Strategies for sequential prediction of stationary time series

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

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

February 13, 2001

Dedicated to the memory of Sid Yakowitz.

Abstract

We present simple procedures for the prediction of a real valued sequence. The algorithms are based on a combination of several simple predictors. We show that if the sequence is a realization of a bounded stationary and ergodic random process then the average of squared errors converges, almost surely, to that of the optimum, given by the Bayes predictor. We offer an analog result for the prediction of stationary gaussian processes.

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

1 Introduction

One of the many themes of Sid's research was the search for prediction and estimation methods for time series that do not necessarily satisfy the classical assumptions for autoregressive markovian and gaussian processes (see, e.g., [18, 19, 27, 28, 29, 30, 31, 32]). He firmly believed that most real-world applications require such robust methods. This note is a contribution to the line of research pursued and promoted by Sid who directed us to this beautiful area of research.

We study the problem of sequential prediction of a real valued sequence. At each time instant $i = 1, 2, \ldots$, the predictor is asked to guess the value of the next outcome y_i of a sequence of real numbers $y_1, y_2, \ldots$ 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 estimate, at time i, is based on the value of y_1^{i-1}. Formally, the strategy of the predictor is a sequence $g = \{g_i\}_{i=1}^{\infty}$ of decision functions

$$g_i : \mathbb{R}^{i-1} \to \mathbb{R}$$

and the prediction formed at time i is $g_i(y_1^{i-1})$. After n time instants, the *normalized cumulative prediction error* on the string y_1^n is

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} (g_i(y_1^{i-1}) - y_i)^2.$$

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 real valued stationary and ergodic process $\{Y_n\}_{-\infty}^{\infty}$. 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 $\{Y_n\}_{-\infty}^{\infty}$,

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

where

$$L^* = \mathbf{E}\left[(Y_0 - \mathbf{E}\{Y_0 | Y_{-\infty}^{-1}\})^2\right]$$

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past $Y_{-\infty}^{-1} = (\ldots, Y_{-3}, Y_{-2}, Y_{-1})$. Note that it follows by stationarity and the martingale convergence theorem (see, e.g., Stout [25]) that $L^* = \lim_{n\to\infty} \mathbf{E}\left[(Y_n - \mathbf{E}\{Y_n | Y_1^{n-1}\})^2\right]$.
This lower bound gives sense to the following definition:

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

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

Universal strategies asymptotically achieve the best possible loss for all ergodic processes in the class. Algoet [1] and Morvai, Yakowitz, Györfi [18] 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.

The purpose of this paper is to introduce several simple prediction strategies which, apart from having the above mentioned universal property of [1] and [18], promise much improved performance for "nice" processes. The algorithms build on a methodology worked out in recent years for prediction of individual sequences, see Vovk [33], Feder, Merhav, and Gutman [8], Littlestone and Warmuth [16], Cesa-Bianchi et al. [5], Kivinen and Warmuth [15], Singer and Feder [23], and Merhav and Feder [17] for a survey.

An approach similar to the one of this paper was adopted by Györfi, Lugosi, and Morvai [14], where prediction of stationary binary sequences was addressed. There we introduced a simple randomized predictor which predicts asymptotically as well as the optimal predictor for all binary ergodic processes. The present setup and results differ in several important points from those of [14]. On the one hand, special properties of the squared loss function considered here allow us to avoid randomization of the predictor, and to define a significantly simpler prediction scheme. On the other hand, possible unboundedness of a real-valued process requires special care, which we demonstrate on the example of gaussian processes. We refer to Nobel [20], Singer and Feder [23], [24], Yang [34], [35] to recent closely related work.

In Section 2 we introduce a universal strategy for bounded ergodic processes which is based on a combination of partitioning estimates. In Section 3, still for bounded processes, we consider, as an alternative, a prediction strategy based on combining generalized linear estimates. In Section 4 we replace the boundedness assumption by assuming that the sequence to predict is an ergodic gaussian process, and show how the techniques of Section 3 may be modified to take care of the difficulties originating in the unboundedness of the process.

The results of the paper are given in an autoregressive framework, that is, the value Y_t is to be predicted based on past observations Y_1^{t-1} of the same process. We may also consider the more general situation when Y_t is predicted based on Y_1^{t-1} and X_1^t, where $\{X_n\}_{-\infty}^{\infty}$ is an $\mathbb{R}^d$-valued process such that $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ is a jointly stationary and ergodic process. The prediction problem is similar to the one defined above 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. Formally, now a prediction strategy is a sequence $g = \{g_i\}_{i=1}^{\infty}$ of functions

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

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

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} (g_i(y_1^{i-1}, x_1^i) - y_i)^2,$$

All results of the paper may be extended, in a straightforward manner to this more general prediction problem. As the extension does not require new ideas, we omit the details.

Another direction for generalizing the results is to consider predicting vector-valued processes. Once again, the extension to $\mathbb{R}^d$-valued processes $\{Y_n\}_{-\infty}^{\infty}$ is obvious, and the details are omitted.

2 Universal prediction by partitioning estimates

In this section we introduce our first prediction strategy for bounded ergodic processes. We assume throughout the section that $|Y_0|$ is bounded by a constant $B > 0$, with probability one. First we assume that the bound B is known. The case of unknown B will be treated later in a remark.

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 $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ 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 $\mathbb{R}$, 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 $y_1^n \in \mathbb{R}^n$, we write $G_\ell(y_1^n)$ for the sequence $G_\ell(y_1), \ldots, G_\ell(y_n)$. Fix positive integers k, ℓ, and for each k-long string s of positive integers, define the partitioning regression function estimate

$$\widehat{E}_n^{(k,\ell)}(y_1^{n-1}, s) = \frac{\sum_{\{k<i<n : G_\ell(y_{i-k}^{i-1}) = s\}} y_i}{\left|\{k < i < n : G_\ell(y_{i-k}^{i-1}) = s\}\right|}, \qquad n > k+1,$$

where $0/0$ is defined to be 0.

Now we define the elementary predictor $h^{(k,\ell)}$ by

$$h_n^{(k,\ell)}(y_1^{n-1}) = \widehat{E}_n^{(k,\ell)}(y_1^{n-1}, G_\ell(y_{n-k}^{n-1})), \qquad n = 1, 2, \ldots$$

That is, $h_n^{(k,\ell)}$ quantizes the sequence y_1^{n-1} according to the partition $\mathcal{P}_\ell$, and looks for all appearances of the last seen quantized strings $G_\ell(y_{n-k}^{n-1})$ of length k in the past. Then it predicts according to the average of the y_i'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, $q_{k,\ell} > 0$. Put $c = 8B^2$, and define the weights

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

and their normalized values

$$v_{t,k,\ell} = \frac{w_{t,k,\ell}}{\sum_{i,j=1}^{\infty} w_{t,i,j}} \ .$$

The prediction strategy g is defined by

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

Theorem 1 *Assume that*
(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) 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

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

Then the prediction scheme g defined above is universal with respect to the class of all ergodic processes such that $\mathbf{P}\{Y_i \in [-B, B]\} = 1$.

One of the main ingredients of the proof is the following lemma, whose proof is a straightforward extension of standard arguments in the prediction theory of individual sequences, see, for example, Kivinen and Warmuth [15], Singer and Feder [24].

Lemma 1 *Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $\tilde{h}_i(y_1^{n-1}) \in [-B, B]$ and $y_1^n \in [-B, B]^n$. Define*

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \geq 8B^2$, and

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

If the prediction strategy $\tilde{g}$ is defined by

$$\tilde{g}_t(y_1^{t-1}) = \sum_{k=1}^{\infty} v_{t,k} \tilde{h}_k(y_1^{t-1}) \qquad t = 1, 2, \ldots$$

then for every $n \geq 1$,

$$L_n(\tilde{g}) \leq \inf_k \left(L_n(\tilde{h}_k) - \frac{c \ln q_k}{n} \right).$$

Here $-\ln 0$ is treated as ∞.

Proof. Introduce $W_1 = 1$ and $W_t = \sum_{k=1}^{\infty} w_{t,k}$ for $t > 1$. First we show that for each $t > 1$,

$$\left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(y_1^{t-1}) \right) \right]^2 \leq -c \ln \frac{W_{t+1}}{W_t} \tag{2}$$

Note that

$$W_{t+1} = \sum_{k=1}^{\infty} w_{t,k} e^{-\left(y_t - \tilde{h}_k(y_1^{t-1})\right)^2/c} = W_t \sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(y_1^{t-1})\right)^2/c},$$

so that

$$-c \ln \frac{W_{t+1}}{W_t} = -c \ln \left(\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(y_1^{t-1})\right)^2/c} \right).$$

Therefore, (2) becomes

$$\exp\left(\frac{-1}{c} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(y_1^{t-1}) \right) \right]^2 \right) \geq \sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(y_1^{t-1})\right)^2/c},$$

which is implied by Jensen's inequality and the concavity of the function

$F_t(z) = e^{-(y_t - z)^2/c}$ for $c \geq 8B^2$. Thus, (2) implies that

$$\begin{aligned}
nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(y_1^{t-1})\right)^2 \\
&= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(y_1^{t-1})\right)\right]^2 \\
&\leq -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\
&= -c \ln W_{n+1} \\
&= -c \ln \left(\sum_{k=1}^{\infty} w_{n+1,k}\right) \\
&= -c \ln \left(\sum_{k=1}^{\infty} q_k e^{-nL_n(\tilde{h}_k)/c}\right) \\
&\leq -c \ln \left(\sup_k q_k e^{-nL_n(\tilde{h}_k)/c}\right) \\
&= \inf_k \left(-c \ln q_k + nL_n(\tilde{h}_k)\right) \ ,
\end{aligned}$$

which concludes the proof. $\square$

Another main ingredient of the proof of Theorem 1 is known as Breiman's generalized ergodic theorem [4], see also Algoet [2].

Lemma 2 (Breiman [4]). *Let $Z = \{Z_i\}_{-\infty}^{\infty}$ be a stationary and ergodic process. 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*

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

Proof of Theorem 1. By a double application of the ergodic theorem, as

$n \to \infty$, almost surely,

$$\begin{aligned}
\widehat{E}_n^{(k,\ell)}(Y_1^{n-1}, s) &= \frac{\frac{1}{n}\sum_{\{k<i<n: G_\ell(Y_{i-k}^{i-1})=s\}} Y_i}{\frac{1}{n}\left|\{k<i<n : G_\ell(Y_{i-k}^{i-1}) = s\}\right|} \\
&\to \frac{\mathbf{E}\{Y_0 I_{\{G_\ell(Y_{-k}^{-1})=s\}}\}}{\mathbf{P}\{G_\ell(Y_{-k}^{-1}) = s,\}} \\
&= \mathbf{E}\{Y_0 | G_\ell(Y_{-k}^{-1}) = s\},
\end{aligned}$$

and therefore

$$\lim_{n\to\infty} \sup_s |\widehat{E}_n^{(k,\ell)}(Y_1^{n-1}, s) - \mathbf{E}\{Y_0|G_\ell(Y_{-k}^{-1}) = s\}| = 0 \qquad \text{almost surely.}$$

Thus, by Lemma 2, as $n \to \infty$, almost surely,

$$\begin{aligned}
L_n(h^{(k,\ell)}) &= \frac{1}{n}\sum_{i=1}^{n}(h^{(k,\ell)}(Y_1^{i-1}) - Y_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\widehat{E}_n^{(k,\ell)}(Y_1^{i-1}, Y_{i-k}^{i-1}) - Y_i)^2 \\
&\to \mathbf{E}\{(Y_0 - \mathbf{E}\{Y_0|G_\ell(Y_{-k}^{-1})\})^2\} \\
&\stackrel{\text{def}}{=} \epsilon_{k,\ell}.
\end{aligned}$$

Since the partitions $\mathcal{P}_\ell$ are nested, $\mathbf{E}\left\{Y_0|\mathcal{G}_\ell(Y_{-k}^{-1})\right\}$ is a martingale indexed by the pair (k, ℓ). Thus, the martingale convergence theorem (see, e.g., Stout [25]) and assumption (b) for the sequence of partitions implies that

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

Now by Lemma 1,

$$L_n(g) \le \inf_{k,\ell}\left(L_n(h^{(k,\ell)}) - \frac{c\ln q_{k,\ell}}{n}\right), \tag{3}$$

and therefore, almost surely,

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

and the proof of the theorem is finished. □

Theorem 1 shows that asymptotically, the predictor g_t defined by (1) predicts as well as the optimal predictor given by the regression function $\mathbf{E}\{Y_t|Y_{-\infty}^{t-1}\}$. In fact, G_t gives a good estimate of the regression function in the following sense:

Corollary 1 *Under the conditions of Theorem 1*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\} - g_i(Y_1^{i-1})\right)^2 = 0 \qquad \textit{almost surely.}$$

Proof. By Theorem 1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g_i(Y_1^{i-1})\right)^2 = L^* \qquad \text{almost surely.}$$

Consider the following decomposition:

$$\begin{aligned}
\left(Y_i - g_i(Y_1^{i-1})\right)^2 &= \left(Y_i - \mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}\right)^2 \\
&\quad +2\left(Y_i - \mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}\right)\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\} - g_i(Y_1^{i-1})\right) \\
&\quad +\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\} - g_i(Y_1^{i-1})\right)^2.
\end{aligned}$$

Then the ergodic theorem implies that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}\right)^2 = L^* \qquad \text{almost surely.}$$

It remains to show that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}\right)\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}-g_i(Y_1^{i-1})\right)=0 \qquad \text{almost surely.} \tag{4}$$

But this is a straightforward consequence of Kolmogorov's classical strong law of large numbers for martingale differences due to Chow [6] (see also Stout [25, Theorem 3.3.1]). It states that if $\{Z_i\}$ is a martingale difference sequence with

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

then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Z_i=0 \qquad \text{almost surely.}$$

Thus, (4) is implied by Chow's theorem since the martingale differences $Z_i = \left(Y_i-\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}\right)\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}-g_i(Y_1^{i-1})\right)$ are bounded by $4B^2$. (To see that the Z_i's indeed form a martingale difference sequence just note that $\mathbf{E}\{Z_i|Y_{-\infty}^{i-1}\}=0$ for all i.) $\square$

Remark. UNKNOWN B. The prediction strategy studied in this section may be easily extended to the case when the process $\{Y_n\}_{-\infty}^{\infty}$ is bounded, but B is unknown, that is, when no upper bound is known to the range of the process. In such a case we may simply start with the hypotheses $B=1$ and predict according to (1) until we find a value Y_n with $|Y_n|>B$. Then we reset the algorithm and start the predictor again but with doubling the value of the previous B, and keep doing this. Then the universal property of Theorem 1 obviously remains valid to this modified strategy.

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 $-\ln q_{k,\ell}/n$ 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.

Remark. SEQUENTIAL GUESSING. If the process takes values from a finite set, one is often interested in the sequential guessing of Y_n upon observing the past Y_1^{n-1}. Such a problem was investigated (among others) by Györfi, Lugosi, and Morvai in [14], where it was assumed that Y_n takes one of two values: $Y_n \in \{0,1\}$. Sequential guessing is then formally defined by a sequence $f = \{f_i\}_{i=1}^{\infty}$ of decision functions

$$f_i : \{0,1\}^{i-1} \to \{0,1\}$$

and the guess formed at time i is $f_i(Y_1^{i-1})$. The normalized cumulative loss of guessing by f on the string Y_1^n is

$$R_1^n(f) = \frac{1}{n}\sum_{i=1}^{n} I_{\{f_i(Y_1^{i-1})\neq Y_i\}},$$

where I denotes the indicator function. Algoet [2] showed that for any guessing strategy f and stationary ergodic binary process $\liminf_{n\to\infty} R_1^n(f) \geq R^*$ almost surely, where

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

is the minimal expected probability of error of guessing Y_0 based on the infinite past $Y_{-\infty}^{-1}$. The existence of a guessing scheme f for which $\lim_{n\to\infty} R_1^n(f) = R^*$ almost surely follows from results of Onrstein [21] and Bailey [3]. In [14] a simple guessing procedure was proposed with the same asymptotic guarantees and with a good finite-sample behavior for Markov processes. The disadvantage of the predictor given in [14] is that it requires randomization. Here we observe that with the help of predictors having the universal convergence property of Theorem 1 we may easily define a nonrandomized guessing scheme with the desired convergence properties. Given a prediction strategy

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

for a binary process $\{Y_n\}$, we simply define the guessing scheme f by the decision functions

$$f_i(y_1^{i-1}) = \begin{cases} 1 & \text{if } g_i(y_1^{i-1}) \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Then we may use the properties of g established in Corollary 1 to conclude that the guessing scheme g defined above has an average number of mistakes $R_1^n(f)$ converging to the optimum R^* almost surely. Indeed, if we define the decision f_i^* based on observing the infinite past $y_{-\infty}^{i-1}$ as the one minimizing the probability of error of guessing Y_i:

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

then we may write

$$\begin{aligned}
&\limsup_{n\to\infty} R_1^n(f) - R^* \\
&= \limsup_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} I_{\{f_i(Y_1^{i-1}) \neq Y_i\}} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{P}\left\{ f_i^*(Y_{-\infty}^{i-1}) \neq Y_i | Y_{-\infty}^{i-1} \right\} \right) \\
&\qquad \text{(by the ergodic theorem)} \\
&= \limsup_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{P}\left\{ f_i(Y_1^{i-1}) \neq Y_i | Y_{-\infty}^{i-1} \right\} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{P}\left\{ f_i^*(Y_{-\infty}^{i-1}) \neq Y_i | Y_{-\infty}^{i-1} \right\} \right) \\
&\qquad \text{(by the martingale convergence theorem)} \\
&\leq \limsup_{n\to\infty} \frac{2}{n} \sum_{i=1}^{n} \left| \mathbf{E}\{Y_i | Y_{-\infty}^{i-1}\} - g_i(Y_1^{i-1}) \right| \\
&\qquad \text{(by Theorem 2.2 in [7])} \\
&\leq \limsup_{n\to\infty} 2 \sqrt{ \frac{1}{n} \sum_{i=1}^{n} \left| \mathbf{E}\{Y_i | Y_{-\infty}^{i-1}\} - g_i(Y_1^{i-1}) \right|^2 } \\
&\to 0 \qquad \text{(by Corollary 1.)}
\end{aligned}$$

Thus, any predictor with the universal property established in Theorem 1 may be converted, in a natural way, into a universal guessing scheme. An alternative proof of the same fact is given by Nobel [20].

3 Universal prediction by generalized linear estimates

This section is devoted to an alternative way of defining a universal predictor for the class of all bounded ergodic processes. Once again, we apply the method described by Lemma 1 to combine elementary predictors, but now, instead of partitioning-based predictors, we use elementary predictors which are generalized linear predictors. Once again, we consider bounded processes, and assume that a positive constant B is known such that $\mathbf{P}\{|Y_0| \leq B\} = 1$. (The case of unknown B may be treated similarly as in Section 2.)

We define an infinite array of elementary predictors $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ as follows. Let $\{\phi_j^{(k)}\}_{j=1}^{\ell}$ be real-valued functions defined on $\mathbb{R}^k$. The elementary predictor $h^{(k,\ell)}$ generates a prediction of form

$$\tilde{h}^{(k,\ell)}(Y_1^{n-1}) = \sum_{j=1}^{l} c_{n,j} \phi_j^{(k)}(Y_{n-k}^{n-1})$$

such that the coefficients $c_{n,j}$ are calculated based on the past observations Y_1^{n-1}. Before defining the coefficients, note that one is tempted to define the $c_{n,i}$'s as the coefficients which minimize

$$\sum_{i=k+1}^{n-1} \left(Y_i - \sum_{j=1}^{\ell} c_j \phi_j^{(k)}(Y_{i-k}^{i-1}) \right)^2$$

if $n > k$, and the all-zero vector otherwise. However, even though the minimum always exists, it is not unique in general, and therefore the minimum is not well-defined. Instead, we define the coefficients by a standard recursive procedure as follows (see, e.g., Tsypkin [26], Györfi [13], Singer and Feder [24]). Introduce

$$X_i = (\phi_1^{(k)}(Y_{i-k}^{i-1}), \ldots \phi_\ell^{(k)}(Y_{i-k}^{i-1}))^T$$

(where the superscript T denotes transpose),

$$A_n = \sum_{i=k+1}^{n-1} X_i X_i^T \ , \quad \text{and} \quad M_n = \sum_{i=k+1}^{n-1} Y_i X_i.$$

Let σ be an arbitrary positive constant and put $B_n = A_n + \sigma I$, where I is the $\ell \times \ell$ identity matrix. Define

$$c_n = (c_{n,1}, \dots, c_{n,\ell}) = (A_n + \sigma I)^{-1} M_n = B_n^{-1} M_n,$$

It easy to see that the inverse can be calculated recursively by

$$B_{i+1}^{-1} = B_i^{-1} - \frac{B_i^{-1} X_i X_i^T B_i^{-1}}{1 + X_i^T B_i^{-1} X_i}\ , \quad i = 2, 3, \dots, n-1\ ,$$

which makes the calculation of c_n easy.

Theorem 2 *Define*

$$h_n^{(k,\ell)}(Y_1^{n-1}) = \begin{cases} B & \textit{if } \tilde{h}^{(k,\ell)}(Y_1^{n-1}) > B \\ -B & \textit{if } \tilde{h}^{(k,\ell)}(Y_1^{n-1}) < -B \\ \tilde{h}^{(k,\ell)}(Y_1^{n-1}) & \textit{otherwise} \end{cases} \qquad n = 1, 2, \dots .$$

Suppose $|\phi_j^{(k)}| \le 1$ *and for any fixed* k *the set*

$$\left\{ \sum_{j=1}^{\ell} c_j \phi_j^{(k)}; \quad (c_1, \dots, c_\ell),\ \ell = 1, 2, \dots \right\}$$

is dense in $C([-B,B]^k)$*. Define a prediction strategy by combining the elementary predictors* $h^{(k,\ell)}$ *given by (1). The obtained predictor is universal with respect to the class of all ergodic processes with* $\mathbf{P}\{Y_i \in [-B,B]\} = 1$*.*

Proof. Let $\overline{A}_n = \mathbf{E} A_n$, and let $(\lambda_{n,j}, \varphi_{n,j})$ $j = 1, 2, \dots, \ell$ be an eigensystem of $\overline{A}_n$ that is, $\{\varphi_{n,j}\}$ are orthogonal solutions of the equation

$$\overline{A}_n \varphi_{n,j} = \lambda_{n,j} \varphi_{n,j}$$

and

$$\lambda_{n,1} \ge \cdots \ge \lambda_{n,\ell} \ge 0.$$

(Note that since $\overline{A}_n$ is symmetric and positive semidefinite, its eigenvalues are all real and nonnegative.) Let $0 \le \ell' \le \ell$ be the integer for which $\lambda_{n,j} > 0$ if $j \le \ell'$ and $\lambda_{n,j} = 0$ if $j > \ell'$. Express the vector $\overline{M}_n = \mathbf{E} M_n$ as

$$\overline{M}_n = \sum_{i=1}^{\ell} u_j \varphi_{n,j},$$

and define

$$c^* = (c_1^*, \ldots, c_\ell^*) = \sum_{j=1}^{\ell'} \frac{u_j}{\lambda_{n,j}} \varphi_{n,j}.$$

(It is easy to see by stationarity that the value of the vector c^* is independent of n.) It is shown by Györfi [13] that

$$\mathbf{E}\left\{\left(Y_0 - \sum_{j=1}^{\ell} c_j^* \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\} = \min_{(c_1,\ldots,c_\ell)} \mathbf{E}\left\{\left(Y_0 - \sum_{j=1}^{\ell} c_j \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\}$$

and moreover

$$\lim_{n\to\infty}(c_{n,1}, \ldots, c_{n,\ell}) = (c_1^*, \ldots, c_\ell^*) \qquad \text{almost surely.} \tag{5}$$

Also, observe that by the ergodic theorem, for any fixed $(c_1, \ldots, c_\ell)$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=k+1}^{n} \left(Y_i - \sum_{j=1}^{\ell} c_j \phi_j^{(k)}(Y_{i-k}^{i-1})\right)^2 = \mathbf{E}\left\{\left(Y_0 - \sum_{j=1}^{\ell} c_j \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\} \tag{6}$$
$$\text{almost surely.}$$

Therefore by (6), (5) and Lemma 2

$$\begin{aligned}
\lim_{n\to\infty} L_n(h^{(k,\ell)}) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=k+1}^{n} \left(Y_i - h^{(k,\ell)}(Y_1^{i-1})\right)^2 \\
&\le \lim_{n\to\infty} \frac{1}{n} \sum_{i=k+1}^{n} \left(Y_i - \tilde{h}^{(k,\ell)}(Y_1^{i-1})\right)^2 \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=k+1}^{n} \left(Y_i - \sum_{j=1}^{\ell} c_{i,j} \phi_j^{(k)}(Y_{i-k}^{i-1})\right)^2 \\
&= \mathbf{E}\left\{\left(Y_0 - \sum_{j=1}^{\ell} c_j^* \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\} \quad \text{almost surely} \\
&\stackrel{\text{def}}{=} \epsilon_{k,\ell}.
\end{aligned}$$

Next define the coefficient vector $\tilde{c}_k = (\tilde{c}_{k,1}, \ldots, \tilde{c}_{k,\ell})$ to be any vector which achieves the minimum in

$$\min_{(c_1,\ldots,c_\ell)} \mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-k}^{-1}\} - \sum_{j=1}^{\ell} c_j \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\}$$

Then

$$\begin{aligned}
\epsilon_{k,\ell} &= \mathbf{E}\left\{\left(Y_0 - \mathbf{E}\{Y_0|Y_{-\infty}^{-1}\}\right)^2\right\} + \mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-\infty}^{-1}\} - \sum_{j=1}^{\ell} c_j^* \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\} \\
&\le L^* + \mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-\infty}^{-1}\} - \sum_{j=1}^{\ell} \tilde{c}_{k,j} \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\} \\
&\le L^* + 2\mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-\infty}^{-1}\} - \mathbf{E}\{Y_0|Y_{-k}^{-1}\}\right)^2\right\} \\
&\quad + 2\mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-k}^{-1}\} - \sum_{j=1}^{\ell} \tilde{c}_{k,j} \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\} \\
&\stackrel{\text{def}}{=} L^* + 2\delta_k + 2\gamma_{k,\ell}.
\end{aligned}$$

It is immediate by the martingale convergence theorem that

$$\lim_{k\to\infty} \delta_k = 0.$$

On the other hand, by the denseness assumption of the theorem, for any fixed k,

$$\lim_{\ell\to\infty} \gamma_{k,\ell} = \lim_{\ell\to\infty} \mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-k}^{-1}\} - \sum_{j=1}^{\ell} \tilde{c}_{k,j} \phi_j^{(k)}(Y_{-k}^{-1})\right)^2\right\} = 0.$$

Thus, we conclude that

$$\inf_{k,\ell} \epsilon_{k,\ell} = \lim_{k,\ell\to\infty} \epsilon_{k,\ell} = L^*.$$

Finally, by Lemma 1,

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

which concludes the proof. $\square$

Again, as in Corollary 1, we may compare the predictor directly to the regression function. By the same argument, we obtain the following result. The details are left to the reader.

Corollary 2 *Under the conditions of Theorem 2*

$$
\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{E}\{Y_i | Y_{-\infty}^{i-1}\} - g_i(Y_1^{i-1}) \right)^2 = 0 \qquad \textit{almost surely.}
$$

4 Prediction of Gaussian processes

Up to this point we have always assumed that the process to predict is bounded. This excludes some traditionally well-studied unbounded processes such as gaussian processes. In this section we define a predictor which is universal with respect to the class of all stationary and ergodic gaussian processes. For gaussian processes the best predictor (i.e., the regression function) is linear, and therefore we may use the techniques of the previous section in the special case when $\phi_j^{(k)}(y_1^k) = y_j$. However, the unboundedness of the process introduces some additional difficulty. To handle it, we use bounded elementary predictors as before, but the bound is increased with n. Also, we need to modify the way of combining these elementary predictors.

The proposed predictor is based on a convex combination of linear predictors of different orders. For each $k = 1, 2, \ldots$ introduce

$$h^{(k)}(y_1^{n-1}) = \sum_{j=1}^{k} c_{n,j} y_{n-j} \ ,$$

where the vector $c_n = (c_{n,1}, \ldots, c_{n,k})$ of coefficients is calculated by the formula introduced in Section 3:

$$c_n = (A_n + \sigma I)^{-1} M_n \ ,$$

where σ is a positive number, $A_n = \sum_{i=k+1}^{n-1} X_i X_i^T$ and $M_n = \sum_{i=k+1}^{n-1} Y_i X_i$ with $X_i = Y_{i-k}^{i-1}$ for $i > k$.

Introduce the notation

$$L_n^N(g) = \frac{1}{N-n} \sum_{i=n+1}^{N} (g_i(y_1^{i-1}) - y_i)^2 \ .$$

Then the predictor g is defined as follows: for all $m = 0, 1, 2, \ldots$, if n is such that $2^m \le n < 2^{m+1}$, then

$$g_n(y_1^{n-1}) = \begin{cases} \sum_{k=1}^{\infty} v_{n,k}^{(m)} h^{(k)}(y_1^{n-1}) & \text{if } \left| \sum_{k=1}^{\infty} v_{n,k}^{(m)} h^{(k)}(y_1^{n-1}) \right| \le \ln(n) \\ \ln(n) \operatorname{sgn}\left(\sum_{k=1}^{\infty} v_{n,k}^{(m)} h^{(k)}(y_1^{n-1}) \right) & \text{otherwise.} \end{cases}$$

where

$$v_{n,k}^{(m)} = \frac{w_{n,k}^{(m)}}{\sum_{k=1}^{\infty} w_{n,k}^{(m)}} \quad \text{and} \quad w_{n,k}^{(m)} = q_{k,m} e^{-(n-1-2^m) L_{2^m}^{n-1}(h^{(k)})/(2m^2)} \ ,$$

with $q_{k,m} = 2^{-(m+1)} I_{k \le 2^{m+1}}$.

Thus, we divide the time instances into intervals of exponentially increasing length and, after initializing the predictor at the beginning of such an interval, we use a different way of combining the elementary predictors $h^{(k)}$ in each such segment. The reason for this is that to be able to combine elementary predictors as in Lemma 1, we need to make sure that the predictor as well as the outcome to predict is appropriately bounded. In our case this can be achieved based on Lemma 3 below which implies that with very large probability, the maximum of n identically distributed normal random variables is at most of the order of $\sqrt{\log n}$.

Theorem 3 *The prediction strategy g defined above is universal with respect to the class of all stationary and ergodic zero-mean gaussian processes. Also,*

$$L_n(g) \le \inf_{k \ge 1} L_n(h^{(k)}) + \frac{2(\log_2 n + 1)^3}{n} + O\left(\frac{1}{n}\right) \qquad \textit{almost surely.}$$

At a key point the proof uses the following well-known properties of gaussian random variables:

Lemma 3 *(Pisier [22])). Let $Z_1, \dots, Z_n$ be zero-mean gaussian random variables with $\mathbf{E}\{Z_i^2\} = \sigma^2$, $i = 1, \dots, n$. Then*

$$\mathbf{E}\left\{\max_{i \le n} |Z_i|\right\} \le \sigma\sqrt{2\ln(2n)},$$

and for each $u > 0$,

$$\mathbf{P}\left\{\max_{i \le n} |Z_i| - \mathbf{E}\left\{\max_{i \le n} |Z_i|\right\} > u\right\} \le e^{-u^2/2\sigma^2}.$$

Proof of Theorem 3. Lemma 3 implies, by taking $u = 2\sigma\sqrt{2\ln(2n)}$,

$$\mathbf{P}\left\{\max_{i \le n} |Y_i| > 3\sigma\sqrt{2\ln(2n)}\right\} \le \frac{1}{(2n)^4}.$$

This implies, by the Borel-Cantelli lemma, that with probability one there exists a finite index T' such that for all $n > T'$, $\max_{i \le n} |Y_i| \le 3\sigma\sqrt{2\ln(2n)}$.

Also, there exists a finite index T such that for all $n > T$, $\max_{i \le n} |Y_i| \le \ln(n)$. Therefore, denoting $\overline{n} = 2^{\lfloor \log_2 n \rfloor + 1}$, we may write

$$\begin{aligned}
nL_n(g) &= \sum_{m=0}^{\lfloor \log_2 n \rfloor - 1} 2^m L_{2^m}^{2^{m+1}-1}(g) + (n - \overline{n}/2 + 1) L_{\overline{n}/2}^{n}(g) \\
&\le 4T \ln^2(T+1) + \sum_{m=\lfloor \log_2 T \rfloor}^{\lfloor \log_2 n \rfloor - 1} 2^m L_{2^m}^{2^{m+1}-1}(g) + (n - \overline{n}/2 + 1) L_{\overline{n}/2}^{n}(g) \\
&\quad \text{(since } \max(|Y_n|, |g(Y_1^{n-1})|) \le \ln(T+1) \text{ if } n \le T) \\
&\le 4T \ln^2(T+1) + \sum_{m=0}^{\lfloor \log_2 n \rfloor - 1} 2^m \inf_{k \ge 1} \left(L_{2^m}^{2^{m+1}-1}(h^{(k)}) - \frac{2m^2 \ln q_{k,m}}{2^m} \right) \\
&\quad + (n - \overline{n}/2 + 1) \inf_{k \ge 1} \left(L_{\overline{n}/2}^{n}(h^{(k)}) - \frac{2\lfloor \log_2 n \rfloor^2 \ln q_{k,m}}{n - \overline{n}/2 + 1} \right) \\
&\quad \text{(by Lemma 1)} \\
&\le 4T \ln^2(T+1) + \inf_{k \ge 1} \left(\sum_{m=0}^{\lfloor \log_2 n \rfloor - 1} 2^m L_{2^m}^{2^{m+1}-1}(h^{(k)}) \right. \\
&\quad \left. + (n - \overline{n}/2 + 1) L_{\overline{n}/2}^{n}(h^{(k)}) \right) + \sum_{m=0}^{\lfloor \log_2 n \rfloor} \frac{2m^2(m+1)}{2^m} \\
&\le 4T \ln^2(T+1) + n \inf_{k \ge 1} L_n(h^{(k)}) + 2(\log_2 n + 1)^3.
\end{aligned}$$

In other words,

$$L_n(g) \le \inf_{k \ge 1} L_n(h^{(k)}) + \frac{2(\log_2 n + 1)^3}{n} + O\left(\frac{1}{n}\right) \qquad \text{almost surely.}$$

This proves the second statement of the theorem. To prove the claimed universality property, it suffices to show that for all ergodic gaussian processes,

$$\limsup_{n \to \infty} \inf_{k \ge 1} L_n(h^{(k)}) = L^*.$$

This can be done similarly to the proof of Theorem 2:

$$\begin{aligned}
\lim_{n\to\infty} L_n(h^{(k)}) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=k+1}^{n}\left(Y_i - h^{(k)}(Y_1^{i-1})\right)^2 \\
&\le \lim_{n\to\infty}\frac{1}{n}\sum_{i=k+1}^{n}\left(Y_i - \sum_{j=1}^{k} c_{i,j}Y_{i-j}\right)^2 \\
&= \mathbf{E}\left\{\left(Y_0 - \sum_{j=1}^{k} c_j^* Y_{-j}\right)^2\right\} \\
&\stackrel{\text{def}}{=} \epsilon_k .
\end{aligned}$$

Define the coefficient vector $(\tilde{c}_{k,1},\ldots,\tilde{c}_{k,\ell})$ such that it minimizes

$$\min_{(c_1,\ldots,c_\ell)} \mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-k}^{-1}\} - \sum_{j=1}^{k} c_j Y_{-j}\right)^2\right\} .$$

(If the minimum is not unique, choose one arbitrarily.) Then

$$\begin{aligned}
\epsilon_k &= \mathbf{E}\left\{\left(Y_0 - \mathbf{E}\{Y_0|Y_{-\infty}^{-1}\}\right)^2\right\} + \mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-\infty}^{-1}\} - \sum_{j=1}^{k} c_j^* Y_{-j}\right)^2\right\} \\
&\le L^* + \mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-\infty}^{-1}\} - \sum_{j=1}^{k} \tilde{c}_{k,j} Y_{-j}\right)^2\right\} \\
&\le L^* + 2\mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-\infty}^{-1}\} - \mathbf{E}\{Y_0|Y_{-k}^{-1}\}\right)^2\right\} \\
&\quad +2\mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-k}^{-1}\} - \sum_{j=1}^{k} \tilde{c}_{k,j} Y_{-j}\right)^2\right\} \\
&= L^* + 2\delta_k
\end{aligned}$$

since

$$\mathbf{E}\left\{\left(\mathbf{E}\{Y_0|Y_{-k}^{-1}\} - \sum_{j=1}^{k} \tilde{c}_{k,j} Y_{-j}\right)^2\right\} = 0.$$

Now the proof can be finished by mimicking the proof of Theorem 2. □

Once again, we may derive a property analogous to Corollary 1:

Corollary 3 *Under the conditions of Theorem 3,*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}-g_i(Y_1^{i-1})\right)^2=0 \qquad \textit{almost surely.}$$

Proof. We proceed exactly as in the proof of Corollary 1. The only thing that needs a bit more care is checking the conditions of Kolmogorov's strong law for sums of martingale differences, since in the gaussian case the corresponding martingale differences are not bounded. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbf{E}Z_i^2 &= \mathbf{E}\left\{\left(Y_i-\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}\right)^2\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}-g_i(Y_1^{i-1})\right)^2\right\}\\
&\le \sqrt{\mathbf{E}\left\{\left(Y_i-\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}\right)^4\right\}\mathbf{E}\left\{\left(\mathbf{E}\{Y_i|Y_{-\infty}^{i-1}\}-g_i(Y_1^{i-1})\right)^4\right\}}\\
&\le \sqrt{8\mathbf{E}\left\{Y_i^4\right\}\left(\mathbf{E}\left\{Y_i^4\right\}+\mathbf{E}\left\{g_i(Y_1^{i-1})^4\right\}\right)}\\
&\le C\sqrt{(1+(\log i)^4)}\\
&\le C'(\log i)^2\ ,
\end{aligned}$$

where C and C' are positive constants, which implies $\sum_{i=1}^{\infty}\mathbf{E}Z_i^2/i^2<\infty$, so the condition of Kolmogorov's theorem is satisfied. □

Remark. Rates of convergence. The inequality of Theorem 3 shows that the rate of convergence of $L_n(g)$ to L^* is determined by the performance of the *best* elementary predictor $h^{(k)}$. The price of adaptation to the best elementary predictor is merely an additional term of the order of $n^{-1}\log^3 n$. This additional term is not much larger than an inevitable estimation error. This is supported by a result of Gerencsér and Rissanen [11] who showed that for any gaussian ARMA$(p+q)$ process and for any predictor g,

$$\mathbf{E}L_n(g)\ge L^*+(1-o(1))L^*(p+q)\frac{\ln n}{n}\ .$$

On the other hand, Gerencsér [10] showed under some mixing conditions for ARMA(p,q) processes that there exists a predictor $g_{p,q}$ such that

$$L_n(g_{p,q})=L^*+(1+o(1))L^*(p+q)\frac{\ln n}{n} \qquad \text{almost surely.}$$

Further rate-of-convergence results under more general conditions for the process were established by Gerencsér [9]. Another general branch of bounds can be found in Goldenshluger and Zeevi [12]. Consider the MA(∞) representation of Y_n

$$Y_n = \sum_{i=1}^{\infty} \psi_i \epsilon_{n-i}$$

with transfer function

$$\psi(z) = \sum_{i=1}^{\infty} \psi_i z^i.$$

Goldenshluger and Zeevi show that if for $0 < l < 1 < L < \infty$ and $\rho > 1$

$$l < |\psi(z)| < L, \text{ for } |z| < \rho,$$

then for large n

$$\mathbf{E}L_n(h^{(k)}) \leq L^* + c(l, L)\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{i^2(\rho-1)2} + \frac{k}{\rho^{2k}(\rho-1)2} + \frac{k}{i}\right)$$

and for $k = k_n = \left\lfloor \frac{\ln n}{2\ln\rho} \right\rfloor$

$$\mathbf{E}L_n(h^{(k)}) \leq L^* + C(l, L, \rho)\frac{1}{n}\sum_{i=1}^{n}\frac{\ln i}{i} \leq L^* + C(l, L, \rho)\frac{(\ln n)^2}{n}.$$

Thus, for the processes investigated by Goldenshluger and Zeevi, the predictor g of Theorem 3 achieves the rate of convergence

$$\mathbf{E}L_n(g) \leq L^* + O\left(\frac{\log^3 n}{n}\right).$$

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 uncertainty. *IEEE Transactions on Information Theory*, 40:609–634, 1994.

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

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

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

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

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

[9] L. Gerencsér. $AR(\infty)$ estimation and nonparametric stochastic complexity. *IEEE Transactions on Information Theory*, 38:1768–1779, 1992.

[10] L. Gerencsér. On Rissanen's predictive stochastic complexity for stationary ARMA processes. *J. of Statistical Planning and Inference*, 41:303–325, 1994.

[11] L. Gerencsér, and J. Rissanen. A prediction bound for Gaussian ARMA processes. *Proc. of the 25th Conference on Decision and Control*, 1487–1490, 1986.

[12] A. Goldenshluger, and A. Zeevi. Non-asymptotic bounds for autoregressive time series modeling. (submitted for publication).

[13] L. Györfi. Adaptive linear procedures under general conditions. *IEEE Transactions on Information Theory*, 30:262–267, 1984.

[14] 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.

[15] J. Kivinen and M. K. Warmuth. Averaging expert predictions. In H. U. Simon P. Fischer, editor, *Computational Learning Theory: Proceedings of the Fourth European Conference, EuroCOLT'99*, pages 153–167. Springer, Berlin, 1999. Lecture Notes in Artificial Intelligence 1572.

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

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

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

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

[20] A. Nobel. Aggregate schemes for sequential prediction of ergodic processes. manuscript, 2000.

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

[22] G. Pisier. Probabilistic methods in the geometry of Banach spaces. In *Probability and Analysis. Lecture Notes in Mathematics, 1206*, pages 167–241. Springer, New York, 1986.

[23] A. Singer and M. Feder. Universal linear prediction by model order weighting. *IEEE Transactions on Signal Processing*, 47:2685–2699, 1999.

[24] A. C. Singer, and M. Feder. Universal linear least-squares prediction. *International Symposium of Information Theory*, 2000.

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

[26] Ya.Z. Tsypkin. *Adaptation and Learning in Automatic Systems.* Academic Press, New York, 1971.

[27] S. Yakowitz. Small-sample hypothesis tests of Markov order, with application to simulated and hydrologic chains. *Journal of the American Statistical Association*, 71:132–136, 1976.

[28] S. Yakowitz. Nonparametric estimation of Markov transition functions. *Annals of Statistics*, 7:671–679, 1979.

[29] S. Yakowitz. Nonparametric density estimation, prediction, and regression for Markov sequences. *Journal of the American Statistical Association*, 80:215–221, 1985.

[30] S. Yakowitz. Nearest-neighbour methods for time series analysis. *Journal of Time Series Analysis*, 8:235–247, 1987.

[31] S. Yakowitz. Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *Journal of Multivariate Analysis*, 30:124–136, 1989.

[32] S. Yakowitz, L. Györfi, J. Kieffer, and G. Morvai. Strongly consistent nonparametric estimation of smooth regression functions for stationary ergodic sequences. *Journal of Multivariate Analysis*, 71:24–41, 1999.

[33] 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.

[34] Y. Yang. Aggregating regression procedures for a better performance. *manuscript*, 1999.

[35] Y. Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74:135–161, 2000.