Gusztáv Morvai; Benjamin Weiss

Intermittent estimation for finite alphabet finitarily Markovian processes with exponential tails

Kybernetika, Vol. 57 (2021), No. 4, 628–646

Persistent URL: http://dml.cz/dmlcz/149212

KYBERNETIKA — VOLUME 57 (2021), NUMBER 4, PAGES 628–646

INTERMITTENT ESTIMATION FOR FINITE ALPHABET FINITARILY MARKOVIAN PROCESSES WITH EXPONENTIAL TAILS

Gusztáv Morvai and Benjamin Weiss

We give some estimation schemes for the conditional distribution and conditional expectation of the the next output following the observation of the first n outputs of a stationary process where the random variables may take finitely many possible values. Our schemes are universal in the class of finitarily Markovian processes that have an exponential rate for the tail of the look back time distribution. In addition explicit rates are given. A necessary restriction is that the scheme proposes an estimate only at certain stopping times, but these have density one so that one rarely fails to give an estimate.

Keywords: nonparametric estimation, stationary processes

Classification: 62G05, 60G25, 60G10

1. INTRODUCTION

In this note we are concerned with the problem of how to best utilize partial information about an unknown stationary process in trying to estimate the conditional distribution of the next output, given that we have observed the first n outputs of the process. Our goal is to obtain explicit bounds for the rate of convergence to zero of the error in our estimation scheme that would be valid for almost any sequence of outputs. The kind of processes that we are interested are best exemplified by binary renewal processes. These are included in the wider class finitarily Markovian (FM) processes over a finite alphabet. To describe this class we need the notion of a memory word w. A word w of length k in the alphabet of the state space of our process is called a memory word if given that the current k outputs constitutes the word w the conditional distribution of the next output is completely determined independently of the rest of the past (a formal definition is given in the next section). A process is FM if with probability one the current outputs for some k form a memory word (cf. [19, 20]). Any process with at least one renewal state has such memory words, and if the process is ergodic it is necessarily FM. We have shown in some earlier works [22, 26, 27, 28] how to use this information to get good estimation schemes for binary renewal processes, but in that case there is prior knowledge of the set of memory words and we would like to extend

DOI: 10.14736/kyb-2021-4-0628

the class of processes and to relax this assumption. We will replace it by an assumption on the tail of the distribution of the **look back time**. For FM processes one defines the look back time, τ, as a function of the entire past $\{X_n, n < 0\}$ to be the minimal k so that $\{X_{-k}...X_{-2}X_{-1}\}$ is a memory word. We will only assume that the tail of the distribution of τ decays exponentially fast - but that otherwise we have no specific knowledge of a sufficient set of memory words.

Even if one would be satisfied with a non quantitative result there is no universal scheme for estimating the conditional distribution of the next output at all time instants in the class of all FM process. This is because there may be infinitely many memory words and the first time a memory word is encountered no meaningful estimation can be made, cf. [31]. It is for this reason **intermittent estimation** has been introduced. Here the idea is to estimate only along a sequence of carefully chosen stopping times. More precisely one defines a sequence of stopping times $\{\lambda_1 < \lambda_2 \cdots < \lambda_n < \cdots\}$ and then gives an estimate for the conditional distribution of X_{λ_n+1} given the values of $\{X_i : 0 \le i \le \lambda_n\}$. Our assumptions will enable us to show that the sequence of stopping times has density one so that we will be giving an estimate "almost all" of the time. In [31] it was shown that a universal scheme for estimating the conditional distribution of the next output does not exist for the class of stationary ergodic processes. He based his proof on a random walk along the Cantor diagonal. The same technique can be applied to prove this for the class of all FM processes. (In fact his counterexample is FM over a ternary alphabet. His powerful technique, the encoding of a countable Markov chain, was used in [2, 5] and [15] where the counterexamples are in fact binary FM.) We will sharpen this result by combining his techniques in [31] with those in [5] and [15].

In section 2 we will describe in detail the class of FM processes. Im section 3 we will define exactly what is intermittent estimation, the stopping times that we need use, define the stopping times and show that they have density one. In section 4 we will give the scheme for estimating the conditional distribution of the next symbol, while in section 5 we suppose that there is a real valued function of the basic states and give a scheme for estimating its conditional expectation given the first n outputs. Finally in the last section we will give our sharpened version of the results in [31] and [15].

For further reading on the topics cf. [2, 3, 24, 28, 30, 32, 34, 35] and [29].

2. CLASSES OF PROCESSES

First let us fix the notation. Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a finite alphabet $\mathcal{X}$. We will use heavily in this paper that the alphabet is finite. 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 $m \le n$. Note that if $m > n$ then X_m^n is the empty string.

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

An important notion is that of a **memory word** which is defined as follows (cf. [4, 19, 20, 21]).

Definition 2.1. We say that the empty word $\emptyset$ with length zero is a memory word if for all $i \geq 1$, all $y \in \mathcal{X}$, all $z^0_{-i+1} \in \mathcal{X}^i$ such that $p(z^0_{-i+1}, y) > 0$:

$$p(y) = p(y|z^0_{-i+1}).$$

For $k \geq 1$ we say that w^0_{-k+1} is a memory word if $p(w^0_{-k+1}) > 0$ and for all $i \geq 1$, all $y \in \mathcal{X}$, all $z^{-k}_{-k-i+1} \in \mathcal{X}^i$ such that $p(z^{-k}_{-k-i+1}, w^0_{-k+1}, y) > 0$:

$$p(y|w^0_{-k+1}) = p(y|z^{-k}_{-k-i+1}, w^0_{-k+1}).$$

Note that the empty word is a memory word if and only if the stationary stochatic process is independent and identically distributed. For more on memory words cf. [25]

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

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

Note that $\mathcal{W}_0$ can contain at most the empty word in which case the stationary process is independent.

Example 2.2. Consider an independent and identically distributed process on a finite alphabet. Assume that every letter in the alphabet has positive probability. Then any word with any length is a memory word, including the empty word.

Example 2.3. Consider an independent and identically distributed process on a finite alphabet. Then the empty word is a memory word. Furthermore, any word with positive length is a memory word if it contains no letter which has probability zero. In other words, the memory words are the empty word and any word which has positive probability.

Example 2.4. Consider the Markov chain with state space $S = \{0, 1, 2\}$ and transition probabilities

$$P(M_2 = 1|M_1 = 0) = P(M_2 = 2|M_1 = 1) = 1,$$
$$P(M_2 = 0|M_1 = 2) = P(M_2 = 1|M_1 = 2) = 0.5.$$

This yields a stationary and ergodic process $\{M_n\}^\infty_{n=-\infty}$. Define

$$Z_n = I_{\{M_n=1\}}.$$

Then $\{Z_n\}^\infty_{n=-\infty}$ is a stationary and ergodic binary Markov chain with order 2. Now we examine the memory words of process $\{Z_n\}^\infty_{n=-\infty}$. The empty word is not a memory word of the process $\{Z_n\}^\infty_{n=-\infty}$ since the process is not independent. Thus $\mathcal{W}_0$ is the empty set. '0' is not a memory word of process $\{Z_n\}^\infty_{n=-\infty}$. The only word with length one which is a memory word of process $\{Z_n\}^\infty_{n=-\infty}$ is '1'. The memory words with length two are the '01', the '10' and the '00'. Note that the word '11' appears in the $\{Z_n\}^\infty_{n=-\infty}$ process with probability zero. Thus '11' is not a memory word. Since process $\{Z_n\}^\infty_{n=-\infty}$ is a Markov chain with order two, any word with length at least two, which has positive probability, is a memory word.

Define the set of all memory words $\mathcal{W}$ as

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

Definition 2.5. A stationary and ergodic process $\{X_n\}$ is said to be finitarily Markovian if

$$P(\exists 0 \le k < \infty : X_{-k+1}^0 \in \mathcal{W}) = 1.$$

Definition 2.6. Define the look back time $\tau(X_{-\infty}^n)$ at time n as

$$\tau(X_{-\infty}^n) = \inf\{t \ge 0 \ : \ X_{n-t+1}^n \in \mathcal{W}\}.$$

Note that the stationary and ergodic process $\{X_n\}$ is finitarily Markovian if and only if

$$P(\tau(X_{-\infty}^0) < \infty) = 1.$$

The class of finitarily Markovian processes includes of course all finite order Markov chains but also many other processes such as the finitarily determined processes of Kalikow, Katznelson and Weiss [7], which serve to represent all isomorphism classes of zero entropy processes. For some concrete examples that are not Markovian consider the following example:

Example 2.7. Let $\{M_n\}$ be any stationary and ergodic first order Markov chain with finite or countably infinite state space S. Let $s \in S$ be an arbitrary state with $P(M_1 = s) > 0$. Now let

$$X_n = I_{\{M_n = s\}}.$$

By Shields [33] Chapter I.2.c.1, the binary time series $\{X_n\}$ is stationary and ergodic. It is also finitarily Markovian. Indeed, the conditional probability $P(X_1 = 1|X_{-\infty}^0)$ does not depend on values beyond the first (going backwards) occurrence of one in $X_{-\infty}^0$ which identifies the first (going backwards) occurrence of state s in the Markov chain $\{M_n\}$. The resulting time series $\{X_n\}$ is not a Markov chain of any order in general. Indeed, consider the Markov chain $\{M_n\}$ with state space $S = \{0, 1, 2\}$ and transition probabilities

$$P(M_2 = 1|M_1 = 0) = P(M_2 = 2|M_1 = 1) = 1,$$
$$P(M_2 = 0|M_1 = 2) = P(M_2 = 1|M_1 = 2) = 0.5.$$

This yields a stationary and ergodic Markov chain $\{M_n\}$, cf. Example I.2.8 in Shields [33]. Clearly, the resulting time series

$$X_n = I_{\{M_n = 0\}}$$

will not be Markov of any order. The conditional probability $P(X_1 = 0|X_{-\infty}^0)$ depends on whether until the first (going backwards) occurrence of one you see even or odd number of zeros. These examples include all stationary and ergodic binary renewal processes with finite expected inter-arrival times, a basic class for many applications. (A stationary and ergodic binary renewal process is defined as a stationary and ergodic binary process such that the times between occurrences of ones are independent and identically distributed with finite expectation, cf. Chapter I.2.c.1 in Shields [33]).

In the previous example any word with positive probability which contains at least one zero is a memory word and determines not only the conditional distribution of X_1 given the past, but also the conditional distribution of X_1^∞ given the past. This need not be the case. For a simple example of such a process we first recall a general construction of stationary processes. Let (Ω, Σ, P) be a probability space and $T : \Omega \to \Omega$ a measurable transformation such that for all $A \subset \Omega$ one has $P(T^{-1}(A)) = P(A)$. Then if X_0 is a any random variable defined on (Ω, Σ, P) a stationary process is defined by setting for all n: $X_n(\omega) = X_0(T^n(\omega))$.

Example 2.8. Let now Ω be the unit circle $\mathbb{R}/\mathbb{Z}$ with Lebesgue measure and for some irrational $0 < \alpha < 1/2$ define $T(\omega) = \omega + \alpha$, where the addition is modulo 1. Let X_0 be the indicator function of the interval $I = [0, 1/2)$ and define the process X_n by $X_n(\omega) = X_0(\omega + n\alpha)$. The values of X_n for $-N \le n \le 0$ determine small intervals that are obtained by intersecting the intervals of the form $[n\alpha, 1/2 + n\alpha)$. Since α is irrational these endpoints form a dense subset of $\mathbb{R}/\mathbb{Z}$. In order to determine the value of $X_1(\omega)$ we need to know whether $\omega + \alpha$ belongs to I or not. If we know the values of X_n for $-N \le n \le 0$ then most of the small intervals determined by these values will be such that when we rotate the interval by α the interval will be either entirely in I or in its complement and so the value of $X_1(\omega)$ will be determined. There are only two such small intervals that contain the endpoints of $I - \alpha$ and for any ω in one of these two small intervals the values of X_{-N}^0 do not determine the value of X_1. It follows that the set of points for which X_{-N}^0 does not determine the value of X_1 has measure which tends to zero as N tends to infinity and this shows that the process is finitarily Markovian.

However, in order to determine the entire future we need to know whether $\omega + n\alpha$ belongs to I or not for all $n \ge 1$. No finite number of observations in the past will completely determine the value of ω so that the memory words that determine X_1 do not determine the complete future. For more on these kinds of examples see [7].

This process has zero entropy but can be easily modified as follows to get a positive entropy process with the same features. One prepares two independent i.i.d. processes Y_n, and Z_n independent of the X_n process, such that $P(Y_n = 0) = 1/3$ and $P(Y_n = 2) = 2/3$ while $P(Z_n = 0) = 2/3$ and $P(Z_n = 2) = 1/3$. Now define a new process U_n by setting $U_n = X_n + Y_n$ when $X_{n-1} = 0$ and $U_n = X_n + Z_n$ when $X_{n-1} = 1$. In this new process the parity is exactly the original X_n process and the memory words for that process are also memory words for this one which clearly has positive entropy.

We note that Morvai and Weiss [16] proved that there is no classification rule for discriminating the class of finitarily Markovian processes from other ergodic processes.

For more on finitarily Markovian processes and intermittent estimation we refer the interested reader to [9, 11, 12, 13, 15, 17, 18, 14, 22, 23].

Throughout in this paper we will assume that the stationary and ergodic process $\{X_n\}$ is finitarily Markovian. (This implies that the look back time will be finite almost surely.)

Now we define the subclass of finitarily Markovian processes for which the tail distribution of the look back time vanishes exponentially fast.

Definition 2.9. Let $0 < R < 1$ be arbitrary. A stationary and ergodic process $\{X_n\}$ is in FMEXPTAIL(R) if it is finitarily Markovian and for some $0 < \rho < R$,

$$P(\tau(X_{-\infty}^0) > n) \leq \rho^n$$

eventually.

Example 2.10. All stationary and ergodic Markov chains with any finite order are in the class $\bigcap_{0<R<1} FMEXPTAIL(R)$.

Example 2.11. Consider the Markov chain $\{M_n\}$ with countably infinite state space $S = \{0, 1, 2, \dots\}$ and transition probabilities

$$P(M_1 = n+1 | M_0 = n) = \left(\frac{1}{2}\right)^{n+1},$$

$$P(M_1 = 0 | M_0 = n) = 1 - \left(\frac{1}{2}\right)^{n+1}$$

where $n \in S$. This yields a stationary and ergodic first order Markov chain $\{M_n\}$. Define $Z_n = I_{\{M_n \neq 0\}}$. Then $\{Z_n\}$ is a stationary and ergodic binary process which is not Markov of any finite order but it is in $\bigcap_{0<R<1} FMEXPTAIL(R)$. In fact, the resulting process $\{Z_n\}$ is a renewal process with renewal state $'0'$.

Example 2.12. Let $0 < R < 1$ and $0 < q < R$ be arbitrary. First we define a Markov-chain which serves as the technical tool for construction of our erexample. Let the state space S be the non-negative integers. From state 0 the process certainly passes to state 1 and then to state 2, at the following epoch. From each state $s \geq 2$, the Markov chain passes either to state 0 with probability $1 - q$ or to state $s + 1$ with probability q. This construction yields a stationary and ergodic Markov chain $\{M_i\}$ with stationary distribution

$$P(M = 0) = P(M = 1) = \frac{1}{2 + \frac{1}{1-q}} = \frac{1-q}{3-2q}$$

and

$$P(M = i) = \frac{1-q}{3-2q} q^{i-2} \quad \text{for } i \geq 2.$$

We will define a binary process $\{X_i\}$ which we denote as $X_i = f(M_i)$ where f is a binary valued function of the state space S. Let $f(0) = 0$, $f(1) = 0$, and $f(s) = 1$ for all even states s. The values $f(s)$ for the odd states $s \geq 3$ can be chosen arbitrarily. The resulting process is stationary and ergodic binary finitarily Markovian (any word with positive probability which contains the sequence 001 is a memory word) and is in $FMEXPTAIL(R)$. Cf. the proof of Theorem 6.1.

Example 2.13. Consider the Markov chain $\{M_n\}$ with countably infinite state space $S = \{0, 1, 2, \dots\}$ and transition probabilities

$$P(M_1 = 1 | M_0 = 0) = P(M_1 = 2 | M_0 = 1) = 1$$

and for $n = 2, 3, \dots$ let

$$P(M_1 = n+1 | M_0 = n) = \left(\frac{n+1}{n+2}\right)^3,$$

$$P(M_1 = 0 | M_0 = n) = 1 - \left(\frac{n+1}{n+2}\right)^3.$$

This yields a stationary and ergodic first order Markov chain $\{M_n\}$. Define $Z_n = I_{\{M_n \ge 2\}}$. Then $\{Z_n\}$ is a stationary and ergodic binary process which is not Markov of any finite order, it is finitarily Markovian (FM) but it is not in $\bigcup_{0<R<1} FMEXPTAIL(R)$.

3. INTERMITTENT SCHEMES AND STOPPING TIMES

An intermittent scheme with respect to a class of processes consists of two parts. A sequence of stopping times which are almost surely finite and strictly increasing $\lambda_1 < \lambda_2 \cdots < \lambda_n < \cdots$ for any process in the class and a sequence of real-valued $\sigma(X_0^{\lambda_n})$-measurable functions $\{h_n\}$. Such schemes are called intermittent because the estimator h_n is defined only for sequences of the form $(X_0, \dots, X_{\lambda_n})$. (h_n gives estimate only where the stopping time λ_n stops.) Often, we will use the notation $h(X_0^{\lambda_n})$ instead of h_n or $h_n(X_0^{\lambda_n})$ for the sake of notational convenience. Note that if $\lambda_n = n$ almost surely for all n then h_n depends solely on X_0^n and we are in the usual sequential estimation settings. In this paper the class of processes we will consider is the $FMEXPTAIL(R)$ for some $0 < R < 1$. (For other cases cf. e. g. [9, 11, 29].)

The purpose of this section is to define the sequence of stopping times $\{\lambda_n\}$ on which we will later estimate different quantities intermittently.

Let $c > 0$ be arbitrary. For $n = 1, 2, \dots$ put

$$k_n = \max\{\lfloor c\log(n)\rfloor, 1\}.$$

Note that all logarithms in this paper are to base 2.

Let $0 < \gamma < 1$ be arbitrary. Define

$$J_k = \left\lceil 2^{\frac{k(1-\gamma)}{c}} \right\rceil.$$

(Note that J_{k_n} grows roughly as $n^{1-\gamma}$.)

We will assume that

$$c\log(|\mathcal{X}|) < \gamma.$$

Define $\zeta_0^{(k,m)} = m$. Define $\zeta_1^{(k,m)}$ as

$$\zeta_1^{(k,m)} = \max\left\{t < m : X_{t-k+1}^t = X_{m-k+1}^m\right\}.$$

Define $\zeta_2^{(k,m)}$ as

$$\zeta_2^{(k,m)} = \max\left\{t < \zeta_1^{(k,m)} : X_{t-k+1}^t = X_{m-k+1}^m\right\}.$$

In general, let $\zeta_i^{(k,m)}$ be defined as

$$\zeta_i^{(k,m)} = \max\left\{t < \zeta_{i-1}^{(k,m)} : X_{t-k+1}^t = X_{m-k+1}^m\right\}.$$

Note that $\zeta_i^{(k,m)}$ is the ith occurrence of the word (seen at position m with length k) X_{m-k+1}^m going backwards in the negative direction. Note that $\zeta_i^{(k,m)}$ is finite with probability one by the Poincaré recurrence theorem for stationary processes.

Define the stopping times λ_n as follows. Let $\lambda_0 = 0$. Define

$$\lambda_1 = \min\left\{\ t > 0\ \ : 0 < k_t < t, \zeta_{J_{k_t}}^{(k_t,t)} \geq k_t - 1\right\}.$$

Define

$$\lambda_2 = \min\left\{\ t > \lambda_1\ \ : 0 < k_t < t, \zeta_{J_{k_t}}^{(k_t,t)} \geq k_t - 1\right\}.$$

In general, define

$$\lambda_n = \min\left\{\ t > \lambda_{n-1}\ \ : 0 < k_t < t, \zeta_{J_{k_t}}^{(k_t,t)} \geq k_t - 1\right\}.$$

Note that the event$\{\lambda_n = t\}$ is measurable with respect to $\sigma(X_0^t)$. Roughly, λ_n is the smallest $t > \lambda_{n-1}$ such that one can find J_{k_t} occurrences of $X_{t-k_t+1}^t$ in the data segment X_0^{t-1}. By Theorem 3.1, λ_n is finite with probability one.

Theorem 3.1. Let $\{X_n\}$ be a stationary and ergodic finitarily Markovian time series taking values from the finite alphabet $\mathcal{X}$. Assume that $0 < \gamma < 1$, $0 < c$, and

$$c\log(|\mathcal{X}|) < \gamma.$$

Then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I_{\{\zeta_{J_{k_n}}^{(k_n,n)} \geq k_n - 1\}} = 1$$

almost surely, the stopping times λ_n are finite with probability one and

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$$

almost surely.

P r o o f. Let $0 < \epsilon$ be arbitrary. By Lemma 7.1 in the Appendix (Theorem 2 in [10]),

$$\limsup_{n\to\infty} \frac{1}{k_n} \log\left(\frac{-\zeta_{J_{k_n}}^{(k,0)}}{J_{k_n}}\right) \leq H$$

almost surely where H is the entropy rate of the process. (The entropy rate of the stationary and ergodic process $\{X_n\}$ is defined as $H = \lim_{k\to\infty} -\frac{1}{k}E(\log(p(X_{-k+1}^0)))$.) Since $H \le \log(|\mathcal{X}|)$ we get that

$$-\zeta_{J_{k_n}}^{(k_n,0)} < J_{k_n} 2^{k_n(\log(|\mathcal{X}|)+\epsilon)}$$

eventually almost surely. Now

$$\begin{aligned} J_{k_n} 2^{k_n(|\mathcal{X}|+\epsilon)} &= \left\lceil 2^{\frac{k_n(1-\gamma)}{c}} \right\rceil 2^{k_n(\log(|\mathcal{X}|)+\epsilon)} \\ &\le \left(2^{(1-\gamma)\log(n)} + 1\right) 2^{c\log(n)(\log(|\mathcal{X}|)+\epsilon)} \\ &\le n^{(1-\gamma+c(\log(|\mathcal{X}|)+\epsilon))} + n^{c(\log(|\mathcal{X}|)+\epsilon)} \end{aligned}$$

which is eventually less than $n - c\log(n)$ provided $0 < c < \frac{\gamma}{\log(|\mathcal{X}|)+\epsilon}$. Since ϵ was arbitrary and by assumption, for some sufficiently small $\epsilon > 0$

$$0 < c < \frac{\gamma}{\log(|\mathcal{X}|)+\epsilon},$$

so we get that

$$\zeta_{J_{k_n}}^{(k_n,0)} - k_n + 1 \ge -n$$

eventually almost surely. That is,

$$I_{\{\zeta_{J_{k_n}}^{(k_n,0)} - k_n + 1 \ge -n\}} = 1$$

eventually almost surely. Now by Maker's generalized ergodic theorem (rediscovered by Breiman, cf. Theorem 1 in Maker [8] or Theorem 12 in Algoet [1])

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I_{\{\zeta_{J_{k_n}}^{(k_n,0)} - k_n + 1 \ge -n\}}(T^n\omega) = 1$$

almost surely, where T denotes the left shift. Thus

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I_{\{\zeta_{J_{k_n}}^{(k_n,n)} \ge k_n - 1\}}(\omega) = 1$$

almost surely. Now it is immediate that the stopping times λ_n are finite with probability one and

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$$

almost surely. The proof of Theorem 3.1 is complete. □

Note that in Theorem 3.1 we used that the alphabet is finite.

4. ESTIMATING THE CONDITIONAL DISTRIBUTION

Define the empirical conditional probability $\hat{p}(x|X_0^{\lambda_n})$ as

$$\hat{p}(x|X_0^{\lambda_n}) = \frac{\sum_{i=1}^{J_{k_{\lambda_n}}} I_{\{X_{\zeta_i^{(k_{\lambda_n},\lambda_n)}+1}=x\}}}{J_{k_{\lambda_n}}}.$$

Roughly, $\hat{p}(x|X_0^{\lambda_n})$ is the ratio of the number of times the letter x follows the $J_{k_{\lambda_n}}$ occurrences of $X_{\lambda_n-k_{\lambda_n}+1}^{\lambda_n}$ to $J_{k_{\lambda_n}}$.

Theorem 4.1. Let $\{X_n\}$ be a stationary and ergodic finitarily Markovian time series taking values from the finite alphabet $\mathcal{X}$. Assume that $0 < \gamma < 1$, $0 < c$, and

$$c \log(|\mathcal{X}|) < \gamma.$$

Assume that the process $\{X_n\}$ is in $FMEXPTAIL(2^{-1/c})$. Assume furthermore that $\epsilon_m > 0$ and

$$\sum_{m=1}^{\infty} e^{-\frac{\epsilon_m^2 J_{k_m}}{2}} < \infty.$$

Then for the stopping times λ_n,

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$$

almost surely and for the estimator $\hat{p}(x|X_0^{\lambda_n})$

$$\max_{x\in\mathcal{X}} \left|\hat{p}(x|X_0^{\lambda_n}) - P(X_{\lambda_n+1} = x|X_0^{\lambda_n})\right| < \epsilon_{\lambda_n}$$

eventually almost surely.

Proof. By Theorem 3.1,

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$$

almost surely.

For a fixed m put $\eta_m = \max\{\tau(X_{-\infty}^m), k_m\}$. Now for $i \geq 0$ define

$$Z_i^{(\eta_m,m)}(x) = I_{\{X_{\zeta_i^{(\eta_m,m)}+1}=x\}}.$$

Clearly, for fixed m and x, $\{Z_i^{(\eta_m,m)}(x)\}_{i=1}^{\infty}$ are conditionally independent and identically distributed given $X_{m-\eta_m+1}^m$. Apply Hoeffding's inequality for sums of independent and identically distributed bounded random variables (cf. Theorem 2 in [6]) to get that

$$P\left(\left|\frac{\sum_{i=1}^{J_{k_m}} Z_i^{(\eta_m,m)}(x)}{J_{k_m}} - P(Z_0^{(\eta_m,m)}(x)|X_{-\infty}^m)\right| \geq \epsilon_m | X_{m-\eta_m+1}^m\right) \leq 2e^{-\frac{\epsilon_m^2 J_{k_m}}{2}}.$$

Now by the union bound

$$P\left(\max_{x\in\mathcal{X}}\left|\frac{\sum_{i=1}^{J_{k_m}} Z_i^{(\eta_m,m)}(x)}{J_{k_m}} - P(Z_0^{(\eta_m,m)}(x)|X_{-\infty}^m)\right| \geq \epsilon_m | X_{m-\eta_m+1}^m\right) \leq 2|\mathcal{X}|e^{-\frac{\epsilon_m^2 J_{k_m}}{2}}.$$

After integrating both sides with respect to the conditioning we get that

$$P\left(\max_{x\in\mathcal{X}}\left|\frac{\sum_{i=1}^{J_{k_m}} Z_i^{(\eta_m,m)}(x)}{J_{k_m}} - P(Z_0^{(\eta_m,m)}(x)|X_{-\infty}^m)\right| \geq \epsilon_m\right) \leq 2|\mathcal{X}|e^{-\frac{\epsilon_m^2 J_{k_m}}{2}}$$

which is summable by assumption and so by the Borel–Cantelli lemma, eventually almost surely,

$$\max_{x\in\mathcal{X}}\left|\frac{\sum_{i=1}^{J_{k_m}} Z_i^{(\eta_m,m)}(x)}{J_{k_m}} - P(Z_0^{(\eta_m,m)}(x)|X_{-\infty}^m)\right| < \epsilon_m.$$

Since, by assumption, for some $0 \leq \rho < 2^{-1/c}$

$$P(\tau(X_{-\infty}^0) > n) < \rho^n$$

eventually, by stationarity for large enough n,

$$\begin{aligned}
P(\tau(X_{-\infty}^n) > k_n) &= P(\tau(X_{-\infty}^0) > k_n) \\
&\leq P(\tau(X_{-\infty}^0) > \lfloor c\log(n)\rfloor) \\
&\leq \rho^{\lfloor c\log(n)\rfloor} \\
&\leq \rho^{(c\log(n))-1} \\
&\leq \frac{2^{(c\log(n)\log(\rho))}}{\rho} \\
&\leq \frac{n^{(c\log(\rho))}}{\rho} \\
&= \frac{1}{\rho n^{(-c\log(\rho))}}
\end{aligned}$$

and the right hand side is summable. Applying the Borel–Cantelli lemma we get

$$\tau(X_{-\infty}^n) \leq k_n$$

eventually almost surely.

Now, almost surely, there exists $N(\omega)$ such that for $n > N(\omega)$, for each n there is an m such that $\lambda_n = m$,

$$\eta_m = k_m$$

and for all $x \in \mathcal{X}$

$$\hat{p}(x|X_0^{\lambda_n}) = \frac{\sum_{i=1}^{J_{k_m}} Z_i^{(k_m,m)}(x)}{J_{k_m}}.$$

Thus

$$\max_{x\in\mathcal{X}} \left|\hat{p}(x|X_0^{\lambda_n}) - P(X_{\lambda_n+1} = x|X_0^{\lambda_n})\right| < \epsilon_{\lambda_n}$$

eventually almost surely. The proof of Theorem 4.1 is complete. □

Since for $2^{1/c} \le n$

$$\begin{aligned} J_{k_n} &= \left\lceil 2^{\frac{k_n(1-\gamma)}{c}} \right\rceil \\ &\ge 2^{\frac{\lfloor c\log(n)\rfloor(1-\gamma)}{c}} \\ &\ge 2^{\frac{(c\log(n)-1)(1-\gamma)}{c}} \\ &\ge 2^{-\frac{(1-\gamma)}{c}} n^{1-\gamma} \end{aligned}$$

we have immediately the following corollary.

Corollary 4.2. Let $\{X_n\}$ be a stationary and ergodic finitarily Markovian time series taking values from the finite alphabet $\mathcal{X}$. Assume that $0 < \gamma < 1$, $0 < c$, and

$$c\log(|\mathcal{X}|) < \gamma.$$

Assume that the process $\{X_n\}$ is in $FMEXPTAIL(2^{-1/c})$. Let $\epsilon > 0$, $\delta > 0$ such that $2\delta + \gamma < 1$. Put $\epsilon_m = \frac{\epsilon}{m^\delta}$. Then for the stopping times λ_n ,

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1 \tag{1}$$

almost surely and for the estimator $\hat{p}(x|X_0^{\lambda_n})$

$$\max_{x\in\mathcal{X}} \left|\hat{p}(x|X_0^{\lambda_n}) - P(X_{\lambda_n+1} = x|X_0^{\lambda_n})\right| < \frac{\epsilon}{(\lambda_n)^\delta}$$

eventually almost surely.

5. ESTIMATING THE CONDITIONAL EXPECTATION

Fix an arbitrary real valued function

$$s : \mathcal{X} \to \Re.$$

We will estimate the the conditional expectation for $s(X)$. Define the empirical conditional expectation $\hat{E}(x|X_0^{\lambda_n})$ as

$$\hat{E}(X_0^{\lambda_n}) = \sum_{x\in\mathcal{X}} s(x)\hat{p}(x|X_0^{\lambda_n}).$$

$\hat{E}(X_0^{\lambda_n})$ is the empirical conditional expectation with respect to the empirical conditional distribution $\hat{p}(\cdot|X_0^{\lambda_n})$.

Theorem 5.1. Let $\{X_n\}$ be a stationary and ergodic finitarily Markovian time series taking values from the finite alphabet $\mathcal{X}$. Assume that $0<\gamma<1$, $0<c$, and

$$c\log(|\mathcal{X}|)<\gamma.$$

Assume that the process $\{X_n\}$ is in $FMEXPTAIL(2^{-1/c})$. Assume furthermore that $\epsilon_m>0$ and

$$\sum_{m=1}^{\infty} e^{-\frac{\epsilon_m^2 J_{k_m}}{2}}<\infty.$$

Then for the stopping times λ_n ,

$$\lim_{n\to\infty}\frac{\lambda_n}{n}=1$$

almost surely and for the estimator $\hat{E}(X_0^{\lambda_n})$

$$\left|\hat{E}(X_0^{\lambda_n})-E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right|<|\mathcal{X}|\max_{x\in\mathcal{X}}|s(x)|\epsilon_{\lambda_n}$$

eventually almost surely.

Proof. By Theorem 4.1,

$$\lim_{n\to\infty}\frac{\lambda_n}{n}=1$$

almost surely and

$$\begin{aligned}
&\left|\hat{E}(X_0^{\lambda_n})-E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right|\\
\le\ &\left|\left(\sum_{x\in\mathcal{X}} s(x)\hat{p}(x|X_0^{\lambda_n})\right)-\left(\sum_{x\in\mathcal{X}} s(x)P(X_{\lambda_n+1}=x|X_0^{\lambda_n})\right)\right|\\
\le\ &|\mathcal{X}|\max_{x\in\mathcal{X}}|s(x)|\max_{x\in\mathcal{X}}\left|\hat{p}(x|X_0^{\lambda_n})-P(X_{\lambda_n+1}=x|X_0^{\lambda_n})\right|\\
<\ &|\mathcal{X}|\max_{x\in\mathcal{X}}|s(x)|\epsilon_{\lambda_n}.
\end{aligned}$$

The proof of Theorem 5.1 is complete. □

Corollary 5.2. Let $\{X_n\}$ be a stationary and ergodic finitarily Markovian time series taking values from the finite alphabet $\mathcal{X}$. Assume that $0<\gamma<1$, $0<c$, and

$$c\log(|\mathcal{X}|)<\gamma.$$

Assume that the process $\{X_n\}$ is in $FMEXPTAIL(2^{-1/c})$. Let $\epsilon>0$, $\delta>0$ such that $2\delta+\gamma<1$. Put $\epsilon_m=\frac{\epsilon}{m^\delta}$. Then for the stopping times λ_n,

$$\lim_{n\to\infty}\frac{\lambda_n}{n}=1$$

almost surely and for the estimator $\hat{E}(X_0^{\lambda_n})$

$$\left|\hat{E}(X_0^{\lambda_n})-E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right|<|\mathcal{X}|\max_{x\in\mathcal{X}}|s(x)|\frac{\epsilon}{(\lambda_n)^\delta}$$

eventually almost surely.

6. THE JUSTIFICATION OF INTERMITTENT ESTIMATION

The next result is an improved, stronger version of the one in Morvai and Weiss [15]. The technique that will be used in the proof, namely the encoding of a countable Markov chain, originates from Ryabko [31]. Note that all stationary and ergodic Markov chains with any finite order are in the class $\bigcap_{0<R<1} FMEXPTAIL(R)$.

More precisely we will show that if for some $0 < R < 1$ an intermittent scheme is consistent for all stationary and ergodic binary time series in $FMEXPTAIL(R)$ then for some stationary and ergodic binary Markov-chain with some finite order $\{\lambda_{n+1} > \lambda_n + 1\}$ happens for infinitely many n with some positive probability.

However, note that if the goal is to estimate merely for the class of all binary Markov chains with some finite order one can choose $\lambda_n = n$. Indeed, one can estimate the order of the Markov chain (cf. e.g. [17, 21] and [25]) and using this estimated order count frequencies of blocks. Eventually almost surely the estimated order will coincide with the real order and there are only finitely many words with that length and the frequency counts for these finitely many words will tend to the real conditional probabilities, almost surely. Cf. [17, 21, 25] and [19].

Theorem 6.1. Consider the binary alphabet $\{0, 1\}$. Let $0 < R < 1$ be arbitrary. For any sequence of stopping times $\{\lambda_n\}$ such that for all stationary and ergodic finitarily Markovian binary time series $\{X_n\}$ in $FMEXPTAIL(R)$ the stopping times λ_n are almost surely finite and $\lambda_1 < \lambda_2 \cdots < \lambda_n < \cdots$ and for all stationary and ergodic binary Markov-chains with arbitrary finite order, eventually almost surely $\lambda_{n+1} = \lambda_n + 1$, and for any sequence of estimators $\{h_n(X_0, \ldots, X_{\lambda_n})\}$ there is a stationary and ergodic finitarily Markovian binary time series $\{X_n\}$ in $FMEXPTAIL(R)$ such that

$$P\left(\limsup_{n\to\infty} |h_n(X_0, \ldots, X_{\lambda_n}) - P(X_{\lambda_n+1} = 1 | X_0, \ldots, X_{\lambda_n})| > 0\right) > 0.$$

P r o o f. The proof mainly follows the footsteps of Ryabko [31], Györfi, Morvai, Yakowitz [5] and especially Morvai and Weiss [15] with alterations where necessary.
Let $0 < q < R$ be arbitrary. First we define a Markov-chain which serves as the technical tool for construction of our counterexample. Let the state space S be the non-negative integers. From state 0 the process certainly passes to state 1 and then to state 2. From each state $s \geq 2$, the Markov chain passes either to state 0 with probability $1 - q$ or to state $s + 1$ with probability q. This construction yields a stationary and ergodic Markov chain $\{M_i\}$ with stationary distribution

$$P(M = 0) = P(M = 1) = \frac{1}{2 + \frac{1}{1-q}} = \frac{1-q}{3-2q}$$

and

$$P(M = i) = \frac{1-q}{3-2q} q^{i-2} \quad \text{for } i \geq 2.$$

Let ψ_k denote the first positive time of occurrence of state $2k$:

$$\psi_k = \min\{i \geq 0 : M_i = 2k\}.$$

Note that if $M_0 = 0$ then $M_i \leq 2k$ for $0 \leq i \leq \psi_k$. For each $0 \leq j < \infty$ we will define a binary-valued Markov-chain $\{X_i^{(j)}\}$ with some finite order, which we denote as $X_i^{(j)} = f^{(j)}(M_i)$ where $f^{(j)}$ will be a $\{0,1\}$ valued function of the state space S. We will also define a process $\{X_i\}$ which we denote as $X_i = f^{(\infty)}(M_i)$ where $f^{(\infty)}$ is also a binary valued function of the state space S, and the time series $\{X_i\}$ will serve as the stationary unpredictable process. For all $0 \leq j \leq \infty$, let $f^{(j)}(0) = 0$, $f^{(j)}(1) = 0$, and $f^{(j)}(s) = 1$ for all even states s. Note that so far we have only defined $f^{(j)}$ partially. We will define the values for the remaining states later on. A feature of this definition of $f^{(j)}(\cdot)$ is that whenever $X_n^{(j)} = 0, X_{n+1}^{(j)} = 0, X_{n+2}^{(j)} = 1$ we know that $M_n = 0$ and vice versa.

Now it is easy to see that if for a certain $0 \leq j \leq \infty$, there is an index K_j such that $f^{(j)}(i) = 1$ for all $i \geq K_j$ then the defined process $\{X_n^{(j)}\}$ is a binary Markov-chain with order not greater than $K_j + 1$.

Now let $f^{(0)}(2k+1) = 1$ for all $k \geq 1$ and so the function $f^{(0)}$ is fully defined. Since $f^{(0)}(i)$ is eventually 1, the defined process $\{X_i^{(0)}\}$ is a stationary ergodic binary Markov-chain with some finite order.

For function $f^{(j)}$ and index $2k$, if $f^{(j)}(i)$ is defined for all $0 \leq i \leq 2k$, then it is easy to see that if $M_0 = 0$ (that is, $f^{(j)}(M_0) = 0$, $f^{(j)}(M_1) = 0$, $f^{(j)}(M_2) = 1$) then $M_i \leq 2k$ for $0 \leq i \leq \psi_k$ and the mapping

$$M_0^{\psi_k} \to (f^{(j)}(M_0), \ldots, f^{(j)}(M_{\psi_k}))$$

is invertible. If we let λ_n operate on process $\{X_i^{(j)}\}$, define

$$A_j(k) = \{M_0 = 0, \psi_k = \lambda_n(X_0^{(j)}, X_1^{(j)}, \ldots) \text{ for some n}\}.$$

Thus as soon as $f^{(j)}(i)$ is defined for all $0 \leq i \leq 2k$ the set $A_j(k)$ is also well defined, it is measurable with respect to $M_0^{\psi_k}$ and depends on state $2k$ and index j which selects the process $\{X_n^{(j)}\}$ on which the stopping times $\{\lambda_n\}$ operate.

Let $N_{-1} = 1$. Notice that $A_0(k)$ is well defined for all k. Now we define $f^{(j)}$ by induction. Assume that for $0 \leq i \leq j-1$ we have already defined a strictly increasing sequence of integers N_{i-1}, and functions $f^{(i)}$ which are eventually constant.

Now we define $f^{(j)}$. Since by assumption $\{X_n^{(j-1)}\}$ is a stationary and ergodic binary-valued Markov process with some finite order, the estimator is assumed to predict eventually on this process and there is a $N_{j-1} > N_{j-2}$ such that

$$P(A_{j-1}(N_{j-1})) > \frac{P(M_0 = 0)}{2} = \frac{1}{2}\frac{1-q}{3-2q}.$$

Now for each $j \leq l \leq \infty$ define $f^{(l)}(2m+1)$ for the segment $N_{j-2} \leq m < N_{j-1}$ as follows,

$$f^{(l)}(2m+1) = f^{(j-1)}(2m+1).$$

Notice that now $A_j(N_{j-1})$ is well defined and coincides with $A_{j-1}(N_{j-1})$. We will define $f^{(j)}(2N_{j-1}+1)$ maliciously. Let

$$B_j^+ = A_j(N_{j-1}) \bigcap \{h_n(f^{(j)}(M_0), \ldots, f^{(j)}(M_{\psi_{N_{j-1}}})) \geq \frac{q}{2}\}$$

and

$$B_j^- = A_j(N_{j-1}) \bigcap \{h_n(f^{(j)}(M_0), \dots, f^{(j)}(M_{\psi_{N_{j-1}}})) < \frac{q}{2}\}.$$

Now notice that the sets B_j^+ and B_j^- do not depend on the future values of $f^{(j)}(2r+1)$ for $r \geq N_{j-1}$. One of the two sets B_j^+, B_j^- has at least probability $\frac{P(M_0=0)}{4} = \frac{1}{4}\frac{1-q}{3-2q}$. Now we specify $f^{(j)}(2N_{j-1}+1)$. Let $f^{(j)}(2N_{j-1}+1) = 1$, $I_j = B_j^-$ if $P(B_j^-) \geq P(B_j^+)$ and let $f^{(j)}(2N_{j-1}+1) = 0$, $I_j = B_j^+$ if $P(B_j^-) < P(B_j^+)$.

Because of the construction of $\{M_i\}$, on event I_j,

$$\begin{aligned} & P(X^{(j)}_{\psi_{N_{j-1}}+1} = 1 | X_0^{(j)}, \dots, X^{(j)}_{\psi_{N_{j-1}}}) \\ = \; & f^{(j)}(2N_{j-1}+1) P(X^{(j)}_{\psi_{N_{j-1}}+1} = f^{(j)}(2N_{j-1}+1) | X_0^{(j)}, \dots, X^{(j)}_{\psi_{N_{j-1}}}) \\ = \; & f^{(j)}(2N_{j-1}+1) P(M_{\psi_{N_{j-1}}+1} = 2N_{j-1}+1 | M_0^{\psi_{N_{j-1}}}) \\ = \; & q f^{(j)}(2N_{j-1}+1). \end{aligned}$$

The difference of the estimate and the conditional probability is at least $\frac{q}{2}$ on set I_j and this event occurs with probability not less than $\frac{1}{4}\frac{1-q}{3-2q}$.

Now for all $N_{j-1} < m$ define

$$f^{(j)}(2m+1) = 1.$$

In this way, $\{X_i^{(j)}\}$ is also a stationary and ergodic binary-valued Markov-chain.

Now by induction, we defined all the functions $f^{(j)}$ for $0 \leq j < \infty$. Since $f^{(\infty)}(m) = f^{(j)}(m) = f^{(j-1)}(m)$ for all $0 \leq m \leq 2N_{j-1}$ so we also defined $f^{(\infty)}$.

Finally by Fatou's Lemma,

$$\begin{aligned} & P(\limsup_{n\to\infty} \{|h_n(X_0^{\lambda_n}) - P(X_{\lambda_n+1} = 1 | X_0^{\lambda_n})| \geq q/2\}) \\ \geq \; & P(\limsup_{j\to\infty} I_j) \geq \limsup_{j\to\infty} P(I_j) \geq \frac{1}{4}\frac{1-q}{3-2q}. \end{aligned}$$

Concerning the conditional probability $P(X_1 = 1 | X_{-\infty}^0)$ observe that as soon as one finds the pattern 001 in the sequence $X_{-\infty}^0$ the conditional probability does not depend on previous values. The probability of the occurence of 001 in the past is one since the original Markov chain is ergodic and our process is therefore also ergodic. Thus the process is finitarily Markovian. Now for $n > 8$,

$$\begin{aligned}
P(\tau(X_{-\infty}^0) > n) &\le P(\tau(X_{-\infty}^0) > n, M_0 = 0)\\
&+ P(\tau(X_{-\infty}^0) > n, M_0 = 1) + P(\tau(X_{-\infty}^0) > n, M_0 \ge n-3)\\
&\le P(M_0 = 0, M_{-1} \ge n-3)\\
&+ P(M_0 = 1, M_{-2} \ge n-3) + P(M_0 \ge n-3)\\
&\le 3P(M_0 \ge n-3)\\
&\le 3\sum_{i=n-3}^{\infty} \frac{1-q}{3-2q} q^{i-2}\\
&\le 3\frac{1-q}{3-2q} q^{n-5} \frac{1}{1-q}\\
&= 3\frac{1-q}{(3-2q)(1-q)q^5} q^n.
\end{aligned}$$

Since $0 < q < R$, for arbitrary $q < \rho < R$,

$$P(\tau(X_{-\infty}^0) > n) \le \rho^n$$

eventually. Thus the unpredictable process $\{X_n\}$ is in $FMEXPTAIL(R)$. The proof of Theorem 6.1 is complete. □

7. APPENDIX

The next lemma can be found in Morvai et al. [10]. Note that the entropy rate of the stationary and ergodic process $\{X_n\}$ is defined as $H = \lim_{k\to\infty} -\frac{1}{k}E(\log(p(X_{-k+1}^0)))$.

Lemma 7.1. (Theorem 2 in Morvai et al. [10]) Let $\{X_n\}$ be a stationary and ergodic process with values in a finite set $\mathcal{X}$ and with entropy rate H. Then for arbitrary integers $j_k \ge 1$,

$$\limsup_{k\to\infty} \frac{1}{k} \log\left(\frac{-\zeta_{j_k}^{(k,0)}}{j_k}\right) \le H$$

almost surely.

(Received November 30, 2020)

REFERENCES

[1] P. Algoet: The strong law of large numbers for sequential decisions under uncertainty. IEEE Trans. Inform. Theory *40* (1994), 609–633. DOI:10.1109/18.335876

[2] 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. DOI:10.1109/18.761258

[3] D. H. Bailey: Sequential Schemes for Classifying and Predicting Ergodic Processes. Ph.D. Thesis, Stanford University, 1976.

[4] I. Csiszár and Zs. Talata: Context tree estimation for not necessarily finite memory processes via BIC and MDL. IEEE Trans. Inform. Theory *52* (2006), 3, 1007–1016. DOI:10.1109/TIT.2005.864431

[5] L. Györfi, G. Morvai, and S. Yakowitz: Limits to consistent on-line forecasting for ergodic time series. IEEE Trans. Inform. Theory *44* (1998), 886–892. DOI:10.1109/18.661540

[6] W. Hoeffding: Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. *58* (1963), 13–30.

[7] S. Kalikow, Y. Katznelson, and B. Weiss: Finitarily deterministic generators for zero entropy systems. Israel J. Math. *79* (1992), 33–45. DOI:10.1007/BF02764801

[8] Ph. T. Maker: The ergodic theorem for a sequence of functions. Duke Math. J. *6* (1940), 27–30.

[9] G. Morvai: Guessing the output of a stationary binary time series. In: Foundations of Statistical Inference (Y. Haitovsky, H. R.Lerche, and Y. Ritov, eds.), Physika-Verlag, pp. 207–215, 2003.

[10] G. Morvai, S. Yakowitz, and P. Algoet: Weakly convergent nonparametric forecasting of stationary time series. IEEE Trans. Inform. Theory *43* (1997), 483–498. DOI:10.1109/18.556107

[11] G. Morvai and B. Weiss: Forecasting for stationary binary time series. Acta Appl. Math. *79* (2003), 25–34. DOI:10.1023/A:1025862222287

[12] G. Morvai and B. Weiss: Intermittent estimation of stationary time series. Test *13* (2004), 525–542. DOI:10.1007/BF02595785

[13] G. Morvai and B. Weiss: Inferring the conditional mean. Theory Stochast. Process. *11* (2005), 1–2, 112–120.

[14] G. Morvai and B. Weiss: Prediction for discrete time series. Probab. Theory Related Fields *132* (2005), 1–12. DOI:10.1007/s00440-004-0386-3

[15] G. Morvai and B. Weiss: Limitations on intermittent forecasting. Statist. Probab. Lett. *72* (2005), 285–290. DOI:10.1016/j.spl.2004.12.016

[16] G. Morvai and B. Weiss: On classifying processes. Bernoulli *11* (2005), 523–532. DOI:10.3150/bj/1120591187

[17] G. Morvai and B. Weiss: Order estimation of Markov chains. IEEE Trans. Inform. Theory *51* (2005), 1496–1497. DOI:10.1109/TIT.2005.844093

[18] G. Morvai and B. Weiss: Forward estimation for ergodic time series. Ann. I. H. Poincaré Probab. Statist. *41* (2005), 859–870. DOI:10.1016/j.anihpb.2004.07.002

[19] G. Morvai and B. Weiss: On estimating the memory for finitarily Markovian processes. Ann. I. H. Poincaré PR *43* (2007), 15–30. DOI:10.1016/j.anihpb.2005.11.001

[20] G. Morvai and B. Weiss: On sequential estimation and prediction for discrete time series. Stoch. Dyn. *7* (2007), 4, 417–437. DOI:10.1142/s021949370700213x

[21] G. Morvai and B. Weiss: Estimating the lengths of memory words. IEEE Trans. Inform. Theory *54* (2008), 8, 3804–3807. DOI:10.1109/TIT.2008.926316

[22] G. Morvai and B. Weiss: On universal estimates for binary renewal processes. Annals Appl. Probab. *18* (2008), 5, 1970–1992. DOI:10.1214/07-aap512

[23] G. Morvai and B. Weiss: Estimating the residual waiting time for binary stationary time series. Proc. ITW2009, Volos 2009, pp. 67–70.

[24] G. Morvai and B. Weiss: A note on prediction for discrete time series. Kybernetika *48* (2012), 4, 809–823.

[25] G. Morvai and B. Weiss: Universal tests for memory words. IEEE Trans. Inform. Theory *59* (2013), 6873–6879. DOI:10.1109/TIT.2013.2268913

[26] G. Morvai and B. Weiss: Inferring the residual waiting time for binary stationary time series. Kybernetika *50* (2014), 869–882. DOI:10.14736/kyb-2014-6-0869

[27] G. Morvai and B. Weiss: A versatile scheme for predicting renewal times. Kybernetika *52* (2016), 348–358. DOI:10.14736/kyb-2016-3-0348

[28] G. Morvai and B. Weiss: Universal rates for estimating the residual waiting time in an intermittent way. Kybernetika *56*, (2020), 4, 601–616. DOI:10.14736/kyb-2020-4-0601

[29] G. Morvai and B. Weiss: On universal algorithms for classifying and predicting stationary processes. Probab. Surveys *18* (2021), 77–131.

[30] G. Morvai and B. Weiss: Consistency, integrability and asymptotic normality for some intermittent estimators. ALEA, Lat. Am. J. Probab. Math. Stat. *18* (2021), 1643–1667. DOI:10.30757/ALEA.v18-60

[31] B. Ya. Ryabko: Prediction of random sequences and universal coding. Problems Inform. Trans. *24* (1988), 87–96. DOI:10.1080/10256018808623911

[32] D. Ryabko: Asymptotic Nonparametric Statistical Analysis of Stationary Time Series. Springer, Cham 2019.

[33] P. C. Shields: The Ergodic Theory of Discrete Sample Paths. In: Graduate Studies in Mathematics. American Mathematical Society *13*, Providence 1996.

[34] J. Suzuki: Universal prediction and universal coding. Systems Computers Japan *34* (2003), 6, 1–11. DOI:10.1002/scj.10357

[35] H. Takahashi: Computational limits to nonparametric estimation for ergodic processes. IEEE Trans. Inform. Theory *57* (2011), 6995–6999. DOI:10.1109/TIT.2011.2165791

Gusztáv Morvai, Budapest. Hungary.
e-mail: morvaigusztav@gmail.com

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