ESAIM: PS
August 2007, Vol. 11, p. 344–364
DOI: 10.1051/ps:2007023

ESAIM: Probability and Statistics
www.edpsciences.org/ps

ON POINTWISE ADAPTIVE CURVE ESTIMATION BASED ON INHOMOGENEOUS DATA

Stéphane Gaïffas[1]

Abstract. We want to recover a signal based on noisy inhomogeneous data (the amount of data can vary strongly on the estimation domain). We model the data using nonparametric regression with random design, and we focus on the estimation of the regression at a fixed point x_0 with little, or much data. We propose a method which adapts both to the local amount of data (the design density is unknown) and to the local smoothness of the regression function. The procedure consists of a local polynomial estimator with a Lepski type data-driven bandwidth selector, see for instance Lepski *et al.* [15]. We assess this procedure in the minimax setup, over a class of function with local smoothness $s > 0$ of Hölder type. We quantify the amount of data at x_0 in terms of a local property on the design density called regular variation, which allows situations with strong variations in the concentration of the observations. Moreover, the optimality of the procedure is proved within this framework.

Mathematics Subject Classification. 62G05, 62G08.

Received February 6, 2006. Revised June 20 and October 10, 2006.

1. Introduction

1.1. The model

We observe n pairs of random variables $(X_i, Y_i) \in \mathbb{R} \times \mathbb{R}$ independent and identically distributed satisfying

$$Y_i = f(X_i) + \xi_i, \tag{1.1}$$

where $f : [0,1] \to \mathbb{R}$ is the unknown signal to be recovered, the variables (ξ_i) are centered Gaussian with known variance σ^2 and independent of the design $X_1, \ldots, X_n$. The variables X_i are distributed with respect to an unknown density μ. We want to recover f at a fixed point x_0.

The classical way of considering the nonparametric regression model is to take deterministic $X_i = i/n$: in this model with an equispaced design, the observations are *homogeneously* distributed over the unit interval. If we take random X_i, we can model cases with *inhomogeneous* observations as the design distribution is "far" from the uniform law. In particular, in order to include situations with little or much data in the model, we allow the density μ to be *degenerate* (vanishing or exploding) at x_0. In this problem, we are interested in the adaptive estimation of f at x_0, both adaptive to the smoothness of f and to the inhomogeneity of the data.

Keywords and phrases. Adaptive estimation, inhomogeneous data, nonparametric regression, random design.

[1] Laboratoire de Probabilités et Modèles Aléatoires, U.M.R. CNRS 7599 and Université Paris 7, 175 rue du Chevaleret, 75013 Paris, France; gaiffas@math.jussieu.fr

1.2. **Motivations**

The adaptive estimation of the regression is a well-developed problem. Several adaptive procedures can be applied for the reconstruction of a signal with unknown smoothness: nonlinear wavelet estimation (thresholding), model selection, kernel estimation with a variable bandwidth (the Lepski method), and so on. Recent results dealing with the adaptive estimation of the regression function when the design is not equispaced or random include Antoniadis *et al.* [1], Baraud [2], Brown and Cai [4], Wong and Zheng [21], Maxim [17], Delouille *et al.* [7], Kerkyacharian and Picard [12], among others.

Here, we focus on a slightly different problem: our aim is to recover the signal locally, based on data which can be eventually very inhomogeneous. More precisely, we want to handle simultaneously situations where the observations are very concentrated at the estimation point, or conversely, very deficient, with the aim to illustrate the consequences of inhomogeneity on the accuracy of estimation within the theory. The minimax rates associated to this estimation problem are computed in Gaïffas [10], under several types of behaviours for the design density. The estimator proposed therein adapts to the inhomogeneity of the data, but not to the smoothness of the regression. Therefore, the results presented here extend Gaïffas [10], since the procedure constructed in the next section has both properties of smoothness adaptation, and “design adaptation”.

1.3. **Organisation of the paper**

In the next section, we construct the adaptive estimator, and we assess this estimator in Section 3. First, we give an upper bound in Theorem 1 which is stated conditionally on the design. Then, we propose in Section 3.2 a way of quantifying the local inhomogeneity of the data with an appropriate assumption on the local behaviour of the design density. Under this assumption, we provide another upper bound in Theorem 2. In Section 4, we discuss the optimality of the estimator, and we prove in Theorem 3 that the convergence rate from Theorem 2 is optimal. Section 5 is devoted to the proofs and some well-known analytic facts are briefly recalled in appendix.

2. Construction of the adaptive procedure

The procedure described here is a local polynomial estimator with an adaptive data-driven selection of the bandwidth (the design density and the smoothness are both unknown). For any $\Delta \subset [0,1]$, we define the empirical sample measure

$$\bar\mu_n(\Delta) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_\Delta(X_i),$$

where $\mathbf{1}_\Delta$ is the indicator of Δ, and if $\bar\mu_n(\Delta) > 0$, we introduce the pseudo-inner product

$$\langle f \,,\, g\rangle_\Delta := \frac{1}{\bar\mu_n(\Delta)}\int_\Delta f g \,\mathrm{d}\bar\mu_n, \tag{2.1}$$

and $\|g\|_\Delta := \langle g \,,\, g\rangle_\Delta^{1/2}$ the corresponding pseudo-norm.

2.1. **Local polynomial estimation**

We fix $K \in \mathbb{N}$ and an interval $I \subset [0,1]$, which is a smoothing parameter that we call *bandwidth.* The idea is to look for the polynomial $\bar f_I$ of order K which is the closest to the data in the least square sense, with respect to the localised design-adapted norm $\|\cdot\|_I$:

$$\bar f_I := \operatorname*{argmin}_{g\in V_K} \|Y - g\|_I^2, \tag{2.2}$$

where V_K is the set of all real polynomials of order at most K. We can rewrite (2.2) in a variational form, in which we look for $\bar f_I \in V_K$ such that for any $\phi \in V_K$,

$$\langle \bar f_I \,,\, \phi\rangle_I = \langle Y \,,\, \phi\rangle_I, \tag{2.3}$$

where it suffices to consider only the power functions $\phi_p(\cdot) = (\cdot - x_0)^p$, $0 \leqslant p \leqslant K$. The coefficients vector $\bar\theta_I \in \mathbb{R}^{K+1}$ of the polynomial $\bar f_I$ is therefore solution, when it makes sense, of the linear system

$$\mathbf{X}_I \theta = \mathbf{Y}_I,$$

where for $0 \leqslant p, q \leqslant K$:

$$(\mathbf{X}_I)_{p,q} := \langle \phi_p \, , \, \phi_q \rangle_I \quad \text{and} \quad (\mathbf{Y}_I)_p := \langle Y \, , \, \phi_p \rangle_I. \tag{2.4}$$

The parameter $f(x_0)$ is then estimated by $\bar f_I(x_0)$. This linear method of estimation, called *local polynomial estimator* is well-known, see for instance Stone [19], Fan and Gijbels [8, 9] and Tsybakov [20] among many others.

In this paper, we work with a slightly modified version of the local polynomial estimator, which is convenient in situations with little or much data. When the smallest eigenvalue of the non-negative matrix $\mathbf{X}_I$ is too small, we add a correcting term allowing to bound it from below: we introduce

$$\bar{\mathbf{X}}_I := \mathbf{X}_I + (n\bar\mu_n(I))^{-1/2} \mathbf{Id}_{K+1} \mathbf{1}_{\Omega_I^c},$$

where $\mathbf{Id}_{K+1}$ is the identity matrix in $\mathbb{R}^{K+1}$ and

$$\Omega_I := \big\{ \lambda(\mathbf{X}_I) > (n\bar\mu_n(I))^{-1/2} \big\}, \tag{2.5}$$

where $\lambda(M)$ stands for the smallest eigenvalue of a matrix M. The quantity $(n\bar\mu_n(I))^{-1/2}$ comes from the variance of $\bar f_I$, and this particular choice preserves the convergence rate of the method. Then, when $\bar\mu(I) > 0$, we consider the solution $\widehat\theta_I$ of the linear system

$$\bar{\mathbf{X}}_I \theta = \mathbf{Y}_I, \tag{2.6}$$

and denote by $\widehat f_I \in V_K$ the polynomial with coefficients $\widehat\theta_I$. When $\bar\mu_n(I) = 0$, we take simply $\widehat f_I := 0$.

The local polynomial estimator is convenient when dealing with a random design, as shown in Fan and Gijbels [9], since its pointwise error has a very tractable decomposition. Conditionally on the design, we can decompose the pointwise error $|\widehat f_I(x_0) - f(x_0)|$ between a bias term of order $|I|^s$ ($|I|$ standing for the length of I) when f is s-Hölder and a variance term of order $(n\bar\mu_n(I))^{-1/2}$, for a general design, in a non-asymptotic way (see Lemma 3 below). Since the optimal bandwidth I makes the balance between the bias and the variance of $\widehat f_I$, it depends on the local smoothness of f *via* the bias term. Therefore, an adaptive technique is required when the smoothness is unknown, which is the case in practical situations.

2.2. Adaptive bandwidth selection

The adaptive procedure described here is based on a method introduced by Lepski [14], see also Lepski *et al.* [15], and Lepski and Spokoiny [16]. If a family of linear estimators can be "well-sorted" by their respective variances (*e.g.* kernel estimators in the white noise model, see Lepski and Spokoiny [16]), the Lepski procedure selects the largest bandwidth such that the corresponding estimator does not differ "significantly" from estimators with a smaller bandwidth. Following this principle, we construct a method which adapts to the unknown smoothness, and additionally to the original Lepski method, to the distribution of the data (the design density is unknown), in particular in cases with little or much data. Bandwidth selection procedures in local polynomial estimation can be found in Fan and Gijbels [8], Goldenshluger and Nemirovski [11] or Spokoiny [18].

The idea of the adaptive procedure is the following: when $\widehat f_I$ is close to f (that is, when I is well-chosen), we have in view of (2.3)

$$\langle \widehat f_J - \widehat f_I \, , \, \phi \rangle_J = \langle Y - \widehat f_I \, , \, \phi \rangle_J \approx \langle Y - f \, , \, \phi \rangle_J = \langle \xi \, , \, \phi \rangle_J,$$

for any $J \subset I$, $\phi \in V_K$, where the right-hand side is a noise term. Then, in order to "remove" this noise, we select the largest I such that this noise term remains smaller than an appropriate threshold, for any $J \subset I$ and $\phi = \phi_p$, $0 \leqslant p \leqslant K$. The bandwidth is selected in a fixed set of intervals G_n called *grid* (which is a tuning parameter of the procedure that we describe below) as follows:

$$\widehat{I}_n := \operatorname*{argmax}_{I \in G_n} \Big\{ \bar{\mu}_n(I) \mid \forall J \in G_n, J \subset I, \ \forall m \in \{0, \ldots, K\}, \\ |\langle \widehat{f}_J - \widehat{f}_I \,,\, \phi_m \rangle_J| \leqslant \|\phi_m\|_J T_n(I,J) \Big\}, \tag{2.7}$$

where

$$T_n(I,J) := \sigma \Big[\Big(\frac{2 \log n}{n \bar{\mu}_n(I)} \Big)^{1/2} + D C_K \Big(\frac{\log(n \bar{\mu}_n(I))}{n \bar{\mu}_n(J)} \Big)^{1/2} \Big], \tag{2.8}$$

with $C_K := 1 + (K+1)^{1/2}$ and $D > 0$ is specified later on. The estimator is then given by

$$\widehat{f}_n(x_0) := \widehat{f}_{\widehat{I}_n}(x_0). \tag{2.9}$$

The threshold choice (2.8) can be understood in the following way: since the variance of $\widehat{f}_I$ is of order $(n\bar{\mu}_n(I))^{-1/2}$, we see that the two terms in $T_n(I,J)$ are ratios between a penalizing log term and the variance of the estimators compared by the rule (2.7). The penalization term is linked with the number of comparisons necessary for selecting the bandwidth.

Within the procedure, we have mainly two choices of grid. The first one is the following: we sort the (X_i, Y_i) into $(X_{(i)}, Y_{(i)})$ such that $X_{(i)} < X_{(i+1)}$. Then, we consider j such that $x_0 \in [X_{(j)}, X_{(j+1)}]$ (if necessary, we take $X_{(0)} = 0$ and $X_{(n+1)} = 1$) and for some $a > 1$ we introduce

$$G_n := \bigcup_{p=0}^{[\log_a(j+1)]} \bigcup_{q=0}^{[\log_a(n-j)]} \Big\{ \big[X_{(j+1-[a^p])}, X_{(j+[a^q])} \big] \Big\}. \tag{2.10}$$

The selection of the bandwidth within this grid is fast, since its cardinality is $O((\log n)^2)$. Another example of grid is given by

$$G_n := \bigcup_{1 \leqslant p < q \leqslant n} \Big\{ \big[x_0 - |X_{(p)} - x_0|, x_0 + |X_{(q)} - x_0| \big] \Big\}, \tag{2.11}$$

where the cardinality is $O(n^2)$, which increases rapidly with the sample size.

Remark. The estimator $\widehat{f}_n(x_0)$ only depends on K and on the grid G_n (which are parameters chosen by the statistician). It does not depend within its construction on μ nor the smoothness of f. In this sense, this estimator is both smoothness-adaptive and design-adaptive.

3. Assessment of the procedure: upper bounds

3.1. Conditionally on the design

When no assumption is made on the local behaviour of the design density μ, we can work conditionally on the design. The procedure is assessed in the following way: first, we consider an ideal *oracle* interval given by

$$I_{n,f}^* := \operatorname*{argmax}_{I \subset [0,1],\, x_0 \in I} \Big\{ \bar{\mu}_n(I) \mid \operatorname{osc} f(I) \leqslant \sigma \Big(\frac{2 \log n}{n \bar{\mu}_n(I)} \Big)^{1/2} \Big\}, \tag{3.1}$$

where $\operatorname{osc} f(I)$ is the local oscillation of f in I, defined by

$$\operatorname{osc} f(I) := \inf_{P \in V_K} \sup_{y \in I} |f(y) - P(y)|, \tag{3.2}$$

where we recall that V_K is the set of all real polynomials with order at most K. The local oscillation is a common way of measuring the smoothness of a function.

The interval $I^*_{n,f}$, which is not necessarily unique, makes the balance between the bias and the $\log n$-penalised variance of $\widehat{f}_I$. Therefore, it can be understood as an *ideal adaptive bandwidth*, see Lepski and Spokoiny [16] and Spokoiny [18]. The $\log n$ term in (3.1) is the *payment for adaptation*, see Section 4.1. We use the word "oracle" since this interval depends on f directly. This oracle interval is used to define

$$R_{n,f} := \sigma\Big(\frac{\log n}{n\bar{\mu}_n(I^*_{n,f})}\Big)^{1/2}, \tag{3.3}$$

which is a random normalisation (it depends on the local amount of data) assessing the adaptive procedure in Theorem 1 below. We introduce also

$$\bar{I}_{n,f} := \operatorname*{argmax}_{I\in G_n}\Big\{\bar{\mu}_n(I) \mid \operatorname{osc} f(I) \leqslant \sigma\Big(\frac{2\log n}{n\mu_n(I)}\Big)^{1/2}\Big\}, \tag{3.4}$$

which is an oracle interval in the grid, and we define the matrices

$$\mathbf{\Lambda}_I := \operatorname{diag}(\|\phi_0\|_I^{-1}, \ldots, \|\phi_K\|_I^{-1}) \text{ and } \mathbf{E}_I := \mathbf{\Lambda}_I\bar{\mathbf{X}}_I\mathbf{\Lambda}_I. \tag{3.5}$$

We denote by $\mathfrak{X}_n$ the sigma-algebra generated by $X_1, \ldots, X_n$, by $\mathbb{E}^n_{f,\mu}$ the expectation with respect to the joint law $\mathbb{P}^n_{f,\mu}$ of the observations (1.1) and by $\lambda(M)$ the smallest eigenvalue of a matrix M. We recall that Ω_I is defined by (2.5).

Theorem 1. *When $\|f\|_\infty < +\infty$, we have on $\Omega_{\bar{I}_{n,f}} \cap \big\{n\bar{\mu}_n(\bar{I}_{n,f}) \geqslant 2\big\}$ for any $p > 0$, $n \geqslant K+1$ and $D \geqslant 4(p+1)^{1/2}$ (see (2.8)):*

$$\mathbb{E}^n_{f,\mu}\big\{|\widehat{f}_n(x_0) - f(x_0)|^p|\mathfrak{X}_n\big\} \leqslant CR^p_{n,f}\big[\lambda(\mathbf{E}_{\bar{I}_{n,f}})^{-p} + (\|f\|_\infty \vee 1)^p\big],$$

where C is a constant depending on p, K, a.

Remark. The fact that Theorem 1 is stated over $\Omega_{\bar{I}_{n,f}} \cap \{n\bar{\mu}_n(\bar{I}_{n,f}) \geqslant 2\}$ put some constraints between f, n and $\bar{\mu}_n$. Indeed, on this set, we have that roughly, the local oscillation of f cannot be too large when the local amount of data is too small. In the next section, we show that for n large enough, this event has a large probability under appropriate assumptions on the design density and the smoothness of f.

Remark. The upper bound in Theorem 1 is non-asymptotic since it holds for any $n \geqslant K+1$. When $n < K+1$, $\mathbf{X}_I$ is degenerate and Ω_I is empty for any I, since $\mathbf{X}_I = \mathbf{F}_I\mathbf{F}'_I$ where $\mathbf{F}_I$ is the matrix of size $n \times (K+1)$ with entries $(\mathbf{F}_I)_{i,m} = (X_i - x_0)^m$ for $0 \leqslant i \leqslant n$ and $0 \leqslant m \leqslant K$.

3.2. How to quantify the local inhomogeneity of the data?

In this section, we propose a way of modeling situations where the amount of data is large or little at the estimation point x_0. The idea is simple: we allow the design density μ to be vanishing or exploding at x_0 with a power function behaviour type, which is quantified by a coefficient β called *index of regular variation.* Regular variation is a well-known notion, commonly used for quantifying the asymptotic behaviour of probability queues. It is also intimately linked with the theory of extreme values. On regular variation, we refer to Bingham *et al.* [3].

Definition 1 (regular variation)**.** A function $g : \mathbb{R}^+ \to \mathbb{R}^+$ is regularly varying at 0 if it is continuous on $(0, +\infty)$, and if there is $\beta \in \mathbb{R}$ such that

$$\forall y > 0, \quad \lim_{h\to 0^+} g(yh)/g(h) = y^\beta. \tag{3.6}$$

We denote by $\mathrm{RV}(\beta)$ the set of all such functions. A function in $\mathrm{RV}(0)$ is *slowly varying.*

Remark. While a function $g \in \mathrm{RV}(\beta)$ goes to 0 at 0 when $\beta > 0$, or to $+\infty$ when $\beta < 0$, a slowing varying function ($\beta = 0$) can go to 0, to $+\infty$ or to some positive constant. Indeed, a typical example of slowly varying function is $(\log(1/h))^b$, which is slowly varying at 0 for any $b \in \mathbb{R}$. Some properties of regularly varying functions are given in appendix.

The assumption on μ is the following: we assume that there is $\nu > 0$ and $\beta > -1$ such that for any x, $|x - x_0| \leqslant \nu$:

$$\mu(x_0 + x) = \mu(x_0 - x) \text{ and } \mu(x_0 + \cdot\,) \in \mathrm{RV}(\beta). \tag{3.7}$$

This assumption means that μ is symmetrical within a neighbourhood of x_0, and varies regularly (on both sides). Note that this assumption includes the classical case where μ is positive and continuous at x_0, and that in this case, $\beta = 0$. The local symmetry assumption is not necessary, but made in order to simplify the presentation of the material.

3.3. Function class

In what follows, we use the notation $I_h := [x_0 - h, x_0 + h]$. For measuring the local smoothness, we consider the class of signals with local oscillation bounded by a function in $\mathrm{RV}(s)$, for $s > 0$.

Definition 2. If $\omega \in \mathrm{RV}(s)$ for some $s > 0$ and $Q, \delta > 0$, we introduce

$$\mathcal{F}_\delta(\omega, Q) := \big\{ f : \mathbb{R} \to \mathbb{R} \text{ s.t. } \|f\|_\infty \leqslant Q \text{ and } \forall h \leqslant \delta,\ \operatorname{osc} f(I_h) \leqslant \omega(h) \big\},$$

where we recall that $\operatorname{osc} f(I)$ is the local oscillation of f around x_0, see (3.2).

This function class contains Hölder balls. Indeed, the set of all the functions f, $\|f\|_\infty \leqslant Q$, such that for the largest integer $k < s$:

$$|f^{(k)}(x) - f^{(k)}(x_0)| \leqslant L|x - x_0|^{s-k}, \quad \forall x,\ |x - x_0| \leqslant \delta, \tag{3.8}$$

where $f^{(k)}$ is the k-th derivative of f is included in $\mathcal{F}_\delta(\omega, Q)$ for $\omega(h) = Lh^s/k!$, which is in particular s-regularly varying. The parameter δ, assumed to be small (eventually going to 0 with n, see Th. 2 below), is the length of the interval in which the smoothness assumption is made: this assumption is local, since we are interested in pointwise estimation. The parameter Q can be arbitrary large, but fixed. We need such a parameter since the upper bound is stated uniformly over a collection of such classes (see Th. 2 below).

3.4. Minimax adaptive upper bound

In this section, we assess the adaptive procedure $\widehat{f}_n$ in the minimax adaptive framework under assumption (3.7), which is an assumption quantifying the local amount the data. Throughout what follows, we use the notation $\mu(I) := \int_I \mu(t)\mathrm{d}t$. We introduce $h_n(\omega, \mu)$ as the smallest solution to

$$\omega(h) = \sigma\Big(\frac{\log n}{n\mu(I_h)}\Big)^{1/2}. \tag{3.9}$$

This quantity is well defined as n is large enough, since $h \mapsto \omega(h)^2\mu(I_h)$ is continuous and vanishing at 0. This equation is the deterministic counterpart (among symmetrical intervals) of the bias-variance equation (3.1). We introduce also

$$r_n(\omega, \mu) := \omega(h_n(\omega, \mu)), \tag{3.10}$$

which is the minimax adaptive convergence rate over the classes $\mathcal{F}_\delta(\omega, Q)$, see Theorem 2 for the upper bound, and Theorem 3 below for the lower bound. When $\omega \in \mathrm{RV}(s)$ and μ satisfies (3.7), we have in view of the properties (A.2) and (A.4) concerning regularly varying functions that $h \mapsto \omega(h)^2\mu(I_h) \in \mathrm{RV}(1+2s+\beta)$. Then, using together (A.2) and (A.5), we can find a slowly varying function $\ell_{\omega,\mu}$ such that

$$r_n(\omega,\mu) = (\log n/n)^{s/(1+2s+\beta)}\ell_{\omega,\mu}(\log n/n). \tag{3.11}$$

The design effect on this rate (*via* the parameter β) is comparable to that of a dimensionality parameter, or to the smoothing degree of an operator "blurring" the original signal f, when considering inverse problems. Note that in the classical case, that is over a s-Hölder ball ($\omega(h) = h^s$) and when μ is positive and continuous at x_0, (3.11) simplifies to

$$r_n(\omega,\mu) = (\log n/n)^{s/(1+2s)},$$

which is the usual pointwise adaptive minimax rate, see Lepski [13] and Brown and Low [5].

Theorem 2. *If*

- μ *satisfies* (3.7);
- $\omega \in \mathrm{RV}(s)$ *for some* $s \in (0, K+1]$;
- $p, Q > 0$ *and* (δ_n) *is a positive sequence such that* $\delta_n \geqslant \rho h_n(\omega,\mu)$, *where* $\rho > 1$ *is a fixed constant*;

the adaptive estimator $\widehat{f}_n(x_0)$ *defined by* (2.9), *with* $D \geqslant 4(p+1/2)^{1/2}$ *and grid choice*

$$G_n := \bigcup_{1\leqslant i\leqslant n} \Big\{ \big[x_0 - |X_i - x_0|, x_0 + |X_i - x_0|\big] \Big\} \tag{3.12}$$

satisfies

$$\sup_{f\in\mathcal{F}_{\delta_n}(\omega,Q)} \mathbb{E}^n_{f,\mu}\big\{|\widehat{f}_n(x_0) - f(x_0)|^p\big\} = O(r_n(\omega,\mu)^p) \tag{3.13}$$

for n *large enough, where* $r_n(\omega,\mu)$ *is given by* (3.10) *and* (3.11).

Remark. In this theorem, we assess the adaptive estimator constructed in Section 2 over classes with smoothness $s \in (0, K+1]$, where K is a tuning constant of the procedure. In the minimax framework considered here, the assumption of knowing an upper bound for s is usual in the study of adaptive methods, and somehow, unavoidable. For instance, when considering adaptive wavelet methods, the "maximum smoothness" corresponds to the number of vanishing moments of the mother wavelet.

Remark. The reason of considering the grid (3.12) in Theorem 2 is linked with the uniform control of the smallest eigenvalue of $\mathbf{E}_{\bar{I}_{n,f}}$, which is necessary for the proof of the upper bound since the method involves the resolution of the linear system (2.6) (see Th. 1). We can prove this theorem with the grid (2.10), which is more convenient in practice, with the extra assumption that $\lambda(\mathbf{E}_{n,f}) \geqslant \lambda$ for some $\lambda > 0$, uniformly over $\mathcal{F}_\delta(\omega, Q)$ for $\omega \in \mathrm{RV}(s)$, $0 < s \leqslant K+1$. However, we have chosen to provide the upper bound under the only assumption (3.7) on the design, which is used to quantify the local amount of data.

Remark. If there are $\beta^-, \beta^+ > -1$ such that $\mu(x_0 + \cdot) \in \mathrm{RV}(\beta^+)$ and $\mu(x_0 - \cdot) \in \mathrm{RV}(\beta^-)$, the result stated in Theorem 2 is the same. The convergence rate still satisfies (3.11), with $\beta = \min(\beta^-, \beta^+)$, which means that the side with the largest amount of data "dominates" (asymptotically) the other one.

3.5. Explicit examples of rates

In this section, we give some explicit rate examples, obtained by solving equation (3.9). Note that each example below is indeed of the form (3.11), and that they are optimal, see Section 4 below.

Example. Let μ be positive and continuous at x_0. Over a s-Hölder ball (see (3.8)), by solving (3.9) with $\omega(h) = h^s$ we find back the usual pointwise minimax adaptive rate (see Lepski [14], Brown and Low [5])

$$(\log n/n)^{s/(1+2s)}.$$

If $\omega(h) = h^s(\log(1/h))^{-s}$ ($\omega \in \mathrm{RV}(s)$), that is, we have locally more smoothness than in the s-Hölder case, the pointwise minimax adaptive rate over $\mathcal{F}_\delta(\omega, Q)$ is

$$n^{-s/(1+2s)}.$$

If G is a continuous function, we denote by $G^{\leftarrow}$ its pseudo-inverse, defined by $G^{\leftarrow}(y) := \inf\{h \geqslant 0 | G(h) \geqslant y\}$. We need the following lemma, which is proved in Gaïffas [10].

Lemma 1. *Let $a \in \mathbb{R}$ and $b > 0$. If $G(h) = h^b(\log(1/h))^a$, we have*

$$G^{\leftarrow}(h) \sim b^{a/b} h^{1/b} (\log(1/h))^{-a/b} \text{ as } h \to 0^+.$$

Example. Let μ be such that $\int_0^h \mu(x_0 + t)\mathrm{d}t = h^{\beta+1}(\log(1/h))^\alpha$ for any h in a neighbourhood of 0 and $\omega(h) = h^s(\log(1/h))^\gamma$ where $\beta > -1$, $s > 0$, $\alpha, \gamma \in \mathbb{R}$. If $G(h) := h^{1+2s+\beta}(\log(1/h))^{2\gamma+\alpha}$, equation (3.9) can be written as $G(h) = t_n$ where $t_n := \sigma^2 \log n/n$. Using Lemma 1 we obtain that

$$\big(n(\log n)^{\alpha-1-\gamma(1+\beta)/s}\big)^{-s/(1+2s+\beta)} \tag{3.14}$$

is the pointwise minimax adaptive rate over $\mathcal{F}_\delta(\omega, Q)$. This rate has to be compared with the pointwise minimax rate from Gaïffas [10]:

$$\big(n(\log n)^{\alpha-\gamma(1+\beta)/s}\big)^{-s/(1+2s+\beta)},$$

where the only difference with (3.14) is α instead of $\alpha - 1$ in the logarithmic exponent. This loss, often called *payment for adaptation* in the literature, is unavoidable in view of Theorem 3 below. Over the set of functions with bounded s-th derivative (s integer), since in this case $\omega(h) = h^s$, the rate (3.14) becomes

$$\big(n(\log n)^{\alpha-1}\big)^{-s/(1+2s+\beta)},$$

again when μ is such that $\int_0^h \mu(x_0 + t)\mathrm{d}t = h^{\beta+1}(\log(1/h))^\alpha$ for any h in a neighbourhood of 0.

4. Minimax adaptive optimality of the estimator

4.1. Payment for adaptation

When μ satisfies (3.7) and if $\omega \in \mathrm{RV}(s)$, we know from Gaïffas [10] that the minimax rate over $\mathcal{F}_\delta(\omega, Q)$ is equal to

$$n^{-s/(2s+1+\beta)}\ell_{\omega,\mu}(1/n), \tag{4.1}$$

where $\ell_{\omega,\mu}$ is a slowly varying function characterized by ω and μ. In Theorem 2, we proved that the adaptive estimator converges with the rate (3.11) which is slower than (4.1) because of the extra $\log n$ term. The aim of this section is to prove that this extra term in unavoidable.

In a model with homogeneous information (for instance white noise or regression with equidistant design), we know that adaptive estimation to the unknown smoothness without loss of efficiency is not possible for pointwise risks, even when we know that the unknown signal belongs to one of two Hölder classes, see Lepski [14], Brown and Low [5] and Lepski and Spokoiny [16]. This means that local adaptation cannot be achieved for free: we have to pay an extra factor in the convergence rate, at least of order $(\log n)^{2s/(1+2s)}$ when estimating a function with Hölder smoothness s. The authors call this phenomenon *payment for adaptation*. Here, we intend to generalize this result to inhomogeneous data.

4.2. A minimax adaptive lower bound

First, let us denote by $H(s, L)$ the Hölder ball with smoothness s and radius L, see (3.8). Then, let $L' > L > 0$ and $s > s' > 0$, where s' and s have the same integer part. We introduce $A := H(s', L')$ and $B := H(s, L)$. We denote by a_n and b_n the minimax rates over A and B respectively, given by

$$a_n = n^{-s'/(1+2s'+\beta)}\ell'(1/n), \quad b_n = n^{-s/(1+2s+\beta)}\ell(1/n),$$

where ℓ' and ℓ are slowly varying, and by α_n the adaptive rate over A, given by

$$\alpha_n = (\log n/n)^{-s'/(1+2s'+\beta)}\ell'(\log n/n).$$

Theorem 3. *If an estimator $\widetilde{f}_n$ satisfies the two following upper bounds for some $p > 1$ (that is, it is asymptotically minimax over A and B):*

$$\text{limsup}_n \sup_{f \in A} \mathbb{E}^n_{f,\mu}\big\{\big(a_n^{-1}|\widetilde{f}_n(x_0) - f(x_0)|\big)^p\big\} < +\infty, \tag{4.2}$$

$$\text{limsup}_n \sup_{f \in B} \mathbb{E}^n_{f,\mu}\big\{\big(b_n^{-1}|\widetilde{f}_n(x_0) - f(x_0)|\big)^p\big\} < +\infty, \tag{4.3}$$

then:

$$\text{liminf}_n \sup_{f \in A} \mathbb{E}^n_{f,\mu}\big\{\big(\alpha_n^{-1}|\widetilde{f}_n(x_0) - f(x_0)|\big)^p\big\} > 0. \tag{4.4}$$

Note that (4.4) contradicts (4.2) since $\lim_n a_n/\alpha_n = 0$. The consequence is that *there is no pointwise minimax adaptive estimator over two such classes A and B and that the best achievable rate is α_n.*

5. Proofs

5.1. Preparatory results and proof of Theorems 1 and 2

For the sake of simplicity, we denote by C a positive constant that can vary from place to place, and which can depend on the parameters K, D, p, σ and Q. We remove also some subscripts from the notations: we write $\bar{I}$ instead of $\bar{I}_{n,f}$ (see (3.4)), $\widehat{I}$ instead of $\widehat{I}_n$ and I^* instead of $I^*_{n,f}$. We denote by $|E|$ the cardinality of a set E and we introduce

$$G_n(I) := \{J \in G_n \text{ such that } J \subset I\}.$$

We denote by $\mathfrak{X}_n$ the sigma-field generated by $X_1, \ldots, X_n$. We recall that $\mathbf{X}_I$ is defined by (2.4) and that $\Omega_I = \{\lambda(\mathbf{X}_I) > (n\bar{\mu}_n(I))^{-1/2}\}$ where $\lambda(\mathbf{X}_I)$ is the smallest eigenvalue of $\mathbf{X}_I$. The following proposition is the main tool for proving Theorems 1 and 2.

Proposition 1. *Let $I \in G_n$ be such that*

$$\operatorname{osc} f(I) \leqslant \sigma\Big(\frac{2\log n}{n\bar{\mu}_n(I)}\Big)^{1/2}, \tag{5.1}$$

where we recall that $\operatorname{osc} f(\cdot)$ stands for the local oscillation of f, see (3.2). When $\|f\|_\infty < +\infty$, we have on $\Omega_I \cap \{n\bar{\mu}_n(I) \geqslant 2\}$ for any $p > 0$ and $n \geqslant K + 1$:

$$\begin{aligned}&\mathbb{E}^n_{f,\mu}\{|\widehat{f}_n(x_0) - f(x_0)|^p|\mathfrak{X}_n\}\\ &\quad\leqslant C\Big(\frac{\log n}{n\bar{\mu}_n(I)}\Big)^{p/2}\Big(\lambda(\mathbf{E}_I)^{-p} + (\|f\|_\infty \vee 1)^p|G_n(I)|^{1/2}\frac{(n\bar{\mu}_n(I))^{p-D^2/16}}{(\log n)^{p/2}}\Big),\end{aligned}$$

where $\lambda(\mathbf{E}_I)$ is the smallest eigenvalue of $\mathbf{E}_I$, see (3.5), and D is a constant from the threshold (2.8).

Proof of Proposition 1. Let us introduce

$$\begin{aligned}\mathcal{T}(I,J,m) &:= \big\{|\langle \widehat{f}_J - \widehat{f}_I\,,\,\phi_m\rangle_J| \leqslant \sigma\|\phi_m\|_J T_n(I,J)\big\},\\ \mathcal{T}(I,J) &:= \cap_{0\leqslant m\leqslant K}\mathcal{T}(I,J,m),\\ \mathcal{T}(I) &:= \cap_{J\in G_n(I)}\mathcal{T}(I,J).\end{aligned}$$

By definition (2.7) of $\widehat{I}$, we have that on $\mathcal{T}(I)$, the bandwidth I is selected if it maximises $\bar{\mu}_n(I)$. Thus, if we introduce

$$\mathrm{T}(I) := \big\{\bar{\mu}_n(I) \leqslant \bar{\mu}_n(\widehat{I})\big\},$$

we have $\mathrm{T}(I)^c \subset \mathcal{T}(I)^c$. When $\|f\|_\infty < +\infty$, we have for any $p>0$ and $J\subset[0,1]$:

$$\mathbb{E}^n_{f,\mu}\big\{|\widehat{f}_J(x_0)|^p|\mathfrak{X}_n\big\} \leqslant C(\|f\|_\infty \vee 1)^p(n\bar{\mu}_n(J))^{p/2}. \tag{5.2}$$

This inequality shows that the estimator cannot be too large in expectation, its proof is given below. We need the following lemma, which is of special importance, since it provides a control on the probability for a bandwidth to be selected by the procedure.

Lemma 2. *If $I\in G_n$ satisfies* (5.1)*, we have on $\Omega_I \cap \big\{n\bar{\mu}_n(I)\geqslant 2\big\}$:*

$$\mathbb{P}^n_{f,\mu}\{\mathcal{T}(I)^c|\mathfrak{X}_n\} \leqslant |G_n(I)|(K+1)(n\bar{\mu}_n(I))^{-D^2/8},$$

where D is a constant from the threshold $T_n(I,J)$, see (2.8)*.*

The proof of this lemma is given below. Together with (5.2), Lemma 2 entails

$$\begin{aligned}&\mathbb{E}^n_{f,\mu}\big\{\big(|\widehat{f}_n(x_0)-f(x_0)|\big)^p\mathbf{1}_{\mathrm{T}(I)^c}|\mathfrak{X}_n\big\}\\ &\quad\leqslant C\big[\big(\mathbb{E}^n_{f,\mu}\{|\widehat{f}_n(x_0)|^{2p}|\mathfrak{X}_n\}\big)^{1/2} + \|f\|_\infty^p\big]\big(\mathbb{P}^n_{f,\mu}\{\mathcal{T}(I)^c|\mathfrak{X}_n\}\big)^{1/2}\\ &\quad\leqslant C(\|f\|_\infty\vee 1)^p|G_n(I)|^{1/2}(n\bar{\mu}_n(I))^{p/2-D^2/16}.\end{aligned} \tag{5.3}$$

The next lemma is a version of the bias-variance decomposition of the local polynomial estimator, which is classical: see for instance Fan and Gijbels [8,9], Goldenshluger and Nemirovski [11], Spokoiny [18] and Tsybakov [20], among others. We recall that the matrix $\mathbf{E}_I$ is defined in (3.5).

Lemma 3. *If I is such that $\bar{\mu}_n(I)>0$ and $x_0\in I$, we have on Ω_I that*

$$|\widehat{f}_I(x_0)-f(x_0)| \leqslant C\lambda(\mathbf{E}_I)^{-1}\big(\operatorname{osc} f(I) + \sigma(n\bar{\mu}_n(I))^{-1/2}|\gamma_I|\big), \tag{5.4}$$

where γ_I is, conditionally on $\mathfrak{X}_n$, centered Gaussian with $\mathbb{E}^n_{f,\mu}\{\gamma_I^2|\mathfrak{X}_n\}\leqslant 1$.

The proof of this lemma is given below. For completing the proof of Proposition 1, we need also the following inequality, which is proven below: if $I\in G_n$ and $J\in G_n(I)$, we have on $\mathcal{T}(I,J)\cap\Omega_J$

$$|\widehat{f}_I(x_0)-\widehat{f}_J(x_0)| \leqslant C\lambda(\mathbf{E}_J)^{-1}\big(\log n/(n\bar{\mu}_n(J))\big)^{1/2}. \tag{5.5}$$

By the definition of $\widehat{I}$, we have

$$\mathrm{T}(I) \subset \mathcal{T}(\widehat{I},I),$$

and using (5.5) we obtain that on $\mathrm{T}(I)\cap\Omega_I$,

$$|\widehat{f}_{\widehat{I}}(x_0)-\widehat{f}_I(x_0)| \leqslant C\lambda(\mathbf{E}_I)^{-1}\big(\log n/(n\bar{\mu}_n(I))\big)^{1/2}.$$

In view of lemma 3 and since I satisfies (5.1) we obtain:

$$\begin{aligned}|\widehat{f}_I(x_0) - f(x_0)| &\leqslant C\lambda(\mathbf{E}_I)^{-1}\big(\operatorname{osc} f(I) + \sigma(n\bar{\mu}_n(I))^{-1/2}|\gamma_I|\big)\\ &\leqslant C\lambda(\mathbf{E}_I)^{-1}\big(\sqrt{2} + (\log n)^{-1/2}|\gamma_I|\big)\big(\log n/(n\bar{\mu}_n(I))\big)^{1/2},\end{aligned}$$

and then, on $\mathrm{T}(I) \cap \Omega_I$, we have

$$|\widehat{f}_n(x_0) - f(x_0)| \leqslant C\lambda(\mathbf{E}_I)^{-1}\big(\sqrt{2} + (\log n)^{-1/2}|\gamma_I|\big)\big(\log n/(n\bar{\mu}_n(I))\big)^{1/2}.$$

Finally, using this inequality together with (5.3), Proposition 1 follows by integrating with respect to $\mathbb{P}^n_{f,\mu}(\cdot|\mathfrak{X}_n)$. □

Proof of Theorem 1. Let j be such that $x_0 \in [X_{(j)}, X_{(j+1)}]$, where $X_{(i)} \leqslant X_{(i+1)}$ for any $1 \leqslant i \leqslant n$ (eventually, we take $X_{(0)} := 0$ and $X_{(n+1)} := 1$). First, we consider the procedure tuned with geometrical grid (2.10). Let I^- be the largest interval in G_n such that $I^- \subset I^*$. Since $\operatorname{osc} f(I)^2\bar{\mu}_n(I)$ increases as I increases, we have

$$\operatorname{osc} f(I^-) \leqslant \sigma\big(2\log n/(n\bar{\mu}_n(I^-))\big)^{1/2},$$

thus $\bar{\mu}_n\big(I^-\big) \leqslant \bar{\mu}_n\big(\bar{I}\big)$, where we recall that $\bar{I}$ is given by (3.4). If p and q are such that

$$I^- = [X_{(j+1-[a^p])}, X_{(j+[a^q])}],$$

where $a > 1$ is the grid parameter, see (2.10), and if u, v are such that $[X_{(u)}, X_{(v)}] \subset I^*$ and $\bar{\mu}_n([X_{(u)}, X_{(v)}]) = \bar{\mu}_n(I^*)$, we have

$$\begin{aligned}\bar{\mu}_n\big([X_{(j+1-[a^p])}, X_{(j+[a^q])}]\big) &\leqslant \bar{\mu}_n\big([X_{(u)}, X_{(v)}]\big)\\ &\leqslant \bar{\mu}_n\big([X_{(j+1-[a^{p+1}])}, X_{(j+[a^{q+1}])}]\big),\end{aligned}$$

thus $\bar{\mu}_n(I^*) \leqslant a^2\bar{\mu}_n(I^-) \leqslant a^2\bar{\mu}_n(\bar{I})$, and

$$\bar{\mu}_n(I^*)/a^2 \leqslant \bar{\mu}_n(\bar{I}) \leqslant \bar{\mu}_n(I^*). \tag{5.6}$$

Moreover, again for the grid choice (2.10), we have

$$|G_n(I)| \leqslant \big(\log(n\bar{\mu}_n(I))/\log a\big)^2, \tag{5.7}$$

then, using Proposition 1, and since $D \geqslant 4(p+1)^{1/2}$, we obtain Theorem 1 when the grid is (2.10). When we use the grid (2.11) in the procedure, we have $|G_n(I)| \leqslant (n\bar{\mu}_n(I))^2$, and $\bar{\mu}_n(\bar{I}) = \bar{\mu}_n(I^*)$, thus using again Proposition 1, we conclude the proof of Theorem 1. □

In the following, we denote by $\mathbf{P}_I$ the projection onto the space V_K of all real polynomials of degree at most K, with respect to for the inner product $\langle\cdot\,,\cdot\rangle_I$, see (2.1). As stated in Section 2.1, we have

$$\widehat{f}_I = \bar{f}_I = \mathbf{P}_I Y \tag{5.8}$$

on the event $\Omega_I = \{\lambda(\mathbf{X}_I) > (n\bar{\mu}_n(I))^{-1/2}\}$. We denote respectively by $\langle\cdot\,,\cdot\rangle$ and by $\|\cdot\|$ the Euclidean inner product and the Euclidean norm in $\mathbb{R}^{K+1}$. We denote by $\|\cdot\|_\infty$ the sup norm in $\mathbb{R}^{K+1}$. We define $e_1 := (1, 0, \ldots, 0) \in \mathbb{R}^{K+1}$.

Proof of Lemma 3. On Ω_I, we have $\widehat{f}_I = \bar{f}_I$ since $\bar{\mathbf{X}}_I = \mathbf{X}_I$, and $\lambda(\mathbf{X}_I) > (n\bar{\mu}_n(I))^{-1/2} > 0$, thus $\mathbf{X}_I$ and $\mathbf{E}_I$ are invertible (see (3.5)). By definition of $\operatorname{osc} f(I)$, we can find a polynomial $P_I^\varepsilon \in V_K$ such that

$$\sup_{x\in I} |f(x) - P_I^\varepsilon(x)| \leqslant \operatorname{osc} f(I) + \varepsilon/\sqrt{n},$$

for any fixed $\varepsilon > 0$. If we denote by $\theta_I \in \mathbb{R}^{K+1}$ the coefficients vector of P_I^ε then

$$\begin{aligned}|\widehat{f}_I(x_0) - f(x_0)| &\leqslant |\langle \mathbf{\Lambda}_I^{-1}(\widehat{\theta}_I - \theta_I)\,,\, e_1\rangle| + \operatorname{osc} f(I) + \varepsilon/\sqrt{n}\\ &= |\langle \mathbf{E}_I^{-1}\mathbf{\Lambda}_I\mathbf{X}_I(\widehat{\theta}_I - \theta_I)\,,\, e_1\rangle| + \operatorname{osc} f(I) + \varepsilon/\sqrt{n}.\end{aligned}$$

In view of (2.3), we have on Ω_I for any $m = 0, \ldots, K$:

$$\begin{aligned}(\mathbf{X}_I(\widehat{\theta}_I - \theta_I))_m &= \langle \widehat{f}_I - P_I^\varepsilon\,,\, \phi_m\rangle_I\\ &= \langle Y - P_I^\varepsilon\,,\, \phi_m\rangle_I\\ &= \langle f - P_I^\varepsilon\,,\, \phi_m\rangle_I + \langle \xi\,,\, \phi_m\rangle_I,\end{aligned}$$

thus, $\mathbf{X}_I(\widehat{\theta}_I - \theta_I) = B_I + V_I$ where $B_{I,m} := \langle f - P_I^\varepsilon\,,\, \phi_m\rangle_I$ and $V_{I,m} := \langle \xi\,,\, \phi_m\rangle_I$, which correspond respectively to bias and variance terms. We have

$$|\langle \mathbf{E}_I^{-1}\mathbf{\Lambda}_I B_I\,,\, e_1\rangle| \leqslant (K+1)^{1/2}\|\mathbf{E}_I^{-1}\|\|\mathbf{\Lambda}_I B_I\|_\infty,$$

and

$$|(\mathbf{\Lambda}_I B_I)_m| = \|\phi_m\|^{-1}|\langle f - P_I^\varepsilon\,,\, \phi_m\rangle_I| \leqslant \|f - P_I^\varepsilon\|_I \leqslant \operatorname{osc} f(I) + \varepsilon/\sqrt{n}$$

for $m \in \{0, \ldots, K\}$. Since $\lambda(M)^{-1} = \|M^{-1}\|$ for any symmetrical and positive matrix M, and since $\|\mathbf{\Lambda}_I^{-1}\| \leqslant 1$, we have $\|\mathbf{E}_I^{-1}\| \leqslant \lambda(\mathbf{X}_I)^{-1} \leqslant (n\bar{\mu}_n(I))^{1/2} \leqslant n^{1/2}$, thus

$$|\langle \mathbf{E}_I^{-1}\mathbf{\Lambda}_I B_I\,,\, e_1\rangle| \leqslant (K+1)^{1/2}\big(\|\mathbf{E}_I^{-1}\| \operatorname{osc} f(I) + \varepsilon\big).$$

Conditionally on $\mathfrak{X}_n$, the random vector V_I is centered Gaussian with covariance matrix $\sigma^2(n\bar{\mu}_n(I))^{-1}\mathbf{X}_I$. Thus $\mathbf{E}_I^{-1}\mathbf{\Lambda}_I V_I$ is again centered Gaussian, with covariance matrix

$$\sigma^2(n\bar{\mu}_n(I))^{-1}\mathbf{E}_I^{-1}\mathbf{\Lambda}_I\mathbf{X}_I\mathbf{\Lambda}_I\mathbf{E}_I^{-1} = \sigma^2(n\bar{\mu}_n(I))^{-1}\mathbf{E}_I^{-1},$$

and $\langle \mathbf{E}_I^{-1}\mathbf{\Lambda}_I V_I\,,\, e_1\rangle$ is then centered Gaussian with variance

$$\sigma^2(n\bar{\mu}_n(I))^{-1}\langle e_1\,,\, \mathbf{E}_I^{-1}e_1\rangle \leqslant \sigma^2(n\bar{\mu}_n(I))^{-1}\|\mathbf{E}_I^{-1}\|.$$

Now, since $\lambda(\mathbf{E}_I) = \inf_{\|x\|=1}\langle x\,,\, \mathbf{E}_I x\rangle \leqslant \|\mathbf{E}_I e_1\| \leqslant (K+1)^{1/2}$, we have $\|\mathbf{E}_I^{-1}\| \leqslant (K+1)^{1/2}\|\mathbf{E}_I^{-1}\|^2$, and the lemma follows. □

Proof of (5.2). If $\bar{\mu}_n(J) = 0$, we have $\widehat{f}_J = 0$ by definition and the result is obvious, thus we assume $\bar{\mu}_n(J) > 0$. Since $\lambda(\bar{\mathbf{X}}_J) \geqslant (n\bar{\mu}_n(J))^{-1/2} > 0$, $\bar{\mathbf{X}}_J$ and $\mathbf{\Lambda}_J$ are invertible and $\mathbf{E}_J$ also is. Thus,

$$\widehat{f}_J(x_0) = \langle \mathbf{\Lambda}_J^{-1}\widehat{\theta}_J\,,\, e_1\rangle = \langle \mathbf{E}_J^{-1}\mathbf{\Lambda}_J\bar{\mathbf{X}}_J\widehat{\theta}_J\,,\, e_1\rangle = \langle \mathbf{E}_J^{-1}\mathbf{\Lambda}_J\mathbf{Y}_J\,,\, e_1\rangle.$$

For any $0 \leqslant m \leqslant K$, we have

$$\begin{aligned}|(\mathbf{\Lambda}_J\mathbf{Y}_J)_m| &\leqslant \|\phi_m\|_J^{-1}\big(|\langle f\,,\, \phi_m\rangle_J| + |\langle \xi\,,\, \phi_m\rangle_J|\big)\\ &\leqslant \|f\|_\infty + \|\phi_m\|_J^{-1}|\langle \xi\,,\, \phi_m\rangle_J|\\ &=: \|f\|_\infty + |V_{J,m}|.\end{aligned}$$

Conditionally on $\mathfrak{X}_n$, the vector V_J with entries $(V_{J,m}; 0 \leqslant m \leqslant K)$ is centered Gaussian with variance $\sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\mathbf{\Lambda}_J\mathbf{X}_J\mathbf{\Lambda}_J$, thus $\langle \mathbf{E}_J^{-1}V_J\,,\,e_1\rangle$ is centered Gaussian with variance

$$\begin{aligned}\sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\langle e_1\,,\,\mathbf{\Lambda}_J^{-1}\bar{\mathbf{X}}_J^{-1}\mathbf{X}_J\bar{\mathbf{X}}_J^{-1}\mathbf{\Lambda}_J^{-1}e_1\rangle &\leqslant \sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\|\mathbf{\Lambda}_J^{-1}\|^2\|\bar{\mathbf{X}}_J^{-1}\|^2\|\mathbf{X}_J\| \\ &\leqslant \sigma^2(K+1),\end{aligned}$$

since $\|\mathbf{X}_J\| \leqslant K+1$, $\|\mathbf{\Lambda}_J^{-1}\| \leqslant 1$ and $\|\bar{\mathbf{X}}_J^{-1}\| = \lambda(\bar{\mathbf{X}}_J)^{-1} \leqslant \big(n\bar{\mu}_n(J)\big)^{1/2}$. Moreover, $\|\mathbf{E}_J^{-1}\| \leqslant \|\mathbf{\Lambda}_J^{-1}\|\|\bar{\mathbf{X}}_J^{-1}\|\|\mathbf{\Lambda}_J^{-1}\| \leqslant \big(n\bar{\mu}_n(J)\big)^{1/2}$, thus

$$|\widehat{f}_J(x_0)| \leqslant (K+1)^{1/2}(\|f\|_\infty \vee 1)\big(n\bar{\mu}_n(J)\big)^{1/2}\big(1+\sigma|\gamma_J|\big),$$

where γ_J is, conditionally on $\mathfrak{X}_n$, centered Gaussian with variance smaller than 1. Then, (5.2) follows by integrating with respect to $\mathbb{P}^n_{f,\mu}(\cdot|\mathfrak{X}_n)$. □

Proof of Lemma 2. Let $0 \leqslant m \leqslant K$ and $J \in G_n(I)$. In view of (2.3) and (5.8), we have on Ω_I:

$$\begin{aligned}\langle \widehat{f}_J - \widehat{f}_I\,,\,\phi_m\rangle_J &= \langle Y - \widehat{f}_I\,,\,\phi_m\rangle_J \\ &= \langle f - \widehat{f}_I\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J \\ &= \langle f - \mathbf{P}_I f\,,\,\phi_m\rangle_J + \langle \mathbf{P}_I f - \widehat{f}_I\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J \\ &= \langle f - \mathbf{P}_I f\,,\,\phi_m\rangle_J + \langle \mathbf{P}_I(f - Y)\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J \\ &= \langle f - \mathbf{P}_I f\,,\,\phi_m\rangle_J - \langle \mathbf{P}_I \xi\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J \\ &:= A + B + C.\end{aligned}$$

By the definition of $\operatorname{osc} f(I)$ we can find a polynomial $P_I^\varepsilon \in V_K$ such that

$$\sup_{x\in I}|f(x) - P_I^\varepsilon(x)| \leqslant \operatorname{osc} f(I) + \varepsilon_n,$$

where $\varepsilon_n := \sigma D C_K(\log 2)/(4n)$, with $C_K = 1+(K+1)^{1/2}$. Thus, since $J \subset I$, $P_I^\varepsilon \in V_K$ and $\mathbf{P}_I$ is an orthogonal projection with respect to $\langle\cdot\,,\,\cdot\rangle_I$,

$$\begin{aligned}|A| \leqslant \|f - \mathbf{P}_I f\|_J\|\phi_m\|_J &\leqslant \|\phi_m\|_J\|f - P_I^\varepsilon - \mathbf{P}_I(f - P_I^\varepsilon)\|_I \\ &\leqslant \|\phi_m\|_J\|f - P_I^\varepsilon\|_I \\ &\leqslant \|\phi_m\|_J(\operatorname{osc} f(I) + \varepsilon_n) \\ &\leqslant \|\phi_m\|_J\big[\sigma(2\log n/(n\bar{\mu}_n(I))^{1/2} + \varepsilon_n\big],\end{aligned} \tag{5.9}$$

where we used (5.1). Conditionally on $\mathfrak{X}_n$, B and C are centered Gaussian. Clearly, C is centered Gaussian with variance $\sigma^2\|\phi_m\|_J^2/(n\bar{\mu}_n(I))$. Since $\mathbf{P}_I\xi$ has covariance matrix $\sigma^2\mathbf{P}_I\mathbf{P}_I' = \sigma^2\mathbf{P}_I$ ($\mathbf{P}_I$ is an orthogonal projection), the variance of B is equal to

$$\begin{aligned}\mathbb{E}^n_{f,\mu}\big\{\langle \mathbf{P}_I\xi\,,\,\phi_m\rangle_J^2|\mathfrak{X}_n\big\} &\leqslant \|\phi_m\|_J^2\mathbb{E}^n_{f,\mu}\{\|\mathbf{P}_I\xi\|_J^2|\mathfrak{X}_n\} \\ &= \|\phi_m\|_J^2\operatorname{Tr}\big(\mathbb{V}\mathrm{ar}(\mathbf{P}_I\xi|\mathfrak{X}_n)\big)/(n\bar{\mu}_n(J)) \\ &= \sigma^2\|\phi_m\|_J^2\operatorname{Tr}(\mathbf{P}_I)/(n\bar{\mu}_n(J)),\end{aligned}$$

where $\operatorname{Tr}(M)$ stands for the trace of a matrix M. Since $\mathbf{P}_I$ is the projection onto V_K, it follows that $\operatorname{Tr}(\mathbf{P}_I) \leqslant K+1$, and that the variance of B is smaller than

$$\sigma^2\|\phi_m\|_J^2(K+1)/(n\bar{\mu}_n(J)).$$

Then,

$$\mathbb{E}^n_{f,\mu}\{(B+C)^2|\mathfrak{X}_n\} \leqslant \sigma^2\|\phi_m\|_J^2 C_K^2/(n\bar{\mu}_n(J)). \tag{5.10}$$

Since $n\bar{\mu}_n(I) \geqslant 2$ and $\bar{\mu}_n(J) \leqslant 1$, we have

$$\varepsilon_n \leqslant \sigma D C_K \big[\log(n\bar{\mu}_n(I))/(4n\bar{\mu}_n(J))\big]^{1/2}. \tag{5.11}$$

Then, the definition of the threshold (2.8) together with (5.9) and (5.11) entail

$$\begin{aligned}\big\{\|\phi_m\|_J^{-1}|\langle \widehat{f}_I - \widehat{f}_J\,,\, \phi_m\rangle_J| > T_n(I,J)\big\}& \\ \subset \Big\{\frac{\|\phi_m\|_J^{-1}|B+C|}{\sigma(n\bar{\mu}_n(J))^{-1/2}C_K} &> D\big[\log(n\bar{\mu}_n(I))\big]^{1/2}/2\Big\}.\end{aligned}$$

Since

$$\mathcal{T}(I,J)^c = \bigcup_{m=0}^{K}\big\{\|\phi_m\|_J^{-1}|\langle \widehat{f}_I - \widehat{f}_J\,,\, \phi_m\rangle_J| > T_n(I,J)\big\},$$

we obtain using (5.10) and the fact that $\mathbb{P}\{|N(0,1)| > x\} \leqslant \exp(-x^2/2)$:

$$\begin{aligned}\mathbb{P}^n_{f,\mu}\{\mathcal{T}(I)^c|\mathfrak{X}_n\} &\leqslant \sum_{J\in G_n(I)}\sum_{m=0}^{K}\exp\big(-D^2\log(n\bar{\mu}_n(I))/8\big) \\ &\leqslant |G_n(I)|(K+1)(n\bar{\mu}_n(I))^{-D^2/8},\end{aligned}$$

which concludes the lemma. □

Proof of (5.5). Let us define $\mathbf{H}_J := \mathbf{\Lambda}_J\mathbf{X}_J$. On Ω_J, we have:

$$\begin{aligned}|\widehat{f}_I(x_0) - \widehat{f}_J(x_0)| &= |(\widehat{\theta}_I - \widehat{\theta}_J)_0| \\ &\leqslant \|\mathbf{\Lambda}_J^{-1}(\widehat{\theta}_I - \widehat{\theta}_J)\|_\infty \\ &\leqslant \|\mathbf{E}_J^{-1}\mathbf{H}_J(\widehat{\theta}_I - \widehat{\theta}_J)\|_\infty \\ &\leqslant (K+1)^{1/2}\lambda(\mathbf{E}_J)^{-1}\|\mathbf{H}_J(\widehat{\theta}_I - \widehat{\theta}_J)\|_\infty.\end{aligned}$$

Since on Ω_J, $\langle \widehat{f}_I - \widehat{f}_J\,,\, \phi_m\rangle_J/\|\phi_m\|_J = (\mathbf{H}_J(\widehat{\theta}_I - \widehat{\theta}_J))_m$, and since $J \subset I$, we obtain on $\mathcal{T}(I,J)$:

$$\begin{aligned}|\widehat{f}_I(x_0) - \widehat{f}_J(x_0)| &\leqslant C\lambda(\mathbf{E}_J)^{-1}T_n(I,J) \\ &\leqslant C\lambda(\mathbf{E}_J)^{-1}\big[\log n/(n\bar{\mu}_n(J))\big]^{1/2},\end{aligned}$$

thus (5.5). □

Let us denote by $\mathbb{P}^n_\mu$ the joint probability of the variables $[X_i; 1 \leqslant i \leqslant n]$ and let us recall the notation $\mu(I) = \int_I \mu(t)\mathrm{d}t$. We recall that $I_h = [x_0 - h, x_0 + h]$ and that $h_n(\omega,\mu)$ is defined by (3.9). We introduce

$$H_n(\omega,\mu) := \operatorname*{argmin}_{h\in[0,1]}\Big\{\omega(h) \geqslant \sigma\Big(\frac{\log n}{n\bar{\mu}_n(I_h)}\Big)^{1/2}\Big\}, \tag{5.12}$$

which is an approximation of $h_n(\omega,\mu)$ when μ is unknown. In what follows, we omit the dependence upon ω and μ to avoid overloaded notations. If $0 < \varepsilon < 1$, we introduce the event

$$\mathrm{C}_n(\varepsilon) := \{(1-\varepsilon)h_n < H_n \leqslant (1+\varepsilon)h_n\},$$

which has probability going to 1 very fast, see Lemma 4 below.

Proof of Theorem 2. For the grid choice (3.12), we have $|G_n(I)| \leqslant n\bar{\mu}_n(I)$. We recall that $f \in \mathcal{F}_{\delta_n}(\omega, Q)$, see Definition 2 and that by assumption, $\delta_n > \rho h_n$ for some fixed $\rho > 1$. We consider $\varepsilon \in (0, \rho - 1]$, so that on $\mathrm{C}_n(\varepsilon)$, we have $\delta_n \geqslant H_n$ and since $f \in \mathcal{F}_{\delta_n}(\omega, Q)$, we have either

$$\begin{aligned} \operatorname{osc} f(I_{H_n}) \leqslant \omega(H_n) = \sigma\Big(\frac{\log n}{n\bar{\mu}_n(I_{H_n})}\Big)^{1/2} \quad \text{or} \\ \leqslant \sigma\Big(\frac{\log n}{n\bar{\mu}_n(I_{H_n}) - 1}\Big)^{1/2}, \end{aligned}$$

which entails that in both cases

$$\operatorname{osc} f(I_{H_n}) \leqslant \sigma\Big(\frac{2\log n}{n\bar{\mu}_n(I_{H_n})}\Big)^{1/2}.$$

Then, using Proposition 1 and since $D \geqslant 4(p+1/2)^{1/2}$, we have for any $f \in \mathcal{F}_{\delta_n}(\omega, Q)$:

$$\mathbb{E}^n_{f,\mu}\big\{|\widehat{f}_n(x_0) - f(x_0)|^p|\mathfrak{X}_n\big\} \leqslant C\Big(\frac{\log n}{n\bar{\mu}_n(I_{H_n})}\Big)^{p/2}\big(\lambda(\mathbf{E}_{I_{H_n}})^{-p} + (Q \vee 1)^p\big) \tag{5.13}$$

on $\mathrm{C}_n(\varepsilon) \cap \Omega_{I_{H_n}} \cap \{n\bar{\mu}_n(I_{H_n}) \geqslant 2\}$. Let us introduce

$$e_{a,b} = \frac{(1+(-1)^a)(b+1)}{a+b+1},$$

and the matrix $\mathbf{E} := \mathbf{\Lambda X \Lambda}$ where $\mathbf{X}$ is the symmetrical matrix with entries $(\mathbf{X})_{p,q} := e_{p+q,\beta}$ for $0 \leqslant p, q \leqslant K$ and

$$\mathbf{\Lambda} := \operatorname{diag}\big[e_{0,\beta}^{-1/2}, e_{2,\beta}^{-1/2}, \ldots, e_{2K,\beta}^{-1/2}\big],$$

where we recall that β is the index of regular variation of μ, see (3.7). $\mathbf{E}$ is the limit in probability of $\mathbf{E}_{H_n}$ as $n \to +\infty$. $\mathbf{\Lambda}$ and $\mathbf{X}$ are invertible thus $\mathbf{E}$ also is. Indeed, we have $\lambda(\mathbf{X}) > 0$, otherwise, defining $\mathbf{p}(t) := (1, t, \ldots, t^K)$, we have

$$0 = \lambda(\mathbf{X}) = \langle \mathbf{x}\,,\, \mathbf{X}\mathbf{x}\rangle = \int_{-1}^{1} (\mathbf{x}'\mathbf{p}(t))^2 |t|^\beta \mathrm{d}t,$$

where $\mathbf{x} \in \mathbb{R}^{K+1}$ is non-zero vector, which leads to a contradiction, since $t \mapsto \mathbf{x}'\mathbf{p}(t)$ is a polynomial ($\mathbf{x}'$ stands for the transposition of $\mathbf{x}$). The following lemma provides some approximations necessary for the proof of Theorem 2, its proof is given below.

Lemma 4. *If $\omega \in \mathrm{RV}(s)$, $s > 0$ and μ satisfies (3.7), we can find an event $\mathrm{S}_n(\varepsilon) \in \mathfrak{X}_n$ such that for any $\varepsilon \in (0, 1/2]$,*

$$\mathrm{S}_n(\varepsilon) \subset \big\{|\lambda(\mathbf{E}_{H_n}) - \lambda(\mathbf{E})| \leqslant \varepsilon\big\} \cap \big\{|\bar{\mu}_n(I_{H_n})/\mu(I_{h_n}) - 1| \leqslant \varepsilon\big\} \cap \mathrm{C}_n(\varepsilon) \tag{5.14}$$

for n large enough, and

$$\mathbb{P}^n_\mu\big\{\mathrm{S}_n(\varepsilon)^c\big\} \leqslant C\exp\big(-D_{\mathrm{S}}\, r_n^{-2}\big), \tag{5.15}$$

where $r_n = r_n(\omega, \mu)$ is given by (3.10) and C, D_{S} are positive constants.

On $\mathrm{S}_n(\varepsilon)$, we have $n\bar{\mu}_n(I_{H_n}) \geqslant (1-\varepsilon)n\mu(I_{h_n}) \to +\infty$ as $n \to +\infty$, thus $\mathrm{S}_n(\varepsilon) \subset \Omega_{I_{H_n}} \cap \{n\bar{\mu}_n(I_{H_n}) \geqslant 2\}$ for n large enough. Then, using together (5.13), (5.14) and (3.9), (3.10), and integrating with respect to $\mathbb{P}^n_\mu$, we have uniformly for $f \in \mathcal{F}_{\delta_n}(\omega, Q)$:

$$\mathbb{E}^n_{f,\mu}\big\{|\widehat{f}_n(x_0) - f(x_0)|^p \mathbf{1}_{\mathrm{S}_n(\varepsilon)}\big\} \leqslant C r_n^p.$$

On the complement $\mathrm{S}_n(\varepsilon)^c$, using together (5.2) and (5.15), , we obtain that uniformly for $f \in \mathcal{F}_{\delta_n}(\omega, Q)$:

$$\begin{aligned}\mathbb{E}^n_{f,\mu}\big\{\big(r_n^{-1}|\widehat{f}_n(x_0) - f(x_0)|\big)^p \mathbf{1}_{\mathrm{S}_n(\varepsilon)^c}\big\} \\ &\leqslant C r_n^{-p}\big[\big(\mathbb{E}^n_{f,\mu}\{|\widehat{f}_n(x_0)|^{2p}\}\big)^{1/2} + Q^p\big]\big(\mathbb{P}^n_\mu\{\mathrm{S}_n(\varepsilon)^c\}\big)^{1/2} \\ &\leqslant C r_n^{-p} n^{p/2}\big(\mathbb{P}^n_\mu\{\mathrm{S}_n(\varepsilon)^c\}\big)^{1/2} = o_n(1),\end{aligned}$$

which concludes the proof of Theorem 2. □

Proof of Lemma 4. The proof of the lemma is divided into several steps. We denote $\mathrm{A}_n(\varepsilon) := \big\{|\lambda(\mathbf{E}_{H_n}) - \lambda(\mathbf{E})| \leqslant \varepsilon\big\}$ and for $a \in \mathbb{N}$ we introduce

$$\mathrm{B}_{n,a}(\varepsilon) := \Big\{\Big|\frac{1}{\mu(I_{h_n})}\int_{I_{H_n}}\Big(\frac{\cdot - x_0}{h_n}\Big)^a \mathrm{d}\bar{\mu}_n - e_{a,\beta}\Big| \leqslant \varepsilon\Big\}.$$

Step 1. Since $\mathbf{E}_{H_n}$ and $\mathbf{E}$ are symmetrical,

$$\bigcap_{0\leqslant p,q\leqslant K}\big\{\big|(\mathbf{E}_{H_n} - \mathbf{E})_{p,q}\big| \leqslant \varepsilon/(K+1)^2\big\} \subset \mathrm{A}_n(\varepsilon),$$

where we used the fact that $\lambda(M) = \inf_{\|x\|=1}\langle x\,,\, Mx\rangle$ for any symmetrical matrix M. Then, if $\varepsilon_1 := \min\big[\varepsilon\ ;\ \varepsilon(\beta+1)/((K+1)^2(2K+\beta+1))\big]$, we have

$$\mathrm{B}_{n,p+q}(\varepsilon_1) \cap \mathrm{B}_{n,2p}(\varepsilon_1) \cap \mathrm{B}_{n,2q}(\varepsilon_1) \subset \big\{\big|(\mathbf{E}_{H_n} - \mathbf{E})_{p,q}\big| \leqslant \varepsilon/(K+1)^2\big\}$$

for any $0 \leqslant p, q \leqslant K$ and then

$$\bigcap_{\alpha=0}^{2K} \mathrm{B}_{n,\alpha}(\varepsilon_1) \subset \mathrm{A}_n(\varepsilon).$$

Step 2. For $a \in \mathbb{N}$, $h > 0$ and $\varepsilon > 0$, we define

$$\mathrm{D}_{n,a}(\varepsilon, h) := \Big\{\Big|\frac{1}{\mu(I_h)}\int_{I_h}\Big(\frac{\cdot - x_0}{h}\Big)^a \mathrm{d}\bar{\mu}_n - e_{a,\beta}\Big| \leqslant \varepsilon\Big\}.$$

We show that, for any $\omega \in \mathrm{RV}(s)$, $s > 0$ and $0 < \varepsilon_2 \leqslant 1/2$ there exists $0 < \varepsilon_3 \leqslant \varepsilon_2$ such that

$$\mathrm{D}_{n,0}(\varepsilon_3, (1-\varepsilon_2)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_3, (1+\varepsilon_2)h_n) \subset \mathrm{C}_n(\varepsilon_2) \tag{5.16}$$

for n large enough. In view of (5.12), we have

$$\big\{H_n \leqslant (1+\varepsilon_2)h_n\big\} = \big\{n\bar{\mu}_n(I_{(1+\varepsilon_2)h_n})/\log n \geqslant \sigma^2\omega((1+\varepsilon_2)h_n)^{-2}\big\}.$$

We introduce $\varepsilon_3 := \min\big[\varepsilon_2\ ;\ 1 - (1-\varepsilon_2^2)^{-2}(1+\varepsilon_2)^{-2s}\big]$, which is positive for ε_2 small enough, and $\ell_\omega(h) := h^{-s}\omega(h)$ which is slowly varying, since $\omega \in \mathrm{RV}(s)$. Since (A.1) holds uniformly over each compact set in $(0, +\infty)$, we have for any $y \in [1/2, 3/2]$

$$(1-\varepsilon_2^2)\ell_\omega(h_n) \leqslant \ell_\omega(yh_n) \leqslant (1+\varepsilon_2^2)\ell_\omega(h_n) \tag{5.17}$$

for n large enough, so (5.17) with $y = 1 + \varepsilon$ ($\varepsilon \leqslant 1/2$) entails together with (3.9) and since $h \mapsto \mu(I_h)$ is increasing:

$$\begin{aligned}(1-\varepsilon_3)n\mu(I_{(1+\varepsilon_2)h_n})/\log n &\geqslant (1-\varepsilon_2^2)^{-2}(1+\varepsilon_2)^{-2s}\sigma^2\omega(h_n)^{-2}\\ &= \sigma^2\big((1+\varepsilon_2)h_n\big)^{-2s}(1-\varepsilon_2^2)^{-2}\ell_\omega(h_n)^{-2}\\ &\geqslant \sigma^2\omega((1+\varepsilon_2)h_n)^{-2}.\end{aligned}$$

Thus

$$\big\{\bar\mu_n(I_{(1+\varepsilon_2)h_n}) \geqslant (1-\varepsilon_3)\mu((1+\varepsilon_2)h_n)\big\} \subset \big\{H_n \leqslant (1+\varepsilon_2)h_n\big\},$$

and similarly on the other side, we have for n large enough:

$$\big\{\bar\mu_n(I_{(1-\varepsilon_2)h_n}) \leqslant (1+\varepsilon_3)\mu((1-\varepsilon_2)h_n)\big\} \subset \{(1-\varepsilon_2)h_n < H_n\},$$

hence (5.16).

Step 3. We prove (5.14). Let us define $\varepsilon_2 := \varepsilon_1/(2(1+\varepsilon_1)^{2K+1})$ and let ε_3 be such that $(1+\varepsilon_3)^{\beta+3}/(1-\varepsilon_3) \leqslant 1+\varepsilon_2$ and $0 < \varepsilon_3 \leqslant \varepsilon_2$. Since $h \mapsto \bar\mu_n(I_h)$ is increasing, we have

$$\mathrm{C}_n(\varepsilon_3) \subset \big\{\bar\mu_n(I_{(1-\varepsilon_3)h_n}) \leqslant \bar\mu_n(I_{H_n}) \leqslant \bar\mu_n(I_{(1+\varepsilon_3)h_n})\big\},$$

and using (5.16), we can find $0 < \varepsilon_4 \leqslant \varepsilon_3$ such that

$$\mathrm{D}_{n,0}(\varepsilon_4,(1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4,(1+\varepsilon_3)h_n) \subset \mathrm{C}_n(\varepsilon_3).$$

In view of (A.1) and since $\ell_\mu(h) := h^{-(\beta+1)}\mu(I_h)$ is slowly varying, we have for any $0 < \varepsilon_3 \leqslant 1/2$:

$$\ell_\mu((1+\varepsilon_3)h_n) \leqslant (1+\varepsilon_3)\ell_\mu(h_n) \text{ and } \ell_\mu((1-\varepsilon_3)h_n) \geqslant (1-\varepsilon_3)\ell_\mu(h_n) \tag{5.18}$$

as n is large enough, thus the previous embeddings entail

$$\mathrm{D}_{n,0}(\varepsilon_4,(1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4,(1+\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_3,h_n) \subset \Big\{\Big|\frac{\bar\mu_n(I_{H_n})}{\bar\mu_n(I_{h_n})} - 1\Big| \leqslant \varepsilon_2\Big\}.$$

In view of the previous embeddings, we have on $\mathrm{D}_{n,0}(\varepsilon_4,(1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4,(1+\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_3,h_n)$:

$$\begin{aligned}\frac{1}{\mu(I_{h_n})}\Big|\int_{I_{H_n}}\Big(\frac{\cdot - x_0}{h_n}\Big)^\alpha \mathrm{d}\bar\mu_n - \int_{I_{h_n}}\Big(\frac{\cdot - x_0}{h_n}\Big)^\alpha \mathrm{d}\bar\mu_n\Big|&\\ \leqslant \Big(\frac{H_n \vee h_n}{h_n}\Big)^\alpha \frac{\bar\mu_n(I_{h_n})}{\mu(I_{h_n})}\Big|\frac{\bar\mu_n(I_{H_n})}{\bar\mu_n(I_{h_n})} - 1\Big|&\\ \leqslant (1+\varepsilon_3)^\alpha(1+\varepsilon_3)\varepsilon_2 \leqslant (1+\varepsilon_1)^{2K+1}\varepsilon_2 = \varepsilon_1/2.&\end{aligned}$$

Then, putting all the previous embeddings together, we obtain

$$\begin{aligned}&\mathrm{D}_{n,0}(\varepsilon_4,(1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4,(1+\varepsilon_3)h_n)\\ &\qquad \cap \mathrm{D}_{n,0}(\varepsilon_4,h_n) \cap \mathrm{D}_{n,\alpha}(\varepsilon_1/2,h_n) \subset \mathrm{B}_{n,\alpha}(\varepsilon_1),\end{aligned}$$

and finally, (5.14) follows if we choose

$$\begin{aligned}\mathrm{S}_n(\varepsilon) := \mathrm{D}_{n,0}(\varepsilon_4,(1-\varepsilon_3)h_n) &\cap \mathrm{D}_{n,0}(\varepsilon_4,(1+\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4,h_n)\\ &\cap \bigcap_{0\leqslant\alpha\leqslant 2K} \mathrm{D}_{n,\alpha}(\varepsilon_4,h_n).\end{aligned}$$

Step 4. We prove (5.15). We show that if μ satisfies (3.7), we have for any positive sequence (γ_n) going to 0 and any $\alpha \in \mathbb{N}, \varepsilon > 0$:

$$\mathbb{P}_\mu^n\big\{\mathrm{D}_{n,\alpha}(\varepsilon, \gamma_n)^c\big\} \leqslant 2\exp\Big(-\frac{\varepsilon^2}{8(1+\varepsilon/3)} n\mu(I_{\gamma_n})\Big), \tag{5.19}$$

when n is large enough. We define $Q_i := \big(\frac{X_i - x_0}{\gamma_n}\big)^\alpha \mathbf{1}_{X_i \in I_{\gamma_n}}$ and $Z_i := Q_i - \mathbb{E}_\mu^n\{Q_i\}$. In view of (3.7), we can find N such that $n \geqslant N$ entails $\gamma_n \leqslant \nu$ and

$$\frac{1}{\mu(I_{\gamma_n})}\mathbb{E}_\mu^n\{Q_i\} = \frac{1+(-1)^\alpha}{2} \frac{\gamma_n^{\beta+1}\ell_\mu(\gamma_n)}{\int_0^{\gamma_n} t^\beta \ell_\mu(t)\mathrm{d}t} \frac{\int_0^{\gamma_n} t^{\alpha+\beta}\ell_\mu(t)\mathrm{d}t}{\gamma_n^{\alpha+\beta+1}\ell_\mu(\gamma_n)},$$

where for $h \leqslant \nu$, $\ell_\mu(h) := h^{-\beta}\mu(x_0+h) = h^{-\beta}\mu(x_0-h)$ is slowly varying. Then, in view of (A.4):

$$\lim_{n\to+\infty} \frac{1}{\mu(I_{\gamma_n})}\mathbb{E}_\mu^n\{Q_i\} = e_{\alpha,\beta},$$

which entails that for n large enough:

$$\mathrm{D}_{n,\alpha}(\varepsilon,\gamma_n)^c \subset \Big\{\frac{1}{n\mu(I_{\gamma_n})}\Big|\sum_{i=1}^n Z_i\Big| > \varepsilon/2\Big\}. \tag{5.20}$$

Note that $\mathbb{E}_\mu^n\{Z_i\} = 0$, $|Z_i| \leqslant 2$, $\sum_{i=1}^n \mathbb{E}_\mu^n\{Z_i^2\} \leqslant n\mathbb{E}_\mu^n\{Q_i^2\} \leqslant n\mu(I_{\gamma_n})$ and that the $[Z_i\ ;\ 1 \leqslant i \leqslant n]$ are independent. Thus, using Bernstein inequality to the sum of the Z_i we obtain (5.19).

Now, using together (3.9), (3.10) and (5.19), we obtain (5.15), which concludes the proof of Lemma 4. □

5.2. Preparatory results and proof of Theorem 3

The proof of Theorem 3 is similar to the proof of Theorem 3 in Brown and Low [5]. It is based on the next theorem which can be found in Cai *et al.* [6]. This result is a general constrained risk inequality which is useful for several statistical problems, for instance superefficiency, adaptation and so on.

Let $p > 1$ and q be such that $1/p + 1/q = 1$ and X be a real random variable having distribution $\mathbb{P}_\theta$ with density f_θ. The parameter θ can take two values θ_1 or θ_2. We want to estimate θ based on X. The risk of an estimator δ based on X is given by

$$R_p(\delta,\theta) := \mathbb{E}_\theta\{|\delta(X) - \theta|^p\}.$$

We define $s(x) := f_{\theta_2}(x)/f_{\theta_1}(x)$ and $\Delta := |\theta_2 - \theta_1|$. Let

$$I_q = I_q(\theta_1,\theta_2) := \big(\mathbb{E}_{\theta_1}\{s^q(X)\}\big)^{1/q}.$$

Theorem 4 (Cai, Low and Zhao [6])**.** *If δ is such that $R_p(\delta,\theta_1) \leqslant \varepsilon^p$ and if $\Delta > \varepsilon I_q$, we have*:

$$R_p(\delta,\theta_2) \geqslant (\Delta - \varepsilon I_q)^p \geqslant \Delta^p\Big(1 - \frac{p\varepsilon I_q}{\Delta}\Big).$$

The next proposition is a generalization of a result by Brown and Low [5] for the random design model, when the data is inhomogeneous. Of course, in the classical case with μ continuous at x_0 and such that $\mu(x_0) > 0$, the result is barely the same as in Brown and Low [5] with the same rates. This proposition is a lower bound for a superefficient estimator which implies directly the adaptive lower bound stated in Theorem 3. Let us recall that a_n is the minimax rate over A and that α_n is the minimax adaptive rate over A, see Section 4.2.

Proposition 2. *If an estimator $\widetilde{f}_n$ based on* (1.1) *is asymptotically minimax over A, that is*:

$$\mathrm{limsup}_n \sup_{f\in A} \mathbb{E}_{f,\mu}^n\big\{\big(a_n^{-1}|\widetilde{f}_n(x_0) - f(x_0)|\big)^p\big\} < +\infty,$$

and if this estimator is superefficient at a function $f_0 \in A$, in the sense that for some $\gamma > 0$:

$$\operatorname{limsup}_n \mathbb{E}^n_{f_0,\mu}\big\{\big(a_n^{-1} n^{\gamma} |\widetilde{f}_n(x_0) - f_0(x_0)|\big)^p\big\} < +\infty, \tag{5.21}$$

then we can find another function $f_1 \in A$ such that

$$\operatorname{liminf}_n \mathbb{E}^n_{f_1,\mu}\big\{\big(\alpha_n^{-1} |\widetilde{f}_n(x_0) - f_1(x_0)|\big)^p\big\} > 0.$$

Proof of Proposition 2. Since $\operatorname{limsup}_n \mathbb{E}^n_{f_0,\mu}\big\{\big(a_n^{-1} n^{\gamma} |\widetilde{f}_n(x_0) - f_0(x_0)|\big)^p\big\} = C < +\infty$, there is N such that for any $n \geqslant N$:

$$\mathbb{E}^n_{f_0,\mu}\big\{\big(|\widetilde{f}_n(x_0) - f_0(x_0)|\big)^p\big\} \leqslant 2C a_n^p n^{-\gamma p}.$$

Let $k' = \lfloor s' \rfloor$ be the largest integer smaller than s'. Let g be k' times differentiable with support included in $[-1, 1]$, and such that $g(0) > 0$ and for any $|x| \leqslant \delta$, $|g^{(k')}(x) - g^{(k')}(0)| \leqslant k'!|x|^{s'-k'}$. Such a function clearly exists. We define

$$f_1(x) := f_0(x) + L' \rho_n^{s'} g\Big(\frac{x - x_0}{\rho_n}\Big),$$

where ρ_n is the smallest solution to

$$L' h^{s'} = \sigma\Big(\frac{b \log n}{n\mu(I_h)}\Big)^{1/2},$$

where $b := 2(p-1)\gamma/g_\infty^2$, $g_\infty := \sup_x |g(x)|$. We clearly have $f_1 \in A$. Let $\mathbb{P}_0^n, \mathbb{P}_1^n$ be the joint laws of the observations (1.1) when respectively $f = f_0$, $f = f_1$. A sufficient statistic for $\{\mathbb{P}_0^n, \mathbb{P}_1^n\}$ is given by $T_n := \log\big(\mathrm{d}\mathbb{P}_0^n / \mathrm{d}\mathbb{P}_1^n\big)$, and

$$T_n \overset{(\text{law})}{=} \begin{cases} N\left(-\dfrac{v_n}{2}, v_n\right) & \text{under } \mathbb{P}_0^n, \\ N\left(\dfrac{v_n}{2}, v_n\right) & \text{under } \mathbb{P}_1^n, \end{cases}$$

where, by definition of ρ_n:

$$\begin{aligned} v_n = \frac{n}{\sigma^2}\|f_0 - f_1\|^2_{L^2(\mu)} &= \frac{n}{\sigma^2}\int (f_0(x) - f_1(x))^2 \mu(x)\mathrm{d}x \\ &\leqslant n L'^2 \rho_n^{2s'} \mu(I_{\rho_n}) g_\infty^2/\sigma^2 = 2(p-1)\gamma \log n. \end{aligned}$$

Since

$$I_q = \exp\big(v_n(q-1)/2\big) \leqslant n^{\gamma},$$

taking $\delta := \widehat{f}_n(x_0)$, $\theta_2 := f_1(x_0)$, $\theta_1 := f_0(x_0)$ and $\varepsilon := a_n$ within Theorem 4 entails

$$R_p(\delta, \theta_2) \geqslant \big(L' \rho_n^{s'} g(0) - 2C a_n n^{-\gamma} n^{\gamma}\big)^p \geqslant C \rho_n^{s'p},$$

since $\lim_n a_n / \rho_n^{s'} \to 0$, and the proposition follows.

□

Proof of Theorem 3. Since $B \subset A$, equations (4.2) and (4.3) entail that $\widetilde{f}_n$ is superefficient at any function $f_0 \in B$. More precisely, $\widetilde{f}_n$ satisfies (5.21) with

$$\gamma = \frac{(s - s')(\beta + 1)}{2(1 + 2s' + \beta)(1 + 2s + \beta)} > 0$$

for any $f_0 \in B$. The conclusion follows from Proposition 2. □

Acknowledgements. I am very grateful to both anonymous referees, who helped me to improve significantly the presentation of the paper, and I wish to thank Marc Hoffmann for helpful advices and encouragements.

Appendix A. Some facts on regular variation

We recall briefly some properties of regularly varying functions. The results stated in this section can be found in Bingham [3]. In all the following, let ℓ be a slowly varying function. An important fact is that the property

$$\lim_{h \to 0^+} \ell(yh)/\ell(h) = 1 \tag{A.1}$$

actually holds *uniformly* for y in any compact set of $(0, +\infty)$. If $R_1 \in \mathrm{RV}(\alpha_1)$ and $R_2 \in \mathrm{RV}(\alpha_2)$, we have

$$R_1 \times R_2 \in \mathrm{RV}(\alpha_1 + \alpha_2) \text{ and } R_1(R_2(\cdot)) \in \mathrm{RV}(\alpha_1 \times \alpha_2). \tag{A.2}$$

If $R \in \mathrm{RV}(\gamma)$ with $\gamma \in \mathbb{R} - \{0\}$, we have

$$R(h) \to \begin{cases} 0 & \text{if } \gamma > 0, \\ +\infty & \text{if } \gamma < 0, \end{cases} \tag{A.3}$$

as $h \to 0^+$. If $\gamma > -1$, we have:

$$\int_0^h t^\gamma \ell(t) \mathrm{d}t \sim (1+\gamma)^{-1} h^{1+\gamma} \ell(h) \text{ as } h \to 0^+, \tag{A.4}$$

and $h \mapsto \int_0^h t^\gamma \ell(t)\mathrm{d}t$ is regularly varying with index $1 + \gamma$. This result is known as the Karamata theorem. Let us define (R is continuous)

$$R^{\leftarrow}(y) = \inf\{h \geqslant 0 \text{ such that } R(h) \geqslant y\},$$

which is the generalized inverse of R. If $R \in \mathrm{RV}(\gamma)$ for some $\gamma > 0$, there exists $R^- \in \mathrm{RV}(1/\gamma)$ such that

$$R(R^-(h)) \sim R^-(R(h)) \sim h \text{ as } h \to 0^+, \tag{A.5}$$

and R^- is unique up to an asymptotic equivalence. Moreover, one version of R^- is $R^{\leftarrow}$.

References

[1] A. Antoniadis, G. Gregoire and P. Vial, Random design wavelet curve smoothing. *Statist. Probab. Lett.* **35** (1997) 225–232.

[2] Y. Baraud, Model selection for regression on a random design. *ESAIM Probab. Statist.* **6** (2002) 127–146 (electronic).

[3] N.H. Bingham, C.M. Goldie and J.L. Teugels, *Regular Variation.* Encyclopedia of Mathematics and its Applications, Cambridge University Press (1989).

[4] L. Brown and T. Cai, Wavelet shrinkage for nonequispaced samples. *Ann. Statist.* **26** (1998) 1783–1799.

[5] L.D. Brown and M.G. Low, A constrained risk inequality with applications to nonparametric functional estimations. *Ann. Statist.* **24** (1996) 2524–2535.

[6] T.T. Cai, M. Low and L.H. Zhao, *Tradeoffs between global and local risks in nonparametric function estimation.* Tech. rep., Wharton, University of Pennsylvania, `http://stat.wharton.upenn.edu/~tcai/paper/html/Tradeoff.html` (2004).

[7] V. Delouille, J. Simoens and R. Von Sachs, Smooth design-adapted wavelets for nonparametric stochastic regression. *J. Amer. Statist. Soc.* **99** (2004) 643–658.

[8] J. Fan and I. Gijbels, Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B. Methodological* **57** (1995) 371–394.

[9] J. Fan and I. Gijbels, *Local polynomial modelling and its applications.* Monographs on Statistics and Applied Probability, Chapman & Hall, London (1996).

[10] S. Gaïffas, Convergence rates for pointwise curve estimation with a degenerate design. *Mathematical Methods of Statistics* **1** (2005) 1–27. Available at `http://hal.ccsd.cnrs.fr/ccsd-00003086/en/`

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

[12] G. Kerkyacharian and D. Picard, Regression in random design and warped wavelets. *Bernoulli*, **10** (2004) 1053–1105.

[13] O.V. Lepski, Asymptotically minimax adaptive estimation i: Upper bounds, optimally adaptive estimates. *Theory Probab. Applic.* **36** (1988) 682–697.

[14] O.V. Lepski, On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, **35** (1990) 454–466.

[15] 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** (1997) 929–947.

[16] O.V. Lepski and V.G. Spokoiny, Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** (1997) 2512–2546.

[17] V. Maxim, *Restauration de signaux bruités sur des plans d'experience aléatoires.* Ph.D. thesis, Université Joseph Fourier, Grenoble 1 (2003).

[18] V.G. Spokoiny, Estimation of a function with discontinuities *via* local polynomial fit with an adaptive window choice. *Ann. Statist.* **26** (1998) 1356–1378.

[19] C.J. Stone, Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** (1980) 1348–1360.

[20] A. Tsybakov, *Introduction à l'estimation non-paramétrique.* Springer (2003).

[21] M.-Y. Wong and Z. Zheng, Wavelet threshold estimation of a regression function with random design. *J. Multivariate Anal.* **80** (2002) 256–284.