ALEA, Lat. Am. J. Probab. Math. Stat. **21**, 1073–1099 (2024)
DOI: 10.30757/ALEA.v21-42

Countable alphabet stationary processes with at least one memory word and intermittent estimation with universal rates

Gusztáv Morvai and Benjamin Weiss

Alfréd Rényi Institute of Mathematics, 13-15 Reáltanoda utca, H-1053, Budapest, Hungary
E-mail address: morvaigusztav@gmail.com
URL: http://math.bme.hu/∼morvai

Hebrew University of Jerusalem, Institute of Mathematics,, Jerusalem 91904 Israel
E-mail address: weiss@math.huji.ac.il
URL: http://mathematics.huji.ac.il/people/benjamin-weiss

Abstract. We present here a number of results that provide universal rates of convergence for certain non parametric estimation problems. For example consider the class $\mathcal{C}$ of all finite order Markov chains on a countable alphabet and the problem of estimating the conditional distribution of X_{n+1} given the first n outputs of the process. We will give a sequence of stopping times with density one and estimators at those times such that almost surely our estimators will eventually differ from the true conditional distribution by no more than a certain fixed sequence tending to zero. Similar results are given for estimating the conditional expectation of X_{n+1} given the first n outputs, but here some additional moment conditions are required. An example shows that this is not possible in general.

1. Introduction

The basic problem that we will be considering in this paper is how to extract information from the first n outputs of a stationary ergodic process $\{X_n\}$ over a discrete (finite or countably infinite) alphabet in order to say as much as possible about the next output. Our interest is in a non-parametric situation where the only information that we have about the process is of a rather general nature. Furthermore, our goal is to obtain point wise quantitative results that would be valid for almost any sequence of outputs. By a quantitative result we mean one in which some rate is given for the convergence to zero of the error in our estimation scheme. In the past such results have been given for the class of iid processes or k-step Markov chains with a finite number of states where the goal was simply to estimate the order of the Markov chain (e.g. cf. Finesso et al. (1996), Csiszár and Shields (2000), Csiszár (2002)). In this paper our goal is to estimate either the

Received by the editors November 5th, 2023; accepted April 3rd, 2024.
2010 *Mathematics Subject Classification.* 62G05, 60G25,60G10.
Key words and phrases. nonparametric estimation, stationary processes.
The first author was supported partly by Alfréd Rényi Institute of Mathematics, Bolyai János Research Scholarship and OTKA grant No. K75143.

conditional distribution or the conditional expectation of the next output. We also plan to deal with a much wider class that includes the finitarily Markovian (FM) processes (cf. Morvai and Weiss (2007a), Csiszár and Talata (2006)). To describe this class we need the notion of a memory word w. A word w of length k in the alphabet (finite or countable) of the state space of our process is called a memory word if given that current k outputs equal 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. Any process with a renewal state has such memory words. We will suppose that we have some information about the class of memory words but that we do not know the joint distribution defining the process. We have treated the special cases of restricted subclasses of finite alphabet FM processes in some detail in earlier works (cf. Morvai and Weiss (2021b), Morvai and Weiss (2007b), Morvai and Weiss (2021c)). The problem with the assumption that the process is FM is that one can not test if the process is FM or not, cf. Morvai and Weiss (2005c). In this paper we will not assume that the process is FM, instead, we will assume that we are given a set of words which may contain some memory words and some words with zero probability. We will also assume that this set contains at least one word with positive probability. It is very important that there is a rule for testing these assumptions, cf. Morvai and Weiss (2013). This fact will make our schemes more practical.

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 in the class of all FM process (even over binary alphabet) in the point wise sense (almost sure convergence), (cf. Bailey (1976), Ryabko (1988), Algoet (1999), Györfi et al. (1998), Morvai and Weiss (2021b), Morvai and Weiss (2005b)) and so intermittent estimation has been introduced (cf. Morvai and Weiss (2007b), Morvai and Weiss (2021c), Morvai and Weiss (2005e), Morvai and Weiss (2021a)). Here the idea is to estimate only along a sequence of carefully chosen stopping times. This is what we will do in this paper. This relatively new approach is in contrast to other more traditional approaches where estimates are given for all time instances but almost sure convergence is modified to convergence in probability or almost sure convergence of Cesaro averages, cf. Bailey (1976), Algoet (1994, 1999), Felber et al. (2013), Merhav and Feder (1998), Györfi and Lugosi (2002), Györfi and Ottucsák (2007), Györfi et al. (2012), Jones et al. (2012), Bunea and Nobel (2008), Hanneke (2020, 2021), and Nobel (2003).

Our first main result will give a sequence of stopping times and a sequence of estimators for the conditional distribution of the next output such that eventually almost surely the maximal deviation between the estimator and the true conditional distribution function is at most an explicit sequence tending to zero. In general the density of the stopping times will equal the probability of encountering a memory word and if this probability is one the density of the stopping times will be one. Our results for estimating conditional expectations will require some moment conditions and while the ones we give may not be optimal we do show that some higher moment condition is necessary. The results we have just described are valid for countable alphabet processes. As an application we give an estimation scheme with explicit rates for the error which will converge almost surely for any finite order countable state Markov chain. To this end we will need a technical extension of these results to the case where we are not given a fixed list of memory words, but rather a sequence of such lists which will eventually stabilize. This convergence will be obtained by using any one of the known schemes for consistent estimation of the order of a Markov chain (see e.g. Morvai and Weiss (2005d) and Morvai and Weiss (2008)).

We will show that existence of a moment is not sufficient by the following construction. Starting from any universal estimation scheme for the conditional mean of the next output which is point wise consistent for all simple Markov chains with state space a finite subset of the integers we construct a countable state Markov chain on a subset of the integers for which the estimation scheme diverge with positive probability.

In Section 2.1 we will introduce the basic definitions e.g. memory words and in Section 2.2 we will give illustrative examples. In Section 3.1 we define the intermittent schemes and the stopping times which will be used for both kinds of estimators. The results for the asymptotic behaviour of the stopping times will be given in Section 3.2, the results for the conditional distribution will be given in Section 3.3 while the ones for the conditional expectations are in Section 3.4. The general results for the case when no prior knowledge on the memory words are assumed are in Section 4.1. Section 4.2 has the result applied to Markov chains on a countable alphabet of unknown order. Section 5 contains the construction demonstrating the necessity of restricting the class of processes in some way. In Section 6 we give a short summary of the key points of the paper.

2. Preliminaries

2.1. *Basic Definitions.*

First let us fix the notation. 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 $m \leq n$. 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.

Definition 2.1. For a $k \geq 0$ 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 $p(z_{-k-i+1}^{-k}, w_{-k+1}^0, y) > 0$.

Note that for $k = 0$, w_{-k+1}^0 is the empty word with $p(w_{-k+1}^0) = 1 > 0$ and $p(y|w_{-k+1}^0) = p(y)$.

Definition 2.2. For a $k \geq 0$ we say that w_{-k+1}^0 is a minimal memory word if it is a memory word and no proper suffix of w_{-k+1}^0 is also a memory word.

Definition 2.3. Define the set $\mathcal{W}_k$ of those memory words w_{-k+1}^0 with length k, that is,

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

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

Definition 2.4. Define the set of all memory words $\mathcal{W}$ as

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

In this paper we will assume that the discrete stationary process $\{X_n\}_{n=-\infty}^{\infty}$ posseses at least one memory word, that is,

$$P(\exists 0 \leq k < \infty : X_{-k+1}^0 \in \mathcal{W}) > 0. \tag{2.1}$$

In our previous paper Morvai and Weiss (2021b) we assumed that the process was from a smaller class of stationary processes (the so called finitarily Markovian processes) where the above probability is one.

We will need a real valued nonnegative strictly monotone increasing continuous function $f(t)$ defined for the real numbers $t > 1$ with the properties

$$\lim_{t\to\infty} f(t) = \infty \tag{2.2}$$

and

$$\lim_{t\to\infty} \frac{f(t)}{t} = 0. \tag{2.3}$$

Let g denote the inverse of f that is

$$g(t) = f^{-1}(t). \tag{2.4}$$

Notice that by (2.2), (2.4) and (2.3) we get that

$$\lim_{t\to\infty} \frac{g(t)}{t} = \lim_{t\to\infty} \frac{g(f(t))}{f(t)} = \lim_{t\to\infty} \frac{t}{f(t)} = \infty. \tag{2.5}$$

The basic examples of such functions are $f(t) = t^{1/\beta}$ for $\beta > 1$ and $f(t) = \log t$. We have written out the results and proofs for general f but the reader can keep in mind one of these examples to get a more concrete picture of the situation.

Fix an arbitrary real valued function

$$s : \mathcal{X} \to \mathbb{R}. \tag{2.6}$$

We will estimate the conditional distribution function and the conditional expectation for $s(X_n)$. Note that since $\mathcal{X}$ is discrete, the function $s(X_n)$ may take on at most countably infinite values. No measurability problem can arise here. In some of the results we will assume some moment conditions for $s(X_n)$.

2.2. *Examples.* Here we give some illustrative examples for countable alphabet stationary and ergodic processes having at least one memory word. (We take some of our constructions from Morvai and Weiss (2021c) and make alterations when necessary in order to get a process over a countably infinite alphabet.)

In the first example the process $\{Z_n\}_{n=-\infty}^{\infty}$ will be independent and identically distributed. Recall that $\mathcal{W}$ is the set of all memory words defined in Section 2.1.

Example 2.5. (Cf. Example 2.1 in Morvai and Weiss Morvai and Weiss (2021c)) Consider an independent and identically distributed process $\{Z_n\}_{n=-\infty}^{\infty}$ on a countable alphabet. Then the empty word is a memory word and it is the only minimal memory word. Any word which has positive probability is a memory word. Thus $P(\exists 0 \le k < \infty : Z_{-k+1}^0 \in \mathcal{W}) = 1$.

In the next example the process $\{Z_n\}_{n=-\infty}^{\infty}$ will be a Markov chain with order 2. Note that for Markov processes $P(\exists 0 \le k < \infty : Z_{-k+1}^0 \in \mathcal{W}) = 1$. (We will use the construction in Example 2.3 in Morvai and Weiss (2021c). The process there is over a binary alphabet while here it will be over a countably infinite alphabet.)

Example 2.6. 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 $\{M_n\}_{n=-\infty}^{\infty}$. Let $\{Y_n\}_{n=-\infty}^{\infty}$ be independent and identically distributed taking values from the set of positive integers such that Y_0 takes each positive

integer for a value with some positive probability. We further assume that the two processes, $\{M_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ are independent from each other. Define

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

Then $\{Z_n\}_{n=-\infty}^{\infty}$ is a stationary and ergodic countably infinite alphabet Markov chain with order 2. The minimal memory words of the process $\{Z_n\}_{n=-\infty}^{\infty}$ are these: all positive integers and all words of the form $x0$ where $x \in \{0, 1, 2, \dots\}$.

In the next example the process $\{Z_n\}_{n=-\infty}^{\infty}$ will not be Markov of any order but still $P(\exists 0 \le k < \infty : Z_{-k+1}^0 \in \mathcal{W}) = 1$. (We will use the method in Example 2.5 in Morvai and Weiss (2021c). The process there is over a binary alphabet while here it will be over a countably infinite alphabet.)

Example 2.7. Consider the Markov chain 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 $\{M_n\}_{n=-\infty}^{\infty}$. Let $\{Y_n\}_{n=-\infty}^{\infty}$ be independent and identically distributed such that Y_0 takes each positive integer for a value with some positive probability. We further assume that the two processes, $\{M_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ are independent from each other. Define $Z_n = I_{\{M_n\neq 0\}}Y_n$. Then $\{Z_n\}_{n=-\infty}^{\infty}$ is a stationary and ergodic countably infinite alphabet process. The minimal memory words of the process $\{Z_n\}_{n=-\infty}^{\infty}$ are these: 0 and those words of the form w_{-k+1}^0 where $k \ge 1$, $w_{-k+1} = 0$, and $w_i \in \{1, 2, \dots\}$ for $-k + 1 < i \le 0$. Note that $\{Z_n\}_{n=-\infty}^{\infty}$ is not Markov of any order but $P(\exists 0 \le k < \infty : Z_{-k+1}^0 \in \mathcal{W}) = 1$.

For a concrete countably infinite alphabet process $\{Z_n\}_{n=-\infty}^{\infty}$ which possesses memory words but

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

consider the following example.

Example 2.8. Consider the periodic Markov chain M_n with four states $\{A, B, C, D\}$ where from state A it passes to B, from B to C, from C to D, from D to A with probability one. This yields a stationary and ergodic process $\{M_n\}_{n=-\infty}^{\infty}$ with uniform marginal distribution on the four states. Define the stationary and ergodic process $\{Z_n\}_{n=-\infty}^{\infty}$ as follows. We randomize at each state of the Markov chain. That is, for $k = 0, 1, 2, 3, \dots$,

$$P(Z_n = k|M_i \ -\infty < i < \infty) = P(Z_n = k|M_n)$$

and $\{Z_n\}_{n=-\infty}^{\infty}$ are conditionally independent given $\{M_n\}_{n=-\infty}^{\infty}$. Now we define the conditional distributions. Let

$$P(Z_n = 0|M_n = B) = P(Z_n = 0|M_n = D) = 1,$$
$$P(Z_n = k|M_n = A) = (0.9)^{k-1}(0.1) \text{ for } k = 1, 2, 3, \dots$$

and

$$P(Z_n = k|M_n = C) = (0.5)^{k-1}(0.5) \text{ for } k = 1, 2, 3, \dots.$$

The minimal memory words of the countably infinite alphabet stationary and ergodic process $\{Z_n\}_{n=-\infty}^{\infty}$ are the positive integers. Note that, since the letter zero has positive probability, we have $P(\exists 0 \le k < \infty : Z_{-k+1}^0 \in \mathcal{W}) < 1$.

For more examples see Morvai and Weiss (2021c).

3. Intermittent Estimators Dependent on Prior Knowledge of a Mnemonic Set $\mathcal{V}$

3.1. *Definition of the Intermittent Schemes.* Recall the notion of memory word and the set $\mathcal{W}$ of all memory words from Section 2.1.

Let $\mathcal{V}$ denote an arbitrary set of words such that it may only contain words with zero probability and memory words. That is,

$$\mathcal{V} \subseteq \mathcal{W} \bigcup \left(\bigcup_{k=1}^{\infty} \{x_{-k+1}^0 \in \mathcal{X}^k : \ p(x_{-k+1}^0) = 0\} \right). \tag{3.1}$$

We will call such set $\mathcal{V}$ a mnemonic set. Our intermittent estimation scheme, the stopping times etc. will depend on this mnemonic set $\mathcal{V}$. Note that one can test if a given set of words has this property or not, cf. Morvai and Weiss (2013). (It is important that we allow words with zero probability to be in a mnemonic set. We do not just have weaker conditions in this way but we have also a test for a set being a mnemonic set. If we required that each and every word in the set posses positive probability then this property could not be tested in case of sets containing infinitely many words. Simply one could not know if a word in the set has not yet appeared becose it has zero probability or it has just small probability.) We will assume that

$$P(\exists 0 \leq k < \infty : X_{-k+1}^0 \in \mathcal{V}) > 0. \tag{3.2}$$

That is, we will assume that the mnemonic set $\mathcal{V}$ contains at least one memory word. In other words, we will assume that the mnemonic set contains at least one word with positive probability. (It is easy to test if a mnemonic set has positive probability. As soon as you see a word from the mnemonic set in the data you know that the mnemonic set has positive probability. As long as no word from the mnemonic set appears in the data, the test says that the mnemonic set has zero probability.)

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

$$\tau(X_{-\infty}^n) = \inf\{t \geq 0 \ : \ X_{n-t+1}^n \in \mathcal{V}\}. \tag{3.3}$$

Note that $\tau(X_{-\infty}^n)$ depends on the mnemonic set $\mathcal{V}$ but we suppress this dependence to keep the notation manageable.

$\tau(X_{-\infty}^n)$ is identifying the first time we see an element of $\mathcal{V}$ as we scan the sequence $X_{-\infty}^n$ from right to left. Note that this is well defined with probability one with the understanding that if there is no such t then the infimum is infinite. We will use the notation $\tau(X_0^n)$ with the understanding that this is defined if $\tau(X_{-\infty}^n) \leq n+1$ and then $\tau(X_0^n) = \tau(X_{-\infty}^n)$.

Note that $P(\tau(X_{-\infty}^n) < \infty) = P(\exists 0 \leq k < \infty : X_{-k+1}^0 \in \mathcal{V})$.

Recall that the function f and its inverse g were defined in Section 2.1. Let $0 < \gamma < 1$ be arbitrary. Now we introduce the stopping times λ_n. First we give an informal definition in words and then a formal one by a mathematical formula. Where the stopping times $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n < \ldots$ stop, there we will be willing to give estimates. Let $\lambda_0 = 1$. For $n \geq 1$, λ_n will be the smallest $t > \lambda_{n-1}$ for which three conditions hold. First, we require that t be large enough so that $f(t) < t$. We require this because we will divide the observed data up to time t

$$X_i : 0 \leq i \leq t$$

into two parts and the upper end point of the first data segment

$$X_i : 0 \leq i < \lfloor f(t) \rfloor$$

will be given by $\lfloor f(t) \rfloor$. (Sometimes we will refer to the data segment by giving the indices, here $[0, \lfloor f(t) \rfloor)$ only.) This is why we demand that $f(t) < t$. Our second condition is that at time t, looking backwards, we should see a word from the mnemonic set $\mathcal{V}$, and this word should appear

fully in the first data segment mentioned above. In this way we will restrict the set of possible words which we allow to be seen at time λ_n. This restriction will be crucial in the proofs. (Instead of a possibly infinite set we will have to deal with a finite set with a bound on the number of elements in it.) Our third condition is that in the second data segment

$$X_i : \lfloor f(t) \rfloor < i < \lfloor g(\lfloor f(t) \rfloor) \rfloor$$

we shall see sufficiently many times the aforementioned word which satisfied our second condition. We require this because our estimate will be based on frequency counts and for a reliable estimate we need sufficiently many occurrences of the word. Now we give the formal definition.

Define the stopping times λ_n as $\lambda_0 = 1$ and for $n \geq 1$,

$$\begin{aligned}\lambda_n = \min\{\ & t > \lambda_{n-1} \ : f(t) < t, \\ & \left(\exists 0 \leq i < \lfloor f(t) \rfloor \ : \tau(X_{-\infty}^t) \leq i,\ \tau(X_0^i) = \tau(X_0^t), X_{i-\tau(X_0^i)+1}^i = X_{t-\tau(X_0^t)+1}^t\right) \text{ and} \\ & \left|\left\{ \lfloor f(t) \rfloor < j < \lfloor g(\lfloor f(t) \rfloor) \rfloor \ : \ \tau(X_0^j) = \tau(X_0^t), X_{j-\tau(X_0^j)+1}^j = X_{t-\tau(X_0^t)+1}^t \right\}\right| \\ & \geq \lceil g(\lfloor f(t) \rfloor)^{(1-\gamma)} \rceil\}. \end{aligned} \tag{3.4}$$

Note that the event $\{\lambda_n = t\}$ is measurable with respect to X_0^t. Note that in the definition of λ_n we require that the word $X_{t-\tau(X_0^t)+1}^t$ seen at time t appear at least once in the first part of the data segment and sufficiently many times in the second part of the data segment. Note also that this sequence of stopping times depends on the set $\mathcal{V}$ but we suppress the dependence to keep the notation manageable. Notice that λ_n is finite almost surely since all three conditions will be true for large enough t. ($f(t) < t$ is true for all large enough t by the assumption $\lim_{t\to\infty} \frac{f(t)}{t} = 0$ in (2.3) in Section 2.1. Since any word with positive probability in $\mathcal{V}$ will appear in the first data segment $[0, \lfloor f(t) \rfloor)$ eventually by ergodicity and the assumption $\lim_{t\to\infty} f(t) = \infty$ in (2.2) in Section 2.1, and any word with positive probability in $\mathcal{V}$ will appear with the frequency equal to its probability eventually almost surely by ergodicity and so more than $\lceil g(\lfloor f(t) \rfloor)^{(1-\gamma)} \rceil < t^{(1-\gamma)} + 1 = t(t^{-\gamma} + t^{-1})$ times eventually almost surely in the second data segment $(\lfloor f(t) \rfloor, \lfloor g(\lfloor f(t) \rfloor) \rfloor)$ by the assumption that the size of the first data segment is negligible, that is, $\lim_{t\to\infty} \frac{f(t)}{t} = 0$ in (2.3) in Section 2.1.)

Put

$$\begin{aligned}\kappa_n = \min\{t : & \left|\left\{ \lfloor f(\lambda_n) \rfloor < j \leq t : \tau(X_0^j) = \tau(X_0^{\lambda_n}), X_{j-\tau(X_0^j)+1}^j = X_{\lambda_n-\tau(X_0^{\lambda_n})+1}^{\lambda_n} \right\}\right| \\ & = \lceil g(\lfloor f(\lambda_n) \rfloor)^{(1-\gamma)} \rceil\}. \end{aligned} \tag{3.5}$$

Note that $\kappa_n \leq \lambda_n$ and $(\lfloor f(\lambda_n) \rfloor, \kappa_n]$ is a time interval in which there are exactly

$$\lceil g(\lfloor f(\lambda_n) \rfloor)^{(1-\gamma)}$$

pieces of occurrences of $X_{\lambda_n-\tau(X_0^{\lambda_n})+1}^{\lambda_n}$. (This will ensure that we will take into consideration exactly $\lceil g(\lfloor f(\lambda_n) \rfloor)^{(1-\gamma)} \rceil$ pieces of occurrences in defining the estimators $\hat{F}_{\lambda_n}(z)$ and $h_n(X_0^{\lambda_n})$.) Define the empirical conditional distribution function $\hat{F}_{\lambda_n}(z)$ as

$$\hat{F}_{\lambda_n}(z) = \frac{\sum_{i=\lfloor f(\lambda_n) \rfloor + 1}^{\kappa_n} I_{\left\{\tau(X_0^i) = \tau(X_0^{\lambda_n}), X_{i-\tau(X_0^i)+1}^i = X_{\lambda_n-\tau(X_0^{\lambda_n})+1}^{\lambda_n}, s(X_{i+1}) \leq z\right\}}}{\lceil g(\lfloor f(\lambda_n) \rfloor)^{(1-\gamma)} \rceil}. \tag{3.6}$$

What we are doing here is to look at all the occurrences of $X_{\lambda_n-\tau(X_0^{\lambda_n})+1}^{\lambda_n}$ in the second part of the data segment $(\lfloor f(\lambda_n) \rfloor, \kappa_n]$, where there are many such occurrences, and observing the value of the immediately following variable and then computing the empirical distribution of these values.

For $n > 0$ define the estimator $h_n(X_0^{\lambda_n})$ for $E(s(X_{\lambda_n+1})|X_0^{\lambda_n})$ at time λ_n as

$$h_n(X_0^{\lambda_n}) = \frac{\sum_{i=\lfloor f(\lambda_n)\rfloor+1}^{\kappa_n} I_{\{\tau(X_0^i)=\tau(X_0^{\lambda_n}),X^i_{i-\tau(X_0^i)+1}=X^{\lambda_n}_{\lambda_n-\tau(X_0^{\lambda_n})+1}\}} s(X_{i+1})}{\lceil g(\lfloor f(\lambda_n)\rfloor)^{(1-\gamma)}\rceil}. \tag{3.7}$$

3.2. *Results on the Asymptotic Behaviour of the Stopping Times.* Recall the notion of memory word from Section 2.1. Recall the mnemonic set $\mathcal{V}$ and the stopping time λ_n from Section 3.1.

Now we examine how fast the stopping times λ_n grow. Note that by construction $\lambda_n \geq n$. Note also that if the mnemonic set $\mathcal{V}$ contains at least one memory word then $0 < P(\bigcup_{k=0}^{\infty}\{X^0_{-k+1} \in \mathcal{V}\})$ since any memory word has positive probability by definition.

Theorem 3.1. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. We assume that the process possesses at least one memory word and the mnemonic set $\mathcal{V}$ contains at least one memory word. Let $0 < \gamma < 1$. Then for the stopping times λ_n,*

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = \frac{1}{P(\bigcup_{k=0}^{\infty}\{X^0_{-k+1} \in \mathcal{V}\})}, \tag{3.8}$$

almost surely.

Proof: Since $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n$ and the stopping times $\lambda_1, \lambda_2, \ldots, \lambda_n$ stop only at certain occurrences of the mnemonic set $\mathcal{V}$, it is immediate that

$$\frac{n}{\lambda_n} \leq \frac{\sum_{i=1}^{\lambda_n} I_{\{\bigcup_{k=0}^{\infty}\{X^i_{i-k+1}\in\mathcal{V}\}\}}}{\lambda_n}.$$

By ergodicity,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{\lambda_n} I_{\{\bigcup_{k=0}^{\infty}\{X^i_{i-k+1}\in\mathcal{V}\}\}}}{\lambda_n} = P(\bigcup_{k=0}^{\infty}\{X^0_{-k+1} \in \mathcal{V}\})$$

almost surely. Thus

$$\limsup_{n\to\infty} \frac{n}{\lambda_n} \leq P(\bigcup_{k=0}^{\infty}\{X^0_{-k+1} \in \mathcal{V}\})$$

almost surely. It is easy to see that

$$\liminf_{n\to\infty} \frac{n}{\lambda_n} \geq P(\bigcup_{k=0}^{\infty}\{X^0_{-k+1} \in \mathcal{V}\}).$$

Indeed since if a certain fixed word in $\mathcal{V}$ has positive probability it will appear with that frequency in the observed sample path. The stopping times $0 < \lambda_1 < \lambda_2 < \ldots$ will stop at all (except finitely many) occurrences of this word eventually almost surely since all the conditions in the definition of the stopping times $0 < \lambda_1 < \lambda_2 < \ldots$ in Section 3.1 will be satisfied at the occurrences of this word eventually almost surely. Namely, $f(t) < t$ is true for all large enough t by the assumption $\lim_{t\to\infty} f(t)/t = 0$ in (2.3) in Section 2.1. This fixed word with positive probability will appear in the first data segment $[0, \lfloor f(t)\rfloor)$ eventually almost surely by ergodicity and the assumption $\lim_{t\to\infty} f(t) = \infty$ in (2.2) in Section 2.1, and this fixed word with positive probability will appear more than $\lceil g(\lfloor f(t)\rfloor)^{(1-\gamma)}\rceil < t^{(1-\gamma)} + 1$ times eventually almost surely in the second data segment $(\lfloor f(t)\rfloor, \lfloor g(\lfloor f(t)\rfloor)\rfloor)$ by ergodicity and the assumption $\lim_{t\to\infty} f(t)/t = 0$ in (2.3) in Section 2.1 (i.e. the length of the first part of the data segment is negligible compared to the second part). Thus the stopping times $0 < \lambda_1 < \lambda_2 < \ldots$ eventually will stop at the occurrences of this word. Thus $\liminf_{n\to\infty} n/\lambda_n$ can not be smaller almost surely than the frequency of this fixed word in the sample path which, by ergodicity, is its probability. This can be said about each word with positive

probability in $\mathcal{V}$. Thus for any finite subset V of $\mathcal{V}$ the stopping times $0 < \lambda_1 < \lambda_2 < \dots$ eventually will stop at all (except finitely many) occurrences of this finite set and so $\liminf_{n\to\infty} n/\lambda_n$ can not be smaller almost surely than the frequency of V in the sample path which, by ergodicity, is its probability $P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in V\})$. Now choose an increasing sequence of finite subsets of $\mathcal{V}$, $V_1 \subseteq V_2 \subseteq V_3 \dots$ such that $\bigcup_{i=1}^{\infty} V_i = \mathcal{V}$. We have already proved that for any $i \geq 1$,

$$P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in V_i\}) \leq \liminf_{n\to\infty} \frac{n}{\lambda_n}$$

almost surely. Now since $V_i \uparrow \mathcal{V}$, we have that

$$P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in V_i\}) \uparrow P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\}).$$

Thus

$$P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\}) \leq \liminf_{n\to\infty} \frac{n}{\lambda_n}$$

almost surely and together with the bound on the lim sup this gives the result. The proof of Theorem 3.1 is complete.

In our earlier papers Morvai (2003) and Morvai and Weiss (2021a), though the schemes there are universal, that is, in those paper there are no other conditions just stationarity and ergodicity, the growth of the stopping times defined there is extremely fast, faster than any exponential growth. It means, that the stopping times there stop very rarely. Thus the schemes in Morvai (2003) and Morvai and Weiss (2021a) are not practical at all. In contrast, the stopping times in this paper have a certain density.

Remark 3.2. Since $0 < \lambda_1 < \lambda_2 < \dots < \lambda_n$, one can interpret the above theorem as

$$P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\})$$

is the asymptotic density of the stopping times $0 < \lambda_1 < \lambda_2 < \dots < \lambda_n$ in $\{1, 2, \dots, \lambda_n\}$, that is,

$$\lim_{n\to\infty} \frac{n}{\lambda_n} = P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\})$$

almost surely. This indicates the growth of the stopping times, that is, for arbitrary $\epsilon > 0$, eventually almost surely, the stopping time λ_n will be between

$$n\left(\frac{1}{P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\})} - \epsilon\right) < \lambda_n < n\left(\frac{1}{P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\})} + \epsilon\right).$$

In our earlier paper Morvai and Weiss (2005e) we treated a special subclass of stationary and ergodic processes and there the asymptotic density of the stopping times is random, it depends on the actual sample path. In the present paper the asymptotic density of the stopping times is not random, it does not depend on the actual sample path.

3.3. *Results for Estimating the Conditional Distribution.* Recall the notion of memory word from Section 2.1. Recall the mnemonic set $\mathcal{V}$, the stopping time λ_n and the estimator $\hat{F}_{\lambda_n}(z)$ from Section 3.1. Recall that function f and its inverse g were defined in Section 2.1.

Theorem 3.3. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. We assume that the process possesses at least one memory word and the mnemonic set $\mathcal{V}$ contains at least one memory word. Let $0 < \gamma < 1$. Assume a sequence of positive real numbers ϵ_m such that*

$$\lim_{m\to\infty} \epsilon_m = 0$$

and

$$me^{-2g(m)^{(1-\gamma)}\epsilon_m^2}$$

is summable. Then for the estimator $\hat{F}_{\lambda_n}(z)$

$$\sup_{-\infty<z<\infty} \left|\hat{F}_{\lambda_n}(z) - P(s(X_{\lambda_n+1}) \le z | X_0^{\lambda_n})\right| \le \epsilon_{\lfloor f(\lambda_n)\rfloor} \tag{3.9}$$

eventually almost surely.

Proof:

Let $0 \le k < m$ be fixed. Recall the definition of the look back time $\tau(X_{-\infty}^k)$ from Section 3.1. Assume $\tau(X_{-\infty}^k) < \infty$. Define $j_0^{(k,m)} = m$ and for $i \ge 0$ let $j_{i+1}^{(k,m)}$ be defined as

$$j_{i+1}^{(k,m)} = \min\left\{t > j_i^{(k,m)} : \tau(X_{-\infty}^t) = \tau(X_{-\infty}^k), X_{t-\tau(X_{-\infty}^t)+1}^t = X_{k-\tau(X_{-\infty}^k)+1}^k\right\}. \tag{3.10}$$

Now for $i \ge 1$ define

$$Z_i^{(k,m)} = s(X_{j_i^{(k,m)}+1}).$$

Clearly, for fixed $k < m$, $\{Z_i^{(k,m)}\}_{i=1}^{\infty}$ are conditionally independent and identically distributed given $X_{-\infty}^k$ with $\tau(X_{-\infty}^k) < \infty$. Apply Lemma 7.1 in the Appendix (Theorem 11.6 in Kosorok Kosorok (2008)) to get that on the set $\{\tau(X_{-\infty}^k) < \infty\}$

$$P\left(\sup_{-\infty<z<\infty}\left|\frac{\sum_{i=1}^{\lceil g(m)^{(1-\gamma)}\rceil} I_{\{Z_i^{(k,m)}\le z\}}}{\lceil g(m)^{(1-\gamma)}\rceil} - P(Z_1^{(k,m)} \le z | X_{-\infty}^k)\right| > \epsilon_m | X_{k-\tau(X_{-\infty}^k)+1}^k\right)$$
$$\le 2e^{-2g(m)^{(1-\gamma)}\epsilon_m^2}.$$

After integrating both sides with respect to the conditioning on the set $\{\tau(X_{-\infty}^k) < \infty\}$ we get that

$$P\left(\sup_{-\infty<z<\infty}\left|\frac{\sum_{i=1}^{\lceil g(m)^{(1-\gamma)}\rceil} I_{\{Z_i^{(k,m)}\le z\}}}{\lceil g(m)^{(1-\gamma)}\rceil} - P(Z_1^{(k,m)} \le z | X_{-\infty}^k)\right| > \epsilon_m, \tau(X_{-\infty}^k) < \infty\right)$$
$$\le 2e^{-2g(m)^{(1-\gamma)}\epsilon_m^2}.$$

Now

$$P(\exists 0 \le k < m :$$
$$\sup_{-\infty<z<\infty}\left|\frac{\sum_{i=1}^{\lceil g(m)^{(1-\gamma)}\rceil} I_{\{Z_i^{(k,m)}\le z\}}}{\lceil g(m)^{(1-\gamma)}\rceil} - P(Z_1^{(k,m)} \le z | X_{-\infty}^k)\right| > \epsilon_m, \tau(X_{-\infty}^k) < \infty)$$
$$\le 2me^{-2g(m)^{(1-\gamma)}\epsilon_m^2}.$$

which is summable and so by the Borel-Cantelli lemma, eventually almost surely, for all $0 \leq k < m$,

$$\tau(X_{-\infty}^{k}) < \infty$$

implies that

$$\sup_{-\infty<z<\infty} \left| \frac{\sum_{i=1}^{\lceil g(m)^{(1-\gamma)} \rceil} I_{\{Z_i^{(k,m)} \leq z\}}}{\lceil g(m)^{(1-\gamma)} \rceil} - P(Z_1^{(k,m)} \leq z | X_{-\infty}^{k}) \right| \leq \epsilon_m.$$

Observe that for suitable $0 \leq k < \lfloor f(\lambda_n) \rfloor$ and $m = \lfloor f(\lambda_n) \rfloor$:

$$\tau(X_{-\infty}^{k}) = \tau(X_0^k) < \infty$$

and for all z

$$\hat{F}_{\lambda_n}(z) = \frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma} \rceil} I_{\{Z_i^{(k,m)} \leq z\}}}{\lceil g(m)^{1-\gamma} \rceil}$$

and

$$P(s(X_{\lambda_n+1}) \leq z | X_0^{\lambda_n}) = P(Z_1^{(k,m)} \leq z | X_{-\infty}^{k}).$$

Thus

$$\sup_{-\infty<z<\infty} \left| \hat{F}_{\lambda_n}(z) - P(s(X_{\lambda_n+1}) \leq z | X_0^{\lambda_n}) \right| \leq \epsilon_{\lfloor f(\lambda_n) \rfloor}$$

eventually almost surely. The proof of Theorem 3.3 is complete.

Remark 3.4. Since $\lim_{m\to\infty} \epsilon_m = 0$, $\lim_{t\to\infty} f(t) = \infty$ and $\lim_{n\to\infty} \lambda_n = \infty$ almost surely thus

$$\lim_{n\to\infty} \epsilon_{\lfloor f(\lambda_n) \rfloor} = 0$$

almost surely.

Remark 3.5. Note that stopping time λ_n is the length of the data segment observed and used by the estimator $\hat{F}_{\lambda_n}$.

Remark 3.6. Note that our results in Theorem 3.3 are given for single orbits of the process and only on carefully selected time instances, (stopping times). These characteristics take our results out of the framework of the traditional statistical literature. For the necessity of intermittent estimation, that is, that one can not have these results for all time instances cf. Ryabko (1988), Algoet (1999), Bailey (1976) and Takahashi (2011).

Remark 3.7. Note that the limit

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = \frac{1}{P(\bigcup_{k=0}^{\infty} \{X_{-k+1}^{0} \in \mathcal{V}\})}$$

depends solely on the distribution of the process and the choice of the mnemonic set $\mathcal{V}$. It does not depend on other parameters as long as the other parameters satisfy the conditions.

Remark 3.8. Since we have no quantification on the "eventually almost sure" part of the statement, we cannot find a "best choice" for γ, ϵ_m and the function f. Despite this, if ϵ_m is monotone decreasing rapidly and the function f increases rapidly (but satisfying the conditions) the upper bound on the error may seem better but since our upper bound on the error is only eventually almost surely true, it may mean that the upper bound will become true slower. (If one had the other type of estimate on the error, one could perhaps make suitable choice under some criterion.)

Corollary 3.9. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. We assume that the process possesses at least one memory word and the mnemonic set $\mathcal{V}$ contains at least one memory word. Let $0 < \gamma < 1$ and $\delta \geq 0$ be arbitrary. Choose $\beta > \frac{2\delta}{(1-\gamma)}$. Let $f(t) = t^{1/\beta}$. Then for arbitrary $\epsilon > 0$*

$$\sup_{-\infty<z<\infty} \left| \hat{F}_{\lambda_n}(z) - P(s(X_{\lambda_n+1}) \leq z | X_0^{\lambda_n}) \right| \leq \frac{\epsilon}{\lfloor (\lambda_n)^{1/\beta} \rfloor^{\delta}}$$

eventually almost surely.

Remark 3.10. Note that neither in Theorem 3.3 nor in Corollary 3.9 there is any moment condition.

3.4. *Results for Estimating the Conditional Expectations.* Recall the notion of memory word from Section 2.1. Recall the mnemonic set $\mathcal{V}$, the stopping time λ_n and the estimator $h_n(X_0^{\lambda_n})$ from Section 3.1. Recall that function f and its inverse g were defined in Section 2.1.

Theorem 3.11. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. We assume that the process possesses at least one memory word and the mnemonic set $\mathcal{V}$ contains at least one memory word. Assume $E(|s(X_1)|^{\alpha}) < \infty$ for some $\alpha > 1$. Put*

$$\omega = \begin{cases} \alpha - 1 & \text{if } 1 < \alpha \leq 2 \\ 0.5\alpha & \text{if } 2 < \alpha. \end{cases}$$

Let $0 < \gamma < 1$, $\delta \geq 0$ be such that

$$\frac{m^{1+\delta\alpha}}{g(m)^{(1-\gamma)\omega}}$$

is summable. Then for the estimator $h_n(X_0^{\lambda_n})$ for arbitrary $\epsilon > 0$,

$$\left| h_n(X_0^{\lambda_n}) - E(s(X_{\lambda_n+1}) | X_0^{\lambda_n}) \right| \leq \epsilon \lfloor f(\lambda_n) \rfloor^{-\delta} \tag{3.11}$$

eventually almost surely. If in addition $E(|s(X_0|^2) < \infty)$ then

$$E\left(\left| s(X_{\lambda_n+1}) - h_n(X_0^{\lambda_n}) \right|^2 | X_0^{\lambda_n} \right) - E\left(\left| s(X_{\lambda_n+1}) - E(s(X_{\lambda_n+1}) | X_0^{\lambda_n}) \right|^2 | X_0^{\lambda_n} \right) \leq \epsilon^2 \lfloor f(\lambda_n) \rfloor^{-2\delta} \tag{3.12}$$

eventually almost surely.

Proof: For $i \geq 1$ define

$$Z_i^{(k,m)} = s(X_{j_i^{(k,m)}+1})$$

where $j_i^{(k,m)}$ is the same as in (3.10). Clearly $\{Z_i^{(k,m)}\}_{i=1}^{\infty}$ are conditionally independent and identically distributed given $X_{-\infty}^k$ with $\tau(X_{-\infty}^k) < \infty$. (Recall the definition of the look back time $\tau(X_{-\infty}^k)$ from Section 3.1.)

First apply Markov's inequality to get that on the set $\{\tau(X_{-\infty}^k) < \infty\}$

$$\begin{aligned} & P\left(\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil} Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right| > \epsilon m^{-\delta}|X_{k-\tau(X_{-\infty}^k)+1}^k\right) \\ = \; & P\left(\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil} Z_i^{(k,m)}}{\lceil g(m)^{(1-\gamma)}\rceil} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right|^\alpha > \frac{\epsilon^\alpha}{m^{\delta\alpha}}|X_{k-\tau(X_{-\infty}^k)+1}^k\right) \\ \le \; & \frac{E\left(\left|\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil}\left(Z_i^{(k,m)} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right)\right|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k\right)}{\epsilon^\alpha m^{-\delta\alpha}\lceil g(m)^{(1-\gamma)}\rceil^\alpha}. \end{aligned}$$

For $1 < \alpha \le 2$ Lemma 7.2 in the Appendix (Theorem 2 of von Bahr and Esseen (1965)) yields that on the set $\{\tau(X_{-\infty}^k) < \infty\}$

$$\begin{aligned} & P\left(\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil} Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right| > \epsilon m^{-\delta}|X_{k-\tau(X_{-\infty}^k)+1}^k\right) \\ \le \; & \frac{E\left(\left|\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil}\left(Z_i^{(k,m)} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right)\right|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k\right)}{\epsilon^\alpha m^{-\delta\alpha}\lceil g(m)^{(1-\gamma)}\rceil^\alpha} \\ \le \; & \frac{2\lceil g(m)^{(1-\gamma)}\rceil E\left(\left|Z_1^{(k,m)} - E(Z_1^{(k,m)}|X_{-\infty}^k)\right|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k\right)}{\epsilon^\alpha m^{-\delta\alpha}\lceil g(m)^{(1-\gamma)}\rceil^\alpha} \\ = \; & \frac{2E\left(\left|Z_1^{(k,m)} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k\right)}{\epsilon^\alpha m^{-\delta\alpha}\lceil g(m)^{(1-\gamma)}\rceil^{(\alpha-1)}}. \end{aligned}$$

Now Jensen's inequality gives that on the set $\{\tau(X_{-\infty}^k) < \infty\}$

$$\begin{aligned} & P\left(\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil} Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right| > \epsilon m^{-\delta}|X_{k-\tau(X_{-\infty}^k)+1}^k\right) \\ \le \; & \frac{2E\left(\left|Z_1^{(k,m)} - E(Z_i^{(k,m)}|X_{-\infty}^k)\right|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k\right)}{\epsilon^\alpha m^{-\delta\alpha}\lceil g(m)^{(1-\gamma)}\rceil^{(\alpha-1)}} \\ \le \; & \frac{2\left(E\left(\left|Z_1^{(k,m)}\right|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k\right) + \left|E(Z_1^{(k,m)}|X_{k-\tau(X_{-\infty}^k)+1}^k)\right|^\alpha\right)}{\epsilon^\alpha m^{-\delta\alpha} g(m)^{(1-\gamma)(\alpha-1)}} \\ \le \; & \frac{2\left(E\left(\left|Z_1^{(k,m)}\right|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k\right) + \left|E(Z_i^{(k,m)}|X_{k-\tau(X_{-\infty}^k)+1}^k)\right|^\alpha\right)}{\epsilon^\alpha m^{-\delta\alpha} g(m)^{(1-\gamma)(\alpha-1)}} \\ \le \; & \frac{4}{\epsilon^\alpha m^{-\delta\alpha} g(m)^{(1-\gamma)(\alpha-1)}} E(|Z_1^{(k,m)}|^\alpha |X_{k-\tau(X_{-\infty}^k)+1}^k). \end{aligned}$$

Integrating both sides on the set $\{\tau(X_{-\infty}^k) < \infty\}$ we get that

$$\begin{aligned} & P\left(\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil} Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil} - E((Z_1^{(k,m)})^\alpha |X_{k-\tau(X_{-\infty}^k)}^k)\right| > \epsilon m^{-\delta}, \tau(X_{-\infty}^k) < \infty\right) \\ \le \; & \frac{4}{\epsilon^\alpha m^{-\delta\alpha} g(m)^{(1-\gamma)(\alpha-1)}} E(|Z_1^{(k,m)}|^\alpha I_{\{\tau(X_{-\infty}^k)<\infty\}}) \end{aligned}$$

and the union bound will give

$$P\left(\exists 0\le k<m:\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil}Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil}-E((Z_1^{(k,m)})^\alpha|X_{k-\tau(X_{-\infty}^k)}^k)\right|>\epsilon m^{-\delta},\tau(X_{-\infty}^k)<\infty\right)$$
$$\le 4\epsilon^{-\alpha}\frac{m^{1+\delta\alpha}}{g(m)^{(1-\gamma)(\alpha-1)}}E(|s(X_1)|^\alpha I_{\{\tau(X_{-\infty}^k)<\infty\}})$$
$$\le 4\epsilon^{-\alpha}\frac{m^{1+\delta\alpha}}{g(m)^{(1-\gamma)(\alpha-1)}}E(|s(X_1)|^\alpha)$$

which is summable.

For $\alpha>2$ apply Markov's inequality, as before and then Lemma 7.3 in the Appendix (Theorem 2.10 of Petrov (1995)) to get that on the set $\{\tau(X_{-\infty}^k)<\infty\}$

$$P\left(\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil}Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil}-E(Z_i^{(k,m)}|X_{-\infty}^k)\right|>\epsilon m^{-\delta}|X_{k-\tau(X_{-\infty}^k)+1}^k\right)$$
$$\le \frac{2C(\alpha)}{\epsilon^\alpha m^{-\delta\alpha}g(m)^{(1-\gamma)\alpha/2}}E(|Z_1^{(k,m)}|^\alpha|X_{k-\tau(X_{-\infty}^k)+1}^k)$$

where $C(\alpha)$ depends only on α. Integrating both sides, just as in the previous case above, we get

$$P\left(\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil}Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil}-E((Z_1^{(k,m)})^\alpha|X_{k-\tau(X_{-\infty}^k)+1}^k)\right|>\epsilon m^{-\delta},\tau(X_{-\infty}^k)<\infty\right)$$
$$\le \frac{2C(\alpha)}{\epsilon^\alpha m^{-\delta\alpha}g(m)^{(1-\gamma)\alpha/2}}E(|Z_1^{(k,m)}|^\alpha I_{\{\tau(X_{-\infty}^k)<\infty\}})$$

and in turn

$$P\left(\exists 0\le k\le m-1:\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil}Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil}-E((Z_1^{(k,m)})^\alpha|X_{k-\tau(X_{-\infty}^k)}^k)\right|>\epsilon m^{-\delta},\tau(X_{-\infty}^k)<\infty\right)$$
$$\le \frac{2C(\alpha)\epsilon^{-\alpha}m^{(1+\delta\alpha)}}{g(m)^{(1-\gamma)0.5\alpha}}E(|s(X_1)|^\alpha I_{\{\tau(X_{-\infty}^k)<\infty\}})$$
$$\le \frac{2C(\alpha)\epsilon^{-\alpha}m^{(1+\delta\alpha)}}{g(m)^{(1-\gamma)0.5\alpha}}E(|s(X_1)|^\alpha)$$

which is summable. Applying the Borel-Cantelli lemma in both cases one gets that eventually almost surely for all $0\le k\le m-1$,

$$\tau(X_{-\infty}^k)<\infty$$

implies that

$$\left|\frac{\sum_{i=1}^{\lceil g(m)^{1-\gamma}\rceil}Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil}-E(Z_1^{(k,m)}|X_{k-\tau(X_{-\infty}^k)}^k)\right|\le\frac{\epsilon}{m^\delta}.$$

Observe that for suitable $0\le k<\lfloor f(\lambda_n)\rfloor$ and $m=\lfloor f(\lambda_n)\rfloor$:

$$\tau(X_{-\infty}^k)=\tau(X_0^k)<\infty,$$
$$h_n(X_0^{\lambda_n})=\frac{\sum_{i=1}^{\lceil g(m)^{(1-\gamma)}\rceil}Z_i^{(k,m)}}{\lceil g(m)^{1-\gamma}\rceil}$$

and

$$E(s(X_{\lambda_n+1})|X_0^{\lambda_n})=E(Z_1^{(k,m)}|X_{k-\tau(X_{-\infty}^k)}^k)$$

and we get that

$$\left|h_n(X_0^{\lambda_n}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right| \le \epsilon \lfloor f(\lambda_n)\rfloor^{-\delta}$$

eventually almost surely which gives (3.11). The first part of Theorem 3.11 is complete. Now we prove (3.12). Since

$$\begin{aligned} & E\left(\left|s(X_{\lambda_n+1}) - h_n(X_0^{\lambda_n})\right|^2 |X_0^{\lambda_n}\right) - E\left(\left|s(X_{\lambda_n+1}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right|^2 |X_0^{\lambda_n}\right) \\ = \;& E\left(s(X_{\lambda_n+1})^2|X_0^{\lambda_n}\right) + (h_n(X_0^{\lambda_n}))^2 \\ - \;& 2h_n(X_0^{\lambda_n})E\left(s(X_{\lambda_n+1})|X_0^{\lambda_n}\right) - E\left(s(X_{\lambda_n+1})^2|X_0^{\lambda_n}\right) \\ - \;& E(s(X_{\lambda_n+1})|X_0^{\lambda_n})^2 + 2E(s(X_{\lambda_n+1})|X_0^{\lambda_n})E\left(s(X_{\lambda_n+1})|X_0^{\lambda_n}\right) \\ = \;& \left(h_n(X_0^{\lambda_n}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right)^2 \end{aligned}$$

apply (3.11). The proof of Theorem 3.11 is complete.

Remark 3.12. Note that in (3.12)

$$E\left(\left|s(X_{\lambda_n+1}) - h_n(X_0^{\lambda_n})\right|^2 |X_0^{\lambda_n}\right) - E\left(\left|s(X_{\lambda_n+1}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right|^2 |X_0^{\lambda_n}\right) \ge 0$$

almost surely since $E(s(X_{\lambda_n+1})|X_0^{\lambda_n})$ minimizes the conditional mean square error. Thus (3.12) says that the conditional mean square error for our estimate $h_n(X_0^{\lambda_n})$ is close to the optimum eventually almost surely.

Corollary 3.13. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. We assume that the process possesses at least one memory word and the mnemonic set $\mathcal{V}$ contains at least one memory word. Assume $E(s(X_1)^\alpha) < \infty$ for some $\alpha > 1$. Let $0 < \gamma < 1$ and $\delta \ge 0$ be arbitrary. Choose*

$$\beta > \max\{\frac{2+\delta\alpha}{(1-\gamma)0.5\alpha}, 1, \frac{2+\delta\alpha}{(1-\gamma)(\alpha-1)}\}.$$

Let $f(t) = t^{1/\beta}$. Then for any $\epsilon > 0$,

$$\left|h_n(X_0^{\lambda_n}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right| \le \epsilon \lfloor (\lambda_n)^{1/\beta}\rfloor^{-\delta}$$

eventually almost surely. If in addition $E(|s(X_0|^2) < \infty)$ then for any $\epsilon > 0$

$$E\left(\left|s(X_{\lambda_n+1}) - h_n(X_0^{\lambda_n})\right|^2 |X_0^{\lambda_n}\right) - E\left(\left|s(X_{\lambda_n+1}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n})\right|^2 |X_0^{\lambda_n}\right) \le \epsilon^2 \lfloor (\lambda_n)^{1/\beta}\rfloor^{-2\delta}$$

eventually almost surely.

Remark 3.14. Note that in Corollary 3.13 the scheme depends on α through the restriction on the choice of β.

Note that in Corollary 3.13 the scheme depends on α while in the next Corollary the scheme is independent of α.

Corollary 3.15. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. We assume that the process possesses at least one memory word and the mnemonic set $\mathcal{V}$ contains at least one memory word. Assume $E(s(X_1)^\alpha) < \infty$ for some $\alpha > 1$. Let $f(t) = \log(t)$. Let $0 < \gamma < 1$ and $\delta \geq 0$ be arbitrary. Then for any $\epsilon > 0$,*

$$\left| h_n(X_0^{\lambda_n}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n}) \right| \leq \epsilon \lfloor \log(\lambda_n) \rfloor^{-\delta}$$

eventually almost surely. If in addition $E(|s(X_0|^2) < \infty)$ then for any $\epsilon > 0$

$$E\left(\left| s(X_{\lambda_n+1}) - h_n(X_0^{\lambda_n}) \right|^2 |X_0^{\lambda_n} \right) - E\left(\left| s(X_{\lambda_n+1}) - E(s(X_{\lambda_n+1})|X_0^{\lambda_n}) \right|^2 |X_0^{\lambda_n} \right) \leq \epsilon^2 \lfloor \log(\lambda_n) \rfloor^{-2\delta}$$

eventually almost surely.

In the next corollary we apply the results to first order Markov chains. This special case we will need in Section 5.

Corollary 3.16. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic first order Markov chain taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. Let the mnemonic set $\mathcal{V}$ be $\mathcal{V} = \mathcal{X}$. Assume $E(s(X_1)^\alpha) < \infty$ for some $\alpha > 1$. Let $f(t) = \log(t)$. Let $0 < \gamma < 1$ and $\delta \geq 0$ be arbitrary. Then*

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

almost surely and for any $\epsilon > 0$,

$$\left| h_n(X_0^{\lambda_n}) - E(s(X_{\lambda_n+1})|X_{\lambda_n}) \right| \leq \epsilon \lfloor \log(\lambda_n) \rfloor^{-\delta}$$

eventually almost surely. If in addition $E(|s(X_0|^2) < \infty)$ then for any $\epsilon > 0$

$$E\left(\left| s(X_{\lambda_n+1}) - h_n(X_0^{\lambda_n}) \right|^2 |X_0^{\lambda_n} \right) - E\left(|s(X_{\lambda_n+1}) - E(s(X_{\lambda_n+1})|X_{\lambda_n})|^2 |X_0^{\lambda_n} \right) \leq \epsilon^2 \lfloor \log(\lambda_n) \rfloor^{-2\delta}$$

eventually almost surely.

Remark 3.17. Note that even though in Corollary 3.15 and Corollary 3.16 the schemes themselves do not depend on α, we have to assume that $E(s(X_1)^\alpha) < \infty$ for some $\alpha > 1$. Cf. Section 5.

The reason for our use of the stopping times λ_n is that they enable us to guarantee that eventually we are doing "almost" as well as the best predictor with an explicit bound on this "almost".

Theorem 2 in Györfi et al. (1998) says that one can not estimate the conditional expectation of the next outcome for all n (with $\lambda_n = n$) for all first order Markov chains taking values in a countable subset of the $[0, 2]$ interval (cf. also Ryabko (1988)). For a restricted class of finite alphabet stationary and ergodic processes when one can estimate for all n see Morvai and Weiss (2005a).

4. Results for Intermittent Estimators not Using Prior Knowledge of a Mnemonic Set $\mathcal{V}$

4.1. *Intermittent Estimators Dependent on a Random Sequence $\mathcal{V}_n$ but not on $\mathcal{V}$.* We can weaken somewhat the condition on the knowing a mnemonic set $\mathcal{V}$ in advance. Assume that we have a random sequence of sets of words $\mathcal{V}_n(X_0^n)$ which will almost surely stabilize to be a fixed a priori unknown non random mnemonic set $\mathcal{V}$. (Our main example will be the class of Markov chains of unknown order with point wise consistent estimator $ORDEST(X_0^n)$ for the order of the chain. Then $\mathcal{V}_n(X_0^n)$ will be $\mathcal{X}^{ORDEST(X_0^n)}$. Since $ORDEST(X_0^n)$ converges to the order of the chain N, $\mathcal{V}_n(X_0^n)$ is eventually a constant equal to $\mathcal{X}^N$ which is a proper mnemonic set.) For notational simplicity we will often write $\mathcal{V}_n$ instead of $\mathcal{V}_n(X_0^n)$ with the understanding that really $\mathcal{V}_n = \mathcal{V}_n(X_0^n)$.

Now we will define $\tau^*_{k,n}$, λ^*_n, κ^*_n, $\hat{F}^*_{\lambda^*_n}(z)$ similarly to τ, λ_n, κ_n, $\hat{F}_{\lambda_n}(z)$ in Section 3.1.

For $k \le n$ we introduce the notation:

$$\tau^*_{k,n} = \inf\{t \ge 0 \ : \ X_{k-t+1}^k \in \mathcal{V}_n\}.$$

Note that this is well defined with probability one with the understanding that if there is no such t then the infimum is infinite. Note that $\tau^*_{k,n}$ depends mainly on X_0^k, its dependence on X_0^n is only through $\mathcal{V}_n$ and once this stabilizes this dependence disappears.

Recall that function f and its inverse g were defined in Section 2.1.

Let $0 < \gamma < 1$ be arbitrary. Define the stopping times λ^*_n as $\lambda^*_0 = 1$ and for $n \ge 1$,

$$\begin{aligned}\lambda^*_n = \min\{\ & t > \lambda^*_{n-1} \ : f(t) < t, \\ & \left(\exists 0 \le i < \lfloor f(t)\rfloor \ : \tau^*_{t,t} \le i,\ \tau^*_{i,t} = \tau^*_{t,t}, X^i_{i-\tau^*_{i,t}+1} = X^t_{t-\tau^*_{t,t}+1}\right) \text{ and} \\ & \left|\left\{\lfloor f(t)\rfloor < j < \lfloor g(\lfloor f(t)\rfloor)\rfloor \ : \ \tau^*_{j,t} = \tau^*_{t,t}, X^j_{j-\tau^*_{j,t}+1} = X^t_{t-\tau^*_{t,t}+1}\right\}\right| \\ & \ge \lceil g(\lfloor f(t)\rfloor)^{(1-\gamma)}\rceil\}.\end{aligned}$$

Note that the event$\{\lambda^*_n = t\}$ is measurable with respect to X_0^t.

Put

$$\begin{aligned}\kappa^*_n = \min\{t : \left|\left\{\lfloor f(\lambda^*_n)\rfloor < j \le t : \tau^*_{\lambda^*_n}(X_0^j) = \tau^*_{\lambda^*_n}(X_0^{\lambda^*_n}), X^j_{j-\tau^*_{j,\lambda^*_n}+1} = X^{\lambda^*_n}_{\lambda^*_n-\tau^*_{\lambda^*_n,\lambda^*_n}+1}\right\}\right| \\ = \lceil g(\lfloor f(\lambda^*_n)\rfloor)^{(1-\gamma)}\rceil\}.\end{aligned}$$

Define the empirical distribution function $\hat{F}^*_{\lambda^*_n}(z)$ as

$$\hat{F}^*_{\lambda^*_n}(z) = \frac{\sum_{i=\lfloor f(\lambda^*_n)\rfloor+1}^{\kappa^*_n} I_{\left\{\tau^*_{i,\lambda^*_n} = \tau^*_{\lambda^*_n,\lambda^*_n}, X^i_{i-\tau^*_{i,\lambda^*_n}+1} = X^{\lambda^*_n}_{\lambda^*_n-\tau^*_{\lambda^*_n,\lambda^*_n}+1}, s(X_{i+1}) \le z\right\}}}{\lceil g(\lfloor f(\lambda^*_n)\rfloor)^{(1-\gamma)}\rceil}.$$

Theorem 4.1. *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. We assume that the process possesses at least one memory word and the (a priori unknown) mnemonic set $\mathcal{V}$ contains at least one memory word. Let $0 < \gamma < 1$. Assume a sequence of positive real numbers ϵ_m such that*

$$\lim_{m\to\infty} \epsilon_m = 0$$

and

$$me^{-2g(m)^{(1-\gamma)}\epsilon_m^2}$$

is summable. Furthermore assume that

$$\mathcal{V}_n = \mathcal{V} \tag{4.1}$$

eventually almost surely. Then for the stopping times λ_n^* *,*

$$\lim_{n\to\infty} \frac{\lambda_n^*}{n} = \frac{1}{P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\})} \tag{4.2}$$

almost surely and for the estimator $\hat{F}^*_{\lambda_n^*}(z)$

$$\sup_{-\infty<z<\infty} \left|\hat{F}^*_{\lambda_n^*}(z) - P(s(X_{\lambda_n^*+1}) \le z | X_0^{\lambda_n^*})\right| \le \epsilon_{\lfloor f(\lambda_n^*)\rfloor} \tag{4.3}$$

eventually almost surely.

Proof: Since $\mathcal{V}_n = \mathcal{V}$ eventually almost surely, for almost all sample paths we can define Ψ as $\Psi = \min\{N > 1 : \mathcal{V}_n = \mathcal{V} \text{ for all } n \ge N\}$. Note that Ψ is a random variable. Now on a given sample path of the process, for any $t \ge \Psi$ either $t \in (\{\lambda_1^*, \lambda_2^*, \dots\} \bigcap \{\lambda_1, \lambda_2, \dots\})$ or $t \notin (\{\lambda_1^*, \lambda_2^*, \dots\} \bigcup \{\lambda_1, \lambda_2, \dots\})$. On a given sample path of the process, let $\phi^* = \max\{n \ge 0 : \lambda_n^* < \Psi\}$ and similarly, let $\phi = \max\{n \ge 0 : \lambda_n < \Psi\}$. Now on the given sample path for all $i \ge 1$, $\lambda^*_{\phi^*+i} = \lambda_{\phi+i}$. Define χ as

$$\chi = \phi - \phi^*. \tag{4.4}$$

Note that χ is a random variable. Now $\lambda_n^* = \lambda_{n+\chi}$ eventually almost surely and in turn $\hat{F}^*_{\lambda_n^*}(z) = \hat{F}_{\lambda_{n+\chi}}(z)$, Theorem 3.3 now completes the proof of Theorem 4.1.

Now we will define $h_n^*(X_0^{\lambda_n^*})$ similarly to $h_n(X_0^{\lambda_n})$ in Section 3.1.

For $n > 0$ define our estimator $h_n^*(X_0^{\lambda_n^*})$ for $E(s(X_{\lambda_n^*+1}|X_0^{\lambda_n^*})$ at time λ_n^* as

$$h_n^*(X_0^{\lambda_n^*}) = \frac{\sum_{i=\lfloor f(\lambda_n^*)\rfloor+1}^{\kappa_n^*} I_{\{\tau^*_{\lambda_n^*}(X_0^i)=\tau^*_{\lambda_n^*}(X_0^{\lambda_n^*}), X^i_{i-\tau^*_{\lambda_n^*}(X_0^i)+1} = X^{\lambda_n^*}_{\lambda_n^*-\tau^*_{\lambda_n^*}(X_0^{\lambda_n^*})+1}\}} s(X_{i+1})}{\lceil g(\lfloor f(\lambda_n^*)\rfloor)^{(1-\gamma)}\rceil}.$$

Theorem 4.2. *Let* $\{X_n\}_{n=-\infty}^{\infty}$ *be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet* $\mathcal{X}$*. We assume that the process possesses at least one memory word and the (a priori unknown) mnemonic set* $\mathcal{V}$ *contains at least one memory word. Assume* $E(|s(X_1)|^\alpha) < \infty$ *for some* $\alpha > 1$*. Put*

$$\omega = \begin{cases} \alpha - 1 & \text{if } 1 < \alpha \le 2 \\ 0.5\alpha & \text{if } 2 < \alpha. \end{cases}$$

Let $0 < \gamma < 1$*,* $\delta \ge 0$ *be such that*

$$\frac{m^{1+\delta\alpha}}{g(m)^{(1-\gamma)\omega}}$$

is summable. Furthermore assume that

$$\mathcal{V}_n = \mathcal{V} \tag{4.5}$$

eventually almost surely. Then for the stopping times λ_n^**,*

$$\lim_{n\to\infty} \frac{\lambda_n^*}{n} = \frac{1}{P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{V}\})} \tag{4.6}$$

almost surely and for the estimator $h_n^*(X_0^{\lambda_n})$ *for arbitrary* $\epsilon > 0$,

$$\left|h_n^*(X_0^{\lambda_n^*}) - E(s(X_{\lambda_n^*+1})|X_0^{\lambda_n^*})\right| \le \epsilon \lfloor f(\lambda_n^*)\rfloor^{-\delta} \tag{4.7}$$

eventually almost surely. If in addition $E(|s(X_0|^2) < \infty)$ *then*

$$E\left(\left|s(X_{\lambda_n^*+1}) - h_n^*(X_0^{\lambda_n^*})\right|^2 |X_0^{\lambda_n^*}\right) - E\left(\left|s(X_{\lambda_n^*+1}) - E(s(X_{\lambda_n^*+1})|X_0^{\lambda_n^*})\right|^2 |X_0^{\lambda_n^*}\right) \le \epsilon^2 \lfloor f(\lambda_n^*)\rfloor^{-2\delta} \tag{4.8}$$

eventually almost surely.

Proof: Similarly to the proof of Theorem 4.1, since $\mathcal{V}_n = \mathcal{V}$ eventually almost surely $\lambda_n^* = \lambda_{n+\chi}$ eventually almost surely and in turn $h_n^*(X_0^{\lambda_n^*}) = h_n(X_0^{\lambda_{n+\chi}})$, where χ is as in (4.4). Thus Theorem 3.11 completes the proof of Theorem 4.2.

Corollary 4.3. *In particular, in Theorem 4.2 let* $0 < \gamma < 1$ *and* $\delta \ge 0$ *be arbitrary and choose* $f(t) = t^{1/\beta}$, *where*

$$\beta > \max\{\frac{2+\delta\alpha}{(1-\gamma)0.5\alpha}, 1, \frac{2+\delta\alpha}{(1-\gamma)(\alpha-1)}.\}$$

Then the summability condition in Theorem 4.2 is satisfied.

Note that in Corollary 4.3 the scheme depends on α while in the next Corollary the scheme does not depend on α.

Corollary 4.4. *In particular, in Theorem 4.2 let* $0 < \gamma < 1$ *and* $\delta \ge 0$ *be arbitrary and choose* $f(t) = \log(t)$. *Then the summability condition in Theorem 4.2 is satisfied.*

For further reading cf. Morvai and Weiss (2007b), Takahashi (2011) and Morvai and Weiss (2021c).

4.2. *An Application: Intermittent Estimation with Universal Rate for Countable Alphabet Markov Chains of Unknown Order.* Finally we can obtain the results for the class of finite order stationary and ergodic Markov chains on a countable alphabet. The theorems follow immediately from the preceding ones.

Now assume that the process $\{X_n\}$ taking values from a countable set $\mathcal{X}$ is a Markov chain of unknown order N. Let $\mathcal{V} = \mathcal{X}^N$. Let $ORDEST(X_0^n)$ be an order estimation scheme such that $ORDEST(X_0^n) \to N$ almost surely (e.g. for the finite alphabet cf. Csiszár (2002) Csiszár and Shields (2000) for the countable alphabet cf. Morvai and Weiss (2005d) or Morvai and Weiss (2008))

Let $\mathcal{V}_n(X_0^n) = \mathcal{X}^{ORDEST(X_0^n)}$. Then obviously, $\mathcal{V}_n(X_0^n) = \mathcal{V}$ eventually almost surely. Recall that function f and its inverse g were defined in Section 2.1. Now, by Theorem 4.1, we have the following result immediately.

Theorem 4.5. *Let* $\{X_n\}_{n=-\infty}^{\infty}$ *be a stationary and ergodic discrete (finite or countably infinite) alphabet Markov process of unknown order. Let* $0 < \gamma < 1$. *Assume a sequence of positive real numbers* ϵ_m *such that* $\lim_{m\to\infty} \epsilon_m = 0$ *and* $me^{-2g(m)^{(1-\gamma)}\epsilon_m^2}$ *is summable. Then for the stopping times* λ_n^* ,

$$\lim_{n\to\infty} \frac{\lambda_n^*}{n} = 1 \tag{4.9}$$

almost surely and for the estimator $\hat{F}^*_{\lambda^*_n}(z)$

$$\sup_{-\infty<z<\infty}\left|\hat{F}^*_{\lambda^*_n}(z)-P(s(X_{\lambda^*_n+1})\le z|X_0^{\lambda^*_n})\right|\le\epsilon_{\lfloor f(\lambda^*_n)\rfloor} \tag{4.10}$$

eventually almost surely.

By Theorem 4.2, we have the following result.

Theorem 4.6. *Let* $\{X_n\}_{n=-\infty}^{\infty}$ *be a stationary and ergodic discrete (finite or countably infinite) alphabet Markov process of unknown order. Assume* $E(|s(X_1)|^\alpha)<\infty$ *for some* $\alpha>1$. *Put*

$$\omega=\begin{cases}\alpha-1 & \text{if } 1<\alpha\le 2\\ 0.5\alpha & \text{if } 2<\alpha.\end{cases}$$

Let $0<\gamma<1$, $\delta\ge 0$ *be such that* $\frac{m^{1+\delta\alpha}}{g(m)^{(1-\gamma)\omega}}$ *is summable. Then for the stopping times* λ^*_n,

$$\lim_{n\to\infty}\frac{\lambda^*_n}{n}=1 \tag{4.11}$$

almost surely and for the estimator $h^*_n(X_0^{\lambda_n})$ *for arbitrary* $\epsilon>0$,

$$\left|h^*_n(X_0^{\lambda^*_n})-E(s(X_{\lambda^*_n+1})|X_0^{\lambda^*_n})\right|\le\epsilon\lfloor f(\lambda^*_n)\rfloor^{-\delta} \tag{4.12}$$

eventually almost surely. If in addition $E(|s(X_0|^2)<\infty)$ *then*

$$E\left(\left|s(X_{\lambda^*_n+1})-h_n(X_0^{\lambda^*_n})\right|^2|X_0^{\lambda^*_n}\right)-E\left(\left|s(X_{\lambda^*_n+1})-E(s(X_{\lambda^*_n+1})|X_0^{\lambda^*_n})\right|^2|X_0^{\lambda^*_n}\right)\le\epsilon^2\lfloor f(\lambda^*_n)\rfloor^{-2\delta} \tag{4.13}$$

eventually almost surely.

Corollary 4.7. *in particular, in Theorem 4.6 let* $0<\gamma<1$ *and* $\delta\ge 0$ *be arbitrary and choose* $f(t)=t^{1/\beta}$, *where*

$$\beta>\max\{\frac{2+\delta\alpha}{(1-\gamma)0.5\alpha},1,\frac{2+\delta\alpha}{(1-\gamma)(\alpha-1)}.\}$$

Then the summability condition in Theorem 4.6 is satisfied.

In Corollary 4.7 the scheme depends on α while in the next Corollary the scheme is independent from α.

Corollary 4.8. *In particular, in Theorem 4.6 let* $0<\gamma<1$ *and* $\delta\ge 0$ *be arbitrary and choose* $f(t)=\log(t)$. *Then the summability condition in Theorem 4.6 is satisfied.*

For further reading cf. Morvai and Weiss (2007b) and Morvai and Weiss (2021c).

5. Limitations on Intermittent Estimation of the Conditional Expectation for First Order Markov Chains

Now assume that the alphabet is the nonnegative integers, that is, $\mathcal{X} = \{0, 1, 2, \dots\}$. Let $s(x) = x$ for all $x \in \mathcal{X}$. We will deal with first order Markov chains.

Instead of formulating a purely negative result we suppose that we have a scheme which does work for finite state Markov chains and show how it fails for a countable state chain.

By Corollary 3.16 there exists h_n and λ_n such that for any stationary and ergodic first order Markov chain $\{M_n\}$ taking values from a finite subset of $\mathcal{X}$,

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

and

$$\lim_{n\to\infty} \left| h_n(M_0^{\lambda_n}) - E(M_{\lambda_n+1}|M_{\lambda_n}) \right| = 0$$

almost surely. (Note that since M_n concentrates on a finite subset of $\mathcal{X}$, all moments of M_n are finite and so the moment condition in Corollary 3.16 is satisfied.)

Theorem 5.1. *For any estimation scheme h_n and stopping times λ_n such that for all stationary and ergodic first order Markov chains $\{M_n\}$ taking values from a finite subset of the nonnegative integers, almost surely,*

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

and

$$\lim_{n\to\infty} \left| h_n(M_0^{\lambda_n}) - E(M_{\lambda_n+1}|M_{\lambda_n}) \right| = 0$$

there exists a stationary and ergodic first order Markov chain $\{M_n\}$ taking values from the nonnegative integers with $E(M_n) < \infty$, such that with positive probability, for infinitely many n,

$$\left| h_n(M_0^{\lambda_n}) - E(M_{\lambda_n+1}|M_{\lambda_n}) \right| > 0.25.$$

Proof: Let $0 = E_0 < E_1 < \dots$ be an increasing sequence of nonnegative integers which will be specified later. Put

$$S = \{E_i : \ 0 \le i < \infty\}$$

and

$$S_k = \{E_i : \ 0 \le i \le k\}.$$

Define the transition probabilities of the Markov chain $\{M_n^{(k)}\}$ taking values in S_k as

$$p^{(k)}_{E_i,E_j} = \begin{cases} 1-\epsilon_i & \text{if } j = 0 \text{ and } i < k \\ 1 & \text{if } j = 0 \text{ and } i = k \\ \epsilon_i & \text{if } j = i+1 \text{ and } i < k \\ 0 & \text{otherwise} \end{cases}$$

and the transition probability of the Markov chain $\{M_n\}$ taking values in S as

$$p_{E_i,E_j} = \begin{cases} 1-\epsilon_i & \text{if } j = 0 \\ \epsilon_i & \text{if } j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

The ϵ_l's will be specified later. Now we start with $S_0 = \{E_0\}$ where $E_0 = 0$. Notice that $M_n^{(0)} = 0$ for all n. Let N_0 be so large that

$$P(h_n(M_0^{(0)}, \dots, M_{\lambda_i}^{(0)}) < 0.5, M_{\lambda_i}^{(0)} = E_0, \text{for some } \lambda_i \le N_0 | M_0^{(0)} = 0) > 1 - \left(\frac{1}{2}\right)^2.$$

Such an N_0 exists since the process is finite valued Markov and $E(M_1^{(0)}|M_0^{(0)}=E_0)=0$. Now let $E_1>2$ be so large that for $\epsilon_0=\frac{1}{E_1}$,

$$\sum_{x_0=0,x_1^{N_0}\in S_0^{N_0}}\left|\prod_{i=0}^{N_0-1}p^{(0)}_{x_i,x_{i+1}}-\prod_{i=0}^{N_0-1}p^{(1)}_{x_i,x_{i+1}}\right|<\left(\frac{1}{2}\right)^2.$$

Now $E(M_1^{(1)}|M_0^{(1)}=E_0)=1$ and so the conditional probability that

$$\left|h_n(M_0^{(1)},\dots,M_{\lambda_i}^{(1)})-E(M_{\lambda_i+1}^{(1)}|M_{\lambda_i}^{(1)}=E_0)\right|\geq 0.25, M_{\lambda_i}^{(1)}=E_0, \text{for some } \lambda_i\leq N_0$$

given the condition that $\{M_0^{(1)}=0\}$ is greater than

$$1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2=1-\left(\frac{1}{2}\right).$$

Now we define $\{M_n^{(k+1)}\}$. Let N_k be so large that

$$P(h_n(M_0^{(k)},\dots,M_{\lambda_i}^{(k)})<0.5, M_{\lambda_i}^{(k)}=E_k, \text{for some } \lambda_i\leq N_k|M_0^{(k)}=0)>1-\left(\frac{1}{2}\right)^{k+2}.$$

Such an N_k exists since the process is finite valued Markov and $E(M_1^{(k)}|M_0^{(k)}=E_k)=0$. Now let $E_{k+1}>E_k$ be so large that for $\epsilon_k=\frac{1}{E_{k+1}}$,

$$\sum_{x_0=0,x_1^{N_k}\in S_k^{N_k}}\left|\prod_{i=0}^{N_k-1}p^{(k)}_{x_i,x_{i+1}}-\prod_{i=0}^{N_k-1}p^{(k+1)}_{x_i,x_{i+1}}\right|<\left(\frac{1}{2}\right)^{k+2}.$$

Now $E(M_1^{(k+1)}|M_0^{(k+1)}=E_k)=1$ and so the conditional probability that

$$\left|h_n(M_0^{(k+1)},\dots,M_{\lambda_i}^{(k+1)})-E(M_{\lambda_i+1}^{(k+1)}|M_{\lambda_i}^{(k+1)})\right|\geq 0.25, M_{\lambda_i}^{(k+1)}=E_k, \text{for some } \lambda_i\leq N_k$$

given the condition that $\{M_0^{(k+1)}=0\}$ is greater than

$$1-\left(\frac{1}{2}\right)^{k+2}-\left(\frac{1}{2}\right)^{k+2}=1-\left(\frac{1}{2}\right)^{k+1}.$$

Now we have defined all the E_i and all ϵ_i and so the Markov chain $\{M_n\}$. Note that for N_k, $1\leq i\leq N_k$, $x_i\in S_k$

$$p_{0,x_1}\prod_{i=1}^{N_k-1}p_{x_i,x_{i+1}}=p^{(k)}_{0,x_1}\prod_{i=0}^{N_k-1}p^{(k+1)}_{x_i,x_{i+1}}$$

and $E(M_1|M_0=E_k)=1$ and so the conditional probability that the event

$$C_k=\{|h_n(M_0,\dots,M_{\lambda_i})-E(M_{\lambda_i+1}|M_{\lambda_i})|\geq 0.25, M_{\lambda_i}=E_k, \text{for some } \lambda_i\leq N_k\}$$

occurs given the condition that $M_0=0$ is greater than

$$1-\left(\frac{1}{2}\right)^{k+1}.$$

The conditional probability of the complement of the event C_k sums so by the Borel-Cantelli Lemma, eventually C_k occurs given the condition $M_0 = 0$. By Example (7.f) Ch. XV.7 in Feller (1968), the stationary probabilities of the countable valued first order Markov chain $\{M_n\}$ are

$$P(M_n = E_0) = \frac{1}{1 + \sum_{i=0}^{\infty} \prod_{j=0}^{i} \epsilon_j}$$

and for $k > 0$,

$$P(M_n = E_k) = \frac{\prod_{l=0}^{k-1} \epsilon_l}{1 + \sum_{i=0}^{\infty} \prod_{j=0}^{i} \epsilon_j}.$$

Note that $1+\sum_{i=0}^{\infty} \prod_{j=0}^{i} \epsilon_j < 2$ since for all $i \geq 0$, $\epsilon_i < 0.5$. This in turn implies that with probability at least 0.5, C_k occurs eventually for all k. What remains to proove is that $E(M_n) < \infty$. Observe that

$$E(M_n) = \sum_{k=0}^{\infty} kP(M_n = E_k) \leq \frac{\sum_{k=0}^{\infty} k(0.5)^k}{1 + \sum_{i=0}^{\infty} \prod_{j=0}^{i} \epsilon_j} < \infty.$$

The proof of Theorem 5.1 is complete.

Remark 5.2. Note that even though $E(M_n) < \infty$ for the counterexample process in the previous theorem, we can not guarantee the existence of higher moments. Cf. Corollary 3.16.

For further reading on the topics cf. Algoet (1999), Bailey (1976), Takahashi (2011), Morvai and Weiss (2007b) and Morvai and Weiss (2021c).

6. Conclusion

In this paper we considered stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet. We allowed countably infinite alphabet in contrast to e.g. our earlier paper Morvai and Weiss (2021b) where the alphabet was finite. The process was assumed to have at least one memory word which is weaker assumption on the process than in our earlier paper Morvai and Weiss (2007a). We estimated both the conditional distribution and the conditional expectation intermittently, that is, along stopping times λ_n. An important parameter of these intermittent schemes is the so called mnemonic set $\mathcal{V}$. In this paper the stopping times had a positive density, that is, $\lim_{n\to\infty} \frac{n}{\lambda_n} > 0$ (in contrast to our earlier result e.g. Morvai and Weiss (2021a) where the stopping times have zero density). The density is determined by the size of the mnemonic set $P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^{0} \in \mathcal{V}\})$. Though this quantity depends on the distribution of the process and the mnemonic set $\mathcal{V}$, it is not a random quantity, it does not depend on the sample path in contrast to e.g. our earlier paper Morvai and Weiss (2005e). We also considered the case where no prior knowledge on the mnemonic set $\mathcal{V}$ is assumed. The primary application to this case was the Markov chains of unknown order. Universal rates were given for the convergence of the error to zero for all our intermittent schemes. These results were given for single orbits of the process and only on carefully selected time instances, (stopping times). These characteristics take our results out of the framework of the traditional statistical literature. We also demonstrated the necessity of restricting the class of processes in some way.

For further reading cf. Morvai and Weiss (2007b) and Morvai and Weiss (2021c), Bühlmann and Wyner (1999), Takahashi (2011), Ryabko (1988), Algoet (1999), Bailey (1976), Ryabko (2019), Suzuki (2003), Ryabko (2010), Berti et al. (2021), Cover (1975), Kalociński and Steifer (2019), Ornstein (1978), Cesa-Bianchi and Lugosi (1999), Delecroix and Rosa (1995), Györfi et al. (1989), Müller et al. (2006), Masry (1989) Rissanen (2010, 2007).

7. Appendix

The next lemma, due to Kosorok (2008), is the generalization of the result of Dvoretzky, Kiefer and Wolfowitz in Dvoretzky et al. (1956), given in their Lemma 2 (refined by Massart (1990) in his Corollary 1) from the special case where the distribution function is continuous to the case which includes when the distribution function may be discontinuous.

Lemma 7.1. *(Theorem 11.6 in Kosorok (2008)) For any independent and identically distributed random varables $X_1, X_2, \ldots, X_n$,*

$$P\left(\sup_{-\infty<z<\infty}\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\le z\}} - P(X_1\le z)\right| > \epsilon\right) \le 2e^{-2\epsilon^2}$$

for all $\epsilon > 0$.

Lemma 7.2. *(Theorem 2 in von Bahr and Esseen (1965)) Let $X_1, X_2, \ldots, X_n$ be random varables satisfying*

$$E(X_{m+1}|\sum_{i=1}^{m} X_i) = 0 \quad \text{for all } 1 \le m \le n-1.$$

Let $1 \le r \le 2$. If

$$E(|X_k|^r < \infty \quad \text{for all } 1 \le k \le n$$

then

$$E(|\sum_{i=1}^{n} X_i|^r) \le 2\sum_{i=1}^{n} E(|X_i|^r).$$

The next lemma is Theorem 2.10 in Petrov Petrov (1995).

Lemma 7.3. *(Theorem 2.10 in Petrov (1995)) Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables with zero means and let $p \ge 2$. Then*

$$E|\sum_{i=1}^{n} Z_i|^p \le C(p)n^{p/2-1}\sum_{i=1}^{n} E|Z_i|^p.$$

where $C(p)$ is a positive constant depending only on p.

As Petrov (1995) pointed out, this is an immediate consequence of the Marcinkiewicz- Zygmund inequality, cf. Marcinkiewicz and Zygmund (1937) (p. 498 in Shiryaev (1996)).

References

Algoet, P. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Trans. Inform. Theory*, **40** (3), 609–633 (1994). DOI: 10.1109/18.335876.

Algoet, P. Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *IEEE Trans. Inform. Theory*, **45** (4), 1165–1185 (1999). MR1686250.

Bailey, D. H. *Sequential Schemes for Classifying and Predicting Ergodic Processes.* Ph.D. thesis, Stanford University (1976).

Berti, P., Pratelli, L., and Rigo, P. A Central Limit Theorem for Predictive Distributions. *Mathematics*, **9** (24) (2021). DOI: 10.3390/math9243211.

Bühlmann, P. and Wyner, A. J. Variable length Markov chains. *Ann. Statist.*, **27** (2), 480–513 (1999). MR1714720.

Bunea, F. and Nobel, A. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory*, **54** (4), 1725–1735 (2008). MR2450298.

Cesa-Bianchi, N. and Lugosi, G. On prediction of individual sequences. *Ann. Statist.*, **27** (6), 1865–1895 (1999). MR1765620.

Cover, T. Open problems in information theory. *1975 IEEE-USSR Joint Workshop on Information Theory*, pp. 35–36 (1975). https://isl.stanford.edu/∼cover/papers/paper37.pdf.

Csiszár, I. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, **48** (6), 1616–1628 (2002). MR1909476.

Csiszár, I. and Shields, P. C. The consistency of the BIC Markov order estimator. *Ann. Statist.*, **28** (6), 1601–1619 (2000). MR1835033.

Csiszár, I. and Talata, Z. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, **52** (3), 1007–1016 (2006). MR2238067.

Delecroix, M. and Rosa, A. Ergodic processes prediction via estimation of the conditional distribution function. *Annales de l'ISUP*, **39** (2), 35–56 (1995). HAL: 03660003.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, **27**, 642–669 (1956). MR83864.

Felber, T., Jones, D., Kohler, M., and Walk, H. Weakly universally consistent static forecasting of stationary and ergodic time series via local averaging and least squares estimates. *J. Statist. Plann. Inference*, **143** (10), 1689–1707 (2013). MR3082227.

Feller, W. *An introduction to probability theory and its applications Vol. I.*, volume 1 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., New York-London-Sydney, third edition (1968). ISBN 978-0-471-25708-0.

Finesso, L., Liu, C.-C., and Narayan, P. The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, **42** (5), 1488–1497 (1996). MR1426225.

Györfi, L., Härdle, W., Sarda, P., and Vieu, P. *Nonparametric curve estimation from time series*, volume 60 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin (1989). ISBN 3-540-97174-2. MR1027837.

Györfi, L. and Lugosi, G. Strategies for sequential prediction of stationary time series. In *Modeling uncertainty*, volume 46 of *Internat. Ser. Oper. Res. Management Sci.*, pp. 225–248. Kluwer Acad. Publ., Boston, MA (2002). MR1893282.

Györfi, L., Morvai, G., and Yakowitz, S. J. Limits to consistent on-line forecasting for ergodic time series. *IEEE Trans. Inform. Theory*, **44** (2), 886–892 (1998). MR1607704.

Györfi, L. and Ottucsák, G. Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory*, **53** (5), 1866–1872 (2007). MR2317147.

Györfi, L., Ottucsák, G., and Walk, H. *Machine Learning for Financial Engineering*. Imperial College Press, London (2012). DOI: 10.1142/p818.

Hanneke, S. Learning whenever learning is possible: Universal learning under general stochastic processes. In *2020 Information Theory and Applications Workshop (ITA)*, pp. 1–95 (2020). DOI: 10.1109/ITA50056.2020.

Hanneke, S. Learning whenever learning is possible: universal learning under general stochastic processes. *J. Mach. Learn. Res.*, **22**, Paper No. 130, 116 (2021). MR4318486.

Jones, D., Kohler, M., and Walk, H. Weakly universally consistent forecasting of stationary and ergodic time series. *IEEE Trans. Inform. Theory*, **58** (2), 1191–1202 (2012). MR2918019.

Kalociński, D. and Steifer, T. An Almost Perfectly Predictable Process with No Optimal Predictor. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2504–2508 (2019). ISBN 978-1-5386-9292-9. DOI: 10.1109/ISIT.2019.8849587.

Kosorok, M. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer-Verlag New York (2008). ISBN 978-0-387-74977-8. DOI: 10.1007/978-0-387-74978-5.

Marcinkiewicz, J. and Zygmund, A. Sur les fonctions indépendantes. *Fund. Math.*, **29** (1), 60–90 (1937). URL http://eudml.org/doc/212925. Available at http://eudml.org/doc/212925.

Masry, E. Nonparametric estimation of conditional probability densities and expectations of stationary processes: strong consistency and rates. *Stochastic Process. Appl.*, **32** (1), 109–127 (1989). MR1008911.

Massart, P. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, **18** (3), 1269–1283 (1990). MR1062069.

Merhav, N. and Feder, M. Universal prediction. *IEEE Trans. Inform. Theory*, **44** (6), 2124–2147 (1998). MR1658815.

Morvai, G. Guessing the Output of a Stationary Binary Time Series. In Haitovsky, Y., Ritov, Y., and Lerche, H. R., editors, *Foundations of Statistical Inference*, pp. 207–215. Physica-Verlag HD, Heidelberg (2003). DOI: 10.1007/978-3-642-57410-8_18.

Morvai, G. and Weiss, B. Forward estimation for ergodic time series. *Ann. Inst. H. Poincaré Probab. Statist.*, **41** (5), 859–870 (2005a). MR2165254.

Morvai, G. and Weiss, B. Limitations on intermittent forecasting. *Statist. Probab. Lett.*, **72** (4), 285–290 (2005b). MR2153125.

Morvai, G. and Weiss, B. On classifying processes. *Bernoulli*, **11** (3), 523–532 (2005c). MR2146893.

Morvai, G. and Weiss, B. Order estimation of Markov chains. *IEEE Trans. Inform. Theory*, **51** (4), 1496–1497 (2005d). MR2241507.

Morvai, G. and Weiss, B. Prediction for discrete time series. *Probab. Theory Related Fields*, **132** (1), 1–12 (2005e). MR2136864.

Morvai, G. and Weiss, B. On estimating the memory for finitarily Markovian processes. *Ann. Inst. H. Poincaré Probab. Statist.*, **43** (1), 15–30 (2007a). MR2288267.

Morvai, G. and Weiss, B. On sequential estimation and prediction for discrete time series. *Stoch. Dyn.*, **7** (4), 417–437 (2007b). MR2378577.

Morvai, G. and Weiss, B. Estimating the lengths of memory words. *IEEE Trans. Inform. Theory*, **54** (8), 3804–3807 (2008). MR2451043.

Morvai, G. and Weiss, B. Universal tests for memory words. *IEEE Trans. Inform. Theory*, **59** (10), 6873–6879 (2013). MR3106870.

Morvai, G. and Weiss, B. Consistency, integrability and asymptotic normality for some intermittent estimators. *ALEA Lat. Am. J. Probab. Math. Stat.*, **18** (2), 1643–1667 (2021a). MR4291036.

Morvai, G. and Weiss, B. Intermittent estimation for finite alphabet finitarily Markovian processes with exponential tails. *Kybernetika (Prague)*, **57** (4), 628–646 (2021b). MR4332885.

Morvai, G. and Weiss, B. On universal algorithms for classifying and predicting stationary processes. *Probab. Surv.*, **18**, 77–131 (2021c). MR4255241.

Müller, U. U., Schick, A., and Wefelmeyer, W. Efficient prediction for linear and nonlinear autoregressive models. *Ann. Statist.*, **34** (5), 2496–2533 (2006). MR2291508.

Nobel, A. B. On optimal sequential prediction for general processes. *IEEE Trans. Inform. Theory*, **49** (1), 83–98 (2003). MR1965889.

Ornstein, D. Guessing the next output of a stationary process. *Israel J. Math.*, **30**, 292–296 (1978). DOI: 10.1007/BF02761077.

Petrov, V. *Limit Theorems of Probability Theory. Sequences of Independent Random Variables*. Oxford Studies in Probability. Clarendon Press, Oxford (1995). ISBN 9780198534990.

Rissanen, J. *Information and Complexity in Statistical Modeling*. Information Science and Statistics. Springer New York (2007). ISBN 978-0-387-36610-4. DOI: 10.1007/978-0-387-68812-1.

Rissanen, J. Basics of estimation. *Front. Electr. Electron. Eng. China*, **5**, 274–280 (2010). DOI: 10.1007/s11460-010-0104-0.

Ryabko, B. Y. Prediction of random sequences and universal coding. *Problems of Inform. Trans.*, **24**, 87–96 (1988). Available at `https://boris.ryabko.net/prinftran88f.pdf`.

Ryabko, D. Discrimination between B-processes is impossible. *J. Theoret. Probab.*, **23** (2), 565–575 (2010). MR2644876.

Ryabko, D. *Asymptotic nonparametric statistical analysis of stationary time series*. SpringerBriefs in Computer Science. Springer Cham (2019). ISBN 978-3-030-12563-9. DOI: 10.1007/978-3-030-12564-6.

Shiryaev, A. N. *Probability*. Graduate Texts in Mathematics. Springer New York, second edition (1996). ISBN 978-1-4757-2539-1. DOI: 10.1007/978-1-4757-2539-1.

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

Takahashi, H. Computational limits to nonparametric estimation for ergodic processes. *IEEE Trans. Inform. Theory*, **57** (10), 6995–6999 (2011). MR2882275.

von Bahr, B. and Esseen, C.-G. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.*, **36**, 299–303 (1965). MR170407.