Sequential prediction of unbounded stationary time series

László Györfi, *Fellow, IEEE,* and György Ottucsák

Abstract

A simple on-line procedure is considered for the prediction of a real valued sequence. The algorithm is based on a combination of several simple predictors. If the sequence is a realization of an unbounded stationary and ergodic random process then the average of squared errors converges, almost surely, to that of the optimum, given by the Bayes predictor. An analog result is offered for the classification of binary processes.

Index Terms

On-line learning, sequential prediction, time series, universal consistency, pattern recognition.

I. Introduction

We study the problem of sequential prediction of a real valued sequence. At each time instant $t = 1, 2, \ldots$, the predictor is asked to guess the value of the next outcome y_t of a sequence of real numbers $y_1, y_2, \ldots$ with knowledge of the pasts $y_1^{t-1} = (y_1, \ldots, y_{t-1})$ (where y_1^0 denotes the empty string) and the side information vectors $x_1^t = (x_1, \ldots, x_t)$, where $x_t \in \mathbb{R}^d$. Thus, the predictor's estimate, at time t, is based on the value of x_1^t and y_1^{t-1}. A prediction strategy is a sequence $g = \{g_t\}_{t=1}^{\infty}$ of functions

$$g_t : \left(\mathbb{R}^d\right)^t \times \mathbb{R}^{t-1} \to \mathbb{R}$$

so that the prediction formed at time t is $g_t(x_1^t, y_1^{t-1})$.

In this paper we assume that $(x_1, y_1), (x_2, y_2), \ldots$ are realizations of the random variables $(X_1, Y_1), (X_2, Y_2), \ldots$ such that $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ is a jointly stationary and ergodic process.

L. Györfi is with Department of Computer Science and Information Theory Budapest University of Technology and Economics, Magyar tudósok körútja 2., Budapest, Hungary, H-1117 (email: gyorfi@szit.bme.hu).

Gy. Ottucsák is with Department of Computer Science and Information Theory Budapest University of Technology and Economics, Magyar tudósok körútja 2., Budapest, Hungary, H-1117 (email: oti@szit.bme.hu).

After n time instants, the *normalized cumulative prediction error* is

$$L_n(g) = \frac{1}{n}\sum_{t=1}^{n}(g_t(X_1^t, Y_1^{t-1}) - Y_t)^2.$$

The results of the paper are given in an autoregressive framework, that is, the value Y_t is predicted based on X_1^t and Y_1^{t-1}. The fundamental limit for the predictability of the sequence can be determined based on a result of Algoet [2], who showed that for any prediction strategy g and stationary ergodic process $\{(X_n, Y_n)\}_{-\infty}^{\infty}$,

$$\liminf_{n\to\infty} L_n(g) \geq L^* \quad \text{almost surely,} \tag{1}$$

where

$$L^* = \mathbb{E}\left\{\left(Y_0 - \mathbb{E}\{Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1}\}\right)^2\right\}$$

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past $X_{-\infty}^0, Y_{-\infty}^{-1}$. Note that it follows by stationarity and the martingale convergence theorem (see, e.g., Stout [22]) that

$$L^* = \lim_{n\to\infty} \mathbb{E}\left\{\left(Y_n - \mathbb{E}\{Y_n | X_1^n, Y_1^{n-1}\}\right)^2\right\}.$$

This lower bound gives sense to the following definition:

Definition 1: A prediction strategy g is called *universally consistent with respect to a class $\mathcal{C}$ of stationary and ergodic processes* $\{(X_n, Y_n)\}_{-\infty}^{\infty}$, if for each process in the class,

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

Universally consistent strategies asymptotically achieve the best possible loss for all ergodic processes in the class. Algoet [1] and Morvai, Yakowitz, and Györfi [17] proved that there exists a prediction strategy universal with respect to the class of all bounded ergodic processes. However, the prediction strategies exhibited in these papers are either very complex or have an unreasonably slow rate of convergence even for well-behaved processes.

Lugosi and Györfi [11] introduced several simple prediction strategies, which are universally consistent with respect to the class of bounded, stationary and ergodic processes. In this paper we extend the results of [11] to unbounded processes. The algorithms build on a methodology worked out in recent years for prediction of individual sequences, see Feder, Merhav, and Gutman [9], Littlestone and Warmuth [13], Cesa-Bianchi and Lugosi [6], Singer and Feder [20], Merhav and Feder [14] for a survey. Most of the result in individual framework holds on for fixed time horizon, which does not suit the asymptotic analysis. Accordingly, the main ingredients of our proof is a lemma which allows us to extend the above cited methods for the asymptotic studies in a simple way. We refer to Nobel [18], Singer and Feder [20], [21] and Yang [25] to recent closely related work. A distinct concept, memory universality is studied by Modha and Masry [15] for bounded and exponentially strongly mixing random process, where the decay coefficients is known.

In Section II we introduce an universally consistent strategy for unbounded ergodic processes which is based on a combination of partitioning estimates. In Section III we consider the $0-1$ loss, i.e., construct a recursive pattern recognition scheme for stationary and ergodic process.

II. UNIVERSAL PREDICTION BY PARTITIONING ESTIMATES

The prediction strategy is defined, at each time instant, as a convex combination of *elementary predictors*, where the weighting coefficients depend on the past performance of each elementary predictor.

We define an infinite array of elementary predictors $h^{(k,\ell)}$, $k,\ell = 1,2,\dots$ as follows. Let $\mathcal{P}_\ell = \{A_{\ell,j}, j = 1,2,\dots,m_\ell\}$ be a sequence of finite partitions of $\mathbb{R}$, and let $\mathcal{Q}_\ell = \{B_{\ell,j}, j = 1,2,\dots,m'_\ell\}$ be a sequence of finite partitions of $\mathbb{R}^d$. Introduce the corresponding quantizers:

$$F_\ell(y) = j, \text{ if } y \in A_{\ell,j}$$

and

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

With some abuse of notation, for any n and $y_1^n \in \mathbb{R}^n$, we write $F_\ell(y_1^n)$ for the sequence $F_\ell(y_1),\dots,F_\ell(y_n)$, and similarly, for $x_1^n \in (\mathbb{R}^d)^n$, we write $G_\ell(x_1^n)$ for the sequence $G_\ell(x_1),\dots,G_\ell(x_n)$.

Fix positive integers k,ℓ, and for each $k+1$-long string z of positive integers, and for each k-long string s of positive integers, define the partitioning regression function estimate

$$\widehat{E}_n^{(k,\ell)}(x_1^n, y_1^{n-1}, z, s) = \frac{\sum_{\{k<t<n: G_\ell(x_{t-k}^t)=z,\, F_\ell(y_{t-k}^{t-1})=s\}} y_t}{\left|\{k<t<n : G_\ell(x_{t-k}^t) = z,\, F_\ell(y_{t-k}^{t-1}) = s\}\right|},$$

for all $n > k+1$ where $0/0$ is defined to be 0.

Introduce the truncation function

$$T_n(z) = \begin{cases} n^\delta & \text{if } z > n^\delta \\ z & \text{if } |z| < n^\delta \\ -n^\delta & \text{if } z < -n^\delta, \end{cases}$$

where

$$0 < \delta < 1/8.$$

Define the elementary predictor $h^{(k,\ell)}$ by

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = T_n\left(\widehat{E}_n^{(k,\ell)}(x_1^n, y_1^{n-1}, G_\ell(x_{n-k}^n), F_\ell(y_{n-k}^{n-1}))\right),$$

for $n = 1, 2, \ldots$. That is, $h_n^{(k,\ell)}$ quantizes the sequence x_1^n, y_1^{n-1} according to the partitions $\mathcal{Q}_\ell$ and $\mathcal{P}_\ell$, and looks for all appearances of the last seen quantized strings $G_\ell(x_{n-k}^n)$ of length $k+1$ and $F_\ell(y_{n-k}^{n-1})$ of length k in the past. Then it predicts according to the truncation of the average of the y_t's following the string.

The proposed prediction algorithm proceeds as follows: let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k,ℓ) of positive integers such that for all k, ℓ, $q_{k,\ell} > 0$. For $\eta_t > 0$, and define the weights

$$w_{t,k,\ell} = q_{k,\ell} e^{-\eta_t (t-1) L_{t-1}(h^{(k,\ell)})}$$

and their normalized values

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{W_t} ,$$

where

$$W_t = \sum_{i,j=1}^{\infty} w_{t,i,j} .$$

The prediction strategy g is defined by

$$g_t(x_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(x_1^t, y_1^{t-1}) , \qquad t = 1, 2, \ldots \tag{2}$$

Theorem 1: Assume that

(a) the sequences of partition $\mathcal{P}_\ell$ 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) the sequences of partition $\mathcal{Q}_\ell$ is nested;

(c) the sequences of partition $\mathcal{P}_\ell$ is asymptotically fine, i.e., 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

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

(d) the sequences of partition $\mathcal{Q}_\ell$ is asymptotically fine;

and choosing parameter of the algorithm as

$$\eta_t = \frac{1}{\sqrt{t}} .$$

Then the prediction scheme g defined above is universally consistent with respect to the class of all ergodic processes such that

$$\mathbf{E}\{Y_1^4\} < \infty.$$

Here we describe two results, which are used in the analysis. The first lemma is a modification of the analysis of Auer *et al.* [3], which allows of the handling the case when the parameter of the algorithm (η_t) is time-dependent and the number of the elementary predictors is infinite.

Lemma 1: Let $h^{(1)}, h^{(2)}, \dots$ be a sequence of prediction strategies (experts). Let $\{q_k\}$ be a probability distribution on the set of positive integers. Denote the normalized loss of the expert $h = (h_1, h_2, \dots)$ by

$$L_n(h) = \frac{1}{n} \sum_{t=1}^{n} \ell_t(h),$$

where

$$\ell_t(h) = \ell(h_t, Y_t)$$

and the loss function ℓ is convex in its first argument h. Define

$$w_{t,k} = q_k e^{-\eta_t (t-1) L_{t-1}(h^{(k)})}$$

where $\eta_t > 0$ is monotonically decreasing, and

$$p_{t,k} = \frac{w_{t,k}}{W_t}$$

where

$$W_t = \sum_{k=1}^{\infty} w_{t,k} \ .$$

If the prediction strategy $g = (g_1, g_2, \dots)$ is defined by

$$g_t = \sum_{k=1}^{\infty} p_{t,k} h_t^{(k)} \qquad t = 1, 2, \dots$$

then for every $n \geq 1$,

$$L_n(g) \leq \inf_k \left(L_n(h^{(k)}) - \frac{\ln q_k}{n \eta_{n+1}} \right) + \frac{1}{2n} \sum_{t=1}^{n} \eta_t \sum_{k=1}^{\infty} p_{t,k} \ell_t^2(h^{(k)}).$$

Proof. Introduce some notations:

$$w'_{t,k} = q_k e^{-\eta_{t-1} (t-1) L_{t-1}(h^{(k)})},$$

which is the weight $w_{t,k}$, where η_t is replaced by η_{t-1} and the sum of these are

$$W'_t = \sum_{k=1}^{\infty} w'_{t,k}.$$

We start the proof with the following chain of bounds:

$$\begin{aligned}\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t} &= \frac{1}{\eta_t}\ln\frac{\sum_{k=1}^{\infty} w_{t,k}e^{-\eta_t\ell_t(h^{(k)})}}{W_t}\\ &= \frac{1}{\eta_t}\ln\sum_{k=1}^{\infty} p_{t,k}e^{-\eta_t\ell_t(h^{(k)})}\\ &\le \frac{1}{\eta_t}\ln\sum_{k=1}^{\infty} p_{t,k}\left(1-\eta_t\ell_t(h^{(k)})+\frac{\eta_t^2}{2}\ell_t^2(h^{(k)})\right)\end{aligned}$$

because of $e^{-x}\le 1-x+x^2/2$ for $x\ge 0$. Moreover,

$$\begin{aligned}&\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t}\\ &\le \frac{1}{\eta_t}\ln\left(1-\eta_t\sum_{k=1}^{\infty} p_{t,k}\ell_t(h^{(k)})+\frac{\eta_t^2}{2}\sum_{k=1}^{\infty} p_{t,k}\ell_t^2(h^{(k)})\right)\\ &\le -\sum_{k=1}^{\infty} p_{t,k}\ell_t(h^{(k)})+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\ell_t^2(h^{(k)}) &(3)\\ &= -\sum_{k=1}^{\infty} p_{t,k}\ell(h_t^{(k)},Y_t)+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\ell_t^2(h^{(k)})\\ &\le -\ell\left(\sum_{k=1}^{\infty} p_{t,k}h_t^{(k)},Y_t\right)+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\ell_t^2(h^{(k)}) &(4)\\ &= -\ell_t(g)+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\ell_t^2(h^{(k)}) &(5)\end{aligned}$$

where (3) follows from the fact that $\ln(1+x)\le x$ for all $x>-1$ and in (4) we used the convexity of the loss $\ell(h,y)$ in its first argument h. From (5) after rearranging we obtain

$$\ell_t(g)\le -\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t}+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\ell_t^2(h^{(k)})\ .$$

Then write a telescope formula:

$$\begin{aligned}\frac{1}{\eta_t}\ln W_t-\frac{1}{\eta_t}\ln W'_{t+1} &= \left(\frac{1}{\eta_t}\ln W_t-\frac{1}{\eta_{t+1}}\ln W_{t+1}\right)\\ &\quad+\left(\frac{1}{\eta_{t+1}}\ln W_{t+1}-\frac{1}{\eta_t}\ln W'_{t+1}\right)\\ &= (A_t)+(B_t).\end{aligned}$$

We have that

$$\begin{aligned}\sum_{t=1}^{n} A_t &= \sum_{t=1}^{n}\left(\frac{1}{\eta_t}\ln W_t - \frac{1}{\eta_{t+1}}\ln W_{t+1}\right)\\ &= \frac{1}{\eta_1}\ln W_1 - \frac{1}{\eta_{n+1}}\ln W_{n+1}\\ &= -\frac{1}{\eta_{n+1}}\ln\sum_{k=1}^{\infty} q_k e^{-\eta_{n+1} n L_n(h^{(k)})}\\ &\le -\frac{1}{\eta_{n+1}}\ln\sup_k q_k e^{-\eta_{n+1} n L_n(h^{(k)})}\\ &= -\frac{1}{\eta_{n+1}}\sup_k\left(\ln q_k - \eta_{n+1} n L_n(h^{(k)})\right)\\ &= \inf_k\left(nL_n(h^{(k)}) - \frac{\ln q_k}{\eta_{n+1}}\right).\end{aligned}$$

$\frac{\eta_{t+1}}{\eta_t} \le 1$, therefore applying Jensen's inequality for concave function, we get that

$$\begin{aligned}W_{t+1} &= \sum_{i=1}^{\infty} q_i e^{-\eta_{t+1} t L_t(h^{(i)})}\\ &= \sum_{i=1}^{\infty} q_i \left(e^{-\eta_t t L_t(h^{(i)})}\right)^{\frac{\eta_{t+1}}{\eta_t}}\\ &\le \left(\sum_{i=1}^{\infty} q_i e^{-\eta_t t L_t(h^{(i)})}\right)^{\frac{\eta_{t+1}}{\eta_t}}\\ &= \left(W'_{t+1}\right)^{\frac{\eta_{t+1}}{\eta_t}}.\end{aligned}$$

Thus,

$$\begin{aligned}B_t &= \frac{1}{\eta_{t+1}}\ln W_{t+1} - \frac{1}{\eta_t}\ln W'_{t+1}\\ &\le \frac{1}{\eta_{t+1}}\frac{\eta_{t+1}}{\eta_t}\ln W'_{t+1} - \frac{1}{\eta_t}\ln W'_{t+1}\\ &= 0.\end{aligned}$$

We can summarize the bounds:

$$L_n(g) \le \inf_k\left(L_n(h^{(k)}) - \frac{\ln q_k}{n\eta_{n+1}}\right) + \frac{1}{2n}\sum_{t=1}^{n}\eta_t\sum_{k=1}^{\infty} p_{t,k}\ell_t^2(h^{(k)}) .$$

□

The next lemma is due to Breiman [5], and its proof may also be found in Györfi *et al.* [10].

Lemma 2: 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

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

almost surely.

Proof of Theorem 1. Because of (1), it is enough to show that

$$\limsup_{n\to\infty} L_n(g) \le L^* \qquad \text{a.s.}$$

By a double application of the ergodic theorem, as $n \to \infty$, a.s.,

$$\begin{aligned}
&\widehat{E}_n^{(k,\ell)}(X_1^n, Y_1^{n-1}, z, s) \\
=\;& \frac{\frac{1}{n}\sum_{\{k<t<n:\, G_\ell(X_{t-k}^t)=z,\, F_\ell(Y_{t-k}^{t-1})=s\}} Y_t}{\frac{1}{n}\left|\{k<t<n:\, G_\ell(X_{t-k}^t)=z,\, F_\ell(Y_{t-k}^{t-1})=s\}\right|} \\
\to\;& \frac{\mathbf{E}\{Y_0 I_{\{G_\ell(X_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\}}\}}{\mathbf{P}\{G_\ell(X_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\}} \\
=\;& \mathbf{E}\{Y_0 \mid G_\ell(X_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\},
\end{aligned}$$

and therefore for all z and s

$$T_n\left(\widehat{E}_n^{(k,\ell)}(X_1^n, Y_1^{n-1}, z, s)\right) \to \mathbf{E}\{Y_0 \mid G_\ell(X_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\}.$$

By Lemma 2, as $n \to \infty$, almost surely,

$$\begin{aligned}
&L_n(h^{(k,\ell)}) \\
&= \frac{1}{n}\sum_{t=1}^{n}(h^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(T_t\left(\widehat{E}_t^{(k,\ell)}(X_1^t, Y_1^{t-1}, G_\ell(X_{t-k}^t), F_\ell(Y_{t-k}^{t-1}))\right) - Y_t\right)^2 \\
&\to \mathbf{E}\{(Y_0 - \mathbf{E}\{Y_0 \mid G_\ell(X_{-k}^0),\, F_\ell(Y_{-k}^{-1})\})^2\} \\
&\stackrel{\text{def}}{=} \quad \epsilon_{k,\ell}.
\end{aligned}$$

$\mathbf{E}\{Y_0 \mid G_\ell(X_{-k}^0),\, F_\ell(Y_{-k}^{-1})\}$ is a martingale indexed by the pair (k,ℓ), since the partitions $\mathcal{P}_\ell$ and $\mathcal{Q}_\ell$ are nested. Thus, the martingale convergence theorem (see, e.g., Stout [22]) and assumptions (c) and (d) for the sequences of partitions implies that

$$\inf_{k,l} \epsilon_{k,l} = \lim_{k,\ell\to\infty} \epsilon_{k,\ell} = \mathbf{E}\left\{\left(Y_0 - \mathbf{E}\{Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1}\}\right)^2\right\} = L^*$$

(cf. Györfi and Lugosi [11]). Apply Lemma 1 with choice $\eta_t = \frac{1}{\sqrt{t}}$ and for the squared loss $\ell_t(h) = (h_t - Y_t)^2$, then the square loss is convex in its first argument h, so

$$L_n(g) \le \inf_{k,\ell}\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right) + \frac{1}{2n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell=1}^{\infty} p_{t,k,\ell}\big(h^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t\big)^4. \tag{6}$$

On the one hand, almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty}\inf_{k,\ell}\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right)\\
&\le \inf_{k,\ell}\limsup_{n\to\infty}\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right)\\
&= \inf_{k,\ell}\limsup_{n\to\infty} L_n(h^{(k,\ell)})\\
&= \inf_{k,\ell}\epsilon_{k,\ell}\\
&= \lim_{k,\ell\to\infty}\epsilon_{k,\ell}\\
&= L^*.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell} p_{t,k,\ell}(h^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t)^4\\
&\le \frac{8}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell} p_{t,k,\ell}\left(h^{(k,\ell)}(X_1^t, Y_1^{t-1})^4 + Y_t^4\right)\\
&\le \frac{8}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell} p_{t,k,\ell}\left(t^{4\delta} + Y_t^4\right)\\
&= \frac{8}{n}\sum_{t=1}^{n}\frac{t^{4\delta} + Y_t^4}{\sqrt{t}},
\end{aligned}$$

therefore, almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell} p_{t,k,\ell}(h^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t)^4\\
&\le \limsup_{n\to\infty}\frac{8}{n}\sum_{t=1}^{n}\frac{Y_t^4}{\sqrt{t}}\\
&= 0,
\end{aligned}$$

where we applied that $\mathbf{E}\{Y_1^4\} < \infty$ and $0 < \delta < \frac{1}{8}$. Summarizing these bounds, we get that, almost surely,

$$\limsup_{n\to\infty} L_n(g) \le L^*$$

and the proof of the theorem is finished. $\square$

Corollary 1: Under the conditions of Theorem 1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \left(\mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1})\right)^2 = 0 \qquad \text{a.s.} \tag{7}$$

Proof. By Theorem 1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 = L^* \qquad \text{a.s.} \tag{8}$$

and by the ergodic theorem we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbf{E}\left\{\left(Y_t - \mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\right\} = L^* \tag{9}$$

almost surely. Now we may write as $n \to \infty$, that

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} \left(\mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1})\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbf{E}\{\left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&\quad -\frac{1}{n}\sum_{t=1}^{n} \mathbf{E}\{\left(Y_t - \mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbf{E}\{\left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&\quad -\frac{1}{n}\sum_{t=1}^{n} \left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 + o(1) \qquad (10) \\
&= 2\frac{1}{n}\sum_{t=1}^{n} g_t(X_1^t, Y_1^{t-1})(Y_t - \mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}) \\
&\quad -\frac{1}{n}\sum_{t=1}^{n} \left(Y_t^2 - \mathbf{E}\{Y_t^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\right) + o(1) \qquad \text{a.s.}
\end{aligned}$$

where (10) holds because of (8) and (9). The second sum is

$$\frac{1}{n}\sum_{t=1}^{n} \left(Y_t^2 - \mathbf{E}\{Y_t^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\right) \to 0 \qquad \text{a.s.}$$

by the ergodic theorem. Put

$$Z_t = g_t(X_1^t, Y_1^{t-1})(Y_t - \mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}).$$

In order to finish the proof it suffices to show

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Z_t = 0 \ . \tag{11}$$

Then

$$\mathbf{E}\{Z_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} = 0,$$

for all t, so the Z_t's form a martingale difference sequence. By the strong law of large numbers for martingale differences due to Chow [7] (see also Stout [22, Theorem 3.3.1]), if $\{Z_t\}$ is a martingale difference sequence with

$$\sum_{n=1}^{\infty} \frac{\mathbf{E}Z_n^2}{n^2} < \infty, \tag{12}$$

then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Z_t = 0 \qquad \text{a.s.}$$

We have to verify (12). By the construction of g_n,

$$\begin{aligned}
\mathbf{E}\left\{Z_n^2\right\} &= \mathbf{E}\left\{\left(g_n(X_1^n, Y_1^{n-1})(Y_n - \mathbf{E}\{Y_n \mid X_{-\infty}^n, Y_{-\infty}^{n-1}\})\right)^2\right\} \\
&\le \mathbf{E}\left\{g_n(X_1^n, Y_1^{n-1})^2 Y_n^2\right\} \\
&\le n^{2\delta} \mathbf{E}\left\{Y_1^2\right\},
\end{aligned}$$

therefore (12) is verified, (11) is proved and the proof of the corollary is finished. □

Remark. CHOICE OF $q_{k,\ell}$. Theorem 1 is true independently of the choice of the $q_{k,\ell}$'s as long as these values are strictly positive for all k and ℓ. In practice, however, the choice of $q_{k,\ell}$ may have an impact on the performance of the predictor. For example, if the distribution $\{q_{k,\ell}\}$ has a very rapidly decreasing tail, then the term $-\ln q_{k,\ell}/\sqrt{n}$ will be large for moderately large values of k and ℓ, and the performance of g will be determined by the best of just a few of the elementary predictors $h^{(k,\ell)}$. Thus, it may be advantageous to choose $\{q_{k,\ell}\}$ to be a large-tailed distribution. For example, $q_{k,\ell} = c_0 k^{-2}\ell^{-2}$ is a safe choice, where c_0 is an appropriate normalizing constant.

III. PREDICTION FOR BINARY LABELS

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 $\mathbb{R}^d \times \{0, 1\}$. The problem is to predict the value of Y_n given the data (X_1^n, Y_1^{n-1}).

We may formalize the prediction (classification) problem as follows. The strategy of the classifier is a sequence $f = \{f_t\}_{t=1}^{\infty}$ of decision functions

$$f_t : \left(\mathbb{R}^d\right)^t \times \{0,1\}^{t-1} \to \{0,1\}$$

so that the classification formed at time t is $f_t(X_1^t, Y_1^{t-1})$. The *normalized cumulative* $0-1$ *loss* for any fixed pair of sequences $X_1^n,\ Y_1^n$ is now

$$R_n(f) = \frac{1}{n}\sum_{t=1}^{n} I_{\{f_t(X_1^t, Y_1^{t-1}) \neq Y_t\}}.$$

In this case there is a fundamental limit for the predictability of the sequence, i.e., Algoet [2] proved that for any classification strategy f and stationary ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\liminf_{n\to\infty} R_n(f) \geq R^* \quad \text{a.s.,} \tag{13}$$

where

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

therefore the following definition is meaningful:

Definition 2: A classification strategy f is called *Cesaro consistent* if for all stationary and ergodic processes $\{X_n, Y_n\}_{-\infty}^{\infty}$,

$$\lim_{n\to\infty} R_n(f) = R^* \quad \text{almost surely.}$$

Therefore, Cesaro consistent strategies asymptotically achieve the best possible loss for all ergodic processes. The first question is, of course, if such a strategy exists. Ornstein [19] and Bailey [4] proved the existence of Cesaro consistent predictors. This was later generalized by Algoet [1]. A simpler estimator with the same convergence property was introduced by Morvai, Yakowitz, and Györfi [17]. Motivated by the need of a practical estimator, Morvai, Yakowitz, and Algoet [16] introduced an even simpler algorithm. However, it is not known whether their predictor is Cesaro consistent. Györfi, Lugosi, and Morvai [12] introduced a simple randomized Cesaro consistent procedure with a practical appeal. Their idea was to combine the decisions of a small number of simple experts in an appropriate way.

The same idea was used in Weissman and Merhav [24]. They studied the consistency in noisy environment. In their model the past of Y_t is not available for the predictor, it has only access to the noisy past X_1^{t-1}. X_t is a noisy function of Y_t, that is, $X_t = u(Y_t, N_t)$, where $u : \{0,1\} \times \mathbb{R} \to \mathbb{R}$ is a function and $\{N_t\}$ is some noise process. A general loss function $\ell(f_t'(X_1^{t-1}), Y_t)$ is considered, where $f_t' : \mathbb{R}^{t-1} \to \mathbb{R}$ and $f_t'(X_1^{t-1})$ is the estimate of Y_t. They used an algorithm based on Vovk [23] to combine the simple experts and used doubling trick to fit the algorithm to infinite time horizon. In case of $0-1$ loss, one may easily modify the results in the sequel such that, they can be applied for the problem of [24].

In this section we present a simple (non-randomized) on-line classification strategy, and prove its Cesaro consistency. Consider the partitioning prediction scheme $g_t(X_1^t, Y_1^{t-1})$ introduced in Section II with

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \widehat{E}_n^{(k,\ell)}(x_1^n, y_1^{n-1}, G_\ell(x_{n-k}^n), y_{n-k}^{n-1}),$$

for $n = 1, 2, \ldots$, and then introduce the corresponding classification scheme:

$$f_t(X_1^t, Y_1^{t-1}) = \begin{cases} 1 & \text{if } g_t(X_1^t, Y_1^{t-1}) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

The main result of this section is the Cesaro consistency of this simple classification scheme:

Theorem 2: Assume that the conditions of Theorem 1 on the sequences of partitions $\mathcal{Q}_\ell$ satisfy and $\eta_t = \frac{1}{\sqrt{t}}$. Then the classification scheme f defined above satisfies

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

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

Proof. Because of (13) we have to show that

$$\limsup_{n\to\infty} R_n(f) \leq R^* \quad \text{a.s.}$$

By Corollary 1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1}) \right)^2 = 0 \qquad \text{a.s.} \tag{14}$$

Introduce the Bayes classification scheme using the infinite past:

$$f_t^*(X_{-\infty}^t, Y_{-\infty}^{t-1}) = \begin{cases} 1 & \text{if } \mathbf{P}\{Y_t = 1 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

and its normalized cumulative $0-1$ loss:

$$R_n(f^*) = \frac{1}{n} \sum_{t=1}^{n} I_{\{f_t^*(X_{-\infty}^t, Y_{-\infty}^{t-1}) \neq Y_t\}}.$$

Put

$$\bar{R}_n(f) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{P}\{f_t(X_1^t, Y_1^{t-1}) \neq Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}$$

and

$$\bar{R}_n(f^*) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{P}\{f_t^*(X_{-\infty}^t, Y_{-\infty}^{t-1}) \neq Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}.$$

Then

$$R_n(f) - \bar{R}_n(f) \to 0 \qquad \text{a.s.}$$

and

$$R_n(f^*) - \bar{R}_n(f^*) \to 0 \qquad \text{a.s.,}$$

since they are the averages of bounded martingale differences. Moreover, by the ergodic theorem

$$\bar{R}_n(f^*) \to R^* \qquad \text{a.s.,}$$

so we have to show that

$$\limsup_{n\to\infty}(\bar{R}_n(f) - \bar{R}_n(f^*)) \le 0 \qquad \text{a.s.}$$

Theorem 2.2 in Devroye, Györfi, and Lugosi [8] implies that

$$\begin{aligned}
&\bar{R}_n(f) - \bar{R}_n(f^*) \\
&= \frac{1}{n}\sum_{t=1}^{n}\Big(\mathbf{P}\{f_t(X_1^t, Y_1^{t-1}) \ne Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&\quad -\mathbf{P}\{f_t^*(X_{-\infty}^t, Y_{-\infty}^{t-1}) \ne Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\Big) \\
&\le 2\frac{1}{n}\sum_{t=1}^{n}\left|\mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1})\right| \\
&\le 2\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left|\mathbf{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1})\right|^2} \\
&\to 0 \qquad \text{a.s.,}
\end{aligned}$$

where in the last step we applied (14). □

References

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

[2] ——, "The strong law of large numbers for sequential decisions under uncertainity," *IEEE Transactions on Information Theory*, vol. 40, pp. 609–634, 1994.

[3] P. Auer, N. Cesa-Bianchi, and C. Gentile, "Adaptive and self-confident on-line learning algorithms," *Journal of Computer and System Sciences*, vol. 64, no. 1, pp. 48–75, 2002, a preliminary version has appeared in *Proc. 13th Ann. Conf. Computational Learning Theory*.

[4] D. H. Bailey, "Sequential schemes for classifying and predicting ergodic processes," Ph.D. dissertation, Stanford University., 1976.

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

[6] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge: Cambridge University Press, 2006.

[7] Y. S. Chow, "Local convergence of martingales and the law of large numbers," *Annals of Mathematical Statistics*, vol. 36, pp. 552–558, 1965.

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

[9] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Transactions on Information Theory*, vol. IT-38, pp. 1258–1270, 1992.

[10] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer, 2002.

[11] L. Györfi and G. Lugosi, "Strategies for sequential prediction of stationary time series," in *Modelling Uncertainty: An Examination of its Theory, Methods and Applications*, M. Dror, P. L'Ecuyer, and F. Szidarovszky, Eds. Kluwer Academic Publishers, 2001, pp. 225–248.

[12] L. Györfi, G. Lugosi, and G. Morvai, "A simple randomized algorithm for consistent sequential prediction of ergodic time series," *IEEE Transactions on Information Theory*, vol. 45, pp. 2642–2650, 1999.

[13] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and Computation*, vol. 108, pp. 212–261, 1994.

[14] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. IT-44, pp. 2124–2147, 1998.

[15] D. Modha and E. Masry, "Memory-universal prediction of stationary random preocesses," *IEEE Transactions on Information Theory*, vol. IT-44, pp. 117–133, 1998.

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

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

[18] A. Nobel, "On optimal sequential prediction for general processes," *IEEE Transactions on Information Theory*, vol. 49, pp. 83–98, 2003.

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

[20] A. C. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2685–2699, 1999.

[21] ——, "Universal linear least-squares prediction," in *Proceedings of the IEEE International Symposium on Information Theory*, 2000.

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

[23] V. Vovk, "A game of prediction with expert advice," *Journal of Computer and System Sciences*, vol. 56, pp. 153–173, 1998.

[24] T. Weissman and N. Merhav, "Universal prediction of random binary sequences in a noisy environment," *Annals of Applied Probability*, vol. 14, no. 1, pp. 54–89, Feb. 2004.

[25] Y. Yang, "Combining different procedures for adaptive regression," *Journal of Multivariate Analysis*, vol. 74, pp. 135–161, 2000.