Journal of Statistical Planning and Inference 139 (2009) 2459–2478

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

The stochastic approximation method for the estimation of a multivariate probability density

Abdelkader Mokkadem*, Mariane Pelletier, Yousri Slaoui

Département de Mathématiques, Université de Versailles-Saint-Quentin, 45, Avenue des Etats-Unis, 78035 Versailles Cedex, France

ARTICLE INFO

Article history:
Received 31 May 2007
Received in revised form
18 November 2008
Accepted 19 November 2008
Available online 6 December 2008

Keywords:
Density estimation
Stochastic approximation algorithm

ABSTRACT

We apply the stochastic approximation method to construct a large class of recursive kernel estimators of a probability density, including the one introduced by Hall and Patil [1994. On the efficiency of on-line density estimators. IEEE Trans. Inform. Theory 40, 1504–1512]. We study the properties of these estimators and compare them with Rosenblatt's nonrecursive estimator. It turns out that, for pointwise estimation, it is preferable to use the nonrecursive Rosenblatt's kernel estimator rather than any recursive estimator. A contrario, for estimation by confidence intervals, it is better to use a recursive estimator rather than Rosenblatt's estimator.

1. Introduction

The advantage of recursive estimators on their nonrecursive version is that their update, from a sample of size n to one of size $n+1$, requires considerably less computations. This property is particularly important in the framework of density estimation, since the number of points at which the function is estimated is usually very large. The first recursive version of Rosenblatt's kernel density estimator—and the most famous one—was introduced by Wolverton and Wagner (1969), and was widely studied; see among many others Yamato (1971), Davies (1973), Devroye (1979), Wegman and Davies (1979) and Roussas (1992). Competing recursive estimators, which may be regarded as weighted versions of Wolverton and Wagner's estimator, were introduced and studied by Deheuvels (1973), Wegman and Davies (1979) and Duflo (1997). Recently, Hall and Patil (1994) defined a large class of weighted recursive estimators, including all the previous recursive estimators. In this paper, we apply the stochastic approximation method to define a class of recursive kernel density estimators, which includes the one introduced by Hall and Patil (1994).

The most famous use of stochastic approximation algorithms in the framework of nonparametric statistics is the work of Kiefer and Wolfowitz (1952), who build up an algorithm which allows the approximation of the maximizer of a regression function. Their well-known algorithm was widely discussed and extended in many directions (see, among many others, Blum, 1954; Fabian, 1967; Kushner and Clark, 1978; Hall and Heyde, 1980; Ruppert, 1982; Chen, 1988; Spall, 1988, 1997; Polyak and Tsybakov, 1990; Duflo, 1996; Dippon and Renz, 1997; Chen et al., 1999; Dippon, 2003, and Mokkadem and Pelletier, 2007a). Stochastic approximation algorithms were also introduced by Révész (1973, 1977) to estimate a regression function, and by Tsybakov (1990) to approximate the mode of a probability density.

Let us recall Robbins–Monro's scheme to construct approximation algorithms of search of the zero z^* of an unknown function $h:\mathbb{R}\to\mathbb{R}$. First, $Z_0\in\mathbb{R}$ is arbitrarily chosen, and then the sequence (Z_n) is recursively defined by setting

$$Z_n = Z_{n-1} + \gamma_n W_n,$$

* Corresponding author.
E-mail addresses: mokkadem@math.uvsq.fr (A. Mokkadem), pelletier@math.uvsq.fr (M. Pelletier), slaoui@math.uvsq.fr (Y. Slaoui).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.11.012

where W_n is an "observation" of the function h at the point Z_{n-1}, and where the stepsize (γ_n) is a sequence of positive real numbers that goes to zero.

Let $X_1,\dots,X_n$ be independent, identically distributed $\mathbb{R}^d$-valued random vectors, and let f denote the probability density of X_1. To construct a stochastic algorithm, which approximates the function f at a given point x, we define an algorithm of search of the zero of the function $h: y \mapsto f(x)-y$. We thus proceed in the following way: (i) we set $f_0(x)\in\mathbb{R}$; (ii) for all $n\geqslant 1$, we set

$$f_n(x)=f_{n-1}(x)+\gamma_n W_n(x),$$

where $W_n(x)$ is an "observation" of the function h at the point $f_{n-1}(x)$. To define $W_n(x)$, we follow the approach of Révész (1973, 1977) and of Tsybakov (1990), and introduce a kernel K (that is, a function satisfying $\int_{\mathbb{R}^d}K(x)\,dx=1$) and a bandwidth (h_n) (that is, a sequence of positive real numbers that goes to zero), and set $W_n(x)=h_n^{-d}K(h_n^{-1}[x-X_n])-f_{n-1}(x)$. The stochastic approximation algorithm we introduce to recursively estimate the density f at the point x can thus be written as

$$f_n(x)=(1-\gamma_n)f_{n-1}(x)+\gamma_n h_n^{-d}K\left(\frac{x-X_n}{h_n}\right). \tag{1}$$

Let (w_n) be a positive sequence such that $\sum w_n=\infty$. When the stepsize (γ_n) is chosen equal to $(w_n[\sum_{k=1}^n w_k]^{-1})$, the estimator f_n defined by (1) can be rewritten as

$$f_n(x)=\frac{1}{\sum_{k=1}^n w_k}\sum_{k=1}^n w_k\frac{1}{h_k^d}K\left(\frac{x-X_k}{h_k}\right). \tag{2}$$

The class of estimators defined by the stochastic approximation algorithm (1) thus includes the general class of recursive estimators expressed as (2), and introduced in Hall and Patil (1994). In particular, the choice $(w_n)=1$ produces the estimator proposed by Wolverton and Wagner (1969), the choice $(w_n)=(h_n^{d/2})$ yields the estimator considered by Wegman and Davies (1979), and the choice $(w_n)=(h_n^d)$ gives the estimator considered by Deheuvels (1973) and Duflo (1997).

The aim of this paper is the study of the properties of the recursive estimator defined by the stochastic approximation algorithm (1), and its comparison with the well-known nonrecursive kernel density estimator introduced by Rosenblatt (1956) (see also Parzen, 1962), and defined as

$$\tilde{f}_n(x)=\frac{1}{nh_n^d}\sum_{k=1}^n K\left(\frac{x-X_k}{h_n}\right). \tag{3}$$

We first compute the bias and the variance of the recursive estimator f_n defined by (1). It turns out that they heavily depend on the choice of the stepsize (γ_n). In particular, for a given bandwidth, there is a trade-off in the choice of (γ_n) between minimizing either the bias or the variance of f_n. To determine the optimal choice of stepsize, we consider two points of view: pointwise estimation and estimation by confidence intervals.

From the pointwise estimation point of view, the criteria we consider to find the optimal stepsize is minimizing the mean squared error (MSE) or the integrated mean squared error (MISE). We display a set of stepsizes (γ_n) minimizing the MSE or the MISE of the estimator f_n defined by (1); we show in particular that the sequence $(\gamma_n)=(n^{-1})$ belongs to this set. The recursive estimator introduced by Wolverton and Wagner (1969) thus belongs to the subclass of recursive kernel estimators which have a minimum MSE or MISE (thanks to an adequate choice of the bandwidth, see Section 2.2). Let us underline that these minimum MSE and MISE are larger than those obtained for Rosenblatt's nonrecursive estimator $\tilde{f}_n$. Thus, for pointwise estimation and when rapid updating is not such important, it is preferable to use Rosenblatt's estimator rather than any recursive estimator defined by the stochastic approximation algorithm (1). Let us also mention that Hall and Patil (1994) introduce a class of on-line estimators, constructed from the class of the recursive estimators defined in (2); their on-line estimators are not recursive any more, but updating them requires much less operations than updating Rosenblatt's estimator, and their MSE and MISE are smaller than those of the recursive estimators (2).

Let us now consider the estimation from confidence interval point of view. Hall (1992) shows that, to minimize the coverage error of probability density confidence intervals, avoiding bias estimation by a slight undersmoothing is more efficient than explicit bias correction. In the framework of undersmoothing, minimizing the MSE comes down to minimizing the variance. We thus display a set of stepsizes (γ_n) minimizing the variance of f_n; we show in particular that, when the bandwidth (h_n) varies regularly with exponent $-a$, the sequence $(\gamma_n)=([1-ad]n^{-1})$ belongs to this set. Let us underline that the variance of the estimator f_n defined with this stepsize is smaller than that of Rosenblatt's estimator. Consequently, even in the case when the on-line aspect is not quite important, it is preferable to use recursive estimators to construct confidence intervals. The simulation results given in Section 3 are corroborating these theoretical results.

To complete the study of the asymptotic properties of the recursive estimator f_n, we give its pointwise strong convergence rate; we compare it with that of Rosenblatt's estimator $\tilde{f}_n$ for which laws of the iterated logarithm were established by Hall (1981) in the case $d=1$ and by Arcones (1997) in the multivariate framework.

The remainder of the paper is organized as follows. In Section 2, we state our main results: the bias and variance of f_n are given in Section 2.1, the pointwise estimation is considered in Section 2.2, the estimation by confidence intervals is developed in Section 2.3, and the strong convergence rate of f_n is stated in Section 2.4. Section 3 is devoted to our simulation results, and Section 4 to the proof of our theoretical results.

2. Assumptions and main results

We consider stepsizes and bandwidths, which belong to the following class of regularly varying sequences.

Definition 1. Let $\gamma \in \mathbb{R}$ and $(v_n)_{n \geqslant 1}$ be a nonrandom positive sequence. We say that $(v_n) \in \mathcal{GS}(\gamma)$ if

$$\lim_{n \to +\infty} n\left[1 - \frac{v_{n-1}}{v_n}\right] = \gamma. \tag{4}$$

Condition (4) was introduced by Galambos and Seneta (1973) to define regularly varying sequences (see also Bojanic and Seneta, 1973), and by Mokkadem and Pelletier (2007a) in the context of stochastic approximation algorithms. Typical sequences in $\mathcal{GS}(\gamma)$ are, for $b \in \mathbb{R}$, $n^{\gamma}(\log n)^b$, $n^{\gamma}(\log \log n)^b$, and so on.

The assumptions to which we shall refer are the following.

(A1) $K : \mathbb{R}^d \to \mathbb{R}$ is a continuous, bounded function satisfying $\int_{\mathbb{R}^d} K(z)\,dz = 1$, and, for all $j \in \{1, \dots d\}$, $\int_{\mathbb{R}} z_j K(z)\,dz_j = 0$ and $\int_{\mathbb{R}^d} z_j^2 |K(z)|\,dz < \infty$.

(A2) (i) $(\gamma_n) \in \mathcal{GS}(-\alpha)$ with $\alpha \in]1/2, 1]$.
(ii) $(h_n) \in \mathcal{GS}(-a)$ with $a \in]0, \alpha/d[$.
(iii) $\lim_{n\to\infty} (n\gamma_n) \in]\min\{2a, (1 - ad)/2\}, \infty]$.

(A3) f is bounded, twice differentiable, and, for all $i, j \in \{1, \dots d\}$, $\partial^2 f/\partial x_i \partial x_j$ is bounded.

Assumption (A2)(iii) on the limit of $(n\gamma_n)$ as n goes to infinity is usual in the framework of stochastic approximation algorithms. It implies in particular that the limit of $([n\gamma_n]^{-1})$ is finite. Throughout this paper we will use the following notation:

$$\xi = \lim_{n\to+\infty} (n\gamma_n)^{-1},$$
$$\mu_j^2 = \int_{\mathbb{R}^d} z_j^2 K(z)\,dz,$$
$$f_{ij}^{(2)}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x). \tag{5}$$

2.1. Bias and variance

Our first result is the following proposition, which gives the bias and the variance of f_n.

Proposition 1 (*Bias and variance of f_n*). *Let Assumptions* (A1)–(A3) *hold, and assume that, for all $i, j \in \{1, \dots d\}$, $f_{ij}^{(2)}$ is continuous at x.*

1. *If $a \leqslant \alpha/(d+4)$, then*

$$\mathbb{E}(f_n(x)) - f(x) = \frac{1}{2(1-2a\xi)} h_n^2 \sum_{j=1}^{d} (\mu_j^2 f_{jj}^{(2)}(x)) + o(h_n^2). \tag{6}$$

If $a > \alpha/(d+4)$, then

$$\mathbb{E}(f_n(x)) - f(x) = o(\sqrt{\gamma_n h_n^{-d}}). \tag{7}$$

2. *If $a \geqslant \alpha/(d+4)$, then*

$$Var(f_n(x)) = \frac{1}{2-(1-ad)\xi} \frac{\gamma_n}{h_n^d} f(x) \int_{\mathbb{R}^d} K^2(z)\,dz + o\left(\frac{\gamma_n}{h_n^d}\right). \tag{8}$$

If $a < \alpha/(d+4)$, then

$$Var(f_n(x)) = o(h_n^4). \tag{9}$$

3. *If $\lim_{n\to\infty} (n\gamma_n) > \max\{2a, (1-ad)/2\}$, then* (6) *and* (8) *hold simultaneously*.

The bias and the variance of the estimator f_n defined by the stochastic approximation algorithm (1) thus heavily depends on the choice of the stepsize (γ_n). Let us recall that the bias and variance of Rosenblatt's estimator $\tilde{f}_n$ are given by

$$\mathbb{E}(\tilde{f}_n(x)) - f(x) = \frac{1}{2}h_n^2\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x)) + o(h_n^2), \tag{10}$$

$$Var(\tilde{f}_n(x)) = \frac{1}{nh_n^d}f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z + o\left(\frac{1}{nh_n^d}\right). \tag{11}$$

To illustrate the results given by Proposition 1, we now give some examples of possible choices of (γ_n), and compare the bias and variance of f_n with those of $\tilde{f}_n$.

Example 1 (*Choices of* (γ_n) *minimizing the bias of* f_n). In view of (6), the asymptotic bias of $f_n(x)$ is minimum when $\xi = 0$, that is, when (γ_n) is chosen such that $\lim_{n\to\infty}(n\gamma_n) = \infty$, and we then have

$$\mathbb{E}[f_n(x)] - f(x) = \frac{1}{2}h_n^2\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x)) + o(h_n^2).$$

In view of (10), the order of the bias of the recursive estimator f_n is thus always greater or equal to that of Rosenblatt's estimator. Let us also mention that choosing the stepsize such that $\lim_{n\to\infty} n\gamma_n = \infty$ (in which case the bias of f_n is equivalent to that of Rosenblatt's estimator) is absolutely unadvised since we then have

$$\lim_{n\to\infty}\frac{Var(\tilde{f}_n(x))}{Var(f_n(x))} = 0.$$

Example 2 (*Choices of* (γ_n) *minimizing the variance of* f_n). As mentioned in the Introduction, it is advised to minimize the variance of f_n for interval estimation.

Corollary 1. *Let the assumptions of Proposition* 1 *hold with* $f(x) > 0$. *To minimize the asymptotic variance of* f_n, α *must be chosen equal to* 1, (γ_n) *must satisfy* $\lim_{n\to\infty} n\gamma_n = 1 - ad$, *and we then have*

$$Var[f_n(x)] = \frac{1-ad}{nh_n^d}f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z + o\left(\frac{1}{nh_n^d}\right).$$

It follows from Corollary 1 and (11) that, thanks to an adequate choice of (γ_n), the variance of the recursive estimator f_n can be smaller than that of Rosenblatt's estimator. To see better the comparison with Rosenblatt's estimator, let us set $(h_n) \in \mathcal{GS}(-1/[d+4])$ (which is the choice leading in particular to the minimum MSE of Rosenblatt's estimator). When (γ_n) is chosen in $\mathcal{GS}(-1)$ and such that $\lim_{n\to\infty} n\gamma_n = 1 - d/[d+4]$, we have

$$\lim_{n\to\infty}\frac{\mathbb{E}(\tilde{f}_n(x)) - f(x)}{\mathbb{E}(f_n(x)) - f(x)} = \frac{1}{2}, \quad \lim_{n\to\infty}\frac{Var(\tilde{f}_n(x))}{Var(f_n(x))} = \frac{d+4}{4}. \tag{12}$$

It is interesting to note that, whatever the dimension d is, the bias of the recursive estimator f_n is equivalent to twice that of Rosenblatt's estimator, whereas the ratio of the variances goes to infinity as the dimension d increases.

To conclude this example, let us mention that the most simple stepsize satisfying the conditions required in Corollary 1 is $(\gamma_n) = ([1-ad]n^{-1})$.

Example 3 (*The class of recursive estimators introduced by Hall and Patil, 1994*). The following lemma ensures that Proposition 1 gives the bias and variance of the recursive estimators defined in (2) and introduced by Hall and Patil (1994) for a large choice of weights (w_n).

Lemma 1. *Set* $(w_n) \in \mathcal{GS}(w^*)$ *and* $(\gamma_n) = (w_n[\sum_{k=1}^n w_k]^{-1})$. *If* $w^* > -1$, *then* $(\gamma_n) \in \mathcal{GS}(-1)$ *and* $\lim_{n\to\infty} n\gamma_n = 1 + w^*$.

Set $(h_n) \in \mathcal{GS}(-a)$; we give explicitly here the bias and variance of three particular recursive estimators.

- When $(w_n) = 1$, f_n is the estimator introduced by Wolverton and Wagner (1969); in view of Lemma 1, Proposition 1 applies with $\xi = 1$, and we have

$$\mathbb{E}(f_n(x)) - f(x) = \frac{1}{2(1-2a)} h_n^2 \sum_{j=1}^{d} (\mu_j^2 f_{jj}^{(2)}(x)) + o(h_n^2),$$

$$Var(f_n(x)) = \frac{1}{1+ad} \frac{1}{nh_n^d} f(x) \int_{\mathbb{R}^d} K^2(z)\,dz + o\left(\frac{\gamma_n}{h_n^d}\right).$$

- When $(w_n) = (h_n^{d/2})$, f_n is the estimator considered by Wegman and Davies (1979); in view of Lemma 1, Proposition 1 applies with $\xi = (1 - ad/2)^{-1}$, and we have

$$\mathbb{E}(f_n(x)) - f(x) = \frac{2-ad}{2(2-[4+d]a)} h_n^2 \sum_{j=1}^{d} (\mu_j^2 f_{jj}^{(2)}(x)) + o(h_n^2),$$

$$Var(f_n(x)) = \frac{(2-ad)^2}{4} \frac{1}{nh_n^d} f(x) \int_{\mathbb{R}^d} K^2(z)\,dz + o\left(\frac{\gamma_n}{h_n^d}\right).$$

- When $(w_n) = (h_n^d)$, f_n is the estimator introduced by Deheuvels (1973) and whose convergence rate was established by Duflo (1997); in view of Lemma 1, Proposition 1 applies with $\xi = (1-ad)^{-1}$, and we have

$$\mathbb{E}(f_n(x)) - f(x) = \frac{1-ad}{2(1-[2+d]a)} h_n^2 \sum_{j=1}^{d} (\mu_j^2 f_{jj}^{(2)}(x)) + o(h_n^2),$$

$$Var(f_n(x)) = \frac{1-ad}{nh_n^d} f(x) \int_{\mathbb{R}^d} K^2(z)\,dz + o\left(\frac{\gamma_n}{h_n^d}\right).$$

Let us underline that the bias and variance of this estimator are equivalent to those of the estimator defined with the stepsize $(\gamma_n = ([1-ad]n^{-1})$ (this choice minimizing the variance of f_n, see Corollary 1), but its updating is less straightforward.

2.2. Choice of the optimal stepsize for point estimation

We first explicit the choices of (γ_n) and (h_n), which minimize the MSE and MISE of the recursive estimator defined by the stochastic approximation algorithm (1), and then provide a comparison with Rosenblatt's estimator.

2.2.1. Choices of (γ_n) minimizing the MSE of f_n

Corollary 2. *Let Assumptions* (A1)–(A3) *hold, assume that* $f(x) > 0$, $\sum_{j=1}^d (\mu_j^2 f_{jj}^{(2)}(x)) \neq 0$, *and that, for all* $i, j \in \{1, \dots d\}$, $f_{ij}^{(2)}$ *is continuous at x. To minimize the MSE of* f_n *at the point x, the stepsize* (γ_n) *must be chosen in* $\mathcal{GS}(-1)$ *and such that* $\lim_{n\to\infty} n\gamma_n = 1$, *the bandwidth* (h_n) *must equal*

$$\left(\left[\frac{d(d+2)}{2(d+4)} \frac{f(x)\int_{\mathbb{R}^d} K^2(z)\,dz}{(\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x))^2}\right]^{1/(d+4)} \gamma_n^{1/(d+4)}\right),$$

and we then have

$$MSE = n^{-4/(d+4)} \frac{(d+4)^{(3d+8)/(d+4)}}{d^{d/(d+4)} 4^{(d+6)/(d+4)} (d+2)^{(2d+4)/(d+4)}} \left[\sum_{j=1}^{d} \mu_j^2 f_{jj}^{(2)}(x)\right]^{2d/(d+4)} \left[f(x)\int_{\mathbb{R}^d} K^2(z)\,dz\right]^{4/(d+4)} [1+o(1)].$$

The most simple example of stepsize belonging to $\mathcal{GS}(-1)$ and such that $\lim_{n\to\infty} n\gamma_n = 1$ is $(\gamma_n) = (n^{-1})$. For this choice of stepsize, the estimator f_n defined by (1) equals the recursive kernel estimator introduced by Wolverton and Wagner (1969). This latest estimator thus belongs to the subclass of recursive kernel estimators, which, thanks to an adequate choice of the bandwidth, have a minimum MSE.

2.2.2. Choices of (γ_n) minimizing the MISE of f_n

The following proposition gives the MISE of the estimator f_n.

Proposition 2. *Let Assumptions* (A1)–(A3) *hold, and assume that, for all* $i,j \in \{1,\dots d\}$, $f_{ij}^{(2)}$ *is continuous and integrable.*

1. *If* $a < \alpha/(d+4)$, *then*

$$MISE = \frac{1}{4(1-2a\xi)^2} h_n^4 \int_{\mathbb{R}^d} \left[\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x)\right]^2 \mathrm{d}x + \mathrm{o}(h_n^4).$$

2. *If* $a = \alpha/(d+4)$, *then*

$$MISE = \frac{1}{4(1-2a\xi)^2} h_n^4 \int_{\mathbb{R}^d} \left[\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x)\right]^2 \mathrm{d}x + \frac{1}{2-(1-ad)\xi}\frac{\gamma_n}{h_n^d}\int_{\mathbb{R}^d} K^2(z)\,\mathrm{d}z + \mathrm{o}\left(h_n^4 + \frac{\gamma_n}{h_n^d}\right).$$

3. *If* $a > \alpha/(d+4)$, *then*

$$MISE = \frac{1}{2-(1-ad)\xi}\frac{\gamma_n}{h_n^d}\int_{\mathbb{R}^d} K^2(z)\,\mathrm{d}z + \mathrm{o}\left(\frac{\gamma_n}{h_n^d}\right).$$

The following corollary ensures that Wolwerton and Wagner's estimator also belongs to the subclass of kernel estimators defined by the stochastic approximation algorithm (1), which, thanks to an adequate choice of the bandwidth, have a minimum MISE.

Corollary 3. *Let Assumptions* (A1)–(A3) *hold, and assume that, for all* $i,j \in \{1,\dots d\}$, $f_{ij}^{(2)}$ *is continuous and integrable. To minimize the MISE of* f_n, *the stepsize* (γ_n) *must be chosen in* $\mathcal{GS}(-1)$ *and such that* $\lim_{n\to\infty} n\gamma_n = 1$, *the bandwidth* (h_n) *must equal*

$$\left(\left[\frac{d(d+2)}{2(d+4)}\frac{\int_{\mathbb{R}^d} K^2(z)\,\mathrm{d}z}{\int_{\mathbb{R}^d}(\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x))^2\,\mathrm{d}x}\right]^{1/(d+4)} \gamma_n^{1/(d+4)}\right),$$

and we then have

$$MISE = n^{-4/(d+4)}\frac{(d+4)^{(3d+8)/(d+4)}}{d^{d/(d+4)}4^{(d+6)/(d+4)}(d+2)^{(2d+4)(d+4)}}\left[\int_{\mathbb{R}^d}\left(\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x)\right)^2 \mathrm{d}x\right]^{d/(d+4)}\left[\int_{\mathbb{R}^d} K^2(z)\,\mathrm{d}z\right]^{4/(d+4)}[1+\mathrm{o}(1)].$$

2.2.3. Comparison with Rosenblatt's estimator

The ratio of the optimal MSE (or MISE) of Rosenblatt's estimator to that of Wolwerton and Wagner's estimator equals

$$\rho(d) = \left[\frac{2^4(d+2)^{2d+4}}{(d+4)^{2d+4}}\right]^{1/(d+4)}.$$

This ratio is always less than one, it at first decreases, and then increases to one as the dimension d increases. This phenomenon is similar to that observed by Hall and Patil (1994). The former authors consider the univariate framework, but look at the efficiency of Wolwerton and Wagner's estimator of the sth-order derivative of f relative to Rosenblatt's one; the ratio $\rho(s)$ varies in s in the same way as $\rho(d)$ does in d. According to pointwise estimation point of view, and when rapid updating is not too important, it is thus preferable to use Rosenblatt's nonrecursive estimator rather than any recursive estimator defined by the stochastic approximation algorithm (1). Let us mention that Hall and Patil (1994) introduce a class of on-line estimators, constructed from the class of the recursive estimators defined in (2); their on-line estimators are not recursive any more, but updating them requires much less operations than updating Rosenblatt's estimator, and their MSE and MISE are smaller than those of the recursive estimators (2).

2.3. Choice of the optimal stepsize for interval estimation

Let us first state the following theorem, which gives the weak convergence rate of the estimator f_n defined in (1).

Theorem 1 (*Weak pointwise convergence rate*). *Let Assumptions* (A1)–(A3) *hold, assume that* $f(x)>0$ *and that, for all* $i,j\in\{1,\dots d\}$, $f_{ij}^{(2)}$ *is continuous at* x.

1. *If there exists* $c\geqslant 0$ *such that* $\gamma_n^{-1}h_n^{d+4}\to c$, *then*

$$\sqrt{\gamma_n^{-1}h_n^d}(f_n(x)-f(x))\xrightarrow{\mathscr{D}}\mathscr{N}\left(\frac{c^{1/2}}{2(1-2a\xi)}\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x)),\frac{1}{2-(1-ad)\xi}f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z\right).$$

2. *If* $\gamma_n^{-1}h_n^{d+4}\to\infty$, *then*

$$\frac{1}{h_n^2}(f_n(x)-f(x))\xrightarrow{\mathbb{P}}\frac{1}{2(1-2a\xi)}\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x)),$$

where $\xrightarrow{\mathscr{D}}$ *denotes the convergence in distribution*, $\mathscr{N}$ *the Gaussian-distribution and* $\xrightarrow{\mathbb{P}}$ *the convergence in probability*.

As mentioned in the Introduction, Hall (1992) shows that, to minimize the coverage error of probability density confidence intervals, avoiding bias estimation by a slight undersmoothing is more efficient than bias correction. Let us recall that, when the bandwidth (h_n) is chosen such that $\lim_{n\to\infty}nh_n^{d+4}=0$ (which corresponds to undersmoothing), Rosenblatt's estimator fulfils the central limit theorem

$$\sqrt{nh_n^d}(\tilde{f}_n(x)-f(x))\xrightarrow{\mathscr{D}}\mathscr{N}\left(0,f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z\right). \tag{13}$$

Now, let Φ denote the distribution function of the $\mathscr{N}(0,1)$, let $t_{\alpha/2}$ be such that $\Phi(t_{\alpha/2})=1-\alpha/2$ (where $\alpha\in]0,1[$), and set

$$I_{g_n}(x)=\left[g_n(x)-t_{\alpha/2}C(g_n)\sqrt{\frac{g_n(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}{nh_n^d}},g_n(x)+t_{\alpha/2}C(g_n)\sqrt{\frac{g_n(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}{nh_n^d}}\right].$$

In view of (13), the asymptotic level of $I_{\tilde{f}_n}(x)$ equals $1-\alpha$ for $C(\tilde{f}_n)=1$. The following corollary gives the values of $C(f_n)$ for which the asymptotic level of $I_{f_n}(x)$ equals $1-\alpha$ too.

Corollary 4. *Let the assumptions of Theorem* 1 *hold with* $\lim_{n\to\infty}n\gamma_n=\gamma_0\in]0,\infty[$ *and* $\lim_{n\to\infty}nh_n^{d+4}=0$. *The asymptotic level of* $I_{f_n}(x)$ *equals* $1-\alpha$ *for*

$$C(f_n)=\sqrt{\gamma_0[2-(1-ad)\gamma_0^{-1}]^{-1}}.$$

Moreover, the minimum of $C(f_n)$ *is reached at* $\gamma_0=1-ad$ *and equals* $\sqrt{1-ad}$.

The optimal stepsizes for interval estimation are thus the sequences $(\gamma_n)\in\mathscr{GS}(-1)$ such that $\lim_{n\to\infty}n\gamma_n=1-ad$, the most simple one being $(\gamma_n)=([1-ad]n^{-1})$. Of course, these stepsizes are those which minimize the variance of f_n (see Corollary 1).

2.4. Strong pointwise convergence rate

The following theorem gives the strong pointwise convergence rate of f_n.

Theorem 2 (*Strong pointwise convergence rate*). *Let Assumptions* (A1)–(A3) *hold, and assume that, for all* $i,j\in\{1,\dots d\}$, $f_{ij}^{(2)}$ *is continuous at* x.

1. *If there exists* $c_1\geqslant 0$ *such that* $\gamma_n^{-1}h_n^{d+4}/(\ln[\sum_{k=1}^n\gamma_k])\to c_1$, *then, with probability one, the sequence*

$$\left(\sqrt{\frac{\gamma_n^{-1}h_n^d}{2\ln[\sum_{k=1}^n\gamma_k]}}(f_n(x)-f(x))\right)$$

is relatively compact and its limit set is the interval

$$\left[\frac{1}{2(1-2a\xi)}\sqrt{\frac{c_1}{2}}\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x)) - \sqrt{\frac{f(x)}{2-(1-ad)\xi}\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z},\right.$$
$$\left.\frac{1}{2(1-2a\xi)}\sqrt{\frac{c_1}{2}}\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x)) + \sqrt{\frac{f(x)}{2-(1-ad)\xi}\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}\right].$$

2. *If* $\gamma_n^{-1}h_n^{d+4}/(\ln[\sum_{k=1}^{n}\gamma_k]) \to \infty$, *then, with probability one,*

$$\lim_{n\to\infty}\frac{1}{h_n^2}(f_n(x)-f(x)) = \frac{1}{2(1-2a\xi)}\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x)).$$

Set (h_n) such that $\lim_{n\to\infty} nh_n^{d+4}/\ln\ln n = 0$. Arcones (1997) proves the following compact law of the iterated logarithm for Rosenblatt's estimator: with probability one, the sequence $(\sqrt{nh_n^d}(\tilde{f}_n(x)-f(x))/\sqrt{2\ln\ln n})$ is relatively compact and its limit set is the interval

$$J = \left[-\sqrt{f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}, \sqrt{f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}\right].$$

Now, set (γ_n) such that $\lim_{n\to\infty} n\gamma_n = \gamma_0 \in]0,\infty[$. The first part of Theorem 2 ensures that, with probability one, the limit set of the sequence $(\sqrt{nh_n^d}(f_n(x)-f(x))/\sqrt{2\ln\ln n})$ is the interval

$$J(\gamma_0) = \left[-A(\gamma_0)\sqrt{f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}, A(\gamma_0)\sqrt{f(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}\right] \quad \text{with } A(\gamma_0) = \sqrt{\frac{\gamma_0}{[2-(1-ad)\gamma_0^{-1}]}}.$$

In particular, for Wolwerton and Wagner's estimator, $A(\gamma_0) = 1/\sqrt{1+ad}$; for the estimator considered by Wegman and Davies (1979), or when $(\gamma_n) = ([1-ad/2]n^{-1})$, $A(\gamma_0) = 1-ad/2$; for the estimator considered by Deheuvels (1973) and Duflo (1997), or when $(\gamma_n) = ([1-ad]n^{-1})$, $A(\gamma_0) = \sqrt{1-ad}$. For all these recursive estimators, the length of the limit interval $J(\gamma_0)$ is smaller than that of J, which shows that they are more concentrated around f than Rosenblatt's estimator is.

3. Simulations

The aim of our simulation studies is to compare the performance of Rosenblatt's estimator defined in (3) with that of the recursive estimators, from confidence interval point of view. Of course, the recursive estimator we consider here is the optimal one according to this criteria (see Corollary 4). We set

$$I_{i,n} = \left[g_n(x) - 1.96C(g_n)\sqrt{\frac{g_n(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}{nh_n^d}}, g_n(x) + 1.96C(g_n)\sqrt{\frac{g_n(x)\int_{\mathbb{R}^d}K^2(z)\,\mathrm{d}z}{nh_n^d}}\right],$$

where:

- if $i=1$, then $g_n = \tilde{f}_n$ is Rosenblatt's estimator, and $C(g_n) = 1$;
- if $i=2$, then $g_n = f_n$ is the optimal recursive estimator defined by the algorithm (1) with the stepsize $(\gamma_n) = ([1-ad]n^{-1})$, and $C(g_n) = \sqrt{1-ad}$.

According to the theoretical results given in Section 2.3, both confidence intervals $I_{1,n}$ and $I_{2,n}$ have the same asymptotic level (equal to 95%), whereas $I_{2,n}$ has a smaller length than $I_{1,n}$. In order to investigate their finite sample behaviours, we consider three sample sizes: $n = 50$, 100, and 200. In each case, the number of simulations is $N = 5000$. Tables 1–4 give (for different values of d, f, x, and (h_n)):

- the empirical levels $\#\{f(x) \in I_{i,n}\}/N$ at each first line concerning $I_{i,n}$.
- the averaged lengths of the intervals $I_{i,n}$ at each second line concerning $I_{i,n}$.

The case $d=1$. In the univariate framework, we consider two densities f: the standard normal $\mathcal{N}(0,1)$ distribution (see Table 1), and the normal mixture $\frac{1}{2}\mathcal{N}(-\frac{1}{2},1)+\frac{1}{2}\mathcal{N}(\frac{1}{2},1)$ distribution (see Table 2). The points at which f is estimated are: $x = 0$, 0.5, and 1. The bandwidth (h_n) is set equal to (n^{-a}) with $a = 0.21$ and 0.23 (the parameter a being chosen slightly larger than $\frac{1}{5}$ to slightly

Table 1
$X \rightsquigarrow \mathcal{N}(0,1)$.

	$x=0$			$x=0.5$			$x=1$		
	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$
$a=0.21$									
$I_{1,n}$	96.74%	96.08%	95.74%	97.1%	96.74%	96.96%	97.72%	97.44%	97.7%
	0.2681	0.2061	0.158	0.2538	0.1948	0.1493	0.2168	0.165	0.126
$I_{2,n}$	99.36%	98%	96.18%	99.76%	98.96%	98.36%	98.86%	98.76%	98.78%
	0.2436	0.184	0.140	0.2332	0.1755	0.1331	0.2068	0.1529	0.1146
$a=0.23$									
$I_{1,n}$	96.58%	96.46%	96.78%	96.78%	97.06%	97.04%	97.32%	97.58%	96.96%
	0.2796	0.2167	0.1674	0.2653	0.205	0.1579	0.225	0.1731	0.1328
$I_{2,n}$	99.46%	98.58%	97.58%	99.6%	99.26%	98.72%	98.68%	98.32%	97.96%
	0.2517	0.1915	0.1467	0.2415	0.1828	0.1393	0.2134	0.159	0.1197

$I_{1,n}$ is the Rosenblatt's interval, $I_{2,n}$ the recursive interval. Two values of a are considered. For each value of a and for each $I_{i,n}$, $i=1,2$, the first line gives the empirical level and the second one the averaged length of $I_{i,n}$.

Table 2
$X \rightsquigarrow \frac{1}{2}\mathcal{N}(-\frac{1}{2},1)+\frac{1}{2}\mathcal{N}(\frac{1}{2},1)$.

	$x=0$			$x=0.5$			$x=1$		
	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$
$a=0.21$									
$I_{1,n}$	96.86%	96.96%	96.86%	96.96%	96.68%	96.8%	97.12%	97.04%	96.94%
	0.2541	0.1949	0.1493	0.2436	0.1866	0.1427	0.2142	0.1642	0.1251
$I_{2,n}$	99.76%	99.04%	98.2%	99.62%	99.28%	98.72%	99.14%	98.94%	98.4%
	0.2334	0.1755	0.1331	0.2257	0.1692	0.1278	0.2045	0.1518	0.1136
$a=0.23$									
$I_{1,n}$	96.92%	97.04%	96.84%	96.56%	96.66%	97.14%	97.02%	97.12%	96.76%
	0.2654	0.2049	0.1579	0.254	0.196	0.151	0.2233	0.1717	0.1321
$I_{2,n}$	99.9%	99.18%	98.76%	99.74%	99.3%	98.92%	98.78%	98.76%	98.2%
	0.2416	0.1826	0.1393	0.2334	0.176	0.1338	0.2116	0.1575	0.1187

$I_{1,n}$ is the Rosenblatt's interval, $I_{2,n}$ the recursive interval. Two values of a are considered. For each value of a and for each $I_{i,n}$, $i=1,2$, the first line gives the empirical level and the second one the averaged length of $I_{i,n}$.

Table 3
$X=AY$ with $Y \rightsquigarrow \mathcal{N}(0,I_2)$.

	$x=(0,0)$			$x=(0.5,0.5)$			$x=(1,1)$		
	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$
$a=0.17$									
$I_{1,n}$	93.82%	94.98%	96.9%	91.06%	92.82%	94.0%	89.48%	86.88%	85.82%
	0.1159	0.0934	0.0757	0.1059	0.0854	0.0686	0.0811	0.0645	0.0515
$I_{2,n}$	97.54%	95.12%	94.34%	96.74%	94.62%	92.86%	97.2%	94.32%	91.16%
	0.0979	0.0765	0.061	0.091	0.0707	0.0558	0.0736	0.0557	0.0432
$a=0.19$									
$I_{1,n}$	95.64%	97.08%	97.28%	93.46%	94.84%	95.82%	91.58%	91.06%	89.04%
	0.1271	0.1042	0.0851	0.1158	0.0946	0.077	0.0883	0.0713	0.0574
$I_{2,n}$	97.5%	97.26%	96.64%	97.22%	96.5%	95.42%	96.74%	95.66%	92.24%
	0.1045	0.0829	0.0666	0.0969	0.0763	0.0609	0.0783	0.0599	0.0469
$a=0.21$									
$I_{1,n}$	96.68%	97.62%	98.24%	95.16%	96.48%	97.16%	92.76%	91.2%	91.04%
	0.1392	0.1157	0.0957	0.1267	0.105	0.0863	0.0962	0.0783	0.0641
$I_{2,n}$	97.16%	97.48%	97.56%	96.96%	96.84%	96.7%	96.72%	96.58%	94.2%
	0.1111	0.0893	0.0726	0.1031	0.0822	0.0662	0.0832	0.0642	0.0509

$I_{1,n}$ is the Rosenblatt's interval, $I_{2,n}$ the recursive interval. Three values of a are considered. For each value of a and for each $I_{i,n}$, $i=1,2$, the first line gives the empirical level and the second one the averaged length of $I_{i,n}$. Note that the recursive interval $I_{2,n}$ computed with $a=0.21$ has a larger empirical level and a smaller length than the Rosenblatt's interval $I_{1,n}$ computed with $a=0.17$.

undersmooth). Both tables show that the recursive estimator performs better than Rosenblatt's one: the empirical levels of the intervals $I_{2,n}$ are greater than those of $I_{1,n}$, whereas their averaged lengths are smaller. Moreover, it can be seen that increasing the undersmoothing (that is, increasing a) has the effect to increase the level, but also to increase the length of the intervals. However, it appears (in view of Tables 1 and 2) that the length of the recursive interval increases less than that of the Rosenblatt's interval; this is expected since the factor $C(g_n)=\sqrt{1-a}$, which appears in the definition of the recursive interval, decreases as a increases.

Table 4
$X=AY$ with $Y\rightsquigarrow\frac{1}{2}\mathcal{N}(-B,I_2)+\frac{1}{2}\mathcal{N}(B,I_2)$.

	$x=(0,0)$			$x=(0.5,0.5)$			$x=(1,1)$		
	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$	$n=50$	$n=100$	$n=200$
$a=0.17$									
$I_{1,n}$	91.84%	91.28%	92.4%	90.06%	89.42%	87.86%	83.24%	80.46%	78.88%
	0.105	0.0847	0.068	0.0976	0.0785	0.063	0.0787	0.0631	0.050
$I_{2,n}$	96.8%	93.76%	91.34%	95.9%	92.32%	86.96%	95.52%	87.6%	82.12%
	0.0903	0.0702	0.0553	0.0851	0.0657	0.0516	0.0716	0.0544	0.0419
$a=0.19$									
$I_{1,n}$	93.54%	93.94%	95.44%	90.72%	91.38%	92.12%	85.46%	84.24%	82.24%
	0.1151	0.094	0.0764	0.1158	0.1069	0.0706	0.0857	0.0692	0.0457
$I_{2,n}$	97.42%	95.92%	94.38%	97.22%	97.06%	91.74%	96.18%	91.26%	86.88%
	0.0964	0.0757	0.0604	0.0969	0.0908	0.0562	0.0762	0.0582	0.0469
$a=0.21$									
$I_{1,n}$	94.82%	96.12%	97.44%	93.14%	93.46%	94.16%	88.72%	86.24%	83.54%
	0.1259	0.1037	0.0858	0.1163	0.0962	0.0793	0.0935	0.0764	0.0624
$I_{2,n}$	97.1%	97.48%	96.96%	96.82%	96.04%	93.96%	96.76%	93.52%	88.24%
	0.1025	0.0813	0.0659	0.0963	0.0762	0.0613	0.0811	0.0627	0.0495
$a=0.24$									
$I_{1,n}$	96.26%	97.48%	98.38%	94.36%	96.16%	96.7%	91.04%	91.08%	89.42%
	0.1435	0.1208	0.1017	0.1325	0.1117	0.0937	0.1058	0.0885	0.0736
$I_{2,n}$	96.18%	97.54%	98.04%	96.68%	97.38%	96.6%	96.98%	95.96%	91.3%
	0.1117	0.0903	0.0743	0.1049	0.0845	0.0691	0.0883	0.0695	0.0558

$I_{1,n}$ is the Rosenblatt's interval, $I_{2,n}$ the recursive interval. Four values of a are considered. For each value of a and for each $I_{i,n}$, $i=1,2$, the first line gives the empirical level and the second one the averaged length of $I_{i,n}$. Note that the recursive intervals $I_{2,n}$ computed with $a=0.21$ and $a=0.24$ have a larger empirical level and a smaller length than the Rosenblatt's intervals $I_{1,n}$ computed with $a=0.17$ and 0.19.

The case $d=2$. In the case when $d=2$, we estimate the density f of the random vector X defined as $X=AY$ with

$$A=\begin{pmatrix}1 & 0\\ 0.5 & 1\end{pmatrix},$$

and where the distribution of the random vector Y is:

- the normal standard distribution $\mathcal{N}(0,I_2)$ (see Table 3);
- the normal mixture $\frac{1}{2}\mathcal{N}(-B,I_2)+\frac{1}{2}\mathcal{N}(B,I_2)$ with $B=\binom{0.5}{0.5}$ (see Table 4).

The points at which f is estimated are: $x=(0,0)$, $(0.5,0.5)$, and $(1,1)$. The bandwidth (h_n) is set equal to (n^{-a}). To slightly undersmooth, the parameter a must be chosen slightly larger than $\frac{1}{6}$; we first chose $a=0.17$ and 0.19. Tables 3 and 4 show that, for these given values of the parameter a, the recursive estimator performs better for the sample sizes $n=50$ and 100, whereas, at first glance, Rosenblatt's estimator performs better in the case when $n=200$. This is explained by the fact that, for this latest sample size, the length of $I_{2,n}$ becomes too small. We have thus added other choices of the parameter a (a=0.21 in Table 3; a=0.21 and 0.24 in Table 4). The larger a is, the larger the length of the intervals $I_{i,n}$ are, and the larger the empirical levels are. Now, Tables 3 and 4 also show that, for the sample size $n=200$, the intervals $I_{2,n}$ computed with $a=0.21$ or 0.24 have a smaller length and a higher level than the intervals $I_{1,n}$ computed with $a=0.17$ or 0.19, so that we can say again that the recursive estimator performs better than Rosenblatt's one.

This simulation study shows the good performance of the recursive estimator defined by the algorithm (1) with the stepsize $(\gamma_n)=([1-ad]n^{-1})$ for interval estimation. The main question which remains open is how to choose the bandwidth (h_n) in $\mathcal{GS}(-a)$, and, in particular, how to determine the parameter a. This problem is not particular to the framework of recursive estimation; in the case when Rosenblatt's estimator is used, Hall (1992) enlightens that criteria to determine the "good undersmoothing" are not easy to determine empirically.

4. Proofs

Throughout this section we use the following notation:

$$\Pi_n=\prod_{j=1}^{n}(1-\gamma_j),$$

$$s_n=\sum_{k=1}^{n}\gamma_k,$$

$$Z_n(x)=\frac{1}{h_n^d}K\left(\frac{x-X_n}{h_n}\right). \tag{14}$$

Let us first state the following technical lemma.

Lemma 2. *Let $(v_n) \in \mathcal{GS}(v^*)$, $(\gamma_n) \in \mathcal{GS}(-\alpha)$, and $m > 0$ such that $m - v^*\xi > 0$ where ξ is defined in* (5). *We have*

$$\lim_{n\to+\infty} v_n \Pi_n^m \sum_{k=1}^n \Pi_k^{-m} \frac{\gamma_k}{v_k} = \frac{1}{m - v^*\xi}. \tag{15}$$

Moreover, for all positive sequence (α_n) such that $\lim_{n\to+\infty} \alpha_n = 0$, *and all $\delta \in \mathbb{R}$,*

$$\lim_{n\to+\infty} v_n \Pi_n^m \left[\sum_{k=1}^n \Pi_k^{-m} \frac{\gamma_k}{v_k} \alpha_k + \delta\right] = 0. \tag{16}$$

Lemma 2 is widely applied throughout the proofs. Let us underline that it is its application, which requires Assumption (A2)(iii) on the limit of $(n\gamma_n)$ as n goes to infinity. Let us mention that, in particular, to prove (8), Lemma 2 is applied with $m = 2$ and $(v_n) = (\gamma_n^{-1} h_n^d)$ (and thus $v^* = \alpha - ad$); the stepsize (γ_n) must thus fulfil the condition $\lim_{n\to\infty} (n\gamma_n) > (\alpha - ad)/2$. Now, since $\lim_{n\to\infty} (n\gamma_n) < \infty$ only if $\alpha = 1$, the condition $\lim_{n\to\infty} (n\gamma_n) \in]\min\{2a, (1 - ad)/2\}, \infty]$ in (A2)(iii) is equivalent to the condition $\lim_{n\to\infty} (n\gamma_n) \in]\min\{2a, (\alpha - ad)/2\}, \infty]$, which appears throughout our proofs. Similarly, since $\xi \neq 0$ only if $\alpha = 1$, the limit $[2 - (\alpha - ad)\xi]^{-1}$ given by the application of Lemma 2 for such m and (v_n) equals the factor $[2 - (1 - ad)\xi]^{-1}$ that stands in the statement of our main results.

Our proofs are now organized as follows. Lemmas 1 and 2 are proved in Section 4.1, Propositions 1 and 2 in Sections 4.2 and 4.3, respectively, Theorems 1 and 2 in Sections 4.4 and 4.5, respectively, and Corollaries 1–4 in Section 4.6.

4.1. Proof of Lemmas 1 and 2

We first prove Lemma 1. Since $(w_n) \in \mathcal{GS}(w^*)$ with $w^* > -1$, we have

$$\lim_{n\to\infty} \frac{n w_n}{\sum_{k=1}^n w_k} = 1 + w^*, \tag{17}$$

which guarantees that $\lim_{n\to\infty} n\gamma_n = 1 + w^*$. Moreover, applying (17), we note that

$$\frac{\sum_{k=1}^{n-1} w_k}{\sum_{k=1}^n w_k} = 1 - \frac{w_n}{\sum_{k=1}^n w_k} = 1 - \frac{1 + w^*}{n} + o\left(\frac{1}{n}\right),$$

so that

$$\lim_{n\to\infty} n\left[1 - \frac{\sum_{k=1}^{n-1} w_k}{\sum_{k=1}^n w_k}\right] = 1 + w^*.$$

It follows that $(\sum_{k=1}^n w_k) \in \mathcal{GS}(1 + w^*)$, and thus that $(\gamma_n) \in \mathcal{GS}(-1)$, which concludes the proof of Lemma 1.

To prove Lemma 2, we first establish (16). Set

$$Q_n = v_n \Pi_n^m \left[\sum_{k=1}^n \Pi_k^{-m} \gamma_k v_k^{-1} \alpha_k + \delta\right].$$

We have

$$Q_n = \frac{v_n}{v_{n-1}} (1 - \gamma_n)^m Q_{n-1} + \gamma_n \alpha_n$$

with, since $(v_n) \in \mathcal{GS}(v^*)$ and in view of (5),

$$\begin{aligned} \frac{v_n}{v_{n-1}} (1 - \gamma_n)^m &= \left(1 + \frac{v^*}{n} + o\left(\frac{1}{n}\right)\right)(1 - m\gamma_n + o(\gamma_n)) \\ &= (1 + v^*\xi\gamma_n + o(\gamma_n))(1 - m\gamma_n + o(\gamma_n)) \\ &= 1 - (m - v^*\xi)\gamma_n + o(\gamma_n). \end{aligned} \tag{18}$$

Set $A \in]0, m - v^*\xi[$; for n large enough, we obtain

$$Q_n \leqslant (1 - A\gamma_n) Q_{n-1} + \gamma_n \alpha_n$$

and (16) follows straightforwardly from the application of Lemma 4.I.1 in Duflo (1996) . Now, let C denote a positive generic constant that may vary from line to line; we have

$$v_n \Pi_n^m \sum_{k=1}^n \Pi_k^{-m} \gamma_k v_k^{-1} - (m - v^* \xi)^{-1} = v_n \Pi_n^m \left[\sum_{k=1}^n \Pi_k^{-m} \gamma_k v_k^{-1} - (m - v^* \xi)^{-1} P_n \right]$$

with, in view of (18),

$$\begin{aligned} P_n &= v_n^{-1} \Pi_n^{-m} \\ &= \sum_{k=2}^n (v_k^{-1} \Pi_k^{-m} - v_{k-1}^{-1} \Pi_{k-1}^{-m}) + C \\ &= \sum_{k=2}^n v_k^{-1} \Pi_k^{-m} \left[1 - \frac{v_k}{v_{k-1}} (1 - \gamma_k)^m \right] + C \\ &= \sum_{k=2}^n v_k^{-1} \Pi_k^{-m} [(m - v^* \xi) \gamma_k + \mathrm{o}(\gamma_k)] + C. \end{aligned}$$

It follows that

$$v_n \Pi_n^m \sum_{k=1}^n \Pi_k^{-m} \gamma_k v_k^{-1} - (m - v^* \xi)^{-1} = v_n \Pi_n^m \left[\sum_{k=1}^n \Pi_k^{-m} v_k^{-1} \mathrm{o}(\gamma_k) + C \right],$$

and (15) follows from the application of (16), which concludes the proof of Lemma 2.

4.2. Proof of Proposition 1

In view of (1) and (14), we have

$$\begin{aligned} f_n(x) - f(x) &= (1 - \gamma_n)(f_{n-1}(x) - f(x)) + \gamma_n (Z_n(x) - f(x)) \\ &= \sum_{k=1}^{n-1} \left[\prod_{j=k+1}^n (1 - \gamma_j) \right] \gamma_k (Z_k(x) - f(x)) + \gamma_n (Z_n(x) - f(x)) + \left[\prod_{j=1}^n (1 - \gamma_j) \right] (f_0(x) - f(x)) \\ &= \Pi_n \sum_{k=1}^n \Pi_k^{-1} \gamma_k (Z_k(x) - f(x)) + \Pi_n (f_0(x) - f(x)). \end{aligned} \tag{19}$$

It follows that

$$\mathbb{E}(f_n(x)) - f(x) = \Pi_n \sum_{k=1}^n \Pi_k^{-1} \gamma_k (\mathbb{E}(Z_k(x)) - f(x)) + \Pi_n (f_0(x) - f(x)).$$

Taylor's expansion with integral remainder ensures that

$$\begin{aligned} \mathbb{E}[Z_k(x)] - f(x) &= \int_{\mathbb{R}^d} K(z) [f(x - z h_k) - f(x)] \, \mathrm{d}z \\ &= \frac{h_k^2}{2} \sum_{j=1}^d (\mu_j^2 f_{jj}^{(2)}(x)) + h_k^2 \delta_k(x) \end{aligned} \tag{20}$$

with

$$\delta_k(x) = \sum_{1 \leqslant i,j \leqslant d} \int_{\mathbb{R}^d} \int_0^1 (1 - s) z_i z_j K(z) [f_{ij}^{(2)}(x - z h_k s) - f_{ij}^{(2)}(x)] \, \mathrm{d}s \, \mathrm{d}z,$$

and, since $f_{ij}^{(2)}$ is bounded and continuous at x for all $i,j \in \{1,\dots,d\}$, we have $\lim_{k\to\infty}\delta_k(x)=0$. In the case $a \leqslant \alpha/(d+4)$, we have $\lim_{n\to\infty}(n\gamma_n) > 2a$; the application of Lemma 2 then gives

$$\begin{aligned}\mathbb{E}[f_n(x)]-f(x) &= \frac{1}{2}\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x))\Pi_n\sum_{k=1}^{n}\Pi_k^{-1}\gamma_k h_k^2[1+o(1)]+\Pi_n(f_0(x)-f(x))\\ &= \frac{1}{2(1-2a\xi)}\sum_{j=1}^{d}(\mu_j^2 f_{jj}^{(2)}(x))[h_n^2+o(1)],\end{aligned}$$

and (6) follows. In the case $a > \alpha/(d+4)$, we have $h_n^2 = o(\sqrt{\gamma_n h_n^{-d}})$; since $\lim_{n\to\infty}(n\gamma_n) > (\alpha-ad)/2$, Lemma 2 then ensures that

$$\begin{aligned}\mathbb{E}[f_n(x)]-f(x) &= \Pi_n\sum_{k=1}^{n}\Pi_k^{-1}\gamma_k o\left(\sqrt{\gamma_k h_k^{-d}}\right)+O(\Pi_n)\\ &= o\left(\sqrt{\gamma_n h_n^{-d}}\right),\end{aligned}$$

which gives (7). Now, we have

$$\begin{aligned}Var[f_n(x)] &= \Pi_n^2\sum_{k=1}^{n}\Pi_k^{-2}\gamma_k^2 Var[Z_k(x)]\\ &= \Pi_n^2\sum_{k=1}^{n}\frac{\Pi_k^{-2}\gamma_k^2}{h_k^d}\left[\int_{\mathbb{R}^d}K^2(z)f(x-zh_k)\,dz-h_k^d\left(\int_{\mathbb{R}^d}K(z)f(x-zh_k)\,dz\right)^2\right]\\ &= \Pi_n^2\sum_{k=1}^{n}\frac{\Pi_k^{-2}\gamma_k^2}{h_k^d}\left[f(x)\int_{\mathbb{R}^d}K^2(z)\,dz+v_k(x)-h_k^d\tilde{v}_k(x)\right]\end{aligned}$$

with

$$\begin{aligned}v_k(x) &= \int_{\mathbb{R}^d}K^2(z)[f(x-zh_k)-f(x)]\,dz,\\ \tilde{v}_k(x) &= \left(\int_{\mathbb{R}^d}K(z)f(x-zh_k)\,dz\right)^2.\end{aligned}$$

Since f is bounded and continuous, we have $\lim_{k\to\infty}v_k(x)=0$ and $\lim_{k\to\infty}h_k^d\tilde{v}_k(x)=0$. In the case $a \geqslant \alpha/(d+4)$, we have $\lim_{n\to\infty}(n\gamma_n) > (\alpha-ad)/2$, and the application of Lemma 2 gives

$$\begin{aligned}Var[f_n(x)] &= \Pi_n^2\sum_{k=1}^{n}\frac{\Pi_k^{-2}\gamma_k^2}{h_k^d}\left[f(x)\int_{\mathbb{R}^d}K^2(z)\,dz+o(1)\right]\\ &= \frac{1}{2-(\alpha-ad)\xi}\frac{\gamma_n}{h_n^d}\left[f(x)\int_{\mathbb{R}^d}K^2(z)\,dz+o(1)\right],\end{aligned}$$

which proves (8). In the case $a < \alpha/(d+4)$, we have $\gamma_n h_n^{-d} = o(h_n^4)$; since $\lim_{n\to\infty}(n\gamma_n) > 2a$, Lemma 2 then ensures that

$$\begin{aligned}Var[f_n(x)] &= \Pi_n^2\sum_{k=1}^{n}\Pi_k^{-2}\gamma_k o(h_k^4)\\ &= o(h_n^4),\end{aligned}$$

which gives (9).

4.3. Proof of Proposition 2

Let us first note that, in view of (20), we have

$$\begin{aligned}
&\int_{\mathbb{R}^d}\left\{\Pi_n\sum_{k=1}^n \Pi_k^{-1}\gamma_k[\mathbb{E}(Z_k(x))-f(x)]\right\}^2 dx\\
&=\frac{1}{4}\int_{\mathbb{R}^d}\left[\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x)\right]^2 dx\left[\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\right]^2+\int_{\mathbb{R}^d}\left[\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\delta_k(x)\right]^2 dx\\
&\quad+\left(\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\right)\left(\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\int_{\mathbb{R}^d}\left[\sum_{j=1}^d\mu_j^2 f_{jj}^{(2)}(x)\right]\delta_k(x)\,dx\right).
\end{aligned}$$

Since $f_{ij}^{(2)}$ is continuous, bounded, and integrable for all $i,j\in\{1,\dots d\}$, the application of Lebesgue's convergence theorem ensures that $\lim_{k\to+\infty}\int_{\mathbb{R}^d}\delta_k^2(x)\,dx=0$ and $\lim_{k\to+\infty}\int_{\mathbb{R}^d}[\sum_{j=1}^d\mu_j^2 f_{jj}^{(2)}(x)]\delta_k(x)\,dx=0$. Moreover, Jensen's inequality gives

$$\begin{aligned}
\int_{\mathbb{R}^d}\left[\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\delta_k(x)\right]^2 dx&\leqslant\left(\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\right)\left(\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\int_{\mathbb{R}^d}\delta_k^2(x)\,dx\right)\\
&\leqslant\left(\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\right)\left(\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k o(h_k^2)\right),
\end{aligned}$$

so that we get

$$\begin{aligned}
&\int_{\mathbb{R}^d}\left\{\Pi_n\sum_{k=1}^n \Pi_k^{-1}\gamma_k[\mathbb{E}(Z_k(x))-f(x)]\right\}^2 dx\\
&=\frac{1}{4}\int_{\mathbb{R}^d}\left[\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x)\right]^2 dx\left[\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\right]^2+O\left(\left[\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k h_k^2\right]\left[\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k o(h_k^2)\right]\right).
\end{aligned}$$

- Let us first consider the case $a\leqslant\alpha/(d+4)$. In this case, $\lim_{n\to\infty}(n\gamma_n)>2a$, and the application of Lemma 2 gives

$$\int_{\mathbb{R}^d}\left\{\Pi_n\sum_{k=1}^n \Pi_k^{-1}\gamma_k[\mathbb{E}(Z_k(x))-f(x)]\right\}^2 dx=\frac{1}{4(1-2a\xi)^2}h_n^4\int_{\mathbb{R}^d}\left[\sum_{j=1}^d\mu_j^2 f_{jj}^{(2)}(x)\right]^2 dx+o(h_n^4),$$

  and ensures that $\Pi_n^2=o(h_n^4)$. In view of (19), we then deduce that

$$\int_{\mathbb{R}^d}\{\mathbb{E}(f_n(x))-f(x)\}^2\,dx=\frac{1}{4(1-2a\xi)^2}h_n^4\int_{\mathbb{R}^d}\left[\sum_{j=1}^d\mu_j^2 f_{jj}^{(2)}(x)\right]^2 dx+o(h_n^4). \tag{21}$$

- Let us now consider the case $a>\alpha/(d+4)$. In this case, we have $h_k^2=o(\sqrt{\gamma_k h_k^{-d}})$ and $\lim_{n\to\infty}(n\gamma_n)>(\alpha-ad)/2$. The application of Lemma 2 then gives

$$\begin{aligned}
\int_{\mathbb{R}^d}\left\{\Pi_n\sum_{k=1}^n \Pi_k^{-1}\gamma_k[\mathbb{E}(Z_k(x))-f(x)]\right\}^2 dx&=O\left(\left[\Pi_n\sum_{k=1}^n\Pi_k^{-1}\gamma_k o(\sqrt{\gamma_k h_k^{-d}})\right]^2\right)\\
&=o(\gamma_n h_n^{-d}),
\end{aligned}$$

  and ensures that $\Pi_n^2=o(\gamma_n h_n^{-d})$. In view of (19), we then deduce that

$$\int_{\mathbb{R}^d}\{\mathbb{E}(f_n(x))-f(x)\}^2\,dx=o(\gamma_n h_n^{-d}). \tag{22}$$

On the other hand, we note that

$$\int_{\mathbb{R}^d} Var[f_n(x)]\,dx = \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2}\gamma_k^2 \int_{\mathbb{R}^d} Var[Z_k(x)]\,dx$$
$$= \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2}\gamma_k^2 \left[\frac{1}{h_k^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K^2(z)f(x-zh_k)\,dz\,dx - \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} K(z)f(x-zh_k)\,dz\right)^2 dx\right]$$

with

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K^2(z)f(x-zh_k)\,dz\,dx = \int_{\mathbb{R}^d} K^2(z)\left(\int_{\mathbb{R}^d} f(x-zh_k)\,dx\right) dz$$
$$= \int_{\mathbb{R}^d} K^2(z)\,dz$$

and

$$\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} K(z)f(x-zh_k)\,dz\right)^2 dx = \int_{\mathbb{R}^{3d}} K(z)K(z')f(x-zh_k)f(x-z'h_k)\,dz\,dz'\,dx$$
$$\leqslant \|f\|_\infty \|K\|_1^2.$$

- In the case $a \geqslant \alpha/(d+4)$, we have $\lim_{n\to\infty}(n\gamma_n) > (\alpha - ad)/2$, and Lemma 2 ensures that

$$\int_{\mathbb{R}^d} Var[f_n(x)]\,dx = \Pi_n^2 \sum_{k=1}^n \frac{\Pi_k^{-2}\gamma_k^2}{h_k^d}\left[\int_{\mathbb{R}^d} K^2(z)\,dz + o(1)\right]$$
$$= \frac{\gamma_n}{h_n^d}\frac{1}{(2-(\alpha-ad)\xi)}\int_{\mathbb{R}^d} K^2(z)\,dz + o\left(\frac{\gamma_n}{h_n^d}\right). \tag{23}$$

- In the case $a < \alpha/(d+4)$, we have $\gamma_n h_n^{-d} = o(h_n^4)$ and $\lim_{n\to\infty}(n\gamma_n) > 2a$, so that Lemma 2 gives

$$\int_{\mathbb{R}^d} Var[f_n(x)]\,dx = \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2}\gamma_k o(h_k^4)$$
$$= o(h_n^4). \tag{24}$$

Part 1 of Proposition 2 follows from the combination of (21) and (24), Part 2 from that of (21) and (23), and Part 3 from that of (22) and (23).

4.4. Proof of Theorem 1

Let us at first assume that, if $a \geqslant \alpha/(d+4)$, then

$$\sqrt{\gamma_n^{-1}h_n^d}(f_n(x) - \mathbb{E}[f_n(x)]) \stackrel{\mathscr{D}}{\to} \mathscr{N}\left(0, \frac{1}{2-(\alpha-ad)\xi} f(x)\int_{\mathbb{R}^d} K^2(z)\,dz\right). \tag{25}$$

In the case when $a > \alpha/(d+4)$, Part 1 of Theorem 1 follows from the combination of (7) and (25). In the case when $a = \alpha/(d+4)$, Parts 1 and 2 of Theorem 1 follow from the combination of (6) and (25). In the case $a < \alpha/(d+4)$, (9) implies that

$$h_n^{-2}(f_n(x) - \mathbb{E}(f_n(x))) \stackrel{\mathbb{P}}{\to} 0,$$

and the application of (6) gives Part 2 of Theorem 1.

We now prove (25). In view of (1), we have

$$f_n(x) - \mathbb{E}[f_n(x)] = (1-\gamma_n)(f_{n-1}(x) - \mathbb{E}[f_{n-1}(x)]) + \gamma_n(Z_n(x) - \mathbb{E}[Z_n(x)])$$
$$= \Pi_n \sum_{k=1}^n \Pi_k^{-1}\gamma_k(Z_k(x) - \mathbb{E}[Z_k(x)]).$$

Set

$$Y_k(x) = \Pi_k^{-1}\gamma_k(Z_k(x) - \mathbb{E}(Z_k(x))). \tag{26}$$

The application of Lemma 2 ensures that

$$\begin{aligned}
v_n^2 &= \sum_{k=1}^n Var(Y_k(x)) \\
&= \sum_{k=1}^n \Pi_k^{-2}\gamma_k^2 Var(Z_k(x)) \\
&= \sum_{k=1}^n \frac{\Pi_k^{-2}\gamma_k^2}{h_k^d}\left[f(x)\int_{\mathbb{R}^d} K^2(z)\,dz + o(1)\right] \\
&= \frac{1}{\Pi_n^2}\frac{\gamma_n}{h_n^d}\left[\frac{1}{2-(\alpha-ad)\xi} f(x)\int_{\mathbb{R}^d} K^2(z)\,dz + o(1)\right].
\end{aligned} \tag{27}$$

On the other hand, we have, for all $p>0$,

$$\mathbb{E}[|Z_k(x)|^{2+p}] = O\left(\frac{1}{h_k^{d(1+p)}}\right), \tag{28}$$

and, since $\lim_{n\to\infty}(n\gamma_n) > (\alpha-ad)/2$, there exists $p>0$ such that $\lim_{n\to\infty}(n\gamma_n) > ((1+p)/(2+p))(\alpha-ad)$. Applying Lemma 2, we get

$$\begin{aligned}
\sum_{k=1}^n \mathbb{E}[|Y_k(x)|^{2+p}] &= O\left(\sum_{k=1}^n \Pi_k^{-2-p}\gamma_k^{2+p}\mathbb{E}[|Z_k(x)|^{2+p}]\right) \\
&= O\left(\sum_{k=1}^n \frac{\Pi_k^{-2-p}\gamma_k^{2+p}}{h_k^{d(1+p)}}\right) \\
&= O\left(\frac{\gamma_n^{1+p}}{\Pi_n^{2+p}h_n^{d(1+p)}}\right),
\end{aligned}$$

and we thus obtain

$$\frac{1}{v_n^{2+p}}\sum_{k=1}^n \mathbb{E}[|Y_k(x)|^{2+p}] = O([\gamma_n h_n^{-d}]^{p/2}) = o(1).$$

The convergence in (25) then follows from the application of Lyapounov's Theorem.

4.5. Proof of Theorem 2

Set

$$S_n(x) = \sum_{k=1}^n Y_k(x),$$

where Y_k is defined in (26), and set $\gamma_0 = h_0 = 1$.

- Let us first consider the case $a \geqslant \alpha/(d+4)$ (in which case $\lim_{n\to\infty}(n\gamma_n) > (\alpha-ad)/2$). We set $H_n^2 = \Pi_n^2\gamma_n^{-1}h_n^d$, and note that, since $(\gamma_n^{-1}h_n^d) \in \mathcal{GS}(\alpha-ad)$, we have

$$\begin{aligned}
\ln(H_n^{-2}) &= -2\ln(\Pi_n) + \ln\left(\prod_{k=1}^n \frac{\gamma_{k-1}^{-1}h_{k-1}^d}{\gamma_k^{-1}h_k^d}\right) \\
&= -2\sum_{k=1}^n \ln(1-\gamma_k) + \sum_{k=1}^n \ln\left(1 - \frac{\alpha-ad}{k} + o\left(\frac{1}{k}\right)\right) \\
&= \sum_{k=1}^n (2\gamma_k + o(\gamma_k)) - \sum_{k=1}^n ((\alpha-ad)\xi\gamma_k + o(\gamma_k)) \\
&= (2-\xi(\alpha-ad))s_n + o(s_n).
\end{aligned} \tag{29}$$

Since $2-\xi(\alpha-ad)>0$, it follows in particular that $\lim_{n\to+\infty} H_n^{-2}=\infty$. Moreover, we clearly have $\lim_{n\to+\infty} H_n^2/H_{n-1}^2=1$, and by (27)

$$\lim_{n\to+\infty} H_n^2\sum_{k=1}^n Var[Y_k(x)] = \frac{1}{2-(\alpha-ad)\xi} f(x)\int_{\mathbb{R}^d} K^2(z)\,dz.$$

Now, in view of (28), $\mathbb{E}[|Y_k(x)|^3] = O(\Pi_k^{-3}\gamma_k^3 h_k^{-2d})$ and, since $\lim_{n\to\infty}(n\gamma_n)>(\alpha-ad)/2$, the application of Lemma 2 and of (29) gives

$$\begin{aligned}\frac{1}{n\sqrt{n}}\sum_{k=1}^n \mathbb{E}(|H_nY_k(x)|^3) &= O\left(\frac{H_n^3}{n\sqrt{n}}\sum_{k=1}^n \Pi_k^{-3}\gamma_k^3 h_k^{-2d}\right)\\ &= O\left(\frac{H_n^3}{n\sqrt{n}}\sum_{k=1}^n \Pi_k^{-3}\gamma_k o([\gamma_k h_k^{-d}]^{3/2})\right)\\ &= o\left(\frac{H_n^3}{n\sqrt{n}}\Pi_n^{-3}[\gamma_n h_n^{-d}]^{3/2}\right)\\ &= o\left(\frac{1}{n\sqrt{n}}\right)\\ &= o([\ln(H_n^{-2})]^{-1}).\end{aligned}$$

The application of Theorem 1 in Mokkadem and Pelletier (2007b) then ensures that, with probability one, the sequence

$$\left(\frac{H_nS_n(x)}{\sqrt{2\ln\ln(H_n^{-2})}}\right) = \left(\frac{\sqrt{\gamma_n^{-1}h_n^d}(f_n(x)-\mathbb{E}(f_n(x)))}{\sqrt{2\ln\ln(H_n^{-2})}}\right)$$

is relatively compact and its limit set is the interval

$$\left[-\sqrt{\frac{f(x)}{2-(\alpha-ad)\xi}\int_{\mathbb{R}^d}K^2(z)\,dz}, \sqrt{\frac{f(x)}{2-(\alpha-ad)\xi}\int_{\mathbb{R}^d}K^2(z)\,dz}\right]. \tag{30}$$

In view of (29), we have $\lim_{n\to\infty} \ln\ln(H_n^{-2})/\ln s_n = 1$. It follows that, with probability one, the sequence

$$(\sqrt{\gamma_n^{-1}h_n^d}(f_n(x)-\mathbb{E}(f_n(x)))/\sqrt{2\ln s_n})$$

is relatively compact, and its limit set is the interval given in (30). The application of (6) (respectively (7)) concludes the proof of Theorem 2 in the case $a=\alpha/(d+4)$ (respectively, $a>\alpha/(d+4)$).

- Let us now consider the case $a<\alpha/(d+4)$ (in which case $\lim_{n\to\infty}(n\gamma_n)>2a$). Set $H_n^{-2}=\Pi_n^{-2}h_n^4(\ln\ln(\Pi_n^{-2}h_n^4))^{-1}$, and note that, since $(h_n^{-4})\in\mathcal{GS}(4a)$, we have

$$\begin{aligned}\ln(\Pi_n^{-2}h_n^4) &= -2\ln(\Pi_n)+\ln\left(\prod_{k=1}^n\frac{h_{k-1}^{-4}}{h_k^{-4}}\right)\\ &= -2\sum_{k=1}^n\ln(1-\gamma_k)+\sum_{k=1}^n\ln\left(1-\frac{4a}{k}+o\left(\frac{1}{k}\right)\right)\\ &= \sum_{k=1}^n(2\gamma_k+o(\gamma_k))-\sum_{k=1}^n(4a\xi\gamma_k+o(\gamma_k))\\ &= (2-4a\xi)s_n+o(s_n).\end{aligned} \tag{31}$$

Since $2-4a\xi>0$, it follows in particular that $\lim_{n\to\infty}\Pi_n^{-2}h_n^4=\infty$, and thus $\lim_{n\to\infty}H_n^{-2}=\infty$. Moreover, we clearly have $\lim_{n\to\infty}H_n^2/H_{n-1}^2=1$. Set $\varepsilon\in]0,\alpha-(d+4)a[$ such that $\lim_{n\to\infty}(n\gamma_n)>2a+\varepsilon/2$; in view of (27), and applying Lemma 2, we get

$$\begin{aligned}H_n^2\sum_{k=1}^n Var[Y_k(x)] &= O\left(\Pi_n^2 h_n^{-4}\ln\ln(\Pi_n^{-2}h_n^4)\sum_{k=1}^n \frac{\Pi_k^{-2}\gamma_k^2}{h_k^d}\right)\\ &= O\left(\Pi_n^2 h_n^{-4}\ln\ln(\Pi_n^{-2}h_n^4)\sum_{k=1}^n \Pi_k^{-2}\gamma_k o(h_k^4 k^{-\varepsilon})\right)\\ &= o(\ln\ln(\Pi_n^{-2}h_n^4)n^{-\varepsilon})\\ &= o(1).\end{aligned}$$

Moreover, applying (28), Lemma 2, and (31), we obtain

$$\begin{aligned}\frac{1}{n\sqrt{n}}\sum_{k=1}^n \mathbb{E}(|H_nY_k(x)|^3) &= O\left(\frac{\Pi_n^3 h_n^{-6}}{n\sqrt{n}}[\ln\ln(\Pi_n^{-2}h_n^4)]^{3/2}\left(\sum_{k=1}^n \Pi_k^{-3}\gamma_k^3 h_k^{-2d}\right)\right)\\ &= O\left(\frac{\Pi_n^3 h_n^{-6}}{n\sqrt{n}}[\ln\ln(\Pi_n^{-2}h_n^4)]^{3/2}\left(\sum_{k=1}^n \Pi_k^{-3}\gamma_k o(h_k^6)\right)\right)\\ &= o\left(\frac{\Pi_n^3 h_n^{-6}}{n\sqrt{n}}\Pi_n^{-3}h_n^6[\ln\ln(\Pi_n^{-2}h_n^4)]^{3/2}\right)\\ &= o([\ln(H_n^{-2})]^{-1}).\end{aligned}$$

The application of Theorem 1 in Mokkadem and Pelletier (2007b) then ensures that, with probability one,

$$\lim_{n\to\infty}\frac{H_nS_n(x)}{\sqrt{2\ln\ln(H_n^{-2})}} = \lim_{n\to\infty} h_n^{-2}\frac{\sqrt{\ln\ln(\Pi_n^{-2}h_n^4)}}{\sqrt{2\ln\ln(H_n^{-2})}}(f_n(x)-\mathbb{E}(f_n(x))) = 0.$$

Noting that (31) ensures that $\lim_{n\to\infty}\ln\ln(H_n^{-2})/\ln\ln(\Pi_n^{-2}h_n^4)=1$, we deduce that

$$\lim_{n\to\infty} h_n^{-2}[T_n(x)-\mathbb{E}(T_n(x))] = 0 \quad \text{a.s.},$$

and Theorem 2 in the case $a<\alpha/(d+4)$ follows from (6).

4.6. *Proof of Corollaries 1–4*

In view of (8), to minimize the variance of f_n, the stepsize (γ_n) must belong to $\mathcal{GS}(-1)$ and satisfy $\lim_{n\to\infty} n\gamma_n=\gamma_0\in]0,\infty[$. For such a choice, $\xi=\gamma_0^{-1}$, so that (8) can be rewritten as

$$Var(f_n(x)) = \frac{\gamma_0}{2-(1-ad)\gamma_0^{-1}}\frac{1}{nh_n^d}f(x)\int_{\mathbb{R}^d}K^2(z)\,dz + o\left(\frac{1}{nh_n^d}\right).$$

The function $\gamma_0\mapsto\gamma_0[2-(1-ad)\gamma_0^{-1}]^{-1}$ reaching its minimum at the point $\gamma_0=1-ad$, Corollary 1 follows.

Let us now prove Corollary 4. When $\lim_{n\to\infty} n\gamma_n=\gamma_0>0$ and $\lim_{n\to\infty} nh_n^d=0$, the first part of Theorem 1 ensures that

$$\sqrt{nh_n^d}(f_n(x)-f(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\frac{\gamma_0^2}{2\gamma_0-(1-ad)}f(x)\int_{\mathbb{R}^d}K^2(z)\,dz\right).$$

Proposition 1 ensuring the consistency of f_n, Corollary 4 follows.

We now show how Corollary 2 can be deduced from Proposition 1. Corollary 3 is deduced from Proposition 2 exactly in the same way, so that its proof is omitted. Set

$$C_1(\xi) = \frac{1}{4(1-2a\xi)^2}\left(\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x)\right)^2,$$

$$C_2(\xi) = \frac{1}{2-(1-ad)\xi}f(x)\int_{\mathbb{R}^d}K^2(z)\,dz.$$

The application of Proposition 1 ensures that

$$MSE = \begin{cases} C_1(\xi)h_n^4 + o(h_n^4) & \text{if } a < \alpha/(d+4), \\ C_1(\xi)h_n^4 + C_2(\xi)\gamma_n h_n^{-d} + o(h_n^4 + \gamma_n h_n^{-d}) & \text{if } a = \alpha/(d+4), \\ C_2(\xi)\gamma_n h_n^{-d} + o(\gamma_n h_n^{-d}) & \text{if } a > \alpha/(d+4). \end{cases} \tag{32}$$

Set $\alpha \in]1/2, 1]$. If $a = \alpha/(d+4)$, $(C_1(\xi)h_n^4 + C_2(\xi)\gamma_n h_n^{-d}) \in \mathcal{GS}(-4\alpha/(d+4))$. If $a < \alpha/(d+4)$, $(h_n^4) \in \mathcal{GS}(-4a)$ with $-4a > -4\alpha/(d+4)$, and, if $a > \alpha/(d+4)$, $(\gamma_n h_n^{-d}) \in \mathcal{GS}(-\alpha + ad)$ with $-\alpha + ad > -4\alpha/(d+4)$. It follows that, for a given α, to minimize the MSE of f_n, the parameter a must be chosen equal to $\alpha/(d+4)$. Moreover, in view of (32), the parameter α must be chosen equal to 1. In other words, to minimize the MSE of f_n, the stepsize (γ_n) must be chosen in $\mathcal{GS}(-1)$, the bandwidth (h_n) in $\mathcal{GS}(-1/(d+4))$ (and, in view of (A2)(iii), the condition $\lim_{n\to\infty} n\gamma_n > 2/(d+4)$ must be fulfilled). For this choice of stepsize and bandwidth, set $\mathcal{L}_n = n\gamma_n$ and $\tilde{\mathcal{L}}_n = n^{1/(d+4)}h_n$. The MSE of f_n can then be rewritten as

$$MSE = n^{-4/(d+4)}[C_1(\xi)\tilde{\mathcal{L}}_n^4 + C_2(\xi)\mathcal{L}_n\tilde{\mathcal{L}}_n^{-d}][1 + o(1)].$$

Now, set $\mathcal{L}_n$. Since the function

$$x \mapsto C_1(\xi)x^4 + C_2(\xi)\mathcal{L}_n x^{-d}$$

reaches its minimum at the point $(dC_2(\xi)\mathcal{L}_n/[4C_1(\xi)])^{1/(d+4)}$, to minimize the MSE of f_n, $\tilde{\mathcal{L}}_n$ must be chosen equal to $(dC_2(\xi)\mathcal{L}_n/[4C_1(\xi)])^{1/(d+4)}$, that is, (h_n) must equal $(dC_2(\xi)/[4C_1(\xi)]\gamma_n)^{1/(d+4)}$. For such a choice, the MSE of f_n can be rewritten as

$$MSE = n^{-4/(d+4)}\mathcal{L}_n^{4/(d+4)}\left(\frac{d}{4}\right)^{-d/(d+4)}\frac{\mathrm{d}+4}{4}[C_1(\xi)]^{d/(d+4)}[C_2(\xi)]^{4/(d+4)}[1 + o(1)].$$

It follows that to minimize the MSE of f_n, the limit of $\mathcal{L}_n$ (that is, of $n\gamma_n$) must be finite (and larger than $2/(d+4)$). Now, set $\gamma_0 > 2/(d+4)$ and $\mathcal{L}_n = \gamma_0\delta_n$ with $\lim_{n\to\infty}\delta_n = 1$ (so that $\lim_{n\to\infty} n\gamma_n = \gamma_0$). In this case, we have $\xi = \gamma_0^{-1}$,

$$C_1(\xi) = \frac{\gamma_0^2}{4\left(\gamma_0 - \dfrac{2}{d+4}\right)^2}c_1, \quad c_1 = \left(\sum_{j=1}^d \mu_j^2 f_{jj}^{(2)}(x)\right)^2,$$

$$C_2(\xi) = \frac{\gamma_0}{2\left(\gamma_0 - \dfrac{2}{d+4}\right)}c_2, \quad c_2 = f(x)\int_{\mathbb{R}^d} K^2(z)\,\mathrm{d}z,$$

‘

and the MSE of f_n can be rewritten as

$$MSE = n^{-4/(d+4)}\delta_n^{4/(d+4)}\frac{d+4}{d^{d/(d+4)}4^{(d+6)/(d+4)}}\frac{\gamma_0^2}{\left(\gamma_0 - \dfrac{2}{d+4}\right)^{(2d+4)/(d+4)}}c_1^{d/(d+4)}c_2^{4/(d+4)}[1 + o(1)].$$

The function $x \mapsto x^2/(x - 2/(d+4))^{(2d+4)/(d+4)}$ reaching its minimum at the point $x = 1$, to minimize the MSE of f_n, γ_0 must be chosen equal to 1. Corollary 2 follows.

Acknowledgements

We are grateful to two referees and an Associate Editor for their helpful comments, which have led to this substantially improved version of the paper.

References

Arcones, M.A., 1997. The law of the iterated logarithm for a triangular array of empirical processes. Electron. J. Probab. 2, 1–39.
Blum, J.R., 1954. Multidimensional stochastic approximation methods. Ann. Math. Statist. 25, 737–744.
Bojanic, R., Seneta, E., 1973. A unified theory of regularly varying sequences. Math. Z. 134, 91–106.
Chen, H., 1988. Lower rate of convergence for locating a maximum of a function. Ann. Statist. 16, 1330–1334.
Chen, H.F., Duncan, T.E., Pasik-Duncan, B., 1999. A Kiefer–Wolfowitz algorithm with randomized differences. IEEE Trans. Automat. Control 44, 442–453.
Davies, H.L., 1973. Strong consistency of a sequential estimator of a probability density function. Bull. Math. Statist. 15, 49–54.
Deheuvels, P., 1973. Sur l'estimation séquentielle de la densité. C.R. Acad. Sci. Paris Ser. A-B 276, 1119–1121.

Devroye, L., 1979. On the pointwise and integral convergence of recursive kernel estimates of probability densities. Utilitas Math. 15, 113–128.
Dippon, J., 2003. Accelerated randomized stochastic optimization. Ann. Statist. 31, 1260–1281.
Dippon, J., Renz, J., 1997. Weighted means in stochastic approximation of minima. SIAM J. Control Optim. 35, 1811–1827.
Duflo, M., 1996. Algorithmes stochastiques. Collection Applications of Mathematics. Springer, Berlin.
Duflo, M., 1997. Random iterative models. Collection Applications of Mathematics. Springer, Berlin.
Fabian, V., 1967. Stochastic approximation of minima with improved asymptotic speed. Ann. Math. Statist. 38, 191–200.
Galambos, J., Seneta, E., 1973. Regularly varying sequences. Proc. Amer. Math. Soc. 41, 110–116.
Hall, P., 1981. Laws of the iterated logarithm for nonparametric density estimators. Z. Warsch. Verw. Gebiete 56, 47–61.
Hall, P., 1992. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann. Statist. 20, 675–694.
Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and its Application. Academic Press, Inc., New York, London.
Hall, P., Patil, P., 1994. On the efficiency of on-line density estimators. IEEE Trans. Inform. Theory 40, 1504–1512.
Kiefer, J., Wolfowitz, J., 1952. Stochastic approximation of the maximum of a regression functions. Ann. Math. Statist. 23, 462–466.
Kushner, H.J., Clark, D.S., 1978. Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer, New York.
Mokkadem, A., Pelletier, M., 2007a. A companion for the Kiefer–Wolfowitz–Blum stochastic approximation algorithm. Ann. Statist. 35 (4), 1749–1772.
Mokkadem, A., Pelletier, M., 2007b. Compact law of the iterated logarithm for matrix-normalized sums of random vectors. Teor. Veroyatn. Primen. 52.
Parzen, E., 1962. On estimation of a probability density and mode. Ann. Math. Statist. 33, 1065–1076.
Polyak, B.T., Tsybakov, A.B., 1990. Optimal orders of accuracy for search algorithms of stochastic optimization. Problems Inform. Transmission 26, 126–133.
Révész, P., 1973. Robbins–Monro procedure in a Hilbert space and its application in the theory of learning processes I. Studia Sci. Math. Hung. 8, 391–398.
Révész, P., 1977. How to apply the method of stochastic approximation in the non-parametric estimation of a regression function. Math. Operationsforsch. Statist. Ser. Statist. 8, 119–126.
Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832–837.
Roussas, G., 1992. Exact rates of almost sure convergence of a recursive kernel estimate of a probability density function: application to regression and hazard rate estimate. J. Nonparam. Statist. 3, 171–195.
Ruppert, D., 1982. Almost sure approximations to the Robbins–Monro and Kiefer–Wolfowitz processes with dependent noise. Ann. Probab. 10, 178–187.
Spall, J.C., 1988. A stochastic approximation algorithm for large-dimensional systems in the Kiefer–Wolfowitz setting. In: Proceedings of the Conference on Decision and Control. IEEE, New York, pp. 1544–1548.
Spall, J.C., 1997. A one-measurement form of simultaneous perturbation stochastic approximation. Automat. J. IFAC 33, 109–112.
Tsybakov, A.B., 1990. Recurrent estimation of the mode of a multidimensional distribution. Problems Inform. Transmission 26, 31–37.
Wegman, E.J., Davies, H.I., 1979. Remarks on some recursive estimators of a probability density. Ann. Statist. 7, 316–327.
Wolverton, C.T., Wagner, T.J., 1969. Asymptotically optimal discriminant functions for pattern classification. IEEE Trans. Inform. Theory 15, 258–265.
Yamato, H., 1971. Sequential estimation of a continuous probability density function and mode. Bull. Math. Statist. 14, 1–12.