Bernoulli **26**(3), 2020, 1989–2020
https://doi.org/10.3150/19-BEJ1180

On estimation of nonsmooth functionals of sparse normal means

O. COLLIER[1,3,*], L. COMMINGES[2,3,†] and A.B. TSYBAKOV[3,‡]

[1]*MODAL'X, Université Paris-Nanterre. E-mail:* [*] *olivier.collier@parisnanterre.fr*
[2]*CEREMADE, Université Paris-Dauphine PSL. E-mail:* [†] *comminges@ceremade.dauphine.fr*
[3]*CREST, ENSAE. E-mail:* [‡] *alexandre.tsybakov@ensae.fr*

We study the problem of estimation of $N_\gamma(\theta) = \sum_{i=1}^d |\theta_i|^\gamma$ for $\gamma > 0$ and of the ℓ_γ-norm of θ for $\gamma \geq 1$ based on the observations $y_i = \theta_i + \varepsilon\xi_i$, $i = 1, \ldots, d$, where $\theta = (\theta_1, \ldots, \theta_d)$ are unknown parameters, $\varepsilon > 0$ is known, and ξ_i are i.i.d. standard normal random variables. We find the non-asymptotic minimax rate for estimation of these functionals on the class of s-sparse vectors θ and we propose estimators achieving this rate.

Keywords: functional estimation; nonsmooth functional; norm estimation; polynomial approximation; sparsity

1. Introduction

In recent years, there has been a growing interest in statistical estimation of non-smooth functionals (cf. Cai and Low [1], Jiao et al. [9], Wu and Yang [15,16], Han et al. [7,8], Carpentier and Verzelen [2], Fukuchi and Sakuma [6]). Some of these papers deal with the normal means model (cf. Cai and Low [1], Carpentier and Verzelen [2]) addressing the problems of estimation of the ℓ_1-norm and of the sparsity index, respectively. In the present paper, we analyze a family of nonsmooth functionals including, in particular, the ℓ_1-norm. We establish non-asymptotic minimax optimal rates of estimation on the classes of sparse vectors and we construct estimators achieving these rates.

Assume that we observe

$$y_i = \theta_i + \varepsilon\xi_i, \quad i = 1, \ldots, d, \tag{1}$$

where $\theta = (\theta_1, \ldots, \theta_d)$ is an unknown vector of parameters, $\varepsilon > 0$ is a known noise level, and ξ_i are i.i.d. standard normal random variables. We consider the problem of estimating the functionals

$$N_\gamma(\theta) = \sum_{i=1}^d |\theta_i|^\gamma, \quad \gamma > 0, \quad \text{and} \quad \|\theta\|_\gamma = \left(\sum_{i=1}^d |\theta_i|^\gamma\right)^{1/\gamma}, \quad \gamma \geq 1,$$

assuming that the vector θ is s-sparse, that is, θ belongs to the class

$$B_0(s) = \left\{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq s\right\}.$$

Here, $\|\theta\|_0$ denotes the number of nonzero components of θ and $s \in \{1, \ldots, d\}$.

Set $n_\gamma(\theta) = N_\gamma(\theta)$ for $0 < \gamma \le 1$ and $n_\gamma(\theta) = \|\theta\|_\gamma$ for $\gamma > 1$. We measure the accuracy of an estimator $\hat{T}$ of $N_\gamma(\theta)$ by the maximal quadratic risk over $B_0(s)$:

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta\big[\big(\hat{T} - n_\gamma(\theta)\big)^2\big].$$

Here and in the sequel, we denote by $\mathbf{E}_\theta$ the expectation with respect to the joint distribution $\mathbf{P}_\theta$ of $(y_1, \dots, y_d)$ satisfying (1).

In this paper, we propose rate-optimal estimators in a non-asymptotic minimax sense, that is, estimators $\hat{T}_\gamma^*$ such that

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta\big[\big(\hat{T}_\gamma^* - n_\gamma(\theta)\big)^2\big] \asymp \inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{E}_\theta\big[\big(\hat{T} - n_\gamma(\theta)\big)^2\big] := \mathcal{R}_{s,d}(\varepsilon, \gamma),$$

where $\inf_{\hat{T}}$ denotes the infimum over all estimators and, for two quantities a and b possibly depending on $s, d, \varepsilon, \gamma$, we write $a \asymp b$ if there exist positive constants c', c'' that may depend only on γ such that $c' \le a/b \le c''$. We establish the following explicit non-asymptotic characterization of the minimax risk:

$$\mathcal{R}_{s,d}(\varepsilon, \gamma) \asymp \begin{cases} \varepsilon^{2\gamma} s^2 \log^\gamma\big(1 + d/s^2\big), & \text{if } s \le \sqrt{d} \text{ and } 0 < \gamma \le 1, \\ \varepsilon^{2\gamma} s^2 \log^{-\gamma}\big(1 + s^2/d\big), & \text{if } s > \sqrt{d} \text{ and } 0 < \gamma \le 1, \end{cases} \tag{2}$$

and

$$\mathcal{R}_{s,d}(\varepsilon, \gamma) \asymp \begin{cases} \varepsilon^2 s^{2/\gamma} \log\big(1 + d/s^2\big), & \text{if } s \le \sqrt{d} \text{ and } \gamma > 1, \\ \varepsilon^2 d^{1/\gamma}, & \text{if } s > \sqrt{d} \text{ and } \gamma \in E, \end{cases} \tag{3}$$

where E is the set of all even integers. We also prove that, in the remaining case $s > \sqrt{d}$ and $\gamma > 1$ such that $\gamma \notin E$, we have

$$c\varepsilon^2 s^{2/\gamma} \log^{1-2\gamma}\big(1 + s^2/d\big) \le \mathcal{R}_{s,d}(\varepsilon, \gamma) \le \bar{c}\varepsilon^2 s^{2/\gamma} \log^{-1}\big(1 + s^2/d\big) \tag{4}$$

for some positive constants $c, \bar{c}$.

The case $s = d$, $\gamma = \varepsilon = 1$ was studied in Cai and Low [1], where it was proved that

$$\mathcal{R}_{d,d}(1, 1) = \inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^d} \mathbf{E}_\theta\big[\big(\hat{T} - N_1(\theta)\big)^2\big] \asymp \frac{d^2}{\log d}.$$

It was also claimed in Cai and Low [1] that $\mathcal{R}_{s,d}(1, 1) \asymp s^2/(\log d)$ for $s \ge d^\beta$ with $\beta > 1/2$, which agrees with the corresponding special case of (2).

We see from (2) and (3) that, for the general sparsity classes $B_0(s)$ there exist two different regimes with an elbow at $s \asymp \sqrt{d}$. We call them the sparse zone and the dense zone. The estimation methods for these two regimes are quite different. In the sparse zone, where s is smaller than $\sqrt{d}$, we show that one can use suitably adjusted thresholding to achieve optimality. In this zone, rate optimal estimators can be obtained based on the techniques developed in Collier et al. [3] to construct minimax optimal estimators of linear and quadratic functionals. In the dense zone,

where s is greater than $\sqrt{d}$, we use another approach. We follow the general scheme of estimation of non-smooth functionals from Lepski et al. [10] and our construction is especially close in the spirit to Cai and Low [1]. Specifically, we consider the best polynomial approximation of the function $|x|^\gamma$ in a neighborhood of the origin and plug in unbiased estimators for each power in the expression of this polynomial. Outside of this neighborhood, for i such that $|y_i|$ is, roughly speaking, greater than the "noise level" of the order of $\varepsilon\sqrt{\log d}$, we use $|y_i|^\gamma$ as an estimator of $|\theta_i|^\gamma$. The main difference from the estimator suggested in Cai and Low [1] for $\gamma = 1$ lies in the fact that, for the polynomial approximation part, we need to introduce a block structure with exponentially increasing blocks and carefully chosen thresholds depending on s. This is needed to achieve optimal bounds for all s in the dense zone and not only for $s = d$ or s comfortably greater than $\sqrt{d}$ as in Cai and Low [1].

In the present work, the variance ε^2 of the noise and the sparsity parameter s need to be known exactly. We conjecture that adaptation to ε^2 can be done without loss of efficiency in the sparse zone $s \le \sqrt{d}$. In the dense zone, the optimal rate can deteriorate dramatically when ε^2 is unknown as shown in Comminges et al. [5] for the case $\gamma = 2$. This contrasts with the results for linear functionals. Indeed, in Collier et al. [4] it is proved that, for linear functionals, adaptation to ε^2 can be done without loss of efficiency, and adaptation to s only brings a logarithmically small deterioration of the rate.

This paper is organized as follows. In Section 2, we introduce the estimators and state the upper bounds for their risks. Section 3 provides the matching lower bounds. The rest of the paper is devoted to the proofs. In particular, some useful results from approximation theory are collected in Section 6.

2. Definition of estimators and upper bounds for their risks

In this section, we propose two different estimators, for the dense and sparse regimes defined by the inequalities $s^2 \ge 4d$ and $s^2 < 4d$, respectively. Recall that, in the Introduction, we used the inequalities $s \ge \sqrt{d}$ and $s < \sqrt{d}$, respectively, to define the two regimes. The factor 4 that we introduce in the definition here is a matter of convenience for the proofs. We note that such a change does not influence the final result since the optimal rate (cf. (3)) is the same, up to a constant, for all s such that $s \asymp \sqrt{d}$.

2.1. Dense zone: $s^2 \ge 4d$

We first study the problem of estimation of $n_\gamma(\theta)$ in the dense zone. Two estimators will be proposed – the first one that achieves optimality when γ is not an even integer, and the second one for even integer γ. They are derived from two estimators of $N_\gamma(\theta)$ that we are going to define now.

We first present the estimator of $N_\gamma(\theta)$ that will be used when γ is not an even integer. For any positive integer K, we denote by $P_{\gamma,K}(\cdot)$ the best approximation of $|x|^\gamma$ by polynomials of degree at most $2K$ on the interval $[-1, 1]$, that is

$$\max_{x\in[-1,1]} \big||x|^\gamma - P_{\gamma,K}(x)\big| = \min_{G\in\mathcal{P}_{2K}} \max_{x\in[-1,1]} \big||x|^\gamma - G(x)\big|,$$

where $\mathcal{P}_{2K}$ is the class of all real polynomials of degree at most $2K$. Since $|x|^\gamma$ is an even function, it suffices to consider approximation by polynomials of even degree. The quality of the best polynomial approximation of $|x|^\gamma$ is described by Lemma 7 in Section 6.

We denote by $a_{\gamma,2k}$ the coefficients of the canonical representation of $P_{\gamma,K}$:

$$P_{\gamma,K}(x) = \sum_{k=0}^{K} a_{\gamma,2k} x^{2k}, \quad x \in \mathbb{R},$$

and by $H_k(\cdot)$ the kth Hermite polynomial

$$H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2}, \quad k \in \mathbb{N}, x \in \mathbb{R}.$$

To construct our first estimator in the dense zone, we use the sample duplication device, that is, we transform y_i into randomized observations $y_{1,i}$, $y_{2,i}$ as follows. Let $z_1, \dots, z_d$ be i.i.d. random variables such that $z_i \sim \mathcal{N}(0, \varepsilon^2)$ and $z_1, \dots, z_d$ are independent of $y_1, \dots, y_d$. Set

$$y_{1,i} = y_i + z_i, \qquad y_{2,i} = y_i - z_i, \quad i = 1, \dots, d.$$

Then, $y_{1,i} \sim \mathcal{N}(\theta_i, \sigma^2)$, $y_{2,i} \sim \mathcal{N}(\theta_i, \sigma^2)$ for $i = 1, \dots, d$, where $\sigma^2 = 2\varepsilon^2$ and the random variables $(y_{1,1}, \dots, y_{1,d}, y_{2,1}, \dots, y_{2,d})$ are mutually independent.

Define the estimator of N_γ as follows:

$$\hat{N}_\gamma = \sum_{i=1}^{d} \xi_\gamma(y_{1,i}, y_{2,i}) \tag{5}$$

where

$$\xi_\gamma(u, v) = \sum_{l=0}^{L} \hat{P}_{\gamma,K_l,M_l}(u) \mathbb{1}_{\sigma t_{l-1} < |v| \le \sigma t_l} + |u|^\gamma \mathbb{1}_{|v| > \sigma t_L},$$

and

$$\begin{cases} \hat{P}_{\gamma,K,M}(u) = \displaystyle\sum_{k=1}^{K} \sigma^{2k} a_{\gamma,2k} M^{\gamma-2k} H_{2k}(u/\sigma), \\ K_l = 4^l c \log\bigl(s^2/d\bigr), \\ M_l = 2^{l+1} \sigma \sqrt{2\log\bigl(s^2/d\bigr)}, \\ t_l = 2^l \sqrt{2\log\bigl(s^2/d\bigr)}, \qquad t_{-1} = 0, \\ L \text{ is the smallest integer such that } 2^L \ge 3\sqrt{\log(d)/\log\bigl(s^2/d\bigr)}. \end{cases} \tag{6}$$

Here and in what follows $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function, and $c > 0$ is a constant that will be chosen small enough (see the proof of Theorem 1 below).

The next theorem provides an upper bound on the risk of $\hat{N}_\gamma$ as estimator of $N_\gamma(\theta)$ in the dense zone.

Theorem 1. *Let the integers d and $s \in \{1, \dots, d\}$ be such that $s^2 \geq 4d$ and let $\gamma > 0$. Then for any $\theta \in B_0(s)$ the estimator defined in (5) satisfies*

$$\mathbf{E}_\theta\big[\big(\hat{N}_\gamma - N_\gamma(\theta)\big)^2\big] \leq C\left(\frac{\varepsilon^{2\gamma} s^2}{\log^\gamma(s^2/d)} + \frac{\varepsilon^2 s^{2/\gamma}}{\log(s^2/d)} \|\theta\|_\gamma^{2\gamma-2} \mathbb{1}_{\gamma>1}\right)$$

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

Inspection of the proof in Section 4.1 shows that the block structure of the estimator is needed to retrieve the sharp logarithmic factor $\log^\gamma(s^2/d)$ in the rate for all $s \geq 2\sqrt{d}$. If s is substantially greater than $\sqrt{d}$, for example, $d^a < s < d$ for some $a > 1/2$, then this logarithmic factor is equivalent to $\log^\gamma(d)$, and it is enough to use the estimator with only two blocks in order to obtain the result.

Although Theorem 1 is valid for all $\gamma > 0$ it will be useful for us only when γ is not an even integer since there exist estimators achieving better rates in the dense zone for even integers γ. We now provide a construction of such an estimator.

Indeed, assume more generally that γ is an integer, not necessarily even. We now use the sample duplication device as above but instead of creating two independent randomized samples, we create γ independent randomized samples $(y_{i,m}, 1, \dots, d)$, $m = 1, \dots, \gamma$, with variance multiplied by γ:

$$y_{i,m} = \theta_i + \varepsilon\sqrt{\gamma}\xi_{i,m},$$

where $\xi_{i,m}$ are i.i.d. standard normal random variables (see Nemirovski [11] for details).

We can now estimate the value $\sum_{i=1}^d \theta_i^\gamma$ by

$$\tilde{N}_\gamma = \sum_{i=1}^d \prod_{m=1}^\gamma y_{i,m}. \tag{7}$$

Since $\mathbf{E}(\prod_{m=1}^\gamma y_{i,m}) = \theta_i^\gamma$ we find immediately that $\tilde{N}_\gamma$ is an unbiased estimator of $\sum_{i=1}^d \theta_i^\gamma$:

$$\mathbf{E}\left(\sum_{i=1}^d \prod_{m=1}^\gamma y_{i,m}\right) = \sum_{i=1}^d \theta_i^\gamma.$$

If γ is an even integer, $\sum_{i=1}^d \theta_i^\gamma = N_\gamma(\theta)$. The risk of $\tilde{N}_\gamma$ admits the following bound.

Theorem 2. *Let γ be an integer. Then, for any $\theta \in \mathbb{R}^d$ we have*

$$\mathbf{E}_\theta\left[\left(\tilde{N}_\gamma - \sum_{i=1}^d \theta_i^\gamma\right)^2\right] \leq C\big(\varepsilon^{2\gamma} d + \varepsilon^2 \|\theta\|_{2\gamma-2}^{2\gamma-2}\big)$$

where $C > 0$ is a constant depending only on γ. In particular, if γ is an even integer we have here $\sum_{i=1}^d \theta_i^\gamma = N_\gamma(\theta)$.

Note that this theorem is valid for any sparsity s but we will use it only in the dense zone since in the sparse zone there exist better estimators achieving the optimal rate, cf. Section 2.2 below.

As a consequence of Theorems 1 and 2, we obtain the following result for estimation of the norm $\|\theta\|_\gamma$.

Theorem 3. (i) *Let the integers d and $s \in \{1, \dots, d\}$ be such that $s^2 \geq 4d$ and let $\gamma > 1$. Set $\hat{n}_\gamma = |\hat{N}_\gamma|^{1/\gamma}$, where $\hat{N}_\gamma$ is defined in (5). Then*

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta \big[\big(\hat{n}_\gamma - \|\theta\|_\gamma \big)^2 \big] \leq C \frac{\varepsilon^2 s^{2/\gamma}}{\log(s^2/d)}, \tag{8}$$

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

(ii) *Let γ be an even integer. Set $\tilde{n}_\gamma = |\tilde{N}_\gamma|^{1/\gamma}$, where $\tilde{N}_\gamma$ is defined in (7). Then*

$$\sup_{\theta \in \mathbb{R}^d} \mathbf{E}_\theta \big[\big(\tilde{n}_\gamma - \|\theta\|_\gamma \big)^2 \big] \leq C \varepsilon^2 d^{1/\gamma}, \tag{9}$$

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

We will prove below that the second bound of Theorem 3 cannot be improved in a minimax sense, and that the first is optimal up to a possible logarithmic factor. Note that, in the dense zone $s^2 \geq 4d$ considered in Theorem 3(i), the right-hand side of (9) is of smaller order than the right-hand side of (8). This privileged position of even powers γ can be explained by the fact that the even power functionals $N_\gamma(\theta)$ admit unbiased estimators converging with much faster rates than the estimators for other values of γ, for which the functionals $N_\gamma(\theta)$ are not smooth.

2.2. Sparse zone: $s^2 \leq 4d$

If s belongs to the sparse zone $s^2 \leq 4d$, we use the estimator

$$\hat{N}^*_\gamma = \sum_{i=1}^d \big\{ |y_i|^\gamma - \varepsilon^\gamma \alpha_\gamma \big\} \mathbb{1}_{y_i^2 > 2\varepsilon^2 \log(1 + d/s^2)}, \tag{10}$$

where

$$\alpha_\gamma = \frac{\mathbf{E}(|\xi|^\gamma \mathbb{1}_{\xi^2 > 2\log(1+d/s^2)})}{\mathbf{P}(\xi^2 > 2\log(1 + d/s^2))} \quad \text{for } \xi \sim \mathcal{N}(0, 1).$$

The next theorem establishes an upper bound on the risk of this estimator.

Theorem 4. *Let the integers d and $s \in \{1, \dots, d\}$ be such that $s^2 \leq 4d$ and $\gamma > 0$. Then for any $\theta \in B_0(s)$ the estimator defined in (10) satisfies*

$$\mathbf{E}_\theta \big[\big(\hat{N}^*_\gamma - N_\gamma(\theta) \big)^2 \big] \leq C \big(\varepsilon^{2\gamma} s^2 \log^\gamma \big(1 + d/s^2 \big) + \varepsilon^2 s^{2/\gamma} \|\theta\|_\gamma^{2\gamma - 2} \mathbb{1}_{\gamma > 1} \big),$$

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

Note that the estimator $\hat{N}_\gamma^*$ can be viewed as an example of applying the following routine developed in Collier et al. [3]. We start from the direct estimator $\sum_{i=1}^d |y_i|^\gamma$ and we threshold every term in this sum. This estimator being biased, we center every term by its mean under the assumption that there is no signal. Finally, we choose the value of the threshold that makes the best compromise between the first and second type errors in the support estimation problem. As opposed to the dense zone, we do not invoke the polynomial approximation. In fact, one can notice that the polynomial approximation is only useful in a neighborhood of 0 but in the sparse zone we renounce estimating small instances of θ_i.

Finally, we derive a consequence of Theorem 4 for estimation of the functional $\|\theta\|_\gamma$.

Theorem 5. *Let the integers d and $s \in \{1, \dots, d\}$ be such that $s^2 \le 4d$ and $\gamma > 1$. Set $\hat{n}_\gamma^* = |\hat{N}_\gamma^*|^{1/\gamma}$, where $\hat{N}_\gamma^*$ is defined in (10). Then*

$$\sup_{\theta \in B_0(s)} \mathbf{E}_\theta \big[\big(\hat{n}_\gamma^* - \|\theta\|_\gamma \big)^2 \big] \le C \varepsilon^2 s^{2/\gamma} \log\big(1 + d/s^2\big)$$

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

3. Lower bounds

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

Theorem 6. *Assume that $\gamma > 0$. Let s, d be integers such that $s \in \{1, \dots, d\}$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. There exist positive constants c_1 and c_2 depending only on γ and $\ell(\cdot)$ such that, for $\phi = \varepsilon^\gamma s \log^{\frac{\gamma}{2}}(1 + d/s^2)$,*

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s) \cap \{\|\theta\|_\gamma \le \phi^{1/\gamma}\}} \mathbf{E}_\theta \ell\big(c_1 \phi^{-1} \big| \hat{T} - N_\gamma(\theta) \big|\big) \ge c_2,$$

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

The proof is omitted since it follows the lines of the proof of the lower bound in Collier et al. [3], Theorem 1, with the only difference that $L(\theta) = \sum_{i=1}^d \theta_i$ should be replaced by $\sum_{i=1}^d \theta_i^\gamma$. The fact that the result is valid not only over $B_0(s)$ but also over the intersection of $B_0(s)$ with $B_\gamma := \{\|\theta\|_\gamma \le \phi^{1/\gamma}\}$ is granted since the support of the prior measure used in the proof of the lower bound in Collier et al. [3], Theorem 1, is included in B_γ for any $\gamma > 0$.

As a corollary of Theorem 6, we obtain the following lower bound.

Theorem 7. *Assume that $\gamma > 1$. Let s, d be integers such that $s \in \{1, \dots, d\}$ and let $\ell(\cdot)$ be any loss function in the class $\mathcal{L}$. There exist positive constants c_1 and c_2 depending only on γ and*

$\ell(\cdot)$ *such that*

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{E}_\theta \ell\big(c_1\big(\varepsilon s^{1/\gamma} \log^{1/2}\big(1 + d/s^2\big)\big)^{-1} \big|\hat{T} - \|\theta\|_\gamma\big|\big) \geq c_2,$$

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

Although Theorems 6 and 7 are valid with no restriction on $s \in \{1, \dots, d\}$, they yield suboptimal bounds in the dense zone. We now turn to the minimax lower bounds with better rates in the dense zone. We state them in the next three theorems of this section.

Theorem 8. *Assume that* $0 < \gamma \leq 1$. *Let* s, d *be integers such that* $s \in \{1, \dots, d\}$ *and let* $\ell(\cdot)$ *be any loss function in the class* $\mathcal{L}$. *There exist positive constants* c_1, c_2 *and* c_3 *depending only on* γ *and* $\ell(\cdot)$ *and a constant* $\bar{C} \geq 4$ *depending only on* γ *such that, if* $s^2 \geq \bar{C}d$ *and* $\phi = c_3 \varepsilon^\gamma s \log^{-\frac{\gamma}{2}}(s^2/d)$, *then*

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{E}_\theta \ell\big(c_1 \phi^{-1} \big|\hat{T} - N_\gamma(\theta)\big|\big) \geq c_2,$$

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

We are now in a position to derive the result on the minimax rate for $0 < \gamma \leq 1$ announced in (2). It is not hard to see that it follows from Theorems 1, 4, 6 and 8 with $\ell(u) = u^2$.

Next, the minimaxity of the rate in the first line of (3) is granted by Theorems 5 and 7 while the second line of (3) follows from Theorem 3(ii) and the next lower bound.

Theorem 9. *Assume that* γ *is an even integer. Let* s, d *be integers such that* $s \in \{1, \dots, d\}$ *and* $s \geq \sqrt{d}$. *Let* $\ell(\cdot)$ *be any loss function in the class* $\mathcal{L}$. *Then there exist positive constants* c_1 *and* c_2 *depending only on* γ *and* $\ell(\cdot)$ *such that*

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{E}_\theta \ell\big(c_1\big(\varepsilon d^{\frac{1}{2\gamma}}\big)^{-1} \big|\hat{T} - \|\theta\|_\gamma\big|\big) \geq c_2 \tag{11}$$

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

In conclusion, we have the following corollary.

Corollary 1. *The minimax risk on* $B_0(s)$ *with loss function* $\ell(u) = u^2$ *satisfies* (2) *and* (3).

Finally, we deduce (4) from Theorem 3(i) and the following lower bound.

Theorem 10. *Assume that* $\gamma > 1$ *is not an even integer. Let* s, d *be integers such that* $s \in \{1, \dots, d\}$ *and let* $\ell(\cdot)$ *be any loss function in the class* $\mathcal{L}$. *There exist positive constants* c_1 *and* c_2 *depending only on* γ *and* $\ell(\cdot)$ *and a constant* $\bar{C} \geq 4$ *depending only on* γ *such that, if* $s^2 \geq \bar{C}d$,

then

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{E}_\theta \ell\left(c_1 \left(\frac{\varepsilon s^{1/\gamma}}{\log^{\gamma-1/2}(s^2/d)} \right)^{-1} \left| \hat{T} - \|\theta\|_\gamma \right| \right) \geq c_2 \tag{12}$$

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

4. Proofs of the upper bounds

Throughout the proofs, we denote by C positive constants that can depend only on γ and may take different values on different appearances.

4.1. Proof of Theorem 1

Denote by S the support of θ. We start with a bias-variance decomposition

$$\begin{aligned}
\left(\hat{N}_\gamma - N_\gamma(\theta)\right)^2 \leq\ & 4\left(\sum_{i \in S} \mathbf{E}_\theta \xi_\gamma(y_{1,i}, y_{2,i}) - \sum_{i \in S} |\theta_i|^\gamma \right)^2 \\
& + 4\left(\sum_{i \in S} \xi_\gamma(y_{1,i}, y_{2,i}) - \sum_{i \in S} \mathbf{E}_\theta \xi_\gamma(y_{1,i}, y_{2,i}) \right)^2 \\
& + 4\left(\sum_{i \notin S} \mathbf{E}_\theta \xi_\gamma(y_{1,i}, y_{2,i}) \right)^2 \\
& + 4\left(\sum_{i \notin S} \xi_\gamma(y_{1,i}, y_{2,i}) - \sum_{i \notin S} \mathbf{E}_\theta \xi_\gamma(y_{1,i}, y_{2,i}) \right)^2
\end{aligned}$$

leading to the bound

$$\begin{aligned}
\mathbf{E}_\theta\left[\left(\hat{N}_\gamma - N_\gamma(\theta)\right)^2\right] \leq\ & 4\left(\sum_{i \in S} B_i \right)^2 + 4 \sum_{i \in S} V_i \\
& + 4d^2 \max_{i \notin S} B_i^2 + 4d \max_{i \notin S} V_i,
\end{aligned} \tag{13}$$

where $B_i = \mathbf{E}_\theta \xi_\gamma(y_{1,i}, y_{2,i}) - |\theta_i|^\gamma$ is the bias of $\xi_\gamma(y_{1,i}, y_{2,i})$ as an estimator of $|\theta_i|^\gamma$ and $V_i = \mathbf{Var}_\theta(\xi_\gamma(y_{1,i}, y_{2,i}))$ is its variance. We now bound separately the four terms in (13). We will show that the first two terms are smaller than

$$C\left(\sigma^{2\gamma} s^2 \log^{-\gamma}\left(s^2/d\right) + \frac{\sigma^2}{\log(s^2/d)} \left(\sum_{i \in S} |\theta_i|^{\gamma-1} \right)^2 \mathbb{1}_{\gamma>1} + \sigma^2 \sum_{i=1}^{d} |\theta_i|^{2\gamma-2} \mathbb{1}_{\gamma>1} \right) \tag{14}$$

while the last two terms are smaller than $C\sigma^{2\gamma}s^2\log^{-\gamma}(s^2/d)$. This proves the theorem since, by Hölder inequality, for any $\theta \in B_0(s)$ and $\gamma > 1$ we have

$$\left(\sum_{i\in S}|\theta_i|^{\gamma-1}\right)^2 \le s^{2/\gamma}\|\theta\|_\gamma^{2\gamma-2}, \tag{15}$$

and

$$\sum_{i=1}^d |\theta_i|^{2\gamma-2} \le s^{1/\gamma}\left(\sum_{i=1}^d |\theta_i|^{2\gamma}\right)^{\frac{\gamma-1}{\gamma}} \le s^{1/\gamma}\|\theta\|_\gamma^{2\gamma-2}. \tag{16}$$

1°. *Bias for $i \notin S$.* For $i \notin S$ using Lemma 2, we obtain

$$|B_i| = \sigma^\gamma \mathbf{E}|\xi|^\gamma \mathbf{P}\big(|\xi| > t_L\big) \le C\sigma^\gamma e^{-t_L^2/2}, \quad \xi \sim \mathcal{N}(0,1).$$

The last exponential is smaller than $1/d$ by the definition of t_L, so that

$$d^2 \max_{i\notin S} B_i^2 \le C\sigma^{2\gamma}d \le C\frac{\sigma^{2\gamma}s^2}{\log^\gamma(s^2/d)}. \tag{17}$$

2°. *Variance for $i \notin S$.* If $i \notin S$, then

$$V_i \le \sum_{l=0}^L \mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(\sigma\xi)\mathbf{P}\big(|\xi| > t_{l-1}\big) + \sigma^{2\gamma}\mathbf{E}|\xi|^{2\gamma}\mathbf{P}\big(|\xi| > t_L\big), \quad \xi \sim \mathcal{N}(0,1). \tag{18}$$

The last term in (18) is bounded from above as in item 1°. Next, in view of Lemma 3,

$$\begin{aligned}\mathbf{E}\hat{P}^2_{\gamma,K_0,M_0}(\sigma\xi) &\le C\sigma^{2\gamma}\frac{6^{2K_0}}{(M_0/\sigma)^{4-2\gamma}} \le C\sigma^{2\gamma}\log^\gamma\big(s^2/d\big)\left(\frac{s^2}{d}\right)^{2c\log 6}\\ &\le C\sigma^{2\gamma}\log^\gamma\big(s^2/d\big)\sqrt{\frac{s^2}{d}}\end{aligned}$$

if c is chosen such that $2c\log 6 \le 1/2$. Here, we use the assumption $s^2 \ge 4d$. For $l \ge 1$, we use Lemma 3 to obtain

$$\begin{aligned}\mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(\sigma\xi)\mathbf{P}\big(|\xi| > t_{l-1}\big) &\le C\sigma^{2\gamma}\frac{6^{2K_l}e^{-t_{l-1}^2/2}}{(M_l/\sigma)^{4-2\gamma}} \le C\sigma^{2\gamma}4^{\gamma l}\log^\gamma\big(s^2/d\big)\left(\frac{s^2}{d}\right)^{(2c\log 6-1/4)4^l}\\ &\le C\sigma^{2\gamma}4^{\gamma l}\log^\gamma\big(s^2/d\big)\left(\frac{s^2}{d}\right)^{-4^l/8}\end{aligned}$$

if we chose c such that $2c\log 6 \le 1/8$. In conclusion, under this choice of c, using the fact that $s^2 \ge 4d$, we get

$$d\max_{i\notin S} V_i \le C\sigma^{2\gamma} d\log^{\gamma}\left(s^2/d\right)\sqrt{\frac{s^2}{d}} \le \frac{C\sigma^{2\gamma}s^2}{\log^{\gamma}(s^2/d)}. \tag{19}$$

3°. *Bias for $i \in S$.* If $i \in S$, the bias has the form

$$B_i = \sum_{l=0}^{L} \mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X)\mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big) + \mathbf{E}|X|^{\gamma}\mathbf{P}\big(|X| > \sigma t_L\big) - |\theta_i|^{\gamma},$$

where $X \sim \mathcal{N}(\theta_i, \sigma^2)$. We will analyze this expression separately in three different ranges of values of $|\theta_i|$.

3.1°. *Case* $0 < |\theta_i| < 2\sigma t_0$. In this case, we use the bound

$$|B_i| \le \max_{l}\big|\mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X) - |\theta_i|^{\gamma}\big| + \big|\mathbf{E}|X|^{\gamma} - |\theta_i|^{\gamma}\big|\mathbf{P}\big(|X| > \sigma t_L\big),$$

where $X \sim \mathcal{N}(\theta_i, \sigma^2)$. Since $|\theta_i| \le M_l$ for all l, we can use Lemma 4 to obtain

$$\big|\mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X) - |\theta_i|^{\gamma}\big| \le C\left(\frac{M_l}{K_l}\right)^{\gamma} \le \frac{C\sigma^{\gamma}}{\log^{\gamma/2}(s^2/d)}. \tag{20}$$

In addition, using Lemma 1 and the inequalities $t_L > 3t_0 \ge 3|\theta_i|/(2\sigma)$, $3\sqrt{2\log(d)} \le t_L \le 6\sqrt{2\log(d)}$ we get

$$\begin{aligned}\big|\mathbf{E}|X|^{\gamma} - |\theta_i|^{\gamma}\big|\mathbf{P}\big(|X| > \sigma t_L\big) &\le C\big(\sigma^{\gamma} + \sigma^2|\theta_i|^{\gamma-2}\mathbb{1}_{|\theta_i|>\sigma}\big)\mathbf{P}\big(|\xi| > t_L - |\theta_i|/\sigma\big)\\ &\le C\sigma^{\gamma}\big(1 + \big(\log^{\gamma/2} d\big)\mathbb{1}_{\gamma>2}\big)\mathbf{P}\big(|\xi| > t_L/3\big) \le \frac{C\sigma^{\gamma}}{\log^{\gamma/2}(s^2/d)}\end{aligned}$$

where $\xi \sim \mathcal{N}(0,1)$. It follows that

$$s^2 \max_{i:0<|\theta_i|<2\sigma t_0} B_i^2 \le \frac{C\sigma^{2\gamma}s^2}{\log^{\gamma}(s^2/d)}. \tag{21}$$

3.2°. *Case* $2\sigma t_0 < |\theta_i| \le 2\sigma t_L$. Let $l_0 \in \{1,\dots,L-1\}$ be the integer such that $\sigma t_{l_0} < |\theta_i| \le \sigma t_{l_0+1}$. We have

$$\begin{aligned}|B_i| \le{}& \sum_{l=0}^{l_0-1}\big|\mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X) - |\theta_i|^{\gamma}\big| \cdot \mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big)\\ &+ \max_{l\ge l_0}\big|\mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X) - |\theta_i|^{\gamma}\big| + \big|\mathbf{E}|X|^{\gamma} - |\theta_i|^{\gamma}\big|,\end{aligned} \tag{22}$$

where $X \sim \mathcal{N}(\theta_i, \sigma^2)$. Analogously to (20), we find

$$\max_{l \ge l_0} \left| \mathbf{E} \hat{P}_{\gamma, K_l, M_l}(X) - |\theta_i|^\gamma \right| \le \frac{C\sigma^\gamma}{\log^{\gamma/2}(s^2/d)}.$$

Next, Lemma 1 and the fact that $|\theta_i| > 2\sigma t_0 = 2\sigma\sqrt{2\log(s^2/d)}$ imply

$$\left| \mathbf{E}|X|^\gamma - |\theta_i|^\gamma \right| \le C\sigma^2 |\theta_i|^{\gamma-2} \le C\left(\frac{\sigma^\gamma}{\log^{\gamma/2}(s^2/d)} \mathbb{1}_{\gamma \le 1} + \frac{\sigma |\theta_i|^{\gamma-1}}{\sqrt{\log(s^2/d)}} \mathbb{1}_{\gamma > 1} \right). \tag{23}$$

Finally, we consider the first sum on the right-hand side of (22). Notice that

$$\mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big) \le e^{-\frac{\theta_i^2}{8\sigma^2}}, \quad l = 0, \dots, l_0 - 1,$$

since $|\theta_i| > \sigma t_{l_0} \ge 2\sigma t_l$ for $l < l_0$. Using these inequalities and Lemma 5 we get

$$\begin{aligned} \sum_{l=0}^{l_0-1} \left| \mathbf{E} \hat{P}_{\gamma, K_l, M_l}(X) \right| \cdot \mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big) &\le C\sigma^\gamma \sum_{l=0}^{l_0-1} 6^{K_l} K_l^{1+\gamma/2} e^{(c-1)\theta_i^2/(8\sigma^2)} \\ &\le C\sigma^\gamma \sum_{l=0}^{l_0-1} t_l^{2+\gamma} e^{(c\log 6 + c - 1) t_l^2/2}. \end{aligned}$$

Choose $c > 0$ such that $c\log 6 + c < 1/4$. As $t_l = 2^l\sqrt{2\log(s^2/d)}$, this yields

$$\begin{aligned} \sum_{l=0}^{l_0-1} \left| \mathbf{E} \hat{P}_{\gamma, K_l, M_l}(X) \right| \cdot \mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big) &\le C\sigma^\gamma e^{-(1/2)\log(s^2/d)} \\ &\le \frac{C\sigma^\gamma}{\log^{\gamma/2}(s^2/d)}. \end{aligned}$$

Furthermore,

$$\begin{aligned} \sum_{l=0}^{l_0-1} |\theta_i|^\gamma \mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big) &\le l_0 |\theta_i|^\gamma e^{-\frac{\theta_i^2}{8\sigma^2}} \\ &\le C \log\left(\frac{\theta_i^2}{2\sigma^2 \log(s^2/d)} \right) |\theta_i|^\gamma e^{-\frac{\theta_i^2}{8\sigma^2}} \\ &\le C\sigma^\gamma e^{-\frac{\theta_i^2}{16\sigma^2}}, \end{aligned} \tag{24}$$

where we have used that $|\theta_i| > \sigma t_{l_0} = \sigma 2^{l_0}\sqrt{2\log(s^2/d)}$. Since $l_0 \ge 1$, this also implies that (24) does not exceed

$$\frac{C\sigma^\gamma}{\log^{\gamma/2}(s^2/d)}.$$

Combining the above arguments yields

$$\Bigg(\sum_{i\in S:2\sigma t_0<|\theta_i|\le 2\sigma t_L} B_i\Bigg)^2 \le C\Bigg(\frac{\sigma^{2\gamma}s^2}{\log^{\gamma}(s^2/d)} + \frac{\sigma^2}{\log(s^2/d)}\Bigg(\sum_{i\in S}|\theta_i|^{\gamma-1}\Bigg)^2 \mathbb{1}_{\gamma>1}\Bigg). \tag{25}$$

3.3°. *Case* $|\theta_i| > 2\sigma t_L$. Recall that the bias B_i has the form

$$B_i = \sum_{l=0}^{L} \mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X)\mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big) + \mathbf{E}|X|^{\gamma}\mathbf{P}\big(|X| > \sigma t_L\big) - |\theta_i|^{\gamma},$$

where $X \sim \mathcal{N}(\theta_i, \sigma^2)$. Using Lemma 5 we get

$$\begin{aligned}\Bigg|\sum_{l=0}^{L} \mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X)\mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big)\Bigg| &\le \max_{l=0,\dots,L}\big|\mathbf{E}\hat{P}_{\gamma,K_l,M_l}(X)\big|\mathbf{P}\big(|X| \le \sigma t_L\big)\\ &\le C\sigma^{\gamma}6^{K_L}K_L^{1+\gamma/2}e^{c\theta_i^2/(8\sigma^2)}e^{-\theta_i^2/(8\sigma^2)}\\ &\le C\sigma^{\gamma}(\log d)^{1+\gamma/2}6^{9c\log d}e^{9(c-1)\log d}\end{aligned}$$

and the last upper bound is smaller than $C\sigma^{\gamma}\log^{-\gamma/2}(s^2/d)$ if $c > 0$ is small enough. On the other hand, using (23) we find that

$$\begin{aligned}\big|\mathbf{E}|X|^{\gamma}\mathbf{P}\big(|X| > \sigma t_L\big) - |\theta_i|^{\gamma}\big| &\le \big|\mathbf{E}|X|^{\gamma} - |\theta_i|^{\gamma}\big| + |\theta_i|^{\gamma}\mathbf{P}\big(|X| \le \sigma t_L\big)\\ &\le C\Bigg(\frac{\sigma^{\gamma}}{\log^{\gamma/2}(s^2/d)}\mathbb{1}_{\gamma\le 1} + \frac{\sigma|\theta_i|^{\gamma-1}}{\sqrt{\log(s^2/d)}}\mathbb{1}_{\gamma>1}\Bigg) + 2|\theta_i|^{\gamma}e^{-\frac{\theta_i^2}{8\sigma^2}}\\ &\le C\Bigg(\frac{\sigma^{\gamma}}{\log^{\gamma/2}(s^2/d)} + \frac{\sigma|\theta_i|^{\gamma-1}}{\sqrt{\log(s^2/d)}}\mathbb{1}_{\gamma>1}\Bigg).\end{aligned}$$

Finally, we get

$$\Bigg(\sum_{i\in S:|\theta_i|>2\sigma t_L} B_i\Bigg)^2 \le C\Bigg(\frac{\sigma^{2\gamma}s^2}{\log^{\gamma}(s^2/d)} + \frac{\sigma^2}{\log(s^2/d)}\Bigg(\sum_{i\in S}|\theta_i|^{\gamma-1}\Bigg)^2 \mathbb{1}_{\gamma>1}\Bigg). \tag{26}$$

4°. *Variance for* $i \in S$. We consider the same three cases as in item 3° above. For the first two cases, it suffices to use a coarse bound granting that, for all $i \in S$,

$$\begin{aligned}V_i &\le \mathbf{E}_{\theta}\big[\xi_{\gamma}^2(y_{1,i}, y_{2,i})\big]\\ &= \sum_{l=0}^{L}\mathbf{E}\hat{P}_{\gamma,K_l,M_l}^2(X)\mathbf{P}\big(\sigma t_{l-1} < |X| \le \sigma t_l\big) + \mathbf{E}|X|^{2\gamma}\mathbf{P}\big(|X| > \sigma t_L\big)\end{aligned} \tag{27}$$

where $X \sim \mathcal{N}(\theta_i, \sigma^2)$.

4.1°. *Case* $0 < |\theta_i| < 2\sigma t_0$. In this case, we deduce from (27) that

$$V_i \le \max_{l=0,\dots,L} \mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(X) + \mathbf{E}|X|^{2\gamma},$$

where $X \sim \mathcal{N}(\theta_i, \sigma^2)$. Lemma 4 and the fact that $\mathbf{E}|X|^{2\gamma} \le C(\sigma^{2\gamma} + \sigma^2|\theta_i|^{2\gamma-2} + |\theta_i|^{2\gamma})$ (cf. Lemma 1) imply

$$\begin{aligned} V_i &\le C\big(M_L^{2\gamma} 2^{8K_L} + \sigma^{2\gamma} + |\theta_i|^{2\gamma}\big) \\ &\le C\big(\sigma^{2\gamma}\log^{\gamma}(d) d^{72c\log 2} + \sigma^{2\gamma}\log^{\gamma}\big(s^2/d\big)\big). \end{aligned}$$

Hence, choosing $c > 0$ small enough and using the assumption $s \ge 2\sqrt{d}$, we conclude that

$$s \max_{i:0<|\theta_i|<2\sigma t_0} V_i \le \frac{C\sigma^{2\gamma}s^2}{\log^{\gamma}(s^2/d)}. \tag{28}$$

4.2°. *Case* $2\sigma t_0 < |\theta_i| \le 2\sigma t_L$. As in item 3.2° above, we denote by $l_0 \in \{1,\dots,L-1\}$ the integer such that $\sigma t_{l_0} < |\theta_i| \le \sigma t_{l_0+1}$. We deduce from (27) that

$$V_i \le \max_{l=0,\dots,l_0-1} \mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(X)\mathbf{P}\big(|X| \le \sigma t_{l_0-1}\big) + \max_{l=l_0,\dots,L} \mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(X) + \mathbf{E}|X|^{2\gamma},$$

where $X \sim \mathcal{N}(\theta_i, \sigma^2)$. The second and third terms on the right-hand side are controlled as in item 4.1°, with the only difference that now we have $\mathbf{E}|X|^{2\gamma} \le C(\sigma^{2\gamma} + |\theta_i|^{2\gamma}) \le C\sigma^{2\gamma}\log^{\gamma}(d)$. For the first term, we find using Lemma 5 that, for $X \sim \mathcal{N}(\theta_i, \sigma^2)$,

$$\begin{aligned} &\max_{l=0,\dots,l_0-1} \mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(X)\mathbf{P}\big(|X| \le \sigma t_{l_0-1}\big) \\ &\quad\le C\sigma^{2\gamma}\big[(\sigma/M_0)^{4-2\gamma} + (\sigma/M_{l_0-1})^{4-2\gamma}\big] 6^{2K_{l_0-1}} e^{c\log(1+4/c)\theta_i^2/(4\sigma^2)} e^{-\theta_i^2/(8\sigma^2)} \\ &\quad\le C\sigma^{2\gamma}\log^{\gamma}(d) e^{(c\log 6 + 4c\log(1+4/c) - 1/2)t^2_{l_0-1}}. \end{aligned} \tag{29}$$

Choosing $c > 0$ small enough allows us to obtain the desired bound

$$s \max_{i:2\sigma t_0<|\theta_i|\le 2\sigma t_L} V_i \le \frac{C\sigma^{2\gamma}s^2}{\log^{\gamma}(s^2/d)}. \tag{30}$$

4.3°. *Case* $|\theta_i| > 2\sigma t_L$. We first note that, for $X \sim \mathcal{N}(\theta_i, \sigma^2)$:

$$\begin{aligned} \mathbf{Var}\big(|y_{1,i}|^{\gamma}\mathbb{1}_{|y_{2,i}|>\sigma t_L}\big) &= \mathbf{P}\big(|X| > \sigma t_L\big)\big[\mathbf{Var}\big(|X|^{\gamma}\big) + \big(\mathbf{E}|X|^{\gamma}\big)^2\mathbf{P}\big(|X| \le \sigma t_L\big)\big] \\ &\le C\big[\sigma^2|\theta_i|^{2\gamma-2} + |\theta|_i^{2\gamma}\mathbf{P}\big(|X| \le \sigma t_L\big)\big], \end{aligned}$$

where we have used the inequalities $\mathbf{Var}(|X|^{\gamma}) \le C\sigma^2|\theta_i|^{2\gamma-2}$ and $(\mathbf{E}|X|^{\gamma})^2 \le \mathbf{E}|X|^{2\gamma} \le C(\sigma^2|\theta_i|^{2\gamma-2} + |\theta_i|^{2\gamma}) \le C|\theta_i|^{2\gamma}$ that are valid due to Lemma 1 and to the fact that $|\theta_i| > \sigma$.

Thus, we obtain

$$
\begin{aligned}
V_i &\le 2\mathbf{Var}\left(\sum_{l=0}^{L}\hat{P}_{\gamma,K_l,M_l}(y_{1,i})\mathbb{1}_{\sigma t_{l-1}<|y_{2,i}|\le\sigma t_l}\right)+2\mathbf{Var}\big(|y_{1,i}|^{\gamma}\mathbb{1}_{|y_{2,i}|>\sigma t_L}\big)\\
&\le 2\sum_{l=0}^{L}\mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(X)\mathbf{P}\big(\sigma t_{l-1}<|X|\le\sigma t_l\big)+C\big[\sigma^2|\theta_i|^{2\gamma-2}+|\theta_i|^{2\gamma}\mathbf{P}\big(|X|\le\sigma t_L\big)\big]\\
&\le C\Big(\max_{l=0,\dots,L}\mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(X)\mathbf{P}\big(|X|\le\sigma t_L\big)+\sigma^{2\gamma}+\sigma^2|\theta_i|^{2\gamma-2}\mathbb{1}_{\gamma>1}+|\theta_i|^{2\gamma}\mathbf{P}\big(|X|\le\sigma t_L\big)\Big)
\end{aligned}
$$

since for $0<\gamma\le 1$ we have $|\theta_i|^{2\gamma-2}\le\sigma^{2\gamma-2}$ due to the fact that $|\theta_i|>\sigma$. In the last display, the term $\max_{l=0,\dots,L}\mathbf{E}\hat{P}^2_{\gamma,K_l,M_l}(X)\mathbf{P}(|X|\le\sigma t_L)$ is controlled via an argument analogous to (29) while

$$
|\theta_i|^{2\gamma}\mathbf{P}\big(|X|\le\sigma t_L\big)\le|\theta_i|^{2\gamma}\mathbf{P}\big(|\xi|\ge|\theta_i|/(2\sigma)\big)\le 2|\theta_i|^{2\gamma}e^{-\frac{\theta_i^2}{8\sigma^2}}\le C\sigma^{2\gamma},\quad \xi\sim\mathcal{N}(0,1),
$$

due to the fact that $t_L<|\theta_i|/(2\sigma)$. This allows us to conclude that

$$
\sum_{i\in S:|\theta_i|>2\sigma t_L}V_i\le C\left(\frac{\sigma^{2\gamma}s^2}{\log^{\gamma}(s^2/d)}+\sigma^2\sum_{i=1}^{d}|\theta_i|^{2\gamma-2}\mathbb{1}_{\gamma>1}\right).\tag{31}
$$

The result of the theorem follows now from (13), (17), (19), (21), (25), (26), (28), (30), and (31).

4.2. Proof of Theorem 2

Set $\sigma_*=\varepsilon\sqrt{\gamma}$. Since $y_{i,m}$ are mutually independent with $\mathbf{E}_\theta[\prod_{m=1}^{\gamma}y_{i,m}]=\theta_i^{\gamma}$ we have

$$
\begin{aligned}
\mathbf{E}_\theta\left[\left(\prod_{m=1}^{\gamma}y_{i,m}-\theta_i^{\gamma}\right)^2\right]&=\mathbf{E}_\theta\left[\prod_{m=1}^{\gamma}y_{i,m}^2\right]-\theta_i^{2\gamma}=\big(\theta_i^2+\sigma_*^2\big)^{\gamma}-\theta_i^{2\gamma}\\
&=\sum_{j=1}^{\gamma}\binom{\gamma}{j}\theta_i^{2(\gamma-j)}\sigma_*^{2j}\le C\big(\sigma_*^2\theta_i^{2(\gamma-1)}+\sigma_*^{2\gamma}\big).
\end{aligned}
$$

The theorem follows from this inequality and the fact that

$$
\mathbf{E}_\theta\big[\big(\tilde{N}_\gamma-N_\gamma(\theta)\big)^2\big]=\mathbf{E}_\theta\left[\left(\sum_{i=1}^{d}\left\{\prod_{m=1}^{\gamma}y_{i,m}-\theta_i^{\gamma}\right\}\right)^2\right]=\sum_{i=1}^{d}\mathbf{E}_\theta\left[\left(\prod_{m=1}^{\gamma}y_{i,m}-\theta_i^{\gamma}\right)^2\right].
$$

4.3. Proof of Theorem 3

Proof of part (i). Set $\phi = \varepsilon^\gamma s \log^{-\gamma/2}(s^2/d)$. First, assume that $\|\theta\|_\gamma \geq \phi^{1/\gamma}$. Then, using the inequality $|a-b|\,|b|^{\gamma-1} \leq |a^\gamma - b^\gamma|$, $\forall a, b > 0$, and Theorem 1 we get

$$\begin{aligned}\mathbf{E}_\theta\big(\hat{n}^*_\gamma - \|\theta\|_\gamma\big)^2 &\leq \frac{\mathbf{E}_\theta(\hat{N}^*_\gamma - N_\gamma(\theta))^2}{\|\theta\|_\gamma^{2\gamma-2}} \\ &\leq C\left(\phi^2 + \frac{\varepsilon^2 s^{2/\gamma}}{\log(s^2/d)}\|\theta\|_\gamma^{2\gamma-2}\right)\big(\|\theta\|_\gamma^{2\gamma-2}\big)^{-1} \\ &\leq C\left(\phi^{2/\gamma} + \frac{\varepsilon^2 s^{2/\gamma}}{\log(s^2/d)}\right) \leq C\phi^{2/\gamma},\end{aligned}$$

which is the desired bound. Next, assume that $\|\theta\|_\gamma < \phi^{1/\gamma}$. Using the inequality $|a-b| \leq |a^\gamma - b^\gamma|^{1/\gamma}$, $\forall a, b > 0$, Jensen's inequality, and Theorem 1 we get

$$\begin{aligned}\mathbf{E}_\theta\big[\big(\hat{n}^*_\gamma - \|\theta\|_\gamma\big)^2\big] &\leq \mathbf{E}_\theta\big[\big|\hat{N}^*_\gamma - N_\gamma(\theta)\big|^{2/\gamma}\big] \\ &\leq C\left(\phi^2 + \frac{\varepsilon^2 s^{2/\gamma}}{\log(s^2/d)}\|\theta\|_\gamma^{2\gamma-2}\right)^{1/\gamma} \\ &\leq C\left(\phi^2 + \frac{\varepsilon^2 s^{2/\gamma}}{\log(s^2/d)}\phi^{2-2/\gamma}\right)^{1/\gamma} \leq C\phi^{2/\gamma}.\end{aligned}$$

Proof of part (ii). We follow the same lines as the proof of part (i) but now we set $\phi = \varepsilon^\gamma d^{1/2}$. If $\|\theta\|_\gamma \geq \phi^{1/\gamma}$ we use the inequality $|a-b|\,|b|^{\gamma-1} \leq |a^\gamma - b^\gamma|$, $\forall a, b > 0$, Theorem 2 and the fact that $\|\theta\|_{2\gamma-2} \leq \|\theta\|_\gamma$ for $\gamma \geq 2$ to obtain

$$\begin{aligned}\mathbf{E}_\theta\big(\hat{n}^*_\gamma - \|\theta\|_\gamma\big)^2 &\leq \frac{\mathbf{E}_\theta(\hat{N}^*_\gamma - N_\gamma(\theta))^2}{\|\theta\|_\gamma^{2\gamma-2}} \\ &\leq C\big(\phi^2 + \varepsilon^2\|\theta\|_{2\gamma-2}^{2\gamma-2}\big)\big(\|\theta\|_\gamma^{2\gamma-2}\big)^{-1} \leq C\big(\phi^{2/\gamma} + \varepsilon^2\big) \leq C\phi^{2/\gamma},\end{aligned}$$

which is the desired bound. On the other hand, if $\|\theta\|_\gamma < \phi^{1/\gamma}$ the inequality $|a-b| \leq |a^\gamma - b^\gamma|^{1/\gamma}$, $\forall a, b > 0$, Jensen's inequality, Theorem 2 and the fact that $\|\theta\|_{2\gamma-2} \leq \|\theta\|_\gamma$ for $\gamma \geq 2$ yield

$$\begin{aligned}\mathbf{E}_\theta\big[\big(\hat{n}^*_\gamma - \|\theta\|_\gamma\big)^2\big] &\leq \mathbf{E}_\theta\big[\big|\hat{N}^*_\gamma - N_\gamma(\theta)\big|^{2/\gamma}\big] \\ &\leq C\big(\phi^2 + \varepsilon^2\|\theta\|_{2\gamma-2}^{2\gamma-2}\big)^{1/\gamma} \leq C\phi^{2/\gamma}.\end{aligned}$$

4.4. Proof of Theorem 4

Denoting by S the support of θ we have

$$\hat{N}_\gamma^* - N_\gamma(\theta) = \sum_{i\in S}\{|y_i|^\gamma - \varepsilon^\gamma\alpha_\gamma - |\theta_i|^\gamma\} - \sum_{i\in S}\{|y_i|^\gamma - \varepsilon^\gamma\alpha_\gamma\}\mathbb{1}_{y_i^2\le 2\varepsilon^2\log(1+d/s^2)} + \sum_{i\notin S}\{|y_i|^\gamma - \varepsilon^\gamma\alpha_\gamma\}\mathbb{1}_{y_i^2>2\varepsilon^2\log(1+d/s^2)},$$

so that

$$\mathbf{E}_\theta\big[\big(\hat{N}_\gamma^* - N_\gamma(\theta)\big)^2\big] \le 4\mathbf{E}_\theta\bigg[\bigg(\sum_{i\in S}\{|y_i|^\gamma - |\theta_i|^\gamma\}\bigg)^2\bigg] + 2^{\gamma+2}\varepsilon^{2\gamma}s^2\log^\gamma\big(1+d/s^2\big) + 4\varepsilon^{2\gamma}s^2\alpha_\gamma^2 + 4d\varepsilon^{2\gamma}\mathbf{E}\big[\big(|\xi|^\gamma-\alpha_\gamma\big)^2\mathbb{1}_{\xi^2>2\log(1+d/s^2)}\big]$$

where $\xi\sim\mathcal{N}(0,1)$. Using Lemma 1, we get

$$\begin{aligned}
&\mathbf{E}_\theta\bigg[\bigg(\sum_{i\in S}\{|y_i|^\gamma - |\theta_i|^\gamma\}\bigg)^2\bigg]\\
&\quad= \sum_{i\in S}\mathbf{E}_\theta\big[\big(|y_i|^\gamma - |\theta_i|^\gamma\big)^2\big] + \sum_{i,j\in S, i\ne j}\big(\mathbf{E}_\theta|y_i|^\gamma - |\theta_i|^\gamma\big)\big(\mathbf{E}_\theta|y_j|^\gamma - |\theta_j|^\gamma\big)\\
&\quad\le C\bigg(\varepsilon^{2\gamma}s + \varepsilon^4\sum_{|\theta_i|>\varepsilon}|\theta_i|^{2\gamma-4}\bigg) + C\bigg(\varepsilon^{2\gamma}s^2 + \varepsilon^4\bigg(\sum_{|\theta_i|>\varepsilon}|\theta_i|^{\gamma-2}\bigg)^2\bigg)\\
&\quad\le C\bigg(\varepsilon^{2\gamma}s^2 + \varepsilon^2\sum_{i=1}^d|\theta_i|^{2\gamma-2}\mathbb{1}_{\gamma\ge1} + \varepsilon^2\bigg(\sum_{i=1}^d|\theta_i|^{\gamma-1}\bigg)^2\mathbb{1}_{\gamma\ge1}\bigg)\\
&\quad\le C\big(\varepsilon^{2\gamma}s^2 + \varepsilon^2 s^{2/\gamma}\|\theta\|_\gamma^{2\gamma-2}\mathbb{1}_{\gamma\ge1}\big)
\end{aligned}$$

where we have used (15) and (16). Next, we use the fact that, for $\xi\sim\mathcal{N}(0,1)$ and any $x>0$, $a\ge0$,

$$\mathbf{E}\big(|\xi|^a\mathbb{1}_{|\xi|>x}\big) \le Cx^{a-1}e^{-x^2/2},$$
$$\mathbf{P}\big(|\xi|>x\big) \ge C(1+x)^{-1}e^{-x^2/2},$$

where C depends only on a. Choosing $x=\sqrt{2\log(1+d/s^2)}\ge\sqrt{2\log(5)}$ (as $d\ge4s^2$), we obtain

$$\alpha_\gamma \le C\frac{x^{\gamma-1}e^{-x^2/2}}{x^{-1}e^{-x^2/2}} \le C\log^{\gamma/2}\big(1+d/s^2\big)$$

The same property implies that

$$\mathbf{E}\big[\big(|\xi|^\gamma - \alpha_\gamma\big)^2 \mathbb{1}_{\xi^2 > 2\log(1+d/s^2)}\big] \le 2\mathbf{E}\big(|\xi|^{2\gamma} \mathbb{1}_{\xi^2 > 2\log(1+d/s^2)}\big) + 2\alpha_\gamma^2 \mathbf{P}\big(\xi^2 > 2\log\big(1+d/s^2\big)\big)$$
$$\le C\frac{s^2}{d}\log^\gamma\big(1+d/s^2\big).$$

Combining the above inequalities proves the theorem.

4.5. Proof of Theorem 5

We act as in the proof of Theorem 3 with suitable modifications. Namely, set $\phi = \varepsilon^\gamma s \times \log^{\gamma/2}(s^2/d)$. If $\|\theta\|_\gamma \ge \phi^{1/\gamma}$ then using Theorem 4 we get

$$\mathbf{E}_\theta\big(\hat{n}_\gamma^* - \|\theta\|_\gamma\big)^2 \le \frac{\mathbf{E}_\theta(\hat{N}_\gamma^* - N_\gamma(\theta))^2}{\|\theta\|_\gamma^{2\gamma-2}} \le C\big(\phi^2 + \varepsilon^2 s^{2/\gamma}\|\theta\|_\gamma^{2\gamma-2}\big)\big(\|\theta\|_\gamma^{2\gamma-2}\big)^{-1}$$
$$\le C\big(\phi^{2/\gamma} + \varepsilon^2 s^{2/\gamma}\big) \le C\phi^{2/\gamma}.$$

On the other hand, if $\|\theta\|_\gamma < \phi^{1/\gamma}$ then using Theorem 4 we get

$$\mathbf{E}_\theta\big[\big(\hat{n}_\gamma^* - \|\theta\|_\gamma\big)^2\big] \le \mathbf{E}_\theta\big[\big|\hat{N}_\gamma^* - N_\gamma(\theta)\big|^{2/\gamma}\big] \le C\big(\phi^2 + \varepsilon^2 s^{2/\gamma}\|\theta\|_\gamma^{2\gamma-2}\big)^{1/\gamma}$$
$$\le C\big(\phi^2 + \varepsilon^2 s^{2/\gamma}\phi^{2-2/\gamma}\big)^{1/\gamma} \le C\phi^{2/\gamma}.$$

5. Lemmas for the proof of Theorem 1

Lemma 1. *If $X \sim \mathcal{N}(\vartheta, \sigma^2)$ and $\gamma > 0$, then*

$$\big|\mathbf{E}\big(|X|^\gamma\big) - |\vartheta|^\gamma\big| \le C\big(\sigma^\gamma \mathbb{1}_{|\vartheta|\le\sigma} + \sigma^2|\vartheta|^{\gamma-2}\mathbb{1}_{|\vartheta|>\sigma}\big),$$
$$\mathbf{Var}\big(|X|^\gamma\big) \le C\big(\sigma^{2\gamma}\mathbb{1}_{|\vartheta|\le\sigma} + \sigma^2|\vartheta|^{2\gamma-2}\mathbb{1}_{|\vartheta|>\sigma}\big).$$

Proof. Set for brevity

$$g(x) = |x|^\gamma, \qquad b_\gamma = \mathbf{E}\big(|X|^\gamma\big) - |\vartheta|^\gamma.$$

First, note that if $|\vartheta| \le \sigma$ we have $|b_\gamma| \le C\sigma^\gamma$. Now, consider the case $|\vartheta| > \sigma$. Then,

$$|b_\gamma| \le \frac{1}{\sqrt{2\pi}\sigma}\Bigg[\bigg|\int_{|x|>|\vartheta|/2}\big(g(x+\vartheta) - g(\vartheta)\big)e^{-\frac{x^2}{2\sigma^2}}\,dx\bigg|$$
$$+ \bigg|\int_{|x|\le|\vartheta|/2}\big(g(x+\vartheta) - g(\vartheta)\big)e^{-\frac{x^2}{2\sigma^2}}\,dx\bigg|\Bigg].$$

We now bound separately the two terms on the right-hand side of this inequality. Using the second order Taylor expansion of g around ϑ and the symmetry of the Gaussian distribution, we get

$$\begin{aligned}\left|\int_{|x|\le|\vartheta|/2}\bigl(g(x+\vartheta)-g(\vartheta)\bigr)e^{-\frac{x^2}{2\sigma^2}}\,dx\right| &\le \frac{1}{2}\int_{|x|\le|\vartheta|/2}\max_{|u|\le|\vartheta|/2}\bigl|g''(\vartheta+u)\bigr|x^2e^{-\frac{x^2}{2\sigma^2}}\,dx\\ &\le C|\vartheta|^{\gamma-2}\int_{|x|\le|\vartheta|/2}x^2e^{-\frac{x^2}{2\sigma^2}}\,dx \le C\sigma^3|\vartheta|^{\gamma-2}.\end{aligned}$$

On the other hand, the first integral in the bound for $|b_\gamma|$ is smaller than

$$\begin{aligned}&C\int_{|x|>|\vartheta|/2}\bigl\{|x|^\gamma+|\vartheta|^\gamma\bigr\}e^{-\frac{x^2}{2\sigma^2}}\,dx\\ &\quad\le C\sigma^{\gamma+1}\int_{|t|>|\vartheta|/(2\sigma)}|t|^\gamma e^{-\frac{t^2}{2}}dt + C\sigma|\vartheta|^\gamma\int_{|t|>|\vartheta|/(2\sigma)}e^{-\frac{t^2}{2}}dt\\ &\quad\le C\sigma^{\gamma+1}\frac{|\vartheta|^{\gamma-1}}{\sigma^{\gamma-1}}e^{-\frac{\vartheta^2}{8\sigma^2}} + C\sigma^2|\vartheta|^{\gamma-1}e^{-\frac{\vartheta^2}{8\sigma^2}}\\ &\quad\le C\sigma^2|\vartheta|^{\gamma-1}e^{-\frac{\vartheta^2}{8\sigma^2}}.\end{aligned}$$

Combining the above inequalities yields the desired bound for the bias. The bound on the variance follows immediately since

$$\mathbf{Var}\bigl(|X|^\gamma\bigr)=\mathbf{E}\bigl(|X|^{2\gamma}\bigr)-\bigl(\mathbf{E}|X|^\gamma\bigr)^2=b_{2\gamma}+|\vartheta|^{2\gamma}-\bigl[b_\gamma+|\vartheta|^\gamma\bigr]^2\le b_{2\gamma}. \qquad \square$$

Lemma 2. *Let $\vartheta\in\mathbb{R}$ and $X\sim\mathcal{N}(\vartheta,1)$. For any $k\in\mathbb{N}$, the k-th Hermite polynomial satisfies*

$$\begin{aligned}\mathbf{E}H_k(X)&=\vartheta^k,\\ \mathbf{E}H_k^2(X)&\le k^k\bigl(1+\vartheta^2/k\bigr)^k.\end{aligned}$$

The proof of this lemma can be found in Cai and Low [1].

Lemma 3. *Let $\hat{P}_{\gamma,K,M}$ be defined in (6) with parameters $K=K_l$ and $M=M_l$ for some $l\in\{0,\dots,L\}$ and small enough $c>0$. If $X\sim\mathcal{N}(0,\sigma^2)$, then*

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X)\le C\sigma^{2\gamma}\frac{6^{2K}}{(M/\sigma)^{4-2\gamma}},$$

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

Proof. Recall that, for the Hermite polynomials, $\mathbf{E}(H_k(\xi)H_j(\xi)) = 0$ if $k \neq j$ and $\xi \sim \mathcal{N}(0,1)$. Using this fact and then Lemmas 8 and 2, we obtain

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X) = M^{2\gamma}\sum_{k=1}^{K} a^2_{\gamma,2k}(\sigma/M)^{4k}\mathbf{E}H^2_{2k}(X/\sigma)$$

$$\leq C6^{2K}M^{2\gamma}\sum_{k=1}^{K}(2k)^{2k}(\sigma/M)^{4k}.$$

Moreover, since $\sigma^2/M^2 = c/(8K)$ we have

$$\sum_{k=1}^{K}(2k)^{2k}(\sigma/M)^{4k} \leq \frac{4\sigma^4}{M^4} + \sum_{2\leq k\leq \log(M/\sigma)}(\sigma/M)^{4k}\big(2\log(M/\sigma)\big)^{2k}$$

$$+ \sum_{\log(M/\sigma)<k\leq K}(c/4)^{2k} \leq \frac{C\sigma^4}{M^4} \tag{32}$$

if c is small enough. The result follows. □

Lemma 4. *Let $\hat{P}_{\gamma,K,M}$ be defined in (6) with parameters $K = K_l$ and $M = M_l$ for some $l \in \{0,\dots,L\}$ and small enough $c > 0$. If $X \sim \mathcal{N}(\vartheta,\sigma^2)$ with $|\vartheta| \leq M$, then*

$$\big|\mathbf{E}\hat{P}_{\gamma,K,M}(X) - |\vartheta|^\gamma\big| \leq C\left(\frac{M}{K}\right)^{\gamma},$$

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X) \leq CM^{2\gamma}2^{8K},$$

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

Proof. To prove the first inequality of the lemma, it is enough to note that, due to Lemma 2,

$$\mathbf{E}\hat{P}_{\gamma,K,M}(X) = \sum_{k=1}^{K} a_{\gamma,2k}M^{\gamma-2k}\vartheta^{2k} \tag{33}$$

and to apply Lemma 7. For the second inequality, we use the bound

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X) \leq M^{2\gamma}\left(\sum_{k=1}^{K}\sigma^{2k}|a_{\gamma,2k}|M^{-2k}\sqrt{\mathbf{E}H^2_{2k}(X/\sigma)}\right)^2. \tag{34}$$

Thus Lemmas 8 and 2 together with the relations $|\vartheta| \leq M$ and $K = (c/8)M^2/\sigma^2$ imply that, for small enough $c > 0$,

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X) \leq CM^{2\gamma}6^{2K}\left(\sum_{k=1}^{K}M^{-2k}\big(2M^2\big)^k\right)^2 \leq CM^{2\gamma}2^{8K}.$$

□

Lemma 5. *Let $\hat{P}_{\gamma,K,M}$ be defined in (6) with parameters $K = K_l$ and $M = M_l$ for some $l \in \{0, \dots, L\}$ and small enough $c > 0$. If $X \sim \mathcal{N}(\vartheta, \sigma^2)$ with $|\vartheta| > 2\sigma t_l$, then*

$$\left|\mathbf{E}\hat{P}_{\gamma,K,M}(X)\right| \le C\sigma^{\gamma} 6^K K^{1+\gamma/2} e^{c\vartheta^2/(8\sigma^2)},$$

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X) \le C\sigma^{2\gamma}(\sigma/M)^{4-2\gamma} 6^{2K} e^{c\log(1+4/c)\vartheta^2/(4\sigma^2)},$$

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

Proof. To prove the first inequality of the lemma, we use (33) and Lemma 8 to obtain

$$\left|\mathbf{E}\hat{P}_{\gamma,K,M}(X)\right| \le CM^{\gamma} K 6^K \left(\frac{\vartheta^2}{M^2}\right)^K.$$

Recall that $M^2 = 8\sigma^2 K/c$ and $|\vartheta| > M$ by assumption of the lemma. Thus,

$$M^{\gamma} K 6^K \left(\frac{\vartheta^2}{M^2}\right)^K \le C\sigma^{\gamma} K^{1+\gamma/2} 6^K e^{K\log(\vartheta^2/M^2)}$$

and the result follows since $K\log(\vartheta^2/M^2) = cM^2/8\sigma^2 \log(\vartheta^2/M^2) \le c\vartheta^2/8\sigma^2$.

We now prove the second inequality of the lemma. Using (34) and then Lemmas 8 and 2 we get

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X) \le CM^{2\gamma} 6^{2K} \left(\sum_{k=1}^{K} (\sigma/M)^{2k} (2k)^k \left(1 + \frac{\vartheta^2}{2\sigma^2 k}\right)^k\right)^2.$$

As $M^2 = 8\sigma^2 K/c$ and $|\vartheta| > M$, we have

$$\frac{\vartheta^2}{2\sigma^2 k} \ge \frac{M^2}{2\sigma^2 K} = \frac{4}{c} \ge 2$$

for $c > 0$ small enough. Using this remark and the fact that the function $x \to x^{-1}\log(1+x)$ is decreasing for $x \ge 2$ we obtain

$$k\log\left(1 + \frac{\vartheta^2}{2\sigma^2 k}\right) \le \frac{c\log(1+4/c)\vartheta^2}{8\sigma^2}.$$

Therefore,

$$\mathbf{E}\hat{P}^2_{\gamma,K,M}(X) \le CM^{2\gamma} 6^{2K} e^{c\log(1+4/c)\vartheta^2/(4\sigma^2)} \left(\sum_{k=1}^{K} (\sigma/M)^{2k} (2k)^k\right)^2.$$

Finally, the result follows by noticing that, by an argument analogous to (32), we have

$$\sum_{k=1}^{K} (\sigma/M)^{2k} (2k)^k \le \frac{C\sigma^2}{M^2}.$$

□

6. Some facts from approximation theory

We start with a proposition relating moment matching to best polynomial approximation. It is similar to several results used in the theory of estimation of non-smooth functionals starting from Lepski et al. [10]. There exist different techniques to prove such results for specific examples. Thus, the proof in Lepski et al. [10] is based on Riesz representation of linear operators, while Wu and Yang [15] provide an explicit construction using Lagrange interpolation. Here, for completeness we give a short proof for a relatively general setting based on optimization arguments.

Let $f : [-1, 1] \to \mathbb{R}$ be a continuous even function. Consider the accuracy of best polynomial approximation of f:

$$\delta_K(f) = \inf_{G \in \mathcal{P}_K} \max_{x \in [-1,1]} \big| f(x) - G(x) \big|$$

where $\mathcal{P}_K$ is the class of all real polynomials of degree at most K.

Proposition 1. *Let $f : [-1, 1] \to \mathbb{R}$ be a continuous even function. For any even integer $K \geq 1$, there exist two probability measures $\tilde{\mu}_0$ and $\tilde{\mu}_1$ on $[-1, 1]$ such that*

(i) *$\tilde{\mu}_0$ and $\tilde{\mu}_1$ are symmetric about 0;*
(ii) *$\int t^l \tilde{\mu}_0(dt) = \int t^l \tilde{\mu}_1(dt)$ for $l = 0, 1, \dots, K$;*
(iii) *$\int f(t)\tilde{\mu}_1(dt) - \int f(t)\tilde{\mu}_0(dt) = 2\delta_K(f)$.*

Proof. Denote by P_{sym} the set of all probability measures on $[-1, 1]$ that are symmetric about 0, and by P_2 be the set of all signed measures on $[-1, 1]$ with total variation not greater than 2. For $K = 2m$, we have

$$\begin{aligned}
&\sup_{(\nu_0,\nu_1) \in P_{\mathsf{sym}} \times P_{\mathsf{sym}} : \int t^l \, d\nu_0(t) = \int t^l \, d\nu_1(t), l=0,\dots,K} \left(\int_{-1}^{1} f(x)\, d\nu_0(x) - \int_{-1}^{1} f(x)\, d\nu_1(x) \right) \\
&\quad = \sup_{\mu \in P_2 : \int t^{2l} d\mu(t) = 0, l=0,\dots,m} \int_{-1}^{1} f(x)\, d\mu(x) \\
&\quad = \sup_{\mu \in P_2} \inf_{\alpha \in \mathbb{R}^{m+1}} \int_{-1}^{1} \left(f(x) - \sum_{l=0}^{m} \alpha_l x^{2l} \right) d\mu(x) \\
&\quad = \inf_{\alpha \in \mathbb{R}^{m+1}} \sup_{\mu \in P_2} \int_{-1}^{1} \left(f(x) - \sum_{l=0}^{m} \alpha_l x^{2l} \right) d\mu(x) \\
&\quad = 2 \min_{\alpha \in \mathbb{R}^{m+1}} \max_{x \in [-1,1]} \left| f(x) - \sum_{l=0}^{m} \alpha_l x^{2l} \right| = 2\delta_K(f),
\end{aligned} \tag{35}$$

where the third equality follows from Sion's minimax theorem, and the second equality uses the fact that f is an even function, so that the maximum over $\mu \in P_2$ in the second line of (35)

is equal to the maximum over symmetric $\mu \in P_2$ satisfying the same moment constraints. Let (ν_0^*, ν_1^*) be the pair of probability measures attaining the maximum in the first line of (35). The proposition follows by setting $\tilde{\mu}_i = \nu_i^*$, $i = 0, 1$. □

As an immediate corollary of Proposition 1 for $f(x) = |x|^\gamma$, we obtain the following result.

Lemma 6. *For any even integer $K \geq 1$ and any $M > 0$, $\gamma > 0$, there exist two probability measures $\tilde{\mu}_0$ and $\tilde{\mu}_1$ on $[-M, M]$ such that*

(i) *$\tilde{\mu}_0$ and $\tilde{\mu}_1$ are symmetric about 0;*
(ii) *$\int t^l \tilde{\mu}_0(dt) = \int t^l \tilde{\mu}_1(dt)$ for $l = 0, 1, \ldots, K$;*
(iii) *$\int |t|^\gamma \tilde{\mu}_1(dt) - \int |t|^\gamma \tilde{\mu}_0(dt) = 2M^\gamma \delta_{K,\gamma}$.*

For the function $f(x) = |x|^\gamma$, the asymptotically exact behavior of the best polynomial approximation $\delta_{K,\gamma}$ as $K \to \infty$ is well known, see, for example, Timan [13], Theorem 7.2.2, implying the following lemma.

Lemma 7. *If $\gamma > 0$ is not an even integer, then there exist positive constants c_* and C^* depending only on γ such that*

$$c_* K^{-\gamma} \leq \delta_{K,\gamma} \leq C^* K^{-\gamma}, \quad \forall K \in \mathbb{N}.$$

Finally, the next lemma provides a useful bound for the coefficients $a_{\gamma,2k}$ in the canonical representation of the best approximation polynomial:

$$P_{\gamma,K}(x) = \sum_{k=0}^{K} a_{\gamma,2k} x^{2k}, \quad x \in \mathbb{R}. \tag{36}$$

Lemma 8. *Let $P_{\gamma,K}(\cdot)$ be the best approximation polynomial of degree $2K$ for $|x|^\gamma$ on $[-1, 1]$. Then the coefficients $a_{\gamma,2k}$ in (36) satisfy*

$$|a_{\gamma,2k}| \leq C 6^K, \quad k = 0, \ldots, K,$$

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

This lemma is an immediate corollary of the following more general fact, which is a consequence of Szegö's theorem on the minimal eigenvalue of a lacunary version of the Hilbert matrix.

Proposition 2. *Let $P(x) = \sum_{k=0}^{N} a_k x^k$ be a polynomial such that $|P(x)| \leq 1$ for all $x \in [-1, 1]$. Then there exists an absolute constant $C > 0$ such that*

$$|a_k| \leq C(\sqrt{2} + 1)^N$$

for all $k \in \{0, \ldots, N\}$.

Proof. We have

$$\int_{-1}^{1}\left(\sum_{k=0}^{N} a_k x^k\right)^2 dx = 2\sum_{i,j=0}^{N}\frac{a_i a_j}{i+j+1}\mathbb{1}_{i+j \text{ even}}. \tag{37}$$

It is easy to see that the quadratic form in (37) is positive definite for all N. Furthermore, as shown by Szegö [12], the minimal eigenvalue $\lambda_{\min}(N)$ of this quadratic form satisfies

$$\lambda_{\min}(N) = 2^{9/4}\pi^{3/2}N^{1/2}(\sqrt{2}-1)^{2N+3}\big(1+o(1)\big) \quad \text{as } N\to\infty.$$

Therefore, there exists an absolute constant $C_0 > 0$ such that $\lambda_{\min}(N) \geq C_0(\sqrt{2}-1)^{2N}$ for all N. This inequality and (37) imply that

$$C_0(\sqrt{2}-1)^{2N}\sum_{k=0}^{N} a_k^2 \leq 1$$

and hence $\max_{k=0,\dots,N}|a_k| \leq C_0^{1/2}(\sqrt{2}-1)^{-N}$. □

7. Construction of the priors for the proof of Theorem 8

The proof of Theorem 8 will be based on Theorem 2.15 in Tsybakov [14]. It proceeds by bounding the minimax risk from below by the Bayes risk with the prior measures on θ that we are going to define in this section.

In what follows we set

$$\Lambda = \sqrt{\log\left(\frac{s^2}{d}\right)}, \qquad M = \varepsilon\Lambda, \tag{38}$$

and we denote by K the smallest even integer such that

$$K \geq \frac{3}{2}e\log\left(\frac{s^2}{d}\right) = \frac{3}{2}e\Lambda^2. \tag{39}$$

We will also write for brevity

$$B = B_0(s).$$

In what follows, unless stated otherwise, $\tilde{\mu}_0$ and $\tilde{\mu}_1$ are the probability measures satisfying Lemma 6 where M is defined in (38) and K is the smallest even integer for which (39) holds.

For $i = 0, 1$, the probability measure μ_i is defined as the distribution of random vector $\theta \in \mathbb{R}^d$ with components θ_j having the form $\theta_j = \epsilon_j\eta_j$, $j = 1,\dots,d$, where ϵ_j is a Bernoulli random

variable with $\mathbf{P}(\epsilon_j = 1) = s/(2d)$, η_j is distributed according to $\tilde{\mu}_i$, and $(\epsilon_1, \dots, \epsilon_d, \eta_1, \dots, \eta_d)$ are mutually independent.

Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be the mixture probability measures defined by the relation

$$\mathbb{P}_i(A) = \int_{\mathbb{R}^d} \mathbf{P}_\theta(A)\mu_i(d\theta), \quad i = 0, 1,$$

for any measurable set A. The densities of $\mathbb{P}_0$ and $\mathbb{P}_1$ with respect to the Lebesgue measure on $\mathbb{R}^d$ have the form

$$f_0(x) = \prod_{i=1}^{d} h(x_i) \quad \text{and} \quad f_1(x) = \prod_{i=1}^{d} g(x_i), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

respectively, where for $x \in \mathbb{R}$ we set

$$h(x) = \frac{s}{2d}\varphi_0(x) + \left(1 - \frac{s}{2d}\right)\varphi(x)$$

and

$$g(x) = \frac{s}{2d}\varphi_1(x) + \left(1 - \frac{s}{2d}\right)\varphi(x)$$

with

$$\varphi_i(x) = \int_{\mathbb{R}} \varphi(x - t)\tilde{\mu}_i(dt), \quad i = 0, 1, \tag{40}$$

where we denote by $\varphi(\cdot)$ the density of the $\mathcal{N}(0, \varepsilon^2)$ distribution.

Note that the measures μ_0 and μ_1 are not supported in B. We associate to them two probability measures $\mu_{0,B}$ and $\mu_{1,B}$ supported in B and the corresponding mixture measures defined by

$$\mu_{i,B}(A) = \frac{\mu_i(A \cap B)}{\mu_i(B)}, \qquad \mathbb{P}_{i,B}(A) = \int_{\mathbb{R}^d} \mathbf{P}_\theta(A)\mu_{i,B}(d\theta), \quad i = 0, 1,$$

for any measurable set A.

8. Proof of Theorem 8

Since we have $\ell(t) \geq \ell(a)\mathbb{1}_{t>a}$ for any $a > 0$, it is enough to prove the theorem for the indicator loss $\ell(t) = \mathbb{1}_{t>a}$. Introduce the following notation:

$$m_i = \int_{\mathbb{R}^d} N_\gamma(\theta)\mu_i(d\theta), \qquad v_i^2 = \int_{\mathbb{R}^d} \big(N_\gamma(\theta) - m_i\big)^2 \mu_i(d\theta), \quad i = 0, 1.$$

Note that Lemmas 6 and 7 imply:

$$\begin{aligned} m_1 - m_0 &= d\left(\int_{\mathbb{R}^d} |\theta_1|^\gamma \mu_1(d\theta) - \int_{\mathbb{R}^d} |\theta_1|^\gamma \mu_0(d\theta)\right) \\ &= \frac{s}{2}\left(\int_{-M}^{M} |t|^\gamma \tilde{\mu}_1(dt) - \int_{-M}^{M} |t|^\gamma \tilde{\mu}_0(dt)\right) \\ &= sM^\gamma \delta_{K,\gamma} \geq c_* s(M/K)^\gamma \geq C_1 \frac{\varepsilon^\gamma s}{\Lambda^\gamma}, \end{aligned} \tag{41}$$

where $C_1 > 0$ is a constant depending only on γ.

Let $V(P, Q)$ denote the total variation distance between two probability measures P and Q. For any $u > 0$ and any $c \in \mathbb{R}$ we have, using Theorem 2.15 in Tsybakov [14],

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{P}_\theta\big(\big|\hat{T} - N_\gamma(\theta)\big| \geq u\big) \geq \frac{1 - V'}{2}, \tag{42}$$

where

$$V' = V(\mathbb{P}_{0,B}, \mathbb{P}_{1,B}) + \mu_{0,B}\big(N_\gamma(\theta) \geq c\big) + \mu_{1,B}\big(N_\gamma(\theta) \leq c + 2u\big).$$

We now apply (42) with the parameters

$$c = m_0 + 3v_0, \qquad u = \frac{m_1 - m_0}{4}.$$

By Chebyshev–Cantelli inequality,

$$\mu_0\big(N_\gamma(\theta) \geq c\big) \leq \frac{v_0^2}{v_0^2 + (c - m_0)^2} = \frac{1}{10}. \tag{43}$$

Next, since the measures $\tilde{\mu}_0$ and $\tilde{\mu}_0$ are supported in $[-M, M]$,

$$\max\big(v_0^2, v_1^2\big) \leq dM^{2\gamma} = d\varepsilon^{2\gamma}\Lambda^{2\gamma}. \tag{44}$$

Thus, we may write

$$\max(v_0, v_1) \leq \left(\frac{\sqrt{d}}{s}\Lambda^{2\gamma}\right)\frac{\varepsilon^\gamma s}{\Lambda^\gamma},$$

where, for $\bar{C}$ large enough, $\frac{\sqrt{d}}{s}\Lambda^{2\gamma} = \frac{\sqrt{d}}{s}\log^\gamma(\frac{s^2}{d}) \leq C_1/12$ (recall that $s^2 \geq \bar{C}d$ by assumption). Therefore,

$$\max(v_0, v_1) \leq \frac{C_1}{12}\frac{\varepsilon^\gamma s}{\Lambda^\gamma}. \tag{45}$$

It follows from (41), (45) and Chebyshev–Cantelli inequality that

$$\begin{aligned}\mu_1\big(N_\gamma(\theta)\le c+2u\big)&=\mu_1\left(N_\gamma(\theta)-m_1\le-\frac{m_1+m_0}{2}+3v_0\right)\\&\le\mu_1\left(N_\gamma(\theta)-m_1\le-\frac{m_1-m_0}{2}+3v_0\right)\\&\le\mu_1\left(N_\gamma(\theta)-m_1\le-\frac{C_1}{4}\frac{\varepsilon^\gamma s}{\Lambda^\gamma}\right)\le\frac{1}{10}.\end{aligned}\tag{46}$$

By Lemma 9, we have $\mu_i(B)\ge 7/8$, $i=0,1$. Combining these inequalities with (43) and (46) we immediately conclude that

$$\mu_{0,B}\big(N_\gamma(\theta)\ge c\big)+\mu_{1,B}\big(N_\gamma(\theta)\le c+2u\big)\le 8/35.\tag{47}$$

Next, we consider the total variation distance $V(\mathbb{P}_{0,B},\mathbb{P}_{1,B})$. Using Lemma 9 we get that, for $\bar{C}$ large enough,

$$\begin{aligned}V(\mathbb{P}_{0,B},\mathbb{P}_{1,B})&\le V(\mathbb{P}_{0,B},\mathbb{P}_0)+V(\mathbb{P}_0,\mathbb{P}_1)+V(\mathbb{P}_1,\mathbb{P}_{1,B})\\&\le V(\mathbb{P}_0,\mathbb{P}_1)+\mu_0\big(B^c\big)+\mu_1\big(B^c\big)\\&\le V(\mathbb{P}_0,\mathbb{P}_1)+1/4\\&\le\sqrt{\chi^2(\mathbb{P}_1,\mathbb{P}_0)/2}+1/4\\&\le(\sqrt{2}+1)/4,\end{aligned}\tag{48}$$

where the last two inequalities are due to Pinsker's inequality and Lemma 11, respectively. Combining (42), (47) and (48) we get that, if $s^2\ge\bar{C}d$ for $\bar{C}>0$ large enough, there exists a constant $C>0$ depending only on γ such that

$$\inf_{\hat{T}}\sup_{\theta\in B_0(s)}\mathbf{P}_\theta\left(\big|\hat{T}-N_\gamma(\theta)\big|\ge C\frac{\varepsilon^\gamma s}{\Lambda^\gamma}\right)>\frac{1}{16}.$$

This completes the proof.

9. Lemmas for the proof of Theorem 8

Lemma 9. *For $i=0,1$, we have*

$$V(\mathbb{P}_i,\mathbb{P}_{i,B})\le\mu_i\big(B^c\big).\tag{49}$$

Furthermore, there exists an absolute constant $\bar{C}>0$ such that, for any $s^2\ge\bar{C}d$,

$$\mu_i\big(B^c\big)\le 1/8,\quad i=0,1.\tag{50}$$

Proof. We can use, for example, Lemma 4 in Comminges et al. [5]. Repeating its argument we get that $V(\mathbb{P}_i, \mathbb{P}_{i,B}) \le \mu_i(B^c) = \mathbf{P}(\mathcal{B}(d, \frac{s}{2d}) > s) \le e^{-\frac{s}{16}}$ where $\mathcal{B}(d, \frac{s}{2d})$ is the binomial random variable with parameters d and $\frac{s}{2d}$. □

Lemma 10. *Let $\tilde{\mu}_0$ and $\tilde{\mu}_1$ be two probability measures on $[-M, M]$ satisfying the moment matching property (ii) of Lemma 6 with some $K \ge 1$. Let φ_0 and φ_1 be defined in (40) where φ is the density of $\mathcal{N}(0, \varepsilon^2)$ distribution. Then*

$$\int \frac{(\varphi_0(x) - \varphi_1(x))^2}{\varphi(x)}\, dx \le \sum_{k=K+1}^{\infty} \frac{\Lambda^{2k}}{k!}$$

where $\Lambda = M/\varepsilon$.

Proof. By rescaling, it suffices to consider the case $\varepsilon = 1$, $M = \Lambda$. Introducing the notation $\mathbb{E}_i(k) = \int t^k \tilde{\mu}_i(dt)$, $i = 0, 1$, it is straightforward to check that

$$\begin{aligned}
\int \frac{(\varphi_0(x) - \varphi_1(x))^2}{\varphi(x)}\, dx &= \int e^{\vartheta\vartheta'} \tilde{\mu}_1(d\vartheta)\tilde{\mu}_1\big(d\vartheta'\big) \\
&\quad + \int e^{\vartheta\vartheta'} \tilde{\mu}_0(d\vartheta)\tilde{\mu}_0\big(d\vartheta'\big) \\
&\quad - 2\int e^{\vartheta\vartheta'} \tilde{\mu}_1(d\vartheta)\tilde{\mu}_0\big(d\vartheta'\big) \\
&= \sum_{k=0}^{\infty} \frac{1}{k!}\big(\big(\mathbb{E}_1(k)\big)^2 + \big(\mathbb{E}_0(k)\big)^2 - 2\mathbb{E}_1(k)\mathbb{E}_0(k)\big) \\
&= \sum_{k=0}^{\infty} \frac{1}{k!}\big(\mathbb{E}_1(k) - \mathbb{E}_0(k)\big)^2.
\end{aligned}$$

It remains to notice that $\mathbb{E}_1(k) = \mathbb{E}_0(k)$ for $k = 0, \dots, K$, by property (ii) of Lemma 6, and $|\mathbb{E}_1(k) - \mathbb{E}_0(k)| \le \Lambda^{2k}$ for all k. □

Lemma 11. *If $s^2 \ge 4d$, then*

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) < 1/4.$$

Proof. Since $\mathbb{P}_0$ and $\mathbb{P}_1$ are product measures we have

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) = \left(1 + \int \frac{(g-h)^2}{h}\right)^d - 1,$$

cf., for example, Tsybakov [14], page 86. It follows from the definition of g and h and from Lemma 10 that

$$\int \frac{(g-h)^2}{h} \le \frac{1}{1-\frac{s}{2d}}\left(\frac{s}{2d}\right)^2 \int \frac{(\varphi_1-\varphi_0)^2}{\varphi} \le 2\left(\frac{s}{2d}\right)^2 \sum_{k=K+1}^{\infty} \frac{\Lambda^{2k}}{k!}.$$

Using the inequalities $k! \ge (k/e)^k$ and $1+x \le e^x$ we get

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) \le \exp\left(\frac{s^2}{2d} \sum_{k=K+1}^{\infty} \left(\frac{e\Lambda^2}{k}\right)^k\right) - 1.$$

Recall that $K \ge 3e\Lambda^2/2$ and $K - 2 < 3e\Lambda^2/2$. Thus,

$$\begin{aligned}\frac{s^2}{2d} \sum_{k=K+1}^{\infty} \left(\frac{e\Lambda^2}{k}\right)^k &\le \frac{s^2}{2d} \sum_{k=K+1}^{\infty} (2/3)^k = \frac{s^2}{d}(2/3)^K \\ &< \frac{4s^2}{9d} \exp\bigl(3e\log(2/3)L^2/2\bigr) = \frac{4}{9}\left(\frac{s^2}{d}\right)^a\end{aligned}$$

where $a = 1 + 3e\log(2/3)/2 < -0.6$. Since $s^2 \ge 4d$ we get $\chi^2(\mathbb{P}_1, \mathbb{P}_0) \le \exp(4^{0.4}/9) - 1 < 1/4$. □

10. Proof of Theorems 7 and 9

Theorems 7 and 9 are obtained as corollaries of Theorem 6 thanks to the following lemma.

Lemma 12. *Let $\gamma > 1$. Then, for any $\phi > 0$, any $\theta \in \mathbb{R}^d$ such that $\|\theta\|_\gamma \le \phi^{1/\gamma}$, and any estimator $\hat{T} \ge 0$,*

$$\mathbf{P}_\theta\bigl(\bigl|\hat{T} - \|\theta\|_\gamma\bigr| \ge \phi^{1/\gamma}\bigr) \ge \mathbf{P}_\theta\bigl(\bigl|\hat{T}^\gamma - \|\theta\|_\gamma^\gamma\bigr| \ge C\phi\bigr).$$

Proof. Since $\|\theta\|_\gamma \le \phi^{1/\gamma}$ we have

$$\begin{aligned}\bigl|\hat{T} - \|\theta\|_\gamma\bigr| &= \bigl|\hat{T} - \|\theta\|_\gamma\bigr| \mathbb{1}_{\{\hat{T} > 2\phi^{1/\gamma}\}} + \bigl|\hat{T} - \|\theta\|_\gamma\bigr| \mathbb{1}_{\{\hat{T} \le 2\phi^{1/\gamma}\}} \\ &\ge \phi^{1/\gamma} \mathbb{1}_{\{\hat{T} > 2\phi^{1/\gamma}\}} + \bigl|\hat{T} - \|\theta\|_\gamma\bigr| \mathbb{1}_{\{\hat{T} \le 2\phi^{1/\gamma}\}} \\ &\ge \phi^{1/\gamma} \mathbb{1}_{\{\hat{T} > 2\phi^{1/\gamma}\}} + \frac{|\hat{T}^\gamma - \|\theta\|_\gamma^\gamma|}{\gamma \max(\hat{T}^{(\gamma-1)/\gamma}, \|\theta\|_\gamma^{\gamma-1})} \mathbb{1}_{\{\hat{T} \le 2\phi^{1/\gamma}\}} \\ &\ge \phi^{1/\gamma} \mathbb{1}_{\{\hat{T} > 2\phi^{1/\gamma}\}} + \frac{|\hat{T}^\gamma - \|\theta\|_\gamma^\gamma|}{2^{\gamma-1}\gamma\phi^{(\gamma-1)/\gamma}} \mathbb{1}_{\{\hat{T} \le 2\phi^{1/\gamma}\}}\end{aligned}$$

where we have used the inequality $|x^\gamma - y^\gamma| \le \gamma \max(x^{\gamma-1}, y^{\gamma-1})|x-y|$, $\forall x, y > 0$. This yields the result of the lemma with $C = 2^{\gamma-1}\gamma$. □

It suffices to prove Theorems 7 and 9 for the indicator loss $\ell(t) = \mathbb{1}_{t \ge a}$, $a > 0$, and to consider the infimum only over non-negative estimators $\hat{T} \ge 0$ since the estimated functional is non-negative. It follows from Lemma 12 that

$$\inf_{\hat{T}\ge 0} \sup_{\theta \in B} \mathbf{P}_\theta\big(|\hat{T} - \|\theta\|_\gamma| \ge \phi^{1/\gamma}\big) \ge \inf_{\hat{T}'\ge 0} \sup_{\theta \in B} \mathbf{P}_\theta\big(|\hat{T}' - \|\theta\|_\gamma^\gamma| \ge C\phi\big),$$

where $B = B_0(s) \cap \{\|\theta\|_\gamma \le \phi^{1/\gamma}\}$. The result of Theorem 7 follows immediately from this inequality with $\phi = c\varepsilon^\gamma s \log^{\gamma/2}(1 + d/s^2)$ and Theorem 6. Here, c is a sufficiently small positive number. To prove Theorem 9, it suffices to apply Theorem 6 with s being the minimal integer greater than or equal to $\sqrt{d}$ and to use the fact that the classes $B_0(s)$ are nested.

11. Proof of Theorem 10

The proof is analogous to that of Theorem 8 subject to a modification that we detail here. Let c and u be as in the proof of Theorem 8:

$$c = m_0 + 3v_0, \qquad u = \frac{m_1 - m_0}{4}.$$

Define

$$c' = c^{1/\gamma}, \qquad u' = \frac{(c+2u)^{1/\gamma} - c^{1/\gamma}}{2}.$$

Analogously to (42), we obtain from Theorem 2.15 in Tsybakov [14] that

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{P}_\theta\big(|\hat{T} - \|\theta\|_\gamma| \ge u'\big) \ge \frac{1 - V'}{2}, \tag{51}$$

where

$$V' = V(\mathbb{P}_{0,B}, \mathbb{P}_{1,B}) + \mu_{0,B}\big(\|\theta\|_\gamma \ge c'\big) + \mu_{1,B}\big(\|\theta\|_\gamma \le c' + 2u'\big).$$

Note that this value is equal to V' defined in the proof of Theorem 8. Hence, V' is bounded from above exactly as in the proof of Theorem 8 and to complete the proof of Theorem 10 we only need to check that $u' \ge C\varepsilon s^{1/\gamma}\log^{1/2-\gamma}(s^2/d)$, which is the desired rate. Using the inequality $|x^\gamma - y^\gamma| \le \gamma \max(x^{\gamma-1}, y^{\gamma-1})|x-y|$, $\forall x, y > 0$, we get

$$u' \ge \frac{2u}{\gamma(c+2u)^{(\gamma-1)/\gamma}}.$$

Next, due to (41), (44) and the assumption that $s \ge 2\sqrt{d}$, we have

$$c + 2u = \frac{m_1 + m_0}{2} + 3v_0 \le sM^\gamma + 3\sqrt{d}M^\gamma \le 3sM^\gamma = 3s\varepsilon^\gamma \Lambda^\gamma.$$

Moreover, (41) implies that $u \geq (C_1/4)s\varepsilon^{\gamma}\Lambda^{-\gamma}$. Thus, $u' \geq C\varepsilon s^{1/\gamma}\Lambda^{1-2\gamma}$ and we conclude by recalling the definition of Λ.

Acknowledgements

The work of Olivier Collier has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The work of A.B. Tsybakov was supported by GENES and by the French National Research Agency (ANR) under the grants IPANEMA (ANR-13-BSH1-0004-02) and Labex Ecodec (ANR-11-LABEX-0047).

References

[1] Cai, T.T. and Low, M.G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346 https://doi.org/10.1214/10-AOS849

[2] Carpentier, A. and Verzelen, N. (2019). Adaptive estimation of the sparsity in the Gaussian vector model. *Ann. Statist.* **47** 93–126. MR3909928 https://doi.org/10.1214/17-AOS1680

[3] Collier, O., Comminges, L. and Tsybakov, A.B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444 https://doi.org/10.1214/15-AOS1432

[4] Collier, O., Comminges, L., Tsybakov, A.B. and Verzelen, N. (2018). Optimal adaptive estimation of linear functionals under sparsity. *Ann. Statist.* **46** 3130–3150. MR3851767 https://doi.org/10.1214/17-AOS1653

[5] Comminges, L., Collier, O., Ndaoud, M. and Tsybakov, A.B. (2018). Adaptive robust estimation in sparse vector model. Available at arXiv:1802.04230.

[6] Fukuchi, K. and Sakuma, J. (2017). Minimax optimal estimators for additive scalar functionals of discrete distributions. Available at arXiv:1701.06381.

[7] Han, Y., Jiao, J., Mukherjee, R. and Weissman, T. (2017). On estimation of L_r-norms in Gaussian white noise models. Available at arXiv:1710.03863.

[8] Han, Y., Jiao, J., Weissman, T. and Wu, Y. (2017). Optimal rates of entropy estimation over Lipschitz balls. Available at arXiv:1711.02141.

[9] Jiao, J., Venkat, K., Han, Y. and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inform. Theory* **61** 2835–2885. MR3342309 https://doi.org/10.1109/TIT.2015.2412945

[10] 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. MR1670867 https://doi.org/10.1007/s004409970006

[11] Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math.* **1738** 85–277. Berlin: Springer. MR1775640

[12] Szegö, G. (1936). On some Hermitian forms associated with two given curves of the complex plane. *Trans. Amer. Math. Soc.* **40** 450–461. MR1501884 https://doi.org/10.2307/1989634

[13] Timan, A.F. (1963). *Theory of Approximation of Functions of a Real Variable*. *International Series of Monographs in Pure and Applied Mathematics*, *Vol.* 34. New York: Pergamon. MR0192238

[14] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794

[15] Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory* **62** 3702–3720. MR3506758 https://doi.org/10.1109/TIT.2016.2548468

[16] Wu, Y. and Yang, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.* **47** 857–883. MR3909953 https://doi.org/10.1214/17-AOS1665

Received May 2018 and revised October 2019