CORE
Metadata, citation and similar papers at core.ac.uk
Provided by HAL Descartes

ADAPTIVE ESTIMATION IN THE FUNCTIONAL NONPARAMETRIC REGRESSION MODEL

Gaëlle Chagny, Angelina Roche

To cite this version:

Gaëlle Chagny, Angelina Roche. ADAPTIVE ESTIMATION IN THE FUNCTIONAL NONPARAMETRIC REGRESSION MODEL. Journal of Multivariate Analysis, Elsevier, 2016, 146, pp.105–118. <10.1016/j.jmva.2015.07.001>. <hal-01099520>

HAL Id: hal-01099520

https://hal.archives-ouvertes.fr/hal-01099520

Submitted on 4 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ADAPTIVE ESTIMATION IN THE FUNCTIONAL NONPARAMETRIC REGRESSION MODEL

GAËLLE CHAGNY[(A)] AND ANGELINA ROCHE[(B)]

Abstract. In this paper, we consider nonparametric regression estimation when the predictor is a functional random variable (typically a curve) and the response is scalar. Starting from a classical collection of kernel estimates, the bias-variance decomposition of a pointwise risk is investigated to understand what can be expected at best from adaptive estimation. We propose a fully data-driven local bandwidth selection rule in the spirit of the Goldenshluger and Lepski method. The main result is a nonasymptotic risk bound which shows the optimality of our tuned estimator from the oracle point of view. Convergence rates are also derived for regression functions belonging to Hölder spaces and under various assumptions on the rate of decay of the small ball probability of the explanatory variable. A simulation study also illustrates the good practical performances of our estimator.

(A) LMRS, UMR CNRS 6085, Université de Rouen, France. gaelle.chagny@gmail.com

(B) MAP5 UMR CNRS 8145, Université Paris Descartes, France. angelina.roche@parisdescartes.fr

Keywords: Functional data analysis. Regression estimation. Nonparametric kernel estimators. Bandwidth selection.

AMS Subject Classification 2010: 62G08; 62H12.

1. Introduction

1.1. **Statistical model.** Nowadays, more and more advanced technologies make possible the recording of observations in such a way that the collected data may be considered as curves (or surfaces). This explains why developing methods for Functional Data Analysis (F.D.A.) is a great challenge for the statisticians, which has become more and more popular in the past decades, as it is highlighted by the monographies of Ramsay and Silverman (2005), Dabo-Niang and Ferraty (2008) and Ferraty and Romain (2011). Regression with functional covariate is one of the most studied problem. The functional linear model has first been widely investigated (see, among all, Ramsay and Dalzell 1991; Cai and Hall 2006; Crambes et al. 2009) with its generalisations (Cardot and Sarda 2005; Müller and Stadtmüller 2005). However, studies of the nonparametric regression model, in which we are interested, are more recent. Consider

$$Y = m(X) + \varepsilon, \tag{1}$$

where Y is a real random variable, X a random variable which takes values in a separable infinite-dimensional Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ (it can be $\mathbb{L}^2(I)$, the set of squared-integrable functions on a subset I of $\mathbb{R}$, or a Sobolev space), and $m : \mathbb{H} \to \mathbb{R}$ the target function to recover. The random variable ε stands for a noise term. We suppose that ε and X are independent and that ε is centred with $\mathbb{E}[\varepsilon^2]^{1/2} = \sigma < \infty$. The specificity thus stands in the dimension which is infinite in two aspects: first, the framework is a functional

Date: December 29, 2014.

one (the covariate X is a stochastic process which lives in an infinite-dimensional space), and then, no assumption is made on the form of the function m to estimate.

The aim of this paper is to provide an adaptive optimal strategy to estimate the regression function m from a data sample $\{(X_i, Y_i), i = 1, ..., n\}$ distributed like the couple (X, Y), that is $Y_i = m(X_i) + \varepsilon_i$, with ε_i's independent identically distributed (*i.i.d.* in the sequel) like ε, and independent from the X_i's. We explore the pointwise risk of a collection of kernel estimates, and propose to select the bandwidth in a data-driven way at a fixed curve so as to obtain a theoretical non-asymptotic risk bound, which proves the optimality of our estimator in the collection.

1.2. **State of the art and motivation.** The starting point of our work is the following collection of estimators. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function (that is $\int_{\mathbb{R}} K(u)du = 1$), $K_h : u \mapsto K(u/h)/h$, for $h > 0$, and

$$\widehat{m}_h(x) := \sum_{i=1}^{n} W_h^{(i)}(x) Y_i \text{ where } W_h^{(i)}(x) := \frac{K_h(\|X_i - x\|)}{\sum_{j=1}^{n} K_h(\|X_j - x\|)}, \tag{2}$$

for any $(x, y) \in \mathbb{H} \times \mathbb{R}$. Inspired by the classical kernel methods of Nadaraya (1964) and Watson (1964), these estimators have first been introduced in the functional context by Ferraty and Vieu (2000) and have then aroused considerable interest and attention. A first order of magnitude for the risk is given by Ferraty and Vieu (2002), almost complete convergence is proved by Ferraty and Vieu (2004), Ferraty et al. (2007) deals with asymptotic normality and Ferraty et al. (2010) examine asymptotic expansions for the mean integrated squared error. Finally, a clear account about convergence rates (upper and lower bounds for the pointwise risk of (2) under concentration assumptions for the process X, as well as minimax rates) is provided by Mas (2012). Moreover, this kind of estimates gives rise to other kernel-based strategies to recover m in the same functional context. Local linear methods in the spirit of local polynomial estimators have been explored by Baíllo and Grané (2009), Barrientos-Marin et al. (2010), Boj et al. (2010) and Berlinet et al. (2011), $k-$nearest neighbor kernels are examined by Burba et al. (2009), a recursive kernel approach is suggested by Amiri et al. (2014) and Reproducing Kernel Hilbert Space-based methods are reported by Avery et al. (2014). Kernel methods have also been successfully extended to estimate the regression function m from dependent data (see, *e.g.* Masry 2005, Delsol 2009 and references therein) or to consider the case of a functional response variable Y (in the model (1)), see Ferraty et al. (2012).

All these methods have one point in common: they heavily depend on the choice of the smoothing parameter h (see (2)), the so-called bandwidth. Heuristically, a large value for h leads to an estimator with large bias (but small variance), while a too small value leads to high variability. It raises the question of defining a bandwidth selection criterion which is proved to automatically balance the bias-variance trade-off. However, not many investigations are concerned with this theoretical (the criterion should be theoretically justified) and practical (the bandwidth has to be chosen for application purpose) problem for F.D.A. The most commonly used method to select the bandwidth of a functional regression kernel estimate is the leave-one-out cross validation. The first algorithm is proposed by Ferraty and Vieu (2002), used for dependent data by Ferraty et al. (2002) and Ferraty and Vieu (2006) and shown to be asymptotically optimal by Rachdi and Vieu (2007). It is a global method that do not depend on the point (curve) of estimation. A local version, also optimal in an asymptotic way, is proposed by Benhenni et al. (2007).

More recently, bayesian strategies have been studied (Shang, 2013, 2014), but only for simulation purposes.

1.3. **Contribution and overview.** Few systematic studies have been undertaken to build a data-driven bandwidth selection method so as to obtain adaptive estimators and nonasymptotic risk bounds for functional regression estimation, while the same question is the subject of a wide litterature in classical nonparametric estimation. Only two papers, to our knowledge, focus on this problem. A method based on empirical prediction-risk minimisation is explored by Antoniadis et al. (2009): oracle inequalities are proved, but the criterion is specific to functional time series prediction. Inspired both by the advances about the Lepski methods (Goldenshluger and Lepski, 2011) and model selection, Chagny and Roche (2014) investigate a global bandwidth selection method to estimate a cumulative distribution function conditionally to a functional covariate, shown to be optimal both in the oracle and minimax sense, for an integrated error.

Motivated by this work and by the observation that a local bandwidth choice can improve significantly the precision of estimation in a functional context (see the results of Benhenni et al. 2007), the goal of the present paper is to define a local bandwidth selection criterion for the estimators (2) in functional regression in such a way that the resulting estimator is optimal in a nonasymptotic point of view. An upper-bound for a pointwise risk is first derived in Section 2. For x_0 a fixed point in $\mathbb{H}$, the risk of an estimator $\widehat{m}(x_0)$ computed at the curve x_0, is, in the sequel,

$$\mathbb{E}\left[(\widehat{m}(x_0) - m(x_0))^2\right].$$

We obtain an exact bias-variance decomposition (Proposition 1) which permits to understand what we can expect at best from adaptive estimation, which is the subject of Section 3. The bandwidth selection, at a fixed curve, is automatically performed in the spirit of Goldenshluger and Lepski (2011). The resulting estimator achieves the same performance as the one which would have been selected if the regularity index of the target function had been known, up to a constant and to a logarithm factor, as it is proved in our main result, Theorem 1. The result holds whatever the sample size. Convergence rates are also deduced in Section 3.4 for functions m belonging to Hölder spaces, and under various concentration assumptions on the process X. These assumptions are on the rate of decay of the small ball probability

$$\varphi^{x_0} : h > 0 \mapsto \mathbb{P}(\|X - x_0\| \leq h),\ x_0 \in \mathbb{H},$$

which influences the rate, as usual. The faster the small ball probability decreases, the smaller the rate of convergence of the estimator is. The rates (Proposition 2) are quite slow when X really lives in an infinite-dimensional space. They nevertheless match with the lower bounds of Mas (2012). This is the so-called "curse of dimensionality". Practical issues are discussed in Section 4 and the performances of our bandwidth selection criterion are compared with cross-validated criteria. Finally, the proofs are gathered in Section 5. Notice that we do not discuss in this work the choice of the norm $\|.\|$ in the kernel estimators (2): we have shown in Chagny and Roche (2014) that the alternative choice of projection semi-norm does not improve the convergence rates, at least under our regularity assumption (H_m defined below). Alternative regularity assumptions with data-driven semi-norms in the kernel could improve in practice the performances of kernel estimators. This is, to our opinion, an interesting perspective but this is far beyond the scope of this paper.

2. Risk of an estimator with fixed bandwidth

We provide in this section upper-bounds for the pointwise risk of any estimator (2) with fixed bandwidth h.

2.1. **Assumptions.** The result will be obtained under the following assumptions.

(H_K) The kernel K is of type I (Ferraty and Vieu, 2006) i.e. its support is in $[0,1]$ and there exist two constants $c_K, C_K > 0$ such that

$$c_K \mathbf{1}_{[0,1]} \leq K \leq C_K \mathbf{1}_{[0,1]}.$$

(H_m) There exists $\beta \in]0,1]$, and a constant $C_m > 0$ such that, for all $x, x' \in \mathbb{H}$,

$$|m(x) - m(x')| \leq C_m \|x - x'\|^\beta.$$

Assumption (H_K) is quite classical in kernel methods for functional data (see Ferraty et al. 2006; Burba et al. 2009; Ferraty et al. 2010). We are aware that this is a strong constraint on the choice of the kernel but alleviate it in a functional data context requires a lot of technical difficulties and it is still, to our knowledge, an open problem. However, since the kernel is chosen by the statistician in practice, this is not a real problem. Assumption (H_m) is an Hölder-type regularity condition on the target function m to estimate, it is required to control the bias term of the risk.

2.2. **Upper bound for the risk.** The following theorem provides a bias-variance decomposition, in a non-asymptotic point of view.

Proposition 1. *If Assumption (H_m) is fulfilled, there exists a constant $C > 0$ which only depends on c_K and C_K such that, for all $h > 0$:*

$$\mathbb{E}\left[(\widehat{m}_h(x_0) - m(x_0))^2\right] \leq C\left(h^{2\beta} + \frac{\sigma^2}{n\varphi^{x_0}(h)}\right). \tag{3}$$

The first term of Inequality (3) is a bias term: it depends on the regularity of the function to estimate m and decreases when h decreases. The second term of Inequality (3) is a variance term which depends on the regularity of the process X through its small ball probability φ^{x_0}. This term increases when h decreases. This result is coherent with Ferraty et al. (2010, Theorem 2).

3. Adaptive estimation

We define in this section a data-driven bandwidth selection method, which leads to an adaptive estimator.

3.1. **Bandwidth selection.** We have at our disposal the estimators $\widehat{m}_h$ defined by (2) for any $h > 0$. Let $\mathcal{H}_n$ be a finite collection of bandwidths, with cardinality depending on n and properties precised below. Denote $h_{\max} = \max \mathcal{H}_n$, $h_{\min} = \min \mathcal{H}_n$. Our aim is to select a bandwidth $\widehat{h}$ in the collection, only from the data, so as to obtain a resulting estimator which has a minimal risk, up to some constants or negligible terms, among the collection $(\widehat{m}_h)_{h \in \mathcal{H}_n}$.

If one has access to the smoothness index β of m (see Assumption (H_m)) and to the knowledge of the small ball probability $\varphi^{x_0}(h)$, the best choice of bandwidth among the collection for a given curve $x_0 \in \mathbb{H}$ would be

$$h^*(x_0) = \arg\min_{h \in \mathcal{H}_n} \left\{h^{2\beta} + \frac{\sigma^2}{n\varphi^{x_0}(h)}\right\}, \tag{4}$$

thanks to the decomposition (3) in Proposition 1. This ideal choice, which realizes the best trade-off between the bias and variance terms is called the "oracle". It cannot be used in practice since it depends both on β and $\varphi^{x_0}(h)$ which are generally unavailable for the statistician. The method we define below is fully data-driven, and is in the spirit of the so-called Lepski methods: the most recent one, developed by Goldenshluger and Lepski (2011), permits to derive the nonasymptotic results we want to obtain. The main idea is the following. Since the oracle h^* depends on unknown quantities, let us "mimic" its choice by selecting

$$\widehat{h}(x_0) = \text{argmin}_{h\in\mathcal{H}_n}\left\{\widehat{A}(h,x_0)+\widehat{V}(h,x_0)\right\}, \tag{5}$$

where $\widehat{V}(h,x_0)$ is an empirical counterpart for the variance term, and $\widehat{A}(h,x_0)$ is an approximation of the bias term. To define them, let us first introduce an empirical version for the shifted small ball probability $\varphi^{x_0}(h)=\mathbb{P}(\|X-x_0\|\leq h)$ as follows

$$\widehat{\varphi}^{x_0}(h) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\|X_i-x_0\|\leq h\}}. \tag{6}$$

It permits to set

$$\widehat{V}(h,x_0) = \begin{cases}\kappa\sigma^2\dfrac{\ln(n)}{n\widehat{\varphi}^{x_0}(h)} \text{ if } \widehat{\varphi}^{x_0}(h)\neq 0\\ +\infty \text{ otherwise,}\end{cases} \tag{7}$$

where κ is a constant specified in the proofs which depends neither on h, nor on n, nor on m which is, in practice, fixed once and for all (see Section 4.1.1). The idea of Goldenshluger and Lepski (2011) is to define $\widehat{A}(h,x_0)$ by comparing two by two the estimators with fixed bandwidth. In our functional context, the comparison can be done as follows:

$$\widehat{A}(h,x_0) = \max_{h'\in\mathcal{H}_n}\left((\widehat{m}_{h'}(x_0)-\widehat{m}_{h\vee h'}(x_0))^2-\widehat{V}(h',x_0)\right)_+. \tag{8}$$

This quantity is proved to be an approximation of the bias term (see Lemma 4). This motivates the following choice of the bandwidth $\hat{h}(x_0)$, and the study of the selected estimator, at the point x_0 is $\widehat{\widehat{m}}(x_0) := \widehat{m}_{\widehat{h}(x_0)}(x_0)$: the theoretical results below establish that it really mimics the oracle. Notice that similar selection criteria, which are based on the used of auxiliary estimators with the bandwidth $h\vee h'$ have been recently used by Chagny and Roche (2014) in a functional setting but for global selection purpose (the selected bandwidth does not depend on the curve x_0), or by Rebelles (2014) for a pointwise slection in multivariate density estimation.

3.2. **Assumptions.** We consider the following assumptions, in addition to the ones introduced in Section 2.1.

$(H_{\mathcal{H}_n})$ The collection $\mathcal{H}_n$ of bandwidths is such that:
- $(H_{\mathcal{H}_n,1})$ its cardinality is bounded by n,
- $(H_{\mathcal{H}_n,2})$ for any $h\in\mathcal{H}_n$, $\varphi^{x_0}(h)\geq C_0\ln(n)/n$, where $C_0>0$ is a purely numerical constant (specified in the proofs).

(H_ε) For any integer $l\geq 2$, $\mathbb{E}[|\varepsilon_1|^l]\leq C_\varepsilon^l\sigma^l l!/2$, for $C_\varepsilon>1$ a constant.

Assumption $(H_{\mathcal{H}_n})$ means that the bandwidth collection should not be too large. Assumption (H_ε) is an integrability condition on the error. Bounded noise or Gaussian noise satisfy such kind of assumption.

3.3. **Theoretical results.**

Theorem 1. *Assume (H_K), (H_m), $(H_{\mathcal{H}_n})$ ($(H_{\mathcal{H}_n,1})$ and $(H_{\mathcal{H}_n,2})$) and (H_ε). There exist two constants $c, C > 0$ depending on c_K, C_K, C_0, C_m, C_ε, $h_{\max}$, κ such that*

$$\mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0) - m(x_0)\right)^2\right] \leq c \min_{h\in\mathcal{H}_n}\left\{h^{2\beta} + \sigma^2 \frac{\ln(n)}{n\varphi^{x_0}(h)}\right\} + \frac{C}{n}. \tag{9}$$

This result proves that our selection rule (5) which permits to define $\widehat{\widehat{m}}(x_0)$ really mimics the oracle one (4). The selected estimator performs as well as the pseudo-estimator $\widehat{m}_{h^*(x_0)}$. The right-hand-side of (9) actually corresponds to the best compromise between the bias and the variance terms of the risk, since the term C/n is negligible with respect to the first term of the right-hand-side of (9). Our procedure is thus adaptive, the selected estimator automatically adapts itself to the unknown smoothness of the target function m. We just have a logarithmic loss in the rate, by comparing the right-hand-side of (9) with (3). The loss is due to adaptation, and is known to be nevertheless adaptive optimal in many estimation problem studied with a pointwise risk. Notice also that the same phenomenon also occured to estimate the conditional cumulative distribution function in the same functional context, but with an integrated criterion, see Chagny and Roche (2014). We see below that it does not affect the convergence rates in most of the cases.

Remark 1. The estimated variance term $\widehat{V}(h, x_0)$ (see (7)) of the selection criterion cannot be used in practice, since it depends on the unknown noise variance $\sigma^2 = \mathbb{E}[\varepsilon_1^2]$. A solution consists in replacing it by an estimator, and to prove that the estimator selected with the new criterion still satisfies the adaptation property. The plug-in does not lead to specific difficulties, and an upper-bound for the risk similar to (9) can be easily proved. We do not give all the theoretical details, since similar results and proofs can be found in Brunel and Comte (2005, Theorem 3.4) (for classical model selection method) and Chagny and Lacour (2014, Theorem 3) (for selection criterion in the spirit of Lepski's methodology). However, all the practical explanations can be found in Section 4.1.2.

3.4. **Rates of convergence.** Using the results of Theorem 1, we compute the rate of decrease of the risk of the selected estimator. Since the the variance term in the upper-bound on the risk (Inequality (9)) depends on the small ball probability $\varphi^{x_0}(h)$, the resulting rate will be related to its rate of decay when $h \to 0$.

We proceed as in Chagny and Roche (2014). Let us define three classes of processes (we denote by c_1, C_1 and c_2 some nonnegative constants):

$H_{X,L}$ There exist some constants $\gamma_1, \gamma_2 \in \mathbb{R}$, and $\alpha > 0$ such that
$c_1 h^{\gamma_1}\exp(-c_2 h^{-\alpha}) \leq \varphi^{x_0}(h) \leq C_1 h^{\gamma_2}\exp(-c_2 h^{-\alpha})$;

$H_{X,M}$ There exist some constants $\gamma_1, \gamma_2 \in \mathbb{R}$, and $\alpha > 1$, such that
$c_1 h^{\gamma_1}\exp(-c_2 \ln^\alpha(1/h)) \leq \varphi^{x_0}(h) \leq C_1 h^{\gamma_2}\exp(-c_2 \ln^\alpha(1/h))$;

$H_{X,F}$ There exists a constant $\gamma > 0$, such that $c_1 h^\gamma \leq \varphi^{x_0}(h) \leq C_1 h^\gamma$.

The class $H_{X,F}$ is typically the class of finite-dimensional processes. We can see this with the Karhunen-Loève decomposition of X,

$$X = \sum_{j\geq 1} \sqrt{\lambda_j}\xi_j\psi_j, \tag{10}$$

which is simply the decomposition of X in the basis $(\psi_j)_{j\geq 1}$ of the eigenfunctions of the covariance operator $\Gamma : f \mapsto \mathbb{E}\left[\langle f, X\rangle X\right]$. Here, $(\xi_j)_{j\geq 1}$ is a sequence of uncorrelated centred standard random variables (the standardised principal component scores) and $(\lambda_j)_{j\geq 1}$ is the sequence of eigenvalues associated to the eigenfunctions $(\psi_j)_{j\geq 1}$. The series (10)

converges in terms of the norm $\|\cdot\|$ of $\mathbb{H}$. Typically, if there is only a finite number of eigenvalues $(\lambda_j)_{j\geq 1}$ which are non null, under mild assumptions on the law of X, we can prove that $H_{X,F}$ is verified and, in that case, X lies a.s. in the finite-dimensional space $\text{span}\{\psi_j,\ \lambda_j > 0\}$ (Chagny and Roche, 2014) (note that, since Γ is a positive operator, $\lambda_j \geq 0$ for all $j \geq 1$). Conversely, Mas (2012, Corollary 1, p. 10) proved that, under general conditions satisfied e.g. by Gaussian processes, the variable X can not verifies Assumption $H_{X,F}$ if the set $\{j,\ \lambda_j > 0\}$ is infinite.

A process X belongs to the class $H_{X,M}$ typically when the eigenvalues $(\lambda_j)_{j\geq 1}$ decrease at an exponential rate. For instance, in the case of a Gaussian process with $c\exp(-2j)/j \leq \lambda_j \leq C\exp(-2j)/j$, we have $c_2 = 1/2$ and $\alpha = 2$ in $H_{X,M}$ (Hoffmann-Jørgensen et al. 1979, Theorem 4.4 and example 4.7, pp. 333 and 336).

Finally, the class $H_{X,L}$ contains the processes such that the eigenvalues $(\lambda_j)_{j\geq 1}$ decrease at a polynomial rate (Hoffmann-Jørgensen et al., 1979, Theorem 4.4 and example 4.5, pp. 333 and 334). This is the case for instance for the Brownian motion (Ash and Gardner, 1975, pp. 41 and 42) or the Brownian bridge (MacNeill, 1978).

Proposition 2. *Suppose that all the assumptions of Theorem 1 are verified. Define, for $\beta \in]0,1]$ and $C > 0$, $\mathcal{F}_\beta^C$ be the following functional classes:*

$$\mathcal{F}_\beta^C = \left\{m : \mathbb{H} \to \mathbb{R}, |m(x) - m(x')| \leq C\|x - x'\|^\beta\right\}.$$

- *If Assumption $H_{X,L}$ is verified, then*

$$\sup_{m\in\mathcal{F}_\beta^C} \mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0) - m(x_0)\right)^2\right] \leq C_1 \left(\ln n\right)^{-\beta/\alpha}.$$

- *If Assumption $H_{X,M}$ is verified, then*

$$\sup_{m\in\mathcal{F}_\beta^C} \mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0) - m(x_0)\right)^2\right] \leq C_2 \exp\left(-\frac{2\beta}{c_2^{1/\alpha}} \ln^{1/\alpha}(n)\right).$$

- *If Assumption $H_{X,F}$ is verified, then*

$$\sup_{m\in\mathcal{F}_\beta^C} \mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0) - m(x_0)\right)^2\right] \leq C_3 \left(\frac{\ln n}{n}\right)^{2\beta/(2\beta+\gamma)}.$$

Here, C_1, C_2 and C_3 are positive real numbers independent of n.

Since the bias-variance decomposition is very similar for the regression function estimation (Inequality (9)) and for the conditional cumulative distribution function estimation, we can easily adapt the proof of Chagny and Roche (2014, Section A.1.1) to prove Proposition 2.

The rates of decay of the risk are quite slow. Mas (2012, Theorem 3 and Corollary 2) has obtained very similar minimax convergence rates with different assumptions. Moreover, although slow, these rates are proved to be the optimal for conditional cumulative distribution function estimation in Chagny and Roche (2014) under the same assumptions on the small ball probability.

4. Numerical results

Our aim in this section is to illustrate the behaviour of our bandwidth selection device studied above and to compare its practical performances with cross-validation methods, to select a kernel estimator of the regression function. We first explain how to implement

our procedure, and detail specifically how to tune the constant appearing in the term $\widehat{V}(h, x_0)$ (which can be compared to a penalty constant).

4.1. **Implementation of the estimator.** For the simulation study, we choose the uniform kernel

$$K(t) = \mathbf{1}_{[0,1[},$$

and the following bandwidth collection

$$\mathcal{H}_n := \{h_{\max}/k, k = 1, \dots, k_{\max}\}.$$

The norm $\|\cdot\|$ is the usual norm of $\mathbb{L}^2([0,1])$: $\|f\|^2 := \int_0^1 f(t)^2 dt$, for all $f \in \mathbb{L}^2([0,1])$.

With the definition of the kernel K, we see that, if $h > \max\{\|X_i - x_0\|, i = 1, \dots, n\}$, $K_h(\|X_i - x_0\|) = 1$. Hence, there is no need to consider bandwidths which are larger than $\max\{\|X_i - x_0\|, i = 1, \dots, n\}$: we thus set $h_{\max} = \max\{\|X_i - x_0\|, i = 1, \dots, n\}$. Note that a bandwidth collection depending on $\max\{\|X_i - x_0\|, i = 1, \dots, n\}$ is also considered in Amiri et al. (2014). Similarly, if h is taken too small, there are too few observations X_i in the ball $\{x \in \mathbb{H}, \|x - x_0\| < h\}$ and the variance of the estimator $\widehat{m}_h(x_0)$ is large. To prevent this, we fix $k_{\max}$ such that $\widehat{\varphi}^{x_0}(h_{\max}/k_{\max}) \geq \ln n/n$ or, worded in a different way, $k_{\max}$ is the largest integer such that $h_{\max}/k_{\max}$ is greater than the empirical quantile of $\{\|X_i - x_0\|\}_{i=1,\dots,n}$ associated to the probability $\ln n/n$.

4.1.1. *Calibration of the constant appearing in the term $\widehat{V}(h, x_0)$.* The constant κ appearing in the term $\widehat{V}(h, x_0)$ is a universal constant in the sense that it does not depend on the model parameters or on the estimation parameters. Hence, we must fix it once and for all.

The method we propose is to evaluate the performances of our estimator on a grid of different values of κ, for different sets of parameters, and to keep the value of κ for which our estimator seems to have a reasonable mean squared error[1]

$$MSE := \mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0) - m(x_0)\right)^2\right].$$

This calibration method is quite classical (see for instance Bertin et al. 2014; Comte and Johannes 2012).

The functional covariate is simulated in the following way:

$$X(t) := \xi_0 + \sum_{j=1}^{J} \xi_j \sqrt{\lambda_j} \psi_j(t),$$

where, for all $j \geq 1$, $\psi_j(t) := \sqrt{2}\sin(\pi(j - 0.5)t)$, $(\lambda_j)_{j\geq 1}$ is a sequence of positive real numbers such that $\sum_{j\geq 1} \lambda_j < +\infty$ and $(\xi_j)_{j\geq 0}$ is an *i.i.d.* sequence of standard random variables and J is a positive integer. In the sequel, we consider three different settings:

(a) $J = 150$ and $\lambda_j = j^{-2}$ for all $j \geq 1$. Here, J is sufficiently large to consider that

$$\sum_{j=1}^{J} \xi_j \sqrt{\lambda_j} \psi_j(t) \approx \sum_{j\geq 1} \xi_j \sqrt{\lambda_j} \psi_j(t).$$

Moreover, the choice of the sequence $(\lambda_j)_{j\geq 1}$, ensures that $\sum_{j\geq 1} \xi_j \sqrt{\lambda_j} \psi_j(t)$ verifies $H_{X,L}$.

[1]Note that the optimal value of κ may vary in practice from one set of parameters to another. Then the idea is to realise a compromise between all the models and to take a "reasonable" value of the constant.

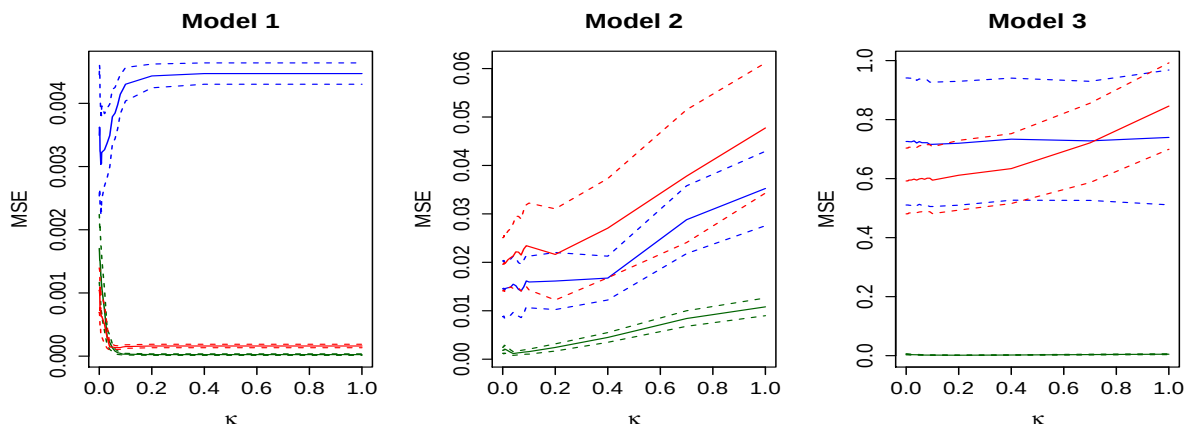

FIGURE 1. Plots of the empirical mean of the squared error calculated over 50 independent samples of size $n = 500$. Dotted lines: 95% confidence interval on the mean squared error. Blue curves: process (a), green curves: process (b) and red curves : process (c).

(b) $J = 150$ and $\lambda_j = e^{-j}/j$ for all $j \geq 1$ which corresponds to a process verifying $H_{X,M}$.

(c) $J = 2$ and $\lambda_j = j^{-2}$ for all $j \geq 1$ which corresponds to a process verifying $H_{X,F}$.

We consider also three regression models:

Model 1 : $Y = \left(\int_0^1 \sin(4\pi t)X(t)dt\right)^2 dt + \varepsilon$ (Chagny and Roche, 2014).

Model 2: $Y = \int_0^1 |X(t)| \ln(|X(t)|)dt + \varepsilon$ (Ferraty and Vieu, 2002).

Model 3: $Y = \int_0^1 X(t)^2 dt + \varepsilon$ (Amiri et al., 2014).

For the three models, we set $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.01$.

Figure 1 represents the empirical mean of the error as a function of κ. We choose the value $\kappa = 0.1$ for which the mean squared error of the estimator seems to be reasonable for all the chosen sets of parameters.

4.1.2. *Estimation of the noise variance* σ^2. Recall that our bandwidth selection criterion (5) depends on the noise variance σ^2, see details in Remark 1. In the simulation study, we first consider that the noise variance is known. However, in practice, we rarely know its value. Hence, we propose a plug-in estimator

$$\widehat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \widehat{m}_{h_{\min}}(X_i)\right)^2.$$

The idea beyond this proposition is the following: we want to propose an estimator which is as close as possible of the quantity $\sum_{i=1}^n \varepsilon_i^2/n$ (we recall that the ε_i's are not observed). We have

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - m(X_i)\right)^2,$$

and we replace $m(X_i)$ by an estimator $\widehat{m}_h(X_i)$ with a bandwidth $h \in \mathcal{H}_n$. A numerical study then allows us to establish that the more h is small, the more the quantity $\sum_{i=1}^n \varepsilon_i^2/n = \sum_{i=1}^n \left(Y_i - m_h(X_i)\right)^2/n$ is close to σ^2.

The estimator $\widehat{m}_{\widehat{h}(x_0)}(x_0)$ with the bandwidth selected by the criterion (5) with σ^2 replaced by $\widehat{\sigma}^2$ is denoted $\widehat{\widehat{m}}^{(ev)}(x_0)$. Note that, in the simulation study, the replacement

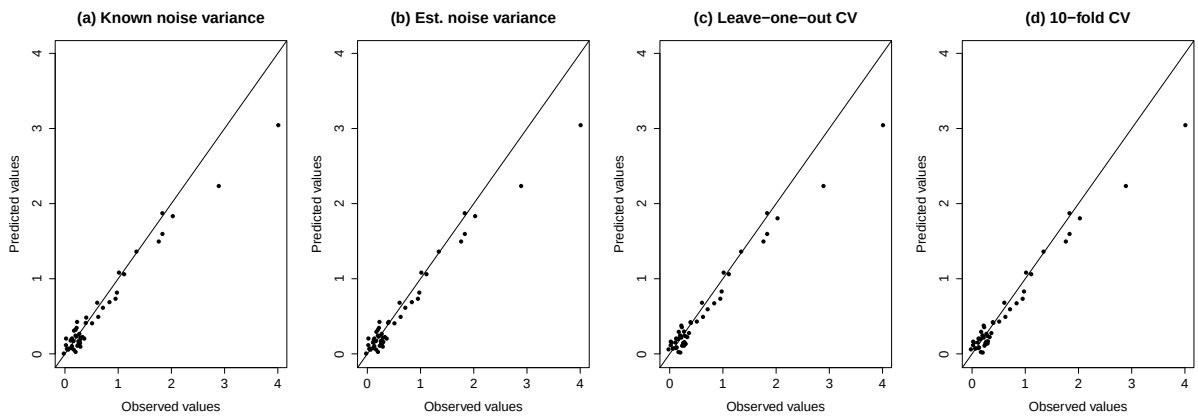

FIGURE 2. Plot of the predicted values $\widehat{Y}^{(j)}$ versus the observed values $Y^{(j)}$. Figure (a): $\widehat{Y}^{(j)} = \widehat{\widehat{m}}(X_i^{(j)})$. Figure (b): $\widehat{Y}^{(j)} = \widehat{\widehat{m}}^{(ve)}(X_i^{(j)})$. Figures (c) and (d): predictions made with the estimator selected by leave-one-out cross-validation and 10-fold cross-validation. Model 3 with X a standard Brownian motion, $n = 500$.

of σ^2 by the plug-in estimator $\widehat{\sigma}^2$ does not influence significantly the performances of the estimator (see Table 1 and Figures 2 (a) and 2 (b)).

4.2. **Simulation results.** We study here the behaviour of our estimator in two settings with Model 3, presented in Subsection 4.1.1:

First setting: The covariate X is simulated with the framework (b) (see Subsection 4.1.1) and the noise variance is fixed to $\sigma^2 = 0.2$.

Second setting: X is a standard Brownian motion and $\sigma^2 = 0.01$.

We made a Monte-Carlo study over 50 replications of the sample $(X_i, Y_i)_{1\leq i\leq n}$ and the curve $x_0 = X_{n+1}$. We compare the performances of our selected kernel estimators $\widehat{\widehat{m}}(x_0)$ and $\widehat{\widehat{m}}^{(ev)}(x_0)$ (bandwidth selected with the method introduced above in the spirit of Goldenshluger and Lepski 2011), with the same kernel estimator with bandwidth selected by leave-one-out and 10-fold cross-validations (CV). Both cross-validated methods consist in splitting the sample in two sub-samples (a learning sample and a test sample), evaluating the estimator at the points $x_0 = X_i$ for all X_i in the test sample using only the data of the learning sample and comparing the predicted values $\widehat{Y}_i^{(-ts)}$ of the Y_i's of the test sample with their real values. This procedure is repeated n times for the leave-one-out CV (the test sample contains only one observation) and 10 times for the 10-fold CV (the test sample contains about $n/10$ observations), in such a way that each observation is in the test sample only one time during the procedure. Finally, the selected bandwidth is the one which minimizes the mean squared prediction error : $\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i^{(-ts)}\right)^2/n$.

4.2.1. *Predicted values.* Figure 2 presents the plot of the predicted values versus the observed values. For the four bandwidth selection devices, the prediction for a new value x_0 is quite accurate for most of the 50 evaluations of the estimators. Some values $m(x_0)$ are badly predicted: these are isolated values, which means that only a very small number of curves X_i's in the sample are close to x_0. This kind of isolated values is hard to estimate with kernel methods (which are local). Few differences are visible between the different bandwidth selection devices.

	First setting Model 3, simulation (b)			**Second setting** Model 3, X brownian motion		
	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
signal-to-noise ratio (snr)	0.055			0.016		
Known noise variance ($\widehat{\widehat{m}}(x_0)$)	0.669	0.387	0.082	0.155	0.115	0.032
Est. noise variance ($\widehat{\widehat{m}}^{(ev)}(x_0)$)	0.669	0.392	0.084	0.155	0.115	0.032
Leave-one-out CV	0.656	0.370	0.080	0.156	0.115	0.032
10-fold CV	0.660	0.369	0.083	0.155	0.114	0.032

TABLE 1. Monte-Carlo study of the MSE over 50 replications.

	First setting Model 3, simulation (b)			**Second setting** Model 3, X brownian motion		
	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
Known noise variance ($\widehat{\widehat{m}}(x_0)$)	0.04	0.11	0.35	0.03	0.06	0.17
Est. noise variance ($\widehat{\widehat{m}}^{(ev)}(x_0)$)	0.39	1.50	9.45	0.33	1.24	7.95
Leave-one-out CV	3.35	17.55	147.61	1.82	8.77	71.20
10-fold CV	3.05	16.06	129.72	1.64	7.96	63.32

TABLE 2. Mean CPU time in seconds (calculated over 50 runs) necessary for the calculation of the estimator (including bandwidth selection).

4.2.2. *Mean Squared Error (MSE).* Table 1 presents the results of the Monte-Carlo study of the MSE of the fourth bandwidth selection devices presented here. We also give an estimation of the signal-to-noise ratio $snr := \sigma^2/\mathrm{Var}(m(X))$.

We first note that, as expected, the mean squared prediction error decreases when the sample size increases. Then, remark that the performances, in terms of mean squared error, of our bandwidth selection devices (with known or estimated noise variance), are very similar to the performances of cross-validated methods. This confirms the first findings we made in view of Figure 2.

4.2.3. *Computational time.* Table 2 presents the mean time necessary to calculate the estimator and to select the bandwidth. We give here the total CPU time (system time and user time) necessary to compute an estimate of the value of m at a point x_0. We performed these computations on a personal computer equipped with a processor Intel Core i5-4200, CPU: 1.60GHz (maximum speed: 2.30GHz), HD: 921Go, Memory: 7.88Go.

We observe that our bandwidth selection device outperforms cross-validation in terms of computation time: even with the estimation of the noise variance, which is time consuming, the computational time of the leave-one-out cross-validation is ranging from about 60 to 420 times that of our bandwidth selection device (50 to 370 times for the 10-fold cross-validation). These differences are due to the fact that cross-validated methods require the calculation of the estimator for each bandwidth and each curve X_i of the sample ($n \times |\mathcal{H}_n|$ calls to the estimation function, where $|\mathcal{H}_n|$ is the cardinality of the bandwidth collection $\mathcal{H}_n$) while our method necessitates only one evaluation of the estimator *per* bandwidth (plus n evaluation of the estimate but for only one bandwidth for the calculation of the noise variance, giving a total of $n + |\mathcal{H}_n|$ calls to the estimation function).

Note that the data-driven collection of bandwidths is larger (most of the time) for the simulation setting (b) than it is when X is a standard brownian motion (the small ball probability decreases faster to 0 when h goes to 0). This explains the observed differences of computational time between the two settings.

5. Proof

5.1. Preliminary results.

5.1.1. *The Bernstein Inequalities.* Let us recall concentration results for empirical process, on which our proofs are based. We shall first state the following useful lemma, which immediatly follows from (Birgé and Massart, 1998, p.366).

Lemma 1. *(Bernstein Inequality) Let $T_1, T_2, \ldots, T_n$ be independent random variables and $S_n(T) = \sum_{i=1}^n (T_i - \mathbb{E}[T_i])$. Assume that*

$$Var(T_1) \leq v^2 \text{ and } \forall l \geq 2, \ \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[|T_i|^l\right] \leq \frac{l!}{2} v^2 b_0^{l-2}.$$

Then, for $\eta > 0$,

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n}|S_n(T)| \geq \eta\right) &\leq 2\exp\left(-\frac{n\eta^2/2}{v^2 + b_0\eta}\right), \\ &\leq 2\max\left\{\exp\left(-\frac{n\eta^2}{4v^2}\right), \exp\left(-\frac{n\eta}{4b_0}\right)\right\}. \end{aligned} \tag{11}$$

We deduce from (11) the following result, the proof of which is based on integration and is omitted.

Lemma 2. *With the same notations and under the same assumptions as the ones of Lemma 1, for any $V > 0$,*

$$\begin{aligned} \mathbb{E}\left[\left(\left(\frac{S_n(T)}{n}\right)^2 - V\right)_+\right] &\leq 2\max\left\{\exp\left(\frac{-n\sqrt{V}}{4b_0}\right)\left(\frac{32b_0^2}{n^2} + \frac{8b_0\sqrt{V}}{n}\right),\right. \\ &\qquad \left.\exp\left(\frac{-nV}{4v^2}\right)\frac{4v^2}{n}\right\}. \end{aligned}$$

5.1.2. *The empirical process.* We apply in this section the previous concentration inequalities to bound the main empirical process involved in our context.

We first remark that Assumption (H_K) implies that, for all integer $l > 0$,

$$c_K^l \varphi^{x_0}(h) \leq \mathbb{E}\left[\left(K_h\left(\|X - x_0\|\right)\right)^l\right] \leq C_K^l \varphi^{x_0}(h). \tag{12}$$

One of the keys of the results is the control, in probability and expectation, of the quantity $R_h^{x_0}$ defined by:

$$R_h^{x_0} := \frac{1}{n}\sum_{i=1}^n \frac{K_h\left(\|X_i - x_0\|\right)}{\mathbb{E}\left[K_h\left(\|X_i - x_0\|\right)\right]}. \tag{13}$$

The following lemma states the bounds.

Lemma 3. *Supposed that (H_K) is fulfilled. Then, the following inequality holds,*

$$\mathbb{P}\left(|R_h^{x_0} - 1| > \frac{1}{2}\right) \leq 2\exp\left(-\frac{n\varphi^{x_0}(h)}{8\left(\frac{C_K^2}{c_K^2} + \frac{C_K}{2c_K}\right)}\right). \tag{14}$$

Moreover if $(H_{\mathcal{H}_n,2})$ is also satisfied, for all $\alpha > 0$,

$$\mathbb{E}\left[\left((R_h^{x_0} - 1)^2 - V_{R^{x_0}}(h)\right)_+\right] \leq \frac{C}{n^\alpha}, \tag{15}$$

with $V_{R^{x_0}}(h) := \kappa_R \ln(n)/(n\varphi^{x_0}(h))$, $\kappa_R > \max\left\{\left(\frac{4C_K\alpha}{c_K\sqrt{C_0}}\right)^2, 4\frac{C_K^2}{c_K^2}\alpha\right\}$ *and* $C > 0$ *depending only on* C_K, c_K, C_0 *and* κ_R.

Inequality (14) follows from Lemma 1, with $T_i = K_h(\|X_i - x_0\|)/\mathbb{E}[K_h(\|X_i - x_0\|)]$ ($\mathbb{E}[T_i] = 1$), which leads to $S_n(T)/n = R_h^{x_0} - 1$. The assumptions are fulfilled, with the parameters (computed mainly thanks to (12)),

$$v^2 = \frac{C_K^2}{c_k^2}\frac{1}{n\varphi^{x_0}(h)} \text{ and } b_0 = \frac{C_K}{c_k}\frac{1}{n\varphi^{x_0}(h)}.$$

Lemma 2 applied with the same process permits then to obtain Inequality (15).

5.2. **Proof of Theorem 1.** A cornerstone decomposition for our main proofs is the following, for $h \in \mathcal{H}_n$,

$$\begin{aligned}(\widehat{m}_h(x_0) - m(x_0))^2 &= \left(\sum_{i=1}^n W_h^{(i)}(x_0)Y_i - m(x_0)\right)^2 \\ &= \left(\sum_{i=1}^n W_h^{(i)}(x_0)m(X_i) + \sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i - m(x_0)\right)^2 \\ &= B_h(x_0) + T_h(x_0) \qquad (16)\end{aligned}$$

with

$$\begin{aligned}B_h(x_0) &= \left(\sum_{i=1}^n W_h^{(i)}(x_0)\left(m(X_i) - m(x_0)\right)\right)^2, \\ T_h(x_0) &= \left(\sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i\right)^2. \qquad (17)\end{aligned}$$

We recall that $\sum_{i=1}^n W_h^{(i)}(x_0) = 1$ and that the independence of the ε_i's with the X_i's implies

$$\mathbb{E}\left[\sum_{i,j=1}^n W_h^{(i)}(x_0)W_h^{(j)}(x_0)\left(m(X_i) - m(x_0)\right)\varepsilon_j\right] = 0,$$

which proves (16).

We first bound the term $B_h(x_0)$. By Assumption (H_m):

$$\begin{aligned}B_h(x_0) &\leq \left(\sum_{i=1}^n W_h^{(i)}(x_0)\left|m(X_i) - m(x_0)\right|\right)^2 \\ &\leq \left(\sum_{i=1}^n W_h^{(i)}(x_0)C_m\|X_i - x_0\|^\beta\right)^2. \qquad (18)\end{aligned}$$

Now we can remark that, by definition of $W_h^{(i)}(x_0)$ and the fact that K is supported by $[0,1]$ (Assumption (H_K)), we have $W_h^{(i)}(x_0) = 0$ if $\|X_i - x_0\| > h$. This and Inequality (18) implies that

$$B_h(x_0) \leq C_m^2 h^{2\beta}\left(\sum_{i=1}^n W_h^{(i)}(x_0)\right)^2 = C_m^2 h^{2\beta}. \qquad (19)$$

We turn now to the term $T_h(x_0)$:

$$\begin{aligned}\mathbb{E}\left[T_h(x_0)\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}<1/2\right\}}\right] \\ &\quad +\mathbb{E}\left[\left(\sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}\geq 1/2\right\}}\right]. \end{aligned} \tag{20}$$

For the first term of Equation (20), remark that, by independence of ε_i with ε_j (for $i \neq j$) and $X_1, \ldots, X_n$, and since ε is centered,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}<1/2\right\}}\right] &= \sum_{i,j=1}^n \mathbb{E}\left[W_h^{(i)}(x_0)W_h^{(j)}(x_0)\varepsilon_i\varepsilon_j \mathbf{1}_{\left\{R_h^{x_0}<1/2\right\}}\right] \\ &= \sigma^2 \sum_{i=1}^n \mathbb{E}\left[\left(W_h^{(i)}(x_0)\right)^2 \mathbf{1}_{\left\{R_h^{x_0}<1/2\right\}}\right].\end{aligned}$$

Now we have, since the quantities $\left(K_h\left(\|X_i - x_0\|\right)\right)_{1\leq i\leq n}$ are non-negative (Assumption (H_K)),

$$\sum_{i=1}^n \left(W_h^{(i)}(x_0)\right)^2 = \frac{\sum_{i=1}^n K_h^2\left(\|X_i - x_0\|\right)}{\left(\sum_{i=1}^n K_h\left(\|X_i - x_0\|\right)\right)^2} \leq 1 \text{ a.s.}$$

This implies that

$$\mathbb{E}\left[\left(\sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}<1/2\right\}}\right] \leq \sigma^2 \mathbb{P}\left(R_h^{x_0} < 1/2\right) \leq \sigma^2 \mathbb{P}\left(|R_h^{x_0} - 1| > 1/2\right).$$

Hence, by Lemma 3, we obtain

$$\mathbb{E}\left[\left(\sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}<1/2\right\}}\right] \leq 2\sigma^2 \exp\left(-\frac{n\varphi^{x_0}(h)}{8\left(\frac{C_K^2}{c_K^2} + \frac{C_K}{c_K}\right)}\right) \leq \frac{C\sigma^2}{n\varphi^{x_0}(h)}, \tag{21}$$

where $C = 16e^{-1}\left(\frac{C_K^2}{c_K^2} + \frac{C_K}{c_K}\right)$ by the fact that $xe^{-x} \leq e^{-1}$ for all $x > 0$.

We turn now to the second term of Equation (20). By definition of $W_h^{(i)}(x_0)$ and $R_h^{x_0}$

$$\begin{aligned}&\mathbb{E}\left[\left(\sum_{i=1}^n W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}\geq 1/2\right\}}\right] \\ &\qquad = \mathbb{E}\left[\left(\sum_{i=1}^n \frac{K_h\left(\|X_i - x_0\|\right)}{n\mathbb{E}\left[K_h\left(\|X - x_0\|\right)\right]}\frac{1}{R_h^{x_0}}\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}\geq 1/2\right\}}\right], \\ &\qquad \leq 4\mathbb{E}\left[\left(\sum_{i=1}^n \frac{K_h\left(\|X_i - x_0\|\right)}{n\mathbb{E}\left[K_h\left(\|X - x_0\|\right)\right]}\varepsilon_i\right)^2\right].\end{aligned}$$

With the same arguments as the ones used for the first term of (20),

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}\geq 1/2\right\}}\right] & \\ &\leq \frac{4}{n^2\left(\mathbb{E}\left[K_h\left(\|X-x_0\|\right)\right]\right)^2}\mathbb{E}\left[\sum_{i=1}^{n} K_h^2\left(\|X_i-x_0\|\right)\varepsilon_i^2\right], \\ &= \frac{4}{n}\mathbb{E}\left[\left(\frac{K_h\left(\|X-x_0\|\right)}{\mathbb{E}\left[K_h\left(\|X-x_0\|\right)\right]}\varepsilon\right)^2\right].\end{aligned}$$

We conclude with Equation (12):

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} W_h^{(i)}(x_0)\varepsilon_i\right)^2 \mathbf{1}_{\left\{R_h^{x_0}\geq 1/2\right\}}\right] \leq \frac{4C_K^2}{c_K^2}\frac{\sigma^2}{n\varphi^{x_0}(h)}. \tag{22}$$

Now Equation (16) and equations (19), (21) and (22) lead us to the expected result.

5.3. **Proof of Theorem 1.** We consider the set

$$\Lambda^{x_0} = \bigcap_{h\in\mathcal{H}_n}\left\{\left|\frac{\widehat{\varphi}^{x_0}(h)}{\varphi^{x_0}(h)}-1\right|<\frac{1}{2}\right\}.$$

To prove Theorem 1, we study the loss function $(\widehat{\widehat{m}}(x_0)-m(x_0))^2$ on the set Λ^{x_0}, and on its complementary $(\Lambda^{x_0})^c$.

- **Step 1. Upper bound for $(\widehat{\widehat{m}}(x_0)-m(x_0))^2\mathbf{1}_{\Lambda^{x_0}}$.**
  Let $h\in\mathcal{H}_n$ be a fixed bandwidth. We first split

$$\begin{aligned}\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2 &\leq 3\left(\widehat{m}_{\widehat{h}(x_0)}(x_0)-\widehat{m}_{\widehat{h}(x_0)\vee h}(x_0)\right)^2+3\left(\widehat{m}_{\widehat{h}(x_0)\vee h}(x_0)-\widehat{m}_h(x_0)\right)^2 \\ &\quad +3\left(\widehat{m}_h(x_0)-m(x_0)\right)^2.\end{aligned}$$

We deduce from the definitions of $\widehat{A}(h,x_0)$, $\widehat{A}(\widehat{h}(x_0),x_0)$ and $\widehat{h}(x_0)$ that

$$\begin{aligned}& 3\left(\widehat{m}_{\widehat{h}(x_0)}(x_0)-\widehat{m}_{\widehat{h}(x_0)\vee h}(x_0)\right)^2+3\left(\widehat{m}_{\widehat{h}(x_0)\vee h}(x_0)-\widehat{m}_h(x_0)\right)^2 \\ \leq\ & 3\left(\widehat{A}(h,x_0)+\widehat{V}\left(\widehat{h}(x_0),x_0\right)\right)+3\left(\widehat{A}\left(\widehat{h}(x_0),x_0\right)+V(h,x_0)\right), \\ \leq\ & 6\left(\widehat{A}\left(h,x_0\right)+\widehat{V}\left(h,x_0\right)\right).\end{aligned}$$

Thus,

$$\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2\mathbf{1}_{\Lambda^{x_0}} \leq \left\{6\widehat{A}(h,x_0)+6\widehat{V}(h,x_0)+3\left(\widehat{m}_h(x_0)-m(x_0)\right)^2\right\}\mathbf{1}_{\Lambda^{x_0}}.$$

The idea is now to come down to the case of known small ball probability. To that aim, we define

$$V(h,x_0)=\frac{2}{3}\kappa\sigma^2\frac{\ln(n)}{n\varphi^{x_0}(h)},\ A(h,x_0)=\max_{h'\in\mathcal{H}_n}\left(\left(\widehat{m}_{h'}(x_0)-\widehat{m}_{h\vee h'}(x_0)\right)^2-V(h',x_0)\right)_+. \tag{23}$$

Compared to the data-driven counterparts (7) and (8), the variance term $V(h,x_0)$ is deterministic here. We then split

$$\begin{aligned}\widehat{A}(h,x_0) &= \max_{h'\in\mathcal{H}_n,\, \widehat{V}(h',x_0)<\infty} \left\{(\widehat{m}_{h\vee h'}(x_0)-\widehat{m}_{h'}(x_0))^2-\widehat{V}(h',x_0)\right\}_+,\\ &\leq \max_{h'\in\mathcal{H}_n,\, \widehat{V}(h',x_0)<\infty} \left\{(\widehat{m}_{h\vee h'}(x_0)-\widehat{m}_{h'}(x_0))^2-V(h',x_0)\right\}_+\\ &\quad+\max_{h'\in\mathcal{H}_n,\, \widehat{V}(h',x_0)<\infty}\left(V(h',x_0)-\widehat{V}(h',x_0)\right)_+,\\ &\leq A(h,x_0)+\max_{h'\in\mathcal{H}_n}\left(V(h',x_0)-\widehat{V}(h',x_0)\right)_+,\end{aligned}$$

which gives

$$\begin{aligned}\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2\mathbf{1}_{\Lambda^{x_0}} &\leq \Big\{6A(h,x_0)+6V(h,x_0)+3\left(\widehat{m}_h(x_0)-m(x_0)\right)^2\\ &+6\max_{h'\in\mathcal{H}_n}\left(V(h',x_0)-\widehat{V}(h',x_0)\right)_+ +6\left(\widehat{V}(h,x_0)-V(h,x_0)\right)\Big\}\mathbf{1}_{\Lambda^{x_0}}. \qquad (24)\end{aligned}$$

But we have, on the set Λ^{x_0}, for any $h'\in\mathcal{H}_n$, $|\widehat{\varphi}^{x_0}(h')-\varphi^{x_0}(h')|<\varphi^{x_0}(h')/2$. In particular, we thus have $\widehat{\varphi}^{x_0}(h')-\varphi^{x_0}(h')<\varphi^{x_0}(h')/2$, that is $\widehat{\varphi}^{x_0}(h')<(3/2)\varphi^{x_0}(h')$. This proves that $V(h',x_0)-\widehat{V}(h',x_0)<0$, and hence $\max_{h'\in\mathcal{H}_n}\left(V(h',x_0)-\widehat{V}(h',x_0)\right)_+=0$. Moreover, on Λ^{x_0},

$$\begin{aligned}\widehat{V}(h,x_0)-V(h,x_0) &= \frac{2}{3}\kappa\sigma^2\frac{\ln(n)}{n}\left(\frac{3}{2}\frac{1}{\widehat{\varphi}^{x_0}(h)}-\frac{1}{\varphi^{x_0}(h)}\right),\\ &\leq \frac{2}{3}\kappa\sigma^2\frac{\ln(n)}{n}2\frac{1}{\varphi^{x_0}(h)}=2V(h,x_0).\end{aligned}$$

Gathering the two bounds in (24), and using that the expectation of the loss of the estimator $\widehat{m}_h$ in the right-hand-side of (24) has already been bounded (see Proposition 1) lead to

$$\mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2\mathbf{1}_{\Lambda^{x_0}}\right] \leq 6\mathbb{E}\left[A\left(h,x_0\right)\right]+18V\left(h,x_0\right)+3C\left(h^{2\beta}+\frac{\sigma^2}{n\varphi^{x_0}(h)}\right),$$

where C is the constant involved in Equation (3).

The proof comes now down to establish an upper-bound for $\mathbb{E}[A(h,x_0)]$: we apply the following lemma (which proof is postponed to the following section).

Lemma 4. *Under the assumptions of Theorem 3, for any $h\in\mathcal{H}_n$, there exists $C>0$ such that*

$$\mathbb{E}\left[A\left(h,x_0\right)\right]\leq 2C_m^2h^{2\beta}+\frac{C}{n}.$$

The constant C depends on σ^2, c_K, C_K, C_0 and C_ε.

We thus obtain

$$\mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2\mathbf{1}_{\Lambda^{x_0}}\right]\leq c\left(h^{2\beta}+\sigma^2\frac{\ln(n)}{n\varphi^{x_0}(h)}\right)+\frac{C}{n}, \qquad (25)$$

with c and C two constants, c depending on C_m, σ^2, c_K and C_K, and C depending on σ^2, c_K, C_K, C_0 and C_ε.

• **Step 2. Upper bound for** $(\widehat{\widehat{m}}(x_0)-m(x_0))^2\mathbf{1}_{(\Lambda^{x_0})^c}$**.** We introduce the terms $B_h(x_0)$ and $T_h(x_0)$ defined by (17):

$$\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2\mathbf{1}_{(\Lambda^{x_0})^c}\leq\left(2B_{\widehat{h}(x_0)}(x_0)+2T_{\widehat{h}(x_0)}(x_0)\right)\mathbf{1}_{(\Lambda^{x_0})^c}.$$

Next, thanks to (19), $B_{\widehat{h}(x_0)}(x_0)\leq\max_{h'\in\mathcal{H}_n}B_{h'}(x_0)\leq\max_{h'\in\mathcal{H}_n}2C_m^2(h')^{2\beta}$. We deduce that

$$\mathbb{E}\left[B_{\widehat{h}(x_0)}(x_0)\mathbf{1}_{(\Lambda^{x_0})^c}\right]\leq 2C_m^2h_{\max}^{2\beta}\mathbb{P}\left((\Lambda^{x_0})^c\right).$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[T_{\widehat{h}(x_0)}(x_0)\mathbf{1}_{(\Lambda^{x_0})^c}\right] &\leq \mathbb{E}\left[\max_{h'\in\mathcal{H}_n}T_{h'}(x_0)\mathbf{1}_{(\Lambda^{x_0})^c}\right]\leq\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[T_{h'}(x_0)\mathbf{1}_{(\Lambda^{x_0})^c}\right],\\ &= \sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(\frac{\sum_{i=1}^n\varepsilon_iK_{h'}(\|X_i-x_0\|)}{\sum_{i=1}^nK_{h'}(\|X_i-x_0\|)}\right)^2\mathbf{1}_{(\Lambda^{x_0})^c}\right],\\ &= \sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\frac{\sum_{i=1}^n\varepsilon_i^2K_{h'}^2(\|X_i-x_0\|)}{\left(\sum_{i=1}^nK_{h'}(\|X_i-x_0\|)\right)^2}\mathbf{1}_{(\Lambda^{x_0})^c}\right],\\ &= \sigma^2\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\frac{\sum_{i=1}^nK_{h'}^2(\|X_i-x_0\|)}{\left(\sum_{i=1}^nK_{h'}(\|X_i-x_0\|)\right)^2}\mathbf{1}_{(\Lambda^{x_0})^c}\right],\\ &\leq \sigma^2\sum_{h'\in\mathcal{H}_n}\mathbb{P}\left((\Lambda^{x_0})^c\right)\leq\sigma^2n\mathbb{P}\left((\Lambda^{x_0})^c\right),\end{aligned}$$

where we have used the properties of ε_i, X_i and Assumption (H_K) in the same way as above, but also Assumption $(H_{\mathcal{H}_n,1})$. Therefore,

$$\mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2\mathbf{1}_{(\Lambda^{x_0})^c}\right]\leq\left(2C_m^2h_{\max}^{2\beta}+\sigma^2n\right)\mathbb{P}\left((\Lambda^{x_0})^c\right). \tag{26}$$

We are reduced to bound $\mathbb{P}((\Lambda^{x_0})^c)$:

$$\mathbb{P}((\Lambda^{x_0})^c)\leq\sum_{h\in\mathcal{H}_n}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\{\|X_i-x_0\|\leq h\}}-\mathbb{E}\left[\mathbf{1}_{\{\|X_i-x_0\|\leq h\}}\right]\right|\geq\frac{\varphi^{x_0}(h)}{2}\right).$$

We apply Bernstein's Inequality (Lemma 1), with $T_i=\mathbf{1}_{\{\|X_i-x_0\|\leq h\}}$ and $\eta=\varphi^{x_0}(h)/2$. Since $0\leq T_i\leq 1$, we set $b_0=1$, and $v^2=\mathrm{Var}(T_1)=\varphi^{x_0}(h)(1-\varphi^{x_0}(h))$. Thus,

$$\begin{aligned}\mathbb{P}((\Lambda^{x_0})^c) &\leq 2\sum_{h\in\mathcal{H}_n}\exp\left(-\frac{n(\varphi^{x_0}(h))^2/8}{\varphi^{x_0}(h)(1-\varphi^{x_0}(h))+\varphi(h)/2}\right)\\ &= 2\sum_{h\in\mathcal{H}_n}\exp\left(-\frac{n\varphi^{x_0}(h)}{8(1-\varphi^{x_0}(h))+4}\right)\leq 2\sum_{h\in\mathcal{H}_n}\exp\left(-\frac{n\varphi^{x_0}(h)}{12}\right).\end{aligned}$$

This leads to $\mathbb{P}((\Lambda^{x_0})^c)\leq 2n^{1-C_0/12}$ thanks to Assumptions $(H_{\mathcal{H}_n,1})$ and $(H_{\mathcal{H}_n,2})$. In (26), we obtain

$$\mathbb{E}\left[\left(\widehat{\widehat{m}}(x_0)-m(x_0)\right)^2\mathbf{1}_{(\Lambda^{x_0})^c}\right]\leq 2\left(2C_m^2h_{\max}^{2\beta}+\sigma^2\right)n^{2-C_0/12}, \tag{27}$$

$$\leq 2\left(2C_m^2h_{\max}^{2\beta}+\sigma^2\right)n^{-1}, \tag{28}$$

as soon as $C_0>36$.

It remains to sum (25) and (27) to obtain Theorem 1.

5.4. **Proof of Lemma 4.** First remark that the definition (8) of $A(h,x_0)$ can be written in the new way

$$A(h,x_0)=\max_{h'\in\mathcal{H}_n,h'\leq h}\left((\widehat{m}_{h'}(x_0)-\widehat{m}_h(x_0))^2-V(h',x_0)\right)_+.$$

For fixed bandwidths $h,h'\in\mathcal{H}_n$ $h\geq h'$, let us introduce $\Omega^{x_0}_{h,h'}=\{R^{x_0}_{h'}\geq 1/2\}\cap\{R^{x_0}_h\geq 1/2\}$ with R^{x_0} defined by (13). We split, for $h'\leq h$,

$$\begin{aligned}(\widehat{m}_{h'}(x_0)-\widehat{m}_h(x_0))^2 &\leq 2\left(\widehat{m}_{h'}(x_0)-m(x_0)\right)^2+2\left(\widehat{m}_h(x_0)-m(x_0)\right)^2\\ &\leq 4\left\{B_{h'}(x_0)+B_h(x_0)+T_{h'}(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}+T_h(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}\right\}\\ &\quad +4T_{h'}(x_0)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}+4T_h(x_0)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c},\end{aligned}$$

with $B_h(x_0)$ and $T_h(x_0)$ defined by (17) (the notations are valid with h or h'). We thus have the following decomposition, for any bandwidth h,

$$\begin{aligned}A(h,x_0) &\leq 4\left(\max_{h'\in\mathcal{H}_n,h'\leq h}B_{h'}(x_0)+B_h(x_0)\right)+4\max_{h'\in\mathcal{H}_n,h'\leq h}\left\{T_{h'}(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}-\frac{V(h',x_0)}{8}\right\}_+\\ &\quad +4\max_{h'\in\mathcal{H}_n,h'\leq h}\left\{T_h(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}-\frac{V(h',x_0)}{8}\right\}_+ +4\max_{h'\in\mathcal{H}_n,h'\leq h}\left(T_{h'}(x_0)+T_h(x_0)\right)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}.\end{aligned}$$

Since $V(h')\geq V(h)$ as soon as $h'\leq h$,

$$\begin{aligned}\max_{h'\in\mathcal{H}_n,h'\leq h}\left\{T_h(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}-\frac{V(h',x_0)}{8}\right\}_+ &\leq \max_{h'\in\mathcal{H}_n,h'\leq h}\left\{T_h(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}-\frac{V(h,x_0)}{8}\right\}_+,\\ &\leq \max_{h'\in\mathcal{H}_n,h'\leq h}\left\{T_{h'}(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}-\frac{V(h',x_0)}{8}\right\}_+.\end{aligned}$$

Thus the splitting becomes

$$\begin{aligned}A(h,x_0) &\leq 4\left(\max_{h'\in\mathcal{H}_n,h'\leq h}B_{h'}(x_0)+B_h(x_0)\right)+8\max_{h'\in\mathcal{H}_n,h'\leq h}\left\{T_{h'}(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}-\frac{V(h',x_0)}{8}\right\}_+\\ &\quad +4\max_{h'\in\mathcal{H}_n,h'\leq h}\left(T_{h'}(x_0)+T_h(x_0)\right)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}.\end{aligned}\tag{29}$$

It remains to bound each of the terms. Two of them have not been centred: the terms involving $B_h(x_0)$ are bias terms, and the term involving $\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}$ will be shown to be directly negligible. The last term, which depends on $T_h(x_0)$, is the more difficult term. We will control it by using Lemma 2.

• **Upper-bound for the terms depending on $B_h(x_0)$.** The term $B_h(x_0)$ is the first term of the right-hand-side of (16) and has thus already been bounded (see (19)): $B_h(x_0)\leq C_m^2h^{2\beta}$, for any bandwidth h. Thus,

$$\max_{h'\in\mathcal{H}_n,h'\leq h}B_{h'}(x_0)+B_h(x_0)\leq\max_{h'\in\mathcal{H}_n,h'\leq h}C_m^2h'^{2\beta}+C_m^2h^{2\beta}\leq 2C_m^2h^{2\beta}.\tag{30}$$

• **Upper-bound for the term depending on $\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}$.**

We roughly bound

$$\max_{h'\in\mathcal{H}_n,h'\leq h}\left(T_{h'}(x_0)+T_h(x_0)\right)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}\leq\sum_{h'\in\mathcal{H}_n}\left\{T_{h'}(x_0)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}+T_h(x_0)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}\right\}.$$

Moreover, by using the independence of X_i and ε_i and $\mathbb{E}[\varepsilon_i]=0$, we have

$$\begin{aligned}\mathbb{E}\left[T_h(x_0)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}\right] &= \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \varepsilon_i K_h(\|X_i-x_0\|)\right)^2}{\left(\sum_{i=1}^n K_h(\|X_i-x_0\|)\right)^2}\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}\right],\\ &= \mathbb{E}\left[\frac{\sum_{i=1}^n \varepsilon_i^2 K_h^2(\|X_i-x_0\|)}{\left(\sum_{i=1}^n K_h(\|X_i-x_0\|)\right)^2}\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}\right],\end{aligned}$$

since the ε_i's are independent. Therefore, since the kernel is non-negative (Assumption (H_K)),

$$\mathbb{E}\left[T_h(x_0)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}\right] \leq \sigma^2\mathbb{P}\left((\Omega^{x_0}_{h,h'})^c\right).$$

The inequality remains evidently valid with $T_h(x_0)$ replaced by $T_{h'}(x_0)$, which leads to

$$\max_{h'\in\mathcal{H}_n, h'\leq h}\left(T_{h'}(x_0)+T_h(x_0)\right)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c} \leq \sigma^2\sum_{h'\in\mathcal{H}_n}\mathbb{P}\left((\Omega^{x_0}_{h,h'})^c\right).$$

Then, by applying twice Inequality (14) of Lemma 3,

$$\begin{aligned}\mathbb{P}\left((\Omega^{x_0}_{h,h'})^c\right) &\leq \mathbb{P}\left(|R^{x_0}_{h'}-1|>\frac{1}{2}\right)+\mathbb{P}\left(|R^{x_0}_{h\vee h'}-1|>\frac{1}{2}\right),\\ &\leq 2\left\{\exp\left(-\frac{n\varphi^{x_0}(h')}{8\left(\frac{C_K^2}{c_K^2}+\frac{C_K}{2c_K}\right)}\right)+\exp\left(-\frac{n\varphi^{x_0}(h\vee h')}{8\left(\frac{C_K^2}{c_K^2}+\frac{C_K}{2c_K}\right)}\right)\right\},\\ &\leq 4n^{-\frac{C_0}{8\left(C_K^2/c_K^2+C_K/2c_K\right)}},\end{aligned}$$

since $\varphi^{x_0}(h)\geq C_0\ln(n)/n$ for any $h\in\mathcal{H}_n$ (Assumption $(H_{\mathcal{H}_n,2})$). Then, as the cardinality of $\mathcal{H}_n$ is bounded by n (Assumption $(H_{\mathcal{H}_n,1})$), we deduce

$$\mathbb{E}\left[\max_{h'\in\mathcal{H}_n, h'\leq h}\left(T_{h'}(x_0)+T_h(x_0)\right)\mathbf{1}_{(\Omega^{x_0}_{h,h'})^c}\right] \leq 4\sigma^2 n^{1-\frac{C_0}{8\left(C_K^2/c_K^2+C_K/2c_K\right)}} \leq \frac{4\sigma^2}{n}, \tag{31}$$

if $C_0\geq 16(C_K^2/c_K^2+C_K/2c_K)$.

• **Upper-bound for the terms depending on $T_{h'}(x_0)$.** First notice that

$$\begin{aligned}T_{h'}(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}} &= \left(\sum_{i=1}^n \frac{K_{h'}(\|X_i-x_0\|)}{n\mathbb{E}[K_{h'}(\|X_i-x_0\|)}\frac{1}{R^{x_0}_{h'}}\varepsilon_i\right)^2\mathbf{1}_{\Omega^{x_0}_{h,h'}},\\ &\leq 4\left(\sum_{i=1}^n \frac{K_{h'}(\|X_i-x_0\|)}{n\mathbb{E}[K_{h'}(\|X_i-x_0\|)}\varepsilon_i\right)^2.\end{aligned}$$

This implies that

$$\begin{aligned}&\mathbb{E}\left[\max_{h'\in\mathcal{H}_n, h'\leq h}\left\{T_{h'}(x_0)\mathbf{1}_{\Omega^{x_0}_{h,h'}}-\frac{V(h',x_0)}{8}\right\}_+\right]\\ &\qquad\leq \sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left\{\left(\frac{1}{n}\sum_{i=1}^n T_{i,h'}(x_0)-\mathbb{E}\left[T_{i,h'}(x_0)\right]\right)^2-\frac{V(h',x_0)}{8}\right\}_+\right],\end{aligned}$$

with $T_{i,h'}(x_0)=(K_{h'}(\|X_i-x_0\|)/\mathbb{E}[K_{h'}(\|X_i-x_0\|))\varepsilon_i$ (we have used that $\mathbb{E}[T_{i,h'}(x_0)]=0$, since ε_i is centred, independent from X_i). Our aim is now to apply the Bernstein Inequality

of Lemma 2. The random variables $T_{i,h'}(x_0)$, $i=1,\dots,n$ are *i.i.d.*, and we compute the following parameters (see the assumptions of Lemma 1 for their definitions)

$$v^2 = \frac{C_\varepsilon^2 C_K^2}{c_K^2}\frac{1}{\varphi^{x_0}(h')} \text{ and } b_0 = \frac{C_\varepsilon C_K}{c_K}\frac{1}{\varphi^{x_0}(h')}.$$

Their values are obtained by using (12) and Assumption (H_ε). Let

$$\overline{V}(h',x_0) = \overline{\kappa}\frac{\ln(n)}{n\varphi^{x_0}(h')},\ \overline{\kappa}>0. \tag{32}$$

Lemma 2 shows that

$$\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left\{\left(\frac{1}{n}\sum_{i=1}^n T_{i,h'}(x_0)-\mathbb{E}\left[T_{i,h'}(x_0)\right]\right)^2-\overline{V}(h',x_0)\right\}_+\right] \le A_1(x_0)+A_2(x_0)+A_3(x_0),$$

with

$$\begin{aligned}
A_1(x_0) &= 2\sum_{h'\in\mathcal{H}_n}\frac{32b_0^2}{n^2}\exp\left(-n\frac{\sqrt{\overline{V}(h',x_0)}}{4b_0}\right),\\
A_2(x_0) &= 2\sum_{h'\in\mathcal{H}_n}\frac{8b_0\sqrt{\overline{V}(h',x_0)}}{n}\exp\left(-n\sqrt{\overline{V}(h',x_0)}4b_0\right),\\
A_3(x_0) &= 2\sum_{h'\in\mathcal{H}_n}\frac{4v^2}{n}\exp\left(-n\sqrt{\overline{V}(h',x_0)}4v^2\right).
\end{aligned}$$

The strategy is now similar for each of these three terms: we use the definition of $\overline{V}$ and Assumption $(H_{\mathcal{H}_n})$ (cardinality of $\mathcal{H}_n$ bounded by n, and lower bound $\varphi(h')\ge C_0\ln(n)/n$, for any h') to prove that the three terms have the order of magnitude $O(1/n)$. More precisely, we prove that

$$\begin{aligned}
A_1(x_0) &\le 64\frac{C_K^2C_\varepsilon^2\sigma^2}{c_k^2C_0^2}\frac{1}{\ln^2(n)}n^{1-\frac{c_K\sqrt{\overline{\kappa}C_0}}{4C_KC_\varepsilon\sigma}},\\
A_2(x_0) &\le 16\frac{C_KC_\varepsilon\sigma\sqrt{\overline{\kappa}}}{c_kC_0^{3/2}}\frac{1}{\ln^{3/2}(n)}n^{1-\frac{c_K\sqrt{\overline{\kappa}C_0}}{4C_KC_\varepsilon\sigma}},\\
A_3(x_0) &\le 8\frac{C_K^2C_\varepsilon^2\sigma^2}{C_0c_k^2}\frac{1}{\ln(n)}n^{1-\frac{c_K^2\overline{\kappa}}{4C_K^2C_\varepsilon^2\sigma^2}}.
\end{aligned}$$

By choosing $\overline{\kappa}\ge\max(64C_K^2C_\varepsilon^2\sigma^2/(C_0c_K^2),8C_K^2C_\varepsilon^2\sigma^2/c_K^2)$, we finally obtain

$$\begin{aligned}
&\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left\{\left(\frac{1}{n}\sum_{i=1}^n T_{i,h'}(x_0)-\mathbb{E}\left[T_{i,h'}(x_0)\right]\right)^2-\overline{V}(h',x_0)\right\}_+\right]\\
&\qquad\le\left(64\frac{C_K^2C_\varepsilon^2}{c_k^2C_0^2}+6\frac{C_KC_\varepsilon\sqrt{\overline{\kappa}}}{c_kC_0^{3/2}}+8\frac{C_K^2C_\varepsilon^2}{C_0c_k^2}\right)\frac{1}{n}.
\end{aligned}$$

If $2\kappa/3$ in the definition of $V(h,x_0)$ (see (7)) is larger than $8\overline{\kappa}$, that is if

$$\kappa\ge 12\times 8C_\varepsilon^2\left(C_K^2/c_K^2\right)\max(8/C_0,1),$$

we also have $V(h,x_0)/8\ge\overline{V}(h,x_0)$ and consequently, we have proved

$$\mathbb{E}\left[\max_{h'\in\mathcal{H}_n,h'\le h}\left\{T_{h'}(x_0)\mathbf{1}_{\Omega_{h,h'}^{x_0}}-\frac{V(h',x_0)}{8}\right\}_+\right]\ \le\ \frac{C}{n}, \tag{33}$$

with C depending on c_K, C_K, C_0, and C_ε.

The proof of Lemma 4 is ended by gathering the inequalities (29), (30), (31), and (33).

References

A. Amiri, C. Crambes, and B. Thiam. Recursive estimation of nonparametric regression with functional covariate. *Comput. Statist. Data Anal.*, 69:154–172, 2014.

A. Antoniadis, E. Paparoditis, and T. Sapatinas. Bandwidth selection for functional time series prediction. *Statist. Probab. Lett.*, 79(6):733–740, 2009.

R. B. Ash and M. F. Gardner. *Topics in stochastic processes.* Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1975. Probability and Mathematical Statistics, Vol. 27.

M. Avery, Y. Wu, H. Helen Zhang, and J. Zhang. RKHS-based functional nonparametric regression for sparse and irregular longitudinal data. *Canad. J. Statist.*, 42(2):204–216, 2014.

A. Baíllo and A. Grané. Local linear regression for functional predictor and scalar response. *J. Multivariate Anal.*, 100(1):102–111, 2009.

J. Barrientos-Marin, F. Ferraty, and P. Vieu. Locally modelled regression and functional data. *J. Nonparametr. Stat.*, 22(5-6):617–632, 2010.

K. Benhenni, F. Ferraty, M. Rachdi, and P. Vieu. Local smoothing regression with functional data. *Comput. Statist.*, 22(3):353–369, 2007.

A. Berlinet, A. Elamine, and A. Mas. Local linear regression for functional data. *Ann. Inst. Statist. Math.*, 63(5):1047–1075, 2011.

K. Bertin, C. Lacour, and V. Rivoirard. Adaptive estimation of conditional density function. preprint, hal-00922555, 2014.

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

E. Boj, P. Delicado, and J. Fortiana. Distance-based local linear regression for functional predictors. *Comput. Statist. Data Anal.*, 54(2):429–437, 2010.

E. Brunel and F. Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, 67(3):441–475, 2005.

F. Burba, F. Ferraty, and P. Vieu. k-nearest neighbour method in functional nonparametric regression. *J. Nonparametr. Stat.*, 21(4):453–469, 2009.

T. T. Cai and P. Hall. Prediction in functional linear regression. *Ann. Statist.*, 34(5): 2159–2179, 2006.

H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.*, 92(1):24–41, 2005.

G. Chagny and C. Lacour. Optimal adaptive estimation of the relative density. preprint, hal-00955161, 2014.

G. Chagny and A. Roche. Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. *Electron. J. Stat.*, 8:2352–2404, 2014.

F. Comte and J. Johannes. Adaptive functional linear regression. *Ann. Statist.*, 40(6): 2765–2797, 2012.

C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. *Ann. Statist.*, 37(1):35–72, 2009.

S. Dabo-Niang and F. Ferraty. *Functional and operatorial statistics.* Springer, 2008.

L. Delsol. Advances on asymptotic normality in non-parametric functional time series analysis. *Statistics*, 43(1):13–33, 2009.

F. Ferraty and Y. Romain. *The Oxford Handbook of Functional Data Analysis*. Oxford Handbooks in Mathematics. OUP Oxford, 2011.

F. Ferraty and P. Vieu. Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(2):139–142, 2000.

F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Comput. Statist.*, 17(4):545–564, 2002.

F. Ferraty and P. Vieu. Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *J. Nonparametr. Stat.*, 16 (1-2):111–125, 2004. The International Conference on Recent Trends and Directions in Nonparametric Statistics.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

F. Ferraty, A. Goia, and P. Vieu. Functional nonparametric model for time series: a fractal approach for dimension reduction. *Test*, 11(2):317–344, 2002.

F. Ferraty, A. Laksaci, and P. Vieu. Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes.*, 9(1):47–76, 2006.

F. Ferraty, A. Mas, and P. Vieu. Nonparametric regression on functional data: inference and practical aspects. *Aust. N. Z. J. Stat.*, 49(3):267–286, 2007.

F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu. Rate of uniform consistency for nonparametric estimates with functional variables. *J. Stat. Plan. Inference*, 140(2):335–352, Feb. 2010.

F. Ferraty, I. Van Keilegom, and P. Vieu. Regression when both response and predictor are functions. *J. Multivariate Anal.*, 109:10–28, 2012.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

J. Hoffmann-Jørgensen, L. A. Shepp, and R. M. Dudley. On the lower tail of Gaussian seminorms. *Ann. Probab.*, 7(2):319–342, 1979.

I. B. MacNeill. Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times. *Ann. Statist.*, 6 (2):422–433, 1978.

A. Mas. Lower bound in regression for functional data by representation of small ball probabilities. *Electron. J. Stat.*, 6:1745–1778, 2012.

E. Masry. Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Process. Appl.*, 115(1):155–177, 2005.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *Ann. Statist.*, 33 (2):774–805, 2005.

E. Nadaraya. On estimating regression. *Theory of Probability and its Application*, 9(4): 141–142, 1964.

M. Rachdi and P. Vieu. Nonparametric regression for functional data: automatic smoothing parameter selection. *J. Statist. Plann. Inference*, 137(9):2784–2801, 2007.

J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, 2005.

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B*, 53(3):539–572, 1991. With discussion and a reply by the authors.

G. Rebelles. Pointwise adaptive estimation of a multivariate density under independence hypothesis. *Bernoulli*, 2014. (to appear).

H. L. Shang. Bayesian bandwidth estimation for a nonparametric functional regression model with unknown error density. *Comput. Statist. Data Anal.*, 67:185–198, 2013.

H. L. Shang. Bayesian bandwidth estimation for a functional nonparametric regression model with mixed types of regressors and unknown error density. *J. Nonparametr. Stat.*, 26(3):599–615, 2014.

G. S. Watson. Smooth regression analysis. *Sankhyā Ser. A*, 26:359–372, 1964.