Bernoulli **11**(3), 2005, 523–532

On classifying processes

GUSZTÁV MORVAI[1] and BENJAMIN WEISS[2]

[1]*Research Group for Informatics and Electronics of the Hungarian Academy of Sciences, 1521 Goldmann György tér 3, Budapest, Hungary. E-mail: morvai@szit.bme.hu*
[2]*Hebrew University of Jerusalem, Jerusalem 91904, Israel. E-mail: weiss@math.huji.ac.il*

We prove several results concerning classifications, based on successive observations $(X_1, \ldots, X_n)$ of an unknown stationary and ergodic process, for membership of a given class of processes, such as the class of all finite-order Markov chains.

Keywords: nonparametric classification; stationary and ergodic processes

1. Introduction and statement of results

If $\mathcal{G}$ is a subclass of all stationary and ergodic binary processes then a sequence of functions $g_n : \{0, 1\}^n \to \{\text{YES, NO}\}$ is a classification for $\mathcal{G}$ in probability if

$$\lim_{n\to\infty} P(g_n(X_1, \ldots, X_n) = \text{YES}) = 1$$

for all processes in $\mathcal{G}$, and

$$\lim_{n\to\infty} P(g_n(X_1, \ldots, X_n) = \text{NO}) = 1$$

for all processes not in $\mathcal{G}$. Similarly, $g_n : \{0, 1\}^n \to \{\text{YES, NO}\}$ is a classification for $\mathcal{G}$ in a pointwise sense if

$$g_n(X_1, \ldots, X_n) = \text{YES eventually almost surely}$$

for all processes in $\mathcal{G}$, and

$$g_n(X_1, \ldots, X_n) = \text{NO eventually almost surely}$$

for all processes not in $\mathcal{G}$. Of course, if g_n is a classification in a pointwise sense then it is a classification in probability, but a classification in probability is not necessarily a classification in a pointwise sense.

For the class $\mathcal{M}_k$ of k-step mixing Markov chains of fixed order k, there is a pointwise classification of the type we have just described. (For mixing Markov chains, see Proposition I.2.10 in Shields 1996.) It was carried out in detail for independent processes by Bailey (1976). (Actually, he proved the result only for independent processes and indicated how to generalize his result for the class of $\mathcal{M}_k$.) For the class $\mathcal{M}_{\text{mix}} = \bigcup_{k=0}^{\infty} \mathcal{M}_{\text{k}}$ of mixing Markov chains of any order, Bailey showed that no such classification exists. See Ornstein and Weiss (1990) for further results on this kind of question. Our concern in this paper is with the class of finitarily Markovian processes which is defined as follows. Let $\{X_n\}_{n=1}^{\infty}$ be a stationary and ergodic binary time series. A one-sided stationary time series

$\{X_n\}_{n=1}^{\infty}$ can always be thought to be a two-sided time series $\{X_n\}_{n=-\infty}^{\infty}$. For $m \leqslant n$ let $X_m^n = (X_m, \ldots, X_n)$.

Definition. *A stationary and ergodic binary time series $\{X_n\}$ is said to be finitarily Markovian if for almost every $x_{-\infty}^{-1}$ there is a finite $K(x_{-\infty}^{-1})$ such that, for all $i > 0$ and y_{-i}^{-1}, if $P(X_0 = 1|X_{-K-i}^{-K-1} = y_{-i}^{-1}, X_{-K}^{-1} = x_{-K}^{-1}) > 0$ then*

$$P(X_0 = 1|X_{-K}^{-1} = x_{-K}^{-1}) = P(X_0 = 1|X_{-K-i}^{-K-1} = y_{-i}^{-1}, X_{-K}^{-1} = x_{-K}^{-1}).$$

This class includes all finite-order Markov chains (mixing or not) and many other processes such as the finitarily deterministic processes of Kalikow, et al. (1992).

Example 1. First we define a Markov process which serves as the technical tool for our construction. Let the state space S be the non-negative integers. The transition probabilities are as follows: with probability one move from 0 to 1 and from 1 to 2; for all $s \geqslant 2$ move with equal probability 0.5 to 0 and $s + 1$. This construction yields a stationary and ergodic Markov process $\{M_i\}$ with stationary distribution

$$P(M_i = 0) = P(M_i = 1) = \frac{1}{4}$$

and

$$P(M_i = j) = \frac{1}{2^j} \qquad \text{for } j \geqslant 2.$$

Now we define the binary hidden Markov chain $\{X_i\}$, which we denote as $X_i = f(M_i)$. Let $f(0) = 0$, $f(1) = 0$, and $f(s) = 1$ for all even states s. A feature of this definition of $f(\cdot)$ is that whenever $X_n = 0$, $X_{n+1} = 0$, $X_{n+2} = 1$, we know that $M_n = 0$ and vice versa. Consider the class of processes of the above form for all possible labellings of the rest of the states by zero and one. (It is easy to see that this class contains Markov chains of order up to $r + 1$, e.g. when, for all $s \geqslant r$, $f(s) = 1$, and processes which are not Markov of any order, e.g. when $f(2^i + 1) = 0$ for $i = 1, 2, \ldots$ and for the rest of the yet unlabelled odd states s, $f(s) = 1$.) This class is a subclass of all stationary and ergodic binary finitarily Markovian processes. (Clearly, the conditional probability $P(X_1 = 1|X_{-\infty}^0)$ does not depend on values beyond the first (going backward) occurrence of 001.) Györfi *et al.* (1998) proved that there is no estimator of the value $P(X_{n+1} = 1|X_1^n)$ from samples X_1^n such that the error tends to zero as n tends to infinity in the pointwise sense for this class of processes.

Example 2. Let $\{M_n\}$ be any stationary and ergodic first-order Markov chain with finite or countably infinite state space S. Let $s \in S$ be an arbitrary state with $P(M_1 = s) > 0$. Now let $X_n = I_{\{M_n = s\}}$. By Shields (1996, Section I.2.c.1), the binary time series $\{X_n\}$ is stationary and ergodic. It is also finitarily Markovian. (Indeed, the conditional probability $P(X_1 = 1|X_{-\infty}^0)$ does not depend on values beyond the first (going backwards) occurrence of one in $X_{-\infty}^0$ which identifies the first (going backwards) occurrence of state s in the Markov chain $\{M_n\}$.) The resulting time series $\{X_n\}$ is not a Markov chain of any order in general. (Indeed, consider the Markov chain $\{M_n\}$ with state space $S = \{0, 1, 2\}$ and

transition probabilities $P(X_2 = 1|X_1 = 0) = P(X_2 = 2|X_1 = 1) = 1$, $P(X_2 = 0|X_1 = 2) = P(X_2 = 1|X_1 = 2) = 0.5$. This yields a stationary and ergodic Markov chain $\{M_n\}$; cf. Example I.2.8 in Shields (1996). Clearly, the resulting time series $X_n = I_{\{M_n=0\}}$ will not be Markov of any order. The conditional probability $P(X_1 = 0|X_{-\infty}^0)$ depends on whether until the first (going backwards) occurrence of one there is an even or odd number of zeros.) These examples include all stationary and ergodic binary renewal processes with finite expected inter-arrival times, a basic class for many applications. (A stationary and ergodic binary renewal process is defined as a stationary and ergodic binary process such that the times between occurrences of ones are independent and identically distributed with finite expectation; cf. Shields 1996, Section I.2.c.1.)

Our main result is that there is no classification for membership in the class of finitarily Markovian processes. As a by-product we will also improve Bailey's result from mixing Markov chains to the class of Markov chains. Our results apply to both pointwise classifications and classifications in probability.

Theorem 1. *Given a sequence of functions* $g_n : \{0, 1\}^n \to \{\text{YES, NO}\}$ *such that*

- *for all stationary and ergodic binary Markov chains* $\{X_n\}$ *with arbitrary finite order*

$$\lim_{n\to\infty} P(g_n(X_1^n) = \text{YES}) = 1, \tag{1}$$

- *for all stationary and ergodic binary non-finitarily Markovian processes*

$$\lim_{n\to\infty} P(g_n(X_1^n) = \text{NO}) = 1, \tag{2}$$

we can construct a single stationary and ergodic binary process $\{X_n\}$ *such that*

$$\limsup_{n\to\infty} P(g_n(X_1^n) = \text{YES}) = 1 \quad \text{and} \quad \limsup_{n\to\infty} P(g_n(X_1^n) = \text{NO}) = 1.$$

Corollary 1. *There is no classification for the class of all stationary and ergodic binary Markov chains with arbitrary finite order, in a pointwise sense or in probability.*

Remark 1. For motivation, consider the universal intermittent estimation problem where the goal is to find stopping times τ_k such that one can estimate $P(X_{\tau_k+1} = 1|X_1^{\tau_k})$ from samples $X_1^{\tau_k}$ in the pointwise sense for all stationary and ergodic binary time series. Such a universal scheme was proposed in Morvai (2003). Unfortunately, the stopping times of Morvai (2003) grow very rapidly. Had one classified the Markov chains from non-Markov chains then one could have improved the scheme of Morvai such that it would have remained universially pointwise consistent for all stationary and ergodic processes; and in particular, if the process turned out to be Markov, one could have estimated the conditional probability $P(X_{k+1} = 1|X_1^k)$ eventually for all k, that is, $\tau_{n+1} = \tau_n + 1$ eventually. Indeed, if $g_n(X_1^n)$ classified the process as Markov then one could simply use a Markov order estimator (e.g. that of Csiszár and Shields 2000) and count frequencies of blocks with length equal to the order, and this estimator is consistent in the pointwise sense for Markov chains. Otherwise one could use the universal estimator of Morvai (2003).

Corollary 2. *There is no classification for the class of all stationary and ergodic binary finitarily Markovian processes, in a pointwise sense or in probability.*

Remark 2. Concerning the above-mentioned intermittent estimation problem, one could have improved the universal estimator of Morvai (2003) for finitarily Markovian processes. Had $g_n(X_1^n)$ classified the process as a finitarily Markovian process one could, for example, have used the stopping times and estimator in Morvai and Weiss (2003); the latter estimator is not universal but works for all finitarily Markovian processes, and the growth of the stopping times is much more moderate than that of the stopping times associated with the universal estimator in Morvai (2003). For non-finitarily Markovian processes one could use the universal estimator of Morvai (2003).

2. Proofs

The following lemma is well known.

Lemma 1. *Let $\{X_n\}$ be a stationary and ergodic binary time series and N a positive integer. Then there is a stationary and ergodic binary Markov chain $\{Z_n\}$ of some finite order up to N such that the N-dimensional distributions of $\{X_n\}$ and $\{Z_n\}$ are identical.*

Proof. Put $P(Z_{N+1} = z|Z_1^N = x_1^N) = P(X_{N+1} = z|X_1^N = x_1^N)$. This yields a stationary and ergodic Markov chain $\{Z_n\}$ of some finite order up to N with the original marginal distribution $P(Z_1^N = x_1^N) = P(X_1^N = x_1^N)$, that is, for $n > N$, define

$$P(Z_1^n = x_1^n) = P(Z_1^N = x_1^N) \prod_{i=N+1}^{n} P(Z_i = z_i|Z_{i-N}^{i-1} = x_{i-N}^{i-1}).$$

Clearly $\{Z_n\}$ is a stationary Markov chain of some finite order up to N since $\{X_n\}$ was stationary. The chain $\{Z_n\}$ can be thought of as a one-step Markov chain by passing to N-tuples. The ergodicity of the $\{X_n\}$ process guarantees that this chain is irreducible when considered as a chain on those N-tuples which have positive measure under the distribution of X_1^N. The process $\{Z_n\}$ is also ergodic since stationary binary irreducible Markov chains of some finite order are ergodic by Proposition I.2.9 in Shields (1996). (See also Kemeny and Snell 1960.) The proof of Lemma 2 is complete. □

Definition. *The entropy rate H associated with a stationary binary time series $\{X_n\}$ is defined as*

$$H = -\mathrm{E}\big\{P(X_0 = 1|X_{-\infty}^{-1})\log_2 P(X_0 = 1|X_{-\infty}^{-1}) + P(X_0 = 0|X_{-\infty}^{-1})\log_2 P(X_0 = 0|X_{-\infty}^{-1})\big\}.$$

Lemma 2. *Given a stationary and ergodic binary process $\{X_n\}$, an integer $N > 0$ and a real number $0 < \delta < 1$, there exists a stationary and ergodic non-finitarily Markovian process $\{Y_n\}$ such that*

$$\sum_{y_1^N \in \{0,1\}^N} |P(X_1^N = y_1^N) - P(Y_1^N = y_1^N)| < \delta. \tag{3}$$

Proof. Let $\{Z_n\}$ be a stationary and ergodic binary time series with zero entropy rate such that all finite words have positive probability. It is well known that such processes exist. For the sake of completeness we supply a proof in Lemma 3 in the Appendix. This process is clearly not finitarily Markovian.

By ergodicity of the $\{X_n\}$ process, there exists an r and a word w_1^r such that the empirical counts of all N blocks from w_1^r are $\delta/2^{N+1}$ close to the probabilities corresponding to the $\{X_n\}$ process.

We would like to define a process in which we alternate between the fixed word w_1^r and the Z_n: $Z_1, w_1^r, Z_2, w_1^r, \ldots$. If we can do this and identify uniquely the position of the Z_n then this process will not be finitarily Markovian. In order to uniquely identify the positions of the Z_n we will add a synchronizing word u_1^m whose length is very small compared to the length of w_1^r and which appears only where we place it. The fact that its length is small means that the finite distributions will remain close to the finite distribution of the $\{X_n\}$ process. For u_1^m to sychronize we need to know that when looking across a string such as $Z_1, u_1^m, w_1^r, Z_2, u_1^m, w_1^r, Z_3$ the word appears only in the two locations where it is written.

Now choose some word u_1^m with length $m = \lceil 10 \log_2 r \rceil$ such that this word does not appear in the word w_1^r and it has no reasonable non-trivial self-overlap. More precisely, there is no non-trivial self-overlap greater than $2/5m$ and there is no overlap with w_1^r greater than $2/5m$. The number of words with length m which have greater self-overlap is at most $2m2^{3/5m}$. The number of words of length m which have overlap with w_1^r greater than $2/5m$ but not completely contained in w_1^r is at most $2m2^{3/5m}$. The number of words with length m completely contained in w_1^r is at most r. Summing the number of these possible bad words, we get

$$r + 4m2^{3/5m} < 2^m.$$

Thus there is at least one word u_1^m with the desired property. The word u_1^m will serve as a synchronizing word.

We will define the desired $\{Y_n\}$ process in two steps. First we will define a non-stationary process $\{W_n\}$ as follows. Consider $n - 1 = \eta(m + r + 1) + \theta$, where $0 \leqslant \theta \leqslant m + r$ and $\eta \geqslant 0$. The process $\{W_n\}$ will be obtained by inserting a fixed block u_1^m, w_1^r of length $m + r$ between successive symbols of the process $\{Z_n\}$. Define the process $\{W_n\}$ as follows:

$$W_n = \begin{cases} Z_{\eta+1} & \text{if } \theta = 0, \\ u_\theta & \text{if } 1 \leqslant \theta \leqslant m, \\ w_{\theta-m} & \text{if } m+1 \leqslant \theta \leqslant m+r. \end{cases}$$

Our assumptions on the synchronizing word imply that such a process will not be stationary, and to ensure stationarity we need to randomize over $m + r + 1$. Here is a formal description. Let ζ be uniformly distributed on $\{0, \ldots, m + r\}$. Let ζ be independent of $\{W_n\}$. Define

$\{Y_n\}$ as $Y_n = W_{n+\zeta}$. (That is, $\{Y_n\}$ is constructed from $\{W_n\}$ by averaging over the $m + r + 1$ shifts of the $\{W_n\}$ process.)

The fact that u_1^m was synchronizing means that ζ is a function of the $\{Y_n\}$ process. Thus from $\{Y_n\}$ one recovers the $\{Z_n\}$ process exactly. Now $\{Y_n\}$ is a stationary and ergodic binary non-finitarily Markovian time series since $\{Z_n\}$ was such. To see that (3) is satisfied one uses the property of w_1^r and takes r sufficiently large that the edge effects caused by u_1^m are negligible. The proof of Lemma 2 is complete. □

Proof of Theorem 1. To construct $\{X_n\}$ we will alternately use the two lemmas to construct a sequence of processes $\{Y_n^{(i)}\}$, which for odd i will be a Markov chain and for even i will not even be finitarily Markovian, but the entire sequence will converge to an ergodic process $\{X_n\}$ which will have the required properties. Here is how this is done. Let $0 < \epsilon_k < 1$ such that $\epsilon_k \to 0$ and $0 < \delta_k < 1$ such that $\sum_{k=1}^{\infty} \delta_k < 0.25$. We construct our process as follows. Let $\{Y_n^{(1)}\}$ be independent and identically distributed random variables assuming the values $\{0, 1\}$ with equal probability. Let $N_1 > 1$ be so large that

$$P(g_{N_1}(Y_1^{(1)}, \ldots, Y_{N_1}^{(1)}) = \text{YES}) \geqslant 1 - \epsilon_1$$

and there exists a set $\mathcal{U}_{N_1} \subseteq \{0, 1\}^{N_1}$ such that $P((Y_1^{(1)}, \ldots, Y_{N_1}^{(1)}) \in \mathcal{U}_{N_1}) > 1 - \epsilon_1$ and

$$\max_{u_1^{N_1}, v_1^{N_1} \in \mathcal{U}_{N_1}} \sum_{x \in \{0,1\}} \frac{1}{N_1} \left| \sum_{i=0}^{N_1 - 1} \left(I_{\{u_{i+1} = x\}} - I_{\{v_{i+1} = x\}} \right) \right| < \epsilon_1.$$

Assume, for $k = 2, \ldots, i-1$, that we have already defined a sequence of stationary and ergodic binary time series $\{Y_n^{(k)}\}$ and positive integers $N_k > k^2$ and sets $\mathcal{U}_{N_k} \subseteq \{0, 1\}^{N_k}$ such that $P((Y_1^{(k)}, \ldots, Y_{N_k}^{(k)}) \in \mathcal{U}_{N_k}) > 1 - \epsilon_k$,

$$\sum_{y_1^{N_{k-1}} \in \{0,1\}^{N_{k-1}}} |P(Y_1^{(k-1)} = y_1, \ldots, Y_{N_{k-1}}^{(k-1)} = y_{N_{k-1}}) - P(Y_1^{(k)} = y_1, \ldots, Y_{N_{k-1}}^{(k)} = y_{N_{k-1}})| < \delta_{k-1},$$

$$\max_{u_k^{N_k}, v_1^{N_k} \in \mathcal{U}_{N_k}} \sum_{x_1^k \in \{0,1\}^k} \frac{1}{N_k - k + 1} \left| \sum_{i=0}^{N_k - k} \left(I_{\{u_{i+1}^{i+k} = x_1^k\}} - I_{\{v_{i+1}^{i+k} = x_1^k\}} \right) \right| < \epsilon_k, \tag{4}$$

and

- if k is even then $\{Y_n^{(k)}\}$ is not finitarily Markovian and

$$P(g_{N_k}(Y_1^{(k)}, \ldots, Y_{N_k}^{(k)}) = \text{NO}) \geqslant 1 - \epsilon_k$$

- if k is odd then $\{Y_n^{(k)}\}$ is a Markov chain with some order and

$$P(g_{N_k}(Y_1^{(k)}, \ldots, Y_{N_k}^{(k)}) = \text{YES}) \geqslant 1 - \epsilon_k.$$

Now we define the same for i. If i is odd then apply Lemma 1 for $\{Y_n^{(i-1)}\}$ with N_{i-1}. Let $\{Y_n^{(i)}\}$ denote the resulting stationary and ergodic binary Markov chain. Now let $N_i > i^2$ be so large that

$$P(g_{N_i}(Y_1^{(i)}, \ldots, Y_{N_i}^{(i)}) = \text{YES}) \geqslant 1 - \epsilon_i$$

and there is a set $\mathcal{U}_{N_i} \subseteq \{0, 1\}^{N_i}$ such that $P((Y_1^{(i)}, \ldots, Y_{N_i}^{(i)}) \in \mathcal{U}_{N_i}) > 1 - \epsilon_i$ and

$$\max_{u_i^{N_i}, v_1^{N_i} \in \mathcal{U}_{N_i}} \sum_{x_1^i \in \{0,1\}^i} \frac{1}{N_i - i + 1} \left| \sum_{j=0}^{N_i - i} \left(I_{\{u_{j+1}^{j+i} = x_1^i\}} - I_{\{v_{j+1}^{j+i} = x_1^i\}} \right) \right| < \epsilon_i.$$

By assumption (1) and the ergodicity of $\{Y_n^{(i)}\}_{n=1}^{\infty}$ there exists such an N_i.

If i is even then apply Lemma 2 for $\{Y_n^{(i-1)}\}$ with N_{i-1} and δ_{i-1}. Let $\{Y_n^{(i)}\}$ denote the resulting non-finitarily Markovian process. Now let $N_i > i^2$ be so large that

$$P(g_{N_i}(Y_1^{(i)}, \ldots, Y_{N_i}^{(i)}) = \text{NO}) \geqslant 1 - \epsilon_i$$

and there is a set $\mathcal{U}_{N_i} \subseteq \{0, 1\}^{N_i}$ such that $P((Y_1^{(i)}, \ldots, Y_{N_i}^{(i)}) \in \mathcal{U}_{N_i}) > 1 - \epsilon_i$ and

$$\max_{u_i^{N_i}, v_1^{N_i} \in \mathcal{U}_{N_i}} \sum_{x_1^i \in \{0,1\}^i} \frac{1}{N_i - i + 1} \left| \sum_{j=0}^{N_i - i} \left(I_{\{u_{j+1}^{j+i} = x_1^i\}} - I_{\{v_{j+1}^{j+i} = x_1^i\}} \right) \right| < \epsilon_i.$$

By assumption (2) and the ergodicity of $\{Y_n^{(i)}\}_{n=1}^{\infty}$ there exists such an N_i.

Now it follows from the construction that for any $n \leqslant N_k$ and $k \leqslant K$,

$$|P(Y_1^{(k)} = y_1, \ldots, Y_n^{(k)} = y_n) - P(Y_1^{(K)} = y_1, \ldots, Y_n^{(K)} = y_n)| \leqslant \sum_{i=k}^{\infty} \delta_i$$

which tends to zero as $k \to \infty$. Now define $\{X_n\}$ in the following way. For each n, let

$$P(X_1^n = x_1^n) = \lim_{k \to \infty} P(Y_1^{(k)} = x_1, \ldots, Y_n^{(k)} = x_n).$$

Clearly $\{X_n\}$ is stationary since all $\{Y_n^{(k)}\}$ were stationary. Since $P((X_1, \ldots, X_{N_k}) \in \mathcal{U}_{N_k}) > 1 - \epsilon_k - \sum_{i=k}^{\infty} \delta_i$, $N_k > k^2$, by (4) and Lemma 4 in the Appendix, $\{X_n\}$ is also ergodic. Now it follows from the construction that

$$|P(X_1^n = x_1^n) - P(Y_1^{(k)} = x_1, \ldots, Y_n^{(k)} = x_n)| \leqslant \sum_{i=k}^{\infty} \delta_i.$$

Thus for k even,

$$P(g_{N_k}(X_1, \ldots, X_{N_k}) = \text{NO}) \geqslant 1 - \epsilon_k - \sum_{i=k}^{\infty} \delta_i$$

and the right-hand side tends to 1 as $k \to \infty$. Similarly, when k is odd,

$$P(g_{N_k}(X_1, \ldots, X_{N_k}) = \text{YES}) \geqslant 1 - \epsilon_k - \sum_{i=k}^{\infty} \delta_i$$

and the right-hand side tends to 1 as $k \to \infty$. The proof of Theorem 1 is complete. □

Appendix

We present now the proofs of two fairly standard lemmas that we used before.

Lemma 3. *There exists a stationary and ergodic time series* $\{Z_n\}$ *with zero entropy rate such that all finite words have positive probability.*

Proof. Let $T : [0, 1] \to [0, 1]$ denote the mapping $x \to x + \alpha \bmod 1$, where α is a fixed irrational. Denote the Lebesgue measure on [0, 1] by μ. For a measurable subset A of [0, 1], let $\tau_A(x) = \min\{n \geqslant 1 : T^n x \in A\}$ denote the first return time to A. Partition A into $A_k = \{x \in A : \tau_A(x) = k\}$. Note that $T^i A_k : 0 \leqslant i < k\}$ are disjoint sets. We will define a particular set A with the property that, for all k, the sets A_k will have positive measure. Indeed, one can choose inductively points $\{x_n\}$ and $\delta_n > 0$, $\sum_{m=n+1}^{\infty} m\delta_m < 0.1\delta_n$ sufficiently small that if $I_n = [x_n - \delta_n, x_n + \delta_n]$ the A defined as follows will have the required property:

$$A = \bigcup_{n=1}^{\infty} \left[\left(I_n \bigcup T^n I_n \right) - \left[\bigcup_{m=n+1}^{\infty} \bigcup_{m-1}^{i=1} T^i I_m \right] \right].$$

It is easy to see that, for all k, $\mu(A_k) > 0$. In this case we can list all binary words with finite length, $\{0, 1, 00, 01, \ldots\} = \{w_1, w_2, \ldots\}$, and denote by $|w_k|$ the length of w_k. Define a partition of [0, 1] into two sets $\{P_0, P_1\}$ by taking the kth word w_k in the list and assigning the first $|w_k|$ sets of $(T^0 A_k)$, $(T^1 A_k)$, $\ldots$, $(T^{k-1} A_k)$ to P_0 or P_1 according to the symbols in w_k and then assign to P_0 all remaining points in [0, 1]. Finally, define a stationary and ergodic binary process as follows. Choose x uniformly on [0, 1] and set

$$Z_n(x) = \begin{cases} 1, & \text{if } T^n x \in P_1 \\ 0, & \text{if } T^n x \in P_0. \end{cases}$$

It is clear that all finite words have positive probability. Furthermore, it is well known that any process defined by an irrational rotation as above is stationary and ergodic and has zero entropy; see Cornfeld *et al.* (1982). The proof of Lemma 3 is complete. □

Lemma 4. *A binary stationary time series* $\{X_n\}$ *is ergodic if there is a sequence of positive integers* $N_k > k^2$ *tending to* ∞, $\epsilon_k > 0$ *tending to zero and a sequence of sets* $\mathcal{U}_{N_k} \subseteq \{0, 1\}^{N_k}$ *with probability greater than* $1 - \epsilon_k$ *such that for all* $u_1^{N_k}, v_1^{N_k} \in \mathcal{U}_{N_k}$,

$$\sum_{x_1^k \in \{0,1\}^k} \frac{1}{N_k - k + 1} \left| \sum_{i=0}^{N_k - k} \left(I_{\{u_{i+1}^{i+k} = x_1^k\}} - I_{\{v_{i+1}^{i+k} = x_1^k\}} \right) \right| < \epsilon_k. \tag{5}$$

Proof. First observe that (5) implies that, for all $u_1^{N_k}, v_1^{N_k} \in \mathcal{U}_{N_k}$, and for all $j \leqslant k$,

$$\sum_{x_1^j \in \{0,1\}^j} \frac{1}{N_k - k + 1} \left| \sum_{i=0}^{N_k - k} \left(I_{\{u_{i+1}^{i+j} = x_1^j\}} - I_{\{v_{i+1}^{i+j} = x_1^j\}} \right) \right| < \epsilon_k. \tag{6}$$

(Indeed,

$$\sum_{x_1^j\in\{0,1\}^j}\frac{1}{N_k-k+1}\left|\sum_{i=0}^{N_k-k}\left(I_{\{u_{i+1}^{i+j}=x_1^j\}}-I_{\{v_{i+1}^{i+j}=x_1^j\}}\right)\right|$$

$$=\sum_{x_1^j\in\{0,1\}^j}\frac{1}{N_k-k+1}\left|\sum_{i=0}^{N_k-k}\sum_{x_{j+1}^k\in\{0,1\}^{k-j}}\left(I_{\{u_{i+1}^{i+k}=x_1^k\}}-I_{\{v_{i+1}^{i+k}=x_1^k\}}\right)\right|$$

$$\leqslant\sum_{x_1^j\in\{0,1\}^j}\sum_{x_{j+1}^k\in\{0,1\}^{k-j}}\frac{1}{N_k-k+1}\left|\sum_{i=0}^{N_k-k}\left(I_{\{u_{i+1}^{i+k}=x_1^k\}}-I_{\{v_{i+1}^{i+k}=x_1^k\}}\right)\right|$$

$$=\sum_{x_1^k\in\{0,1\}^k}\frac{1}{N_k-k+1}\left|\sum_{i=0}^{N_k-k}\left(I_{\{u_{i+1}^{i+k}=x_1^k\}}-I_{\{v_{i+1}^{i+k}=x_1^k\}}\right)\right|$$

which is, by assumption, less than ϵ_k.) Now for any $M\leqslant k$ and $u_1^{N_k}$, $v_1^{N_k}\in\mathcal{U}_{N_k}$,

$$\sum_{x_1^M\in\{0,1\}^M}\frac{1}{N_k-M+1}\left|\sum_{i=0}^{N_k-M}\left(I_{\{u_{i+1}^{i+M}=x_1^M\}}-I_{\{v_{i+1}^{i+M}=x_1^M\}}\right)\right|$$

$$\leqslant\sum_{x_1^M\in\{0,1\}^M}\frac{1}{N_k-k+1}\left|\sum_{i=0}^{N_k-k}\left(I_{\{u_{i+1}^{i+M}=x_1^M\}}-I_{\{v_{i+1}^{i+M}=x_1^M\}}\right)\right|\frac{N_k-k+1}{N_k-M+1}$$

$$+\frac{k-M}{N_k-M+1}2^M$$

$$\leqslant\epsilon_k+\frac{k-M}{N_k-M+1}2^M,$$

where we used (6). Thus, for any $M\leqslant k$ and $u_1^{N_k}$, $v_1^{N_k}\in\mathcal{U}_{N_k}$,

$$\sum_{x_1^M\in\{0,1\}^M}\frac{1}{N_k-M+1}\left|\sum_{i=0}^{N_k-M}\left(I_{\{u_{i+1}^{i+M}=x_1^M\}}-I_{\{v_{i+1}^{i+M}=x_1^M\}}\right)\right|\leqslant\epsilon_k+\frac{k-M}{N_k-M+1}2^M. \quad (7)$$

Assume the process $\{X_n\}$ is stationary but not ergodic. Then, for some M and for some $a_1^M\in\{0,1\}^M$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}I_{\{X_{i+1}^{i+M}=a_1^M\}}$$

almost surely exists, but the limit is not a constant on any set of probability one (cf. Theorem 7.2.1 in Gray 1988). This means that there exist $\delta>0$ and a positive integer n_0 such that, for all $n>n_0$, there will be sets E_n, $F_n\subseteq\{0,1\}^n$ of probability $>10\delta$ such that, for all $u_1^n\in E_n$ and $v_1^n\in F_n$,

$$\frac{1}{n-M+1}\left|\sum_{i=0}^{n-M}\left(I_{\{u_{i+1}^{i+M}=a_1^M\}}-I_{\{v_{i+1}^{i+M}=a_1^M\}}\right)\right|>10\delta.$$

For M and δ above choose k large enough that $M<k$, $\epsilon_k<0.5\delta$, $2^M(k-M)/(N_k-M+1)<0.5\delta$, and $N_k>n_0$. (Such a k exists since $\epsilon_k\to 0$ and $k/N_k<1/k\to 0$.) However, this leads to a contradiction since $\mathcal{U}_{N_k}$ fills all but δ, while on sets E_{N_k} and F_{N_k}, which have probability at least 10δ, the empirical distributions differ. ($\mathcal{U}_{N_k}$ should have non-empty intersection with both E_{N_k} and F_{N_k}, and so on $\mathcal{U}_{N_k}$ the emprical distribution should differ by 10δ which contradicts (7) and the fact that $\epsilon_k+2^M(k-M)/(N_k-M+1)<\delta$.) The proof of Lemma 4 is complete. □

References

Bailey, D.H. (1976) Sequential schemes for classifying and predicting ergodic processes. Ph.D. thesis, Stanford University.

Cornfeld, I.P., Fomin, S.F. and Sinai, Ya.G. (1982) *Ergodic Theory*. New York: Springer-Verlag.

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

Gray, R.M. (1988) *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag.

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

Kalikow, S., Katznelson, Y. and Weiss, B. (1992) Finitarily deterministic generators for zero entropy systems. *Israel J. Math.*, **79**, 33–45.

Kemeny, J.G. and Snell, J.L. (1960) *Finite Markov Chains*. Princeton, NJ: Van Nostrand Reinhold.

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

Morvai, G. and Weiss, B. (2003) Forecasting for stationary binary time series. *Acta Appl. Math.*, **79**, 25–34.

Ornstein, D. and Weiss, B. (1990) How sampling reveals a process. *Ann Probab.*, **18**, 905–930.

Shields, P.C. (1996) *The Ergodic Theory of Discrete Sample Paths*. Providence, RI: American Mathematical Society.

Received October 2003 and revised June 2004