A simple randomized algorithm for sequential prediction of ergodic time series

László Györfi
Department of Computer Science
and Information Theory
Technical University of Budapest
1521 Stoczek u. 2,
Budapest, Hungary
gyorfi@inf.bme.hu

Gábor Lugosi *
Department of Economics,
Pompeu Fabra University
Ramon Trias Fargas 25-27,
08005 Barcelona, Spain,
lugosi@upf.es

Gusztáv Morvai
Research Group for Informatics and Electronics
Hungarian Academy of Sciences
1521 Goldmann György tér 3,
Budapest, Hungary
morvai@inf.bme.hu

June 24, 1999

Abstract

We present a simple randomized procedure for the prediction of a binary sequence. The algorithm uses ideas from recent developments of the theory of the prediction of individual sequences. We show that if the sequence is a realization of a stationary and ergodic random process then the average number of mistakes converges, almost surely, to that of the optimum, given by the Bayes predictor. The desirable finite-sample properties of the predictor are illustrated by its performance for Markov processes. In such cases the predictor exhibits near optimal behavior even without knowing the order of the Markov process. Prediction with side information is also considered.

*The work of the second author was supported by DGES grant PB96-0300

1 Introduction

We address the problem of sequential prediction of a binary sequence. A sequence of bits $y_1, y_2, \ldots \in \{0,1\}$ is hidden from the predictor. At each time instant $i = 1, 2, \ldots$, the predictor is asked to guess the value of the next outcome y_i with knowledge of the past $y_1^{i-1} = (y_1, \ldots, y_{i-1})$ (where y_1^0 denotes the empty string). Thus, the predictor's decision, at time i, is based on the value of y_1^{i-1}. We also assume that the predictor has access to a sequence of independent, identically distributed (i.i.d.) random variables $U_1, U_2, \ldots$, uniformly distributed on $[0,1]$, so that the predictor can use U_i in forming a randomized decision for y_i. Formally, the strategy of the predictor is a sequence $g = \{g_i\}_{i=1}^\infty$ of decision functions

$$g_i : \{0,1\}^{i-1} \times [0,1] \to \{0,1\}$$

and the randomized prediction formed at time i is $g_i(y_1^{i-1}, U_i)$. The predictor pays a unit penalty each time a mistake is made. After n rounds of play, the *normalized cumulative loss* on the string y_1^n is

$$L_1^n(g, U_1^n) = \frac{1}{n} \sum_{i=1}^n I_{\{g_i(y_1^{i-1}, U_i) \neq y_i\}},$$

where I denotes the indicator function. When no confusion is caused, or when the predictor does not randomize, we will simply write $L_1^n(g) = L_1^n(g, U_1^n)$. In general, we denote the average number of mistakes between times m and n by

$$L_m^n(g, U_m^n) = \frac{1}{n-m+1} \sum_{i=m}^n I_{\{g_i(y_1^{i-1}, U_i) \neq y_i\}}.$$

We also write

$$\widehat{L}_1^n(g) = \mathbf{E} L_1^n(g, U_1^n) \quad \text{and} \quad \widehat{L}_m^n(g) = \mathbf{E} L_m^n(g, U_m^n)$$

for the expected loss of the randomized strategy g. (Here the expectation is taken with respect to the randomization U_1^n.)

In this paper we assume that $y_1, y_2, \ldots$ are realizations of the random variables $Y_1, Y_2, \ldots$ drawn from the binary-valued stationary and ergodic process $\{Y_n\}_{-\infty}^\infty$. We assume that the randomizing variables $U_1, U_2, \ldots$ are independent of the process $\{Y_n\}_{-\infty}^\infty$.

In this case there is a fundamental limit for the predictability of the sequence. This is stated in the next theorem whose proof may be found in Algoet [2].

Theorem 1 *(Algoet [2]) For any prediction strategy g and stationary ergodic process $\{Y_n\}_{-\infty}^\infty$,*

$$\liminf_{n\to\infty} L_1^n(g) \geq L^* \quad \textit{almost surely,}$$

where

$$L^* = \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0 | Y_{-\infty}^{-1}\}\right)\right]$$

is the minimal (Bayes) probability of error of any decision for the value of Y_0 based on the infinite past $Y_{-\infty}^{-1} = (\ldots, Y_{-3}, Y_{-2}, Y_{-1})$.

Based on Theorem 1, the following definition is meaningful:

Definition 1 *A prediction strategy g is called* universal *if for all stationary and ergodic processes $\{Y_n\}_{-\infty}^{\infty}$,*

$$\lim_{n\to\infty} L_1^n(g) = L^* \quad \textit{almost surely.}$$

Therefore, universal strategies asymptotically achieve the best possible loss for all ergodic processes. The first question is, of course, if such a strategy exists. The affirmative answer follows from a more general result of Algoet [2]. Here we give an alternative proof which is based on earlier results of Ornstein and Bailey.

Theorem 2 *(Algoet [2]) There exists a universal prediction scheme.*

Proof. Ornstein [19] proved that there exists a sequence of functions $f_i : \{0,1\}^i \to [0,1]$, $i = 1, 2, \ldots$ such that for all ergodic processes $\{Y_n\}_{-\infty}^{\infty}$,

$$\lim_{n\to\infty} f_n(Y_{-n}^{-1}) = \mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1}\} \quad \text{almost surely.} \tag{1}$$

Bailey [4] observed that for such estimators, for all ergodic processes

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} |(f_{i-1}(Y_1^{i-1}) - \mathbf{P}\{Y_i = 1|Y_{-\infty}^{i-1}\}| = 0 \quad \text{almost surely.} \tag{2}$$

Indeed, (1) and Breiman's generalized ergodic theorem (see Lemma 4 in the Appendix) yield (2).

Once such a sequence $\{f_i\}$ of estimators is available, we may define a (non-randomized) prediction scheme by the plug-in predictor

$$g_n(y_1^{n-1}) = \begin{cases} 1 & \text{if } f_{n-1}(y_1^{n-1}) \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

It is well-known that the probability of error of such a plug-in predictor may be bounded by the L_1 error of the estimator it is based on. In particular, by a simple inequality appearing in the proof of [9, Theorem 2.2],

$$\mathbf{P}\left\{g_n(Y_1^{n-1}) \neq Y_n | Y_{-\infty}^{n-1}\right\} - \mathbf{P}\left\{g^*(Y_{-\infty}^{n-1}) \neq Y_n|Y_{-\infty}^{n-1}\right\} \leq 2\left|f_{n-1}(Y_1^{n-1}) - \mathbf{P}\left\{Y_n = 1|Y_{-\infty}^{n-1}\right\}\right|.$$

Therefore,

$$\begin{aligned}
|L_1^n(g) - L^*| \quad \leq \quad & \left|L_1^n(g) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}\left\{g_i(Y_1^{i-1}) \neq Y_i|Y_1^{i-1}\right\}\right| \\
& + \frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{P}\left\{g_i(Y_1^{i-1}) \neq Y_i|Y_1^{i-1}\right\} - \mathbf{P}\left\{g^*(Y_{-\infty}^{i-1}) \neq Y_i|Y_{-\infty}^{i-1}\right\}\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{P}\left\{g^*(Y_{-\infty}^{i-1}) \neq Y_i|Y_{-\infty}^{i-1}\right\} - L^*\right|
\end{aligned}$$

$$\begin{aligned}
\leq \quad & \left|L_1^n(g) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}\left\{g_i(Y_1^{i-1}) \neq Y_i|Y_1^{i-1}\right\}\right| \\
& + \frac{2}{n}\sum_{i=1}^{n}\left|f_{i-1}(Y_1^{i-1}) - \mathbf{P}\left\{Y_i = 1|Y_{-\infty}^{i-1}\right\}\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{P}\left\{g^*(Y_{-\infty}^{i-1}) \neq Y_i|Y_{-\infty}^{i-1}\right\} - L^*\right|.
\end{aligned}$$

The first term of the right-hand side tends to zero almost surely by the Hoeffding-Azuma inequality (Lemma 5 in the Appendix) and the Borel-Cantelli lemma. The second one converges to zero almost surely by (2) and the third term tends to zero almost surely by the ergodic theorem. □

It was Ornstein [19] who first proved the existence of estimators satisfying (1). This was later generalized by Algoet [1]. A simpler estimator with the same convergence property was introduced by Morvai, Yakowitz, and Györfi [17]. Unfortunately, even the simpler estimator needs so large amounts of data that its practical use is unrealistic. By this we mean that even for "simple" i.i.d. or Markov processes the rate of convergence of the estimator is very slow. Motivated by the need of a practical estimator, Morvai, Yakowitz, and Algoet [18] introduced an even simpler algorithm. However, it is not known whether their estimator satisfies (1), and we do not even know whether the corresponding predictor is universal. The purpose of this paper is to introduce a new simple universal predictor whose finite-sample performance for Markov processes promise practical applicability.

2 A simple universal algorithm

In this section we present a simple prediction strategy, and prove its universality. It is motivated by some recent developments from the theory of the prediction of individual sequences (see, e.g., Vovk [22], Feder, Merhav, and Gutman [10], Littlestone and Warmuth [12], Cesa-Bianchi et al. [7]). These methods predict according to a combination of several predictors, the so-called *experts*.

The main idea in this paper is that if the sequence to predict is drawn from a stationary and ergodic process, combining the predictions of a small and simple set of appropriately chosen predictors (the so-called experts) suffices to achieve universality.

First we define an infinite sequence of experts $h^{(1)}, h^{(2)}, \ldots$ as follows: Fix a positive integer k, and for each $n \geq 1$, $s \in \{0,1\}^k$ and $y \in \{0,1\}$ define the function $\widehat{P}_n^k : \{0,1\} \times \{0,1\}^{n-1} \times \{0,1\}^k \to [0,1]$ by

$$\widehat{P}_n^k(y, y_1^{n-1}, s) = \frac{\left|\{k < i < n : y_{i-k}^{i-1} = s, y_i = y\}\right|}{\left|\{k < i < n : y_{i-k}^{i-1} = s\}\right|} \quad \text{for } n > k+1, \tag{3}$$

where $0/0$ is defined to be $1/2$. Also, for $n \leq k+1$ we define $\widehat{P}_n^k(y, y_1^{n-1}, s) = 1/2$. In other words, $\widehat{P}_n^k(y, y_1^{n-1}, s)$ is the proportion of the appearances of the bit y following the string s among all appearances of s in the sequence y_1^{n-1}.
The expert $h^{(k)}$ is a sequence of functions $h_n^{(k)} : \{0,1\}^{n-1} \to \{0,1\}$, $n = 1, 2, \ldots$ defined by

$$h_n^{(k)}(y_1^{n-1}) = \begin{cases} 0 & \text{if } \widehat{P}_n^k(0, y_1^{n-1}, y_{n-k}^{n-1}) > \frac{1}{2} \\ 1 & \text{otherwise,} \end{cases} \qquad n = 1, 2, \ldots.$$

That is, expert $h^{(k)}$ is a (nonrandomized) prediction strategy, which looks for all appearances of the last seen string y_{n-k}^{n-1} of length k in the past and predicts according to the larger of the relative frequencies of 0's and 1's following the string. We may call $h^{(k)}$ a *k-th order empirical Markov strategy.*

The proposed prediction algorithm proceeds as follows: Let $m = 0, 1, 2, \ldots$ be a nonnegative integer. For $2^m \leq n < 2^{m+1}$, the prediction is based upon a weighted majority of predictions of the experts $h^{(1)}, \ldots, h^{(2^{m+1})}$ as follows:

$$g_n(y_1^{n-1}, u) = \begin{cases} 0 & \text{if } u > \dfrac{\sum_{k=1}^{2^{m+1}} h_n^{(k)}(y_1^{n-1}) w_n(k)}{\sum_{k=1}^{2^{m+1}} w_n(k)} \\ 1 & \text{otherwise,} \end{cases} \qquad n = 1, 2, \ldots,$$

where $w_n(k)$ is the weight of expert $h^{(k)}$ defined by the past performance of $h^{(k)}$ as

$$w_{2^m}(k) = 1 \quad \text{and} \quad w_n(k) = e^{-\eta_m (n-2^m) L_{2^m}^{n-1}(h^{(k)})} \text{ for } 2^m < n < 2^{m+1},$$

where $\eta_m = \sqrt{8 \ln(2^{m+1})/2^m}$. Recall that

$$L_{2^m}^{n-1}(h^{(k)}) = \frac{1}{n-2^m} \sum_{i=2^m}^{n-1} I_{\left\{h_i^{(k)}(y_1^{i-1}) \neq y_i\right\}}$$

is the average number of mistakes made by expert $h^{(k)}$ between times 2^m and $n-1$. The weight of each expert is therefore exponentially decreasing with the number of its mistakes on this part of the data.

Remarks. 1. The above-mentioned estimator of Morvai, Yakowitz, and Algoet [18] selects a value of k in a certain data-dependent manner, and uses the corresponding estimate $\widehat{P}_n^k$. The new estimate, however, takes a mixture (weighted average) of all possible values of k, with exponential weights depending on the past performance of each component estimator. As Lemma 1 below suggests, this technique guarantees a number of errors almost as small as that of the *best* expert (i.e., best value of k).

2. Ryabko [21] proposed an estimator somewhat similar in spirit to the predictor defined here. Ryabko used a mixture of empirical Markov predictors, and proved its universality for all stationary and ergodic processes in a sense related to the Kullback-Leibler divergence. The idea of diversifying Markov strategies also appears in Algoet [1].

3. Each time n equals a power of two, all weights are reset to 1, and a simple majority vote is taken among the experts. This is necessary to make the algorithm sequential and to be able to incorporate more and more experts in the decision. If the total length of the sequence to be predicted was finite (say n) and known in advance, then no such resetting would be necessary, one could just use the first n experts as Lemma 1 below describes. However, to achieve universality, an infinite class of experts is necessary. As the first part of the proof of Theorem 3 below shows, we do not loose much by such a resetting of the weights.

4. Related prediction schemes have been proposed by Feder, Merhav, and Gutman [10] for individual sequences. Their computationally quite simple methods are shown to predict asymptotically as well as any finite-state predictor.

The main result of this section is the universality of this simple prediction scheme:

Theorem 3 *The prediction scheme g defined above is universal.*

In the proof we use a beautiful result of Cesa-Bianchi et al. [7]. It states that, given a set of K experts, and a sequence of fixed length n, there exists a randomized predictor whose number of mistakes is not more than that of the best predictor plus $\sqrt{(n/2)\ln K}$ for *all possible* sequences y_1^n. The simpler algorithm and statement cited below is due to Cesa-Bianchi [6]:

Lemma 1 *Let $\tilde{h}^{(1)}, \ldots, \tilde{h}^{(K)}$ be a finite collection of prediction strategies (experts), and let $\eta > 0$. Then if the prediction strategy $\tilde{g}$ is defined by*

$$\tilde{g}_i(y_1^{i-1}, u) = \begin{cases} 0 & \textit{if } u > \dfrac{\sum_{k=1}^{K} \mathbf{P}\left\{\tilde{h}^{(k)}(y_1^{i-1}, U_i) = 1\right\} \tilde{w}_i(k)}{\sum_{k=1}^{K} \tilde{w}_i(k)} \\ 1 & \textit{otherwise,} \end{cases}$$

$i = 1, 2, \ldots$, where for all $k = 1, \ldots, K$

$$\tilde{w}_1(k) = 1 \quad \textit{and} \quad \tilde{w}_i(k) = e^{-\eta(i-1)\widehat{L}_1^{i-1}(\tilde{h}^{(k)})}, \quad i > 1$$

then for every $n \geq 1$ and $y_1^n \in \{0,1\}^n$,

$$\widehat{L}_1^n(\tilde{g}) \leq \min_{k=1,\dots,K} \widehat{L}_1^n(\tilde{h}^{(k)}) + \frac{\ln K}{\eta n} + \frac{\eta}{8}.$$

In particular, if N is a positive integer, and $\eta = \sqrt{8N^{-1}\ln K}$, then

$$\widehat{L}_1^n(\tilde{g}) \leq \min_{k=1,\dots,K} \widehat{L}_1^n(\tilde{h}^{(k)}) + \frac{\sqrt{N}}{n}\sqrt{\frac{\ln K}{2}}, \quad n \leq N.$$

Proof of Theorem 3. Taking $K = 2^{m+1}$ and $N = 2^m$ in Lemma 1, we have that the expected number of errors committed by g on a segment $2^m, \dots, 2^{m+1}-1$ is bounded, for any $y_{2^m}^{2^{m+1}-1} \in \{0,1\}^{2^m}$, as

$$\begin{aligned}
\widehat{L}_{2^m}^{2^{m+1}-1}(g) &= \mathbf{E}\left[\frac{1}{2^m}\sum_{i=2^m}^{2^{m+1}-1} I_{\{g_i(y_1^{i-1},U_i)\neq y_i\}}\right] \\
&\leq \min_{k\leq 2^{m+1}} L_{2^m}^{2^{m+1}-1}(h^{(k)}) + \sqrt{\frac{\ln(2^{m+1})}{2\cdot 2^m}} \\
&= \min_{k=1,2,\dots} L_{2^m}^{2^{m+1}-1}(h^{(k)}) + \sqrt{\frac{\ln(2^{m+1})}{2\cdot 2^m}},
\end{aligned}$$

where the last equality follows from the fact that for all $i < 2^{m+1}$, all experts $h^{(k)}$ with $k \geq 2^{m+1}$ predict identically to $h^{(2^{m+1})}$. (Note that since the predictors $h^{(k)}$ are deterministic, for every m, $\widehat{L}_{2^m}^{2^{m+1}-1}(h^{(k)}) = L_{2^m}^{2^{m+1}-1}(h^{(k)})$.)

Similarly, denoting $\overline{n} = 2^{\lfloor \log_2 n\rfloor+1}$, and invoking Lemma 1 with $K = \overline{n}$ and $N = \overline{n}/2$,

$$L_{\overline{n}/2}^n(g) \leq \left(\min_{k=1,2,\dots} L_{\overline{n}/2}^n(h^{(k)}) + \frac{\sqrt{\overline{n}/2}}{n-\overline{n}/2+1}\sqrt{\frac{\ln(\overline{n})}{2}}\right).$$

Therefore, for any sequence $y_1, y_2, \dots$,

$$\begin{aligned}
n\widehat{L}_1^n(g) &= \sum_{m=0}^{\lfloor \log_2 n\rfloor -1} 2^m L_{2^m}^{2^{m+1}-1}(g) + (n-\overline{n}/2+1)L_{\overline{n}/2}^n(g) \\
&\leq \sum_{m=0}^{\lfloor \log_2 n\rfloor -1} 2^m \left(\min_{k=1,2,\dots} L_{2^m}^{2^{m+1}-1}(h^{(k)}) + \sqrt{\frac{\ln(2^{m+1})}{2\cdot 2^m}}\right) \\
&\quad +(n-\overline{n}/2+1)\left(\min_{k=1,2,\dots} L_{\overline{n}/2}^n(h^{(k)}) + \frac{\sqrt{\overline{n}/2}}{n-\overline{n}/2+1}\sqrt{\frac{\ln(\overline{n})}{2}}\right) \\
&\leq n\min_{k=1,2,\dots} L_1^n(h^{(k)}) + \sum_{m=0}^{\lfloor \log_2 n\rfloor -1}\sqrt{\frac{2^m\ln(2^{m+1})}{2}} + \sqrt{\frac{\overline{n}/2\ln\overline{n}}{2}} \\
&= n\min_{k=1,2,\dots} L_1^n(h^{(k)}) + \sum_{m=0}^{\lfloor \log_2 n\rfloor}\sqrt{\frac{2^m\ln(2^{m+1})}{2}}.
\end{aligned}$$

Denoting $\mu = \lfloor \log_2 n\rfloor$, we may write

$$\begin{aligned}
\sum_{m=0}^{\mu}\sqrt{\frac{2^m\ln(2^{m+1})}{2}} &\leq \sqrt{\frac{\ln(2^{\mu+1})}{2}}\sum_{m=0}^{\mu}2^{m/2} \\
&< \sqrt{\frac{\ln(2^{\mu+1})}{2}}\cdot\frac{2^{(\mu+1)/2}}{\sqrt{2}-1} \\
&= c\sqrt{\frac{\overline{n}\log_2\overline{n}}{2}},
\end{aligned}$$

where

$$c = \frac{\sqrt{\ln 2}}{\sqrt{2}-1} \approx 2.01.$$

Thus, we obtain

$$\begin{aligned}
\widehat{L}_1^n(g) &\leq \min_{k=1,2,\dots} L_1^n(h^{(k)}) + \frac{c}{n}\sqrt{\frac{\overline{n}\log_2\overline{n}}{2}} \\
&\leq \min_{k=1,2,\dots} L_1^n(h^{(k)}) + c\sqrt{\frac{\log_2 n+1}{n}}.
\end{aligned}$$

Noting that for any fixed sequence y_1^n, $L_1^n(g,U_1^n)$ is a sum of $[0,1]$-valued independent random variables whose expectation is $\widehat{L}_1^n(g)$, we may use Hoeffding's inequality [11] to see that for any sequence y_1^n, and $\epsilon > 0$,

$$\mathbf{P}\left\{\left|L_1^n(g,U_1^n) - \widehat{L}_1^n(g)\right| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}. \tag{4}$$

Therefore, if L is now evaluated on the random sequence $Y_1, Y_2, \dots$, we obtain

$$\begin{aligned}
\limsup_{n\to\infty} L_1^n(g,U_1^n) &\leq \limsup_{n\to\infty}\left(\min_{k=1,2,\dots} L_1^n(h^{(k)}) + c\sqrt{\frac{\log_2 n+1}{n}}\right). \\
&= \limsup_{n\to\infty}\min_{k=1,2,\dots} L_1^n(h^{(k)}) \quad \text{almost surely.}
\end{aligned}$$

Thus, it remains to show that for any ergodic process $Y_1, Y_2, \dots$,

$$\limsup_{n\to\infty}\min_{k=1,2,\dots} L_1^n(h^{(k)}) \leq L^* \quad \text{almost surely.} \tag{5}$$

This will follow easily from the following lemma:

Lemma 2 *For any $k \geq 1$,*

$$\limsup_{n\to\infty} L_1^n(h^{(k)}) \leq L^* + \epsilon_k \quad \textit{almost surely,}$$

where $\epsilon_k > 0$ is such that $\lim_{k\to\infty}\epsilon_k = 0$.

Remark. If the process $\{Y_n\}$ happens to be m-th order Markov, then it is easy to see that $\epsilon_k = 0$ for all $k \geq m$. The performance of the predictor for such processes is investigated in the next section.

Proof. Introduce

$$\tilde{L}_1^n(h^{(k)}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{P}\{Y_i \neq h_i^{(k)}(Y_1^{i-1})|Y_{-\infty}^{i-1}\}.$$

By Lemma 5 in the Appendix we immediately obtain

$$\lim_{n\to\infty} \left| L_1^n(h^{(k)}) - \tilde{L}_1^n(h^{(k)}) \right| = 0 \quad \text{almost surely.}$$

Therefore, it suffices to show that $\limsup_{n\to\infty} \tilde{L}_1^n(h^{(k)}) \leq L^* + \epsilon_k$ almost surely. To this end, first we study the asymptotic behavior of the quantity $\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\}$. Notice that

$$\begin{aligned}
\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \;\leq\; & I_A\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \\
& + I_{B_k}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \\
& + I_{C_k}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \\
& + I_{D_k^c}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \qquad (6)
\end{aligned}$$

where

$$A = \left\{\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\} = \frac{1}{2}\right\},$$

$$B_k = \left\{\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\} < \frac{1}{2} \text{ and } \mathbf{P}\{Y_0 = 1|Y_{-k}^{-1}\} < \frac{1}{2}\right\},$$

$$C_k = \left\{\mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\} < \frac{1}{2} \text{ and } \mathbf{P}(Y_0 = 0|Y_{-k}^{-1}) < \frac{1}{2}\right\},$$

and $D_k = A \cup B_k \cup C_k$. Notice that

$$\begin{aligned}
&\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \\
&= \mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\} I_{\{\hat{P}_n^k(1,Y_{-n+1}^{-1}|Y_{-k}^{-1}) \leq \frac{1}{2}\}} + \mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\} I_{\{\hat{P}_n^k(1,Y_{-n+1}^{-1}|Y_{-k}^{-1}) > \frac{1}{2}\}}. \qquad (7)
\end{aligned}$$

Now we examine the four terms on the right-hand side of (6). For the first term (7) yields

$$\begin{aligned}
I_A\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} &= I_A\frac{1}{2} \\
&= I_A \min\left(\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\}\right).
\end{aligned}$$

For the second term observe that under B_k, for sufficiently large n,

$$\hat{P}_n^k(1, Y_{-n+1}^{-1}, Y_{-k}^{-1}) < \frac{1}{2} \quad \text{almost surely,}$$

and therefore by (7) we have

$$\lim_{n\to\infty} I_{B_k}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}) = I_{B_k} \min\left(\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\}\right) \quad \text{a.s.}$$

For the third term we obtain similarly

$$\lim_{n\to\infty} I_{C_k}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}) = I_{C_k} \min\left(\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\}\right) \quad \text{a.s.}$$

The last term is simply bounded by

$$I_{D_k^c}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \leq I_{D_k^c}.$$

Combining all these bounds, we obtain

$$\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} \leq I_{D_k}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} + I_{D_k^c} \qquad (8)$$

and

$$\lim_{n\to\infty} I_{D_k}\mathbf{P}\{Y_0 \neq h_n^{(k)}(Y_{-n+1}^{-1})|Y_{-\infty}^{-1}\} = I_{D_k} \min\left(\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\}\right) \quad \text{a.s.} \quad (9)$$

From (8) it is immediate that

$$\limsup_{n\to\infty} \tilde{L}_1^n(h^{(k)}) \leq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I_{D_k}(T^iY_{-\infty}^{\infty})\mathbf{P}\{Y_i \neq h_i^{(k)}(Y_1^{i-1})|Y_{-\infty}^{i-1}\} + \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I_{D_k^c}(T^iY_{-\infty}^{\infty}),$$

where T denotes the left shift operator defined on doubly infinite binary sequences $y_{-\infty}^{\infty} \in \{0,1\}_{-\infty}^{\infty}$. By this inequality and (9), Breiman's generalized ergodic theorem (see Lemma 4 in the Appendix) implies

$$\begin{aligned}
\limsup_{n\to\infty} \tilde{L}_1^n(h^{(k)}) &\leq \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\}\right)\right] + \mathbf{P}\{D_k^c\} \\
&= L^* + \mathbf{P}\{D_k^c\} \quad \text{almost surely.}
\end{aligned}$$

Since by the martingale convergence theorem

$$\lim_{k\to\infty} \mathbf{P}\{Y_0 = 1|Y_{-k}^{-1}\} = \mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\} \quad \text{almost surely,}$$

we have

$$\lim_{k\to\infty} \mathbf{P}(D_k^c) = 0.$$

Taking $\epsilon_k = \mathbf{P}\{D_k^c\}$, the proof of the lemma is complete. □

Now we return to the proof of Theoren 3. By Lemma 2, for arbitrary K,

$$\begin{aligned}
\limsup_{n\to\infty} \min_{k=1,2,\ldots} L_1^n(h^{(k)}) &\leq \limsup_{n\to\infty} L_1^n(h^{(K)}) \\
&\leq L^* + \epsilon_K.
\end{aligned}$$

Since K is arbitrary and $\epsilon_K \to 0$, (5) is established, and the proof of the theorem is finished. □

Remarks. 1. The proposed estimate is clearly easy to compute. One merely has to keep track of the expected cumulative losses $L_{2^m}^{n-1}(h^{(k)})$ for $k = 1, 2, \ldots, n$. However, for large n, storing the entire data history may be problematic. In such cases, more efficient tree-based data structures, such as the ones described by Feder, Merhav, and Gutman [10], may be applied. We do not investigate this issue further here.

2. We see from the analysis that for *any* sequence $y_1, y_2, \ldots$ and for all n,

$$\widehat{L}_1^n(g) \le \min_{k=1,2,\ldots} L_1^n\left(h^{(k)}\right) + 2.01\sqrt{\frac{\log_2 n + 1}{n}},$$

and that the difference $|L_1^n(g, U_1^n) - \widehat{L}_1^n(g)|$ between the actual loss and the expected loss is $O_p(n^{-1/2})$. (For a sequence of random variables $\{X_n\}$ and sequence of nonnegative numbers $\{a_n\}$ we say that $X_n = O_p(a_n)$ if for every $\epsilon > 0$ there exists a constant $c > 0$ such that $\limsup_{n\to\infty} \mathbf{P}\{|X_n| \ge ca_n\} < \epsilon$.) The rate of convergence to L^* depends on the behavior of the best expert for the time segment up to n. For example, in the next section we show that for m-th order Markov processes the m-th expert predicts very well, and this fact will suffice to derive performance bounds for the proposed predictor.

3. The proposed predictor is by no means the only possibility. Different sets of experts may be combined in a similar fashion, and universality only depends on the behavior of the best expert. If some additional information is known (or suspected) about the process to be predicted, this information may be built in the definition of the experts. We chose the empirical Markov strategies as experts for convenience, and as we'll see it in the next section, this choice pays off whenever the process happens to be finite order Markov.

3 Markov processes

In this section we assume that the process to predict $\{Y_n\}_{-\infty}^{\infty}$ is (in addition to being stationary and ergodic) m-th order Markov, that is, for any binary sequence $y_{-\infty}^{-1} = (\ldots, y_{-2}, y_{-1})$,

$$\mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1} = y_{-\infty}^{-1}\} = \mathbf{P}\{Y_0 = 1 | Y_{-m}^{-1} = y_{-m}^{-1}\},$$

where m is a positive integer. We show that the proposed predictor achieves a nearly optimal performance for any m and for any such process, even though the predictor does not use the knowledge that the process is m-th order Markov. The intuitive reason for such a behavior is the following: we have seen it in the previous section that for *any* sequence,

$$L_1^n(g, U_1^n) \le \min_{k=1,2,\ldots} L_1^n\left(h^{(k)}\right) + 3\sqrt{\frac{\log_2 n + 1}{n}} + O_p\left(\sqrt{\frac{1}{n}}\right).$$

On the other hand, if the sequence is m-th order Markov, then there exists an expert, namely $h^{(m)}$ with very good performance.

In order to simplify our analysis, we modify the experts somewhat. They are defined as before but the probability estimates of (3) are now replaced by

$$\bar{P}_n^k(y, y_1^{n-1}, s) = \frac{\left|\{k < i < n : y_{i-k}^{i-1} = s, y_i = y\}\right| + 1}{\left|\{k < i < n : y_{i-k}^{i-1} = s\}\right| + 2}. \tag{10}$$

In other words, the simple empirical frequency counts are now replaced by the corresponding Laplace estimates. It is easy to see that all results of Section 2 remain valid for the modified predictor.

Remark. The reason for this modification is that this way we can appeal to a result of Rissanen [20] which simplifies our analysis. We believe that similar performance bounds are true for the original predictor of Section 2.

In the next theorem we compare the performance of our predictor to the universal lower bound L^*. The statement only gives information about the expected loss, but we believe this result already illustrates the good behavior of the proposed predictor for Markov processes.

Theorem 4 *If the process to be predicted is a stationary and ergodic m-th order Markov process, then the cumulative loss $L_1^n(g) = L_1^n(g, U_1^n)$ of the prediction strategy of Section 2 (with the modified estimates of (10)) satisfies*

$$\mathbf{E}L_1^n(g) \le L^* + 2\sqrt{\frac{2^{m-1}\log n}{n}} + 3\sqrt{\frac{\log_2 n + 1}{n}} + \sqrt{\frac{c}{n}},$$

where $c > 0$ is a universal constant.

Proof. First note that (4) implies

$$\mathbf{E}\left[|L_1^n(g,U_1^n)-\widehat{L}_1^n(g)||Y_{-\infty}^{\infty}\right]\le\int_0^\infty 2e^{-2n\epsilon^2}d\epsilon\le\sqrt{\frac{\ln(2e)}{2n}}$$

(see, e.g., [9, page 208]), and therefore it suffices to investigate $\widehat{L}_1^n(g)$. Recall also from the proof of Theorem 3 that for any input sequence,

$$\widehat{L}_1^n(g)\le\min_{k=1,2,\ldots}L_1^n(h^{(k)})+3\sqrt{\frac{\log_2 n+1}{n}},$$

and, in particular,

$$\widehat{L}_1^n(g)\le L_1^n(h^{(m)})+3\sqrt{\frac{\log_2 n+1}{n}}.$$

Thus, it suffices to show that for m-th order Markov processes the performance of the m-th expert $h^{(m)}$ satisfies

$$\mathbf{E}L_1^n(h^{(m)})\le L^*+2\sqrt{\frac{2^{m-1}\log n}{n}}+\sqrt{\frac{c}{n}}$$

for some constant c. To this end, observe that, on the one hand,

$$\begin{aligned}\mathbf{E}L_1^n(h^{(m)}) &= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n I_{\{h_i^{(m)}(Y_1^{i-1})\neq Y_i\}}\right]\\ &= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n\mathbf{P}\{h_i^{(m)}(Y_1^{i-1})\neq Y_i|Y_{-\infty}^{i-1}\}\right],\end{aligned}$$

and on the other hand, by the Markov property,

$$\begin{aligned}L^* &= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n\min\left(\mathbf{P}\left\{Y_i=1|Y_{i-m}^{i-1}\right\},\mathbf{P}\left\{Y_i=1|Y_{i-m}^{i-1}\right\}\right)\right]\\ &= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n\mathbf{P}\left\{h^{(m,*)}(Y_{i-m}^{i-1})\neq Y_i|Y_{-\infty}^{i-1}\right\}\right],\end{aligned}$$

where $h^{(m,*)}$ is the Bayes decision, given, for any $s\in\{0,1\}^m$, by

$$h^{(m,*)}(s)=\begin{cases}1 & \text{if } \mathbf{P}\{Y_0=1|Y_{-m}^{-1}=s\}\ge 1/2\\ 0 & \text{otherwise.}\end{cases}$$

(Note that the optimal predictor, that is, the one which minimizes the probability of error at every step predicts according to $h^{(m,*)}$.)

The above equalities imply that

$$\begin{aligned}\mathbf{E}L_1^n(h^{(m)})-L^* &\le \frac{1}{n}\sum_{i=1}^n\mathbf{E}\left|\mathbf{P}\left\{h_i^{(m)}(Y_1^{i-1})\neq Y_i|Y_{-\infty}^{i-1}\right\}-\mathbf{P}\left\{h^{(m,*)}(Y_{i-m}^{i-1})\neq Y_i|Y_{-\infty}^{i-1}\right\}\right|\\ &\le \frac{2}{n}\sum_{i=1}^n\mathbf{E}\left|\bar{P}_i^m(1,Y_1^{i-1},Y_{i-m}^{i-1})-\mathbf{P}\left\{Y_i=1|Y_1^{i-1}\right\}\right|,\end{aligned}$$

where the second inequality follows by [9, Theorem 2.2]. In the rest of the proof we simply apply some known results from the theory of universal prediction. First, by applications of Jensen's and Pinsker's inequalities (see Merhav and Feder [15, eq. (20)]) we obtain

$$\begin{aligned}&\frac{2}{n}\sum_{i=1}^n\mathbf{E}\left|\bar{P}_i^m(1,Y_1^{i-1},Y_{i-m}^{i-1})-\mathbf{P}\left\{Y_i=1|Y_1^{i-1}\right\}\right|\\ &\le 2\sqrt{\frac{1}{n}\sum_{i=1}^n\sum_{y_1^{i-1}\in\{0,1\}^{i-1}}\mathbf{P}\{Y_1^{i-1}=y_1^{i-1}\}\sum_{j=0}^1\mathbf{P}\{Y_i=j|Y_1^{i-1}=y_1^{i-1}\}\log\frac{\mathbf{P}\{Y_i=j|Y_1^{i-1}=y_1^{i-1}\}}{\bar{P}_i^m(j,y_1^{i-1},y_{i-m}^{i-1})}}.\end{aligned}$$

Observe that on the right-hand side, under the square root sign, we have the normalized Kullback-Leibler divergence between the probability measure of Y_1^n and its estimate constructed as a product of the Laplace estimates (10). But this divergence, for m-th order Markov sources, is well-known to be bounded by

$$\frac{2^m}{2n}\log n+O\left(\frac{1}{n}\right),$$

see Rissanen [20]. This concludes the proof. □

Remarks. 1. As Theorem 4 shows, by exponential weighting of the empirical Markov strategies, the predictor automatically adapts to the unknown Markov order. Similar results, though in different setup, are achieved by Modha and Masry [13],[14] by complexity regularization.

2. Merhav, Feder, and Gutman, [16] showed that if the process is m-th order Markov, then the randomized predictor $\tilde{h}^{(m)}$ defined by

$$\tilde{h}_i^{(m)}(y_1^{i-1},U)=\begin{cases}0 & \text{if } \widehat{P}_n^m(0,y_1^{n-1},y_{n-m}^{n-1})>\frac{1}{2}\\ 1 & \text{if } \widehat{P}_n^m(0,y_1^{n-1},y_{n-m}^{n-1})<\frac{1}{2}\\ I_{\{U\ge 1/2\}} & \text{otherwise}\end{cases}$$

achieves $\mathbf{E}L_1^n(\tilde{h}^{(m)})-L^*\le C/n$, where C is a constant depending of the distribution of the process. However, in an interesting contrast, the best *distribution-free* upper bound for all m-th order Markov processes is of the order of $n^{-1/2}$. To illustrate this, consider the case $m=0$, that is, when $\{Y_n\}$ is an i.i.d. process with $\mathbf{P}\{Y_1=1\}=1/2+\theta$, and the predictor $\tilde{h}^{(0)}$ is based on a majority vote of the bits appeared in the past. In this case, for every n,

$$\sup_{\theta\in[-1/2,1/2]}\left(\mathbf{E}L_1^n(\tilde{h}^{(0)})-L^*\right)\ge c_1 n^{-1/2},$$

where c_1 is a universal constant. (This is straightforward to see by considering $\theta=cn^{-1/2}$

for some small constant c, and writing

$$\begin{aligned}\mathbf{E}L_1^n(\tilde{h}^{(0)}) - L^* &= \frac{1}{n}\sum_{i=1}^n \left[\mathbf{P}\{\tilde{h}^{(0)}(Y_1^{i-1}, U_i) \neq Y_i\} - \left(\frac{1}{2} - \theta\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^n 2\theta \mathbf{P}\{\tilde{h}^{(0)}(Y_1^{i-1}, U_i) = 0\} \\ &\geq \frac{1}{n}\sum_{i=1}^n 2\theta \mathbf{P}\left\{\sum_{j=1}^{i-1} Y_j < \frac{i-1}{2}\right\} \\ &= \frac{1}{n}\sum_{i=1}^n 2\theta \mathbf{P}\left\{\sum_{j=1}^{i-1}(Y_j - \mathbf{E}Y_j) < -(i-1)\theta\right\}.\end{aligned}$$

Finally, invoke the Berry-Esséen theorem (see, e.g., [8]) to deduce that there exists a universal constant c_2 such that $\mathbf{P}\left\{\sum_{j=1}^{i-1}(Y_j - \mathbf{E}Y_j) < -(i-1)\theta\right\} \geq c_2$ for every $2 \leq i \leq n$.) Thus, even though for every single value of θ, $\mathbf{E}L_1^n(\tilde{h}^{(m)}) - L^*$ converges to zero at a rate of $O(1/n)$, the *minimax* rate of convergence is, in fact, $O(1/\sqrt{n})$. Since the upper bound in Theorem 4 is independent of the distribution, we see that, in this sense, (ignoring logarithmic factors) the order of magnitude of the bound is the best possible.

4 Prediction with side information

In this section we apply the same ideas to the seemingly more difficult classification (or pattern recognition) problem. The setup is the following: let $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ be a stationary and ergodic sequence of pairs taking values in $\mathcal{R}^d \times \{0,1\}$. The problem is to predict the value of Y_n given the data $(X_n, \mathcal{D}^{n-1})$, where we denote $\mathcal{D}^{n-1} = (X_1^{n-1}, Y_1^{n-1})$. The prediction problem is similar to the one studied in Section 2 with the exception that the sequence of X_i's is also available to the predictor. One may think about the X_i's as side information.

We may formalize the prediction problem as follows. A (randomized) prediction strategy is a sequence $g = \{g_i\}_{i=1}^{\infty}$ of decision functions

$$g_i : \{0,1\}^{i-1} \times \left(\mathcal{R}^d\right)^i \times [0,1] \to \{0,1\}$$

so that the prediction formed at time i is $g_i(y_1^{i-1}, x_1^i, U_i)$. The *normalized cumulative loss* for any fixed pair of sequences x_1^n, y_1^n is now

$$R_1^n(g, U_1^n) = \frac{1}{n}\sum_{i=1}^n I_{\{g_i(y_1^{i-1}, x_1^i, U_i) \neq y_i\}},$$

We also use the short notation $R_1^n(g) = R_1^n(g, U_1^n)$. Denote the expected loss of the randomized strategy g by

$$\widehat{R}_1^n(g) = \mathbf{E}R_1^n(g, U_1^n).$$

We assume that the randomizing variables $U_1, U_2, \ldots$ are independent of the process $\{(X_n, Y_n)\}$.

Just like in the case of prediction without side information, the fundamental limit is given by the Bayes probability of error:

Theorem 5 *For any prediction strategy g and stationary ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$,*

$$\liminf_{n\to\infty} R_1^n(g) \geq R^* \quad \textit{almost surely,}$$

where

$$R^* = \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1}, X_{-\infty}^0\}, \mathbf{P}\{Y_0 = 0 | Y_{-\infty}^{-1}, X_{-\infty}^0\}\right)\right].$$

The proof of this lower bound is similar to that of Theorem 1, the details are omitted. It follows from results of Morvai, Yakowitz, and Györfi [17] that there exists a prediction strategy g such that for all ergodic processes, $R_1^n(g) \to R^*$ almost surely. (We omit the details here.) The algorithm of Morvai, Yakowitz, and Györfi, however, has a very slow rate of convergence even for i.i.d. processes. The main message of this section is a simple universal procedure with a practical appeal. The idea, again, is to combine the decisions of a small number of simple experts in an appropriate way.

We define an infinite array of experts $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ as follows. Let $\mathcal{P}_\ell = \{A_{\ell,j}, j = 1, 2, \ldots, m_\ell\}$ be a sequence of finite partitions of the feature space $\mathcal{R}^d$, and let G_ℓ be the corresponding quantizer:

$$G_\ell(x) = j, \text{ if } x \in A_{\ell,j}.$$

With some abuse of notation, for any n and $x_1^n \in \left(\mathcal{R}^d\right)^n$, we write $G_\ell(x_1^n)$ for the sequence $G_\ell(x_1), \ldots, G_\ell(x_n)$. Fix positive integers k, ℓ, and for each $s \in \{0,1\}^k$, $z \in \{1, 2, \ldots, m_\ell\}^{k+1}$, and $y \in \{0,1\}$ define

$$\hat{P}_n^{(k,\ell)}(y, y_1^{n-1}, x_1^n, s, z) = \frac{\left|\{k < i < n : y_{i-k}^{i-1} = s, G_\ell(x_{i-k}^i) = z, y_i = y\}\right|}{\left|\{k < i < n : y_{i-k}^{i-1} = s, G_\ell(x_{i-k}^i) = z, \}\right|}, \quad n > k+1. \quad (11)$$

0/0 is defined to be 1/2. Also, for $n \le k+1$ we define $\hat{P}_n^{(k,\ell)}(y, y_1^{n-1}, x_1^n, s, z) = 1/2$.

The expert $h^{(k,\ell)}$ is now defined by

$$h_n^{(k,\ell)}(y_1^{n-1}, x_1^n) = \begin{cases} 0 & \text{if } \hat{P}_n^{(k,\ell)}(0, y_1^{n-1}, x_1^n, y_{n-k}^{n-1}, G_\ell(x_{n-k}^n)) < \frac{1}{2} \\ 1 & \text{otherwise,} \end{cases} \qquad n = 1, 2, \ldots$$

That is, expert $h^{(k,\ell)}$ quantizes the sequence x_1^n according to the partition $\mathcal{P}_\ell$, and looks for all appearances of the last seen quantized strings $y_{n-k}^{n-1}, G_\ell(x_{n-k}^n)$ of length k in the past. Then it predicts according to the larger of the relative frequencies of 0's and 1's following the string.

The proposed algorithm combines the predictions of these experts similarly to that of Section 2. This way both the length of the string to be matched and the resolution of the quantizer are adjusted depending on the data. The formal definition is as follows: For any $m = 0, 1, 2, \ldots$, if $2^m \le n < 2^{m+1}$, the prediction is based upon a weighted majority of predictions of the $(2^{m+1})^2$ experts $h^{(k,\ell)}$, $k, l \le 2^{m+1}$ as follows:

$$g_n(y_1^{n-1}, x_1^n, u) = \begin{cases} 0 & \text{if } u > \dfrac{\sum_{k,\ell \le 2^{m+1}} h_n^{(k,\ell)}(y_1^{n-1}, x_1^n) w_n(k,\ell)}{\sum_{k,\ell \le 2^{m+1}} w_n(k,\ell)} \\ 1 & \text{otherwise,} \end{cases}$$

where $w_n(k,\ell)$ is the weight of expert $h^{(k,\ell)}$ defined by the past performance of $h^{(k,\ell)}$ as

$$w_{2^m}(k,\ell) = 1 \quad \text{and} \quad w_n(k,\ell) = e^{-\eta_m (n-2^m) R_{2^m}^{n-1}(h^{(k,\ell)})} \text{ for } 2^m < n < 2^{m+1},$$

where $\eta_m = \sqrt{8 \ln (2^{m+1})^2 / 2^m}$.

To prove the universality of the method, we need some natural conditions on the sequence of partitions. We assume the following:

(a) the sequence of partitions is nested, that is, any cell of $\mathcal{P}_{\ell+1}$ is a subset of a cell of $\mathcal{P}_\ell$, $\ell = 1, 2, \ldots$;
(b) each partition $\mathcal{P}_\ell$ is finite;
(c) if $\operatorname{diam}(A) = \sup_{x,y \in A} \|x - y\|$ denotes the diameter of a set, then for each sphere S centered at the origin

$$\lim_{\ell \to \infty} \max_{j : A_{\ell,j} \cap S \ne \emptyset} \operatorname{diam}(A_{\ell,j}) = 0.$$

Remark. The next theorem states the universality of the proposed pattern recognition scheme. The definition of the algorithm is somewhat arbitrary, we just chose one of the

many possibilities. In this version, at time n, only partitions with indices at most n are taken into account. It is easy to see that the universality property remains valid if the number of partitions considered at time n is an arbitrary, polynomially increasing function of n. The conditions for the sequence of partitions again give a lot of liberty to the user. In applications, the partitions may be chosen to incorporate some prior knowledge about the process. In this paper we merely prove universality of the scheme. Performance bounds in the style of Section 3 for special types of proceses may be derived, thanks to the powerful individual sequence bounds. Here, however, the analysis may be substantially more complicated.

Theorem 6 *Assume that the sequence of partitions $\mathcal{P}_\ell$ satisfies the three conditions above. Then the pattern recognition scheme g defined above satisfies*

$$\lim_{n \to \infty} R_1^n(g) = R^* \quad \textit{almost surely}$$

for any stationary and ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$.

Proof. As in the proof of Theorem 3, we obtain that for any stationary and ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\begin{aligned} \limsup_{n \to \infty} R_1^n(g, U_1^n) &\le \limsup_{n \to \infty} \left(\min_{\substack{k = 1,2,\ldots \\ \ell = 1,2,\ldots,n-1}} R_1^n(h^{(k,\ell)}) + 2c\sqrt{\frac{\log_2 n + 1}{n}} \right) \\ &= \limsup_{n \to \infty} \min_{\substack{k = 1,2,\ldots \\ \ell = 1,2,\ldots,n-1}} R_1^n(h^{(k,\ell)}) \quad \text{almost surely.} \end{aligned}$$

Thus, it remains to show that

$$\limsup_{n \to \infty} \min_{\substack{k = 1,2,\ldots \\ \ell = 1,2,\ldots,n-1}} R_1^n(h^{(k,\ell)}) \le R^* \quad \text{almost surely.}$$

To prove this, we use the following lemma, whose proof is easily obtained by copying that of Lemma 2:

Lemma 3 *For each $k, \ell \ge 1$, there exists a positive number $\epsilon_{k,\ell}$ such that for any fixed ℓ, $\lim_{k \to \infty} \epsilon_{k,\ell} = 0$ and*

$$\limsup_{n \to \infty} R_1^n(h^{(k,\ell)}) \le R_{(\ell)}^* + \epsilon_{k,\ell},$$

where

$$R_{(\ell)}^* = \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1}, G_\ell(X_{-\infty}^0)\}, \mathbf{P}\{Y_0 = 0 | Y_{-\infty}^{-1}, G_\ell(X_{-\infty}^0)\}\right)\right].$$

Now we return to the proof of Theorem 6. Since the sequence of partitions $\mathcal{P}_\ell$ is nested, and by (c), the sequences

$$\mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1}, G_\ell(X_{-\infty}^0)\} \quad \text{and} \quad \mathbf{P}\{Y_0 = 0 | Y_{-\infty}^{-1}, G_\ell(X_{-\infty}^0)\} \quad l = 1, 2, \ldots$$

are martingales and they converge almost surely to

$$\mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1}, X_{-\infty}^{0}\} \quad \text{and} \quad \mathbf{P}\{Y_0 = 0 | Y_{-\infty}^{-1}, X_{-\infty}^{0}\}.$$

Thus, it follows from Lebesgue's dominated convergence theorem that

$$\lim_{l \to \infty} R^*_{(\ell)} = \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1 | Y_{-\infty}^{-1}, X_{-\infty}^{0}\}, \mathbf{P}\{Y_0 = 0 | Y_{-\infty}^{-1}, X_{-\infty}^{0}\}\right)\right] = R^*.$$

Now it follows easily that

$$\limsup_{n \to \infty} \min_{\substack{k = 1,2,\ldots \\ \ell = 1,2,\ldots,n-1}} R_1^n(h^{(k,\ell)}) \le R^* \quad \text{almost surely,}$$

and the proof of the theorem is finished. □

5 Appendix

Here we describe two results which are used in the analysis. The first is due to Breiman [5], and its proof may also be found in Algoet [2].

Lemma 4 Breiman's generalized ergodic theorem [5]. *Let $Z = \{Z_i\}_{-\infty}^{\infty}$ be a stationary and ergodic time series. Let T denote the left shift operator. Let f_i be a sequence of real-valued functions such that for some function f, $f_i(Z) \to f(Z)$ almost surely. Assume that $\mathbf{E}\sup_i |f_i(Z)| < \infty$. Then*

$$\lim_{t \to \infty} \frac{1}{n} \sum_{i=1}^{n} f_i(T^i Z) = \mathbf{E} f(Z)$$

almost surely.

The second is the Hoeffding-Azuma inequality for sums of bounded martingale differences:

Lemma 5 Hoeffding [11], Azuma [3]. *Let $X_1, X_2, \ldots$ be a sequence of random variables, and assume that $V_1, V_2, \ldots$ is a martingale difference sequence with respect to $X_1, X_2, \ldots$. Assume furthermore that there exist random variables $Z_1, Z_2, \ldots$ and nonnegative constants $c_1, c_2, \ldots$ such that for every $i > 0$ Z_i is a function of $X_1, \ldots, X_{i-1}$, and*

$$Z_i \le V_i \le Z_i + c_i \quad \textit{with probability one.}$$

Then for any $\epsilon > 0$ and n

$$\mathbf{P}\left\{\sum_{i=1}^{n} V_i \ge \epsilon\right\} \le e^{-2\epsilon^2 / \sum_{i=1}^{n} c_i^2}$$

and

$$\mathbf{P}\left\{\sum_{i=1}^{n} V_i \le -\epsilon\right\} \le e^{-2\epsilon^2 / \sum_{i=1}^{n} c_i^2}.$$

Acknowledgement. We thank Nicoló Cesa-Bianchi for teaching us all wee needed to know about prediction with expert advise. We are also grateful to Sid Yakowitz for illuminating discussions and to the referees for a very careful reading of the manuscript and for valuable suggestions. We also thank Márta Horváth for useful conversations.

References

[1] P. Algoet. Universal schemes for prediction, gambling, and portfolio selection. *Annals of Probability*, 20:901–941, 1992.

[2] P. Algoet. The strong law of large numbers for sequential decisions under uncertainity. *IEEE Transactions on Information Theory*, 40:609–634, 1994.

[3] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[4] D.H. Bailey. *Sequential schemes for classifying and predicting ergodic processes.* PhD thesis, Stanford University, 1976.

[5] L. Breiman. The individual ergodic theorem of information theory. *Annals of Mathematical Statistics*, 28:809–811, 1957. Correction. *Annals of Mathematical Statistics*, 31:809–810, 1960.

[6] N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 163–170. ACM Press, 1997.

[7] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[8] Y.S. Chow and H. Teicher. *Probability Theory, Independence, Interchangeability, Martingales* (2nd edition). Springer-Verlag, New York, 1988.

[9] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[10] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

[11] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[12] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[13] D.S. Modha and E. Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42:2133–2145, 1996.

[14] D.S. Modha and E. Masry. Memory-universal prediction of stationary random processes. *IEEE Transactions on Information Theory*, 44:117–133, 1998.

[15] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[16] N. Merhav, M. Feder, and M. Gutman. Some properties of sequential predictors for binary Markov sources. *IEEE Transactions on Information Theory*, 39:887–892, 1993.

[17] G. Morvai, S. Yakowitz, and L. Györfi. Nonparametric inference for ergodic, stationary time series. *Annals of Statistics*, 24:370–379, 1996.

[18] G. Morvai, S. Yakowitz, and P. Algoet. Weakly Convergent Stationary Time Series. *IEEE Transactions on Information Theory*, 43:483–498, 1997.

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

[20] J. Rissanen. Complexity of strings in the class of Markov sources. *IEEE Transactions on Information Theory*, 32:526–532, 1986.

[21] B.Ya. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24:87–96, 1988.

[22] V.G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 372–383. Association of Computing Machinery, New York, 1990.