Gusztáv Morvai and Benjamin Weiss:

Forecasting for stationary binary time series.

Acta Appl. Math. 79 (2003), no. 1-2, 25–34.

Abstract

The forecasting problem for a stationary and ergodic binary time series $\{X_n\}_{n=0}^{\infty}$ is to estimate the probability that $X_{n+1}=1$ based on the observations X_i, $0 \le i \le n$ without prior knowledge of the distribution of the process $\{X_n\}$. It is known that this is not possible if one estimates at all values of n. We present a simple procedure which will attempt to make such a prediction infinitely often at carefully selected stopping times chosen by the algorithm. We show that the proposed procedure is consistent under certain conditions, and we estimate the growth rate of the stopping times.

1 Introduction

T. Cover [3] posed two fundamental problems concerning estimation 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 $\{X_n\}_{n=-\infty}^{\infty}$.) Cover's first problem was on backward estimation.

Problem 1 *Is there an estimation scheme f_{n+1} for the value $P(X_1 = 1|X_{-n}, \dots, X_0)$ such that f_{n+1} depends solely on the observed data segment $(X_{-n}, \dots, X_0)$ and*

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

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

This problem was solved by Ornstein [13] by constructing such a scheme. (See also Bailey [2].) Ornstein's scheme is not a simple one and the proof of consistency is rather sophisticated. A much simpler scheme and proof of consistency were provided by Morvai, Yakowitz, Györfi [12]. (See also Weiss [18].)

Cover's second problem was on forward estimation (forecasting).

Problem 2 *Is there an estimation scheme f_{n+1} for the value $P(X_{n+1} = 1|X_0, \dots, X_n)$ such that f_{n+1} depends solely on the data segment $(X_0, \dots, X_n)$ and*

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

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

This problem was answered by Bailey [2] in a negative way, that is, he showed that there is no such scheme. (Also see Ryabko [16], Györfi, Morvai, Yakowitz [7] and Weiss [18].) Bailey used the technique of cutting and stacking developed by Ornstein [14] (see also Shields [17]). Ryabko's construction was based on a function of an infinite state Markov-chain. This negative result can be interpreted as follows. Consider a market analyst whose task it is to predict the probability of the event 'the price of a certain share will go up tomorrow' given the observations up to the present day. Bailey's result says that the difference between the estimate and the true conditional probability cannot eventually be small for all stationary and ergodic market processes. The difference will be big infinitely often. These results show that there is a great difference between Problems 1 and 2. Problem 1 was addressed by Morvai, Yakowitz, Algoet [11] and a very simple estimation scheme was given which satisfies the statement in Problem 1 *in probability* instead of *almost surely*. However, for the class of all stationary and ergodic binary Markov-chains of some finite order Problem 2 can be solved. Indeed, if the time series is a Markov-chain of some finite (but unknown) order, we can estimate the order (e.g. as in Csiszár, Shields [5]) and count frequencies of blocks with length equal to the order.

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

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

Let $d(\cdot,\cdot)$ be the Hamming distance (that is for $x,y \in \{0,1\}$, $d(x,y)=0$ if and only if $x=y$ and $d(x,y)=1$ otherwise), and define the distance on sequences $(\dots,x_{-1},x_0,)$ and $(\dots,y_{-1},y_0)$ as follows. Let

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

(For details see Gray [6] p. 51.)

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

We note that from the proof of Ryabko [16] and Györfi, Morvai, Yakowitz [7] it is clear that even for the class of all stationary and ergodic binary time-series with almost surely continuous conditional probability $P(X_1=1|\dots,X_{-1},X_0)$ one can not solve Problem 2.

For $n\ge 1$, let the function $p_n(\cdot)$ be defined as

$$p_n(x_{-n+1},\dots,x_0) = P(X_{-n+1}=x_{-n+1},\dots,X_0=x_0) \tag{2}$$

where $x_{-i}\in\{0,1\}$ for $0\le i\le n-1$.

The entropy rate H associated with a stationary binary time-series $\{X_n\}_{-\infty}^{\infty}$ is defined as $H = \lim_{n\to\infty} -\frac{1}{n} E \log_2 p_n(X_{-n+1},\dots,X_{-1},X_0)$. We note that the entropy rate of a stationary binary time-series always exists. For details cf. Cover, Thomas [4], pp. 63-64.

Now we may pose our problem.

Problem 3 *Is there a sequence of strictly increasing stopping times $\{\lambda_n\}$ with*

$$\lambda_n \le 2^{n(H+\epsilon)}$$

and an estimation scheme $f_n(X_0,\dots,X_{\lambda_n})$ which depends on the observed data segment $(X_0,\dots,X_{\lambda_n})$ such that

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

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$ with almost surely continuous conditional probability $P(X_1=1|\dots,X_{-1},X_0)$?

It turns out that the answer is affirmative and such a scheme will be exhibited below. This result can be interpreted as if the market analyst can refrain from predicting, that is, he may say that he does not want to predict today, but will predict at infinitely many time instances, and not too rarely, since $\lambda_n \le 2^{n(H+\epsilon)}$, and the difference between the prediction and the true conditional probability will vanish almost surely at these stopping times. We note that the stationary processes with almost surely continuous conditional distribution generalize the processes for which the conditional distribution is actually continuous, these are essentially the Random Markov Processes of Kalikow [8], or the continuous g-measures studied by Mike Keane in [9]. Morvai [10] proposed a different estimator which is consistent on a certain stopping time sequence, but those stopping times grow like an exponential tower which is unrealistic and much faster growth than the mere exponential one in Problem 3.

2 The Proposed Estimator

Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary time series taking values from a binary alphabet $\mathcal{X} = \{0, 1\}$. (Note that all stationary time series $\{X_n\}_{n=0}^{\infty}$ can be thought to be a two sided time series, that is, $\{X_n\}_{n=-\infty}^{\infty}$.) Now we exhibit an estimator which is consistent on a certain stopping time sequence for a restricted class of stationary time series. For notational convenience, let $X_m^n = (X_m, \dots, X_n)$, where $m \le n$.

Define the stopping times as follows. Set $\zeta_0 = 0$. For $k = 1, 2, \dots$, define sequence η_k and ζ_k recursively. Let

$$\eta_k = \min\{t > 0 : X_{\zeta_{k-1}-(k-1)+t}^{\zeta_{k-1}+t} = X_{\zeta_{k-1}-(k-1)}^{\zeta_{k-1}}\} \text{ and } \zeta_k = \zeta_{k-1} + \eta_k.$$

One denotes the kth estimate of $P(X_{\zeta_k+1} = 1|X_0^{\zeta_k})$ by g_k, and defines it to be

$$g_k = \frac{1}{k}\sum_{j=0}^{k-1} X_{\zeta_j+1}. \tag{3}$$

It will be useful to define other processes $\{\tilde{X}_n\}_{n=-\infty}^{0}$ and $\{\hat{X}_n^{(k)}\}_{n=-\infty}^{\infty}$ for $k \ge 0$ as follows. Let

$$\tilde{X}_{-n} = X_{\zeta_n - n} \text{ for } n \ge 0, \text{ and } \hat{X}_n^{(k)} = X_{\zeta_k+n} \text{ for } -\infty < n < \infty. \tag{4}$$

For an arbitrary stationary binary time series $\{Y_n\}$, and for all $k \ge 1$ and $1 \le i \le k$ define $\hat{\zeta}_0^k(Y_{-\infty}^0) = 0$ and

$$\hat{\eta}_i^k(Y_{-\infty}^0) = \min\{t > 0 : Y_{\hat{\zeta}_{i-1}^k-(k-i)-t}^{\hat{\zeta}_{i-1}^k-t} = Y_{\hat{\zeta}_{i-1}^k-(k-i)}^{\hat{\zeta}_{i-1}^k}\}$$

and

$$\hat{\zeta}_i^k(Y_{-\infty}^0) = \hat{\zeta}_{i-1}^k(Y_{-\infty}^0) - \hat{\eta}_i^k(Y_{-\infty}^0).$$

When it is obvious on which time series $\hat{\eta}_i^k(Y_{-\infty}^0)$ and $\hat{\zeta}_i^k(Y_{-\infty}^0)$ are evaluated, we will use the notation $\hat{\eta}_i^k$ and $\hat{\zeta}_i^k$. Let T denote the left shift operator, that is, $(Tx_{-\infty}^{\infty})_i = x_{i+1}$. It is easy to see that if $\zeta_k(x_{-\infty}^{\infty}) = l$ then $\hat{\zeta}_k^k(T^l x_{-\infty}^{\infty}) = -l$.

We will need the next lemma for later use.

Lemma 1 *Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary binary process. Then the time series $\{\hat{X}_n^{(k)}\}_{n=-\infty}^{\infty}$, $\{\tilde{X}_n\}_{n=-\infty}^{0}$ and $\{X_n\}_{n=-\infty}^{\infty}$ have identical distribution. Thus all these time series are stationary, and $\{\tilde{X}_n\}_{n=-\infty}^{0}$ can be thought to be two sided stationary time series $\{\tilde{X}_n\}_{n=-\infty}^{\infty}$.*

Let $k \ge 0$, $n \ge 0$, $m \ge 0$, $x_{m-n}^m \in \mathcal{X}^{n+1}$ be arbitrary. It is immediate that for $l \ge 0$,

$$T^l\{X_{\zeta_k+m-n}^{\zeta_k+m} = x_{m-n}^m, \zeta_k = l\} = \{X_{m-n}^m = x_{m-n}^m, \hat{\zeta}_k^k(X_{-\infty}^0) = -l\}. \tag{5}$$

First we prove that for $k \geq 0$, $P((\hat{X}^{(k)}_{m-n},\ldots,\hat{X}^{(k)}_{m})=(x_{m-n},\ldots,x_m))=P(X^m_{m-n}=x^m_{m-n})$. By the construction in (4), the stationarity of the time series $\{X_n\}$, and (5) we have

$$\begin{aligned}
&P((\hat{X}^{(k)}_{m-n},\ldots,\hat{X}^{(k)}_{m})=(x_{m-n},\ldots,x_m))\\
&\quad= P(X^{\zeta_k+m}_{\zeta_k+m-n}=x^m_{m-n})\\
&\quad= \sum_{l=0}^{\infty}P(X^{\zeta_k+m}_{\zeta_k+m-n}=x^m_{m-n},\zeta_k=l)\\
&\quad= \sum_{l=0}^{\infty}P(X^{m}_{m-n}=x^m_{m-n},\hat{\zeta}^k_k(X^0_{-\infty})=-l)\\
&\quad= P(X^m_{m-n}=x^m_{m-n}).
\end{aligned}$$

Now we prove that $P(\tilde{X}^0_{-n}=x^0_{-n})=P(X^0_{-n}=x^0_{-n})$. By the construction in (4), the stationarity of the time series $\{X_n\}$, and (5) (with $m=0$) we have

$$\begin{aligned}
&P(\tilde{X}^0_{-n}=x^0_{-n})=P(X^{\zeta_n}_{\zeta_n-n}=x^0_{-n})\\
&\quad= \sum_{l=0}^{\infty}P(X^{\zeta_n}_{\zeta_n-n}=x^0_{-n},\zeta_n=l)\\
&\quad= \sum_{l=0}^{\infty}P(X^{0}_{-n}=x^0_{-n},\hat{\zeta}^n_n(X^0_{-\infty})=-l)\\
&\quad= P(X^0_{-n}=x^0_{-n}).
\end{aligned}$$

The proof of the Lemma is complete.
Now we show the consistency of our estimate g_k defined in (3).

Theorem 1 *Let $\{X_n\}$ be a stationary binary time series. For the estimator defined in (3),*

$$\lim_{k\to\infty}\left|g_k-P(X_{\zeta_k+1}=1|X^{\zeta_k}_0)\right|=0 \quad \textit{almost surely}$$

provided that the conditional probability $P(X_1=1|X^0_{-\infty})$ is almost surely continuous. Moreover, under the same conditions,

$$\lim_{k\to\infty}g_k=\lim_{k\to\infty}P(X_{\zeta_k+1}=1|X^{\zeta_k}_0)=P(\tilde{X}_1=1|\tilde{X}^0_{-\infty}) \quad \textit{almost surely.}$$

Recalling (3) we can write

$$\begin{aligned}
g_k &= \frac{1}{k}\sum_{j=0}^{k-1}[X_{\zeta_j+1}-P(X_{\zeta_j+1}=1|X^{\zeta_j}_{-\infty})]+\frac{1}{k}\sum_{j=0}^{k-1}P(X_{\zeta_j+1}=1|X^{\zeta_j}_{-\infty})\\
&= \frac{1}{k}\sum_{j=0}^{k-1}\Gamma_j+\frac{1}{k}\sum_{j=0}^{k-1}P(X_{\zeta_j+1}=1|X^{\zeta_j}_{-\infty}). \qquad (6)
\end{aligned}$$

Observe that $\{\Gamma_j,\sigma(X^{\zeta_j+1}_{-\infty})\}$ is a bounded martingale difference sequence for $0\leq j<\infty$. To see this notice that $\sigma(X^{\zeta_j+1}_{-\infty})$ is monotone increasing, and Γ_j is measurable with respect to

$\sigma(X_{-\infty}^{\zeta_j+1})$, and $E(\Gamma_j | X_{-\infty}^{\zeta_{j-1}+1}) = 0$ for $0 \le j < \infty$ (where you may define $\zeta_{-1} = -1$). Now apply Azuma's exponential bound for bounded martingale differences in Azuma [1] to get that for any $\epsilon > 0$,

$$P\left(\left| \frac{1}{k} \sum_{j=0}^{k-1} \Gamma_j \right| > \epsilon \right) \le 2 \exp(-\epsilon^2 k/2).$$

After summing the right hand side over k, and appealing to the Borel-Cantelli lemma for a sequence of ϵ's tending to zero we get $\frac{1}{k}\sum_{j=0}^{k-1}\Gamma_j \to 0$ almost surely.
Define the function $p : \mathcal{X}^{*-} \to [0,1]$ as $p(x_{-\infty}^0) = P(X_1 = 1 | X_{-\infty}^0 = x_{-\infty}^0)$.
For arbitrary $j \ge 0$, by the construction in (4),

$$X_{\zeta_j - j}^{\zeta_j} = (\hat{X}_{-j}^{(j)}, \dots, \hat{X}_0^{(j)}) = \tilde{X}_{-j}^0 \quad \text{and} \quad \lim_{j\to\infty} d^*(\tilde{X}_{-\infty}^0, (\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)})) = 0 \qquad (7)$$

almost surely. By assumption, the function $p(\cdot)$ is continuous on a set $C \subseteq \mathcal{X}^{*-}$ with $P(X_{-\infty}^0 \in C) = 1$, and by the Lemma, $P(\tilde{X}_{-\infty}^0 \in C) = 1$, and for each $j \ge 0$, $P((\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) \in C) = 1$, and finally,

$$P(\tilde{X}_{-\infty}^0 \in C, (\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) \in C \ \text{ for all } j \ge 0) = 1.$$

By the Lemma, the construction in (4), the continuity of $p(\cdot)$ on the set C, and by (7)

$$P(X_{\zeta_j+1} = 1 | X_{-\infty}^{\zeta_j}) = p(\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) \to p(\tilde{X}_{-\infty}^0) = P(\tilde{X}_1 = 1 | \tilde{X}_{-\infty}^0)$$

and $\frac{1}{k}\sum_{j=0}^{k-1} P(X_{\zeta_j+1} = 1 | X_{-\infty}^{\zeta_j}) \to P(\tilde{X}_1 = 1 | \tilde{X}_{-\infty}^0)$ almost surely. We have proved that $g_k \to P(\tilde{X}_1 = 1 | \tilde{X}_{-\infty}^0)$ almost surely.
Now observe that by (1) and the continuity of $p(\cdot)$ on the set C, almost surely, for all $\epsilon > 0$, there is a $J(\epsilon, \tilde{X}_{-\infty}^0)$, such that for all $z_{-\infty}^0 \in C$, if $z_{-J}^0 = \tilde{X}_{-J}^0$ then $|p(z_{-\infty}^0) - p(\tilde{X}_{-\infty}^0)| < \epsilon$.
By (7), and since $\epsilon > 0$ was arbitrary, almost surely,

$$\begin{aligned}
\lim_{j\to\infty} P(X_{\zeta_j+1} = 1 | X_0^{\zeta_j}) &= \lim_{j\to\infty} E\{P(X_{\zeta_j+1} = 1 | X_{-\infty}^{\zeta_j}) | X_0^{\zeta_j}\} \\
&= \lim_{j\to\infty} E\{p(X_{-\infty}^{\zeta_j}) | X_0^{\zeta_j}\} \\
&= p(\tilde{X}_{-\infty}^0) = P(\tilde{X}_1 = 1 | \tilde{X}_{-\infty}^0).
\end{aligned}$$

The proof of Theorem 1 is complete.
Remark. We note that for all stationary binary time-series, the estimation scheme described above is consistent in probability. This may be seen as follows:

$$\begin{aligned}
&E\left| g_k - P(X_{\zeta_k+1} = 1 | X_0^{\zeta_k}) \right| \\
&\le \ E\left| \frac{1}{k}\sum_{j=0}^{k-1} [X_{\zeta_j+1} - P(X_{\zeta_j+1} = 1 | X_{-\infty}^{\zeta_j})] \right| \\
&+ \ \frac{1}{k}\sum_{j=0}^{k-1} E\left| P(\hat{X}_1^{(j)} = 1 | \dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) - P(\hat{X}_1^{(j)} = 1 | \hat{X}_{-j}^{(j)}, \dots, \hat{X}_0^{(j)}) \right| \\
&+ \ E\left| \frac{1}{k}\sum_{j=0}^{k-1} P(\hat{X}_1^{(k)} = 1 | \hat{X}_{-j}^{(k)}, \dots, \hat{X}_0^{(k)}) - P(\hat{X}_1^{(k)} = 1 | \hat{X}_{\hat{\zeta}_k^k}^{(k)}, \dots, \hat{X}_0^{(k)}) \right|,
\end{aligned}$$

where we used (7) and the Lemma. The first term converges to zero since $X_{\zeta_j+1} - P(X_{\zeta_j+1} = 1|X_{-\infty}^{\zeta_j})$ is a martingale difference sequence with respect to $\sigma(X_{-\infty}^{\zeta_j+1})$ and an average of bounded martingale differences converges to zero almost surely cf. Azuma [1]. Applying (4), (7) and the Lemma, the sum of the last two terms can be estimated by the sum

$$\begin{aligned}
&\frac{1}{k}\sum_{j=0}^{k-1} E\left|P(X_1 = 1|X^0_{-\infty}) - P(X_1 = 1|X^0_{-j})\right| \\
+ \quad & E\left|\frac{1}{k}\sum_{j=0}^{k-1} P(X_1 = 1|X^0_{-j}) - P(X_1 = 1|X^0_{\hat{\zeta}^k_k})\right|
\end{aligned}$$

and both terms converge to zero since by the martingale convergence theorem $\lim_{j\to\infty} P(X_1 = 1|X^0_{-j}) = P(X_1 = 1|X^0_{-\infty})$ almost surely, and thus the limit in fact exists and equals zero. Next we will give some universal estimates for the growth rate of the stopping times ζ_k in terms of the entropy rate of the process. This is natural since the ζ_k are defined by recurrence times for blocks of length k, and these are known to grow exponentially with the entropy rate. (Cf. Ornstein and Weiss [15].)

Theorem 2 *Let $\{X_n\}$ be a stationary and ergodic binary time series. Then for arbitrary $\epsilon > 0$,*

$$\zeta_k < 2^{k(H+\epsilon)} \quad \textit{eventually almost surely,}$$

where H denotes the entropy rate associated with time series $\{X_n\}$.

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

$$\mathcal{X}^* = \{(\ldots, x_{-1}, x_0, x_1, \ldots) : x_i \in \{0,1\} \text{ for all } -\infty < i < \infty\}.$$

Define $B_k \subseteq \{0,1\}^k$ as

$$B_k = \{x^0_{-k+1} \in \{0,1\}^k : 2^{-k(H+0.5\epsilon)} < p_k(x^0_{-k+1})\},$$

where $p_k(\cdot)$ is as in (2). Note that there is a trivial bound on the cardinality of the set B_k, namely,

$$|B_k| \le 2^{k(H+0.5\epsilon)}. \tag{8}$$

By the Lemma, the distribution of the time series $\{\tilde{X}_n\}$ is the same as the distribution of $\{X_n\}$ and by the Shannon-McMillan-Breiman Theorem (cf. Cover, Thomas [4], p. 475),

$$P\left(\bigcup_{k=1}^{\infty}\bigcap_{i\ge k}\{\tilde{X}^0_{-i+1} \in B_i\}\right) = 1. \tag{9}$$

Define the set $Q_k(y^0_{-k+1})$ as follows:

$$Q_k(y^0_{-k+1}) = \{z^{\infty}_{-\infty} \in \mathcal{X}^* : -\hat{\zeta}^k_k(z^0_{-\infty}) \ge 2^{k(H+\epsilon)}, z^0_{-k+1} = y^0_{-k+1})\}.$$

We will estimate the probability of $Q_k(y^0_{-k+1})$ by means of the ergodic theorem. Let $x^\infty_{-\infty} \in \mathcal{X}^*$ be a typical sequence of the time series $\{X_n\}$. Define $\alpha_0(y^0_{-k+1}) = 0$ and for $i \geq 1$ let

$$\alpha_i(y^0_{-k+1}) = \min\{l > \alpha_{i-1}(y^0_{-k+1}) : T^{-l}x^\infty_{-\infty} \in Q_k(y^0_{-k+1})\}.$$

Define also $\beta_0(y^0_{-k+1}) = 0$ and for $i \geq 1$ let

$$\beta_i(y^0_{-k+1}) = \min\{l > \beta_{i-1}(y^0_{-k+1}) + 2^{k(H+\epsilon)} : T^{-l}x^\infty_{-\infty} \in Q_k(y^0_{-k+1})\}.$$

Observe that for arbitrary $l > 0$,

$$\sum_{j=1}^{\infty} 1_{\{\beta_{l-1}(y^0_{-k+1}) < \alpha_j(y^0_{-k+1}) \leq \beta_l(y^0_{-k+1})\}} \leq k+1.$$

By the Lemma and the ergodicity of the time series $\{X_n\}$,

$$\begin{aligned}
&P((\dots,\hat{X}^{(k)}_{-1},\hat{X}^{(k)}_0,\hat{X}^{(k)}_1,\dots) \in Q_k(y^0_{-k+1})) = P(X^\infty_{-\infty} \in Q_k(y^0_{-k+1})) \\
&= \lim_{t\to\infty} \frac{1}{\beta_t(y^0_{-k+1})} \sum_{j=1}^{\infty} 1_{\{\alpha_j(y^0_{-k+1}) \leq \beta_t(y^0_{-k+1})\}} \\
&= \lim_{t\to\infty} \frac{1}{\beta_t(y^0_{-k+1})} \sum_{l=1}^{t}\sum_{j=1}^{\infty} 1_{\{\beta_{l-1}(y^0_{-k+1}) < \alpha_j(y^0_{-k+1}) \leq \beta_l(y^0_{-k+1})\}} \\
&\leq \lim_{t\to\infty} \frac{t(k+1)}{t2^{k(H+\epsilon)}} = \frac{(k+1)}{2^{k(H+\epsilon)}}. \qquad (10)
\end{aligned}$$

By the construction in (4), $-\hat{\zeta}^k_k(\dots,\hat{X}^{(k)}_{-1},\hat{X}^{(k)}_0) = \zeta_k(X^\infty_0)$, and $(\hat{X}^{(k)}_{-k+1},\dots,\hat{X}^{(k)}_0) = \tilde{X}^0_{-k+1}$ and by the upper bound on the cardinality of set B_k in (8) and by (10), we get

$$\begin{aligned}
&P(\zeta_k(X^\infty_0) \geq 2^{k(H+\epsilon)}, \tilde{X}^0_{-k+1} \in B_k) \\
&= P(-\hat{\zeta}^k_k(\dots,\hat{X}^{(k)}_{-1},\hat{X}^{(k)}_0) \geq 2^{k(H+\epsilon)}, \tilde{X}^0_{-k+1} \in B_k) \\
&= P(-\hat{\zeta}^k_k(\dots,\hat{X}^{(k)}_{-1},\hat{X}^{(k)}_0) \geq 2^{k(H+\epsilon)}, (\hat{X}^{(k)}_{-k+1},\dots,\hat{X}^{(k)}_0) \in B_k) \\
&= \sum_{y^0_{-k+1}\in B_k} P((\dots,\hat{X}^{(k)}_{-1},\hat{X}^{(k)}_0,\hat{X}^{(k)}_1,\dots) \in Q_k(y^0_{-k+1})) \leq (k+1)2^{-k0.5\epsilon}.
\end{aligned}$$

The right hand side sums, the Borel-Cantelli Lemma and the Shannon-McMillan-Breiman Theorem in (9) together yield that $\zeta_k < 2^{k(H+\epsilon)}$eventually almost surely and Theorem 2 is proved.

References

[1] K. Azuma, "Weighted sums of certain dependent random variables," in *Tohoku Mathematical Journal,* vol. 37, pp. 357–367, 1967.

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

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

[4] T.M. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 1991.

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

[6] R.M. Gray, *Probability, Random Processes, and Ergodic Properties.* Springer-Verlag, New York, 1988.

[7] 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.

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

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

[10] G. Morvai "Guessing the output of a stationary binary time series" *In: Foundations of statistical inference (Shoresh)*, pp. 207–215, Contrib. Statist., Physica, Heidelberg, 2003.

[11] G. Morvai, S. Yakowitz, and P. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Transactions on Information Theory*, vol. 43, pp. 483-498, 1997.

[12] G. Morvai, S. Yakowitz, and L. Györfi, "Nonparametric inferences for ergodic, stationary time series," *Annals of Statistics.*, vol. 24, pp. 370–379, 1996.

[13] D. S. Ornstein, "Guessing the next output of a stationary process," *Israel Journal of Mathematics,* vol. 30, pp. 292–296, 1978.

[14] D. S. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems.* Yale University Press, 1974.

[15] D. S. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Transactions on Information Theory*, vol. 39, pp. 78–83, 1993.

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

[17] P.C. Shields, "Cutting and stacking: a method for constructing stationary processes," *IEEE Transactions on Information Theory,* vol. 37, pp. 1605–1614, 1991.

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