HAL open science

Recursive estimation of the conditional geometric median in Hilbert spaces

Hervé Cardot, Peggy Cénac, Pierre-André Zitt

To cite this version:

Hervé Cardot, Peggy Cénac, Pierre-André Zitt. Recursive estimation of the conditional geometric median in Hilbert spaces. Electronic Journal of Statistics , 2012, 6, p. 2535-2562. 10.1214/12-EJS759 . hal-00687762

HAL Id: hal-00687762

https://hal.science/hal-00687762v1

Submitted on 14 Apr 2012

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.

Recursive estimation of the conditional geometric median in Hilbert spaces

Hervé CARDOT, Peggy CÉNAC, Pierre-André ZITT
Institut de Mathématiques de Bourgogne, Université de Bourgogne,
9 avenue Alain Savary, 21078 Dijon Cédex, France
email: {Herve.Cardot, Peggy.Cenac, Pierre-Andre.Zitt}@u-bourgogne.fr

April 14, 2012

Abstract

A recursive estimator of the conditional geometric median in Hilbert spaces is studied. It is based on a stochastic gradient algorithm whose aim is to minimize a weighted L_1 criterion and is consequently well adapted for robust online estimation. The weights are controlled by a kernel function and an associated bandwidth. Almost sure convergence and L^2 rates of convergence are proved under general conditions on the conditional distribution as well as the sequence of descent steps of the algorithm and the sequence of bandwidths. Asymptotic normality is also proved for the averaged version of the algorithm with an optimal rate of convergence. A simulation study confirms the interest of this new and fast algorithm when the sample sizes are large. Finally, the ability of these recursive algorithms to deal with very high-dimensional data is illustrated on the robust estimation of television audience profiles conditional on the total time spent watching television over a period of 24 hours.

Keywords: asymptotic normality, averaging, CLT, kernel regression, Mallows-Wasserstein distance, online data, Robbins-Monro, robust estimator, sequential estimation, Stochastic gradient.

1 Introduction

It is not unusual nowadays to get large samples of high-dimensional or functional data together with real covariates that are correlated with the functional variable under study. The estimation of how the shape of the functional response may depend on real or functional covariates has been deeply studied in the statistical literature : linear models for functional response have been proposed by Faraway (1997), Cuevas et al. (2002) or Bosq (2000) (see also Ramsay and Silverman (2005)) and Greven et al. (2010) whereas nonlinear relationships are studied in Lecoutre (1990), Chiou et al. (2004), Lian (2007), Cardot (2007), Lian (2011) and Ferraty et al. (2011).

The main drawback of all the above mentioned estimators, whose target is the conditional expectation, is that they all rely, explicitly or not, on least squares and are consequently sensitive to outliers. In such a context of large samples of high dimensional data, outlying observations, which may not be uncommon, might be hard to detect with automatic procedures. Directly considering robust indicators of centrality such as medians is a way to deal with this issue. If Y be a random variable taking values in a Hilbert space H, its geometric median m (also called spatial median or L_1-median, see Small (1990) for a

survey) is defined as follows

$$m := \mathbf{argmin}_{\alpha \in H} \mathbb{E}\left[\|Y - \alpha\| - \|Y\|\right]. \tag{1}$$

The median m is uniquely defined under simple conditions when the dimension of H is larger than or equal to 2, it has a 0.5 breakdown point (Kemperman (1987)) as well as a bounded gross sensitivity error (Cardot et al. (2011)). When one has a sample at hand, algorithms based on the minimization of the empirical version of risk (1) have been proposed by Vardi and Zhang (2000) and properties of such robust estimators can be found in the recent review by Möttönen et al. (2010). Nevertheless, these computational techniques may not be able to handle very large samples of high-dimensional data since they require to store all the data. An alternative approach, developed by Chaouch and Goga (2012) and which can cope with this issue, consists in considering unequal probability sampling techniques in order to select, in a effective way, subsamples with sizes much smaller than the initial sample size.

We suggest in this work another direction based on recursive techniques which do not require to store all the data. Another interest of these recursive approaches is that they allow automatic update of the estimators if, for example, the data arrive sequentially. Recently, a simple recursive algorithm which gives efficient estimates of the geometric median in separable Hilbert spaces has been proposed by Cardot et al. (2011). It is shown that averaged versions of classic stochastic gradient algorithms have a limiting normal distribution that is the same as the distribution of the static estimator based on a direct minimization of the empirical version of risk (1).

In a finite dimension context, Cadre and Gannoun (2000) and Cheng and De Gooijer (2007) proposed to introduce a kernel function K in the empirical version of (1) in order to take covariate effects into account. The kernel weights are controlled by a sequence of bandwidth values that tends to zero when the sample size increases in order to build consistent estimates of the conditional geometric median. With the same ideas of local approximation of the conditional distribution, we study, in this work, a modification of the recursive algorithm suggested in Cardot et al. (2011). It consists in introducing weights, controlled by a kernel function, in order to build consistent recursive estimators of the conditional geometric median. The response variable is also allowed to take values in a separable Hilbert space. For real response, recursive estimators of the regression function based on kernel weights have been introduced by Révész (1977) whereas a deep study of their asymptotic properties, which also includes averaged estimation procedures, is proposed in Mokkadem et al. (2009).

The paper is organized as follows. In Section 2, we first define the stochastic gradient recursive estimator as well as its averaged version for the case of a real covariate. Note that our results could be extended to multidimensional covariates. We state the asymptotic normality, under general conditions, of the averaged algorithm in separable Hilbert spaces, with an optimal rate of convergence. The regularity hypotheses, which are much weaker than those of Cadre and Gannoun (2000), are also expressed in terms of the Wasserstein distance between the conditional distributions.

In Section 3, a comparison of the static approach, which consists in minimizing the empirical version of risk (1), with the stochastic gradient estimator and its averaged version is performed on a simulation study. It confirms the good behavior as well as the stability, with respect to the descent steps, of the averaged algorithm. The ability of this estimator to deal with large samples of very high-dimensional data is then illustrated on the estimation of television audience profiles given the total time spent watching television. Proofs are gathered in Section 4.

2 Notations, hypotheses and main results

Let (Y, X) be a pair of random variables taking values in $H \times \mathbb{R}$, where H is a Hilbert space whose norm is denoted by $\|\cdot\|$. Suppose that X is continuous, and denote by $p(x)$ its density at $x \in \mathbb{R}$. For any x in the support of X, denote by μ_x the conditional law of Y given $X = x$. Consider, for $(\alpha, x) \in H \times \mathbb{R}$, the following functional

$$G(\alpha, x) := p(x)\, \mathbb{E}\left[\|Y - \alpha\| - \|Y\| \,|X = x\right]. \tag{2}$$

The geometric median of Y given $X = x$, denoted by $m(x)$, is defined as the solution of the following optimization problem:

$$m(x) := \mathbf{argmin}_{\alpha \in H}\, G(\alpha, x). \tag{3}$$

The solution of (3) is unique provided that the conditional distribution μ_x is not supported by a straight line (Kemperman (1987)). We suppose from now on the following assumption.

A1. For every x in the support of the probability density function p of the random variable X, μ_x is not concentrated on a straight line: for all $v \in H$, there is $w \in H$ such that $\langle v, w\rangle = 0$ and

$$\mathbf{Var}(\langle w, Y\rangle \,|X = x) > 0. \tag{4}$$

Suppose we have a sequence $(X_n, Y_n)_{n\geq 1}$ of independent copies of (X, Y). In the unconditional case where the X variable is not taken into account, one can look for the unconditional median, *i.e.* the minimum m defined by (1). Under weak hypotheses, the median is uniquely defined as the zero of the derivative:

$$-\mathbb{E}\left[\frac{Y - \alpha}{\|Y - \alpha\|}\right].$$

We introduced in Cardot et al. (2011) the following recursive estimator of m:

$$Z_{n+1} = Z_n + \gamma_n \frac{Y_{n+1} - Z_n}{\|Y_{n+1} - Z_n\|}, \tag{5}$$

where γ_n was a well-chosen deterministic sequence. In the present case, the law of Y_n is not the conditional law μ_x, so this idea does not work directly. However, it is natural to see Y_n as an approximate sample of μ_x if X_n happens to be very close to x. Therefore, a simple estimator can be built by introducing weights, through a kernel function K, whose properties will be specified later. We modify (5) as follows to take the weights into account, and define our recursive estimator of $m(x)$:

$$Z_{n+1}(x) = Z_n(x) + \gamma_n \frac{Y_{n+1} - Z_n(x)}{\|Y_{n+1} - Z_n(x)\|} \frac{1}{h_n} K\left(\frac{X_{n+1} - x}{h_n}\right) \tag{6}$$

with two deterministic sequences of tuning parameters h_n and γ_n whose properties are given below.

For a constant sequence (h_n), this algorithm converges towards the minimum of the modified objective function:

$$G_h(\alpha, x) := \mathbb{E}\left[\left(\|Y - \alpha\| - \|Y\|\right) \frac{1}{h} K\left(\frac{X - x}{h}\right)\right]. \tag{7}$$

The partial derivative of G_h with respect to α is an element of H defined by

$$\begin{aligned}\Phi_h(\alpha) &:= \nabla_\alpha G_h(\alpha, x) \\ &= -\mathbb{E}\left[\frac{Y-\alpha}{\|Y-\alpha\|}\frac{1}{h}K\left(\frac{X-x}{h}\right)\right].\end{aligned} \tag{8}$$

We will see in Proposition 4.1 that, under suitable hypotheses, when h goes to zero, Φ_h goes to the gradient Φ of G, defined by:

$$\Phi(x,\alpha) = -p(x)\mathbb{E}\left[\frac{Y-\alpha}{\|Y-\alpha\|}\middle| X = x\right]. \tag{9}$$

The idea of using a kernel, and of assigning a large weight to Y_n when X_n is close to x can only work if the conditional law $\mu_{x'}$ varies, in some sense, regularly. A natural way of expressing this regularity is through the Mallows-Wasserstein distance. Let us recall its definition.

Definition 1. *Let μ and ν be two probability measures on H with finite second order moments. Let $\mathcal{C}$ be the set of couplings of μ and ν,* i.e. *the set of measures π on $H \times H$ whose first marginal is μ and whose second marginal is ν.*

The Wasserstein distance between μ and ν is given by:

$$\mathcal{W}_2(\mu,\nu) = \left(\inf_{\pi\in\mathcal{C}}\int \|x-y\|^2\, d\pi(x,y)\right)^{1/2}.$$

We may now state our assumptions.

A2. The probability density function p of the random variable X is bounded and satisfies a uniform Hölder condition : there are two constants $\beta > 0$ and $C_2 > 0$ such that

$$\forall (x,x') \in \mathbb{R}^2, \quad |p(x) - p(x')| \leq C_2|x-x'|^\beta.$$

We denote by $p_{\max} = \sup_{x\in\mathbb{R}} p(x)$.

A3. The gradient $\Phi(x,\alpha)$ defined by (9) satisfies a uniform Hölder condition with coefficient β. There is $C_3 > 0$ such that

$$\forall (x,x') \in \mathbb{R}^2, \forall \alpha \in H, \quad \left\|\Phi(\alpha,x) - \Phi(\alpha,x')\right\| \leq C_3|x-x'|^\beta. \tag{10}$$

A4. The conditional law $\mu_x = \mathcal{L}(Y|X=x)$ varies regularly with x: there are two constants C_4 and β such that

$$\mathcal{W}_2(\mu_x,\mu_{x'}) \leq C_4\left|x-x'\right|^\beta. \tag{11}$$

A5. The kernel function K is positive, bounded with compact support and satisfies

$$\int_{\mathbb{R}} K(u)du = 1.$$

A6. There is a constant C_6 such that:

$$\forall \alpha \in H, \forall x, \quad \mathbb{E}\left[\|Y-\alpha\|^{-2}|X=x\right] \leq C_6. \tag{12}$$

Remark 1. *Without loss of generality, we suppose that the constant β in **A2**, **A3** and **A4** has always the same value.*

*Assumption **A3** is a regularity assumption that is required to control the approximation error and to prove the convergence of the algorithm. Assumption **A4** seems to be more natural, and we prove in section 4.1 that, together with **A6**, it implies **A3**.*

*Hypotheses **A2** and **A5** are classical in nonparametric estimation and could be weakened at the expense of more complicated proofs. For classical properties of kernel estimators under general hypotheses, see for example Wand and Jones (1995).*

*Similarly, Hypothesis **A6** is stated quite strongly here, in order to avoid additional technicalities in the proof of the asymptotic normality if the averaged algorithm. See Cardot et al. (2011) for a relaxed version, under which the same results should hold. Informally it forces the law to be "spread out" and this avoids pathological behaviors of the algorithm.*

We have three main results. The first one states the almost sure convergence of the algorithm.

Theorem 2.1. *Under assumptions **A1–A3** and **A5**, and if $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 h_n^{-1} < \infty$ as well as $\sum_n \gamma_n h_n^{\beta} < \infty$, then, for all x such that $p(x) > 0$,*

$$\lim_{n\to\infty} \|Z_n(x) - m(x)\| = 0 \quad a.s.$$

Remark 2. *In the following, for simplicity, we choose the step size and window size as inverse powers of n:*

$$\gamma_n = \frac{c_\gamma}{n^\gamma}, \qquad h_n = \frac{c_h}{n^h}. \tag{13}$$

With these choices the assumptions on the step sizes are:

$$\gamma \leq 1, \qquad 2\gamma - h > 1, \qquad \gamma + \beta h > 1. \tag{14}$$

The assumptions on h and γ are always satisfied if we choose $\gamma = 1$ and $h < 1$. However, as shown in the simulation study, the performances of algorithm (6) strongly depend on the choice of the steps γ_n and particularly on the constant c_γ. Therefore, we also introduce the following averaged algorithm which is less sensitive to the choice of the step sizes γ_n and has nice convergence properties,

$$\overline{Z}_{n+1}(x) = \frac{1}{n}\sum_{k=1}^{n} Z_k(x). \tag{15}$$

Our main result is a central limit theorem on this averaged algorithm. To adapt the proof of the corresponding CLT from Cardot et al. (2011), we need a good *a priori* bound on the error $Z_n(x) - m$.

Proposition 2.2. *Suppose that x is such that $p(x) > 0$ and that $\gamma \leq 1$, $2\gamma - h > 1$, $\gamma + \beta h > 1$, and $h(1 + 2\beta) \geq \gamma$. Under Assumptions **A1–A3** and **A5**, there exist an increasing sequence of events $(\Omega_N)_{N\in\mathbb{N}}$, and constants C_N, such that $\Omega = \bigcup_{N\in\mathbb{N}} \Omega_N$, and*

$$\forall N, \quad \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_n - m(x)\|^2\right] \leq C_N \frac{\ln(n)}{n^{\gamma-h}}.$$

This proposition tells us that, up to a logarithmic factor, the optimal rates of convergence in nonparametric estimation can be attained for well chosen values of the parameter γ and h. If $\gamma = 1$ and $h = (1+2\beta)^{-1}$, then,

$$\|Z_n - m(x)\|^2 = O_p\left(\ln(n)\, n^{-2\beta/(2\beta+1)}\right). \tag{16}$$

Finally our main result is the following central limit theorem for the averaged algorithm.

Theorem 2.3. *Assume **A1**, **A2** and **A4–A6**. Let x satisfy $p(x) > 0$. If $\gamma < 1$, $2\gamma - h > 1$, $\gamma + \beta h > 1$ and $h > (2\beta+1)^{-1}$, then:*

$$\frac{n}{\sqrt{\sum_{k=1}^{n} \frac{1}{h_k}}} \left(\overline{Z}_n - m(x)\right) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \Gamma^{-1}\Sigma\Gamma^{-1}\right),$$

where

$$\Sigma = p(x)\left(\int K^2(u)du\right) \mathbb{E}\left[\frac{(Y - m(x)) \otimes (Y - m(x))}{\|Y - m(x)\|^2}\middle| X = x\right], \tag{17}$$

$$\Gamma = \mathbb{E}\left[\frac{1}{\|Y - m(x)\|}\left(\mathbf{I}_H - \frac{(Y - m(x)) \otimes (Y - m(x))}{\|Y - m(x)\|^2}\right)\middle| X = x\right]. \tag{18}$$

As shown in Cardot et al. (2011) in the unconditional framework, the operator Γ has a bounded inverse under assumption **A1**, so that the asymptotic variance operator is well defined. Let us also remark that with our assumptions on the sequence of bandwidths, we have

$$\sum_{k=1}^{n} \frac{1}{h_k} = \frac{n}{h_n} \cdot \frac{1}{1+h} + o\left(n h_n^{-1}\right). \tag{19}$$

Consequently, the rate of convergence in the CLT is of order $\sqrt{nh_n}$, which is the usual rate of convergence in distribution for nonparametric regression, provided that the bias term is negligible compared to the variance. This latter condition is ensured by the additional condition $h < (2\beta+1)^{-1}$ and we have, with Theorem 2.3,

$$\sqrt{nh_n}\left(\overline{Z}_n - m(x)\right) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{1+h}\Gamma^{-1}\Sigma\Gamma^{-1}\right).$$

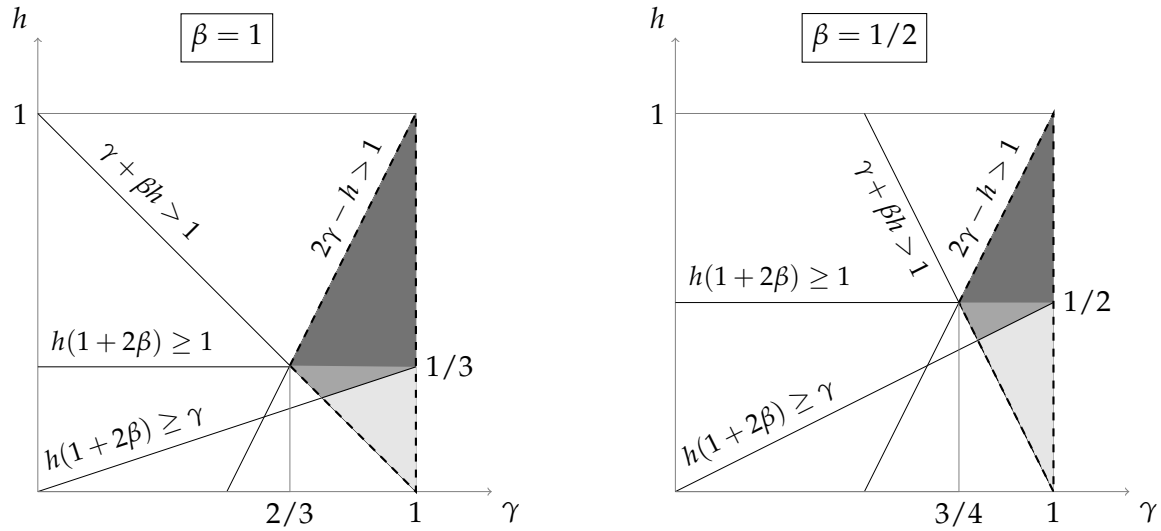
As in the real regression case (see Mokkadem et al. (2009)) it turns out that the averaged estimator has a smaller asymptotic variance, with in our case a factor $(1+h)^{-1}$, than the classical kernel estimator which minimizes the empirical version of risk (7).

Remark 3. *Proceeding exactly as in the proof of Theorem 2.3, it is possible to establish a CLT for another weighted version of the algorithm $\widetilde{Z}_n = \frac{1}{n}\sum_{k=1}^{n} \sqrt{h_k}(Z_k - m)$, which is the empirical mean of $\sqrt{h_n}(Z_n - m)$. Under the same assumptions of Theorem 2.3, one has:*

$$\sqrt{n}\widetilde{Z}_n \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \Gamma^{-1}\Sigma\Gamma^{-1}\right).$$

3 Examples

We first consider a simple simulated example in order to compare the performances of the averaged algorithm with the more classic static one as well as the recursive Robbins-Monro estimator without averaging. Then, the ability of our recursive averaged estimator to deal with large samples of very high-dimensional data is illustrated on the robust estimation of television audience profiles, measured at a minute scale over a period of 24 hours, given the total time spent watching television. All functions are coded in R (R Development Core Team (2010)) and are available on request to the authors.



In this picture we represent the possible choices for the parameters h and γ, when β varies. On the left is the most regular case where $\beta = 1$, on the right we set $\beta = 1/2$. In both cases, if (γ, h) lies in the lighter region, Theorem 2.1 holds and the algorithm converges. In the middle region, the algorithm converges and the additional convergence estimate of Proposition 2.2 holds. Finally, if (γ, h) is in the darker region, the CLT of Theorem 2.3 holds. All these two regions get smaller when β is small. Note that even in the most regular case $\beta = 1$, in order to fulfill the hypotheses of Theorem 2.3, it is necessary to choose γ larger than $2/3$ and h larger than $1/3$.

Figure 1: Possible choices for h and γ.

3.1 A simulated example

Consider a Brownian motion Y measured at d equispaced time points in the interval $[0, 1]$, so that we have $\mathbf{Y} = (Y(t_1), \ldots, Y(t_d))$. Besides, suppose that we know the mean value $X = \int_0^1 Y(t)dt$ of each trajectory Y. We can look for the conditional (geometric) median of vector $\mathbf{Y}$ given X. The joint distribution of $(\mathbf{Y}, X)$ is clearly Gaussian with $\mathbb{E}\mathbf{Y} = 0$, $\mathbb{E}X = 0$,

$$\mathbf{Cov}(Y(t_j), Y(t_\ell)) = \min(t_j, t_\ell), \quad \mathbf{Var}(X) = \frac{1}{3} \quad \text{and } \mathbf{Cov}(X, Y(t_j)) = t_j \left(1 - \frac{t_j}{2}\right).$$

Consequently, the distribution of $\mathbf{Y}$ given $X = x$ is Gaussian with conditional expectation, for $j = 1, \ldots, p$,

$$\mathbb{E}\left[Y(t_j) \middle| X = x\right] = \frac{3}{2} t_j (2 - t_j) x,$$

and a covariance matrix that does not depend on x. By symmetry of the Gaussian distribution, it is also clear that the conditional expectation is equal to the conditional geometric median, when $H = \mathbb{R}^d$ equipped with the usual Euclidean norm, so that

$$m(t_j, x) = \frac{3}{2} t_j (2 - t_j) x. \tag{20}$$

The hypotheses on the density p are clearly satisfied since X is a Gaussian random variable. Furthermore, the Wasserstein distance between two Gaussian laws with expectations m_1 and m_2 and the same covariance matrix is simply $\|m_1 - m_2\|$, (see *e.g.* Givens and Shortt (1984)) so that we can deduce, with (20), that $\beta = 1$ in Assumption **A4**.

We draw n i.i.d. copies of $(\mathbf{Y}, X)$ and we focus in this simulation study on the geometric median of Y given $x = 0.39$, which corresponds to the value of the third quartile of X. Note that our conclusions remain unchanged for other non extreme values of X.

We first compute the static estimator, named "static kernel" in the following. It is based on a direct minimization, with the Weiszfeld's algorithm (see Vardi and Zhang (2000) and Möttönen et al. (2010)), of

$$\boldsymbol{\alpha} \mapsto \sum_{i=1}^{n} w_i \left\| \mathbf{Y}_i - \boldsymbol{\alpha} \right\|, \tag{21}$$

where $w_i = \left[\sum_{\ell=1}^{n} K(h_n^{-1}(X_\ell - x))\right]^{-1} K(h_n^{-1}(X_i - x))$ and K is the Gaussian kernel.

The Robbins Monro estimator Z_n, defined in (6), and the averaged estimator $\overline{Z}_n$, defined in (15), are run for 10 starting points chosen randomly in the sample. Among the 10 estimations, we retain the one with the smallest empirical risk (21).

The accuracy of the different estimators $\widehat{m}$ are compared, for different values of the bandwidth h and sample sizes n, with the quadratic criterion,

$$R(\widehat{m}) = \frac{1}{d} \sum_{j=1}^{d} \left(m(t_j) - \widehat{m}(t_j) \right)^2. \tag{22}$$

Since $\beta = 1$, we can choose $\gamma = 9/10$ and $h = 3/10$, $c_h = 1$, so that the quadratic estimation error for the Robbins-Monro algorithm, will be, up to the $\ln(n)$ factor, of order $n^{-6/10}$ (see Proposition 2.2).

Note that, for simplicity of comparison with the static kernel estimator, we also consider fixed values for $h_n \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ and take in this case $\gamma = 2/3$. We are aware that the assumptions needed for the asymptotic convergence are not satisfied but the sample size is fixed in advance here.

We first present in Table 1 the mean value, over 500 replications, of the MSE defined in (22), when estimating the conditional median with a sample size of $n = 500$ in dimension $d = 100$. For comparison and interpretability of the results, note that $100R(0) = 18.4$.

We note that, when the sample size is moderate (*i.e.* $n = 500$), the interest of considering the averaged recursive estimation procedure is less evident than in the unconditional case (see Cardot et al. (2010)) since the Robbins-Monro estimator Z_n defined in (6) can perform, for well chosen values of the tuning parameters c_γ and h_n, nearly as well as the static estimator. Nevertheless, we can remark that Z_n is highly sensitive to the values of the tuning parameters and its performances deteriorate much with small variations of these parameters as seen in Table 1. This is not the case of the averaged estimator $\overline{Z}_n$, defined in (15), which is much less sensitive and thus allows less sharp choices of the values of the tuning parameters provided the descent steps do not force the algorithm to converge too rapidly. We note again (see Cardot et al. (2010)) that for too small values of c_γ (i.e. $c_\gamma = 0.1$), the algorithm converges too quickly and averaging leads to estimations that are outperformed by the direct Robbins-Monro approach. A way to deal with this drawback is to perform averaging only after a certain number of iterations. All these remarks are clearly illustrated in Figure 2 which presents the estimation error, defined in (22), for both algorithms and for different values of c_γ.

When the sample size gets larger the interest of the averaging step becomes clearer since the estimation error of the Robbins-Monro estimator are always larger as soon as $c_\gamma \geq 1$ (see Table 2). Furthermore, the estimation errors of the static kernel estimator and the averaged recursive one are also now very close to each other.

Table 1: Mean estimation errors ($\times$ 100) of the different estimators, for $n = 500$, $d = 100$, and descent parameter $\gamma = 2/3$ when h_n has a constant value and $\gamma = 0.9$ when $h_n = n^{-h}$ with $h = \gamma/3 = 0.3$.

	Bandwidth h_n					
	0.05	0.10	0.15	0.20	0.25	$n^{-0.3}$
Static kernel	0.349	0.179	0.148	0.172	0.245	
Robbins Monro						
$c_\gamma = 0.1$	0.689	0.625	0.659	0.769	0.912	2.458
$c_\gamma = 0.3$	0.370	0.194	0.159	0.178	0.253	0.332
$c_\gamma = 1$	0.590	0.297	0.229	0.240	0.297	0.183
$c_\gamma = 3$	1.177	0.647	0.486	0.425	0.453	0.248
Averaged						
$c_\gamma = 0.1$	1.047	1.000	1.051	1.160	1.336	2.995
$c_\gamma = 0.3$	0.406	0.213	0.178	0.202	0.287	0.534
$c_\gamma = 1$	0.402	0.195	0.160	0.182	0.252	0.192
$c_\gamma = 3$	0.443	0.209	0.163	0.252	0.256	0.170

Table 2: Mean estimation errors ($\times$ 100) of the different estimators, for $n = 2000$, $d = 100$, and descent parameter $\gamma = 2/3$ when h_n has a constant value and $\gamma = 0.9$ when $h_n = n^{-h}$ with $h = \gamma/3 = 0.3$.

	Bandwidth h_n					
	0.05	0.10	0.15	0.20	0.25	$n^{-0.3}$
Static kernel	0.082	0.053	0.060	0.099	0.176	
Robbins Monro						
$c_\gamma = 0.1$	0.139	0.128	0.149	0.205	0.324	1.321
$c_\gamma = 0.3$	0.095	0.061	0.065	0.103	0.181	0.083
$c_\gamma = 1$	0.173	0.104	0.098	0.126	0.194	0.061
$c_\gamma = 3$	0.403	0.230	0.175	0.192	0.253	0.096
Averaged						
$c_\gamma = 0.1$	0.240	0.237	0.270	0.332	0.484	1.712
$c_\gamma = 0.3$	0.091	0.058	0.065	0.102	0.183	0.138
$c_\gamma = 1$	0.090	0.057	0.063	0.101	0.178	0.060
$c_\gamma = 3$	0.097	0.058	0.064	0.101	0.180	0.057

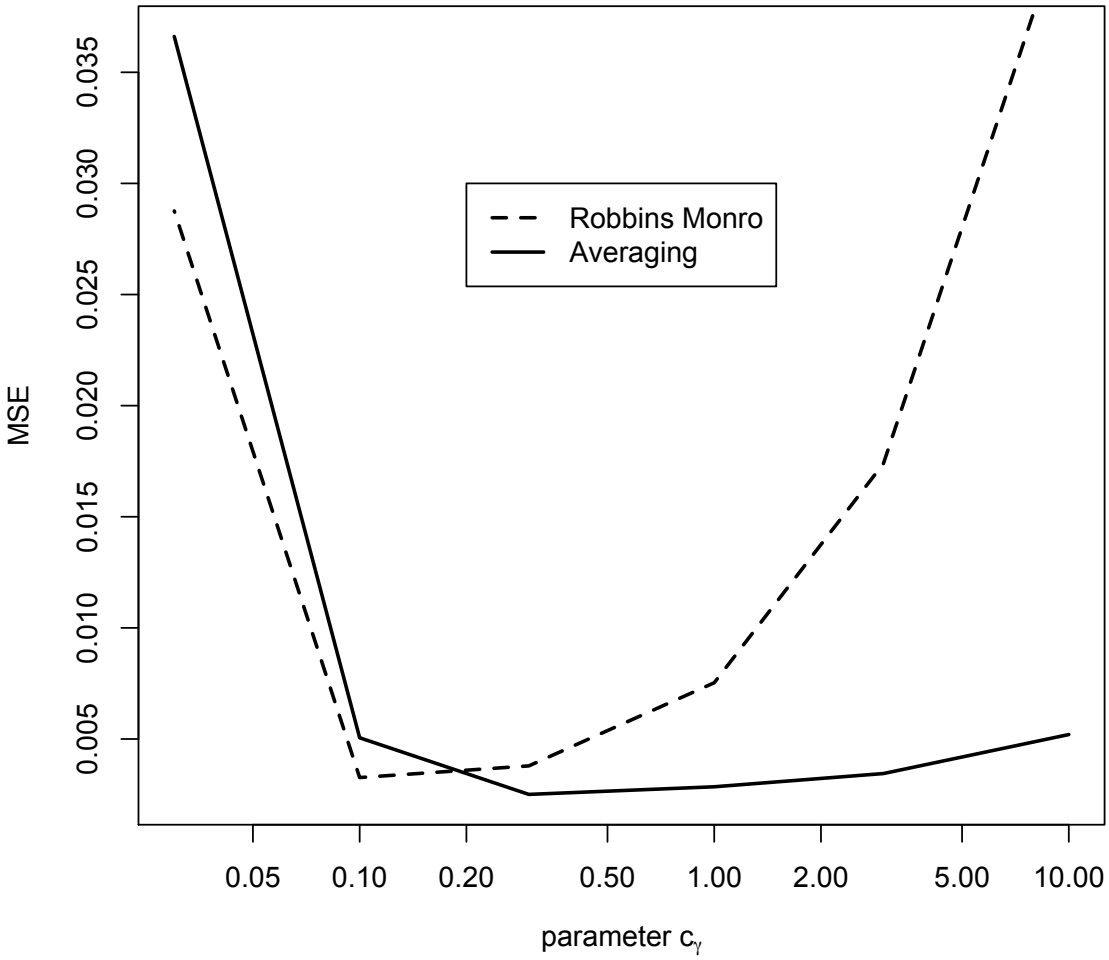


Figure 2: Comparison of the two recursive algorithms according to the mean square error of estimation for different values of c_γ (with a logarithmic scale). The sample size is $n = 500$ and $d = 100$.

3.2 Television audience data

We have a sample of $n = 5422$ individual audiences measured every minute over a period of 24 hours and by the Médiamétrie company in France. For $j = 1, \ldots, 1440$, an observation $Y_i(t_j)$ represents the proportion of time spent by the individual i watching television during the jth minute of this day. Thus, each vector $\mathbf{Y}_i$ belongs to $[0,1]^{1440}$. Note that in fact the first measurement t_1 is made at 3 AM of day d and the last one just before 3 AM of day $d+1$ (see Figure 3). A more detailed description of these data can be found in Cardot et al. (2011).

We are interested in estimating television consumption behaviors, over a 24 hours period, according to the total time spent watching television. The covariate X, is the proportion of time spent watching television over the considered period, $X_i = (\sum_j Y_i(t_j))/1440$, for $i = 1, \ldots, n = 5422$. We consider the quantile values of X which are, in the sample, $q_{25} = 0.0599$, $q_{50} = 0.128$, $q_{75} = 0.225$ and $q_{90} = 0.348$. This means for example, that the ten percent of consumers with the highest consumption levels spend more than 34.8 % of their time watching television whereas the 25 % of consumers with the lowest consumption levels spend less than 6% of their time watching television.

We have drawn in Figure 3 the estimated conditional median profiles with a bandwidth value set to $h_n = 0.05$ and a descent parameter $c_\gamma = 0.5$, for $x \in \{q_{25}, q_{50}, q_{75}, q_{90}\}$. For comparison and better interpretation, we have also plotted the overall geometric median as well as the mean profile. One can note that the shape of the conditional profiles strongly depend on the value of the covariate and that multiplicative models that could be thought to be natural (see the simulation study), are in fact not adapted for modeling the conditional audience median profiles. This is clear if we compare, for example, the levels of the conditional median curves for $x = q_{75}$ and $x = q_{90}$ at time 15 and at time 21. Around 21, their values are approximately the same and are close to the global maximum whereas at time 15 the value of the conditional median for $x = q_{90}$ is about twice the value of the conditional median for $x = q_{75}$.

From a computational speed point of view, for one starting point, our algorithm, which takes less than two seconds, is about 70 times faster than the static estimator which requires 140 seconds to converge.

4 Proofs

Notation. *In all the proofs, x will be a fixed point in $\mathbb{R}$ satisfying $p(x) > 0$. Since x will not vary, we will abuse notation and drop it from various quantities. In particular, in the following m will denote the median $m(x)$ of the conditional law μ_x, and we will write $Z_n = Z_n(x)$ and $\Phi(\alpha) = \Phi(x, \alpha)$.*

4.1 About the assumptions

We begin by a simple geometric result on unit vectors. For a, b two points in H, let $D(a,b)$ be the unit vector "starting" from a in the direction of b. Now if a, b, c are three points in H, such that $\|a-b\| \leq \|a-c\|$, Thales' theorem shows that:

$$\frac{\|D(a,b) - D(a,c)\|}{\|b-c'\|} = \frac{\|a + D(a,b) - a\|}{\|a-b\|} = \frac{1}{\|a-b\|},$$

so

$$\|D(a,b) - D(a,c)\| \leq \frac{\|b-c'\|}{\|a-b\|} \leq \frac{\|b-c\|}{\|a-b\|}.$$

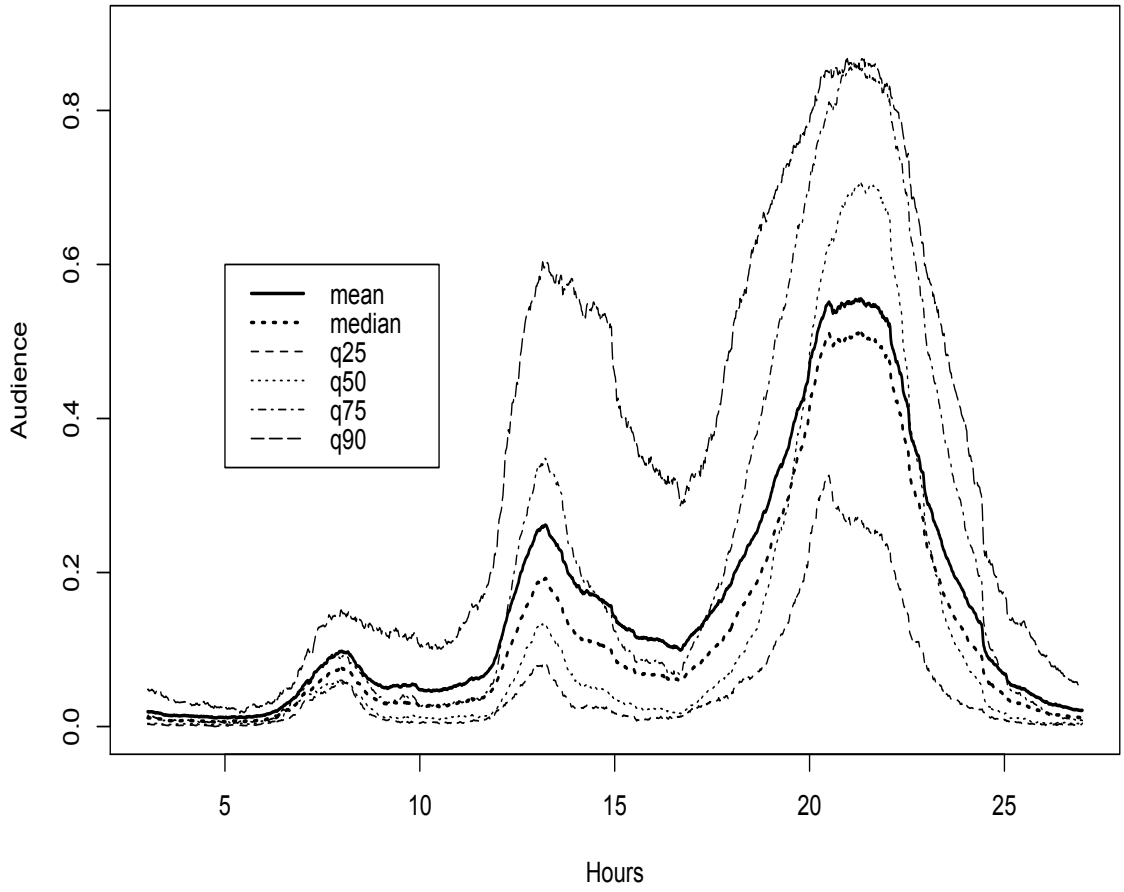
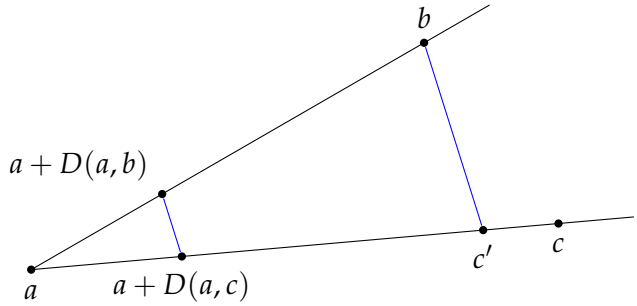


Figure 3: Estimation of the conditional median profile for different levels of total time spent watching television, on the 6th September 2010.



In any case,

$$\|D(a,b) - D(a,c)\| \leq \frac{\|b-c\|}{\min(\|a-b\|, \|a-c\|)}. \tag{23}$$

We will need a "decoupled" version of this inequality:

$$\|D(a,b) - D(a,c)\| \leq \frac{\|b-c\|}{\|a-b\|} + \frac{\|b-c\|}{\|a-c\|}. \tag{24}$$

We can now prove that **A4** and **A6** imply **A3**. Let x, x' be two real numbers in the support of p. Recall that μ_x denotes the law $\mathcal{L}(Y|X=x)$. Let Y and Y' be two random variables with respective laws μ_x and $\mu_{x'}$, such that their joint law π achieves the Wasserstein distance. Let us first show that:

$$\forall \alpha \in H, \quad \left\|\mathbb{E}\left[D(\alpha, Y)\right] - \mathbb{E}\left[D(\alpha, Y')\right]\right\| \leq C\left|x - x'\right|^{\beta}. \tag{25}$$

Fix an $\alpha \in H$. We have:

$$\begin{aligned}\left\|\mathbb{E}\left[D(\alpha, Y)\right] - \mathbb{E}\left[D(\alpha, Y')\right]\right\| &\leq \left\|\mathbb{E}_{\pi}\left[D(\alpha, Y) - D(\alpha, Y')\right]\right\| \\ &\leq \mathbb{E}_{\pi}\left[\left\|D(\alpha, Y) - D(\alpha, Y')\right\|\right].\end{aligned}$$

Now we use the geometric bound (24), and Hölder's inequality:

$$\begin{aligned}\left\|\mathbb{E}\left[D(\alpha, Y)\right] - \mathbb{E}\left[D(\alpha, Y')\right]\right\| &\leq \mathbb{E}_{\pi}\left[\frac{\|Y-Y'\|}{\|Y-\alpha\|}\right] + \mathbb{E}_{\pi}\left[\frac{\|Y-Y'\|}{\|Y'-\alpha\|}\right] \\ &\leq \left(\sqrt{\mathbb{E}\left[\frac{1}{\|Y-\alpha\|^2}\right]} + \sqrt{\mathbb{E}\left[\frac{1}{\|Y'-\alpha\|^2}\right]}\right)\sqrt{\mathbb{E}_{\pi}\left[\|Y-Y'\|^2\right]}.\end{aligned}$$

The first term is bounded by $2\sqrt{C_6}$ thanks to **A6**. The second one is, by definition, the Wasserstein distance, and is bounded by $C_4\left|x-x'\right|^{\beta}$ thanks to **A4**, therefore (25) holds. Since p is $\mathcal{C}^2$ with compact support, the product $\Phi(x,\alpha) = p(x)\mathbb{E}\left[D(\alpha, Y)|X=x\right]$ is itself uniformly β-Hölder continuous; in other words **A3** holds.

4.2 First properties

Recall that, for $z \in H$, $\Phi_h(z)$ is defined by (8) as the conditional expectation of the step, with window size h. When h goes to zero, this "expected step" converges.

Proposition 4.1. *The expected step is bounded:*

$$\exists C, \forall h > 0, \forall \alpha \in H, \quad \|\Phi_h(\alpha)\| \leq p_{\max}. \tag{26}$$

*Moreover, under hypotheses **A2**, **A3** and **A5**, there exists a constant C such that:*

$$\|\Phi_h(\alpha) - \Phi(\alpha)\| \leq Ch^{\beta}, \tag{27}$$

where $\Phi(x,\alpha)$ is defined by (9).

Proof. With our strong hypotheses this result is easy to prove. Indeed

$$\Phi_h(\alpha) = \int \frac{1}{h} K\left(\frac{x'-x}{h}\right) \Phi(x',\alpha) dx',$$

so that by Jensen's inequality

$$\|\Phi_h(\alpha)\| \leq p_{\max} \int_{x'} \frac{1}{h} K\left(\frac{x-x'}{h}\right) dx' = p_{\max}.$$

Moreover,

$$\|\Phi_h(\alpha) - \Phi(\alpha)\| \leq \int \frac{1}{h} K\left(\frac{x-x'}{h}\right) \left\|\Phi(x',\alpha) - \Phi(x,\alpha)\right\| dx'.$$

Now we use Assumption **A3** to bound the norm by $C_3 \left|x'-x\right|^\beta$, the compact support of the bounded function K (Assumption **A5**) and we integrate:

$$\begin{aligned}
\|\Phi_h(\alpha) - \Phi(\alpha)\| &\leq C_3 \int \frac{1}{h} K\left(\frac{x-x'}{h}\right) \left|x-x'\right|^\beta dx' \\
&\leq C_3 \int K(t) h^\beta t^\beta dt \\
&\leq C h^\beta.
\end{aligned}$$

□

Thanks to this result, we have a natural decomposition of algorithm (6). Let us introduce the two following quantities:

$$D_h(z) = \Phi_h(z) - \Phi(z), \tag{28}$$

$$\xi_{n+1} = \left[-\frac{Y_{n+1}-Z_n}{\|Y_{n+1}-Z_n\|} \frac{1}{h_n} K\left(\frac{X_{n+1}-x}{h_n}\right) - \Phi_{h_n}(Z_n)\right]. \tag{29}$$

In terms of these quantities, we can rewrite (6) as:

$$Z_{n+1} = Z_n - \gamma_n \Phi(Z_n) - \gamma_n D_{h_n}(Z_n) - \gamma_n \xi_{n+1}. \tag{30}$$

The first term $D_{h_n}(Z_n)$ will be controlled by Proposition 4.1. The second term ξ_{n+1} defines a sequence of martingale differences, since the conditional expectation given the sequence of σ-algebra $\mathcal{F}_n = \sigma(Z_1,\ldots,Z_n) = \sigma(Y_1,X_1,\ldots,Y_n,X_n)$ satisfies

$$\mathbb{E}\left[\xi_{n+1}|\mathcal{F}_n\right] = 0, \ a.s.$$

For future reference, let us note the following bound on ξ_n:

$$\begin{aligned}
\mathbb{E}\left[\|\xi_{n+1}\|^2 \middle| \mathcal{F}_n\right] &= \mathbb{E}\left[\left\|\frac{Y_{n+1}-Z_n}{\|Y_{n+1}-Z_n\|} \frac{1}{h_n} K\left(\frac{X_{n+1}-x}{h_n}\right)\right\|^2 \middle| \mathcal{F}_n\right] - \|\Phi_{h_n}(Z_n)\|^2, \ a.s \\
&\leq \frac{1}{h_n^2} \mathbb{E}\left[K^2\left(\frac{X_{n+1}-x}{h_n}\right)\right] - \|\Phi_{h_n}(Z_n)\|^2, \ a.s \\
&\leq \frac{C}{h_n}, \ a.s.
\end{aligned} \tag{31}$$

4.3 Almost sure convergence

In this section we prove Theorem 2.1. Define $V_n = \|Z_n - m\|^2$. By (30), we have:

$$\begin{aligned} V_{n+1} = V_n &+ \gamma_n^2 \|\Phi(Z_n)\|^2 + \gamma_n^2 \|D_{h_n}(Z_n) + \xi_{n+1}\|^2 \\ &+ 2 \langle Z_n - m, \gamma_n \Phi(Z_n) \rangle + 2\gamma_n \langle Z_n - m, D_{h_n}(Z_n) + \xi_{n+1} \rangle \\ &+ 2\gamma_n^2 \langle \Phi(Z_n), D_{h_n}(Z_n) + \xi_{n+1} \rangle . \end{aligned}$$

The first scalar product is non positive, we denote it by $(-\eta_n)$. We condition by $\mathcal{F}_n$: the ξ_{n+1} in the scalar products disappear by the martingale property. Then we use Hölder's inequality:

$$\begin{aligned} \mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n &+ \gamma_n^2 \|\Phi(Z_n)\|^2 + 2\gamma_n^2 \|D_{h_n}(Z_n)\|^2 + 2\gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 \middle| \mathcal{F}_n\right] \\ &- \eta_n + 2\gamma_n \|Z_n - m\| \|D_{h_n}(Z_n)\| + 2\gamma_n^2 \|\Phi(Z_n)\| \|D_{h_n}(Z_n)\| . \end{aligned}$$

On the last term we use $2xy \leq x^2 + y^2$ to get:

$$\begin{aligned} \mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n &+ 2\gamma_n^2 \|\Phi(Z_n)\|^2 + 3\gamma_n^2 \|D_{h_n}(Z_n)\|^2 + 2\gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 \middle| \mathcal{F}_n\right] \\ &- \eta_n + 2\gamma_n \|Z_n - m\| \|D_{h_n}(Z_n)\| . \end{aligned}$$

On the last term we bound $\|Z_n - m\|$ by $(1 + V_n)$ to get:

$$\begin{aligned} \mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq (1 + 2\gamma_n \|D_{h_n}(Z_n)\|) V_n &+ 2\gamma_n^2 \|\Phi(Z_n)\|^2 + 3\gamma_n^2 \|D_{h_n}(Z_n)\|^2 \\ &+ 2\gamma_n^2 \mathbb{E}\left[\|\xi_{n+1}\|^2 \middle| \mathcal{F}_n\right] + 2\gamma_n \|D_{h_n}(Z_n)\| - \eta_n. \end{aligned}$$

Finally we bound $D_{h_n}(Z_n)$ by Ch_n^β thanks to (27), $\mathbb{E}\left[\|\xi_{n+1}\|^2 \middle| \mathcal{F}_n\right]$ by C/h_n thanks to (31) and $\|\Phi(Z_n)\|$ by $p(x)$. This yields:

$$\begin{aligned} \mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &\leq \left(1 + 2C\gamma_n h_n^\beta\right) V_n + 2p(x)\gamma_n^2 + 3C^2\gamma_n^2 h_n^{2\beta} \\ &\quad + 2C\frac{\gamma_n^2}{h_n} + 2C\gamma_n h_n^\beta - \eta_n \\ &\leq (1 + b_n) V_n + \chi_n - \eta_n, \end{aligned}$$

where $b_n = 2C\gamma_n h_n^\beta$ and $\chi_n = 3C^2\gamma_n^2 h_n^{2\beta} + 2C\gamma_n^2 h_n^{-1} + 2C\gamma_n h_n^\beta$ satisfy:

$$\sum b_n < \infty, \qquad \sum \chi_n < \infty.$$

Therefore by the Robbins–Siegmund Lemma (Theorem 1.3.12 of Duflo (1997)), V_n converges almost surely and $\sum_n \eta_n < \infty$. This implies that the limit of V_n is zero, by the same argument than in Cardot et al. (2011), assuming that $\sum_n \gamma_n = \infty$.

4.4 Proof of proposition 2.2

For the sake of clarity, we follow the same steps as the proof of Proposition 3.2 in Cardot et al. (2011), and emphasize the necessary changes.

Step 1 — a spectral decomposition. This step is exactly the same as in Cardot et al. (2011): thanks to a spectral decomposition of Γ, we can define the operators:

$$\alpha_k = \mathbf{I}_H - \gamma_k \Gamma, \qquad \beta_n = \alpha_n \alpha_{n-1} \cdots \alpha_1.$$

Introducing the sequence of real functions, for $n \in \mathbb{N}$,

$$f_n(x) = \prod_{k=1}^{n} (1 - \gamma_k x),$$

we see that each operator β_n can be also expressed as follows:

$$\beta_n x = \sum_{\lambda \in \Lambda} f_n(\lambda) \langle e_\lambda, x \rangle e_\lambda, \quad x \in H,$$

their inverses are bounded operators, and satisfy: $\beta_n^{-1} x = \sum_{\lambda \in \Lambda} f_n^{-1}(\lambda) \langle e_\lambda, x \rangle e_\lambda$.

Moreover there exist constants $\kappa_1, \kappa_2, \kappa_3$ such that:

$$\begin{aligned} \forall x \in \sigma(\Gamma), \quad \kappa_1 \exp(-s_n x) \leq f_n(x) \leq \kappa_2 \exp(-s_n x), \\ \left| s_n - \frac{c_\gamma}{1-\gamma} n^{1-\gamma} \right| \leq \kappa_3, \end{aligned} \tag{32}$$

where we recall that $s_n = \sum_{k=1}^{n} \gamma_k$, and $\gamma_k = c_\gamma k^{-\gamma}$.

Step 2 — Decomposition of the algorithm. Recall the decomposition (30) , and rewrite the algorithm as follows:

$$\begin{aligned} Z_{n+1} &= Z_n - \gamma_n \xi_{n+1} - \gamma_n \Phi(Z_n) - \gamma_n D_{h_n}(Z_n) \\ &= Z_n - \gamma_n \xi_{n+1} - \gamma_n (\Gamma(Z_n - m) + \delta_n) - \gamma_n D_{h_n}(Z_n) \end{aligned} \tag{33}$$

where $\delta_n = \Phi(Z_n) - \Gamma(Z_n - m)$ is the difference between the gradient of G and the gradient of its quadratic approximation. Compared to Cardot et al. (2011), there are two differences: the martingale difference ξ_n has changed, and there is an additional term $\gamma_n D_{h_n}(Z_n)$. Therefore:

$$\forall k, \quad Z_{k+1} - m = \alpha_k (Z_k - m) - \gamma_k \xi_{k+1} - \gamma_k \delta_k - \gamma_k D_{h_k}(Z_k). \tag{34}$$

Rewriting $\alpha_{n-1}\alpha_{n-2} \cdots \alpha_{k+1}$ as $\beta_{n-1}\beta_k^{-1}$, we get by induction,

$$Z_n - m = \beta_{n-1}(Z_1 - m) + \beta_{n-1} M_n - \beta_{n-1} R_{n-1} - \beta_{n-1} R'_{n-1}, \tag{35}$$

where

$$\begin{aligned} R_n &= \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k \\ M_n &= -\sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1} \\ R'_n &= \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} D_{h_k}(Z_k). \end{aligned}$$

At this point, the first and third term are the same as in Cardot et al. (2011), the martingale has changed and there is an additional remainder term R'_n.

Step 3 — The deterministic term. Just as in Cardot et al. (2011), we get:

$$\mathbb{E}\left[\|\beta_{n-1}(Z_1-m)\|^2\right] \leq C\exp\left(-2n^{1-\gamma}\right)\mathbb{E}\left[\|Z_1-m\|^2\right]. \tag{36}$$

Step 4 — The martingale. Still following Cardot et al. (2011), we use the spectral decomposition to deal with the martingale part. The changes appear just before eq. (41) in that paper, where the bound on $\mathbb{E}[\|\xi_k\|^2]$ has to be changed (from 1 to C/h_n, using the new bound (31)). Then we use the bounds (32) to get:

$$\begin{aligned}\mathbb{E}\left[\|\beta_{n-1}M_n\|^2\right] &\leq C\sum_{k\leq n-1}\frac{\gamma_k^2}{h_k}\left(\frac{f_{n-1}(\lambda_{min})}{f_k(\lambda_{min})}\right)^2 \\ &\leq C\sum_{k\leq n-1}\frac{\gamma_k^2}{h_k}\exp\left(-\frac{1}{1-\gamma}\left(n^{1-\gamma}-k^{1-\gamma}\right)\right).\end{aligned} \tag{37}$$

Once more, the first terms in the sum are negligible (thanks to the exponential), and we isolate the last terms, for $k \geq l(n)$, where $l(n)$ is given by

$$l(n)^{1-\gamma} = n^{1-\gamma} - c_\alpha \ln(n)\,, \tag{38}$$

for some constant c_α. Choosing c_α large enough, the arguments from Cardot et al. (2011) ensure that the main contribution comes from the last terms. The number of terms, that is $n-l(n)$, is of the order $\ln(n)n^\gamma$, and $\gamma_{l(n)}^2/h_{l(n)}$ is equivalent to $cn^{h-2\gamma}$. Therefore

$$\mathbb{E}\left[\|\beta_{n-1}M_n\|^2\right] \leq C\frac{\ln(n)}{n^{\gamma-h}}. \tag{39}$$

Step 5 — the error terms.

The first error term is $R_n = \beta_{n-1}\sum_{k=1}^n \gamma_k\beta_k^{-1}\delta_k$, where $\delta_k = \Phi(Z_k) - \Gamma(Z_k - m)$. This one can be treated exactly as in Cardot et al. (2011). We recall the definition of the event Ω_N:

$$\Omega_N = \left\{\omega, \begin{array}{ll} \forall n \geq N, \forall k \geq n-l(n), & \|Z_k(\omega)-m\| \leq 1/K \\ & \text{and } \|\delta_k(\omega)\| \leq C_r\|Z_k(\omega)-m\|^2 \\ \forall k, \|\delta_k(\omega)\| \leq N. & \end{array}\right\},$$

for a value of K to be chosen later, and $l(n)$ defined by (38). Then, for any power of n (say n^{-42}) there is a C such that, on Ω_N and for $n \geq N$,

$$\|\beta_{n-1}R_n\|^2 \leq \frac{CN^2}{n^{42}} + \frac{C}{K^2}\sum_{k=l(n)+1}^n \gamma_k\|Z_k-m\|^2. \tag{40}$$

We now turn to the bound of the new error term $R_n' = \beta_{n-1}\sum_{k=1}^{n-1}\gamma_k\beta_k^{-1}D_{h_k}(Z_k)$. To bound D_h, we use (27):

$$\|D_h(z)\| \leq Ch^\beta.$$

Therefore for N large enough, and for $k \geq l(n)$,

$$\left\|\mathbf{1}_{\Omega_N}D_{h_k}(Z_k)\right\| \leq Ch_k^\beta.$$

For k smaller than $l(n)$, we use the crude bound $D_{h_k}(Z_k) \leq p_{\max}+1$. Finally we get:

$$\left\|\mathbf{1}_{\Omega_N}R_n'\right\| \leq \frac{C}{n^{42}} + (n-l(n))\gamma_{l(n)}h_{l(n)}^\beta.$$

The last term is bounded by $Cn^{-\beta h}$ and dominates the first term.

Finally, since by assumption, $h(1+2\beta) \geq \gamma$, one gets

$$\left\|\mathbf{1}_{\Omega_N} R'_n\right\|^2 \leq \frac{C}{n^{\gamma-h}}. \tag{41}$$

Now we use (36), (39), (40) and (41) to bound the four terms that appear in (35). We get, for $n \geq N$ and some new constant C:

$$\mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_n - m\|^2\right] \leq \frac{C\ln(n)}{n^{\gamma-h}} + \frac{C'}{K^2} \sup_{l(n)<k\leq n} \mathbb{E}\left[\mathbf{1}_{\Omega_N} \|Z_k - m\|^2\right].$$

By the same induction than in Cardot et al. (2011), we obtain the bound announced in Proposition 2.2.

4.5 Proof of Theorem 2.3

The following proof follows the same guidelines as the proof of Theorem 3.4 in Cardot et al. (2011). Again we emphasize the necessary changes due to the introduction of the kernel and of the conditional distribution. We first linearize the target function around the conditional median m as in (33):

$$\forall n, \quad Z_{n+1} - m = (\mathbf{I}_H - \gamma_n \Gamma)(Z_n - m) - \gamma_n \xi_{n+1} - \gamma_n \delta_n - \gamma_n D_{h_n}(Z_n),$$

where (ξ_n) is a martingale difference sequence. Therefore, for all k,

$$\Gamma(Z_k - m) = \gamma_k^{-1}\left((Z_k - m) - (Z_{k+1} - m)\right) - \xi_{k+1} - \delta_k - D_{h_k}(Z_k). \tag{42}$$

Define now,

$$T_n := Z_n - m, \quad \overline{T}_n := \overline{Z}_n - m \quad \text{and} \quad M_{n+1} := \sum_{k=1}^{n} \xi_{k+1},$$

and sum (42) over k

$$n\Gamma\overline{T}_n = \sum_{k=1}^{n} \frac{1}{\gamma_k}\left(T_k - T_{k+1}\right) - \sum_{k=1}^{n}\left(\delta_k + D_{h_k}(Z_k)\right) - M_{n+1},$$

so that

$$\frac{n}{\sqrt{\sum_{k=1}^{n}\frac{1}{h_k}}}\Gamma\overline{T}_n = \frac{n}{\sqrt{\sum_{k=1}^{n}\frac{1}{h_k}}}\left(\frac{T_1\sqrt{h_1}}{\gamma_1} - A_n + A'_n - A''_n\right) + \frac{n}{\sqrt{\sum_{k=1}^{n}\frac{1}{h_k}}}M_{n+1}, \tag{43}$$

where

$$\begin{aligned}
A_n &:= \frac{T_{n+1}}{\gamma_n}, \\
A'_n &:= \sum_{k=2}^{n} T_k\left[\frac{1}{\gamma_k} - \frac{1}{\gamma_{k+1}}\right], \\
A''_n &:= \sum_{k=1}^{n}\left(\delta_k + D_{h_k}(Z_k)\right).
\end{aligned}$$

Step Zero — convergence of covariance operators Our first task is to establish a central limit theorem for the last term of (43):

$$\frac{n}{\sqrt{\sum_{k=1}^{n}\frac{1}{h_k}}}M_n \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0,\Sigma\right), \tag{44}$$

where Σ is the limiting covariance defined by (17). On the space of linear operators on H we consider two classical norms, the (strong) operator norm and the Hilbert-Schmidt norm:

$$\|A\|_{\mathrm{op}} = \sup\left\{\|Ay\|_H\,;\|y\|\leq 1\right\},$$
$$\|A\|_{\mathrm{H.S.}} = \left(\sum_{i=0}^{\infty}\left\langle Ae_j,e_j\right\rangle^2\right)^{1/2},$$

where e_j is an orthonormal base of H. The following lemma will be useful.

Lemma 4.2. *Define a random covariance operator Σ_n by:*

$$\Sigma_n = h_n\mathbb{E}\left[\xi_{n+1}\otimes\xi_{n+1}|\mathcal{F}_n\right]. \tag{45}$$

Then:

$$\sqrt{\Sigma_n}\xrightarrow[n\to\infty]{H.\text{-}S.}\sqrt{\Sigma},\quad a.s. \tag{46}$$

In particular, Σ_n converges to Σ a.s. in the operator norm. Moreover, if $\overline{\Sigma}_n$ denotes the following averaged version of Σ_n:

$$\overline{\Sigma}_n = \frac{1}{\sum_{k=1}^{n}\frac{1}{h_k}}\sum_{k=1}^{n}\frac{1}{h_k}\Sigma_k = \frac{1}{\sum_{k=1}^{n}\frac{1}{h_k}}\sum_{k=1}^{n}\mathbb{E}\left[\xi_{k+1}\otimes\xi_{k+1}|\mathcal{F}_k\right],$$

then

$$\sqrt{\overline{\Sigma}_n}\xrightarrow[n\to\infty]{H.\text{-}S.}\sqrt{\Sigma},\quad a.s. \tag{47}$$

Finally, for any orthogonal projection operator P,

$$\mathbb{E}\left[\mathrm{Tr}\left(\overline{\Sigma}P\right)\right]\xrightarrow[n\to\infty]{}\mathbb{E}\left[\mathrm{Tr}\left(\Sigma P\right)\right]. \tag{48}$$

Remark 4. *Let us note that the convergence of square roots of covariance operators is equivalent to the convergence of the centered Gaussian laws with these covariances; see e.g. (Bogachev, 1998), Example 3.8.13.*

Proof. We first show that the convergence (46) holds in operator norm. Recall that $D(x,y)$ denotes the unit vector $(y-x)/\|y-x\|$. Let us rewrite Σ_n.

$$\Sigma_n = \frac{1}{h_n}\mathbb{E}\left[K^2\left(\frac{X_{n+1}-x}{h_n}\right)D(Z_n,Y_{n+1})\otimes D(Z_n,Y_{n+1})\middle|\mathcal{F}_n\right] - h_n\Phi_{h_n}(Z_n)\otimes\Phi_{h_n}(Z_n).$$

Denote by (X,Y) a couple of random variables with the original joint law, and Y_x be a random variable with law μ_x, independent from (X,Y).

We decompose the difference $\Sigma_n-\Sigma = D_1+D_2+D_3+D_4$ where

$$\begin{aligned}
D_1 &= \frac{1}{h_n}\mathbb{E}\left[K^2\left(\frac{X_{n+1}-x}{h_n}\right)D(Z_n,Y_{n+1})\otimes D(Z_n,Y_{n+1})\middle|\mathcal{F}_n\right]\\
&\quad-\frac{1}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)D(m,Y)\otimes D(m,Y)\right]\\
D_2 &= \frac{1}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)D(m,Y)\otimes D(m,Y)\right]-\frac{1}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)D(m,Y_x)\otimes D(m,Y_x)\right]\\
D_3 &= \frac{1}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)D(m,Y_x)\otimes D(m,Y_x)\right]-\Sigma\\
D_4 &= -h_n\Phi_{h_n}(Z_n)\otimes\Phi_{h_n}(Z_n).
\end{aligned}$$

Note that only the first and the last terms are random; the others depend on n only through the quantity h_n. For $(a,b,c) \in H^3$, it is easy to see that:

$$\begin{aligned}\|D(a,b) \otimes D(a,b) - D(a,c) \otimes D(a,c)\|_{\text{op}} &\leq (\|D(a,b)\| + \|D(a,c)\|)\, \|D(a,b) - D(a,c)\| \\ &\leq 2\left(\frac{1}{\|a-b\|} + \frac{1}{\|a-c\|}\right) \|b-c\|,\end{aligned}$$

where we used (24) in the last line. Therefore:

$$\|D_1\|_{\text{op}} \leq \frac{2}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)\left(\frac{1}{\|Y-m\|} + \frac{1}{\|Y-Z_n\|}\right)\right] \|Z_n - m\|.$$

Conditioning on X and using Assumption **A6**, we get:

$$\|D_1\|_{\text{op}} \leq 4\sqrt{C_6}\frac{1}{h_n}\mathbb{E}\left[K^2((X-x)/h_n)\right] \|Z_n - m\|.$$

The boundedness of p and the finiteness of $v^2 = \int K^2(u)du$ ensure

$$\frac{1}{h_n}\mathbb{E}\left[K^2((X-x)/h_n)\right] \xrightarrow[\text{n}\to\infty]{} p(x)v^2, \tag{49}$$

by dominated convergence; therefore the sequence $(h_n)^{-1}\mathbb{E}\left[K^2((X-x)/h_n)\right]$ is bounded. Since Z_n converges a.s. to m, $\|D_1\|_{\text{op}}$ converges a.s. to zero.

The second term D_2 is treated similarly; we get:

$$\|D_2\|_{\text{op}} \leq \frac{2}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)\left(\frac{1}{\|Y-m\|} + \frac{1}{\|Y_x-m\|}\right) \|Y - Y_x\|\right].$$

Recall that $\mu_x = \mathcal{L}(Y|X=x)$ and let $\mu_{x,x'}$ be a coupling of μ_x and $\mu_{x'}$ that achieves the Wasserstein distance. We condition on the value of X and we apply Hölder's inequality in order to bound the first integral with Assumption **A6** and the second one with Assumption **A4**:

$$\begin{aligned}\|D_2\|_{\text{op}} &\leq \frac{2}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)\left(\int \frac{1}{\|y-m\|^2}d\mu_X(y)\right)^{1/2}\left(\int \|y-y'\|^2 d\mu_{x,X}(y,y')\right)^{1/2}\right] \\ &\quad + \frac{2}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)\left(\int \frac{1}{\|y-m\|^2}\mu_x(dy)\right)^{1/2}\left(\int \|y-y'\|^2 d\mu_{x,X}(y,y')\right)^{1/2}\right] \\ &\leq \frac{4\sqrt{C_6}}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)\mathcal{W}_2(\mu_X,\mu_x)\right] \\ &\leq \frac{4C_4\sqrt{C_6}}{h_n}\mathbb{E}\left[K^2\left(\frac{X-x}{h_n}\right)|X-x|^\beta\right] \\ &= 4C_4\sqrt{C_6}h_n^\beta \int K^2(y)\,|y|^\beta\, dy = \mathcal{O}(h_n^\beta) \xrightarrow[\text{n}\to\infty]{} 0.\end{aligned}$$

In the third term D_3, since Y_x is independent of X we may write

$$D_3 = \left(\frac{1}{h_n}\mathbb{E}\left[K^2((X-x)/h_n)\right] - p(x)v^2\right)\Sigma.$$

Thanks to (49), this converges to zero. Finally, by Proposition 4.1, $\Phi_{h_n}(Z_n)$ is almost surely bounded, and since $\|a \otimes b\|_{\text{op}} \leq \|a\|\,\|b\|$,

$$\|h_n\Phi_{h_n}(Z_n) \otimes \Phi_{h_n}(Z_n)\|_{\text{op}} \leq h_n\,\|\Phi_{h_n}(Z_n)\|^2 \xrightarrow[\text{n}\to\infty]{} 0, \quad a.s.$$

Therefore, Σ_n converges to Σ in the operator norm.

To prove the convergence of $\overline{\Sigma}_n$ in operator norm, observe that

$$\left\|\overline{\Sigma}_n - \Sigma\right\|_{\text{op}} \leq \frac{1}{\sum \frac{1}{h_k}} \sum \frac{1}{h_k} \left\|\Sigma_k - \Sigma\right\|_{\text{op}}.$$

Since $\left\|\Sigma_k - \Sigma\right\|_{\text{op}}$ converges to zero, the conclusion follows by the Toeplitz lemma.

Let us show that these convergences hold in the Hilbert–Schmidt norm. For any a, $\text{Tr}(a \otimes a) = \|a\|$. Therefore:

$$\begin{aligned}
\left\|\sqrt{\Sigma_n}\right\|_{H.S.}^2 &= \text{Tr}(\Sigma_n) \\
&= \frac{1}{h_n}\mathbb{E}\left[K^2((X-x)/h_n)\right] - h_n \left\|\Phi_{h_n}(Z_n)\right\|^2 \\
&\xrightarrow[\text{n}\to\infty]{} p(x)v^2 = \text{Tr}(\Sigma) = \left\|\sqrt{\Sigma}\right\|_{\text{H.S.}}^2.
\end{aligned}$$

Another application of the Toeplitz lemma shows that

$$\left\|\sqrt{\overline{\Sigma}_n}\right\|_{H.S.}^2 = \text{Tr}(\overline{\Sigma}_n) = \frac{1}{\sum_{k=1}^n \frac{1}{h_k}} \sum_{k=1}^n \text{Tr}\left(\frac{1}{h_k}\Sigma_k\right) \xrightarrow[n\to\infty]{} \text{Tr}(\Sigma) = \left\|\sqrt{\Sigma}\right\|_{\text{H.S.}}^2.$$

By the same reasoning as in Example 3.8.15 of (Bogachev, 1998), this implies the H.-S. convergences (46) and (47).

Finally, let P be an orthogonal projection operator. Choose a basis $(e_i)_{i\in\mathbb{N}}$ of orthonormal eigenvectors of P: $Pe_i = 0$ or $Pe_i = e_i$. Since $\overline{\Sigma}_n$ is trace-class, so is $\overline{\Sigma}_n P$ and:

$$\begin{aligned}
\text{Tr}\left(\overline{\Sigma}_n P\right) &= \sum_i \left\langle e_i, \overline{\Sigma}_n P e_i\right\rangle = \sum_i \left\langle P e_i, \overline{\Sigma}_n P e_i\right\rangle \\
&= \sum_i \left\langle \sqrt{\overline{\Sigma}_n} P e_i, \sqrt{\overline{\Sigma}_n} P e_i\right\rangle = \sum_i \left\|\sqrt{\overline{\Sigma}_n} P e_i\right\|^2 \\
&= \left\|\sqrt{\overline{\Sigma}_n} P\right\|_{\text{H.S.}}^2 \\
&\xrightarrow[n\to\infty]{} \|\Sigma P\|_{\text{H.S.}}^2 = \text{Tr}(\Sigma P).
\end{aligned}$$

This convergence is almost sure. Since $\text{Tr}(\Sigma_k) \leq \frac{1}{h_k}\mathbb{E}\left[K^2((X-x)/h_k)\right] \leq C$, the convergence also holds in L^1 by dominated convergence, and (48) holds. □

Step 1 — The CLT for the martingale. To prove the CLT (44), let us check that the assumptions of Theorem 5.1 in (Jakubowski, 1988) are fulfilled. Reminding (19), translated in our context, these assumptions are:

$$\forall \eta > 0 \quad \lim_{n\to\infty} \mathbb{P}\left(\sup_{1\leq k\leq n} \sqrt{\frac{h_n}{n}} \left\|\xi_{k+1}\right\| > \eta\right) = 0, \tag{50}$$

$$\text{a.s.} \quad \lim_{n\to\infty} \frac{1}{\sum_{k=1}^n \frac{1}{h_k}} \sum_{k=1}^n \left\langle \xi_{k+1}, e_i\right\rangle \left\langle \xi_{k+1}, e_j\right\rangle = \psi_{i,j}, \tag{51}$$

$$\forall \varepsilon > 0 \quad \lim_{N\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\frac{h_n}{n} \sum_{i=1}^n \sum_{j=N}^\infty \left\langle \xi_{i+1}, e_j\right\rangle^2 > \varepsilon\right) = 0, \tag{52}$$

where $(e_n)_{n\in\mathbb{N}}$ is an orthonormal basis of H and $\psi_{i,j} := \langle \Sigma e_i, e_j \rangle$.

We deal with condition (50) by applying Markov's inequality. Let $\eta > 0$.

$$\begin{aligned}\mathbb{P}\left(\sup_{1\le k\le n}\sqrt{\frac{h_n}{n}}\,\|\xi_{k+1}\| > \eta\right) &\le \sum_{k=1}^{n}\mathbb{P}\left(\sqrt{\frac{h_n}{n}}\,\|\xi_{k+1}\| > \eta\right)\\ &\le \frac{1}{\eta^p}\sum_{k=1}^{n}\mathbb{E}\left[\left(\frac{h_n}{n}\right)^{p/2}\|\xi_{k+1}\|^p\right],\end{aligned}$$

for any $p \ge 1$. We chose an integer p such that $p > 2$. By convexity of the function $x \mapsto x^p$, we have, for any n,

$$\|\xi_{n+1}\|^p \le 2^{p-1}\left(\frac{1}{h_n^p}K^p\left(\frac{X_{n+1}-x}{h_n}\right) + \|\Phi_{h_n}(Z_n)\|^p\right).$$

Thus an easy computation yields

$$\mathbb{E}\left[\|\xi_{n+1}\|^p\right] \le \frac{2^{p-1}p_{\max}\int_{\mathbb{R}}K^p(z)dz}{h_n^{p-1}} + 2^{p-1}\mathbb{E}\left[\|\Phi_{h_n}(Z_n)\|^p\right].$$

In the last term, $\Phi_{h_n}(Z_n)$ is bounded, thanks to (26). Consequently, there exists a constant $C(p)$ (independent of n) such that

$$\mathbb{E}\left[\|\xi_{n+1}\|^p\right] \le \frac{C(p)}{h_n^{p-1}}.$$

Hence we have, for a constant $C'(p)$ independent of n,

$$\mathbb{P}\left(\sup_{1\le k\le n}\sqrt{\frac{h_n}{n}}\,\|\xi_{k+1}\| > \eta\right) \le \frac{C(p)h_n^{p/2}}{n^{p/2}\eta^p}\sum_{k=1}^{n}h_k^{-p+1} \le \frac{C'(p)}{n^{p/2-h(p/2-1)-1}}.$$

Since $p > 2$, one has $p/2 - h(p/2-1) - 1 > 0$ and thus (50) holds.

Condition (51) is a consequence of the law of large numbers for martingales. Let us consider $(e_n)_{n\in\mathbb{N}}$ an orthonormal basis of H. From the decomposition

$$\langle \xi_{n+1}, e_i\rangle\langle \xi_{n+1}, e_j\rangle = \mathbb{E}\left[\langle \xi_{n+1}, e_i\rangle\langle \xi_{n+1}, e_j\rangle \middle| \mathcal{F}_n\right] + \varepsilon_{n+1},$$

with $\varepsilon_{n+1} := \langle \xi_{n+1}, e_i\rangle\langle \xi_{n+1}, e_j\rangle - \mathbb{E}\left[\langle \xi_{n+1}, e_i\rangle\langle \xi_{n+1}, e_j\rangle \middle| \mathcal{F}_n\right]$, we have

$$\begin{aligned}\frac{1}{\sum_{k=1}^{n}\frac{1}{h_k}}\sum_{k=1}^{n}\langle \xi_{k+1}, e_i\rangle\langle \xi_{k+1}, e_j\rangle &= \frac{1}{\sum_{k=1}^{n}\frac{1}{h_k}}\sum_{k=1}^{n}\mathbb{E}\left[\langle \xi_{k+1}, e_i\rangle\langle \xi_{k+1}, e_j\rangle \,|\mathcal{F}_k\right] + \frac{1}{\sum_{k=1}^{n}\frac{1}{h_k}}\sum_{k=1}^{n}\varepsilon_{k+1}\\ &= \langle e_i, \overline{\Sigma}_n e_j\rangle + \frac{1}{\sum_{k=1}^{n}\frac{1}{h_k}}\sum_{k=1}^{n}\varepsilon_{k+1}.\end{aligned}$$

By Lemma 4.2, the matrix element $\langle e_i, \overline{\Sigma}_n e_j\rangle$ converges to $\psi_{i,j}$. The law of large numbers for the martingale $(\sum_{k=1}^{n}\varepsilon_{k+1})_{n\in\mathbb{N}}$ whose increasing process is of order n^{1+3h} yields

$$\lim_{n\to\infty}\frac{1}{\sum_{k=1}^{n}\frac{1}{h_k}}\sum_{k=1}^{n}\varepsilon_{k+1} = 0 \quad \text{a.s.},$$

since $h < 1$, and condition (51) is satisfied.

It remains to check condition (52). Let $\varepsilon > 0$. Applying Markov's inequality, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{h_n}{n}\sum_{k=1}^{n}\sum_{j=N}^{\infty}\left\langle \xi_{k+1}, e_j\right\rangle^2 > \varepsilon\right) &\leq \frac{h_n}{n\varepsilon}\sum_{k=1}^{n}\sum_{j=N}^{\infty}\mathbb{E}\left[\left\langle \xi_{k+1}, e_j\right\rangle^2\right] \\
&\leq \frac{h_n}{n\varepsilon}\frac{\sum\frac{1}{h_k}}{\sum\frac{1}{h_k}}\sum_{k=1}^{n}\sum_{j=N}^{\infty}\mathbb{E}\left[\mathbb{E}\left[\left\langle \xi_{k+1}, e_j\right\rangle^2\middle|\mathcal{F}_k\right]\right] \\
&\leq \frac{h_n}{n\varepsilon}\left(\sum_{k=1}^{n}\frac{1}{h_k}\right)\mathbb{E}\left[\sum_{j=N}^{\infty}\left\langle e_j, \overline{\Sigma}_n e_j\right\rangle\right].
\end{aligned}$$

Call P_N the orthogonal projection on the e_i, $i \geq N$.

$$\mathbb{P}\left(\frac{h_n}{n}\sum_{k=1}^{n}\sum_{j=N}^{\infty}\left\langle \xi_{k+1}, e_j\right\rangle^2 > \varepsilon\right) \leq \frac{h_n}{n\varepsilon}\left(\sum_{k=1}^{n}\frac{1}{h_k}\right)\mathbb{E}\left[\mathrm{Tr}(\overline{\Sigma}_n P_N)\right].$$

Therefore

$$\limsup_n \mathbb{P}\left(\frac{h_n}{n}\sum_{k=1}^{n}\sum_{j=N}^{\infty}\left\langle \xi_{k+1}, e_j\right\rangle^2 > \varepsilon\right) \leq \frac{1}{(1+h)\varepsilon}\mathbb{E}\left[\mathrm{Tr}(\Sigma P_N)\right],$$

and (52) follows.

Step 2 — The remaining terms are negligible. Now, it remains to prove that all the other terms in (43) converge in probability to zero. Due to the equivalence (19), we have to prove the convergence in probability to zero of

$$\sqrt{\frac{h_n}{n}}\left(\frac{T_1\sqrt{h_1}}{\gamma_1} - A_n + A'_n - A''_n\right).$$

Recall that $\mathbb{E}\left[\mathbf{1}_{\Omega_N}\|T_n\|^2\right] \leq C_N\frac{\ln(n)}{n^\gamma h_n}$, thanks to Proposition 2.2. For the first term $A_n = \frac{T_{n+1}}{\gamma_n}$, we have:

$$\mathbb{E}\left[\mathbf{1}_{\Omega_N}\frac{h_n}{n}\|A_n\|^2\right] \leq \frac{C'_N\ln(n)}{n^{1-\gamma}},$$

therefore $\sqrt{\frac{h_n}{n}}.A_n \xrightarrow[\mathrm{n}\to\infty]{\mathrm{P}} 0$. Let us turn to the second term $A'_n = \sum_{k=2}^{n} T_k\left[\frac{1}{\gamma_k} - \frac{1}{\gamma_{k+1}}\right]$. Since there exists a constant C such that

$$\left|\frac{1}{\gamma_k} - \frac{1}{\gamma_{k+1}}\right| \leq Ck^{\gamma-1},$$

by applying Jensen's inequality together with Proposition 2.2, there is a positive constant C such that

$$\mathbb{E}\left[\sqrt{\frac{h_n}{n}}\|A'_n\|\mathbf{1}_{\Omega_N}\right] \leq C\sqrt{\ln(n)}n^{\gamma/2-1/2}.$$

Therefore $\sqrt{\frac{h_n}{n}}.A'_n \xrightarrow[\mathrm{n}\to\infty]{\mathrm{P}} 0$ since $\gamma < 1$.

Finally, for the last term $A_n'' = \sum_{k=1}^n \left(\delta_k + D_{h_k}(Z_k)\right)$, since on Ω_N, $\|\delta_k\| \leq C_r \|Z_k - m\|^2$, we have for the part in δ_k:

$$\mathbb{E}\left[\mathbf{1}_{\Omega_N}\left\|\sqrt{\frac{h_n}{n}}\sum_{k=1}^n \delta_k\right\|\right] \leq C\ln(n)n^{h/2-\gamma+1/2},$$

For the additional term, due to (27), we have $\mathbb{E}\left[\mathbf{1}_{\Omega_N}\left\|D_{h_k}(Z_k)\right\|\right] \leq Ch_k^\beta$ so that for some positive constant C,

$$\mathbb{E}\left[\mathbf{1}_{\Omega_N}\sqrt{\frac{h_n}{n}}\left\|A_n''\right\|\right] \leq Cn^{1/2-h/2-h\beta}.$$

The end of the proof follows the same guidelines as in Cardot et al. (2011).

Acknowledgements. We thank the company Médiamétrie for allowing us to illustrate our methodologies with their data.

References

Bogachev, V. I. (1998). *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Bosq, D. (2000). *Linear processes in function spaces*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York. Theory and applications.

Cadre, B. and Gannoun, A. (2000). Asymptotic normality of consistent estimate of the condition L_1-median. *Pub. Inst. Stat. Univ. Paris*, XXXXIV:13–33.

Cardot, H. (2007). Conditional functional principal components analysis. *Scandinavian J. of Statistics*, 34:317–335.

Cardot, H., Cénac, P., and Chaouch, M. (2010). Stochastic approximation to the multivariate and the functional median. In Lechevallier, Y. and Saporta, G., editors, *Compstat 2010*, pages 421–428. Physica Verlag, Springer.

Cardot, H., Cénac, P., and Zitt, P.-A. (2011). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. To appear in Bernoulli.

Chaouch, M. and Goga, C. (2012). Using complex surveys to estimate the L_1-median of a functional variable: application to electricity load curves. *International Statistical Review / Revue Internationale de Statistique*, 80:40–59.

Cheng, Y. and De Gooijer, J. (2007). On the uth geometric conditional quantile. *J. Statist. Plann. Inference*, 137:1914–1930.

Chiou, J. M., Müller, H. G., and Wang, J. L. (2004). Functional response models. *Statistica Sinica*, 14:675–693.

Cuevas, A., Febrero, M., and Fraiman, R. (2002). Linear functional regression : the case of fixed design and functional response. *Canadian J. of Statistics*, 30:285–300.

Duflo, M. (1997). *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, 39:254–261.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2011). Kernel regression with functional response. *Electronic J. of Statist.*, 5:159–171.

Givens, C. and Shortt, R. (1984). A class of Wassertstein metrics for probability distributions. *Michigan Math. J.*, 31:231–240.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic J. of Statist.*, 4:1022–1054.

Jakubowski, A. (1988). Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probab. Math. Statist.*, 9(1):95–114.

Kemperman, J. H. B. (1987). The median of a finite measure on a Banach space. In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam.

Lecoutre, J. (1990). Uniform consistency of a class of regression function estimators for Banach-space valued variables. *Statist. and Probability Letters*, 10:145–149.

Lian, H. (2007). Nonlinear functional models for functional response in reproducing kernel Hilbert spaces. *Canadian J. of Statistics*, 35:597–606.

Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic J. of Statist.*, 5:31–40.

Mokkadem, A., Pelletier, M., and Slaoui, Y. (2009). Revisiting Révész stochastic approximation method for the estimation of a regression function. *Alea*, 6:63–114.

Möttönen, J., Nordhausen, K., and Oja, H. (2010). Asymptotic theory of the spatial median. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, volume 7, pages 182–193. IMS Collection.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, second edition.

Révész, P. (1977). How to apply the method of stochastic approximation in the nonparametric estimation of a regression function. *Statistics*, 8:119–126.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*. Number 60 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.