HAL
open science

Minimax estimation of linear and quadratic functionals on sparsity classes

Olivier Collier, Laëtitia Comminges, Alexandre Tsybakov

To cite this version:

Olivier Collier, Laëtitia Comminges, Alexandre Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. 2015. hal-01215833

HAL Id: hal-01215833

https://hal.science/hal-01215833v1

Preprint submitted on 15 Oct 2015

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.

Minimax estimation of linear and quadratic functionals on sparsity classes

Olivier Collier[1], Laëtitia Comminges[2] and Alexandre B. Tsybakov[3]

[1]Modal'X, Université Paris-Ouest
[2]Université Paris-Dauphine
[3]CREST-ENSAE

Abstract

For the Gaussian sequence model, we obtain non-asymp- totic minimax rates of estimation of the linear, quadratic and the ℓ_2-norm functionals on classes of sparse vectors and construct optimal estimators that attain these rates. The main object of interest is the class $B_0(s)$ of s-sparse vectors $\theta = (\theta_1, \dots, \theta_d)$, for which we also provide completely adaptive estimators (independent of s and of the noise variance σ) having logarithmically slower rates than the minimax ones. Furthermore, we obtain the minimax rates on the ℓ_q-balls $B_q(r) = \{\theta \in \mathbb{R}^d : \|\theta\|_q \le r\}$ where $0 < q \le 2$, and $\|\theta\|_q = \left(\sum_{i=1}^d |\theta_i|^q\right)^{1/q}$. This analysis shows that there are, in general, three zones in the rates of convergence that we call the sparse zone, the dense zone and the degenerate zone, while a fourth zone appears for estimation of the quadratic functional. We show that, as opposed to estimation of θ, the correct logarithmic terms in the optimal rates for the sparse zone scale as $\log(d/s^2)$ and not as $\log(d/s)$. For the class $B_0(s)$, the rates of estimation of the linear functional and of the ℓ_2-norm have a simple elbow at $s = \sqrt{d}$ (boundary between the sparse and the dense zones) and exhibit similar performances, whereas the estimation of the quadratic functional $Q(\theta)$ reveals more complex effects: the minimax risk on $B_0(s)$ is infinite and the sparseness assumption needs to be combined with a bound on the ℓ_2-norm. Finally, we apply our results on estimation of the ℓ_2-norm to the problem of testing against sparse alternatives. In particular, we obtain a non-asymptotic analog of the Ingster-Donoho-Jin theory revealing some effects that were not captured by the previous asymptotic analysis.

Keywords: nonasymptotic minimax estimation; linear functional; quadratic functional; sparsity; unknown noise variance; thresholding

1 Introduction

In this paper, 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 linear and quadratic functionals

$$L(\theta) = \sum_{i=1}^d \theta_i, \qquad \text{and} \qquad Q(\theta) = \sum_{i=1}^d \theta_i^2,$$

and of the ℓ_2-norm

$$\|\theta\|_2 = \sqrt{Q(\theta)}$$

based on the observations $y_1, \dots, y_d$.

In this paper, we assume that θ belongs to a given subset Θ of $\mathbb{R}^d$. We will be considering classes Θ with elements satisfying the sparsity constraints $\|\theta\|_0 \le s$ where $\|\theta\|_0$ denotes the number of non-zero components of θ, or $\|\theta\|_q \le r$ where

$$\|\theta\|_q = \left(\sum_{i=1}^{d} |\theta_i|^q \right)^{1/q}.$$

Here, $r, q > 0$ and the integer $s \in [1, d]$ are given constants.

Let $T(\theta)$ be one of the functionals $L(\theta)$, $Q(\theta)$ or $\sqrt{Q(\theta)}$. As a measure of quality of an estimator $\hat{T}$ of the functional $T(\theta)$, we consider the maximum squared risk

$$\sup_{\theta \in \Theta} \mathbf{E}_\theta (\hat{T} - T(\theta))^2,$$

where $\mathbf{E}_\theta$ denotes the expectation with respect to the probability measure $\mathbf{P}_\theta$ of the vector of observations $(y_1, \dots, y_d)$ satisfying (1). The best possible quality is characterized by the minimax risk

$$R_T^*(\Theta) = \inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbf{E}_\theta (\hat{T} - T(\theta))^2,$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators. In this paper, we find minimax optimal estimators of $T(\theta)$, i.e., estimators $\tilde{T}$ such that

$$\sup_{\theta \in \Theta} \mathbf{E}_\theta (\tilde{T} - T(\theta))^2 \asymp R_T^*(\Theta). \tag{2}$$

Here and below, we write $a \asymp b$ if $c \le a/b \le C$ for some absolute positive constants c and C. Note that the minimax optimality is considered here in the non-asymptotic sense, i.e., (2) should hold for all d and σ.

The literature on minimax estimation of linear and quadratic functionals is rather extensive. The analysis of estimators of linear functionals from the minimax point of view was initiated in [21] while for the quadratic functionals we refer to [15]. These papers, as well as the subsequent publications [10, 11, 14, 16, 18, 19, 26, 27, 29, 30, 31, 32, 33, 35, 36], focus on minimax estimation of functionals on the classes Θ describing the smoothness properties of functions in terms of their Fourier or wavelet coefficients. Typical examples are Sobolev ellipsoids, hyperrectangles or Besov bodies while a typical example of linear functional is the value of a smooth function at a point. In this framework, a deep analysis of estimation of functionals is now available including the minimax rates (and in some cases the minimax constants), oracle inequalities and adaptation. Extensions to linear inverse problems have been considered in detail by [7, 8, 17]. Note that classes Θ studied in this literature are convex classes. Estimation of functionals on the non-convex sparsity classes $B_0(s) = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \le s\}$ or $B_q(r) = \{\theta \in \mathbb{R}^d : \|\theta\|_q \le r\}$ with $0 < q < 1$ has received much less attention. We are only aware of the paper [9], which establishes upper and lower bounds on the minimax risk for estimators of the linear functional $L(\theta)$ on the class $B_0(s)$. However, that paper considers the special case when $s < d^a$ for some $a < 1/2$, and $\sigma = 1/\sqrt{d}$ and there is a logarithmic gap between the upper and lower bounds. Minimax rates for the estimation of $Q(\theta)$ and of the

ℓ_2-norm on the classes $B_0(s)$ and $B_q(r)$, $0 < q < 2$, were not studied. Note, that estimation the ℓ_2-norm is closely related to minimax optimal testing of hypotheses under the ℓ_2 separation distance in the spirit of [24]. Indeed, the optimal tests for this problem are based on estimators of the ℓ_2-norm. A non-asymptotic study of minimax rates of testing for the classes $B_0(s)$ and $B_q(r)$, $0 < q < 2$, is given in [4] and [40]. But for the testing problem, the risk function is different and these papers do not provide results on the estimation of the ℓ_2-norm. Note also that the upper bounds on the minimax rates of testing in [4] and [40] depart from the lower bounds by a logarithmic factor.

In this paper, we find non-asymptotic minimax rates of estimation of the above three functionals on the sparsity classes $B_0(s)$, $B_q(r)$ and construct optimal estimators that attain these rates. We deal with non-convex classes B_q $(0 < q < 1)$ for the linear functional and with the classes that are not quadratically convex $(0 < q < 2)$ for $Q(\theta)$ and of the ℓ_2-norm. Our main object of interest is the class $B_0(s)$, for which we also provide completely adaptive estimators (independent of σ and s) having logarithmically slower rates than the minimax ones. Some interesting effects should be noted. First, we show that, for the linear functional and the ℓ_2-norm there are, in general, three zones in the rates of convergence that we call the sparse zone, the dense zone and the degenerate zone, while for the quadratic functional an additional fourth zone appears. Next, as opposed to estimation of the vector θ in the ℓ_2-norm, cf. [13, 5, 1, 28, 37, 40], the correct logarithmic terms in the optimal rates for the sparse zone scale as $\log(d/s^2)$ and not as $\log(d/s)$. Noteworthy, for the class $B_0(s)$, the rates of estimation of the linear functional and of the ℓ_2-norm have a simple elbow at $s = \sqrt{d}$ (boundary between the sparse and the dense zones) and exhibit similar performances, whereas the estimation of the quadratic functional $Q(\theta)$ reveals more complex effects and is not possible only on the basis of sparsity described by the condition $\theta \in B_0(s)$. Finally, we apply our results on estimation of the ℓ_2-norm to the problem of testing against sparse alternatives. In particular, we obtain a non-asymptotic analog of Ingster-Donoho-Jin theory revealing some effects that were not captured by the previous asymptotic analysis.

2 Minimax estimation of the linear functional

In this section, we study the minimax rates of estimation of the linear functional $L(\theta)$ and we construct minimax optimal estimators.

Assume first that Θ is the class of s-sparse vectors $B_0(s) = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \le s\}$ where s is a given integer, $1 \le s \le d$. Consider the estimator

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

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function.

The following theorem shows that

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

is the minimax rate of estimation of the linear functional on the class $B_0(s)$ and that $\hat{L}$ is a minimax optimal estimator.

Theorem 1. *There exist absolute constants $c > 0, C > 0$ such that, for any integers s, d satisfying $1 \le s \le d$, and any $\sigma > 0$,*

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta(\hat{L} - L(\theta))^2 \le C\psi_\sigma^L(s,d), \tag{3}$$

and

$$R_L^*(B_0(s)) \ge c\psi_\sigma^L(s,d). \tag{4}$$

Proofs of (3) and of (4) are given in Sections 8 and 7 respectively. Note that since $\log(1+u) \ge u/2$ for $0 < u \le 1$, and $\log(1+u) \le u$ we have

$$\sigma^2 s^2 \log(1 + d/s^2) \asymp \min(\sigma^2 s^2 \log(1 + d/s^2), \sigma^2 d) \tag{5}$$

for all $1 \le s \le d$. This writing clarifies the fact that the rate exhibits a "hidden" elbow at $s = \sqrt{d}$. Thus,

$$R_L^*(B_0(s)) \asymp \min(\sigma^2 s^2 \log(1 + d/s^2), \sigma^2 d). \tag{6}$$

We consider now the classes $B_q(r) = \{\theta \in \mathbb{R}^d : \|\theta\|_q \le r\}$, where $0 < q \le 1$, and r is a positive number. For any $r, \sigma, q > 0$ any integer $d \ge 1$, we define the intege

$$m = \max\{s \in \{1, \dots, d\} : \sigma^2 \log(1 + d/s^2) \le r^2 s^{-2/q}\} \tag{7}$$

if the set $\{s \in \{1, \dots, d\} : \sigma^2 \log(1 + d/s^2) \le r^2 s^{-2/q}\}$ is non-empty, and we put $m = 0$ if this set is empty. The next two theorems show that the optimal rate of convergence of estimators of the linear functional on the class $B_q(r)$ is of the form:

$$\psi_{\sigma,q}^L(r,d) = \begin{cases} \sigma^2 m^2 \log(1 + d/m^2) & \text{if} \quad m \ge 1, \\ r^2 & \text{if} \quad m = 0. \end{cases}$$

The following theorem shows that $\psi_{\sigma,q}^L(r,d)$ is a lower bound on the convergence rate of the minimax risk of the linear functional on the class $B_q(r)$.

Theorem 2. *If $0 < q \le 1$, then there exists a constant $c > 0$ such that, for any integer $d \ge 1$ and any $r, \sigma > 0$, we have*

$$R_L^*(B_q(r)) \ge c\psi_{\sigma,q}^L(r,d). \tag{8}$$

The proof of Theorem 2 is given in Section 7.

We now turn to the construction of minimax optimal estimators on $B_q(r)$. For $0 < q \le 1$, define the following statistic

$$\hat{L}_q = \begin{cases} \sum_{j=1}^d y_j & \text{if} \quad m > \sqrt{d}, \\ \sum_{j=1}^d y_j \, \mathbb{1}_{\{|y_j| > 2\sigma\sqrt{2\log(1+d/m^2)}\}} & \text{if} \quad 1 \le m \le \sqrt{d}, \\ 0 & \text{if} \quad m = 0. \end{cases}$$

Theorem 3. *Let $0 < q \le 1$. There exists a constant $C > 0$ such that, for any integer $d \ge 1$ and any $r, \sigma > 0$, we have*

$$\sup_{\theta \in B_q(r)} \mathbf{E}_\theta(\hat{L}_q - L(\theta))^2 \le C\psi_{\sigma,q}^L(r,d). \tag{9}$$

The proof of Theorem 3 is given in Section 8. Theorems 2 and 3 imply that $\psi^L_{\sigma,q}(r,d)$ is the minimax rate of estimation of the linear functional on the ball $B_q(r)$ and that $\hat{L}_q$ is a minimax optimal estimator.

Some remarks are in order here. Apart from the degenerate case $m = 0$ when the zero estimator is optimal, we obtain on $B_q(r)$ the same expression for the optimal rate as on the class $B_0(s)$, with the difference that the sparsity s is now replaced by the "effective sparsity" m. Heuristically, m is obtained as a solution of

$$\sigma^2 m^2 \log(1 + d/m^2) \asymp r^2 m^{2-2/q}$$

where the left hand side represents the estimation error for m-sparse signals established in Theorem 1 and the right hand side gives the error of approximating a vector from $B_q(r)$ by an m-sparse vector in squared ℓ_1-norm. Note also that, in view of (5), we can equivalently write the optimal rate in the form

$$\psi^L_{\sigma,q}(r,d) \asymp \begin{cases} \sigma^2 d & \text{if} \quad m > \sqrt{d}, \\ \sigma^2 m^2 \log(1 + d/m^2) & \text{if} \quad 1 \le m \le \sqrt{d}, \\ r^2 & \text{if} \quad m = 0. \end{cases}$$

Thus, the optimal rate on $B_q(r)$ has in fact three regimes that we will call the dense zone ($m > \sqrt{d}$), the sparse zone ($1 \le m \le \sqrt{d}$), and the degenerate zone ($m = 0$). Furthermore, it follows from the definition of m that the rate $\psi^L_{\sigma,q}(r,d)$ in the sparse zone is of the order $\sigma^2 (r/\sigma)^{2q} \log^{1-q}(1 + d(\sigma/r)^{2q})$, which leads to

$$\psi^L_{\sigma,q}(r,d) \asymp \begin{cases} \sigma^2 d & \text{if} \quad m > \sqrt{d}, \\ \sigma^2 (r/\sigma)^{2q} \log^{1-q}(1 + d(\sigma/r)^{2q}) & \text{if} \quad 1 \le m \le \sqrt{d}, \\ r^2 & \text{if} \quad m = 0. \end{cases}$$

In particular, for $q = 1$, the logarithmic factor disappears from the rate, and the optimal rates in the sparse and degenerate zones are both equal to r^2. Therefore, for $q = 1$, there is no need to introduce thresholding in the definition of $\hat{L}_q$, and it is enough to use only the zero estimator for $m \le \sqrt{d}$ and the estimator $\sum_{j=1}^d y_j$ for $m > \sqrt{d}$ to achieve the optimal rate.

Remark 1. In this section and throughout the paper, theorems on the minimax lower bounds are stated for the squared loss function only. However, the proofs in Section 7 are given for the indicator loss function, which is more general. For each of the considered classes Θ, they have the form

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbf{P}_\theta\big(|\hat{T} - T(\theta)| \ge \psi\big) \ge c, \tag{10}$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators, ψ is the corresponding minimax optimal rate and $c > 0$ is an absolute constant. Clearly, (10) implies lower bounds for the minimax risk with any monotone non-decreasing loss function on $\mathbb{R}_+$ taking value 0 at 0.

3 Minimax estimation of the quadratic functional

Consider now the problem of estimation of the quadratic functional $Q(\theta) = \sum_{i=1}^d \theta_i^2$. For any integers s, d satisfying $1 \le s \le d$, and any $\sigma > 0$, we introduce the notation

$$\bar{\psi}_\sigma(s,d) = \begin{cases} \sigma^4 s^2 \log^2(1 + d/s^2) & \text{if} \quad s < \sqrt{d}, \\ \sigma^4 d & \text{if} \quad s \ge \sqrt{d}. \end{cases}$$

The following theorem shows that

$$\psi_\sigma^Q(s,d,\kappa) = \min\{\kappa^4, \max\{\sigma^2\kappa^2, \bar\psi_\sigma(s,d)\}\}$$

is a lower bound on the convergence rate of the minimax risk of the quadratic functional on the class $B_2(\kappa)\cap B_0(s)$, where $B_2(\kappa) = \{\theta\in\mathbb{R}^d : \|\theta\|_2 \le \kappa\}$.

Theorem 4. *There exists an absolute constant $c>0$ such that, for any integers s,d satisfying $1\le s\le d$, and any $\kappa,\sigma>0$, we have*

$$R_Q^*(B_2(\kappa)\cap B_0(s)) \ge c\,\psi_\sigma^Q(s,d,\kappa). \tag{11}$$

The proof of Theorem 4 is given in Section 7.

Remark 2. Note that the minimax risk $R_Q^*(B_2(\kappa)\cap B_0(s))$ is monotone non-decreasing in s while the right hand side of (11) is not monotone in s. Nevertheless, there is no problem since $\bar\psi_\sigma(s,d)$ is equivalent, up to absolute constants, to a monotone function of s, for which (11) remains valid with another constant c. For example, we have

$$\bar\psi_\sigma(s,d) \asymp d\, f^2\big(\min\{1, s/\sqrt{d}\}\big), \tag{12}$$

where $f(t) = t\log(1+4/t^2)$. It is easy to check that the right hand side of (12) is monotone in s.

One of the consequences of Theorem 4 is that $R_Q^*(B_0(s)) = \infty$ (set $\kappa=\infty$ in (11)). Thus, only smaller classes than $B_0(s)$ are of interest when estimating the quadratic functional. The class $B_2(\kappa)\cap B_0(s)$ naturally arises in this context but other classes can be considered as well.

We now turn to the construction of minimax optimal estimator on $B_2(\kappa)\cap B_0(s)$. Set

$$\alpha_s = \mathbf{E}\big(X^2 \,|\, X^2 > 2\log(1+d/s^2)\big) = \frac{\mathbf{E}\big(X^2 \mathbb{1}_{\{|X|>\sqrt{2\log(1+d/s^2)}\}}\big)}{\mathbf{P}\big(|X| > \sqrt{2\log(1+d/s^2)}\,\big)},$$

where $X\sim\mathcal{N}(0,1)$ denotes the standard normal random variable. Introduce the notation

$$\psi_\sigma(s,d,\kappa) = \max\{\sigma^2\kappa^2, \bar\psi_\sigma(s,d)\}.$$

Thus,

$$\psi_\sigma^Q(s,d,\kappa) = \min\{\kappa^4, \psi_\sigma(s,d,\kappa)\}. \tag{13}$$

Define the following statistic

$$\hat Q = \begin{cases} \sum_{j=1}^d (y_j^2 - \alpha_s\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\sqrt{2\log(1+d/s^2)}\}} & \text{if} \quad s<\sqrt{d} \text{ and } \kappa^4 \ge \psi_\sigma(s,d,\kappa),\\ \sum_{j=1}^d y_j^2 - d\sigma^2 & \text{if} \quad s\ge\sqrt{d} \text{ and } \kappa^4 \ge \psi_\sigma(s,d,\kappa),\\ 0 & \text{if} \quad \kappa^4 < \psi_\sigma(s,d,\kappa). \end{cases}$$

Theorem 5. *There exists an absolute constant $C>0$ such that, for any integers s,d satisfying $1\le s\le d$, and any $\kappa,\sigma>0$, we have*

$$\sup_{\theta\in B_2(\kappa)\cap B_0(s)} \mathbf{E}_\theta(\hat Q - Q(\theta))^2 \le C\,\psi_\sigma^Q(s,d,\kappa). \tag{14}$$

The proof of Theorem 5 is given in Section 8. Theorems 4 and 5 imply that $\psi_\sigma^Q(s,d,\kappa)$ is the minimax rate of estimation of the quadratic functional on the class $B_2(\kappa)\cap B_0(s)$ and that $\hat{Q}$ is a minimax optimal estimator.

As a corollary, we obtain the minimax rate of convergence on the class $B_2(\kappa)$ (set $s=d$ in Theorems 4 and 5). In this case, the estimator $\hat{Q}$ takes the form

$$\hat{Q}_* = \begin{cases} \sum_{j=1}^d y_j^2 - d\sigma^2 & \text{if} \quad \kappa^4 \geq \max\{\sigma^2\kappa^2, \sigma^4 d\}, \\ 0 & \text{if} \quad \kappa^4 < \max\{\sigma^2\kappa^2, \sigma^4 d\}. \end{cases}$$

Corollary 1. *There exist absolute constants $c, C > 0$ such that, for any $\kappa, \sigma > 0$, we have*

$$\sup_{\theta\in B_2(\kappa)} \mathbf{E}_\theta(\hat{Q}_* - Q(\theta))^2 \leq C\min\{\kappa^4, \max(\sigma^2\kappa^2, \sigma^4 d)\}, \tag{15}$$

and

$$R_Q^*(B_2(\kappa)) \geq c\min\{\kappa^4, \max(\sigma^2\kappa^2, \sigma^4 d)\}. \tag{16}$$

Note that the upper bounds of Theorem 5 and Corollary 1 obviously remain valid for the positive part estimators $\hat{Q}_+ = \max\{\hat{Q}, 0\}$, and $\hat{Q}_{*,+} = \max\{\hat{Q}_*, 0\}$. The upper rate as in (15) on the class $B_2(\kappa)$ with an extra logarithmic factor is obtained for different estimators in [26, 27].

Alternatively, we consider the classes $B_q(r)$, where r is a positive number and $0 < q < 2$. As opposed to the case of $B_0(s)$, we do not need to consider intersection with $B_2(\kappa)$. Indeed, it is granted that the ℓ_2-norm of θ is uniformly bounded thanks to the inclusion $B_q(r) \subseteq B_2(r)$. For any $r, \sigma > 0$, $0 < q < 2$, and any integer $d \geq 1$ we set

$$\psi_{\sigma,q}^Q(r,d) = \begin{cases} \max\{\sigma^2 r^2, \sigma^4 d\} & \text{if} \quad m > \sqrt{d}, \\ \max\{\sigma^2 r^2, \sigma^4 m^2 \log^2(1+d/m^2)\} & \text{if} \quad 1 \leq m \leq \sqrt{d}, \\ r^4 & \text{if} \quad m = 0, \end{cases}$$

where m is the integer defined above (cf. (7)) and depending only on d, r, σ, q. The following theorem shows that $\psi_{\sigma,q}^Q(r,d)$ is a lower bound on the convergence rate of the minimax risk of the quadratic functional on the class $B_q(r)$.

Theorem 6. *Let $0 < q < 2$. There exists a constant $c > 0$ such that, for any integer $d \geq 1$, and any $r, \sigma > 0$, we have*

$$R_Q^*(B_q(r)) \geq c\,\psi_{\sigma,q}^Q(r,d). \tag{17}$$

We now turn to the construction of minimax optimal estimators on $B_q(r)$. Consider the following statistic

$$\hat{Q}_q = \begin{cases} \sum_{j=1}^d y_j^2 - d\sigma^2 & \text{if} \quad m > \sqrt{d}, \\ \sum_{j=1}^d (y_j^2 - \tilde{\alpha}_m \sigma^2)\, \mathbb{1}_{\{|y_j| > 2\sigma\sqrt{2\log(1+d/m^2)}\}} & \text{if} \quad 1 \leq m \leq \sqrt{d}, \\ 0 & \text{if} \quad m = 0, \end{cases}$$

where $\tilde{\alpha}_m = \mathbf{E}\big(X^2 \,|\, X^2 > 8\log(1+d/m^2)\big)$, $X \sim \mathcal{N}(0,1)$.

Theorem 7. *Let $0 < q < 2$. There exists a constant $C > 0$ such that, for any integer $d \geq 1$, and any $r, \sigma > 0$, we have*

$$\sup_{\theta \in B_q(r)} \mathbf{E}_\theta(\hat{Q}_q - Q(\theta))^2 \leq C\psi^Q_{\sigma,q}(r,d). \tag{18}$$

The proof of Theorem 7 is given in Section 8. Theorems 6 and 7 imply that $\psi^Q_{\sigma,q}(r,d)$ is the minimax rate of estimation of the quadratic functional on the class $B_q(r)$ and that $\hat{Q}_q$ is a minimax optimal estimator.

Notice that, in view of the definition of m, in the sparse zone we have

$$\sigma^4 m^2 \log^2(1 + d/m^2) \asymp \sigma^4 (r/\sigma)^{2q} \log^{2-q}(1 + d(\sigma/r)^{2q}),$$

which leads to

$$\psi^Q_{\sigma,q}(r,d) \asymp \begin{cases} \max\{\sigma^2 r^2, \sigma^4 d\} & \text{if} \quad m > \sqrt{d}, \\ \max\{\sigma^2 r^2, \sigma^4 (r/\sigma)^{2q} \log^{2-q}(1 + d(\sigma/r)^{2q})\} & \text{if} \quad 1 \leq m \leq \sqrt{d}, \\ r^4 & \text{if} \quad m = 0. \end{cases}$$

One can check that for $q = 2$ this rate is of the same order as the rate obtained in Corollary 1.

4 Minimax estimation of the ℓ_2-norm

Interestingly, the minimax rates of estimation of the ℓ_2-norm $\|\theta\|_2 = \sqrt{Q(\theta)}$ do not degenerate as the radius κ grows to infinity, as opposed to the rates for $Q(\theta)$ established above. It turns out that the restriction to $B_2(\kappa)$ is not needed to get meaningful results for estimation of $\sqrt{Q(\theta)}$ on the sparsity classes. We drop this restriction and assume that $\Theta = B_0(s)$. Consider the estimator

$$\hat{N} = \sqrt{\max\{\hat{Q}_\bullet, 0\}}$$

where

$$\hat{Q}_\bullet = \begin{cases} \sum_{j=1}^d (y_j^2 - \alpha_s \sigma^2)\, \mathbb{1}_{\{|y_j| > \sigma\sqrt{2\log(1+d/s^2)}\}} & \text{if} \quad s < \sqrt{d}, \\ \sum_{j=1}^d y_j^2 - d\sigma^2 & \text{if} \quad s \geq \sqrt{d}. \end{cases}$$

The following theorem shows that $\hat{N}$ is a minimax optimal estimator of the ℓ_2-norm $\|\theta\|_2 = \sqrt{Q(\theta)}$ on the class $B_0(s)$ and that the corresponding minimax rate of convergence is

$$\psi^{\sqrt{Q}}_\sigma(s,d) = \begin{cases} \sigma^2 s \log(1 + d/s^2) & \text{if} \quad s < \sqrt{d}, \\ \sigma^2 \sqrt{d} & \text{if} \quad s \geq \sqrt{d}. \end{cases}$$

Theorem 8. *There exist absolute constants $c > 0, C > 0$ such that, for any integers s, d satisfying $1 \leq s \leq d$, and any $\sigma > 0$,*

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta(\hat{N} - \|\theta\|_2)^2 \leq C\psi^{\sqrt{Q}}_\sigma(s,d), \tag{19}$$

and

$$R^*_{\sqrt{Q}}(B_0(s)) \geq c\psi^{\sqrt{Q}}_\sigma(s,d). \tag{20}$$

Proofs of (19) and of (20) are given in Sections 8 and 7 respectively.

Our next step is to analyze the classes $B_q(r)$. For any $r, \sigma > 0$, $0 < q < 2$, and any integer $d \geq 1$ we set

$$\psi_{\sigma,q}^{\sqrt{Q}}(r,d) = \begin{cases} \sigma^2\sqrt{d} & \text{if} \quad m > \sqrt{d}, \\ \sigma^2 m \log(1 + d/m^2) & \text{if} \quad 1 \leq m \leq \sqrt{d}, \\ r^2 & \text{if} \quad m = 0, \end{cases}$$

where m is the integer defined above (cf. (7)) and depending only on d, r, σ, q. The estimator that we consider when θ belongs to the class $B_q(r)$ is

$$\hat{N}_q = \sqrt{\max\{\hat{Q}_q, 0\}}.$$

Theorem 9. *Let $0 < q < 2$. There exist constants $C, c > 0$ such that, for any integer $d \geq 1$, and any $r, \sigma > 0$, we have*

$$\sup_{\theta \in B_q(r)} \mathbf{E}_\theta(\hat{N}_q - \|\theta\|_2)^2 \leq C\psi_{\sigma,q}^{\sqrt{Q}}(r,d), \tag{21}$$

and

$$R^*_{\sqrt{Q}}(B_q(r)) \geq c\psi_{\sigma,q}^{\sqrt{Q}}(r,d). \tag{22}$$

Proofs of (21) and of (22) are given in Sections 8 and 7 respectively.

As in the case of linear and quadratic functionals, we have an equivalent expression for the optimal rate:

$$\psi_{\sigma,q}^{\sqrt{Q}}(r,d) \asymp \begin{cases} \sigma^2\sqrt{d} & \text{if} \quad m > \sqrt{d}, \\ \sigma^2 (r/\sigma)^q \log^{1-q/2}(1 + d(\sigma/r)^{2q}) & \text{if} \quad 1 \leq m \leq \sqrt{d}, \\ r^2 & \text{if} \quad m = 0. \end{cases}$$

Though we formally did not consider the case $q = 2$, note that the logarithmic factor disappears from the above expression when $q = 2$, and the optimal rates in the sparse and degenerate zones are both equal to r^2. This suggests that, for $q = 2$, there is no need to introduce thresholding in the definition of $\hat{N}_q$, and it is enough to use only the zero estimator for $m \leq \sqrt{d}$ and the estimator $\big(\max\big\{\sum_{j=1}^d y_j^2 - d\sigma^2, 0\big\}\big)^{1/2}$ for $m > \sqrt{d}$ to achieve the optimal rate.

5 Estimation with unknown noise level

In this section, we discuss modifications of the above estimators when the noise level σ is unknown. A general idea leading to our construction is that the smallest y_j^2 are likely to correspond to zero components of θ, and thus to contain information on σ not corrupted by θ. Here, we will demonstrate this idea only for estimation of s-sparse vectors in the case $s \leq \sqrt{d}$. Then, not more than $d - \sqrt{d}$ smallest y_j^2 can be used for estimation of the variance. Throughout this section, we assume that $d \geq 3$.

We start by considering estimation of the linear functional. Then it is enough to replace σ in the definition of $\hat{L}$ by the following statistic

$$\hat{\sigma} = 3\Big(\frac{1}{d} \sum_{j \leq d - \sqrt{d}} y_{(j)}^2\Big)^{1/2}$$

where $y_{(j)}^2 \le \cdots \le y_{(d)}^2$ are the order statistics associated to $y_1^2, \dots, y_d^2$. Note that $\hat\sigma$ is not a good estimator of σ but rather an over-estimator. The resulting estimator of $L(\theta)$ is

$$\tilde L = \sum_{j=1}^d y_j \ \mathbb{1}_{\{|y_j| > \hat\sigma\sqrt{2\log(1+d/s^2)}\}}.$$

Theorem 10. *There exists an absolute constant C such that, for any integers s and d satisfying $s \le \sqrt{d}$, and any $\sigma > 0$,*

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta(\tilde L - L(\theta))^2 \le C\psi_\sigma^L(s,d).$$

The proof of Theorem 10 is given in Section 8.

Note that the estimator $\tilde L$ depends on s. To turn it into a completely data-driven one, we may consider

$$\tilde L' = \sum_{j=1}^d y_j \ \mathbb{1}_{\{|y_j| > \hat\sigma\sqrt{2\log d}\}}.$$

Inspection of the proof of Theorem 10 leads to the conclusion that

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta(\tilde L' - L(\theta))^2 \le C\sigma^2 s^2 \log d. \tag{23}$$

Thus, the rate for the data-driven estimator $\tilde L'$ is not optimal but the deterioration is only in the expression under the logarithm.

A data-driven estimator of the quadratic functional can be taken in the form:

$$\tilde Q = \sum_{j=1}^d y_j^2 \ \mathbb{1}_{\{|y_j| > \hat\sigma\sqrt{2\log d}\}}.$$

The following theorem shows that the estimator $\tilde Q$ is nearly minimax on $B_2(\kappa) \cap B_0(s)$ for $s \le \sqrt{d}$.

Theorem 11. *There exists an absolute constant C such that, for any integers s and d satisfying $s \le \sqrt{d}$, and any $\sigma > 0$,*

$$\sup_{\theta \in B_2(\kappa) \cap B_0(s)} \mathbf{E}_\theta(\tilde Q - Q(\theta))^2 \le C \max\Big\{\sigma^2\kappa^2, \sigma^4 s^2 \log^2 d\Big\}.$$

The proof of Theorem 11 is given in Section 8.

6 Consequences for the problem of testing

The results on estimation of the ℓ_2-norm stated above allow us to obtain the solution of the problem of non-asymptotic minimax testing on the classes $B_0(s)$ and $B_q(r)$ under the ℓ_2 separation distance. For $q \ge 0$, $u > 0$, and $\delta > 0$, consider the set

$$\Theta_{q,u}(\delta) = \{\theta \in B_q(u) : \ \|\theta\|_2 \ge \delta\}.$$

Assume that we wish to test the hypothesis $\mathbf{H}_0 : \theta = 0$ against the alternative

$$\mathbf{H}_1 : \ \theta \in \Theta_{q,u}(\delta).$$

Let Δ be a test statistic with values in $\{0,1\}$. We define the risk of test Δ as the sum of the first type error and the maximum second type error:

$$\mathbf{P}_0(\Delta = 1) + \sup_{\theta \in \Theta_{q,u}(\delta)} \mathbf{P}_\theta(\Delta = 0).$$

A benchmark value is the minimax risk of testing

$$\mathcal{R}_{q,u}(\delta) = \inf_{\Delta} \Big\{ \mathbf{P}_0(\Delta = 1) + \sup_{\theta \in \Theta_{q,u}(\delta)} \mathbf{P}_\theta(\Delta = 0) \Big\}$$

where $\inf_\Delta$ is the infimum over all $\{0,1\}$-valued statistics. The *minimax rate of testing on* $\Theta_{q,u}$ is defined as $\lambda > 0$, for which the following two facts hold:

(i) for any $\varepsilon \in (0,1)$ there exists $A_\varepsilon > 0$ such that, for all $A > A_\varepsilon$,

$$\mathcal{R}_{q,u}(A\lambda) \le \varepsilon, \tag{24}$$

(ii) for any $\varepsilon \in (0,1)$ there exists $a_\varepsilon > 0$ such that, for all $0 < A < a_\varepsilon$,

$$\mathcal{R}_{q,u}(A\lambda) \ge 1 - \varepsilon. \tag{25}$$

Note that this defines a non-asymptotic minimax rate of testing as opposed to the classical asymptotic definition that can be found, for example, in [24]. A non-asymptotic minimax study of testing for the classes $B_0(s)$ and $B_q(r)$ is given by [4] and [40]. However, those papers derive the minimax rates of testing on $\Theta_{q,u}$ only up to a logarithmic factor. The next theorem provides the exact expression for the minimax rates in the considered testing setup.

Theorem 12. *For any integers s and d satisfying $1 \le s \le d$, and any $\sigma > 0$, the minimax rate of testing on $\Theta_{0,s}$ is equal to $\lambda = (\psi_\sigma^{\sqrt{Q}}(s,d))^{1/2}$. For any $0 < q < 2$, and any $r, \sigma > 0$, the minimax rate of testing on $\Theta_{q,r}$ is equal to $\lambda = (\psi_{\sigma,q}^{\sqrt{Q}}(r,d))^{1/2}$.*

The proof of this theorem consists in establishing the upper bounds (24) and the lower bounds (25). We note first that the lower bounds (25) are essentially proved in [4] and [40]. However, in those papers they are stated in somewhat different form, so for completeness we give a brief proof in Section 7, which is very close to the proofs of the lower bounds (20) and (22). The upper bounds (24) are straightforward in view of (19) and (21). Indeed, for example, to prove (24) with $q = 0$ and $u = s$, we fix some $A > 0$ and consider the test

$$\Delta^* = \mathbb{1}_{\{\hat{N} > (A/2)(\psi_\sigma^{\sqrt{Q}}(s,d))^{1/2}\}}. \tag{26}$$

Then, writing for brevity $\psi = \psi_\sigma^{\sqrt{Q}}(s,d)$ and applying Chebyshev's inequality, we have

$$\begin{aligned} \mathcal{R}_{0,s}(A\psi) &\le \mathbf{P}_0(\Delta^* = 1) + \sup_{\theta \in \Theta_{0,s}(A\sqrt{\psi})} \mathbf{P}_\theta(\Delta^* = 0) \\ &\le \mathbf{P}_0(\hat{N} > A\sqrt{\psi}/2) + \sup_{\theta \in B_0(s)} \mathbf{P}_\theta(\hat{N} - \|\theta\|_2 \le -A\sqrt{\psi}/2) \\ &\le 2 \sup_{\theta \in B_0(s)} \frac{\mathbf{E}_\theta(\hat{N} - \|\theta\|_2)^2}{(A/2)^2 \psi} \le C_* A^{-2} \end{aligned} \tag{27}$$

for some absolute constant $C_* > 0$, where the last inequality follows from (19). Choosing A_ε as a solution of $C_* A_\varepsilon^{-2} = \varepsilon$ we obtain (24). The case $0 < q < 2$ is treated analogously by introducing the test

$$\Delta_q^* = \mathbb{1}_{\{\hat{N} > (A/2)(\psi_{\sigma,q}^{\sqrt{Q}}(r,d))^{1/2}\}}$$

and using (21) rather than (19) to get the upper bound (24).

Furthermore, as a simple corollary we obtain a non-asymptotic analog of the Ingster-Donoho-Jin theory. Consider the problem of testing the hypothesis $\mathbf{H}_0 : \theta = 0$ against the alternative $\mathbf{H}_1 : \ \theta \in \Theta_s(\delta)$ where

$$\Theta_s(\delta) = \{\theta \in \mathbb{R}^d : \ \|\theta\|_0 = s, \ \theta_j \in \{0, \delta\}, \ j = 1, \dots, d\} \tag{28}$$

for some integer $s \in [1, d]$ and some $\delta > 0$. [22] and [12] studied a slightly different but equivalent problem (with θ_j taking values 0 and δ at random) assuming in addition that $s = d^a$ for some $a \in (0, 1/2)$. In an asymptotic setting when $\sigma \to 0$ and $d = d_\sigma \to \infty$, [22] obtained the *detection boundary* in the exact minimax sense, that is the value $\lambda = \lambda_\sigma$ such that asymptotic analogs of (24) and (25) hold with $A_\varepsilon = a_\varepsilon$ and $\varepsilon = 0$. [12] proved that the detection boundary is attained at the Higher Criticism test. Extensions to the regression and classification problems and more references can be found in [23], [25], [3]. Note that the alternatives in these papers are defined not exactly in the same way as in (28).

A natural non-asymptotic analog of these results consists in establishing the minimax rate of testing on $\Theta_s(\delta)$ in the sense of the definition (24) - (25). This is done in the next corollary that covers not only $\Theta_s(\delta)$ but also the following more general class:

$$\Theta_s^*(\delta) = \Big\{\theta \in \mathbb{R}^d : \ \|\theta\|_0 = s, \ \min_{j:\ \theta_j \neq 0} |\theta_j| \geq \delta\Big\}.$$

We define the minimax rate of testing on the classes Θ_s and Θ_s^* similarly as such rate was defined for $\Theta_{q,u}$, by modifying (24) - (25) in an obvious way.

Corollary 2. *Let s and d be integers satisfying $1 \leq s \leq d$, and let $\sigma > 0$. The minimax rate of testing on Θ_s is equal $\lambda = \sigma\sqrt{\log(1 + d/s^2)}$ for $s \leq \sqrt{d}$. Furthermore, the minimax rate of testing on Θ_s^* is equal to*

$$\lambda = \begin{cases} \sigma\sqrt{\log(1 + d/s^2)} & \text{if} \quad s < \sqrt{d}, \\ \sigma d^{1/4}/\sqrt{s} & \text{if} \quad s \geq \sqrt{d}. \end{cases}$$

The proof of the upper bound in this corollary is essentially the same as in Theorem 12. We take the same test statistic Δ^* and then act as in (27) using that $\Theta_s(A\lambda)$ and $\Theta_s^*(A\lambda)$ are included in $\Theta_{0,s}(A\lambda\sqrt{s})$. The proof of the lower bound for the case $s \leq \sqrt{d}$ is also the same as in Theorem 12 since the measure μ_ρ used in the proofs (cf. Section 7) is supported on s-sparse vectors θ with all coefficients taking the same value. For $s > \sqrt{d}$ we need a slightly different lower bound argument - see Section 7 for the details.

[22] and [12] derived the asymptotic rate of testing in the form $\lambda = c(a)\sigma\sqrt{\log d}$ where the exact value $c(a) > 0$ is explicitly given as a function of a appearing in the relation $s = d^a$, $0 < a < 1/2$. Corollary 2 allows us to explore more general behavior of s leading to other types of rates. For example, we find that the minimax rate of testing is of the order σ if $s = \sqrt{d}$ and it is of the order $\sigma\sqrt{\log\log d}$ if $s \asymp \sqrt{d}/(\log d)^\gamma$ for any $\gamma > 0$. Such effects are not captured by the previous asymptotic results. Note also that the test Δ^* (cf. (26))

that achieves the minimax rates in Corollary 2 is very simple - it is a plug-in test based on the estimator of the ℓ_2-norm. We do not need to invoke refined techniques as the Higher Criticism test. However, we do not prove that our method achieves the exact constant $c(a)$ in the specific regime considered by [22] and [12].

7 Proofs of the lower bounds

7.1 General tools

The proofs of the lower bounds in this section use a technique based on a reduction to testing between two probability measures, one of which is a mixture measure. This is a special case of what is called the method of fuzzy hypotheses or Le Cam's method since Le Cam [34] was apparently the first to consider this kind of argument.

Let μ be a probability measure on Θ. Denote by $\mathbb{P}_\mu$ the mixture probability measure

$$\mathbb{P}_\mu = \int_\Theta \mathbf{P}_\theta \, \mu(d\theta).$$

A vector $\theta \in \mathbb{R}^d$ is called s-sparse if $\|\theta\|_0 = s$. For an integer s such that $1 \le s \le d$ and $\rho > 0$, we denote by μ_ρ the uniform distribution on the set of s-sparse vectors in $\mathbb{R}^d$ with all nonzero coefficients equal to $\sigma\rho$. Let

$$\chi^2(P', P) = \int (dP'/dP)^2 dP - 1$$

be the chi-square divergence between two mutually absolutely continuous probability measures P' and P.

The following lemma is obtained by combining arguments of [4] and [9].

Lemma 1. *For all $\sigma > 0, \rho > 0$, $1 \le s \le d$, we have*

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

For completeness, the proof of this lemma is given in the Appendix. We will also need a second lemma, which is a special case of Theorem 2.15 in [39]:

Lemma 2. *Let Θ be a subset of $\mathbb{R}^d$ containing 0. Assume that there exists a probability measure μ on Θ and numbers $v > 0, \beta > 0$ such that $T(\theta) = 2v$ for all $\theta \in \mathrm{supp}(\mu)$ and $\chi^2(\mathbb{P}_\mu, \mathbf{P}_0) \le \beta$, Then*

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbf{P}_\theta\big(|\hat{T} - T(\theta)| \ge v\big) \ge \frac{1}{4} \exp(-\beta),$$

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

7.2 Proof of the lower bound (4) in Theorem 1

Set $\rho = \sqrt{\log(1 + d/s^2)}$. Then, by Lemma 1,

$$\chi^2(\mathbb{P}_{\mu_\rho}, \mathbf{P}_0) \le \left(1 - \frac{s}{d} + \frac{s}{d}\left(1 + \frac{d}{s^2}\right)\right)^s - 1 = \left(1 + \frac{1}{s}\right)^s - 1 \le e - 1. \tag{29}$$

Next, $L(\theta) = \sigma s \rho$ for all $\theta \in \mathrm{supp}(\mu_\rho)$, and also $\mathrm{supp}(\mu_\rho) \subseteq B_0(s)$. Thus, the assumptions of Lemma 2 are satisfied with $\Theta = B_0(s)$, $\beta = e - 1$, $v = \sigma s \rho/2 = (1/2)\sigma s \sqrt{\log(1 + d/s^2)}$ and $T(\theta) = L(\theta)$. An application of Lemma 2 yields

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{P}_\theta \left(|\hat{T} - L(\theta)| \geq (1/2)\sigma s \sqrt{\log(1 + d/s^2)} \right) \geq \frac{1}{4} \exp(1 - e),$$

which implies (4).

7.3 Proof of Theorem 4

We start by rewriting in a more convenient form the lower rates we need to prove. For this, consider separately the cases $s \geq \sqrt{d}$ and $s < \sqrt{d}$.

Case $s \geq \sqrt{d}$. The lower rate we need to prove in this case is $\min\{\kappa^4, \max(\sigma^2\kappa^2, \sigma^4 d)\}$. It is easy to check that we can write it as follows:

$$\min\{\kappa^4, \max(\sigma^2\kappa^2, \sigma^4 d)\} = \begin{cases} \sigma^2\kappa^2 & \text{if} \quad \kappa^4 > \sigma^4 d^2, \\ \sigma^4 d & \text{if} \quad \sigma^4 d < \kappa^4 \leq \sigma^4 d^2, \\ \kappa^4 & \text{if} \quad \kappa^4 \leq \sigma^4 d. \end{cases} \tag{30}$$

Note that the lower rate $\sigma^4 d$ for $\sigma^4 d < \kappa^4 \leq \sigma^4 d^2$ follows from the lower rate κ^4 for $\kappa^4 < \sigma^4 d$ and the fact that the minimax risk is a non-decreasing function of κ. Therefore, to prove Theorem 4 for $s \geq \sqrt{d}$, it is enough to show that $R^*_Q(B_2(\kappa) \cap B_0(s)) \geq c(\text{lower rate})$, where $c > 0$ is an absolute constant, and

$$\text{lower rate} = \begin{cases} \sigma^2\kappa^2 & \text{if} \quad \kappa^4 > \sigma^4 d^2 \quad \text{and } s = \sqrt{d}, \\ \kappa^4 & \text{if} \quad \kappa^4 \leq \sigma^4 d \quad \text{and } s = \sqrt{d}. \end{cases} \tag{31}$$

In (31), we assume without loss of generality that $\sqrt{d}$ is an integer and we replace without loss of generality the condition $s \geq \sqrt{d}$ by $s = \sqrt{d}$ since the minimax risk is a non-decreasing function of s.

Case $s < \sqrt{d}$. The lower rate we need to prove in this case is

$$\min\{\kappa^4, \max(\sigma^2\kappa^2, \sigma^4 s^2 \log^2(1 + d/s^2))\}.$$

The same argument as above shows that the analog of representation (30) holds with d replaced by $s^2 \log^2(1 + d/s^2)$, and that it is enough to prove the lower rate of the form:

$$\text{lower rate} = \begin{cases} \sigma^2\kappa^2 & \text{if} \quad \kappa^4 > \sigma^4 s^4 \log^4(1 + d/s^2) \quad \text{and } s < \sqrt{d}, \\ \kappa^4 & \text{if} \quad \kappa^4 \leq \sigma^4 s^2 \log^2(1 + d/s^2) \quad \text{and } s < \sqrt{d}. \end{cases} \tag{32}$$

Thus, to prove Theorem 4 it remains to establish (31) and (32). This is done in the following two propositions. Proposition 1 is used with $b = \log 2$ and it is a more general fact than the first lines in (31) and (32) since $B_2(\kappa) \cap B_0(s) \supseteq B_2(\kappa) \cap B_0(1)$, and $s \log(1 + d/s^2) \geq \log 2$ for $1 \leq s \leq \sqrt{d}$. Proposition 2 is applied with $b = 1/(\log 2)$.

Proposition 1. *Let $b > 0$. If $\kappa > b\sigma$, then*

$$\inf_{\hat{T}} \sup_{\theta \in B_2(\kappa) \cap B_0(1)} \mathbf{P}_\theta \left(|\hat{T} - Q(\theta)| \geq (3b/8)\sigma\kappa \right) \geq \frac{1}{4} \exp(-b^2/4),$$

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

Proposition 2. *Let $b > 0$. If $\kappa^4 \leq b^2\sigma^4 s^2 \log^2(1 + d/s^2)$ and $1 \leq s \leq d$, then*

$$\inf_{\hat{T}} \sup_{\theta \in B_2(\kappa) \cap B_0(s)} \mathbf{P}_\theta \left(|\hat{T} - Q(\theta)| \geq \kappa^2/(2\max(b,1)) \right) \geq \frac{1}{4}\exp(1-e),$$

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

Remark 3. At first sight, the proof of the lower bound seems to exhibit a paradox: the proof for the rate $\sigma^2\kappa^2$ involves a two-point comparison, while the trivial rate κ^4 needs a more elaborate proof. But, in fact it is not surprising since the rate $\sigma^2\kappa^2$ is independent from the dimension d, so that it is natural that the proof only uses simple arguments that also hold for $d = 1$. On the other hand, the bound κ^4 needs a construction based on multiple hypotheses, since the dimension-dependent rate $\sigma^4 d$ derives from it in view of the above argument.

7.4 Proof of Proposition 1

Consider the vectors $\theta = (\kappa, 0, \ldots, 0)$ and $\theta' = (\kappa - b\sigma/2, 0, \ldots, 0)$. Clearly, θ and θ' belong to $B_2(\kappa) \cap B_0(1)$. We have

$$d(\theta, \theta') \triangleq \left| Q(\theta) - Q(\theta') \right| = |\sigma^2 b^2/4 - \kappa\sigma b| > 3\sigma\kappa b/4,$$

and the Kullback-Leibler divergence between $\mathbf{P}_\theta$ and $\mathbf{P}_{\theta'}$ satisfies

$$K(\mathbf{P}_\theta, \mathbf{P}_{\theta'}) = \frac{\|\theta - \theta'\|_2^2}{2\sigma^2} = \frac{b^2}{8}.$$

We now apply Theorem 2.2 and (2.9) in [39] to obtain the result.

7.5 Proof of Proposition 2

Set $\rho = \kappa/(\sigma\sqrt{\max(b,1)s})$. Then $\rho^2 \leq \log(1+d/s^2)$ and due to (29) we have $\chi^2(\mathbb{P}_{\mu_\rho}, \mathbf{P}_0) \leq e-1$. Next, $Q(\theta) = \|\theta\|_2^2 = s\sigma^2\rho^2 = \kappa^2/\max(b,1)$ for all $\theta \in \mathrm{supp}(\mu_\rho)$, which implies $\mathrm{supp}(\mu_\rho) \subseteq B_2(\kappa)$. We also have $\mathrm{supp}(\mu_\rho) \subseteq B_0(s)$ by construction. Therefore, the assumptions of Lemma 2 are satisfied with $\Theta = B_2(\kappa) \cap B_0(s)$, $\beta = e - 1$, $v = \kappa^2/(2\max(b,1))$ and $T(\theta) = Q(\theta)$. An application of Lemma 2 yields the result.

7.6 Proof of Theorem 2

In order to prove Theorem 2, we will need the following proposition.

Proposition 3. *Let $b > 0$. If $\kappa^2 \leq b^2\sigma^2 s^2 \log(1 + d/s^2)$ and $1 \leq s \leq d$, then*

$$\inf_{\hat{T}} \sup_{\theta \in B_1(\kappa) \cap B_0(s)} \mathbf{P}_\theta \left(|\hat{T} - L(\theta)| \geq \kappa/(2\max(b,1)) \right) \geq \frac{1}{4}\exp(1-e),$$

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

Proof. We proceed as in the proof of Proposition 2 with the following modifications. We now set $\rho = \kappa/(\max(b,1)\sigma s)$. Then $\chi^2(\mathbb{P}_{\mu_\rho}, \mathbf{P}_0) \leq e - 1$ and $L(\theta) = \|\theta\|_1 = s\sigma\rho = \kappa/\max(b,1)$ for all $\theta \in \mathrm{supp}(\mu_\rho)$, so that $\mathrm{supp}(\mu_\rho) \subseteq \Theta = B_1(\kappa) \cap B_0(s)$ and Lemma 2 applies with $\beta = e - 1$, $v = \kappa/(2\max(b,1))$ and $T(\theta) = L(\theta)$. □

Proof of Theorem 2. First notice that, for an integer $s \in [1, d]$, and $0 < q < 1$, $\kappa > 0$,

$$B_1(\kappa) \cap B_0(s) \subset B_q(r) \quad \text{if} \quad s^{1-q}\kappa^q = r^q. \tag{33}$$

We will prove the theorem by considering separately the cases $m = 0$ and $m \geq 1$.

Case $m = 0$. Then, $r^2 < \sigma^2 \log(1 + d)$ and the assumption of Proposition 3 is satisfied with $s = 1$, $b = 1$, and $\kappa = r$. Applying Proposition 3 with these parameters and using (33) with $s = 1$ we easily deduce that $R_L^*(B_q(r)) \geq Cr^2$.

Case $m \geq 1$. We now use the embedding (33) with $s = m$. Then

$$\kappa = rm^{1-1/q} \geq \sigma m \sqrt{\log(1 + d/m^2)} \tag{34}$$

where the last inequality follows from the definition of m. Furthermore, the fact that $m \geq 1$ and the definition of m imply

$$2^{-2/q} r^2 m^{-2/q} \leq r^2 (m + 1)^{-2/q} < \sigma^2 \log(1 + d/(m + 1)^2) \leq \sigma^2 \log(1 + d/m^2). \tag{35}$$

This proves that for κ defined in (34) we have $\kappa^2 \leq 2^{2/q}\sigma^2 m^2 \log(1 + d/m^2)$. Thus, the assumption of Proposition 3 is satisfied with $s = m$, $b = 2^{1/q}$ and κ defined in (34). Applying Proposition 3 with these parameters and using (33) with $s = m$ we deduce that $R_L^*(B_q(r)) \geq C\kappa^2$. This and (34) yield $R_L^*(B_q(r)) \geq C\sigma^2 m^2 \log(1 + d/m^2)$, which is the desired lower bound.

7.7 Proof of Theorem 6

First notice that, for an integer $s \in [1, d]$, and $0 < q < 2$, $\kappa > 0$,

$$B_2(\kappa) \cap B_0(s) \subset B_q(r) \quad \text{if} \quad s^{1-q/2}\kappa^q = r^q. \tag{36}$$

Consider separately the cases $m = 0$, $1 \leq m \leq \sqrt{d}$, and $m > \sqrt{d}$.

Case $m = 0$. Then, $r^2 < \sigma^2 \log(1+d)$ so that the assumption of Proposition 2 is satisfied with $s = 1$, $b = 1$, and $\kappa = r$. Applying Proposition 2 with these parameters and using (36) with $s = 1$ and $\kappa = r$ we get that $R_Q^*(B_q(r)) \geq Cr^4$.

Case $1 \leq m \leq \sqrt{d}$. We start by using (36) with $s = m$. Then

$$\kappa = rm^{1/2-1/q} \geq \sigma \sqrt{m \log(1 + d/m^2)} \tag{37}$$

where the last inequality follows from the definition of m. For this κ, using (35) we obtain $\kappa^2 \leq 2^{2/q}\sigma^2 m \log(1 + d/m^2)$. Thus, the assumption of Proposition 2 is satisfied with $s = m$, $b = 2^{2/q}$ and κ defined in (37). Applying Proposition 2 with these parameters and using (36) with $s = m$ we deduce that $R_Q^*(B_q(r)) \geq C\kappa^4$. This and (37) prove the lower bound $R_Q^*(B_q(r)) \geq C\sigma^4 m^2 \log^2(1 + d/m^2)$.

To show that $R_Q^*(B_q(r)) \geq C\sigma^2 r^2$, we use (36) with $s = 1$ and $\kappa = r$. Now, $m \geq 1$, which implies $r^2 \geq \sigma^2 \log(1+d) \geq \sigma^2(\log 2)$. Thus, the assumption of Proposition 1 is satisfied with $s = 1$, $\kappa = r$, and any $0 < b < \sqrt{\log 2}$, leading to the bound $R_Q^*(B_2(\kappa) \cap B_0(1)) \geq C\sigma^2 r^2$. This inequality and the embedding in (36) with $s = 1$ yield the result.

Case $m > \sqrt{d}$. It suffices to note that the argument used above in the case $1 \leq m \leq \sqrt{d}$ remains valid for $m > \sqrt{d}$ and $s = \sqrt{d}$ instead of $s = m$ (assuming without loss of generality that $\sqrt{d}$ is an integer).

7.8 Proof of the lower bound (20) in Theorem 8

Let $s < \sqrt{d}$. Set $\rho = \sqrt{\log(1+d/s^2)}$. Due to (29) we have $\chi^2(\mathbb{P}_{\mu_\rho}, \mathbf{P}_0) \le e-1$. Next, $\|\theta\|_2 = \sigma\rho\sqrt{s} = \sigma\sqrt{s\log(1+d/s^2)}$ for all $\theta \in \mathrm{supp}(\mu_\rho)$, and $\mathrm{supp}(\mu_\rho) \subseteq B_0(s)$ by construction. Therefore, the assumptions of Lemma 2 are satisfied with $\Theta = B_0(s)$, $\beta = e-1$, $v = \sigma\sqrt{s\log(1+d/s^2)}/2$ and $T(\theta) = \|\theta\|_2$. An application of Lemma 2 yields the result for $s < \sqrt{d}$. To obtain the lower bound for $s \ge \sqrt{d}$, it suffices to consider the case $s = \sqrt{d}$ (assuming without loss of generality that $\sqrt{d}$ is an integer) and to repeat the above argument with this value of s.

7.9 Proof of the lower bound (22) in Theorem 9

If $m = 0$ we have $r^2 < \sigma^2\log(1+d)$. In this case, set $\rho = r/\sigma$, $s = 1$. Then, $\rho < \sqrt{\log(1+d)}$ and due to (29) with $s = 1$ we have $\chi^2(\mathbb{P}_{\mu_\rho}, \mathbf{P}_0) \le 1$. Next, $\|\theta\|_2 = \|\theta\|_q = r$ for all $\theta \in \mathrm{supp}(\mu_\rho)$. Thus, $\mathrm{supp}(\mu_\rho) \subseteq B_q(r)$ and the assumptions of Lemma 2 are satisfied with $\Theta = B_q(r)$, $\beta = 1$, $v = r/2$ and $T(\theta) = \|\theta\|_2$, which implies the bound $R^*_{\sqrt{Q}}(B_q(r)) \ge Cr^2$ for $m = 0$.

Case $1 \le m \le \sqrt{d}$. Use the same construction as in the proof of (20) replacing there s with m. Then, $\|\theta\|_2 = \sigma\sqrt{m\log(1+d/m^2)}$, and $\|\theta\|_q = \sigma\rho m^{1/q} = \sigma m^{1/q}\sqrt{\log(1+d/m^2)}$ for all $\theta \in \mathrm{supp}(\mu_\rho)$. By definition of m, we have $\sigma m^{1/q}\sqrt{\log(1+d/m^2)} \le r$ guaranteeing that $\mathrm{supp}(\mu_\rho) \subseteq B_q(r)$. Other elements of the argument remain as in the proof of (20).

Case $m > \sqrt{d}$. Use the same construction as in the proof of (20) with $s = \sqrt{d}$ (assuming without loss of generality that $\sqrt{d}$ is an integer). Then $\rho = \sqrt{\log 2}$, $\|\theta\|_2 = \sigma d^{1/4}\sqrt{\log 2}$, and $\|\theta\|_q = \sigma d^{1/(2q)}\sqrt{\log 2} \le r$ (by definition of m) for all $\theta \in \mathrm{supp}(\mu_\rho)$. Other elements of the argument remain as in the proof of (20).

7.10 Proof of the lower bounds in Theorem 12 and in Corollary 2

The following lemma reduces the proof to the argument, which is very close to that of the previous two proofs.

Lemma 3. *If μ is a probability measure on Θ, then*

$$\inf_{\Delta}\Big\{\mathbf{P}_0(\Delta = 1) + \sup_{\theta\in\Theta}\mathbf{P}_\theta(\Delta = 0)\Big\} \ge 1 - \sqrt{\chi^2(\mathbb{P}_\mu, \mathbf{P}_0)}$$

where $\inf_\Delta$ is the infimum over all $\{0,1\}$-valued statistics.

Proof. For any $\{0,1\}$-valued statistic Δ,

$$\begin{aligned}\mathbf{P}_0(\Delta = 1) + \sup_{\theta\in\Theta}\mathbf{P}_\theta(\Delta = 0) &\ge \mathbf{P}_0(\Delta = 1) + \int_\Theta \mathbf{P}_\theta(\Delta = 0)\mu(d\theta)\\ &= \mathbf{P}_0(\Delta = 1) + \mathbb{P}_\mu(\Delta = 0) \ge 1 - V(\mathbb{P}_\mu, \mathbf{P}_0) \ge 1 - \sqrt{\chi^2(\mathbb{P}_\mu, \mathbf{P}_0)}\end{aligned}$$

where $V(\cdot,\cdot)$ denotes the total variation distance and the last two inequalities follow from the standard properties of this distance (cf. Theorem 2.2(i) and (2.27) in [39]). □

Proof of the lower bound in Theorem 12 for $q = 0$. We use a slightly modified argument of Subsection 7.8. As in Subsection 7.8, it suffices to prove the result in the case $s < \sqrt{d}$. Then,

$\psi_\sigma^{\sqrt{Q}}(s,d) = \sigma^2 s \log(1+d/s^2)$, so that our aim is to show that the lower rate of testing on $B_0(s)$ is $\lambda = \sigma\sqrt{s\log(1+d/s^2)}$. Fix $A \in (0,1)$. We use Lemma 3 with $\Theta = \Theta_{0,s}(A\lambda)$ and $\mu = \mu_\rho$ where we take $\rho = A\sqrt{\log(1+d/s^2)}$. For all $\theta \in \mathrm{supp}(\mu_\rho)$ we have $\|\theta\|_2 = \sigma\rho\sqrt{s} = A\lambda$ while $\mathrm{supp}(\mu_\rho) \subseteq B_0(s)$ by construction. Hence $\mathrm{supp}(\mu_\rho) \subseteq \Theta_{0,s}(A\lambda)$, so that we can apply Lemma 3. Next, by Lemma 1,

$$\chi^2(\mathbb{P}_{\mu_\rho}, \mathbf{P}_0) \le \left(1 - \frac{s}{d} + \frac{s}{d}\left(1+\frac{d}{s^2}\right)^{A^2}\right)^s - 1 \le \left(1+\frac{A^2}{s}\right)^s - 1 \le \exp(A^2) - 1 \tag{38}$$

where we have used that $(1+x)^{A^2} - 1 \le A^2 x$ for $0 < A < 1$, $x > 0$. The last display and Lemma 3 imply that $\mathcal{R}_{0,s}(A\lambda) \ge 1 - \sqrt{\exp(A^2)-1}$. Choosing a_ε such that $\sqrt{\exp(a_\varepsilon^2)-1} = \varepsilon$ proves (25).

Proof of the lower bound in Theorem 12 for $0 < q < 2$ follows along similar lines but now we modify, in the same spirit, the argument of Subsection 7.9 rather than that of Subsection 7.8. The corresponding ρ in Subsection 7.9 is multiplied by a suitable $A \in (0,1)$ and then Lemma 3 is applied. We omit the details.

Proof of the lower bound in Corollary 2. As explained after the statement of Corollary 2, we need only to consider the case $s > \sqrt{d}$ for the class Θ_s^*. Then, $\lambda = \sigma d^{1/4}/\sqrt{s}$. Instead of μ_ρ we consider now a slightly different measure $\bar\mu_\rho$, which is the uniform distribution on the set of s-sparse vectors in $\mathbb{R}^d$ with nonzero coefficients taking values in $\{-\sigma\rho, \sigma\rho\}$. Then, similarly to Lemma 1,

$$\chi^2(\mathbb{P}_{\bar\mu_\rho}, \mathbf{P}_0) \le \left(1 - \frac{s}{d} + \frac{s}{d}\cosh(\rho^2)\right)^s - 1, \tag{39}$$

cf. formula (27) in [4]. Fix $A \in (0,1)$. We now use Lemma 3 with $\Theta = \Theta_s^*(A\lambda)$ and $\mu = \bar\mu_\rho$ where we take $\rho = Ad^{1/4}/\sqrt{s}$. For all $\theta \in \mathrm{supp}(\bar\mu_\rho)$ we have $|\theta_j| = \sigma\rho = A\sigma d^{1/4}/\sqrt{s} = A\lambda$ and also $\mathrm{supp}(\bar\mu_\rho) \subseteq \{\|\theta\|_0 = s\}$ by construction. Hence $\mathrm{supp}(\bar\mu_\rho) \subseteq \Theta_s^*(A\lambda)$, so that we can apply Lemma 3. Since $s > \sqrt{d}$ we have $\rho < 1$. Using (39) and the fact that $\cosh(x) \le 1 + x^2$ for $0 < x < 1$ we obtain

$$\chi^2(\mathbb{P}_{\bar\mu_\rho}, \mathbf{P}_0) \le \left(1 + \frac{s\rho^4}{d}\right)^s - 1 \le \exp(A^4) - 1$$

and we conclude the proof in the same way as it is done after (38).

8 Proofs of the upper bounds

We will use the following lemma.

Lemma 4. *For $X \sim \mathcal{N}(0,1)$ and any $x > 0$ we have*

$$\frac{4}{\sqrt{2\pi}(x+\sqrt{x^2+4})} e^{-x^2/2} \le \mathbf{P}\big(|X| > x\big) \le \frac{4}{\sqrt{2\pi}(x+\sqrt{x^2+2})} e^{-x^2/2}, \tag{40}$$

$$\mathbf{E}\Big[X^2 \mathbb{1}_{\{|X|>x\}}\Big] \le \sqrt{\frac{2}{\pi}}\Big(x + \frac{2}{x}\Big) e^{-x^2/2}, \tag{41}$$

$$\mathbf{E}\Big[X^4 \mathbb{1}_{\{|X|>x\}}\Big] \le \sqrt{\frac{2}{\pi}}\Big(x^3 + 3x + \frac{1}{x}\Big) e^{-x^2/2}. \tag{42}$$

Inequality (40) is due to [6] and [38]. Inequalities (41) and (42) follow from integration by parts.

In this section, we will use the notation

$$x = \sqrt{2\log(1+d/s^2)}, \quad \hat{S} = \{j : |y_j| > \sigma x\}, \quad S = \{j : \theta_j \neq 0\}. \tag{43}$$

We also recall that the observations are of the form $y_j = \theta_i + \sigma\xi_j, \quad j = 1, \dots, d$, with i.i.d. errors $\xi_j \sim \mathcal{N}(0,1)$. We will denote by $C_i, i = 1, 2, \dots$, absolute positive constants, and by C absolute positive constants that can vary from line to line.

8.1 Proof of the bound (3) in Theorem 1

Clearly, $\mathbf{E}_\theta(\sum_{j=1}^d y_j - L(\theta))^2 = \sigma^2 d$. Thus, in view of (5), to prove (3) it is enough to show that for $s \le \sqrt{d}$ we have

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta(\hat{L}_* - L(\theta))^2 \le C\sigma^2 s^2 \log(1 + d/s^2) \tag{44}$$

where

$$\hat{L}_* = \sum_{j=1}^d y_j \, \mathbb{1}_{\{|y_j| > \sigma\sqrt{2\log(1+d/s^2)}\}}$$

and $C > 0$ is an absolute constant. Recalling the notation set in (43) we have

$$\hat{L}_* - L(\theta) = \sum_{j \in S}(y_j - \theta_j) - \sum_{j \in S \setminus \hat{S}} y_j + \sum_{j \in \hat{S} \setminus S} y_j. \tag{45}$$

Thus, for $\theta \in B_0(s)$, we obtain

$$\begin{aligned}
\mathbf{E}_\theta(\hat{L}_* - L(\theta))^2 &\le 3\,\mathbf{E}\Big(\sum_{j \in S} \sigma\xi_j\Big)^2 + 3\,\mathbf{E}_\theta\Big(\sum_{j \in S} y_j \, \mathbb{1}_{\{|y_j| \le \sigma x\}}\Big)^2 + 3\,\mathbf{E}\Big(\sum_{j \in S^c} \sigma\xi_j \, \mathbb{1}_{\{|\xi_j| > x\}}\Big)^2 \\
&\le 3\sigma^2\Big\{(s + s^2x^2) + \sum_{j \in S^c} \mathbf{E}\Big(\xi_j^2 \, \mathbb{1}_{\{|\xi_j| > x\}}\Big)\Big\} \\
&\le 3\sigma^2\Big\{(s + s^2x^2) + d\sqrt{\frac{2}{\pi}}\Big(x + \frac{2}{x}\Big)e^{-x^2/2}\Big\} \quad \text{(by (41))} \\
&\le 3\sigma^2\Big\{(s + s^2x^2) + s^2\sqrt{\frac{2}{\pi}}\Big(x + \frac{2}{x}\Big)\Big\},
\end{aligned}$$

and (44) follows since $x \ge \sqrt{2\log 2}$ for $s \le \sqrt{d}$.

8.2 Proof of Theorem 3

We will consider only the sparse zone $1 \le m \le \sqrt{d}$ since the cases $m = 0$ and $m > \sqrt{d}$ are trivial. Fix $\theta \in B_q(r)$. We will use the notation

$$\tilde{d} = 1 + d/m^2, \quad \tilde{x} = 2\sqrt{2\log\tilde{d}}, \quad \tilde{S} = \{j : |\theta_j| > \sigma\tilde{x}/2\}.$$

Note that

$$\mathrm{Card}(\tilde{S}) \le \left(\frac{2r}{\sigma\tilde{x}}\right)^q < 2^{-q/2}(m+1) \le 2^{1-q/2}m, \tag{46}$$

where the first inequality is due to the fact that $\theta \in B_q(r)$ and the second follows from the definition of m.

Consider first the bias of $\hat{L}_q$. Lemma 5 yields

$$\begin{aligned}\big(\mathbf{E}_\theta(\hat{L}_q) - L(\theta)\big)^2 &\le C\Big(\sum_{j=1}^d \min(|\theta_j|, \sigma\tilde{x})\Big)^2 \le C\Big(\sum_{j=1}^d |\theta_j|^q(\sigma\tilde{x})^{1-q}\Big)^2 \\ &\le C\left(\frac{r}{\sigma\tilde{x}}\right)^{2q}\sigma^2\log\tilde{d} \\ &\le C\sigma^2 m^2\log\tilde{d},\end{aligned} \tag{47}$$

where we have used (46). Next, the variance of $\hat{L}_q$ has the form

$$\mathrm{Var}_\theta(\hat{L}_q) = \sum_{j=1}^d \mathrm{Var}_\theta(y_j\ \mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}).$$

Here, for indices j belonging to $\tilde{S}$, using (46) we have

$$\begin{aligned}\sum_{j\in\tilde{S}} \mathrm{Var}_\theta(y_j\ \mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}) &\le 2\sum_{j\in\tilde{S}} \mathrm{Var}_\theta(y_j) + 2\sum_{j\in\tilde{S}} \mathrm{Var}_\theta(y_j\ \mathbb{1}_{\{|y_j|\le\sigma\tilde{x}\}}) \\ &\le 2\mathrm{Card}(\tilde{S})\sigma^2(1+\tilde{x}^2) \\ &\le C\sigma^2 m\log\tilde{d}.\end{aligned} \tag{48}$$

For indices j belonging to $\tilde{S}^c$, we have

$$\begin{aligned}\sum_{j\in\tilde{S}^c} \mathrm{Var}_\theta(y_j\ \mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}) &\le \sum_{j\in\tilde{S}^c} \mathbf{E}_\theta(y_j^2\ \mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}) \\ &\le 2\sum_{j\in\tilde{S}^c}\theta_j^2 + 2\sigma^2\sum_{j\in\tilde{S}^c}\mathbf{E}_\theta(\xi_j^2\ \mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}) \\ &\le 2\Big(\sum_{j\in\tilde{S}^c}|\theta_j|\Big)^2 + 2\sigma^2\sum_{j\in\tilde{S}^c}\mathbf{E}(\xi_j^2\ \mathbb{1}_{\{|\xi_j|>\sqrt{2\log\tilde{d}}\}}).\end{aligned} \tag{49}$$

Using the same argument as in (47) we find

$$\Big(\sum_{j\in\tilde{S}^c}|\theta_j|\Big)^2 \le C\Big(\sum_{j=1}^d \min(|\theta_j|, \sigma\tilde{x})\Big)^2 \le C\sigma^2 m^2\log\tilde{d}. \tag{50}$$

Finally, (41) implies

$$\sigma^2\sum_{j\in\tilde{S}^c}\mathbf{E}(\xi_j^2\ \mathbb{1}_{\{|\xi_j|>\sqrt{2\log\tilde{d}}\}}) \le C\sigma^2(d/\tilde{d})\sqrt{\log\tilde{d}} \le C\sigma^2 m^2\log\tilde{d} \tag{51}$$

where for the last inequality we have used that $\log \tilde{d} \geq \log 2$ for $m \leq \sqrt{d}$. Combining (48) – (51) we obtain that

$$\mathrm{Var}_\theta(\hat{L}_q) \leq C\sigma^2 m^2 \log \tilde{d}.$$

Together with (41), this yields the desired result:

$$\sup_{\theta \in B_q(r)} \mathbf{E}_\theta(\hat{L}_q - L(\theta))^2 \leq C\sigma^2 m^2 \log \tilde{d}.$$

8.3 Proof of Theorem 5

We will use the notation set in (43). Moreover, we recall the definition of the estimator studied here:

$$\hat{Q}_* = \begin{cases} \sum_{j=1}^d y_j^2 - d\sigma^2 & \text{if} \quad \kappa^4 \geq \max\{\sigma^2\kappa^2, \sigma^4 d\}, \\ 0 & \text{if} \quad \kappa^4 < \max\{\sigma^2\kappa^2, \sigma^4 d\}. \end{cases}$$

The upper bound κ^4 for $\kappa^4 < \psi_\sigma(s,d,\kappa)$ is trivial since the risk of the zero estimator is equal to κ^4. Let now $\kappa^4 \geq \psi_\sigma(s,d,\kappa)$. We analyze separately the cases $s \geq \sqrt{d}$, $\kappa^4 \geq \psi_\sigma(s,d,\kappa)$, and $s < \sqrt{d}$, $\kappa^4 \geq \psi_\sigma(s,d,\kappa)$.

Case $s \geq \sqrt{d}$ and $\kappa^4 \geq \psi_\sigma(s,d,\kappa)$. Then, $\hat{Q} = \hat{Q}_*$ and Theorem 5 claims a bound with the rate $\psi_\sigma^Q(s,d,\kappa) = \psi_\sigma(s,d,\kappa) = \max(\sigma^2\kappa^2, \sigma^4 d)$. To prove this bound, note that

$$\hat{Q}_* - Q(\theta) = 2\sigma \sum_{j=1}^d \theta_j \xi_j + \sigma^2 \sum_{j=1}^d (\xi_j^2 - 1).$$

Thus, for all $\theta \in B_2(\kappa)$,

$$\begin{aligned} \mathbf{E}_\theta(\hat{Q}_* - Q(\theta))^2 &= 4\sigma^2 \mathbf{E}\Big(\sum_{j=1}^d \theta_j \xi_j\Big)^2 + \sigma^4 \mathbf{E}\Big(\sum_{j=1}^d (\xi_j^2 - 1)\Big)^2 \\ &= 4\sigma^2 \|\theta\|_2^2 + 2\sigma^4 d \leq 6 \max(\sigma^2\kappa^2, \sigma^4 d). \end{aligned} \tag{52}$$

Case $s < \sqrt{d}$ and $\kappa^4 \geq \psi_\sigma(s,d,\kappa)$. Then, $\hat{Q} = \hat{Q}'$ where

$$\hat{Q}' = \sum_{j=1}^d (y_j^2 - \alpha\sigma^2)\, \mathbb{1}_{\{|y_j| > \sigma\sqrt{2\log(1+d/s^2)}\}}$$

and $\psi_\sigma^Q(s,d,\kappa) = \max(\sigma^2\kappa^2, \sigma^4 s^2 \log^2(1+d/s^2))$. Here and below in this proof, we set for brevity $\alpha = \alpha_s$.

Let x be defined in (43). Since $s < \sqrt{d}$, we have $x \geq \sqrt{2\log 2}$. Using Lemma 4, we find that, for $s \leq \sqrt{d}$,

$$\alpha = \frac{\mathbf{E}\Big(X^2 \mathbb{1}_{\{|X|>x\}}\Big)}{\mathbf{P}\big(|X| > x\big)} \leq (x + 2/x)(x+1) \leq 5x^2 = 10\, \log(1 + d/s^2). \tag{53}$$

Similarly to (45), we get

$$\hat{Q}' - Q(\theta) = \sum_{j \in S} (y_j^2 - \alpha\sigma^2 - \theta_j^2) - \sum_{j \in S \setminus \hat{S}} (y_j^2 - \alpha\sigma^2) + \sum_{j \in \hat{S} \setminus S} (y_j^2 - \alpha\sigma^2),$$

where S and $\hat{S}$ are defined in (43). Thus,

(54)

$$\mathbf{E}_\theta\big(\hat{Q}' - Q(\theta)\big)^2 \le 3\,\mathbf{E}_\theta\Big[\Big(\sum_{j\in S}(y_j^2 - \alpha\sigma^2 - \theta_j^2)\Big)^2 + \Big(\sum_{j\in S\setminus\hat{S}}(y_j^2 - \alpha\sigma^2)\Big)^2 + \Big(\sum_{j\in \hat{S}\setminus S}(y_j^2 - \alpha\sigma^2)\Big)^2\Big].$$

For $\theta \in B_2(\kappa)\cap B_0(s)$, the first term on the right-hand side satisfies

$$\begin{aligned}\mathbf{E}_\theta\Big(\sum_{j\in S}(y_j^2 - \alpha\sigma^2 - \theta_j^2)\Big)^2 &= \mathbf{E}\Big(\sum_{j\in S}(2\sigma\theta_j\xi_j + \sigma^2(\xi_j^2 - \alpha))\Big)^2 \\ &\le 4\sigma^2\|\theta\|_2^2 + 2\sigma^4 s^2(\alpha^2+3) \\ &\le 4\sigma^2\|\theta\|_2^2 + 2\sigma^4 s^2(25x^4+3),\end{aligned} \tag{55}$$

where the last inequality derives from (53). Hence, using the definition of x in (43) we find

$$\begin{aligned}\mathbf{E}_\theta\Big(\sum_{j\in S}(y_j^2 - \alpha\sigma^2 - \theta_j^2)\Big)^2 &\le C_1\,\big(\sigma^2\|\theta\|_2^2 + \sigma^4 s^2\log^2(1+d/s^2)\big) \\ &\le C_1\,\big(\sigma^2\kappa^2 + \sigma^4 s^2\log^2(1+d/s^2)\big).\end{aligned} \tag{56}$$

Furthermore, by definition of $\hat{S}$,

$$\begin{aligned}\mathbf{E}_\theta\Big(\sum_{j\in S\setminus\hat{S}}(y_j^2 - \alpha\sigma^2)\Big)^2 &\le 4\sigma^4 s^2\log^2(1+d/s^2) + 2\sigma^4 s^2\alpha^2 \\ &\le C_2\sigma^4 s^2\log^2(1+d/s^2)\end{aligned}$$

for any $\theta \in B_0(s)$. Finally, α was chosen such that, for any $j \notin S$,

$$\mathbf{E}_\theta\Big[\big(y_j^2 - \alpha\sigma^2\big)\mathbb{1}_{\{|y_j|>\sigma x\}}\Big] = \sigma^2\mathbf{E}\Big[\big(X^2-\alpha\big)\mathbb{1}_{\{|X|>x\}}\Big] = 0,$$

where $X \sim \mathcal{N}(0,1)$. Thus, by independence we have

$$\begin{aligned}\mathbf{E}_\theta\Big(\sum_{j\in\hat{S}\setminus S}(y_j^2 - \alpha\sigma^2)\Big)^2 &= \sum_{j\notin S}\mathbf{E}_\theta\Big[\big(y_j^2-\alpha\sigma^2\big)^2\mathbb{1}_{\{|y_j|>\sigma x\}}\Big] \\ &\le \sigma^4 d\,\mathbf{E}\Big[\big(X^2-\alpha\big)^2\mathbb{1}_{\{|X|>x\}}\Big] \\ &\le 16\sigma^4 d\,\mathbf{E}\Big[X^4\mathbb{1}_{\{|X|>x\}}\Big]\end{aligned} \tag{57}$$

since $\alpha \le 5X^2$ on the event $\{|X|>x\}$, cf. (53). Now, Lemma 4 implies

$$\mathbf{E}_\theta\Big(\sum_{j\in\hat{S}\setminus S}(y_j^2 - \alpha\sigma^2)\Big)^2 \le C_3\sigma^4 d\,x^3 e^{-x^2/2},$$

and by definition of x,

$$\mathbf{E}_\theta\Big(\sum_{j\in\hat{S}\setminus S}(y_j^2 - \alpha\sigma^2)\Big)^2 \le C_4\sigma^4 s^2 x^3 \le (C_4/\sqrt{2\log 2})\sigma^4 s^2 x^4 \le C_5\sigma^4 s^2\log^2(1+d/s^2),$$

where we have used the fact that $x \ge \sqrt{2\log 2}$. Combining the above displays yields

$$\sup_{\theta\in B_2(\kappa)\cap B_0(s)}\mathbf{E}_\theta\big(\hat{Q}' - Q(\theta)\big)^2 \le C_6\max(\sigma^2\kappa^2, \sigma^4 s^2\log^2(1+d/s^2)).$$

Remark 4. This proof elucidates why we have chosen the threshold x in the form (43). In (54), the three terms on the right hand side are of the order respectively $\sigma^2\kappa^2 + \sigma^4 s^2 x^4$, $\sigma^4 s^2 x^4$, and $\sigma^4 d x^3 e^{-x^2/2}$. Among these, the expressions containing x are balanced if $\sigma^4 s^2 x^4 \asymp \sigma^4 d x^3 e^{-x^2/2}$, which is equivalent to $x e^{x^2/2} \asymp d/s^2$. This leads to a choice of x in the form $\sqrt{2\log(d/s^2) - \log\log(d/s^2)} \asymp \sqrt{2\log(d/s^2)}$.

8.4 Proof of Theorem 7

Fix $\theta \in B_q(r)$. We will prove the theorem only for $1 \le m \le \sqrt{d}$ since the case $m = 0$ is trivial and the result for the case $m > \sqrt{d}$ follows from (52) and the fact that $\|\theta\|_2 \le \|\theta\|_q \le r$. In this proof, we will write for brevity $\alpha = \tilde{\alpha}_m$, $\tilde{d} = 1 + d/m^2$, $\tilde{x} = 2(2\log\tilde{d})^{1/2}$. Let $J \subseteq \{1, \dots, d\}$ be the set of indices corresponding to the m largest in absolute value components of θ, and let $|\theta|_{(j)}$ denote the jth largest absolute value of the components of θ. It is easy to see that

$$|\theta|_{(j)} \le \frac{\|\theta\|_q}{j^{1/q}}.$$

This implies

$$\sum_{j\in J^c} \theta_j^2 = \sum_{j\ge m+1} |\theta|_{(j)}^2 \le |\theta|_{(m)}^{2-q} \sum_{j\ge m+1} |\theta|_{(j)}^q \le \left(\frac{\|\theta\|_q}{m^{1/q}}\right)^{2-q} \|\theta\|_q^q = \|\theta\|_q^2 m^{1-2/q}.$$

Therefore, since $\theta \in B_q(r)$ and due to the definition of m,

$$\sum_{j\in J^c} \theta_j^2 \le r^2 m^{1-2/q} \le \sigma^2 m \log\tilde{d}, \tag{58}$$

and

$$\forall\, j \in J^c: \quad |\theta_j| \le r m^{-1/q} \le \sigma\sqrt{\log\tilde{d}} \le \sigma\tilde{x}/2. \tag{59}$$

We have

$$\begin{aligned} \hat{Q}_q - Q(\theta) = &\sum_{j\in J} \{y_j^2 - \alpha\sigma^2 - \theta_j^2\} - \sum_{j\in J\setminus\tilde{S}} \{y_j^2 - \alpha\sigma^2\} \\ &+ \sum_{j\in\tilde{S}\setminus J} \{y_j^2 - \alpha\sigma^2\} - \sum_{j\in J^c} \theta_j^2, \end{aligned} \tag{60}$$

where $\tilde{S} = \{j : |\theta_j| > \sigma\tilde{x}/2\}$. Consider the first sum on the right hand side of (60). Since $\mathrm{Card}(J) = m$, and $\alpha \le 40\log\tilde{d}$ (which is obtained analogously to (53) recalling that now $\alpha = \tilde{\alpha}_m$ instead of $\alpha = \alpha_s$), the same argument as in (56) leads to

$$\mathbf{E}_\theta\Big(\sum_{j\in J} \{y_j^2 - \alpha\sigma^2 - \theta_j^2\}\Big)^2 \le C(\sigma^2\|\theta\|_2^2 + \sigma^4 m^2 \log^2\tilde{d}). \tag{61}$$

Next, consider the second sum on the right hand side of (60). By definition of $\tilde{S}$,

$$\mathbf{E}_\theta\Big(\sum_{j\in J\setminus\tilde{S}} \{y_j^2 - \alpha\sigma^2\}\Big)^2 \le \Big(\sum_{j\in J} \sigma^2(\tilde{x} + \alpha)\Big)^2 \le C\sigma^4 m^2 \log^2\tilde{d}. \tag{62}$$

Let us now turn to the third sum on the right hand side of (60). The bias-variance decomposition yields

$$\mathbf{E}_\theta\Big(\sum_{j\in\tilde{S}\setminus J}\{y_j^2-\alpha\sigma^2\}\Big)^2=\mathbf{E}_\theta\Big(\sum_{j\in J^c}(y_j^2-\alpha\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)^2$$
$$=\sum_{j\in J^c}\mathrm{Var}_\theta\Big((y_j^2-\alpha\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)+\Big[\sum_{j\in J^c}\mathbf{E}_\theta\Big((y_j^2-\alpha\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)\Big]^2.$$

Here,

$$\begin{aligned}\mathrm{Var}_\theta\Big((y_j^2-\alpha\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)&\le\mathbf{E}_\theta\Big((y_j^2-\alpha\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)^2\\&\le C\mathbf{E}_\theta\Big((\theta_j^4+\sigma^4\xi_j^4+\alpha^2\sigma^4)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)\\&\le C\Big[\theta_j^4+\alpha^2\sigma^4+\sigma^4\mathbf{E}\Big(\xi_j^4\,\mathbb{1}_{\{|\xi_j|>\tilde{x}/2\}}\Big)\Big]\quad\text{(by (59))}.\end{aligned}$$

Using now the same argument as in (57) to bound $\mathbf{E}\big(\xi_j^4\,\mathbb{1}_{\{|\xi_j|>\tilde{x}/2\}}\big)$ we obtain

$$\begin{aligned}\sum_{j\in J^c}\mathrm{Var}_\theta\Big((y_j^2-\alpha\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)&\le C\big(\sum_{j\in J^c}\theta_j^4+\sigma^4m^2\log^2\tilde{d}\big)\\&\le C\Big(\Big(\sum_{j\in J^c}\theta_j^2\Big)^2+\sigma^4m^2\log^2\tilde{d}\Big).\end{aligned}$$

Furthermore, by Lemma 6,

$$\Big|\sum_{j\in J^c}\mathbf{E}_\theta\Big((y_j^2-\alpha\sigma^2)\,\mathbb{1}_{\{|y_j|>\sigma\tilde{x}\}}\Big)\Big|\le C\sum_{j\in J^c}\theta_j^2.$$

Combining the above displays leads to the following bound :

$$\mathbf{E}_\theta\Big(\sum_{j\in\tilde{S}\setminus J}\{y_j^2-\alpha\sigma^2\}\Big)^2\le C\Big(\Big(\sum_{j\in J^c}\theta_j^2\Big)^2+\sigma^4m^2\log^2\tilde{d}\Big).\tag{63}$$

From (60) - (63) we deduce that

$$\mathbf{E}_\theta(\hat{Q}_q-Q(\theta))^2\le C\Big(\sigma^2\|\theta\|_2^2+\Big(\sum_{j\in J^c}\theta_j^2\Big)^2+\sigma^4m^2\log^2\tilde{d}\Big).$$

The result now follows if we use (58) and note that $\|\theta\|_2\le\|\theta\|_q\le r$.

8.5 Proof of the upper bound (19) in Theorem 8

Fix $\theta\in B_0(s)$ and set for brevity $\tau=\big(\psi_\sigma^{\sqrt{Q}}(s,d)\big)^{1/2}$. We will bound the risk $\mathbf{E}_\theta(\hat{N}-\|\theta\|_2)^2$ separately for the cases $\|\theta\|_2\le\tau$ and $\|\theta\|_2>\tau$.

Case $\|\theta\|_2\le\tau$. Using the elementary inequality $(a-b)^2\le2(a^2-b^2)+4b^2$, we find

$$\mathbf{E}_\theta(\hat{N}-\|\theta\|_2)^2\le2\,\mathbf{E}_\theta(\max\{\hat{Q}_\bullet,0\}-Q(\theta))+4Q(\theta)\le2\,\Big(\mathbf{E}_\theta(\hat{Q}_\bullet-Q(\theta))^2\Big)^{1/2}+4\tau^2.$$

Note that $\hat{Q}_\bullet = \hat{Q}$ if we set $\kappa = \tau$ in the definition of $\hat{Q}$. Furthermore, $\theta \in B_0(s)$ and, in the case under consideration θ belongs to $B_2(\tau)$. Now, use that for all $\theta \in B_2(\tau) \cap B_0(s)$, due to Theorem 5, we have

$$\mathbf{E}_\theta(\hat{Q}_\bullet - Q(\theta))^2 \le C\psi_\sigma^Q(s,d,\tau).$$

Using this inequality and the fact that $\psi_\sigma^Q(s,d,\tau) = \big(\psi_\sigma^{\sqrt{Q}}(s,d)\big)^2$, we obtain the desired rate:

$$\mathbf{E}_\theta(\hat{N} - \|\theta\|_2)^2 \le C_7\psi_\sigma^{\sqrt{Q}}(s,d) + 4\tau^2 = (C_7+4)\psi_\sigma^{\sqrt{Q}}(s,d).$$

Case $\|\theta\|_2 > \tau$. Using the elementary inequality $\forall\ a > 0, b \ge 0$, $(a-b)^2 \le (a^2-b^2)^2/a^2$, we find

$$\mathbf{E}_\theta(\hat{N} - \|\theta\|_2)^2 \le \frac{\mathbf{E}_\theta(\hat{Q}_\bullet - Q(\theta))^2}{\|\theta\|_2^2}.$$

Now, we bound $\mathbf{E}_\theta(\hat{Q}_\bullet - Q(\theta))^2$ along the lines of the proof of Theorem 5. In particular, if $s \ge \sqrt{d}$ we have $\hat{Q}_\bullet = \hat{Q}_*$, $\tau = \sigma d^{1/4}$ and using (52) we obtain

$$\frac{\mathbf{E}_\theta(\hat{Q}_\bullet - Q(\theta))^2}{\|\theta\|_2^2} \le 4\sigma^2 + \frac{2\sigma^4 d}{\|\theta\|_2^2} \le 4\sigma^2 + \frac{2\sigma^4 d}{\tau^2} \le C_8\sigma^2\sqrt{d},$$

which is the desired rate. If $s < \sqrt{d}$, we have $\hat{Q}_\bullet = \hat{Q}'$, $\tau = \sigma\sqrt{s\log(1+d/s^2)}$ and using (56) and the subsequent bounds in the proof of Theorem 5, we obtain

$$\begin{aligned}\frac{\mathbf{E}_\theta(\hat{Q}_\bullet - Q(\theta))^2}{\|\theta\|_2^2} &\le \frac{3\big(C_1\sigma^2\|\theta\|_2^2 + (C_1+C_2+C_5)\sigma^4 s^2\log^2(1+d/s^2)\big)}{\|\theta\|_2^2} \\ &\le C_9\left(\sigma^2 + \frac{\sigma^4 s^2\log^2(1+d/s^2)}{\tau^2}\right) \le C_{10}\sigma^2 s\log(1+d/s^2),\end{aligned} \tag{64}$$

which is again the desired rate.

8.6 Proof of the upper bound (21) in Theorem 9

The case $m = 0$ is trivial. For $m \ge 1$, we use the same method of reduction to the risk of estimators of Q as in the proof of (19). The difference is that now we set $\tau = \big(\psi_{\sigma,q}^{\sqrt{Q}}(r,d)\big)^{1/2}$, we replace s by m, and we apply Theorem 7 rather than to Theorem 5. In particular, an analog of (64) with $s = m$ is obtained using (61).

8.7 Proof of Theorem 10

Here, we will use the notation set in (43). As in the proof of the bound (3) and with the same notation, we have, for $\theta \in B_0(s)$,

$$\begin{aligned}\mathbf{E}_\theta(\tilde{L} - L(\theta))^2 &\le 3\mathbf{E}\Big(\sum_{j\in S}\sigma\xi_j\Big)^2 + 3\mathbf{E}_\theta\Big(\sum_{j\in S} y_j\ \mathbb{1}_{\{|y_j|\le\hat{\sigma}x\}}\Big)^2 + 3\mathbf{E}\Big(\sum_{j\in S^c}\sigma\xi_j\ \mathbb{1}_{\{\sigma|\xi_j|>\hat{\sigma}x\}}\Big)^2 \\ &\le 3\Big\{(s\sigma^2 + s^2\mathbf{E}_\theta(\hat{\sigma}^2)x^2) + \sigma^2\sum_{j\in S^c}\mathbf{E}\Big(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat{\sigma}x\}}\Big)\Big\}.\end{aligned}$$

Here,

$$\mathbf{E}_\theta\Big(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat\sigma\sqrt{2\log(1+d/s^2)}\}}\Big)$$
$$= \mathbf{E}_\theta\Big(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat\sigma\sqrt{2\log(1+d/s^2)}\}}\mathbb{1}_{\{\hat\sigma>\sigma\}}\Big) + \mathbf{E}_\theta\Big(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat\sigma\sqrt{2\log(1+d/s^2)}\}}\mathbb{1}_{\{\hat\sigma\le\sigma\}}\Big).$$

The first term on the right hand side satisfies

$$\mathbf{E}_\theta\Big(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat\sigma\sqrt{2\log(1+d/s^2)}\}}\mathbb{1}_{\{\hat\sigma>\sigma\}}\Big) \le \mathbf{E}_\theta\Big(\xi_j^2\ \mathbb{1}_{\{|\xi_j|>\sqrt{2\log(1+d/s^2)}\}}\Big)$$
$$\le \frac{Cs^2}{d}\sqrt{\log(1+d/s^2)} \quad \text{(by (41))}.$$

For the second term, we use Lemma 7 to get

$$\mathbf{E}_\theta\Big(\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat\sigma\sqrt{2\log(1+d/s^2)}\}}\mathbb{1}_{\{\hat\sigma\le\sigma\}}\Big) \le \sqrt{\mathbf{E}(\xi_1^4)}\sqrt{\mathbf{P}_\theta(\hat\sigma\le\sigma)} \le C\sqrt{d}\exp(-\sqrt{d}/C).$$

Combining the above displays and using Lemma 7 to bound $\mathbf{E}_\theta(\hat\sigma^2)$ we obtain

$$\mathbf{E}_\theta(\tilde L - L(\theta))^2 \le C\sigma^2 s^2\log(1+d/s^2).$$

8.8 Proof of Theorem 11

Set $\tilde S = \{j : |y_j| \ge \hat\sigma\sqrt{2\log d}\}$ and recall that $S = \{j : \theta_j \ne 0\}$. As in the proof of Theorem 5 we get

$$\mathbf{E}_\theta\big(\tilde Q - Q(\theta)\big)^2 \le 3\ \mathbf{E}_\theta\Big[\Big(\sum_{j\in S}(y_j^2-\theta_j^2)\Big)^2 + \Big(\sum_{j\in S\setminus\tilde S} y_j^2\Big)^2 + \Big(\sum_{j\in\tilde S\setminus S} y_j^2\Big)^2\Big].$$

We bound separately the three terms on the right hand side. For $\theta\in B_2(\kappa)\cap B_0(s)$, the first term on the right-hand side satisfies, due to (56) with $\alpha=0$,

$$\mathbf{E}_\theta\Big(\sum_{j\in S}(y_j^2-\theta_j^2)\Big)^2 \le C\ \big(\sigma^2\|\theta\|_2^2+\sigma^4 s^2\big) \le C\ \big(\sigma^2\kappa^2+\sigma^4 s^2\big). \tag{65}$$

Using Lemma 7 we find

$$\mathbf{E}_\theta\Big(\sum_{j\in S\setminus\tilde S} y_j^2\Big)^2 = \mathbf{E}_\theta\Big(\sum_{j\in S} y_j^2\ \mathbb{1}_{\{|y_j|<\hat\sigma\sqrt{2\log d}\}}\Big)^2$$
$$\le s^2\mathbf{E}_\theta(\hat\sigma^4)(2\log d)^2 \le C\sigma^4 s^2\log^2 d. \tag{66}$$

Finally, we write the third term as follows

$$\mathbf{E}_\theta\Big(\sum_{j\in\tilde S\setminus S} y_j^2\Big)^2 = \mathbf{E}_\theta\Big(\sum_{j\notin S}\sigma^2\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat\sigma\sqrt{2\log d}\}}\Big)^2 \le 2(A_1+A_2) \tag{67}$$

where

$$A_1 = \mathbf{E}_\theta\Big(\sum_{j=1}^d \sigma^2\xi_j^2\ \mathbb{1}_{\{\sigma|\xi_j|>\hat\sigma\sqrt{2\log d}\}}\mathbb{1}_{\{\hat\sigma>\sqrt{2}\sigma\}}\Big)^2,$$
$$A_2 = \mathbf{E}_\theta\Big(\sum_{j=1}^d \sigma^2\xi_j^2\ \mathbb{1}_{\{\hat\sigma\le\sqrt{2}\sigma\}}\Big)^2.$$

Using (42) we obtain

$$A_1 \le \sigma^4 \mathbf{E}_\theta\Big(\sum_{j=1}^d \xi_j^2 \ \mathbb{1}_{\{|\xi_j|>2\sqrt{\log d}\}}\Big)^2 \le 2\sigma^4 d^2 \mathbf{E}\big(X^4 \ \mathbb{1}_{\{|X|>2\sqrt{\log d}\}}\big) \le C\sigma^4(\log d)^{3/2} \tag{68}$$

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

$$A_2 \le \sigma^4 \mathbf{E}_\theta\Big(\sum_{j=1}^d \xi_j^2 \ \mathbb{1}_{\{\hat\sigma\le\sqrt{2}\sigma\}}\Big)^2 \le \sigma^4 d^2 \max_{1\le j\le d} \mathbf{E}_\theta\big(\xi_j^4 \ \mathbb{1}_{\{\hat\sigma\le\sqrt{2}\sigma\}}\big).$$

Using (42) we find

$$\begin{aligned}\mathbf{E}_\theta\big(\xi_j^4 \ \mathbb{1}_{\{\hat\sigma\le\sqrt{2}\sigma\}}\big) &\le \mathbf{E}_\theta\big(\xi_j^4 \ \mathbb{1}_{\{|\xi_j|>2\sqrt{\log d}\}}\big) + \mathbf{E}_\theta\big(\xi_j^4 \ \mathbb{1}_{\{|\xi_j|\le 2\sqrt{\log d}\}} \ \mathbb{1}_{\{\hat\sigma\le\sqrt{2}\sigma\}}\big)\\ &\le \frac{C}{d^2}(\log d)^{3/2} + 16(\log d)^2 \mathbf{P}_\theta(\hat\sigma \le \sqrt{2}\sigma).\end{aligned}$$

The last two displays and the bound for $\mathbf{P}_\theta(\hat\sigma \le \sqrt{2}\sigma)$ from Lemma 7 yield

$$A_2 \le C\sigma^4(\log d)^{3/2}. \tag{69}$$

Combining (65) - (69) proves the theorem.

9 Appendix: Auxiliary lemmas

Proof of Lemma 1. We first follow the lines of the proof of Theorem 7 in [9] and then apply a result of [2] (*cf.* also Section 6 in [20]) in the same spirit as it was done in [4]. Let φ_σ be a density of normal distribution with mean 0 and variance σ^2. For $I \in \mathcal{S}(s,d)$, let

$$g_I(y_1,\dots,y_d) = \prod_{j=1}^d \varphi_\sigma(y_j - f_j)$$

where $f_j = \sigma\rho \mathbb{1}_{j\in I}$. The density of $\mathbb{P}_{\mu_\rho}$ is

$$g = \frac{1}{\binom{d}{s}} \sum_{I\in\mathcal{S}(s,d)} g_I$$

and we can write

$$\chi^2(\mathbb{P}_{\mu_\rho}, \mathbf{P}_0) = \int \Big(\frac{d\mathbb{P}_{\mu_\rho}}{d\mathbf{P}_0}\Big)^2 d\mathbf{P}_0 - 1 = \int \frac{g^2}{f} - 1$$

where f is a density of n i.i.d. normal random variables with mean 0 and variance σ^2. Now,

$$\int \frac{g^2}{f} = \frac{1}{\binom{d}{s}^2} \sum_{I\in\mathcal{S}(s,d)} \sum_{I'\in\mathcal{S}(s,d)} \int \frac{g_I g_{I'}}{f}.$$

It is easy to see that

$$\int \frac{g_I g_{I'}}{f} = \exp(\rho^2 \mathrm{Card}(I \cap I')),$$

which implies

$$\int \frac{g^2}{f} = \mathbf{E} \exp(\rho^2 J)$$

where J is a random variable with hypergeometric distribution,

$$\mathbf{P}(J = j) = \frac{\binom{s}{j}\binom{d-s}{s-j}}{\binom{d}{s}}.$$

As shown in [2], J coincides in distribution with the conditional expectation $\mathbf{E}[Z|\mathcal{B}]$ where Z is a binomial random variable with parameters $(s,\ s/d)$ and $\mathcal{B}$ is a suitable σ-algebra. This fact and Jensen's inequality lead to the following bound implying the lemma:

$$\int \frac{g^2}{f} \leq \mathbf{E} \exp(\rho^2 Z) = \left(1 - \frac{s}{d} + \frac{s}{d} e^{\rho^2}\right)^s.$$

□

In the next two lemmas, we will use the notation $D_i(t) = \mathbf{E}\big(X^i \mathbb{1}_{\{X>t\}}\big)$ for $i \geq 0, t > 0$, where $X \sim \mathcal{N}(0,1)$. Clearly, $D_0(t) = 1 - \Phi(t)$, and $D_1(t) = \phi(t)$ where Φ and ϕ are the standard normal c.d.f. and density respectively. For $i \geq 2$ integration by parts gives $D_i(t) = t^{i-1}\phi(t) + (i-1)D_{i-2}(t)$. It follows that $D_i(t) = O(t^{i-1}e^{-t^2/2})$ as $t \to \infty$, and each D_i as well as each of its derivatives is uniformly bounded.

Lemma 5. *Let $y \sim \mathcal{N}(a, \sigma^2)$ and $\hat{T} = y\,\mathbb{1}_{\{|y|>\sigma\tau\}}$ where $\tau > 0$. Set $B(a) = \mathbf{E}(\hat{T}) - a$. Then there exists $C > 0$ such that*

$$|B(a)| \leq C \min(|a|, \sigma\tau).$$

Proof. Note that $B(a) = \mathbf{E}(y\ \mathbb{1}_{\{|y|\leq\sigma\tau\}})$, so that $|B(a)| \leq \sigma\tau$. Thus, it remains to show that there exists $C > 0$ such that $|B(a)| \leq C|a|$. We have

$$B(a) = a\big(D_0(\tau + a/\sigma) + D_0(\tau - a/\sigma)\big) + \sigma\big(D_1(\tau + a/\sigma) - D_1(\tau - a/\sigma)\big).$$

Since all D_i and their derivatives are uniformly bounded the result follows. □

Lemma 6. *Let $y \sim \mathcal{N}(a, \sigma^2)$, and $\tau > 0$. Let α be such that $\mathbf{E}\Big[(X^2 - \alpha)\ \mathbb{1}_{\{|X|>\tau\}}\Big] = 0$, where $X \sim \mathcal{N}(0,1)$. Then there exists $C > 0$ such that*

$$\Big|\mathbf{E}\Big[(y^2 - \alpha\sigma^2)\ \mathbb{1}_{\{|y|>\sigma\tau\}}\Big]\Big| \leq Ca^2.$$

Proof. We have

$$\mathbf{E}\Big[(y^2 - \alpha\sigma^2)\ \mathbb{1}_{\{|y|>\sigma\tau\}}\Big] = \sigma^2\Big(D_2(\tau + a/\sigma) + D_2(\tau - a/\sigma)\Big) + 2a\sigma\Big(D_1(\tau + a/\sigma) - D_1(\tau - a/\sigma)\Big)$$
$$+ (a^2 - \alpha\sigma^2)\Big(D_0(\tau + a/\sigma) + D_0(\tau - a/\sigma)\Big).$$

Using that D_0 is bounded and D_1 is Lipschitz continuous, we see that it is enough to check the condition $|f(a)| \le Ca^2$ for

$$f(a) = \sigma^2\Big[D_2(\tau + a/\sigma) + D_2(\tau - a/\sigma) - \alpha\Big(D_0(\tau + a/\sigma) + D_0(\tau - a/\sigma)\Big)\Big].$$

Now, $f(0) = 0$ by definition of α and $f'(0) = 0$ because f is symmetric. Since the second derivatives of D_2 and D_0 are uniformly bounded Taylor's theorem gives the result. □

Lemma 7. *For any θ such that $\|\theta\|_0 \le \sqrt{d}$ we have*

$$\mathbf{E}_\theta(\hat\sigma^2) \le 9\sigma^2, \qquad \mathbf{E}_\theta(\hat\sigma^4) \le C\sigma^4, \tag{70}$$

and

$$\mathbf{P}_\theta(\hat\sigma \le \sigma) \le Cd\exp(-\sqrt{d}/C) \tag{71}$$

for some absolute constant $C > 0$.

Proof. Since $\|\theta\|_0 \le \sqrt{d}$ we have

$$\hat\sigma^2 \le \frac{9}{d}\sum_{j=1}^{d-\|\theta\|_0} y_{(j)}^2.$$

Denote by F the set of indices i corresponding to the $d - \|\theta\|_0$ smallest values y_i^2. Then

$$\sum_{j=1}^{d-\|\theta\|_0} y_{(j)}^2 = \sum_{i\in F} y_i^2 = \sigma^2\sum_{i\in S^c}\xi_i^2 + \sum_{i\in S\cap F} y_i^2 - \sigma^2\sum_{i\in S^c\cap F^c}\xi_i^2$$

where $S = \{j : \theta_j \ne 0\}$. For any $i \in S\cap F$ and any $j \in S^c\cap F^c$, we have

$$y_i^2 \le \sigma^2\xi_j^2.$$

Furthermore, $\mathrm{Card}(S\cap F) = \mathrm{Card}(S^c\cap F^c)$. Therefore,

$$\hat\sigma^2 \le \frac{9\sigma^2}{d}\sum_{i\in S^c}\xi_i^2.$$

This implies (70). We now prove (71). Let G be the set of indices i corresponding to the $\lfloor d - \sqrt{d}\rfloor$ smallest y_i^2. Here, $\lfloor x\rfloor$ denotes the largest integer less than or equal to x. Then we have

$$\sum_{j\le d-\sqrt{d}} y_{(j)}^2 = \sum_{i\in G} y_i^2 \ge \sigma^2\sum_{i\in S^c\cap G}\xi_i^2 \ge \sigma^2\sum_{i\in S^c}\xi_i^2 - 2\sqrt{d}\,\sigma^2\max_{i\in S^c}\xi_i^2,$$

where we have used that $\mathrm{Card}(G^c) \le 2\sqrt{d}$. This implies:

$$\hat\sigma^2 \ge \frac{9\sigma^2}{d}\sum_{i\in S^c}\xi_i^2 - \frac{18\sigma^2}{\sqrt{d}}\max_{i\in S^c}\xi_i^2.$$

Thus,

$$\mathbf{P}_\theta(\hat\sigma \le \sqrt{2}\sigma) \le \mathbf{P}\Big(9\sigma^2 \sum_{i\in S^c} \xi_i^2 - 18\sqrt{d}\sigma^2 \max_{i\in S^c} \xi_i^2 \le 2d\sigma^2\Big)$$

$$\le \mathbf{P}\Big(9 \sum_{i\in S^c} \xi_i^2 \le 3d\Big) + \mathbf{P}\Big(18 \max_{i\in S^c} \xi_i^2 \ge \sqrt{d}\Big). \tag{72}$$

The first term on the right hand side of (72) satisfies

$$\mathbf{P}\Big(3 \sum_{i\in S^c} \xi_i^2 \le d\Big) \le \mathbf{P}\left(U_D - D \le -2d/3 + \sqrt{d}\right)$$

where $D = \mathrm{Card}(S^c)$, and U_D is a χ^2 random variable with D degrees of freedom. A standard bound on the tails of χ^2 random variables (see, e.g. [33]) yields

$$\mathbf{P}(U_D - D \le -t) \le \exp(-t^2/(4D)), \qquad \forall\, t > 0.$$

Thus, for $d > 2$, we obtain

$$\mathbf{P}\Big(3 \sum_{i\in S^c} \xi_i^2 \le d\Big) \le \exp(-(2d/3 - \sqrt{d})^2/(4D)) \le \exp(-d/C)$$

where $C > 0$ is an absolute constant. Finally, the second term on the right hand side of (72) satisfies

$$\mathbf{P}\left(\max_{i\in S^c} \xi_i^2 \ge \frac{\sqrt{d}}{18}\right) \le d \exp\left(-\frac{\sqrt{d}}{36}\right)$$

in view of (40). Plugging the last two displays in (72) we obtain (71). □

Acknowledgement

We would like to thank Nicolas Verzelen for remarks on the text that helped to improve the presentation. 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), Labex ECODEC (ANR - 11-LABEX-0047), and ANR -11- IDEX-0003-02. 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] Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. *Electron. J. Stat.* **4** 932–949.

[2] Aldous, D.J. (1985). *Exchangeability and Related Topics, École d'été de Saint-Flour XIII – 1983. Lecture Notes in Mathematics,* **1117**. Springer, New York.

[3] Arias-Castro, E. Candès, E. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556.

[4] Baraud, Y. (2002). Non asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606.

[5] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203–268.

[6] Birnbaum, Z.W. (1942). An inequality for Mills ratio. *Ann. Math. Statist.* **13** 245–246.

[7] Butucea, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *Ann. Statist.* **35** 1907–1930.

[8] Butucea, C. and Comte, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli.* **15** 69–98.

[9] Cai, T. T. and Low, M.L. (2004). Minimax Estimation of Linear Functionals Over Nonconvex Parameter Spaces. *Ann. Statist.* **32** 552–576.

[10] Cai, T. T. and Low, M.L. (2005a). On adaptive estimation of linear functionals. *Ann. Statist.* **33** 2311–2343.

[11] Cai, T. T. and Low, M.L. (2005b). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33** 2930–2956.

[12] Donoho, D.L. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994.

[13] Donoho, D.L. and Johnstone, I.M. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99** 277–303.

[14] Donoho, D.L. and Liu, R. (1991). Geometrizing rates of convergence. III. *Ann. Statist.* **19** 668–701.

[15] Donoho, D.L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

[16] Efromovich, S. and Low, M.L. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24** 1106–1125.

[17] Goldenshluger, A. and Pereverzev, S.V. (2003). On adaptive inverse estimation of linear functionals. *Bernoulli* **9** 783–807.

[18] Golubev, G.K. (2004). The method of risk envelopes in the estimation of linear functionals. *Problemy Peredachi Informatsii* **40** 58–72.

[19] 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.

[20] W. Hoeffding *Probability inequalities for sums of bounded random variables.* J. Amer. Statist. Assoc., 58, 13-30, 1963.

[21] Ibragimov, I.A. and Hasminskii, R.Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18–32.

[22] Ingster, Y.I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6** 47–49.

[23] Ingster, Y.I., Pouet, C. and Tsybakov, A.B. (2009). Classification of sparse high-dimensional vectors. *Phil. Transactions of the Royal Soc., A.* **367** 4427–4448.

[24] Ingster, Y.I. and Suslina, I.A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models.* Springer, New York.

[25] Ingster, Y.I., Tsybakov, A.B. and Verzélen, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526.

[26] Johnstone, I.M. (2001a). Chi-square oracle inequalities. *Lecture Notes-Monograph Series* **36** 399-418.

[27] Johnstone, I.M. (2001b). Thresholding for weighted χ^2. *Statist. Sinica* **11** 691–704.

[28] Johnstone, I.M. (2013). *Gaussian Estimation: Sequence and Wavelet Models.* Book draft.

[29] Juditsky, A. and Nemirovski, A. (2009). Nonparametric estimation via convex programming. *Ann. Statist.* **37** 2278–2300.

[30] Klemelä, J. (2006). Sharp adaptive estimation of quadratic functionals. *Probab. Theory Related Fields* **134** 539–564.

[31] Klemelä, J. and Tsybakov, A.B. (2001). Sharp adaptive estimation of linear functionals. *Ann. Statist.* **29** 1567–1600.

[32] Laurent, B., Ludena, C. and Prieur, C. (2008). Adaptive estimation of linear functionals by model selection. *Electron. J. Stat.* **2** 993–1020.

[33] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338.

[34] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

[35] Lepski, O., Nemirovski, A. and Spokoiny, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253.

[36] Nemirovski, A. (2000). *Topics in Nonparametric Statistics. Ecole d'été de Probabilités de Saint Flour 1998. Lecture Notes in Mathematics* **1738**. Springer, New York.

[37] Rigollet, P. and Tsybakov, A.B. (2011). Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771.

[38] Sampford, M.R. (1953). Some inequalities on Mills ratio and related functions. *Ann. Math. Statist.* **24** 132–134.

[39] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics, New York, 2009.

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