2_{\mathsf{med}}$ is the dependence on the parameter β. It reflects the fact that the estimator is tailored for a given and known distribution of noise F. Furthermore, as shown below, the rate (14) cannot

be achieved if it is only known that F belongs to one of the classes of distributions that we consider in this paper.

Instead of using one particular quantile, like the median in Section 4.1, one can estimate σ^2 by an integral over all quantiles, which allows one to avoid considering distribution-dependent quantities like (13).

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_α by the empirical quantile $\hat{q}_\alpha$ of level α we obtain the following estimator

$$\hat{\sigma}^2 = \int_0^{1-s/d} \hat{q}_\alpha \, d\alpha = \frac{1}{d}\sum_{k=1}^{d-s} Y^2_{(k)}, \tag{16}$$

where $Y^2_{(1)} \le \ldots \le Y^2_{(d)}$ 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.* [14]. Also, up to a constant factor, $\hat{\sigma}^2$ coincides with the statistic used in Collier, Comminges and Tsybakov [8] .

The following theorem provides an upper bound on the risk of the estimator $\hat{\sigma}^2$ under the assumption that the noise belongs to the class $\mathcal{G}_{a,\tau}$. Set

$$\phi_{\mathsf{exp}}(s,d) = \max\left(\frac{1}{\sqrt{d}}, \frac{s}{d}\log^{2/a}\left(\frac{ed}{s}\right)\right).$$

THEOREM 3. *Let $\tau > 0$, $a > 0$, and let s, d be integers satisfying $1 \le s < d/2$. Then, the estimator $\hat{\sigma}^2$ defined in (16) satisfies*

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\hat{\sigma}^2 - \sigma^2\right)^2}{\sigma^4} \le C\phi^2_{\mathsf{exp}}(s,d) \tag{17}$$

where $C > 0$ is a constant depending only on a and τ.

The next theorem establishes the performance of variance estimation 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 4. *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 (16) satisfies*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\hat{\sigma}^2 - \sigma^2\right)^2}{\sigma^4} \le C\phi^2_{\mathsf{pol}}(s,d), \tag{18}$$

where $C > 0$ is a constant depending only on a and τ.

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 expected squared error of a quadratic function of the observations.

We now state the lower bounds matching the results of Theorems 3 and 4.

THEOREM 5. *Let $\tau > 0$, $a > 0$, 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,*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\Big(c(\phi_{\mathsf{exp}}(s,d))^{-1}\Big|\frac{\hat{T}}{\sigma^2} - 1\Big|\Big) \ge c', \tag{19}$$

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

Theorems 3 and 5 imply that the estimator $\hat{\sigma}^2$ is rate optimal in a minimax sense when the noise belongs to $\mathcal{G}_{a,\tau}$, in particular when it is sub-Gaussian. Interestingly, an extra logarithmic factor appears in the optimal rate when passing from the pure Gaussian distribution of ξ_i's (*cf.* Proposition 7) 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. Also note that this logarithmic factor vanishes as $a \to \infty$.

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

THEOREM 6. *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,*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\, \ell\Big(c(\phi_{\mathsf{pol}}(s,d))^{-1}\Big|\frac{\hat{T}}{\sigma^2} - 1\Big|\Big) \ge c' \tag{20}$$

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

This theorem shows that the rate $\phi_{\mathsf{pol}}(s,d)$ obtained in Theorem 4 cannot be improved in a minimax sense.

A drawback of the estimator defined in (16) is in the lack of adaptivity to the sparsity parameter s. At first sight, it may seem that the estimator

$$\hat{\sigma}^2_* = \frac{2}{d} \sum_{1 \le k \le d/2} Y^2_{(k)} \tag{21}$$

could be taken as its adaptive version. However, $\hat{\sigma}^2_*$ is not a good estimator of σ^2 as can be seen from the following proposition.

PROPOSITION 8. *Define $\hat{\sigma}^2_*$ as in (21). Let $\tau > 0$, $a \ge 2$, and let s, d be integers satisfying $1 \le s \le d$, and $d = 4k$ for an integer k. Then,*

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \frac{\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\big(\hat{\sigma}^2_* - \sigma^2\big)^2}{\sigma^4} \ge \frac{1}{64}.$$

On the other hand, it turns out that a simple plug-in estimator

$$\hat{\sigma}^2 = \frac{1}{d}\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|_2^2 \tag{22}$$

with $\hat{\boldsymbol{\theta}}$ chosen as in Section 2 achieves rate optimality adaptively to the noise distribution and to the sparsity parameter s. This is detailed in the next theorem.

THEOREM 7. *Let s and d be integers satisfying $1 \le s < \lfloor \gamma d \rfloor/4$, where $\gamma \in (0, 1/2]$ is the tuning parameter in the definition of $\tilde{\sigma}^2$. Let $\hat{\sigma}^2$ be the estimator defined by (22) where $\hat{\boldsymbol{\theta}}$ is defined in (5). Then the following properties hold.*

1. *Let $\tau > 0, a > 0$. There exist constants $c, C > 0$ and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\lambda_j = c\log^{1/a}(ed/j), j = 1, \dots, d$, we have*

$$\sup_{P_\xi \in \mathcal{P}_{a,\tau}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta}, P_\xi, \sigma}\big|\hat{\sigma}^2 - \sigma^2\big| \le C\sigma^2 \phi_{\mathsf{exp}}(s, d).$$

2. *Let $\tau > 0, a > 4$. There exist constants $c, C > 0$ and $\gamma \in (0, 1/2]$ depending only on (a, τ) such that if $\lambda_j = c(d/j)^{1/a}, j = 1, \dots, d$, we have*

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

5. Proofs of the upper bounds.

5.1. *Proof of Proposition 1.* 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. Denote by J a set of $m - s$ indices i, for which B_i contains no outliers.

As $a > 2$, there exist constants $L = L(a, \tau)$ and $r = r(a, \tau) \in (1, 2]$ such that $\mathbf{E}|\xi_1^2 - 1|^r \le L$. Using von Bahr-Esseen inequality (*cf.* [18]) and the fact that $|B_i| \ge k$ we get

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

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

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\bar{\sigma}_i^2 \notin I) \le \frac{1}{4}, \quad i = 1, \dots, m, \tag{23}$$

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

$$\mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}(\tilde{\sigma}^2 \notin I) \le \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\Big(\sum_{i=1}^m \mathbb{1}_{\bar{\sigma}_i^2 \notin I} \ge \frac{m}{2}\Big) \le \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\Big(\sum_{i \in J} \mathbb{1}_{\bar{\sigma}_i^2 \notin I} \ge \frac{m}{2} - s\Big). \tag{24}$$

Now, $s \le \frac{\lfloor \gamma d \rfloor}{4} = \frac{m}{4}$, so that $\frac{m}{2} - s \ge \frac{m-s}{3}$. Set $\eta_i = \mathbb{1}_{\bar{\sigma}_i^2 \notin I}$, $i \in J$. Due to (23) we have $\mathbf{E}(\eta_i) \le 1/4$, and $(\eta_i, i \in J)$ are independent. Using these remarks and Hoeffding's inequality we find

$$\mathbf{P}\Big(\sum_{i \in J} \eta_i \ge \frac{m}{2} - s\Big) \le \mathbf{P}\Big(\sum_{i \in J}(\eta_i - \mathbf{E}(\eta_i)) \ge \frac{m-s}{12}\Big) \le \exp(-C(m-s)).$$

Note that $|J| = m - s \geq 3m/4 = 3\lfloor \gamma d \rfloor/4$. Thus, if γ is chosen small enough depending only on a and τ then

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

This proves the desired bound in probability. To obtain the bounds in expectation, set $Z = \left|\tilde{\sigma}^2 - \sigma^2\right|$. Let first $a > 4$ and take some $r \in (1, a/4)$. Then

$$\begin{aligned}
\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(Z^2\right) &\leq \frac{\sigma^4}{4} + \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(Z^2 \mathbb{1}_{Z \geq \frac{\sigma^2}{2}}\right) \\
&\leq \frac{9\sigma^4}{4} + 2\left(\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\tilde{\sigma}^{4r}\right)\right)^{1/r}\left(\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\left(Z \geq \sigma^2/2\right)\right)^{1-1/r} \\
&\leq \frac{9\sigma^4}{4} + 2\left(\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\tilde{\sigma}^{4r}\right)\right)^{1/r}\exp(-Cd).
\end{aligned}$$

Since $m \geq 4s$, we can easily argue that $\tilde{\sigma}^{4r} \leq \sum_{i\in J}\bar{\sigma}_i^{4r}$. It follows that

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\tilde{\sigma}^{4r}\right) \leq C\sigma^{4r}d^2.$$

Hence $\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(Z^2\right) \leq C\sigma^4$. Similarly, if $a > 2$, then $\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(Z\right) \leq C\sigma^2$.

5.2. *Proof of Theorem 1.* 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 kth order statistic of $|u_1|, \ldots, |u_d|$. Combining these two inequalities we get

$$2\|\boldsymbol{u}\|_2^2 \leq 2\sigma\sum_{j=1}^d \xi_j u_j + \tilde{\sigma}\Big\{\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\|\boldsymbol{u}\|_2 - \sum_{j=s+1}^d \lambda_j |u|_{(d-j+1)}\Big\}. \tag{25}$$

For some permutation $(\varphi(1), \ldots, \varphi(d))$ of $(1, \ldots, d)$, we have

$$\Big|\sum_{i=1}^d \xi_j u_j\Big| \leq \sum_{j=1}^d |\xi|_{(d-j+1)}|u_{\varphi(j)}| \leq \sum_{j=1}^d |\xi|_{(d-j+1)}|u|_{(d-j+1)}, \tag{26}$$

where the last inequality is due to the fact that the sequence $|\xi|_{(d-j+1)}$ is non-increasing. Hence

$$\begin{aligned}
2\|\boldsymbol{u}\|_2^2 &\leq 2\sigma\sum_{j=1}^s |\xi|_{(d-j+1)}|u|_{(d-j+1)} + \tilde{\sigma}\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\|\boldsymbol{u}\|_2 + \sum_{j=s+1}^d \left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)|u|_{(d-j+1)} \\
&\leq \left\{2\sigma\Big(\sum_{j=1}^s |\xi|_{(d-j+1)}^2\Big)^{1/2} + \tilde{\sigma}\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2} + \Big(\sum_{j=s+1}^d \left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\Big)^{1/2}\right\}\|\boldsymbol{u}\|_2.
\end{aligned}$$

This implies

$$\|\boldsymbol{u}\|_2^2 \le C\left\{\sigma^2\sum_{j=1}^{s}|\xi|^2_{(d-j+1)} + \tilde{\sigma}^2\sum_{j=1}^{s}\lambda_j^2 + \sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right\}.$$

From Lemmas 1 and 2 we have $\mathbf{E}(|\xi|^2_{(d-j+1)}) \le C\lambda_j^2$. Using this and Proposition 1 we obtain

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\|\boldsymbol{u}\|_2^2\right) \le C\left(\sigma^2\sum_{j=1}^{s}\lambda_j^2 + \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right)\right). \tag{27}$$

Define the events $\mathcal{A}_j = \left\{|\xi|_{(d-j+1)} \le \lambda_j/4\right\} \cap \left\{1/2 \le \tilde{\sigma}^2/\sigma^2 \le 3/2\right\}$ for $j = s+1,\dots,d$. Then

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right) \le 4\sigma^2\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}|\xi|^2_{(d-j+1)}\mathbb{1}_{\mathcal{A}_j^c}\right).$$

Fixing some $1 < r < a/2$ we get

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right) \le 4\sigma^2\sum_{j=s+1}^{d}\mathbf{E}\left(|\xi|^{2r}_{(d-j+1)}\right)^{1/r}\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\mathcal{A}_j^c\right)^{1-1/r}.$$

Lemmas 1, 2 and the definitions of parameters λ_j imply that

$$\mathbf{E}\left(|\xi|^{2r}_{(d-j+1)}\right)^{1/r} \le C\lambda_s^2, \quad j = s+1,\dots,d.$$

Furthermore, it follows from the proofs of Lemmas 1 and 2 that if the constant c in the definition of λ_j is chosen large enough, then $\mathbf{P}(|\xi|_{(d-j+1)} > \lambda_j/4) \le q^j$ for some $q < 1/2$ depending only on a and τ. This and Proposition 1 imply that $\mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}(\mathcal{A}_j^c) \le e^{-cd} + q^j$. Hence,

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\sum_{j=s+1}^{d}\left(2\sigma|\xi|_{(d-j+1)} - \tilde{\sigma}\lambda_j\right)_+^2\right) \le C\sigma^2\lambda_s^2\sum_{j=s+1}^{d}(e^{-cd}+q^j)^{1-1/r} \le C'\sigma^2\sum_{j=1}^{s}\lambda_j^2.$$

Combining this inequality with (27) we obtain

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\left(\|\boldsymbol{u}\|_2^2\right) \le C\sigma^2\sum_{j=1}^{s}\lambda_j^2. \tag{28}$$

To complete the proof, it remains to note that $\sum_{j=1}^{s}\lambda_j^2 \le C(\phi^*_{\mathsf{pol}}(s,d))^2$ in the polynomial case and $\sum_{j=1}^{s}\lambda_j^2 \le C(\phi^*_{\mathsf{exp}}(s,d))^2$ in the exponential case, cf. Lemma 3.

5.3. *Proof of part (i) of Proposition 3.* We consider separately the "dense" zone $s > \sqrt{d}$ and the "sparse" zone $s \le \sqrt{d}$. Let first $s > \sqrt{d}$. Then the rate $\phi^*_{\mathcal{N}(0,1)}(s,d)$ is of order $\sqrt{\frac{s}{1+\log_+(s^2/d)}}$. Thus, for $s > \sqrt{d}$ we need to prove that

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\left(\left|\frac{\hat{N}^*-\|\boldsymbol{\theta}\|_2}{\sigma}\right|^2\right)\le\frac{Cs}{1+\log_+(s^2/d)}. \tag{29}$$

Denoting $\boldsymbol{\xi}=(\xi_1,\dots,\xi_d)$ we have

$$\begin{aligned}\left|\hat{N}^*-\|\boldsymbol{\theta}\|_2\right| &= \left|\Big|\sum_{j=1}^d Y_j^2-d\hat{\sigma}^2\Big|^{1/2}-\|\boldsymbol{\theta}\|_2\right| \\ &= \left|\sqrt{\big|\|\boldsymbol{\theta}\|_2^2+2\sigma(\boldsymbol{\theta},\boldsymbol{\xi})+\sigma^2\|\boldsymbol{\xi}\|_2^2-d\hat{\sigma}^2\big|}-\|\boldsymbol{\theta}\|_2\right| \\ &\le \left|\sqrt{\big|\|\boldsymbol{\theta}\|_2^2+2\sigma(\boldsymbol{\theta},\boldsymbol{\xi})\big|}-\|\boldsymbol{\theta}\|_2\right|+\sigma\sqrt{\big|\|\boldsymbol{\xi}\|_2^2-d\big|}+\sqrt{d|\sigma^2-\hat{\sigma}^2|}.\end{aligned} \tag{30}$$

The first term in the last line vanishes if $\boldsymbol{\theta}=0$, while for $\boldsymbol{\theta}\neq 0$ it is bounded as follows:

$$\left|\sqrt{\big|\|\boldsymbol{\theta}\|_2^2+2\sigma(\boldsymbol{\theta},\boldsymbol{\xi})\big|}-\|\boldsymbol{\theta}\|_2\right|=\|\boldsymbol{\theta}\|_2\left|\sqrt{\Big|1+\frac{2\sigma(\boldsymbol{\theta},\boldsymbol{\xi})}{\|\boldsymbol{\theta}\|_2^2}\Big|}-1\right|\le\frac{2\sigma|(\boldsymbol{\theta},\boldsymbol{\xi})|}{\|\boldsymbol{\theta}\|_2} \tag{31}$$

where we have used the inequality $|\sqrt{|1+x|}-1|\le|x|$, $\forall x\in\mathbb{R}$. Since here $|(\boldsymbol{\theta},\boldsymbol{\xi})|/\|\boldsymbol{\theta}\|_2\sim\mathcal{N}(0,1)$ we have, for all $\boldsymbol{\theta}$,

$$\mathbf{E}\left(\left|\sqrt{\big|\|\boldsymbol{\theta}\|_2^2+2\sigma(\boldsymbol{\theta},\boldsymbol{\xi})\big|}-\|\boldsymbol{\theta}\|_2\right|^2\right)\le 4\sigma^2, \tag{32}$$

and since $\|\boldsymbol{\xi}\|_2^2$ has a chi-square distribution with d degrees of freedom we have

$$\mathbf{E}\Big(\big|\|\boldsymbol{\xi}\|_2^2-d\big|\Big) \le \Big(\mathbf{E}\Big(\big|\|\boldsymbol{\xi}\|_2^2-d\big|^2\Big)\Big)^{1/2}=\sqrt{2d}.$$

Next, by Proposition 7 we have that, for $s>\sqrt{d}$,

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\left(\left|\frac{\hat{\sigma}^2}{\sigma^2}-1\right|\right)\le\frac{Cs}{d(1+\log_+(s^2/d))} \tag{33}$$

for some absolute constant $C>0$. Combining (30) – (33) yields (29).

Let now $s\le\sqrt{d}$. Then the rate $\phi^*_{\mathcal{N}(0,1)}(s,d)$ is of order $\sqrt{s\log(1+d/s^2)}$. Thus, for $s\le\sqrt{d}$ we need to prove that

$$\sup_{\sigma>0}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma}\left(\left|\frac{\hat{N}^*-\|\boldsymbol{\theta}\|_2}{\sigma}\right|^2\right)\le Cs\log(1+d/s^2). \tag{34}$$

We have

$$\begin{aligned}
(35)\quad \left|\hat{N}^* - \|\boldsymbol{\theta}\|_2\right| &= \left|\Big|\sum_{j=1}^{d}(Y_j^2\ \mathbb{1}_{\{|Y_j|>\rho_j\}}) - d\alpha\hat{\sigma}^2\Big|^{1/2} - \|\boldsymbol{\theta}\|_2\right| \\
&= \left|\Big|\sum_{j\in S}(Y_j^2\ \mathbb{1}_{\{|Y_j|>\rho_j\}}) + \sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}}) - d\alpha\hat{\sigma}^2\Big|^{1/2} - \|\boldsymbol{\theta}\|_2\right| \\
&\le \left|\sqrt{\sum_{j\in S}(Y_j^2\ \mathbb{1}_{\{|Y_j|>\rho_j\}})} - \|\boldsymbol{\theta}\|_2\right| + \left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}}) - d\alpha\hat{\sigma}^2\right|^{1/2}.
\end{aligned}$$

Here,

$$\begin{aligned}
(36)\qquad \left|\sqrt{\sum_{j\in S}(Y_j^2\ \mathbb{1}_{\{|Y_j|>\rho_j\}})} - \|\boldsymbol{\theta}\|_2\right| &\le \left|\sqrt{\sum_{j\in S}(Y_j\ \mathbb{1}_{\{|Y_j|>\rho_j\}} - \theta_j)^2}\right| \\
&\le \sqrt{\sum_{j\in S}\rho_j^2} + \sigma\sqrt{\sum_{j\in S}\xi_j^2}.
\end{aligned}$$

Hence, writing for brevity $\mathbf{E}_{\boldsymbol{\theta},\mathcal{N}(0,1),\sigma} = \mathbf{E}$, we get

$$\begin{aligned}
\mathbf{E}\left(\left|\sqrt{\sum_{j\in S}(Y_j^2\ \mathbb{1}_{\{|Y_j|>\rho_j\}})} - \|\boldsymbol{\theta}\|_2\right|^2\right) &\le 16\mathbf{E}\left(\hat{\sigma}^2_{\mathsf{med},1} + \hat{\sigma}^2_{\mathsf{med},2}\right) s\log\left(1 + d/s^2\right) + 2\sigma^2 s \\
&\le C\sigma^2 s\log(1+d/s^2),
\end{aligned}$$

where we have used the fact that $\mathbf{E}\left(|\hat{\sigma}^2_{\mathsf{med},\mathsf{k}} - \sigma^2|\right) \le C\sigma^2$, $\mathsf{k} = 1, 2$, by Proposition 6. Next, we study the term $\Gamma = \left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}}) - d\alpha\hat{\sigma}^2\right|$. We first write

$$(37)\qquad \Gamma \le \left|\sigma^2\sum_{j\notin S}\xi_j^2(\ \mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}} - \ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}})\right| + \left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat{\sigma}^2\right|,$$

where $t_* = 2\sigma\sqrt{2\log(1+d/s^2)}$. For the second summand on the right hand side of (37) we have

$$\left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat{\sigma}^2\right| \le \sigma^2\left|\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - (d-|S|)\alpha\right| + \left|\sigma^2 - \hat{\sigma}^2\right| d\alpha + |S|\alpha\sigma^2,$$

where $|S|$ denotes the cardinality of S. By Proposition 7 we have $\mathbf{E}(|\hat{\sigma}^2 - \sigma^2|) \le C/\sqrt{d}$ for $s \le \sqrt{d}$. Hence,

$$\mathbf{E}\left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat{\sigma}^2\right| \le \sigma^2\sqrt{d\mathbf{E}\left(\xi_1^4\ \mathbb{1}_{\{|\xi_1|>\sqrt{2\log(1+d/s^2)}\}}\right)} + C\alpha\sigma^2\left(\sqrt{d} + s\right).$$

It is not hard to check (cf., e.g., [8, Lemma 4]) that, for $s \le \sqrt{d}$,

$$\alpha \le C(\log\left(1+d/s^2\right))^{1/2}\frac{s^2}{d},$$

and

$$\mathbf{E}\left(\xi_1^4\ \mathbb{1}_{\{|\xi_1|>\sqrt{2\log(1+d/s^2)}\}}\right) \le C(\log\left(1+d/s^2\right))^{3/2}\frac{s^2}{d},$$

so that

$$\mathbf{E}\left|\sigma^2\sum_{j\notin S}(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>t_*\}}) - d\alpha\hat\sigma^2\right| \le C\sigma^2 s\log(1+d/s^2).$$

Thus, to complete the proof it remains to show that

$$\sigma^2\sum_{j\notin S}\mathbf{E}\left|\xi_j^2(\mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}} - \mathbb{1}_{\{\sigma|\xi_j|>t_*\}})\right| \le C\sigma^2 s\log(1+d/s^2). \tag{38}$$

Recall that ρ_j is independent from ξ_j. Hence, conditioning on ρ_j we obtain

$$\sigma^2\mathbf{E}\left(\left|\xi_j^2(\mathbb{1}_{\{\sigma|\xi_j|>\rho_j\}} - \mathbb{1}_{\{\sigma|\xi_j|>t_*\}})\right|\rho_j\right) \le |\rho_j^2 - t_*^2|e^{-t_*^2/(8\sigma^2)} + \sigma^2\mathbb{1}_{\{\rho_j<t_*/2\}}, \tag{39}$$

where we have used the fact that, for $b>a>0$,

$$\int_a^b x^2e^{-x^2/2}dx \le \int_a^b xe^{-x^2/4}dx \le |b^2-a^2|e^{-\min(a^2,b^2)/4}/2.$$

Using Proposition 6 and definitions of ρ_j and t_*, we get that, for $s\le\sqrt{d}$,

$$\begin{aligned}\mathbf{E}\left(|\rho_j^2 - t_*^2|\right)e^{-t_*^2/(8\sigma^2)} &\le 8\max_{\mathsf{k}=1,2}\mathbf{E}(|\hat\sigma^2_{\mathsf{med},\mathsf{k}} - \sigma^2|)\frac{s^2}{d}\log(1+d/s^2)\\ &\le C\sigma^2\frac{s}{d}\log(1+d/s^2).\end{aligned} \tag{40}$$

Next, it follows from Proposition 6 that there exists $\gamma\in(0,1/8)$ small enough such that for $s\le\gamma d$ we have $\max_{\mathsf{k}=1,2}\mathbf{P}(\hat\sigma^2_{\mathsf{med},\mathsf{k}} < \sigma^2/2) \le 2e^{-c_\gamma d}$ where $c_\gamma>0$ is a constant. Thus, $\sigma^2\mathbf{P}(\rho_j < t_*/2) \le 2\sigma^2e^{-c_\gamma d} \le C\sigma^2(s/d)\log(1+d/s^2)$. Combining this with (39) and (40) proves (38).

5.4. *Proof of part (i) of Proposition 4 and part (i) of Proposition 5.* We only prove Proposition 4 since the proof of Proposition 5 is similar taking into account that $\mathbf{E}(\xi_1^4)<\infty$. We consider separately the "dense" zone $s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$ and the "sparse" zone $s \le \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$. Let first $s > \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$. Then the rate $\phi^\circ_{\mathsf{exp}}(s,d)$ is of order $d^{1/4}$ and thus we need to prove that

$$\sup_{P_\xi\in\mathcal{G}_{a,\tau}}\sup_{\|\boldsymbol{\theta}\|_0\le s}\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat N^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2\big) \le C\sigma^2\sqrt{d}.$$

Since σ is known, arguing similarly to (30) - (31) we find

$$|\hat N^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2| \le \left|\frac{2\sigma|(\boldsymbol{\theta},\boldsymbol{\xi})|}{\|\boldsymbol{\theta}\|_2}\right|\mathbb{1}_{\boldsymbol{\theta}\neq0} + \sigma\sqrt{\big|\|\boldsymbol{\xi}\|_2^2 - d\big|}.$$

As $\mathbf{E}(\xi_1^4) < \infty$, this implies

$$\mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2\big) \le 8\sigma^2 + C\sigma^2\sqrt{d},$$

which proves the result in the dense case. Next, in the sparse cas $s \le \frac{\sqrt{d}}{\log^{\frac{2}{a}}(ed)}$, we need to prove that

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2\big) \le C\sigma^2 s \log^{\frac{2}{a}}(ed).$$

This is immediate by Theorem 1 and the fact that $|\hat{N}^\circ_{\mathsf{exp}} - \|\boldsymbol{\theta}\|_2|^2 \le \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2$ for the plug-in estimator $\hat{N}^\circ_{\mathsf{exp}} = \|\hat{\boldsymbol{\theta}}\|_2$.

5.5. *Proof of Proposition 6.* 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 M be the median of G, that is $G(M) = 1/2$. By the definition of $\hat{M}$,

$$|F_d(\hat{M}) - 1/2| \le |F_d(M) - 1/2|. \tag{41}$$

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

$$|G_d(\hat{M}) - 1/2| \le |G_d(M) - 1/2| + 2s/d. \tag{42}$$

The DKW inequality [24, 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 $t > 0$ such that $\sqrt{\frac{t}{d}} + \frac{s}{d} \le 1/8$, 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

$$|G(\hat{M}) - 1/2| \le |G(M) - 1/2| + 2\left(\sqrt{\frac{t}{2(d-s)}} + \frac{s}{d}\right) \le 2\left(\sqrt{\frac{t}{d}} + \frac{s}{d}\right) \le \frac{1}{4}, \tag{43}$$

where the last two inequalities are due to the fact that $G(M) = 1/2$ and to the assumption about t. Notice that

$$|G(\hat{M}) - 1/2| = |G(\hat{M}) - G(M)| = \big|F(\hat{M}/\sigma^2) - F(M/\sigma^2)\big|. \tag{44}$$

Using (43), (44) and the fact that $M = \sigma^2 F^{-1}(1/2)$ we obtain that, on the event $\mathcal{A}$,

$$F^{-1}(1/4) \le \hat{M}/\sigma^2 \le F^{-1}(3/4). \tag{45}$$

This and (44) imply

$$|G(\hat{M}) - 1/2| \ge c_{**}\big|\hat{M}/\sigma^2 - M/\sigma^2\big| = c_{**}\beta\,\big|\hat{\sigma}^2_{\mathsf{med}}/\sigma^2 - 1\big|. \tag{46}$$

where $c_{**} = \min_{x\in[F^{-1}(1/4),F^{-1}(3/4)]} F'(x) > 0$, and $\beta = F^{-1}(1/2)$. Combining the last inequality with (43) we get that, on the event $\mathcal{A}$,

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

Recall that we assumed that $\sqrt{\frac{t}{d}} + \frac{s}{d} \le 1/8$. Thus, there exists a constant $c_* > 0$ depending only on F such that for $t > 0$ and integers s, d satisfying $\sqrt{\frac{t}{d}} + \frac{s}{d} \le 1/8$ we have

$$\sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},F,\sigma} \left(\left| \frac{\hat{\sigma}^2_{\mathsf{med}}}{\sigma^2} - 1 \right| \ge c_* \left(\sqrt{\frac{t}{d}} + \frac{s}{d} \right) \right) \le 2e^{-t}. \tag{47}$$

This and the assumption that $\frac{s}{d} \le \gamma < 1/8$ imply the result of the proposition in probability. We now prove the result in expectation. Set $Z = \left|\hat{\sigma}^2_{\mathsf{med}} - \sigma^2\right| / \sigma^2$. We have

$$\mathbf{E}_{\boldsymbol{\theta},F,\sigma}(Z) \le c_* s/d + \int_{c_* s/d}^{c_*/8} \mathbf{P}_{\boldsymbol{\theta},F,\sigma}(Z > u)\, du + \mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z \mathbb{1}_{Z \ge c_*/8}\right).$$

Using (47), we get

$$\int_{c_* s/d}^{c_*/8} \mathbf{P}_{\boldsymbol{\theta},F,\sigma}(Z > u)\, du \le \frac{2c_*}{\sqrt{d}} \int_0^\infty e^{-t^2} dt \le \frac{C}{\sqrt{d}}.$$

As $s < d/2$, one may check that $\hat{\sigma}^{2+\epsilon}_{\mathsf{med}} \le \left(\max_{i \notin S} (\sigma \xi_i)^2 / \beta\right)^{1+\epsilon/2} \le (\sigma^2/\beta)^{1+\epsilon/2} \sum_{i=1}^d |\xi_i|^{2+\epsilon}$. Since $\mathbf{E}|\xi_1|^{2+\epsilon} < \infty$ this yields $\mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z^{1+\epsilon}\right) \le Cd$. It follows that

$$\mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z \mathbb{1}_{Z \ge c_*/8}\right) \le \left(\mathbf{E}_{\boldsymbol{\theta},F,\sigma}\left(Z^{1+\epsilon}\right)\right)^{1/(1+\epsilon)} \mathbf{P}_{\boldsymbol{\theta},F,\sigma}(Z \ge c_*/8)^{\epsilon/(1+\epsilon)} \le Cde^{-d/C}.$$

Combining the last three displays yields the desired bound in expectation.

5.6. *Proof of part (i) of Proposition 7.* In this proof, we write for brevity $\mathbf{E} = \mathbf{E}_{\boldsymbol{\theta},\sigma,\mathcal{N}(0,1)}$ and $\mathbf{P} = \mathbf{P}_{\boldsymbol{\theta},\sigma,\mathcal{N}(0,1)}$. Set

$$\varphi_d(t) = \frac{1}{d} \sum_{i=1}^d e^{itY_j}, \quad \varphi(t) = \mathbf{E}(\varphi_d(t)), \quad \varphi_0(t) = e^{-\frac{t^2\sigma^2}{2}}.$$

Since $s/d < 1/8$ and $\varphi(t) = \varphi_0(t)\big(1 - \frac{|S|}{d} + \frac{1}{d}\sum_{j \in S} \exp(i\theta_j t)\big)$, we have

$$\frac{3}{4}\varphi_0(t) \le \Big(1 - \frac{2s}{d}\Big)\varphi_0(t) \le |\varphi(t)| \le \varphi_0(t). \tag{48}$$

Consider the events

$$\mathcal{B}_1 = \Big\{\sigma^2/2 \le \tilde{\sigma}^2 \le 3\sigma^2/2\Big\} \quad \text{and} \quad \mathcal{A}_u = \Big\{ \sup_{v \in \mathbb{R}} |\varphi_d(v) - \varphi(v)| \le \sqrt{\frac{u}{d}} \Big\}, \quad u > 0.$$

By Proposition 1, $\mathcal{B}_1$ holds with probability at least $1 - e^{-cd}$ if the tuning parameter γ in the definition of $\tilde{\sigma}^2$ is small enough. Using Hoeffding's inequality, it is not hard to check that $\mathcal{A}_u$ holds with probability at least $1 - 4e^{-u}$. Moreover,

$$\mathbf{E}\Big(\sqrt{d} \sup_{v \in \mathbb{R}} |\varphi_d(v) - \varphi(v)|\Big) \le C. \tag{49}$$

Notice that on the event $\mathcal{D} = \{|\varphi_d(\hat{t}_1)| > (es/\sqrt{d} + 1)^{-1}/4\}$ we have $\hat{\sigma}^2 = \tilde{v}^2 \le 2\tilde{\sigma}^2$. First, we bound the risk restricted to $\mathcal{D} \cap \mathcal{B}_1^c$. We have

$$\mathbf{E}\big(|\hat{\sigma}^2 - \sigma^2| \mathbb{1}_{\mathcal{D} \cap \mathcal{B}_1^c}\big) \le \mathbf{E}\big(|2\tilde{\sigma}^2 + \sigma^2| \mathbb{1}_{\mathcal{B}_1^c}\big).$$

Thus, using the Cauchy-Schwarz inequality and Proposition 1 we find

$$\mathbf{E}\big(|\hat\sigma^2-\sigma^2|\mathbb{1}_{\mathcal{D}\cap\mathcal{B}_1^c}\big)\le C\sigma^2 e^{-d/C}\le\frac{C'\sigma^2}{\sqrt{d}}. \tag{50}$$

Next, we bound the risk restricted to $\mathcal{D}^c$. It will be useful to note that $\mathcal{A}_{\log d}\cap\mathcal{B}_1\subset\mathcal{D}$. Indeed, on $\mathcal{A}_{\log d}\cap\mathcal{B}_1$, using the assumption $s<d/8$ we have

$$|\varphi_d(\hat t_1)|\ge\frac34\varphi_0(\hat t_1)-\sqrt{\frac{\log d}{d}}\ge\frac{3}{4(es/\sqrt{d}+1)^{1/3}}-\sqrt{\frac{\log d}{d}}>\frac{1}{4(es/\sqrt{d}+1)}.$$

Thus, applying again the Cauchy-Schwarz inequality and Proposition 1 we find

$$\begin{aligned}\mathbf{E}\big(|\hat\sigma^2-\sigma^2|\mathbb{1}_{\mathcal{D}^c}\big)=\mathbf{E}\big(|\tilde\sigma^2-\sigma^2|\mathbb{1}_{\mathcal{D}^c}\big)&\le\big(\mathbf{E}\big(|\tilde\sigma^2-\sigma^2|^2\big)\big)^{1/2}\left(\mathbf{P}(\mathcal{D}^c)\right)^{1/2}\\&\le C\sigma^2\sqrt{\mathbf{P}(\mathcal{A}_{\log d}^c)+\mathbf{P}(\mathcal{B}_1^c)}\le C\sigma^2\sqrt{\frac{4}{d}+e^{-cd}}\le\frac{C'\sigma^2}{\sqrt{d}}.\end{aligned} \tag{51}$$

To complete the proof, it remains to handle the risk restricted to the event $\mathcal{C}=\mathcal{D}\cap\mathcal{B}_1$. We will use the following decomposition

$$|\hat\sigma^2-\sigma^2|\le\Big|\frac{2\log(|\varphi_d(\hat t_1)|)}{\hat t_1^2}-\frac{2\log(|\varphi(\hat t_1)|)}{\hat t_1^2}\Big|+\Big|-\frac{2\log(|\varphi(\hat t_1)|)}{\hat t_1^2}-\sigma^2\Big|. \tag{52}$$

Since $-2\log(|\varphi_0(\hat t_1)|)/\hat t_1^2=\sigma^2$, it follows from (48) that

$$\Big|-\frac{2\log(|\varphi(\hat t_1)|)}{\hat t_1^2}-\sigma^2\Big|\le\frac{Cs}{d\,\hat t_1^2}=\frac{Cs\tilde\sigma^2}{d\log(4(es/\sqrt{d}+1))}.$$

Therefore,

$$\mathbf{E}\Big(\Big|-\frac{2\log(|\varphi(\hat t_1)|)}{\hat t_1^2}-\sigma^2\Big|\mathbb{1}_{\mathcal{C}}\Big)\le\frac{Cs\sigma^2}{d\log(es/\sqrt{d}+1)}. \tag{53}$$

Next, using the inequality

$$\big|\log(|\varphi_d(t)|)-\log(|\varphi(t)|)\big|\le\frac{|\varphi_d(t)-\varphi(t)|}{|\varphi(t)|\wedge|\varphi_d(t)|},\quad\forall t\in\mathbb{R},$$

we find

$$\begin{aligned}\Big|\frac{\log(|\varphi_d(\hat t_1)|)}{\hat t_1^2}-\frac{\log(|\varphi(\hat t_1)|)}{\hat t_1^2}\Big|\mathbb{1}_{\mathcal{C}}&\le\frac{\sup_{v\in\mathbb{R}}|\varphi_d(v)-\varphi(v)|}{\hat t_1^2|\varphi(\hat t_1)|\wedge|\varphi_d(\hat t_1)|}\mathbb{1}_{\mathcal{C}}\\&\le\frac{C\sigma^2U}{\sqrt{d}\,\log(es/\sqrt{d}+1)}\left(\frac{es}{\sqrt{d}}+1\right),\end{aligned}$$

where $U=\sqrt{d}\,\sup_{v\in\mathbb{R}}|\varphi_d(v)-\varphi(v)|$. Bounding $\mathbf{E}(U)$ by (49) we finally get

$$\mathbf{E}\left[\Big|\frac{\log(|\varphi_d(\hat t_1)|)}{\hat t_1^2}-\frac{\log(|\varphi(\hat t_1)|)}{\hat t_1^2}\Big|\mathbb{1}_{\mathcal{C}}\right]\le C\sigma^2\max\left(\frac{1}{\sqrt{d}},\frac{s}{d\log(es/\sqrt{d}+1)}\right). \tag{54}$$

We conclude by combining inequalities (50) - (54).

5.7. *Proof of Theorems 3 and 4.* Let $\|\boldsymbol{\theta}\|_0 \le s$ and denote by S the support of $\boldsymbol{\theta}$. Note first that, by the definition of $\hat\sigma^2$,

$$\frac{\sigma^2}{d}\sum_{i=1}^{d-2s}\xi^2_{(i)} \le \hat\sigma^2 \le \frac{\sigma^2}{d}\sum_{i\in S^c}\xi_i^2, \tag{55}$$

where $\xi^2_{(1)} \le \cdots \le \xi^2_{(d)}$ are the ordered values of $\xi_1^2,\dots,\xi_d^2$. Indeed, the right hand inequality in (55) follows from the relations

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

To show the left hand inequality in (55), notice that at least $d-2s$ among the $d-s$ order statistics $Y^2_{(1)},\dots,Y^2_{(d-s)}$ 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 $\sigma^2\xi^2_{(1)},\dots,\sigma^2\xi^2_{(d-2s)}$ among $\sigma^2\xi_1^2,\dots,\sigma^2\xi_d^2$.

Using (55) we get

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

so that

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

Then

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

Note that under assumption (2) we have $\mathbf{E}(\xi_1^4)<\infty$ and Lemmas 1 and 3 yield

$$\sum_{i=1}^{2s}\sqrt{\mathbf{E}\xi^4_{(d-i+1)}} \le \sqrt{C}\sum_{i=1}^{2s}\log^{2/a}\big(ed/i\big) \le C'\sqrt{C}s\log^{2/a}\Big(\frac{ed}{2s}\Big).$$

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

5.8. *Proof of Theorem 7.* With the same notation as in the proof of Theorem 1, we have

$$\hat\sigma^2-\sigma^2 = \frac{\sigma^2}{d}\big(\|\boldsymbol{\xi}\|_2^2-d\big) + \frac{1}{d}\left(\|\boldsymbol{u}\|_2^2 - 2\sigma\boldsymbol{u}^T\boldsymbol{\xi}\right). \tag{56}$$

It follows from (25) that

$$\|\boldsymbol{u}\|_2^2 + 2\sigma|\boldsymbol{u}^T\boldsymbol{\xi}| \le 3\sigma|\boldsymbol{u}^T\boldsymbol{\xi}| + \frac{\tilde\sigma}{2}\Big\{\Big(\sum_{j=1}^s\lambda_j^2\Big)^{1/2}\|\boldsymbol{u}\|_2 - \sum_{j=s+1}^d\lambda_j|u|_{(d-j+1)}\Big\}.$$

Arguing as in the proof of Theorem 1, we obtain

$$\|\boldsymbol{u}\|_2^2 + 2\sigma|\boldsymbol{u}^T\boldsymbol{\xi}| \le \Big(U_1 + \frac{\tilde\sigma}{2}\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2} + U_2\Big)\|\boldsymbol{u}\|_2,$$

where

$$U_1 = 3\sigma\Big(\sum_{j=1}^s |\xi|^2_{(d-j+1)}\Big)^{1/2}, \quad U_2 = \Big(\sum_{j=s+1}^d \Big(3\sigma|\xi|_{(d-j+1)} - \frac{\tilde\sigma}{2}\lambda_j\Big)_+^2\Big)^{1/2}$$

Using the Cauchy-Schwarz inequality, Proposition 1 and (28) and writing for brevity $\mathbf{E} = \mathbf{E}_{\boldsymbol{\theta},P_\xi,\sigma}$ we find

$$\mathbf{E}\Big(\tilde\sigma\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\|\boldsymbol{u}\|_2\Big) \le \Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\sqrt{\mathbf{E}(\tilde\sigma^2)}\sqrt{\mathbf{E}(\|\boldsymbol{u}\|_2^2)} \le C\sigma^2\sum_{j=1}^s \lambda_j^2.$$

Since $\mathbf{E}(\xi_1^4) < \infty$ we also have $\mathbf{E}\big|\|\boldsymbol{\xi}\|_2^2 - d\big| \le C\sqrt{d}$. Finally, using again (28) we get, for $k = 1, 2$,

$$\mathbf{E}(U_k\|\boldsymbol{u}\|_2) \le \sqrt{\mathbf{E}(\|\boldsymbol{u}\|_2^2)}\sqrt{\mathbf{E}(U_k^2)} \le \sigma\Big(\sum_{j=1}^s \lambda_j^2\Big)^{1/2}\sqrt{\mathbf{E}(U_k^2)} \le C\sigma^2\sum_{j=1}^s \lambda_j^2,$$

where the last inequality follows from the same argument as in the proof of Theorem 1. These remarks together with (56) imply

$$\mathbf{E}\left(|\hat\sigma^2 - \sigma^2|\right) \le \frac{C}{d}\Big(\sigma^2\sqrt{d} + \sigma^2\sum_{j=1}^s \lambda_j^2\Big).$$

We conclude the proof by bounding $\sum_{j=1}^s \lambda_j^2$ in the same way as in the end of the proof of Theorem 1.

6. Proofs of the lower bounds.

6.1. *Proof of Theorems 5 and 6 and part (ii) of Proposition 7.* Since we have $\ell(t) \ge \ell(A)\mathbb{1}_{t\ge A}$ for any $A > 0$, it is enough to prove the theorems for the indicator loss $\ell(t) = \mathbb{1}_{t\ge 1}$. This remark is valid for all the proofs of this section and will not be further repeated.

(i) We first prove the lower bounds with the rate $1/\sqrt{d}$ in Theorems 5 and 6. 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 7. 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 $\mathcal{G}_{a,\tau}$ with any $\tau > 0$, $a > 0$, and to $\mathcal{P}_{a,\tau}$ with any $\tau > 0$, $a \ge 2$. 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

$$H^2(\mathbf{P}_1, \mathbf{P}_0) = 2\big(1 - (1 - h^2/2)^d\big) \tag{57}$$

where $h^2 = \int(\sqrt{f_0(x)} - \sqrt{f_0(x/\sigma_1)/\sigma_1})^2dx$. By Theorem 7.6. in Ibragimov and Hasminskii [15],

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

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 $h^2 \le \bar{c} c_0^2/d$ where $\bar{c} > 0$ is a constant. This and (57) imply that for c_0 small enough we have $H(\mathbf{P}_1, \mathbf{P}_0) \le 1/2$. Finally, choosing such a small c_0 and using Theorem 2.2(ii) in Tsybakov [21] we obtain

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

(ii) We now prove the lower bound with the rate $\frac{s}{d}\log^{2/a}(ed/s)$ in Theorem 5. It is enough to conduct the proof for $s \ge s_0$ where $s_0 > 0$ is an arbitrary absolute constant. Indeed, for $s \le s_0$ we have $\frac{s}{d}\log^{2/a}(ed/s) \le C/\sqrt{d}$ where $C > 0$ is an absolute constant and thus Theorem 5 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 $s \ge 32$.

We take $P_\xi = U$ where U is the Rademacher distribution, that is the uniform distribution on $\{-1, 1\}$. Clearly, $U \in \mathcal{G}_{a,\tau}$. Let $\delta_1, \dots, \delta_d$ be i.i.d. Bernoulli random variables with probability of success $\mathbf{P}(\delta_1 = 1) = \frac{s}{2d}$, and let $\epsilon_1, \dots, \epsilon_d$ be i.i.d. Rademacher random variables that are independent of $(\delta_1, \dots, \delta_d)$. Denote by μ the distribution of $(\alpha\delta_1\epsilon_1, \dots, \alpha\delta_d\epsilon_d)$ where $\alpha = (\tau/2)\log^{1/a}(ed/s)$. Note that μ is not necessarily supported on $\Theta_s = \{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_0 \le s\}$ as the number of nonzero components of a vector drawn from μ can be larger than s. Therefore, we consider a restricted to Θ_s version of μ defined by

$$\bar{\mu}(A) = \frac{\mu(A \cap \Theta_s)}{\mu(\Theta_s)} \tag{58}$$

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

$$\mathbb{P}_\mu = \int \mathbf{P}_{\boldsymbol{\theta},U,1}\, \mu(d\boldsymbol{\theta}) \quad \text{and} \quad \mathbb{P}_{\bar{\mu}} = \int \mathbf{P}_{\boldsymbol{\theta},U,1}\, \bar{\mu}(d\boldsymbol{\theta}). \tag{59}$$

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

$$\mathbb{P}_\mu = \mathbf{P}_{0,\tilde{P},\sigma_0} \tag{60}$$

where $\sigma_0 > 0$ is defined by

$$\sigma_0^2 = 1 + \frac{\tau^2 s}{8d}\log^{2/a}(ed/s) \le 1 + \frac{\tau^2}{8}. \tag{61}$$

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 ϵ, ξ, δ are jointly independent. Thus, to prove (60) it is enough to show that, for all $t \ge 2$,

$$\mathbf{P}\big((\tau/2)\log^{1/a}(ed/s)\,\delta\epsilon + \xi > t\sigma_0\big) \le e^{-(t/\tau)^a}. \tag{62}$$

But this inequality immediately follows from the fact that for $t \ge 2$ the probability in (62) is smaller than

$$\mathbf{P}(\epsilon = 1, \delta = 1)\,\mathbb{1}_{(\tau/2)\log^{1/a}(ed/s) > t-1} \le \frac{s}{4d}\mathbb{1}_{\tau\log^{1/a}(ed/s) > t} \le e^{-(t/\tau)^a}. \tag{63}$$

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

$$\begin{aligned}
&\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\Big(\Big|\frac{\hat{T}}{\sigma^2} - 1\Big| \ge u\Big) \\
&\qquad \ge \max\Big\{\mathbf{P}_{0,\tilde{P},\sigma_0}(|\hat{T} - \sigma_0^2| \ge \sigma_0^2 u), \int \mathbf{P}_{\boldsymbol{\theta},U,1}(|\hat{T} - 1| \ge u)\bar{\mu}(d\boldsymbol{\theta})\Big\} \\
&\qquad \ge \max\Big\{\mathbb{P}_\mu(|\hat{T} - \sigma_0^2| \ge \sigma_0^2 u), \mathbb{P}_{\bar{\mu}}(|\hat{T} - 1| \ge \sigma_0^2 u)\Big\} \qquad (64)
\end{aligned}$$

where the last inequality uses (60). Write $\sigma_0^2 = 1 + 2\phi$ where $\phi = \frac{\tau^2 s}{16d}\log^{2/a}(ed/s)$ and choose $u = \phi/\sigma_0^2 \ge \phi/(1 + \tau^2/8)$. Then, the expression in (64) 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 [21] yields

$$\max\Big\{\mathbb{P}_\mu(|\hat{T} - \sigma_0^2| \ge \phi), \mathbb{P}_{\bar{\mu}}(|\hat{T} - 1| \ge \phi)\Big\} \ge \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar{\mu}})}{2} \qquad (65)$$

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 (65) and Lemma 5 for any $s \ge 32$.

(iii) Finally, we prove the lower bound with the rate $\tau^2(s/d)^{1-2/a}$ in Theorem 6. Again, we do not consider the case $s \le 32$ since in this case the rate $1/\sqrt{d}$ is dominating and Theorem 6 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 σ_0^2 there exists $\tilde{P} \in \mathcal{P}_{a,\tau}$ such that (60) holds. Indeed, analogously to (62) we now have, for all $t \ge 2$,

$$\mathbf{P}\big(\alpha\,\delta\epsilon + \xi > t\sigma_0\big) \le \mathbf{P}(\epsilon = 1, \delta = 1)\,\mathbb{1}_{(\tau/2)(d/s)^{1/a} > t-1} \le \frac{s}{4d}\mathbb{1}_{\tau(d/s)^{1/a} > t} \le (t/\tau)^a. \qquad (66)$$

To finish the proof, it remains to repeat the argument of (64) and (65) with $\phi = \frac{\tau^2(s/d)^{1-2/a}}{16}$.

6.2. *Proof of Theorem 2.* We argue similarly to the proof of Theorems 5 and 6, in particular, we set $\alpha = (\tau/2)\log^{1/a}(ed/s)$ when proving the bound on the class $\mathcal{G}_{a,\tau}$, and $\alpha = (\tau/2)(d/s)^{1/a}$ when proving the bound on $\mathcal{P}_{a,\tau}$. In what follows, we only deal with the class $\mathcal{G}_{a,\tau}$ since the proof for $\mathcal{P}_{a,\tau}$ is analogous. Consider the measures μ $\bar{\mu}$, $\mathbb{P}_\mu$, $\mathbb{P}_{\bar{\mu}}$ and $\tilde{P}$ defined in Section 6.1. Similarly to (64), for any estimator $\hat{T}$ and any $u > 0$ we have

$$\begin{aligned}
&\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta},P_\xi,\sigma}\big(|\hat{T} - \|\boldsymbol{\theta}\|_2| \ge \sigma u\big) \\
&\qquad \ge \max\Big\{\mathbf{P}_{0,\tilde{P},\sigma_0}(|\hat{T}| \ge \sigma_0 u), \int \mathbf{P}_{\boldsymbol{\theta},U,1}(|\hat{T} - \|\boldsymbol{\theta}\|_2| \ge u)\bar{\mu}(d\boldsymbol{\theta})\Big\} \\
&\qquad \ge \max\Big\{\mathbb{P}_\mu(|\hat{T}| \ge \sigma_0 u), \mathbb{P}_{\bar{\mu}}(|\hat{T} - \|\boldsymbol{\theta}\|_2| \ge \sigma_0 u)\Big\} \\
&\qquad \ge \max\Big\{\mathbb{P}_\mu(|\hat{T}| \ge \sigma_0 u), \mathbb{P}_{\bar{\mu}}(|\hat{T}| < \sigma_0 u, \|\boldsymbol{\theta}\|_2 \ge 2\sigma_0 u)\Big\} \\
&\qquad \ge \min_B \max\big\{\mathbb{P}_\mu(B), \mathbb{P}_{\bar{\mu}}(B^c) - \bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u)\big\} \\
&\qquad \ge \min_B \frac{\mathbb{P}_\mu(B) + \mathbb{P}_{\bar{\mu}}(B^c)}{2} - \frac{\bar{\mu}(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u)}{2} \qquad (67)
\end{aligned}$$

where σ_0 is defined in (61), 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 (60) and to the inequality $\sigma_0 \geq 1$. Since $\min_B \left\{\mathbb{P}_\mu(B) + \mathbb{P}_{\bar\mu}(B^c)\right\} = 1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu})$, we get

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\sigma>0} \sup_{\|\boldsymbol{\theta}\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}\big(|\hat T - \|\boldsymbol{\theta}\|_2|/\sigma \geq u\big) \geq \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) - \bar\mu(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u)}{2}. \tag{68}$$

Consider first the case $s \geq 32$. Set $u = \frac{\alpha\sqrt{s}}{4\sigma_0}$. Then (77) and (80) imply that

$$V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) \leq e^{-\frac{3s}{16}}, \quad \bar\mu(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u) \leq 2e^{-\frac{s}{16}},$$

which, together with (68) and the fact that $s \geq 32$ yields the result.

Let now $s < 32$. Then we set $u = \frac{\alpha\sqrt{s}}{8\sqrt{2}\sigma_0}$. It follows from (78) and (81) that

$$1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) - \bar\mu(\|\boldsymbol{\theta}\|_2 < 2\sigma_0 u) \geq \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) = 1\Big) = \frac{s}{2}\Big(1 - \frac{s}{2d}\Big)^{d-1}.$$

It is not hard to check that the minimum of the last expression over all integers s, d such that $1 \leq s < 32$, $s \leq d$, is bounded from below by a positive number independent of d. We conclude by combining these remarks with (68).

6.3. *Proof of part (ii) of Proposition 4 and part (ii) of Proposition 5.* We argue similarly to the proof of Theorems 5 and 6, in particular, we set $\alpha = (\tau/2)\log^{1/a}(ed/s)$ when proving the bound on the class $\mathcal{G}_{a,\tau}$, and $\alpha = (\tau/2)(d/s)^{1/a}$ when proving the bound on $\mathcal{P}_{a,\tau}$. In what follows, we only deal with the class $\mathcal{G}_{a,\tau}$ since the proof for $\mathcal{P}_{a,\tau}$ is analogous. Without loss of generality we assume that $\sigma = 1$.

To prove the lower bound with the rate $\phi^\circ_{\mathsf{exp}}(s,d)$, we only need to prove it for s such that $(\phi^\circ_{\mathsf{exp}}(s,d))^2 \leq c_0\sqrt{d}/\log^{2/a}(ed)$ with any small absolute constant $c_0 > 0$, since the rate is increasing with s.

Consider the measures μ $\bar\mu$, $\mathbb{P}_\mu$, $\mathbb{P}_{\bar\mu}$ defined in Section 6.1 with $\sigma_0 = 1$. Let ξ_1 be distributed with c.d.f. F_0 defined in item (i) of the proof of Theorems 5 and 6. Using the notation as in the proof of Theorems 5 and 6, we define $\tilde P$ as the distribution of $\tilde\xi_1 = \sigma_1\xi_1 + \alpha\delta_1\epsilon_1$ with $\sigma_1^2 = (1 + \alpha^2 s/(2d))^{-1}$ where now δ_1 is the Bernoulli random variable with $\mathbf{P}(\delta_1 = 1) = \frac{s}{2d}(1 + \alpha^2 s/(2d))^{-1}$. By construction, $\mathbf{E}\tilde\xi_1 = 0$ and $\mathbf{E}\tilde\xi_1^2 = 1$. Since the support of F_0 is in $[-3/2, 3/2]$ one can check as in item (ii) of the proof of Theorems 5 and 6 that $\tilde P \in \mathcal{G}_{a,\tau}$. Next, analogously to (67) - (68) we obtain that, for any $u > 0$,

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\theta\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, 1}\big(|\hat T - \|\boldsymbol{\theta}\|_2| \geq u\big) \geq \frac{1 - V(\mathbb{P}_{\bar\mu}, P_{0,\tilde P,1}) - \bar\mu(\|\boldsymbol{\theta}\|_2 < 2u)}{2}.$$

Let $\mathbf{P}_0$ and $\mathbf{P}_1$ denote the distributions of $(\xi_1, \dots, \xi_d)$ and of $(\sigma_1\xi_1, \dots, \sigma_1\xi_d)$, respectively. Acting as in item (i) of the proof of Theorems 5 and 6 and using the bound

$$|1 - \sigma_1| \leq \alpha^2 s/d = \frac{\tau^2}{4}\frac{s}{d}\log^{2/a}(ed/s) \leq Cc_0/\sqrt{d}$$

we find that $V(\mathbf{P}_0, \mathbf{P}_1) \leq H(\mathbf{P}_0, \mathbf{P}_1) \leq 2\kappa c_0^2$ for some $\kappa > 0$. Therefore, $V(\mathbb{P}_\mu, P_{0,\tilde P,1}) = V(\mathbf{P}_0 * \mathbf{Q}, \mathbf{P}_1 * \mathbf{Q}) \leq V(\mathbf{P}_0, \mathbf{P}_1) \leq 2\kappa c_0^2$ where $\mathbf{Q}$ denotes the distribution of $(\alpha\delta_1\epsilon_1, \dots, \alpha\delta_d\epsilon_d)$. This bound and the fact that $V(\mathbb{P}_{\bar\mu}, P_{0,\tilde P,1}) \leq V(\mathbb{P}_{\bar\mu}, \mathbb{P}_\mu) + V(\mathbb{P}_\mu, P_{0,\tilde P,1})$ imply

$$\sup_{P_\xi \in \mathcal{G}_{a,\tau}} \sup_{\|\theta\|_0 \leq s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, 1}\big(|\hat T - \|\boldsymbol{\theta}\|_2| \geq u\big) \geq \frac{1 - V(\mathbb{P}_\mu, \mathbb{P}_{\bar\mu}) - \bar\mu(\|\boldsymbol{\theta}\|_2 < 2u)}{2} - \kappa c_0^2.$$

We conclude by repeating the argument after (68) in the proof of Theorem 2 and choosing $c_0 > 0$ small enough to guarantee that the right hand side of the last display is positive.

6.4. *Proof of part (ii) of Proposition 7.* The lower bound with the rate $1/\sqrt{d}$ follows from the argument as in item (i) of the proof of Theorems 5 and 6 if we replace there F_0 by the standard Gaussian distribution. The lower bound with the rate $\frac{s}{d(1+\log_+(s^2/d))}$ follows from Lemma 8 and the lower bound for estimation of $\|\boldsymbol{\theta}\|_2$ in Proposition 3.

6.5. *Proof of Proposition 8.* Assume that $\boldsymbol{\theta} = 0$, $\sigma = 1$ and set

$$\xi_i = \sqrt{3}\epsilon_i u_i,$$

where the ϵ_i's and the u_i are independent, with Rademacher and uniform distribution on $[0,1]$ respectively. Then note that

$$\mathbf{E}_{0,P_\xi,1}\big(\hat\sigma_*^2 - 1\big)^2 \geq \big(\mathbf{E}_{0,P_\xi,1}(\hat\sigma_*^2) - 1\big)^2 = \Big(\mathbf{E}_{0,P_\xi,1}\Big\{\hat\sigma_*^2 - \frac{3}{d}\sum_{i=1}^d u_i^2\Big\}\Big)^2, \tag{69}$$

since $\mathbf{E}(u_i^2) = 1/3$. Note also that $\hat\sigma_*^2 = \frac{3}{d/2}\sum_{i=1}^{d/2} u_{(i)}^2$. Now,

$$\begin{aligned}
\frac{1}{d/2}\sum_{i=1}^{d/2} u_{(i)}^2 - \frac{1}{d}\sum_{i=1}^{d} u_i^2 &= \frac{1}{d}\sum_{i=1}^{d/2} u_{(i)}^2 - \frac{1}{d}\sum_{i=d/2+1}^{d} u_{(i)}^2 \\
&\leq \frac{1}{d}\sum_{i=1}^{d/4} u_{(i)}^2 - \frac{1}{d}\sum_{i=3d/4+1}^{d} u_{(i)}^2 \\
&\leq \frac{1}{4}(u_{(d/4)}^2 - u_{(3d/4)}^2).
\end{aligned}$$

Since $u_{(i)}$ follows a Beta distribution with parameters $(i, d-i+1)$ we have $\mathbf{E}(u_{(i)}^2) = \frac{i(i+1)}{(d+1)(d+2)}$, and

$$\mathbf{E}_{0,P_\xi,1}\Big(\frac{1}{d/2}\sum_{i=1}^{d/2} u_{(i)}^2 - \frac{1}{d}\sum_{i=1}^{d} u_i^2\Big) \leq \frac{1}{4}\mathbf{E}_{0,P_\xi,1}(u_{(d/4)}^2 - u_{(3d/4)}^2) = -\frac{d}{8(d+2)} \leq -\frac{1}{24}.$$

This and (69) prove the proposition.

7. Lemmas.

7.1. *Lemmas for the upper bounds.*

LEMMA 1. *Let $z_1, \dots, z_d \overset{iid}{\sim} P$ with $P \in \mathcal{G}_{a,\tau}$ for some $a, \tau > 0$ and let $z_{(1)} \leq \cdots \leq z_{(d)}$ be the order statistics of $|z_1|, \dots, |z_d|$. Then for $u > 2^{1/a}\tau \vee 2$, we have*

$$\mathbf{P}\Big(z_{(d-j+1)} \leq u\log^{1/a}\big(ed/j\big), \forall\, j = 1, \dots, d\Big) \geq 1 - 4e^{-u^a/2}, \tag{70}$$

and, for any $r > 0$,

$$\mathbf{E}\big(z_{(d-j+1)}^r\big) \leq C\log^{r/a}\big(ed/j\big), \qquad j = 1, \dots, d, \tag{71}$$

where $C > 0$ is a constant depending only on τ, a and r.

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

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

Thus, for $v \geq 2^{1/a} \vee (2/\tau)$ we have

$$\mathbf{P}(z_{(d-j+1)} \geq v\tau \log^{1/a}(ed/j)) \leq 2\Big(\frac{ed}{j}\Big)^{j(1-v^a)} \leq 2e^{-jv^a/2}, \qquad j = 1, \dots, d, \tag{72}$$

and

$$\mathbf{P}\Big(\exists\, j \in \{1, \dots, d\} : z_{(d-j+1)} \geq v\tau \log^{1/a}(ed/j)\Big) \leq 2\sum_{j=1}^{d} e^{-jv^a/2} \leq 4e^{-v^a/2}$$

implying (70). Finally, (71) follows by integrating (72). □

LEMMA 2. *Let $z_1, \dots, z_d \overset{iid}{\sim} P$ with $P \in \mathcal{P}_{a,\tau}$ for some $a, \tau > 0$ and let $z_{(1)} \leq \dots \leq z_{(d)}$ be the order statistics of $|z_1|, \dots, |z_d|$. Then for $u > (2e)^{1/a}\tau \vee 2$, we have*

$$\mathbf{P}\Big(z_{(d-j+1)} \leq u\Big(\frac{d}{j}\Big)^{1/a}, \forall\, j = 1, \dots, d\Big) \geq 1 - \frac{2e\tau^a}{u^a} \tag{73}$$

and, for any $r \in (0, a)$,

$$\mathbf{E}\big(z^r_{(d-j+1)}\big) \leq C\Big(\frac{d}{j}\Big)^{r/a}, \qquad j = 1, \dots, d, \tag{74}$$

where $C > 0$ is a constant depending only on τ, a and r.

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

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

Set $t_j = u\Big(\frac{d}{j}\Big)^{1/a}$ and $q = e(\tau/u)^a$. The assumption on u yields that $q < 1/2$, so that

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

This proves (73). The proof of (74) is analoguous to that of (71). □

LEMMA 3. *For all $a > 0$ and all integers $1 \leq s \leq d$,*

$$\sum_{i=1}^{s} \log^{2/a}\big(ed/i\big) \leq Cs \log^{2/a}\Big(\frac{ed}{s}\Big)$$

where $C > 0$ depends only on a.

The proof is simple and we omit it.

7.2. *Lemmas for the lower bounds.* For two probability measures P_1 and P_2 on a measurable space $(\Omega,\mathcal{U})$, we denote by $V(\mathrm{P}_1,\mathrm{P}_2)$ the total variation distance between P_1 and P_2:

$$V(\mathrm{P}_1,\mathrm{P}_2)=\sup_{B\in\mathcal{U}}|\mathrm{P}_1(B)-\mathrm{P}_2(B)|\,.$$

LEMMA 4 (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 $\lambda>0$,*

$$\mathbf{P}\big(\mathcal{B}(d,p)\geq\lambda\sqrt{d}+dp\big)\leq\exp\bigg(-\frac{\lambda^2}{2p(1-p)\big(1+\frac{\lambda}{3p\sqrt{d}}\big)}\bigg), \tag{75}$$

$$\mathbf{P}\big(\mathcal{B}(d,p)\leq-\lambda\sqrt{d}+dp\big)\leq\exp\bigg(-\frac{\lambda^2}{2p(1-p)}\bigg). \tag{76}$$

Inequality (75) is a combination of formulas (3) and (10) on pages 440–441 in [19]. Inequality (76) is formula (6) on page 440 in [19].

LEMMA 5. *Let $\mathbb{P}_\mu$ and $\mathbb{P}_{\bar\mu}$ be the probability measures defined in (59). The total variation distance between these two measures satisfies*

$$V(\mathbb{P}_\mu,\mathbb{P}_{\bar\mu})\leq\mathbf{P}\Big(\mathcal{B}\Big(d,\frac{s}{2d}\Big)>s\Big)\leq e^{-\frac{3s}{16}}, \tag{77}$$

and

$$V(\mathbb{P}_\mu,\mathbb{P}_{\bar\mu})\leq 1-\mathbf{P}\Big(\mathcal{B}\Big(d,\frac{s}{2d}\Big)=0\Big)-\mathbf{P}\Big(\mathcal{B}\Big(d,\frac{s}{2d}\Big)=1\Big). \tag{78}$$

PROOF. We have

$$V(\mathbb{P}_\mu,\mathbb{P}_{\bar\mu})=\sup_B\bigg|\int\mathbf{P}_{\boldsymbol{\theta},U,1}(B)d\mu(\boldsymbol{\theta})-\int\mathbf{P}_{\boldsymbol{\theta},U,1}(B)d\bar\mu(\boldsymbol{\theta})\bigg|\leq\sup_{|f|\leq1}\bigg|\int f d\mu-\int f d\bar\mu\bigg|=V(\mu,\bar\mu).$$

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,

$$V(\mathbb{P}_\mu,\mathbb{P}_{\bar\mu})\leq\mu(\Theta_s^c)=\mathbf{P}\Big(\mathcal{B}\Big(d,\frac{s}{2d}\Big)>s\Big). \tag{79}$$

Combining this inequality with (75) we obtain (77). To prove (78), we use again (79) and notice that $\mathbf{P}\Big(\mathcal{B}\Big(d,\frac{s}{2d}\Big)>s\Big)\leq\mathbf{P}\Big(\mathcal{B}\Big(d,\frac{s}{2d}\Big)\geq2\Big)$ for any integer $s\geq1$. □

LEMMA 6. *Let $\bar{\mu}$ be defined in* (58) *with some $\alpha > 0$. Then*

$$\bar{\mu}\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha}{2}\sqrt{s}\Big) \le 2e^{-\frac{s}{16}}, \tag{80}$$

and, for all $s \le 32$,

$$\bar{\mu}\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha\sqrt{s}}{4\sqrt{2}}\Big) = \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) = 0\Big). \tag{81}$$

PROOF. First, note that

$$\mu\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha}{2}\sqrt{s}\Big) = \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) < \frac{s}{4}\Big) \le e^{-\frac{s}{16}} \tag{82}$$

where the last inequality follows from (76). Next, inspection of the proof of Lemma 5 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 (82) proves (80). To prove (81), it suffices to note that $\mu\Big(\|\boldsymbol{\theta}\|_2 < \frac{\alpha\sqrt{s}}{4\sqrt{2}}\Big) = \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) < \frac{s}{32}\Big)$. □

LEMMA 7. *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 $[-3/2, 3/2]$. □

LEMMA 8. *Let $\tau > 0$, $a > 4$ and let s, d be integers satisfying $1 \le s \le d$. Let $\mathcal{P}$ be any subset of $\mathcal{P}_{a,\tau}$. Assume that for some function $\phi(s, d)$ of s and d and for some positive constants c_1, c_2, c_1', c_2' we have*

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\left| \frac{\hat{T}}{\sigma^2} - 1 \right| \ge \frac{c_1}{\sqrt{d}} \right) \ge c_1', \tag{83}$$

and

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\left| \frac{\hat{T} - \|\boldsymbol{\theta}\|_2}{\sigma} \right| \ge c_2 \phi(s, d) \right) \ge c_2'. \tag{84}$$

Then

$$\inf_{\hat{T}} \sup_{P_\xi \in \mathcal{P}} \sup_{\sigma > 0} \sup_{\|\boldsymbol{\theta}\|_0 \le s} \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma} \left(\left| \frac{\hat{T}}{\sigma^2} - 1 \right| \ge c_3 \max\left(\frac{1}{\sqrt{d}}, \frac{\phi^2(s, d)}{d} \right) \right) \ge c_3'$$

for some constants $c_3, c_3' > 0$.

PROOF. Let $\hat{\sigma}^2$ be an arbitrary estimator of σ^2. Based on $\hat{\sigma}^2$, we can construct an estimator $\hat{T} = \hat{N}^*$ of $\|\boldsymbol{\theta}\|_2$ defined by formula (11), case $s > \sqrt{d}$. It follows from (30), (31) and (84) that

$$c_2' \le \mathbf{P}\left(2|(\boldsymbol{\theta}, \boldsymbol{\xi})| \ge c_2\|\boldsymbol{\theta}\|_2\phi(s,d)/3\right) + \mathbf{P}\left(\sqrt{|\|\boldsymbol{\xi}\|_2^2 - d|} \ge c_2\phi(s,d)/3\right)$$
$$+ \mathbf{P}\left(\sqrt{d\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right|} \ge c_2\phi(s,d)/3\right),$$

where we write for brevity $\mathbf{P} = \mathbf{P}_{\boldsymbol{\theta}, P_\xi, \sigma}$. Hence

$$\mathbf{P}\left(\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| \ge c_2^2\phi^2(s,d)/(9d)\right) \ge c_2' - c^* \max\left(\frac{d}{\phi^4(s,d)}, \frac{1}{\phi^2(s,d)}\right)$$

for some constant $c^* > 0$ depending only on a and τ. If $\phi^2(s,d) > \max\left(\sqrt{\frac{2c^*d}{c_2'}}, \frac{2c^*}{c_2'}\right)$, then

$$\mathbf{P}\left(\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| \ge C\max\left(\frac{1}{\sqrt{d}}, \frac{\phi^2(s,d)}{d}\right)\right) \ge c_2'/2.$$

If $\phi^2(s,d) \le \max\left(\sqrt{\frac{2c^*d}{c_2'}}, \frac{2c^*}{c_2'}\right)$, then $\max\left(\frac{1}{\sqrt{d}}, \frac{\phi^2(s,d)}{d}\right)$ is of order $\frac{1}{\sqrt{d}}$ and the result follows from (83). □

8. Acknowledgements. The work of O. Collier has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The work of M.Ndaoud and 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. *Annals of Statistics*, **46**, 3603–3642, 2018.
[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] D. BELOMESTNY, M. TRABS AND A.B. TSYBAKOV Sparse covariance matrix estimation in high-dimensional deconvolution. arXiv preprint, arXiv:1710.10870. To appear in *Bernoulli*.
[4] C. BUTUCEA AND C. MATIAS. Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, **11**, 309–340, 2005.
[5] T.T. CAI AND J. JIN. Optimal rates of convergence for estimating the null and proportion of non-null effects in large-scale multiple testing. *Annals of Statistics*, **38**, 100–145, 2010.
[6] A. CARPENTIER AND N. VERZELEN. Adaptive estimation of the sparsity in the Gaussian vector model. *Annals of Statistics*, **47**, 93–126, 2019.
[7] M. CHEN, C. GAO AND Z. REN. Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics*, **46**, 1932–1960, 2018.
[8] 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.
[9] O. COLLIER, L. COMMINGES, A.B. TSYBAKOV AND N. VERZELEN. Optimal adaptive estimation of linear functionals under sparsity. *Annals of Statistics*, **46**, 3130–3150, 2018.
[10] D.L. DONOHO, I.M. JOHNSTONE, J.C. HOCH AND A.S. STERN. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B* , **54**, 41–81, 1992.
[11] 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.

[12] Y. Golubev and E. Krymova. On estimation of the noise variance in high-dimensional linear models. arXiv preprint, arXiv:1711.09208.
[13] Z. Guo, W. Wang, T.T. Cai and H. Li. Optimal estimation of genetic relatedness in high-dimensional linear models. *J. of the American Statist. Assoc.*, published online, 2018.
[14] P.J. Huber. *Robust Statistics,* J. Wiley, 1981.
[15] I.A.Ibragimov and R.Z.Hasminskii. *Statistical Estimation. Asymptotic Theory,* Springer, New York, 1981.
[16] L.Janson, R. Foygel Barber and E.Candes. EigenPrism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society. Series B* , **79**, 1037–1065, 2017.
[17] S. Minsker and X. Wei. Estimation of the covariance structure of heavy-tailed distributions. arXiv preprint, arXiv:1708.00502.
[18] V.V. Petrov. *Limit Theorems of Probability Theory.* Clarendon Press, Oxford, 1995.
[19] G. Shorack and J. Wellner. *Empirical Processes with Applications to Statistics,* John Wiley, New York, 1986.
[20] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, **99**, 879–898, 2012.
[21] A.B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, New York, 2009.
[22] N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic J. of Statist.* **6** 38–90, 2012.
[23] N. Verzelen and E. Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli* **24**, 3683–3710, 2018.
[24] L. Wasserman. *All of Statistics.* Springer, New York, 2005.

CEREMADE, Université Paris-Dauphine
PSL Research University
75016 Paris, France
E-mail: laetitia.comminges@dauphine.fr

Modal'X, UPL, Université Paris Nanterre
92000 Nanterre France
E-mail: olivier.collier@parisnanterre.fr

CREST (UMR CNRS 9194), ENSAE
5, av. Henry Le Chatelier, 91764 Palaiseau, France
E-mail: ndaoudm@gmail.com; alexandre.tsybakov@ensae.fr