Universal Tests for Memory Words

Gusztáv Morvai, Benjamin Weiss

***Abstract*—The main result is a universal point wise test that when presented with a set of words S on a finite or countable alphabet $\mathcal{X}$ that purports to be a set of memory words for a stationary process will eventually almost surely return the value YES precisely when all positive probability words in S are memory words. For example, if S consists of all of the single letters in $\mathcal{X}$ then the test will eventually say yes if and only if the process is a Markov chain. Various further positive and negative results of this type are also given.**

***Index Terms*—Statistical learning, stationary processes, stochastic processes**

I. Introduction

It is a basic question to decide what one can learn about a stationary ergodic process from observing the first n outputs of the process. In various degrees of generality this question has been studied by many people. For particular classes of processes such as Markov chains or finitarily Markovian processes (for their definition see Definition 3) more detailed results have been obtained especially concerning the memory length and the memory words. Recall that for a process taking values in $\mathcal{X}$, which may be finite or infinite, a word w in this alphabet is called a memory word if having seen this word at time t the conditional probability distribution of the next output, X_{t+1}, is independent of what we observed before the occurrence of w (for the formal definition see Definition 1). Our main result is a universal point wise test that when presented with a set of words S that purports to be a set of memory words for a process will eventually almost surely return the value YES precisely when all positive probability words in S are memory words. For example, if S consists of all of the single letters in $\mathcal{X}$ then the test will eventually say yes if and only if the process is a Markov chain. This result has been proven in [10] and our present paper is a far reaching generalization of this. We should emphasize that the fact that we are allowing the alphabet to be infinite and even in the binary case we are allowing for the lengths of the memory words to be unbounded. This complicates matters since one loses any hope of universal estimates for the convergence of the usual empirical estimators. Another immediate consequence of our main result is a test, which when presented with a two valued process can decide whether or not the process is a renewal process.

For a stationary time series $\{X_n\}$, the memory length of the past, $K(\dots, X_{-1}, X_0)$, is the smallest $k \geq 0$ such that $(X_{-k+1}, \dots, X_0)$ is a memory word and infinity if there is no such k (for the formal definition see Definition 2). For d-step Markov chains this value is always less than or equal to d. A process is finitarily Markovian (FM) if the memory length is almost surely finite. Now it is known [9] that there is no test for deciding whether or not a process is FM.

For minimal memory words, which are those memory words which have no proper suffix which is a memory word, we have a negative result already in case the alphabet consists of two symbols, say $\{0, 1\}$ and S_0 is the set of all words of the form $\{0, 01, 011, 0111, \dots\}$. Namely there is no test designed to work only for binary renewal processes and which would return a positive answer only in case all of the words of S_0 are **minimal** memory words. This negative result can be used to deduce an earlier result of ours from [10] concerning the non existence of a sequence of good stopping times $\{\lambda_n\}$ together with a sequence of estimators $\{h_n(X_0, \dots, X_{\lambda_n})\}$, such that for all FM processes, almost surely,

$$\lim_{n\to\infty} |h_n(X_0, \dots, X_{\lambda_n}) - K(X_0, \dots, X_{\lambda_n})| = 0.$$

For more results of similar vein see [2], [3], [4], [6], [8], [11], [1].

II. Results

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 $X_m^n = (X_m, \dots, X_n)$, where $m \leq n$. Note that if $m > n$ then X_m^n is the empty string. Let

$$\mathcal{X}^* = \bigcup_{k=0}^{\infty} \mathcal{X}^k$$

where $\mathcal{X}^0$ is a set that contains exactly the empty string $\emptyset$.

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

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

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

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

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

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

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

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

Let

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

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

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

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

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

Let $S \subseteq \mathcal{X}^*$ and $S_k = \mathcal{X}^k \bigcap S$. We define some explicit statistics. The first is a measurement of the failure of S to be a set of memory words.

For a given set $S \subseteq \mathcal{X}^*$ define

$$U(S) = \{(k, w_{-k+1}^0, i, z_{-k-i+1}^{-k}, x) : 0 \leq k < \infty, w_{-k+1}^0 \in S_k, \\ 1 \leq i, z_{-k-i+1}^{-k} \in \mathcal{X}^i, x \in \mathcal{X}, p(z_{-k-i+1}^{-k}, w_{-k+1}^0, x) > 0\}$$

and

$$\Gamma(S) = \sup_{U(S)} \left|p(x|w_{-k+1}^0) - p(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right|.$$

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

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

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

where $\frac{0}{0} = 0$.

Note that we have the empirical conditional probabilities in this strange way because we want to ahcieve conditional independence so that the Hoeffding inequality would be applicable in the proof of Theorem 1.

These empirical distributions, as well as the sets we are about to introduce are functions of X_0^n, but we suppress the dependence to keep the notation manageable.

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

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

Finally, define the empirical version of Γ as follows. Let

$$\hat{U}_n(S) = \{(k, w_{-k+1}^0, i, z_{-k-i+1}^{-k}, x) : 0 \leq k < \infty, w_{-k+1}^0 \in S_k, \\ 1 \leq i \leq n, z_{-k-i+1}^{-k} \in \mathcal{X}^i, x \in \mathcal{X}, z_{-k-i+1}^{-k} w_{-k+1}^0 x \in \mathcal{L}_{k+i}^n\}$$

and

$$\hat{\Gamma}_n(S) = \max_{\hat{U}_n(S)} \left|\hat{p}_n(x|w_{-k+1}^0) - \hat{p}_n(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right|.$$

Let us agree by convention that if the set over which we are maximizing is empty then $\hat{\Gamma}_n = 0$. Observe, that by ergodicity, for any $w_{-k+1}^0 \in \mathcal{X}^k$,$z_{-k-i+1}^{-k} \in \mathcal{X}^i$,$x \in \mathcal{X}$ if $p(z_{-k-i+1}^{-k}, w_{-k+1}^0, x) > 0$ then

$$z_{-k-i+1}^{-k} w_{-k+1}^0 x \in \mathcal{L}_{k+i}^n$$

eventually almost surely and

$$\lim_{n\to\infty} \left|\hat{p}_n(x|w_{-k+1}^0) - \hat{p}_n(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right| = \left|p(x|w_{-k+1}^0) - p(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right|$$

almost surely which facts together in turn imply that

$$\liminf_{n\to\infty} \hat{\Gamma}_n(S) \geq \Gamma(S) \ \text{ almost surely.}$$

With this in hand we can give a scheme for testing if a given set S consists entirely of memory words or if there is a word in S which is not memory word.

Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. For example we can take $\beta = \gamma = 0.25$. Let

$$TEST_n(S) = \begin{cases} YES & \text{if } \hat{\Gamma}_n(S) \leq n^{-\beta} \\ NO & \text{otherwise.} \end{cases}$$

Note that $TEST_n$ depends on X_0^n.

Theorem 1: Let $\{X_n\}$ be an arbitrary stationary and ergodic process taking values from $\mathcal{X}$. Let S be an arbitrary collection of words from $\mathcal{X}^*$. If some word in S has positive probability and is not a memory word then eventually almost surely

$$TEST_n(S) = NO.$$

If all words in S with positive probability are memory words then, for $n > 2^{\frac{1}{(1-\gamma-2\beta)}}$,

$$P(TEST_n(S) = NO) \leq 12n^6 e^{\frac{-n^{1-\gamma-2\beta}}{2}}.$$

It then follows that in this case eventually almost surely

$$TEST_n(S) = YES.$$

Proof:

We only have to deal with words with positive probability. Indeed, if $p(w_{-k+1}^0)=0$ then

$$\max_{(z_{-k-i+1}^{-k},w_{-k+1}^0,x)\in\mathcal{L}_{k+i}^n}\left|\hat{p}_n(x|w_{-k+1}^0)-\hat{p}_n(x|z_{-k-i+1}^{-k},w_{-k+1}^0)\right|$$

is zero, since w_{-k+1}^0 never appears and so the maximum is taken over the empty set.

If there is a $w_{-k+1}^0\in S$ which is not a memory word, then there are z_{-k-i+1}^{-k} and x such that $p(x|w_{-k+1}^0)\neq p(x|z_{-k-i+1}^{-k}w_{-k+1}^0)$ and $p(z_{-k-i+1}^{-k}w_{-k+1}^0x)>0$. By ergodicity, $TEST_n(S)=NO$ eventually almost surely.

Now assume S is a collection of memory words. We will estimate the probability of the undesirable event as follows.

Set $\lambda_{\ell,k,0}^+=0$, $\lambda_{\ell,k,0}^-=0$ and for $i>0$ define

$$\lambda_{\ell,k,i}^+=\lambda_{\ell,k,i-1}^+ + \min\{t>0: X_{\ell+\lambda_{\ell,k,i-1}^+-k+1+t}^{\ell+\lambda_{\ell,k,i-1}^+}=X_{\ell+\lambda_{\ell,k,i-1}^+-k+1}^{\ell+\lambda_{\ell,k,i-1}^+}\}$$

and

$$\lambda_{\ell,k,i}^-=\lambda_{\ell,k,i-1}^- + \min\{t>0: X_{\ell-\lambda_{\ell,k,i-1}^--k+1-t}^{\ell-\lambda_{\ell,k,i-1}^-}=X_{\ell-\lambda_{\ell,k,i-1}^--k+1}^{\ell-\lambda_{\ell,k,i-1}^-}\}.$$

Informally, if $\lambda_{\ell,k,i}^+=s$ then the i-th occurrence (going forward in positive direction from position ℓ) of the block $X_{\ell-k+1}^{\ell}$ (with length k starting at position $\ell-k+1$ and ending at position ℓ) can be found at X_{s-k+1}^s. Similarly, $\lambda_{\ell,k,i}^-$ is the position of the i-th occurrence of the block $X_{\ell-k+1}^{\ell}$ going backward in negative direction from position ℓ.

For a given pair (k,ℓ) such that $0\le k<n$, $k-1\le\ell\le n-1$, assume that $X_{\ell-k+1}^{\ell+1}=w_{-k+1}^0x$ and w_{-k+1}^0 is a memory word. Since w_{-k+1}^0 is a memory word, by Lemma 1 in the Appendix for any $i,j\ge1$,

$$X_{\ell-\lambda_{\ell,k,i}^-+1},\dots,X_{\ell-\lambda_{\ell,k,1}^-+1},X_{\ell+\lambda_{\ell,k,1}^+-1},\dots,X_{\ell+\lambda_{\ell,k,j}^+-1}$$

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

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

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

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

Summing over all pairs (k,ℓ) such that $0\le k<n$ and all $k-1\le\ell\le n-1$ and over all pairs (i,j) such that $i\ge0$, $j\ge0$, $i+j\ge\lceil n^{1-\gamma}\rceil$ we get that

$$\begin{aligned}&P\Bigg(\text{ For some } 0\le k<n, k-1\le\ell\le n-1:\\&\quad X_{\ell-k+1}^{\ell+1}\in\mathcal{L}_{k+1}^n, K(X_{-\infty}^{\ell})\le k,\\&\quad\left|\hat{p}_n(X_{\ell+1}|X_{\ell-k+1}^{\ell})-p(X_{\ell+1}|X_{\ell-k+1}^{\ell})\right|>n^{-\beta}/2\Bigg)\\&\le\ n^2\sum_{h=\lceil n^{1-\gamma}\rceil}^{\infty}h2e^{-0.5n^{-2\beta}h}.\end{aligned}$$

Now

$$\begin{aligned}&P(TEST_n(S)=NO)=P(\hat{\Gamma}_n(S)>n^{-\beta})\le\\&P\Bigg(\max_{0\le k<\infty}\max_{w_{-k+1}^0\in\mathcal{W}_k}\max_{1\le i\le n}\max_{(z_{-k-i+1}^{-k},w_{-k+1}^0,x)\in\mathcal{L}_{k+i}^n}\\&\left|\hat{p}_n(x|w_{-k+1}^0)-\hat{p}_n(x|z_{-k-i+1}^{-k},w_{-k+1}^0)\right|>n^{-\beta}\Bigg)\le\\&\sum_{k=0}^{n-1}P\Bigg(\max_{w_{-k+1}^0\in\mathcal{W}_k}\max_{1\le i\le n}\max_{(z_{-k-i+1}^{-k},w_{-k+1}^0,x)\in\mathcal{L}_{k+i}^n}\\&\left|\hat{p}_n(x|w_{-k+1}^0)-\hat{p}_n(x|z_{-k-i+1}^{-k},w_{-k+1}^0)\right|>n^{-\beta}\Bigg)\\&\le\sum_{k=0}^{n-1}\sum_{i=1}^{n}P\Bigg(\max_{w_{-k+1}^0\in\mathcal{W}_k}\max_{(z_{-k-i+1}^{-k},w_{-k+1}^0,x)\in\mathcal{L}_{k+i}^n}\\&\left|\hat{p}_n(x|w_{-k+1}^0)-\hat{p}_n(x|z_{-k-i+1}^{-k},w_{-k+1}^0)\right|>n^{-\beta}\Bigg)\\&\le\sum_{k=0}^{n-1}\sum_{i=1}^{n}P\Bigg(\max_{w_{-k+1}^0\in\mathcal{W}_k}\max_{(z_{-k-i+1}^{-k},w_{-k+1}^0,x)\in\mathcal{L}_{k+i}^n}\\&\left|\hat{p}_n(x|w_{-k+1}^0)-p(x|w_{-k+1}^0)\right|>n^{-\beta}/2\Bigg)\\&+\sum_{k=0}^{n-1}\sum_{i=1}^{n}P\Bigg(\max_{w_{-k+1}^0\in\mathcal{W}_k}\max_{(z_{-k-i+1}^{-k},w_{-k+1}^0,x)\in\mathcal{L}_{k+i}^n}\\&\left|p(x|z_{-k-i+1}^{-k},w_{-k+1}^0)-\hat{p}_n(x|z_{-k-i+1}^{-k},w_{-k+1}^0)\right|\\&>n^{-\beta}/2\Bigg)\\&\le 2n^4\sum_{h=\lceil n^{1-\gamma}\rceil}^{\infty}h2e^{\frac{-n^{-2\beta}h}{2}}.\end{aligned}$$

By Lemma 3 in the Appendix, for $n>2^{\frac{1}{(1-\gamma-2\beta)}}$, the right

hand side is at most

$$2n^4 6n^2 e^{\frac{-n^{1-\gamma-2\beta}}{2}}$$

which is summable provided $2\beta+\gamma < 1$ and the Borel-Cantelli Lemma yields that

$$P(\hat{\Gamma}_n(S) \leq n^{-\beta} \text{ eventually}) = 1$$

and so $TEST_n(S) = YES$ eventually almost surely. The proof of Theorem 1 is complete.

Remark 2: Although if all words in S with positive probability are memory words then the probability of error

$$P(TEST_n(S) = NO) \leq 12n^6 e^{\frac{-n^{1-\gamma-2\beta}}{2}}$$

tends to zero faster for small γ and β, in this case we require much bigger difference $\hat{\Gamma}_n(S) > n^{-\beta}$ on much greater number of occurrences $n^{1-\gamma}+1$ in order to say NO.

Corollary 1: Our result yields an order estimation scheme for countable alphabet Markov chains. (Cf. Theorem 2.1 in [7].) Indeed, choose the smallest $0 \leq k_n \leq n$ such that $TEST_n(\mathcal{X}^{k_n}) = YES$ and let $k_n = n$ otherwise. Clearly, for a stationary and ergodic Markov chain of order K, $k_n = K$ eventually almost surely and $k_n \to \infty$ if the process is not Markov of any order.

Remark 3: Theorem 1 has been proved in [10] for singleton sets S.

Corollary 2: If $S = \{\emptyset\}$ that is S consists of the empty string then $TEST_n(S)$ yields a test for independence.

Theorem 2: One can eventually decide if a binary stationary and ergodic process is a classical renewal process or not.

Proof:

Indeed, apply Theorem 1 for

$$A = \{0, 01, 011, 0111, 01111, \ldots\}$$

and

$$B = \{1, 10, 100, 1000, 10000, \ldots\}.$$

If for some $0 \leq i \leq n$ $X_i = 0$ and $TEST_n(A) = YES$ eventually almost surely then the process is a classical renewal process with renewal state 0 . If for some $0 \leq i \leq n$ $X_i = 1$ and $TEST_n(B) = YES$ eventually almost surely then the process is a classical renewal process with renewal state 1 . (If both $TEST_n(A) = YES$ and $TEST_n(B) = YES$ eventually almost surely then the process is either a first order Markov chain or an independent identically distributed process.) If both $TEST_n(A) = NO$ and $TEST_n(B) = NO$ eventually almost surely then the process is not a classical renewal process. The proof of Theorem 2 is complete.

Definition 3: A stationary and ergodic time series $\{X_n\}$ is called finitarily Markovian if the set of memory words has probability one, that is,

$$P(\bigcup_{n=0}^{\infty} X_{-n+1}^0 \in \mathcal{W}_n) = 1.$$

Remark 4: The stationary time series $\{X_n\}$ is finitarily Markovian if and only if $K(X_{-\infty}^0)$ is finite (though not necessarily bounded) almost surely.

Remark 5: It has been proved in [9] that there is no test for the finitarily Markovian property.

Theorem 3: Let $S = \{0, 01, 011, 0111, \ldots\}$. There is no $testm_n(S)$ such that for all stationary and ergodic binary classical renewal processes with renewal state 0 and with the property that for all $k \geq 1$, $z_{-k+1}^0 \in \{0,1\}^k$: $p(z_{-k+1}^0) > 0$, eventually almost surely, $testm_n(S) = YES$ if and only if all words in S are minimal memory words.

Proof:

The proof will proceed by contradiction. Namely we will assume that we have been given some $testm_n(S)$ that satisfies the properties stated in the theorem and construct a process for which will it will fail. The construction itself is motivated by a construction in Ryabko [12]. At each stage k, we will define a binary process $\{X_n^{(k)}\}$ such that S will consists of memory words - but not all of them will be minimal memory words while for the limiting process they will be minimal. On the other hand the NO values that the finite stages will have will be preserved in the limit which will contradict the assumed property of the test. Let $f(0) = 0$ and for all positive integer j put $f(j) = 1$.

Let $k = 1$. Consider a Markov chain $\{M_n^{(1)}\}$ with state space on the nonnegative integers with transition probabilities

$$p_{i,i+1}^{(1)} = p_{i,0}^{(1)} = \frac{1}{2}$$

for all $i \geq 0$.

Define $X_n^{(1)} = f(M_n^{(1)})$. Note that $\{X_n^{(1)}\}$ is independent and identically distributed process. Since 1 is a memory word, the word 01 can not be minimal and therefore there exists an N_1 large enough such that

$$P\left(testm_{N_1}(S)(X_0^{(1)}, \ldots, X_{N_1}^{(1)}) = NO | X_0^{(1)} = 0\right) > 1 - \frac{1}{2}.$$

For the next stage, $k = 2$ we will define a first order countable alphabet Markov chain $\{M_n^{(2)}\}$ and the higher order binary Markov chain $\{X_n^{(2)}\}$ as follows.

Define the transition probabilities of $\{M_n^{(2)}\}$ as

$$p_{i,i+1}^{(2)} = p_{i,i+1}^{(1)}$$

and

$$p_{i,0}^{(2)} = p_{i,0}^{(1)}$$

for $i \leq N_1$ and let

$$p_{i,i+1}^{(2)} = \left(\frac{1}{2}\right)^2$$

and

$$p_{i,0}^{(2)} = 1 - \left(\frac{1}{2}\right)^2$$

for $i > N_1$. Notice that we haven't changed the transition probabilities in the range that gave us the NO value for the test .
Put $X_n^{(2)} = f(M_n^{(2)})$. Now $\{X_n^{(2)}\}$ is an $N_1 + 1$ order binary Markov chain. Observe that for all the conditional distribution of

$$(X_0^{(2)}, \ldots, X_{N_1}^{(2)})$$

given $\{X_0^{(2)} = 0\}$ is the same as the conditional distribution of

$$(X_0^{(1)}, \ldots, X_{N_1}^{(1)})$$

given $\{X_0^{(1)} = 0\}$, so that the NO value for this process will still occur with probability greater than $1/2$. Now the run of one's with length $N_1 + 1$ is a memory word of the process $\{X_n^{(2)}\}$ and so this word preceded by a zero at the left is also a memory word of the same process, but not a minimal one. Therefore, for a sufficiently large $N_2 > N_1$ we will have:

$$P\left(testm_{N_2}(S)(X_0^{(2)}, \ldots, X_{N_2}^{(2)}) = NO | X_0^{(2)} = 0\right) > 1 - \left(\frac{1}{2}\right)^2.$$

For the general step $k+1$ we proceed inductively, defining the first order countable alphabet Markov chain $\{M_n^{(k+1)}\}$ and the higher order binary Markov chain $\{X_n^{(k+1)}\}$ as follows.
Define the transition probabilities of $\{M_n^{(k+1)}\}$ as

$$p_{i,i+1}^{(k+1)} = p_{i,i+1}^{(k)}$$

and

$$p_{i,0}^{(k+1)} = p_{i,0}^{(k)}$$

for $i \leq N_k$ and let

$$p_{i,i+1}^{(k+1)} = \left(\frac{1}{2}\right)^{k+1}$$

and

$$p_{i,0}^{(k+1)} = 1 - \left(\frac{1}{2}\right)^{k+1}$$

for $i > N_k$.
Put $X_n^{(k+1)} = f(M_n^{(k+1)})$. Now $\{X_n^{(k+1)}\}$ is an $N_k + 1$ order binary Markov chain. Observe that for all $1 \leq i \leq k$, the conditional distribution of

$$(X_0^{(i)}, \ldots, X_{N_i}^{(i)})$$

given $\{X_0^{(i)} = 0\}$ is the same as the conditional distribution of

$$(X_0^{(k+1)}, \ldots, X_{N_i}^{(k+1)})$$

given $\{X_0^{(k+1)} = 0\}$. Now clearly the run of one's with length $N_k + 1$ is a memory word of the process $\{X_n^{(k+1)}\}$ and so this word preceded by a zero at the left is also a memory word of the same process, but not a minimal one. Therefore, there exists $N_{k+1} > N_k$ large enough such that

$$P\left(testm_{N_{k+1}}(S)(X_0^{(k+1)}, \ldots, X_{N_{k+1}}^{(k+1)}) = NO | X_0^{(k+1)} = 0\right) > 1 - \left(\frac{1}{2}\right)^{k+1}.$$

Now we define a stationary and ergodic binary classical renewal process $\{X_n\}$ which is not a Markov chain of any order. First we define the transition probabilities of the first order countable alphabet Markov chain $\{M_n\}$ as

$$p_{i,i+1} = p_{i,i+1}^{(k)}$$

and

$$p_{i,0} = p_{i,0}^{(k)}$$

for $i \leq N_k$. Now put

$$X_n = f(M_n).$$

Now observe that for all $1 \leq i \leq k$, the conditional distribution of

$$(X_0^{(i)}, \ldots, X_{N_i}^{(i)})$$

given $\{X_0^{(i)} = 0\}$ is the same as the conditional distribution of

$$(X_0, \ldots, X_{N_i})$$

given $\{X_0 = 0\}$. Thus

$$P\left(testm_{N_k}(S)(X_0, \ldots, X_{N_k}) = NO | X_0 = 0\right) > 1 - \left(\frac{1}{2}\right)^{k+1}.$$

Thus by the Borel-Cantelli Lemma,

$$testm_{N_k}(S)(X_0, \ldots, X_{N_k}) = NO$$

infinitely often with some positive probability. But S does consist of minimal memory words for the $\{X_n\}$ process. For this observe that

$$p_{i,i+1} > 0,$$

$$p_{i,i+1} \geq p_{i+1,i+2}$$

and

$$\lim_{i \to \infty} p_{i,i+1} = 0,$$

so that no matter how long a string of consecutive 1's one sees, one cannot tell what the conditional probability for the next symbol to be a one is. This means that the test should eventually be returning a YES value, whereas infinitely often it returns a NO. This contradiction concludes the proof of Theorem 3.

Remark 6: Of course, for some special sets one can make tests if the particular set contains only minimal memory words or not. For example, assume the alphabet size is finite. Then for all finite sets of words there is a test. Indeed, one must check that the set is a collection of memory words (cf. Theorem 1) and that each of those finitely many suffixes which has positive probabilities (this can be checked) is not a memory word (and

this can be also done by applying Theorem 1 for singleton sets).

Remark 7: Those sets which contain a word with positive probability and its suffix, are not collections of minimal memory words.

Theorem 4: Let $A = \{\emptyset, 1, 11, 111, \ldots\}$ where $\emptyset$ denotes the empty string. There is no $NTEST_n(A)$ such that for all stationary and ergodic binary classical renewal processes with renewal state 0 and with the property that for all $k \geq 1$, $z_{-k+1}^0 \in \{0,1\}^k$: $p(z_{-k+1}^0) > 0$, eventually almost surely, $NTEST_n(A) = YES$ if no word in A is a memory word while if even one word in A is a memory word then eventually almost surely $NTEST_n(A) = NO$.

Proof:

We argue by contradiction. If there were such a test $NTEST_n(A)$ then we could check if $S = \{0, 01, 011, 0111, \ldots\}$ is a collection of minimal memory words. Indeed, we know by assumption that all words in S are memory words. We also know by assumption that the suffixes can be checked to be non memory words by $NTEST_n(A)$. So S consists of minimal memory words if and only if $NTEST_n(A) = YES$ eventually almost surely. However this would contradict Theorem 3. The proof of Theorem 4 is complete.

Remark 8: In Theorem 7 in [10] it has been shown that for any strictly increasing sequence of stopping times $\{\lambda_n\}$ and sequence of estimators $\{h_n(X_0, \ldots, X_{\lambda_n})\}$, such that for all stationary and ergodic binary Markov chains with arbitrary finite order, almost surely,

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

and

$$\lim_{n\to\infty} |h_n(X_0, \ldots, X_{\lambda_n}) - K(X_0, \ldots, X_{\lambda_n})| = 0$$

there is a stationary, ergodic finitarily Markovian binary time series such that on a set of positive measure of process realizations

$$h_n(X_0, \ldots, X_{\lambda_n}) \neq K(X_{-\infty}^{\lambda_n})$$

infinitely often.

We can now strengthen this negative result by showing that there is no density one stopping time estimator even if we only demand that it succeed on binary renewal processes - rather than on all finitarily Markovian processes.

Corollary 3: There is no pair of strictly increasing sequence of stopping times $\{\lambda_n\}$ and sequence of estimators $\{h_n(X_0, \ldots, X_{\lambda_n})\}$ such that for all stationary and ergodic binary classical renewal processes with renewal state 0 and with the property that for all $k \geq 1$, $z_{-k+1}^0 \in \{0,1\}^k$: $p(z_{-k+1}^0) > 0$:

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

and

$$\lim_{n\to\infty} |h_n(X_0, \ldots, X_{\lambda_n}) - K(X_0, \ldots, X_{\lambda_n})| = 0$$

almost surely. The proof follows easily from Theorem 3, indeed let $S = \{0, 01, 011, 0111, \ldots\}$. If one could estimate the length of a minimal memory word on a stopping time sequence $\{\lambda_n\}$ with density one, then one could construct $testm_n(S)$ such that evetually almost surely $testm_n(S) = YES$ if and only if all words in S are minimal memory words. Indeed one could define $testm_n(S) = YES$ if for all i such that $0.5n < \lambda_i < n$, for all $0 \leq k \leq \lambda_i$,

$$X_{\lambda_i - k+1}^{\lambda_i} \in S \quad \text{implies} \quad h_n(X_0, \ldots, X_{\lambda_i}) = k.$$

Otherwise define $testm_n(S) = NO$. This test would be correct eventually. But this would contradict Theorem 3.

III. Appendix

Recall that $\lambda_{\ell,k,0}^+ = 0$, $\lambda_{\ell,k,0}^- = 0$ and for $i > 0$

$$\lambda_{\ell,k,i}^+ = \lambda_{\ell,k,i-1}^+ + \min\{t > 0 : X_{\ell+\lambda_{\ell,k,i-1}^+ - k+1+t}^{\ell+\lambda_{\ell,k,i-1}^+ + t} = X_{\ell+\lambda_{\ell,k,i-1}^+ - k+1}^{\ell+\lambda_{\ell,k,i-1}^+}\}$$

and

$$\lambda_{\ell,k,i}^- = \lambda_{\ell,k,i-1}^- + \min\{t > 0 : X_{\ell-\lambda_{\ell,k,i-1}^- - k+1-t}^{\ell-\lambda_{\ell,k,i-1}^- - t} = X_{\ell-\lambda_{\ell,k,i-1}^- - k+1}^{\ell-\lambda_{\ell,k,i-1}^-}\}.$$

The next result is Lemma 1 in Morvai and Weiss [10].

Lemma 1: Let $\{X_n\}$ be an arbitrary stationary and ergodic process taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. Assume $w_{-k+1}^0 \in \mathcal{X}^k$ is a memory word and $x \in \mathcal{X}$ is a letter. Then for any $i, j \geq 1$,

$$X_{\ell-\lambda_{\ell,k,i}^- + 1}, \ldots, X_{\ell-\lambda_{\ell,k,1}^- + 1}, X_{\ell+\lambda_{\ell,k,1}^+ + 1}, \ldots, X_{\ell+\lambda_{\ell,k,j}^+ + 1}$$

are conditionally independent and identically distributed random variables given $X_{\ell-k+1}^{\ell} = w_{-k+1}^0, X_{\ell+1} = x$, where the identical distribution is $p(\cdot|w_{-k+1}^0)$.

The next lemma is due to Hoeffding, cf. [5].

Lemma 2: (Hoeffding's inequality, Hoeffding 1963) Let $X_1, X_2, \ldots, X_n$ be independent real valued random variables, and $a_1, b_1, \ldots, a_n, b_n$ be real numbers such that $a_i \leq X_i \leq b_i$ with probability one for all $1 \leq i \leq n$. Then, for all $\epsilon > 0$,

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

Lemma 3: Assume $0 < \gamma < 1$ and $0 < \beta < \frac{1-\gamma}{2}$. Then

$$\sum_{h=\lceil n^{1-\gamma}\rceil}^{\infty} h e^{\frac{-n^{-2\beta}h}{2}} \leq 6n^2 e^{\frac{-n^{1-\gamma-2\beta}}{2}}$$

for $n > 2^{\frac{1}{(1-\gamma-2\beta)}}$.

Proof: Observe that $he^{\frac{-n^{-2\beta}h}{2}}$ is monotone decreasing in h as soon as the derivative $e^{\frac{-n^{-2\beta}h}{2}} - h0.5n^{-2\beta}e^{\frac{-n^{-2\beta}h}{2}}$ is

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

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

Integrating by parts we get that

$$\begin{aligned}
&\int_{n^{1-\gamma}}^{\infty} h e^{\frac{-n^{-2\beta}h}{2}} dh \\
&= \left[h \frac{-1}{\frac{n^{-2\beta}}{2}} e^{\frac{-n^{-2\beta}h}{2}} \right]_{n^{1-\gamma}}^{\infty} - \int_{n^{1-\gamma}}^{\infty} \frac{-1}{\frac{n^{-2\beta}}{2}} e^{\frac{-n^{-2\beta}h}{2}} dh \\
&= \frac{n^{1-\gamma}}{\frac{n^{-2\beta}}{2}} e^{\frac{-n^{-2\beta}n^{1-\gamma}}{2}} - \left[\frac{1}{\left(\frac{n^{-2\beta}}{2}\right)^2} e^{\frac{-n^{-2\beta}h}{2}} \right]_{n^{1-\gamma}}^{\infty} \\
&= \frac{n^{1-\gamma}}{\frac{n^{-2\beta}}{2}} e^{\frac{-n^{1-\gamma-2\beta}}{2}} + \frac{1}{\left(\frac{n^{-2\beta}}{2}\right)^2} e^{\frac{-n^{1-\gamma-2\beta}}{2}} \\
&= \left(2n^{1-\gamma+2\beta} + 4n^{4\beta}\right) e^{\frac{-n^{1-\gamma-2\beta}}{2}} \\
&\leq \left(2n^2 + 4n^2\right) e^{\frac{-n^{1-\gamma-2\beta}}{2}} \\
&\leq 6n^2 e^{\frac{-n^{1-\gamma-2\beta}}{2}}
\end{aligned}$$

since by assumption $0 < \gamma < 1$ and $0 < \beta < \frac{1-\gamma}{2}$. The proof of Lemma 3 is complete.

REFERENCES

[1] P. Bühlmann and A. J. Wyner, "Variable- length Markov chains," *Annals of Statistics* , vol. 27, pp. 480–513, 1999.

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

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

[4] I. Csiszár and Zs. Talata, "Context tree estimation for not necessarily finite memory processes via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, 2006.

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

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

[7] G. Morvai and B. Weiss, "Order estimation of Markov chains," *IEEE Transactions on Information Theory*, vol. 51, pp. 1496-1497, 2005.

[8] G. Morvai and B. Weiss, "Limitations on intermittent forecasting," *Statistics and Probability Letters*, vol. 72, pp. 285-290, 2005.

[9] G. Morvai and B. Weiss, "On classifying processes," *Bernoulli*, vol. 11, pp. 523-532, 2005.

[10] G. Morvai and B. Weiss, "On estimating the memory for finitarily Markovian processes," *Ann. I.H.Poincaré-PR*, vol. 43, pp. 15-30, 2007.

[11] G. Morvai and B. Weiss, "Estimating the Lengths of Memory Words". *IEEE Transactions on Information Theory*, vol.54 no. 8 pp. 3804-3807, 2008.

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