HAL open science

Minimax optimal estimators for general additive functional estimation

Olivier Collier, Laëtitia Comminges

▶ **To cite this version:**

Olivier Collier, Laëtitia Comminges. Minimax optimal estimators for general additive functional estimation. 2019. hal-02273511

HAL Id: hal-02273511

https://hal.science/hal-02273511v1

Preprint submitted on 29 Aug 2019

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 optimal estimators for general additive functional estimation

Olivier Collier olivier.collier@parisnanterre.fr
MODAL'X
Université Paris-Nanterre
Nanterre, 92000, France
and
CREST, ENSAE
Palaiseau, 91120

Laëtitia Comminges comminges@ceremade.dauphine.fr
CEREMADE
Université Paris-Dauphine
Paris, 75775, France
and
CREST, ENSAE
Palaiseau, 91120

Editor:

Abstract

In this paper, we observe a sparse mean vector through Gaussian noise and we aim at estimating some additive functional of the mean in the minimax sense. More precisely, we generalize the results of (Collier et al., 2017, 2019) to a very large class of functionals. The optimal minimax rate is shown to depend on the polynomial approximation rate of the marginal functional, and optimal estimators achieving this rate are built.

Keywords: Minimax estimation, additive functional, sparsity, polynomial approximation

1. Introduction

In the general problem of functional estimation, one is interested in estimating some quantity $\boldsymbol{F}(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is an unknown parameter and $\boldsymbol{F}$ is a known function. Information on this quantity is provided by an observation $\boldsymbol{y} \sim \mathbf{P}_{\boldsymbol{\theta}}$, where $(\mathbf{P}_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$ is some family of probability distributions.

An exhaustive bibliography on the subject of functional estimation is out of the scope of this paper, but typical examples include functionals of a density function, e.g. the integrals of its square Bickel and Ritov (1988), of smooth functionals of its derivatives Birgé and Massart (1995) or of nonlinear functionals Kerkyacharian and Picard (1996).

In this work, we focus on the case where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a finite vector and $\boldsymbol{F}$ is an additive functional, *i.e.*,

$$\boldsymbol{F}(\boldsymbol{\theta}) = \sum_{i=1}^{d} F(\theta_i), \tag{1}$$

which has now been well studied in the literature. For example, in the case when $\mathbf{P}_{\boldsymbol{\theta}}$ is the multinomial distribution $\mathcal{M}(n, p_1, \ldots, p_d)$, Shannon's or Rnyi's entropy, which correspond respectively to marginal functionals $F(t) = -t\log(t)$ and $F(t) = t^\alpha$, are considered in Han et al. (2015a,b); Wu and Yang (2016) among others. The distinct elements problem, *i.e.*, finding how many different colors are present among at most d colored balls in an urn, can also be expressed in this form Polyanskiy and Wu (2019); Wu and Yang (2018). Moreover, the quadratic functional defined by $F(t) = t^2$ is key in the problem of signal detection Carpentier et al. (2018), and when the vector $\boldsymbol{\theta}$ is assumed to be sparse, *i.e.*, when most of its coefficients are assumed to be exactly 0, it also plays a crucial role for noise variance estimation Comminges et al. (2019). Finally, robust estimation of the mean is shown in Collier and Dalalyan (2019) to be related with a linear functional of the outliers.

Here, our aim is not to focus on some particular functional, but to exhibit optimal minimax rates over large classes of functionals. Furthermore, we consider the Gaussian mean model, *i.e.*,

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\theta}, I_d) \quad \Rightarrow \quad y_i = \theta_i + \xi_i, \quad \xi_i \overset{iid}{\sim} \mathcal{N}(0,1) \tag{2}$$

and we measure the quality of an estimator by the minimax risk defined by

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbf{E}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2, \tag{3}$$

where $\boldsymbol{\Theta}$ is some set of parameters. This framework was also used in Collier et al. (2017); Cai and Low (2011); Collier et al. (2019), where respectively the cases when $F(t) = t$ or $F(t) = t^2$, $F(t) = |t|$ and $F(t) = |t|^\gamma$ for $0 < \gamma \leq 1$ are studied. It is clear from the last two papers that for rapidly growing functionals, it is relevant to restrict the set of $\boldsymbol{\theta}'s$ to a bounded subset of $\mathbb{R}^d$. Therefore, we assume that each component of $\boldsymbol{\theta}$ belongs to a segment, which we take for simplicity sake in the form $[-M, M]$. Finally, we place ourselves in a sparse context, which means that we assume the number of nonzero coefficients of $\boldsymbol{\theta}$ – its l_0-norm – to be bounded by a known quantity, and we define

$$\boldsymbol{\Theta} \triangleq \boldsymbol{\Theta}_{s,M} = \big\{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_0 \leq s, \|\boldsymbol{\theta}\|_\infty \leq M\big\}. \tag{4}$$

In this paper, we build minimax rate-optimal estimators when the functional F is not too regular in the sense of polynomial approximation and does not grow too fast, when s is at least of the order of $\sqrt{d}$ and M is at most of order $\sqrt{\log(s^2/d)}$, showing that the polynomial approximation based method developed in Collier et al. (2019) can be extended to a very broad class of functionals. More precisely, we make the following assumptions, where we use the notation $\delta_{K,M}$ that is introduced in (9) below:

(A0) F is continuous on $[-\sqrt{\log(s^2/d)}, \sqrt{\log(s^2/d)}]$.

(A0') F is continuous on $[-\sqrt{\log(s)}, \sqrt{\log(s)}]$.

(A1) There exist positive real numbers ϵ_1, C_1 such that

$$\sup_{\sqrt{2\log(s^2/d)} \leq M \leq \sqrt{2\log(d)}} \frac{\|F - F(0)\|_{\infty,[-M,M]}}{e^{\epsilon_1 M^2}} \leq C_1. \tag{5}$$

(A2) There exist positive real numbers ϵ_2, C_2 such that

$$\sup_{\sqrt{2\log(s^2/d)} \leq M \leq \sqrt{2\log(d)}} \frac{\delta_{M^2,M}^{-1}}{e^{\epsilon_2 M^2}} \leq C_2. \tag{6}$$

(A3) $\forall \alpha > 0$, $\exists f_\alpha > 0$ such that if $|1 - K_1/K_2| \vee |1 - M_1/M_2| \leq \alpha$, then

$$f_\alpha^{-1} \leq \frac{\delta_{K_1,M_1}}{\delta_{K_2,M_2}} \leq f_\alpha.$$

We make the first assumption on the continuity of F for simplicity sake. Indeed, it would not be hard to extend the result to the case of a functional that is piecewise continuous with a finite number of discontinuities, even if discontinuous functionals might not be very important in practice. The second assumption is very mild, since estimation of rapidly growing functionals leads to very large minimax rates, making such problems uninteresting in practice. However the third assumption is essential: it expresses how the polynomial approximation rate drives the quality of estimation of the associated additive functional. Assumption (A2) thus requires that F is not smooth enough to be very quickly approximated by polynomials. In Section 3, we recall the relation between polynomial approximation and differentiability. Finally, the last assumption is convenient to show that our lower and upper bounds match up to a constant. We believe that it is satisfied for all reasonable functionals.

Our theorems allow to recover some of the results implied by Cai and Low (2011); Collier et al. (2019), but cover a large part of all possible functionals. Note that some papers have already tackled the problem of general functionals. In Fukuchi and Sakuma (2019), the authors give optimal rates of convergence for additive functionals in the discrete distribution case, when the fourth-derivative of the marginal functional is close in sup-norm to an inverse power function. In Koltchinskii and Zhilova (2018), the case of general, not necessarily additive, functionals is considered in the Gaussian mean model with arbitrary covariance matrix. However, their results differ significantly from ours since they consider minimax risk over all marginal functionals belonging to some relatively small set of bounded and smooth functions in the Hlder sense. For example, none of the results obtained in Collier et al. (2017, 2019); Cai and Low (2011) can be recovered. Finally, the minimax rate for even larger classes of functionals, under constraints in the form $\sum_{i=1}^d c(\theta_i) \leq 1$ which includes sparsity, is obtained in Polyanskiy and Wu (2019) in term of the quantity

$$\sup_{\pi_1,\pi_2} \Big\{ \Big| \int F(\theta)\, \pi_1(d\theta) - \int F(\theta)\, \pi_2(d\theta) \Big| \,|\, \chi^2(\mathbf{P}_{\pi_1}, \mathbf{P}_{\pi_2}) \leq \frac{1}{d}, \mathbf{E}_{\pi_i} \sum_{i=1}^d c(\theta_i) \leq 1 \Big\} \tag{7}$$

where $\mathbf{P}_\pi = \int \mathbf{P}_\theta\, \pi(d\theta)$, $\chi^2(\mathbf{P}_{\pi_1}, \mathbf{P}_{\pi_2})$ is the chi-square divergence between probabilities $\mathbf{P}_{\pi_1}$ and $\mathbf{P}_{\pi_2}$ and the supremum is taken over all probability distributions on $\mathbf{\Theta}$. Their theorems allow for example to recover the minimax rate from Cai and Low (2011) when Θ is bounded, and may also allow to get the minimax rates from this paper. However, they do not exhibit generic estimators achieving the minimax risk. This paper fills in this gap in some cases.

Organization of the paper

In Section 2, we build rate-optimal estimators of the additive functional and assess their performance. In Section 3, we prove their optimality up to constants, and discuss conditions under which Assumption (A2) is satisfied. The proofs of the theorems are postponed to Section 4, while technical lemmas can be found in Section 5.

2. Upper bounds

2.1 Polynomial approximation

Here, we set the notation on polynomial approximation that will be used throughout this paper. First denote $\mathcal{P}_K$ the set of polynomials of degree at most K, then define the polynomial of best approximation of F on $[a,b]$ by

$$P_{K,[a,b]} = \underset{P\in\mathcal{P}_K}{\arg\min}\|F-P\|_{\infty,[a,b]} \tag{8}$$

and the polynomial approximation rate by

$$\delta_{K,[a,b]} = \|F - P_{K,[a,b]}\|_{\infty,[a,b]}. \tag{9}$$

In the following, we write $P_{K,M} = P_{K,[-M,M]}$, $\delta_{K,M} = \delta_{K,[-M,M]}$, and we decompose $P_{K,M}$ in the canonical base as

$$P_{K,M} = \sum_{k=0}^{K} a_{k,K,M} X^k. \tag{10}$$

2.2 Definition of the estimator and main theorem

First, we use the sample duplication trick to transform observation y_i into independent randomized observations $y_{1,i}, y_{2,i}$ while keeping the same mean. Let us consider random variables $z_1,\dots,z_d \overset{iid}{\sim} \mathcal{N}(0,1)$ independent of $\boldsymbol{y}$ and define

$$y_{1,i} = y_i + z_i, \quad y_{2,i} = y_i - z_i, \tag{11}$$

so that $y_{1,i}, y_{2,i} \overset{iid}{\sim} \mathcal{N}(\theta_i, 2)$. Yet for convenience, we will assume that $y_{1,i}, y_{2,i} \overset{iid}{\sim} \mathcal{N}(\theta_i, 1)$.

Then, we recall the definition of the Hermite polynomials H_k defined by

$$H_k(x) = (-1)^k e^{x^2/2}\frac{\partial^k}{\partial x^k}\big(e^{-x^2/2}\big), \tag{12}$$

which have in particular the property that $\mathbf{E}_{X\sim\mathcal{N}(\theta,1)}H_k(X) = \theta^k$.

Finally, we define our estimator of $\boldsymbol{F}(\boldsymbol{\theta})$ as

$$\hat{\boldsymbol{F}} = \sum_{i=1}^{d} \hat{F}(y_{1,i}, y_{2,i}) \tag{13}$$

where

$$\hat{F}(u,v) = \sum_{l=0}^{L} \hat{P}_{K_l,M_l}(u)\mathbb{1}_{t_{l-1}<|v|\le t_l} + \hat{P}_{K_{L+1},M_{L+1}}(u)\mathbb{1}_{t_L<|v|}, \tag{14}$$

and for an arbitrary constant $c > 0$,

$$\begin{cases} \hat{P}_{K,M}(u) = \sum_{k=1}^{K} a_{k,K,M} H_k(u), \\ M_l = 2^l \sqrt{2\log(s^2/d)}, \quad K_l = \frac{c}{8} M_l^2, \\ t_l = M_l/2,\, t_{-1} = 0, \\ L \text{ is the largest integer such that } 2^L < \sqrt{\log(d)/\log(s^2/d)}, \\ M_{L+1} = \sqrt{2\log(d)}. \end{cases} \tag{15}$$

The next theorem is a slight modification of Theorem 1 in Collier et al. (2019). It states the performance of our estimator in the case when the signal $\boldsymbol{\theta}$ is not too sparse.

Theorem 1 *Assume that $2\sqrt{d} \le s \le d$ and that F satisfies Assumptions (A1-A2) with $\epsilon_1 + \epsilon_2$ small enough. Then the estimator defined in (13) with small enough c, depending on ϵ_1 and ϵ_2, satisfies*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{s,\sqrt{2\log(d)}}} \mathbf{E}_{\boldsymbol{\theta}} \big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 \le C_3\, s^2 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}, \tag{16}$$

where C_3 is some positive constant, depending only on C_1 and C_2.

Furthermore, in the case when no sparsity is assumed ($s = d$), we can derive a simpler statement for every segment $[-M, M]$ included in $[-\sqrt{\log(d)}, \sqrt{\log(d)}]$. To this end, we define the simplified estimator

$$\tilde{\boldsymbol{F}} = \sum_{i=0}^{d} \hat{P}_{K,M}(y_i), \quad \hat{P}_{K,M}(u) = \sum_{k=0}^{K} a_{k,K,M} H_k(u), \tag{17}$$

with $K = c\log(d)/\log(e\log(d)/M^2)$ for an arbitrary constant $c > 0$.

Theorem 2 *Assume that $0 < M \le \sqrt{\log(d)}$, that for some constants $C_1', C_2' > 0$*

$$\|F - F(0)\|_{\infty,[-M,M]} \le C_1' d^{\epsilon_1}, \quad \delta^{-1}_{K,M} \le C_2' d^{\epsilon_2}, \tag{18}$$

and let $\tilde{\boldsymbol{F}}$ be the estimator defined by (17). Then if $2\epsilon_1 + 2\epsilon_2 < 1$ and if c is chosen small enough, depending only on $\epsilon_1 + \epsilon_2$, then

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d,M}} \mathbf{E}_{\boldsymbol{\theta}} \big(\tilde{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 \le C_3' d^2 \delta^2_{K,M}, \tag{19}$$

where C_3' is some positive constant, depending only on C_1' and C_2'.

3. Optimality results

The next theorem, which is a slight modification of Theorem 4 in Collier et al. (2019), states a lower bound on the minimax rate.

Theorem 3 *Assume that* $0 < M \leq \sqrt{\log(s^2/d)}$ *and that for some constants* $C_1'', C_2'' > 0$

$$\|F - F(0)\|_{\infty,[-M,M]} \leq C_1''\Big(\frac{s^2}{d}\Big)^{\epsilon_1}, \quad \delta^{-1}_{e^2\log(s^2/d)/\log(e\log(s^2/d)/M^2),M} \leq C_2''\Big(\frac{s^2}{d}\Big)^{\epsilon_2}, \tag{20}$$

and that Assumption (A0) holds. Then there exists an absolute positive constant $C \geq 1$ *such that if* $s^2 \geq Cd$*, if* $2\epsilon_1 + 2\epsilon_2 < 1$ *and if* c *is chosen small enough, depending on* $\epsilon_1 + \epsilon_2$*, we have*

$$\inf_{\hat{\boldsymbol{F}}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{s,M}} \mathbf{E}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 \geq C_3'' s^2 \delta^2_{e^2\log(s^2/d)/\log(e\log(s^2/d)/M^2),M}, \tag{21}$$

for some positive constant C_3''*, depending only on* C_1'' *and* C_2''*.*

But our estimation problem is more difficult than the problem where we know in advance that the nonzero coefficients belong to the first k components of $\boldsymbol{\theta}$ for $k \in \{s, \dots, d\}$, and the last theorem gives lower bounds for these problems as well by replacing d by k. This argument leads to the following corollary:

Corollary 4 *Let Assumptions (A0'-A1-A2) hold. Then there exist an absolute positive constant* $C \geq 1$ *such that if* $s^2 \geq Cd$*, if* $2\epsilon_1 + 2\epsilon_2 < 1$ *and if* c *is chosen small enough, depending only on* $\epsilon_1 + \epsilon_2$*, we have*

$$\inf_{\hat{\boldsymbol{F}}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{s,\sqrt{\log(s)}}} \mathbf{E}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 \geq C_4 s^2 \max_{s\leq k\leq d} \delta^2_{e^2\log(s^2/k),\sqrt{\log(s^2/k)}}, \tag{22}$$

for some positive constant C_4*, depending only on* C_1*,* C_2*.*

Furthermore, the next theorem states that Assumption (A3) is sufficient to prove that the upper bound from Theorem 1 matches with the lower bound from Corollary 4.

Theorem 5 *Let Assumptions (A0'-A1-A2-A3) hold. Then there exist an absolute positive constant* $C \geq \sqrt{2}$ *such that if* $s^2 \geq Cd$*, if* ϵ_1, ϵ_2 *are small enough and* c *is chosen small enough, depending only on* ϵ_1 *and* ϵ_2*, we have*

$$C_5 \leq \frac{\inf_{\hat{\boldsymbol{F}}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{s,\sqrt{\log(d)}}} \mathbf{E}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2}{s^2 \max_{s\leq k\leq d} \delta^2_{\log(s^2/k),\sqrt{\log(s^2/k)}}} \leq C_5', \tag{23}$$

for some positive constants C_5, C_5'*, depending only on* C_1, C_2 *and* c*.*

This means in particular that for non-regular functionals satisfying the conditions (A0'-A1-A2-A3), the rate appearing in Theorem 5 must be the same as the rate found in Polyanskiy and Wu (2019). More precisely, let us denote $\delta_{\chi^2}(\frac{1}{\sqrt{d}})$ the following quantity

$$\sup_{\pi_1,\pi_2} \Big\{ \Big| \int F(\theta)\,\pi_1(d\theta) - \int F(\theta)\,\pi_2(d\theta) \Big| \,|\, \chi^2(\mathbf{P}_{\pi_1}, \mathbf{P}_{\pi_2}) \leq \frac{1}{d}, \mathbf{E}_{\pi_i}\|\boldsymbol{\theta}\|_0 \leq s \Big\} \tag{24}$$

where the supremum is taken over all distribution probabilities on $[-M, M]^d$. According to Theorem 8 in Polyanskiy and Wu (2019), if $\delta_{\chi^2}(\frac{1}{\sqrt{d}}) \geq 31\frac{\|F\|_{\infty,[-M,M]}}{\sqrt{d}}$, then $\delta_{\chi^2}(\frac{1}{\sqrt{d}})$ is the minimax rate for estimating $\sum_{i=1}^d F(\theta_i)$ over $\mathbf{\Theta}_{s,M}$, hence it is of the order of

$$s^2 \max_{s\leq k\leq d} \delta^2_{\log(s^2/k),\sqrt{\log(s^2/k)}}. \tag{25}$$

Moreover, similar results as in Cai and Low (2011); Collier et al. (2019) (with bounded parameter space) can be easily deduced since for the function $x \to |x|^\gamma$, the approximation rate $\delta_{K,M}$ is of the order of $(M/K)^\gamma$ (*cf.* for example Theorem 7.2.2 in Timan (1963)).

Finally, Assumption (A2) is strongly related to the differentiability of the marginal functional F. Indeed, the following properties can be found in Timan (1963), Sections 5.1.5 and 6.2.4:

- If F has a bounded derivative of order r on $[-1, 1]$, then
$$\forall n \geq 1, \quad \delta_{n,[-1,1]} \leq \frac{C}{n^r}, \tag{26}$$
for some positive contant C.
- F is infinitely derivable on $[a, b]$ if and only if for any $r > 0$,
$$n^r \delta_{n,[a,b]} \to 0. \tag{27}$$

This suggests that many not infinitely differentiable functionals satisfy Assumption (A2).

4. Proof of theorems

In the whole section, we denote by A a positive constant the value of which may vary from line to line. This constant only depends on C_1 and C_2 (Theorem 1) and Theorem 5), C_1' and C_2' (Theorem 2), C_1'' and C_2'' (Theorem 3). Moreover, since

$$\mathbf{E}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 = \mathbf{E}_{\boldsymbol{\theta}}\Big[\big(\hat{\boldsymbol{F}} - dF(0)\big) - \big(\boldsymbol{F}(\boldsymbol{\theta}) - dF(0)\big)\Big]^2, \tag{28}$$

we can assume without loss of generality that $F(0) = 0$, which we do throughout this section.

4.1 Proof of Theorem 1

Denote by S the support of $\boldsymbol{\theta}$. We start with a bias-variance decomposition

$$\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 \leq 4\Big(\sum_{i\in S} \mathbf{E}_{\boldsymbol{\theta}}\hat{F}(y_{1,i}, y_{2,i}) - \sum_{i\in S} F(\theta_i)\Big)^2 \tag{29}$$

$$+ 4\Big(\sum_{i\in S} \hat{F}(y_{1,i}, y_{2,i}) - \sum_{i\in S} \mathbf{E}_{\boldsymbol{\theta}}\hat{F}(y_{1,i}, y_{2,i})\Big)^2 \tag{30}$$

$$+ 4\Big(\sum_{i\notin S} \mathbf{E}_{\boldsymbol{\theta}}\hat{F}(y_{1,i}, y_{2,i})\Big)^2 \tag{31}$$

$$+ 4\Big(\sum_{i\notin S} \hat{F}(y_{1,i}, y_{2,i}) - \sum_{i\notin S} \mathbf{E}_{\boldsymbol{\theta}}\hat{F}(y_{1,i}, y_{2,i})\Big)^2 \tag{32}$$

leading to the bound

$$\mathbf{E}_{\boldsymbol{\theta}}(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta}))^2 \leq 4s^2 \max_{i\in S} B_i^2 + 4s \max_{i\in S} V_i + 4d^2 \max_{i\notin S} B_i^2 + 4d \max_{i\notin S} V_i, \tag{33}$$

where $B_i = \mathbf{E}_{\boldsymbol{\theta}}\hat{F}(y_{1,i}, y_{2,i}) - F(\theta_i)$ is the bias of $\hat{F}(y_{1,i}, y_{2,i})$ as an estimator of $F(\theta_i)$ and $V_i = \mathbf{Var}_{\boldsymbol{\theta}}(\hat{F}(y_{1,i}, y_{2,i}))$ is its variance. We now bound separately the four terms in (33).

1°. *Bias for $i \notin S$.* If $i \notin S$, then $B_i = 0$.

2°. *Variance for $i \notin S$.* If $i \notin S$, then using in particular Lemma 9,

$$V_i \leq \sum_{l=0}^{L+1} \mathbf{E}\hat{P}^2_{K_l,M_l}(\xi)\, \mathbf{P}(t_{l-1} < |\xi|), \quad \xi \sim \mathcal{N}(0,1), \tag{34}$$

$$\leq A \sum_{l=0}^{L+1} \|F\|^2_{\infty,[-M_l,M_l]} 6^{K_l} e^{-t^2_{l-1}/2}. \tag{35}$$

For $l = 0$, we have by Assumptions (A1-A2), if $\epsilon_1 + \epsilon_2 < 1/4$ and for c small enough

$$\|F\|^2_{\infty,[-M_0,M_0]} 6^{K_0} e^{-t^2_{-1}/2} \delta^{-2}_{K_0,M_0} \leq A\Big(\frac{s^2}{d}\Big)^{4\epsilon_1+4\epsilon_2+c\log(6)/4} \leq A\frac{s^2}{d}. \tag{36}$$

Then, if $l > 0$,

$$\|F\|^2_{\infty,[-M_l,M_l]}\, 6^{K_l} e^{-t^2_{l-1}/2} \delta^{-2}_{K_l,M_l} \leq A\Big(\frac{s^2}{d}\Big)^{4^l(4\epsilon_1+4\epsilon_2+c\log(6)/4-\frac{1}{16})} \tag{37}$$

so that for small enough c, ϵ_1 and ϵ_2 and since $s^2 \geq 4d$,

$$dV_i \leq As^2 \max_{l=0,\dots,L} \delta^2_{K_l,M_l}. \tag{38}$$

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

$$B_i = \sum_{l=0}^{L} \big\{\mathbf{E}\hat{P}_{K_l,M_l}(\xi) - F(\theta_i)\big\}\, \mathbf{P}(t_{l-1} < |\xi| \leq t_l) \tag{39}$$

$$+ \big\{\mathbf{E}\hat{P}_{K_{L+1},M_{L+1}}(\xi) - F(\theta_i)\big\}\, \mathbf{P}(t_L < |\xi|), \quad \xi \sim \mathcal{N}(\theta_i, 1). \tag{40}$$

We will analyze this expression separately in different ranges of $|\theta_i|$.

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

$$|B_i| \leq \max_l \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi) - F(\theta_i)\big|, \quad \xi \sim \mathcal{N}(\theta_i, 1). \tag{41}$$

Since $|\theta_i| \leq M_l$ for all l, we have by the definition of P_{K_l,M_l} and since $F(0) = 0$,

$$\big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi) - F(\theta_i)\big| \leq \big|P_{K_l,M_l}(\theta_i) - a_{0,K_l,M_l} - F(\theta_i)\big| \tag{42}$$

$$\leq \big|P_{K_l,M_l}(\theta_i) - F(\theta_i)\big| + |F(0) - P_{K_l,M_l}(0)| \tag{43}$$

$$\leq 2\delta_{K_l,M_l}, \tag{44}$$

so that

$$s^2 \max_{0<|\theta_i|\le 2t_0} B_i^2 \le 4s^2 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}. \tag{45}$$

3.2°. *Case* $2t_0 < |\theta_i| \le 2t_L$. Let $l_0 \in \{0,\dots,L-1\}$ be the integer such that $2t_{l_0} < |\theta_i| \le 2t_{l_0+1}$. We have

$$\begin{aligned} |B_i| \le \sum_{l=0}^{l_0} \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi) - F(\theta_i)\big| \cdot \mathbf{P}(t_{l-1} < |\xi| \le t_l) \\ + \max_{l>l_0} \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi) - F(\theta_i)\big|, \quad \xi \sim \mathcal{N}(\theta_i, 1) \end{aligned} \tag{46}$$

The arguments in (42) yield that

$$\max_{l>l_0} \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi) - F(\theta_i)\big| \le 4 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}. \tag{47}$$

Furthermore, using the triangular inequality,

$$\sum_{l=0}^{l_0} \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi) - F(\theta_i)\big| \cdot \mathbf{P}(t_{l-1} < |\xi| \le t_l) \tag{48}$$

$$\le \sum_{l=0}^{l_0} \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi)\big| \cdot \mathbf{P}(|\xi| \le t_l) + \sum_{l=0}^{l_0} \big|F(\theta_i)\big| \cdot \mathbf{P}(t_{l-1} < |\xi| \le t_l). \tag{49}$$

The first sum in the right-hand side can be bounded using Lemma 11, since

$$\big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi)\big| \, \mathbf{P}(|\xi| \le t_l) \le A\|F\|_{\infty,[-M_l,M_l]} \, 3^{K_l} e^{c\theta_i^2/16} \, \mathbf{P}(|\xi| \le t_l), \tag{50}$$

so that, as $|\theta_i| > 2t_{l_0} \ge 2t_l$ for $l \le l_0$,

$$\begin{aligned} \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi)\big| \, \mathbf{P}(|\xi| \le t_l)\delta^{-1}_{K_l,M_l} &\le A3^{K_l} e^{(c-2)\theta_i^2/16} e^{(\epsilon_1+\epsilon_2)M_l^2} && (51)\\ &\le Ae^{(8\epsilon_1+8\epsilon_2+c\log(3)+(c-2)/2)t_l^2/2} && (52)\\ &= A\Big(\frac{s^2}{d}\Big)^{2^{2l-2}(8\epsilon_1+8\epsilon_2+c\log(3)+(c-2)/2)} && (53) \end{aligned}$$

Again, if $\epsilon_1 + \epsilon_2 < \frac{1}{8}$, choosing c small enough yields that

$$\sum_{l=0}^{l_0} \big|\mathbf{E}\hat{P}_{K_l,M_l}(\xi)\big| \, \mathbf{P}(|\xi| \le t_l) \le A \max_{l=0,\dots,L} \delta_{K_l,M_l}. \tag{54}$$

Moreover, similar arguments lead to the fact that if $4\epsilon_1 + \epsilon_2 < \frac{1}{8}$

$$\sum_{l=0}^{l_0} |F(\theta_i)|\mathbf{P}(t_{l-1} < |\xi| \le t_l) \le \|F\|_{\infty,[-M_{l_0+1},M_{l_0+1}]}\mathbf{P}(|\xi| \le t_{l_0}) \le A\delta_{K_{l_0},M_{l_0}}, \tag{55}$$

and we conclude that

$$s^2 \max_{2t_0<|\theta_i|\le 2t_L} B_i^2 \le As^2 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}. \tag{56}$$

3.3°. *Case* $2t_L < |\theta_i| \le \sqrt{2\log(d)}$. Similar arguments as in the previous case yield that

$$s^2 \max_{2t_L<|\theta_i|\le\sqrt{2\log(d)}} B_i^2 \le As^2 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}. \tag{57}$$

4°. *Variance for* $i \in S$. We consider the same cases as in item 3° above. In all cases, it suffices to bound the variance by the second-order moment, which grants that, for all $i \in S$,

$$V_i \le \sum_{l=0}^{L} \mathbf{E}\hat{P}^2_{K_l,M_l}(\xi)\,\mathbf{P}(t_{l-1} < |\xi| \le t_l) + \mathbf{E}\hat{P}^2_{K_{L+1},M_{L+1}}(\xi)\,\mathbf{P}(t_L < |\xi|), \quad \xi \sim \mathcal{N}(\theta_i, 1). \tag{58}$$

4.1°. *Case* $0 < |\theta_i| \le 2t_0$. In this case, we deduce from (58) that

$$V_i \le \max_{l=0,\dots,L+1} \mathbf{E}\hat{P}^2_{K_l,M_l}(\xi), \quad \xi \sim \mathcal{N}(\theta_i, 1). \tag{59}$$

Lemma 10 implies

$$V_i \le A\|F\|^2_{\infty,[-M_{L+1},M_{L+1}]} 12^{K_{L+1}} \le Ad^{4\epsilon_1 + c\log(12)/4}, \tag{60}$$

which, as $\sqrt{d} \le s$, is sufficient to conclude that

$$s \max_{0<|\theta_i|\le 2t_0} V_i \le As^2 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}, \tag{61}$$

for $c, \epsilon_1, \epsilon_2$ small enough.

4.2°. *Case* $2t_0 < |\theta_i| \le 2t_L$. As in item 3.2° above, we denote by $l_0 \in \{0, \dots, L-1\}$ the integer such that $2t_{l_0} < |\theta_i| \le 2t_{l_0+1}$. We deduce from (58) that

$$V_i \le (l_0+1) \max_{l=0,\dots,l_0} \mathbf{E}\hat{P}^2_{K_l,M_l}(\xi)\,\mathbf{P}(|\xi| \le t_{l_0}) + \max_{l=l_0+1,\dots,L+1} \mathbf{E}\hat{P}^2_{K_l,M_l}(\xi), \quad \xi \sim \mathcal{N}(\theta_i, 1). \tag{62}$$

The last term on the right hand side is controlled as in item 4.1°. For the first term, we find using Lemma 11 that, for $\xi \sim \mathcal{N}(\theta_i, 1)$,

$$\max_{l=0,\dots,l_0} \mathbf{E}\hat{P}^2_{K_l,M_l}(\xi)\,\mathbf{P}(|\xi| \le t_{l_0}) \le A\|F\|^2_{\infty,[-M_{l_0},M_{l_0}]}\, 6^{K_{l_0}} e^{\frac{c\log(1+8/c)}{8}\theta_i^2}\, e^{-\theta_i^2/8} \tag{63}$$

$$\le Ae^{(8\epsilon_1 + \frac{c\log 6}{2} + \frac{c\log(1+8/c)}{2} - \frac{1}{2})t_{l_0}^2}. \tag{64}$$

Choosing $c, \epsilon_1, \epsilon_2$ small enough allows us to obtain the desired bound

$$s \max_{2t_0<|\theta_i|\le 2t_L} V_i \le As^2 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}. \tag{65}$$

4.3°. *Case* $2t_L < |\theta_i| \le \sqrt{2\log(d)}$. Similar arguments as in the previous case yield that

$$s \max_{2t_L<|\theta_i|\le\sqrt{2\log(d)}} V_i \le As^2 \max_{l=0,\dots,L+1} \delta^2_{K_l,M_l}. \tag{66}$$

The result of the theorem follows.

4.2 Proof of Theorem 2

By construction, we have

$$\mathbf{E}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 \le d^2\delta_{K,M}^2 + \mathbf{Var}\big(\hat{\boldsymbol{F}}\big). \tag{67}$$

To bound the variance, we write

$$\hat{\boldsymbol{F}} = \sum_{k=0}^{K} a_{k,K,M} S_k, \quad S_k = \sum_{i=1}^{d} H_k(y_i), \tag{68}$$

so that

$$\mathbf{Var}(\hat{\boldsymbol{F}}) \le \Big(\sum_{k=0}^{K} |a_{k,K,M}| \sqrt{\mathbf{Var}(S_k)}\Big)^2, \tag{69}$$

since for any random variables $X_1, \dots, X_n$, we have

$$\mathbf{E}\Big(\sum_{i=1}^{n} X_i\Big)^2 \le \Big(\sum_{i=1}^{n} \sqrt{\mathbf{E}(X_i^2)}\Big)^2. \tag{70}$$

Furthermore, by Lemmas 6 and 8,

$$\sum_{k=0}^{K} |a_{k,K,M}| \sqrt{\mathbf{Var}(S_k)} \le A\sqrt{d}\|F\|_{\infty,[-M,M]}\, K(1+\sqrt{2})^K \Big(1 + \frac{K}{M^2}\Big)^{K/2}. \tag{71}$$

Using the definition of K, we have

$$K \log(1 + K/M^2) \le Ac \log(d), \tag{72}$$

hence, taking c small enough implies that

$$\mathbf{Var}(\hat{\boldsymbol{F}})\delta_{K,M}^{-2} \le AK^2(1+\sqrt{2})^{2K} d^{2\epsilon_1 + 2\epsilon_2 + 1 + Ac} \le Ad^2. \tag{73}$$

The result follows.

4.3 Proof of Theorem 3

Preliminary: By Markov's inequality, we have for every $K > 0$

$$\inf_{\hat{\boldsymbol{F}}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbf{E}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})\big)^2 \ge \frac{s^2\delta_{K,M}^2}{4} \inf_{\hat{\boldsymbol{F}}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbf{P}_{\boldsymbol{\theta}}\Big(|\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})| \ge s\delta_{K,M}/2\Big), \tag{74}$$

and Theorem 2.15 in Tsybakov (2009) implies that for any prior measures $\bar{\mu}_0$ and $\bar{\mu}_1$ concentrated on $\boldsymbol{\Theta}$

$$\inf_{\hat{\boldsymbol{F}}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbf{P}_{\boldsymbol{\theta}}\Big(|\hat{\boldsymbol{F}} - \boldsymbol{F}(\boldsymbol{\theta})| \ge \frac{m_1 - m_0}{4}\Big) \ge \frac{1 - V}{2} \tag{75}$$

with

$$V = \mathrm{TV}(\bar{\mathbf{P}}_0, \bar{\mathbf{P}}_1) + \bar{\mu}_0\big(\boldsymbol{F}(\boldsymbol{\theta}) \ge m_0 + 3v_0\big) + \bar{\mu}_1\big(\boldsymbol{F}(\boldsymbol{\theta}) \le \frac{m_0 + m_1}{2} + 3v_0\big), \tag{76}$$

where TV denotes the total-variation distance, and for $i = 0, 1$, $\bar{\mathbf{P}}_i$ is defined for every measurable set by

$$\bar{\mathbf{P}}_i(A) = \int_{\mathbb{R}^d} \mathbf{P}_{\boldsymbol{\theta}}(A)\, \bar{\mu}_i(d\boldsymbol{\theta}) \tag{77}$$

and m_0, m_1, v_0 are to be chosen later.

Construction of the prior measures: First we choose

$$K = \frac{e^2 \log(s^2/d)}{\log(e \log(s^2/d)/M^2)}, \tag{78}$$

and we denote μ_i for $i \in \{0, 1\}$ the distribution of the random vector $\boldsymbol{\theta} \in \mathbb{R}^d$ with independent components distributed as $\epsilon\eta_i$, where ϵ and η_i are independent, $\epsilon \sim \mathcal{B}\big(s/(2d)\big)$ and η_i is distributed according to ν_i defined in Lemma 7. Then, we define probabilities $\mathbf{P}_0$ and $\mathbf{P}_1$ by

$$\mathbf{P}_i(A) = \int_{\mathbb{R}^d} \mathbf{P}_{\boldsymbol{\theta}}(A)\, \mu_i(d\boldsymbol{\theta}), \tag{79}$$

for all measurable sets A. The densities of these probabilities with respect to the Lebesgue measure on $\mathbb{R}^d$ are given by

$$f_i(x) = \prod_{i=1}^{d} g_i(x_i), \tag{80}$$

where

$$g_i(x) = \frac{s}{2d}\phi_i(x) + \Big(1 - \frac{s}{2d}\Big)\phi(x), \tag{81}$$

and

$$\phi_i(x) = \int_{\mathbb{R}} \phi(x - t)\, \nu_i(dt), \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{82}$$

But as the μ_i's are not supported on $\boldsymbol{\Theta}$, we define counterparts $\bar{\mu}_i$'s by

$$\bar{\mu}_i(A) = \frac{\mu_i(A \cap \boldsymbol{\Theta})}{\mu_i(\boldsymbol{\Theta})}. \tag{83}$$

Finally, we denote

$$m_i = \int_{\mathbb{R}^d} \boldsymbol{F}(\boldsymbol{\theta})\, \mu_i(d\boldsymbol{\theta}), \quad v_i^2 = \int_{\mathbb{R}^d} (\boldsymbol{F}(\boldsymbol{\theta}) - m_i)^2\, \mu_i(d\boldsymbol{\theta}). \tag{84}$$

Bounding the probabilities in (76): According to Lemma 7, we have

$$m_1 - m_0 = d \times \frac{s}{2d} \times \Big(\int_{-M}^{M} F(t)\, \nu_1(dt) - \int_{-M}^{M} F(t)\, \nu_0(dt) \Big) = s\delta_{K,M}. \tag{85}$$

Using Lemma 9 in Collier et al. (2019) and Chebyshev-Cantelli's inequality, we have for d large enough

$$\bar{\mu}_0\big(\boldsymbol{F}(\boldsymbol{\theta}) \geq m_0 + 3v_0\big) \leq \mu_0\big(\boldsymbol{F}(\boldsymbol{\theta}) \geq m_0 + 3v_0\big) + e^{-s/16} \tag{86}$$

$$\leq \frac{v_0^2}{v_0^2 + (3v_0)^2} + e^{-s/16} < \frac{1}{5}. \tag{87}$$

Now, we notice that for $i \in \{0,1\}$, we have

$$v_i^2 \le d\|F\|^2_{\infty,[-M,M]}, \tag{88}$$

so that for C large enough,

$$\frac{m_0+m_1}{2} + 3v_0 - m_1 \le 3\sqrt{d}\|F\|_{\infty,[-M,M]} - \frac{s\delta_{K,M}}{2} \le -\frac{s\delta_{K,M}}{3}, \tag{89}$$

since the assumptions of the theorem imply that

$$\frac{\sqrt{d}}{s\delta_{K,M}}\|F\|_{\infty,[-M,M]} \le A\Big(\frac{s^2}{d}\Big)^{\epsilon_1+\epsilon_2-1/2}. \tag{90}$$

Consequently,

$$\bar{\mu}_1\big(\boldsymbol{F}(\boldsymbol{\theta}) \le \frac{m_0+m_1}{2} + 3v_0\big) \le \mu_1\big(\boldsymbol{F}(\boldsymbol{\theta}) - m_1 \le -\frac{s\delta_{K,M}}{3}\big) + e^{-s/16} \tag{91}$$

$$\le \frac{9v_1^2}{9v_1^2 + s^2\delta^2_{K,M}} + e^{-s/16}, \tag{92}$$

by Chebyshev-Cantelli's inequality, and the last quantity is smaller than

$$\frac{9d\|F\|^2_{\infty,[-M,M]}}{9d\|F\|^2_{\infty,[-M,M]} + s^2\delta^2_{K,M}} + e^{-s/16}. \tag{93}$$

Finally, we use again the fact that $d\|F\|^2_{\infty,[-M,M]}/(s^2\delta^2_{K,M}) \le A(d/s^2)^{1-2\epsilon_1-2\epsilon_2}$ with $s^2/d > C$, so that for C large enough,

$$\bar{\mu}_1\big(\boldsymbol{F}(\boldsymbol{\theta}) \le \frac{m_0+m_1}{2} + 3v_0\big) < \frac{1}{5}. \tag{94}$$

Bounding the total-variation distance in (76): We can upper bound the total-variation distance as follows:

$$\mathrm{TV}(\bar{\mathbf{P}}_0, \bar{\mathbf{P}}_1) \le \mathrm{TV}(\bar{\mathbf{P}}_0, \mathbf{P}_0) + \mathrm{TV}(\mathbf{P}_0, \mathbf{P}_1) + \mathrm{TV}(\mathbf{P}_1, \bar{\mathbf{P}}_1) \tag{95}$$

$$\le \sqrt{\chi^2(\mathbf{P}_0, \mathbf{P}_1)/2} + \mu_0(\boldsymbol{\Theta}^{\complement}) + \mu_1(\boldsymbol{\Theta}^{\complement}), \tag{96}$$

where $\boldsymbol{\Theta}^{\complement}$ denotes the complement of $\boldsymbol{\Theta}$. As before,

$$\mu_i(\boldsymbol{\Theta}^{\complement}) \le \mathbf{P}\Big(\mathcal{B}\big(d, \frac{s}{2d}\big) > s\Big) \le e^{-\frac{s}{16}}. \tag{97}$$

Furthermore, since the $\mathbf{P}_i$'s are product measures, we have

$$\chi^2(\mathbf{P}_0, \mathbf{P}_1) = \Big(1 + \int \frac{(g_1-g_0)^2}{g_0}\Big)^d - 1, \tag{98}$$

and by the definition of g_0, g_1,

$$\int \frac{(g_1-g_0)^2}{g_0} \le \frac{1}{1-\frac{s}{2d}}\Big(\frac{s}{2d}\Big)^2 \int \frac{(\phi_1-\phi_0)^2}{\phi} \le \frac{s^2}{2d^2}\int \frac{(\phi_1-\phi_0)^2}{\phi}. \tag{99}$$

Then

$$\int \frac{(\phi_1-\phi_0)^2}{\phi} = \int e^{\theta\theta'}\nu_1(d\theta)\nu_1(d\theta') + \int e^{\theta\theta'}\nu_0(d\theta)\nu_0(d\theta') - 2\int e^{\theta\theta'}\nu_0(d\theta)\nu_1(d\theta') \tag{100}$$

$$= \sum_{k\geq 0}\frac{1}{k!}\Big(\int t^k\nu_1(dt) - \int t^k\nu_0(dt)\Big)^2 \tag{101}$$

$$\leq 4\sum_{k\geq K+1}\frac{M^{2k}}{k!}, \tag{102}$$

and the choice of K along with the condition on M imply that $eM^2/K \leq 1/e$, so that

$$\int \frac{(\phi_1-\phi_0)^2}{\phi} \leq 4\sum_{k\geq K+1}\Big(\frac{eM^2}{k}\Big)^k \leq 4\Big(\frac{eM^2}{K}\Big)^K. \tag{103}$$

Coming back to the χ^2-distance and using the fact that $1+x\leq e^x$ for every $x\in\mathbb{R}$, we get

$$\chi^2(\mathbf{P}_0,\mathbf{P}_1) \leq \exp\Big[\frac{2s^2}{d}\Big(\frac{eM^2}{K}\Big)^K\Big] - 1. \tag{104}$$

Finally, we compute

$$K\log\Big(\frac{eM^2}{K}\Big) = -e^2\log(s^2/d)\times g\Big(e\log(s^2/d)/M^2\Big), \tag{105}$$

where

$$g(x) = \frac{\log\Big(\frac{x}{\log(x)}\Big)}{\log(x)}, \tag{106}$$

and it can be shown that $g>0.5$, so that $\chi^2(\mathbf{P}_0,\mathbf{P}_1)\leq e^{2(d/s^2)^{e^2/2-1}}-1$. This inequality, combined with (97), yields

$$\mathrm{TV}(\bar{\mathbf{P}}_0,\bar{\mathbf{P}}_1) < 3/5 \tag{107}$$

if C and d are large enough.

The proof is completed by gathering (74), (85), (75), (76), (86), (94) and (107).

4.4 Proof of Theorem 5

If $l\in\{0,\dots,L+1\}$, then by definition of K_l in (15), we have

$$K_l \leq \frac{c}{4}\log(d) \leq \frac{c}{4}\log(s^2/C) \leq \log(s) \tag{108}$$

for c small enough. Besides, if $l_0 = \lfloor\frac{\log_2(4/c)}{2}\rfloor + 1$, where $\lfloor\cdot\rfloor$ denotes the integer part, then

$$\forall l\geq l_0,\quad K_l \geq \log(s^2/d). \tag{109}$$

On the other hand, when $k \in \{s, \dots, d\}$, the quantity $\log(s^2/k)$ ranges from $\log(s^2/d)$ to $\log(s)$ and the consecutive differences satisfy

$$\log\left(s^2/k\right) - \log\left(s^2/(k+1)\right) = \log(1+1/k) \in [0,1], \tag{110}$$

so that for every $l \in \{l_0, \dots, L+1\}$, there exists an integer $k_l \in \{s, \dots, d\}$ such that

$$|K_l - \log(s^2/k_l)| \leq 1. \tag{111}$$

Now note that $\log(s^2/k_l) \geq \log(C)$, which yields that, for every $l \in \{l_0, \dots, L+1\}$,

$$\frac{K_l}{\log(s^2/k_l)} = 1 + \frac{K_l - \log(s^2/k_l)}{\log(s^2/k_l)} \in \Big[1 - \frac{1}{\log(C)}, 1 + \frac{1}{\log(C)}\Big]. \tag{112}$$

But for $l \in \{0, \dots, l_0 - 1\}$, we have

$$1 \leq \frac{K_l}{K_0} \leq \frac{4}{c}, \tag{113}$$

so that the last two displays, combined with Assumption (A3), entail that

$$\max_{l=0,\dots,L+1} \delta^2_{K_l,M_l} \leq A \max_{l=0,\dots,L+1} \delta^2_{\log(s^2/k_l),\sqrt{\log(s^2/k_l)}} \leq A \max_{k=s,\dots,d} \delta^2_{\log(s^2/k),\sqrt{\log(s^2/k)}}. \tag{114}$$

Finally, we conclude by Assumption (A3) again, since

$$\max_{k=s,\dots,d} \delta^2_{\log(s^2/k),\sqrt{\log(s^2/k)}} \leq A \max_{k=s,\dots,d} \delta^2_{e^2\log(s^2/k),\sqrt{\log(s^2/k)}}. \tag{115}$$

5. Lemmas

In the whole section, we denote by A an absolute positive constant that precise value may vary from line to line.

The following lemma is a direct consequence of Proposition 2 in Collier et al. (2019).

Lemma 6 *Let $P_{K,M}$ be the polynomial defined in (8). Then the coefficients $a_{k,K,M}$ in (10) satisfy*

$$|a_{k,K,M}| \leq A\|F\|_{\infty,[-M,M]} M^{-k}(1+\sqrt{2})^K, \quad k = 0, \dots, K. \tag{116}$$

The following lemma is a slight modification of Lemma 1 in Cai and Low (2011):

Lemma 7 *Assume that F is continuous on $[-M, M]$, then for every positive integer K, if $\delta_{K,M} > 0$, there exist measures ν_0, ν_1 on $[-M, M]$ such that*

$$\begin{cases} \int t^l \nu_0(dt) = \int t^l \nu_1(dt), \quad l = 0, \dots, K \\ \int F(t)\nu_0(dt) - \int F(t)\nu_1(dt) = 2\delta_{K,M}. \end{cases} \tag{117}$$

Proof Denote $\mathcal{C}$ the set of continuous functions on $[-M, M]$ equipped with the uniform norm, and $\mathcal{F}_k$ be the linear space spanned by $\mathcal{P}_K$ (the set of polynomials of degree smaller than K) and F. Note that F does not belong to $\mathcal{P}_K$, since by assumption, $\delta_{K,M} > 0$. Then every element g of $\mathcal{F}_K$ can be represented as $g = cF + P$, where $P \in \mathcal{P}_K$ and $c \in \mathbb{R}$. Then

we can define the linear functional T on $\mathcal{F}_K$ by $T(g) = c\delta_{K,M}$. We then compute the norm of T defined as

$$\|T\| = \sup\{T(g) \mid \|g\|_\infty = 1\}. \tag{118}$$

Now, every $g \in \mathcal{F}_K$ satisfying $\|g\|_\infty = 1$ can be written as

$$g = \frac{cF + P}{\|cF + P\|_\infty}, \quad P \in \mathcal{P}_K, \tag{119}$$

so that

$$\|T\| = \sup_{c,P} \frac{c\delta_{K,M}}{\|cF + P\|_\infty} = \sup_P \frac{\delta_{K,M}}{\|F - P\|_\infty} = 1 \tag{120}$$

by definition of $\delta_{K,M}$. Then, using Hahn-Banach and Riesz representation theorems, we can extend T on $\mathcal{C}$ without changing its norm, and represent this extension $\tilde{T}$ as

$$\tilde{T}(g) = \int_{-M}^{M} g(t)\,\tau(dt), \tag{121}$$

where τ is a signed measure with total variation 1. Then, using Jordan decomposition, we can write τ as a difference of two positive measures

$$\tau = \tau^+ - \tau^-. \tag{122}$$

Denoting $\nu_0 = 2\tau^+$ and $\nu_1 = 2\tau^-$, which are probability measures since 2τ has total variation 2 and $\int_{-M}^{M} \tau(dt) = 0$, the last properties of the lemma follow from the properties of τ. ■

The proof of the next lemma can be found in Cai and Low (2011).

Lemma 8 *Let $\theta \in \mathbb{R}$ and $X \sim \mathcal{N}(\theta, 1)$. For any $k \in \mathbb{N}$, the k-th Hermite polynomial satisfies*

$$\mathbf{E}H_k(X) = \theta^k, \tag{123}$$

$$\mathbf{E}H_k^2(X) \le \left(k + \theta^2\right)^k. \tag{124}$$

Lemma 9 *Let $\hat{P}_{K,M}$ be defined in (15) with $K \le M^2$. If $\xi \sim \mathcal{N}(0,1)$, then*

$$\mathbf{E}\hat{P}_{K,M}^2(\xi) \le A\|F\|_{\infty,[-M,M]}^2\, 6^K. \tag{125}$$

Proof Recall that, for the Hermite polynomials, $\mathbf{E}(H_k(\xi)H_j(\xi)) = 0$ if $k \ne j$ and $\xi \sim \mathcal{N}(0,1)$. Using this fact and then Lemmas 6 and 8 we obtain

$$\mathbf{E}\hat{P}_{K,M}^2(\xi) = \sum_{k=1}^{K} a_{k,K,M}^2 \mathbf{E}H_k^2(\xi) \le A\|F\|_{\infty,[-M,M]}^2\,(1+\sqrt{2})^{2K} \sum_{k=1}^{K} (k/M^2)^k. \tag{126}$$

Moreover, since $K/M^2 \le 1$, we have $\sum_{k=1}^{K}(k/M^2)^k \le K$. The result follows. ■

Lemma 10 *Let $\hat{P}_{K,M}$ be defined in (15) with parameters $K = cM^2/8$ and $c \leq 8$. If $\xi \sim \mathcal{N}(\theta, 1)$ with $|\theta| \leq M$, then*

$$\mathbf{E}\hat{P}^2_{K,M}(X) \leq A\|F\|^2_{\infty,[-M,M]}\, 12^K. \tag{127}$$

Proof We use the bound

$$\mathbf{E}\hat{P}^2_{K,M}(\xi) \leq \left(\sum_{k=1}^{K} |a_{k,K,M}| \sqrt{\mathbf{E}H_k^2(\xi)}\right)^2. \tag{128}$$

Thus Lemma 8 in particular and the fact that $K \leq M^2$ imply that

$$\mathbf{E}\hat{P}^2_{K,M}(\xi) \leq A\|F\|^2_{\infty,[-M,M]}(1+\sqrt{2})^{2K}\left(\sum_{k=1}^{K} M^{-k}2^{k/2}M^k\right)^2 \leq A\|F\|^2_{\infty,[-M,M]}\, 12^K. \tag{129}$$

■

Lemma 11 *Let $\hat{P}_{K,M}$ be defined in (15) with $K = cM^2/8$ and $c \leq 8$. If $\xi \sim \mathcal{N}(\theta, 1)$ with $|\theta| > M$, then*

$$\left|\mathbf{E}\hat{P}_{K,M}(\xi)\right| \leq A\|F\|_{\infty,[-M,M]}\, 3^K e^{c\theta^2/16}, \tag{130}$$

$$\mathbf{E}\hat{P}^2_{K,M}(\xi) \leq A\|F\|^2_{\infty,[-M,M]}\, 6^K e^{\frac{c\log(1+8/c)}{8}\theta^2}. \tag{131}$$

Proof To prove the first inequality of the lemma, we use Lemma 6 to obtain

$$\left|\mathbf{E}\hat{P}_{K,M}(\xi)\right| \leq A\|F\|_{\infty,[-M,M]} K(1+\sqrt{2})^K \left(\frac{|\theta|}{M}\right)^K, \tag{132}$$

and the result follows from

$$K\log(|\theta|/M) = \frac{cM^2}{8}\log(|\theta|/M) \leq c\theta^2/16. \tag{133}$$

We now prove the second inequality of the lemma. Using (128) and then Lemmas 6 and 8 we get

$$\mathbf{E}\hat{P}^2_{K,M}(\xi) \leq A\|F\|^2_{\infty,[-M,M]}\,(1+\sqrt{2})^{2K}\left(\sum_{k=1}^{K} M^{-k}(k+\theta^2)^{k/2}\right)^2. \tag{134}$$

But as $\frac{\theta^2}{k} \geq \frac{M^2}{K} = \frac{8}{c} \geq 1$, we can use the fact that the function $x \to x^{-1}\log(1+x)$ is decreasing on $\mathbb{R}_+^*$ to obtain that

$$k\log\left(1+\frac{\theta^2}{k}\right) \leq \frac{c\theta^2\log(1+8/c)}{8}. \tag{135}$$

Therefore,

$$\mathbf{E}\hat{P}^2_{K,M}(\xi) \leq A\|F\|^2_{\infty,[-M,M]}\,(1+\sqrt{2})^{2K} e^{\frac{c\log(1+8/c)}{8}\theta^2}\left(\sum_{k=1}^{K}(k/M^2)^{k/2}\right)^2. \tag{136}$$

Finally, the result follows since $K \leq M^2$. ■

Acknowledgments

We thank A.B. Tsybakov for fruitful discussions during the redaction of this paper. Olivier Colliers research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

References

P. Bickel and Y. Ritov. Estimating integrated squares density derivatives: sharp best order of convergence estimates. *Sankhya Ser. A*, 50:381–393, 1988.

L. Birgé and P. Massart. Estimation of integral functionals of a density. *Annals of Statistics*, 23:11–29, 1995.

T.T. Cai and M.G. Low. Testing composite hypothesis, Hermite polynomials and optimal estimation of a nonsmooth functional. *Annals of Statistics*, 39(2):1012–41, 2011.

A. Carpentier, O. Collier, L. Comminges, A.B. Tsybakov, and Y. Wang. Minimax rate of testing in sparse linear regression. *To appear in Automation and Remote Control*, 2018.

O. Collier and A. Dalalyan. Rate-optimal estimation of p-dimensional linear functionals in a sparse Gaussian model. *Preprint*, 2019.

O. Collier, L. Comminges, and A.B. Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *Annals of Statistics*, 45(3):923–58, 2017.

O. Collier, L. Comminges, and A.B. Tsybakov. On estimation of nonsmooth functionals of sparse normal means. *Preprint*, 2019.

L. Comminges, O. Collier, M. Ndaoud, and A.B. Tsybakov. Adaptive robust estimation in sparse vector model. *Preprint*, 2019.

K. Fukuchi and J. Sakuma. Minimax optimal estimators for additive scalar functionals of discrete distributions. *Preprint*, 2019.

Y. Han, J. Jiao, and T. Weissman. Minimax estimation of discrete distributions under l1 loss. *IEEE Transactions on Information Theory*, 61(5):2835–85, 2015a.

Y. Han, J. Jiao, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(11):6343–54, 2015b.

G. Kerkyacharian and D. Picard. Estimating nonquadratic functionals of a density using Haar wavelets. *Annals of Statistics*, 24(2):485–507, 1996.

V. Koltchinskii and M. Zhilova. Efficient estimation of smooth functionals in Gaussian shift models. *Preprint*, 2018.

Y. Polyanskiy and Y. Wu. Dualizing Le Cam's method, with applications to estimating the unseens. *Preprint*, 2019.

A.F. Timan. *Theory of approximation of functions of a real variable*. Pergamon, 1st edition, 1963. ISBN 9781483184814.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 0387790519, 9780387790510.

Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–20, 2016.

Y. Wu and P. Yang. Sample complexity of the distinct element problem. *Mathematical Statistics and Learning*, 1(1):37–72, 2018.