J. Austral. Math. Soc. (*Series A*) 67 (1999), 1–14

UNIFORM CONSISTENCY OF THE PARTITIONING ESTIMATE UNDER ERGODIC CONDITIONS

NAÂMANE LAIB

(Received 18 March 1998; revised 23 December 1998)

Communicated by R. A. Maller

Abstract

We establish the uniform almost sure convergence of the partitioning estimate, which is a histogram-like mean regression function estimate, under ergodic conditions for a stationary and unbounded process. The main application of our results concerns time series analysis and prediction in the Markov processes case.

1991 *Mathematics subject classification* (*Amer. Math. Soc.*): primary 62G05; secondary 60F15, 62M10.
Keywords and phrases: Autoregression, ergodic processes, martingale difference, partitioning estimate, prediction, regression, stationarity.

1. Introduction

Let $Z_i = (X_i, Y_i)$, $i = 1, 2, \dots$, be a strictly stationary sequence of $\mathbb{R}^d \times \mathbb{R}^p$-valued random vectors ($d, p \geq 1$). We denote by X_1 a random variable (rv) having the same distribution as the X_i's, $i \geq 1$, and similarly for Y_1 and the Y_i's. Let $\rho(\cdot)$ be an integrable Borel real-valued function defined on $\mathbb{R}^p$ and set $\Psi(x) = \mathrm{E}[\rho(Y_1) \mid X_1 = x]$, $x \in \mathbb{R}^d$, the conditional mean function of $\rho(Y_1)$ given $X_1 = x$.

The function $\Psi(\cdot)$, which includes the regression function $r(x) = \mathrm{E}(Y_1 \mid X_1 = x)$ as a special case, can be estimated by a partitioning estimate in the following way. To simplify the notations, we begin with the case where the X's are $\mathbb{R}$-valued. Let $\cdots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \cdots$ be a partition of the real line into equal intervals of length a_n. For definiteness, we specify each interval to be right closed and left open. Let K_n be a nonnegative integer, we denote $I_n^r = [x_n^r, x_n^{r+1}[= [(r-1)/K_n, r/K_n[$ with $r \in \{\dots -2, -1, 0, 1, 2, \dots\}$. For each $x \in \mathbb{R}$, let $I_n^r(x)$ be the interval containing x.

https://doi.org/10.1017/S1446788700000835 Published online by Cambridge University Press

The quantity $\Psi(x)$ is estimated by

$$(1.1)\qquad \Psi_n(x) = \begin{cases} \left[\sum_{i=1}^n \rho(Y_i) I\{X_i \in I_n^r(x)\}\right]\left[\sum_{i=1}^n I\{X_i \in I_n^r(x)\}\right]^{-1} \\ \qquad\qquad \text{if } \sum_{i=1}^n I\{X_i \in I_n^r(x)\} \neq 0, \\ 0 \qquad\qquad \text{otherwise.} \end{cases}$$

Here $I\{A\}$ denotes the indicator function of the set A.

The partitioning estimate is useful in time series analysis. It has been used among others by Diebolt and Laïb [10] to construct nonparametric tests for nonlinear autoregressive processes. It has been recently studied by Carbonez *et al.* [5] under random censoring. Results on weak and strong universal consistency in the independent and identically distributed case are obtained by Devroye and Györfi [9], Györfi [14], Gordon and Olshen [11, 12].

The consistency of partitioning and kernel estimates has been considered by many authors, under a variety of mixing conditions, such as strong mixing (α-mixing), β-mixing, uniform mixing (ϕ-mixing). See, for instance, Yakowitz [34], Györfi and Masry [17], Tran [30], Roussas [29], Laïb [22] and Bosq [3]. The monograph by Györfi *et al.* [18] gives a large coverage of the literature in nonparametric inference for dependent series.

Our aim in this paper is to prove the uniform almost sure convergence of the partitioning estimate, which is a histogram-like mean regression function, under very weak assumptions on the dependence structure of the vector of random variable $Z = (X, Y)$. These results allow us to construct nonparametric predictors when $\{X_i;\ i \in \mathbb{Z}\}$ is a real Markovian process of finite order.

The originality of our result is in the very weak dependence structure imposed on the observation process $\{X_i;\ i \in \mathbb{Z}\}$, which is only assumed ergodic. This type of dependence has not been investigated much before. The ergodicity condition is very general, it is less restrictive than any mixing condition (α-mixing, β-mixing, ρ-mixing, ... : see, for example, Ash and Gardner [2, p. 120] and Rosenblatt [28]). It is the minimal condition that one may expect when dealing with problems such as the strong law of large numbers or consistency of functional estimators. For example, the ARMA processes are not ϕ-mixing in general and some linear processes are not α-mixing (see, Pham and Tran [26]). The following example illustrates the theory. Let (ζ_i) be the first-order autoregressive process defined by

$$(1.2)\qquad \zeta_i = \frac{1}{2}\zeta_{i-1} + \epsilon_i, \quad i \in \mathbb{Z},$$

where the ϵ_i's are independent symmetric Bernoulli random variables taking values -1 and 1. This Markov process is ergodic. Its distribution is absolutely continuous with respect to Lebesgue measure and does not satisfy the strong α-mixing condition, as it follows from Whithers [33, Theorem 5.2], see also Andrews [1]. However, it is ergodic. To see this, it suffices (see, for instance, Breiman [4, p. 119]) to rewrite (1.2) as a general linear process:

$$\zeta_i = \sum_{j=-\infty}^{\infty} a_j \epsilon_{i-j} \quad \text{with} \quad a_j = \begin{cases} 0 & \text{for } j \le -1, \\ 2^{-j} & \text{for } j \ge 0. \end{cases} \tag{1.3}$$

In the setting of ergodic observations, Györfi [13], Györfi and Masry [17], Yakowitz [34] have obtained the L^1 and L^2 convergence for density estimates. Rosa [27] established the convergence of the conditional median. Delecroix *et al.* [7] have studied the mean uniform convergence and the almost sure convergence of a family of recursive estimates of the density. Delecroix and Rosa [8] have studied the uniform almost sure convergence of the kernel autoregression function estimate and its derivatives for Markovian processes. The uniform almost sure convergence of kernel estimates of the regression function and their robust versions have been established by Laïb and Ould-Saïd [23]. Ould-Saïd [25] has studied a nonparametric predictor based on estimation of the conditional mode function.

The paper is organized as follows. In Section 2, we list the assumptions that we need. In Section 3, we give some preliminary results. In Section 4, we establish Theorem 1, which states that $\Psi_n(x)$ converges completely to $\Psi(x)$ as $n \to \infty$, uniformly with respect to x in a fixed compact subset Δ of $\mathbb{R}$. The difficulty arising from the unboundedness of the function $\rho(\cdot)$ is handled by a truncation argument similar to those used by Mack and Silverman [24]. Generalization of this result to the d-dimensional case is considered in Section 5. Finally, Section 6 is devoted to the application of Theorem 1 to prediction of Markovian time series analysis.

2. Notation, definitions and assumptions

Throughout the paper, the following notations will be used: $\|\phi\| = \sup_{t\in\mathbb{R}} |\phi(t)|$ denotes the sup-norm of the bounded function $\phi : \mathbb{R} \to \mathbb{R}$, $\Phi_x(X)$ denotes $I\{X \in I_n^r(x)\}$, where $I\{A\}$ is the indicator function of the set A, and $a_n = \mu(I_n^r(x))$, where μ stands for the Lebesgue measure. The compact $[a, b] = \bigcup_{-m_n}^{m_n} I_n^r$ with $m_n = [(b-a)/a_n] + 1$ (where $[t]$ is the integer part of t) will be denoted by Δ. The symbol C will denote a generic positive constant, whose value may change from one part of calculation to another.

DEFINITION. Let $\{\mathscr{A}_i;\ i \ge 1\}$ be a nondecreasing sequence of σ-fields and let $\{X_i;\ i \in \mathbb{Z}\}$ be a sequence of random variables such that Z_i is $\mathscr{A}_i$-measurable. The sequence of real random variables $\{Z_i\}$ is called a *martingale difference process* with respect to the σ-fields $\mathscr{A}_i$, if

$$\mathrm{E}\,(X_i \mid \mathscr{A}_{i-1}) \overset{\text{a.s.}}{=} 0.$$

It is ergodic if it is stationary; and every invariant event has probability zero or one. In the sequel we denote by $\mathscr{F}_i = \sigma((X_1, Y_1), \ldots, (X_i, Y_i))$ the σ-field generated by $(X_1, Y_1), \ldots, (X_i, Y_i)$.

Note that Györfi and Lugosi [16] have pointed out that the ergodic condition alone is not sufficient to ensure the L^1 consistency of kernel or histogram density estimates. A complementary assumption is therefore needed. Györfi *et al.* [15] suppose that the conditional distribution of X_m given $\mathscr{F}_{-\infty}^0 = \sigma(X_0, X_{-1}, \ldots)$ (the σ-field generated by the entire past of the process) is absolutely continuous almost surely, for some positive integer m.

Following the work of these authors, we assume that:

(A1) For all $i \in \mathbb{N}$, the conditional density $f_{X_i}^{\mathscr{F}_{i-1}}(\cdot)$ of X_i with respect to $\mathscr{F}_{i-1}$ and the corresponding marginal density $f(\cdot)$ exist.

(A2) For all $i \in \mathbb{N}$, both densities $f_{X_i}^{\mathscr{F}_{i-1}}(\cdot)$ and $f(\cdot)$ belong to $\mathscr{C}_0(\mathbb{R})$, where $\mathscr{C}_0(\mathbb{R})$ is the space of continuous functions going to zero at infinity.

(A3) The sequence $\{n^{-1}\sum_{i=1}^n f_{X_i}^{\mathscr{F}_{i-1}}(x)\}$ converges uniformly in x to $f(x)$, that is

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n f_{X_i}^{\mathscr{F}_{i-1}}(x) - f(x) \right| \overset{\text{a.s.}}{=} 0.$$

We also need the following assumptions on the probability distribution of $\{Z\}$:

(A4) The random variables $Z_i = (X_i, Y_i)$, $i = 1, 2, \ldots$, form a strictly stationary ergodic sequence $\{Z_i\}$.

(A5) There exists a compact subset Δ of $\mathbb{R}$ such that $\inf\{f(x); x \in \Delta\} > 0$ and $\sup\{f(x); x \in \Delta\} < \infty$.

(A6) $\rho(\cdot)$ is a real-valued Borel function defined on $\mathbb{R}$ such that $\mathrm{E}\,(|\rho(Y)|^\gamma) < \infty$ for some $\gamma > 1$.

(A7) The conditional mean of $\rho(Y_i)$ given X_i and $\mathscr{F}_{i-1}$ only depends on X_i, that is, for all $i \geq 1$, $\mathrm{E}\,(\rho(Y_i) \mid \mathscr{G}(X_i)) = \Psi(X_i)$, where $\mathscr{G}(X_i) = \sigma(X_i, \mathscr{F}_{i-1})$.

(A8) The function $\Psi(\cdot)$ is bounded on the real line.

(A9) $\Psi(\cdot)$ is a c_Ψ-Lipschitzian function of order one for some $0 < c_\Psi < \infty$.

(A10) The sequence $\{a_n : n \geq 1\}$ satisfies:

- (i) $a_n \to 0$, $na_n \to \infty$ as $n \to \infty$,
- (ii) $na_n/\log n \to \infty$ as $n \to \infty$,
- (iii) There exists $\beta \in \,]2b/\gamma, 1/3[$, where $b > 1$ is a real number, such that $a_n^3 n^{1-3\beta}/\log n \to \infty$ as $n \to \infty$.

3. Some preliminary results

The estimator defined in (1.1) can be written as $\Psi_n(x) = g_n(x)/f_n(x)$, where

$$g_n(x) = \frac{1}{na_n}\sum_{i=1}^{n}\rho(Y_i)\Phi_x(X_i), \quad f_n(x) = \frac{1}{na_n}\sum_{i=1}^{n}\Phi_x(X_i). \tag{3.1}$$

For all $n \in \mathbb{N}$, let $h_n(x) = (na_n)^{-1}\sum_{i=1}^{n} \mathrm{E}\left[\rho(Y_i)\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right]$. Clearly,

$$\begin{aligned}\sup_{x\in\Delta}\left|\Psi_n(x) - \Psi(x)\right| \le \left(\left|\inf_{x\in\Delta} f(x)\right|\right)^{-1}\Big[&\sup_{x\in\Delta}\left|g_n(x) - h_n(x)\right| \\ &+ \sup_{x\in\Delta}\left|h_n(x) - f(x)\Psi(x)\right| \\ &+ \sup_{x\in\Delta}\left|\Psi_n(x)\right|\sup_{x\in\Delta}\left|f_n(x) - f(x)\right|\Big].\end{aligned} \tag{3.2}$$

The limiting behaviors of the terms in the right-hand side of (3.2) are given in the following propositions.

PROPOSITION 1. *Suppose that the assumptions* (A6) *and* (A10) *hold. Then, we have*

$$\lim_{n\to\infty}\sup_{x\in\Delta}\left|g_n(x) - h_n(x)\right| \overset{\text{a.s.}}{=} 0. \tag{3.3}$$

The proof of Proposition 1 is presented in Appendix I.

PROPOSITION 2. *Suppose that* (X_i, Y_i) *satisfies the assumptions* (A1)–(A5) *and* (A7)–(A9). *Then, we have*

$$\lim_{n\to\infty}\sup_{x\in\Delta}\left|h_n(x) - \Psi(x)f(x)\right| \overset{\text{a.s.}}{=} 0. \tag{3.4}$$

PROOF. Decompose $h_n(x) - \Psi(x)f(x)$ as follows:

$$h_n(x) - \Psi(x)f(x) = \left[h_n(x) - \Psi(x)l_n(x)\right] + \left[\Psi(x)l_n(x) - f(x)\Psi(x)\right] \tag{3.5}$$

with

$$l_n(x) = \frac{1}{na_n}\sum_{i=1}^{n}\mathrm{E}\left(\Phi_x(X_i)\middle|\mathscr{F}_{i-1}\right).$$

We then have

$$h_n(x) - \Psi(x)l_n(x) = \frac{1}{na_n}\sum_{i=1}^{n}W_{i,n}(x),$$

where

$$W_{i,n}(x) \overset{\text{a.s.}}{=} \mathrm{E}\left[\rho(Y_i)\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right] - \Psi(x)\left[\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right].$$

It follows from the properties of the conditional expectation and (A7) that

$$\begin{aligned} W_{i,n} &\overset{\text{a.s.}}{=} \mathrm{E}\left[\mathrm{E}\left(\rho(Y_i)\Phi_x(X_i) \mid \mathscr{G}(X_i)\right) \mid \mathscr{F}_{i-1}\right] - \Psi(x)\,\mathrm{E}\left[\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right] \\ &= \mathrm{E}\left[\Phi_x(X_i)\,\mathrm{E}\left(\rho(Y_i) \mid \mathscr{G}(X_i)\right) \mid \mathscr{F}_{i-1}\right] - \Psi(x)\,\mathrm{E}\left[\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right] \\ &= \mathrm{E}\left[(\Psi(X_i) - \Psi(x))\,\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right] \\ &= \int_{I_n^r(x)} (\Psi(u) - \Psi(x))\,\mathrm{P}\left(X_i \in du \mid \mathscr{F}_{i-1}\right). \\ &= \int_{I_n^r(x)} (\Psi(u) - \Psi(x))\, f_{X_i}^{\mathscr{F}_{i-1}}(u)du, \end{aligned}$$

since $\mathscr{F}_{i-1} \subset \mathscr{G}(X_i)$. Moreover, since $\Psi(\cdot)$ is a c_Ψ-Lipschitzian function, we obtain by (A5)

$$\begin{aligned} \left|h_n(x) - \Psi(x)l_n(x)\right| &\le c_\Psi \int_{I_n^r(x)} \left[\frac{1}{n}\sum_{i=1}^{n} f_{X_i}^{\mathscr{F}_{i-1}}(u)du - f(x)\right] + c_\Psi C a_n \\ &\le c_\Psi a_n \left[\left\|\frac{1}{n}\sum_{i=1}^{n} f_{X_i}^{\mathscr{F}_{i-1}} - f\right\| + C\right]. \end{aligned}$$

To treat the second member of the right-hand side of (3.5), observe that

$$\left|\Psi(x)l_n(x) - f(x)\Psi(x)\right| \le \|\Psi\| \left\|\frac{1}{n}\sum_{i=1}^{n} f_{X_i}^{\mathscr{F}_{i-1}} - f\right\|.$$

The last upper bounds being independent of x, it follows that

$$\sup_{x\in\Delta}\left|h_n(x) - \Psi(x)l_n(x)\right| \le \left(c_\Psi a_n + \|\Psi\|\right)\left\|\frac{1}{n}\sum_{i=1}^{n} f_{X_i}^{\mathscr{F}_{i-1}} - f\right\| + c_\Psi C a_n.$$

The result follows then from (2.1) since $a_n \to 0$ as $n \to \infty$. □

PROPOSITION 3. *Assume that* (A6) *and* (A10) *hold. Then, we have*

$$\lim_{n\to\infty}\sup_{x\in\Delta}\left|\Psi_n(x)\right|\sup_{x\in\Delta}\left|f_n(x) - f(x)\right| \overset{\text{a.s.}}{=} 0. \tag{3.6}$$

PROOF. Since the function $\rho(\cdot)$ is not assumed to be bounded, write

$$\Psi_n(x) = \Psi_{n,1}(x) + \Psi_{n,2}(x),$$

where

$$\Psi_{n,1}(x) = \sum_{i=1}^{n} Q_{n,i}(x)\rho(Y_i)I\left\{|\rho(Y_i)| > n^{\beta}\right\},$$

$$\Psi_{n,2}(x) = \sum_{i=1}^{n} Q_{n,i}(x)\rho(Y_i)I\left\{|\rho(Y_i)| \le n^{\beta}\right\}$$

with

$$Q_{n,i}(x) = \Phi_x(X_i)\left(\sum_{i=1}^{n}\Phi_x(X_i)\right)^{-1}.$$

Now observe that $\left\{\Psi_{n,1}(x) \neq 0\right\} \subset \bigcup_{i=1}^{n}\left\{|\rho(Y_i)| > n^{\beta}\right\}$. Stationarity and Markov's inequality imply, under (A6), that

$$\mathrm{P}\left\{\Psi_{n,1}(x) \neq 0\right\} \le n\,\mathrm{P}\left\{|\rho(Y_0)| > n^{\beta}\right\} \le Cn^{1-\beta\gamma}.$$

Borel-Cantelli's Lemma implies that $\mathrm{P}\left\{\overline{\lim}\{\Psi_{n,1}(x) \neq 0\}\right\} = 0$, since $\beta\gamma > 2$, and that $\mathrm{P}\left\{\underline{\lim}\{|\Psi_n(x)| \le n^{\beta}\}\right\} = 1$. Consequently, for all $\omega \in \underline{\lim}\left\{|\Psi_n(x)| \le n^{\beta}\right\}$ and $n \ge n_0(\omega)$,

$$\sup_{x\in\Delta}\left|\Psi_n(x)\right|\sup_{x\in\Delta}\left|f_n(x) - f(x)\right| \le n^{\beta}\sup_{x\in\Delta}\left|f_n(x) - f(x)\right| \text{ almost surely.}$$

The remainder of the proof is similar to the proof of Proposition 1 (it is a particular case of Proposition 1, where we take $\rho(y) = 1$ for all $y \in \mathbb{R}^p$). □

4. Main result and concluding remarks

The main result of this paper is the following theorem, whose proof is a consequence of the preliminary results established in Section 3.

THEOREM 1. *Suppose that* (X_i, Y_i) *satisfies the assumptions* (A1)–(A10). *Then, we have*

$$\lim_{n\to\infty}\sup_{x\in\Delta}\left|\Psi_n(x) - \Psi(x)\right| \overset{\text{a.s.}}{=} 0. \tag{4.1}$$

PROOF. By (A5), $\sup_{x\in\Delta} f^{-1}(x)$ is bounded. Taking the limits as $n \to \infty$, the desired result follows then from Proposition 1, Proposition 2 and Proposition 3 . □

REMARKS. 1. Our theorem applies to many general situations. First, the ergodic condition is very weak and is satisfied by many important time series. Second, our approach removes the restriction that the time series is bounded.

2. Our result allows us to construct estimates of quantities including the estimate of the regression function $r(\cdot)$ as a special case. Referring to the initial set-up of real-valued random variables $Z_j, j \geq 1$, one may estimate interesting quantities such as $(Z_{i+s}^m \mid Z_{i-u+1}, \ldots, Z_i)$, $m > 0$ and $s \geq 0$ both integers, or a function of a finite number of Z's beyond the present Z_i.

3. The condition (2.1) is obviously satisfied by ergodic processes with sufficiently smooth densities (see, Györfi [13] and Györfi and Masry [17]).

4. We emphasize that without (A7) the partitioning estimate is not consistent Györfi *et al.* [18, Theorem 2]. Note that the assumption (A7) holds when, for instance, we take $Y_i = X_{i+1}$ and $\{X_i\}$ a Markov process. In particular, it holds in the case of a nonlinear autoregressive process $X_{i+1} = T(X_i) + \epsilon_{i+1}$, where $T(X_i) = \mathrm{E}(X_{i+1} \mid X_i = x)$, $\{\epsilon_i\}$ is an independent and identically distributed sequence with mean zero and unit variance and the X_i is independent of ϵ_{i+1}.

5. The Lipschitz condition imposed on $\Psi(\cdot)$ is commonly introduced in parametric regression estimation.

6. The assumptions (A1) and (A4) allow us to obtain the uniform convergence of the bias term. For the variance term, under the ergodic condition we do not have any exponential inequality at hand as we do for the independent and identically distributed or mixing case. To treat the variance term, we have used an exponential type inequality for martingale difference processes due to Laïb [21] , in the proof of Proposition 1.

7. A crucial point, in this context, concerns the choice of an appropriate smoothing parameter. The basic tools for data-driven bandwidth selection rules (see, for example, Hart and View [19]) are based on asymptotic optimality properties usually derived from inequalities on moments of sums of the underlying variables. As far as we know, we do not have such inequalities for ergodic processes.

8. The results obtained under general ergodicity conditions are theoretically interesting, since they are stated under such an unrestrictive dependence condition. But they do not provide convergence rates (see, Krengel [20, p. 14]). There is no Law of the Iterated Logarithm for the quantities in (2.1), which can go to zero at an arbitrary rate.

5. Generalisation to the d-dimensional case

We briefly discuss a generalization of our result to the d-dimensional case. Let $\mathscr{P}_n = \{A_{n,j}, j = 1, 2, \ldots\}$ be a sequence of partitions of $\mathbb{R}^d$ into cubes of the form $\prod_{i=1}^d [c_i k_{i,j} a_n, c_i(k_{i,j} + 1)a_n)$, where the c_i's are reals, the $k_{i,j}$'s are integers and a_n is the bandwidth length. For a given x, let $A_{n,j}(x)$ be the cube containing x. Take the mean of $\rho(Y_i)$'s for which the X_i's corresponding to the Y_i's fall in $A_{n,j}(x)$. The partitioning estimate $\Psi_n(x)$ is defined by

$$\Psi_n(x) = g_n(x)/f_n(x),$$

https://doi.org/10.1017/S1446788700000835 Published online by Cambridge University Press

where $g_n(x) = (na_n^d)^{-1} \sum_{i=1}^n \rho(Y_i) I\{X_i \in A_{n,j}(x)\}$ and $f_n(x) = (na_n^d)^{-1} \sum_{i=1}^n I\{X_i \in A_{n,j}(x)\}$. We have the following result which extends Theorem 1 to the d-dimensional case.

THEOREM 2. *If all the assumptions mentioned in Theorem* 1 *hold with in addition* (A10) *strengthened to*

(A11)-(i) $a_n^d \to 0$, $na_n^d \to \infty$ *it as* $n \to \infty$,
(A11)-(ii) $na_n^d / \log n \to \infty$, *as* $n \to \infty$,
(A11)-(iii) *there exists* $\beta \in]2b/\gamma, 1/3[$, *where* $b > 1$ *is a real number, such that* $a_n^{3d} n^{1-3\gamma} / \log n \to \infty$, *as* $n \to \infty$,

then (4.1) *still holds.*

6. Application to prediction for Markov processes

The main application of our results concerns time series analysis and prediction in the Markovian case. Let $\{Z_i;\ i \in \mathbb{Z}\}$ be a real-valued stationary process. Given observations $Z_1, \dots, Z_N$, we wish to construct a predictor of Z_{N+1}. For this purpose, we choose a positive integer s and try to approximate

$$Z_{N+1}^* = \mathrm{E}\left[\rho(Z_{N+1}) \mid Z_1, Z_2, \dots, Z_{N-s+1}, \dots, Z_N\right].$$

In the special case when $\{Z_i;\ i \in \mathbb{Z}\}$ is a real-valued Markovian process of order s, we know that the real random variable $\Psi(Z_{N-s+1}, \dots, Z_N)$ is the best approximation of Z_{N+1}. If we take $N \geq s$, $X_i = (Z_i, \dots, Z_{i+s-1})$, $Y_i = \rho(Z_{i+s})$, $n = N - s + 1$, we then obtain the previous scheme and define the predictor as $\hat{Z}_{N+1} = \Psi_N(X_N)$ with respect to the loss function $\eta(\cdot)$ defined by

$$\eta(x) = \int_{-\infty}^{x} \rho(u)du.$$

COROLLARY 1. *Under the conditions of Theorem* 1, *if the process* $\{Z_i;\ i \in \mathbb{Z}\}$ *is stationary and Markovian, then we have*

$$\left|\hat{Z}_{N+1} - Z_{N+1}^*\right| I\{(Z_{N-s+1}, \dots, Z_N) \in I_N(x)\} \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to \infty.$$

Appendix I

In order to establish Proposition 1, we need the following exponential type inequality due to Laïb [21] for martingale difference processes.

https://doi.org/10.1017/S1446788700000835 Published online by Cambridge University Press

LEMMA 1 (Laïb [21]). *Let* $\{(X_n, \mathscr{F}_n);\ n \geq 1\}$ *be a martingale difference with* $|X_i| \leq b$ *almost surely for all* $i \geq 1$. *Set* $S_n = \sum_{i=1}^{n} X_i$ *and* $M_n = \max_{1\leq k\leq n} |S_k|$. *Then, for all* $\lambda > 0$, *we have*

$$\mathrm{P}\,(M_n \geq \lambda) \leq 2\exp\left(-\frac{\lambda}{2b} h\left(\frac{\lambda}{nb}\right)\right),$$

where $h(x) = 2x^{-1}[(1+x)\log(1+x) - x]$ *is an increasing function.*

PROOF OF PROPOSITION 1. Let

$$g_n(x) - h_n(x) = \frac{1}{na_n}\sum_{i=1}^{n} \phi_x(X_i, Y_i),$$

where

$$\phi_x(X_i, Y_i) = \rho(Y_i)\Phi_x(X_i) - \mathrm{E}\,(\rho(Y_i)\Phi_x(X_i) \mid \mathscr{F}_{i-1}).$$

In order to deal with large values of the function $\rho(\cdot)$, decompose $\phi_x(X_i, Y_i)$ as follows $\phi_x(X_i, Y_i) = \phi_x^+(X_i, Y_i) + \phi_x^-(X_i, Y_i)$, where

$$\begin{aligned}\phi_x^-(X_i, Y_i) = {} & \rho(Y_i)I\left\{|\rho(Y_i)| \leq n^\beta\right\}\Phi_x(X_i)\\ & - \mathrm{E}\left[\rho(Y_i)I\left\{|\rho(Y_i)| \leq n^\beta\right\}\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right],\end{aligned}$$

and

$$\begin{aligned}\phi_x^+(X_i, Y_i) = {} & \rho(Y_i)I\left\{|\rho(Y_i)| > n^\beta\right\}\Phi_x(X_i)\\ & - \mathrm{E}\left[\rho(Y_i)I\left\{|\rho(Y_i)| > n^\beta\right\}\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right].\end{aligned}$$

Since

$$|g_n(x) - h_n(x)| \leq \frac{1}{na_n}\left|\sum_{i=1}^{n}\phi_x^-(X_i, Y_i)\right| + \frac{1}{na_n}\left|\sum_{i=1}^{n}\phi_x^+(X_i, Y_i)\right|,$$

we have for any $\epsilon > 0$ that

$$\begin{aligned}\text{(AI.1)}\quad \mathrm{P}\left\{\sup_{x\in\Delta}|g_n(x) - h_n(x)| \geq \epsilon\right\} \leq {} & \mathrm{P}\left\{\sup_{x\in\Delta}\frac{1}{na_n}\left|\sum_{i=1}^{n}\phi_x^-(X_i, Y_i)\right| \geq \epsilon/2\right\}\\ & + \mathrm{P}\left\{\sup_{x\in\Delta}\frac{1}{na_n}\left|\sum_{i=1}^{n}\phi_x^+(X_i, Y_i)\right| \geq \epsilon/2\right\}.\end{aligned}$$

We now proceed to evaluate the probabilities on the right-hand side of (AI.1). We begin with $\mathrm{P}\{\sup_{x\in\Delta}(na_n)^{-1}\left|\sum_{i=1}^n \phi_x^-(X_i, Y_i)\right| \geq \epsilon/2\}$. Let x_r be a point in I_n^r. We have then

$$\mathrm{P}\left\{\sup_{x\in\Delta}\frac{1}{na_n}\left|\sum_{i=1}^n \phi_x^-(X_i, Y_i)\right| \geq \epsilon/2\right\} \leq \mathrm{P}\left\{\max_{-m_n\leq r\leq m_n}\frac{1}{na_n}\left|\sum_{i=1}^n \phi_{x_r}^-(X_i, Y_i)\right| \geq \epsilon/2\right\}.$$

Since for each fixed $1 \leq i \leq n$ and $-m_n \leq r \leq m_n$, $\phi_{x_r}^-(X_i, Y_i)$ is $\mathscr{F}_i$-measurable and $\mathrm{E}[\phi_{x_r}^-(X_i, Y_i) \mid \mathscr{F}_{i-1}] = 0$ almost surely, then $\left\{(\phi_{x_r}^-(X_i, Y_i), \mathscr{F}_i)\right\}$ forms a martingale difference process. Furthermore, since $\left|\phi_{x_r}^-(X_i, Y_i)\right| \leq 2n^\beta = b_n$, we obtain with the choice $\lambda = \lambda_n = \epsilon n a_n/2$ in Lemma 1

(AI.2)
$$\mathrm{P}\left\{\max_{-m_n\leq r\leq m_n}\frac{1}{na_n}\left|\sum_{i=1}^n \phi_{x_r}^-(X_i, Y_i)\right| \geq \epsilon/2\right\} \leq 4m_n \exp\left(-\frac{\epsilon a_n n^{1-\beta}}{8}h\left(\frac{\epsilon a_n n^{-\beta}}{4}\right)\right).$$

Since $a_n n^{-\beta} \to 0$ as $n \to \infty$, a Taylor expansion of order one of the function $x \to h(x)$ around 0 shows that the right-hand side of (AI.2) is bounded by

$$4m_n \exp\left(-\frac{\epsilon^2 a_n^2 n^{1-2\beta}}{32}\left[1 - \frac{\epsilon a_n n^{-\beta}}{12}\right]\right).$$

It results from Borel-Cantelli's Lemma, (A10)-(i) and A(10)-(iii) that

(AI.3)
$$\limsup_{n\to\infty}\sup_{x\in\Delta}\frac{1}{na_n}\left|\sum_{i=1}^n \phi_x^-(X_i, Y_i)\right| \overset{\text{a.s.}}{=} 0.$$

Now we evaluate the probability in the second member of the right-hand side of (AI.1). We have

$$\sup_{x\in\Delta}\frac{1}{na_n}\left|\sum_{i=1}^n \phi_x^+(X_i, Y_i)\right| \leq \mathrm{I}_n + \mathrm{II}_n,$$

where

$$\mathrm{I}_n = \sup_{x\in\Delta}\frac{1}{na_n}\left|\sum_{i=1}^n \rho(Y_i) I\left\{|\rho(Y_i)| > n^\beta\right\}\Phi_x(X_i)\right|$$

and

$$\mathrm{II}_n = \sup_{x\in\Delta}\frac{1}{na_n}\left|\sum_{i=1}^n \mathrm{E}\left[\rho(Y_i) I\left\{|\rho(Y_i)| > n^\beta\right\}\Phi_x(X_i) \mid \mathscr{F}_{i-1}\right]\right|.$$

Observe that $\{\mathrm{I}_n \neq 0\} \subset \bigcup_{i=1}^n \{|\rho(Y_i)| > n^\beta\}$. We obtain, by the same arguments as in the proof of Proposition 3,

$$\text{(AI.4)} \qquad \mathrm{P}\left\{\underline{\lim}\{\mathrm{I}_n \neq 0\}\right\} = 0 \quad \text{since } \beta\gamma > 2.$$

Now, we give an estimate for $\mathrm{E}\left(|\,\mathrm{II}_n\,|\right)$. Let $a > 1$ and $b > 1$ be conjugate exponents, $a^{-1} + b^{-1} = 1$ $(a \leq \gamma)$. Stationarity and Hölder's inequality imply, under (A10)–(ii), that

$$\begin{aligned}
\mathrm{E}\left(|\,\mathrm{II}_n\,|\right) &\leq \frac{1}{na_n} \sum_{i=1}^n \left(\mathrm{E}\left(|\rho(Y_i)|^a\right)\right)^{1/a} \left(\mathrm{P}\{|\rho(Y_i)| > n^\beta\}\right)^{1/b} \\
&\leq C a_n^{-1} n^{-\beta\gamma/b} \left(\mathrm{E}\left(|\rho(Y_0)|^\gamma\right)\right)^{1/b} \\
&\leq C a_n^{-1} n^{-\beta\gamma/b} \\
&\leq \frac{C n^{1-\beta\gamma/b}}{\log n}.
\end{aligned}$$

Hence, we obtain by Markov's inequality that for any $\epsilon > 0$,

$$\mathrm{P}\left\{|\,\mathrm{II}_n\,| > \epsilon\right\} \leq C n^{1-\beta\gamma/b}/\epsilon \log n.$$

Thus, $\sum_{n \geq 1} \mathrm{P}\{|\,\mathrm{II}_n\,| > \epsilon\} < \infty$ since $\beta\gamma > 2b$. Borel-Cantelli's Lemma yields

$$\text{(AI.5)} \qquad \mathrm{P}\left\{\overline{\lim}\{|\,\mathrm{II}_n\,| > \epsilon\}\right\} = 0.$$

It follows from (AI.4) and (AI.5) that

$$\text{(AI.6)} \qquad \lim_{n\to\infty} \sup_{x \in \Delta} \frac{1}{na_n} \left| \sum_{i=1}^n \phi_x^+(X_i, Y_i) \right| \overset{\text{a.s.}}{=} 0.$$

The proof of Proposition 1 results from (AI.3) and (AI.6). □

References

[1] D. W. K. Andrews, 'Non-strong mixing autoregressive processes', *J. Appl. Probab.* **21** (1984), 930–934.

[2] R. Ash and M. Gardner, *Topics in stochastic processes* (Academic Press, New York, 1975).

[3] D. Bosq, 'Nonparametric prediction for unbounded almost stationary processes', in: *Nonparametric functional estimation and related topics* (Kluwer Acad. Publ., Amsterdam, 1991) pp. 389–403.

[4] L. Breiman, *Probability* (Addison-Wesley, Reading, MA, 1968).

[5] A. Carbonez, L. Györfi and E. C. Van der Meulen, 'Partitioning estimates of regression function under random censoring', *Statist. Decisions* **13** (1995), 21–37.

[6] G. Collomb and W. Härdle, 'Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations', *Stochastic Process. Appl.* **23** (1986), 77–89.

[7] M. Delecroix, M. N. Nogueira and A. C. Martins Rosa, 'Sur l'estimationn de la densité pour des observations ergodiques', *Stat. et Anal des données* **16** (1991), 25–38.

[8] M. Delecroix and A. C. Rosa, 'Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis', *J. Nonparametric Statistics* **6** (1996), 367–382.

[9] L. Devroye and L. Györfi, 'Distribution-free exponential bound for the L_1 error of partitioning estimates of a regression function', in: *Probability and statistical decision theory (Bad Tatzmannsdorf, 1983)* (Reidel, Boston, 1985) pp. 67–76.

[10] J. Diebolt and N. Laïb, 'A weak invariance principle for cumulated functionals of the regressogram estimator with dependent data', *J. Nonparametr. Statist.* **4** (1994), 149–163.

[11] L. Gordon and R. A. Olshen, 'Consistent nonparametric regression from recursive partitioning schemes', *J. Multivariate Anal.* **10** (1980), 611–627.

[12] ———, 'Almost sure consistent nonparametric regression from recursive partitioning schemes', *J. Multivariate Anal.* **15** (1984), 147–163.

[13] L. Györfi, 'Strong consistent density estimate from ergodic sample', *J. Multivariate Anal.* **11** (1981), 81–84.

[14] ———, 'Universal consistencies of a regression estimate for unbounded regression function', in: *Nonparametric functional estimation and related topics* (ed. Roussas) (Kluwer Acad. Publ., Amsterdam, 1991) pp. 329–338.

[15] L. Györfi, W. Härdle, P. Sarda and P. Vieu, *Nonparametric curve estimation from time series*, Lecture Notes in Statistics 60 (Springer, New York, 1989).

[16] L. Györfi and G. Lugosi, 'Kernel density estimation from ergodic sample is not universally consistent', *Comput. Statist. Data Anal.* **14** (1992), 437–442.

[17] L. Györfi and E. Masry, 'The L_1 and L_2 strong consistency of recursive kernel density estimation from dependent samples', *IEEE Trans. Inform. Theory* **36** (1990), 531–539.

[18] L. Györfi, G. Morvai and S. Yakowitz, 'Limits to consistent on line forecasting for ergodic time series', *IEEE Trans. Inform. Theory* **44** (1998), 868–892.

[19] J. Hart and P. Vieu, 'Data driven bandwidth choice for density estimation based on dependent data', *Ann. Statist.* **18** (1990), 837–890.

[20] U. Krengel, *Ergodic theorems*, De Gruyter Studies in Mathematics 6 (Walter De Gruyter, Berlin, 1985).

[21] N. Laïb, 'Exponential-type inequalities for martingale difference sequences. Application to nonparametric regression estimation', *Comm. Statist. Theory Methods*, to appear.

[22] ———, 'Estimation non-paramétrique de la régression pour des données dépendantes. Application à la prévision', *C. R. Acad. Sci. Paris* **317** (1993), 1173–1177.

[23] N. Laïb and E. Ould-Saïd, 'Estimation non-paramétrique robuste de la fonction de régression pour des observations ergodiques', *C. R. Acad. Sci. Paris* **320** (1996), 271–276.

[24] Y. P. Mack and B. W. Silverman, 'Weak and strong uniform consistency of kernel regression estimates', *Z. Wahrsch. Verw. Gebiete* **61** (1982), 405–415.

[25] E. Ould-Saïd, 'A note on ergodic processes prediction via estimation of the conditional mode function', *Scand. J. Statist.* **2** (1997), 231–240.

[26] T. D. Pham and T. Tran, 'Some mixing properties of time series models', *Stochastic Process. Appl.* **19** (1985), 279–303.

[27] A. C. Rosa, 'Estimation non-paramétrique de la médianne conditionnelle sous hypothèse ergodique. Application à la prévision des processus', in: *XXIV^e Journées de statistiques,*' Bruxelles, (1992) pp. 325–328.

[28] M. Rosenblatt, 'Dependence and asymptotic independence from random processes', *M.A.A. Stud. Math.* **18** (1978), 24–45.

[29] G. G. Roussas, 'Nonparametric regression estimation under mixing conditions', *Stochastic Process. Appl.* **36** (1990), 107–116.

[30] L. T. Tran, 'Nonparametric function estimation for time series by local average estimators', *Ann. Statist.* **21** (1993), 1040–1057.

[31] Y. K. Truong, 'Robust nonparametric regression in time series', *J. Multivariate Anal.* **41** (1992), 163–177.

[32] Y. K. Truong and C. J. Stone, 'Nonparametric function estimation involving time series', *Ann. Statist.* **20** (1992), 77–97.

[33] C. S. Withers, 'Central limit theorems for dependent variables', *Z. Wahrsch. Verw. Gebiete* **57** (1981b), 509–534.

[34] S. Yakowitz, 'Nonparametric density and regression estimation for Markov sequences without mixing assumptions', *J. Multivariate Anal.* **30** (1989), 124–136.

L.S.T.A. Université Paris 6
Aile 45-55, 3ème étage
4, Place Jussieu
75252 Paris Cedex 05
France
e-mail: nal@ccr.jussieu.fr