

Uniform estimation of a signal based on inhomogeneous data

Stéphane Gaïffas

Université Pierre et Marie Curie – Paris 6
Laboratoire de Statistique Théorique et Appliquée

Abstract: The aim of this paper is to recover a signal based on inhomogeneous noisy data (the amount of data can vary strongly from one point to another.) In particular, we focus on the understanding of the consequences of the inhomogeneity of the data on the accuracy of estimation. For that purpose, we consider the model of regression with a random design, and we consider the minimax framework. Using the uniform metric weighted by a spatially-dependent rate in order to assess the accuracy of estimators, we are able to capture the deformation of the usual minimax rate in situations where local lacks of data occur (the latter are modelled by a design density with vanishing points). In particular, we construct an estimator both design and smoothness adaptive, and we develop a new criterion to prove the optimality of these deformed rates.

Key words and phrases: nonparametric regression, adaptive estimation, minimax theory, random design.

1. Introduction

Motivations. A problem particularly prominent in statistical literature is the adaptive reconstruction of a signal based on irregularly sampled noisy data. In several practical situations, the statistician cannot obtain “nice” regularly sampled observations, because of various constraints linked with the source of the data, or the way the data is obtained. For instance, in signal or image processing, the irregular sampling can be due to the process of motion or disparity compensation (used in advanced video processing), while in topography, measurement constraints are linked with the properties of the ground. See [Feichtinger and Gröchenig \(1994\)](#) for a survey on irregular sampling, [Almansa et al. \(2003\)](#), [Vázquez et al. \(2000\)](#) for applications concerning respectively satellite image and stereo imaging, and [Jansen et al. \(2004\)](#) for examples of geographical constraints.

Such constraints can result in a lack of data, that can be locally very strong. As a consequence, the accuracy of a procedure based on such data can become very poor locally. The aim of the paper is to study from a theoretical point of view the consequences of the *inhomogeneity* of the data on the reconstruction of a univariate signal. Natural questions arise: how does the inhomogeneity impact on the accuracy of estimation? What does the

optimal convergence rate become in such situations? Can the rate vary strongly from one point to another, and how?

The model. The widest spread way to model such observations is as follows. We model the available data $[(X_i, Y_i); 1 \leq i \leq n]$ by

$$Y_i = f(X_i) + \sigma \xi_i, \quad (1)$$

where ξ_i are i.i.d. Gaussian standard and independent of the X_i 's and where $\sigma > 0$ is the noise level. The design variables X_i are i.i.d. with density μ with respect to the Lebesgue measure. The density μ is unknown to the statistician, and for simplicity, we assume that its support is $[0, 1]$. The more the density μ is “far” from the uniform law, the more the data drawn from (1) is inhomogeneous. A simple way to include situations with local lacks of data within the model (1) is to allow the density μ to be vanishing at some points. This kind of behaviour is not commonly used in literature, since most papers assume μ to be uniformly bounded away from zero. Below, we give references where this kind of design is handled.

In practice, we don't know μ , since it requires to know in a precise way the constraints making the observation irregularly sampled, neither do we know the smoothness of f . Therefore, a convenient procedure shall adapt both to the design and to the smoothness of f . Such a procedure (that is proved to be optimal) is constructed here.

Methodology. We want to reconstruct f globally with sup norm loss. The choice of sup norm for measuring the error of estimation is crucial for the study conducted here. Indeed, it appears that it allows to capture in a very simple way the consequences of inhomogeneity directly on the convergence rate: when the data is inhomogeneous, the optimal rate is deformed (in comparison with the usual rate), see Theorem 1 and 2 in Section 2.

The sup norm choice leads to a particular adaptive estimation method, that can handle “very” inhomogeneous design. This method involves an interpolation transform, where the scaling coefficients are estimated by local polynomials with a smoothing parameter selected by a Lepski type procedure, see for instance Lepski et al. (1997). The Lepski type procedure developed here is adapted to the random design setting, when the design law is unknown. Note that the original adaptive method from Lepski, see for instance Lepski (1990), was developed only in the Gaussian white noise model, which is an idealized version of (1) when the design is uniform: see for instance Brown and Low (1996) and Brown et al. (2002).

If we measure the error of estimation with \mathbb{L}^2 -norm, which is more standard in non-parametric literature, the phenomenon of deformation of the rate does not occur: see for

instance the results from Chesneau (2007), which includes design densities that can vanish. Moreover, in \mathbb{L}^2 estimation, more standard tools shall be used, like orthogonal series, splines, or wavelets, see for instance Green and Silverman (1994), Efromovich (1999) and Härdle et al. (1998).

Literature. Pointwise estimation at a point where the design vanishes is studied in Hall et al. (1997), with the use of a local linear procedure. This design behaviour is given as an example in Guerre (1999), where a more general setting for the design is considered, with a Lipschitz regression function. In Gaïffas (2005a), pointwise minimax rates over Hölder classes are computed for several design behaviours, and an adaptive estimator for the pointwise risk is constructed in Gaïffas (2005b). In these papers, it appears that, depending on the design behaviour at the estimation point, the range of minimax rates is very wide: from very slow (logarithmic) rates to very fast quasi-parametric rates. Many adaptive techniques have been developed in literature for handling irregularly sampled data. Among wavelet methods, see Hall et al. (1997) for interpolation; Antoniadis et al. (1997), Antoniadis and Pham (1998), Brown and Cai (1998), Hall et al. (1998) and Wong and Zheng (2002) for transformation and binning; Antoniadis and Fan (2001) for a penalization approach; Delouille et al. (2001) and Delouille et al. (2004) for the construction of design-adapted wavelet via lifting; Pensky and Wiens (2001) for projection-based techniques; Kerkycharian and Picard (2004) for warped wavelets. For model selection, see Baraud (2002). See also the PhD manuscripts from Maxim (2003) and Delouille (2002).

2. Results

To measure the smoothness of f , we consider the standard Hölder class $H(s, L)$, where $s, L > 0$, which is defined as the set of all the functions $f : [0, 1] \rightarrow \mathbb{R}$ such that

$$|f^{(\lfloor s \rfloor)}(x) - f^{(\lfloor s \rfloor)}(y)| \leq L|x - y|^{s - \lfloor s \rfloor}, \quad \forall x, y \in [0, 1],$$

where $\lfloor s \rfloor$ is the largest integer smaller than s . Minimax theory over such classes is standard: we know from Stone (1982) that in model (1), the minimax rate is equal to $(\log n/n)^{s/(2s+1)}$ over $H(s, L)$ whenever μ is continuous and uniformly bounded away from zero.

We use the notation $\mu(I) := \int_I \mu(t)dt$. We recall that μ is the common density of the X_i (wrt the Lebesgue measure). If $F = H(s, L)$ is fixed, we consider the sequence of positive curves $h_n(\cdot) = h_n(\cdot; F, \mu)$ satisfying

$$Lh_n(x)^s = \sigma \left(\frac{\log n}{n\mu([x - h_n(x), x + h_n(x)])} \right)^{1/2} \quad (2)$$

for all $x \in [0, 1]$, and we define

$$r_n(x; F, \mu) := Lh_n(x; F, \mu)^s.$$

Since $h \mapsto h^{2s} \mu([x-h, x+h])$ is increasing for any x , these curves are well-defined (for n large enough) and unique.

In Theorem 1 below, we show that $r_n(\cdot) = r_n(\cdot; F, \mu)$ is an upper bound over F . This spatially-dependent rate is achievable by an adaptive estimator, over a whole family of Hölder classes. In Theorem 2 below, we prove that, in some sense, this rate is optimal. In the next example, we give an explicit example of such a spatially-dependent rate.

Example. When $s = 1$, $\sigma = L = 1$ and $\mu(x) = 4|x-1/2|\mathbf{1}_{[0,1]}(x)$, the solution to Equation (2) can be written as

$$r_n(x) = (\log n/n)^{\alpha_n(x)},$$

where

$$\alpha_n(x) = \begin{cases} \frac{1}{3} \left(1 - \frac{\log(1-2x)}{\log(\log n/n)}\right) & \text{when } x \in \left[0, \frac{1}{2} - \left(\frac{\log n}{2n}\right)^{1/4}\right], \\ \frac{\log \left(((x-1/2)^4 + 4 \log n/n)^{1/2} - (x-1/2)^2 \right) - \log 2}{2 \log(\log n/n)} & \text{when } x \in \left[\frac{1}{2} - \left(\frac{\log n}{2n}\right)^{1/4}, \frac{1}{2} + \left(\frac{\log n}{2n}\right)^{1/4}\right], \\ \frac{1}{3} \left(1 - \frac{\log(2x-1)}{\log(\log n/n)}\right) & \text{when } x \in \left[\frac{1}{2} + \left(\frac{\log n}{2n}\right)^{1/4}, 1\right]. \end{cases}$$

In this example, the amount of data is low at the middle of the unit interval. The consequence is that the convergence rate has two “regimes”. Indeed, $r_n(1/2) = (\log n/n)^{1/4}$ is slower than the rate at the boundaries $r_n(0) = r_n(1) = (\log n/n)^{1/3}$, which comes from the standard minimax rate $(\log n/n)^{s/(2s+1)}$ with $s = 1$. Hence, in this example, $r_n(\cdot)$ switches from one “regime” to another. In view of Theorem 2 below, we know that, in some sense, this phenomenon is unavoidable. We show the shape of this deformed rate for several sample sizes in Figure 1.

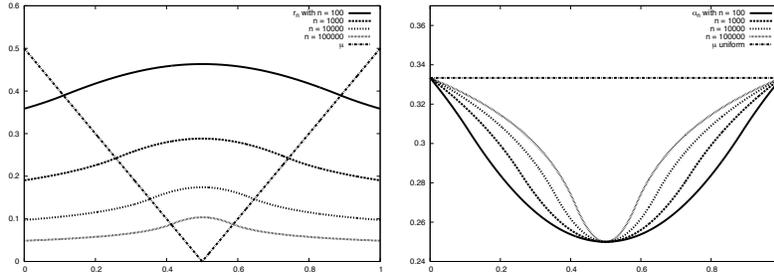


FIGURE 1. $r_n(\cdot)$ and $\alpha_n(\cdot)$ for several sample sizes

In what follows, $a_n \lesssim b_n$ means that $a_n \leq Cb_n$ for any n , where $C > 0$ is independent of n . From now on, C stands for a generic constant that can vary from place to place. It can depend on the parameters of the setting, namely $R, L, Q, w(\cdot)$, but not on f nor n . Let $\mathbf{E}_{f\mu}$ denote the expectation with respect to the joint law $\mathbf{P}_{f\mu}$ of $[(X_i, Y_i); 1 \leq i \leq n]$. Let $w(\cdot)$ be a loss function, namely a non-negative and non-decreasing function such that $w(0) = 0$ and $w(x) \leq A(1 + |x|^b)$ for some $A, b > 0$. If $Q > 0$, we define $H^Q(s, L) := H(s, L) \cap \{f \mid \|f\|_\infty \leq Q\}$ (the constant Q needs not to be known). Let R be a fixed natural integer.

Upper bound. In this section, we show that the spatially-dependent rate $r_n(\cdot)$ defined by (2) is an upper bound over Hölder classes. Let us consider the following

Assumption D. We assume that μ is continuous, and that whether $\mu(x) > 0$ for any x , or $\mu(x) = 0$ for a finite number of x . Moreover, for any x such that $\mu(x) = 0$, we assume that there exists $\beta(x) \geq 0$ such that $\mu(y) = |y - x|^{\beta(x)}$ for any y in a neighbourhood of x .

Theorem 1. *Let us assume that the density μ satisfies Assumption D. Then, for any $F = H^Q(s, L)$ where $s \in (0, R + 1]$, the estimator \hat{f}_n given by (11) satisfies*

$$\sup_{f \in F} \mathbf{E}_{f\mu} \left[w \left(\sup_{x \in [0, 1]} r_n(x)^{-1} |\hat{f}_n(x) - f(x)| \right) \right] \leq C \quad (3)$$

as $n \rightarrow +\infty$, where $r_n(\cdot) = r_n(\cdot; F, \mu)$ is given by (2) and where $C > 0$ is a fixed constant, depending on the parameters $R, L, Q, w(\cdot)$.

This theorem assesses the estimator \hat{f}_n (constructed in Section 3 below) over function sets F in a family of Hölder classes. This estimator is smoothness adaptive, since it converges with the spatially-dependent rate $r_n(\cdot, F, \mu)$ uniformly over F , which is the optimal rate in view of Theorem 2 below. Moreover, this estimator is also “design-adaptive”, since it does not depend within its construction on the design density (which is unknown to the statistician).

Remark. Within Theorem 1, there are two situations:

- $\mu(x) > 0$ for any x : we have $r_n(x) \asymp (\log n/n)^{s/(2s+1)}$ for any x , which is the standard minimax rate over $H(s, L)$ ($a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$). However, this result is new since adaptive estimators over Hölder balls in regression with random design were not previously constructed.
- $\mu(x) = 0$ for one or several x : the rate $r_n(\cdot)$ can vary strongly, depending on the behaviour of μ . For instance, in the example below, $r_n(\cdot)$ changes *in order* from one point to another, from $(\log n/n)^{1/4}$ up to $(\log n/n)^{1/3}$.

Remark. For the statement of Theorem 1, we need to assume that $\|f\|_\infty \leq Q$ for some $Q > 0$ (unknown to the statistician). This assumption is necessary, since the upper bound is uniform over Hölder classes, for the sup norm risk.

Remark. Implicitly, we assumed in Theorem 1 that $s \in (0, R + 1]$, where R is a known parameter. Indeed, in the minimax framework considered here, the fact of knowing an upper bound for the smoothness s is usual in the study of adaptive methods.

Optimality of $r_n(\cdot)$. We have seen that the rate $r_n(\cdot)$ defined by (2) is an upper bound over Hölder classes, see Theorem 1. In Theorem 2 below, we prove that this rate is indeed optimal. In order to show that $r_n(\cdot)$ is optimal in the minimax sense over some class F , the classical criterion consists in showing that for some constant $C > 0$,

$$\inf_{\widehat{f}_n} \sup_{f \in F} \mathbf{E}_{f\mu} [w(\sup_{x \in [0,1]} r_n(x)^{-1} |\widehat{f}_n(x) - f(x)|)] \geq C, \quad (4)$$

where the infimum is taken among all estimators based on the observations (1). However, this criterion does not exclude the existence of another normalisation $\rho_n(\cdot)$ that can improve $r_n(\cdot)$ in some regions of $[0, 1]$. Indeed, (4) roughly consists in a minoration of the uniform risk over the whole unit interval and then, only over some particular points. Therefore, we need a new criterion that strengthens the usual minimax one to prove the optimality of $r_n(\cdot)$. The idea is simple: we localize (4) by replacing the supremum over $[0, 1]$ by a supremum over any (small) interval $I_n \subset [0, 1]$, that is

$$\inf_{\widehat{f}_n} \sup_{f \in F} \mathbf{E}_{f\mu} [w(\sup_{x \in I_n} r_n(x)^{-1} |\widehat{f}_n(x) - f(x)|)] \geq C, \quad \forall I_n. \quad (5)$$

It is noteworthy that in (5), the length of the intervals cannot be arbitrarily small. Actually, if an interval I_n has a length smaller than a given limit, (5) does not hold anymore. Indeed, beyond this limit, we can improve $r_n(\cdot)$ for the risk localized over I_n : we can construct an estimator \widehat{f}_n such that

$$\sup_{f \in F} \mathbf{E}_{f\mu} [w(\sup_{x \in I_n} r_n(x)^{-1} |\widehat{f}_n(x) - f(x)|)] = o(1), \quad (6)$$

see Proposition 1 below. The phenomenon described in this section, which concerns the uniform risk, is related to the results from Cai and Low (2005) for shrunk \mathbb{L}^2 risks. In what follows, $|I|$ stands for the length of an interval I . We recall that we use the notation $\mu(I) = \int_I \mu(x) dx$.

Theorem 2. *Suppose that*

$$\mu(I) \gtrsim |I|^{\beta+1} \quad (7)$$

uniformly for any interval $I \subset [0, 1]$, where $\beta \geq 0$ and let $F = H(s, L)$. Then, for any interval $I_n \subset [0, 1]$ such that

$$|I_n| \sim n^{-\alpha} \quad (8)$$

with $\alpha \in (0, (1 + 2s + \beta)^{-1})$, we have

$$\inf_{\hat{f}_n} \sup_{f \in F} \mathbf{E}_{f\mu} \left[w \left(\sup_{x \in I_n} r_n(x)^{-1} |\hat{f}_n(x) - f(x)| \right) \right] \geq C \quad (9)$$

as $n \rightarrow +\infty$, where $r_n(\cdot) = r_n(\cdot; F, \mu)$ is given by (2).

Corollary 1. If $v_n(\cdot)$ is an upper bound over $F = H(s, L)$ in the sense of (3), we have

$$\sup_{x \in I_n} v_n(x)/r_n(x) \geq C$$

for any interval I_n as in Theorem 2. Hence, $r_n(\cdot)$ cannot be improved uniformly over an interval with length $n^{\varepsilon-1/(1+2s+\beta)}$, for any arbitrarily small $\varepsilon > 0$.

Proposition 1. Let $F = H^Q(s, L)$ and ℓ_n be a positive sequence satisfying $\log \ell_n = o(\log n)$.

a) Let μ be such that $0 < \mu(x) < +\infty$ for any $x \in [0, 1]$. If I_n is an interval satisfying

$$|I_n| \sim (\ell_n/n)^{1/(1+2s)},$$

we can construct an estimator \hat{f}_n such that

$$\sup_{f \in F} \mathbf{E}_{f\mu} \left[w \left(\left(\frac{n}{\log n} \right)^{s/(2s+1)} \sup_{x \in I_n} |\hat{f}_n(x) - f(x)| \right) \right] = o(1).$$

b) Let $\mu(x_0) = 0$ for some $x_0 \in [0, 1]$ and $\mu([x_0 - h, x_0 + h]) = h^{\beta+1}$ where $\beta \geq 0$ for any h in a fixed neighbourhood of 0. If

$$I_n = [x_0 - (\ell_n/n)^{1/(1+2s+\beta)}, x_0 + (\ell_n/n)^{1/(1+2s+\beta)}],$$

we can construct an estimator \hat{f}_n such that

$$\sup_{f \in F} \mathbf{E}_{f\mu} \left[w \left(\sup_{x \in I_n} r_n(x)^{-1} |\hat{f}_n(x) - f(x)| \right) \right] = o(1).$$

Remark. Note that in case a), $r_n(x) \asymp (\log n/n)^{s/(2s+1)}$ for any $x \in [0, 1]$ and that (7) holds with $\beta = 0$.

This proposition entails that $r_n(\cdot)$ can be improved for localized risks (6) over intervals I_n with size $(\ell_n/n)^{1/(1+2s+\beta)}$ where ℓ_n can be a slow term such as $(\log n)^\gamma$ for any $\gamma \geq 0$. A consequence is that the lower bound in Theorem 2 cannot be improved, since (9) does not hold anymore when I_n has a length smaller than (8). This phenomenon is linked both to the choice of the uniform metric for measuring the error of estimation, and to the nature of the noise within the model (1). It is also a consequence of the minimax paradigm: it is well-known that the minimax risk actually concentrates on some critical functions of the

considered class (that we rescale and place within I_n here, hence the critical length for I_n), which is a property allowing to prove lower bounds such as the one in Theorem 2.

3. Construction of an adaptive estimator

The adaptive method proposed here differs from the techniques mentioned in Introduction. Indeed, it is not appropriate to apply a wavelet decomposition of the scaling coefficients at the finest scale here, since it is a \mathbb{L}^2 -transform, while the criterion (3) considered here uses the uniform metric. This is the reason why here, the analysis is focused on a precise estimation of the scaling coefficients. Each scaling coefficient is estimated by a local polynomial estimator (LPE) of f with an adaptively selected bandwidth.

Let $(V_j)_{j \geq 0}$ be a multiresolution analysis of $\mathbf{L}^2([0, 1])$ with a scaling function ϕ compactly supported and R -regular (the parameter R comes from Theorem 1), which ensures that

$$\|f - P_j f\|_\infty \lesssim 2^{-js} \quad (10)$$

for any $f \in H(s, L)$ with $s \in (0, R + 1]$, where P_j denotes the projection onto V_j . We use P_j as an interpolation transform. Interpolation transforms in the unit interval are constructed in Donoho (1992) and Cohen et al. (1993). We have $P_j f = \sum_{k=0}^{2^j-1} \alpha_{jk} \phi_{jk}$, where $\phi_{jk}(\cdot) = 2^{j/2} \phi(2^j \cdot - k)$ and $\alpha_{jk} = \int f \phi_{jk}$. We consider the largest integer J such that $N := 2^J \leq n$, and we estimate the scaling coefficients $(\alpha_{jk})_{0 \leq k \leq 2^j}$ at the high resolution level $j = J$. If $\hat{\alpha}_{Jk}$ are estimators of α_{Jk} , we simply consider

$$\hat{f}_n := \sum_{k=0}^{2^J-1} \hat{\alpha}_{Jk} \phi_{Jk}. \quad (11)$$

Let us denote by Pol_R the set of all real polynomials with degree at most R . Suppose for the moment that we are given some accurate estimators $\bar{f}_k \in \text{Pol}_R$ of f over the support of ϕ_{Jk} . Then $\alpha_{Jk} = \int f \phi_{Jk} \approx \int \bar{f}_k \phi_{Jk}$. In the particular situation where the scaling function ϕ has R moments, that is

$$\int \phi(t) t^p dt = \mathbf{1}_{p=0}, \quad p \in \{0, \dots, R\}, \quad (12)$$

and when f is s -Hölder for $s \in (0, R + 1]$, accurate estimators of α_{Jk} are given by the statistic

$$\hat{\alpha}_{Jk} := 2^{-J/2} \bar{f}_k(k2^{-J}). \quad (13)$$

This comes from the fact that when $f \in H(s, L)$, we have $\int f \phi_{Jk} \approx \int \bar{f}_k \phi_{Jk} = 2^{-J/2} \int \bar{f}_k(t) \phi(t) dt$, where \bar{f}_k is the Taylor expansion of f at $k2^{-J}$ up to the order $\lfloor s \rfloor$. If ϕ does not satisfies (12), $\int \bar{f}_k \phi_{Jk}$ can be computed exactly using a quadrature formula, in the same way

as in [Delyon and Juditsky \(1995\)](#). Indeed, there is a matrix \mathbf{Q}_J (characterized by ϕ) with entries (q_{Jkm}) for $(k, m) \in \{0, \dots, 2^J - 1\}^2$ such that

$$\int P \phi_{Jk} = 2^{-J/2} \sum_{m \in \Gamma_{Jk}} q_{Jkm} P(m/2^J) \quad (14)$$

for any $P \in \text{Pol}_R$. Within this equation, the entries of the quadrature matrix \mathbf{Q}_J satisfy

$$q_{Jkm} \neq 0 \rightarrow |k - m| \leq L_\phi \text{ and } m \in \Gamma_{Jk}, \quad (15)$$

where $L_\phi > 0$ is the support length of ϕ (the matrix \mathbf{Q}_J is band-limited). For instance, if we consider the Coiflets basis, which satisfies the moment condition (12), we have $q_{Jkm} = \mathbf{1}_{k=m}$, and we can use directly (13). If the $(\phi(\cdot - k))_k$ are orthogonal, then $q_{Jkm} = \phi(m - k)$, see [Delyon and Juditsky \(1995\)](#).

For the sake of simplicity, we assume in what follows that ϕ satisfies the moment condition (12), thus the coefficients α_{Jk} are estimated by (13). Each polynomial \bar{f}_k in (13) is a local polynomial estimator computed at $k2^{-J}$ with smoothing parameter $\hat{\Delta}_k$ (the so-called “bandwidth”, which is, here, an interval included in $[0, 1]$ containing the point $k2^{-J}$). Hence, we write $\bar{f}_k = \bar{f}_k^{(\hat{\Delta}_k)}$. The smoothing parameters $\hat{\Delta}_k$ are selected via an adaptive rule. Below, we describe the computation of the local polynomial estimators and we define the selection rule for the $\hat{\Delta}_k$.

Local polynomials. The polynomials used to estimate each scaling coefficients are defined via a slightly modified version of the local polynomial estimator (LPE). This linear method of estimation is standard, see for instance [Fan and Gijbels \(1995, 1996\)](#), among many others. For any interval $\delta \subset [0, 1]$, we define the empirical sample measure $\bar{\mu}_n(\delta) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \in \delta}$, where $\mathbf{1}_{X_i \in \delta}$ equals one if $X_i \in \delta$, and zero otherwise. If $\bar{\mu}_n(\delta) > 0$, we introduce the pseudo-inner product

$$\langle f, g \rangle_\delta := \frac{1}{\bar{\mu}_n(\delta)} \int_\delta f g d\bar{\mu}_n, \quad (16)$$

and $\|g\|_\delta := \langle g, g \rangle_\delta^{1/2}$ the corresponding pseudo-norm. A local polynomial estimator is computed for each point of the regular grid $\{k2^{-J}; 0 \leq k \leq 2^J\}$. Let δ be an interval containing $k2^{-J}$. The standard LPE at $k2^{-J}$ is defined as the polynomial $\bar{f}_k^{(\delta)}$ of degree R which is the closest to the data in the least square sense, with respect to the localized empirical norm $\|\cdot\|_\delta$. More precisely, if $\varphi_{kp}(\cdot) := (\cdot - k2^{-J})^p$, $0 \leq p \leq R$, we look for $\bar{f}_k^{(\delta)} \in \text{Span}\{\varphi_{kp}(\cdot); 0 \leq p \leq R\}$ satisfying

$$\langle \bar{f}_k^{(\delta)}, \varphi \rangle_\delta = \langle Y, \varphi \rangle_\delta \quad (17)$$

for any $\varphi(\cdot) \in \{\varphi_{kp}(\cdot); 0 \leq p \leq R\}$. The coefficients vector $\bar{\theta}_k^{(\delta)} \in \mathbb{R}^{R+1}$ of the polynomial $\bar{f}_k^{(\delta)}$ is therefore solution, when it makes sense, to the linear system

$$\mathbf{X}_k^{(\delta)} \theta = \mathbf{Y}_k^{(\delta)},$$

where for $0 \leq p, q \leq R$:

$$(\mathbf{X}_k^{(\delta)})_{p,q} := \langle \varphi_{kp}, \varphi_{kq} \rangle_\delta \quad \text{and} \quad (\mathbf{Y}_k^{(\delta)})_p := \langle Y, \varphi_{kp} \rangle_\delta. \quad (18)$$

This is the standard definition of the LPE. Moreover, whenever $\bar{\mu}_n(\delta) = 0$, we simply take $\bar{f}_k^{(\delta)} = 0$. We modify this linear system as follows: when the smallest eigenvalue of $\mathbf{X}_k^{(\delta)}$ (which is non-negative) is too small, we add a correcting term allowing to bound it from below. We introduce

$$\bar{\mathbf{X}}_k^{(\delta)} := \mathbf{X}_k^{(\delta)} + (n\bar{\mu}_n(\delta))^{-1/2} \mathbf{Id}_{R+1} \mathbf{1}_{\Omega_k(\delta)^c},$$

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

$$\Omega_k(\delta) := \{\lambda(\mathbf{X}_k^{(\delta)}) > (n\bar{\mu}_n(\delta))^{-1/2}\}, \quad (19)$$

where $\lambda(M)$ stands for the smallest eigenvalue of a matrix M . The quantity $(n\bar{\mu}_n(\delta))^{-1/2}$ comes from the variance of $\bar{f}_k^{(\delta)}$, and this particular choice preserves the convergence rate of the method. This modification of the classical LPE is convenient in situations with little data. Below is a precise definition of the LPE at $k2^{-J}$ that we consider here.

Definition 1. When $\bar{\mu}_n(\delta) > 0$, we consider the solution $\bar{\theta}_k^{(\delta)}$ of the linear system

$$\bar{\mathbf{X}}_k^{(\delta)} \theta = \mathbf{Y}_k^{(\delta)}, \quad (20)$$

and we introduce $\bar{f}_k^{(\delta)}(x) := (\bar{\theta}_k^{(\delta)})_0 + (\bar{\theta}_k^{(\delta)})_1(x - k2^{-J}) + \dots + (\bar{\theta}_k^{(\delta)})_R(x - k2^{-J})^R$. When $\bar{\mu}_n(\delta) = 0$, we take simply $\bar{f}_k^{(\delta)} := 0$.

Adaptive bandwidth selection. The adaptive procedure selecting the intervals $\widehat{\Delta}_k$ is based on a method introduced by Lepski (1990), see also Lepski et al. (1997), and Lepski and Spokoiny (1997). 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 (1997)), 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). Bandwidth selection procedures in local polynomial estimation can be found in Fan and Gijbels (1995), Goldenshluger and Nemirovski (1997) or Spokoiny (1998), among others.

The idea of the adaptive rule for selecting the interval δ at the point $k2^{-J}$ is the following: when $\bar{f}_k^{(\delta)}(x)$ is close to $f(x)$ for $x \in \delta$ (that is, when δ is well-chosen), we have in view of (17):

$$\langle \bar{f}_k^{(\delta')} - \bar{f}_k^{(\delta)}, \varphi \rangle_{\delta'} = \langle Y - \bar{f}_k^{(\delta)}, \varphi \rangle_{\delta'} \approx \langle Y - f, \varphi \rangle_{\delta'} = \langle \xi, \varphi \rangle_{\delta'}$$

for any $\delta' \subset \delta$ and $\varphi(\cdot) \in \{\varphi_{kp}(\cdot); 0 \leq p \leq R\}$, where the right-hand side is a noise term. Hence, in order to “remove” this noise, we select the largest δ such that this noise term remains smaller than an appropriate threshold, for any $\delta' \subset \delta$ and $\varphi(\cdot) \in \{\varphi_{kp}(\cdot); 0 \leq p \leq R\}$. At each point of the regular grid $\{k2^{-J}; 0 \leq k \leq 2^J\}$, the bandwidth $\hat{\Delta}_k$ is selected in a fixed set of intervals G_k called *grid* (which is defined below) as follows:

$$\hat{\Delta}_k := \operatorname{argmax}_{\delta \in G_k} \left\{ \bar{\mu}_n(\delta) \mid \forall \delta' \in G_k, \delta' \subset \delta, \forall p \in \{0, \dots, R\}, \right. \\ \left. |\langle \bar{f}_k^{(\delta')} - \bar{f}_k^{(\delta)}, \varphi_{kp} \rangle_{\delta'}| \leq \|\varphi_{kp}\|_{\delta'} T_n(\delta, \delta') \right\}, \quad (21)$$

where

$$T_n(\delta, \delta') := \sigma \left[\left(\frac{\log n}{n \bar{\mu}_n(\delta)} \right)^{1/2} + DC_R \left(\frac{\log(n \bar{\mu}_n(\delta))}{n \bar{\mu}_n(\delta')} \right)^{1/2} \right], \quad (22)$$

with $C_R := 1 + (R + 1)^{1/2}$ and $D > (2(b + 1))^{1/2}$, if we want to prove Theorem 1 with a loss function satisfying $w(x) \leq A(1 + |x|^b)$. The threshold choice (22) can be understood in the following way: since the variance of $\bar{f}_k^{(\delta)}$ is of order $(n \bar{\mu}_n(\delta))^{-1/2}$, we see that the two terms in $T_n(\delta, \delta')$ are ratios between a penalizing log term and the variance of the estimators compared by the rule (21). The penalization term is linked with the number of comparisons necessary to select the bandwidth. To prove Theorem 1, we use the grid

$$G_k := \bigcup_{1 \leq i \leq n} \left\{ [k2^{-J} - |X_i - k2^{-J}|, k2^{-J} + |X_i - k2^{-J}|] \right\}, \quad (23)$$

and we recall that the scaling coefficients are estimated by

$$\hat{\alpha}_{Jk} := 2^{-J/2} \bar{f}_k^{(\hat{\Delta}_k)}(k2^{-J}).$$

Remark. In this form, the adaptive estimator has a complexity $O(n^2)$. This can be decreased using a smaller grid. An example of such a grid is the following: first, we sort the (X_i, Y_i) into $(X_{(i)}, Y_{(i)})$ such that $X_{(i)} < X_{(i+1)}$. Then, we consider $i(k)$ such that $k2^{-J} \in [X_{(i(k))}, X_{(i(k)+1)}]$ (if necessary, we take $X_{(0)} = 0$ and $X_{(n+1)} = 1$) and for some $a > 1$ (to be chosen by the statistician) we introduce

$$G_k := \bigcup_{p=0}^{\lfloor \log_a(i(k)+1) \rfloor} \bigcup_{q=0}^{\lfloor \log_a(n-i(k)) \rfloor} \left\{ [X_{(i(k)+1-[a^p])}, X_{(i(k)+[a^q])}] \right\}. \quad (24)$$

With this grid, the selection of the bandwidth is fast, and the complexity of the procedure is $O(n(\log n)^2)$. We can use this grid in practice, but we need extra assumptions on the design if we want to prove Theorem 1 with this grid choice.

4. Proofs

We recall that the weight function $w(\cdot)$ is non-negative, non-decreasing and such that $w(x) \leq A(1 + |x|)^b$ for some $A, b > 0$. We denote by μ^n the joint law of X_1, \dots, X_n and \mathfrak{X}_n the sigma-field generated by X_1, \dots, X_n . $|A|$ denotes both the length of an interval A and the cardinality of a finite set A . M^\top is the transpose of M , and $\xi = (\xi_1, \dots, \xi_n)^\top$. We introduce $x_k := k2^{-J}$ for $k \in \{0, \dots, 2^J\}$. As previously, C stands for a generic constant that can vary from place to place.

Proof of Theorem 1. To prove the upper bound, we use the estimator defined by (11) where ϕ is a scaling function satisfying (12) (for instance the Coifflets basis), and where the scaling coefficients are estimated by (13). In view of (2) and since μ is continuous on $[0, 1]$, we have

$$r_n(x) \gtrsim (\log n/n)^{s/(1+2s)}. \quad (25)$$

Together with (10), this entails

$$\sup_{x \in [0,1]} r_n(x)^{-1} \|f - P_J f\|_\infty \lesssim n^{s/(2s+1)} 2^{-Js} = o(1),$$

since $2^J \asymp n^{-1}$. Hence,

$$\begin{aligned} \sup_{x \in [0,1]} r_n(x)^{-1} |\widehat{f}_n(x) - f(x)| &\lesssim \sup_{x \in [0,1]} r_n(x)^{-1} \left| \sum_{k=0}^{2^J-1} (\widehat{\alpha}_{Jk} - \alpha_{Jk}) \phi_{Jk}(x) \right| \\ &\lesssim \max_{0 \leq k \leq 2^J-1} \sup_{x \in S_k} r_n(x)^{-1} 2^{J/2} |\widehat{\alpha}_{Jk} - \alpha_{Jk}|, \end{aligned}$$

where S_k denotes the support of ϕ_{Jk} . Let f_k be the Taylor polynomial of f at x_k up to the order $\lfloor s \rfloor$. Using (12), we have $\int f_k \phi_{Jk} = 2^{-J/2} f(x_k)$, and since $f \in H(s, L)$, we have $|\alpha_{Jk} - f(x_k)| = |\int f \phi_{Jk} - f(x_k)| \lesssim 2^{-J(s+1/2)}$. Together with (13) and (25), this entails

$$\sup_{x \in [0,1]} r_n(x)^{-1} |\widehat{f}_n(x) - f(x)| \lesssim \max_{0 \leq k \leq 2^J-1} \sup_{x \in S_k} r_n(x)^{-1} |\widehat{f}_k^{\widehat{\Delta}_k}(x_k) - f(x_k)|. \quad (26)$$

Since μ is continuous, $r_n(\cdot)$ is continuously differentiable. Hence, since $|S_k| = 2^{-J} \asymp n^{-1}$, we have $\sup_{x \in S_k} |r_n(x)^{-1} - r_n(x_k)^{-1}| \leq 2^{-J} \|(r_n^{-1})'\|_\infty$, where g' stands for the derivative of g . Moreover, $|(r_n(x)^{-1})'| \lesssim h'_n(x) h_n(x)^{-(s+1)} \lesssim n^{-1}$, since $h'_n(x)$ is uniformly bounded and $h_n(x) \gtrsim (\log n/n)^{1/(2s+1)}$. This entails

$$\sup_{x \in S_k} r_n(x)^{-1} \lesssim r_n(x_k)^{-1}. \quad (27)$$

In what follows, $\|\cdot\|_\infty$ denotes the supremum norm in \mathbb{R}^{R+1} . The following lemma is a version of the bias-variance decomposition of the local polynomial estimator, which is classical: see for instance [Fan and Gijbels \(1995, 1996\)](#), [Goldenshluger and Nemirovski \(1997\)](#), [Spokoiny \(1998\)](#), among others. We define the matrix

$$\mathbf{E}_k^{(\delta)} := \mathbf{\Lambda}_k^{(\delta)} \bar{\mathbf{X}}_k^{(\delta)} \mathbf{\Lambda}_k^{(\delta)},$$

where $\bar{\mathbf{X}}_k$ is given by [\(18\)](#) and $\mathbf{\Lambda}_k^{(\delta)} := \text{diag}[\|\varphi_{k0}\|_\delta^{-1}, \dots, \|\varphi_{kR}\|_\delta^{-1}]$.

Lemma 1. *Conditionally on \mathfrak{X}_n , for any $f \in H(s, L)$ and $\delta \in G_k$, we have*

$$|\bar{f}_k^{(\delta)}(x_k) - f(x_k)| \lesssim \lambda(\mathbf{E}_k^{(\delta)})^{-1} (L|\delta|^s + \sigma(n\bar{\mu}_n(\delta))^{-1/2} \|\mathbf{U}_k^{(\delta)} \xi\|_\infty)$$

on $\Omega_k(\delta)$, where $\mathbf{U}_k^{(\delta)}$ is a \mathfrak{X}_n -measurable matrix of size $(R+1) \times (n\bar{\mu}_n(\delta))$ satisfying $\mathbf{U}_k^{(\delta)} (\mathbf{U}_k^{(\delta)})^\top = \mathbf{Id}_{R+1}$.

The proof of Lemma 1 is given later on. Note that within this lemma, the bandwidth δ can change from one point x_k to another. We denote $\mathbf{U}_k := \mathbf{U}_k^{(\delta_k)}$ for short. Let us define $W := \mathbf{U} \xi$ where $\mathbf{U} := (\mathbf{U}_0^\top, \dots, \mathbf{U}_{2^J}^\top)^\top$. In view of Lemma 1, W is conditionally on \mathfrak{X}_n a centered Gaussian vector such that $\mathbf{E}_{f\mu}[W_k^2 | \mathfrak{X}_n] = 1$ for any $k \in \{0, \dots, (R+1)2^J\}$. We introduce $W^N := \max_{0 \leq k \leq (R+1)2^J} |W_k|$ and the event $\mathcal{W}_N := \{|W^N - \mathbf{E}[W^N | \mathfrak{X}_n]| \leq L_W (\log n)^{1/2}\}$, where $L_W > 0$. We recall the following classical results about the supremum of a Gaussian vector (see for instance in [Ledoux and Talagrand \(1991\)](#)):

$$\mathbf{E}_{f\mu}[W^N | \mathfrak{X}_n] \lesssim (\log N)^{1/2} \lesssim (\log n)^{1/2},$$

and

$$\mathbf{P}_{f\mu}[\mathcal{W}_N^c | \mathfrak{X}_n] \lesssim \exp(-L_W^2 (\log n)/2) = n^{-L_W^2/2}. \quad (28)$$

Let us define the event

$$\mathbb{T}_k := \{\bar{\mu}_n(\Delta_k) \leq \bar{\mu}_n(\hat{\Delta}_k)\}$$

and

$$R_k := \sigma\left(\frac{\log n}{n\bar{\mu}_n(\Delta_k)}\right)^{1/2},$$

where the intervals Δ_k are given by

$$\Delta_k := \operatorname{argmax}_{\delta \in G_k} \left\{ \bar{\mu}_n(\delta) \mid L|\delta|^s \leq \sigma\left(\frac{\log n}{n\bar{\mu}_n(\delta)}\right)^{1/2} \right\}.$$

There is an event $S_n \in \mathfrak{X}_n$ such that $\mu^n[S_n^c] = o(1)$ faster than any power of n , and such that $R_k \asymp r_n(x_k)$ and $\lambda(\mathbf{E}_k^{(\Delta_k)}) \geq \lambda_0$ for some constant λ_0 , uniformly for any $k \in \{0, \dots, 2^J - 1\}$. This event is constructed later on. We decompose

$$|\bar{f}_k^{(\hat{\Delta}_k)}(x_k) - f(x_k)| \leq A_k + B_k + C_k + D_k,$$

where

$$\begin{aligned} A_k &:= |\bar{f}_k^{(\widehat{\Delta}_k)}(x_k) - f(x_k)| \mathbf{1}_{\mathcal{W}_N^c \cup \mathcal{S}_n^c}, \\ B_k &:= |\bar{f}_k^{(\widehat{\Delta}_k)}(x_k) - f(x_k)| \mathbf{1}_{\mathcal{T}_k^c \cap \mathcal{W}_N \cap \mathcal{S}_n}, \\ C_k &:= |\bar{f}_k^{(\widehat{\Delta}_k)}(x_k) - \bar{f}_k^{(\Delta_k)}(x_k)| \mathbf{1}_{\mathcal{T}_k \cap \mathcal{S}_n}, \\ D_k &:= |\bar{f}_k^{(\Delta_k)}(x_k) - f(x_k)| \mathbf{1}_{\mathcal{W}_N \cap \mathcal{S}_n}. \end{aligned}$$

Term A_k . For any $\delta \in G_k$, we have

$$|\bar{f}_k^{(\delta)}(x_k)| \lesssim (n\bar{\mu}_n(\delta))^{1/2} \|f\|_\infty (1 + W^N). \quad (29)$$

This inequality is proved later on. Hence, using together (25), (29) and $\|f\|_\infty \leq Q$, we obtain

$$\mathbf{E}_{f\mu} \left[\left(\max_{0 \leq k \leq 2^J} r_n(x_k)^{-1} |\bar{f}_k^{(\widehat{\Delta}_k)}(x_k)| \right)^{2b} | \mathfrak{X}_n \right] \lesssim n^{2sb/(2s+1)+b+2}.$$

Then, using $w(x) \leq A(1 + |x|^b)$ and the Cauchy-Schwarz inequality, we obtain

$$\mathbf{E}_{f\mu} \left[w \left(\max_{0 \leq k \leq 2^J} r_n(x_k)^{-1} A_k \right) \right] \lesssim n^{sb/(2s+1)+b/2+1} \mathbf{P}_{f\mu} [\mathcal{W}_N^c \cup \mathcal{S}_n^c]^{1/2} = o(1),$$

since $\mu^n[\mathcal{S}_n^c] = o(1)$ faster than any power of n , and L_W can be chosen arbitrarily large in (28).

Term D_k . Using Lemma 1, together with the definition of Δ_k and the fact that $W^N \lesssim (\log n)^{1/2}$ on \mathcal{W}_N , we have

$$\begin{aligned} |\bar{f}_k^{(\Delta_k)}(x_k) - f(x_k)| &\leq \lambda(\mathbf{E}_k^{(\Delta_k)})^{-1} (L|\Delta_k|^s + \sigma(n\bar{\mu}_n(\Delta_k))^{-1/2} W^N) \\ &\leq \lambda(\mathbf{E}_k^{(\Delta_k)})^{-1} R_k (1 + (\log n)^{-1/2} W^N) \\ &\lesssim \lambda(\mathbf{E}_k^{(\Delta_k)})^{-1} r_n(x_k) \end{aligned}$$

on $\mathcal{W}_N \cap \mathcal{S}_n$, thus

$$\mathbf{E}_{f\mu} \left[w \left(\max_{0 \leq k \leq 2^J} r_n(x_k)^{-1} D_k \right) \right] \leq C.$$

Term C_k . We introduce $G_k(\delta) := \{\delta' \in G_k | \delta' \subset \delta\}$ and the following events:

$$\begin{aligned} \mathcal{T}_k(\delta, \delta', p) &:= \{ |\langle \bar{f}_k^{(\delta)} - \bar{f}_k^{(\delta')}, \varphi_{kp} \rangle_{\delta'}| \leq \sigma \|\varphi_{kp}\|_{\delta'} T_n(\delta, \delta') \}, \\ \mathcal{T}_k(\delta, \delta') &:= \cap_{0 \leq p \leq R} \mathcal{T}_k(\delta, \delta', p), \\ \mathcal{T}_k(\delta) &:= \cap_{\delta' \in G_k(\delta)} \mathcal{T}_k(\delta, \delta'). \end{aligned}$$

By the definition (21) of the selection rule, we have $\mathcal{T}_k \subset \mathcal{T}_k(\widehat{\Delta}_k, \Delta_k)$. Let $\delta \in G_k, \delta' \in G_k(\delta)$. On $\mathcal{T}_k(\delta, \delta') \cap \Omega_k(\delta')$ we have (a proof is given later on)

$$|\bar{f}_k^{(\delta)}(x_k) - \bar{f}_k^{(\delta')}(x_k)| \lesssim \lambda(\mathbf{E}_k^{(\delta')})^{-1} \left(\frac{\log n}{n\bar{\mu}_n(\delta')} \right)^{1/2}. \quad (30)$$

Thus, using (30), we obtain

$$\mathbf{E}_{f\mu} \left[w \left(\max_{0 \leq k \leq 2^J} r_n(x_k)^{-1} C_k \right) \right] \leq C.$$

Term B_k . By the definition (21) of the selection rule, we have $\mathbf{T}_k^{\mathfrak{G}} \subset \mathcal{T}_k(\Delta_k)^{\mathfrak{G}}$. We need the following lemma.

Lemma 2. *If $\delta \in G_k$ satisfies*

$$L|\delta|^s \leq \sigma \left(\frac{\log n}{n\bar{\mu}_n(\delta)} \right)^{1/2} \quad (31)$$

and $f \in H(s, L)$, we have

$$\mathbf{P}_{f\mu} [\mathcal{T}_k(\delta)^{\mathfrak{G}} | \mathfrak{X}_n] \leq (R+1)(n\bar{\mu}_n(\delta))^{1-D^2/2}$$

on $\Omega_k(\delta)$, where D is the constant from the threshold (22).

Using together Lemma 2, $\|f\|_{\infty} \leq Q$ and (29), we obtain

$$\mathbf{E}_{f\mu} \left[w \left(\max_{0 \leq k \leq 2^J} R_k^{-1} |\bar{f}_k^{(\hat{\Delta}_k)}(x_k) - f(x_k)| \mathbf{1}_{\mathbf{T}_k^{\mathfrak{G}} \cap \mathcal{W}_N} \right) | \mathfrak{X}_n \right] \leq C,$$

thus

$$\mathbf{E}_{f\mu} \left[w \left(\max_{0 \leq k \leq 2^J} r_n(x_k)^{-1} B_k \right) \right] \leq C,$$

and Theorem 1 follows. \square

Proof of Lemma 1. On $\Omega_k(\delta)$, we have $\bar{\mathbf{X}}_k^{(\delta)} = \mathbf{X}_k^{\delta}$, and $\lambda(\mathbf{X}_k^{(\delta)}) > (n\bar{\mu}_n(\delta))^{-1/2} > 0$, thus $\mathbf{X}_k^{(\delta)}$ and $\mathbf{E}_k^{(\delta)}$ are invertible. Let f_k be the Taylor polynomial of f at x_k up to the order $\lfloor s \rfloor$ and $\theta_k \in \mathbb{R}^{R+1}$ be the coefficient vector of f_k . Using $f \in H(s, L)$, we obtain

$$\begin{aligned} |\bar{f}_k^{(\delta)}(x_k) - f(x_k)| &\lesssim | \langle (\mathbf{\Lambda}_k^{(\delta)})^{-1} (\bar{\theta}_k^{(\delta)} - \theta_k), e_1 \rangle | + |\delta|^s \\ &= | \langle (\mathbf{E}_k^{(\delta)})^{-1} \mathbf{\Lambda}_k^{(\delta)} \mathbf{X}_k^{(\delta)} (\bar{\theta}_k^{(\delta)} - \theta_k), e_1 \rangle | + |\delta|^s. \end{aligned}$$

In view of (17), we have on $\Omega_k(\delta)$ for any $p \in \{0, \dots, R\}$:

$$\begin{aligned} (\mathbf{X}_k^{(\delta)} (\bar{\theta}_k^{(\delta)} - \theta_k))_p &= \langle \bar{f}_k^{(\delta)} - f_k, \varphi_{kp} \rangle \delta \\ &= \langle Y - f_k, \varphi_{kp} \rangle \delta \end{aligned}$$

thus, $\mathbf{X}_k^{(\delta)} (\bar{\theta}_k^{(\delta)} - \theta_k) = B_k^{(\delta)} + V_k^{(\delta)}$ where $(B_k^{(\delta)})_p := \langle f - f_k, \varphi_{kp} \rangle \delta$ and $(V_k^{(\delta)})_p := \langle \xi, \varphi_{kp} \rangle \delta$, which correspond respectively to bias and variance terms. Since $f \in H(s, L)$ and $\lambda(M)^{-1} = \|M^{-1}\|$ for any symmetrical and positive matrix M , we have

$$| \langle (\mathbf{E}_k^{(\delta)})^{-1} \mathbf{\Lambda}_k^{(\delta)} B_k^{(\delta)}, e_1 \rangle | \lesssim \lambda(\mathbf{E}_k^{(\delta)})^{-1} L |\delta|^s.$$

Since $(V_k^{(\delta)})_p = (n\bar{\mu}_n(\delta))^{-1}\mathbf{D}_k^{(\delta)}\xi$ where $\mathbf{D}_k^{(\delta)}$ is the $(R+1) \times (n\bar{\mu}_n(\delta))$ matrix with entries $(\mathbf{D}_k^{(\delta)})_{i,p} := (X_i - x_k)^p$, $X_i \in \delta$, we can write

$$|\langle (\mathbf{E}_k^{(\delta)})^{-1}\mathbf{\Lambda}_k^{(\delta)}V_k^{(\delta)}, e_1 \rangle_\delta| \lesssim \sigma(n\bar{\mu}_n(\delta))^{-1/2} \|(\mathbf{E}_k^{(\delta)})^{-1/2}\| \| \mathbf{U}_k^{(\delta)}\xi \|_\infty,$$

where $\mathbf{U}_k^{(\delta)} := (n\bar{\mu}_n(\delta))^{-1/2}(\mathbf{E}_k^{(\delta)})^{-1/2}\mathbf{\Lambda}_k^{(\delta)}\mathbf{D}_k^{(\delta)}$ satisfies $\mathbf{U}_k^{(\delta)}(\mathbf{U}_k^{(\delta)})^\top = \mathbf{Id}_{R+1}$ since $\mathbf{E}_k^{(\delta)} = \mathbf{\Lambda}_k^{(\delta)}\mathbf{X}_k^{(\delta)}\mathbf{\Lambda}_k^{(\delta)}$ and $\mathbf{X}_k^{(\delta)} = (n\bar{\mu}_n(\delta))^{-1}\mathbf{D}_k^{(\delta)}(\mathbf{D}_k^{(\delta)})^\top$, thus the lemma. \square

Proof of (29). If $\bar{\mu}_n(\delta) = 0$, we have $\bar{f}_k^{(\delta)} = 0$ by definition and the result is obvious, thus we assume $\bar{\mu}_n(\delta) > 0$. Since $\lambda(\bar{\mathbf{X}}_k^{(\delta)}) \geq (n\bar{\mu}_n(\delta))^{-1/2} > 0$, the matrices $\bar{\mathbf{X}}_k^{(\delta)}$ and $\mathbf{\Lambda}_k^{(\delta)}$ are invertible thus $\mathbf{E}_k^{(\delta)}$ also is, and we have

$$\begin{aligned} |\bar{f}_k^{(\delta)}(x_k)| &= |\langle (\mathbf{\Lambda}_k^{(\delta)})^{-1}\bar{\theta}_k^{(\delta)}, e_1 \rangle| \\ &\leq \|(\mathbf{E}_k^{(\delta)})^{-1}\mathbf{\Lambda}_k^{(\delta)}\bar{\mathbf{X}}_k^{(\delta)}\bar{\theta}_k^{(\delta)}\| \\ &= \|(\mathbf{E}_k^{(\delta)})^{-1}\mathbf{\Lambda}_k^{(\delta)}\mathbf{Y}_k^{(\delta)}\|. \end{aligned}$$

Moreover, we have by the definition of $\bar{\mathbf{X}}_k^{(\delta)}$:

$$\|(\mathbf{E}_k^{(\delta)})^{-1}\| \leq \|(\mathbf{\Lambda}_k^{(\delta)})^{-1}\|^2 \|(\bar{\mathbf{X}}_k^{(\delta)})^{-1}\| \lesssim \|(\bar{\mathbf{X}}_k^{(\delta)})^{-1}\| = \lambda(\bar{\mathbf{X}}_k^{(\delta)})^{-1} \leq (n\bar{\mu}_n(\delta))^{1/2}.$$

Let us denote $\tilde{\mathbf{E}}_k^{(\delta)} := \mathbf{\Lambda}_k^{(\delta)}\mathbf{X}_k^{(\delta)}\mathbf{\Lambda}_k^{(\delta)}$. With the same notations as in the proof of Lemma 1, we have

$$\begin{aligned} |(\mathbf{\Lambda}_k^{(\delta)}\mathbf{Y}_k^{(\delta)})_p| &= \| \|\varphi_{kp}\|_\delta^{-1} \langle f, \varphi_{kp} \rangle_\delta + (\mathbf{\Lambda}_k^{(\delta)})_p \langle \xi, \varphi_{kp} \rangle_\delta \| \\ &\leq \|f\|_\infty + |((n\bar{\mu}_n(\delta))^{-1}\tilde{\mathbf{E}}_k^{(\delta)}(\tilde{\mathbf{E}}_k^{(\delta)})^{-1}\mathbf{\Lambda}_k^{(\delta)}\mathbf{D}_k^{(\delta)}\xi)_p| \\ &= \|f\|_\infty + |((n\bar{\mu}_n(\delta))^{-1/2}\tilde{\mathbf{E}}_k^{(\delta)}\mathbf{U}_k^{(\delta)}\xi)_p|, \end{aligned}$$

thus (29), since $\|\tilde{\mathbf{E}}_k^{(\delta)}\| \leq R+1$. \square

Proof of (30). Let us define $\mathbf{H}_k^{(\delta)} := \mathbf{\Lambda}_k^{(\delta)}\mathbf{X}_k^{(\delta)}$. On $\Omega_k(\delta')$, we have:

$$\begin{aligned} |\bar{f}_k^{(\delta)}(x_k) - \bar{f}_k^{(\delta')}(x_k)| &= |(\bar{\theta}_k^{(\delta)} - \bar{\theta}_k^{(\delta')})_0| \\ &\leq \|\mathbf{\Lambda}_k^{(\delta')}(\bar{\theta}_k^{(\delta)} - \bar{\theta}_k^{(\delta')})\|_\infty \\ &= \|(\mathbf{E}_k^{(\delta')})^{-1}\mathbf{H}_k^{(\delta')}(\bar{\theta}_k^{(\delta)} - \bar{\theta}_k^{(\delta')})\|_\infty \\ &\lesssim \lambda(\mathbf{E}_k^{(\delta')})^{-1} \|\mathbf{H}_k^{(\delta')}(\bar{\theta}_k^{(\delta)} - \bar{\theta}_k^{(\delta')})\|_\infty. \end{aligned}$$

Since on $\Omega_k(\delta')$, $(\mathbf{H}_k^{(\delta')}(\bar{\theta}_k^{(\delta)} - \bar{\theta}_k^{(\delta')}))_p = \langle \bar{f}_k^{(\delta)} - \bar{f}_k^{(\delta')}, \varphi_{kp} \rangle_{\delta'} / \|\varphi_{kp}\|_{\delta'}$, and since $\delta' \subset \delta$, we obtain (30) on $\mathcal{I}_k(\delta, \delta')$. \square

Proof of Lemma 2. We denote by $\mathbf{P}_k^{(\delta)}$ the projection onto $\text{Span}\{\varphi_{k0}, \dots, \varphi_{kR}\}$ with respect to the inner product $\langle \cdot, \cdot \rangle_\delta$. Note that on $\Omega_k(\delta)$, we have $\bar{f}_k^{(\delta)} = \mathbf{P}_k^{(\delta)}Y$. Let $\delta \in G_k$ and $\delta' \in G_k(\delta)$. In view of (17), we have on $\Omega_k(\delta)$ for any $\varphi = \varphi_{kp}$, $p \in \{0, \dots, R\}$:

$$\begin{aligned} \langle \bar{f}_k^{(\delta')} - \bar{f}_k^{(\delta)}, \varphi \rangle_{\delta'} &= \langle Y - \bar{f}_k^{(\delta)}, \varphi \rangle_{\delta'} \\ &= \langle f - \mathbf{P}_k^{(\delta)}Y, \varphi \rangle_{\delta'} + \langle \xi, \varphi \rangle_{\delta'} \\ &= A_k - B_k + C_k, \end{aligned}$$

where $A_k := \langle f - \mathbf{P}_k^{(\delta)}f, \varphi \rangle_{\delta'}$, $B_k := \sigma \langle \mathbf{P}_k^{(\delta)}\xi, \varphi \rangle_{\delta'}$ and $C_k := \sigma \langle \xi, \varphi \rangle_{\delta'}$. If f_k is the Taylor polynomial of f at x_k up to the order $\lfloor s \rfloor$, since $\delta' \subset \delta$ and $f \in H(s, L)$ we have:

$$|A_k| \leq \|\varphi\|_{\delta'} \|f - f_k + \mathbf{P}_k^{(\delta)}(f_k - f)\|_\delta \leq \|\varphi\|_{\delta'} \|f - f_k\|_\delta \lesssim \|\varphi\|_{\delta'} L |\delta|^s,$$

and using (31), we obtain

$$|A_k| \lesssim \|\varphi\|_{\delta'} \sigma \left(\frac{\log n}{n \bar{\mu}_n(\delta)} \right)^{1/2}.$$

Since $\mathbf{P}_k^{(\delta)}$ is an orthogonal projection, the variance of B_k is equal to

$$\begin{aligned} \sigma^2 \mathbf{E}_{f\mu} [\langle \mathbf{P}_k^{(\delta)}\xi, \varphi \rangle_{\delta'}^2 | \mathfrak{X}_n] &\leq \sigma^2 \|\varphi\|_{\delta'}^2 \mathbf{E}_{f\mu} [\|\mathbf{P}_k^{(\delta)}\xi\|_{\delta'}^2 | \mathfrak{X}_n] \\ &= \sigma^2 \|\varphi\|_{\delta'}^2 \text{Tr}(\mathbf{P}_k^{(\delta)}) / (n \bar{\mu}_n(\delta')), \end{aligned}$$

where $\text{Tr}(M)$ stands for the trace of a matrix M . Since $\mathbf{P}_k^{(\delta)}$ is the projection onto Pol_R , $\text{Tr}(\mathbf{P}_k^{(\delta)}) \leq R + 1$, and the variance of B_k is smaller than $\sigma^2 \|\varphi\|_{\delta'}^2 (R + 1) / (n \bar{\mu}_n(\delta'))$. Then,

$$\mathbf{E}_{f\mu} [(B + C)^2 | \mathfrak{X}_n] \leq \sigma^2 \|\varphi\|_{\delta'}^2 C_R^2 / (n \bar{\mu}_n(\delta')). \quad (32)$$

In view of the threshold choice (22), we have

$$\begin{aligned} &\{ |\langle \bar{f}_k^{(\delta')} - \bar{f}_k^{(\delta)}, \varphi \rangle_{\delta'}| > \|\varphi\|_{\delta'} T_n(\delta, \delta') \} \\ &\subset \left\{ \frac{\|\varphi\|_{\delta'}^{-1} |B_k + C_k|}{\sigma(n \bar{\mu}_n(\delta'))^{-1/2} C_R} > D (\log(n \bar{\mu}_n(\delta)))^{1/2} \right\}, \end{aligned}$$

and using (32) together with $\mathbf{P}[|N(0, 1)| > x] \leq \exp(-x^2/2)$ and $|G_k(\delta)| \leq (n \bar{\mu}_n(\delta))$, we obtain

$$\begin{aligned} \mathbf{P}_{f\mu} [T(\delta)^{\mathfrak{C}} | \mathfrak{X}_n] &\leq \sum_{\delta' \in G_k(\delta)} \sum_{p=0}^R \exp(-D^2 \log(n \bar{\mu}_n(\delta)) / 2) \\ &\leq (R + 1) (n \bar{\mu}_n(\delta))^{1-D^2/2}, \end{aligned}$$

which concludes the proof. \square

Construction of S_n . We construct an event $S_n \in \mathfrak{X}_n$ such that $\mu^n [S_n^{\mathfrak{C}}] = o(1)$ faster than any power of n , and such that on this event, $R_k \asymp r_n(x_k)$ and $\lambda(\mathbf{E}_k^{(\Delta_k)}) \geq \lambda_0$ uniformly for

any $k \in \{0, \dots, 2^J\}$. We need preliminary approximation results, linked with the approximation of μ by $\bar{\mu}_n$. The following deviation inequalities use Bernstein inequality for the sum of independent random variables, which is standard. We have

$$\mu^n \left[\left| \frac{\bar{\mu}_n(\delta)}{\mu(\delta)} - 1 \right| > \epsilon \right] \lesssim \exp(-\epsilon^2 n \mu(\delta)) \quad (33)$$

for any interval $\delta \subset [0, 1]$ and $\epsilon \in (0, 1)$. Let us define the events

$$D_{n,a}^{(\delta)}(x, \epsilon) := \left\{ \left| \frac{1}{\mu(\delta)} \int_{\delta} \left(\frac{\cdot - x}{|\delta|} \right)^a d\bar{\mu}_n - e_a(x, \mu) \right| \leq \epsilon \right\}$$

where $e_a(x, \mu) := (1 + (-1)^a)(\beta(x) + 1)/(a + \beta(x) + 1)$ (a is a natural integer) where we recall that $\beta(x)$ comes from Assumption **D** (if x is such that $\mu(x) > 0$ then $\beta(x) = 0$). Using together Bernstein inequality and the fact that

$$\frac{1}{\mu(\delta)} \int_{\delta} \left(\frac{t - x}{|\delta|} \right)^a \mu(t) dt \rightarrow e_a(x, \mu)$$

as $|\delta| \rightarrow 0$, we obtain

$$\mu^n \left[(D_{n,a}^{(\delta)}(x, \epsilon))^{\mathbb{G}} \right] \lesssim \exp(-\epsilon^2 n \mu(\delta)). \quad (34)$$

By definition (23) of G_k , we have $\Delta_k = [x_k - H_n(x_k), x_k + H_n(x_k)]$ where

$$H_n(x) := \operatorname{argmin}_{h \in [0,1]} \left\{ Lh^s \geq \sigma \left(\frac{\log n}{n \bar{\mu}_n([x - h, x + h])} \right)^{1/2} \right\} \quad (35)$$

is an approximation of $h_n(x)$ (see (2)). Since $\bar{\mu}_n$ is “close” to μ , these quantities are close to each other for any x . Indeed, if $\delta_n(x) := [x - h_n(x), x + h_n(x)]$ and $\Delta_n(x) := [x - H_n(x), x + H_n(x)]$ we have using together (35) and (2):

$$\{H_n(x) \leq (1 + \epsilon)h_n(x)\} = \left\{ \frac{\bar{\mu}_n[(1 + \epsilon)\delta_n(x)]}{\mu[\delta_n(x)]} \geq (1 - \epsilon)^{-2} \right\} \quad (36)$$

for any $\epsilon \in (0, 1)$, where $(1 + \epsilon)\delta_n(x) := [x - (1 + \epsilon)h_n(x), x + (1 + \epsilon)h_n(x)]$. Hence, for each $x = x_k$, the left hand side event of (36) has a probability that can be controlled under Assumption **D** by (33), and the same argument holds for $\{H_n(x) > (1 - \epsilon)h_n(x)\}$. Combining (33), (34) and (36), we obtain that the event

$$B_{n,a}(x, \epsilon) := \left\{ \left| \frac{1}{\bar{\mu}_n(\Delta_n(x))} \int_{\Delta_n(x)} \left(\frac{\cdot - x}{|\delta_n(x)|} \right)^a d\bar{\mu}_n - e_a(x, \mu) \right| \leq \epsilon \right\}$$

satisfies also (34) for n large enough. This proves that $(\mathbf{X}_k^{(\Delta_k)})_{p,q}$ and $(\mathbf{\Lambda}_k^{(\Delta_k)})_p$ are close to $e_{p+q}(x_k, \mu)$ and $e_{2p}(x_k, \mu)^{-1/2}$ respectively on the event

$$S_n := \bigcap_{a \in \{0, \dots, 2R\}} \bigcap_{k \in \{0, \dots, 2^J - 1\}} B_{n,a}(x_k, \epsilon).$$

Using the fact that $\lambda(M) = \inf_{\|x\|=1} x^\top M x$ for a symmetrical matrix M , where $\lambda(M)$ denotes the smallest eigenvalue of M , we can conclude that for n large enough,

$$\lambda(\mathbf{\Lambda}_k^{(\Delta_k)} \mathbf{X}_k^{(\Delta_k)} \mathbf{\Lambda}_k^{(\Delta_k)}) \gtrsim \min_{x \in [0,1]} \lambda(\mathbf{E}(x, \mu)) =: \lambda_0,$$

where $\mathbf{E}(x, \mu)$ has entries $(\mathbf{E}(x, \mu))_{p,q} = e_{p+q}(x, \mu) / (e_{2p}(x, \mu) e_{2q}(x, \mu))^{1/2}$. Since $\mathbf{E}(x, \mu)$ is definite positive for any $x \in [0, 1]$, we obtain that on S_n , $\lambda(\mathbf{X}_k^{(\Delta_k)}) \geq \lambda_1$ for some constant $\lambda_1 > 0$, thus $S_n \subset \Omega_n(\Delta_k)$ and $\lambda(\mathbf{E}_k^{(\Delta_k)}) \gtrsim \lambda_0$ uniformly for any $k \in \{0, \dots, 2^J - 1\}$, since $\mathbf{E}_k^{(\Delta_k)} = \mathbf{\Lambda}_k^{(\Delta_k)} \mathbf{X}_k^{(\Delta_k)} \mathbf{\Lambda}_k^{(\Delta_k)}$ on $\Omega_n(\Delta_k)$. Moreover, since $R_k = L H_n(x_k)^s$, using together (33) and (36), we obtain $R_k \asymp r_n(x_k)$ uniformly for $k \in \{0, \dots, 2^J - 1\}$. \square

Proof of Theorem 2. The main features of the proof are first, a reduction to the Bayesian risk over an hardest cubical subfamily of functions for the \mathbb{L}^∞ metrics, which is standard: see Korostelev (1993), Donoho (1994), Korostelev and Nussbaum (1999) and Bertin (2004), and the choice of rescaled hypothesis with design-adapted bandwidth $h_n(\cdot)$, necessary to achieve the rate $r_n(\cdot)$.

Let us consider $\varphi \in H(s, L; \mathbb{R})$ (the extension of $H(s, L)$ to the whole real line) with support $[-1, 1]$ and such that $\varphi(0) > 0$. We define

$$a := \min \left[1, \left(\frac{2}{\|\varphi\|_\infty^2} \left(\frac{1}{1 + 2s + \beta} - \alpha \right) \right)^{1/(2s)} \right]$$

and

$$\Xi_n := 2a(1 + 2^{1/(s - \lfloor s \rfloor)}) \sup_{x \in [0,1]} h_n(x),$$

where we recall that $\lfloor s \rfloor$ is the largest integer smaller than s . Note that (7) entails

$$\Xi_n \lesssim (\log n/n)^{1/(1+2s+\beta)}. \quad (37)$$

If $I_n = [c_n, d_n]$, we introduce $x_k := c_n + k \Xi_n$ for $k \in K_n := \{1, \dots, \lfloor |I_n| \Xi_n^{-1} \rfloor\}$, and denote for the sake of simplicity $h_k := h_n(x_k)$. We consider the family of functions

$$f(\cdot; \theta) := \sum_{k \in K_n} \theta_k f_k(\cdot), \quad f_k(\cdot) := L \alpha^s h_k^s \varphi\left(\frac{\cdot - x_k}{h_k}\right),$$

which belongs to $H(s, L)$ for any $\theta \in [-1, 1]^{|K_n|}$. Using Bernstein inequality, we can see that

$$\mathbf{H}_n := \bigcap_{k \in K_n} \left\{ \begin{array}{l} \bar{\mu}_n([x_k - h_k, x_k + h_k]) \\ \mu([x_k - h_k, x_k + h_k]) \end{array} \geq 1/2 \right\}$$

satisfies

$$\mu^n[\mathbf{H}_n] = 1 - o(1). \quad (38)$$

Let us introduce $b := c^s \varphi(0)$. For any distribution \mathbf{B} on $\Theta_n \subset [-1, 1]^{|K_n|}$, by a minoration of the minimax risk by the Bayesian risk, and since w is non-decreasing, the left hand side of (9) is smaller than

$$\begin{aligned} w(b) \inf_{\hat{\theta}} \int_{\Theta_n} \mathbf{P}_{\hat{\theta}}^n [\max_{k \in K_n} |\hat{\theta}_k - \theta_k| \geq 1] \mathbf{B}(d\theta) \\ \geq w(b) \int_{\mathfrak{X}_n} \inf_{\hat{\theta}} \int_{\Theta_n} \mathbf{P}_{\hat{\theta}}^n [\max_{k \in K_n} |\hat{\theta}_k - \theta_k| \geq 1 | \mathfrak{X}_n] \mathbf{B}(d\theta) d\mu^n. \end{aligned}$$

Hence, together with (38), Theorem 2 follows if we show that on \mathfrak{H}_n

$$\sup_{\hat{\theta}} \int_{\Theta_n} \mathbf{P}_{\hat{\theta}}^n [\max_{k \in K_n} |\hat{\theta}_k - \theta_k| < 1 | \mathfrak{X}_n] \mathbf{B}(d\theta) = o(1). \quad (39)$$

We denote by $L(\theta; Y_1, \dots, Y_n)$ the conditional on \mathfrak{X}_n likelihood function of the observations Y_i from (1) when $f(\cdot) = f(\cdot; \theta)$. Conditionally on \mathfrak{X}_n , we have

$$L(\theta; Y_1, \dots, Y_n) = \prod_{1 \leq i \leq n} g_{\sigma}(Y_i) \prod_{k \in K_n} \frac{g_{v_k}(y_k - \theta_k)}{g_{v_k}(y_k)},$$

where g_v is the density of $N(0, v^2)$, $v_k^2 := \mathbf{E}\{y_k^2 | \mathfrak{X}_n\}$ and

$$y_k := \frac{\sum_{i=1}^n Y_i f_k(X_i)}{\sum_{i=1}^n f_k^2(X_i)}.$$

Thus, choosing

$$\mathbf{B} := \bigotimes_{k \in K_n} \mathbf{b}, \quad \mathbf{b} := (\delta_{-1} + \delta_1)/2, \quad \Theta_n := \{-1, 1\}^{|K_n|},$$

the left hand side of (39) is smaller than

$$\int \frac{\prod_{1 \leq i \leq n} g_{\sigma}(Y_i)}{\prod_{k \in K_n} g_{v_k}(y_k)} \left(\prod_{k \in K_n} \sup_{\hat{\theta}_k} \int_{\{-1, 1\}} \mathbf{1}_{|\hat{\theta}_k - \theta_k| < 1} g_{v_k}(y_k - \theta_k) \mathbf{b}(d\theta_k) \right) dY_1 \times \dots \times dY_n,$$

and $\hat{\theta}_k = \mathbf{1}_{y_k \geq 0} - \mathbf{1}_{y_k < 0}$ are strategies reaching the supremum. Then, in (39), it suffices to take the supremum over estimators $\hat{\theta}$ with coordinates $\hat{\theta}_k \in \{-1, 1\}$ measurable with respect to y_k only. Since conditionally on \mathfrak{X}_n , y_k is in law $N(\theta_k, v_k^2)$, the left hand side of (39) is smaller than

$$\prod_{k \in K_n} \left(1 - \inf_{\hat{\theta}_k \in \{-1, 1\}} \int_{\{-1, 1\}} \int \mathbf{1}_{|\hat{\theta}_k(u) - \theta_k| \geq 1} g_{v_k}(u - \theta_k) du \mathbf{b}(d\theta_k) \right).$$

Moreover, if $\Phi(x) := \int_{-\infty}^x g_1(t) dt$

$$\begin{aligned} \inf_{\hat{\theta}_k \in \{-1, 1\}} \int_{\{-1, 1\}} \int \mathbf{1}_{|\hat{\theta}_k(u) - \theta_k| \geq 1} g_{v_k}(u - \theta_k) du \mathbf{b}(d\theta_k) \\ \geq \frac{1}{2} \int \min(g_{v_k}(u - 1), g_{v_k}(u + 1)) du = \Phi(-1/v_k). \end{aligned}$$

On H_n , we have in view of (2)

$$v_k^2 = \frac{\sigma^2}{\sum_{i=1}^n f_k^2(X_i)} \geq \frac{2}{(1-\delta)\|\varphi\|_\infty^2 c^{2s} \log n},$$

and since $\Phi(-x) \geq \exp(-x^2/2)(x\sqrt{2\pi})$ for any $x > 0$, we obtain

$$\Phi(-1/v_k) \gtrsim (\log n)^{-1/2} n^{\{\alpha-1/(1+2s+\beta)\}/2} =: L_n.$$

Thus, the left hand side of (39) is smaller than $(1-L_n)^{|K_n|}$, and since

$$|I_n| \Xi_n^{-1} L_n \gtrsim n^{\{1/(1+2s+\beta)-\alpha\}/2} (\log n)^{1/2-1/(1+2s+\beta)} \rightarrow +\infty$$

as $n \rightarrow +\infty$, Theorem 2 follows. \square

Proof of Corollary 1. Let us consider the loss function $w(\cdot) = |\cdot|$, and let \widehat{f}_n^v be an estimator converging with rate $v_n(\cdot)$ over F in the sense of (3). Hence,

$$\begin{aligned} 1 &\lesssim \sup_{f \in F} \mathbf{E}_{f\mu} \left[\sup_{x \in I_n} r_n(x)^{-1} |\widehat{f}_n^v(x) - f(x)| \right] \\ &\leq \sup_{x \in I_n} \frac{v_n(x)}{r_n(x)} \sup_{f \in F} \mathbf{E}_{f\mu} \left[\sup_{x \in I_n} v_n(x)^{-1} |\widehat{f}_n^v(x) - f(x)| \right] \lesssim \sup_{x \in I_n} \frac{v_n(x)}{r_n(x)}, \end{aligned}$$

where we used Theorem 2. \square

Proof of Proposition 1. Without loss of generality, we consider the loss $w(\cdot) = |\cdot|$. In order to prove Proposition 1, we use the linear LPE. If we denote by $\partial^m f$ the m -th derivative of f , a slight modification of the proof of Lemma 1 gives for $f \in H(s, L)$ with $s > m$,

$$|\partial^m \widehat{f}_k^{(\delta)}(x_k) - \partial^m f(x_k)| \lesssim \lambda(\mathbf{E}_k^{(\delta)})^{-1} |\delta|^{-m} (L|\delta|^s + \sigma(n\bar{\mu}_n(\delta)))^{-1/2} W^N,$$

where in the same way as in the proof of Theorem 1, W^N satisfies

$$\mathbf{E}_{f\mu}[W^N | \mathfrak{X}_n] \lesssim (\log N)^{1/2}, \quad (40)$$

with N depending on the size of the supremum, to be specified below. First, we prove a). Since $|I_n| \sim (\ell_n/n)^{1/(2s+1)}$, if $I_n = [a_n, b_n]$, the points

$$x_k := a_n + (k/n)^{1/(2s+1)}, \quad k \in \{0, \dots, N\},$$

where $N := \lfloor \ell_n \rfloor$ belong to I_n . We consider the bandwidth

$$h_n = \left(\frac{\log \ell_n}{n} \right)^{1/(2s+1)}, \quad (41)$$

and we take $\delta_k := [x_k - h_n, x_k + h_n]$. Note that since $\mu(x) > 0$ for any x , $\bar{\mu}_n(\delta) \asymp |\delta|$ as $|\delta| \rightarrow 0$ with probability going to 1 faster than any power of n (using Bernstein inequality,

for instance). We consider the estimator defined by

$$\widehat{f}_n(x) := \sum_{m=0}^r \partial^m \bar{f}_k^{(\delta_k)}(x_k)(x - x_k)^m/m! \quad \text{for } x \in [x_k, x_{k+1}), \quad k \in \{0, \dots, [\ell_n]\}, \quad (42)$$

where $r := \lfloor s \rfloor$. Using a Taylor expansion of f up to the degree r together with (41) gives

$$(n/\log n)^{s/(1+2s)} \sup_{x \in I_n} |\widehat{f}_n(x) - f(x)| \lesssim \left(\frac{\log \ell_n}{\log n} \right)^{s/(1+2s)} (1 + (\log \ell_n)^{-1/2} W^N).$$

Then, integrating with respect to $\mathbf{P}_{f\mu}(\cdot | \mathfrak{X}_n)$ and using (40) where $N = \lfloor \ell_n \rfloor$ entails a), since $\log \ell_n = o(\log n)$.

The proof of b) is similar to that of a). In this setting, the rate $r_n(\cdot)$ (see (2)) can be written as $r_n(x) = (\log n/n)^{\alpha_n(x)}$ for x in I_n (for n large enough) where $\alpha_n(x_0) = s/(1 + 2s + \beta)$ and $\alpha_n(x) > s/(1 + 2s + \beta)$ for $x \in I_n - \{x_0\}$. We define

$$x_{k+1} = \begin{cases} x_k + n^{-\alpha_n(x_k)/s} & \text{for } k \in \{-N, \dots, -1\} \\ x_k + n^{-\alpha_n(x_{k+1})/s} & \text{for } k \in \{0, \dots, N\}, \end{cases}$$

where $N := \lfloor \ell_n \rfloor$. All the points fit in I_n , since

$$|x_{-N} - x_N| \leq \sum_{-N \leq k \leq N} n^{-\min(\alpha_n(x_k), \alpha_n(x_{k+1}))/s} \leq 2(\ell_n/n)^{1/(1+2s+\beta)}.$$

We consider the bandwidths

$$h_k := (\log \ell_n/n)^{\alpha_n(x_k)/s},$$

and the intervals $\delta_k = [x_k - h_k, x_k + h_k]$. We keep the same definition (42) for \widehat{f}_n . Since x_0 is a local extremum of $r_n(\cdot)$, we have in the same way as in the proof of a) that

$$\begin{aligned} \sup_{x \in I_n} r_n(x)^{-1} |\widehat{f}_n(x) - f(x)| &\lesssim \left[\max_{-N \leq k \leq -1} \left(\frac{\log \ell_n}{\log n} \right)^{\alpha_n(x_k)} \right. \\ &\quad \left. + \max_{0 \leq k \leq N-1} \left(\frac{\log \ell_n}{\log n} \right)^{\alpha_n(x_{k+1})} \right] (1 + (\log \ell_n)^{-1/2} W^N), \end{aligned}$$

hence

$$\mathbf{E}_{f\mu} \left[\sup_{x \in I_n} r_n(x)^{-1} |\widehat{f}_n(x) - f(x)| \right] \lesssim \left(\frac{\log \ell_n}{\log n} \right)^{s/(1+2s+\beta)} = o(1),$$

which concludes the proof of Proposition 1. \square

REFERENCES

- ALMANSA, A., ROUGE, B. and JAFFARD, S. (2003). Irregular sampling in satellite images and reconstruction algorithms. In *CANUM 2003*. CANUM 2003, <http://www.math.univ-montp2.fr/canum03/communications/ms/andres.almansa.pdf>.
- ANTONIADIS, A. and FAN, J. Q. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, **96** 939–967.
- ANTONIADIS, A., GREGOIRE, G. and VIAL, P. (1997). Random design wavelet curve smoothing. *Statistics and Probability Letters*, **35** 225–232.

- ANTONIADIS, A. and PHAM, D. T. (1998). Wavelet regression for random or irregular design. *Comput. Statist. Data Anal.*, **28** 353–369.
- BARAUD, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.*, **6** 127–146 (electronic).
- BERTIN, K. (2004). Minimax exact constant in sup-norm for nonparametric regression with random design. *J. Statist. Plann. Inference*, **123** 225–242.
- BROWN, L. and CAI, T. (1998). Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics*, **26** 1783–1799.
- BROWN, L. D., CAI, T., LOW, M. G. and ZHANG, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of Statistics*, **30** 688 – 707.
- BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, **24** 2384–2398.
- CAI, T. T. and LOW, M. G. (2005). Nonparametric estimation over shrinking neighborhoods: superefficiency and adaptation. *Ann. Statist.*, **33** 184–213.
- CHESNEAU, C. (2007). Regression with random design: A minimax study. *Statistics and Probability Letters*, **77** 40–53.
- COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelets transforms. *Appl. Comput. Harmon. Anal.*, **1** 54–81.
- DELOUILLE, V. (2002). *Nonparametric stochastic regression using design-adapted wavelets*. Ph.D. thesis, Université catholique de Louvain.
- DELOUILLE, V., FRANKE, J. and VON SACHS, R. (2001). Nonparametric stochastic regression with design-adapted wavelets. *Sankhyā Ser. A*, **63** 328–366. Special issue on wavelets.
- DELOUILLE, V., SIMOENS, J. and VON SACHS, R. (2004). Smooth design-adapted wavelets for nonparametric stochastic regression. *Journal of the American Statistical Society*, **99** 643–658.
- DELYON, B. and JUDITSKY, A. (1995). Estimating wavelet coefficients. In *Lecture notes in Statistics* (A. Antoniadis and G. Oppenheim, eds.), vol. 103. Springer-Verlag, New York, 151–168.
- DONOHO, D. (1992). Interpolating wavelet transforms. Tech. rep., Department of Statistics, Stanford University, <http://www-stat.stanford.edu/~donoho/Reports/1992/interpol.ps.Z>.
- DONOHO, D. L. (1994). Asymptotic minimax risk for sup-norm loss: Solution via optimal recovery. *Probability Theory and Related Fields*, **99** 145–170.
- EFROMOVICH, S. (1999). *Nonparametric curve estimation*. Springer Series in Statistics, Springer-Verlag, New York. Methods, theory, and applications.
- FAN, J. and GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B. Methodological*, **57** 371–394.
- FAN, J. and GIJBELS, I. (1996). *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability, Chapman & Hall, London.
- FEICHTINGER, H. G. and GRÖCHENIG, K. (1994). Theory and practice of irregular sampling. In *Wavelets: mathematics and applications*. Stud. Adv. Math., CRC, Boca Raton, FL, 305–363.
- GAÏFFAS, S. (2005a). Convergence rates for pointwise curve estimation with a degenerate design. *Mathematical Methods of Statistics*, **1** 1–27. Available at <http://hal.ccsd.cnrs.fr/ccsd-00003086/en/>.
- GAÏFFAS, S. (2005b). On pointwise adaptive curve estimation based on inhomogeneous data. Preprint LPMA no 974 available at <http://hal.ccsd.cnrs.fr/ccsd-00004605/en/>.

- GOLDENSHLUGER, A. and NEMIROVSKI, A. (1997). On spatially adaptive estimation of nonparametric regression. *Mathematical Methods of Statistics*, **6** 135–170.
- GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric regression and generalized linear models*, vol. 58 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London. A roughness penalty approach.
- GUERRE, E. (1999). Efficient random rates for nonparametric regression under arbitrary designs. Personal communication.
- HALL, P., MARRON, J. S., NEUMANN, M. H. and TETTERINGTON, D. M. (1997). Curve estimation when the design density is low. *The Annals of Statistics*, **25** 756–770.
- HALL, P., PARK, B. U. and TURLACH, B. A. (1998). A note on design transformation and binning in nonparametric curve estimation. *Biometrika*, **85** 469–476.
- HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, approximation, and statistical applications*, vol. 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York.
- JANSEN, M., NASON, P. G. and SILVERMAN, B. W. (2004). Multivariate nonparametric regression using lifting. Tech. rep., University of Bristol, UK, <http://www.stats.ox.ac.uk/~silverma/pdf/jansennasonsilverman.pdf>.
- KERKYACHARIAN, G. and PICARD, D. (2004). Regression in random design and warped wavelets. *Bernoulli*, **10** 1053–1105.
- KOROSTELEV, A. and NUSSBAUM, M. (1999). The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli*, **5** 1099–1118.
- KOROSTELEV, V. (1993). An asymptotically minimax regression estimator in the uniform norm up to exact constant. *Theory of Probability and its Applications*, **38** 737–743.
- LEDoux, M. and TALAGRAND, M. (1991). *Probability in Banach spaces*, vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.
- LEPSKI, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, **35** 454–466.
- LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, **25** 929–947.
- LEPSKI, O. V. and SPOKOINY, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, **25** 2512–2546.
- MAXIM, V. (2003). *Restauration de signaux bruités sur des plans d'expérience aléatoires*. Ph.D. thesis, Université Joseph Fourier, Grenoble 1.
- PENSKY, M. and WIENS, D. P. (2001). On non-equally spaced wavelet regression. *Advances in Soviet Mathematics*, **53** 681–690.
- SPOKOINY, V. G. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *The Annals of Statistics*, **26** 1356–1378.
- STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **10** 1040–1053.
- VÀZQUEZ, C., KONRAD, J. and DUBOIS, E. (2000). Wavelet-based reconstruction of irregularly-sampled images: application to stereo imaging. In *Proc. Int. Conf. on Image Processing, ICIP-2000*. IEEE, <http://iss.bu.edu/jkonrad/Publications/local/cpapers/Vazq00icip.pdf>.

WONG, M.-Y. and ZHENG, Z. (2002). Wavelet threshold estimation of a regression function with random design. *Journal of Multivariate Analysis*, **80** 256–284.

Laboratoire de Statistique Théorique et Appliquée

Université Pierre et Marie Curie – Paris 6

175 rue du Chevaleret

75013 PARIS

E-mail: gaiffas@ccr.jussieu.fr