Pointwise Adaptive M-estimation in Nonparametric Regression

Michaël CHICHIGNOUD[1]

[1] *Université Aix-Marseille 1, LATP*
39, rue F. Joliot Curie, 13453 Marseille Cedex 13, France.
E-mail: chichign@cmi.univ-mrs.fr

Abstract

This paper deals with the nonparametric estimation in heteroscedastic regression $Y_i = f(X_i) + \xi_i,\ i = 1, \ldots, n$, with incomplete information, i.e. each real random variable ξ_i has a density g_i which is unknown to the statistician. The aim is to estimate the regression function f at a given point. Using a local polynomial fitting from M-estimator denoted $\hat{f}^h$ and applying Lepski's procedure for the bandwidth selection, we construct an estimator $\hat{f}^{\hat{h}}$ which is adaptive over the collection of isotropic Hölder classes. In particular, we establish new exponential inequalities to control deviations of local M-estimators allowing to construct the minimax estimator. The advantage of this estimator is that it does not depend on densities of random errors and we only assume that the probability density functions are symmetric and monotonically on $\mathbb{R}_+$. It is important to mention that our estimator is robust compared to extreme values of the noise.

Key words: adaptation, Huber function, Lepski's method, M-estimation, minimax estimation, nonparametric regression, robust estimation, pointwise estimation.

AMS 2000 subject classification: Primary 62G08; secondary 62G20, 62G35.

1 Introduction

Let the statistical experiment be generated by the observation $Z^{(n)} = (X_i, Y_i)_{i=1,\dots n}, n \in \mathbb{N}^*$, where each (X_i, Y_i) satisfies the equation

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n. \tag{1.1}$$

Here $f : [0,1]^d \to \mathbb{R}$ is an unknown function to be estimated at a given point $x_0 \in [0,1]^d$ from the observation $Z^{(n)}$.

The real random variables $(\xi_i)_{i\in 1,\dots,n}$ (the noise) are supposed to be independent and each variable ξ_i has a symmetric density $g_i(\cdot)$, with respect to the Lebesgue measure on $\mathbb{R}$. We also assumed that g_i is monotonically on $\mathbb{R}_+$ for any i.

The design points $(X_i)_{i\in 1,\dots,n}$ are independent and uniformly distributed on $[0,1]^d$. The random vectors $(X_i)_{i\in 1,\dots,n}$ and $(\xi_i)_{i\in 1,\dots,n}$ are independent.

Along the paper, the unknown function f is supposed to be smooth, in particular, it belongs to an isotropic Hölder ball of functions $\mathbb{H}_d(\beta, L, M)$ (cf. Definition 1 below). Here $\beta > 0$ is the smoothness of f, $L > 0$ is the Lipschitz constant and M is an upper bound of f and its partial derivatives.

Motivation. In this paper, the considered problem is the robust nonparametric estimation, i.e. the estimation of the regression function f in the presence of a heavy-tailed noise (cf. Rousseeuw and Leroy [1987] and Huber and Ronchetti [2009]). Well-known examples are when the noise distribution is for instance Laplace (no finite exponential's moment) or Cauchy (no finite order's moments). Moreover, we assume that the noise densities $\left(g_i\right)_{i=1,\dots,n}$ are unknown to the statistician. This problem has popular applications, for example in relative GPS positioning (cf. Chang and Guo [2005]) or in robust image denoising (cf. Astola, Egiazarian, Foi, and Katkovnik [2010]).

In parametric case, we consider f as a constant parameter $\theta \in \mathbb{R}$. The use of empiric criteria is very popular, i.e. the minimization of the following *contrast function* ρ:

$$\hat\theta = \arg\min_{t\in[-M,M]} \sum_{i=1}^{n} \rho(Y_i - t),$$

The most famous contrast functions are the square function $\rho(z) = z^2$ ($\hat\theta$ become the empiric mean), the absolute value function $\rho(z) = |z|$ ($\hat\theta$ become the empiric mean) and the *Huber function*, as defined in (5.3), without an explicit expression of $\hat{\theta}$ (cf. Huber [1964]). It is well known that the square function leads to the empiric mean which does not fit with a heavy-tailed noise. Thus the square function is not suitable in the model (1.1).

In nonparametric estimation, we propose a *local parametric approach* (LPA) to estimate the regression function at a given point $x_0 \in [0,1]^d$ in the model (1.1). We suppose that f is locally almost polynomial (with degree $b \in \mathbb{N}$) and we use the parametric estimator on a neighborhood denoted $V_{x_0}(h)$ The parameter is reconstructed from the following criterion, for any $x_0 \in [0,1]^d$ and $h \in [0,1]$

$$\hat{\theta} = \arg \min_{t \in [-M,M]^{N_b}} \sum_{i : X_i \in V_{x_0}(h)} \rho\big(Y_i - f_t(X_i)\big)\, K\left(\frac{X_i - x_0}{h}\right). \tag{1.2}$$

where $f_t(\cdot)$ is a polynomial of degree b with coefficients t, $K(\cdot)$ is a *kernel function* and N_b is the number of partial derivatives of f of order smaller than b. We refer to $\hat{f}^h(x_0) = f_{\hat{\theta}}(x_0)$ as the ρ*-LPA estimator*. It belongs to the family of M-estimators and it relies on a local scale parameter h, called the *bandwidth*. A crucial issue is the optimal choice of the parameter h. To adress it we use quite standard arguments based on the *bias/variance trade-off* (cf. (1.5) below) in minimax case and the *Lepski's rule* for the *data-driven selection* in adaptation. First, since f is smooth ($f \in \mathbb{H}_d(\beta, L, M)$), cf. Definition 1 below) we notice that

$$\exists \theta = \theta(f, x_0, h) \in \Theta\big(M\big) : \quad b_h := \sup_{x \in V_{x_0}(h)} \big|f(x) - f_\theta(x)\big| \leq L d h^\beta. \tag{1.3}$$

We can choose θ as the coefficients of Taylor polynomial as defined in (2.6). Thus, if h is chosen sufficiently small our original model (1.1) is well approximated inside of $V_{x_0}(h)$ by the "parametric" model

$$\mathcal{Y}_i = f_\theta(X_i) + \xi_i, \quad \forall i \,:\, X_i \in V_{x_0}(h). \tag{1.4}$$

With this model, the ρ*-LPA estimator* $\hat{\theta}$ achieves the usual parametric rate of convergence $1/\sqrt{nh^d}$, where nh^d is the number of the observations in the neighborhood $V_{x_0}(h)$ (See Theorem 1, Section 3).

This approach has been introduced by Katkovnik [1985] and used for the first time in robust nonparametric estimation by Tsybakov [1986], Härdle and

Tsybakov [1988] and Hall and Jones [1990] to obtain asymptotic normality and minimax results. We also notice that Tsybakov[1982a,1982b,1983] obtained similar results to estimate the locally almost constant functions.

Minimax Estimation. To guarantee good performance of the ρ-LPA estimator in the minimax sense, we assume that ρ' is bounded and Lipschitz. On the other hand, the Huber function satisfies these assumptions, making it suitable for our problem. Moreover, it is commonly used in practice (see for instance Petrus [1999] and Chang and Guo [2005]). As for *linear estimators* (kernel estimators, least square estimators, etc.), a good choice of the bandwidth $h = \bar{h}_n(\beta, L)$ provides an optimal ρ-LPA estimator over the Hölder space $\mathbb{H}_d(\beta, L, M)$. Finally, $\bar{h}_n(\beta, L) = (L^2 n)^{-\frac{1}{2\beta+d}}$ is chosen as the solution of the following bias/variance trade-off

$$\left(nh^d\right)^{-1/2} + Lh^\beta \to \min_h . \tag{1.5}$$

In the model (1.1), we show that the corresponding estimator $\hat{f}^{\bar{h}_n(\beta,L)}(x_0)$ achieves the rate of convergence $n^{-\beta/(2\beta+d)}$ (cf. Definition 2) for $f(x_0)$ on $\mathbb{H}_d(\beta, L, M)$ (See Theorem 1). We should point out that both the knowledge of β and L is required to the statistician in order to built the optimal bandwidth $\bar{h}_n(\beta, L)$.

Adaptive Estimation. In nonparametric statistics, an important problem is the adaptation compared to the smoothness parameters β and L that are unknown in practice. This requests to develop a data-driven (adaptive) selection to choose the bandwidth. Then, the interesting feature is the selection of estimators from a given family (cf. Barron, Birgé, and Massart [1999], Lepski, Mammen, and Spokoiny [1997], Goldenshluger and Lepski [2008]). In this context, several approaches to the selection from the family of linear estimators were recently proposed, see for instance Goldenshluger and Lepski [2008], Goldenshluger and Lepski [2009], Juditsky, Lepski, and Tsybakov [2009] and the references therein. However, those methods strongly rely on the linearity property. Robust estimators are generally non-linear, there standard arguments (like the bias/variance trade-off) cannot be applied straightforwardly. For instance, Brown, Cai, and Zhou [2008] use the asymptotic normality of the median to approximate the model (1.1) by the *wavelet sequence data* and they use *BlockJS wavelet thresholding* for adaptation over Besov spaces with the integrated risk. Recently, Reiss, Rozenholc, and Cuenod [2011] have considered the pointwise estimation for locally almost constant functions in the homoscedastic regression with a heavy-tailed noise. That corresponds to $\beta \le 1$ for the Hölder functions in the model (1.1) (cf. also Definition 1). They have considered the symmetric and continuous density with $g(0) > 0$.

In the context of adaptation, other new points are developed in this paper:
– adaptative pointwise estimation for any regularity β of isotropic functions,
– random design and heteroscedastic model,
– unknown and heavy-tailed noise.

For it, we construct an adaptive estimator (cf. Definition 3) using general adaptation scheme due to Lepski [1990] (*Lepski's method*). This method is applied to choose the bandwidth of the ρ-LPA estimator in the model (1.1).

We remind that M, the upper bound of f and its partial derivatives, is involved in the construction of the ρ-LPA estimator (1.2). Then, we assume that the parameter M is known and we do not study the adaptation compared to it. Contrary to the constants β, L, one could estimate M to "inject" it in the procedure without loss of generality in the performance of our estimator (cf. Härdle and Tsybakov [1988]).

Exponential Inequality. Lepski's procedure requires, in particular to establish the exponential inequality for the deviations of ρ-LPA estimator. As far as we know, these results seems to be new.

Denote by $\mathbb{P}_f$ the probability law of the observations $Z^{(n)}$ satisfying (1.1). As we mentioned above, we need to establish the following inequality, for any $\varepsilon > 0$ and $h \in (n^{-1/d}, 1)$:

$$\mathbb{P}_f\left(\left|\hat{f}^h(x_0) - f(x_0)\right| \ge \frac{\varepsilon}{\sqrt{nh^d}}\right) \le \mathcal{C}\exp\left\{-\frac{\varepsilon^2}{A + B\varepsilon/\sqrt{nh^d}}\right\}, \qquad (1.6)$$

where $\mathcal{C}$, A, B are positive constants and A, B must be "known". Details are given in Proposition 1. All results of this paper are based on (1.6).

The main difficulty in establishing (1.6) is that the explicit expression of ρ-LPA estimator is not typically available. Let us briefly discuss the main ingredients of M-estimation allowing to prove (1.6). If the derivative of contrast function ρ' is continuous, then solving the minimization problem (1.2) can be viewed as solving the following system of equations in t (first order condition):

$$\forall p, \quad \tilde{D}_h^p(t) := \sum_{i:X_i \in V_{x_0}(h)} \frac{\partial}{\partial t_p} \rho\big(Y_i - f_t(X_i)\big)\, K\left(\frac{X_i - x_0}{h}\right) = 0, \qquad (1.7)$$

where t_p is the p^{th} component of the vector t. Since ρ' is bounded, the partial derivatives $\tilde{D}_h^p(\cdot)$ can be viewed as an empirical process (i.e. a sum of independent and bounded random variables).

Denote $\tilde{D}_h(\cdot)$ the vector of partial derivatives and $D_h(\cdot) = \mathbb{E}_{f_\theta}\tilde{D}_h(\cdot)$ where $\mathbb{E}_{f_\theta} = \mathbb{E}^n_{f_\theta}$ is the mathematical expectation with respect to the probability law $\mathbb{P}_{f_\theta}$ of the "parametric" observations $(X_i, \mathcal{Y}_i)_{i=1,\dots,n}$.

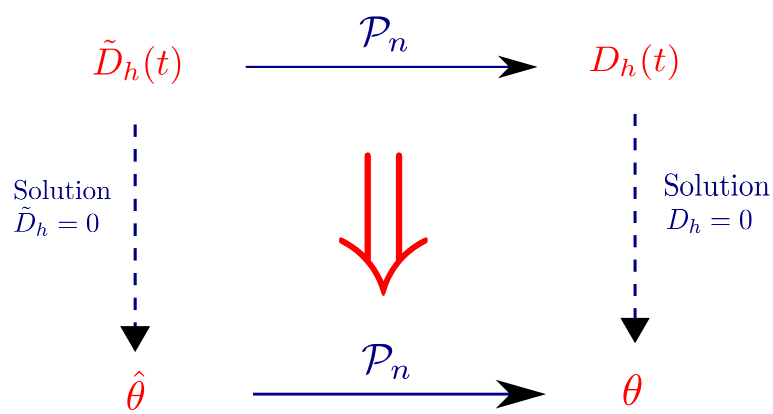

Figure 1: Illustration of the deviations' control. $\mathcal{P}_n$ represent the probability convergence.

Properties of the function $D_h(\cdot)$ allow us to prove that θ is the unique solution of $D_h(\cdot) = 0$. We also notice that $|\hat{f}^h(x_0) - f(x_0)| \leq \|\hat{\theta} - \theta\|_1$. The idea (presented in Figure 1) is to deduce the exponential inequality for $\|\hat{\theta} - \theta\|_1$ from the exponential inequality for $\sup_t \big|\tilde{D}_h(t) - D_h(t)\big|$. As we mentioned above, we notice that $\sup_t \big|\tilde{D}_h(t) - D_h(t)\big|$ can be viewed as the supremum of an empirical process.

Now, classical arguments in probability tools can be used. To control $\sup_t \big|\tilde{D}_h(t) - D_h(t)\big|$, we could used standard tools developed by Talagrand

[1996a, 1996b], Massart [2000] or Bousquet [2002]. But the obtained exponential inequalities (like (1.6)) contain unknown constants or require the knowledge of an expectation's bound of $\sup_t \left|\tilde{D}_h(t) - D_h(t)\right|$. To obtain this bound, we can use the maximal inequalities developed by Van der Vaart and Wellner [1996] (Chapter 2, Section 2.2) for *sub-gaussian processes*. But here again, there are universal constants (and unknown) in the bound of the expectation. Massart [2007] (Chapter 6) gives exponential inequalities for $\sup_t \left|\tilde{D}_h(t) - D_h(t)\right|$ without the expectation, but some constants are very big in our case. In this paper, we choose to apply standard *chaining argument* and *Bernstein's inequality* (cf. (7.8)) directly on $\sup_t \left|\tilde{D}_h(t) - D_h(t)\right|$ (cf. Proof of Lemma 3). That allows us to have constants smaller the ones cited in the papers above.

Perspectives.

– We think that conditions on the noise densities could be reduced. We could consider the densities not necessary monotonically on $\mathbb{R}_+$, only the symmetric assumption seems necessary.
– A possible perspective of this work is the study of estimating anisotropic functions. Indeed, the method developed by Kerkyacharian, Lepski, and Picard [2001], Klutchnikoff [2005] and Goldenshluger and Lepski [2008, 2009] are based on the linear properties and the machinery considered in those works can not adapt straightforwardly to nonlinear estimators.
– Another perspective is to prove an oracle inequality for the family of ρ-LPA estimators indexed by the bandwidth with the integrated risk. It could be interesting to introduce some criterion for choosing the optimal contrast function.
– Finally, we should also study the heteroscedastic model (1.1) with a degenerate design when the design density is vanishing or exploding.

This paper is organized as follows. We present exponential inequalities in Section 2, in order to control deviations of ρ-LPA estimator. In Section 3, we present the results concerning minimax estimation and Section 4 is devoted to the adaptive estimation. An application of ρ-LPA estimator with Huber function is proposed in Section 5. The proofs of the main results (exponential inequalities and upper bounds) are given in Section 6, technical lemmas are postponed to the appendix.

2 Exponential inequality for ρ-LPA estimator

Construction. To construct our estimator, we use the so-called *local polynomial approach* (LPA) which consists in the following. Let

$$V_{x_0}(h) = \bigotimes_{j=1}^{d} \big[(x_0)_j - h/2, (x_0)_j + h/2\big] \cap [0,1]^d,$$

be a neighborhood around x_0 of width $h \in (0,1)$. Fix $b > 0$ (without loss of generality we will assume that b is an integer), let

$$\mathcal{S}_b = \{p = (p_1, \ldots, p_d) \in \mathbb{N}^d : \; 0 \le |p| \le b, \; |p| = p_1 + \ldots + p_d\},$$

and we denote N_b the cardinal of $\mathcal{S}_b$. Let $U(z)$, $z \in \mathbb{R}^d$ be the N_b-dimensional vector of monomials of the following type (the sign $\top$ below denotes the transposition):

$$U^\top(z) = \left(\prod_{j=1}^{d} z_j^{p_j}, \quad p \in \mathcal{S}_b \right). \tag{2.1}$$

For any $t^\top = \big(t_{p_1,\ldots,p_d} \in \mathbb{R} \; : \; p \in \mathcal{S}_b\big) \in \mathbb{R}^{N_b}$, we define the local polynomial in a neighborhood of x_0 as for any $x \in [0,1]^d$

$$f_t(x) \;\; = \;\; t^\top U\left(\frac{x - x_0}{h}\right) \mathbb{I}_{V_{x_0}(h)}(x) = \sum_{p \in \mathcal{S}_b} t_p \, \left(\frac{x - x_0}{h}\right)^p \mathbb{I}_{V_{x_0}(h)}(x), \tag{2.2}$$

where $z^p = z_1^{p_1} \cdots z_d^{p_d}$ for $z = (z_1, \ldots, z_d)$ and $\mathbb{I}$ denotes the indicator function. For any $M > 0$, introduce the following subset of $\mathbb{R}^{N_b}$

$$\Theta\big(M\big) = \big\{ t \in \mathbb{R}^{N_b} : \;\; \|t\|_1 \le M \big\}, \tag{2.3}$$

where $\|\cdot\|_1$ is ℓ_1-norm on $\mathbb{R}^{N_b}$. We notice that for any $t \in \Theta(M)$, $\|f_t\|_\infty \le M$ where $\|\cdot\|_\infty$ is the sup-norm on $[0,1]^d$.

The function ρ is called *contrast function* if it has the following properties.

Assumption 1.

1. *$\rho : \mathbb{R} \to \mathbb{R}_+$ is symmetric, convex and $\rho(0) = 0$,*
2. *the derivative ρ' is 1-Lipschitz on $\mathbb{R}$ and bounded: $\dot\rho_\infty = \|\rho'\|_\infty < \infty$,*

3. the second derivative ρ'' is defined almost everywhere and there exist $L_\rho > 0$ and $\alpha > 0$ such that

$$\sup_{i=\overline{1,n}} \int_{\mathbb{R}} \left|\rho''(z+u) - \rho''(z+v)\right| g_i(z)\, dz \leq L_\rho\, |u-v|^\alpha, \quad \forall u, v \in \mathbb{R}.$$

where $\overline{1,n} = 1, \dots, n$.

A well-known example of a contrast function ρ satisfying Assumption 1 above is the Huber function (cf. Huber [1964]) presented in Section 5.

Let K be a kernel function, i.e. a positive function with a compact support included in $[-1/2, 1/2]^d$ such that $K_\infty := \|K\|_\infty < \infty$ and $\int_{\mathbb{R}} K(z)dz = 1$. We will construct the ρ-LPA estimator for $f(x_0)$ using *local ρ-criterion* which is defined as follows:

$$\tilde{\pi}_h(t) = \tilde{\pi}_h\big(t, Z^{(n)}\big) = \frac{1}{nh^d} \sum_{i=1}^{n} \rho\big(Y_i - f_t(X_i)\big)\, K\left(\frac{X_i - x_0}{h}\right). \tag{2.4}$$

Let $\hat{\theta}(h)$ be the solution of the following minimization problem:

$$\hat{\theta}(h) = \arg\min_{t \in \Theta(M)} \tilde{\pi}_h(t). \tag{2.5}$$

The *ρ-LPA estimator* $\hat{f}^h(x_0)$ of $f(x_0)$ is defined as $\hat{f}^h(x_0) = \hat{\theta}_{0,\dots,0}(h)$. We notice that this local approach can be considered as the estimation for successive derivatives of the function f. However in the present paper, we focus on the estimation of $f(x_0)$.

Exponential inequality. Later on, we will only consider values of $h \in [n^{-1/d}, 1]$. Put $\theta = \theta(f, x_0, h) = \big\{\theta_p : \ p \in \mathcal{S}_b\big\}$, where $\theta_0 = \theta_{0,\dots,0} = f(x_0)$ and

$$\forall p \in \mathcal{S}_b \ : \ p \neq 0, \quad \theta_p = \frac{\partial^{|p|} f(x_0)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \frac{h^{|p|}}{p_1! \dots p_d!}. \tag{2.6}$$

Here, we do not assume the existence of partial derivatives of f. To define θ properly, the following agreement will be used in the sequel: if the function f and the vector p are such that $\partial^{|p|} f$ does not exist, we put $\theta_p = 0$.

Set $\mathcal{B}\big(\theta, z\big) = \{t \in \Theta(M) : \ \|t - \theta\|_2 \leq z\}$ the Euclidean ball with radius z and center θ and define the event for any $h, z > 0$

$$G_z^h \ = \ \big\{\hat{\theta}(h) \in \mathcal{B}\big(\theta, z\big)\big\}, \tag{2.7}$$

where $\hat{\theta}(h)$ is given by (2.5). Let

$$\Sigma = 2 + 2\sum_{l=1}^{\infty} d^2 10^{2l-1} \exp\left\{-\frac{18\,10^l}{\pi^4\,l^4}\frac{(8K_\infty)^{-1}}{K_\infty + 1/3}\right\}, \tag{2.8}$$

be some finite constants and let the constant λ be the smallest eigenvalue of matrix

$$\int_{[-1/2,1/2]^d} U(x)\,U^\top(x)\,K(x)\,dx.$$

Tsybakov [2008] (Lemma 1.4) showed that λ is positive, on the hand the last matrix is strictly positive definite. Finally, put

$$c\big(\rho,(g_i)_i\big) = \inf_{i=1,\ldots,n}\int_{\mathbb{R}} \rho''(z)\,g_i(z)\,dz, \tag{2.9}$$

and define the set of sequences of symmetric densities which are monotonically on $\mathbb{R}_+$

$$\mathcal{G}_\rho^{(c)} = \big\{(g_i)_i \,:\, c\big(\rho,(g_i)_i\big) \ge c\big\}, \quad c > 0. \tag{2.10}$$

Denote for all $a, b \in \mathbb{R}$, $a \vee b = \max(a,b)$. The next proposition is the milestone for all results proved in the paper.

Proposition 1. *Let ρ be a contrast function and let $c > 0$. Then, for any $n \in \mathbb{N}^*$, $(g_i)_i \in \mathcal{G}_\rho^{(c)}$, $h > n^{-1/d}$, $M > 0$, $x_0 \in [0,1]^d$ and any f such that $\|\theta(f,x_0,h)\|_1 \le M$, we have for any $\varepsilon \ge \frac{4N_b}{c\lambda}\big(1 \vee b_h\sqrt{nh^d}\big)$*

$$\begin{aligned}&\mathbb{P}_f\left(\sqrt{nh^d}\big|\hat{f}^h(x_0) - f(x_0)\big| \ge \varepsilon,\, G_\delta^h\right)\\ &\quad\le N_b \Sigma \exp\left\{-\frac{\left(\frac{c\lambda}{2N_b}\,\varepsilon - \big(1 \vee b_h\sqrt{nh^d}\big)\right)^2}{8K_\infty^2(1 \vee \dot{\rho}_\infty^2) + \frac{4\,K_\infty}{3N_b}\,(1 \vee \dot{\rho}_\infty)\frac{c\lambda\,\varepsilon}{\sqrt{nh^d}}}\right\}.\end{aligned} \tag{2.11}$$

The proof of this proposition is given in Section 6.

Remark 1. *The control of the deviations of $\hat{f}^h$ is realized under the event G_δ^h that the estimator $\hat{\theta}(h)$ is contained in a ball centered at θ whereas its radius does not depend on n, else it could change the rate of convergence. In Section 6 we give an exponential inequality to control the probability of the complementary of G_δ^h (cf. Lemma 4).*

Remark 2. *In the minimax case, the knowledge of constants in (2.11) is not required. However for adaptation, the constant c is involved in the construction of adaptive estimator. This restricted the consideration of the noise densities which satisfy (2.10). We notice that this problem is simplified to the calibration of an alone constant with a dataset.*

3 Minimax Results on $\mathbb{H}_d(\beta, L, M)$

In this section, we present several results concerning maximal and minimax risks on $\mathbb{H}_d(\beta, L, M)$. We propose the estimator which bound the maximal risk on this class of functions without restriction imposed on these parameters.

Preliminaries.

Definition 1. *Fix $\beta > 0$, $L > 0$, $M > 0$ and let $\lfloor \beta \rfloor$ be the largest integer strictly smaller than β. The* isotropic Hölder class *$\mathbb{H}_d(\beta, L, M)$ is the set of functions $f : [0,1]^d \to \mathbb{R}$ admitting on $[0,1]^d$ all partial derivatives of order $\lfloor \beta \rfloor$ and such that for any $x, y \in [0,1]^d$*

$$\begin{aligned}\left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} - \frac{\partial^{|p|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \right| &\leq L \big[\|x - y\|_1 \big]^{\beta - \lfloor \beta \rfloor}, \quad \forall \, |p| = \lfloor \beta \rfloor, \\ \sum_{p \in \mathcal{S}_{\lfloor \beta \rfloor}} \sup_{x \in [0,1]^d} \left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| &\leq M.\end{aligned}$$

where x_j and y_j are the j^{th} components of x and y.

Let $\mathbb{E}_f = \mathbb{E}_f^n$ be the mathematical expectation with respect to the probability law $\mathbb{P}_f$ of the observation $Z^{(n)}$ satisfying (1.1). Firstly, we define the maximal risk on $\mathbb{H}_d(\beta, L, M)$ corresponding to the estimation of the function f at a given point $x_0 \in [0,1]^d$.

Let $\tilde{f}_n$ be an arbitrary estimator built from the observation $Z^{(n)}$. For any r let

$$R_{n,r}\big[\tilde{f}_n, \mathbb{H}_d(\beta, L, M)\big] = \sup_{f \in \mathbb{H}_d(\beta, L, M)} \mathbb{E}_f \big| \tilde{f}_n(x_0) - f(x_0) \big|^r. \tag{3.1}$$

This quantity is called *maximal risk* of the estimator $\tilde{f}_n$ on $\mathbb{H}_d(\beta, L, M)$ and the *minimax risk* on $\mathbb{H}_d(\beta, L, M)$ is defined as

$$R_{n,r}\big[\mathbb{H}_d(\beta, L, M)\big] = \inf_{\tilde{f}} R_{n,r}\big[\tilde{f}, \mathbb{H}_d(\beta, L, M)\big], \tag{3.2}$$

where the infimum is taken over the set of all estimators.

Definition 2. *The normalizing sequence ψ_n is called minimax rate of convergence and the estimator $\hat{f}$ is called minimax (asymptotically minimax) if*

$$\liminf_{n\to\infty} \psi_n^{-r}\, R_{n,r}\big[\hat{f}, \mathbb{H}_d(\beta, L, M)\big] \;>\; 0, \tag{3.3}$$

$$\limsup_{n\to\infty} \psi_n^{-r}\, R_{n,r}\big[\hat{f}, \mathbb{H}_d(\beta, L, M)\big] \;<\; \infty. \tag{3.4}$$

Upper bound for maximal risk. Let the minimizer of the bias/variance trade-off (1.5) be given by

$$\bar{h} = (L^2 n)^{-\frac{1}{2\beta+d}}. \tag{3.5}$$

The next theorem shows how to construct the estimator based on locally parametric approach which achieves the following rate of convergence in the model (1.1)

$$\varphi_n(\beta) = n^{-\frac{\beta}{2\beta+d}}. \tag{3.6}$$

Let $\hat{f}^{\bar{h}}(x_0) = \hat{\theta}_{0,\dots,0}\big(\bar{h}\big)$ be given by (2.3), (2.4) and (2.5) with $h = \bar{h}$ and $b = \lfloor\beta\rfloor$.

Theorem 1. *Let $\beta > 0$, $L > 0$, $M > 0$, $x_0 \in [0,1]^d$, $c > 0$ and let ρ be a fixed contrast function. Then for any $(g_i)_i \in \mathcal{G}_\rho^{(c)}$*

$$\limsup_{n\to\infty} \varphi_n^{-r}(\beta)\, R_{n,r}\Big[\hat{f}^{\bar{h}}(x_0), \mathbb{H}_d(\beta, L, M)\Big] < \infty, \quad \forall r \geq 1.$$

This theorem will be deduced from Proposition 1 and the proof is given in Section 6.2.

Remark 3. *Tsybakov [1982a] showed lower bounds (3.3) for rate $n^{-\frac{\beta}{2\beta+d}}$ with the following assumption on Kullback distance on the noise density g, i.e. it exists $v_0 > 0$ such that*

$$\int g(u)\, \ln \frac{g(u)}{g(u+v)}\, du \leq o\big(v^2\big), \quad \forall v \,:\, |v| \leq v_0.$$

We notice that Gaussian and Cauchy densities verify this assumption (cf. also Tsybakov [2008] Chapter 2). In this case, we conclude that $\hat{f}^{\bar{h}}$ is minimax and $\varphi_n(\beta)$ is the minimax rate on $\mathbb{H}_d(\beta, L, M)$.

4 Bandwidth Selection of ρ-LPA Estimator

This section is devoted to the adaptive estimation over the collection of classes $\Big\{\mathbb{H}_d(\beta, L, M)\Big\}_{\beta,L}$. Here we suppose M known, as we mentioned in the introduction, the parameter M could be estimated and used with a "Plug-in" method (cf. Härdle and Tsybakov [1992]). We will not impose any restriction on the possible value of L, but we will assume that $\beta \in (0, b]$, where b as previously, is an arbitrary chosen integer.

We start by remarking that there is not optimally adaptive estimator. Well-known disadvantage of maximal approach is the dependence of the estimator on the smoothness parameters describing the functional class on which the maximal risk is determined (cf. (3.1)). In particular, $\bar{h}_n(\beta, L)$, optimally chosen in view of (1.5), depends explicitly on β and L. To overcome this drawback, a maximal adaptive approach has been proposed by Lepski [1990] for pointwise estimation. The first question arising in the adaptation (reduced to the problem at hand) can be formulated as follows.

Does there exist an estimator which would be minimax on $\mathbb{H}(\beta, L, M)$ simultaneously for all values of β and L belonging to some given set $\mathfrak{B} \subseteq \big[\mathbb{R}_+ \setminus 0\big] \times \big[\mathbb{R}_+ \setminus 0\big]$?

For integrated risks, the answer is positive (cf. Lepski [1991], Donoho, Johnstone, Kerkyacharian, and Picard [1995], Lepski and Spokoiny [1997], Goldenshluger and Nemirovski [1997] and Juditsky [1997]). For the estimation of the function at a given point, it is typical that the price to pay is not null (cf. Lepski [1990], Brown and Low [1996], Lepski and Spokoiny [1997], Tsybakov [1998], Klutchnikoff [2005], Reiss, Rozenholc, and Cuenod [2011], Chichignoud [2011]). Mostly, the price to pay is a power of $(b-\beta)\ln n$ for pointwise estimation.

Let $\Psi = \{\psi_n(\beta)\}_{\beta \in (0,b]}$ be a given family of normalizations.

Definition 3. *The family Ψ is called admissible if there exists an estimator $\hat{f}_n$ such that for some $L, M > 0$*

$$\limsup_{n\to\infty} \psi_n^{-r}(\beta)\; R_{n,r}\big(\hat{f}_n, \mathbb{H}_d(\beta, L, M)\big) < \infty, \quad \forall \beta \in (0, b]. \tag{4.1}$$

The estimator $\hat{f}_n$ satisfying (4.1) is called Ψ-attainable. The estimator $\hat{f}_n$ is called Ψ-adaptive if (4.1) holds for any $L > 0$.

Lepski [1990] showed that the family of rates $\{\varphi_n(\beta)\}_{\beta \in (0,b]}$, defined in (3.6), is not admissible in the white noise model. With other tools, Brown and Low [1996] extend this result for density estimation and nonparametric Gaussian regression. It means that there is no-estimator which would be minimax simultaneously for several values of parameter β, for pointwise estimation, even if L is supposed to be fixed. This result does not require any restriction on β as well.

Now, we need to find another family of normalizations for maximal risk which would be attainable and, moreover, optimal in view of some criterion of optimality. Let Φ be the following family of normalizations, for any $\beta \in (0, b]$

$$\phi_n(\beta) = \left(\frac{\varrho_n(\beta)}{n}\right)^{\frac{\beta}{2\beta+d}}, \quad \varrho_n(\beta) = 1 + \frac{2(b-\beta)}{(2\beta+d)(2b+d)} \ln n. \tag{4.2}$$

We notice that $\phi_n(b) = \varphi_n(b)$ and for n large enough $\varrho_n(\beta) \sim (b-\beta)\ln n$ for any $\beta \neq b$. It is possible to show that this family Φ is adaptive optimal using the most recent criterion developed by Klutchnikoff [2005] used for the white noise model and used by Chichignoud [2011] for the multiplicative uniform regression. On the other hand, the so-called *price to pay for adaptation* $\varrho_n(\beta)$ could be considered as optimal.

Construction of Φ-adaptive estimator. We begin by stating that the construction of our estimation procedure is decomposed in several steps. First, we determine the family of ρ-LPA estimators. Next, based on Lepski's method, we propose a data-driven selection from this family.

Let ρ be a fixed contrast function. In the model (1.1), we recall that the sequence of densities $(g_i)_i$ is "unknown" for the statistician. We take $\hat{f}^h$ the estimator given by (2.3), (2.4) and (2.5), so the family of ρ-LPA estimators $\hat{\mathcal{F}}$ is defined now as follows. Put

$$h_{\min} = (\ln n)^{2/d} n^{-1/d}, \quad h_{\max} = n^{-\frac{1}{2b+d}}, \tag{4.3}$$

and

$$h_k = 2^{-k} h_{\max}, \quad k = \overline{0, \mathbf{k}_n} := 0, \ldots, \mathbf{k}_n,$$

where $\mathbf{k}_n$ is the largest integer such that $h_{\mathbf{k}_n} \geq h_{\min}$. Set

$$\hat{\mathcal{F}} = \left\{ \hat{f}^{(k)}(x_0) = \hat{\theta}_{0,\dots,0}(h_k), \quad k = \overline{0, \mathbf{k}_n} \right\}. \tag{4.4}$$

We put $\hat{f}^*(x_0) = \hat{f}^{(\hat{k})}(x_0)$, where $\hat{f}^{(\hat{k})}(x_0)$ is selected from $\hat{\mathcal{F}}$ in accordance with the rule:

$$\hat{k} = \inf \left\{ k = \overline{0, \mathbf{k}_n} : \quad \left| \hat{f}^{(k)}(x_0) - \hat{f}^{(l)}(x_0) \right| \leq C S_n(l), \quad l = \overline{k+1, \mathbf{k}_n} \right\}. \tag{4.5}$$

Here we have used the following notations. Let $c > 0$ be fixed and

$$\begin{aligned} C &= \frac{4N_b}{c\lambda} \left(1 + 2\, K_\infty \, (1 \vee \dot{\rho}_\infty) \sqrt{rd} \right), \\ S_n(l) &= \left[\frac{1 + l \ln 2}{n \left(h_l \right)^d} \right]^{1/2}, \quad l = \overline{0, \mathbf{k}_n}, \end{aligned} \tag{4.6}$$

where $r \geq 1$ is the power of the risk and c is defined in (2.9), $\dot{\rho}_\infty$ and K_∞ are respectively bounds of $\rho'(\cdot)$ and $K(\cdot)$, and the positive constant λ is the smallest eigenvalue of the matrix $\int_{[-0.5,0.5]^d} U(x)\, U^\top(x)\, K(x)\, dx$. We will see that this matrix is strictly positive definite (cf. Lemma 1).

Main Result. The next theorem is the main result of this paper. It allows us to guarantee a good performance of our adaptive ρ-LPA estimator $\hat{f}^*$.

Theorem 2. *Let $b > 0, M > 0$ and ρ be a fixed contrast function. Then, for any $(g_i)_i \in \mathcal{G}_\rho^{(c)}$, $\beta \in (0, b]$, $L > 0$ and $r \geq 1$*

$$\limsup_{n \to \infty} \phi_n^{-r}(\beta)\; R_{n,r} \left[\hat{f}^*(x_0), \mathbb{H}_d(\beta, L, M) \right] < \infty.$$

The proof (given in Section 6.3) is based on the scheme due to Lepski, Mammen, and Spokoiny [1997].

Remark 4. *The assertion of the theorem means that the proposed estimator $\hat{f}^*(x_0)$ is Φ-adaptive in the model (1.1) (cf. Definition 3). It implies in particular that the family of normalizations Φ is admissible.*

Remark 5. *In the present paper, we do not give the explicit expression of the constant in the upper bound of the risk with the proof given in this paper. But it is possible to solve this problem. In the proof of Lepski's method, we notice that the upper bound polynomially depends on the parameter C and it is important to minimize this constant. We see that this constant depends on the contrast function ρ and it is easy to see that minimizing $C = C(\rho)$ can be viewed as minimizing the following Huber variance (cf. Huber and Ronchetti [2009] Page 74)*

$$\sigma_\rho^2 = \frac{\int (\rho')^2 \, dg}{\left(\int \rho'' \, dg\right)^2} \to \min_\rho,$$

where g is the noise density in the homeoscedastic model.

Remark 6. *The limitation concerning the consideration of isotropic classes of functions is due to the use of Lepski's procedure. It seems that to be able to treat the adaptation over the scale of anisotropic classes (i.e. d-dimensional functions with different regularities β for each variable). Another scheme should be applied as in Lepski and Levit [1999], Kerkyacharian, Lepski, and Picard [2001], Klutchnikoff [2005] and Goldenshluger and Lepski [2008]. As we have mentioned above, these latter procedures cannot be used with ρ-LPA estimators, and for the model (1.1) this problem is still open.*

5 Application: Huber function

Consider the model (1.1), with following additional assumptions.

$$g_i(\cdot) = g(\cdot/\sigma_i)/\sigma_i, \quad i = 1, ..., n, \tag{5.1}$$

where the density g is symmetric and monotonically on $\mathbb{R}_+$. $(\sigma_i)_i$ is a sequence of real values such that for any i, $0 < \sigma_{\min} \leq \sigma_i < \infty$ where $\sigma_{\min}$ is known. The model (1.1) with (5.1) can be written as

$$Y_i = f(X_i) + \sigma_i \, \xi_i, \quad i = 1, ..., n, \tag{5.2}$$

where (ξ_i) are i.i.d. with the density g.

Let

$$\rho_\gamma(z) = \gamma(z - 0.5\,\gamma)\, \mathbb{I}_{|z|>\gamma} + 0.5\, z^2\, \mathbb{I}_{|z|\leq\gamma},\ z \in \mathbb{R},\ \gamma \geq 0. \tag{5.3}$$

the Huber function (Huber [1964]). We construct the ρ_γ-LPA estimator from (2.3), (2.4) and (2.5). The function ρ_γ is a contrast function verifying Assumption 1. Recall that the constant $c = c\big(\rho_\gamma, (g_i)_i\big)$ defined in (2.9) must be positive. We notice that the second derivative can be written as $\rho''_\gamma(\cdot) = \mathbb{I}_{[-\gamma,\gamma]}(\cdot)$ and that

$$c\big(\rho_\gamma, (g_i)_i\big) \geq c_\gamma := 2\int_0^{\gamma\sigma_{\min}} g(z)dz.$$

We formulate the following assertion: for any $\sigma_{\min} > 0$ and any g a symmetric density and monotonically on $\mathbb{R}_+$, there exists a constant $\gamma_0 > 0$ such that for any $\gamma \geq \gamma_0$, $c_\gamma > 0$.

We propose the adaptive ρ_{γ_0}-LPA estimator $\hat{f}^*_{\gamma_0}(x_0)$ selected with the data-driven selection proposed in Section 4 with the constant

$$C = \frac{2N_b}{\lambda\int_0^{\gamma_0\sigma_{\min}} g(z)dz}\left(1 + 2\,K_\infty\,(1 \vee \gamma_0)\sqrt{rd}\right).$$

The next result is a direct consequence of Theorem 2.

Corollary 1. *Let $b > 0$, $M > 0$ be some fixed constants and consider the model (5.2). Then, for any $(g_i)_i \in \mathcal{G}^{(c_{\gamma_0})}_{\rho_{\gamma_0}}$, $\beta \in (0, b]$, $L > 0$ and $r \geq 1$*

$$\limsup_{n\to\infty} \phi_n^{-r}(\beta) R_{n,r}\left[\hat{f}^*_{\gamma_0}(x_0), \mathbb{H}_d(\beta, L, M)\right] < \infty.$$

Remark 7. *We notice that the threshold parameter C explicitly depends on the minoration $\sigma_{\min}$ of the noises variances $(\sigma_i)_i$. Contrary to linear estimators (C polynomially depends on $(\sigma_i)_i$), we can see that the influence of $(\sigma_i)_i$ is very limited for ρ-LPA estimators.*

Remark 8. *Corollary 1 only guarantees that asymptotically for any $\gamma \geq \gamma_0$, ρ_γ-LPA estimators have the same performance. In the future, an important question to adress is: how one can choose the parameter γ? In theory, there is yet no criterion for choosing an optimal γ, but we can make the following remarks. If $\gamma = \infty$, then the ρ_∞-LPA estimator is the least square estimator (sensitive to extreme values of the noise) and if $\gamma = 0$ then the ρ_0-LPA estimator becomes the median estimator (robust estimator). It is well-known that least squares estimator and median estimator respectively suffer from undersmoothing and oversmoothing. This phenomenon is highlighted by Reiss,* *Rozenholc, and Cuenod [2011]. We believe that a better choice of parameter γ should give a "semi-robust" estimator. Locally this could reduce the above mentioned issue. In practice, it will be interesting to select the parameter γ as a measurable function of observations which adapts to extreme values of the noise. This problem is related to the estimation of the noise variance and to the minimization of the Huber variance (cf. Huber and Ronchetti [2009] Page 74).*

6 Proofs of Main Results: Exponential inequalies and Upper Bounds

6.1 Proof of Proposition 1

Notations. Recall that $\mathcal{S}_b = \left\{q \in \mathbb{N}^d \ : \ |q| \leq b, \ |q| = q_1 + \ldots + q_d\right\}$ and N_b its cardinal. We consider the *partial derivative* of the local ρ-criterion

$$\tilde{D}_h(\cdot) = \left(\frac{\partial}{\partial t_p}\, \tilde{\pi}_h(\cdot)\right)^{\top}_{p \in \mathcal{S}_b}, \tag{6.1}$$

where $\tilde{\pi}_h(\cdot)$ is the local ρ-criterion defined in (2.4). Let also

$$\mathcal{E}_h(\cdot) = \mathbb{E}_f\left[\tilde{D}_h(\cdot)\right] \quad \text{and} \quad D_h(\cdot) = \mathbb{E}_{f_\theta}\big[\tilde{D}_h(\cdot)\big], \tag{6.2}$$

where f_θ is the Taylor polynomial defined in (2.6), $\mathbb{E}_{f_\theta} = \mathbb{E}^n_{f_\theta}$ be the mathematical expectation with respect to the probability law $\mathbb{P}_{f_\theta}$ of the "parametric" observations $(X_i, \mathcal{Y}_i)_{i=1,\ldots,n}$ (cf. (1.4)) and $\mathbb{E}_f = \mathbb{E}^n_f$ be the mathematical expectation with respect to the probability law $\mathbb{P}_f$ of the observation $Z^{(n)}$.

We call the *Jacobian matrix* J_D of D_h such that

$$\big(J_D(\cdot)\big)_{p,q \in \mathcal{S}_b} := \left(\frac{\partial}{\partial t_q} D^p_h(\cdot)\right)_{p,q \in \mathcal{S}_b} = \left(\mathbb{E}_{f_\theta} \frac{\partial^2}{\partial t_p \partial t_q}\, \tilde{\pi}_h(\cdot)\right)_{p,q \in \mathcal{S}_b}, \tag{6.3}$$

where $D^p_h(\cdot)$ is the p^{th} component of $D_h(\cdot)$.

Auxiliary lemmas. We give the following lemma concerning the *deterministic criterion* D_h defined in (6.2). Denote $\|\cdot\|_2$ the ℓ_2-norm on $\mathbb{R}^{N_b}$.

Lemma 1. *Let ρ be a contrast function, for any $(g_i)_i \in \mathcal{G}_\rho^{(c)}$ we have the following assertions:*

1. *the matrix $J_D(\theta)$ is strictly positive definite and θ is the unique solution of the equation $D_h(\cdot) = (0, ..., 0)$ on $\Theta(M)$,*
2. *there exists $\delta > 0$ which only depends on the contrast function ρ such that for any $\tilde{\theta} \in \mathcal{B}(\theta, \delta)$, we have*

$$\|\tilde{\theta} - \theta\|_2 \leq \frac{2}{c\lambda} \inf_{h>0} \left\| D_h(\tilde{\theta}) - D_h(\theta) \right\|_2 .$$

Recall that $b_h = \sup_{x \in V_{x_0}(h)} \left| f_\theta(x) - f(x) \right|$ corresponds to the approximation error (bias) and denote $\mathcal{E}_h^p(\cdot)$ the p^{th} component of $\mathcal{E}_h(\cdot)$. Let us give a lemma which allows us to control the *bias term.*

Lemma 2. *For any contrast function ρ, $h > n^{-1/d}$ and any f such that $\|\theta\|_1 \leq M$, we have*

$$\max_{p \in \mathcal{S}_b} \sup_{t \in \Theta(M)} \left| \mathcal{E}_h^p(t) - D_h^p(t) \right| \leq b_h .$$

The next result allows us to control deviations of *partial derivatives of ρ-criterion* $\tilde{D}_h$ defined in (6.1).

Lemma 3. *For any contrast function ρ, any f such that $\|\theta\|_1 \leq M$ and any $h > n^{-1/d}$, we have for any $z \geq 2\big(1 \vee b_h \sqrt{nh^d}\big)$*

$$\max_{p \in \mathcal{S}_b} \mathbb{P}_f \left(\sqrt{nh^d} \sup_{t \in \Theta(M)} \left| \tilde{D}_h^p(t) - \mathcal{E}_h^p(t) \right| \geq z \right)$$
$$\leq \varSigma \exp \left\{ - \frac{\left(z - b_h \sqrt{nh^d} \right)^2}{8K_\infty^2 \left(1 \vee \dot{\rho}_\infty^2\right) + \frac{4K_\infty}{3\sqrt{nh^d}} \left(1 \vee \dot{\rho}_\infty\right) z} \right\} .$$

As we mentioned above, the partial derivatives $\big(\tilde{D}_h^p(\cdot)\big)_p$ can be considered as empirical processes. Thus the proof (given in Appendix) is based on a chaining argument and Bernstein's inequality (cf. (7.8)). In particular, it is required that the derivative ρ' of the contrast function is bounded and Lipschitz.

Denote by $\bar{G}_\delta^h$ the complementary of G_δ^h (defined in (2.7)) where the radius δ is defined in Lemma 1 and let $\varkappa_\delta = \inf_{t \in \Theta(M) \setminus \mathcal{B}(\theta, \delta)} \|D_h(t)\|_2 / 2$ be a positive constant. The next lemma allows us to control the probability of the event that "the ρ-LPA estimator does not belong to the ball centered on θ with radius δ".

Lemma 4. *For any contrast function ρ, $f \in \mathbb{H}_d(\beta, L, M)$, $\delta > 0$ and $n \in \mathbb{N}^*$ such that*

$$\frac{\varkappa_\delta}{\sqrt{N_b}} \geq 2 \sup_{h \in [h_{\min}, h_{\max}]} \left(b_h \vee 1/\sqrt{nh^d}\right),$$

we have

$$\mathbb{P}_f\left[\bar{G}_\delta^h\right] \leq N_b \varSigma \exp\left\{-\frac{nh^d\left(\varkappa_\delta/2\sqrt{N_b}\right)^2}{8K_\infty^2\left(1 \vee \dot{\rho}_\infty^2\right) + \frac{4\varkappa_\delta}{3\sqrt{N_b}} K_\infty\left(1 \vee \dot{\rho}_\infty\right)}\right\}.$$

Proofs of those lemmas are given in Appendix.

Proof of Proposition 1. Definitions of $\hat{\theta}(h)$ and $\theta = \theta(f, x_0, h)$ imply that for any $\varepsilon \geq \frac{4N_b}{c\lambda}\left(1 \vee b_h\sqrt{nh^d}\right)$

$$\begin{aligned}
\mathbb{P}_f\left(\sqrt{nh^d}\left|\hat{f}^h(x_0) - f(x_0)\right| \geq \varepsilon,\ G_\delta^h\right) &\leq \mathbb{P}_f\left(\sqrt{nh^d}\left|\hat{\theta}_{0,\ldots,0}(h) - \theta_{0,\ldots,0}\right| \geq \varepsilon,\ G_\delta^h\right) \\
&\leq \mathbb{P}_f\left(\sqrt{nh^d}\sqrt{N_b}\left\|\hat{\theta}(h) - \theta\right\|_2 \geq \varepsilon,\ G_\delta^h\right),
\end{aligned}$$

where $\|\cdot\|_2$ is the ℓ_2-norm on $\mathbb{R}^{N_b}$. Under the event G_δ^h we have $\hat{\theta}(h) \in \mathcal{B}(\theta, \delta)$ for the specific choice of δ given in Lemma 1 and depending on the contrast ρ. According to Lemma 1 *(2)* we obtain that

$$\begin{aligned}
&\mathbb{P}_f\left(\sqrt{nh^d}\left|\hat{f}^h(x_0) - f(x_0)\right| \geq \varepsilon,\ G_\delta^h\right) \\
&\leq \mathbb{P}_f\left(\sqrt{nh^d}\frac{2\sqrt{N_b}}{c\lambda}\left\|D_h\left(\hat{\theta}(h)\right) - D_h(\theta)\right\|_2 \geq \varepsilon\right).
\end{aligned}$$

Using Lemma 1 *(1)* and the definition of $\hat{\theta}(h)$ in (2.5), reminding that $D_h(\theta) = \tilde{D}_h\left(\hat{\theta}(h)\right) = 0$ and using the well-known inequality $\|\cdot\|_2 \leq \sqrt{N_b}\|\cdot\|_\infty$ (where $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are respectively ℓ_2-norm and ℓ_∞-norm on $\mathbb{R}^{N_b}$), we get with the last inequality:

$$\begin{aligned}
&\mathbb{P}_f\left(\sqrt{nh^d}\left|\hat{f}^h(x_0)-f(x_0)\right|\geq\varepsilon,\,G^h_\delta\right)\\
&\quad\leq\mathbb{P}_f\left(\sqrt{nh^d}\frac{2\sqrt{N_b}}{c\lambda}\left\|\tilde{D}_h\big(\hat{\theta}(h)\big)-D_h\big(\hat{\theta}(h)\big)\right\|_2\geq\varepsilon\right)\\
&\quad\leq\sum_{p\in\mathcal{S}_b}\mathbb{P}_f\left(\sqrt{nh^d}\sup_{t\in\Theta(M)}\left|\tilde{D}^p_h(t)-D^p_h(t)\right|\geq\frac{c\lambda\,\varepsilon}{2N_b}\right).
\end{aligned}$$

Applying Lemma 3 with $z=\frac{c\lambda\,\varepsilon}{2N_b}$ and the last inequality, finally we obtain the assertion of Proposition 1

$$\begin{aligned}
&\mathbb{P}_f\left(\sqrt{nh^d}\left|\hat{f}^h(x_0)-f(x_0)\right|\geq\varepsilon,\,G^h_\delta\right)\\
&\quad\leq N_b\varSigma\exp\left\{-\frac{\left(\frac{c\lambda\,\varepsilon}{2N_b}-\left(1\vee b_h\sqrt{nh^d}\right)\right)^2}{8K^2_\infty\left(1\vee\dot{\rho}^2_\infty\right)+\frac{4\,K_\infty}{3N_b}\left(1\vee\dot{\rho}_\infty\right)\frac{c\lambda\,\varepsilon}{\sqrt{nh^d}}}\right\}.
\end{aligned}$$

∎

6.2 Proof of Theorem 1

Before starting Proofs of the main results of this paper, let us define auxiliary results. The next proposition provides us with upper bound for the risk of a ρ-LPA estimator. Put

$$\begin{aligned}
\bar{C}_r &= \left(\frac{4N_b}{c\lambda}\right)^r+N_b\varSigma\int_{\frac{4N_b}{c\lambda}}^{\infty}rz^{r-1}\\
&\quad\times\exp\left\{-\frac{\left(\frac{z}{N_b}\frac{c\lambda}{2}-1\right)^2}{8K^2_\infty(1\vee\dot{\rho}^2_\infty)+\frac{4\delta}{3N_b}c\lambda\,K_\infty\left(1\vee\dot{\rho}_\infty\right)}\right\}dz,\quad r\geq1. \qquad (6.4)
\end{aligned}$$

Proposition 2. *Let ρ be a contrast function. Then, for any $n\in\mathbb{N}^*$, $h>n^{-1/d}$, $x_0\in[0,1]^d$ and any f such that $\|\theta\|_1<\infty$, we have*

$$\mathbb{E}_f\left|\hat{f}^h(x_0)-f(x_0)\right|^r\mathbb{I}_{G^h_\delta}\leq\bar{C}_r\left(1\vee b_h\sqrt{nh^d}\right)^r\left(nh^d\right)^{-r/2},\quad r\geq1.$$

The proof of Proposition 2 is deduced from Proposition 1 by integration.

Proof of Theorem 1 By definition of $\mathbb{H}_d(\beta, L, M)$, the approximation error (bias) b_h as defined in (1.3) verified $b_h \leq Ldh^\beta$ for any $h > 0$. Moreover by definition of $\bar{h} = (L^2 n)^{-\frac{1}{2\beta+d}}$ in (3.5), we have that $b_{\bar{h}} \sqrt{n\bar{h}^d} \leq d$ and $\left(n\bar{h}^d\right)^{-1/2} = L^{\frac{d}{2\beta+d}} \varphi_n(\beta)$. We get

$$\mathbb{E}_f \left|\hat{f}^{\bar{h}}(x_0) - f(x_0)\right|^r = \mathbb{E}_f \left|\hat{f}^{\bar{h}}(x_0) - f(x_0)\right|^r \mathbb{I}_{G_\delta^{\bar{h}}} + \mathbb{E}_f \left|\hat{f}^{\bar{h}}(x_0) - f(x_0)\right|^r \mathbb{I}_{\bar{G}_\delta^{\bar{h}}}. \quad (6.5)$$

The right hand side is controlled by Lemma 4. Indeed, we can use the Cauchy-Schwarz inequality,

$$\begin{aligned} &\mathbb{E}_f \left|\hat{f}^{\bar{h}}(x_0) - f(x_0)\right|^r \mathbb{I}_{\bar{G}_\delta^{\bar{h}}} \\ &\leq \left(\mathbb{E}_f \left|\hat{f}^{\bar{h}}(x_0) - f(x_0)\right|^{2r} \mathbb{P}_f \left\{\bar{G}_\delta^{\bar{h}}\right\}\right)^{1/2} \\ &\leq (2M)^r \sqrt{N_b \varSigma}\, \exp\left\{ - \frac{n\bar{h}^d \left(\varkappa_\delta / 2\sqrt{N_b}\right)^2}{16 K_\infty^2 \left(1 \vee \dot{\rho}_\infty^2\right) + \frac{8\varkappa_\delta}{3\sqrt{N_b}} K_\infty \left(1 \vee \dot{\rho}_\infty\right)} \right\}. \quad (6.6) \end{aligned}$$

The last inequality is obtained because M is a upper bound of f and $\hat{f}^{\bar{h}}$ (cf. Definition 1 and (2.3)). Using Proposition 2, (6.5) and (6.6), we obtain

$$\begin{aligned} &\mathbb{E}_f \left|\hat{f}^{\bar{h}}(x_0) - f(x_0)\right|^r \\ &\quad \leq \bar{C}_r d^r\, L^{\frac{rd}{2\beta+d}} \varphi_n^r(\beta) \\ &\qquad + (2M)^r \sqrt{N_b \varSigma}\, \exp\left\{ - \frac{n\bar{h}^d \left(\varkappa_\delta / 2\sqrt{N_b}\right)^2}{16 K_\infty^2 \left(1 \vee \dot{\rho}_\infty^2\right) + \frac{8\varkappa_\delta}{3\sqrt{N_b}} K_\infty \left(1 \vee \dot{\rho}_\infty\right)} \right\}. \end{aligned}$$

When n tends towards $+\infty$, Theorem 1 is proved. ∎

6.3 Proof of Theorem 2

We start the proof with formulating some auxiliary results whose proofs are given in Appendix. Define

$$h^* = \left[\frac{\varrho_n(\beta)}{L^2 d^2\, n}\right]^{\frac{1}{2\beta+d}}, \quad (6.7)$$

where $\varrho_n^2(\beta)$ is defined in (4.2). Let κ be an integer defined as follows:

$$2^{-\kappa} h_{\max} \leq h^* < 2^{-\kappa+1} h_{\max}. \quad (6.8)$$

For any n large enough, we have $h_{\min} \leq h^* \leq h_{\max}$.

Lemma 5. *For any $f \in \mathbb{H}_d(\beta, L, M)$, any n large enough and any $k \geq \kappa+1$*

$$\mathbb{P}_f\big(\hat{k} = k,\ G_\delta^{h_k}\big) \leq J\, 2^{-2(k-1)rd},$$

where $J = N_b \Sigma\big(1 + (1-2^{-2rd})^{-1}\big)$.

Proof of Theorem 2. This proof is based on the scheme due to Lepski, Mammen, and Spokoiny [1997]. The definition of h^* (6.7) and κ (6.8) implies that for any n large enough

$$\left(1 \vee b_{h_k}\sqrt{n h_k^d}\right) \leq 2\sqrt{1 + k\ln 2}, \quad \forall k \geq \kappa. \tag{6.9}$$

Using Proposition 2, the last inequality yields

$$\mathbb{E}_f\big|\hat{f}^{(k)}(x_0) - f(x_0)\big|^r \mathbb{I}_{G_\delta^{h_k}} \leq \bar{C}_r\, S_n^r(k), \quad \forall k \geq \kappa. \tag{6.10}$$

To get this result we have applied Proposition 2 with $h = h_k$ and (6.9). We also have

$$\begin{aligned}
&\mathbb{E}_f\big|\hat{f}^{(\hat{k})}(x_0) - f(x_0)\big|^r \\
&\quad = \mathbb{E}_f\big|\hat{f}^{(\hat{k})}(x_0) - f(x_0)\big|^r \mathbb{I}_{\hat{k}\leq\kappa, G_\delta^{h_{\hat{k}}}} + \mathbb{E}_f\big|\hat{f}^{(\hat{k})}(x_0) - f(x_0)\big|^r \mathbb{I}_{\hat{k}>\kappa, G_\delta^{h_{\hat{k}}}} \\
&\qquad + \mathbb{E}_f\big|\hat{f}^{(\hat{k})}(x_0) - f(x_0)\big|^r \mathbb{I}_{\bar{G}_\delta^{h_{\hat{k}}}} \\
&\quad := R_1(f) + R_2(f) + R_3(f).
\end{aligned} \tag{6.11}$$

First we control R_1. By convexity of $|\cdot|^r$, $r \geq 1$ and with the triangular inequality, we have

$$\big|\hat{f}^{(\hat{k})}(x_0) - f(x_0)\big|^r \leq 2^{r-1}\big|\hat{f}^{(\hat{k})}(x_0) - \hat{f}^{(\kappa)}(x_0)\big|^r + 2^{r-1}\big|\hat{f}^{(\kappa)}(x_0) - f(x_0)\big|^r.$$

The definition of $\hat{k}$ in (4.5) yields

$$\big|\hat{f}^{(\hat{k})}(x_0) - \hat{f}^{(\kappa)}(x_0)\big|^r \mathbb{I}_{\hat{k}\leq\kappa, G_\delta^{h_{\hat{k}}}} \leq C^r S_n^r(\kappa),$$

where the constant C is defined in (4.6). In view of (6.10), the definitions of h_κ lead to

$$\mathbb{E}_f\big|\hat{f}^{(\kappa)}(x_0) - f(x_0)\big|^r \mathbb{I}_{\hat{k}\leq\kappa, G_\delta^{h_{\hat{k}}}} \leq \bar{C}_r\, S_n^r(\kappa),$$

where $\bar{C}_r$ is defined in (6.4). Noting that the right hand side of the obtained inequality is independent of f and taking into account the definition of κ and h^* we obtain

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-r}(\beta) R_1(f) < \infty. \tag{6.12}$$

Now, let us bounded from above R_2. Applying Cauchy-Schwartz inequality, in view of Lemma 5 we have for n large enough

$$\begin{aligned} R_2(f) &= \sum_{k=\kappa}^{\mathbf{k}_n} \mathbb{E}_f \big| \hat{f}^{(k)}(x_0) - f(x_0) \big|^r \mathbb{I}_{G_\delta^{h_k}} \\ &\leq \sum_{k>\kappa} \Big(\mathbb{E}_f \big| \hat{f}^{(k)}(x_0) - f(x_0) \big|^{2r} \mathbb{I}_{G_\delta^{h_k}} \Big)^{1/2} \sqrt{\mathbb{P}_f\big\{ \hat{k} = k, G_\delta^{h_k} \big\}} \\ &= \sqrt{J} \sum_{k>\kappa} \Big(\mathbb{E}_f \big| \hat{f}^{(k)}(x_0) - f(x_0) \big|^{2r} \mathbb{I}_{G_\delta^{h_k}} \Big)^{1/2} 2^{-(k-1)rd}. \end{aligned}$$

We obtain from (6.10) and the last inequality

$$R_2(f) \leq \frac{\bar{C}_{2r}^r \, 2^{rd} \sqrt{J}}{\big(n h_{\max}^d \big)^{r/2}} \sum_{s\geq 0} (1 + s\ln 2)^{r/2} 2^{-srd}.$$

It remains to note that the right hand side of the last inequality is independent of f. Thus, we have

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-r}(\beta) R_2(f) < \infty. \tag{6.13}$$

It remains to bound $R_3(f)$. By definition, note that $|\hat{f}^{(\hat{k})}(x_0)| \leq M$, this allows us to state that $\big| \hat{f}^{(\hat{k})}(x_0) - f(x_0) \big| \leq 2M$. Finally we obtain

$$R_3(f) \leq 2^r M^r \, \mathbb{P}_f \big\{ \bar{G}_\delta^{h_{\hat{k}}} \big\}.$$

Since $n h_{\min}^d = \big(\ln n \big)^{2d}$, then

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}_d(\beta,L,A,M)} \phi_n^{-r}(\beta) R_3(f) < \infty, \tag{6.14}$$

follows now from Lemma 4. Theorem 2 is proved from (6.11), (6.12), (6.13) and (6.14). ∎

7 Appendix

7.1 Proof of Lemma 1

1. By definition of the Jacobian matrix $J_D(\cdot)$ in (6.3), we can write for any $p, q \in \mathcal{S}_b$

$$\Big[J_D\big(\tilde{\theta}\big)\Big]_{p,q} = \frac{1}{n}\sum_{i=1}^{n}\int_{[-1/2,1/2]^d} x^{p+q}\, K(x) \int_{\mathbb{R}} \rho''\big(z - f_{\tilde{\theta}-\theta}(y+hx)\big)\, g_i(z)dz\, dx.$$

Applying this formula when $\tilde{\theta} = \theta$, the term $f_{\tilde{\theta}-\theta}$ vanishes, so:

$$J_D\big(\theta\big) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \rho''(z)\, g_i(z)dz \int_{[-1/2,1/2]^d} U(x)\, U^\top(x)\, K(x)dx.$$

where $U(\cdot)$ is defined in (2.1). Since $(g_i)_i \in \mathcal{G}_\rho^{(c)}$, the definition of c in (2.9) implies that $\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \rho''(z)\, g_i(z)\, dz \geq c > 0$.

Now we show that $J_D(\theta)$ is a strictly positive definite matrix, indeed for any $\tau \in \mathbb{R}^{N_b}\backslash 0$

$$\begin{aligned}
\tau^\top J_D(\theta)\tau &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \rho''(z)\, g_i(z)dz\, \tau^\top \int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d} U(x)\, U^\top(x)\, K(x)\, dx\, \tau \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \rho''(z)\, g_i(z)dz \int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d} \left[\tau^\top U(x)\right]^2 K(x)\, dx \\
&\geq c\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d} \left[\tau^\top U(x)\right]^2 K(x)\, dx > 0. \qquad (7.1)
\end{aligned}$$

Let us show that for any $h > n^{-1/d}$, θ is the unique solution of $D_h(\cdot) = (0, \ldots, 0)$. By definition in (6.2), D_h can be written as

$$D_h^p(t) = -\frac{1}{n}\sum_{i=1}^{n}\int_{[-0.5,0.5]^d} x^p K(x) \int_{\mathbb{R}} \rho'\big(z - f_{t-\theta}(y+hx)\big)\, g_i(z)\, dz\, dx. \quad (7.2)$$

Moreover, we have that

$$D_h(t) = (0, \ldots, 0) \Longrightarrow \sum_{p\in\mathcal{S}_b} \big(t_p - \theta_p\big) D_h^p(t) = 0.$$

Denote $u(\cdot) = f_{t-\theta}(y + h\cdot)$. Since for any i, g_i is monotonically on $\mathbb{R}_+$ and symmetric, then

$$\forall x \in [-0.5, 0.5]^d, \quad \inf_{z>0} \inf_{i=1,\dots,n} g_i\big(z - |u(x)|\big) - g_i\big(z + |u(x)|\big) \geq 0. \tag{7.3}$$

Since $(g_i)_i$ are symmetric, K is positive and ρ' is odd and positive on $\mathbb{R}_+^*$, the last inequality and (7.2) imply

$$\begin{aligned}
& \int_{[-0.5,0.5]^d} u(x)K(x) \int_{\mathbb{R}} \rho'\big(z - u(x)\big) \frac{\sum_{i=1}^n g_i(z)}{n} \, dz \, dx = 0 \\
\Leftrightarrow \; & \int_{[-0.5,0.5]^d} u(x)K(x) \int_{\mathbb{R}} \rho'(z) \sum_{i=1}^n g_i\big(z + u(x)\big) \, dz \, dx = 0 \\
\Leftrightarrow \; & \int_{[-0.5,0.5]^d} u(x)K(x) \int_0^\infty \rho'(z) \sum_{i=1}^n g_i\big(z - u(x)\big) - g_i\big(z + u(x)\big) \, dz \, dx = 0 \\
\Leftrightarrow \; & \int_{[-0.5,0.5]^d} |u(x)|K(x) \int_0^\infty \rho'(z) \sum_{i=1}^n g_i\big(z - |u(x)|\big) - g_i\big(z + |u(x)|\big) \, dz \, dx = 0 \\
\Leftrightarrow \; & \forall x \in [-0.5, 0.5]^d, \; z > 0, \quad \sum_{i=1}^n g_i\big(z - |u(x)|\big) - g_i\big(z + |u(x)|\big) = 0 \qquad (7.4)
\end{aligned}$$

Assume that there exists $x_0 \in [-0.5, 0.5]^d$ such that $u(x_0) \neq 0$. In particular for any i, since g_i is monotonically on $\mathbb{R}_+$, there exists $z_{i,x_0} > 0$ such that

$$g_i(z_{i,x_0} - |u(x_0)|) - g_i\big(z_{i,x_0} + |u(x_0)|\big) > 0.$$

That leads a contradiction in view of (7.3) and (7.4), thus for any $x \in [-0.5, 0.5]^d$, we have $u(x) = 0$. By definition of $u(\cdot)$, we get

$$\forall x \in [-0.5, 0.5]^d, \; h > n^{-1/d}, \quad |f_{t-\theta}(y + hx)| = 0 \Longrightarrow t = \theta.$$

Then, θ is the unique solution of $D_h(\cdot) = 0$.

2. Let $|||\cdot|||_2$ be the euclidian matrix norm, $\lambda_{\max}(A)$ the spectral ray of the matrix A and $\lambda_0(A)$ the smallest eigenvalue of A. According to Lemma 1 *(1)*, there exists a radius $\delta > 0$, which only depends of ρ such that

$$\inf_{\tilde{\theta} \in \mathcal{B}(\theta,\delta)} \lambda_0\left(J_D(\tilde{\theta})\right) \geq \lambda_0\left(J_D(\theta)\right)/2 > 0. \tag{7.5}$$

This assertion can be explained as follows. In view of Assumption 1 (*3*), $\lambda_0\left(J_D(\cdot)\right)$ is a continuous function. So, there exists a radius $\delta > 0$ expected such that (7.5) is true.

According to the local inverse function theorem, we can deduced that for any $\tilde{\theta} \in \mathcal{B}(\theta, \delta)$,

$$|||J_{D^{-1}}(\tilde{\theta})|||_2 = |||J_D^{-1}(\tilde{\theta})|||_2 = \lambda_{\max}\big(J_D^{-1}(\tilde{\theta})\big) = 1/\lambda_0\big(J_D(\tilde{\theta})\big), \tag{7.6}$$

By definition of $\mathcal{G}_\rho^{(c)}$, we have for any $(g_i)_i \in \mathcal{G}_\rho^{(c)}$, $c = c(\rho, (g_i)_i)$ and according to (7.1) $\lambda = \lambda_0\big(J_D(\tilde{\theta})\big) > 0$. The smallest eigenvalue of $J_D(\theta)$ is bigger than $c\lambda > 0$. Indeed we have

$$J_D\big(\theta\big) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \rho''(z)\, g_i(z) dz \int_{[-1/2,1/2]^d} U(x)\, U^\top(x)\, K(x)\, dx,$$

and

$$\begin{aligned}\lambda_0\big(J_D(\theta)\big) &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \rho''(z)\, g_i(z) dz \;\; \lambda_0\left(\int_{[-1/2,1/2]^d} U(x)\, U^\top(x)\, K(x)\, dx\right)\\ &\geq c\lambda.\end{aligned}$$

By definition of δ in (7.5), using (7.6) and the last inequality, we have for any $\tilde{\theta} \in \mathcal{B}(\theta, \delta)$

$$|||J_{D^{-1}}(\tilde{\theta})|||_2 \leq \frac{2}{c\lambda}. \tag{7.7}$$

As D_h is differentiable and each partial derivative is continuous (cf. Assumption 1 *(3)*), we use the local inverse function theorem and (7.7) which give for any $\tilde{\theta} \in \mathcal{B}(\theta, \delta)$ the following inequality

$$\|\tilde{\theta} - \theta\|_2 = \left\|D_h^{-1} \circ D_h\big(\tilde{\theta}\big) - D_h^{-1} \circ D_h(\theta)\right\|_2 \leq \frac{2}{c\lambda}\left\|D_h\big(\tilde{\theta}\big) - D_h(\theta)\right\|_2$$

∎

7.2 Proof of Lemma 2

By definition of $\mathcal{E}_h^p$ and D_h^p in (6.2), we have for any $t \in \Theta(M)$

$$\begin{aligned}\left|\mathcal{E}_h^p(t) - D_h^p(t)\right| &\leq \frac{1}{nh^d}\sum_{i=1}^{n}\int_{[-1/2,1/2]^d}\left|\frac{x - x_0}{h}\right|^p K\left(\frac{x - x_0}{h}\right)\\ &\quad\times \int_{\mathbb{R}}\left|\rho'\big(z + f(x) - f_t(x)\big) - \rho'\big(z - f_{t-\theta}(x)\big)\right|\, g_i(z) dz\, dx.\end{aligned}$$

Since ρ' is 1-Lipschitz (cf. Assumption 1 (2)) and $\int K = 1$, then with the last inequality, it yields

$$\forall h > n^{-1/d}, \quad \max_{p \in \mathcal{S}_b} \sup_{t \in \Theta(M)} \left| \mathcal{E}_h^p(t) - D_h^p(t) \right| \leq b_h.$$

■

7.3 Proof of Lemma 3

Bernstein's Inequality. To prove this lemma, we use the following well-known Bernstein's inequality which can be found in Massart [2007] (Section 2.2.3, Proposition 2.9). Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent square integrable random variables such that for some nonnegative constant $\mathcal{X}_\infty$, $\mathcal{X}_i \leq \mathcal{X}_\infty$ almost surely for all $i = 1, \cdots, n$. Then for any positive ϵ, we have

$$\mathbb{P}\left(\sum_{i=1}^n (\mathcal{X}_i - \mathbb{E}\mathcal{X}_i) \geq \epsilon \right) \leq \exp\left\{ -\frac{\epsilon^2}{2\sum_{i=1}^n \mathbb{E}\mathcal{X}_i^2 + 2\mathcal{X}_\infty \epsilon/3} \right\}, \tag{7.8}$$

where $\mathbb{E} = \mathbb{E}^n$ is the mathematical expectation with respect to the probability law $\mathbb{P}$ of $\mathcal{X}_1, \ldots, \mathcal{X}_n$. The latter inequality is so-called *Bernstein's inequality.*

Proof of Lemma 3. We have for any $p \in \mathcal{S}_b$

$$\sup_{t \in \Theta(M)} \left| \tilde{D}_h^p(t) - D_h^p(t) \right| \leq \sup_{t \in \Theta(M)} \left| \tilde{D}_h^p(t) - \mathcal{E}_h^p(t) \right| + \sup_{t \in \Theta(M)} \left| \mathcal{E}_h^p(t) - D_h^p(t) \right|.$$

In view of Lemma 2, we get

$$\sup_{t \in \Theta(M)} \left| \tilde{D}_h^p(t) - D_h^p(t) \right| \leq \sup_{t \in \Theta(M)} \left| \tilde{D}_h^p(t) - \mathcal{E}_h^p(t) \right| + b_h. \tag{7.9}$$

Set $L(\cdot) = \tilde{D}_h^p(\cdot) - \mathcal{E}_h^p(\cdot)$. To establish the assertion of the lemma, we use a chaining argument on $L(\cdot)$. Remember that $\Theta(M)$ is a compact of $\mathbb{R}^{N_b}$ with ℓ_1-norm. Let $t_0 \in \Theta(M)$ be fixed and for any $l \in \mathbb{N}^*$ put Γ_l a 10^{-l}-net on $\Theta(M)$. We introduce the following notations

$$u_0(t) = t_0, \quad u_l(t) = \arg\inf_{u \in \Gamma_l} \|u - t\|_1, \quad l \in \mathbb{N}^*.$$

Since ρ' is continuous, $L(\cdot)$ is stochastically continuous which allows us to use the following chaining argument

$$L(t) = L(t_0) + \sum_{l=1}^{\infty} L\big(u_l(t)\big) - L\big(u_{l-1}(t)\big), \quad \forall t \in \Theta(M). \tag{7.10}$$

Using (7.9) and (7.10), we obtain

$$\begin{aligned}
&\mathbb{P}_f\left(\sqrt{nh^d} \sup_{t\in\Theta(M)} \left|\tilde{D}_h^p(t) - D_h^p(t)\right| \geq z\right) \\
&\leq \mathbb{P}_f\left(\sup_{t\in\Theta(M)} |L(t)| \geq \frac{z}{\sqrt{nh^d}} - b_h\right) \\
&\leq \mathbb{P}_f\left(|L(t_0)| + \sup_{t\in\Theta(M)} \sum_{l=1}^{\infty} \left|L\big(u_l(t)\big) - L\big(u_{l-1}(t)\big)\right| \geq \frac{z}{\sqrt{nh^d}} - b_h\right). \quad (7.11)
\end{aligned}$$

We can control the second term as follows.

$$\sup_{t\in\Theta(M)} \sum_{l=1}^{\infty} \left|L\big(u_l(t)\big) - L\big(u_{l-1}(t)\big)\right| \leq \sum_{l=1}^{\infty} \sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1} \\ \|u-v\|_1\leq 10^{-l}}} |L(u) - L(v)|,$$

where $\Gamma_0 = \{t_0\}$. Using (7.11) and last inequality, we get

$$\begin{aligned}
&\mathbb{P}_f\left(\sqrt{nh^d} \sup_{t\in\Theta(M)} |L(t)| \geq z - b_h\sqrt{nh^d}\right) \\
&\leq \mathbb{P}_f\left(\sqrt{nh^d}\,|L(t_0)| \geq z/2 - b_h\sqrt{nh^d}/2\right) \\
&\quad + \mathbb{P}_f\left(\sqrt{nh^d} \sum_{l=1}^{\infty} \sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1} \\ \|u-v\|_1\leq 10^{-l}}} \left|L(u) - L(v)\right| \geq z/2 - b_h\sqrt{nh^d}/2\right). \quad (7.12)
\end{aligned}$$

By Definition of $\tilde{D}_h^p$ in (6.1), we can write:

$$\tilde{D}_h^p(t) = \frac{1}{nh^d} \sum_{i=1}^{n} \rho'\big(Y_i - f_t(X_i)\big) \left(\frac{X_i - x_0}{h}\right)^p K\left(\frac{X_i - x_0}{h}\right).$$

We define the function $\mathcal{W}_t(x,z) = \frac{1}{\sqrt{nh^d}} \rho'\big(z + f(x) - f_t(x)\big) \left(\frac{x-x_0}{h}\right)^p K\left(\frac{x-x_0}{h}\right)$ for all $x \in [0,1]^d$ and $z \in \mathbb{R}$. Since for any i, $Y_i = f(X_i) + \xi_i$, the process

$\sqrt{nh^d}\, L(\cdot)$ can be written as an empirical process (sum of independent, zero-mean and bounded random variables).

$$\sqrt{nh^d}\, L(t) = \sum_{i=1}^{n} \mathcal{W}_t\big(X_i,\xi_i\big) - \mathbb{E}_f \mathcal{W}_t\big(X_i,\xi_i\big), \quad t \in \Theta(M) \tag{7.13}$$

At a fixed point t_0, we can use classical exponential inequalities for empirical process. By definition of $\mathcal{W}_t(.,.)$ above, we have

$$\sum_{i=1}^{n} \mathbb{E}_f \mathcal{W}_t^2(X_i,\xi_i) \le \dot{\rho}_\infty^2\, K_\infty^2, \qquad \|\mathcal{W}_t(.,.)\|_\infty \le \dot{\rho}_\infty\, K_\infty/\sqrt{nh^d}, \tag{7.14}$$

where $\|\cdot\|_\infty$ is the sup-norm.

For the control of the first probability of (7.12), we use the Bernstein's inequality (7.8), then

$$\mathbb{P}_f\left(\sqrt{nh^d}\left|\tilde{D}_h^p(t_0) - \mathcal{E}_h^p(t_0)\right| \ge \frac{z}{2} - \frac{b_h\sqrt{nh^d}}{2}\right)$$
$$\le 2\exp\left\{-\frac{\left(z - b_h\sqrt{nh^d}\right)^2}{8\dot{\rho}_\infty^2\, K_\infty^2 + \frac{4\dot{\rho}_\infty K_\infty}{3\sqrt{nh^d}}\left(z - b_h\sqrt{nh^d}\right)}\right\}. \tag{7.15}$$

The second probability can be bounded as follows:

$$\mathbb{P}_f\left(\sqrt{nh^d}\sum_{l=1}^{\infty}\sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\le 10^{-l}}}\big|L(u)-L(v)\big| \ge \frac{z}{2} - \frac{b_h\sqrt{nh^d}}{2}\right)$$
$$\le \mathbb{P}_f\left(\sqrt{nh^d}\sum_{l=1}^{\infty}\frac{1}{l^2}\sup_{l\ge1} l^2 \sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\le 10^{-l}}}\big|L(u)-L(v)\big| \ge \frac{z}{2} - \frac{b_h\sqrt{nh^d}}{2}\right)$$
$$\le \sum_{l=1}^{\infty}\sum_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\le 10^{-l}}}\mathbb{P}_f\left(\sqrt{nh^d}\,\frac{\pi^2}{6}l^2\,\big|L(u)-L(v)\big| \ge \frac{z}{2} - \frac{b_h\sqrt{nh^d}}{2}\right). \tag{7.16}$$

In view of (7.13), we notice that

$$\sqrt{nh^d}\big[L(u)-L(v)\big] = \sum_{i=1}^{n}\mathcal{W}_u\big(X_i,\xi_i\big)-\mathcal{W}_v\big(X_i,\xi_i\big)-\mathbb{E}_f\left[\mathcal{W}_u\big(X_i,\xi_i\big)-\mathcal{W}_v\big(X_i,\xi_i\big)\right],$$

then we have a sum of independent zero-mean random variables with finite variance and bounded. Since ρ' is assumed Lipschitz, we have the following assertions.

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}_f\left[\mathcal{W}_u(X_i,\xi_i)-\mathcal{W}_v(X_i,\xi_i)\right]^2 &\leq K_\infty^2\|u-v\|_1^2,\\ \|\mathcal{W}_u(.,.)-\mathcal{W}_v(.,.)\|_\infty &\leq K_\infty\|u-v\|_1/\sqrt{nh^d},\end{aligned}$$

Using (7.16), the Bernstein's inequality (7.8) and the last three inequalities, we obtain

$$\begin{aligned}&\mathbb{P}_f\left(\sqrt{nh^d}\sum_{l=1}^{\infty}\sup_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq 10^{-l}}}\big|L(u)-L(v)\big|\geq\frac{z}{2}-\frac{b_h\sqrt{nh^d}}{2}\right)\\ &\leq 2\sum_{l=1}^{\infty}\sum_{\substack{u,v\in\Gamma_l\times\Gamma_{l-1}\\ \|u-v\|_1\leq 10^{-l}}}\exp\left\{-\frac{36\|u-v\|_1^{-1}}{\pi^4\,l^4}\right.\\ &\qquad\qquad\left.\times\frac{\left(z-b_h\sqrt{nh^d}\right)^2}{8K_\infty^2\,\|u-v\|_1+\frac{4K_\infty}{3\sqrt{nh^d}}\left(z-b_h\sqrt{nh^d}\right)}\right\}\\ &\leq 2\sum_{l=1}^{\infty}\#(\Gamma_l)\,\#(\Gamma_{l-1})\,\exp\left\{-\frac{36\;10^l}{\pi^4\,l^4}\frac{\left(z-b_h\sqrt{nh^d}\right)^2}{8K_\infty^2+\frac{4K_\infty}{3\sqrt{nh^d}}\left(z-b_h\sqrt{nh^d}\right)}\right\},\end{aligned}\tag{7.17}$$

where $\#(\Gamma_l)$ is the cardinal of Γ_l. Moreover, we notice that $\#(\Gamma_l)\leq d10^l$. Recall that $z\geq 2\big(1\vee b_h\sqrt{nh^d}\big)$ and we notice that $\min_{l\in\mathbb{N}^*}\frac{18\;10^l}{\pi^4\,l^4}>1$. The last assertions allows us to write that

$$\begin{aligned}&\exp\left\{-\frac{36\;10^l}{\pi^4\,l^4}\frac{\left(z-b_h\sqrt{nh^d}\right)^2}{8K_\infty^2+\frac{4K_\infty}{3\sqrt{nh^d}}\left(z-b_h\sqrt{nh^d}\right)}\right\}\\ &\quad\leq\exp\left\{-\frac{18\;10^l}{\pi^4\,l^4}\frac{(8K_\infty)^{-1}}{K_\infty+1/3}\right\}\times\exp\left\{-\frac{\left(z-b_h\sqrt{nh^d}\right)^2}{8K_\infty^2+\frac{4K_\infty}{3\sqrt{nh^d}}\left(z-b_h\sqrt{nh^d}\right)}\right\}.\end{aligned}$$

Using (7.11), (7.12), (7.15), (7.17) and the last inequality, we have for any $p \in \mathcal{S}_b$

$$\mathbb{P}_f\left(\sqrt{nh^d}\sup_{t\in\Theta(M)}\left|\tilde{D}_h^p(t)-\mathcal{E}_h^p(t)\right|\geq z\right)$$
$$\leq \Sigma\exp\left\{-\frac{\left(z-b_h\sqrt{nh^d}\right)^2}{4K_\infty^2\left(1\vee\dot{\rho}_\infty^2\right)+\frac{4K_\infty}{3\sqrt{nh^d}}\left(1\vee\dot{\rho}_\infty\right)z}\right\},$$

where Σ is defined in (2.8). This concludes the proof of Lemma 3. ∎

7.4 Proof of Lemma 4

Remember that the event $\bar{G}_\delta^h$ can be written as $\bar{G}_\delta^h = \left\{\hat{\theta}(h)\notin\mathcal{B}(\theta,\delta)\right\}$ and $\hat{\theta}(h)$ and θ are respectively the solutions of equations $\tilde{D}_h(\cdot)=0$ and $D_h(\cdot)=0$. Moreover θ is the unique solution of $D_h(\cdot)=0$, then we can notice the following inclusion

$$\left\{\hat{\theta}(h)\notin\mathcal{B}(\theta,\delta)\right\}\subseteq\left\{\sup_{t\in\Theta(M)\backslash\mathcal{B}(\theta,\delta)}\|\tilde{D}_h(t)-D_h(t)\|_2\geq\varkappa_\delta\right\}, \qquad (7.18)$$

where $\varkappa_\delta=\inf_{t\in\Theta(M)\backslash\mathcal{B}(\theta,\delta)}\|D_h(t)\|_2/2$. The latter inclusion can be interpreted as follows. In view of Lemma 1 *(1)*, θ is the unique solution of $D_h(\cdot)=0$ thus $D_h(\cdot)$ is not null on $\Theta(M)\backslash\mathcal{B}(\theta,\delta)$. Moreover, $D_h(\cdot)$ does not depend of n, then $\varkappa_\delta$ is positive and does not depend on n. The event $\left\{\hat{\theta}(h)\notin\mathcal{B}(\theta,\delta)\right\}$ implies that there exists $\tilde{\theta}\in\Theta(M)\backslash\mathcal{B}(\theta,\delta)$ such that $\tilde{D}_h\big(\tilde{\theta}\big)=0$, then on a neighborhood of $\tilde{\theta}$, $D_h(\cdot)$ and $\tilde{D}_h(\cdot)$ are not closed. So, there exists $\bar{\theta}\in\Theta(M)\backslash\mathcal{B}(\theta,\delta)$ such that

$$\|\tilde{D}_h\big(\bar{\theta}\big)-D_h\big(\bar{\theta}\big)\|_2\geq\varkappa_\delta.$$

Then, the latter inequality implies (7.18) by passing to the supremum.

Applying the inclusion (7.18), we obtain

$$\mathbb{P}_f\left(\bar{G}_\delta^h\right)\leq\sum_{p\in\mathcal{S}_b}\mathbb{P}_f\left(\sqrt{nh^d}\sup_{t\in\Theta(M)\backslash\mathcal{B}(\theta,\delta)}\left|\tilde{D}_h^p(t)-D_h^p(t)\right|>\frac{\sqrt{nh^d}\varkappa_\delta}{\sqrt{N_b}}\right)$$

Assumptions on n,h in Lemma 4 allow us to show that $\sqrt{nh^d}\varkappa_\delta/\sqrt{N_b} \geq 2\big(1 \vee\ b_h\sqrt{nh^d}\big)$. Using Lemma 3 with $z = \sqrt{nh^d}\varkappa_\delta/\sqrt{N_b}$, we have

$$\mathbb{P}_f\left(\bar{G}^h_\delta\right) \leq N_b \varSigma\ \exp\left\{-\frac{nh^d\left(\varkappa_\delta/2\sqrt{N_b}\right)^2}{8K^2_\infty\left(1 \vee \dot{\rho}^2_\infty\right) + \frac{4\varkappa_\delta}{3\sqrt{N_b}}K_\infty\left(1 \vee \dot{\rho}_\infty\right)}\right\}.$$

The lemma is proved. ∎

7.5 Proof of Lemma 5

Note that by definition of $\hat{k}$ in (4.5)

$$\forall k \geq \kappa+1, \quad \big\{\hat{k}=k\big\} = \cup_{l \geq k}\left\{\big|\hat{f}^{(k-1)}(x_0) - \hat{f}^{(l)}(x_0)\big| > C\ S_n(l)\right\}.$$

Note that $S_n(l)$ is monotonically increasing in l and, therefore,

$$\begin{aligned}\big\{\hat{k}=k\big\} \ \subseteq\ & \left\{\big|\hat{f}^{(k-1)}(x_0) - f(x_0)\big| > 2^{-1}C\ S_n(k-1)\right\} \\ & \cup\left[\cup_{l \geq k}\left\{\big|\hat{f}^{(l)}(x_0) - f(x_0)\big| > 2^{-1}C\ S_n(l)\right\}\right].\end{aligned}$$

We come to the following inequality: for any $k \geq \kappa+1$

$$\begin{aligned}\mathbb{P}\left(\hat{k}=k, G^{h_k}_\delta\right) \ \leq\ & \mathbb{P}\left\{\big|\hat{f}^{(k-1)}(x_0) - \hat{f}(x_0)\big| > 2^{-1}C\ S_n(k-1), G^{h_k}_\delta\right\} \\ & + \sum_{l \geq k}\mathbb{P}\left\{\big|\hat{f}^{(l)}(x_0) - f(x_0)\big| > 2^{-1}C\ S_n(l), G^{h_k}_\delta\right\}. \quad (7.19)\end{aligned}$$

Notice that the definition of $S_n(l)$ yields $N_{h_l}\ S_n(l) = \big[1 + l\ln 2\big]^{1/2}$. Thus, applying Proposition 1 with $\varepsilon = C\big[1 + l\ln 2\big]^{1/2}$ and $h = h_l$ and using the inequality (6.9), we obtain by definition of C in (4.6), for any $l \geq k-1$ and n large enough

$$\mathbb{P}\left\{\big|\hat{f}^{(l)}(x_0) - f(x_0)\big| > 2^{-1}C\ S_n(l)\right\} \ \leq\ N_b\varSigma\, 2^{-2rd\,l}. \qquad (7.20)$$

We obtain from (7.19) and (7.20) that $k \geq \kappa+1$

$$\mathbb{P}\left(\hat{k}=k, G^{h_k}_\delta\right) \leq J2^{-2(k-1)rd},$$

where $J = N_b\varSigma\big(1 + (1-2^{-2rd})^{-1}\big)$. ∎

References

[1] J. Astola, K. Egiazarian, A. Foi, and V. Katkovnik. From local kernel to nonlocal multiple-model image denoising. *Int. J. Comput. Vision*, 86(1):1–32, 2010.

[2] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[3] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[4] L. Brown and M. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, 24(6):2524–2535, 1996.

[5] L.D. Brown, T. Tony Cai, and H.H. Zhou. Robust nonparametric estimation via wavelet median regression. *Ann. Statist.*, 36(5):2055–2084, 2008.

[6] X.W. Chang and Y. Guo. Huber's m-estimation in relative gps positioning: computational aspects. *Journal of Geodesy.*, 2005.

[7] M. Chichignoud. Minimax and minimax adaptive estimation in multiplicative regression : locally bayesian approach. *Probab. Theory Related Fields*, 2011. to appear.

[8] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. With discussion and a reply by the authors.

[9] A. Goldenshluger and O.V. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(3):1150–1190, 2008.

[10] A. Goldenshluger and O.V. Lepski. Structural adaptation via lp-norm oracle inequalities. *Probab. Theory and Related Fields*, 143:41–71, 2009.

[11] A. Goldenshluger and A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2):135–170, 1997.

[12] Peter Hall and M. C. Jones. Adaptive M-estimation in nonparametric regression. *Ann. Statist.*, 18(4):1712–1728, 1990.

[13] W. Härdle and A.B. Tsybakov. Robust nonparametric regression with simultaneous scale curve estimation. *Ann. Statist.*, 16(1):120–135, 1988.

[14] W. Härdle and A.B. Tsybakov. Robust locally adaptive nonparametric regression. In *Data analysis and statistical inference*, pages 127–144. Eul, Bergisch Gladbach, 1992.

[15] P. Huber and E. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2009.

[16] P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35: 73–101, 1964.

[17] A. Juditsky. Wavelet estimators: adapting to unknown smoothness. *Math. Methods Statist.*, 6(1):1–25, 1997.

[18] A.B. Juditsky, O.V. Lepski, and A.B. Tsybakov. Nonparametric estimation of composite functions. *Ann. Statist.*, 37(3):1360–1404, 2009.

[19] V. Katkovnik. *Nonparametric identification and data smoothing.* "Nauka", Moscow (in Russian), 1985. The method of local approximation.

[20] G. Kerkyacharian, O.V. Lepski, and D. Picard. Non linear estimation in anisotropic multi-index denoising. *Probab. Theory and Related Fields*, 121:137–170, 2001.

[21] N. Klutchnikoff. *On the adaptive estimation of anisotropic functions.* PhD thesis, Aix-Masrseille 1, 2005.

[22] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997.

[23] O.V. Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35(3):454–466, 1990.

[24] O.V. Lepski. Asymptotically minimax adaptive estimation i. upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697, 1991.

[25] O.V. Lepski and B.Y. Levit. Adaptive nonparametric estimation of smooth multivariate functions. *Mathematicals methods of statistics*, 1999.

[26] O.V. Lepski and V.G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Annals of statistics*, 25(6):2512–2546, 1997.

[27] P. Massart. Some applications of concentration inequalities to statistics. *Probability Theory*, volume spécial dédié à Michel Talagrand(2):245–303, 2000.

[28] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[29] P. Petrus. Robust huber adaptive filter. *IEEE Transactions on Signal Processing.*, 47:1129–1133, 1999.

[30] M. Reiss, Y. Rozenholc, and C. Cuenod. Pointwise adaptive estimation for robust and quantile regression. 2011. Source: Arxiv.

[31] P. Rousseeuw and A. Leroy. *Robust regression and outlier detection.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987.

[32] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996a.

[33] M. Talagrand. A new look at independence. *Annals of Probability*, 24:1–34, 1996b.

[34] A. B. Tsybakov. Nonparametric signal estimation when there is incomplete information on the noise distribution. *Problems of Information Transmission*, 18(2):116–130, 1982a.

[35] A. B. Tsybakov. Robust estimates of a function. *Problems of Information Transmission*, 18(3):39–52, 1982b.

[36] A. B. Tsybakov. Convergence of nonparametric robust algorithms of reconstruction of functions. *Automation and Remote Control*, (12):66–76, 1983.

[37] A. B. Tsybakov. Robust reconstruction of functions by a local approximation method. *Problems of Information Transmission*, 22(2):69–84, 1986.

[38] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated, 2008.

[39] A.B. Tsybakov. Pointwise and sup-norm sharp adaptive estimation of function on the sobolev classes. *Annals of statistics*, 26(6):2420–2469, 1998.

[40] Aad W. Van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.