A simple randomized algorithm for consistent 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

April 22, 1998

Classification: C13.

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

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.

1 Introduction

We address the problem of sequential prediction of a binary sequence. A sequence of bits $y_0, y_1, y_2, \ldots \in \{0,1\}$ is hidden from the predictor. At each time instant $i = 1, 2, \ldots$, the bit y_{i-1} is revealed and the predictor is asked to guess the value of next outcome y_i. Thus, the predictor's decision, at time i, is based on the value of $y_1^{i-1} = (y_1, \ldots, y_{i-1})$. We also assume that the predictor has access to a sequence of 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, we will simply write $L_n(g) = L_1^n(g, U_1^n)$. In general, we denote

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

We also write

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

for the expected loss of the randomized strategy g.

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 ergodic process $\{Y_n\}_{-\infty}^{\infty}$ (which is independent of the randomizing variables $U_1, U_2, \ldots$).

In this case there is a fundamental limit for the predictability of the sequence. This is stated in the next lemma:

Theorem 1 *For any prediction strategy g and stationary ergodic process $\{Y_n\}_{-\infty}^{\infty}$,*

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

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 $Y_{-\infty}^{-1} = (\ldots, Y_{-3}, Y_{-2}, Y_{-1})$.

Proof. An easy application of the Azuma-Hoeffding inequality for sums of bounded martingale differences (Hoeffding [8], Azuma [2]) shows that for any prediction strategy g and $\epsilon > 0$,

$$\mathbf{P}\left\{\left|L_n(g) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}\{g_i(Y_1^{i-1}, U_i) \neq Y_i | Y_{-\infty}^{i-1}\}\right| > \epsilon \middle| U_1, \ldots, U_n\right\} \leq 2e^{-2n\epsilon^2}.$$

In particular,

$$\lim_{n\to\infty}\left(L_n(g) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}\{g_i(Y_1^{i-1}, U_i) \neq Y_i | Y_{-\infty}^{i-1}\}\right) = 0 \quad \text{almost surely.}$$

It is well known (see, e.g., [6]) that for any predictor g_i,

$$\begin{aligned}\mathbf{P}\{g_i(Y_1^{i-1}, U_i) \neq Y_i | Y_{-\infty}^{i-1}\} &\geq \mathbf{P}\{g_i^*(Y_{-\infty}^{i-1}) \neq Y_i | Y_{-\infty}^{i-1}\} \\ &= \min\left(\mathbf{P}\{Y_i = 1 | Y_{-\infty}^{i-1}\}, \mathbf{P}\{Y_i = 0 | Y_{-\infty}^{i-1}\}\right),\end{aligned}$$

where g_i^* is the optimal (Bayes) decision for Y_i based on $Y_{-\infty}^{i-1}$ given by

$$g_i^*(y_{-\infty}^{i-1}) = \begin{cases} 1 & \text{if } \mathbf{P}\{Y_i = 1 | Y_{-\infty}^{i-1} = y_{-\infty}^{i-1}\} \geq 1/2 \\ 0 & \text{otherwise.}\end{cases}$$

Note that by stationarity, $g_1^* = g_2^* = \cdots \stackrel{\text{def}}{=} g^*$. Therefore,

$$\liminf_{n\to\infty} L_n(g) \geq \liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\min\left(\mathbf{P}\{Y_i = 1 | Y_{-\infty}^{i-1}\}, \mathbf{P}\{Y_i = 0 | Y_{-\infty}^{i-1}\}\right) \quad \text{almost surely.}$$

Finally, we note that by the ergodic theorem (see, e.g., Stout [13]) the average on the right-hand side converges almost surely to L^*, so the proof is finished. □

Based on Theorem 1 the following definition is meaningful.

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

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

Therefore, consistent 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 may be easily deduced from earlier results of Ornstein and Bailey as follows:

Theorem 2 *There exists a consistent prediction scheme.*

Proof. Ornstein [12] 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 $\{Y_n\}_{-\infty}^{\infty}$,

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

(A simpler estimator with the same convergence property was introduced by Morvai, Yakowitz, and Györfi [11].) Bailey [3] showed that for such estimators, for all ergodic processes

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

Indeed, (1) and Breiman's generalized ergodic theorem (see, e.g., Algoet [1]) yield (2).

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

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

By [6, 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

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

The first term of the right hand side tends to zero almost surely by the Hoeffding-Azuma inequality [8], [2] by a similar argument that was used in the proof of Theorem 1. The second one converges to zero almost surely by (2) and the third term tends to zero almost surely by the ergodic theorem. □

Unfortunately, all known estimators satisfying (1) are either very complicated or need so large amounts of data that their practical use is unrealistic. Therefore, designing a simple direct algorithm is called for.

2 A simple consistent algorithm

In this section we present a simple prediction strategy, and prove its consistency. It is motivated by some recent developments from the theory of the prediction of individual sequences (see, e.g., Vovk [14], Feder, Merhav, and Gutman [7], Littlestone and Warmuth [9], Cesa-Bianchi et al. [5]). 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 consistency.

First we define an infinite sequence of experts $h^{(1)}, h^{(2)}, \ldots$ as follows: Fix a positive integer k, and for each $s \in \{0,1\}^k$ and $y \in \{0,1\}$ define

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

$0/0$ is defined to be $1/2$. Also, for $n \le k+1$ 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}.
Also introduce the function

$$F_k(z) = I_{\{z \in [0.5-1/k, 0.5+1/k]\}} \frac{z - 0.5 + 1/k}{2/k} + I_{\{z > 0.5+1/k\}}.$$

The expert $h^{(k)}$ is a sequence of functions $h_n^{(k)} : \{0,1\}^{n-1} \times [0,1] \to \{0,1\}$, $n = 1, 2, \ldots$ defined by

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

That is, expert $h^{(k)}$ 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. The function F_k only plays a role if these frequencies are close to $1/2$. In such a case a randomized prediction is made. (Note that $F_k(z)$ is continuous and it differs from $I_{\{z \ge 1/2\}}$ only if $|z - 1/2| < 1/k$.)

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

$$g_n(y_1^{n-1}, u) = \begin{cases} 0 & \text{if } u < \dfrac{\sum_{k=1}^{2^{m+1}} F_k(\widehat{P}_n^k(0, y_1^{n-1}|y_{n-k}^{n-1}))w_n(k)}{\sum_{k=1}^{2^{m+1}} w_n(k)} \\ 1 & \text{otherwise,} \end{cases} \qquad n = 1, 2, \ldots,$$

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

$$w_{2^m}(k) = 1 \quad \text{and} \quad w_n(k) = e^{-\eta_m \widehat{L}_{2^m}^{n-1}(h^{(k)})} \quad n > 2^m,$$

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

$$\widehat{L}_{2^m}^{n-1}(h^{(k)}) = \frac{1}{n-2^m}\sum_{i=2^m}^{n-1}\mathbf{P}\left\{h_i^{(k)}(y_1^{i-1},U_i)\neq y_i\right\}$$

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.

Our main result is the consistency of this simple prediction scheme:

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

In the proof we use a beautiful result of Cesa-Bianchi et al. [5]. It states that, given a set of N 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 about $\sqrt{(n/2)\ln N}$ for *all possible* sequences y_1^n. The simpler algorithm and statement cited below is due to Cesa-Bianchi [4]:

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

$$\tilde{g}_t(y_1^{t-1},u) = \begin{cases} 0 & \text{if } u < \dfrac{\sum_{k=1}^N \mathbf{P}\left\{\tilde{h}^{(k)}(y_1^{t-1},U)=0\right\}\tilde{w}_t(k)}{\sum_{k=1}^N \tilde{w}_t(k)} \\ 1 & \text{otherwise,}\end{cases}$$

$t=1,2,\ldots,n$, where for all $k=1,\ldots,N$

$$\tilde{w}_1(k)=1 \quad \text{and} \quad \tilde{w}_t(k) = e^{-\eta\widehat{L}_1^{t-1}(\tilde{h}^{(k)})}, \quad t>1$$

with $\eta=\sqrt{8\ln N/n}$, then for every $y_1^n\in\{0,1\}^n$,

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

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

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

where the last equality follows from the fact that since $n < 2^{m+1}$, all experts $h^{(k)}$ with $k \geq 2^{m+1}$ predict zero with probability $1/2$ up to time n (and therefore they are identical to $h^{(2^{m+1})}$).

Therefore, denoting $\overline{n} = 2^{\lfloor \log_2 n \rfloor + 1}$, for any sequence $y_1, y_2, \ldots$,

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

therefore

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

where

$$c = \frac{\sqrt{2\ln 2}}{\sqrt{2}-1} \approx 2.84.$$

It follows from McDiarmid's inequality (McDiarmid [10]; see also [6, Theorem 9.2]) that for any sequence y_1^n,

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

Therefore, if L and $\widehat{L}$ are now evaluated on the random sequence $Y_1, Y_2, \ldots$, we obtain

$$\begin{aligned} \limsup_{n\to\infty} L_1^n(g, U_1^n) &\leq \limsup_{n\to\infty} \left(\min_{k=1,2,\ldots} \widehat{L}_1^n(h^{(k)}) + c\sqrt{\frac{\log_2 n + 1}{n}} \right). \\ &= \limsup_{n\to\infty} \min_{k=1,2,\ldots} \widehat{L}_1^n(h^{(k)}) \quad \text{almost surely.} \end{aligned}$$

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

$$\limsup_{n\to\infty} \min_{k=1,2,\ldots} \widehat{L}_1^n(h^{(k)}) \leq L^* \quad \text{almost surely.} \tag{4}$$

This will follow easily from the following lemma:

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

$$\limsup_{n\to\infty} \left| \widehat{L}_1^n(h^{(k)}) - \mathbf{P}\left\{ g^{(k)}(Y_{-k}^{-1}) \neq Y_0 \right\} \right| \leq \frac{2}{k} \quad \textit{almost surely,}$$

where for any $s \in \{0,1\}^k$,

$$g^{(k)}(s) = \begin{cases} 1 & \textit{if } \mathbf{P}\left\{ Y_0 = 1 | Y_{-k}^{-1} = s \right\} \geq \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases}$$

is the Bayes decision for Y_0 given Y_{-k}^{-1}.

Proof. Note that

$$\begin{aligned} & \widehat{L}_1^n(h^{(k)}) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[I_{\{h_i^{(k)}(Y_1^{i-1},U_i)\neq Y_i\}} | Y_0^\infty \right] \\ &= \left(\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[I_{\{h_i^{(k)}(Y_1^{i-1},U_i)\neq Y_i\}} | Y_0^\infty \right] - \frac{1}{n}\sum_{i=1}^n I_{\{g^{(k)}(Y_{i-k}^{i-1})\neq Y_i\}} \right) + \frac{1}{n}\sum_{i=1}^n I_{\{g^{(k)}(Y_{i-k}^{i-1})\neq Y_i\}}. \end{aligned}$$

For the second term on the right-hand side, it follows from the ergodic theorem that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n I_{\{g^{(k)}(Y_{i-k}^{i-1})\neq Y_i\}} = \mathbf{P}\left\{ g^{(k)}(Y_{-k}^{-1}) \neq Y_0 \right\} \quad \text{almost surely.}$$

Therefore, it suffices to show that

$$\limsup_{n\to\infty} \left| \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[I_{\{h_i^{(k)}(Y_1^{i-1},U_i)\neq Y_i\}} | Y_0^\infty \right] - \frac{1}{n}\sum_{i=1}^n I_{\{g^{(k)}(Y_{i-k}^{i-1})\neq Y_i\}} \right| \leq \frac{2}{k} \quad \text{almost surely.}$$

To see this, write

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[I_{\{h_i^{(k)}(Y_1^{i-1},U_i)\neq Y_i\}} | Y_0^\infty \right] - \frac{1}{n}\sum_{i=1}^n I_{\{g^{(k)}(Y_{i-k}^{i-1})\neq Y_i\}} \\ &= \frac{1}{n}\sum_{i=1}^n \left(\mathbf{E}\left[Y_i + h_i^{(k)}(Y_1^{i-1},U_i) - 2Y_i h_i^{(k)}(Y_1^{i-1},U_i) | Y_0^\infty \right] - Y_i - g^{(k)}(Y_{i-k}^{i-1}) + 2Y_i g^{(k)}(Y_{i-k}^{i-1}) \right) \\ & \quad \text{(using that if } a,b \in \{0,1\} \text{ then } I_{\{a\neq b\}} = a+b-2ab) \\ &= \frac{1}{n}\sum_{i=1}^n \left(Y_i + (1-2Y_i)\mathbf{E}\left[h_i^{(k)}(Y_1^{i-1},U_i) | Y_0^\infty \right] - Y_i - (1-2Y_i) g^{(k)}(Y_{i-k}^{i-1}) \right) \\ &= \frac{1}{n}\sum_{i=1}^n (1-2Y_i)\left(F_k(\widehat{P}_i^k(1, Y_1^{i-1} | Y_{i-k}^{i-1})) - g^{(k)}(Y_{i-k}^{i-1}) \right) \\ &= \left(\frac{1}{n}\sum_{i=1}^n (1-2Y_i)\left(F_k(\widehat{P}_i^k(1, Y_1^{i-1} | Y_{i-k}^{i-1})) - F_k(\mathbf{P}\{Y_i = 1 | Y_{i-k}^{i-1}\}) \right) \right) \\ & \quad + \left(\frac{1}{n}\sum_{i=1}^n (1-2Y_i)\left(F_k(\mathbf{P}\{Y_i = 1 | Y_{i-k}^{i-1})\} - g^{(k)}(Y_{i-k}^{i-1}) \right) \right). \end{aligned} \tag{5}$$

Now it follows from the ergodic theorem that

$$\lim_{n\to\infty} \max_{y\in\{0,1\},s\in\{0,1\}^k} \left|\widehat{P}_n^k(y, Y_1^{n-1}|s) - \mathbf{P}\left\{Y_0 = y|Y_{-k}^{-1} = s\right\}\right| = 0 \quad \text{almost surely,}$$

and therefore

$$\lim_{i\to\infty} \left|\widehat{P}_i^k(1, Y_1^{i-1}|Y_{i-k}^{i-1})) - \mathbf{P}\left\{Y_i = 1|Y_{i-k}^{i-1}\right\}\right| = 0 \quad \text{almost surely,}$$

so by the continuity of F_k, we have, for the first term on the right-hand side of (5), that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(1-2Y_i)\left(F_k(\widehat{P}_i^k(1, Y_1^{i-1}|Y_{i-k}^{i-1})) - F_k(\mathbf{P}\{Y_i = 1|Y_{i-k}^{i-1}\})\right) = 0 \quad \text{almost surely.}$$

For the second term on the right-hand side of (5), note that by the ergodic theorem, almost surely,

$$\begin{aligned}
&\lim_{n\to\infty} \left|\frac{1}{n}\sum_{i=1}^{n}(1-2Y_i)\left(F_k\left(\mathbf{P}\left\{Y_i = 1|Y_{i-k}^{i-1}\right\}\right) - g^{(k)}(Y_{i-k}^{i-1})\right)\right| \\
&= \left|\mathbf{E}\left[(1-2Y_0)\left(F_k\left(\mathbf{P}\left\{Y_0 = 1|Y_{-k}^{-1}\right\}\right) - g^{(k)}(Y_{-k}^{-1})\right)\right]\right| \\
&= \left|\mathbf{E}\left[\mathbf{E}\left[(1-2Y_0)\left(F_k\left(\mathbf{P}\left\{Y_0 = 1|Y_{-k}^{-1}\right\}\right) - g^{(k)}(Y_{-k}^{-1})\right)\Big|Y_{-k}^{-1}\right]\right]\right| \\
&= \left|\mathbf{E}\left[\left(1-2\mathbf{P}\left\{Y_0 = 1|Y_{-k}^{-1}\right\}\right)\left(F_k\left(\mathbf{P}\left\{Y_0 = 1|Y_{-k}^{-1}\right\}\right) - g^{(k)}(Y_{i-k}^{i-1})\right)\right]\right| \\
&\le \mathbf{E}\left|\left(1-2\mathbf{P}\left\{Y_0 = 1|Y_{-k}^{-1}\right\}\right) I_{\left\{\left|\mathbf{P}\left\{Y_0=1|Y_{-k}^{-1}\right\}-1/2\right|\le 1/k\right\}}\right| \\
&\le \mathbf{E}\left[\frac{2}{k} I_{\left\{\left|\mathbf{P}\left\{Y_0=1|Y_{-k}^{-1}\right\}-1/2\right|\le 1/k\right\}}\right] \\
&\le \frac{2}{k},
\end{aligned}$$

and Lemma 2 is proved. □

Now we return to the proof of Theorem 3. Since

$$\mathbf{P}\left\{g^{(k)}(Y_{-k}^{-1}) \neq Y_0\right\} = \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1|Y_{-k}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-k}^{-1}\}\right)\right],$$

it follows from the martingale convergence theorem and Lebesgue's dominated convergence theorem that

$$\begin{aligned}
\lim_{k\to\infty} \mathbf{P}\left\{g^{(k)}(Y_{-k}^{-1}) \neq Y_0\right\} &= \lim_{k\to\infty} \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1|Y_{-k}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-k}^{-1}\}\right)\right] \\
&= \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1|Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0|Y_{-\infty}^{-1}\}\right)\right] \\
&= L^*. \qquad (6)
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
&\lim_{k\to\infty} \limsup_{n\to\infty} \left|\widehat{L}_1^n(h^{(k)}) - L^*\right| \\
&\le \lim_{k\to\infty} \limsup_{n\to\infty} \left(\left|\widehat{L}_1^n(h^{(k)}) - \mathbf{P}\left\{g^{(k)}(Y_{-k}^{-1}) \neq Y_0\right\}\right| + \left|\mathbf{P}\left\{g^{(k)}(Y_{-k}^{-1}) \neq Y_0\right\} - L^*\right|\right) \\
&\le \lim_{k\to\infty} \frac{2}{k} \text{ (By Lemma 2 and (6)).} \\
&= 0.
\end{aligned}$$

Finally, for a fixed $\epsilon > 0$, choose the positive integer K such that $\limsup_{n\to\infty} \left|\widehat{L}_1^n(h^{(K)}) - L^*\right| < \epsilon$. Then

$$\begin{aligned}\limsup_{n\to\infty} \min_{k=1,2,\ldots} \widehat{L}_1^n(h^{(k)}) &\leq \limsup_{n\to\infty} \widehat{L}_1^n(h^{(K)}) \\ &\leq L^* + \epsilon.\end{aligned}$$

Since ϵ was arbitrary, (4) 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 $\widehat{L}_{2^m}^{n-1}(h^{(k)})$ for $k = 1, 2, \ldots, n$.

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

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

In other words, the algorithm is guaranteed to perform almost well as the best expert. The rate of convergence to L^* depends on the behavior of the best expert.

3. The function F_k is defined somewhat arbitrarily. All that's needed for consistency is that $F_k(z)$ is continuous and it only differs from $I_{\{z\geq 1/2\}}$ in a shrinking neighborhood of $1/2$ as $k \to \infty$. For finite-sample behavior the choice of F_k may be an important issue, however, we cannot offer any good intuition on this. Actually, $F_k(z) = I_{\{z\geq 1/2\}}$ is the most natural choice, but continuity of F_k is needed in our analysis. We do not know if it is necessary.

3 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 $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$ 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 the previous section 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 risk* 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 $R_n(g) = R_1^n(g, U_1^n)$. Denote the expected risk of the randomized strategy g by

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

Similarly to the notation of the previous section, we write

$$R_m^n(g, U_1^n) = \frac{1}{n-m+1} \sum_{i=m}^{n} I_{\{g_i(y_1^{i-1}, x_1^i, U_i) \neq y_i\}}, \quad \text{and} \quad \widehat{R}_m^n(g) = \mathbf{E} R_m^n(g, U_1^n).$$

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

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

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 [11] that there exists a prediction strategy g such that for all ergodic processes, $R_n(g) \to R^*$ almost surely. (The result of [11] should be complemented with an argument similar appearing in the proof of Theorem 2 to obtain the above statement. To avoid repetition, the details are again omitted.) The algorithm of Morvai, Yakowitz, and Györfi, however, is not useful in practice, as it requires

an astronomical data size. The main message of this section is a simple consistent 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 $s \in \{0,1\}^k$ and $z \in \{1,2,\ldots,m_\ell\}^{k+1}$ and $y \in \{0,1\}$ define

$$\widehat{P}_n^{(k,\ell)}(y,y_1^{n-1},x_1^n|s,z) = \frac{\left|\{k<i<n : y_{i-k}^{i-1} = s, G_\ell(x_{i-k}^i) = z, y_i = y\}\right|}{\left|\{k<i<n : y_{i-k}^{i-1} = s, G_\ell(x_{i-k}^i) = z,\}\right|}, \quad n > k+1. \tag{7}$$

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

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

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

where F_k is defined in the previous section. 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 lenght 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 the previous section. 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:

$$g_n(y_1^{n-1},x_1^n,u) = \begin{cases} 0 & \text{if } u < \dfrac{\sum_{k,\ell\le 2^{m+1}} F_k(\widehat{P}_n^{(k,\ell)}(0,y_1^{n-1},x_1^n|y_{n-k}^{n-1},G_\ell(x_{n-k}^n)))w_n(k,\ell)}{\sum_{k,\ell\le 2^{m+1}} w_n(k,\ell)} \\ 1 & \text{otherwise,} \end{cases}$$

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

$$w_{2^m}(k,\ell) = 1 \quad \text{and} \quad w_n(k,\ell) = e^{-\eta_m \widehat{R}_{2^m}^{n-1}(h^{(k,\ell)})} \quad n > 2^m,$$

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

For the consistency 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 $\mathrm{diam}(A) = \sup_{x,y \in A} \|x - y\|$ denotes the diameter of a set, then for each sphere S centered at the origin

$$\lim_{\ell \to \infty} \max_{j; A_{\ell,j} \cap S \neq \emptyset} \mathrm{diam}(A_{\ell,j}) = 0.$$

Theorem 5 *Assume that the sequence of partitions $\mathcal{P}_\ell$ satisfies the three conditions above. Then the pattern recognition scheme g defined above satisfies*

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

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

Proof of Theorem 5. Exactly the same way as in the first part of the proof of Theorem 3, we obtain that for any stationary and ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\begin{aligned} \limsup_{n \to \infty} R_1^n(g, U_1^n) &\leq \limsup_{n \to \infty} \left(\min_{\substack{k=1,2,\ldots \\ \ell = 1,2,\ldots,n-1}} \widehat{R}_1^n(h^{(k,\ell)}) + 2c\sqrt{\frac{\log_2 n + 1}{n}} \right) \\ &= \limsup_{n \to \infty} \min_{\substack{k=1,2,\ldots \\ \ell = 1,2,\ldots,n-1}} \widehat{R}_1^n(h^{(k,\ell)}) \quad \text{almost surely.} \end{aligned}$$

Thus, it remains to show that

$$\limsup_{n \to \infty} \min_{\substack{k=1,2,\ldots \\ \ell = 1,2,\ldots,n-1}} \widehat{R}_1^n(h^{(k,\ell)}) \leq R^* \quad \text{almost surely.}$$

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

Lemma 3 *For each $k \geq 1$,*

$$\limsup_{n \to \infty} \left| \widehat{R}(h^{(k,\ell)}) - \mathbf{P}\left\{ g^{(k,\ell)}(Y_{-k}^{-1}, X_{-k}^0) \neq Y_0 \right\} \right| \leq \frac{2}{k} \quad \textit{almost surely,}$$

where for any $s \in \{0,1\}^k$ and $z \in \{1, 2, \ldots, m_\ell\}^{k+1}$,

$$g^{(k,\ell)}(s, z) = \begin{cases} 1 & \textit{if } \mathbf{P}\left\{ Y_0 = 1 | Y_{-k}^{-1} = s, G_\ell(X_{-k}^0) = z \right\} \geq \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases}$$

is the Bayes decision for Y_0 given $Y_{-k}^{-1}, G_\ell(X_{-k}^0)$.

Now we return to the proof of Theorem 5. For fix ℓ, the sequences

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

are martingales, and they converge almost surely to

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

respectively. 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

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

Since

$$\mathbf{P}\left\{g^{(k,\ell)}(Y_{-k}^{-1}, X_{-k}^0) \neq Y_0\right\} = \mathbf{E}\left[\min\left(\mathbf{P}\{Y_0 = 1|Y_{-k}^{-1}, G_\ell(X_{-k}^0)\}, \mathbf{P}\{Y_0 = 0|Y_{-k}^{-1}, G_\ell(X_{-k}^0)\}\right)\right],$$

we conclude that

$$\begin{aligned} &\lim_{\ell\to\infty}\lim_{k\to\infty}\limsup_{n\to\infty}\left|\widehat{R}_1^n(h^{(k,\ell)}) - R^*\right| \\ &\leq \lim_{\ell\to\infty}\lim_{k\to\infty}\limsup_{n\to\infty}\left(\left|\widehat{R}_1^n(h^{(k,\ell)}) - \mathbf{P}\left\{g^{(k,\ell)}(Y_{-k}^{-1}, X_{-k}^0) \neq Y_0\right\}\right|\right. \\ &\quad \left. + \left|\mathbf{P}\left\{g^{(k,\ell)}(Y_{-k}^{-1}, X_{-k}^0) \neq Y_0\right\} - R^*\right|\right) \\ &= 0 \quad \text{almost surely.} \end{aligned}$$

Now it follows easily that

$$\limsup_{n\to\infty} \min_{\substack{k=1,2,\ldots \\ \ell=1,2,\ldots,n-1}} \widehat{R}_1^n(h^{(k,\ell)}) \leq R^* \quad \text{almost surely,}$$

and the proof of the theorem is finished. □

Acknowledgement. We thank Nicoló Cesa-Bianchi for teaching us all wee needed to know about prediction with expert advise.

References

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

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

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

[4] 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.

[5] 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.

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

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

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

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

[10] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

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

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

[13] W.F. Stout. *Almost sure convergence.* Academic Press, New York, 1974.

[14] 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.