Probab. Theory Relat. Fields 132, 1–12 (2005)
Digital Object Identifier (DOI) 10.1007/s00440-004-0386-3

Gusztáv Morvai · Benjamin Weiss

Prediction for discrete time series

Received: 6 September 2003 / Revised version: 30 June 2004 /
Published online: 9 October 2004 –

Abstract. Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite or countably infinite set $\mathcal{X}$. Assume that the distribution of the process is otherwise unknown. We propose a sequence of stopping times λ_n along which we will be able to estimate the conditional probability $P(X_{\lambda_n+1} = x|X_0, \ldots, X_{\lambda_n})$ from data segment $(X_0, \ldots, X_{\lambda_n})$ in a pointwise consistent way for a restricted class of stationary and ergodic finite or countably infinite alphabet time series which includes among others all stationary and ergodic finitarily Markovian processes. If the stationary and ergodic process turns out to be finitarily Markovian (among others, all stationary and ergodic Markov chains are included in this class) then $\lim_{n\to\infty} \frac{n}{\lambda_n} > 0$ almost surely. If the stationary and ergodic process turns out to possess finite entropy rate then λ_n is upperbounded by a polynomial, eventually almost surely.

1. Introduction

Bailey [1] and Ryabko [14] considered the problem of estimating the conditional probability $P(X_{n+1} = 1|X_0, \ldots, X_n)$ for binary time series. They showed that one cannot estimate this quantity from the data $(X_0, \ldots, X_n)$ such that the difference tends to zero almost surely as n increases, for all stationary and ergodic binary time series.

It is well known, that if one knows in advance that the process is Markov with arbitrary (unknown) order, then one can estimate the order (c.f. Csiszár and Shields [4], Csiszár [5]), and using this estimate for the order, one can count empirical averages of blocks with lengths one plus the order for estimating $P(X_{n+1} = 1|X_0, \ldots, X_n)$ in a pointwise consistent way. In the present paper we will consider the case when it is not known in advance if the process is Markov or not.

Morvai [11] exhibited a sequence of stopping times η_n such that $P(X_{\eta_n+1} = 1|X_0, \ldots, X_{\eta_n})$ can be estimated from data segment $(X_0, \ldots, X_{\eta_n})$ in a pointwise consistent way, that is, the error vanishes as n increases. The disadvantage of that scheme is that the stopping times grow very fast. Another, more reasonable scheme was proposed by Morvai and Weiss [12] for a subclass of stationary and ergodic

G. Morvai: Research Group for Informatics and Electronics of the Hungarian Academy of Sciences, Budapest, 1521 Goldmann György tér 3, Hungary.
e-mail: `morvai@szit.bme.hu`

B. Weiss: Hebrew University of Jerusalem, Jerusalem 91904 Israel.
e-mail: `weiss@math.huji.ac.il`

Mathematics Subject Classification (2000): 62G05, 60G25, 60G10

Key words or phrases: Nonparametric estimation – Stationary processes

binary time series. There the stopping times still grow exponentially, though not so fast as in Morvai [11].

Bailey [1] proved that there is no test for the Markov property, that is, there is no algorithm which could tell you eventually if the process is Markov with any order or not, over all stationary and ergodic binary time series.

In this paper discrete (finite or countably infinite) alphabet stationary and ergodic processes are treated. We propose a much denser (compared to Morvai and Weiss [12]) sequence of stopping times λ_n along which we will be able to estimate $P(X_{\lambda_n+1}=x|X_0,\dots,X_{\lambda_n})$ from samples $(X_0,\dots,X_{\lambda_n})$ in a pointwise consistent way for those processes whose conditional distribution is almost surely continuous (see the precise definition below). This class includes all Markov processes with arbitrary order and the much wider class of finitarily Markovian processes. Despite Bailey's result, for the proposed stopping times λ_n, if the stationary and ergodic process turns out to be finitarily Markovian (which includes all stationary and ergodic Markov chains with arbitrary order) then $\lim_{n\to\infty}\frac{n}{\lambda_n}>0$ almost surely. If the stationary and ergodic process turns out to possess finite entropy rate then λ_n is upperbounded by a polynomial, eventually almost surely.

2. The proposed algorithm

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 $m\le n$. Note that if $m>n$ then X_m^n is the empty string.

For $k\ge 1$, let $1\le l_k\le k$ be a nondecreasing unbounded sequence of integers, that is, $1=l_1\le l_2\dots$ and $\lim_{k\to\infty} l_k=\infty$.

Define auxiliary stopping times (similarly to Morvai and Weiss [12]) as follows. Set $\zeta_0=0$. For $n=1,2,\dots,$ let

$$\zeta_n=\zeta_{n-1}+\min\{t>0: X_{\zeta_{n-1}-(l_n-1)+t}^{\zeta_{n-1}+t}=X_{\zeta_{n-1}-(l_n-1)}^{\zeta_{n-1}}\}. \tag{1}$$

Among other things, using ζ_n and l_n we can define a very useful process $\{\tilde{X}_n\}_{n=-\infty}^{0}$ as a function of X_0^{∞} as follows. Let $J(n)=\min\{j\ge 1: l_{j+1}>n\}$ and define

$$\tilde{X}_{-i}=X_{\zeta_{J(i)}-i} \text{ for } i\ge 0. \tag{2}$$

As we will see in the proof of the Theorem, the $\{\tilde{X}\}_{n=-\infty}^{0}$ has the same distribution as the original process. For notational convenience let $p_k(x_{-k}^0)$ and $p_k(y|x_{-k}^0)$ denote the distribution $P(X_{-k}^0=x_{-k}^0)$ and the conditional distribution $P(X_1=y|X_{-k}^0=x_{-k}^0)$, respectively.

Definition 1. *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 $0\le K<\infty$ such that for all $i\ge 1$, all $y\in\mathcal{X}$, all $z_{-K-i+1}^{-K}\in\mathcal{X}^i$*

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

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

Definition 2. *The stationary time series* $\{X_n\}$ *is said to be finitarily Markovian if* $K(X_{-\infty}^{0})$ *is finite (though not necessarily bounded) almost surely.*

In order to estimate $K(\tilde{X}_{-\infty}^{0})$ we need to define some explicit statistics.

Define

$$\Delta_k(\tilde{X}_{-k+1}^{0}) = \sup_{1\le i} \sup_{\{z_{-k-i+1}^{-k}\in\mathcal{X}^i, x\in\mathcal{X}: p_{k+i}(z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^{0}, x)>0\}} \left| p_{k-1}(x|\tilde{X}_{-k+1}^{0}) - p_{k+i-1}(x|(z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^{0})) \right|.$$

We will divide the data segment X_0^n into two parts: $X_0^{\lceil \frac{n}{2}\rceil -1}$ and $X_{\lceil \frac{n}{2}\rceil}^{n}$. Let $\mathcal{L}_{n,k}^{(1)}$ denote the set of strings with length $k+1$ which appear at all in $X_0^{\lceil \frac{n}{2}\rceil -1}$. That is,

$$\mathcal{L}_{n,k}^{(1)} = \{x_{-k}^{0} \in \mathcal{X}^{k+1} : \exists k \le t \le \lceil \frac{n}{2} \rceil - 1 : X_{t-k}^{t} = x_{-k}^{0}\}.$$

For a fixed $0 < \gamma < 1$ let $\mathcal{L}_{n,k}^{(2)}$ denote the set of strings with length $k+1$ which appear more than $n^{1-\gamma}$ times in $X_{\lceil \frac{n}{2}\rceil}^{n}$. That is,

$$\mathcal{L}_{n,k}^{(2)} = \{x_{-k}^{0} \in \mathcal{X}^{k+1} : \#\{\lceil \frac{n}{2} \rceil + k \le t \le n : X_{t-k}^{t} = x_{-k}^{0}\} > n^{1-\gamma}\}.$$

Let

$$\mathcal{L}_k^n = \mathcal{L}_{n,k}^{(1)} \bigcap \mathcal{L}_{n,k}^{(2)}.$$

We define the empirical version of Δ_k as follows:

$$\begin{aligned}
&\hat{\Delta}_k^n(\tilde{X}_{-k+1}^{0}) \\
&\quad = \max_{1\le i\le n} \max_{(z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^{0}, x)\in\mathcal{L}_{k+i}^n} 1_{\{\zeta_{J(k)}\le \lceil \frac{n}{2}\rceil -1\}} \\
&\qquad \times \left| \frac{\#\{\lceil \frac{n}{2}\rceil + k \le t \le n : X_{t-k}^{t} = (\tilde{X}_{-k+1}^{0}, x)\}}{\#\{\lceil \frac{n}{2}\rceil + k - 1 \le t \le n-1 : X_{t-k+1}^{t} = \tilde{X}_{-k+1}^{0}\}} \right. \\
&\qquad \left. - \frac{\#\{\lceil \frac{n}{2}\rceil + k + i \le t \le n : X_{t-k-i}^{t} = (z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^{0}, x)\}}{\#\{\lceil \frac{n}{2}\rceil + k + i - 1 \le t \le n-1 : X_{t-k-i+1}^{t} = (z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^{0})\}} \right|.
\end{aligned}$$

Note that the cut off $1_{\{\zeta_{J(k)}\le \lceil \frac{n}{2}\rceil -1\}}$ ensures that $\tilde{X}_{-k+1}^{0}$ is defined from $X_0^{\lceil \frac{n}{2}\rceil -1}$.

Observe, that by ergodicity, for any fixed k,

$$\liminf_{n\to\infty} \hat{\Delta}_k^n \ge \Delta_k \text{ almost surely.} \tag{3}$$

We define an estimate χ_n for $K(\tilde{X}_{-\infty}^{0})$ from samples X_0^n as follows. Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. Set $\chi_0 = 0$, and for $n \ge 1$ let χ_n be the smallest $0 \le k_n < n$ such that $\hat{\Delta}_{k_n}^n \le n^{-\beta}$.

Observe that if $\zeta_j \le \lceil \frac{n}{2} \rceil - 1 < \zeta_{j+1}$ then $\chi_n \le l_{j+1}$.

Here the idea is (cf. the proof of the Theorem) that if $K(\tilde{X}_{-\infty}^0) < \infty$ then χ_n will be equal to $K(\tilde{X}_{-\infty}^0)$ eventually and if $K(\tilde{X}_{-\infty}^0) = \infty$ then $\chi_n \to \infty$.

Now we define the sequence of stopping times λ_n along which we will be able to estimate. Set $\lambda_0 = \zeta_0$, and for $n \ge 1$ if $\zeta_j \le \lambda_{n-1} < \zeta_{j+1}$ then put

$$\lambda_n = \min\{t > \lambda_{n-1} : X_{t-\chi_t+1}^t = X_{\zeta_j-\chi_t+1}^{\zeta_j}\} \tag{4}$$

and

$$\kappa_n = \chi_{\lambda_n}. \tag{5}$$

Observe that if $\zeta_j \le \lambda_{n-1} < \zeta_{j+1}$ then $\zeta_j \le \lambda_{n-1} < \lambda_n \le \zeta_{j+1}$. If $\chi_{\lambda_{n-1}+1} = 0$ then $\lambda_n = \lambda_{n-1} + 1$. Note that λ_n is a stopping time and κ_n is our estimate for $K(\tilde{X}_{-\infty}^0)$ from samples $X_0^{\lambda_n}$.

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

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

Let $f : \mathcal{X} \to (-\infty, \infty)$ be bounded, otherwise arbitrary. Define the function $F : \mathcal{X}^{*-} \to (-\infty, \infty)$ as

$$F(x_{-\infty}^0) = E(f(X_1)|X_{-\infty}^0 = x_{-\infty}^0).$$

E.g. if $f(x) = 1_{\{x=z\}}$ for a fixed $z \in \mathcal{X}$ then $F(y_{-\infty}^0) = P(X_1 = z|X_{-\infty}^0 = y_{-\infty}^0)$. If $\mathcal{X}$ is a finite or countably infinite subset of the reals and $f(x) = x$ then $F(y_{-\infty}^0) = E(X_1|X_{-\infty}^0 = y_{-\infty}^0)$.

One denotes the nth estimate of $E(f(X_{\lambda_n+1})|X_0^{\lambda_n})$ from samples $X_0^{\lambda_n}$ by f_n, and defines it to be

$$f_n = \frac{1}{n} \sum_{j=0}^{n-1} f(X_{\lambda_j+1}). \tag{6}$$

3. Main results

Define the distance $d^*(\cdot, \cdot)$ on $\mathcal{X}^{*-}$ as follows. For $x_{-\infty}^0, y_{-\infty}^0 \in \mathcal{X}^{*-}$ let

$$d^*(x_{-\infty}^0, y_{-\infty}^0) = \sum_{i=0}^{\infty} 2^{-i-1} 1_{\{x_{-i} \ne y_{-i}\}}. \tag{7}$$

Definition 3. *We say that $F(X_{-\infty}^0)$ is almost surely continuous if for some set $C \subseteq \mathcal{X}^{*-}$ which has probability one the function $F(X_{-\infty}^0)$ restricted to this set C is continuous with respect to metric $d^*(\cdot, \cdot)$ (Cf. Morvai and Weiss [12]).*

The processes with almost surely continuous conditional expectation generalizes the processes for which it is actually continuous, cf. Kalikow [9] and Keane [10]. The stationary finitarily Markovian processes are included in the class of stationary processes with almost surely continuous $E(f(X_1)|X_{-\infty}^0)$ for arbitrary bounded $f(\cdot)$.

Note that Ryabko [14], and Györfi, Morvai, Yakowitz [7] showed that one cannot estimate $P(X_{n+1}=1|X_0^n)$ for all n in a pointwise consistent way even for the class of all stationary and ergodic binary finitarily Markovian time series.

The entropy rate H associated with a stationary finite or countably infinite alphabet time series $\{X_n\}$ is defined as $H = \lim_{n\to\infty} \frac{-1}{n+1}\sum_{x_{-n}^0\in\mathcal{X}^{n+1}} p_n(x_{-n}^0) \log_2 p_n(x_{-n}^0)$. We note that the entropy rate of a stationary finite alphabet time series is finite. For details cf. Cover, Thomas [3], pp. 63-64.

Fix positive real numbers $0<\beta,\gamma<1$ such that $2\beta+\gamma<1$, fix a sequence l_n that $1=l_1\le l_2,\dots,l_n\to\infty$ and fix a bounded function $f(\cdot):\mathcal{X}\to(-\infty,\infty)$ and with these numbers, sequence and function define ζ_n, χ_n, κ_n, λ_n and $F(\cdot)$ as described in the previous section. For the resulting f_n we have the following theorem:

Theorem. *Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite or countably infinite set $\mathcal{X}$. If the conditional expectation $F(X_{-\infty}^0)$ is almost surely continuous then almost surely,*

$$\lim_{n\to\infty} f_n = F(\tilde{X}_{-\infty}^0) \quad \textit{and} \quad \lim_{n\to\infty}\left| f_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n})\right| = 0.$$

The l_n may be chosen in such a fashion that whenever the stationary and ergodic time series $\{X_n\}$ has finite entropy rate then the λ_n grow no faster than a polynomial in n.

If the stationary and ergodic time series $\{X_n\}$ turns out to be finitarily Markovian then

$$\lim_{n\to\infty}\frac{\lambda_n}{n} = \frac{1}{p_{K(\tilde{X}_{-\infty}^0)-1}(\tilde{X}^0_{-K(\tilde{X}_{-\infty}^0)+1})} < \infty \ \textit{almost surely.}$$

Moreover, if the stationary and ergodic time series $\{X_n\}$ turns out to be independent and identically distributed then $\lambda_n=\lambda_{n-1}+1$ eventually almost surely.

Proof of the Theorem. **Step 1.** The time series $\{\tilde{X}_n\}_{n=-\infty}^0$ and $\{X_n\}_{n=-\infty}^0$ have identical distribution.

For all $k\ge 1$ and $1\le i\le k$ define (similarly to Morvai and Weiss [12]) $\hat{\zeta}_0^k=0$ and

$$\hat{\zeta}_i^k = \hat{\zeta}_{i-1}^k - \min\{t>0 : X_{\hat{\zeta}_{i-1}^k-(l_{k-i+1}-1)-t}^{\hat{\zeta}_{i-1}^k-t} = X_{\hat{\zeta}_{i-1}^k-(l_{k-i+1}-1)}^{\hat{\zeta}_{i-1}^k}\}.$$

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 $\hat{\zeta}_k^k(T^l x_{-\infty}^{\infty})=-l$.

Now the statement follows from stationarity and the fact that for $k\ge 0, n\ge 0$, $x_{-n}^0\in\mathcal{X}^{n+1}, l\ge 0$,

$$T^l\{X_{\zeta_k-n}^{\zeta_k}=x_{-n}^0, \zeta_k=l\} = \{X_{-n}^0=x_{-n}^0, \hat{\zeta}_k^k(X_{-\infty}^0)=-l\}. \tag{8}$$

Step 2. We show that $P(\chi_n = K(\tilde{X}^0_{-\infty})$ eventually $|K(\tilde{X}^0_{-\infty}) < \infty) = 1$ and $P(\lim_{n\to\infty} \chi_n = \infty | K(\tilde{X}^0_{-\infty}) = \infty) = 1$.

By Step 1, $\{\tilde{X}_n\}^0_{n=-\infty}$ is stationary and ergodic with the same distribution as $\{X_n\}^0_{n=-\infty}$. We may assume that the sample path $\tilde{X}^0_{-\infty}$ is such that all finite blocks that appear have positive probability. It is immediate that if $K(\tilde{X}^0_{-\infty}) < \infty$ then for all $k \ge K(\tilde{X}^0_{-\infty})$, $\Delta_k = 0$ and $\Delta_{K(\tilde{X}^0_{-\infty})-1} > 0$ (otherwise the length of the memory would be not greater than $K(\tilde{X}^0_{-\infty}) - 1$). If $K(\tilde{X}^0_{-\infty}) = \infty$ then $\Delta_k > 0$ for all k, (otherwise $K(\tilde{X}^0_{-\infty})$ would be finite). Thus by (3) if $K(\tilde{X}^0_{-\infty}) = \infty$ then $\chi_n \to \infty$ and if $K(\tilde{X}^0_{-\infty}) < \infty$ then $\chi_n \ge K(\tilde{X}^0_{-\infty})$ eventually almost surely. We have to show that $\chi_n \le K(\tilde{X}^0_{-\infty})$ eventually almost surely provided that $K(\tilde{X}^0_{-\infty}) < \infty$.

Fix now $k < n$. We will estimate the probability of the undesirable event as follows:

$$
\begin{aligned}
&P(\hat{\Delta}^n_k > n^{-\beta}, K(\tilde{X}^0_{-\infty}) = k | X_0^{\lceil \frac{n}{2} \rceil}) \\
&\le \sum_{i=1}^{n} P\Bigg(\max_{(z^{-k}_{-k-i+1}, \tilde{X}^0_{-k+1}, x) \in \mathcal{L}^n_{k+i}} 1_{\{\zeta_{J(k)} \le \lceil \frac{n}{2} \rceil - 1\}} \\
&\quad \times \Bigg| \frac{\#\{\lceil \frac{n}{2} \rceil + k \le t \le n : X^t_{t-k} = (\tilde{X}^0_{-k+1}, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k - 1 \le t \le n-1 : X^t_{t-k+1} = \tilde{X}^0_{-k+1}\}} \\
&\quad - \frac{\#\{\lceil \frac{n}{2} \rceil + k + i \le t \le n : X^t_{t-k-i} = (z^{-k}_{-k-i+1}, \tilde{X}^0_{-k+1}, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k + i - 1 \le t \le n-1 : X^t_{t-k-i+1} = (z^{-k}_{-k-i+1}, \tilde{X}^0_{-k+1})\}} \Bigg| \\
&\quad > n^{-\beta}, K(\tilde{X}^0_{-\infty}) = k | X_0^{\lceil \frac{n}{2} \rceil} \Bigg).
\end{aligned}
$$

Define $\mathcal{M}_{k-1}$ as the set of all $x^0_{-k+1} \in \mathcal{X}^k$ such that for all $i \ge 1$, $z \in \mathcal{X}$, and $y^{-k}_{-k-i+1} \in \mathcal{X}^i$, $p_{k+i}(y^{-k}_{-k-i+1}, x^0_{-k+1}, z) > 0$ implies that $p_{k-1}(z | x^0_{-k+1}) = p_{k+i-1}(z | y^{-k}_{-k-i+1}, x^0_{-k+1})$. By the definition of $\hat{\Delta}^n_k$ and since $K(\tilde{X}^0_{-\infty}) = k$ we have easily that

$$
\begin{aligned}
&P\Bigg(\max_{(z^{-k}_{-k-i+1}, \tilde{X}^0_{-k+1}, x) \in \mathcal{L}^n_{k+i}} 1_{\{\zeta_{J(k)} \le \lceil \frac{n}{2} \rceil - 1\}} \\
&\quad \times \Bigg| \frac{\#\{\lceil \frac{n}{2} \rceil + k \le t \le n : X^t_{t-k} = (\tilde{X}^0_{-k+1}, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k - 1 \le t \le n-1 : X^t_{t-k+1} = \tilde{X}^0_{-k+1}\}} \\
&\quad - \frac{\#\{\lceil \frac{n}{2} \rceil + k + i \le t \le n : X^t_{t-k-i} = (z^{-k}_{-k-i+1}, \tilde{X}^0_{-k+1}, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k + i - 1 \le t \le n-1 : X^t_{t-k-i+1} = (z^{-k}_{-k-i+1}, \tilde{X}^0_{-k+1})\}} \Bigg| \\
&\quad > n^{-\beta}, K(\tilde{X}^0_{-\infty}) = k | X_0^{\lceil \frac{n}{2} \rceil} \Bigg)
\end{aligned}
$$

$$\le P\Bigg(\max_{y_{-k+1}^0 \in \mathcal{M}_{k-1}, (z_{-k-i+1}^{-k}, y_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \left| \frac{\#\{\lceil \frac{n}{2} \rceil + k \le t \le n : X_{t-k}^t = (y_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k - 1 \le t \le n-1 : X_{t-k+1}^t = y_{-k+1}^0\}} - \frac{\#\{\lceil \frac{n}{2} \rceil + k + i \le t \le n : X_{t-k-i}^t = (z_{-k-i+1}^{-k}, y_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k + i - 1 \le t \le n-1 : X_{t-k-i+1}^t = (z_{-k-i+1}^{-k}, y_{-k+1}^0)\}} \right| > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil} \Bigg).$$

We can estimate this last probability as the sum of two terms:

$$P\Bigg(\max_{y_{-k+1}^0 \in \mathcal{M}_{k-1}, (z_{-k-i+1}^{-k}, y_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \left| \frac{\#\{\lceil \frac{n}{2} \rceil + k \le t \le n : X_{t-k}^t = (y_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k - 1 \le t \le n-1 : X_{t-k+1}^t = y_{-k+1}^0\}} - \frac{\#\{\lceil \frac{n}{2} \rceil + k + i \le t \le n : X_{t-k-i}^t = (z_{-k-i+1}^{-k}, y_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k + i - 1 \le t \le n-1 : X_{t-k-i+1}^t = (z_{-k-i+1}^{-k}, y_{-k+1}^0)\}} \right| > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil} \Bigg)$$

$$\le P\Bigg(\max_{y_{-k+1}^0 \in \mathcal{M}_{k-1}, (z_{-k-i+1}^{-k}, y_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \left| \frac{\#\{\lceil \frac{n}{2} \rceil + k \le t \le n : X_{t-k}^t = (y_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k - 1 \le t \le n-1 : X_{t-k+1}^t = y_{-k+1}^0\}} - p_{k-1}(x | y_{-k+1}^0) \right| > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil} \Bigg)$$

$$+ P\Bigg(\max_{y_{-k+1}^0 \in \mathcal{M}_{k-1}, (z_{-k-i+1}^{-k}, y_{-k+1}^0, x) \in \mathcal{L}_{k+i}^n} \left| p_{k-1}(x | y_{-k+1}^0) - \frac{\#\{\lceil \frac{n}{2} \rceil + k + i \le t \le n : X_{t-k-i}^t = (z_{-k-i+1}^{-k}, y_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2} \rceil + k + i - 1 \le t \le n-1 : X_{t-k-i+1}^t = (z_{-k-i+1}^{-k}, y_{-k+1}^0)\}} \right| > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil} \Bigg).$$

We overestimate these probabilities. For any $m \ge 0$ and x_{-m}^0 define $\sigma_i^m(x_{-m}^0)$ as the time of the i-th ocurrence of the string x_{-m}^0 in the data segment $X_{\lceil \frac{n}{2} \rceil}^n$, that is, let $\sigma_0^m(x_{-m}^0) = \lceil \frac{n}{2} \rceil + m - 1$ and for $i \ge 1$ define

$$\sigma_i^m(x_{-m}^0) = \min\{t > \sigma_{i-1}^m(x_{-m}^0) : X_{t-m}^t = x_{-m}^0\}.$$

Now

$$P\Bigg(\max_{y^0_{-k+1}\in\mathcal{M}_{k-1},(z^{-k}_{-k-i+1},y^0_{-k+1},x)\in\mathcal{L}^n_{k+i}}$$
$$\Bigg|\frac{\#\{\lceil\frac{n}{2}\rceil+k\le t\le n: X^t_{t-k}=(y^0_{-k+1},x)\}}{\#\{\lceil\frac{n}{2}\rceil+k-1\le t\le n-1: X^t_{t-k+1}=y^0_{-k+1}\}}$$
$$-\frac{\#\{\lceil\frac{n}{2}\rceil+k+i\le t\le n: X^t_{t-k-i}=(z^{-k}_{-k-i+1},y^0_{-k+1},x)\}}{\#\{\lceil\frac{n}{2}\rceil+k+i-1\le t\le n-1: X^t_{t-k-i+1}=(z^{-k}_{-k-i+1},y^0_{-k+1})\}}\Bigg|$$
$$>n^{-\beta}|X_0^{\lceil\frac{n}{2}\rceil}\Bigg)$$
$$\le P\Bigg(\max_{y^0_{-k+1}\in\mathcal{M}_{k-1},(y^0_{-k+1},x)\in\mathcal{L}^{(1)}_{n,k}}\ \sup_{j>n^{1-\gamma}}$$
$$\left|\frac{1}{j}\sum_{r=1}^{j}1_{\{X_{\sigma_r^{k-1}(y^0_{-k+1})}=x\}}-p_{k-1}(x|y^0_{-k+1})\right|>0.5n^{-\beta}|X_0^{\lceil\frac{n}{2}\rceil}\Bigg)$$
$$+P\Bigg(\max_{y^0_{-k+1}\in\mathcal{M}_{k-1},(z^{-k}_{-k-i+1},y^0_{-k+1},x)\in\mathcal{L}^{(1)}_{n,k+i}}\ \sup_{j>n^{1-\gamma}}$$
$$\left|\frac{1}{j}\sum_{r=1}^{j}1_{\{X_{\sigma_r^{k+i-1}(z^{-k}_{-k-i+1},y^0_{-k+1},x)}=x\}}-p_{k-1}(x|y^0_{-k+1})\right|>0.5n^{-\beta}|X_0^{\lceil\frac{n}{2}\rceil}\Bigg)$$

Since both $\mathcal{L}^{(1)}_{n,k}$ and $\mathcal{L}^{(1)}_{n,k+i}$ depend solely on $X_0^{\lceil\frac{n}{2}\rceil}$ we get

$$P\Bigg(\max_{y^0_{-k+1}\in\mathcal{M}_{k-1},(z^{-k}_{-k-i+1},y^0_{-k+1},x)\in\mathcal{L}^n_{k+i}}$$
$$\Bigg|\frac{\#\{\lceil\frac{n}{2}\rceil+k\le t\le n: X^t_{t-k}=(y^0_{-k+1},x)\}}{\#\{\lceil\frac{n}{2}\rceil+k-1\le t\le n-1: X^t_{t-k+1}=y^0_{-k+1}\}}$$
$$-\frac{\#\{\lceil\frac{n}{2}\rceil+k+i\le t\le n: X^t_{t-k-i}=(z^{-k}_{-k-i+1},y^0_{-k+1},x)\}}{\#\{\lceil\frac{n}{2}\rceil+k+i-1\le t\le n-1: X^t_{t-k-i+1}=(z^{-k}_{-k-i+1},y^0_{-k+1})\}}\Bigg|>n^{-\beta}|X_0^{\lceil\frac{n}{2}\rceil}\Bigg)$$
$$\le\sum_{y^0_{-k+1}\in\mathcal{M}_{k-1},(y^0_{-k+1},x)\in\mathcal{L}^{(1)}_{n,k}}\ \sum_{j=\lceil n^{1-\gamma}\rceil}^{\infty}P\Bigg(\left|\frac{1}{j}\sum_{r=1}^{j}1_{\{X_{\sigma_r^{k-1}(y^0_{-k+1})}=x\}}-p_{k-1}(x|y^0_{-k+1})\right|$$
$$>0.5n^{-\beta}|X_0^{\lceil\frac{n}{2}\rceil}\Bigg)$$
$$+\sum_{y^0_{-k+1}\in\mathcal{M}_{k-1},(z^{-k}_{-k-i+1},y^0_{-k+1},x)\in\mathcal{L}^{(1)}_{n,k+i}}\ \sum_{j=\lceil n^{1-\gamma}\rceil}^{\infty}P\Bigg(\Bigg|\frac{1}{j}\sum_{r=1}^{j}1_{\{X_{\sigma_r^{k+i-1}(z^{-k}_{-k-i+1},y^0_{-k+1})}=x\}}$$
$$-p_{k-1}(x|y^0_{-k+1})\Bigg|>0.5n^{-\beta}|X_0^{\lceil\frac{n}{2}\rceil}\Bigg).$$

Each of these represents the deviation of an empirical count from its mean. The variables in question are independent since whenever the block y^0_{-k+1} occurs the next term is chosen using the same distribution $p_{k-1}(x|y^0_{-k+1})$. Thus by Hoeffding's inequality (cf. Hoeffding [8] or Theorem 8.1 of Devroye et al. [6]) for sums of bounded independent random variables and since the cardinality of both $\mathcal{L}^{(1)}_{n,k}$ and $\mathcal{L}^{(1)}_{n,k+i}$ is not greater than $(n+2)/2$, we have

$$P\Bigg(\max_{y^0_{-k+1}\in\mathcal{M}_{k-1},(z^{-k}_{-k-i+1},y^0_{-k+1},x)\in\mathcal{L}^n_{k+i}} \Bigg|\frac{\#\{\lceil\frac{n}{2}\rceil+k\le t\le n : X^t_{t-k}=(y^0_{-k+1},x)\}}{\#\{\lceil\frac{n}{2}\rceil+k-1\le t\le n-1 : X^t_{t-k+1}=y^0_{-k+1}\}} - \frac{\#\{\lceil\frac{n}{2}\rceil+k+i\le t\le n : X^t_{t-k-i}=(z^{-k}_{-k-i+1},y^0_{-k+1},x)\}}{\#\{\lceil\frac{n}{2}\rceil+k+i-1\le t\le n-1 : X^t_{t-k-i+1}=(z^{-k}_{-k-i+1},y^0_{-k+1})\}}\Bigg| > n^{-\beta}|X_0^{\lceil\frac{n}{2}\rceil}\Bigg) \le 2\frac{n+2}{2}\sum_{j=\lceil n^{1-\gamma}\rceil}^{\infty} 2e^{-2n^{-2\beta}j}.$$

Thus

$$P(\hat{\Delta}^n_k > n^{-\beta}, K(\tilde{X}^0_{-\infty})=k|X_0^{\lceil\frac{n}{2}\rceil}) \le n(n+2)2e^{-2n^{-2\beta+1-\gamma}}.$$

Integrating both sides we get

$$P(\hat{\Delta}^n_k > n^{-\beta}, K(\tilde{X}^0_{-\infty})=k) \le n(n+2)2e^{-2n^{-2\beta+1-\gamma}}.$$

The right hand side is summable provided $2\beta+\gamma<1$ and the Borel-Cantelli Lemma yields that

$$P(\hat{\Delta}^n_k \le n^{-\beta} \text{ eventually}, K(\tilde{X}^0_{-\infty})=k) = P(K(\tilde{X}^0_{-\infty})=k).$$

Thus $\chi_n\le k$ eventually almost surely on $K(\tilde{X}^0_{-\infty})=k$.

Step 3. We show the first part of the Theorem.

Recalling (6) we can write

$$f_n = \frac{1}{n}\sum_{j=0}^{n-1}[f(X_{\lambda_j+1}) - E(f(X_{\lambda_j+1})|X^{\lambda_j}_{-\infty})] + \frac{1}{n}\sum_{j=0}^{n-1}E(f(X_{\lambda_j+1})|X^{\lambda_j}_{-\infty}) \quad (9)$$

Observe that the first term is an average of orthogonal bounded random variables and by Theorem 3.2.2 in Révész [13], it tends to zero.

Now we deal with the second term. If $K(\tilde{X}^0_{-\infty})<\infty$ then by Step 2, $\chi_n = K(\tilde{X}^0_{-\infty})$ eventually and by (1), (2), (4) and Step 1, eventually,

$$E(f(X_{\lambda_j+1})|X^{\lambda_j}_{-\infty}) = E(f(X_{\lambda_j+1})|X^{\lambda_j}_0) = F(\tilde{X}^0_{-\infty}).$$

We may deal with the case when $K(\tilde{X}^0_{-\infty}) = \infty$ and by Step 2, $\chi_n \to \infty$. For arbitrary $j \geq 0$, by (5) and (4) and the construction in (2),

$$X^{\lambda_j}_{\lambda_j-\kappa_j+1} = \tilde{X}^0_{-\kappa_j+1} \text{ and } \lim_{j\to\infty} d^*(\tilde{X}^0_{-\infty}, X^{\lambda_j}_{-\infty}) = 0 \text{ almost surely.} \tag{10}$$

Be Step 1, and the almost sure continuity of $F(\cdot)$, for some set $C \subseteq \mathcal{X}^{*-}$ with full measure, $F(\cdot)$ is continuous on C and

$$\tilde{X}^0_{-\infty} \in C, X^n_{-\infty} \in C \text{ for all } n \geq 0 \text{ almost surely.} \tag{11}$$

By the continuity of $F(\cdot)$ on the set C and (10), $E(f(X_{\lambda_j+1})|X^{\lambda_j}_{-\infty}) = F(X^{\lambda_j}_{-\infty}) \to F(\tilde{X}^0_{-\infty})$ and $f_n \to F(\tilde{X}^0_{-\infty})$ almost surely.

Define the random neighbourhood $\mathcal{N}_j(X_0^{\lambda_j})$ of $X_0^{\lambda_j}$ depending on the random data segment $X_0^{\lambda_j}$ itself as

$$\mathcal{N}_j(X_0^{\lambda_j}) = \{z^0_{-\infty} \in \mathcal{X}^{*-} \ : \ z_{-\kappa_j+1} = X_{\lambda_j-\kappa_j+1}, \ldots, z_0 = X_{\lambda_j}\}.$$

Note that by (1), (2), (5) and (4), $\tilde{X}^0_{-\infty} \in \mathcal{N}_j(X_0^{\lambda_j})$ and by (11) and the continuity of $F(\cdot)$ on the set C, and since $\kappa_j \to \infty$, by (10), almost surely,

$$\begin{aligned}
&\lim_{j\to\infty} \left| E(f(X_{\lambda_j+1})|X_0^{\lambda_j}) - F(\tilde{X}^0_{-\infty}) \right| = \lim_{j\to\infty} \left| E\{F(X^{\lambda_j}_{-\infty})|X_0^{\lambda_j}\} - F(\tilde{X}^0_{-\infty}) \right| \\
&\leq \lim_{j\to\infty} \sup_{y^0_{-\infty}, z^0_{-\infty} \in \mathcal{N}_j(X_0^{\lambda_j}) \bigcap C} |F(y^0_{-\infty}) - F(z^0_{-\infty})| = 0.
\end{aligned}$$

Step 4. We show the second part of the Theorem.

Now we assume that the stationary and ergodic finite or countably infinite alphabet time series $\{X_n\}$ possesses finite entropy rate H. (A stationary finite alphabet time series always has finite entropy rate.)

We will in fact obtain a more precise estimate, namely, if for some $0 < \epsilon_2 < \epsilon_1$, $\sum_{k=1}^{\infty}(k+1)2^{-l_k(\epsilon_1-\epsilon_2)} < \infty$ then

$$\lambda_n < 2^{l_n(H+\epsilon_1)} \text{ eventually almost surely.}$$

In particular, for arbitrary $\delta > 0$, $0 < \epsilon_2 < \epsilon_1$,
if $l_n = \min\left(n, \max\left(1, \lfloor \frac{2+\delta}{\epsilon_1-\epsilon_2} \log_2 n \rfloor\right)\right)$ then

$$\lambda_n < n^{\frac{2+\delta}{\epsilon_1-\epsilon_2}(H+\epsilon_1)}$$

eventually almost surely, and the upper bound is a polynomial.

Since $\lambda_n \leq \zeta_n$, it is enough to prove the result for ζ_n. Let $\mathcal{X}^*$ be the set of all two-sided sequences, that is,

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

Define $B_k \subseteq \mathcal{X}^{l_k}$ as $B_k = \{x^0_{-l_k+1} \in \mathcal{X}^{l_k} : 2^{-l_k(H+\epsilon_2)} < p_{l_k-1}(x^0_{-l_k+1})\}$. Note that there is a trivial bound on the cardinality of the set B_k, namely,

$$|B_k| \le 2^{l_k(H+\epsilon_2)}. \tag{12}$$

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

$$\Upsilon_k(y^0_{-l_k+1}) = \{z^\infty_{-\infty} \in \mathcal{X}^* : -\hat{\zeta}^k_k(z^0_{-\infty}) \ge 2^{l_k(H+\epsilon_1)}, z^0_{-l_k+1} = y^0_{-l_k+1})\}.$$

We will estimate the probability of $\Upsilon_k(y^0_{-l_k+1})$ by a frequency argument. Let $x^\infty_{-\infty} \in \mathcal{X}^*$ be a typical sequence of the time series $\{X_n\}$. Define $\rho_0(y^0_{-l_k+1}, x^\infty_{-\infty}) = 0$ and for $i \ge 1$ let

$$\rho_i(y^0_{-l_k+1}, x^\infty_{-\infty}) = \min\{l > \rho_{i-1}(y^0_{-l_k+1}, x^\infty_{-\infty}) : T^{-l}x^\infty_{-\infty} \in \Upsilon_k(y^0_{-l_k+1})\}.$$

Define also $\tau_0(y^0_{-l_k+1}, x^\infty_{-\infty}) = 0$ and for $i \ge 1$ let

$$\tau_i(y^0_{-l_k+1}, x^\infty_{-\infty}) = \min\{l \ge \tau_{i-1}(y^0_{-l_k+1}, x^\infty_{-\infty}) + 2^{l_k(H+\epsilon_1)} : T^{-l}x^\infty_{-\infty} \in \Upsilon_k(y^0_{-l_k+1})\}.$$

Notice that if $\tau_{i-1} = \rho_m$ then $\tau_i \le \rho_{m+k+1}$. (Indeed, since there are at least $k+1$ occurrences of the block $y^0_{-l_k+1}$ in the data segment $X^{\rho_m+1}_{-\rho_{m+k+1}-l_k+1}$ hence $2^{l_k(H+\epsilon_1)} \le -\hat{\zeta}^k_k(T^{-\rho_m}x^\infty_{-\infty}) \le \rho_{m+k+1} - \tau_{i-1}$.) By the ergodicity of the time series $\{X_n\}$,

$$\begin{aligned}
P(X^\infty_{-\infty} \in \Upsilon_k(y^0_{-l_k+1})) &= \lim_{t\to\infty} \frac{\#\{j \ge 1 : \rho_j(y^0_{-l_k+1}, x^\infty_{-\infty}) \le \tau_t(y^0_{-l_k+1}, x^\infty_{-\infty})\}}{\tau_t(y^0_{-l_k+1}, x^\infty_{-\infty})} \\
&= \lim_{t\to\infty} \frac{\sum_{l=1}^t \#\{j \ge 1 : \tau_{l-1}(y^0_{-l_k+1}, x^\infty_{-\infty}) < \rho_j(y^0_{-l_k+1}, x^\infty_{-\infty}) \le \tau_l(y^0_{-l_k+1}, x^\infty_{-\infty})\}}{\tau_t(y^0_{-l_k+1}, x^\infty_{-\infty})} \\
&\le \lim_{t\to\infty} \frac{t(k+1)}{t2^{l_k(H+\epsilon_1)}} = \frac{(k+1)}{2^{l_k(H+\epsilon_1)}}.
\end{aligned} \tag{13}$$

Since

$$T^l\{\zeta_k = l, X^{\zeta_k}_{\zeta_k-l_k+1} \in B_k\} = \{\hat{\zeta}^k_k = -l, X^0_{-l_k+1} \in B_k\}$$

by stationarity and the upper bound on the cardinality of the set B_k in (12) and by (13), we get

$$\begin{aligned}
P(\zeta_k \ge 2^{l_k(H+\epsilon_1)}, \tilde{X}^0_{-l_k+1} \in B_k) &= P(\zeta_k \ge 2^{l_k(H+\epsilon_1)}, X^{\zeta_k}_{\zeta_k-l_k+1} \in B_k) \\
&= P(-\hat{\zeta}^k_k \ge 2^{l_k(H+\epsilon_1)}, X^0_{-l_k+1} \in B_k) \\
&= \sum_{y^0_{-l_k+1} \in B_k} P(X^\infty_{-\infty} \in \Upsilon_k(y^0_{-l_k+1})) \\
&\le (k+1)2^{-l_k(\epsilon_1-\epsilon_2)}.
\end{aligned}$$

By assumption, the right hand side sums and the Borel-Cantelli Lemma yields that the event $\{\zeta_k \ge 2^{l_k(H+\epsilon_1)}, \tilde{X}^0_{-l_k+1} \in B_k\}$ cannot happen infinitely many times. By Step 1, 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. Chung [2]) $\tilde{X}^0_{-l_k+1} \in B_k$ eventually almost surely and so $\zeta_k \ge 2^{l_k(H+\epsilon_1)}$ cannot happen infinitely many times.

Step 5. We show the rest of the Theorem.

By Step 2, if $1 \le K(\tilde{X}^0_{-\infty}) < \infty$ then $\chi_n = K(\tilde{X}^0_{-\infty})$ eventually, and by ergodicity, $\frac{n}{\lambda_n} \to p_{K(\tilde{X}^0_{-\infty})-1}(\tilde{X}^0_{-K(\tilde{X}^0_{-\infty})+1}) > 0$. If $K(\tilde{X}^0_{-\infty}) = 0$ then by Step 2, $\chi_n = 0$ eventually, and by (4), $\lambda_n = \lambda_{n-1} + 1$ eventually. The proof of the Theorem is complete. □

References

1. Bailey, D.H.: Sequential schemes for classifying and predicting ergodic processes. Ph.D. thesis, Stanford University, 1976
2. Chung, K.L.: A note on the ergodic theorem of information theory. Ann. Math. Stat. **32**, 612–614 (1961)
3. Cover, T.M., Thomas, J.: Elements of information theory. Wiley, 1991
4. Csiszár, I., Shields, P.: The consistency of the BIC Markov order estimator. Ann. Stat. **28**, 1601–1619 (2000)
5. Csiszár, I.: Large-scale typicality of Markov sample paths and consistency of MDL order estimators. IEEE Trans. Inf. Theory **48**, 1616–1628 (2002)
6. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition. Springer-Verlag, New York, 1996
7. Györfi, L., Morvai, G., Yakowitz, S.: Limits to consistent on-line forecasting for ergodic time series. IEEE Trans. Inf. Theory **44**, 886–892 (1998)
8. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Association **58**, 13–30 (1963)
9. Kalikow, S.: Random Markov processes and uniform martingales. Israel J. Math. **71**, 33–54 (1990)
10. Keane, M.: Strongly mixing g-measures. Invent. Math. **16**, 309–324 (1972)
11. Morvai, G.: Guessing the output of a stationary binary time series. In: Foundations of Statistical Inference, Y. Haitovsky, H.R.Lerche, Y. Ritov (eds.), Physika-Verlag, 2003, pp. 207–215
12. Morvai, G., Weiss, B.: Forecasting for stationary binary time series. Acta Applicandae Mathematicae **79**, 25–34 (2003)
13. Révész, P.: The law of large numbers. Academic Press, 1968
14. Ryabko, Ya.B.: Prediction of random sequences and universal coding. Problems of Inform. Trans. **24**, 87–96 (1988)