Sequential prediction of binary sequence with side information only

György Ottucsák and László Györfi
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
H-1117 Magyar tudósok körútja 2, Budapest, Hungary
{oti,gyorfi}@szit.bme.hu

***Abstract*— A simple on-line procedure is considered for the prediction of a binary-valued sequence in the setup introduced and studied by Weissman and Merhav [13], [14], where only side information is available for the algorithm. The (non-randomized) algorithm is based on a convex combination of several simple predictors. If the side information is also binary-valued (i.e. original sequence is corrupted by a binary sequence) and both processes are realizations of stationary and ergodic random processes then the average of the loss converges, almost surely, to that of the optimum, given by the Bayes predictor. An analog result is offered for the classification of binary processes.**

I. Introduction

We study the problem of sequential prediction of a binary-valued sequence, when only side information is available for the algorithm. 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 binary numbers $y_1, y_2, \ldots$ with knowledge of side information $x_1^{t-1} = (x_1, \ldots, x_{t-1})$, where $x_t \in \{0, 1\}$. By definition x_1^0 is empty. Thus, the predictor's estimate, at time t, is based on the value of x_1^{t-1}. A prediction strategy is a sequence $g = \{g_t\}_{t=1}^{\infty}$ of functions

$$g_t : \{0, 1\}^{t-1} \to \mathbb{R}$$

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

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.

After n time instants, the *normalized cumulative loss* is

$$\begin{aligned} L_n(g) &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \ell(g_t(X_1^{t-1}), Y_t) \\ &= \frac{1}{n}\sum_{t=1}^{n} \big[(1 - Y_t)\ell(g_t(X_1^{t-1}), 0) + Y_t\ell(g_t(X_1^{t-1}), 1)\big], \end{aligned}$$

where $\ell : \mathbb{R} \times \{0, 1\} \to [0, K]$ is a bounded loss function, which is convex in its first argument. This model was introduced and studied in Weissman and Merhav [13], [14].

The case when also the past of the sequence Y_1^{t-1} is available for the predictor is well-studied. The fundamental limit for the predictability of the prediction strategy $g'_t(X_1^t, Y_1^{t-1})$ was determined by Algoet [2] who showed that for any prediction strategy g' and stationary ergodic process $\{(X_n, Y_n)\}_{-\infty}^{\infty}$,

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

where

$$\begin{aligned} \overline{L}^* = \mathbf{E}\big\{(1 - Y_0)\ell(\mathbf{E}\{Y_0 \mid X_{-\infty}^0, Y_{-\infty}^{-1}\}, 0) \\ + Y_0\ell(\mathbf{E}\{Y_0 \mid X_{-\infty}^0, Y_{-\infty}^{-1}\}, 1)\big\} \end{aligned}$$

is the minimal Bayes error of any prediction for the binary value of Y_0 based on the infinite past $X_{-\infty}^0$ and $Y_{-\infty}^{-1}$. Universally consistent strategies asymptotically achieve the best possible loss for all ergodic processes in the class. Algoet [1] and Morvai, Yakowitz and Györfi [10] proved that there exists a prediction strategy universal with respect to the class of all bounded ergodic processes. Györfi, Lugosi and Morvai [8] gave a simple universal algorithm for 0-1 loss and Györfi and Lugosi [7] introduced several simple prediction strategies, which are universally consistent for squared loss with respect to the class of bounded, stationary and ergodic processes.

However, the prediction with side information only is a delicate problem, because Y_t neither in the learning, nor in the prediction is available. In that case the fundamental limit for the predictability of the sequence can be determined as follows. Let

$$g_t^*(X_1^{t-1}) = \mathbf{E}(Y_t | X_1^{t-1})$$

be the Bayes-optimal predictor and its normalized cumulative loss is

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

Now define

$$\delta_t = \ell(g_t(X_1^{t-1}), Y_t) - \mathbf{E}\left(\ell(g_t(X_1^{t-1}), Y_t) | X_1^{t-1}\right)$$

then we can write

$$\begin{aligned} L_n(g) &= \frac{1}{n}\sum_{t=1}^{n} \delta_t + \frac{1}{n}\sum_{t=1}^{n} \mathbf{E}\left(\ell(g_t(X_1^{t-1}), Y_t) | X_1^{t-1}\right) \\ &\geq \frac{1}{n}\sum_{t=1}^{n} \delta_t + \frac{1}{n}\sum_{t=1}^{n} \mathbf{E}\left(\ell(g_t^*(X_1^{t-1}), Y_t) | X_1^{t-1}\right) \ . \end{aligned}$$

Weissman and Merhav [14, Lemma 1] proved

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

under the condition that $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$ is conditionally mixing in the sense that

$$\sum_{s=1}^{\infty} \sup_{t\geq 1} \mathbf{E}\left|\mathbf{P}\{Y_{t+s}=a|Y_t=a,X_1^{t+s-1}\}- \mathbf{P}\{Y_{t+s}=a|X_1^{t+s-1}\}\right|<\infty, \quad (2)$$

where $a\in\{0,1\}$. Therefore, we get

$$\liminf_{n\to\infty} L_n(g) \geq \liminf_{n\to\infty} L_n(g^*) = L^* \ , \quad (3)$$

with

$$L^* = \mathbf{E}\left\{(1-Y_0)\ell(\mathbf{E}\{Y_0 \mid X_{-\infty}^{-1}\},0) + Y_0\ell(\mathbf{E}\{Y_0 \mid X_{-\infty}^{-1}\},1)\right\} \ . \quad (4)$$

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,

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

Weissman and Merhav [14] introduced a universally consistent predictor for the above described setup. They used an algorithm based on Vovk [12] to combine the simple predictors and used doubling trick to fit the algorithm to infinite time horizon. In this paper we give a simple universally consistent predictor which does not use the doubling trick and the only assumption on the loss function that it is convex in its first argument. The algorithm builds on a methodology worked out in recent years for prediction of individual sequences (see e.g. Cesa-Bianchi and Lugosi [4] for a survey). We also managed to extend the result for the seemingly more difficult classification problem ($0-1$ loss).

In Section II we introduce a universally consistent strategy based on a combination of simple predictors for bounded loss function which is convex in its first argument in case of ergodic process. 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 for a Binary Memoryless Channel: general convex loss

Henceforth, we assume that the connection between Y_t and X_t are characterized by an *binary memoryless channel* as, e.g., binary symmetric channel or binary erasure channel. It means that Y_t is the input of the channel and X_t is the output of the channel, and based on the past outputs X_1^{t-1} we want to estimate input Y_t. We suppose also that the crossover probabilities of the channel are *known* for the algorithm. This assumption is indeed a realistic in many applications, where noisy medium is well-characterized statistically.

Then the algorithm is able to construct a random variable $\tilde{r}(X_t,\mathbf{C})$ which is an efficient estimate of original bit Y_t where $\mathbf{C}$ is the channel matrix:

$$\mathbf{C} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} \ ,$$

and $0\leq p,q<\frac{1}{2}$ are the crossover probabilities of the channel. More precisely, let

$$\tilde{r}(X_t,\mathbf{C}) = \frac{X_t-p}{1-p-q}$$

and it is an conditionally unbiased estimate of Y_t respect to X_1^{t-1}. Namely,

$$\begin{aligned}\mathbf{E}\{X_t|Y_t\} &= I_{\{Y_t=0\}}[(1-p)Y_t+p(1-Y_t)] \\ &\quad +I_{\{Y_t=1\}}[(1-q)Y_t+q(1-Y_t)] \\ &= I_{\{Y_t=0\}}[p(1-Y_t)]+I_{\{Y_t=1\}}[(1-q)Y_t] \\ &= p+Y_t(1-p-q)\end{aligned}$$

and therefore

$$\begin{aligned}\mathbf{E}\{\tilde{r}(X_t,\mathbf{C})|X_1^{t-1}\} &= \mathbf{E}\left\{\frac{X_t-p}{1-p-q}\middle| X_1^{t-1}\right\} \\ &= \mathbf{E}\left\{\frac{\mathbf{E}\left\{X_t|Y_t,X_1^{t-1}\right\}-p}{1-p-q}\middle| X_1^{t-1}\right\} \\ &= \mathbf{E}\left\{\frac{\mathbf{E}\left\{X_t|Y_t\right\}-p}{1-p-q}\middle| X_1^{t-1}\right\} \\ &= \mathbf{E}\{Y_t|X_1^{t-1}\} \ ,\end{aligned}$$

where the third equation follows from the memoryless property of the channel.

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

We define an infinite array of elementary predictors $h^{(k)}$, $k=1,2,\ldots$ as follows. Let $J_n^{(k)}$ be the locations of the matches of the last seen binary string x_{n-k}^{n-1} of length k in the past:

$$J_n^{(k)} = \{k<t<n : x_{t-k}^{t-1} = x_{n-k}^{n-1}\} \ .$$

Now define the elementary predictor $h^{(k)}$ by

$$h^{(k)}(x_1^{n-1}) = \tilde{r}\left(\frac{\sum_{\{t\in J_n^{(k)}\}} x_t}{|J_n^{(k)}|},\mathbf{C}\right) \ ,$$

$n>k+1$, where $0/0$ is defined to be 0. Note that $h^{(k)}(x_1^{n-1})\in\left[\frac{-p}{1-p-q};\frac{1-p}{1-p-q}\right]$.

Since, the predictor has no access to the "clean" sequence Y_t thus to measure its own performance (loss) it must use another type of the loss function based on X_t only. Define the following loss function introduced by Weissman and Merhav [13]: let $\tilde{\ell}:\mathbb{R}\times\{0,1\}\to[\frac{-pK}{1-p-q},\frac{(1-p)K}{1-p-q}]$ be the estimated loss, where K is the upper bound of $\ell(\cdot,\cdot)$. More precisely, let

$$\tilde{\ell}(h^{(k)}(X_1^{t-1}),X_t) \stackrel{\text{def}}{=} \tilde{r}(1-X_t,\mathbf{C})\ell(h^{(k)}(X_1^{t-1}),0) + \tilde{r}(X_t,\mathbf{C})\ell(h^{(k)}(X_1^{t-1}),1) \ ,$$

which is an (conditionally) unbiased estimate of the k-th expert's true loss. The cumulative estimated loss of the k-th

expert is given by

$$\tilde{L}_n(h^{(k)}) = \frac{1}{n}\sum_{t=1}^{n} \tilde{\ell}(h^{(k)}(X_1^{t-1}), X_t) \ .$$

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

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

and their normalized values

$$p_{t,k} = \frac{w_{t,k}}{W_t} \ , \quad \text{where} \quad W_t \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} w_{t,i} \ .$$

The prediction strategy g is defined by

$$g_t(x_1^{t-1}) = \sum_{k=1}^{\infty} p_{t,k} h^{(k)}(x_1^{t-1}) \ , \qquad t = 1, 2, \dots \ . \tag{5}$$

Theorem 1: Assume that $\{Y_t\}$ is stationary ergodic, and $\{X_t\}$ is the output sequence of a binary memoryless channel if $\{Y_t\}$ is the input sequence. The prediction scheme g defined above is universally consistent with respect to the class of all ergodic processes satisfying (2).

Here we describe three lemmas, which are used in the analysis. The first lemma allows us to handle the case when the number of the elementary predictors is infinite.

Lemma 1 (Györfi and Ottucsák [9]): 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

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

and the loss function ℓ is convex in its first argument h and $\ell(\cdot,\cdot) \in [0, B]$, where $B \in \mathbb{R}^+$. Define

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

with $\eta_t = \frac{1}{B\sqrt{t}}$, and

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

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 $n \geq 1$,

$$L_n(g) \leq \inf_k \left(L_n(h^{(k)}) - \frac{2B \ln q_k}{\sqrt{n}} \right) + \frac{B}{2\sqrt{n}} \ .$$

Lemma 2 (Weissman and Merhav [13], Lemma 2): If $\ell(\cdot,\cdot) \in [0, B]$ then for any predictor g

$$\limsup_{n\to\infty} \frac{\sqrt{n}|L_n(g) - \tilde{L}_n(g)|}{\sqrt{\log\log n}} \leq C(\mathbf{C}) \quad \text{a.s.},$$

where $C(\mathbf{C})$ is a deterministic constant depending on the channel matrix.

The next lemma is due to Breiman [3], and its proof may also be found in Györfi *et al.* [6].

Lemma 3 (Breiman [3]): 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

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

Proof of Theorem 1. Because of (2) we have (3), therefore it is enough to show that

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

Now we can write

$$\begin{aligned}
&\limsup_{n\to\infty} L_n(g) - L^* \\
&\quad \leq \limsup_{n\to\infty} |L_n(g) - \tilde{L}_n(g)| && (6)\\
&\qquad + \limsup_{n\to\infty} \tilde{L}_n(g) - \inf_k \limsup_{n\to\infty} \tilde{L}_n(h^{(k)}) && (7)\\
&\qquad + \inf_k \limsup_{n\to\infty} \tilde{L}_n(h^{(k)}) - \inf_k \limsup_{n\to\infty} L_n(h^{(k)}) && (8)\\
&\qquad + \inf_k \limsup_{n\to\infty} L_n(h^{(k)}) - L^*. && (9)
\end{aligned}$$

(6) and (8) goes to zero because of Lemma 2. For (7), we can apply Lemma 1 with $\overline{\ell}(\cdot,\cdot) = \tilde{\ell}(\cdot,\cdot) + \frac{pK}{1-p-q}$, where the last additive term ensures that $\overline{\ell}(\cdot,\cdot) \geq 0$. Then we have

$$\begin{aligned}
\limsup_{n\to\infty} \tilde{L}_n(g) &\leq \limsup_{n\to\infty} \inf_k \left(\tilde{L}_n(h^{(k)}) - \frac{2B\ln q_k}{\sqrt{n}} \right) \\
&\leq \inf_k \limsup_{n\to\infty} \left(\tilde{L}_n(h^{(k)}) - \frac{2B\ln q_k}{\sqrt{n}} \right) \\
&\leq \inf_k \limsup_{n\to\infty} \tilde{L}_n(h^{(k)}) \ .
\end{aligned}$$

Thus it remains to show that (9) is smaller than zero:

$$\inf_k \limsup_{n\to\infty} L_n(h^{(k)}) - L^* \leq 0 \ .$$

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

$$\begin{aligned}
h_n^{(k)}(X_1^{n-1}) &= \tilde{r}\left(\frac{\sum_{\{t \in J_n^{(k)}\}} X_t}{|J_n^{(k)}|}, \mathbf{C} \right) \\
&\to \tilde{r}\left(\mathbf{E}\{X_0 | X_{-k}^{-1}\}, \mathbf{C} \right) \\
&= \mathbf{E}\{\tilde{r}(X_0, \mathbf{C}) | X_{-k}^{-1}\} \\
&= \mathbf{E}\{Y_0 | X_{-k}^{-1}\} \ .
\end{aligned}$$

By Lemma 3, as $n \to \infty$, almost surely,

$$
\begin{aligned}
L_n(h^{(k)}) &= \frac{1}{n}\sum_{t=1}^{n}\ell(h^{(k)}(X_1^{t-1}),Y_t)\\
&\to \mathbf{E}\{\ell(\mathbf{E}\{Y_0 \mid X_{-k}^{-1}\},Y_0)\\
&= \mathbf{E}\{(1-Y_0)\ell(\mathbf{E}\{Y_0 \mid X_{-k}^{-1}\},0)\\
&\quad +Y_0\ell(\mathbf{E}\{Y_0 \mid X_{-k}^{-1}\},1)\\
&\stackrel{\text{def}}{=} \epsilon_k.
\end{aligned}
$$

Thus, the martingale convergence theorem (see, e.g., Stout [11, Theorem 2.8.6.]) implies that

$$
\begin{aligned}
\inf_k \epsilon_k &= \lim_{k\to\infty}\epsilon_k\\
&= \mathbf{E}\left\{(1-Y_0)\ell(\mathbf{E}\{Y_0 \mid X_{-\infty}^{-1}\},0)\right.\\
&\quad \left.+\, Y_0\ell(\mathbf{E}\{Y_0 \mid X_{-\infty}^{-1}\},1)\right\}\\
&= L^*
\end{aligned}
$$

as desired. □

Remark 1: (Prediction under channel uncertainty) If we assume that sometimes the algorithm has access to the original bit Y_t, then we may construct a universal consistent prediction scheme. However in a number of cases there are expensive to obtain Y_t, therefore the forecaster has the option to query this information. For query it used i.i.d. sequence $S_1, S_2, \ldots, S_n$ of Bernoulli random variables such that $\mathbf{P}\{S_t = 1\} = \epsilon$ and asks label Y_t if $S_t = 1$. Then the algorithm can construct an efficient estimate of the crossover probabilities:

$$
\tilde{p}_n = \frac{\sum_{t=1}^{n} I_{\{X_t=1,Y_t=0\}}S_t}{\sum_{t=1}^{n} I_{\{Y_t=0\}}S_t}
$$

and

$$
\tilde{q}_n = \frac{\sum_{t=1}^{n} I_{\{X_t=0,Y_t=1\}}S_t}{\sum_{t=1}^{n} I_{\{Y_t=1\}}S_t}\ ,
$$

where $\tilde{p}_n \to p$ and $\tilde{q}_n \to q$. Now using these estimates in $\tilde{\ell}(\cdot,\cdot)$ and $\tilde{r}(\cdot,\cdot)$ we obtain a universal prediction scheme. The above described situation appears when the algorithm is supported by a human expert or we have a second no noisy-channel. For example, in case of natural language processing (e.g. 8 bits represent a character), the human observer select the best possible reconstruction, which e.g, it can be found in the "dictionary" and fits in with the context.

III. Universal prediction for a Binary Memoryless channel: Zero-One loss

In this section we apply the same ideas to the seemingly more difficult classification (or pattern recognition) problem. 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

$$
f_t : \{0,1\}^t \to \{0,1\}
$$

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

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

(2) implies (3) such that

$$
\liminf_{n\to\infty} R_n(f) \geq R^* \tag{10}
$$

where

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

Consider the prediction scheme $g_t(X_1^{t-1})$ with squared loss $\ell(x,y) = (x-y)^2$, introduced in the previous section, and then introduce the corresponding classification scheme:

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

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

Theorem 2: Assume that $\{Y_t\}$ is stationary ergodic, and $\{X_t\}$ is the output sequence of a binary memoryless channel if $\{Y_t\}$ is the input sequence. 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 $\{(X_n,Y_n)\}_{n=-\infty}^{\infty}$ satisfying (2).

For the proof of Theorem 2 we need the following corollary of Theorem 1.

Corollary 1: Under the conditions of Theorem 1,

$$
\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{E}\{Y_t \mid X_{-\infty}^{t-1}\} - g_t(X_1^{t-1})\right)^2 = 0 \quad \text{a.s.} \tag{11}
$$

where g_t is the predictor for squared loss $\ell(x,y) = (x-y)^2$.

Proof. The ergodic theorem implies that

$$
\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}\left\{\left(Y_t - \mathbf{E}\{Y_t \mid X_{-\infty}^{t-1}\}\right)^2 \Big| X_{-\infty}^{t-1}\right\} = L^* \qquad \text{a.s.}
$$

and note that

$$
\begin{aligned}
&\mathbf{E}\left\{ \left(Y_t - g_t(X_1^{t-1})\right)^2 \big| X_{-\infty}^{t-1}\right\}\\
&\quad = \mathbf{E}\{\left(Y_t - \mathbf{E}\{Y_t \mid X_{-\infty}^{t-1}\}\right)^2 \mid X_{-\infty}^{t-1}\}\\
&\qquad + \left(\mathbf{E}\{Y_t \mid X_{-\infty}^{t-1}\} - g_t(X_1^{t-1})\right)^2,
\end{aligned}
$$

therefore in order to finish the proof it suffices to show

$$
\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}\left\{\left(Y_t - g_t(X_1^{t-1})\right)^2 \Big| X_{-\infty}^{t-1}\right\} = L^* \qquad \text{a.s.} \tag{12}
$$

By Theorem 1 with squared loss, we have

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

Thus, for (12), we have to prove that

$$\frac{1}{n}\sum_{t=1}^{n}\Big(\big(Y_t-g_t(X_1^{t-1})\big)^2-\mathbf{E}\{\big(Y_t-g_t(X_1^{t-1})\big)^2\mid X_{-\infty}^{t-1}\}\Big)$$
$$=\frac{1}{n}\sum_{t=1}^{n}\big(Y_t^2-\mathbf{E}\{Y_t^2\mid X_{-\infty}^{t-1}\}\big)$$
$$-2\frac{1}{n}\sum_{t=1}^{n}g_t(X_1^{t-1})(Y_t-\mathbf{E}\{Y_t\mid X_{-\infty}^{t-1}\})\to 0\qquad\text{a.s.}$$

Because of assumption (2)

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

and the corollary is proved. □

Proof of Theorem 2. Because of (10) we have to show that

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

Introduce the Bayes classification scheme using the infinite past:

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

and its normalized cumulative $0-1$ loss:

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

Put

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

and

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

Because of assumption (2) we have

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

and

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

moreover, by the Breiman 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 [5] implies that

$$\bar{R}_n(f)-\bar{R}_n(f^*)\;=\;\frac{1}{n}\sum_{t=1}^{n}\Big(\mathbf{P}\{f_t(X_1^{t-1})\neq Y_t\mid X_{-\infty}^{t-1}\}$$
$$-\mathbf{P}\{f_t^*(X_{-\infty}^{t-1})\neq Y_t\mid X_{-\infty}^{t-1}\}\Big)$$
$$\le\;2\frac{1}{n}\sum_{t=1}^{n}\big|\mathbf{E}\{Y_t\mid X_{-\infty}^{t-1}\}-g_t(X_1^{t-1})\big|$$
$$\le\;2\sqrt{\frac{1}{n}\sum_{t=1}^{n}\big|\mathbf{E}\{Y_t\mid X_{-\infty}^{t-1}\}-g_t(X_1^{t-1})\big|^2}$$
$$\to 0\qquad\text{a.s.,}$$

where in the last step we applied the result of Corollary 1.□

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] 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.

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

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

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

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

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

[9] L. Györfi and Gy. Ottucsák. Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory (to appear)*, 53, 2007.

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

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

[12] V. Vovk. A game of prediction with expert advice. In *Proc. Third Annual Workshop on COLT*, pages 371–383, San Mateo,CA, 1995. Kaufmann.

[13] T. Weissman and N. Merhav. Universal prediction of binary individual sequences in the presence of noise. *IEEE Trans. Inform. Theory*, 47(6):2151–2173, July 2001.

[14] T. Weissman and N. Merhav. Universal prediction of random binary sequences in a noisy environment. *Annals of Applied Probability*, 14(1):54–89, Feb. 2004.