OPTIMAL ADAPTIVE ESTIMATION OF LINEAR FUNCTIONALS UNDER SPARSITY

O Collier, Laëtitia Comminges, A B Tsybakov, Nicolas Verzelen

OPTIMAL ADAPTIVE ESTIMATION OF LINEAR FUNCTIONALS UNDER SPARSITY

O. COLLIER, L. COMMINGES, A.B. TSYBAKOV AND N. VERZELEN

Abstract. We consider the problem of estimation of a linear functional in the Gaussian sequence model where the unknown vector $\theta \in \mathbb{R}^d$ belongs to a class of s-sparse vectors with unknown s. We suggest an adaptive estimator achieving a non-asymptotic rate of convergence that differs from the minimax rate at most by a logarithmic factor. We also show that this optimal adaptive rate cannot be improved when s is unknown. Furthermore, we address the issue of simultaneous adaptation to s and to the variance σ^2 of the noise. We suggest an estimator that achieves the optimal adaptive rate when both s and σ^2 are unknown.

1. Introduction

We consider the model

$$y_j = \theta_j + \sigma\xi_j, \quad j = 1, \dots, d, \tag{1}$$

where $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ is an unknown vector of parameters, ξ_j are i.i.d. standard normal random variables, and $\sigma > 0$ is the noise level. We study the problem of estimation of the linear functional

$$L(\theta) = \sum_{i=1}^{d} \theta_i,$$

based on the observations $y = (y_1, \dots, y_d)$.

For $s \in \{1, \dots, d\}$, we denote by Θ_s the class of all $\theta \in \mathbb{R}^d$ satisfying $\|\theta\|_0 \le s$, where $\|\theta\|_0$ denotes the number of non-zero components of θ. We assume that θ belongs to Θ_s for some $s \in \{1, \dots, d\}$. Parameter s characterizes the sparsity of vector θ. The problem of estimation of $L(\theta)$ in this context arises, for example, if one wants to estimate the value of a function f at a fixed point from noisy observations of its Fourier coefficients knowing that the function admits a sparse representation with respect to the first d functions of the Fourier basis. Indeed, in this case the value $f(0)$ is equal to the sum of Fourier coefficients of f with even indices.

As a measure of quality of an estimator $\hat{T}$ of the functional $L(\theta)$ based on the sample $(y_1, \dots, y_d)$, we consider the maximum squared risk

$$\psi_s^{\hat{T}} \triangleq \sup_{\theta \in \Theta_s} \mathbf{E}_\theta (\hat{T} - L(\theta))^2,$$

where $\mathbf{E}_\theta$ denotes the expectation with respect to the distribution $\mathbf{P}_\theta$ of $(y_1, \dots, y_d)$ satisfying (1). For each fixed $s \in \{1, \dots, d\}$, the best quality of estimation is characterized by the minimax risk

$$\psi_s^* \triangleq \inf_{\hat{T}} \sup_{\theta \in \Theta_s} \mathbf{E}_\theta (\hat{T} - L(\theta))^2,$$

Date: November 29, 2016.

where the infimum is taken over all estimators. An estimator T^* is called rate optimal on Θ_s if $\psi_s^{T^*} \asymp \psi_s^*$. Here and in the following we write $a(d,s,\sigma) \asymp b(d,s,\sigma)$ for two functions $a(\cdot)$ and $b(\cdot)$ of d, s and σ if there exist absolute constants $c > 0$ and $c' > 0$ such that $c < a(d,s,\sigma)/b(d,s,\sigma) < c'$ for all d, all $s \in \{1,\dots,d\}$ and all $\sigma > 0$.

The problem of estimation of the linear functional from the minimax point of view has been analyzed in [6, 1, 2, 4, 5, 8] among others. Most of these papers study minimax estimation of linear functionals on classes of vectors θ different from Θ_s. Namely, θ is considered as a vector of first d Fourier or wavelet coefficients of functions belonging to some smoothness class, such as Sobolev or Besov classes. In particular, the class of vectors θ is assumed to be convex, which is not the case of class Θ_s. Cai and Low [1] were the first to address the problem of constructing rate optimal estimators of $L(\theta)$ on the sparsity class Θ_s and evaluating the minimax risk ψ_s^*. They studied the case $s < d^a$ for some $a < 1/2$, with $\sigma = 1/\sqrt{d}$, and established upper and lower bounds on ψ_s^* that are accurate up to a logarithmic factor in d. The sharp non-asymptotic expression for the minimax risk ψ_s^* is derived in [3] where it is shown that, for all d, all $s \in \{1,\dots,d\}$ and all $\sigma > 0$

$$\psi_s^* \asymp \sigma^2 s^2 \log(1 + d/s^2).$$

Furthermore, [3] proves that a simple estimator of the form

$$\hat{L}_s^* = \begin{cases} \sum_{j=1}^d y_j \mathbb{1}_{y_j^2 > 2\sigma^2 \log(1+d/s^2)}, & \text{if } s < \sqrt{d}, \\ \sum_{j=1}^d y_j, & \text{otherwise,} \end{cases} \tag{2}$$

is rate optimal. Here and in the following, $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function.

Note that the minimax risk ψ_s^* critically depends on the parameter s that in practice is usually unknown. More importantly, the rate optimal estimator $\hat{L}_s^*$ depends on s as well, which makes it inaccessible in practice.

In this paper, we suggest adaptive estimators of $L(\theta)$ that do not depend on s and achieve a non-asymptotic rate of convergence $\Phi^L(\sigma, s)$ that differs from the minimax rate ψ_s^* at most by a logarithmic factor. We also show that this rate cannot be improved when s is unknown in the sense of the definition that we give in Section 2 below. Furthermore, in Section 3 we address the issue of simultaneous adaptation to s and σ. We suggest an estimator that achieves the best rate of adaptive estimation $\Phi^L(\sigma, s)$ when both s and σ are unknown.

2. Main results

We assume throughout the paper that $d \geq 3$. Our aim is to show that the optimal adaptive rate of convergence is of the form

$$\Phi^L(\sigma, s) = \sigma^2 s^2 \log(1 + d(\log d)/s^2)$$

and to construct an adaptive estimator attaining this rate. Note that

$$\Phi^L(\sigma, s) \asymp \sigma^2 d(\log d), \quad \text{for all } \sqrt{d \log d} \leq s \leq d. \tag{3}$$

Indeed, since the function $x \mapsto x \log(1 + 1/x)$ is increasing for $x > 0$,

$$d(\log d)/2 \leq s^2 \log(1 + d(\log d)/s^2) \leq d(\log d), \quad \forall\ \sqrt{d \log d} \leq s \leq d,\ d \geq 3. \tag{4}$$

To construct an adaptive estimator, we first consider a collection of non-adaptive estimators indexed by $s = 1, \dots, d$:

$$\hat{L}_s = \begin{cases} \sum_{j=1}^d y_j \mathbb{1}_{y_j^2 > \alpha\sigma^2 \log(1+d(\log d)/s^2)}, & \text{if } s \le \sqrt{d \log d}, \\ \sum_{j=1}^d y_j, & \text{otherwise,} \end{cases} \tag{5}$$

where $\alpha > 0$ is a constant that will be chosen large enough. Note that if in definition (5) we replace $d(\log d)$ by d, and α by 2, we obtain the estimator $\hat{L}_s^*$ suggested in [3], cf. (2). It is proved in [3] that the estimator $\hat{L}_s^*$ is rate optimal in the minimax non-adaptive sense. The additional $\log d$ factor is necessary to achieve adaptivity as it will be clear from the subsequent arguments.

We obtain an adaptive estimator via data-driven selection in the collection of estimators $\{\hat{L}_s\}$. The selection is based on a Lepski type scheme. For $s = 1, \dots, d$, consider the thresholds $\omega_s > 0$ given by

$$\omega_s^2 = \beta\sigma^2 s^2 \log(1 + d(\log d)/s^2) = \beta\Phi^L(\sigma, s),$$

where $\beta > 0$ is a constant that will be chosen large enough. We define the selected index $\hat{s}$ by the relation

$$\hat{s} \triangleq \min\left\{ s \in \{1, \dots, \lfloor\sqrt{d \log d}\rfloor\} : |\hat{L}_s - \hat{L}_{s'}| \le \omega_{s'} \text{ for all } s' > s \right\} \tag{6}$$

with the convention that $\hat{s} = \lfloor\sqrt{d \log d}\rfloor + 1$ if the set in (6) is empty. Here, $\lfloor\sqrt{d \log d}\rfloor$ denotes the largest integer less than $\sqrt{d \log d}$. Finally, we define an adaptive to s estimator of L as

$$\hat{L} \triangleq \hat{L}_{\hat{s}}. \tag{7}$$

The following theorem exhibits an upper bound on its risk.

Theorem 1. *There exists an absolute constant $\alpha_0 > 0$ such that the following holds. Let $\hat{L}$ be the estimator defined in (7) with parameters $\alpha > \alpha_0$ and $\beta > 37\alpha$. Then, for all $\sigma > 0$, and all integers $d \ge 3$, $s \in \{1, \dots, d\}$ we have*

$$\sup_{\theta \in \Theta_s} \mathbf{E}_\theta (\hat{L} - L(\theta))^2 \le C\Phi^L(\sigma, s),$$

where $C > 0$ is an absolute constant.

Observe that for small s (such that $s \le d^b$ for $b < 1/2$), we have $1 \le \Phi^L(\sigma, s)/\psi_s^* \le c'$ where $c' > 0$ is an absolute constant. Therefore, for such s our estimator $\hat{L}$ attains the best possible rate on Θ_s given by the minimax risk ψ_s^* and it cannot be improved, even by estimators depending on s. Because of this, the only issue is to check that the rate $\Phi^L(\sigma, s)$ cannot be improved if s is greater than d^b with $b < 1/2$. For definiteness, we consider below the case $b = 1/4$ but with minor modifications the argument applies to any $b < 1/2$. Specifically, we prove that any estimator whose maximal risk over Θ_s is smaller (within a small constant) than $\Phi^L(\sigma, s)$ for some $s \ge d^{1/4}$, must have a maximal risk over Θ_1 of power order in d instead of the logarithmic order $\Phi^L(\sigma, 1)$ corresponding to our estimator. In other words, if we find an estimator that improves upon our estimator only slightly (by a constant factor) for some $s \ge d^{1/4}$, then this estimator inevitably loses much more for small s, such as $s = 1$, since there the ratio of maximal risks of the two estimators behaves as a power of d.

Theorem 2. *Let $d \geq 3$ and $\sigma > 0$. There exist two small absolute constants $C_0 > 0$ and $C_1 > 0$ such that the following holds. Any estimator $\widehat{T}$ that satisfies*

$$\sup_{\theta \in \Theta_s} \mathbf{E}_\theta \big[\big(\widehat{T} - L(\theta) \big)^2 \big] \leq C_0 \Phi^L(\sigma, s) \quad \text{for some } s \geq d^{1/4}$$

has a degenerate maximal risk over Θ_1, that is

$$\sup_{\theta \in \Theta_1} \mathbf{E}_\theta \big[\big(\widehat{T} - L(\theta) \big)^2 \big] \geq C_1 \sigma^2 d^{1/4} \ .$$

The property obtained in Theorem 2 can be paraphrased in an asymptotic context to conclude that $\Phi^L(\sigma, s)$ is the adaptive rate of convergence on the scale of classes $\{\Theta_s, s = 1, \dots, d\}$ in the sense of the definition in [10]. Indeed, assume that $d \to \infty$. Following [10], we call a function $s \mapsto \Psi_d(s)$ the *adaptive rate of convergence on the scale of classes* $\{\Theta_s, s = 1, \dots, d\}$ if the following holds.

(i) There exists an estimator $\hat{L}$ such that, for all d,

$$\max_{s=1,\dots,d} \sup_{\theta \in \Theta_s} \mathbf{E}_\theta (\hat{L} - L(\theta))^2 / \Psi_d(s) \leq C, \tag{8}$$

where $C > 0$ is a constant (clearly, such an estimator $\hat{L}$ is adaptive since it cannot depend on s).

(ii) If there exist another function $s \mapsto \Psi'_d(s)$ and a constant $C' > 0$ such that, for all d,

$$\inf_{\hat{T}} \max_{s=1,\dots,d} \sup_{\theta \in \Theta_s} \mathbf{E}_\theta (\hat{T} - L(\theta))^2 / \Psi'_d(s) \leq C', \tag{9}$$

and

$$\min_{s=1,\dots,d} \frac{\Psi'_d(s)}{\Psi_d(s)} \to 0 \text{ as } d \to \infty, \tag{10}$$

then there exists $\bar{s} \in \{1, \dots, d\}$ such that

$$\frac{\Psi'_d(\bar{s})}{\Psi_d(\bar{s})} \min_{s=1,\dots,d} \frac{\Psi'_d(s)}{\Psi_d(s)} \to \infty \text{ as } d \to \infty. \tag{11}$$

In words, this definition states that the adaptive rate of convergence $\Psi_d(s)$ is such that any improvement of this rate for some s (cf. (10)) is possible only at the expense of much greater loss for another $\bar{s}$ (cf. (11)).

Corollary 1. *The rate $\Phi^L(\sigma, s)$ is the adaptive rate of convergence on the scale of classes $\{\Theta_s, s = 1, \dots, d\}$.*

It follows from the above results that the rate $\Phi^L(\sigma, s)$ cannot be improved when adaptive estimation on the family of sparsity classes $\{\Theta_s, s = 1, \dots, d\}$ is considered. The ratio between the best rate of adaptive estimation $\Phi^L(\sigma, s)$ and the minimax rate ψ_s^* is equal to

$$\phi_s^* = \frac{\Phi^L(\sigma, s)}{\psi_s^*} = \frac{\log(1 + d(\log d)/s^2)}{\log(1 + d/s^2)}.$$

As mentioned above, $\phi_s^* \asymp 1$ if $s \leq d^b$ for $b < 1/2$. In a vicinity of $s = \sqrt{d}$ we have $\phi_s^* \asymp \log \log d$, whereas for $s \geq \sqrt{d \log d}$ the behavior of this ratio is logarithmic: $\phi_s^* \asymp \log d$. Thus, there are three different regimes and we see that, in all of them, rate adaptive estimation of the linear functional on the sparsity classes is impossible without loss of efficiency as compared to the minimax estimation. However, this loss is at most logarithmic in d.

3. Adaptation to s when σ is unknown

In this section we discuss a generalization of our adaptive estimator to the case when the standard deviation σ of the noise is unknown.

To treat the case of unknown σ, we first construct an estimator $\hat{\sigma}$ of σ such that, with high probability, $\sigma \le \hat{\sigma} \le 10\sigma$. Then, we consider the family of estimators defined by a relation analogous to (5):

$$\hat{L}'_s = \begin{cases} \sum_{j=1}^d y_j \mathbb{1}_{y_j^2 > \alpha\hat{\sigma}^2 \log(1+d(\log d)/s^2)}, & \text{if } s \le \sqrt{d\log d}, \\ \sum_{j=1}^d y_j, & \text{otherwise,} \end{cases} \tag{12}$$

where $\alpha > 0$ is a constant to be chosen large enough. The difference from (5) consists in the fact that we replace the unknown σ by $\hat{\sigma}$. Then, we define a random threshold $\omega'_s > 0$ as

$$(\omega'_s)^2 = \beta\hat{\sigma}^2 s^2 \log(1 + d(\log d)/s^2),$$

where $\beta > 0$ is a constant to be chosen large enough. The selected index $\hat{s}'$ is defined by the formula analogous to (6):

$$\hat{s}' \triangleq \min\big\{s \in \{1, \dots, \lfloor\sqrt{d\log d}\rfloor\} : |\hat{L}'_s - \hat{L}'_{s'}| \le \omega'_{s'} \text{ for all } s' > s\big\}. \tag{13}$$

Finally, the adaptive estimator when σ is unknown is defined as

$$\hat{L}' \triangleq \hat{L}'_{\hat{s}'}.$$

The aim of this section is to show that the risk of the estimator $\hat{L}'$ admits an upper bound with the same rate as in Theorem 1 for all d large enough. Consequently, $\hat{L}'$ attains the best rate of adaptive estimation as follows from Section 2.

Different estimators $\hat{\sigma}$ can be used. By slightly modifying the method suggested in [3], we consider the statistic

$$\hat{\sigma} = 9\Big(\frac{1}{\lfloor d/2\rfloor}\sum_{j\le d/2} y^2_{(j)}\Big)^{1/2} \tag{14}$$

where $y^2_{(1)} \le \cdots \le y^2_{(d)}$ are the order statistics associated to $y_1^2, \dots, y_d^2$. This statistic has the properties stated in the next proposition. In particular, $\hat{\sigma}$ overestimates σ but it turns out to be without prejudice to the attainment of the best rate by the resulting estimator $\hat{L}'_s$.

Proposition 1. *There exists an absolute constant $d_0 \ge 3$ such that the following holds. Let $\hat{\sigma}$ be the estimator defined in* (14)*. Then, for all integers $d \ge d_0$ and $s < d/2$ we have*

$$\inf_{\theta\in\Theta_s} \mathbf{P}_\theta(\sigma \le \hat{\sigma} \le 10\sigma) \ge 1 - d^{-5}, \tag{15}$$

and

$$\sup_{\theta\in\Theta_s} \mathbf{E}_\theta(\hat{\sigma}^4) \le \bar{C}\sigma^4, \tag{16}$$

where $\bar{C}$ is an absolute constant.

The proof of this proposition is given in Section 4. Using Proposition 1 we establish the following bound on the risk of the estimator $\hat{L}'$.

Theorem 3. *There exist large enough absolute constants $\alpha > 0$, $\beta > 0$, and $d_0 \geq 3$ such that the following holds. Let $\hat{\sigma}$ be the estimator defined in* (14)*. Then, for the estimator $\hat{L}'$ with tuning parameters α and β, for all $\sigma > 0$, and all integers $d \geq d_0$ and $s < d/2$ we have*

$$\sup_{\theta \in \Theta_s} \mathbf{E}_\theta (\hat{L}' - L(\theta))^2 \leq C \Phi^L(\sigma, s), \tag{17}$$

where $C > 0$ is an absolute constant.

Thus, the estimator $\hat{L}'$, which is independent of both s and σ achieves the rate $\Phi^L(\sigma, s)$ that is the best possible rate of adaptive estimation established in Section 2.

The condition $s < d/2$ in this theorem can be generalized to $s \leq \zeta d$ for some $\zeta \in (0, 1)$. In fact, for any $\zeta \in (0, 1)$, we can modify the definition of (14) by summing only over the $(1-\zeta)d$ smallest values of y_i^2. Then, changing the numerical constants α and β in the definition of ω_s', we obtain that the corresponding estimator $\hat{L}'$ achieves the best possible rate simultaneously for all $s \leq \zeta d$ with a constant C in (17) that would depend on ζ. However, we cannot set $\zeta = 1$. Indeed, the following proposition shows that it is not possible to construct an estimator, which is simultaneously adaptive to all $\sigma > 0$ and to all $s \in [1, d]$.

Proposition 2. *Let $d \geq 3$ and $\sigma > 0$. There exists a small absolute constant $C_0 > 0$ such that the following holds. Any estimator $\widehat{T}$ that satisfies*

$$\sup_{\theta \in \Theta_1} \mathbf{E}_\theta \big[\big(\widehat{T} - L(\theta) \big)^2 \big] \leq C_0 \sigma^2 d \ , \qquad \forall \sigma > 0, \tag{18}$$

has a degenerate maximal risk over Θ_d, that is, for any fixed $\sigma > 0$,

$$\sup_{\theta \in \Theta_d} \mathbf{E}_\theta \big[\big(\widehat{T} - L(\theta) \big)^2 \big] = \infty \ . \tag{19}$$

In other words, when σ is unknown, any estimator, for which the maximal risk over Θ_d is finite for all σ, cannot achieve over Θ_1 a risk of smaller order than $\sigma^2 d$, and hence cannot be minimax adaptive. Indeed, as shown above, the adaptive minimax rate over Θ_1 is of the order $\sigma^2 \log d$.

4. Proofs of the upper bounds

In the following, we will denote by $c_1, c_2, \ldots$ absolute positive constants. We will write for brevity L instead of $L(\theta)$.

4.1. Proof of Theorem 1.

Let $s \in \{1, \ldots, d\}$ and assume that θ belongs to Θ_s. We have

$$\mathbf{E}_\theta (\hat{L} - L)^2 = \mathbf{E}_\theta \big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s} \leq s} \big] + \mathbf{E}_\theta \big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s} > s} \big]. \tag{20}$$

Consider the first summand on the right hand side of (20). Set for brevity $s_0 = \lfloor \sqrt{d \log d} \rfloor + 1$. Using the definition of $\hat{s}$ we obtain, on the event $\{\hat{s} \leq s\}$,

$$(\hat{L}_{\hat{s}} - L)^2 \leq 2\omega_s^2 + 2(\hat{L}_s - L)^2 \text{ if } s < s_0 \text{ or } s \geq s_0, \hat{s} < s_0.$$

Thus,

$$\forall\ s < s_0 : \quad \mathbf{E}_\theta \big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s} \leq s} \big] \leq 2\beta^2 \Phi^L(\sigma, s) + 2\mathbf{E}_\theta (\hat{L}_s - L)^2, \tag{21}$$

$$\begin{aligned} \forall\ s \geq s_0 : \quad \mathbf{E}_\theta \big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s} \leq s} \big] &\leq \mathbf{E}_\theta \big[(\hat{L}_{\hat{s}} - L)^2 (\mathbb{1}_{\hat{s} \leq s, \hat{s} < s_0} + \mathbb{1}_{\hat{s} = s_0}) \big] \\ &\leq 2\beta^2 \Phi^L(\sigma, s) + 2\mathbf{E}_\theta (\hat{L}_s - L)^2 + \mathbf{E}_\theta (\hat{L}_{s_0} - L)^2. \end{aligned} \tag{22}$$

By Lemma 6 proved at the end of this section, for any tuning constant $\alpha > 0$ large enough we have

$$\sup_{\theta\in\Theta_s} \mathbf{E}_\theta(\hat{L}_s - L)^2 \le c_1\Phi^L(\sigma, s), \quad s = 1, \dots, s_0 - 1.$$

Note that, in view of (3), for all $s \in [s_0, d]$ we have

$$\Phi^L(\sigma, s_0) \le \sigma^2 d \log d \le 2\sigma^2 s^2 \log(1 + (d \log d)/s^2) = 2\Phi^L(\sigma, s),$$

and by definition of $\hat{L}_s$, for all $s \in [s_0, d]$ and all $\theta \in \mathbb{R}^d$, we have $\mathbf{E}_\theta(\hat{L}_s - L)^2 = \sigma^2 d \le 2\Phi^L(\sigma, s)$. Combining these remarks with (21) and (22) yields

$$\sup_{\theta\in\Theta_s} \mathbf{E}_\theta\big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s}\le s}\big] \le c_2\Phi^L(\sigma, s), \quad s = 1, \dots, d. \tag{23}$$

Consider now the second summand on the right hand side of (20). Since $\hat{s} \le s_0$ we obtain the following two facts. First,

$$\sup_{\theta\in\Theta_s} \mathbf{E}_\theta\big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s}>s}\big] = 0, \quad \forall\ s \ge s_0. \tag{24}$$

Second, on the event $\{\hat{s} > s\}$,

$$(\hat{L}_{\hat{s}} - L)^4 \le \sum_{s<s'\le s_0} (\hat{L}_{s'} - L)^4.$$

Thus,

$$\begin{aligned}
\sup_{\theta\in\Theta_s} \mathbf{E}_\theta\big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s}>s}\big] &\le \sup_{\theta\in\Theta_s} \Big[\sqrt{\mathbf{P}_\theta(\hat{s} > s)}(d \log d)^{1/4} \max_{s<s'\le s_0} \sqrt{\mathbf{E}_\theta(\hat{L}_{s'} - L)^4}\Big] \\
&\le (d \log d)^{1/4} \sup_{\theta\in\Theta_s} \sqrt{\mathbf{P}_\theta(\hat{s} > s)} \max_{s'\le s_0} \Big[\sup_{\theta\in\Theta_{s'}} \sqrt{\mathbf{E}_\theta(\hat{L}_{s'} - L)^4}\Big] \qquad (25)
\end{aligned}$$

where for the second inequality we have used that $\Theta_s \subset \Theta_{s'}$ for $s < s'$. To evaluate the right hand side of (25) we use the following two lemmas.

Lemma 1. *For all $s \le s_0$ we have*

$$\sup_{\theta\in\Theta_s} \mathbf{E}_\theta\big(\hat{L}_s - L\big)^4 \le c_3\sigma^4 d^4 (\log d)^2, \quad \sup_{\theta\in\Theta_s} \mathbf{E}_\theta\big(\hat{L}'_s - L\big)^4 \le c_3\sigma^4 d^4 (\log d)^2.$$

Lemma 2. *There exist absolute constants $c_4 > 0$ and $\alpha_0 > 0$ such that the following holds.*
(i) For all $\alpha > \alpha_0$ and $\beta > 37\alpha$ we have

$$\max_{s\le\sqrt{d\log d}} \sup_{\theta\in\Theta_s} \mathbf{P}_\theta(\hat{s} > s) \le c_4 d^{-5}. \tag{26}$$

(ii) There exist $\alpha > \alpha_0$ and $\beta > \alpha_0$ such that

$$\max_{s\le\sqrt{d\log d}} \sup_{\theta\in\Theta_s} \mathbf{P}_\theta(\hat{s}' > s) \le c_4 d^{-5}.$$

From (24), (25), the first inequality in Lemma 1, and part (i) of Lemma 2 we find that

$$\sup_{\theta\in\Theta_s} \mathbf{E}_\theta\big[(\hat{L}_{\hat{s}} - L)^2 \mathbb{1}_{\hat{s}>s}\big] \le c_5\sigma^2 \le c_6\Phi^L(\sigma, s), \quad s = 1, \dots, d.$$

Combining this inequality with (20) and (23) we obtain the theorem.

4.2. **Proofs of the lemmas.**

Proof of Lemma 1. For $s = s_0$, $\hat{L}_s - L = \hat{L}'_s - L = \sigma \sum_{i=1}^d \xi_i$. As a consequence

$$\mathbf{E}_\theta(\hat{L}_s - L)^4 = \mathbf{E}_\theta(\hat{L}'_s - L)^4 = 3\sigma^4 d^2 \le 3\sigma^4 d^4 (\log d)^2 \ .$$

Henceforth, we focus on the case $s \le \sqrt{d \log(d)}$. We have

$$\hat{L}_s - L = \sigma \sum_{i=1}^d \xi_i - \sum_{i=1}^d y_i \mathbb{1}_{y_i^2 \le \alpha\sigma^2 \log(1+d(\log d)/s^2)}. \tag{27}$$

Thus,

$$\mathbf{E}_\theta(\hat{L}_s - L)^4 \le 8\Big(\sigma^4 \mathbf{E}\Big(\sum_{i=1}^d \xi_i\Big)^4 + d^4\alpha^2\sigma^4 \log^2(1 + d(\log d)/s^2)\Big) \le c_3 \sigma^4 d^4 (\log d)^2.$$

In a similar way,

$$\hat{L}'_s - L = \sigma \sum_{i=1}^d \xi_i - \sum_{i=1}^d y_i \mathbb{1}_{y_i^2 \le \alpha\hat{\sigma}^2 \log(1+d(\log d)/s^2)}, \tag{28}$$

and

$$\mathbf{E}_\theta(\hat{L}'_s - L)^4 \le 8\Big(\sigma^4 \mathbf{E}\Big(\sum_{i=1}^d \xi_i\Big)^4 + d^4\alpha^2 \mathbf{E}_\theta(\hat{\sigma}^4) \log^2(1 + d(\log d)/s^2)\Big).$$

The desired bound for $\mathbf{E}_\theta(\hat{L}'_s - L)^4$ follows from this inequality and (16). □

Proof of Lemma 2. We start by proving part (i) of Lemma 2. Note first that, for $s \le \sqrt{d \log d}$ and all θ we have

$$\mathbf{P}_\theta\big(|\hat{L}_s - L| > \omega_{s'}/2\big) \le \mathbf{P}_\theta\big(|\hat{L}_s - L| > \omega_s/3\big), \quad \forall\ s < s' \le d. \tag{29}$$

Indeed, if $s < s'$ we have $\omega_{s'} > \omega_s$ since the function $t \mapsto \omega_t$ is increasing for $t > 0$. If $\sqrt{d \log d} \le s' \le d$, we use (4), which yields $\omega_{s'}^2 \ge \beta\sigma^2 d(\log d)/2 \ge \omega^2_{\sqrt{d \log d}}/2$ and since $s \le \sqrt{d \log d}$ we obtain $\omega_{s'}^2 \ge \omega_s^2/2$ using again the fact that the function $t \mapsto \omega_t$ is increasing.

From (29) we obtain that, for $s \le \sqrt{d \log d}$, all s' such that $s < s' \le d$ and all θ,

$$\mathbf{P}_\theta\big(|\hat{L}_{s'} - \hat{L}_s| > \omega_{s'}\big) \le \mathbf{P}_\theta\big(|\hat{L}_{s'} - L| > \omega_{s'}/2\big) + \mathbf{P}_\theta\big(|\hat{L}_s - L| > \omega_s/3\big).$$

This inequality and the definition of $\hat{s}$ imply that, for all $s \le \sqrt{d \log d}$ and all θ,

$$\begin{aligned} \mathbf{P}_\theta(\hat{s} > s) &\le \sum_{s < s' \le d} \mathbf{P}_\theta\big(|\hat{L}_{s'} - \hat{L}_s| > \omega_{s'}\big) \\ &\le d\mathbf{P}_\theta\big(|\hat{L}_s - L| > \omega_s/3\big) + \sum_{s < s' \le d} \mathbf{P}_\theta\big(|\hat{L}_{s'} - L| > \omega_{s'}/2\big). \end{aligned} \tag{30}$$

Note that, for $\sqrt{d \log d} < s' \le d$, we have $\hat{L}_{s'} = \sum_{i=1}^d y_i$, and $\omega_{s'} \ge \sigma\sqrt{\beta d \log d}/\sqrt{2}$ due to (4). Hence, for $\sqrt{d \log d} < s' \le d$, and all θ,

$$\mathbf{P}_\theta\big(|\hat{L}_{s'} - L| > \omega_{s'}/2\big) \le \mathbf{P}\Big(\Big|\sum_{i=1}^d \xi_i\Big| > \frac{\sqrt{\beta d \log d}}{2\sqrt{2}}\Big) \le 2d^{-\beta/16},$$

where we have used that ξ_i are i.i.d. standard Gaussian random variables. This inequality and (30) imply that, for $s \le \sqrt{d\log d}$, and all θ,

$$\mathbf{P}_\theta(\hat{s} > s) \le \sqrt{d\log d} \max_{s<s'\le\sqrt{d\log d}} \mathbf{P}_\theta\big(|\hat{L}_{s'} - L| > \omega_{s'}/2\big) + d\mathbf{P}_\theta\big(|\hat{L}_s - L| > \omega_s/3\big) + 2d^{1-\beta/16}. \tag{31}$$

As $\Theta_s \subset \Theta_{s'}$ for $s < s'$, we have

$$\max_{s<s'\le\sqrt{d\log d}} \sup_{\theta\in\Theta_s} \mathbf{P}_\theta\big(|\hat{L}_{s'} - L| > \omega_{s'}/2\big) \le \max_{s'\le\sqrt{d\log d}} \sup_{\theta\in\Theta_{s'}} \mathbf{P}_\theta\big(|\hat{L}_{s'} - L| > \omega_{s'}/2\big).$$

Together with (31) this implies

$$\max_{s\le\sqrt{d\log d}} \sup_{\theta\in\Theta_s} \mathbf{P}_\theta(\hat{s} > s) \le 2d \max_{s'\le\sqrt{d\log d}} \sup_{\theta\in\Theta_{s'}} \mathbf{P}_\theta\big(|\hat{L}_{s'} - L| > \omega_{s'}/3\big) + 2d^{1-\beta/16}.$$

For $\beta > 96$ the last summand in this inequality does not exceed $2d^{-5}$. Thus, to prove (26) it is enough to show that

$$\max_{s\le\sqrt{d\log d}} \sup_{\theta\in\Theta_s} \mathbf{P}_\theta\big(|\hat{L}_s - L| > \omega_s/3\big) \le c_7 d^{-6}. \tag{32}$$

The rest of this proof consists in demonstrating that (32) is satisfied if the tuning constants α and β are properly chosen. Fix $s \le \sqrt{d\log d}$ and let θ belong to Θ_s. We will denote by S the support of θ and we set for brevity

$$a \triangleq \sqrt{\log(1 + d(\log d)/s^2)}.$$

From (27) and the fact that $y_i = \theta_i + \sigma\xi_i$ we have

$$\begin{aligned} |\hat{L}_s - L| &= \Big|\sigma \sum_{i\in S} \xi_i - \sum_{i\in S} y_i \mathbb{1}_{y_i^2 \le \alpha\sigma^2 a^2} + \sigma \sum_{i\notin S} \xi_i \mathbb{1}_{\xi_i^2 > \alpha a^2}\Big| \\ &\le \sigma\Big|\sum_{i\in S} \xi_i\Big| + \sigma\Big|\sum_{i\notin S} \xi_i \mathbb{1}_{\xi_i^2>\alpha a^2}\Big| + \sqrt{\alpha}\sigma s a. \end{aligned} \tag{33}$$

In the following we assume that $\beta > 37\alpha$. Using this and recalling that $\omega_s = \sqrt{\beta}\sigma s a$ we find

$$\begin{aligned} \mathbf{P}_\theta\big(|\hat{L}_s - L| > \omega_s/3\big) &\le \mathbf{P}\Big(\Big|\sum_{i\notin S} \xi_i \mathbb{1}_{\xi_i^2>\alpha a^2}\Big| > \sqrt{\beta} s a/6\Big) + \mathbf{P}\Big(\Big|\sum_{i\in S} \xi_i\Big| > (\sqrt{\beta}/6 - \sqrt{\alpha}) s a\Big) \\ &\le \mathbf{P}\Big(\Big|\sum_{i\notin S} \xi_i \mathbb{1}_{\xi_i^2>\alpha a^2}\Big| > \sqrt{\alpha} s a\Big) + \mathbf{P}\Big(\Big|\sum_{i\in S} \xi_i\Big| > c_8\sqrt{\alpha} s a\Big). \end{aligned} \tag{34}$$

Since ξ_i are i.i.d. $\mathcal{N}(0,1)$ random variables, we have

$$\mathbf{P}\Big(\Big|\sum_{i\in S} \xi_i\Big| > c_8\sqrt{\alpha} s a\Big) \le 2\exp\Big(-\frac{c_8^2}{2}\alpha s a^2\Big). \tag{35}$$

We now use the relation

$$s a^2 = s\log(1 + d(\log d)/s^2) \ge (\log d)/2 \quad \text{for all } s \in [1, \sqrt{d\log d}]. \tag{36}$$

Indeed, for $s \in [1, \sqrt{d\log d}/3]$ the left hand side of (36) is monotone increasing in s, while for $s \in [\sqrt{d\log d}/3, \sqrt{d\log d}]$ the inequality in (36) is trivial. It follows from (35) and (36) that,

for all $\beta > 37\alpha$ and all suitably large α,

$$\mathbf{P}\Big(\Big|\sum_{i\in S}\xi_i\Big| > c_8\sqrt{\alpha}sa\Big) \le d^{-6}. \tag{37}$$

Next, consider the first probability on the right hand side of (34). To bound it from above, we invoke the following lemma.

Lemma 3. *For any absolute constant $\alpha > 0$ large enough, for all $s \le \sqrt{d\log d}$, and all $U \subseteq \{1,\dots,d\}$,*

$$\mathbf{P}\Big(\sup_{t\in[1,10]}\Big|\sum_{i\in U}\xi_i\mathbb{1}_{|\xi_i|>\sqrt{\alpha}at}\Big| > \sqrt{\alpha}sa\Big) \le c_9 d^{-6}.$$

Combining (34), (37) and Lemma 3 we obtain (32). Thus, part (i) of Lemma 2 follows.

We now proceed to the proof of part (ii) of Lemma 2. Proposition 1 implies that, for $s \le \sqrt{d\log d}$ and $\theta \in \Theta_s$,

$$\mathbf{P}_\theta(\hat{s}' > s) \le \mathbf{P}_\theta(\hat{s}' > s, \hat{\sigma} \in [\sigma, 10\sigma]) + d^{-5}.$$

On the event $\{\hat{\sigma} \in [\sigma, 10\sigma]\}$, we can replace $\hat{\sigma}$ in the definition of $\hat{s}'$ either by σ or by 10σ according to cases, thus making the analysis of $\mathbf{P}_\theta(\hat{s}' > s, \hat{\sigma} \in [\sigma, 10\sigma])$ equivalent, up to the values of numerical constants, to the analysis of $\mathbf{P}_\theta(\hat{s} > s)$ given below. The only non-trivial difference consists in the fact that the analog of (33) when $\hat{L}_s$ is replaced by $\hat{L}'_s$ contains the term $\sigma\Big|\sum_{i\notin S}\xi_i\mathbb{1}_{\xi_i^2>\alpha\hat{\sigma}^2a^2/\sigma^2}\Big|$ instead of $\sigma\Big|\sum_{i\notin S}\xi_i\mathbb{1}_{\xi_i^2>\alpha a^2}\Big|$ while $\hat{\sigma}$ depends on $\xi_1,\dots,\xi_d$. This term is evaluated using Lemma 3 and the fact that

$$\mathbf{P}\Big(\Big|\sum_{i\notin S}\xi_i\mathbb{1}_{|\xi_i|>\sqrt{\alpha}\hat{\sigma}a/\sigma}\Big| > \sqrt{\alpha}sa,\ \hat{\sigma}\in[\sigma,10\sigma]\Big) \le \mathbf{P}\Big(\sup_{t\in[1,10]}\Big|\sum_{i\notin S}\xi_i\mathbb{1}_{|\xi_i|>\sqrt{\alpha}at}\Big| > \sqrt{\alpha}sa\Big).$$

We omit further details that are straightforward from inspection of the proof of part (i) of Lemma 2 given above. Thus, part (ii) of Lemma 2 follows. □

For the proof of Lemma 3, recall the following fact about the tails of the standard Gaussian distribution.

Lemma 4. *Let $X \sim \mathcal{N}(0,1)$, $x > 1$ and $q \in \mathbb{N}$. There exists a constant C_q^* depending only on q such that*

$$\mathbf{E}\Big[X^{2q}\mathbb{1}_{|X|>x}\Big] \le C_q^* x^{2q-1}e^{-x^2/2}.$$

We will also use the Fuk-Nagaev inequality [9, page 78] that we state here for reader's convenience.

Lemma 5 (Fuk-Nagaev inequality)**.** *Let $p > 2$ and $v > 0$. Assume that $X_1,\dots,X_n$ are independent random variables with $\mathbf{E}(X_i) = 0$ and $\mathbf{E}|X_i|^p < \infty$, $i = 1,\dots,n$. Then,*

$$\mathbf{P}\Big(\sum_{i=1}^n X_i > v\Big) \le (1+2/p)^p\sum_{i=1}^n \mathbf{E}|X_i|^p v^{-p} + \exp\left(-\frac{2v^2}{(p+2)^2e^p\sum_{i=1}^n \mathbf{E}X_i^2}\right).$$

Proof of Lemma 3. We have

$$\begin{aligned} p_0 &\triangleq \mathbf{P}\Big(\sup_{t\in[1,10]}\big|\sum_{i\in U}\xi_i\mathbb{1}_{|\xi_i|>\sqrt{\alpha}at}\big| > \sqrt{\alpha}sa\Big) \\ &= \mathbf{E}\Big[\mathbf{P}\Big(\sup_{t\in[1,10]}\big|\sum_{i\in U}\epsilon_i|\xi_i|\mathbb{1}_{|\xi_i|>\sqrt{\alpha}at}\big| > \sqrt{\alpha}sa\ \Big|\ |\xi_i|, i\in U\Big)\Big] \end{aligned}$$

where ϵ_i denotes the sign of ξ_i. Consider the function $g(x) = \sup_{t\in[1,10]} \big|\sum_{i\in U} x_i|\xi_i|\mathbb{1}_{|\xi_i|>\sqrt{\alpha}at}\big|$ where $x = (x_i, i \in U)$ with $x_i \in \{-1, 1\}$. For any $i_0 \in U$, let $g_{i_0,u}(x)$ denote the value of this function when we replace x_{i_0} by $u \in \{-1, 1\}$. Note that, for any fixed $(|\xi_i|, i \in U)$, we have the bounded differences condition:

$$\sup_x |g(x) - g_{i_0,u}(x)| \le 2|\xi_i|\mathbb{1}_{|\xi_i|>\sqrt{\alpha}a} \triangleq 2Z_i \quad \forall u \in \{-1, 1\}, i_0 \in U.$$

The vector of Rademacher random variables $(\epsilon_1, \dots, \epsilon_d)$ is independent from $(|\xi_1|, \dots, |\xi_d|)$.Thus, for any fixed $(|\xi_i|, i \in U)$ we can use the bounded differences inequality, which yields

$$p_0 \le \mathbf{E}\Big[\exp\Big(-\frac{\alpha s^2a^2}{2\sum_{i\in U} Z_i^2}\Big)\Big] \le \exp\Big(-\frac{\alpha s^2a^2}{2\Delta}\Big) + \mathbf{P}\Big(\sum_{i\in U} Z_i^2 > \Delta\Big), \quad \forall\ \Delta > 0. \tag{38}$$

We now set $\Delta = \sum_{i\in U} \mathbf{E}Z_i^2 + d\exp\big(-\alpha a^2/(2p)\big)$ for some integer $p > 2$ that will be chosen large enough.

To finish the proof, it remains to show that each of the two summands on the right hand side of (38) does not exceed $c_9d^{-6}/2$ if p and α are large enough. To bound from above the probability $\mathbf{P}\Big(\sum_{i\in U} Z_i^2 > \Delta\Big)$ we apply Lemma 5 with $X_i = Z_i^2 - \mathbf{E}(Z_i^2)$ and $v = d\exp\big(-\alpha a^2/(2p)\big)$. The random variables X_i are centered and satisfy, in view of Lemma 4,

$$\mathbf{E}|X_i|^p \le 2^{p-1}\mathbf{E}|Z_i|^{2p} \le 2^{p-1}C_p^*(\sqrt{\alpha}a)^{2p-1}e^{-\alpha a^2/2} \tag{39}$$

for any $p \in \mathbb{N}^*$. Thus, Lemma 5 yields

$$\mathbf{P}\Big(\sum_{i\in U} Z_i^2 > \Delta\Big) \ \le\ C_p^* 2^{p-1}(1 + 2/p)^p \frac{(\sqrt{\alpha}a)^{2p-1}}{d^p} + \exp\left(-\frac{d\exp(\alpha a^2(1/2 - 1/p))}{(p + 2)^2e^pC_2^*(\sqrt{\alpha}a)^3}\right).$$

The expression in the last display can be rendered smaller than $c_9d^{-6}/2$ for some absolute constant $c_9 > 0$ and all $d \ge 3$ by choosing p large enough.

Finally, using (39) we find

$$\frac{\alpha s^2a^2}{2\Delta} \ge \frac{\alpha s^2a^2}{2d(C_1^*\sqrt{\alpha}a\exp(-\alpha a^2/2) + \exp(-\alpha a^2/(2p)))} \ge \frac{c_{10}\sqrt{\alpha}as^2\exp(\alpha a^2/(2p))}{d},$$

whereas

$$\frac{s^2\exp(\alpha a^2/(2p))}{d} = \frac{s^2}{d}\Big(1 + \frac{d\log d}{s^2}\Big)^{\alpha/(2p)} \ge \log d\,\Big(\frac{d\log d}{s^2}\Big)^{\alpha/(2p)-1} \ge \log d$$

for any $\alpha \ge 2p$ and any $s \le \sqrt{d\log d}$. Hence, for such α and s,

$$\exp\Big(-\frac{\alpha s^2a^2}{2\Delta}\Big) \le \exp\big(-c_{10}\sqrt{\alpha}a\log d\big)$$

that is smaller than $c_9d^{-6}/2$ for all $d \ge 3$ provided that the absolute constant α is large enough. Thus, Lemma 3 follows.

□

Lemma 6. *There exists an absolute constant $c_1 > 0$ such that, for all $\alpha > 0$ large enough,*

$$\sup_{\theta\in\Theta_s} \mathbf{E}_\theta(\hat{L}_s - L)^2 \le c_1\Phi^L(\sigma, s), \quad \sup_{\theta\in\Theta_s} \mathbf{E}_\theta(\hat{L}'_s - L)^2 \le c_1\Phi^L(\sigma, s), \quad \forall s \le \sqrt{d\log d}.$$

Proof. We easily deduce from (33) that

$$\mathbf{E}_\theta(\hat{L}_s - L)^2 \le 3\sigma^2\Big(s + d\mathbf{E}\left[X^2 \mathbb{1}_{X^2>\alpha a^2}\right] + \alpha s^2 a^2\Big),$$

where $X \sim \mathcal{N}(0,1)$. By (39),

$$d\mathbf{E}\left[X^2 \mathbb{1}_{X^2>2a^2}\right] \le 2C_1^* ad\exp(-a^2) = \frac{2C_1^* ads^2}{s^2 + d\log d} \le \frac{2C_1^* s^2 a}{\log d},$$

which implies that the desired bound for $\mathbf{E}_\theta(\hat{L}_s - L)^2$ holds whenever $\alpha \ge 2$. Next, we prove the bound of the lemma for $\mathbf{E}_\theta(\hat{L}'_s - L)^2$. Similarly to (33),

$$\hat{L}'_s - L = \sigma\sum_{i\in S}\xi_i - \sum_{i\in S} y_i \mathbb{1}_{y_i^2 \le \alpha\hat{\sigma}^2 a^2} + \sigma\sum_{i\notin S}\xi_i \mathbb{1}_{\sigma^2\xi_i^2 > \alpha\hat{\sigma}^2 a^2}.$$

This implies

$$\begin{aligned}\mathbf{E}_\theta\left[(\hat{L}'_s - L)^2 \mathbb{1}_{\hat{\sigma}\in[\sigma,10\sigma]}\right] &\le \mathbf{E}_\theta\Big(\sigma\Big|\sum_{i\in S}\xi_i\Big| + \sqrt{\alpha}\hat{\sigma}sa + \sigma W\Big)^2 \\ &\le 3\Big(\sigma^2 s + \alpha\mathbf{E}_\theta(\hat{\sigma}^2)a^2 s^2 + \sigma^2\mathbf{E}(W^2)\Big),\end{aligned} \tag{40}$$

where $W \triangleq \sup_{t\in[1,10]}\Big|\sum_{i\notin S}\xi_i \mathbb{1}_{|\xi_i|>\sqrt{\alpha}at}\Big|$. Using Lemma 3 we find that, for all $\alpha > 0$ large enough,

$$\begin{aligned}\mathbf{E}(W^2) &\le (\sqrt{\alpha}sa)^2 + \mathbf{E}\Big(\sum_{i\notin S}|\xi_i|\Big)^2 \mathbb{1}_{W>\sqrt{\alpha}sa} \\ &\le \alpha s^2 a^2 + \Big[\mathbf{E}\Big(\sum_{i\notin S}|\xi_i|\Big)^4\Big]^{1/2} c_9 d^{-3} \le \alpha s^2 a^2 + c_9\sqrt{3}d^{-1}.\end{aligned}$$

Plugging this bound in (40) and using (16) we get

$$\mathbf{E}_\theta\left[(\hat{L}'_s - L)^2 \mathbb{1}_{\hat{\sigma}\in[\sigma,10\sigma]}\right] \le c_{11}\Phi^L(\sigma, s).$$

On the other hand, by virtue of Lemma 1 and (15),

$$\mathbf{E}_\theta\left[(\hat{L}'_s - L)^2 \mathbb{1}_{\hat{\sigma}\notin[\sigma,10\sigma]}\right] \le \sqrt{\mathbf{P}_\theta(\hat{\sigma}\notin[\sigma,10\sigma])}\sqrt{\mathbf{E}_\theta(\hat{L}'_s - L)^4} \le \frac{\sqrt{c_3}\sigma^2\log d}{d^{1/2}} \le c_{12}\Phi^L(\sigma, s).$$

The desired bound for $\mathbf{E}_\theta(\hat{L}'_s - L)^2$ follows from the last two displays.

□

4.3. **Proofs of Proposition 1 and of Theorem 3.**

Proof of Proposition 1. Since $s \le d/2$, there exists a subset T of size $\lfloor d/2\rfloor$ such that $T\cap S = \emptyset$. By Definition of $\hat{\sigma}^2$, we obtain that

$$\hat{\sigma}^2 \le \frac{81\sigma^2}{\lfloor d/2\rfloor}\sum_{i\in T}\xi_i^2 \ .$$

This immediately implies (16). To prove (15), note that the Gaussian concentration inequality (*cf.* [7]) yields

$$\mathbf{P}\Big(\Big(\sum_{i\in T}\xi_i^2\Big)^{1/2} > \sqrt{100\lfloor d/2\rfloor/81}\Big) \le \exp(-cd) \ ,$$

for a positive constant c. Therefore,

$$\mathbf{P}_\theta(\hat\sigma \le 10\sigma) \ge 1 - \exp(-cd). \tag{41}$$

Next, let $\mathcal{G}$ be the collection of all subsets of $\{1,\dots,d\}$ of cardinality $\lfloor d/2\rfloor$. We now establish a bound on the deviations of random variables $Z_G = \frac{1}{\sigma^2}\sum_{i\in G} y_i^2$ uniformly over all $G\in\mathcal{G}$. Fix any $G\in\mathcal{G}$. The random variable Z_G has a chi-square distribution with $\lfloor d/2\rfloor$ degrees of freedom and non-centrality parameter $\sum_{i\in G}\theta_i^2$. In particular, this distribution is stochastically larger than a central chi-square distribution with $d' = \lfloor d/2\rfloor$ degrees of freedom. Let Z be a random variable with this central chi-square distribution. By Lemma 11.1 in [12],

$$\mathbf{P}\Big(Z \le \frac{d'}{e}x^{2/d'}\Big) \le x, \qquad \forall\ x>0.$$

Take $x = \binom{d}{d'}^{-1} e^{-d'/2}$. Using the bound $\log\binom{d}{d'} \le d'\log(ed/d')$ it follows that $\log(1/x) \le d'(\frac{3}{2} + \log(\frac{d}{d'})) \le d'(\frac{3}{2}+\log 2)+1$. Taking the union bound over all $G\in\mathcal{G}$ we conclude that

$$\mathbf{P}\Big(\inf_{G\in\mathcal{G}} Z_G \le \frac{d'}{4e^3}\Big(1-\frac{2}{d'}\Big)\Big) \le e^{-d'/2} < d^{-5}/2$$

for all d large enough. Since $\widehat\sigma^2 = \sigma^2\frac{81}{d'}\inf_{G\in\mathcal{G}} Z_G^2$, we obtain that $\widehat\sigma^2 \ge \sigma^2$ with probability at least $1-d^{-5}/2$ for all d large enough. Combining this with (41), we get (15) for all d large enough.

□

Proof of Theorem 3. We repeat the proof of Theorem 1 replacing there $\hat L_s$ by $\hat L_s'$ and $\hat s$ by $\hat s'$. The difference is that, in view of (16), the relation (21) now holds with $c_{15}\beta^2\Phi^L(\sigma,s)$ instead of $\beta^2\Phi^L(\sigma,s)$, and we use the results of Lemmas 1, 2 and 6 related to $\hat L_s'$ rather than to $\hat L_s$. □

5. Proofs of the lower bounds

5.1. **Proof of Theorem 2.** Theorem 2 is an immediate consequence of the following lemma.

Lemma 7. *For all $d\ge 3$ and all $s\ge d^{1/4}$,*

$$R(s) \triangleq \inf_{\tilde L}\left\{\sup_{\theta\in\Theta_1}\mathbf{E}_\theta(\tilde L - L)^2\sigma^{-2}d^{-1/4} + \sup_{\theta\in\Theta_s}\mathbf{E}_\theta(\tilde L-L)^2\big(\Phi^L(\sigma,s)\big)^{-1}\right\} \ge \frac{1}{160}. \tag{42}$$

Proof. We first introduce some notation. For a probability measure μ on Θ_s, we denote by $\mathbb{P}_\mu$ the mixture probability measure $\mathbb{P}_\mu = \int_{\Theta_s}\mathbf{P}_\theta\mu(d\theta)$. Let $\mathcal{S}(s,d)$ denote the set of all subsets of $\{1,\dots,d\}$ of size s, and let S be a set-valued random variable uniformly distributed on $\mathcal{S}(s,d)$. For any $\rho>0$, denote by μ_ρ the distribution of the random variable $\sigma\rho\sum_{j\in S} e_j$ where e_j is the jth canonical basis vector in $\mathbb{R}^d$. Next, let $\chi^2(Q,P) = \int(dQ/dP)^2dP - 1$ denote the chi-square divergence between two probability measures Q and P such that $Q\ll P$, and $\chi^2(Q,P)=+\infty$ if $Q\not\ll P$.

Take any $s\ge d^{1/4}$ and set

$$\rho \triangleq \sqrt{\log(1+d(\log d)/s^2)}/2 = \big(\Phi^L(\sigma,s)\big)^{1/2}/(2s\sigma).$$

Consider the mixture distribution $\mathbb{P}_{\mu_\rho}$ with this value of ρ. For any estimator $\tilde{L}$, we have $\sup_{\theta\in\Theta_s} \mathbf{E}_\theta(\tilde{L}-L)^2 \geq \mathbb{E}_{\mu_\rho}(\tilde{L}-L)^2 \geq \mathbb{E}_{\mu_\rho}(\tilde{L}-\mathbb{E}_{\mu_\rho}(L))^2 = \mathbb{E}_{\mu_\rho}(\tilde{L}-\sigma s\rho)^2$. Therefore,

$$\begin{aligned} R(s) &\geq \inf_{\tilde{L}} \left\{ \mathbf{E}_0(\tilde{L}^2)\sigma^{-2}d^{-1/4} + \mathbb{E}_{\mu_\rho}(\tilde{L}-\sigma s\rho)^2 \big(\Phi^L(\sigma,s)\big)^{-1} \right\} \\ &\geq \frac{1}{16}\inf_{\tilde{L}} \left\{ \mathbf{P}_0(\tilde{L} > \sigma s\rho/2)\sigma^{-2}d^{-1/4}\Phi^L(\sigma,s) + \mathbb{P}_{\mu_\rho}(\tilde{L} < \sigma s\rho/2) \right\} \\ &\geq \frac{1}{16}\inf_{\mathcal{A}} \left\{ \mathbf{P}_0(\mathcal{A})\sigma^{-2}d^{-1/4}\Phi^L(\sigma,s) + \mathbb{P}_{\mu_\rho}(\mathcal{A}^c) \right\}, \end{aligned} \tag{43}$$

where $\inf_{\mathcal{A}}$ denotes the infimum over all measurable events $\mathcal{A}$, and $\mathcal{A}^c$ denotes the complement of $\mathcal{A}$. It remains to prove that the expression in (43) is not smaller than $1/160$. This will be deduced from the following lemma, the proof of which is given at the end of this section.

Lemma 8. *Let P and Q be two probability measures on a measurable space $(X,\mathcal{U})$. Then, for any $q>0$,*

$$\inf_{\mathcal{A}\in\mathcal{U}} \{P(\mathcal{A})q + Q(\mathcal{A}^c)\} \geq \max_{0<\tau<1} \left[\frac{q\tau}{1+q\tau}\big(1-\tau(\chi^2(Q,P)+1)\big) \right].$$

We now apply Lemma 8 with $P=\mathbf{P}_0$, $Q=\mathbb{P}_{\mu_\rho}$, and

$$q = \sigma^{-2}d^{-1/4}\Phi^L(\sigma,s) = s^2 d^{-1/4}\log\Big(1+\frac{d(\log d)}{s^2}\Big). \tag{44}$$

By Lemma 1 in [3], the chi-square divergence $\chi^2(\mathbb{P}_{\mu_\rho},\mathbf{P}_0)$ satisfies

$$\chi^2(\mathbb{P}_{\mu_\rho},\mathbf{P}_0) \leq \Big(1-\frac{s}{d}+\frac{s}{d}e^{\rho^2}\Big)^s - 1 \leq \Big(1+\frac{s}{d}\Big(e^{\rho^2}-1\Big)\Big)^s \ .$$

Since $\rho^2 = \Big(\log\big(1+\frac{d(\log d)}{s^2}\big)\Big)/4$, we find

$$\begin{aligned} \chi^2(\mathbb{P}_{\mu_\rho},\mathbf{P}_0) &\leq \exp\left[s\log\left[1+\frac{s}{d}\left(\Big(1+\frac{d(\log d)}{s^2}\Big)^{1/4} - 1 \right)\right]\right] \\ &\leq \exp\left[s\log\Big(1+\frac{\log d}{4s}\Big)\right] \leq d^{1/4}\ , \end{aligned} \tag{45}$$

where we have used that $(1+x)^{1/4}\leq 1+x/4$ for $x>0$. Take

$$\tau = (d^{1/4}+1)^{-1}/2. \tag{46}$$

Then, using (44) and the inequality $s\geq d^{1/4}$ we find

$$q\tau = \frac{s^2\log\Big(1+\frac{d(\log d)}{s^2}\Big)}{2d^{1/4}(d^{1/4}+1)} \geq \frac{d^{1/2}\log(1+d^{1/2}(\log d))}{2d^{1/4}(d^{1/4}+1)} > \frac{1}{4}, \quad \forall\ d\geq 3. \tag{47}$$

Lemma 8 and inequalities (45) – (47) imply

$$\inf_{\mathcal{A}} \left\{ \mathbf{P}_0(\mathcal{A})\sigma^{-2}d^{-1/4}\Phi^L(\sigma,s) + \mathbb{P}_{\mu_\rho}(\mathcal{A}^c) \right\} \geq \frac{q\tau}{2(1+q\tau)} \geq \frac{1}{10}.$$

□

Proof of Lemma 8. We follow the same lines as in the proof of Proposition 2.4 in [11]. Thus, for any $\tau \in (0,1)$,

$$P(\mathcal{A}) \geq \tau(Q(\mathcal{A}) - v), \quad \text{where } v = Q\left(\frac{dP}{dQ} < \tau\right) \leq \tau(\chi^2(Q,P) + 1).$$

Then,

$$\begin{aligned}\inf_{\mathcal{A}} \{P(\mathcal{A})q + Q(\mathcal{A}^c)\} &\geq \inf_{\mathcal{A}} \{q\tau(Q(\mathcal{A}) - v) + Q(\mathcal{A}^c)\} \\ &\geq \min_{0\leq t\leq 1} \max(q\tau(t-v), 1-t) = \frac{q\tau(1-v)}{1+q\tau}.\end{aligned}$$

□

5.2. **Proof of Corollary 1.** First, note that condition (8) with $\Psi_d(s) = C\Phi^L(\sigma,s)$ is satisfied due to Theorem 1. Next, the minimum in condition (10) with $\Psi_d(s) = C\Phi^L(\sigma,s)$ can be only attained for $s \geq d^{1/4}$, since for $s < d^{1/4}$ we have $\Phi^L(\sigma,s) \asymp \psi_s^*$ where ψ_s^* is the minimax rate on Θ_s. Thus, it is not possible to achieve a faster rate than $\Phi^L(\sigma,s)$ for $s < d^{1/4}$, and therefore (10) is equivalent to the condition

$$\min_{s\geq d^{1/4}} \frac{\Psi_d'(s)}{\Phi^L(\sigma,s)} \to 0,$$

and

$$\min_{s=1,\dots,d} \frac{\Psi_d'(s)}{\Phi^L(\sigma,s)} \asymp \min_{s\geq d^{1/4}} \frac{\Psi_d'(s)}{\Phi^L(\sigma,s)}.$$

Obviously, $\Psi_d'(s)$ cannot be of smaller order than the minimax rate ψ_s^*, which implies that

$$\min_{s\geq d^{1/4}} \frac{\Psi_d'(s)}{\Phi^L(\sigma,s)} \geq \min_{s\geq d^{1/4}} \frac{c\psi_s^*}{\Phi^L(\sigma,s)} = \min_{s\geq d^{1/4}} \frac{c\log(1+d/s^2)}{\log(1+d(\log d)/s^2)} \geq \frac{c'}{\log d}$$

where $c, c' > 0$ are absolute constants. On the other hand, Theorem 2 yields

$$\frac{C'\Psi_d'(1)}{\Phi^L(\sigma,1)} \geq \frac{C'C_1\sigma^2 d^{1/4}}{\Phi^L(\sigma,1)} = \frac{C'C_1 d^{1/4}}{\log(1+d(\log d))}\ .$$

Combining the last three displays, we find

$$\frac{\Psi_d'(1)}{\Phi^L(\sigma,1)} \min_{s=1,\dots,d} \frac{\Psi_d'(s)}{\Phi^L(\sigma,s)} \geq \frac{c'C'C_1 d^{1/4}}{(\log d)\log(1+d(\log d))} \to \infty,$$

as $d \to \infty$, thus proving (11) with $\bar{s} = 1$.

5.3. **Proof of Proposition 2.** Since in this proof we consider different values of σ, we denote the probability distribution of $(y_1,\dots,y_d)$ satisfying (1) by $\mathbf{P}_{\theta,\sigma^2}$. Let $\mathbf{E}_{\theta,\sigma^2}$ be the corresponding expectation. Assume that $\widehat{T}$ satisfies (18) with $C_0 = 1/512$. We will prove that (19) holds for $\sigma = 1$. The extension to arbitrary $\sigma > 0$ is straightforward and is therefore omitted.

Let $a > 1$ be a positive number and let μ be the d-dimensional normal distribution with zero mean and covariance matrix $a^2\mathbf{I}_d$ where $\mathbf{I}_d$ is the identity matrix. In what follows, we consider the mixture probability measure $\mathbb{P}_\mu = \int_{\Theta_d} \mathbf{P}_{\theta,1}\mu(d\theta)$. Observe that $\mathbb{P}_\mu = \mathbf{P}_{0,1+a^2}$.

Fixing $\theta = 0$ and $\sigma^2 = 1 + a^2$ in (18), we get $\mathbf{E}_{0,1+a^2}\big[\widehat{T}^2\big] \le 2C_0 a^2 d$ and therefore $\mathbf{P}_{0,1+a^2}(|\widehat{T}| \ge \frac{1}{8}a\sqrt{d}) \le \frac{1}{4}$. Since $\mathbb{P}_\mu = \mathbf{P}_{0,1+a^2}$, this implies

$$\mathbb{P}_\mu\big(|\widehat{T}| < \frac{1}{8}a\sqrt{d}\big) > \frac{3}{4}. \tag{48}$$

For θ distributed according to μ, $L(\theta)$ has a normal distribution with mean 0 and variance $a^2 d$. Hence, using the table of standard normal distribution, we find

$$\mu\Big(|L(\theta)| \le \frac{a}{4}\sqrt{d}\Big) < \frac{1}{4}\ .$$

Combining this with (48), we conclude that, with $\mathbb{P}_\mu$-probability greater than $1/2$, we have simultaneously $|L(\theta)| > a\sqrt{d}/4$ and $|\widehat{T}| < a\sqrt{d}/8$. Hence,

$$\sup_{\theta\in\Theta_d} \mathbf{E}_{\theta,1}\big[\big(\widehat{T} - L(\theta)\big)^2\big] \ge \mathbb{E}_\mu\big[(\widehat{T} - L(\theta))^2\big] \ge \frac{1}{128}a^2 d$$

where $\mathbb{E}_\mu$ denotes the expectation with respect to $\mathbb{P}_\mu$. The result now follows by letting a tend to infinity.

Acknowledgement. The work of A.B.Tsybakov was supported by GENES and by the French National Research Agency (ANR) under the grants IPANEMA (ANR-13-BSH1-0004-02) and Labex Ecodec (ANR-11-LABEX-0047). It was also supported by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

References

[1] Cai, T. T. and Low, M.L. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist.* **32** 552–576.
[2] Cai, T. T. and Low, M.L. (2005). On adaptive estimation of linear functionals. *Ann. Statist.* **33** 2311–2343.
[3] Collier, O., Comminges, L., and Tsybakov, A.B. (2016). Minimax estimation of linear and quadratic functionals under sparsity constraints. *Ann. Statist.*, to appear.
[4] Golubev, G.K. (2004). The method of risk envelopes in the estimation of linear functionals. *Problemy Peredachi Informatsii* **40** 58–72.
[5] Golubev, Y. and Levit, B. (2004). An oracle approach to adaptive estimation of linear functionals in a Gaussian model. *Math. Methods Statist.* **13** 392–408.
[6] Ibragimov, I.A. and Hasminskii, R.Z. *Nonparametric estimation of the value of a linear functional in Gaussian white noise.* Theory Probab. Appl., 29, 18-32, 1984.
[7] Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces.* Springer, Berlin, Heidelberg.
[8] Laurent, B., Ludena, C. and Prieur, C. (2008). Adaptive estimation of linear functionals by model selection. *Electron. J. Stat.* **2** 993–1020.
[9] Petrov, V.V. (1995). *Limit Theorems of Probability Theory.* Clarendon Press, Oxford.
[10] Tsybakov A.B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** 2420–2469.
[11] Tsybakov A.B. (2009). *Introduction to Nonparameric Estimation.* Springer, New York.
[12] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90.