Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ann. I. H. Poincaré – PR 41 (2005) 859–870

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES

www.elsevier.com/locate/anihpb

Forward estimation for ergodic time series

Gusztáv Morvai [a], Benjamin Weiss [b,*]

[a] Research Group for Informatics and Electronics of the Hungarian Academy of Sciences, 1521 Goldmann György tér 3, Budapest, Hungary
[b] Hebrew University of Jerusalem, Jerusalem 91904, Israel

Received 29 August 2003; accepted 15 June 2004

Available online 5 February 2005

Abstract

The forward estimation problem for stationary and ergodic time series $\{X_n\}_{n=0}^{\infty}$ taking values from a finite alphabet $\mathcal{X}$ is to estimate the probability that $X_{n+1}=x$ based on the observations X_i, $0 \leqslant i \leqslant n$ without prior knowledge of the distribution of the process $\{X_n\}$. We present a simple procedure g_n which is evaluated on the data segment $(X_0,\dots,X_n)$ and for which, $\mathrm{error}(n)=|g_n(x)-P(X_{n+1}=x|X_0,\dots,X_n)|\to 0$ almost surely for a subclass of all stationary and ergodic time series, while for the full class the Cesaro average of the error tends to zero almost surely and moreover, the error tends to zero in probability.

Résumé

Le problème de l'estimation future d'une série temporelle ergodique et stationnaire $\{X_n\}_{n=0}^{\infty}$, prenant ses valeurs dans un alphabet fini $\mathcal{X}$, est d'estimer la probabilité que $X_{n+1}=x$, connaissant les X_i pour $0 \leqslant i \leqslant n$ mais sans connaissance préalable de la distribution du processus $\{X_i\}$. Nous présentons un procédé simple g_n, évalué sur les données $(X_0,\dots,X_n)$, pour lequel $\mathrm{erreur}(n)=|g_n(x)-P(X_{n+1}=x|X_0,\dots,X_n)|\to 0$ presque sûrement pour une sous-classe de toutes les séries temporelles ergodiques et stationnaires, tandis que pour la classe entière la moyenne de Cesaro de l'erreur tend vers zéro presque sûrement. De plus, l'erreur tend vers zéro en probabilité.

MSC: 62G05; 60G25; 60G10

Keywords: Nonparametric estimation; Stationary processes

[*] Corresponding author.
E-mail addresses: morvai@szit.bme.hu (G. Morvai), weiss@math.huji.ac.il (B. Weiss).

0246-0203/$ – see front matter © 2004 Elsevier SAS. All rights reserved.
doi:10.1016/j.anihpb.2004.07.002

1. Introduction

T. Cover [6] posed two fundamental problems concerning estimation for stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$. (Note that a stationary time series $\{X_n\}_{n=0}^{\infty}$ can be extended to be a two sided stationary time series $\{X_n\}_{n=-\infty}^{\infty}$.) Cover's first problem was on backward estimation.

Problem 1. Is there an estimation scheme f_n for the value $P(X_1 = 1|X_{-n}, \dots, X_0)$ such that f_n depends solely on the observed data segment $(X_{-n}, \dots, X_0)$ and

$$\lim_{n\to\infty} \left| f_n(X_{-n}, \dots, X_0) - P(X_1 = 1|X_{-n}, \dots, X_0) \right| = 0$$

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$?

This problem was solved by Ornstein [20] by constructing such a scheme. (See also Bailey [5].) Ornstein's scheme is not a simple one and the proof of consistency is rather sophisticated. For an even more general case, a much simpler scheme and proof of consistency were provided by Morvai, Yakowitz and Györfi [19]. (See also Algoet [1] and Weiss [24].) Note that none of these schemes are reasonable from the data consumption point of view.

Cover's second problem was on forward estimation.

Problem 2. Is there an estimation scheme f_n for the value $P(X_{n+1} = 1|X_0, \dots, X_n)$ such that f_n depends solely on the data segment $(X_0, \dots, X_n)$ and

$$\lim_{n\to\infty} \left| f_n(X_0, \dots, X_n) - P(X_{n+1} = 1|X_0, \dots, X_n) \right| = 0$$

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$?

This problem was answered by Bailey [5] in a negative way, that is, he showed that there is no such scheme. (Also see Ryabko [22], Györfi, Morvai and Yakowitz [11] and Weiss [24].) Bailey used the technique of cutting and stacking developed by Ornstein [21] and Shields [23]. Ryabko's construction was based on a function of an infinite state Markov-chain.

Morvai [16] addressed a modified version of Problem 2. There one is not required to predict for all time instances rather he may refuse to predict for certain values of n. However, he is expected to predict infinitely often. Morvai [16] proposed a sequence of stopping times λ_n and he managed to estimate the conditional probability $P(X_{\lambda_n+1} = 1|X_0, \dots, X_{\lambda_n})$ in the pointwise sense, that is, for his estimator along the proposed stopping time sequence, the error tends to zero as n increases, almost surely. Another estimator was proposed for this modified Problem 2 by Morvai and Weiss [17] for which the λ_n grow more slowly, but the consistency only holds for a certain subclass of all stationary binary time series.

In this paper we consider the original Problem 2 but we shall impose an additional restriction on the possible time series. The conditional probability $P(X_1 = 1|\dots, X_{-1}, X_0)$ is said to be continuous if a version of it is continuous with respect to metric $\sum_{i=0}^{\infty} 2^{-i-1}|x_{-i} - y_{-i}|$, where $x_{-i}, y_{-i} \in \{0, 1\}$.

Problem 3. Is there an estimation scheme f_n for the value $P(X_{n+1} = 1|X_0, \dots, X_n)$ such that f_n depends solely on the data segment $(X_0, \dots, X_n)$ and

$$\lim_{n\to\infty} \left| f_n(X_0, \dots, X_n) - P(X_{n+1} = 1|X_0, \dots, X_n) \right| = 0$$

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$ with continuous conditional probability $P(X_1 = 1|\dots, X_{-1}, X_0)$?

We will answer this question in the affirmative. This class includes all k-step Markov chains. It is not known if the schemes proposed by Bailey [5], Ornstein [20], Morvai, Yakowitz and Györfi [19] solve Problem 3 or not.

Problem 4. Is there an estimation scheme f_n for the value $P(X_{n+1}=1|X_0,\dots,X_n)$ such that f_n depends solely on the data segment $(X_0,\dots,X_n)$ and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\bigl|f_i(X_0,\dots,X_i)-P(X_{i+1}=1|X_0,\dots,X_i)\bigr|=0$$

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$?

Bailey [5] (cf. Algoet [2] also) showed that any scheme that solves Problem 1 can be easily modified to solve Problem 4 (indeed, just exchange the data segment $(X_{-n},\dots,X_0)$ for $(X_0,\dots,X_n)$, but apparently not all solutions of Problem 4 arise in this fashion. For further reading cf. Algoet [1,3], Morvai, Yakowitz and Györfi [19], Györfi et al. [8], Györfi, Lugosi and Morvai [10], Györfi and Lugosi [9] and Weiss [24].

Problem 5. Is there an estimation scheme f_n for the value $P(X_{n+1}=1|X_0,\dots,X_n)$ such that f_n depends solely on the data segment $(X_0,\dots,X_n)$ and for arbitrary $\epsilon>0$,

$$\lim_{n\to\infty}P\bigl(\bigl|f_n(X_0,\dots,X_n)-P(X_{n+1}=1|X_0,\dots,X_n)\bigr|>\epsilon\bigr)=0$$

for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$?

By stationarity, for any scheme that solves Problem 1, the shifted version of it solves Problem 5. (Just replace the data segment $(X_{-n},\dots,X_0)$ by $(X_0,\dots,X_n)$.)

There are existing schemes that solve Problem 4 (e.g. Bailey [5], Ornstein [20], and even for a more general case Morvai, Yakowitz and Györfi [19], Algoet [1], Györfi and Lugosi [9]) and there are schemes that solve Problem 5 (e.g. Bailey [5], Ornstein [20] and for even more general case Morvai, Yakowitz and Györfi [19], Algoet [1], Morvai, Yakowitz and Algoet [18]). In this paper we propose a reasonable, very simple algorithm that *simultanously* solves Problem 3, 4 and 5. Note that the schemes given by Bailey [5], Ornstein [20], Morvai, Yakowitz and Györfi [19], Algoet [1] and Weiss [24] are not reasonable at all, they consume data extremely rapidly, cf. Morvai [15] and it is not known if their schemes solve Problem 3 or not.

2. Preliminaries and main results

Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary time series taking values from a finite alphabet $\mathcal{X}$. (Note that all stationary time series $\{X_n\}_{n=0}^{\infty}$ can be thought to be a two sided time series, that is, $\{X_n\}_{n=-\infty}^{\infty}$.) For notational convenience, let $X_m^n=(X_m,\dots,X_n)$, where $m\leqslant n$. Note that if $m>n$ then X_m^n is the empty string.

Let $g:\mathcal{X}\to(-\infty,\infty)$ be arbitrary.

Our goal is to estimate the conditional expectation $E(g(X_{n+1})|X_0^n)$ from samples X_0^n.

For $k\geqslant 1$ define the stopping times $\tau_i^k(n)$ which indicate where the k-block X_{n-k+1}^n occurs previously in the time series $\{X_n\}$. Formally we set $\tau_0^k(n)=0$ and for $i\geqslant 1$ let

$$\tau_i^k(n)=\min\bigl\{t>\tau_{i-1}^k(n)\colon\ X_{n-k+1-t}^{n-t}=X_{n-k+1}^n\bigr\}. \tag{1}$$

Let $K_n\geqslant 1$ and $J_n\geqslant 1$ be sequences of nondecreasing positive integers tending to ∞ which will be fixed later.

Define κ_n as the largest $1\leqslant k\leqslant K_n$ such that there are at least J_n occurrences of the block X_{n-k+1}^n in the data segment X_0^n, that is,

$$\kappa_n=\max\bigl\{1\leqslant k\leqslant K_n\colon\ \tau_{J_n}^k(n)\leqslant n-k+1\bigr\} \tag{2}$$

if there is such k and 0 otherwise.

Define λ_n as the number of occurrences of the block $X_{n-\kappa_n+1}^n$ in the data segment X_0^n, that is,

$$\lambda_n = \max\{1 \leqslant j:\ \tau_j^{\kappa_n} \leqslant n-\kappa_n+1\} \tag{3}$$

if $\kappa_n > 0$ and zero otherwise. Observe that if $\kappa_n > 0$ then $\lambda_n \geqslant J_n$.

Our estimate g_n for $E(g(X_{n+1})|X_0^n)$ is defined as $g_0 = 0$ and for $n \geqslant 1$,

$$g_n = \frac{1}{\lambda_n}\sum_{i=1}^{\lambda_n} g(X_{n-\tau_i^{\kappa_n}(n)+1}) \tag{4}$$

if $\kappa_n > 0$ and zero otherwise.

Let $\mathcal{X}^{*-}$ be the set of all one-sided sequences, that is,

$$\mathcal{X}^{*-} = \{(\ldots, x_{-1}, x_0):\ x_i \in \mathcal{X} \text{ for all } -\infty < i \leqslant 0\}.$$

Define the function $G : \mathcal{X}^{*-} \to (-\infty, \infty)$ as

$$G(x_{-\infty}^0) = E(g(X_1)|X_{-\infty}^0 = x_{-\infty}^0).$$

Note that as a conditional expectation this is only defined almost surely. E.g. if $g(x) = 1_{\{x=z\}}$ for a fixed $z \in \mathcal{X}$ then $G(y_{-\infty}^0) = P(X_1 = z|X_{-\infty}^0 = y_{-\infty}^0)$.

Define a distance on $\mathcal{X}^{*-}$ as

$$d^*(x_{-\infty}^0, y_{-\infty}^0) = \sum_{i=0}^{\infty} 2^{-i-1} 1_{\{x_{-i} \neq y_{-i}\}}.$$

Definition. The conditional expectation $G(X_{-\infty}^0)$ is said to be continuous if a version of it is continuous on the set $\mathcal{X}^{*-}$ with respect to metric $d^*(\cdot,\cdot)$. Since this space is compact, in fact, continuity is equivalent to uniform continuity.

The processes with continuous conditional expectation are essentially the Random Markov Processes of Kalikow [12], or the continuous g-measures studied by Mike Keane [13].

Theorem. *Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite alphabet $\mathcal{X}$. Assume $K_n = \max(1, \lfloor 0.1 \log_{|\mathcal{X}|} n \rfloor)$ and $J_n = \max(1, \lceil n^{0.5} \rceil)$. Then*

(A) *if the conditional expectation $G(X_{-\infty}^0)$ is continuous with respect to metric $d^*(\cdot,\cdot)$ then*

$$\lim_{n\to\infty} \left|g_n - E(g(X_{n+1})|X_0^n)\right| = 0 \quad \textit{almost surely,}$$

(B) *without any continuity assumption,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \left|g_i - E(g(X_{i+1})|X_0^i)\right| = 0 \quad \textit{almost surely,}$$

(C) *without any continuity assumption, for arbitrary $\epsilon > 0$,*

$$\lim_{n\to\infty} P(\left|g_n - E(g(X_{n+1})|X_0^n)\right| > \epsilon) = 0.$$

Remarks. Note that these results are valid without the ergodic assumption. One may use the ergodic decomposition throughout the proofs, cf. Gray [7], p. 268.

We note that from the proof of Ryabko [22] and Györfi, Morvai and Yakowitz [11] it is clear that the continuity condition in the first part of the theorem cannot be relaxed. Even for the class of all stationary and ergodic binary time-series with merely almost surely continuous conditional probability $P(X_1 = 1|\ldots, X_{-1}, X_0)$ one cannot solve Problem 2 in the Introduction. (An almost surely continuous conditional probability is such that as a function restricted to a set C with full measure, it is continuous on C.)

We do not know if the shifted version of our proposed scheme g_n solves Problem 1 or not. (That is, in the case when g_n is evaluated on $(X_{-n}, \ldots, X_0)$ rather than on $(X_0, \ldots, X_n)$.)

If $\mathcal{X}$ is a countably infinite alphabet then there is no scheme that could achieve similar result to part (A) in the theorem for all bounded $g(\cdot)$, even if you assume that the resulting $G(\cdot)$ is continuous, and the time series is in fact a first order Markov chain. Indeed, whenever a new state appears which has not occurred before, you are unable to predict, cf. Györfi, Morvai and Yakowitz [11].

3. Auxiliary results

For $k \geqslant 1$, $n \geqslant 0$ and $j \geqslant 0$ it will be useful to define auxiliary processes $\{\tilde{X}_i^{(k,n,j)}\}_{i=-\infty}^{\infty}$ as follows. Let

$$\widetilde{X}_i^{(k,n,j)} = X_{n-\tau_j^k(n)+i} \quad \text{for } -\infty < i < \infty. \tag{5}$$

For an arbitrary stationary time series $\{Y_n\}$ for $k \geqslant 1$ let $\tilde{\tau}_0^k(Y_{-\infty}^{\infty}) = 0$ and for $i \geqslant 1$ define

$$\tilde{\tau}_i^k(Y_{-\infty}^{\infty}) = \min\bigl\{t > \tilde{\tau}_{i-1}^k(Y_{-\infty}^{\infty}) \colon\ Y_{-k+1+t}^t = Y_{-k+1}^0\bigr\}. \tag{6}$$

If it is obvious on which time series $\tilde{\tau}_i^k(Y_{-\infty}^{\infty})$ is evaluated, we will write $\tilde{\tau}_i^k$. Let T denote the left shift, that is, $(Tx_{-\infty}^{\infty})_i = x_{i+1}$.

We will need the next lemmas for later use.

Lemma 1. *Let $\{X_n\}$ be a stationary time series taking values from a finite alphabet $\mathcal{X}$. For $k \geqslant 1$, $n \geqslant 0$ and $j \geqslant 0$, the time series $\{\widetilde{X}_i^{(k,n,j)}\}_{i=-\infty}^{\infty}$ has the same distribution as $\{X_i\}_{i=-\infty}^{\infty}$.*

Proof. Note that by (1), and (6),

$$T^{n-s}\{X_{n-\tau_j^k(n)+l}^{n-\tau_j^k(n)+m} = x_l^m, \tau_j^k(n) = s\} = \{X_l^m = x_l^m, \tilde{\tau}_j^k = s\}$$

where $\tilde{\tau}_j^k$ is evaluated on time series $\{X_i\}_{i=-\infty}^{\infty}$. Now by (5) and stationarity,

$$\begin{aligned} P(\widetilde{X}_l^{(k,n,j)} = x_l, \ldots, \widetilde{X}_m^{(k,n,j)} = x_m) &= \sum_{s=0}^{\infty} P(X_{n-\tau_j^k(n)+l}^{n-\tau_j^k(n)+m} = x_l^m, \tau_j^k(n) = s) = \sum_{s=0}^{\infty} P(X_l^m = x_l^m, \tilde{\tau}_j^k = s) \\ &= P(X_l^m = x_l^m) \end{aligned}$$

and the proof of Lemma 1 is complete. □

Lemma 2. *Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite alphabet $\mathcal{X}$. Assume $K_n \to \infty$, $J_n \to \infty$ and $\lim_{n\to\infty}(J_n/n) = 0$. Then*

$$\lim_{n\to\infty} \kappa_n = \infty \quad \textit{almost surely.}$$

Proof. We argue by contradiction. Suppose, that $\kappa_{n_i}=K$, $X_{n_i-K}^{n_i}=x_{-K}^0$ for a subsequence n_i. Then a simple frequency count (in the data segment $X_0^{n_i}$ there are less than J_{n_i} occurrences of block x_{-K}^0) yields that

$$P(X_{-K}^0=x_{-K}^0)\leqslant \lim_{n\to\infty}\frac{J_n}{n}=0.$$

The set of sequences that contain a block with zero probability has zero probability and thus Lemma 2 is proved. □

4. Pointwise consistency

Proof of Theorem (A). By Lemma 2, for large n,

$$\begin{aligned}\bigl|g_n(x)-E\bigl(g(X_{n+1})|X_0^n\bigr)\bigr| &= \left|\frac{1}{\lambda_n}\sum_{j=1}^{\lambda_n} g(X_{n-\tau_j^{\kappa_n}(n)+1}) - E\bigl(g(X_{n+1})|X_0^n\bigr)\right|\\
&\leqslant \max_{J=J_n,\ldots,n}\ \max_{k=1,\ldots,K_n}\left|\frac{1}{J}\sum_{j=1}^{J}\bigl[g(X_{n-\tau_j^k(n)+1})-G(X_{-\infty}^{n-\tau_j^k(n)})\bigr]\right|\\
&\quad+\left|\frac{1}{\lambda_n}\sum_{j=1}^{\lambda_n}G(X_{-\infty}^{n-\tau_j^{\kappa_n}(n)})-E\bigl(g(X_{n+1})|X_0^n\bigr)\right|.\end{aligned}$$

Concerning the first term, by (1), (6) and (5),

$$\frac{1}{J}\sum_{j=1}^{J}\bigl[g(X_{n-\tau_j^k(n)+1})-G(X_{-\infty}^{n-\tau_j^k(n)})\bigr]=\frac{1}{J}\sum_{j=0}^{J-1}\bigl[g(\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k+1})-G(\ldots,\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k-1},\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k})\bigr] \tag{7}$$

where $\tilde\tau_j^k$ is evaluated on $\{\widetilde{X}_i^{(k,n,J)}\}_{i=-\infty}^{\infty}$. Since by Lemma 1

$$G(\ldots,\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k-1},\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k})=E\bigl(g(\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k+1})|\ldots,\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k-1},\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k}\bigr),$$

the pair ($\Gamma_j=g(\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k+1})-G(\ldots,\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k-1},\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k}),\mathcal{F}_j=\sigma(X_{-\infty}^{\tilde\tau_j^k})$) forms a martingale difference sequence ($E(\Gamma_j|\mathcal{F}_j)=0$ and Γ_j is measurable with respect to $\mathcal{F}_{j+1}$) for which Azuma's exponential bound (cf. Azuma [4]) yields

$$P\left(\left|\frac{1}{J}\sum_{j=0}^{J-1}\bigl[g(\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k+1})-G(\ldots,\widetilde{X}^{(k,n,J)}_{\tilde\tau_j^k})\bigr]\right|>\epsilon\right)\leqslant 2\,\mathrm{e}^{-\epsilon^2 J/B}$$

for any B such that $\max_{x\in\mathcal{X}}|g(x)|<B$. Now by (7)

$$P\left(\max_{J=J_n,\ldots,n}\ \max_{k=1,\ldots,K_n}\left|\frac{1}{J}\sum_{j=1}^{J}\bigl[g(X_{n-\tau_j^k(n)+1})-G(X_{-\infty}^{n-\tau_j^k(n)})\bigr]\right|>\epsilon\right)\leqslant nK_n2\,\mathrm{e}^{-\epsilon^2 J_n/B}$$

and by assumption $nK_n2\,\mathrm{e}^{-\epsilon^2 J_n/B}$ sums up and the Borel–Cantelli Lemma yields almost sure convergence to zero. Concerning the second term,

$$\left|\frac{1}{\lambda_n}\sum_{j=1}^{\lambda_n}\bigl[G(X_{-\infty}^{n-\tau_j^{\kappa_n}(n)})-E\bigl(G(X_{-\infty}^n)|X_0^n\bigr)\bigr]\right|\to 0\quad\text{almost surely}$$

since κ_n tends to infinity by Lemma 2, $X_{n-\tau_j^{\kappa_n}(n)-\kappa_n+1}^{n-\tau_j^{\kappa_n}(n)} = X_{n-\kappa_n+1}^n$ for $0 \leqslant j \leqslant \lambda_n$, and the conditional expectation $G(\cdot)$ is in fact uniformly continuous on $\mathcal{X}^{*-}$ with respect to $d^*(\cdot,\cdot)$. The proof of Theorem (A) is complete. □

5. Time average performance

If the process does not have continuous conditional expectations then the last step in the proof of Theorem (A) is not valid. It can be carried out for most time instances n by using the typical behaviour of almost every realization $x_{-\infty}^{\infty}$. More specifically, for every $\delta > 0$, the probability of the set of those $x_{-\infty}^0$ for which

$$\bigl|E\bigl(g(X_1)|X_{-k+1}^0 = x_{-k+1}^0\bigr) - G(x_{-\infty}^0)\bigr| < \delta \quad \text{for all } k \geqslant K$$

tends to one as K tends to infinity. The typical behaviour we are after is the statement that most of the times $t = n - \tau_j^{\kappa_n}(n)$ the sequence $T^t x_{-\infty}^t$ belongs to the above mentioned set. While this need not be the case for all n, it is true for most n's and the next lemma makes this precise. For the analysis we will fix a value of κ_n at k.

Define the set of good indexes $M_n(\delta, K) \subseteq \{K-1,\dots,n-1\}$ as

$$M_n(\delta,K) = \bigl\{K-1 \leqslant i \leqslant n-1\colon\ \bigl|E\bigl(g(X_{i+1})|X_{i-k+1}^i\bigr) - G(X_{-\infty}^i)\bigr| < \delta \quad \text{for all } k \geqslant K\bigr\}.$$

We will analyze the behaviour of our algorithm for $\kappa_n = k$ for each $i \leqslant n$ by first dividing up the indices $\{1,2,\dots,n\}$ according to the value of $X_{i-k+1}^i = y_{-k+1}^0$, and considering what happens for each of these.

Let $y_{-k+1}^0 \in \mathcal{X}^k$. Define the set of indexes $I_n^k(y_{-k+1}^0) \subseteq \{k-1,\dots,n-1\}$, where you can find the pattern y_{-k+1}^0, that is,

$$I_n^k(y_{-k+1}^0) = \bigl\{k-1 \leqslant i \leqslant n-1\colon\ X_{i-k+1}^i = y_{-k+1}^0\bigr\}.$$

Define $D_k(i)$ as

$$D_k(i) = \begin{cases} \{\tau_j^k(i)\colon\ \tau_j^k(i) \leqslant i-k+1 \text{ and } 1 \leqslant j \leqslant i+1\} & \text{if } \tau_{J_i}^k(i) \leqslant i-k+1,\\ \emptyset & \text{otherwise.}\end{cases}$$

Let $E_n^k(\delta,K)$ be defined as

$$E_n^k(\delta,K) = \bigl\{0 \leqslant i \leqslant n-1\colon\ \bigl|D_k(i) \cap M_n(\delta,K)\bigr| > (1-\delta^{0.5})|D_k(i)|\bigr\}.$$

If the number of occurrences of y_{-k+1}^0 prior to i was not enough for our algorithm then $D_k(i)$ will be empty. This is rare, and can be expressed as follows: Let

$$F_n^k = \bigl\{0 \leqslant i \leqslant n-1\colon\ D_k(i) = \emptyset\bigr\}.$$

It is immediate that

$$|F_n^k| \leqslant |\mathcal{X}|^k J_n. \tag{8}$$

Lemma 3. *Assume $|M_n(\delta,K)| \geqslant (1-\delta)n$. Then*

$$\bigl|\bigl\{0 \leqslant l \leqslant n-1\colon\ l \notin E_n^k(\delta,K) \text{ and } l \notin F_n^k\bigr\}\bigr| \leqslant \delta^{0.5} n.$$

Proof. Fix δ, K, k and $x \in \mathcal{X}$. Temporarily fix also $y_{-k+1}^0 \in \mathcal{X}^k$. Let $z = |I_n^k(y_{-k+1}^0)|$ and let $k \leqslant i_1 \leqslant i_2 \leqslant \cdots \leqslant i_z$ denote the elements of $I_n^k(y_{-k+1}^0)$. Let $i_j(y_{-k+1}^0)$ be the largest element $i_{j'}$ of $I_n^k(y_{-k+1}^0)$ such that $D_k(i_{j'}) \neq \emptyset$ and

$$\bigl|\bigl\{0 \leqslant l \leqslant n-1\colon\ l \in D_k(i_{j'}) \text{ and } l \notin M_n(\delta,K)\bigr\}\bigr| \geqslant \delta^{0.5}\bigl|D_k(i_{j'})\bigr|.$$

Define S to be the set of these indexes as y_{-k+1}^0 varies over all element $\mathcal{X}^k$. It is clear that if $i, j \in S$, $i \neq j$ then $D_k(i) \cap D_k(j) = \emptyset$ since different blocks y_{-k+1}^0 are involved. It follows from the construction that $\{D_k(i)\}_{i \in S}$ is a disjoint cover of $\{0 \leqslant l \leqslant n-1\colon\ l \notin E_n^k(\delta, K)$ and $l \notin F_n^k\}$. It follows that

$$n\delta \geqslant \bigl|\bigl\{0 \leqslant l \leqslant n-1\colon\ l \notin M_n(\delta, K)\bigr\}\bigr| \geqslant \sum_{i \in S} \bigl|\bigl\{0 \leqslant l \leqslant n-1\colon\ l \in D_k(i) \text{ and } l \notin M_n(\delta, K)\bigr\}\bigr|$$
$$\geqslant \delta^{0.5} \sum_{i \in S} \bigl|D_k(i)\bigr| = \delta^{0.5} \left|\bigcup_{i \in S} D_k(i)\right|.$$

Now

$$\bigl|\bigl\{0 \leqslant l \leqslant n-1\colon\ l \notin E_n^k(\delta, K) \text{ and } l \notin F_n^k\bigr\}\bigr| \leqslant \left|\bigcup_{i \in S} D_k(i)\right| \leqslant \delta^{0.5} n$$

and the proof of Lemma 3 is complete. □

Proof of Theorem (B). Consider

$$\frac{1}{n}\sum_{i=0}^{n-1} \bigl|g_i(x) - E\bigl(g(X_{i+1})|X_0^i\bigr)\bigr|$$
$$\leqslant \frac{|0 - E(g(X_1)|X_0)|}{n} + \frac{1}{n}\sum_{i=1}^{n-1} 1_{\{\kappa_i < K_i\}} + \frac{1}{n}\sum_{i=1}^{n-1} \max_{J=J_i,\ldots,i} \left|\frac{1}{J}\sum_{j=1}^{J} \bigl[g(X_{i-\tau_j^{K_i}(i)+1}) - G(X_{-\infty}^{i-\tau_j^{K_i}(i)})\bigr]\right|$$
$$+ \frac{1}{n}\sum_{i=1}^{n-1} \left|\frac{1}{\lambda_i}\sum_{j=1}^{\lambda_i} G(X_{-\infty}^{i-\tau_j^{K_i}(i)}) - E\bigl(g(X_{i+1})|X_{i-K_i+1}^i\bigr)\right| 1_{\{\kappa_i = K_i\}}$$
$$+ \frac{1}{n}\sum_{i=1}^{n-1} \bigl|E\bigl(g(X_{i+1})|X_{i-K_i+1}^i\bigr) - E\bigl(g(X_{i+1})|X_0^i\bigr)\bigr|.$$

The first term tends to zero. The second term tends to zero since by (8) $|F_n^{K_n}|/n \leqslant |\mathcal{X}|^{K_n} J_n/n \to 0$.

Concerning the third term, by (7) and by Azuma's exponential bound (cf. Azuma [4])

$$\sum_{J=J_i}^{i} P\left(\left|\frac{1}{J}\sum_{j=1}^{J}\bigl[g(\widetilde{X}_{\tilde{\tau}_j^{K_i}+1}^{(K_i,i,J)}) - G(\ldots, \widetilde{X}_{\tilde{\tau}_j^{K_i}-1}^{(K_i,i,J)}, \widetilde{X}_{\tilde{\tau}_j^{K_i}}^{(K_i,i,J)})\bigr]\right| > \epsilon\right) \leqslant 2i\mathrm{e}^{-\epsilon^2 J_i/B}$$

(where B is any real such that $2\max_{x \in \mathcal{X}} |g(x)| < B$) and the right-hand side is summable, hence the Borel–Cantelli Lemma yields almost sure convergence to zero. By Toeplitz lemma the average also converges to zero.

Now we deal with the fourth term. Let $0 < \epsilon < 1$ be arbitrary. Choose the integer d large enough such that $|\mathcal{X}|^{-10(d-1)} < \epsilon$. Let $\delta = \epsilon/d^2$. Let K and N_0 be so large that $|M_n(\delta, K)|/n > (1-\delta)$ for all $n \geqslant N_0$. (There exist such K and N_0 since by the ergodic theorem and the martingale convergence theorem $\lim_{k\to\infty}\lim_{n\to\infty}\frac{|M_n(\delta,k)|}{n} = 1$ almost surely.) Now let $N_1 \geqslant N_0$ be so large that $K_n - d + 2 \geqslant K$ and $|\mathcal{X}|^{10(K_n-d+1)} \geqslant N_0$ for all $n \geqslant N_1$. Assume $n \geqslant N_1$. The sum

$$\frac{1}{n}\sum_{i=1}^{n-1} \left|\frac{1}{\lambda_i}\sum_{j=1}^{\lambda_i} G(X_{-\infty}^{i-\tau_j^{K_i}(i)}) - E\bigl(g(X_{i+1})|X_{i-K_i+1}^i\bigr)\right| 1_{\{\kappa_i = K_i\}}$$

that we are trying to estimate will be divided into blocks according to the value of K_i. In fact only values in the range $[K_n - d + 2, K_n]$ need be considered since the sum up to $|\mathcal{X}|^{10K_n - d+1}$ can be estimated by $|\mathcal{X}|^{10(K_n-d+1)}2\max_{y\in\mathcal{X}}|g(y)|$ and so by our assumption on d, after dividing by n this will be at most $\epsilon 2\max_{y\in\mathcal{X}}|g(y)|$. For i in the range $[|\mathcal{X}|^{10(k-1)}, |\mathcal{X}|^{10k})$ for $K_n - \delta + 2 \leqslant k \leqslant K_n$, and $\kappa_i = K_i$, if $i \in E^{k-1}_{|\mathcal{X}|^{10k}}(\delta, K)$ then we get for more than $(1-\sqrt{\delta})|D_k(i)|$ terms an upper bound of δ while for the rest we may use $2\max_{y\in\mathcal{X}}|g(y)|$. This gives an upper bound of

$$\frac{\delta|D_k(i)| + \sqrt{\delta}|D_k(i)|2\max_{y\in\mathcal{X}}|g(y)|}{|D_k(i)|}.$$

Using Lemma 3 we can estimate the sum over all i in the interval $[|\mathcal{X}|^{10(k-1)}, |\mathcal{X}|^{10k})$ by

$$n\Bigl(\delta + \sqrt{\delta}2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr|\Bigr) + \sqrt{\delta}n2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr|.$$

Dividing by n, we have an upper bound:

$$\delta + \sqrt{\delta}2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr| + \sqrt{\delta}2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr|.$$

The same argument yields the same upper bound for the i's in the range $[|\mathcal{X}|^{10K_n}, n)$.

Summing over k in the range $[K_n - d + 2, K_n + 1]$ yields an upper bound:

$$d\delta + d\sqrt{\delta}2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr| + d\sqrt{\delta}2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr|.$$

Recall that $\sqrt{\delta}d = \sqrt{\epsilon}$ and this yields an upper bound:

$$\epsilon + \sqrt{\epsilon}2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr| + \sqrt{\epsilon}2\max_{y\in\mathcal{X}}\bigl|g(y)\bigr|.$$

Since ϵ was arbitrary, the fourth term tends to zero.

Now we deal with the last term. Since by the martingale convergence theorem, $E(g(X_1)|X^0_{-i}) \to G(X^0_{-\infty})$ almost surely, thus

$$\lim_{i\to\infty}\bigl|E\bigl(g(X_1)|X^0_{-K_i+1}\bigr) - E\bigl(g(X_1)|X^0_{-i}\bigr)\bigr| = 0$$

and applying Breiman's generalized ergodic theorem, cf. Maker [14] (or Algoet [2]),

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\bigl|E\bigl(g(X_{i+1})|X^i_{i-K_i+1}\bigr) - E\bigl(g(X_{i+1})|X^i_0\bigr)\bigr| = 0$$

almost surely and the proof of Theorem (B) is complete. □

6. Weak consistency

Proof of Theorem (C). In order to show that for all ergodic stationary processes our estimate g_n converges in probability we follow the steps in the proof of Theorem (A). The probability that

$$\bigl(\bigl|g_n(x) - E\bigl(g(X_{n+1})|X^n_0\bigr)\bigr| > 3\epsilon\bigr)$$

can be estimated as the sum of the probability of several sets,

$$P\left(\max_{J=J_n,\ldots,n}\max_{k=1,\ldots,K_n}\left|\frac{1}{J}\sum_{j=1}^{J}\bigl[g(X_{n-\tau^k_j(n)+1}) - G(X^{n-\tau^k_j(n)}_{-\infty})\bigr]\right| > \epsilon\right),$$

$$P(\kappa_n < K_n),$$

$$P\bigl(\bigl|E\bigl(g(X_{n+1})|X^n_0\bigr) - E\bigl(g(X_{n+1})|X^n_{n-K_n+1}\bigr)\bigr| > \epsilon\bigr)$$

and

$$P\left(\left|\frac{1}{\lambda_n}\sum_{j=1}^{\lambda_n} G\left(X_{-\infty}^{n-\tau_j^{\kappa_n}(n)}\right) - E\big(g(X_{n+1})|X_{n-\kappa_n+1}^{n}\big)\right| > \epsilon, \kappa_n = K_n\right).$$

For the first, the argument given there suffices. Concerning the second, it tends to zero by Lemma 4 in the Appendix. (Apply it with $A = \{X_{n-K_n+1}^n = x_{n-K_n+1}^n\}$, $D = J_n$. Then sum over all possible $x_{n-K_n+1}^n$ to get that this second probability in question is not greater than $|\mathcal{X}|^{K_n} J_n/n$ which tends to zero.) For the third, it is easy to see that it tends to zero by stationarity and by the martingale convergence theorem which implies that

$$\lim_{n\to\infty} P\left(\left|E\big(g(X_1)|X_{-n}^0\big) - E\big(g(X_1)|X_{-K_n+1}^0\big)\right| > \epsilon\right) = 0.$$

We concentrate on the last probability. Recall the notations from the proof of Theorem (B). The main thing is to show that with probability at least $1-\epsilon$, for n sufficiently large, most of the elements $l \in I_n^{K_n}(X_{-K_n+1}^0)$ are such that $T^l x_{-\infty}^{\infty}$ does not belong to the set

$$\widetilde{M}_n(\epsilon) = \left\{x_{-\infty}^{\infty}\colon \left|E\big(g(X_1)|X_{-k+1}^0 = x_{-k+1}^0\big) - G(x_{-\infty}^0)\right| > \epsilon \text{ for some } k \geqslant K_n\right\}$$

as neither does $T^n x_{-\infty}^{\infty}$ itself. By the martingale convergence theorem, the probability of the set $\widetilde{M}_n(\epsilon)$ tends to zero as n tends to infinity. Let n be so large that this probability in question is less than $\epsilon^2/2$. Let

$$B_n = \left\{x_{-\infty}^{\infty}\colon \left|\left\{l \in I_n^{K_n}(x_{n-K_n+1}^n)\colon x_{-\infty}^{\infty} \in T^{-l}\widetilde{M}_n(\epsilon)\right\}\right| > \epsilon\left|I_n^{K_n}(x_{n-K_n+1}^n)\right|\right\}.$$

The probability of B_n will be evaluated using the ergodic theorem along the orbit of a typical point. Let $x_{-\infty}^{\infty}$ be such a typical orbit and N be a very large number. Fix $y_{-K_n+1}^0$, and note those elements in $I_N^{K_n}(y_{-K_n+1}^0)$ that belong to B_n. We will cover them with disjoint blocks of length K_n, beginning on the right end $N-1$ in the obvious way. These sets (subsets of $I_N^k(y_{-K_n+1}^0)$) we call $C_r(y_{-K_n+1}^0)$ where $r = 1, 2, \ldots$. Formally, let $\cdots < l_2 < l_1$ denote the elements of $I_N^{K_n}(y_{-K_n+1}^0)$. Let $C_0(y_{-K_n+1}^0) = \emptyset$. For $r \geqslant 1$ we define $C_r(y_{-K_n+1}^0)$ recursively. Let l be the largest index such that $l \geqslant n$, $l \notin \bigcup_{r'<r} C_{r'}(y_{-K_n+1}^0)$ and $x_{-\infty}^{\infty} \in T^{-n+l}B_n$. If there is such l then set $C_r(y_{-K_n+1}^0) = \{l-n+K_n-1 \leqslant l_i \leqslant l$ for $i = 1, 2, \ldots\}$. Let $R(y_{-K_n+1}^0)$ be the largest r for which $C_r(y_{-K_n+1}^0)$ is defined. Let

$$I_N\big(\widetilde{M}_n(\epsilon)\big) = \left\{0 \leqslant l \leqslant N-1\colon T^l x_{-\infty}^{\infty} \in \widetilde{M}_n(\epsilon)\right\}.$$

Then by the construction of $C_r(y_{-K_n+1}^0)$, for each $1 \leqslant r \leqslant R(y_{-K_n+1}^0)$,

$$\left|\left\{l \in C_r(y_{-K_n+1}^0)\colon T^l x_{-\infty}^{\infty} \in \widetilde{M}_n(\epsilon)\right\}\right| > \epsilon\left|C_r(y_{-K_n+1}^0)\right|.$$

Since $x_{-\infty}^{\infty}$ is typical, for large N, $|I_N(\widetilde{M}_n(\epsilon))| \leqslant \epsilon^2 N$ and

$$\epsilon^2 N \geqslant \sum_{y_{-K_n+1}^0 \in \mathcal{X}^{K_n}} \sum_{r=1}^{R(y_{-K_n+1}^0)} \left|\left\{l \in C_r(y_{-K_n+1}^0)\colon T^l x_{-\infty}^{\infty} \in \widetilde{M}_n(\epsilon)\right\}\right| \geqslant \epsilon \sum_{y_{-K_n+1}^0 \in \mathcal{X}^{K_n}} \sum_{r=1}^{R(y_{-K_n+1}^0)} \left|C_r(y_{-K_n+1}^0)\right|.$$

Let

$$I_N(B_n) = \{n \leqslant l \leqslant N-1\colon T^{l-n}x_{-\infty}^{\infty} \in B_n\}.$$

But those $n \leqslant l \leqslant N-1$, such that $T^{l-n}x_{-\infty}^{\infty} \in B_n$ are covered by this union – thus

$$\epsilon\left|I_N(B_n)\right| \leqslant \epsilon^2 N$$

and thus

$$P(B_n) = \lim_{N\to\infty} \frac{|I_N(B_n)|}{N} \leqslant \epsilon$$

since $x_{-\infty}^{\infty}$ was typical. The proof of the Theorem is complete. □

Appendix A

Lemma 4. *Let $\{X_n\}$ be stationary and ergodic. For an arbitrary set A measurable with respect to $\sigma(X_0^n)$, the probability of the event*

$$\tilde{A}(n, D) = \left\{ x_{-\infty}^{\infty} \in A\colon \sum_{i=0}^{n-1} 1_A(T^i x_{-\infty}^{\infty}) < D \right\}$$

is not greater than D/n.

Proof. Fix a typical orbit $x_{-\infty}^{\infty}$. Let

$$I_N\big(\tilde{A}(n, D)\big) = \big\{ n \leqslant l \leqslant N-1\colon\ T^l x_{-\infty}^{\infty} \in \tilde{A}(n, D) \big\}.$$

We make a disjoint cover. Let $\ldots, l_2 < l_1$ denote the elements of $I_N(\tilde{A}(n, D))$. Set $E_r = \emptyset$ and for $r = 1, 2, \ldots,$ define E_r recursively. Let l denote the largest element of $I_N(\tilde{A}(n, D))$ such that $l \notin \bigcup_{r'<r} E_{r'}$ if there is such and let

$$E_r = \{l - n \leqslant l_i \leqslant l\colon\ \text{for } i = 1, 2, \ldots\}.$$

Now let R denote the largest r for which E_r has been defined. Since the cover is disjoint, $R(n+1) \leqslant N$. Then clearly,

$$\frac{I_N(\tilde{A}(n, D))}{N} \leqslant \frac{RD}{R(n+1)} \leqslant \frac{D}{(n+1)}$$

and the left-hand side tends to $P(\tilde{A}(n, D))$. The proof of Lemma 4 is complete. □

References

[1] P. Algoet, Universal schemes for prediction, gambling and portfolio selection, Ann. Probab. 20 (1992) 901–941, Correction: Ann. Probab. 23 (1995) 474–478.

[2] P. Algoet, The strong low of large numbers for sequential decisions under uncertainty, IEEE Trans. Inform. Theory 40 (1994) 609–634.

[3] P. Algoet, Universal schemes for learning the best nonlinear predictor given the infinite past and side information, IEEE Trans. Inform. Theory 45 (1999) 1165–1185.

[4] K. Azuma, Weighted sums of certain dependent random variables, Tohoku Math. J. 37 (1967) 357–367.

[5] D.H. Bailey, Sequential Schemes for Classifying and Predicting Ergodic Processes, Ph.D. thesis, Stanford University, 1976.

[6] T.M. Cover, Open problems in information theory, in: 1975 IEEE Joint Workshop on Information Theory, IEEE Press, New York, 1975, pp. 35–36.

[7] R.M. Gray, Probability, Random Processes, and Ergodic Properties, Springer-Verlag, New York, 1988.

[8] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, A Distribution Free Theory of Nonparametric Regression, Springer-Verlag, New York, 2002.

[9] L. Györfi, G. Lugosi, Strategies for sequential prediction of stationary time series, in: M. Dror, P. L'Ecuyer, F. Szidarovszky (Eds.), Modeling Uncertainty an Examination of Stochastic Theory, Methods, and Applications, Kluwer Academic, 2002, pp. 225–248.

[10] L. Györfi, G. Lugosi, G. Morvai, A simple randomized algorithm for consistent sequential prediction of ergodic time series, IEEE Trans. Inform. Theory 45 (1999) 2642–2650.

[11] L. Györfi, G. Morvai, S. Yakowitz, Limits to consistent on-line forecasting for ergodic time series, IEEE Trans. Inform. Theory 44 (1998) 886–892.

[12] S. Kalikow, Random Markov processes and uniform martingales, Israel J. Math. 71 (1990) 33–54.
[13] M. Keane, Strongly mixing g-measures, Invent. Math. 16 (1972) 309–324.
[14] Ph.T. Maker, The ergodic theorem for a sequence of functions, Duke Math. J. 6 (1940) 27–30.
[15] G. Morvai, Estimation of conditional distribution for stationary time series, Ph.D. thesis, Technical University of Budapest, 1994.
[16] G. Morvai, Guessing the output of a stationary binary time series, in: Y. Haitovsky, H.R. Lerche, Y. Ritov (Eds.), Foundations of Statistical Inference, Physika-Verlag, 2003, pp. 207–215.
[17] G. Morvai, B. Weiss, Forecasting for stationary binary time series, Acta Appl. Math. 79 (2003) 25–34.
[18] G. Morvai, S. Yakowitz, P. Algoet, Weakly convergent nonparametric forecasting of stationary time series, IEEE Trans. Inform. Theory 43 (1997) 483–498.
[19] G. Morvai, S. Yakowitz, L. Györfi, Nonparametric inferences for ergodic, stationary time series, Ann. Statist. 24 (1996) 370–379.
[20] D.S. Ornstein, Guessing the next output of a stationary process, Israel J. Math. 30 (1978) 292–296.
[21] D.S. Ornstein, Ergodic Theory, Randomness, and Dynamical Systems, Yale University Press, 1974.
[22] B.Ya. Ryabko, Prediction of random sequences and universal coding, Problems Inform. Transmission 24 (April–June 1988) 87–96.
[23] P.C. Shields, Cutting and stacking: a method for constructing stationary processes, IEEE Trans. Inform. Theory 37 (1991) 1605–1614.
[24] B. Weiss, Single Orbit Dynamics, American Mathematical Society, 2000.