G. Morvai and B. Weiss: Limitations on intermittent forecasting.

Appeared in : Statist. Probab. Lett. 72 (2005), no. 4, pp. 285–290.

Abstract

Bailey showed that the general pointwise forecasting for stationary and ergodic time series has a negative solution. However, it is known that for Markov chains the problem can be solved. Morvai showed that there is a stopping time sequence $\{\lambda_n\}$ such that $P(X_{\lambda_n+1} = 1|X_0, \ldots, X_{\lambda_n})$ can be estimated from samples $(X_0, \ldots, X_{\lambda_n})$ such that the difference between the conditional probability and the estimate vanishes along these stoppping times for all stationary and ergodic binary time series. We will show it is not possible to estimate the above conditional probability along a stopping time sequence for all stationary and ergodic binary time series in a pointwise sense such that if the time series turns out to be a Markov chain, the predictor will predict eventually for all n.

Key words: Nonparametric estimation, prediction theory, stationary and ergodic processes, finite order Markov chains
Mathematics Subject Classifications (2000): 62G05, 60G25, 60G10

1 Introduction and Statement of Results

Cover [2] posed the following fundamental problem concerning forecasting for stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$. (Note that a stationary time series $\{X_n\}_{n=0}^{\infty}$ can be extended to be a two sided stationary time series $\{X_n\}_{n=-\infty}^{\infty}$.)

Problem 1

Is there an estimation scheme f_n for the value $P(X_{n+1} = 1|X_0, X_1, \ldots, X_n)$ such that f_n depends solely on the data segment $(X_0, X_1, \ldots, X_n)$ and

$$\lim_{n\to\infty} |f_n(X_0, X_1, \ldots, X_n) - P(X_{n+1} = 1|X_0, X_1, \ldots, X_n)| = 0$$

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$?

This problem was answered by Bailey [1] in a negative way, that is, he showed that there is no such scheme. (Also see Ryabko [10], Györfi, Morvai, Yakowitz [5] and Weiss [11].)

Morvai [8] considered the following modification of Problem 1.

Problem 2

Are there a strictly increasing sequence of stopping times $\{\lambda_n\}$ and estimators $\{h_n(X_0, \ldots, X_{\lambda_n})\}$ such that for all stationary ergodic binary time series $\{X_n\}$ the estimator h_n is consistent at stopping times λ_n, that is,

$$\lim_{n\to\infty} |h_n(X_0, \ldots, X_{\lambda_n}) - P(X_{\lambda_n+1} = 1|X_0, \ldots, X_{\lambda_n})| = 0$$

almost surely ?

Morvai [8] constructed a scheme that solves Problem 2. Unfortunatelly, his stopping times grow extremly rapidly and so that scheme is not practical at all.

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

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

Define the distance $d^*(\cdot, \cdot)$ on $\mathcal{X}^{*-}$ as follows. Let

$$d^*((\ldots, x_{-1}, x_0), (\ldots, y_{-1}, y_0)) = \sum_{i=0}^{\infty} 2^{-i-1}|x_{-i} - y_{-i}|.$$

DEFINITION The conditional probability $P(X_1 = 1|\ldots, X_{-1}, X_0)$ is almost surely continuous if to some set $C \subseteq \mathcal{X}^{*-}$ which has probability one the conditional probability $P(X_1 = 1|\ldots, X_{-1}, X_0)$ restricted to this set C is continuous with respect to metric $d^*(\cdot, \cdot)$.

The processes with almost surely continuous conditional probability generalizes the processes for which it is actually continuous, these are essentially the Random Markov Processes of Kalikow [6], or the continuous g-measures studied by Mike Keane [7].

A more moderate growth (compared to Morvai [8]) was achieved by Morvai and Weiss [9] but the consistency was secured only for the subclass of all stationary and ergodic binary time series with almost surely continuous conditional probability $P(X_1 = 1|\ldots, X_{-1}, X_0)$.

However for the class of all stationary and ergodic Markov-chains of some finite order Problem 1 can be solved. Indeed, if the time series is a Markov-chain of some finite order, we can estimate the order (e.g. as in Csiszár, Shields [3] and Csiszár [4]) and count frequencies of blocks with length equal to the order. Bailey showed that one can't test for being in the class.

It is conceivable that one can improve the result of Morvai [8] or Morvai and Weiss [9] so that if the process happens to be Markovian then one eventually estimates at all times. Our purpose in this paper is to show that this is not possible. This puts some new restrictions on what can be achieved in estimating along stopping times.

Theorem 1 *For any strictly increasing sequence of stopping times* $\{\lambda_n\}$ *such that for all stationary and ergodic binary Markov-chains with arbitrary finite order, eventually* $\lambda_{n+1} = \lambda_n + 1$, *and for any sequence of estimators* $\{h_n(X_0, \ldots, X_{\lambda_n})\}$ *there is a stationary and ergodic binary time series* $\{X_n\}$ *with almost surely continuous conditional probability* $P(X_1 = 1|\ldots, X_{-1}, X_0)$, *such that*

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

Remark: Bailey [1] among other things proved that there is no sequence of functions $\{e_n(X_0^{n-1})\}$ which for all stationary and ergodic time series, if

it turns out to be a Markov-chain, would be eventually 1 and 0 otherwise. (That is, there is no test for the Markov property.) This result does not imply ours. On the other hand, our result implies Bailey's. (Indeed, if there were a test for Markov-chains in the above sense, we could apply the estimator in Morvai [8] or Morvai and Weiss [9] if the time series is not a Markov-chain of some finite order, and if the time series is a Markov-chain of some finite order we can estimate the order of the Markov chain (e.g. as in Csiszár, Shields [3] or Csiszár [4]) and count frequencies of blocks with length equal to the order.

Bailey [1] and Ryabko [10] proved less than our Theorem 1. They proved the nonexistence of the desired estimator when the estimator should work for all stationary and ergodic binary time series and when all $\lambda_n = n$, that is, when we always require good prediction.

2 Proof of Theorem 1

PROOF:

The proof mainly follows the footsteps of Ryabko [10] and Györfi, Morvai, Yakowitz [5] with alterations where necessary. For $m \leq n$ let $X_m^n = (X_m, \ldots, X_n)$. First we define the same Markov-chain as in Ryabko [10] which serves as the technical tool for construction of our counterexample. Let the state space S be the non-negative integers. From state 0 the process certainly passes to state 1 and then to state 2, at the following epoch. From each state $s \geq 2$, the Markov chain passes either to state 0 or to state $s+1$ with equal probabilities 0.5. This construction yields a stationary and ergodic Markov chain $\{M_i\}$ with stationary distribution

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

and

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

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

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

Note that if $M_0 = 0$ then $M_i \leq 2k$ for $0 \leq i \leq \psi_k$. For each $0 \leq j < \infty$ we will define a binary-valued Markov-chain $\{X_i^{(j)}\}$ with some finite order,

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

Now observe that if for a certain $0 \leq j \leq \infty$, there is an index K_j such that $f^{(j)}(i) = 1$ for all $i \geq K_j$ then the defined process $\{X_n^{(j)}\}$ is a binary Markov-chain with order not greater than K_j. (Indeed, the probabilities $P(X_n^{(j)} = 1|X_0^{(j)}, \ldots, X_{n-1}^{(j)})$ are determined by the last K_j bits $(X_{n-K_j}^{(j)}, \ldots, X_{n-1}^{(j)})$. To see this consider the following cases.

I. If for some $1 \leq i \leq K_j - 2$ $X_{n-i}^{(j)} = 1$ and $X_{n-1-i}^{(j)} = X_{n-2-i}^{(j)} = 0$ than we can detect that $M_{n-i} = 2$, $M_{n-1-i} = 1$ and $M_{n-2-i} = 0$ and the conditional probability does not depend on previous values.

II. If there is no $1 \leq i \leq K_j - 2$ such that $X_{n-i}^{(j)} = 1$ and $X_{n-1-i}^{(j)} = X_{n-2-i}^{(j)} = 0$ we have three sub-cases.

II/1. If $X_{n-1}^{(j)} = 1$ then $M_{n-1} \geq K_j$. In this case the conditional probability is 0.5.

II/2. If $X_{n-2}^{(j)} = X_{n-1}^{(j)} = 0$ then $M_{n-1} = 1$ and the conditional probability is 1.

II/3. If $X_{n-2}^{(j)} = 1$ and $X_{n-1}^{(j)} = 0$ then $M_{n-1} = 0$ and so the conditional probability is 0.)

Now let $f^{(0)}(2k+1) = 1$ for all $k \geq 1$ and so the function $f^{(0)}$ is fully defined. Since $f^{(0)}(i)$ is eventually 1, the defined process $\{X_i^{(0)}\}$ is a stationary ergodic binary Markov-chain with some finite order.
For function $f^{(j)}$ and index $2k$, if $f^{(j)}(i)$ is defined for all $0 \leq i \leq 2k$, then it is easy to see that if $M_0 = 0$ (that is, $f^{(j)}(M_0) = 0$, $f^{(j)}(M_1) = 0$, $f^{(j)}(M_2) = 1$) then $M_i \leq 2k$ for $0 \leq i \leq \psi_k$ and the mapping

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

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

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

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

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

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

$$P(A_{j-1}(N_{j-1})) > 1/8.$$

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

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

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

$$B_j^+ = A_j(N_{j-1}) \bigcap \{h_n(f^{(j)}(M_0), \dots, f^{(j)}(M_{\psi_{N_{j-1}}})) \ge \frac{1}{4}\}$$

and

$$B_j^- = A_j(N_{j-1}) \bigcap \{h_n(f^{(j)}(M_0), \dots, f^{(j)}(M_{\psi_{N_{j-1}}})) < \frac{1}{4}\}.$$

Now notice that the sets B_j^+ and B_j^- do not depend on the future values of $f^{(j)}(2r+1)$ for $r \ge N_{j-1}$. One of the two sets B_j^+, B_j^- has at least probability $1/16$. Now we specify $f^{(j)}(2N_{j-1}+1)$. Let $f^{(j)}(2N_{j-1}+1) = 1$, $I_j = B_j^-$ if $P(B_j^-) \ge P(B_j^+)$ and let $f^{(j)}(2N_{j-1}+1) = 0$, $I_j = B_j^+$ if $P(B_j^-) < P(B_j^+)$.

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

$$\begin{aligned} & P(X^{(j)}_{\psi_{N_{j-1}}+1} = 1 | X^{(j)}_0, \ldots, X^{(j)}_{\psi_{N_{j-1}}}) \\ &= f^{(j)}(2N_{j-1}+1) P(X^{(j)}_{\psi_{N_{j-1}}+1} = f(2N_{j-1}+1) | X^{(j)}_0, \ldots, X^{(j)}_{\psi_{N_{j-1}}}) \\ &= f^{(j)}(2N_{j-1}+1) P(M_{\psi_{N_{j-1}+1}} = 2N_{j-1}+1 | M_0^{\psi_{N_{j-1}}}) \\ &= 0.5 f^{(j)}(2N_{j-1}+1). \end{aligned}$$

The difference of the estimate and the conditional probability is at least $\frac{1}{4}$ on set I_j and this event occurs with probability not less than $1/16$.

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

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

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

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

Finally by Fatou's Lemma,

$$\begin{aligned} & P(\limsup_{n\to\infty} \{|h_n(X_0^{\lambda_n}) - P(X_{\lambda_n+1} = 1 | X_0^{\lambda_n})| \geq 1/4\}) \\ \geq \;& P(\limsup_{j\to\infty} I_j) \geq \limsup_{j\to\infty} P(I_j) \geq \frac{1}{16}. \end{aligned}$$

Concerning the conditional probability $P(X_1 = 1 | X_{-\infty}^0)$ observe that as soon as one finds the pattern '001' in the sequence $X_{-\infty}^0$ the conditional probability does not depend on previous values. The probability of the occurence of $'001'$ in the past is one since the original Markov chain is ergodic and our process is therefore also ergodic. Thus the conditional probabilities are almost surely continuous. The proof of Theorem 1 is complete.

References

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

[2] T. M. Cover, "Open problems in information theory," in *1975 IEEE Joint Workshop on Information Theory*, pp. 35–36. New York: IEEE Press, 1975.

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

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

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

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

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

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

[9] G. Morvai and B. Weiss, "Forecasting for stationary binary time series," *Acta Applicandae Mathematicae*, vol. 79, pp. 25–34, 2003.

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

[11] B. Weiss, *Single Orbit Dynamics*, American Mathematical Society, 2000.