A Note on Prediction for Discrete Time Series

Gusztáv Morvai, Benjamin Weiss *†

Abstract

Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite or countably infinite set $\mathcal{X}$ and that $f(X)$ is a function of the process with finite second moment. Assume that the distribution of the process is otherwise unknown. We construct a sequence of stopping times λ_n along which we will be able to estimate the conditional expectation $E(f(X_{\lambda_n+1})|X_0,\ldots,X_{\lambda_n})$ from the observations $(X_0,\ldots,X_{\lambda_n})$ in a point wise consistent way for a restricted class of stationary and ergodic finite or countably infinite alphabet time series which includes among others all stationary and ergodic finitarily Markovian processes. If the stationary and ergodic process turns out to be finitarily Markovian (in particular, all stationary and ergodic Markov chains are included in this class) then $\lim_{n\to\infty}\frac{n}{\lambda_n}>0$ almost surely. If the stationary and ergodic process turns out to possess finite entropy rate then λ_n is upper bounded by a polynomial, eventually almost surely.

Keywords: Nonparametric estimation, stationary processes
*Mathematics Subject Classifications (2000)*62G05, 60G25, 60G10

1 Introduction

I. Vajda wrote a number of papers on estimating certain parameters and functions from observations [21, 15, 2, 7] and on statistical tests [8, 14, 10, 22]. In

*Gusztáv Morvai is with MTA-BME Stochastics Research Group, 1 Egry József utca , Building H, Budapest,1111, Hungary, e-mail: morvai@math.bme.hu The first author was supported by OTKA Grant No. K75143 and Bolyai János Research Scholarship.

†Benjamin Weiss is with Hebrew University of Jerusalem, Jerusalem 91904 Israel, e-mail: weiss@math.huji.ac.il.

this paper, which we dedicate to his memory, we will consider the particular problem of estimating the conditional expectations by an algorithm which will involve a scheme for discriminating processes.

One of the basic problems in nonparametric estimation is the problem of estimating the conditional probability $P(X_{n+1} = 1|X_0, \ldots, X_n)$ for a binary time series. While there are universal schemes that converge in probability Bailey [1] showed that one cannot estimate this quantity from the data $(X_0, \ldots, X_n)$ such that the difference tends to zero almost surely as n increases, for all stationary and ergodic binary time series (a simpler proof was given later by Ryabko [19]).

For special classes of processes universal point wise schemes can be given. For example, if one knows in advance that the process is Markov of some order , then one can estimate the order (c.f. Csiszár and Shields [5], Csiszár [6]), and using this estimate for the order, one can count empirical averages of blocks with lengths one plus the order and obtain in this way a point wise consistent estimator. In an earlier paper (Morvai and Weiss [16]) we took up the case when it is not known in advance if the process is Markov or not. To circumvent the negative results we used the idea of restricting the estimation to stopping times and treated processes whose conditional distribution is almost surely continuous. This class includes all Markov processes with arbitrary order and the much wider class of finitarily Markovian processes.

Compared to the earlier results we were able to show that in this case the sequence of stopping times grows less rapidly and if it is indeed finitarily Markov then it will have positive density, while if the process has finite entropy then the growth rate is upper bounded by a polynomial, eventually almost surely.

While we did include the possibility of a countably infinite alphabet we assumed that the function, $f(x)$,whose conditional probability we are estimating is bounded. In this note we will give a somewhat simpler scheme and allow an unbounded function of the process, $f(X_i)$, as long as it has a finite second moment. Our main result is the construction of a sequences of stopping times λ_n and corresponding estimator f_n such that for any process with almost sure conditional probabilities,

$$\lim_{n\to\infty} \left| f_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n}) \right| = 0.$$

The parameters defining these stopping times may be chosen in such a fashion that whenever the stationary and ergodic time series $\{X_n\}$ has finite entropy rate then λ_n grows no faster than a polynomial in n.

If the stationary and ergodic time series $\{X_n\}$ turns out to be finitarily Markovian then

$$\lim_{n\to\infty} \frac{\lambda_n}{n} < \infty \text{ almost surely.}$$

Moreover, if the stationary and ergodic time series $\{X_n\}$ turns out to be independent and identically distributed then $\lambda_n = \lambda_{n-1}+1$ eventually almost surely.

For related problems see Ryabko [20].

2 Results

Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) 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 $X_m^n = (X_m, \ldots, X_n)$, where $m \le n$. Note that if $m > n$ then X_m^n is the empty string and $\mathcal{X}^0$ is a set that contains exactly the empty string $\emptyset$.

For $k \ge 1$, let $1 \le l_k \le k$ be a nondecreasing unbounded sequence of integers, that is, $1 = l_1 \le l_2 \ldots$ and $\lim_{k\to\infty} l_k = \infty$.
Define auxiliary stopping times as follows. Set $\zeta_0 = 0$. For $n = 1, 2, \ldots$, let

$$\zeta_n = \zeta_{n-1} + \min\{t > 0 : X_{\zeta_{n-1}-(l_n-1)+t}^{\zeta_{n-1}+t} = X_{\zeta_{n-1}-(l_n-1)}^{\zeta_{n-1}}\}. \tag{1}$$

Among other things, using ζ_n and l_n we can define a very useful process $\{\tilde{X}_n\}_{n=-\infty}^{0}$ as a function of X_0^{∞} as follows. Let $J(n) = \min\{j \ge 1 :\ l_{j+1} > n\}$ and define

$$\tilde{X}_{-i} = X_{\zeta_{J(i)}-i} \text{ for } i \ge 0. \tag{2}$$

As we will see in the next lemma, the $\{\tilde{X}_n\}_{n=-\infty}^{0}$ has the same distribution as the original process.

Lemma 1 *The time series $\{\tilde{X}_n\}_{n=-\infty}^{0}$ and $\{X_n\}_{n=-\infty}^{0}$ have identical distribution and for $j = 0, 1, 2, \ldots$ the random variables X_{ζ_j+1} are identically distributed.*

Proof: For all $k \geq 1$ and $1 \leq i \leq k$ define $\hat{\zeta}_0^k = 0$ and

$$\hat{\zeta}_i^k = \hat{\zeta}_{i-1}^k - \min\{t > 0 : X_{\hat{\zeta}_{i-1}^k - (l_{k-i+1}-1)-t}^{\hat{\zeta}_{i-1}^k - t} = X_{\hat{\zeta}_{i-1}^k - (l_{k-i+1}-1)}^{\hat{\zeta}_{i-1}^k}\}.$$

Let T denote the left shift operator, that is, $(Tx_{-\infty}^{\infty})_i = x_{i+1}$. It is easy to see that if $\zeta_k(x_{-\infty}^{\infty}) = l$ then $\hat{\zeta}_k^k(T^l x_{-\infty}^{\infty}) = -l$.
Now the statement follows from stationarity and the fact that for $k \geq 0$, $m \geq 0$, $n \geq 0$, $x_{-n}^m \in \mathcal{X}^{m+n+1}$, $l \geq 0$,

$$T^l\{X_{\zeta_k - n}^{\zeta_k + m} = x_{-n}^m, \zeta_k = l\} = \{X_{-n}^m = x_{-n}^m, \hat{\zeta}_k^k(X_{-\infty}^0) = -l\}. \tag{3}$$

The proof of Lemma 1 is complete.

For convenience let $p(x_{-k+1}^0)$ and $p(y|x_{-k+1}^0)$ denote the distribution $P(X_{-k+1}^0 = x_{-k+1}^0)$ and the conditional distribution $P(X_1 = y|X_{-k+1}^0 = x_{-k+1}^0)$, respectively. Note that $P(X_{t+1}^t = \emptyset) = 1$, $P(X_1 = y|X_1^0 = \emptyset) = P(X_1 = y)$.

Definition 1 *For some $0 \leq k$ and $w_{-k+1}^0 \in \mathcal{X}^k$ We say that w_{-k+1}^0 is a memory word if $p(w_{-k+1}^0) > 0$ and for all $i \geq 1$, all $y \in \mathcal{X}$, all $z_{-k-i+1}^{-k} \in \mathcal{X}^i$*

$$p(y|w_{-k+1}^0) = p(y|z_{-k-i+1}^{-k}, w_{-k+1}^0)$$

provided $p(z_{-k-i+1}^{-k}, w_{-k+1}^0, y) > 0$. If no proper suffix of w is a memory word then w is called a minimal memory word.

Define the set $\mathcal{W}_k$ of those memory words w_{-k+1}^0 with length k, that is,

$$\mathcal{W}_k = \{w_{-k+1}^0 \in \mathcal{X}^k : \ w_{-k+1}^0 \text{ is a memory word}\}.$$

Let

$$\mathcal{W}^* = \bigcup_{k=0}^{\infty} \mathcal{W}_k.$$

Definition 2 *For a stationary time series* $\{X_n\}$ *the (random) length* $K(X_{-\infty}^0)$ *of the memory of the sample path* $X_{-\infty}^0$ *is the smallest possible* $0 \leq K < \infty$ *such that for all* $i \geq 1$*, all* $y \in \mathcal{X}$*, all* $z_{-K-i+1}^{-K} \in \mathcal{X}^i$

$$p(y|X_{-K+1}^0) = p(y|z_{-K-i+1}^{-K}, X_{-K+1}^0)$$

provided $p(z_{-K-i+1}^{-K}, X_{-K+1}^0, y) > 0$*, and* $K(X_{-\infty}^0) = \infty$ *if there is no such* K*.*

Remark 1 *For stationary and ergodic time series* $\{X_n\}$*,* $K(x_{-\infty}^0)$ *is the smallest* $k \geq 0$ *such that* $x_{-k+1}^0 \in \mathcal{W}_k$ *and* $K(x_{-\infty}^0) = \infty$ *if there is no such* k*.*

In order to estimate $K(\tilde{X}_{-\infty}^0)$ we need to define some explicit statistics. Define

$$\Delta^{r,k} = \sup_{1 \leq i} \sup_{\{z_{r-k-i+1}^{r-k} \in \mathcal{X}^i, x \in \mathcal{X} : p(z_{r-k-i+1}^{r-k}, X_{r-k+1}^r, x) > 0\}} \left| p(x|X_{r-k+1}^r) - p(x|z_{-k-i+1}^{-k}, X_{r-k+1}^r) \right|$$

and

$$\Gamma^k(\tilde{X}_{-k+1}^0) = \sup_{1 \leq i} \sup_{\{z_{-k-i+1}^{-k} \in \mathcal{X}^i, x \in \mathcal{X} : p(z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^0, x) > 0\}} \left| p(x|\tilde{X}_{-k+1}^0) - p(x|z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^0) \right|.$$

Let us agree that if the set over which the sup is taken is empty then the sup is zero.

We need to define an empirical version of this based on the observation of a finite data segment X_0^n. To this end first define the empirical version of the conditional probability as

$$\hat{p}_n(x|w_{-k+1}^0) = \frac{\left(\#\{k-1 \leq t \leq n-1 : X_{t-k+1}^{t+1} = (w_{-k+1}^0, x)\} - 1\right)^+}{\left(\#\{k-1 \leq t \leq n-1 : X_{t-k+1}^t = w_{-k+1}^0\} - 1\right)^+}$$

where $\frac{0}{0} = 0$.
These empirical distributions, as well as the sets we are about to introduce are functions of X_0^n, but we suppress the dependence to keep the notation manageable.

For a fixed $0 < \gamma < 1$ let $\mathcal{L}_k^n$ denote the set of strings with length $k+1$ which appear more than $n^{1-\gamma}$ times in X_0^n. That is,

$$\mathcal{L}_k^n = \{x_{-k}^0 \in \mathcal{X}^{k+1} : \#\{k \le t \le n-1 : X_{t-k}^t = x_{-k}^0\} > n^{1-\gamma} + 1\}.$$

Define

$$\hat{\Delta}_n^{r,k}(X_0^n) = \max_{1 \le i \le n} \max_{(z_{-k-i+1}^{-k}, X_{r-k+1}^r, x) \in \mathcal{L}_{k+i}^n} \left| \hat{p}_n(x|X_{r-k+1}^r) - \hat{p}_n(x|z_{-k-i+1}^{-k}, X_{r-k+1}^r) \right|.$$

Let us agree by convention that if the smallest of the sets over which we are maximizing is empty then $\hat{\Delta}_n^{r,k} = 0$.

Observe, that by ergodicity, the ergodic theorem implies that almost surely the empirical distributions $\hat{p}_n$ converge to the true distributions p and so for any r, k,

$$\liminf_{n\to\infty} \hat{\Delta}_n^{r,k} \ge \Delta^{r,k} \text{ almost surely.}$$

Finally, define the empirical version of Γ^k as follows:

$$\hat{\Gamma}_n^k(X_0^n) = I_{\{k \le l_{(\max\{j:\zeta_j \le n\})}\}} \hat{\Delta}_n^{\max\{\zeta_j \le n : j=0,1,2,\ldots\},k}.$$

Since

$$\liminf_{n\to\infty} \hat{\Delta}_n^{r,k} \ge \Delta^{r,k} \text{ almost surely.}$$

thus

$$\liminf_{n\to\infty} \hat{\Gamma}_n^k \ge \Gamma^k \text{ almost surely.} \tag{4}$$

We define an estimate χ_n for $K(\tilde{X}_{-\infty}^0)$ from samples X_0^n as follows. Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. Set $\chi_0 = 0$, and for $n \ge 1$ let

$$\chi_n(X_0^n) = \min\{0 \le k \le l_{(\max\{j:\zeta_j \le n\})} : \hat{\Gamma}_n^k \le n^{-\beta} \text{ or } k = l_{(\max\{j:\zeta_j \le n\})}\}. \quad (5)$$

Here the idea is (cf. the proof of Theorem 1) that if $K(\tilde{X}_{-\infty}^0) < \infty$ then χ_n will be equal to $K(\tilde{X}_{-\infty}^0)$ eventually and if $K(\tilde{X}_{-\infty}^0) = \infty$ then $\chi_n \to \infty$.

Now we define the sequence of stopping times λ_n along which we will be able to estimate. Set $\lambda_0 = \zeta_0$, and for $n \ge 1$ if $\zeta_j \le \lambda_{n-1} < \zeta_{j+1}$ then put

$$\lambda_n = \min\{t > \lambda_{n-1} : X_{t-\chi_t+1}^t = X_{\zeta_j-\chi_t+1}^{\zeta_j}\} \quad (6)$$

and

$$\kappa_n = \chi_{\lambda_n}. \quad (7)$$

Observe that if $\zeta_j \le \lambda_{n-1} < \zeta_{j+1}$ then $\zeta_j \le \lambda_{n-1} < \lambda_n \le \zeta_{j+1}$. If $\chi_{\lambda_{n-1}+1} = 0$ then $\lambda_n = \lambda_{n-1} + 1$. Note that λ_n is a stopping time and κ_n is our estimate for $K(\tilde{X}_{-\infty}^0)$ from samples $X_0^{\lambda_n}$.

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

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

Let $f : \mathcal{X} \to (-\infty, \infty)$ be arbitrary. Define the function $F : \mathcal{X}^{*-} \to (-\infty, \infty)$ as

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

E.g. if $f(x) = 1_{\{x=z\}}$ for a fixed $z \in \mathcal{X}$ then $F(y_{-\infty}^0) = P(X_1 = z|X_{-\infty}^0 = y_{-\infty}^0)$. If $\mathcal{X}$ is a finite or countably infinite subset of the reals and $f(x) = x$ then $F(y_{-\infty}^0) = E(X_1|X_{-\infty}^0 = y_{-\infty}^0)$.

One denotes the nth auxiliary estimate of $E(f(X_{\zeta_n+1})|X_0^{\zeta_n})$ from samples $X_0^{\zeta_n}$ by g_n, and defines it to be

$$g_n = \frac{1}{n}\sum_{j=0}^{n-1} f(X_{\zeta_j+1}). \quad (8)$$

One denotes the nth estimate of $E(f(X_{\lambda_n+1})|X_0^{\lambda_n})$ from samples $X_0^{\lambda_n}$ by f_n, and defines it to be $f_0 = g_0$ and for $n > 0$

$$f_n(X_0^{\lambda_n}) = g_{\min\{t \ge 0 : \zeta_t \le \lambda_n < \zeta_{t+1}\}}. \quad (9)$$

Note that the same estimate will be used for a while and this will help us to get rid of the boundedness condition in Morvai and Weiss in [16].
Define the distance $d^*(\cdot,\cdot)$ on $\mathcal{X}^{*-}$ as follows. For $x^0_{-\infty}$, $y^0_{-\infty} \in \mathcal{X}^{*-}$ let

$$d^*(x^0_{-\infty}, y^0_{-\infty}) = \sum_{i=0}^{\infty} 2^{-i-1} 1_{\{x_{-i} \neq y_{-i}\}}. \tag{10}$$

Definition 3 *We say that $F(X^0_{-\infty})$ is almost surely continuous if for some set $C \subseteq \mathcal{X}^{*-}$ which has probability one the function $F(X^0_{-\infty})$ restricted to this set C is continuous with respect to metric $d^*(\cdot,\cdot)$.*

The processes with almost surely continuous conditional expectation generalizes the processes for which it is actually continuous, cf. Kalikow [12] and Keane [13]. The stationary finitarily Markovian processes with $E(|f(X_1)|) < \infty$ are included in the class of stationary processes with almost surely continuous $E(f(X_1)|X^0_{-\infty})$.

Note that Ryabko [19], and Györfi, Morvai, Yakowitz [9] showed that one cannot estimate $P(X_{n+1} = 1|X_0^n)$ for all n in a pointwise consistent way even for the class of all stationary and ergodic binary finitarily Markovian time series.

The entropy rate H associated with a stationary finite or countably infinite alphabet time series $\{X_n\}$ is defined as

$$H = \lim_{n\to\infty} \frac{-1}{n+1} \sum_{x^0_{-n} \in \mathcal{X}^{n+1}} p_n(x^0_{-n}) \log_2 p_n(x^0_{-n}).$$

We note that the entropy rate of a stationary finite alphabet time series is finite. For details cf. Cover, Thomas [4], pp. 63-64.

Before stating the theorem let us recapitulate what we have done. For a fixed sequence $1 = l_1 \leq l_2, \ldots, l_n \to \infty$ and for any stationary and ergodic process $\{X_n\}$ we define a sequence of stopping times $\{\zeta_n\}$ in such a way that $\{X_{\zeta_n - i}\}_{i=0}^{\infty}$ converges to the process $\{\tilde{X}_{-i}\}_{i=0}^{\infty}$ which has the same distribution as $\{X_{-i}\}_{i=0}^{\infty}$. If f is a real valued function defined on the alphabet $\mathcal{X}$ of the process $\{X_n\}$ then the estimators of the conditional expectation

of $E(f(X_{\zeta_n+1})|X_0^{\zeta_n})$ are defined in (8). Now fix $0 < \beta, \gamma < 1$ such that $2\beta + \gamma < 1$ and define the estimators χ_n for the memory length $K(\tilde{X}^0_{-\infty})$ as in (5). In turn these are used to define the stopping times λ_n as in (6) and finally the estimators $f_n(X_0^{\lambda_n})$ are defined in (9).

Theorem 1 *Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite or countably infinite set $\mathcal{X}$. Assume that f is a real valued function defined on $\mathcal{X}$ such that*

$$E\left(f(X_1)^2\right) < \infty.$$

1. *If the conditional expectation $F(X^0_{-\infty}) = E(f(X_1)|X^0_{-\infty})$ is almost surely continuous then for the stopping times λ_n (see (6)) and the estimators f_n (see (9)) we have*

$$\lim_{n\to\infty} f_n = F(\tilde{X}^0_{-\infty}) \quad \textit{and} \quad \lim_{n\to\infty} \left| f_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n}) \right| = 0$$

   almost surely.

2. *The l_n may be chosen in such a fashion that whenever the stationary and ergodic time series $\{X_n\}$ has finite entropy rate then the stopping times λ_n grow no faster than a polynomial in n.*

3. *If the stationary and ergodic time series $\{X_n\}$ turns out to be finitarily Markovian then*

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = \frac{1}{p(\tilde{X}^0_{-K(\tilde{X}^0_{-\infty})+1})} < \infty \quad \textit{almost surely}$$

   where $K(\tilde{X}^0_{-\infty})$ is the length of the memory word. Moreover, if the stationary and ergodic time series $\{X_n\}$ turns out to be independent and identically distributed then $\lambda_n = \lambda_{n-1} + 1$ eventually almost surely.

Proof:
Step 1. *We show that $P(\chi_n = K(\tilde{X}^0_{-\infty})$ eventually $|K(\tilde{X}^0_{-\infty}) < \infty) = 1$ and $P(\lim_{n\to\infty} \chi_n = \infty | K(\tilde{X}^0_{-\infty}) = \infty) = 1$.*
By Lemma 1, $\{\tilde{X}_n\}^0_{n=-\infty}$ is stationary and ergodic with the same distribution as $\{X_n\}^0_{n=-\infty}$. We may assume that the sample path $\tilde{X}^0_{-\infty}$ is such that all

finite blocks that appear have positive probability. It is immediate that if $K(\tilde{X}^0_{-\infty}) < \infty$ then for all $k \geq K(\tilde{X}^0_{-\infty})$, $\Gamma^k = 0$ and $\Gamma^{(K(\tilde{X}^0_{-\infty})-1)} > 0$ (otherwise the length of the memory would be not greater than $K(\tilde{X}^0_{-\infty})-1$). If $K(\tilde{X}^0_{-\infty}) = \infty$ then $\Gamma^k > 0$ for all k, (otherwise $K(\tilde{X}^0_{-\infty})$ would be finite). Thus by (4) if $K(\tilde{X}^0_{-\infty}) = \infty$ then $\chi_n \to \infty$ and if $K(\tilde{X}^0_{-\infty}) < \infty$ then $\chi_n \geq K(\tilde{X}^0_{-\infty})$ eventually almost surely. We have to show that $\chi_n \leq K(\tilde{X}^0_{-\infty})$ eventually almost surely provided that $K(\tilde{X}^0_{-\infty}) < \infty$.
Fix now $k < n$. We will estimate the probability of the undesirable event as follows:

Set $\psi^+_{l,k,0} = 0$, $\psi^-_{l,k,0} = 0$ and for $i > 0$ define

$$\psi^+_{l,k,i} = \psi^+_{l,k,i-1} + \min\{t > 0 : X^{l+\psi^+_{l,k,i-1}+t}_{l+\psi^+_{l,k,i-1}-k+1+t} = X^{l+\psi^+_{l,k,i-1}}_{l+\psi^+_{l,k,i-1}-k+1}\}$$

and

$$\psi^-_{l,k,i} = \psi^-_{l,k,i-1} + \min\{t > 0 : X^{l-\psi^-_{l,k,i-1}-t}_{l-\psi^-_{l,k,i-1}-k+1-t} = X^{l-\psi^-_{l,k,i-1}}_{l-\psi^-_{l,k,i-1}-k+1}\}.$$

For a given $0 \leq k < n$, $k-1 \leq l \leq n-1$ assume that $X^{l+1}_{l-k+1} = w^0_{-k+1}x$ and w^0_{-k+1} is a memory word. Since w^0_{-k+1} is a memory word, by Lemma 1 in [17] for any $i, j \geq 1$,

$$X_{l-\psi^-_{l,k,i}+1}, \ldots, X_{l-\psi^-_{l,k,1}+1}, X_{l+\psi^+_{l,k,1}+1}, \ldots, X_{l+\psi^+_{l,k,j}+1}$$

are conditionally independent and identically distributed random variables given $X^l_{l-k+1} = w^0_{-k+1}, X_{l+1} = x$, where the identical distribution is $p(\cdot|w^0_{-k+1})$. By Hoeffding's inequality for sums of bounded independent random variables (cf. Lemma 2 in the Appendix) the probability that

$$\left| \frac{\sum_{h=1}^i 1_{\{X_{l-\psi^-_{l,k,h}+1}=x\}} + \sum_{h=1}^j 1_{\{X_{l+\psi^+_{l,k,h}+1}=x\}}}{i+j} - p(x|w^0_{l-k+1}) \right|$$

is greater than $0.5n^{-\beta}$ is not greater than $2e^{-0.5n^{-2\beta}(i+j)}$ given the condition $X_{l-k+1}^{l+1} = w_{-k+1}^0 x$. Multiplying both sides by $P(X_{l-k+1}^{l+1} = w_{-k+1}^0 x)$ and summing over all possible memory words w_{-k+1}^0 and x we get that

$$\begin{aligned} & P(K(X_{-\infty}^{l}) \le k, X_{l-k+1}^{l+1} \in \mathcal{L}_{k+1}^n, \\ & \quad \Big| \frac{\sum_{h=1}^{i} 1_{\{X_{l-\psi_{l,k,h}^-+1} = X_{l+1}\}} + \sum_{h=1}^{j} 1_{\{X_{l+\psi_{l,k,h}^+ +1} = X_{l+1}\}}}{i+j} \\ & \quad - p(X_{l+1}|X_{l-k+1}^{l})\Big| > n^{-\beta}/2) \\ \le \; & 2e^{-0.5n^{-2\beta}(i+j)}. \end{aligned}$$

Summing over all pairs (k,l) such that $0 \le k < n$ and all $k-1 \le l \le n-1$ and over all pairs (i,j) such that $i \ge 0$, $j \ge 0$, $i+j \ge \lceil n^{1-\gamma} \rceil$ we get that

$$\begin{aligned} & P\,(\text{ For some } 0 \le k < n, k-1 \le l \le n-1 : \\ & \quad X_{l-k+1}^{l+1} \in \mathcal{L}_{k+1}^n, K(X_{-\infty}^{l}) \le k, \\ & \quad \left|\hat{p}_n(X_{l+1}|X_{l-k+1}^{l}) - p(X_{l+1}|X_{l-k+1}^{l})\right| > n^{-\beta}/2\Big) \\ \le \; & n^2 \sum_{h=\lceil n^{1-\gamma} \rceil}^{\infty} h 2e^{-0.5n^{-2\beta}h}. \end{aligned}$$

Now

$$\begin{aligned} & P(I_{\{k \le l_{(\max\{j:\zeta_j \le n\})}\}} \hat{\Delta}_n^{\max\{\zeta_j \le n: j=0,1,2,\ldots\},k} > n^{-\beta}, K(\tilde{X}_{-\infty}^{0}) \le k) \le \\ & \quad P(\max_{0 \le k < \infty} \max_{w_{-k+1}^0 \in \mathcal{W}_k} \max_{1 \le i \le n} \max_{(z_{-k-i+1}^{-k}, w_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \\ & \quad \left|\hat{p}_n(x|w_{-k+1}^0) - \hat{p}_n(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right| > n^{-\beta}) \le \\ & \quad \sum_{k=0}^{n-1} P(\max_{w_{-k+1}^0 \in \mathcal{W}_k} \max_{1 \le i \le n} \max_{(z_{-k-i+1}^{-k}, w_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \\ & \quad \left|\hat{p}_n(x|w_{-k+1}^0) - \hat{p}_n(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right| > n^{-\beta}) \\ & \quad \le \sum_{k=0}^{n-1} \sum_{i=1}^{n} P(\max_{w_{-k+1}^0 \in \mathcal{W}_k} \max_{(z_{-k-i+1}^{-k}, w_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \\ & \quad \left|\hat{p}_n(x|w_{-k+1}^0) - \hat{p}_n(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right| > n^{-\beta}) \\ & \quad \le \sum_{k=0}^{n-1} \sum_{i=1}^{n} P(\max_{w_{-k+1}^0 \in \mathcal{W}_k} \max_{(z_{-k-i+1}^{-k}, w_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \end{aligned}$$

$$
\begin{aligned}
&\left|\hat{p}_n(x|w^0_{-k+1}) - p(x|w^0_{-k+1})\right| > n^{-\beta}/2) \\
&+\sum_{k=0}^{n-1}\sum_{i=1}^{n} P(\max_{w^0_{-k+1}\in\mathcal{W}_k}\ \max_{(z^{-k}_{-k-i+1},w^0_{-k+1},x)\in\mathcal{L}^n_{k+i}} \\
&\left|p(x|z^{-k}_{-k-i+1},w^0_{-k+1}) - \hat{p}_n(x|z^{-k}_{-k-i+1},w^0_{-k+1})\right| \\
&> n^{-\beta}/2) \\
&\leq 2n^4 \sum_{h=\lceil n^{1-\gamma}\rceil}^{\infty} h2e^{\frac{-n^{-2\beta}h}{2}}.
\end{aligned}
$$

Now we give an upper bound on the sum on the right hand side. Observe that $he^{\frac{-n^{-2\beta}h}{2}}$ is monotone decreasing in h as soon as the derivative

$$e^{\frac{-n^{-2\beta}h}{2}} - h0.5n^{-2\beta}he^{\frac{-n^{-2\beta}h}{2}}$$

is negative for $h > n^{1-\gamma}$ which is the case for $n > 2^{\frac{1}{(1-\gamma-2\beta)}}$. Using this fact, we bound the sum by the integral

$$\sum_{h=\lceil n^{1-\gamma}\rceil}^{\infty} he^{\frac{-n^{-2\beta}h}{2}} \leq \int_{n^{1-\gamma}}^{\infty} he^{\frac{-n^{-2\beta}h}{2}}dh.$$

Integrating by parts we get that

$$
\begin{aligned}
\int_{n^{1-\gamma}}^{\infty} he^{\frac{-n^{-2\beta}h}{2}}dh &= \left[h\frac{-1}{\frac{n^{-2\beta}}{2}}e^{\frac{-n^{-2\beta}h}{2}}\right]_{n^{1-\gamma}}^{\infty} - \int_{n^{1-\gamma}}^{\infty}\frac{-1}{\frac{n^{-2\beta}}{2}}e^{\frac{-n^{-2\beta}h}{2}}dh \\
&= \frac{n^{1-\gamma}}{\frac{n^{-2\beta}}{2}}e^{\frac{-n^{-2\beta}n^{1-\gamma}}{2}} - \left[\frac{1}{\left(\frac{n^{-2\beta}}{2}\right)^2}e^{\frac{-n^{-2\beta}h}{2}}\right]_{n^{1-\gamma}}^{\infty} \\
&= \frac{n^{1-\gamma}}{\frac{n^{-2\beta}}{2}}e^{\frac{-n^{1-\gamma-2\beta}}{2}} + \frac{1}{\left(\frac{n^{-2\beta}}{2}\right)^2}e^{\frac{-n^{1-\gamma-2\beta}}{2}} \\
&= \left(2n^{1-\gamma+2\beta} + 4n^{4\beta}\right)e^{\frac{-n^{1-\gamma-2\beta}}{2}} \\
&\leq \left(2n^2 + 4n^2\right)e^{\frac{-n^{1-\gamma-2\beta}}{2}}
\end{aligned}
$$

since by assumption $0 < \gamma < 1$ and $0 < \beta < \frac{1-\gamma}{2}$.

The right hand side is summable provided $2\beta + \gamma < 1$ and the Borel-Cantelli Lemma yields that

$$I_{\{K(\tilde{X}^0_{-\infty})\leq k\}} I_{\{k\leq l_{(\max\{j:\zeta_j\leq n\})}\}} \hat{\Delta}_n^{(\max_{j:\zeta_j\leq n}\zeta_j),k} \leq n^{-\beta}$$

eventually almost surely. Since

$$I_{\{k\leq l_{(\max\{j:\zeta_j\leq n\})}\}} = 1$$

eventually almost surely thus

$$I_{\{K(\tilde{X}^0_{-\infty})\leq k\}} \hat{\Gamma}_n^k(X_0^n) \leq n^{-\beta}$$

eventually almost surely. Thus $\chi_n \leq k$ eventually almost surely on $K(\tilde{X}^0_{-\infty}) = k$.

Step 2. *We show that* $g_n \to F(\tilde{X}^0_{-\infty})$ *almost surely.*
Recalling (8) we can write

$$g_n = \frac{1}{n}\sum_{j=0}^{n-1}[f(X_{\zeta_j+1}) - E(f(X_{\zeta_j+1})|X_{-\infty}^{\zeta_j})] + \frac{1}{n}\sum_{j=0}^{n-1} E(f(X_{\zeta_j+1})|X_{-\infty}^{\zeta_j}) \quad (11)$$

Consider the first term and observe that $\{\Theta_j = f(X_{\zeta_j+1}) - E(f(X_{\zeta_j+1})|X_{-\infty}^{\zeta_j})\}$ is a sequence of orthogonal random variables with $E\Theta_j = 0$ and, by Lemma 1, X_{ζ_j+1} has the same distribution as X_1. Now by Theorem 3.2.2 in [18] (cf. Lemma 3 in the Appendix),

$$\frac{1}{k}\sum_{j=0}^{k-1}\Theta_j \to 0 \ \text{ almost surely.}$$

Now we deal with the second term. For arbitrary $j \geq 0$, by (7) and (6) and the construction in (2),

$$X_{\zeta_j-l_j+1}^{\zeta_j} = \tilde{X}^0_{-l_j+1} \ \text{ and } \ \lim_{j\to\infty} d^*(\tilde{X}^0_{-\infty}, X_{-\infty}^{\zeta_j}) = 0 \ \text{ almost surely.} \quad (12)$$

By Lemma 1 and the almost sure continuity of $F(\cdot)$, for some set $C \subseteq \mathcal{X}^{*-}$ with full measure, $F(\cdot)$ is continuous on C and

$$\tilde{X}^0_{-\infty} \in C, X^n_{-\infty} \in C \ \text{ for all } n \geq 0 \text{ almost surely.} \quad (13)$$

By the continuity of $F(\cdot)$ on the set C and (12),

$$E(f(X_{\zeta_j+1})|X_{-\infty}^{\zeta_j}) = F(X_{-\infty}^{\zeta_j}) \to F(\tilde{X}_{-\infty}^0)$$

and $g_n \to F(\tilde{X}_{-\infty}^0)$ almost surely.

Step 3. *We prove the first part of Theorem 1.*

By (9) and the fact that $\zeta_n \to \infty$ Step 2 yields immediately $f_n \to F(\tilde{X}_{-\infty}^0)$ almost surely. What remains to be proven is that

$$E(f(X_{\lambda_j+1})|X_0^{\lambda_j}) \to F(\tilde{X}_{-\infty}^0).$$

If $K(\tilde{X}_{-\infty}^0) < \infty$ then by Step 1, $\chi_n = K(\tilde{X}_{-\infty}^0)$ eventually and by (1), (2), (6) and Lemma 1, eventually,

$$E(f(X_{\lambda_j+1})|X_0^{\lambda_j}) = E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) = F(\tilde{X}_{-\infty}^0).$$

We may deal with the case when $K(\tilde{X}_{-\infty}^0) = \infty$ and by Step 1, $\chi_n \to \infty$. For arbitrary $j \geq 0$, by (7) and (6) and the construction in (2),

$$X_{\lambda_j-\kappa_j+1}^{\lambda_j} = \tilde{X}_{-\kappa_j+1}^0 \text{ and } \lim_{j\to\infty} d^*(\tilde{X}_{-\infty}^0, X_{-\infty}^{\lambda_j}) = 0 \text{ almost surely.} \quad (14)$$

By Lemma 1 and the almost sure continuity of $F(\cdot)$, for some set $C \subseteq \mathcal{X}^{*-}$ with full measure, $F(\cdot)$ is continuous on C and

$$\tilde{X}_{-\infty}^0 \in C, X_{-\infty}^n \in C \text{ for all } n \geq 0 \text{ almost surely.} \quad (15)$$

By the continuity of $F(\cdot)$ on the set C and (14),

$$E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) = F(X_{-\infty}^{\lambda_j}) \to F(\tilde{X}_{-\infty}^0).$$

Define the random neighborhood $\mathcal{N}_j(X_0^{\lambda_j})$ of $X_0^{\lambda_j}$ depending on the random data segment $X_0^{\lambda_j}$ itself as

$$\mathcal{N}_j(X_0^{\lambda_j}) = \{z_{-\infty}^0 \in \mathcal{X}^{*-} \ : \ z_{-\kappa_j+1} = X_{\lambda_j-\kappa_j+1}, \ldots, z_0 = X_{\lambda_j}\}.$$

Note that by (1), (2), (7) and (6), $\tilde{X}_{-\infty}^0 \in \mathcal{N}_j(X_0^{\lambda_j})$ and by (15) and the continuity of $F(\cdot)$ on the set C, and since $\kappa_j \to \infty$, by (12), almost surely,

$$\lim_{j\to\infty} \left|E(f(X_{\lambda_j+1})|X_0^{\lambda_j}) - F(\tilde{X}_{-\infty}^0)\right| =$$
$$\lim_{j\to\infty} \left|E\{F(X_{-\infty}^{\lambda_j})|X_0^{\lambda_j}\} - F(\tilde{X}_{-\infty}^0)\right| \leq$$
$$\lim_{j\to\infty} \sup_{y_{-\infty}^0, z_{-\infty}^0 \in \mathcal{N}_j(X_0^{\lambda_j}) \bigcap C} |F(y_{-\infty}^0) - F(z_{-\infty}^0)| = 0.$$

Step 4. *We prove the second part of Theorem 1.*
Now we assume that the stationary and ergodic finite or countably infinite alphabet time series $\{X_n\}$ possesses finite entropy rate H. (A stationary finite alphabet time series always has finite entropy rate.)
We will in fact obtain a more precise estimate, namely, if for some $0 < \epsilon_2 < \epsilon_1$,

$$\sum_{k=1}^{\infty}(k+1)2^{-l_k(\epsilon_1-\epsilon_2)} < \infty$$

then

$$\lambda_n < 2^{l_n(H+\epsilon_1)} \quad \text{eventually almost surely.}$$

In particular, for arbitrary $\delta > 0$, $0 < \epsilon_2 < \epsilon_1$, if

$$l_n = \min\left(n, \max\left(1, \lfloor \frac{2+\delta}{\epsilon_1-\epsilon_2}\log_2 n \rfloor\right)\right)$$

then

$$\lambda_n < n^{\frac{2+\delta}{\epsilon_1-\epsilon_2}(H+\epsilon_1)}$$

eventually almost surely, and the upper bound is a polynomial.

Since $\lambda_n \le \zeta_n$, it is enough to prove the result for ζ_n. Let $\mathcal{X}^*$ be the set of all two-sided sequences, that is,

$$\mathcal{X}^* = \{(\ldots, x_{-1}, x_0, x_1, \ldots) : x_i \in \mathcal{X} \text{ for all } -\infty \le i < \infty\}.$$

Define $B_k \subseteq \mathcal{X}^{l_k}$ as

$$B_k = \{x_{-l_k+1}^0 \in \mathcal{X}^{l_k} : 2^{-l_k(H+\epsilon_2)} < p_{l_k-1}(x_{-l_k+1}^0)\}.$$

Note that there is a trivial bound on the cardinality of the set B_k, namely,

$$|B_k| \le 2^{l_k(H+\epsilon_2)}. \tag{16}$$

Define the set $\Upsilon_k(y_{-k+1}^0)$ as follows:

$$\Upsilon_k(y_{-l_k+1}^0) = \{z_{-\infty}^{\infty} \in \mathcal{X}^* : -\hat{\zeta}_k^k(z_{-\infty}^0) \ge 2^{l_k(H+\epsilon_1)}, z_{-l_k+1}^0 = y_{-l_k+1}^0)\}.$$

We will estimate the probability of $\Upsilon_k(y_{-l_k+1}^0)$ by a frequency argument. Let $x_{-\infty}^{\infty} \in \mathcal{X}^*$ be a typical sequence of the time series $\{X_n\}$. Define $\rho_0(y_{-l_k+1}^0, x_{-\infty}^{\infty}) = 0$ and for $i \ge 1$ let

$$\rho_i(y_{-l_k+1}^0, x_{-\infty}^{\infty}) = \min\{l > \rho_{i-1}(y_{-l_k+1}^0, x_{-\infty}^{\infty}) : T^{-l}x_{-\infty}^{\infty} \in \Upsilon_k(y_{-l_k+1}^0)\}.$$

Define also $\tau_0(y^0_{-l_k+1}, x^\infty_{-\infty}) = 0$ and for $i \geq 1$ let

$$\tau_i(y^0_{-l_k+1}, x^\infty_{-\infty}) = \min\{l \geq \tau_{i-1}(y^0_{-l_k+1}, x^\infty_{-\infty}) + 2^{l_k(H+\epsilon_1)} : T^{-l}x^\infty_{-\infty} \in \Upsilon_k(y^0_{-l_k+1})\}.$$

Notice that if $\tau_{i-1} = \rho_m$ then $\tau_i \leq \rho_{m+k+1}$. Indeed, since there are at least $k+1$ occurrences of the block $y^0_{-l_k+1}$ in the data segment $X^{\rho_m+1}_{-\rho_{m+k+1}-l_k+1}$ hence

$$2^{l_k(H+\epsilon_1)} \leq -\hat{\zeta}^k_k(T^{-\rho_m}x^\infty_{-\infty}) \leq \rho_{m+k+1} - \tau_{i-1}.$$

By the ergodicity of the time series $\{X_n\}$,

$$\begin{aligned}
&P(X^\infty_{-\infty} \in \Upsilon_k(y^0_{-l_k+1})) = \\
&\quad \lim_{t\to\infty} \frac{\#\{j \geq 1 : \rho_j(y^0_{-l_k+1}, x^\infty_{-\infty}) \leq \tau_t(y^0_{-l_k+1}, x^\infty_{-\infty})\}}{\tau_t(y^0_{-l_k+1}, x^\infty_{-\infty})} = \\
&\quad \lim_{t\to\infty} \frac{\sum_{l=1}^t \#\{j \geq 1 : \tau_{l-1}(y^0_{-l_k+1}, x^\infty_{-\infty}) < \rho_j(y^0_{-l_k+1}, x^\infty_{-\infty}) \leq \tau_l(y^0_{-l_k+1}, x^\infty_{-\infty})\}}{\tau_t(y^0_{-l_k+1}, x^\infty_{-\infty})} \leq \\
&\quad \lim_{t\to\infty} \frac{t(k+1)}{t2^{l_k(H+\epsilon_1)}} = \frac{(k+1)}{2^{l_k(H+\epsilon_1)}}. \qquad (17)
\end{aligned}$$

Since

$$T^l\{\zeta_k = l, X^{\zeta_k}_{\zeta_k-l_k+1} \in B_k\} = \{\hat{\zeta}^k_k = -l, X^0_{-l_k+1} \in B_k\}$$

by stationarity and the upper bound on the cardinality of the set B_k in (16) and by (17), we get

$$\begin{aligned}
P(\zeta_k \geq 2^{l_k(H+\epsilon_1)}, \tilde{X}^0_{-l_k+1} \in B_k) &= P(\zeta_k \geq 2^{l_k(H+\epsilon_1)}, X^{\zeta_k}_{\zeta_k-l_k+1} \in B_k) \\
&= P(-\hat{\zeta}^k_k \geq 2^{l_k(H+\epsilon_1)}, X^0_{-l_k+1} \in B_k) \\
&= \sum_{y^0_{-l_k+1} \in B_k} P(X^\infty_{-\infty} \in \Upsilon_k(y^0_{-l_k+1})) \\
&\leq (k+1)2^{-l_k(\epsilon_1-\epsilon_2)}.
\end{aligned}$$

By assumption, the right hand side sums and the Borel-Cantelli Lemma yields that the event

$$\{\zeta_k \geq 2^{l_k(H+\epsilon_1)}, \tilde{X}^0_{-l_k+1} \in B_k\}$$

cannot happen infinitely many times. By Lemma 1, the distribution of the time series $\{\tilde{X}_n\}$ is the same as the distribution of $\{X_n\}$ and by the Shannon-McMillan-Breiman Theorem (cf. Chung [3]) $\tilde{X}^0_{-l_k+1} \in B_k$ eventually almost surely and so $\zeta_k \geq 2^{l_k(H+\epsilon_1)}$ cannot happen infinitely many times.

Step 5. *We prove the third part of Theorem 1.*
By Step 1, if $1 \leq K(\tilde{X}^0_{-\infty}) < \infty$ then $\chi_n = K(\tilde{X}^0_{-\infty})$ eventually, and by ergodicity,

$$\frac{n}{\lambda_n} \to p_{K(\tilde{X}^0_{-\infty})-1}(\tilde{X}^0_{-K(\tilde{X}^0_{-\infty})+1}) > 0.$$

If $K(\tilde{X}^0_{-\infty}) = 0$ then by Step 1, $\chi_n = 0$ eventually, and by (6),

$$\lambda_n = \lambda_{n-1} + 1$$

eventually almost surely. The proof of Theorem 1 is complete.

3 Appendix

In the appendix we give the statement of two of the less familiar results thar we used. The first is due to Hoeffding, cf. [11].

Lemma 2 *(Hoeffding's inequality, Hoeffding [11]) Let $X_1, X_2, \ldots, X_n$ be independent real valued random variables, and $a_1, b_1, \ldots, a_n, b_n$ be real numbers such that $a_i \leq X_i \leq b_i$ with probability one for all $1 \leq i \leq n$. Then, for all $\epsilon > 0$,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i)\right| > \epsilon\right) \leq 2e^{-\left(2n\epsilon^2/\frac{1}{n}\sum_{i=1}^{n}|b_i-a_i|^2\right)}.$$

The next is due to Révész, cf. [18].

Lemma 3 *(Theorem 3.2.2 in Révész [18]) Let $X_1, X_2, \ldots, X_n$ be a sequence of square integrable random variables such that*

$$E(X_i) = E(X_i X_j) = 0 \quad \textit{for } i < j, i, j = 1, 2, \ldots$$

and

$$\sum_{i=1}^{\infty} \frac{E(X_i^2)}{i^2} \log^2 i < \infty.$$

Then

$$\frac{\sum_{i=1}^{n} X_i}{n} \to 0 \quad \textit{almost surely.}$$

References

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

[2] A. Berlinet, I. Vajda, E.C. van der Meulen, "About the asymptotic accuracy of Barron density estimates." *IEEE Transactions on Information Theory* vol.44, 3 (1998), p. 999-1009 (1998)

[3] K.L. Chung, "A note on the ergodic theorem of information theory," *The Annals of Mathematical Statistics.*, vol. 32, pp. 612-614, 1961.

[4] T.M. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 1991.

[5] I. Csiszár and P. Shields, "The consistency of the BIC Markov order estimator," *Annals of Statistics.*, vol. 28, pp. 1601-1619, 2000.

[6] I. Csiszár, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators ," *IEEE Transactions on Information Theory*, vol. 48, pp. 1616-1628, 2002.

[7] G.A. Darbellay and I. Vajda, " Estimation of the information by an adaptive partitioning of the observation space." *IEEE Transactions on Information Theory vol.45, 4 (1999), p. 1315-1321 (1999)*

[8] J. Feistauerov and I. Vajda, "Testing system entropy and prediction error probability" , *IEEE Transactions on Systems Man and Cybernetics* vol.23, 5 (1993), p. 1352-1358 (1993)

[9] L. Györfi, G. Morvai, and S. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Transactions on Information Theory*, vol. 44, pp. 886–892, 1998.

[10] L. Györfi, G. Morvai and I. Vajda, "Information-theoretic methods in testing the goodness of fit" , *Proceedings of the 2000 IEEE International Symposium on Information Theory*, p. 28, IEEE, (New York 2000) , ISIT 2000, (Sorrento, IT, 25.06.2000-30.06.2000) (2000)

[11] W. Hoeffding, "Probability inequalities for sums of bounded random variables ," *Journal of the American Statistical Association*, vol. 58, pp. 13-30, 1963.

[12] S. Kalikow "Random Markov processes and uniform martingales ," *Israel Journal of Mathematics*, vol. 71, pp. 33–54, 1990.

[13] M. Keane "Strongly mixing g-measures," *Invent. Math.* , vol. 16, pp. 309–324, 1972.

[14] H. Luschgy, L. A. Rukhin and I. Vajda, "Adaptive Tests for Stochastic Processes in the Ergodic Case" , *Stochastic Processes and their Applications* vol.45, 1 (1993), p. 45-59 (1993)

[15] G. Morvai and I. Vajda, "A Survay on Log-Optimum Portfolio Selection." *In: Second European Congress on Systems Science.* Afcet, Paris 1993, pp. 936-944.

[16] G. Morvai and B. Weiss, "Prediction for discrete time series." *Probability Theory and Related Fields,* Vol. 132, pp.1-12, 2005.

[17] G. Morvai and B. Weiss, "Estimating the memory for finitarily Markovian processes." *Ann. I.H.Poincaré Probabilités et Statistiqoes*, vol. 43, pp. 15-30, 2007.

[18] P. Révész, *The Law of Large Numbers*, Academic Press, 1968.

[19] B. Ya. Ryabko, "Prediction of random sequences and universal coding," *Problems of Inform. Trans.,* vol. 24, pp. 87-96, Apr.-June 1988.

[20] B. Ryabko, "Compression-based methods for nonparametric prediction and estimation of some characteristics of time series." *IEEE Trans. Inform. Theory* vol. 55 no. 9, pp. 43094315, 2009.

[21] I. Vajda, F. Österreicher, "Existence, Uniqueness and Evaluation of Log-Optimal Investment Portfolio" , *Kybernetika* vol.29, 2 (1993), p. 105-120 (1993)

[22] I. Vajda and P. Harremoës, "On the Bahadur-efficient testing of uniformity by means of entropy" , *IEEE Transactions on Information Theory* vol.54, p. 321-331 (2008)