Nonparametric regression with martingale increment errors

Sylvain Delattre, Stéphane Gaïffas

Nonparametric regression with martingale increment errors

Sylvain Delattre[1], Stéphane Gaïffas[2]

October 29, 2010

Abstract

We consider the problem of adaptive estimation of the regression function in a framework where we replace ergodicity assumptions (such as independence or mixing) by another structural assumption on the model. Namely, we propose adaptive upper bounds for kernel estimators with data-driven bandwidth (Lepski's selection rule) in a regression model where the noise is an increment of martingale. It includes, as very particular cases, the usual i.i.d. regression and auto-regressive models. The cornerstone tool for this study is a new result for self-normalized martingales, called "stability", which is of independent interest. In a first part, we only use the martingale increment structure of the noise. We give an adaptive upper bound using a random rate, that involves the occupation time near the estimation point. Thanks to this approach, the theoretical study of the statistical procedure is disconnected from usual ergodicity properties like mixing. Then, in a second part, we make a link with the usual minimax theory of deterministic rates. Under a β-mixing assumption on the covariates process, we prove that the random rate considered in the first part is equivalent, with large probability, to a deterministic rate which is the usual minimax adaptive one.

Keywords. Nonparametric regression ; Adaptation ; Kernel estimation ; Lepski's method ; Self-normalized martingales ; Random rates ; Minimax rates ; β-Mixing.

1 Introduction

1.1 Motivations

In the theoretical study of statistical or learning algorithms, stationarity, ergodicity and concentration inequalities are assumptions and tools of first importance. When one wants to obtain asymptotic results for some procedure, stationarity and ergodicity of the random process generating the data is mandatory. Using extra assumptions, like moments and boundedness conditions, concentration inequalities can be used to

[1]Université Paris Diderot - Paris 7, Laboratoire de Probabilités et Modèles Aléatoires. *email*: delattre@math.jussieu.fr

[2]Université Pierre et Marie Curie - Paris 6, Laboratoire de Statistique Théorique et Appliquée. *email*: stephane.gaiffas@upmc.fr

[3]This work is supported in part by French Agence Nationale de la Recherche (ANR) ANR Grant "Prognostic" ANR-09-JCJC-0101-01. (http://www.lsta.upmc.fr/prognostic/index.php)

obtain finite sample results. Such tools are standard when the random process is assumed to be i.i.d., like Bernstein's or Talagrand's inequality (see [20], [31] and [28], among others). To go beyond independence, one can use a mixing assumption in order to "get back" independence using coupling, see [9], so that, roughly, the "independent data tools" can be used again. This approach is widely used in nonparametric statistics, statistical learning theory and time series analysis.

The aim of this paper is to replace stationarity and ergodicity assumptions (such as independence or mixing) by another structural assumption on the model. Namely, we consider a regression model where the noise is an increment of martingale. It includes, as very particular cases, the usual i.i.d. regression and the auto-regressive models. The cornerstone tool for this study is a new result, called "stability", for self-normalized martingales, which is of independent interest. In this framework, we study kernel estimators with a data-driven bandwidth, following the Lepski's selection rule, see [22], [24].

The Lepski's method is a statistical algorithm for the construction of optimal adaptive estimators. It was introduced in [21, 22, 23], and it provides a way to select the bandwidth of a kernel estimator from the data. It shares the same kind of adaptation properties to the inhomogeneous smoothness of a signal as wavelet thresholding rules, see [24]. It can be used to construct an adaptive estimator of a multivariate anisotropic signal, see [18], and recent developments shows that it can be used in more complex settings, like adaptation to the semi-parametric structure of the signal for dimension reduction, or the estimation of composite functions, see [13], [17]. In summary, it is commonly admitted that Lepski's idea for the selection of a smoothing parameter works for many problems. However, theoretical results for this procedure are mostly stated in the idealized model of Gaussian white noise, excepted for [12], where the model of regression with a random design was considered. As far as we know, nothing is known on this procedure in other settings: think for instance of the auto-regressive model or models with dependent data.

Our approach is in two parts: in a first part, we consider the problem of estimation of the regression function. We give an adaptive upper bound using a random rate, that involves the occupation time at the estimation point, see Theorem 1. In this first part, we only use the martingale increment structure of the noise, and not stationarity or ergodicity assumptions on the observations. Consequently, even if the underlying random process is transient (e.g. there are few observations at the estimation point), the result holds, but the occupation time is typically small, so that the random rate is large (and eventually not going to zero as the sample size increases). The key tool is a new result of stability for self-normalized martingales stated in Theorem 2, see Section 3. It works surprisingly well for the statistical application proposed here, but it might give new results for other problems as well, like upper bounds for procedures based on minimization of the empirical risk, model selection (see [26]), etc. In a second part (Section 4), we make a link with the usual minimax theory of deterministic rates. Using a β-mixing assumption, we prove that the random rate used in Section 2 is equivalent, with a large probability, to a deterministic rate which is the usual adaptive minimax one, see Proposition 1.

The message of this paper is twofold. First, we show that the kernel estimator and Lepski's method are very robust with respect to the statistical properties of the model: they does not require stationarity or ergodicity assumptions, such as independence or mixing to "do the job of adaptation", see Theorem 1. The second part of the message

is that, for the theoretical assessment of an estimator, one can use advantageously a theory involving random rates of convergence. Such a random rate naturally depends on the occupation time at the point of estimation (=the local amount of data), and it is "almost observable" if the smoothness of the regression were to be known. An ergodicity property, such as mixing, shall only be used in a second step of the theory, for the derivation of the asymptotic behaviour of this rate (see Section 4). Of course, the idea of random rates for the assessment of an estimator is not new. It has already been considered in [15, 14] for discrete time and in [8] for diffusion models. However, this work contains, as far as we know, the first result concerning adaptive estimation of the regression with a martingale increment noise.

1.2 The model

Consider sequences $(X_k)_{k\geq 0}$ and $(Y_k)_{k\geq 1}$ of random variables respectively in $\mathbb{R}^d$ and $\mathbb{R}$, both adapted to a filtration $(\mathscr{F}_k)_{k\geq 0}$, and such that for all $k \geq 1$:

$$Y_k = f(X_{k-1}) + \varepsilon_k, \tag{1}$$

where the sequence $(\varepsilon_k)_{k\geq 1}$ is a $(\mathscr{F}_k)$-martingale increment:

$$\mathbb{E}(|\varepsilon_k||\mathscr{F}_{k-1}) < \infty \ \text{ and } \ \mathbb{E}(\varepsilon_k|\mathscr{F}_{k-1}) = 0,$$

and where $f : \mathbb{R}^d \to \mathbb{R}$ is the unknown function of interest. We study the problem of estimation of f at a point $x \in \mathbb{R}^d$ based on the observation of $(Y_1, \ldots, Y_N)$ and $(X_0, \ldots, X_{N-1})$, where $N \geq 1$ is a finite $(\mathscr{F}_k)$-stopping time. This allows for "sample size designing", see Remark 1 below. The analysis is conducted under the following assumption on the sequence $(\varepsilon_k)_{k\geq 1}$:

Assumption 1. *There is a $(\mathscr{F}_k)$-adapted sequence $(\sigma_k)_{k\geq 0}$, assumed to be observed, of positive random variables and $\mu, \gamma > 0$ such that*:

$$\mathbb{E}\Big[\exp\Big(\mu\frac{\varepsilon_k^2}{\sigma_{k-1}^2}\Big) \mid \mathscr{F}_{k-1}\Big] \leq \gamma \quad \forall k \geq 1.$$

This assumption means that the martingale increment ε_k, normalized by σ_{k-1}, is uniformly subgaussian. In the case where ε_k is Gaussian conditionally to $\mathscr{F}_{k-1}$, Equation (1) is satisfied if (σ_k) is such that $\mathrm{Var}\,(\varepsilon_k|\mathscr{F}_{k-1}) \leq c\sigma_{k-1}^2$ for any $k \geq 0$, where $c > 0$ is a deterministic constant not depending on k. If one assumes that $\mathrm{Var}\,(\varepsilon_k|\mathscr{F}_{k-1}) \leq \bar{\sigma}^2$ for a known constant $\bar{\sigma} > 0$, one can take simply $\sigma_k \equiv \bar{\sigma}$. Note that σ_{k-1} is not necessarily the conditional variance of ε_k, but an observed upper bound of it.

Particular cases of model (1) are the regression and the auto-regressive model.

Example 1. In the regression model, one observes $(Y_k, X_{k-1})_{k=1}^n$ satisfying

$$Y_k = f(X_{k-1}) + s(X_{k-1})\zeta_k,$$

where (ζ_k) is i.i.d. centered, such that $\mathbb{E}(\exp(\mu\zeta_k^2)) \leq \gamma$ and independent of $\mathscr{F}_k = \sigma(X_0, \ldots, X_k)$, and where $f : \mathbb{R}^d \to \mathbb{R}$ and $s : \mathbb{R}^d \to \mathbb{R}^+$. This model is a particular case of (1) with $\sigma_k^2 \geq s(X_k)^2$.

Example 2. In the auto-regressive model, one observes a sequence $(X_k)_{k=0}^n$ in $\mathbb{R}^d$ satisfying

$$X_k = \vec{f}(X_{k-1}) + S(X_{k-1})\vec{\zeta_k}, \tag{2}$$

where $\vec{f} = (f_1, \dots, f_d) : \mathbb{R}^d \to \mathbb{R}^d$, where $S : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and where $\vec{\zeta_k} = (\zeta_{k,1}, \dots, \zeta_{k,d})$ is a sequence of centered i.i.d. vectors in $\mathbb{R}^d$ independent of X_0, with covariance matrix I_d and such that $\mathbb{E}(\exp(\mu\zeta_{k,j}^2)) \leq \gamma$. The problem of estimation of each coordinate f_j is a particular case of (1) with $Y_k = (X_k)_j$, $\mathscr{F}_k = \sigma(X_0, \vec{\zeta_1}, \dots, \vec{\zeta_k})$ and $\sigma_k^2 \geq S_{j,j}(X_k)^2$.

Let us mention that these two examples are very particular. The analysis conducted here allows to go way beyond the i.i.d. case, as long as (ζ_k) is a martingale increment.

Remark 1. The results given in Section 2 are stated in a setting where one observes $(X_{k-1}, Y_k)_{k=1}^N$ with N a stopping time. Of course, this contains the usual case $N \equiv n$, where n is a fixed sample size. This framework includes situations where the statistician decides to stop the sampling according to some design of experiment rule. This is the case when obtaining data has a cost, that cannot be more than a maximum value, for instance.

Remark 2. Note that while $\zeta_k = \varepsilon_k/\sigma_{k-1}$ is conditionally subgaussian, ε_k is not in general, (see [6] for examples).

1.3 The Lepski's method

In what follows, $|x|$ stands for the Euclidean norm of $x \in \mathbb{R}^d$. An object of importance in the analysis conducted below is the following. For $h > 0$, we define

$$L(h) = \sum_{k=1}^{N} \frac{1}{\sigma_{k-1}^2} \mathbf{1}_{|X_{k-1}-x|\leq h},$$

which is the occupation time of $(X_k)_{k\geq 0}$ at x renormalized by (σ_k). Then, if h is such that $L(h) > 0$ (there is at least one observations in the interval $[x-h, x+h]$), we define the kernel estimator

$$\hat{f}(h) = \frac{1}{L(h)} \sum_{k=1}^{N} \frac{1}{\sigma_{k-1}^2} \mathbf{1}_{|X_{i-1}-x|\leq h} Y_k.$$

Let $(h_i)_{i\geq 0}$ be a decreasing sequence of positive numbers, called *bandwidths*, and define the following set, called *grid*, as

$$\mathcal{H} := \{h_j : L(h_j) > 0\}.$$

For the sake of simplicity, we will consider only on a geometrical grid, where

$$h_j = h_0 q^j$$

for some parameters $h_0 > 0$ and $q \in (0, 1)$. The Lepski's method selects one of the bandwidths in $\mathcal{H}$. Let $b > 0$ and for any $h > 0$, define

$$\psi(h) := 1 + b\log(h_0/h).$$

For $u > 0$, define, on the event $\{L(h_0)^{-1/2} \leq u\}$, the bandwidth

$$H_u = \min\Big\{h \in \mathcal{H} : \Big(\frac{\psi(h)}{L(h)}\Big)^{1/2} \leq u\Big\}, \tag{3}$$

and let $u_0 > 0$. The estimator of $f(x)$ is $\hat{f}(\hat{H})$ defined on the set $\{L(h_0)^{-1/2} \leq u_0\}$, where $\hat{H}$ is selected according to the following rule:

$$\hat{H} := \max\Big\{h \in \mathcal{H} : h \geq H_{u_0} \text{ and } \forall h' \in [H_{u_0}, h] \cap \mathcal{H}, \\ |\hat{f}(h) - \hat{f}(h')| \leq \nu\Big(\frac{\psi(h')}{L(h')}\Big)^{1/2}\Big\}, \tag{4}$$

where ν is a positive constant. This is the standard Lepski's procedure, see [22, 23, 24, 25]. In the next Section, we give an upper bound for $\hat{f}(\hat{H})$, with a normalization (convergence rate) that involves $L(h)$. This result is stated without any further assumptions on the model.

Remark 3. The number u_0 is a fixed constant such that the largest bandwidth h_0 in the grid satisfies $L(h_0)^{-1/2} \leq u_0$. This deterministic constraint is very mild: if we have some data close to x, and if h_0 is large enough (this is the largest bandwidth in the grid), then $L(h_0)$ should be large, at least such that $L(h_0)^{-1/2} \leq u_0$. Consider the following basic example: $X_k \in [-1, 1]^d$ almost surely for any k and $\sigma_k \equiv 1$, then by taking $h_0 = \sqrt{d}$ and $u_0 = 1$ the event $\{L(h_0)^{-1/2} \leq u_0\}$ has probability one. In Section 4 (see Proposition 1) we prove that a mixing assumption on $(X_k)_{k\geq 0}$ entails that this event has an overwhelming probability.

2 Adaptive upper bound

The usual way of stating an adaptive upper bound for $\hat{f}(\hat{H})$, see for instance [24], is to prove that it has the same convergence rate as the *oracle* estimator $\hat{f}(H^*)$, which is the "best" among a collection $\{\hat{f}(h) : h \in \mathcal{H}\}$. The oracle bandwidth H^* realizes a bias-variance trade-off, that involves explicitly the unknown f. For $h \in \mathcal{H}$ define

$$\widetilde{f}(h) := \frac{1}{L(h)} \sum_{k=1}^{N} \frac{1}{\sigma_{k-1}^2} \mathbf{1}_{|X_{k-1}-x|\leq h} f(X_{k-1}). \tag{5}$$

Consider a family of non-negative random variables $(W(h); h \in \mathcal{H})$ that bounds from above the local smoothness of f (measured by its increments):

$$\sup_{h' \in [H_{u_0}, h] \cap \mathcal{H}} \big|\widetilde{f}(h') - f(x)\big| \leq W(h), \quad \forall h \in \mathcal{H}. \tag{6}$$

Nothing is required on $(W(h) : h \in \mathcal{H})$ for the moment, one can perfectly choose it as the left hand side of (6) for each $h \in \mathcal{H}$ for instance. However, for the analysis conducted here, we need to bound W from below and above (see Remark 5): introduce

$$\bar{W}(h) := [W(h) \vee (\delta_0 (h/h_0)^{\alpha_0})] \wedge u_0, \tag{7}$$

where δ_0 and α_0 are positive constants. On the set

$$\big\{L(h_0)^{-1/2} \leq \bar{W}(h_0)\big\},$$

define the random *oracle* bandwidth

$$H^* := \min\Big\{h \in \mathcal{H} : \Big(\frac{\psi(h)}{L(h)}\Big)^{1/2} \le \bar{W}(h)\Big\}, \tag{8}$$

and consider the event

$$\Omega' := \big\{L(h_0)^{-1/2} \le \bar{W}(h_0), W(H^*) \le u_0\big\}.$$

The event Ω' is the "minimal" requirement for the proof of an upper bound for $\hat{f}(\hat{H})$, see Remarks 5 and 6 below.

Theorem 1. *Let Assumption 1 hold and let $\hat{f}(\hat{H})$ be the procedure given by the Lepski's rule (4). Then, for any $\rho \in (0, b\mu\nu^2/(64\alpha_0(1+\gamma)))$, we have*

$$\mathbb{P}\Big[\Big\{|\hat{f}(\hat{H}) - f(x)| \ge t\bar{W}(H^*)\Big\} \cap \Omega'\Big] \le C_0 \frac{(\log(t+1))^{1+\rho/2}}{t^\rho}$$

for any $t \ge t_0$, where $C_0, t_0 > 0$ are constants depending on $\rho, \mu, \gamma, q, b, u_0, \delta_0, \alpha_0, \nu$.

The striking fact in this Theorem is that we don't use any stationarity, ergodicity or concentration property. In particular, we cannot give at this point the behaviour of the random normalization $\bar{W}(H^*)$. It does not go to 0 in probability with $N \to +\infty$ when $L(h_0)$ does not go to $+\infty$ in probability, which happens if $(X_k)_{k\ge 0}$ is a transient Markov chain for instance. Hence, without any further assumption, Theorem 1 does not entail that $\hat{f}(\hat{H})$ is close to $f(x)$. On the other hand, when $(X_k)_{k\ge 0}$ is mixing, we prove that $\bar{W}(H^*)$ behaves as the deterministic minimax optimal rate, see Section 4. The cornerstone of the proof of this Theorem is a new result concerning the stability of self-normalized martingales, see Theorem 2 in Section 3 below.

Remark 4. The parameter ρ of decay of the probability in Theorem 1 is increasing with the threshold parameter ν from (4). So, for any $p > 0$ and ν large enough, Theorem 1 entails that the expectation of $(\bar{W}(H^*)^{-1}|\hat{f}(\hat{H}) - f(x)|)^p \mathbf{1}_{\Omega'}$ is finite.

Remark 5. The definition of $\bar{W}$ is related to the fact that since nothing is required on the sequence (X_k), the occupation time $L(h)$ can be small, even if h is large. In particular, $L(h)$ has no reason to be close to its expectation. So, without the introduction of $\bar{W}$ above, that bounds from below W by a power function, we cannot give a lower estimate of H^* (even rough), which is mandatory for the proof of Theorem 1.

Remark 6. On the event Ω', we have $\{L(h_0)^{-1/2} \le \bar{W}(h_0)\}$, meaning that the bandwidth h_0 (the largest in $\mathcal{H}$) is large enough to contain enough points in $[x-h_0, x+h_0]$, so that $L(h_0) \ge \bar{W}(h_0)^2$. This is not a restriction when $W(h) = Lh^s$ [f has a local Hölder exponent s] for instance, see Section 4.

Remark 7. In the definition of $\hat{f}(\hat{H})$, we use kernel estimation with the rectangular kernel $K(x) = \mathbf{1}_{[-1,1]}(x)/2$. This is mainly for technical simplicity, since the proof of Theorem 1 is already technically involved. Consequently, Theorem 1 does not give, on particular cases (see Section 4), the adaptive minimax rate of convergence for regression functions with an Hölder exponent s larger than 1. To improve this, one can consider the Lepski's method applied to local polynomials (LP) (see [12], and see [10] about (LP)). This would lead, in the framework considered here, to strong technical difficulties.

3 Stability for self-normalized martingales

We consider a local martingale $(M_n)_{n\in\mathbb{N}}$ with respect to a filtration $(\mathcal{G}_n)_{n\in\mathbb{N}}$, and for $n \geq 1$ denote its increment by $\Delta M_n := M_n - M_{n-1}$. The predictable quadratic variation of M_n is

$$\langle M\rangle_n := \sum_{k=1}^{n} \mathbb{E}[\Delta M_k^2|\mathcal{G}_{k-1}].$$

Concentration inequalities for martingales have a long history. The first ones are the Azuma-Hoeffding's inequality (see [1], [16]) and the Freedman's inequality (see [11]). The latter states that, if (M_n) is a square integrable martingale such that $|\Delta M_k| \leq c$ a.s. for some constant $c > 0$ and $M_0 = 0$, then for any $x, y > 0$:

$$\mathbb{P}[M_n \geq x, \langle M\rangle_n \leq y] \leq \exp\Big(-\frac{x^2}{2(y+cx)}\Big). \tag{9}$$

Later on, an alternative to the assumption $|\Delta M_k| \leq c$ was proposed. This is the so-called Bernstein's condition, which requires that there is some constant $c > 0$ such that for any $p \geq 2$:

$$\sum_{k=1}^{n} \mathbb{E}\big[|\Delta M_k|^p|\mathcal{G}_{k-1}\big] \leq \frac{p!}{2}c^{p-2}\langle M\rangle_n, \tag{10}$$

see [7], and [27]. In [30] (see Chapter 8), inequality (9) is proved with $\langle M\rangle_n$ replaced by a $\mathcal{G}_{n-1}$-measurable random variable nR_n^2, under the assumption that

$$\sum_{k=1}^{n} \mathbb{E}\big[|\Delta M_k|^p|\mathcal{G}_{k-1}\big] \leq \frac{p!}{2}c^{p-2}nR_n^2 \tag{11}$$

holds for any $p \geq 2$. There are many other very recent deviation inequalities for martingales, in particular inequalities involving the quadratic variation $[M]_n = \sum_{k=1}^{n} \Delta M_k^2$, see for instance [7] and [4].

For the proof of Theorem 1, a Bernstein's type of inequality is not enough: note that in (9), it is mandatory to work on the event $\{\langle M\rangle_n \leq y\}$. A control of the probability of this event usually requires an extra assumption on $(X_k)_{k\geq 0}$, such as independence or mixing (see Section 4), and this is precisely what we wanted to avoid here. Moreover, for the proof of Theorem 1, we need a result concerning M_T, where T is an arbitrary finite stopping-time.

In order to tackle this problem, a first idea is to try to give a deviation for the self-normalized martingale $M_T/\sqrt{\langle M\rangle_T}$. It is well-known that this is not possible, a very simple example is given in Remark 8 below. In the next Theorem 2, we give a simple solution to this problem. Instead of $M_T/\sqrt{\langle M\rangle_T}$, we consider $\sqrt{a}M_T/(a + \langle M\rangle_T)$, where $a > 0$ is an arbitrary real number, and we prove that the exponential moments of this random variable are uniformly bounded under Assumption 2 below. The result stated in Theorem 2 is of independent interest, and we believe that it can be useful for other statistical problems.

Assumption 2. *Assume that* $M_0 = 0$ *and that*

$$\Delta M_n = s_{n-1}\zeta_n \tag{12}$$

for any $n \geq 1$, where $(s_n)_{n\in\mathbb{N}}$ is a $(\mathcal{G}_n)$-adapted sequence of random variables and $(\zeta_n)_{n\geq 1}$ is a sequence of $(\mathcal{G}_n)$-martingale increments such that for $\alpha = 1$ or $\alpha = 2$ and some $\mu > 0, \gamma > 1$:

$$\mathbb{E}\big[\exp(\mu|\zeta_k|^\alpha)|\mathcal{G}_{k-1}\big] \leq \gamma \quad \textit{for any} \quad k \geq 1. \tag{13}$$

Let us define

$$V_n := \sum_{k=1}^{n} s_{k-1}^2.$$

Note that if $(\zeta_n)_{n\geq 1}$ is a conditionally normalized sequence (ie $\mathbb{E}(\zeta_n^2|\mathcal{G}_{n-1}) = 1$) then (12) entails that $V_n = \langle M\rangle_n$. Moreover, if Assumption 2 holds, we have $\langle M\rangle_n \leq c_\mu V_n$ for any $n \geq 1$ with $c_\mu = \ln 2/\mu$ when $\alpha = 2$ and $c_\mu = 2/\mu^2$ when $\alpha = 1$. Denote $\cosh(x) = (e^x + e^{-x})/2$ for any $x \in \mathbb{R}$.

Theorem 2. *Let Assumption 2 holds.*
• *If $\alpha = 2$, we have for any $\lambda \in [0, \frac{\mu}{2(1+\gamma)})$, any $a > 0$ and any finite stopping-time T:*

$$\mathbb{E}\Big[\exp\Big(\lambda \frac{aM_T^2}{(a+V_T)^2}\Big)\Big] \leq 1 + c_\lambda, \tag{14}$$

where $c_\lambda := \exp\big(\frac{\lambda\Gamma_\lambda}{2(1-2\lambda\Gamma_\lambda)}\big)(\exp(\lambda\Gamma_\lambda) - 1)$ and $\Gamma_\lambda := \frac{1+2\gamma}{2(\mu-\lambda)}$.
• *If $\alpha = 1$, we have for any $\lambda \in (-\mu, \mu)$, any $a > 0$ and any finite stopping-time T:*

$$\mathbb{E}\Big[\cosh\Big(\lambda \frac{\sqrt{a}M_T}{a+V_T}\Big)\Big] \leq 1 + c'_\lambda, \tag{15}$$

where $c'_\lambda = (\gamma-1)\lambda^2 \exp\big((\gamma-1)\lambda^2/\mu^2\big)\cosh\big(2\log 2 + 2(\gamma-1)\lambda^2/\mu^2\big)/\mu^2$.

The proof of Theorem 2 is given in Section 5. Theorem 2 shows that when ζ_k is subgaussian (resp. sub-exponential) conditionally to $\mathcal{G}_{k-1}$, then $\sqrt{a}|M_T|/(a+V_T)$ is also subgaussian (resp. sub-exponential), hence the name *stability*. Indeed, we cannot expect an improvement in the tails of $\sqrt{a}|M_T|/(a+V_T)$ due to the summation, since the s_{k-1} are arbitrary (for instance, it can be equal to zero for every k excepted for one).

Remark 8. It is tempting to take "$a = V_T$" in Theorem 2. However, the following basic example shows that it is not possible. Take $(B_t)_{t\geq 0}$ a standard Brownian motion, consider $M_n = B_n$ and define the stopping time $T_c = \inf\{n \geq 1 : B_n/\sqrt{n} \geq c\}$, where $c > 0$. For any $c > 0$, T_c is finite a.s. (use the law of iterated logarithm for instance). So, in this example, one has $M_{T_c}/\sqrt{\langle M\rangle_{T_c}} = M_{T_c}/\sqrt{T_c} \geq c$, for any $c > 0$.

4 Consistency with the minimax theory of deterministic rates

In this Section, we prove that, when $(X_k)_{k\geq 0}$ is mixing, then Theorem 1 gives the adaptive minimax upper bound. Let us consider again sequences $(X_k)_{k\geq 0}$ and $(Y_k)_{k\geq 1}$ of random variables satisfying (1), where $(\varepsilon_k)_{k\geq 0}$ an $(\mathscr{F}_k)_{k\geq 0}$-martingale increment. For the sake of simplicity, we work under the following simplified version of Assumption 1.

Assumption 3. *There is a known $\sigma > 0$ and $\mu, \gamma > 0$ such that*:

$$\mathbb{E}\Big[\exp\Big(\mu\frac{\varepsilon_k^2}{\sigma^2}\Big) \mid \mathscr{F}_{k-1}\Big] \leq \gamma \quad \forall k \geq 1.$$

Moreover, we consider the setting where we observe $(Y_1, \ldots, Y_n)$ and $(X_0, \ldots, X_{n-1})$, namely the stopping-time N is simply equal to n (the results in this section are proved for n large enough). Note that in this setting, we have $L(h) = \sigma^{-2} \sum_{k=1}^{n} \mathbf{1}_{|x_{k-1}-x|\leq h}$. We assume also that $(X_k)_{k\geq 0}$ is a strictly stationary sequence.

4.1 Some preliminaries

A function $\ell : \mathbb{R}^+ \to \mathbb{R}^+$ is slowly varying if it is continuous and if

$$\lim_{h\to 0^+} \ell(yh)/\ell(h) = 1, \quad \forall y > 0.$$

Fix $\tau \in \mathbb{R}$. A function $g : \mathbb{R}^+ \to \mathbb{R}^+$ is τ-regularly varying if $g(y) = y^\tau \ell(y)$ for some slowly varying ℓ. Regular variation is a standard and useful notion, of importance in extreme values theory for instance. We refer to [5] on this topic.

Below we will use the notion of β-mixing to measure the dependence of the sequence $(X_k)_{k\geq 0}$. This measure of dependence was introduced by Kolmogorov, see [19], and we refer to [9] for topics on dependence. Introduce the σ-field $\mathscr{X}_u^v = \sigma(X_k : u \leq k \leq v)$, where u, k, v are integers. A strictly stationary process $(X_k)_{k\in\mathbb{Z}}$ is called β-mixing or absolutely regular if

$$\beta_q := \frac{1}{2} \sup \Big(\sum_{i=1}^{I} \sum_{j=1}^{J} \Big| \mathbb{P}[U_i \cap V_j] - \mathbb{P}[U_i]\mathbb{P}[V_j] \Big| \Big) \to 0 \text{ as } q \to +\infty, \tag{16}$$

where the supremum is taken among all finite partitions $(U_i)_{i=1}^I$ and $(V_j)_{j=1}^J$ of Ω that are, respectively, $\mathscr{X}_{-\infty}^0$ and $\mathscr{X}_q^{+\infty}$ measurable. This notion of dependence is convenient in statistics because of a coupling result by Berbee, see [3], that allows to construct, among β-mixing observations, independent blocks, on which one can use Bernstein's or Talagrand's inequality (for a supremum) for instance. This strategy has been adopted in a series of papers dealing with dependent data, see [32, 2, 29] among others. In this section, we use this approach to give a deterministic equivalent to the random rate used in Section 2. This allows to prove that Theorem 1 is consistent with the usual minimax theory of deterministic rates, when one assumes that the sequence $(X_k)_{k\geq 0}$ is β-mixing.

4.2 Deterministic rates

We assume that f has Hölder-type smoothness in a neighbourhood of x. Let us fix two constants $\delta_0, u_0 > 0$ and recall that h_0 is the maximum bandwidth used in the Lepski's procedure (see Section 1.3).

Assumption 4 (Smoothness of f)**.** *There is $0 < s \leq 1$ and a slowly varying function ℓ_w such that the following holds:*

$$\sup_{y:|y-x|\leq h} |f(y) - f(x)| \leq w(h), \text{ where } w(h) := h^s \ell_w(h)$$

for any $h \leq h_0$, w is increasing on $[0, h_0]$, $w(h) \geq \delta_0(h/h_0)^2$ and $w(h) \leq u_0$ for any $h \in [0, h_0]$.

This is slightly more general than an Hölder assumption because of the slowly varying term ℓ_w. The usual Hölder assumption is recovered by taking $\ell_w \equiv r$, where $r > 0$ is some constant (the radius in Hölder smoothness).

Under Assumption 4, one has that

$$\sup_{h' \in [H_{u_0}, h] \cap \mathcal{H}} |\widetilde{f}(h') - f(x)| \leq w(h) \quad \forall h \in \mathcal{H}.$$

Under this assumption, one can replace $\bar{W}$ by w in the statement of Theorem 1 and from the definition of the oracle bandwidth H^* (see (8)). An oracle bandwidth related to the modulus of continuity w can be defined in the following way: on the event

$$\Omega_0 = \{L(h_0)^{-1/2} \leq w(h_0)\},$$

let us define

$$H_w := \min \Big\{ h \in]0, h_0] : \Big(\frac{\psi(h)}{L(h)}\Big)^{1/2} \leq w(h) \Big\}. \tag{17}$$

Under some ergodicity condition (using β-mixing) on $(X_k)_{k \geq 0}$, we are able to give a deterministic equivalent to $w(H_w)$. Indeed, in this situation, the occupation time $L(h)$ concentrates around its expectation $\mathbb{E}L(h)$, so a natural deterministic equivalent to (17) is given by

$$h_w := \min \Big\{ h \in]0, h_0] : \Big(\frac{\psi(h)}{\mathbb{E}L(h)}\Big)^{1/2} \leq w(h) \Big\}. \tag{18}$$

Note that h_w is well defined and unique when $(\mathbb{E}L(h_0))^{-1/2} \leq w(h_0)$, ie when $n \geq \sigma^2/(P_X([x-h_0, x+h_0])w(h_0)^2)$, where P_X stands for the distribution of X_0. We are able to give the behaviour of h_w under the following assumption.

Assumption 5 (Local behaviour of P_X)**.** *There is $\tau \geq -1$ and a slowly varying function ℓ_X such that*

$$P_X([x-h, x+h]) = h^{\tau+1} \ell_X(h) \quad \forall h \leq h_0.$$

This is an extension of the usual assumption on P_X which requires that it has a continuous density f_X wrt the Lebesgue measure such that $f_X(x) > 0$ (see also [12]). It is met when $f_X(y) = c|y-x|^\tau$ for y close to x for instance (in this case ℓ_X is constant).

Lemma 1. *Grant Assumptions 4 and 5. Then h_w is well defined by (18) and unique when n is large enough and such that*

$$h_w = (\sigma^2/n)^{1/(2s+\tau+1)} \ell_1(\sigma^2/n) \text{ and } w(h_w) = (\sigma^2/n)^{s/(2s+\tau+1)} \ell_2(\sigma^2/n),$$

where ℓ_1 and ℓ_2 are slowly varying functions that depend on s, τ and ℓ_X, ℓ_w.

The proof of this lemma easily follows from basic properties of regularly varying functions, so it is omitted. Explicit examples of such rates are given in [12]. Note that in the i.i.d. regression setting, we know from [12] that $w(h_w)$ is the minimax adaptive rate of convergence. Now, under the following mixing assumption, we can prove that the random rate $w(H_w)$ and the deterministic rate $w(h_w)$ have the same order of magnitude with a large probability.

Assumption 6. *Let $(\beta_q)_{q\geq 1}$ be the sequence of β-mixing coefficients of $(X_k)_{k\geq 0}$, see (16), and let $\eta, \kappa > 0$. We assume that for any $q \geq 1$:*

$$\beta_q \leq \frac{1}{\psi^{-1}(2q)},$$

where $\psi(u) = \eta(\log u)^\kappa$ (geometric mixing) or $\psi(u) = \eta u^\kappa$ (arithmetic mixing).

Proposition 1. *Let Assumptions 4, 5 and 6 hold. On Ω_0, let H_w be given by (17) and let (for n large enough) h_w be given by (18). Then, if (X_k) is geometrically β-mixing, or if it is arithmetically β-mixing with a constant $\kappa < 2s/(\tau+1)$, we have*

$$\mathbb{P}\Big[\Big\{\frac{w(h_w)}{4} \leq w(H_w) \leq 4w(h_w)\Big\} \cap \Omega_0\Big] \geq 1 - \varphi_n \quad \text{and} \quad \mathbb{P}[\Omega_0^\complement] = o(\varphi_n)$$

for n large enough, where in the geometrically β-mixing case:

$$\varphi_n = \exp(-C_1 n^{\delta_1} \ell_1(1/n)) \text{ where } \delta_1 = \frac{2s}{(2s+\tau+1)(\kappa+1)}$$

and in the arithmetically β-mixing case:

$$\varphi_n = C_2 n^{-\delta_2} \ell_2(1/n) \text{ where } \delta_2 = \frac{2s}{2s+\tau+1}\Big(\frac{1}{\kappa} - \frac{\tau+1}{2s}\Big),$$

where C_1, C_2 are positive constants and ℓ_1, ℓ_2 are slowly varying functions that depends on $\eta, \kappa, \tau, s, \sigma$ and ℓ_X, ℓ_w.

The proof of Proposition 1 is given in Section 5 below. The assumption used in Proposition 1 allows a geometric β-mixing, or an arithmetic β-mixing, up to a certain order, for the sequence (X_k). This kind of restriction on the coefficient of arithmetic mixing is standard, see for instance [29, 32, 2].

The next result is a direct corollary of Theorem 1 and Proposition 1. It says that when $(X_k)_{k\geq 0}$ is mixing, then the deterministic rate $w(h_w)$ is an upper bound for the risk of $\hat{f}(\hat{H})$.

Corollary 1. *Let Assumptions 3, 4 and 5 hold. Let Assumption 6 hold, with the extra assumption that $\kappa < 2s/(s+\tau+1)$ in the arithmetical β-mixing case. Moreover, assume that $|f(x)| \leq Q$ for some known constant $Q > 0$. Let us fix $p > 0$. If $\nu > 0$ satisfies $b\mu\nu^2 > 128p(1+\tau)$ (recall that ν is the constant in front the threshold in the Lepski's procedure, see (4)) then we have*

$$\mathbb{E}[|\widetilde{f}(\hat{H}) - f(x)|^p] \leq C_1 w(h_w)^p$$

for n large enough, where $\widetilde{f}(\hat{H}) = -Q \vee \hat{f}(\hat{H}) \wedge Q$ and where $C_1 > 0$ depends on $q, p, s, \mu, \gamma, b, u_0, \delta_0, \nu, Q$.

The proof of Corollary 1 is given in Section 5 below. Let us recall that in the i.i.d regression model with gaussian noise, we know from [12] that $w(h_w)$ is the minimax adaptive rate of convergence. So, Corollary 1 proves that Theorem 1 is consistent with the minimax theory of deterministic rates, when (X_k) is β-mixing.

Example 3. Assume that f is s-Hölder, ie Assumption 4 holds with $w(h) = Lh^s$ so $\ell_w(h) \equiv L$ and assume that P_X has a density f_X which is continuous and bounded away from zero on $[x - h_0, x + h_0]$, so that Assumption 5 is satisfied with $\tau = 0$. In this setting, one easily obtains that $w(h_w)$ is equal (up to some constant) to $(\log n/n)^{s/(2s+1)}$, which is the pointwise minimax adaptive rate of convergence, see [25, 23, 24] for the white-noise model and [12] for the regression model.

5 Proof of the main results

5.1 Proof of Theorem 2 for $\alpha = 2$

Let $a > 0$ and $\lambda \in [0, \frac{\mu}{2(1+\gamma)})$. Define $Y_0 := 0$ and for $n \geq 1$:

$$Y_n := \frac{aM_n^2}{(a+V_n)^2} \quad \text{and} \quad H_n := \mathbb{E}\big[\exp\big(\lambda(Y_n - Y_{n-1})\big) \mid \mathcal{G}_{n-1}\big].$$

Assume for the moment that H_n is finite a.s, hence we can define the local martingale

$$S_n := \sum_{k=1}^{n} e^{\lambda Y_{k-1}} \big(e^{\lambda(Y_k - Y_{k-1})} - H_k\big),$$

so that

$$\begin{aligned}\exp(\lambda Y_n) &= 1 + \sum_{k=1}^{n} e^{\lambda Y_{k-1}} \big(e^{\lambda(Y_k - Y_{k-1})} - 1\big) \\ &= 1 + S_n + \sum_{k=1}^{n} e^{\lambda Y_{k-1}} (H_k - 1).\end{aligned}$$

Using the sequence of localizing stopping times

$$T_p := \min\Big\{n \geq 0 : \sum_{k=1}^{n+1} \mathbb{E}(e^{\lambda Y_k} | \mathcal{G}_{k-1}) > p\Big\}$$

for $p > 0$, the process $(S_{n\wedge T_p})_{n\geq 0}$ is a uniformly integrable martingale. So using Fatou's Lemma, one easily gets that

$$\begin{aligned}\mathbb{E}(e^{\lambda Y_T}) \leq \liminf_{p\to+\infty} \mathbb{E}(e^{\lambda Y_{T\wedge T_p}}) &\leq \liminf_{p\to+\infty} \Big\{1 + \mathbb{E}(S_{T\wedge T_p}) + \mathbb{E}\Big(\sum_{k=1}^{T\wedge T_p} e^{\lambda Y_{k-1}}(H_k - 1)\Big)\Big\} \\ &= 1 + \liminf_{p\to+\infty} \mathbb{E}\Big(\sum_{k=1}^{T\wedge T_p} e^{\lambda Y_{k-1}}(H_k - 1)\Big).\end{aligned}$$

This entails (14) if we prove that

$$\sum_{i=1}^{n} e^{\lambda Y_{k-1}}(H_k - 1) \leq c_\lambda \tag{19}$$

for all $n \geq 1$. First, we prove that

$$H_n \leq \exp\Big[\frac{\lambda a s_{n-1}^2}{(a+V_n)^2}\Big(\Gamma_\lambda + \frac{2M_{n-1}^2}{a+V_{n-1}}(2\lambda\Gamma_\lambda - 1)\Big)\Big], \tag{20}$$

which entails that H_n is finite almost surely. We can write

$$\begin{aligned} Y_n - Y_{n-1} &= a\frac{M_n^2 - M_{n-1}^2}{\big(a+V_n\big)^2} + aM_{n-1}^2\frac{\big(a+V_{n-1}\big)^2 - \big(a+V_n\big)^2}{\big(a+V_n\big)^2\big(a+V_{n-1}\big)^2} \\ &= a\frac{\big(M_n - M_{n-1}\big)^2 + 2M_{n-1}\big(M_n - M_{n-1}\big)}{\big(a+V_n\big)^2} \\ &\quad - \frac{aM_{n-1}^2 s_{n-1}^2(2a + V_{n-1} + V_n)}{\big(a+V_n\big)^2\big(a+V_{n-1}\big)^2} \\ &\leq \frac{a\big(s_{n-1}^2\zeta_n^2 + 2M_{n-1}s_{n-1}\zeta_n\big)}{\big(a+V_n\big)^2} - \frac{2aM_{n-1}^2 s_{n-1}^2}{\big(a+V_n\big)^2\big(a+V_{n-1}\big)} \end{aligned}$$

where we used that $V_{n-1} \leq V_n$. In other words

$$\exp\big(\lambda(Y_n - Y_{n-1})\big) \leq \exp\big(\mu_n\zeta_n^2 + \rho_n\zeta_n - \delta_n\big),$$

with:

$$\mu_n = \frac{\lambda a s_{n-1}^2}{\big(a+V_n\big)^2}, \quad \rho_n = \frac{2\lambda a s_{n-1}M_{n-1}}{\big(a+V_n\big)^2}, \quad \delta_n = \frac{2\lambda a s_{n-1}^2 M_{n-1}^2}{\big(a+V_n\big)^2\big(a+V_{n-1}\big)}.$$

The random variables μ_n, ρ_n and δ_n are $\mathcal{G}_{n-1}$-measurable and one has $0 \leq \mu_n \leq \lambda$. We need the following Lemma.

Lemma 2. *Let ζ be a real random variable such that $\mathbb{E}[\zeta] = 0$ and such that*

$$\mathbb{E}[\exp(\mu\zeta^2)] \leq \gamma$$

for some $\mu > 0$ and $\gamma > 1$. Then, for any $\rho \in \mathbb{R}$ and $m \in [0, \mu)$, we have

$$\mathbb{E}[e^{m\zeta^2 + \rho\zeta}] \leq \exp\Big(\frac{(1+2\gamma)(\rho^2 + m)}{2(\mu - m)}\Big).$$

The proof of this Lemma is given in Section 6. Conditionally to $\mathcal{G}_{n-1}$, we apply Lemma 2 to ζ_n. This gives

$$H_n \leq \mathbb{E}[\exp(\mu_n\zeta_n^2 + \rho_n\zeta_n - \delta_n) \mid \mathcal{G}_{n-1}] \leq \exp\Big(\Gamma_\lambda\big(\rho_n^2 + \mu_n\big) - \delta_n\Big),$$

that can be be written

$$H_n \leq \exp\Big[\frac{\lambda a s_{n-1}^2}{(a+V_n)^2}\Big(\Gamma_\lambda + 2M_{n-1}^2\Big(\frac{2\lambda\Gamma_\lambda a}{(a+V_n)^2} - \frac{1}{a+V_{n-1}}\Big)\Big)\Big]$$

which yields (20) using $a/(a+V_n)^2 \leq 1/(a+V_{n-1})$. Since $\lambda < \mu/[2(1+\gamma)]$, we have $2\lambda\Gamma_\lambda - 1 < 0$, so (20) entails

$$H_n - 1 \leq \exp\Big[\frac{\lambda\Gamma_\lambda a s_{n-1}^2}{(a+V_n)^2}\Big] - 1 \leq (\exp(\lambda\Gamma_\lambda) - 1)\frac{a s_{n-1}^2}{(a+V_n)^2},$$

where we used the fact that $e^{\mu x} - 1 \leq (e^\mu - 1)x$ for any $x \in [0, 1/2]$, and $\mu > 0$. Note that (20) entails also the following inclusion:

$$\{H_n > 1\} \subset \left\{\frac{2M_{n-1}^2}{a+V_{n-1}} < \frac{\Gamma_\lambda}{1 - 2\lambda\Gamma_\lambda}\right\} \subset \left\{e^{\lambda Y_{n-1}} < \exp\Big(\frac{\lambda\Gamma_\lambda}{2(1-2\lambda\Gamma_\lambda)}\Big)\right\}.$$

It follows that

$$\sum_{k=1}^{n} e^{\lambda Y_{k-1}}(H_k - 1) \leq c_\lambda \sum_{k=1}^{n} \frac{a s_{k-1}^2}{(a+V_k)^2},$$

so (19) follows, since

$$\sum_{k=1}^{n} \frac{a s_{k-1}^2}{(a+V_k)^2} \leq \int_0^{V_n} \frac{a}{(a+x)^2} dx \leq 1.$$

This concludes the proof of (14) for $\alpha = 2$. □

5.2 Proof of Theorem 2 for $\alpha = 1$

First, note that (13) and the fact that the ζ_k are centered entails that for any $|\lambda| < \mu$, we have

$$\mathbb{E}[\exp(\lambda \zeta_k) \mid \mathcal{G}_{k-1}] \leq \exp(\mu' \lambda^2) \tag{21}$$

for any $k \geq 1$, where $\mu' = (\gamma - 1)/\mu^2$. Now, we use the same mechanism of proof as for the case $\alpha = 2$. Let $a > 0$ and $\lambda \in (-\mu, \mu)$ be fixed. Define

$$Y_n = \frac{\sqrt{a} M_n}{a + V_n} \text{ and } H_n = \mathbb{E}\big[\cosh(\lambda Y_n) - \cosh(\lambda Y_{n-1}) \mid \mathcal{G}_{n-1}\big].$$

Assuming for the moment that H_n is finite almost surely, we define the local martingale

$$S_n := \sum_{k=1}^{n} \Big(\cosh(\lambda Y_k) - \cosh(\lambda Y_{k-1}) - H_k \Big).$$

Thus, inequality (15) follows if we prove that for all $n \geq 1$:

$$\cosh(\lambda Y_n) \leq 1 + S_n + \mu' \lambda^2 \exp\big(\mu' \lambda^2\big) \cosh\big(2 \log 2 + 2 \mu' \lambda^2\big).$$

We can write

$$Y_n - Y_{n-1} = -\frac{\sqrt{a} M_{n-1} s_{n-1}^2}{(a+V_n)(a+V_{n-1})} + \frac{\sqrt{a} s_{n-1} \zeta_n}{a + V_n},$$

which gives, together with (21):

$$\mathbb{E}\Big[\exp\big(\pm\lambda(Y_n - Y_{n-1})\big) \mid \mathcal{G}_{n-1}\Big] \leq \exp\Big(\pm \frac{\lambda \sqrt{a} M_{n-1} s_{n-1}^2}{(a+V_n)(a+V_{n-1})} + \frac{\mu' \lambda^2 a s_{n-1}^2}{(a+V_n)^2} \Big).$$

As we have

$$\cosh(\lambda Y_n) = \frac{1}{2} e^{\lambda Y_{n-1}} e^{\lambda(Y_n - Y_{n-1})} + \frac{1}{2} e^{-\lambda Y_{n-1}} e^{-\lambda(Y_n - Y_{n-1})},$$

we derive:

$$\begin{aligned}
\mathbb{E}\Big[\cosh(\lambda Y_n) \mid \mathcal{G}_{n-1}\Big] &\leq \frac{1}{2} \exp\Big(\lambda Y_{n-1} - \frac{\lambda \sqrt{a} M_{n-1} s_{n-1}^2}{(a+V_n)(a+V_{n-1})} + \frac{\mu' \lambda^2 a s_{n-1}^2}{(a+V_n)^2} \Big) \\
&\quad + \frac{1}{2} \exp\Big(-\lambda Y_{n-1} + \frac{\lambda \sqrt{a} M_{n-1} s_{n-1}^2}{(a+V_n)(a+V_{n-1})} + \frac{\mu' \lambda^2 a s_{n-1}^2}{(a+V_n)^2} \Big), \\
&= \exp\Big(\frac{\mu' \lambda^2 a s_{n-1}^2}{(a+V_n)^2} \Big) \cosh\Big((1 - \frac{s_{n-1}^2}{a+V_n}) \lambda Y_{n-1} \Big).
\end{aligned}$$

So, it remains to prove that

$$\sum_{k=1}^{n}\Big(\exp\Big(\frac{\mu'\lambda^2 a s_{k-1}^2}{(a+V_k)^2}\Big)\cosh\Big((1-\frac{s_{k-1}^2}{a+V_k})\lambda Y_{k-1}\Big)-\cosh(\lambda Y_{k-1})\Big)$$
$$\leq \mu'\lambda^2\exp\big(\mu'\lambda^2\big)\cosh\big(2\log 2+2\mu'\lambda^2\big).$$

We need the following lemma.

Lemma 3. *If $A>0$, one has*

$$\sup_{\eta\in[0,1]}\sup_{z\geq 0}\big(e^{A\eta}\cosh((1-\eta)z)-\cosh(z)\big)\leq A\eta e^{A\eta}\cosh(2\log 2+2A).$$

The proof of this Lemma is given in Section 6. Using Lemma 3 with $\eta=s_{k-1}^2/(a+V_k)$ and $A=\mu'\lambda^2 a/(a+V_k)$, we obtain

$$\exp\Big(\frac{\mu'\lambda^2 a s_{k-1}^2}{(a+V_k)^2}\Big)\cosh\Big((1-\frac{s_{k-1}^2}{a+V_k})\lambda Y_{k-1}\Big)-\cosh(\lambda Y_{k-1})$$
$$\leq \frac{\mu'\lambda^2 a s_{k-1}^2}{(a+V_k)^2}e^{\mu'\lambda^2}\cosh\big(2\log 2+2\lambda^2\mu'\big),$$

and (15) follows, since

$$\sum_{k=1}^{n}\frac{a s_{k-1}^2}{(a+V_k)^2}\leq\int_0^{V_n}\frac{a}{(a+x)^2}dx\leq 1.$$

This concludes the proof of Theorem 2. □

5.3 Proof of Theorem 1

5.3.1 Notations

Let us fix $\lambda\in(0,\frac{\mu}{2(1+\gamma)})$, to be chosen later. In the following we denote by C any constant which depends only on (λ,μ,γ). Let us recall that on the event

$$\Omega':=\{L(h_0)^{-1/2}\leq\bar{W}(h_0)\}\cap\{W(H^*)\leq u_0\},$$

the bandwidths H^* and $\hat{H}$ are well defined, and let us we set for short

$$\mathbb{P}'(A)=\mathbb{P}(\Omega'\cap A).$$

We use the following notations: for $h>0$ and $a>0$, take

$$M(h):=\sum_{k=1}^{N}\frac{1}{\sigma_{k-1}^2}\mathbf{1}_{|X_{k-1}-x|\leq h^{\varepsilon}k},\quad Z(h,a):=\frac{\sqrt{a}\,|M(h)|}{a+L(h)}.\tag{22}$$

If $h=h_j\in\mathcal{H}$, we denote $h_-:=h_{j+1}$ and $h_+:=h_{j-1}$ if $j\geq 1$. We will use repeatedly the following quantity: for $i_0\in\mathbb{N}$ and $t>0$, consider

$$\pi(i_0,t):=\mathbb{P}\Big[\sup_{i\geq i_0}\psi^{-1/2}(h_i)\sup_{a\in I(h_i)}Z\big(h_i,a\psi(h_i)\big)>t\Big],\tag{23}$$

where

$$I(h):=[u_0^{-2},\delta_0^{-2}(h/h_0)^{-2\alpha_0}].$$

Note that this interval is related to the definition of $\bar{W}$, see (7). The proof of Theorem 1 contains three main steps. Namely,

1. the study of the risk of the ideal estimator $\bar{W}(H^*)^{-1}|\hat{f}(H^*) - f(x)|$,
2. the study of the risk $\bar{W}(H^*)^{-1}|\hat{f}(\hat{H}) - f(x)|$ when $\{H^* \leq \hat{H}\}$,
3. the study of the risk $\bar{W}(H^*)^{-1}|\hat{f}(\hat{H}) - f(x)|$ when $\{H^* > \hat{H}\}$.

These are the usual steps in the study of the Lepski's method, see [22, 23, 24, 25]. However, the context (and consequently the proof) proposed here differs significantly from the "usual" proof.

5.3.2 On the event $\{H^* \leq \hat{H}\}$

Recall that $\nu > 0$ is the constant in front of the Lepski's threshold, see (4). Let us prove the following.

Lemma 4. *For all $t > 0$ one has*

$$\mathbb{P}'\Big[\bar{W}(H^*)^{-1}|\hat{f}(H^*) - f(x)| > t\Big] \leq \pi(0, (t-1)/2), \tag{24}$$

and

$$\mathbb{P}'\Big[H^* \leq \hat{H}, \bar{W}(H^*)^{-1}\big|\hat{f}(\hat{H}) - f(x)\big| > t\Big] \leq \pi(0, (t-\nu-1)/2). \tag{25}$$

Proof. First, use the decomposition

$$|\hat{f}(H^*) - f(x)| \leq |\widetilde{f}(H^*) - f(x)| + \frac{|M(H^*)|}{L(H^*)},$$

where we recall that $\widetilde{f}(h)$ is given by (5), and the fact that $|\widetilde{f}(H^*) - f(x)| \leq \bar{W}(H^*)$, since $W(H^*) \leq \bar{W}(H^*)$ on $\{W(H^*) \leq u_0\}$. Then, use (8) to obtain $L(H^*)^{1/2} \geq \psi(H^*)^{1/2}\bar{W}(H^*)^{-1}$, so that

$$\begin{aligned}\frac{|M(H^*)|}{L(H^*)} &\leq \frac{2|M(H^*)|}{L(H^*) + \psi(H^*)\bar{W}(H^*)^{-2}} \\ &\leq 2\bar{W}(H^*)\psi^{-1/2}(H^*)Z\big(H^*, \bar{W}^{-2}(H^*)\psi(H^*)\big),\end{aligned}$$

and

$$\begin{aligned}\bar{W}^{-1}(H^*)\frac{|M(H^*)|}{L(H^*)} &\leq 2\psi^{-1/2}(H^*) \sup_{a \in I(H^*)} Z\big(H^*, a\psi(H^*)\big) \\ &\leq 2 \sup_{j \geq 0} \psi^{-1/2}(h_j) \sup_{a \in I(h_j)} Z\big(h_j, a\psi(h_j)\big),\end{aligned} \tag{26}$$

this concludes the proof of (24). On $\{H^* \leq \hat{H}\}$, one has using (4) and (8):

$$|\hat{f}(\hat{H}) - \hat{f}(H^*)| \leq \nu(\psi(H^*)/L(H^*))^{1/2} \leq \nu\bar{W}(H^*).$$

Hence, since $W(H^*) \leq \bar{W}(H^*)$ on $\{W(H^*) \leq u_0\}$, we have for all $t > 0$:

$$\begin{aligned}&\mathbb{P}'\Big[H^* \leq \hat{H}, \bar{W}(H^*)^{-1}\big|\hat{f}(\hat{H}) - f(x)\big| > t\Big] \\ &\leq \mathbb{P}'\Big[H^* \leq \hat{H}, \bar{W}(H^*)^{-1}\big|\hat{f}(H^*) - f(x)\big| > t - \nu\Big],\end{aligned}$$

and (25) follows using (24). □

5.3.3 On the event $\{H^* > \hat{H}\}$

Lemma 5. *For any $t, \eta > 0$, we have*

$$\mathbb{P}'\Big(H^* \le \eta, \sup_{H_{u_0} \le h < H^*, h \in \mathcal{H}} \frac{|M(h)|}{(L(h)\psi(h))^{1/2}} > t\Big) \le \pi(i_0(\eta), t/2),$$

where we put

$$i_0(\eta) = \min\{i \in \mathbb{N} : h_i < \eta\}.$$

Proof. Note that $u(h) := (\psi(h)/L(h))^{1/2}$ is decreasing, so $h = H_{u(h)}$ for $h \in \mathcal{H}$, and note that

$$\frac{|M(h)|}{(L(h)\psi(h))^{1/2}} = u(h)^{-1} \frac{|M(H_{u(h)})|}{L(H_{u(h)})}.$$

If $h < H^*$ then $u(h) = (\psi(h)/L(h))^{1/2} \ge \bar{W}(h)$ using (8), and $\bar{W}(h) \ge \varepsilon_0 (h/h_0)^{\alpha_0}$. So, $u(h) \ge \varepsilon_0 (H_{u(h)}/h_0)^{\alpha_0}$ when $h < H^*$. If $h \ge H_{u_0}$, then $u(h) \le u_0$ using the definition of H_{u_0}. This entails

$$\begin{aligned}
&\sup_{H_{u_0} \le h < H^*, h \in \mathcal{H}} \frac{|M(h)|}{(L(h)\psi(h))^{1/2}} \\
&\qquad \le \sup\Big\{u^{-1} \frac{|M(H_u)|}{L(H_u)}; u : H_u < H^* \text{ and } \delta_0 (H_u/h_0)^{\alpha_0} < u \le u_0\Big\}.
\end{aligned}$$

Hence, for any u such that $\delta_0 (H_u/h_0)^{\alpha_0} < u \le u_0$ and $H_u < H^* \le \eta$, one has using (3):

$$\begin{aligned}
u^{-1} \frac{|M(H_u)|}{L(H_u)} &\le 2u^{-1} \frac{|M(H_u)|}{L(H_u) + u^{-2}\psi(H_u)} \\
&= 2\psi(H_u)^{-1/2} Z\big(H_u, u^{-2}\psi(H_u)\big) \\
&\le 2 \sup_{i : h_i < \eta} \psi(h_i)^{-1/2} \sup_{\delta_0 (h_i/h_0)^{\alpha_0} \le u \le u_0} Z\big(h_i, u^{-2}\psi(h_j)\big). \qquad \square
\end{aligned}$$

Lemma 6. *For any $s, t > 0$ define*

$$\eta_{s,t} := h_0 \Big(\frac{u_0 s}{\delta_0 t}\Big)^{1/\alpha_0}. \tag{27}$$

Then, for all $0 < s < t$, we have:

$$\begin{aligned}
&\mathbb{P}'\Big[H^* > \hat{H}, \bar{W}(H^*)^{-1} \big|\hat{f}(\hat{H}) - f(x)\big| > t\Big] \\
&\quad \le \pi\Big(0, \frac{s-1}{2}\Big) + \pi\Big(i_0(\eta_{s,t}), \frac{1}{4}\Big(\nu - \frac{2s}{t}\Big)\Big) + \pi\Big(0, \frac{1}{2}\Big(\frac{\nu t}{2s} - 1\Big)\Big).
\end{aligned}$$

Proof. Let $0 < s < t$. One has

$$\begin{aligned}
&\mathbb{P}'\big[H^* > \hat{H}, \bar{W}(H^*)^{-1} \big|\hat{f}(\hat{H}) - f(x)\big| > t\big] \\
&\qquad \le \mathbb{P}'\big[H^* > \hat{H}, (L(\hat{H})/\psi(\hat{H}))^{1/2} \big|\hat{f}(\hat{H}) - f(x)\big| > s\big] \\
&\qquad + \mathbb{P}'\Big[H^* > \hat{H}, (\psi(\hat{H})/L(\hat{H}))^{1/2} > (t/s)\bar{W}(H^*)\Big].
\end{aligned}$$

The first term is less than $\pi(0,(s-1)/2)$, indeed, on $\{W(H^*)\le u_0, H^*>\hat H\}$ one has

$$\begin{aligned}(L(\hat H)/\psi(\hat H))^{1/2}\big|\hat f(\hat H)-f(x)\big| &\le (L(\hat H)/\psi(\hat H))^{1/2}\big|\widetilde f(\hat H)-f(x)\big|\\ &\quad+(L(\hat H)\psi(\hat H))^{-1/2}|M(\hat H)|\\ &\le (L(\hat H)/\psi(\hat H))^{1/2}W(\hat H)+(L(\hat H)\psi(\hat H))^{-1/2}|M(\hat H)|\\ &\le 1+(L(\hat H)\psi(\hat H))^{-1/2}|M(\hat H)|,\end{aligned}$$

and the desired upper-bound follows from Lemma 5. Let us bound the second term. Consider

$$\omega\in\big\{W(H^*)\le u_0, H^*>\hat H, (\psi(\hat H)/L(\hat H))^{1/2}>(t/s)\bar W(H^*)\big\}.$$

Due to the definition of $\hat H$, see (4), there exits $h'=h'_\omega\in[H_{u_0},\hat H]$ such that

$$\big|\hat f(h')-\hat f(\hat H_+)\big|>\nu(\psi(h')/L(h'))^{1/2}.$$

But since $h'\le\hat H<H^*$, one has

$$\begin{aligned}\nu\Big(\frac{\psi(h')}{L(h')}\Big)^{1/2}<\big|\hat f(h')-\hat f(\hat H_+)\big| &\le \big|\widetilde f(h')-\widetilde f(\hat H_+)\big|+\frac{|M(h')|}{L(h')}+\frac{|M(\hat H_+)|}{L(\hat H_+)}\\ &\le 2\bar W(H^*)+\frac{|M(h')|}{L(h')}+\frac{|M(\hat H_+)|}{L(\hat H_+)}\\ &\le \frac{2s}{t}\Big(\frac{\psi(\hat H)}{L(\hat H)}\Big)^{1/2}+\frac{|M(h')|}{L(h')}+\frac{|M(\hat H_+)|}{L(\hat H_+)}\\ &\le \frac{2s}{t}\Big(\frac{\psi(h')}{L(h')}\Big)^{1/2}+\frac{|M(h')|}{L(h')}+\frac{|M(\hat H_+)|}{L(\hat H_+)}.\end{aligned}$$

So, since $h'\le\hat H$ entails (for such an ω) that $(\psi(h')/L(h'))^{1/2}\ge(\psi(\hat H)/L(\hat H))^{1/2}>(t/s)\bar W(H^*)$, we obtain

$$\begin{aligned}\frac{|M(h')|}{L(h')}+\frac{|M(\hat H_+)|}{L(\hat H_+)} &>\Big(\nu-\frac{2s}{t}\Big)\Big(\frac{\psi(h')}{L(h')}\Big)^{1/2}\\ &\ge\Big(\nu-\frac{2s}{t}\Big)\max\Big[\Big(\frac{\psi(h')}{L(h')}\Big)^{1/2},\frac{t}{s}\bar W(H^*)\Big],\end{aligned}$$

and therefore

$$\begin{aligned}\omega\in\Big\{\sup_{H_{u_0}\le h<H^*,h\in\mathcal H}\frac{|M(h)|}{(L(h)\psi(h))^{1/2}}>\frac12\Big(\nu-\frac{2s}{t}\Big)\Big\}\\ \cup\Big\{\frac{|M(H^*)|}{L(H^*)}\ge\frac{t}{2s}\Big(\nu-\frac{2s}{t}\Big)\bar W(H^*)\Big\}.\end{aligned}$$

In addition, because of $\hat H\ge H_{u_0}$ one has

$$\delta_0(H^*/h_0)^{\alpha_0}\le\bar W(H^*)<(s/t)(\psi(\hat H)/L(\hat H))^{1/2}\le(s/t)u_0,$$

so $H^* \le \eta_{s,t}$, where $\eta_{s,t}$ is given by (27). We have shown that

$$\Big\{W(H^*) \le u_0, H^* > \hat{H}, \Big(\frac{\psi(\hat{H})}{L(\hat{H})}\Big)^{1/2} > \frac{t}{s}\bar{W}(H^*)\Big\}$$
$$\subset \Big\{H^* \le \eta_{s,t}, \sup_{H_{u_0} \le h < H^*, h \in \mathcal{H}} \frac{|M(h)|}{(L(h)\psi(h))^{1/2}} > \frac{1}{2}\Big(\nu - \frac{2s}{t}\Big)\Big\}$$
$$\cup \Big\{\frac{|M(H^*)|}{L(H^*)} \ge \Big(\frac{\nu t}{2s} - 1\Big)\bar{W}(H^*)\Big\},$$

and we conclude using Lemma 5 and (26). □

5.3.4 Finalization of the proof

In order to conclude the proof of Theorem 1, we need the following uniform version of Theorem 2: under the same assumptions as in Theorem 2, we have for any $0 < a_0 < a_1$:

$$\mathbb{E}\Big[\sup_{a \in [a_0, a_1]} \exp\Big(\frac{\lambda}{2}\frac{aM_N^2}{(a+V_N)^2}\Big)\Big] \le (1 + c_\lambda)(1 + \log(a_1/a_0)). \tag{28}$$

Indeed, since

$$\Big|\frac{\partial}{\partial a}\frac{aM_N^2}{(a+V_N)^2}\Big| = \Big|\frac{M_N^2}{(a+V_N)^3}(V_N - a)\Big| \le a^{-1}\frac{aM_N^2}{(a+V_N)^2} = Y^a/a,$$

we have

$$\sup_{a \in [a_0, a_1]} \exp(\lambda Y^a/2) \le \exp(\lambda Y^{a_0}/2) + \int_{a_0}^{a_1} a^{-1}\exp(\lambda Y^a/2)\lambda Y^a/2\, da$$
$$\le \exp(\lambda Y^{a_0}) + \int_{a_0}^{a_1} a^{-1}\exp(\lambda Y^a)\, da,$$

so (28) follows taking the expectation and using Theorem 2. Now, using (28) with

$$s_k = \frac{1}{\sigma_{k-1}}\mathbf{1}_{|X_{k-1}-x| \le h}, \quad \zeta_k = \varepsilon_k/\sigma_{k-1}$$

we obtain

$$\mathbb{E}\Big[\exp\big((\lambda/2)\sup_{a \in [a_0, a_1]} Z(h,a)^2\big)\Big] \le C(1 + \log(a_1/a_0)),$$

where we recall that $Z(h,a)$ is given by (22). So, using Markov's inequality, we arrive, for all $h > 0$, $a_1 > a_0 > 0$ and $t \ge 0$, at:

$$\mathbb{P}\Big[\sup_{a \in [a_0, a_1]} Z(h,a) \ge t\Big] \le C(1 + \log(a_1/a_0))e^{-\lambda t^2/2}. \tag{29}$$

A consequence of (29), together with an union bound, is that for all $i_0 \in \mathbb{N}$ and $t > 0$:

$$\pi(i_0, t) \le Ce^{-\lambda t^2/2}\sum_{i \ge i_0}(h_i/h_0)^{b\lambda t^2/2}\big(1 + 2\log(u_0/\delta_0) + 2\alpha_0\log(h_0/h_i)\big), \tag{30}$$

where we recall that $\pi(i_0, t)$ is given by (23).

Now, it remains to use what the grid $\mathcal{H}$ is. Recall that for some $q \in (0,1)$, we have $h_i = h_0 q^i$ and we denote by C any positive number which depends only on $\lambda, \mu, \gamma, q, b, u_0, \delta_0, \alpha_0, \nu$. Using together (25) and Lemma 6, one gets for $0 < s < t$:

$$\mathbb{P}'\big[\bar{W}(H^*)^{-1}|\hat{f}(\hat{H}) - f(x)| > t\big] \leq \pi\Big(0, \frac{t-\nu-1}{2}\Big) + \pi\Big(0, \frac{s-1}{2}\Big) + \pi\Big(i_0(\eta_{s,t}), \frac{1}{4}\Big(\nu - \frac{2s}{t}\Big)\Big) + \pi\Big(0, \frac{1}{2}\Big(\frac{\nu t}{2s} - 1\Big)\Big),$$

and using (30), we have for any $u > 0$, $i_0 \in \mathbb{N}$:

$$\pi(i_0, u) \leq C e^{-\lambda u^2/2}(i_0+1) q^{i_0 b \lambda u^2/2}.$$

Recalling that $\eta_{s,t}$ is given by (27) and that $i_0(\eta) = \min\{i \in \mathbb{N} : h_i < \eta\}$, we have

$$\frac{\log(\delta_0/u_0) + \log(t/s)}{\alpha_0 \log(1/q)} < i_0(\eta_{s,t}) \leq \frac{\log(\delta_0/u_0) + \log(t/s)}{\alpha_0 \log(1/q)} + 1. \tag{31}$$

Now, recall that $0 < \rho < \frac{b\mu\nu^2}{64\alpha_0(1+\gamma)}$ and consider $s = \sqrt{(8\rho \log t)/\lambda} + 1$. When t is large enough, we have $s < t$ and:

$$\pi\Big(0, \frac{s-1}{2}\Big) \leq C_1 t^{-\rho}, \quad \pi\Big(0, \frac{t-\nu-1}{2}\Big) \leq C_2 \exp(-C_2' t^2),$$
$$\pi\Big(0, \frac{1}{2}\Big(\frac{\nu t}{2s} - 1\Big)\Big) \leq C_3 \exp\big(-C_3'(t/\log t)^2\big),$$

for constants C_i, C_i' that depends on $\lambda, b, \nu, \delta_0, u_0, \alpha_0, q$. For the last probability, we have:

$$\pi\Big(i_0(\eta_{s,t}), \frac{1}{4}\Big(\nu - \frac{2s}{t}\Big)\Big) \leq C \exp\Big(-\frac{\lambda(\nu - 2s/t)^2}{32}\Big)(i_0(\eta_{s,t}) + 1) \times \exp\Big(-\frac{i_0(\eta_{s,t}) b\lambda(\nu - 2s/t)^2 \log(1/q)}{32}\Big),$$

and by taking $\lambda \in (0, \frac{\mu}{2(1+\gamma)})$ and t large enough, one has

$$\frac{b\lambda(\nu - 2s/t)^2}{32\alpha_0} > \rho,$$

so we obtain together with (31):

$$\pi\Big(i_0(\eta_{s,t}), \frac{1}{4}\Big(\nu - \frac{2s}{t}\Big)\Big) \leq C \frac{(\log(t+1))^{1+\rho/2}}{t^\rho},$$

when t is large enough. This concludes the proof of Theorem 1. □

5.4 Proof of Proposition 1

Let us denote for short $I_h = [x-h, x+h]$. Recall that h_w is well-defined when $n \geq \sigma^2/(P_X[I_{h_0}] w(h_0)^2)$, and that H_w is well defined on the event

$$\Omega_0 = \{L(h_0) \geq w(h_0)^{-2}\}.$$

So, from now on, we suppose that n is large enough, and we work on Ω_0. We need the following Lemma, which says that, when $L(h_w)$ and $\mathbb{E}L(h_w)$ are close, then H_w and h_w are close.

Lemma 7. *If Assumption 4 holds, we have for any $0 < \varepsilon < 1$ that on Ω_0:*

$$\Big\{L(h_w) \geq \frac{\mathbb{E}L(h_w)}{(1+\varepsilon)^s}\Big\} \subset \big\{H_w \leq (1+\varepsilon)h_w\big\} \quad \text{and}$$
$$\Big\{L(h_w) \leq \frac{\mathbb{E}L(h_w)}{(1-\varepsilon)^s}\Big\} \subset \big\{H_w > (1-\varepsilon)h_w\big\},$$

when n is large enough.

The proof of Lemma 7 is given in Section 6 below. We use also the next Lemma from [2] (see Claim 2, p. 858). It is a corollary of Berbee's coupling lemma [3], that uses a construction from the proof of Proposition 5.1 in [32], see p. 484.

Lemma 8. *Grant Assumption 6. Let q, q_1 be integers such that $0 \leq q_1 \leq q/2$, $q_1 \geq 1$. Then, there exist random variables $(X_i^*)_{i=1}^n$ satisfying the following:*

- *For $j = 1, \ldots, J := [n/q]$, the random vectors*

  $$U_{j,1} := (X_{(j-1)q+1}, \ldots, X_{(j-1)q+q_1}) \text{ and } U_{j,1}^* := (X_{(j-1)q+1}^*, \ldots, X_{(j-1)q+q_1}^*)$$

  have the same distribution, and so have the random vectors

  $$U_{j,2} := (X_{(j-1)q+q_1+1}, \ldots, X_{jq}) \text{ and } U_{j,2}^* := (X_{(j-1)q+q_1+1}^*, \ldots, X_{jq}^*).$$

- *For $j = 1, \ldots, J$,*

  $$\mathbb{P}[U_{j,1} \neq U_{j,1}^*] \leq \beta_{q-q_1} \text{ and } \mathbb{P}[U_{j,2} \neq U_{j,2}^*] \leq \beta_{q_1}.$$

- *For each $k = 1, 2$, the random vectors $U_{1,k}^*, \ldots, U_{J,k}^*$ are independent.*

In what follows, we take simply $q_1 = [q/2] + 1$, where $[x]$ stands for the integral part of x, and introduce the event $\Omega^* = \{X_i = X_i^*, \forall i = 1, \ldots, n\}$. Assume to simplify that $n = Jq$. Lemma 8 gives

$$\mathbb{P}[(\Omega^*)^{\complement}] \leq J(\beta_{q-q_1} + \beta_{q-q_1}) \leq 2J\beta_{[q/2]} \leq \frac{2n\beta_{[q/2]}}{q}. \tag{32}$$

Then, denote for short $L^*(h) = \sum_{i=1}^n \mathbf{1}_{|X_{i-1}^* - x| \leq h}$, and note that, using Lemma 7, we have, for $z := 1 - 1/(1+\varepsilon)^s$:

$$\begin{aligned}\big\{H_w > (1+\varepsilon)^s h_w\big\} \cap \Omega^* \cap \Omega_0 &\subset \big\{L^*(h_w) - \mathbb{E}L(h_w) \geq z\mathbb{E}L(h_w)\big\} \\ &= \Big\{\frac{1}{n}\sum_{i=1}^n (\mathbf{1}_{|X_{i-1}^* - x| \leq h_w} - P_X[I_{h_w}]) \geq z P_X[I_{h_w}]\Big\}.\end{aligned}$$

Use the following decomposition of the sum:

$$\frac{1}{n}\sum_{i=1}^n (\mathbf{1}_{|X_{i-1}^* - x| \leq h_w} - P_X[I_{h_w}]) \leq \frac{1}{J}\sum_{j=1}^J (Z_{j,1} + Z_{j,2}),$$

where for $k \in \{1, 2\}$, we put

$$Z_{j,k} := \frac{1}{q}\sum_{i \in I_{j,k}} (\mathbf{1}_{|X_{i-1}^* - x| \leq h_w} - P_X[I_{h_w}]),$$

where $I_{j,1} := \{(j-1)q+1, \dots, (j-1)q+q_1\}$ and $I_{j,2} := \{(j-1)q+q_1+1, \dots, jq\}$. For $k \in \{1,2\}$, we have using Lemma 8 that the variables $(Z_{j,k})_{j=1}^J$ are independent, centered, such that $\|Z_{j,k}\|_\infty \leq 1/2$ and $\mathbb{E}[Z_{j,k}^2] \leq P_X[I_{h_w}]/4$. So, Bernstein's inequality gives

$$\mathbb{P}\Big[\Big\{H_w > \frac{1}{(1-z)^{1/s}} h_w\Big\} \cap \Omega^* \cap \Omega_0\Big] \leq 2\exp\Big(-\frac{z^2}{2(1+z/3)} \frac{nP_X[I_{h_w}]}{q}\Big),$$

and doing the same on the other side gives for any $z \in (0,1)$:

$$\mathbb{P}\Big[\Big\{\frac{1}{(1+z)^{1/s}} h_w \leq H_w \leq \frac{1}{(1-z)^{1/s}} h_w\Big\}^{\complement} \cap \Omega^*\Big] \leq 4\exp\Big(-\frac{z^2}{2(1+z/3)} \frac{nP_X[I_{h_w}]}{q}\Big).$$

So, when n is large enough, we have

$$\mathbb{P}[h_w/2 \leq H_w \leq 2h_w] \geq 1 - 4\exp\Big(-\frac{CnP_X[I_{h_w}]}{q}\Big) - \frac{2n\beta_{[q/2]}}{q}. \tag{33}$$

But, since on $[0, h_0]$ w is increasing and $w(h) = h^s \ell_w(h)$ where ℓ_w is slowly varying, we have $\{h_w/2 \leq H_w \leq 2h_w\} \subset \{w(h_w)/4 \leq w(H_w) \leq 4w(h_w)\}$ when n is large enough. Now, Lemma 1 and Assumption 5 gives that

$$nP_X[I_{h_w}] = n^{2s/(2s+\tau+1)} \ell(1/n),$$

where ℓ is a slowly varying function that depends on ℓ_X, ℓ_w, s, τ and σ. When the β-mixing is geometric, we have $\psi^{-1}(p) = \exp((p/\eta)^{1/\kappa})$, so the choice $q = n^{2s\kappa/((2s+\tau+1)(\kappa+1))}$ implies

$$\mathbb{P}\Big[\{\frac{w(h_w)}{4} \leq w(H_w) \leq 4w(h_w)\} \cap \Omega_0\Big] \geq 1 - \exp(-C_1 n^{\delta_1} \ell_1(1/n)).$$

When the mixing is arithmetic, we have $\psi^{-1}(p) = (p/\eta)^{1/\kappa}$, so the choice $q = n^{2s/(2s+\tau+1)} \ell(1/n)/(\log n)^2$ implies

$$\mathbb{P}\Big[\{\frac{w(h_w)}{4} \leq w(H_w) \leq 4w(h_w)\} \cap \Omega_0\Big] \geq 1 - C_2 n^{-\delta_2} \ell_2(1/n).$$

So, it only remains to control the probability of Ω_0. Using the same coupling argument as before together with Bernstein's inequality, we have when n is large enough:

$$\begin{aligned}
\mathbb{P}[L(h_0) < w(h_0)^{-2}] &= \mathbb{P}[L(h_0) - \mathbb{E}L(h_0) < w(h_0)^{-2} - \mathbb{E}L(h_0)] \\
&\leq \mathbb{P}\Big(L(h_0) - \mathbb{E}L(h_0) < -\frac{nP_X[I_{h_0}]}{2}\Big) \\
&\leq \exp\Big(-C_2 \frac{nP_X[I_{h_0}]}{q}\Big) + \frac{2n\beta_{[q/2]}}{q}.
\end{aligned}$$

So, when the β-mixing is geometric, the choice $q = n^{\kappa/(\kappa+1)}$ implies that $\mathbb{P}[\Omega_0^{\complement}] \leq \exp(-C_1 n^{1/(\kappa+1)}) = o(\varphi_n)$. When the mixing is arithmetic, we have $\psi^{-1}(p) = (p/\eta)^{1/\kappa}$, so the choice $q = n/(\log n)^2$ gives $\mathbb{P}[\Omega_0^{\complement}] \leq C_2 (\log n)^2 n^{-1/\kappa} = o(\varphi_n)$. This concludes the proof of Proposition 1. □

5.5 Proof of Corollary 1

Let us fix $\rho \in (p, \frac{b\mu\nu^2}{128(1+\gamma)})$ (note that $\alpha_0 = 2$ under Assumption 4). Using Assumption 4, one can replace $\bar{W}$ by w in the statement of Theorem 1. This gives

$$\mathbb{P}\Big[\Big\{|\hat{f}(\hat{H}) - f(x)| \geq tw(H^*)\Big\} \cap \Omega_0\Big] \leq C_0 \frac{(\log(t+1))^{\rho/2+1}}{t^\rho}$$

for any $t \geq t_0$, where we recall that $\Omega_0 = \{L(h_0)^{-1/2} \leq w(h_0)\}$, and where

$$H^* := \min\Big\{h \in \mathcal{H} : \Big(\frac{\psi(h)}{L(h)}\Big)^{1/2} \leq w(h)\Big\}.$$

Recall the definition (17) of H_w, and note that by construction of $\mathcal{H}$, one has that $H_w \leq H^* \leq q^{-1}H_w$. So, on the event $\{H_w \leq 2h_w\}$, one has, using the fact that w is s-regularly varying, that $w(H^*) \leq w(2q^{-1}h_w) \leq 2(2/q)^s w(h_w)$ for n large enough. So, putting for short $A := \{H_w \leq 2h_w\} \cap \Omega_0$, we have

$$\mathbb{P}\Big[\Big\{|\hat{f}(\hat{H}) - f(x)| \geq c_1 tw(h_w)\Big\} \cap A\Big] \leq C_0 \frac{(\log(t+1))^{\rho/2+1}}{t^\rho}$$

for any $t \geq t_0$, where $c_1 = 2(2/q)^s$. Since $\rho > p$, we obtain, by integrating with respect to t, that

$$\mathbb{E}\big[|w(h_w)^{-1}(\hat{f}(\hat{H}) - f(x))|^p \mathbf{1}_A\big] \leq C_1,$$

where C_1 is a constant depending on C_0, t_0, q, ρ, s, p. Now, it only remains to observe that using Proposition 1, $\mathbb{P}(A^\complement) \leq 2\varphi_n$, and that $\varphi_n = o(w(h_w))$ in the geometrically β-mixing case, and in the arithmetically β-mixing when $\kappa < 2s/(s+\tau+1)$. □

6 Proof of the Lemmas

6.1 Proof of Lemma 7

For n large enough, we have $\psi((1+\varepsilon)h_w)/\ell_w((1+\varepsilon)h_w)^2 \leq (1+\varepsilon)^s \psi(h_w)/\ell_w(h_w)^2$ since ψ/ℓ_w^2 is slowly varying. So,

$$\frac{\psi((1+\varepsilon)h_w)}{w((1+\varepsilon)h_w)^2} \leq \frac{1}{(1+\varepsilon)^s}\frac{\psi(h_w)}{w(h_w)^2} = \frac{1}{(1+\varepsilon)^s}\mathbb{E}L(h_w).$$

On the other hand, by definition of H_w, we have

$$\{H_w \leq (1+\varepsilon)h_w\} = \Big\{L((1+\varepsilon)h_w) \geq \frac{\psi((1+\varepsilon)h_w)}{w((1+\varepsilon)h_w)^2}\Big\},$$

and $L((1+\varepsilon)h_w) \geq L(h_w)$, so we proved that the embedding

$$\Big\{\frac{L(h_w)}{\mathbb{E}L(h_w)} \geq \frac{1}{(1+\varepsilon)^s}\Big\} \subset \{H_w \leq (1+\varepsilon)h_w\}$$

holds when n is large enough. The same argument allows to prove that

$$\Big\{\frac{L(h_w)}{\mathbb{E}L(h_w)} \leq \frac{1}{(1-\varepsilon)^s}\Big\} \subset \{H_w > (1-\varepsilon)h_w\},$$

which concludes the proof of the Lemma. □

6.2 Proof of Lemma 2

Take $m \in [0, \mu)$ and $\rho \in \mathbb{R}$. Note that $e^y \leq 1 + ye^y \leq 1 + y + y^2 e^y$ for any $y \geq 0$, so

$$\begin{aligned} e^{m\zeta^2+\rho\zeta} &\leq e^{\rho\zeta} + m\zeta^2 e^{m\zeta^2+\rho\zeta} \\ &\leq 1 + \rho\zeta + (\rho^2 + m)\zeta^2 e^{m\zeta^2+\rho\zeta}, \end{aligned}$$

and

$$\mathbb{E}[e^{m\zeta^2+\rho\zeta}] \leq 1 + (\rho^2 + m)\mathbb{E}[\zeta^2 e^{m\zeta^2+\rho\zeta}], \tag{34}$$

since $\mathbb{E}\zeta = 0$. Take $m_1 \in (m, \mu)$. Since $\rho\zeta \leq \varepsilon\rho^2/2 + \zeta^2/(2\varepsilon)$ for any $\varepsilon > 0$, we obtain for $\varepsilon = [2(m_1 - m)]^{-1}$:

$$e^{m\zeta^2+\rho\zeta} \leq \exp\Big(\frac{\rho^2}{4(m_1 - m)}\Big) e^{m_1\zeta^2}.$$

Together with

$$\zeta^2 \leq \frac{1}{\mu - m_1} e^{(\mu - m_1)\zeta^2}$$

and the definition of μ, this entails

$$\mathbb{E}[\zeta^2 e^{m\zeta^2+\rho\zeta}] \leq \frac{\gamma}{\mu - m_1} \exp\Big(\frac{\rho^2}{4(m_1 - m)}\Big).$$

Thus,

$$\begin{aligned} \mathbb{E}[e^{m\zeta^2+\rho\zeta}] &\leq 1 + \frac{\gamma(\rho^2 + m)}{\mu - m_1} \exp\Big(\frac{\rho^2}{4(m_1 - m)}\Big) \\ &\leq 1 + \frac{\gamma(\rho^2 + m)}{\mu - m_1} \exp\Big(\frac{\rho^2 + m}{4(m_1 - m)}\Big). \end{aligned}$$

For the choice $m_1 = \mu/(1 + 2\gamma) + 2\gamma m/(1 + 2\gamma)$ one has $\gamma/(\mu - m_1) = 1/[2(m_1 - m)]$, so the Lemma follows using that $1 + ye^{y/2} \leq e^y$ for all $y \geq 0$. This concludes the proof of the Lemma. □

6.3 Proof of Lemma 3

Let $\eta \in [0, 1]$ and $z \in \mathbb{R}_+$ be such that $e^{A\eta}\cosh((1 - \eta)z) - \cosh(z) \geq 0$. Let us show that one has

$$z \leq 2\log 2 + 2A. \tag{35}$$

Since $\cosh(z)/\cosh((1 - \eta)z) \geq e^{\eta z}/2$ one has $z \leq \eta^{-1}\log 2 + A$. Thus (35) holds if $\eta \geq 1/2$. If $\eta < 1/2$ and $z \geq \log(3)$, it is easy to check that the derivative of $x \mapsto \cosh((1 - x)z)e^{\eta x/2}$ is non-positive, hence $\cosh(z) \geq e^{\eta z/2}\cosh((1 - \eta)z)$ in this case. Thus, we have either $z \leq \log(3)$ or $z \leq 2A$ which yields (35) in every case. Finally, from (35), we easily derive

$$\begin{aligned} e^{A\eta}\cosh((1 - \eta)z) - \cosh(z) &= \cosh((1 - \eta)z)\Big(e^{A\eta} - \frac{\cosh(z)}{\cosh((1 - \eta)z)}\Big) \\ &\leq \cosh(z)(e^{A\eta} - 1) \\ &\leq \cosh\big(2\log(2) + 2A\big) A\eta e^{A\eta}. \end{aligned}$$

This concludes the proof of the Lemma. □

References

[1] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tôhoku Math. J. (2)*, 19:357–367, 1967.

[2] Y. Baraud, F. Comte, and G. Viennet. Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, 29(3):839–875, 2001.

[3] Henry C. P. Berbee. *Random walks with stationary increments and renewal theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1979.

[4] Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *Ann. Appl. Probab.*, 18(5):1848–1869, 2008.

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

[6] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31(3):307–327, 1986.

[7] Victor H. de la Peña. A general class of exponential inequalities for martingales and ratios. *Ann. Probab.*, 27(1):537–564, 1999.

[8] Sylvain Delattre, Marc Hoffmann, and Mathieu Kessler. Dynamics adaptive estimation of a scalar diffusion. Technical report, Universités Paris 6 et Paris 7, 2002. http://www.proba.jussieu.fr/mathdoc/textes/PMA-762.pdf.

[9] Paul Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. Properties and examples.

[10] Jianqing Fan and Irène Gijbels. *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1996.

[11] David A. Freedman. On tail probabilities for martingales. *Ann. Probability*, 3:100–118, 1975.

[12] Stéphane Gaïffas. On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Stat.*, 11:344–364 (electronic), 2007.

[13] Alexander Goldenshluger and Oleg Lepski. Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields*, 143(1-2):41–71, 2009.

[14] Guerre. Design-adaptive pointwise nonparametric regression estimation for recurrent markov time series. Econometrics 0411007, EconWPA, November 2004.

[15] Emmanuel Guerre. Design adaptive nearest neighbor regression estimation. *J. Multivariate Anal.*, 75(2):219–244, 2000.

[16] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[17] Anatoli Juditsky, Oleg Lepski, and Alexandre Tsybakov. Nonparametric estimation of composite functions. 2007.

[18] Gérard Kerkyacharian, Oleg Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields*, 121(2):137–170, 2001.

[19] A. N. Kolmogorov and Ju. A. Rozanov. On a strong mixing condition for stationary Gaussian processes. *Teor. Verojatnost. i Primenen.*, 5:222–227, 1960.

[20] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[21] O. V. Lepski. Asymptotically minimax adaptive estimation i: Upper bounds, optimally adaptive estimates. *Theory of Probability and its Applications*, 36(4):682–697, 1988.

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

[23] O. V. Lepski. On problems of adaptive estimation in white gaussian noise. *Advances in Soviet Mathematics*, 12:87–106, 1992.

[24] 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. *The Annals of Statistics*, 25(3):929–947, 1997.

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

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

[27] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994.

[28] Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[29] Karine Tribouley and Gabrielle Viennet. $\mathbf{L}_p$ adaptive density estimation in a β mixing framework. *Ann. Inst. H. Poincaré Probab. Statist.*, 34(2):179–208, 1998.

[30] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

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

[32] Gabrielle Viennet. Inequalities for absolutely regular sequences: application to density estimation. *Probab. Theory Related Fields*, 107(4):467–492, 1997.