An open problem on strongly consistent learning of the best prediction for Gaussian processes

László Györfi and Alessio Sancetta

For Gaussian process, we present an open problem whether or not there is a data driven predictor of the conditional expectation of the current value given the past such that the difference between the predictor and the conditional expectation tends to zero almost surely for all stationary, ergodic, Gaussian process. We show some related negative and positive findings.

1 Open problem

Let $\{Y_n\}_{-\infty}^{\infty}$ be a stationary, ergodic, mean zero Gaussian process. The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$. It is an open problem whether it is possible to learn the best predictor from the past data in a strongly consistent way, i.e., whether there exists a prediction rule g such that

$$\lim_{n\to\infty} \left(\mathbf{E}\{Y_n \mid Y_1^{n-1}\} - g_n(Y_1^{n-1})\right) = 0 \quad \text{almost surely} \tag{1}$$

for all stationary and ergodic Gaussian processes. (Here Y_1^{n-1} denotes the string $Y_1, \ldots, Y_{n-1}$.)

Bailey [3] and Ryabko [31] proved that just stationarity and ergodicity is not enough, i.e., for any predictor g, there is a binary valued stationary ergodic process such that

László Györfi
Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Stoczek u.2, 1521 Budapest, Hungary e-mail: gyorfi@cs.bme.hu. This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013).

Alessio Sancetta
Department of Economics, Royal Holloway University of London, e-mail: asancetta@gmail.com

$$\mathbf{P}\left\{\limsup_{n\to\infty}|g_n(Y_1^{n-1})-\mathbf{E}\{Y_n\mid Y_1^{n-1}\}|\geq 1/2\right\}\geq 1/8,$$

(cf. Györfi, Morvai, Yakowitz [18]).

In this paper we try to collect some related results such that the main aim is to have as mild conditions on the stationary, ergodic, Gaussian process $\{Y_n\}_{-\infty}^{\infty}$ as possible.

Concerning ergodicity of stationary Gaussian processes, L_2 ergodicity means that

$$\mathbf{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)^2\right\}\to 0, \tag{2}$$

which is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}r(i)\to 0, \tag{3}$$

where

$$r(i)=cov(Y_i,Y_0),$$

(cf. Karlin, Taylor [22]). Moreover, because of stationarity the ergodic theorem implies that

$$\frac{1}{n}\sum_{i=1}^{n}Y_i\to\mathbf{E}\{Y_1\mid\mathscr{F}\} \tag{4}$$

a.s. such that $\mathscr{F}$ is the σ-algebra of invariant sets. From (2) and (4) we get that

$$\mathbf{E}\{Y_1\mid\mathscr{F}\}=0 \tag{5}$$

a.s. Thus, from (3) we get the strong law of large numbers, and so (3) is a necessary condition for ergodicity of a stationary Gaussian process. Maruyama [25] and Grenander [15] proved that the necessary and sufficient condition for ergodicity of a stationary Gaussian process is that the spectral distribution function F is everywhere continuous. Lindgren [24] showed that

$$\frac{1}{n}\sum_{i=1}^{n}r(i)^2\to 0 \tag{6}$$

is a necessary condition for ergodicity, while

$$r(i)\to 0 \tag{7}$$

is a sufficient condition. Because of Jensen inequality, we get that

$$\left(\frac{1}{n}\sum_{i=1}^{n}r(i)\right)^2\leq\frac{1}{n}\sum_{i=1}^{n}r(i)^2,$$

therefore (6) implies (3). Cornfeld et al. [8] showed that for stationary Gaussian process with absolutely continuous spectral distribution, (7) is a necessary, too.

In the theory of prediction of stationary Gaussian process, the Wold decomposition plays an important role. It says that we have $Y_n = U_n + V_n$, where the stationary Gaussian processes $\{U_n\}_{-\infty}^{\infty}$ and $\{V_n\}_{-\infty}^{\infty}$ are independent, $\{U_n\}_{-\infty}^{\infty}$ has the MA(∞) representation

$$\sum_{j=0}^{\infty} a_j^* Z_{n-j}, \tag{8}$$

with i.i.d. Gaussian innovations $\{Z_n\}$ and with

$$\sum_{i=1}^{\infty} |a_i^*|^2 < \infty, \tag{9}$$

while the process $\{V_n\}_{-\infty}^{\infty}$ is deterministic: $V_n = \mathbf{E}\{V_n \mid V_{-\infty}^{n-1}\}$.

For a stationary, ergodic Gaussian process, we may get a similar decomposition, if we write the continuous spectral distribution function in the form $F(\lambda) = F^{(a)}(\lambda) + F^{(s)}(\lambda)$, where $F^{(a)}(\lambda)$ is an absolutely continuous distribution function with density function f and $F^{(s)}(\lambda)$ is singular continuous distribution function. Then we have the decomposition $Y_n = U_n' + V_n'$, where the stationary, ergodic Gaussian processes $\{U_n'\}_{-\infty}^{\infty}$ and $\{V_n'\}_{-\infty}^{\infty}$ are independent, $\{U_n'\}_{-\infty}^{\infty}$ has the spectral distribution function $F^{(a)}$, while $\{V_n'\}_{-\infty}^{\infty}$ has the spectral distribution function $F^{(s)}$. If $\int_{-\pi}^{\pi} \ln f(\lambda) d\lambda > -\infty$, then $U_n = U_n'$ and $V_n = V_n'$ (cf. Lindgren [24]).

In the analysis of stationary Gaussian processes one often assumes the MA(∞) or the AR(∞) representations such that these representations imply various type of mixing properties. The AR(∞) representation of the process $\{Y_n\}$ means that

$$Y_n = Z_n + \sum_{j=1}^{\infty} c_j^* Y_{n-j}, \tag{10}$$

with the vector $c^* = (c_1^*, c_2^*, \dots)$. Bierens [4] introduced a non invertible MA(1) process such that

$$Y_n = Z_n - Z_{n-1}, \tag{11}$$

where the innovations $\{Z_n\}$ are i.i.d. standard Gaussian. Bierens [4] proved that this process has no AR(∞) representation.

The rest of the paper is organized as follows. In Section 2 we summarize the basic concepts of predicting Gaussian time series, while Section 3 contains some positive and negative findings concerning universally consistent prediction. The current machine learning techniques in Section 4 may result in universal consistency.

2 Prediction of Gaussian processes

In this section we consider the classical problem of Gaussian time series prediction (cf. Brockwell and Davis [6]). In this context, parametric models based on distributional assumptions and structural conditions such as AR(p), MA(q), ARMA(p,q) and ARIMA(p,d,q) are usually fitted to the data (cf. Gerencsér and Rissanen [14], Gerencsér [12, 13]). However, in the spirit of modern nonparametric inference, we try to avoid such restrictions on the process structure. Thus, we only assume that we observe a string realization Y_1^{n-1} of a zero mean, stationary and ergodic Gaussian process $\{Y_n\}_{-\infty}^{\infty}$, and try to predict Y_n, the value of the process at time n.

For Gaussian time series and for any integer $k > 0$, $\mathbf{E}\{Y_n \mid Y_{n-k}^{n-1}\}$ is a linear function of Y_{n-k}^{n-1}:

$$\mathbf{E}\{Y_n \mid Y_{n-k}^{n-1}\} = \sum_{j=1}^{k} c_j^{(k)} Y_{n-j}, \tag{12}$$

where the coefficients $c_j^{(k)}$ minimize the risk

$$\mathbf{E}\left\{\left(\sum_{j=1}^{k} c_j Y_{-j} - Y_0\right)^2\right\},$$

therefore the main ingredient is the estimate of the coefficients $c_1^{(k)}, \ldots, c_k^{(k)}$ from the data Y_1^{n-1}. Such an estimate is called elementary predictor, it is denoted by $\tilde{h}^{(k)}$ generating a prediction of form

$$\tilde{h}^{(k)}(Y_1^{n-1}) = \sum_{j=1}^{k} C_{n,j}^{(k)} Y_{n-j}$$

such that the coefficients $C_{n,j}^{(k)}$ minimize the empirical risk

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

if $n > k$, and the all-zero vector otherwise. Even though the minimum always exists, it is not unique in general, and therefore the minimum is not well-defined. It is shown by Györfi [16] that there is a unique vector $C_n^{(k)} = (C_{n,1}^{(k)}, \ldots, C_{n,k}^{(k)})$ such that

$$\sum_{i=k+1}^{n-1}\left(\sum_{j=1}^{k} C_{n,j}^{(k)} Y_{i-j} - Y_i\right)^2 = \min_{(c_1,\ldots,c_k)} \sum_{i=k+1}^{n-1}\left(\sum_{j=1}^{k} c_j Y_{i-j} - Y_i\right)^2,$$

and it has the smallest Euclidean norm among the minimizer vectors.

For fixed k, an elementary predictor

$$\tilde{h}^{(k)}(Y_1^{n-1}) = \sum_{j=1}^{k} C_{n,j}^{(k)} Y_{n-j}$$

cannot be consistent. In order to get consistent predictions there are three main principles:

- k is a deterministic function of n,
- k depends on the data Y_1^{n-1},
- aggregate the elementary predictors $\{\tilde{h}^{(k)}(Y_1^{n-1}), k = 1,2,\ldots,n-2\}$.

3 Deterministic k_n

Schäfer [32] investigated the following predictor: for $a > 0$, introduce the truncation function

$$T_a(z) = \begin{cases} a & \text{if } z > a; \\ z & \text{if } |z| < a; \\ -a & \text{if } z < -a. \end{cases}$$

Choose $L_n \uparrow \infty$, then his predictor is

$$\bar{g}_n(Y_1^{n-1}) = \sum_{j=1}^{k_n} C_{n,j}^{(k_n)} T_{L_n}(Y_{n-j}).$$

Schäfer [32] proved that, under some conditions on the Gaussian process, we have that

$$\lim_{n\to\infty} \left(\mathbf{E}\{Y_n \mid Y_{n-k_n}^{n-1}\} - \bar{g}_n(Y_1^{n-1}) \right) = 0 \quad \text{a.s.}$$

His conditions include that the process has the MA(∞) representation (8) such that

$$\sum_{i=1}^{\infty} |a_i^*| < \infty, \tag{13}$$

and therefore it is purely nondeterministic and the spectral density exists. Moreover, he assumed that

$$\mathbf{E}\{Y_n \mid Y_{-\infty}^{n-1}\} - \mathbf{E}\left\{ Y_n | Y_{n-k_n}^{n-1} \right\} \to 0$$

a.s. For example, he proved the strong consistency with $k_n = n^{1/4}$ if the spectral density is bounded away from zero. The question left is how to avoid these conditions such that we pose conditions only on the covariances just slightly stronger than (7).

For a deterministic sequence $k_n, n = 1,2,\ldots$, consider the predictor

$$\tilde{g}_n(Y_1^{n-1}) = \tilde{h}^{(k_n)}(Y_1^{n-1}) = \sum_{j=1}^{k_n} C_{n,j}^{(k_n)} Y_{n-j}.$$

For the prediction error $\mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \tilde{g}_n(Y_1^{n-1})$ we have the decomposition

$$\mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \tilde{g}_n(Y_1^{n-1}) = I_n + J_n,$$

where

$$I_n = \mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\}$$

is the approximation error, and

$$J_n = \mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} - \tilde{g}_n(Y_1^{n-1}) = \sum_{j=1}^{k_n} (c_j^{(k_n)} - C_{n,j}^{(k_n)}) Y_{n-j}$$

is the estimation error. In order to have small approximation error, we need $k_n \to \infty$, while the control of the estimation error is possible if this convergence to ∞ is slow.

We guess that the following is true:

Conjecture 1. For any deterministic sequence k_n, there is a stationary, ergodic Gaussian process such that the prediction error $\mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \sum_{j=1}^{k_n} C_{n,j}^{(k_n)} Y_{n-j}$ does not converge to 0 a.s.

Next we show that the approximation error tends to zero in L_2 without any condition:

Lemma 1. *For any sequence $k_n \to \infty$ and for any stationary process $\{Y_n\}_{-\infty}^{\infty}$,*

$$\lim_{n\to\infty} \mathbf{E}\{(I_n)^2\} = 0.$$

Proof. We follow the argument from Doob [10]. Because of stationarity,

$$\mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\}$$

and

$$\mathbf{E}\{Y_0 \mid Y_{-n+1}^{-1}\} - \mathbf{E}\left\{Y_0 | Y_{-k_n}^{-1}\right\}$$

have the same distribution. The sequence $\mathbf{E}\{Y_0 \mid Y_{-n+1}^{-1}\}$, $n = 1, 2, \ldots$ is a martingale such that $\mathbf{E}\{Y_0 \mid Y_{-n+1}^{-1}\} \to \mathbf{E}\{Y_0 \mid Y_{-\infty}^{-1}\}$ a.s. and in L_2, too. Similarly, if $k_n \to \infty$ then $\mathbf{E}\{Y_0 \mid Y_{-k_n}^{-1}\} \to \mathbf{E}\{Y_0 \mid Y_{-\infty}^{-1}\}$ a.s. and in L_2. These imply that

$$\mathbf{E}\{Y_0 \mid Y_{-n+1}^{-1}\} - \mathbf{E}\left\{Y_0 | Y_{-k_n}^{-1}\right\} \to 0$$

a.s. and in L_2, therefore for the variance of the approximation error, $k_n \to \infty$ implies that

$$\mathbf{Var}(I_n) = \mathbf{Var}(\mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\}) \to 0. \tag{14}$$

□

Next we consider the problem of strong convergence of the approximation error. First we show a negative finding:

Proposition 1. *Put $k_n = (\ln n)^{1-\delta}$ with $0 < \delta < 1$. Then for the MA(1) process defined in (11), the approximation error does not converge to zero a.s.*

Proof. For the MA(1) process defined in (11), we get that

$$\mathbf{E}\left\{Y_n | Y_1^{n-1}\right\} = \sum_{j=1}^{n-1} \left(\frac{j}{n} - 1\right) Y_{n-j},$$

(see equation (5) in Bierens [4]). Similarly,

$$\mathbf{E}\left\{Y_{k_n+1} | Y_1^{k_n}\right\} = \sum_{j=1}^{k_n} \left(\frac{j}{k_n+1} - 1\right) Y_{k_n+1-j},$$

so stationarity implies that

$$\mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} = \sum_{j=1}^{k_n} \left(\frac{j}{k_n+1} - 1\right) Y_{n-j}.$$

On the one hand

$$\begin{aligned}
\mathbf{E}\left\{Y_n | Y_1^{n-1}\right\} &= \sum_{j=1}^{n-1} \left(\frac{j}{n} - 1\right) Y_{n-j} \\
&= \sum_{j=1}^{n-1} \left(\frac{j}{n} - 1\right) (Z_{n-j} - Z_{n-j-1}) \\
&= \left(\frac{1}{n} - 1\right) Z_{n-1} + \frac{1}{n} \sum_{j=0}^{n-2} Z_j,
\end{aligned}$$

and on the other hand

$$\begin{aligned}
\mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} &= \sum_{j=1}^{k_n} \left(\frac{j}{k_n+1} - 1\right) Y_{n-j} \\
&= \sum_{j=1}^{k_n} \left(\frac{j}{k_n+1} - 1\right) (Z_{n-j} - Z_{n-j-1}) \\
&= \left(\frac{1}{k_n+1} - 1\right) Z_{n-1} + \frac{1}{k_n+1} \sum_{j=n-k_n-1}^{n-2} Z_j.
\end{aligned}$$

Thus

$$\mathbf{E}\left\{Y_n | Y_1^{n-1}\right\} - \mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\}$$

$$= \left(\frac{1}{n} - \frac{1}{k_n+1}\right) Z_{n-1} + \frac{1}{n}\sum_{j=0}^{n-2} Z_j - \frac{1}{k_n+1}\sum_{j=n-k_n-1}^{n-2} Z_j$$
$$= \frac{1}{n}\sum_{j=0}^{n-1} Z_j - \frac{1}{k_n+1}\sum_{j=n-k_n-1}^{n-1} Z_j.$$

The strong law of large numbers implies that $\frac{1}{n}\sum_{j=0}^{n-1} Z_j \to 0$ a.s., therefore we have to prove that

$$\limsup_n \frac{1}{k_n+1}\sum_{j=n-k_n-1}^{n-1} Z_j = \infty$$

a.s. Let $n_m = \lfloor m \ln m \rfloor$ be a subsequence of the positive integers, then we show that

$$\limsup_m \frac{1}{k_{n_m}+1}\sum_{j=n_m-k_{n_m}-1}^{n_m-1} Z_j = \infty$$

a.s. One can check that $n_m - k_{n_m} > n_{m-1}$, therefore the intervals $[n_m - k_{n_m} - 1, n_m - 1]$, $m = 1,2,\dots$ are disjoint, and so for $C > 0$ the error events

$$A_m := \left\{ \frac{1}{k_{n_m}+1}\sum_{j=n_m-k_{n_m}-1}^{n_m-1} Z_j > C \right\}$$

$m = 1,2,\dots$ are independent. If φ and Φ denote the density and the distribution function of a standard normal distribution, then the tail probabilities of the standard Gaussian satisfy

$$\frac{\varphi(z)}{z}\left(1 - \frac{1}{z^2}\right) \le \Phi(-z) \le \frac{\varphi(z)}{z}$$

for $z > 0$, (cf. Feller [11, p. 179]). These imply that

$$\mathbf{P}\{A_m\} = \mathbf{P}\left\{ \frac{1}{k_{n_m}+1}\sum_{j=n_m-k_{n_m}-1}^{n_m-1} Z_j > C \right\}$$
$$= \Phi(-C\sqrt{k_{n_m}+1})$$
$$\ge \frac{\varphi\left(C\sqrt{k_{n_m}+1}\right)}{C\sqrt{k_{n_m}+1}}\left(1 - \frac{1}{C^2(k_{n_m}+1)}\right).$$

Because of the choice of k_n, we get that

$$\sum_{m=1}^{\infty} \mathbf{P}\{A_m\} \ge \sum_{m=1}^{\infty} \frac{\varphi\left(C\sqrt{k_{n_m}+1}\right)}{C\sqrt{k_{n_m}+1}}\left(1 - \frac{1}{C^2(k_{n_m}+1)}\right) = \infty,$$

so the (second) Borel-Cantelli lemma for independent events implies that

$$\mathbf{P}\left\{\limsup_m A_m\right\} = 1,$$

and the proof of the proposition is finished. □

Proposition 2. *Assume that for all $n > k$,*

$$\sum_{j=k+1}^{n-1} c_j^{(n-1)} r(j) \le C_1 k^{-\gamma}, \tag{15}$$

and

$$\sum_{j=1}^{k} (c_j^{(n-1)} - c_j^{(k)}) r(j) \le C_2 k^{-\gamma}, \tag{16}$$

with $\gamma > 0$. If

$$k_n = (\ln n)^{(1+\delta)/\gamma} \tag{17}$$

($\delta > 0$), then for the approximation error, we have that $I_n = \mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} \to 0$ a.s.

Proof. The approximation error I_n is a zero mean Gaussian random variable. Buldygin, Donchenko [7] proved that $I_n \to 0$ a.s. if and only if $\mathbf{Var}(I_n) \to 0$ and for any $\varepsilon > 0$,

$$\mathbf{P}\left\{\limsup_{n\to\infty} I_n < \varepsilon\right\} > 0. \tag{18}$$

Because of (14), we have to verify (18), which is equivalent to

$$\mathbf{P}\left\{\limsup_{n\to\infty} I_n \ge \varepsilon\right\} < 1.$$

Next we show that under the conditions of the proposition we have that

$$\mathbf{Var}(I_n) \le \frac{c}{(\ln n)^{1+\delta}} \tag{19}$$

with some constants $c > 0$ and $\delta > 0$. In order to show (19), consider the representations $\mathbf{E}\{Y_n \mid Y_1^{n-1}\} = \sum_{j=1}^{n-1} c_j^{(n-1)} Y_{n-j}$ and $\mathbf{E}\{Y_n \mid Y_{n-k_n}^{n-1}\} = \sum_{j=1}^{k_n} c_j^{(k_n)} Y_{n-j}$. Introduce the vectors $X_i^{(k)} = (Y_{i-k}, \dots, Y_{i-1})^T$ (where the superscript T denotes transpose), and the empirical covariance matrix $R_n^{(k)} = \frac{1}{n-k-1} \sum_{i=k+1}^{n-1} X_i^{(k)} (X_i^{(k)})^T$, and the vector of empirical covariances $M_n^{(k)} = \frac{1}{n-k-1} \sum_{i=k+1}^{n-1} Y_i X_i^{(k)}$. If $r(n) \to 0$ then the covariance matrix $R^{(k)} = \mathbf{E}\{R_n^{(k)}\}$ is not singular, and the optimal mean squared error of the prediction is

$$\begin{aligned}\mathbf{E}\{(Y_0 - \mathbf{E}\{Y_0 \mid Y_{-k}^{-1}\})^2\} &= \mathbf{E}\{Y_0^2\} - \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-k}^{-1}\}^2\} \\ &= \mathbf{E}\{Y_0^2\} - (M^{(k)})^T (R^{(k)})^{-1} M^{(k)},\end{aligned}$$

(cf. Proposition 5.1.1 in Brockwell, Davis [6]), where $M^{(k)} = \mathbf{E}\{M_n^{(k)}\}$. Thus,

$$\begin{aligned}\mathbf{Var}(I_n) &= \mathbf{E}\{I_n^2\} \\ &= \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-(n-1)}^{-1}\}^2\} - \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-k_n}^{-1}\}^2\} \\ &= (M^{(n-1)})^T (R^{(n-1)})^{-1} M^{(n-1)} - (M^{(k_n)})^T (R^{(k_n)})^{-1} M^{(k_n)}.\end{aligned}$$

Moreover, we have that $c^{(n-1)} = (R^{(n-1)})^{-1} M^{(n-1)}$ and $c^{(k_n)} = (R^{(k_n)})^{-1} M^{(k_n)}$. Applying the conditions of the proposition, we get that

$$\begin{aligned}\mathbf{Var}(I_n) &= \sum_{j=1}^{n-1} c_j^{(n-1)} M_j^{(n-1)} - \sum_{j=1}^{k_n} c_j^{(k_n)} M_j^{(k_n)} \\ &= \sum_{j=1}^{n-1} c_j^{(n-1)} r(j) - \sum_{j=1}^{k_n} c_j^{(k_n)} r(j) \\ &= \sum_{j=1}^{k_n} (c_j^{(n-1)} - c_j^{(k_n)}) r(j) + \sum_{j=k_n+1}^{n-1} c_j^{(n-1)} r(j) \\ &\le (C_1 + C_2) k_n^{-\gamma}\end{aligned}$$

with $\gamma > 0$, then for the choice (17), (19) is proved. Thus, (19) implies that

$$\mathbf{P}\{I_n \ge \varepsilon\} = \Phi\left(-\frac{\varepsilon}{\sqrt{\mathbf{Var}(I_n)}}\right) \le e^{-\frac{\varepsilon^2}{2\mathbf{Var}(I_n)}} \le e^{-\frac{\varepsilon^2 (\ln n)^{1+\delta}}{2c}} = n^{-\frac{\varepsilon^2 (\ln n)^{\delta}}{2c}}$$

therefore

$$\sum_{n=1}^{\infty} \mathbf{P}\{I_n \ge \varepsilon\} < \infty,$$

so the Borel-Cantelli Lemma implies that

$$\limsup_{n\to\infty} I_n < \varepsilon$$

a.s. □

The partial autocorrelation function of Y_n is $\alpha(j) := c_j^{(j)}$ where $c_j^{(j)}$ is as defined before, i.e. the j^{th} coefficient from the AR(j) approximation of Y_n. It is possible to explicitly bound the approximation error I_n using $\alpha(j)$. The asymptotic behavior of $\alpha(j)$ has been studied extensively in the literature. For example, $|\alpha(j)| \le c/j$ for fractionally integrated ARIMA processes (e.g. Inoue, [20]). This includes Gaussian processes such that $|r(i)| \le c(j+1)^{-\beta}$ under the sole condition that $\beta > 0$, as conjectured in Remark 1 below. It is unknown whether all stationary and ergodic purely non deterministic Gaussian processes have partial correlation function satisfying $|\alpha(j)| \le c/j$.

Proposition 3. *Suppose that*

$$\sum_{j=k+1}^{\infty} \alpha^2(j) \le ck^{-\gamma}$$

with $\gamma > 0$, $c > 0$. For the choice (17), we have that $I_n = \mathbf{E}\{Y_n \mid Y_1^{n-1}\} - \mathbf{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} \to 0$ *a.s.*

Proof. We follow the line of the proof of Proposition 2 such that verify (19). To ease notation, let σ_k^2 be the optimal mean square prediction error of the AR(k) approximation: $\sigma_k^2 := \mathbf{E}\{(Y_0 - \mathbf{E}\{Y_0 \mid Y_{-k}^{-1}\})^2\}$. By the Durbin-Levinson Algorithm (cf. Brockwell and Davis [6], Proposition 5.2.1),

$$\sigma_k^2 = \sigma_{k-1}^2(1-\alpha^2(k)) = r(0)\prod_{j=1}^{k}(1-\alpha^2(j)),$$

iterating the recursion and noting that $\sigma_0^2 = r(0) = \mathbf{Var}(Y_0)$. Since, as in the proof of Proposition 2,

$$\begin{aligned}
\mathbf{E}\{(I_n)^2\} =& \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-(n-1)}^{-1}\}^2\} - \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-k_n}^{-1}\}^2\} \\
=& \sigma_{k_n}^2 - \sigma_{n-1}^2 \\
=& r(0)\prod_{j=1}^{k_n}(1-\alpha^2(j))\left(1 - \prod_{j=k+1}^{n-1}(1-\alpha^2(j))\right).
\end{aligned}$$

Without loss of generality assume that $\alpha(j)^2 \le C < 1$. For $0 < x \le C < 1$ apply the inequality $-\frac{\ln C}{C}x \le \ln(1-x)$, then

$$\begin{aligned}
\mathbf{E}\{(I_n)^2\} &\le r(0)\left(1 - \prod_{j=k_n+1}^{\infty}(1-\alpha^2(j))\right) \\
&= r(0)\left(1 - e^{\sum_{j=k_n+1}^{\infty}\ln(1-\alpha^2(j))}\right) \\
&\le r(0)\left(1 - e^{-\frac{\ln C}{C}\sum_{j=k_n+1}^{\infty}\alpha^2(j)}\right) \\
&\le r(0)\frac{\ln C}{C}\sum_{j=k_n+1}^{\infty}\alpha^2(j) \\
&\le r(0)\frac{\ln C}{C}ck_n^{-\gamma}.
\end{aligned}$$

Hence, with the choice (17), (19) is verified. □

Remark 1. We conjecture that under the condition

$$|r(i)| \le c(|i|+1)^{-\beta},$$

$c < \infty$, $\beta > 0$, the conditions of Propositions 2 and 3 are satisfied. Notice that for any MA(p), the conditions of Proposition 2 are met, while for any AR(q), the conditions of Proposition 3 are satisfied.

Remark 2. Notice that the MA(1) example in the proof of Proposition 1 satisfies the conditions of the Propositions 2 and 3 with $\gamma = 1$, since $\alpha(j) = c_j^{(j)} = j^{-1}$, and $r(0) = 2$, $r(1) = -1$ and $r(i) = 0$ if $i \geq 2$, from which one gets that

$$\mathbf{Var}(I_n) = \frac{1}{k_n + 1} - \frac{1}{n}.$$

Moreover, the choice $k_n = (\ln n)^{1+\delta}$ is just slightly larger than in the proof of Proposition 1.

Remark 3. Under the AR(∞) representation (10), the derivation and the conditions of Proposition 2 can be simplified. Multiplying both sides of (10) by Y_n and taking expectations,

$$\begin{aligned}
\mathbf{E}Y_n^2 &= \sum_{i=1}^{\infty} c_i^* \mathbf{E}\{Y_n Y_{n-i}\} + \mathbf{E}\{Y_n Z_n\} \\
&= \sum_{i=1}^{\infty} c_i^* r(i) + \mathbf{Var}(Z_0) \\
&= \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-\infty}^{-1}\}^2\} + \mathbf{Var}(Z_0).
\end{aligned}$$

It implies that $\sum_{i=1}^{\infty} c_i^* r(i) < \infty$ and

$$\begin{aligned}
\mathbf{Var}(I_n) &\leq \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-\infty}^{-1}\}^2\} - \mathbf{E}\{\mathbf{E}\{Y_0 \mid Y_{-k_n}^{-1}\}^2\} \\
&\leq \sum_{j=1}^{k_n} (c_j^* - c_j^{(k_n)}) r(j) + \sum_{j=k_n+1}^{\infty} c_j^* r(j) \\
&\leq (C_1 + C_2) k_n^{-\gamma},
\end{aligned}$$

if the conditions $\sum_{i=k}^{\infty} c_i^* r(i) \leq C_1 k^{-\gamma}$ and $\sum_{j=1}^{k} (c_j^* - c_j^{(k)}) r(j) \leq C_2 k^{-\gamma}$ are satisfied.

Remark 4. If the process is has the MA(∞) representation (8), then $r(i) = \sum_{j=0}^{\infty} a_j^* a_{j+i}^*$ assuming the innovations have variance one. The Cauchy-Schwarz inequality implies that

$$|r(i)| \leq \sqrt{\sum_{j=0}^{\infty} (a_j^*)^2 \sum_{j=0}^{\infty} (a_{j+i}^*)^2} = \sqrt{\sum_{j=0}^{\infty} (a_j^*)^2 \sum_{j=i}^{\infty} (a_j^*)^2} \to 0.$$

We show that for $a_j^* > 0$, $\beta > 1$ implies (13). To see this, note that $r(i) > 0$ and

$$\left(\sum_{j=0}^{\infty} a_j^* \right)^2 \geq \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_j^* a_{j+i}^* = \sum_{i=1}^{\infty} r(i).$$

Moreover, $\sum_{i=1}^{\infty} r(i) < \infty$ implies that $\sum_{j=0}^{\infty} a_j^* < \infty$. Notice that without any conditions on $\{a_j^*\}$, $\sum_{i=1}^{\infty} |r(i)| < \infty$ does not imply that $\sum_{j=0}^{\infty} |a_j^*| < \infty$.

Remark 5. Concerning the estimation error the main difficulty is the possible slow rate of convergence of averages. For an arbitrary ergodic process, the rate of convergence of an average can be arbitrary slow, which means that for any sequence $a_n \downarrow 0$, there is a zero mean, stationary, ergodic process such that

$$\limsup_n \frac{\mathbf{E}\{(\frac{1}{n}\sum_{i=1}^n Y_i)^2\}}{a_n} > 0. \tag{20}$$

The question here is whether or not for arbitrary sequence $a_n \downarrow 0$, there is a zero mean, stationary, ergodic Gaussian process with covariances $\{r(i)\}$ such that (20) is satisfied. To see this, let $\{Y_i\}$ have the MA(∞)representation (10) with Z_j standard normal and $a_0^* = 1$, $a_j^* = j^{-\alpha}$ for $j > 0$, $1 > \alpha > 1/2$. Then $a_j^* \downarrow$, therefore $r(j) \downarrow$ and so we get that

$$\begin{aligned}
\limsup_n \frac{\mathbf{E}\{(\frac{1}{n}\sum_{i=1}^n Y_i)^2\}}{a_n} &\geq \limsup_n \frac{1}{na_n} \sum_{i=0}^n r(i) \left(1 - \frac{|i|}{n}\right) \\
&\geq \limsup_n \frac{1}{2na_n} \sum_{i=0}^n r(i) \\
&\geq \limsup_n \frac{1}{2a_n} r(n) \\
&= \limsup_n \frac{1}{2a_n} \left(\sum_{j=0}^{\infty} a_j^* a_{j+n}^*\right).
\end{aligned}$$

Then $a_j^* \downarrow$ implies that

$$\begin{aligned}
\limsup_n \frac{\mathbf{E}\{(\frac{1}{n}\sum_{i=1}^n Y_i)^2\}}{a_n} &\geq \limsup_n \frac{1}{2a_n} \left(\sum_{j=0}^{\infty} (a_{j+n}^*)^2\right) \\
&\geq \limsup_n \frac{n^{(1-2\alpha)}}{2a_n}.
\end{aligned}$$

For $\alpha \to 1/2$ the sequence can be made to diverge for any $a_n \to 0$ polynomially. We can make it logarithmic using $a_j^* = \left(j \ln^{1+\varepsilon}(j)\right)^{-1/2}$ for $j > 1$ and some $\varepsilon > 0$, but it is a bit more complex. Similar slow rate results can be derived for empirical covariances.

If the process $\{Y_n\}_{-\infty}^{\infty}$ satisfies some mixing conditions, then Meir [27], Alquier and Wintenberger [2] and McDonald et al. [26] analyzed the predictor $\tilde{h}^{(k)}(Y_1^{n-1})$. If the process $\{Y_n\}_{-\infty}^{\infty}$ is stationary and ergodic, then Klimo and Nelson [23] proved that

$$C_n^{(k)} \to c^{(k)} \tag{21}$$

a.s. Unfortunately, the convergence (21) does not imply that $\sum_{j=1}^{k}(C_{n,j}^{(k)}-c_j^{(k)})Y_{n-j}\to 0$ a.s.

Conjecture 2. For any fixed k, there is a stationary, ergodic Gaussian process such that $\sum_{j=1}^{k}(C_{n,j}^{(k)}-c_j^{(k)})Y_{n-j}$ does not converge to 0 a.s.

Lemma 2. *Let*

$$r_n(i) := n^{-1}\sum_{j=1}^{n} Y_j Y_{j+i}$$

be the empirical autocovariance. Suppose that $|r(i)|\le c(|i|+1)^{-\beta}$, $c<\infty$, $\beta>0$. *Then, for the sequence* $a_n=(n^{\alpha}/k_n)$ *with* $\alpha\in(0,\beta\wedge(1/2))$,

$$a_n \max_{i\le k_n}|r_n(i)-r(i)|\to 0$$

a.s.

Proof. At first, we show that

$$n^2\mathbf{E}|r_n(k)-r(k)|^2\le c^2c_k n^{2-2\beta} \tag{22}$$

with $c_k<\infty$. Note that

$$\mathbf{E}|r_n(k)-r(k)|^2=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\left\{\left(Y_iY_{i+k}-\mathbf{E}Y_iY_{i+k}\right)\left(Y_jY_{j+k}-\mathbf{E}Y_jY_{j+k}\right)\right\}.$$

To this end,

$$\begin{aligned}&\mathbf{E}\left\{\left(Y_iY_{i+k}-\mathbf{E}Y_iY_{i+k}\right)\left(Y_jY_{j+k}-\mathbf{E}Y_jY_{j+k}\right)\right\}\\ &=\mathbf{E}\left\{Y_iY_{i+k}Y_jY_{j+k}\right\}-\mathbf{E}Y_iY_{i+k}\mathbf{E}Y_jY_{j+k}.\end{aligned}$$

By Isserlis Theorem ([21]),

$$\begin{aligned}&\mathbf{E}\left\{Y_iY_{i+k}Y_jY_{j+k}\right\}\\ &=\mathbf{E}Y_iY_{i+k}\mathbf{E}Y_jY_{j+k}+\mathbf{E}Y_iY_j\mathbf{E}Y_{i+k}Y_{j+k}+\mathbf{E}Y_{i+k}Y_j\mathbf{E}Y_iY_{j+k}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbf{E}\left\{\left(Y_iY_{i+k}-\mathbf{E}Y_iY_{i+k}\right)\left(Y_jY_{j+k}-\mathbf{E}Y_jY_{j+k}\right)\right\}\\ &=\mathbf{E}Y_iY_j\mathbf{E}Y_{i+k}Y_{j+k}+\mathbf{E}Y_{i+k}Y_j\mathbf{E}Y_iY_{j+k}\\ &=r^2(i-j)+r(i-j+k)\,r(i-j-k).\end{aligned}$$

Hence,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\left\{\left(Y_iY_{i+k}-\mathbf{E}Y_iY_{i+k}\right)\left(Y_jY_{j+k}-\mathbf{E}Y_jY_{j+k}\right)\right\}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{n} r^2(i-j) + \sum_{i=1}^{n}\sum_{j=1}^{n} r(i-j+k)r(i-j-k) \\
&= nr^2(0) + 2\sum_{i=1}^{n-1}(n-i)r^2(i) + nr^2(k) + 2\sum_{i=1}^{n-1}(n-i)r(i+k)r(i-k) \\
&\le n\left(2r^2(0) + 2\sum_{i=1}^{n-1} r^2(i) + \sum_{i=1}^{n-1}(r^2(i+k)+r^2(i-k))\right) \\
&\le nc^2\left(2 + 2\sum_{i=1}^{n-1}(|i|+1)^{-2\beta} + \sum_{i=1}^{n-1}\left((|i+k|+1)^{-2\beta} + (|i-k|+1)^{-2\beta}\right)\right) \\
&\le c^2 c_k n^{2-2\beta},
\end{aligned}
$$

and so (22) is proved. Ninness [29] proved that if an arbitrary sequence of random variables X_n, $n = 1,2,\dots$ satisfies

$$
\mathbf{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right\} \le Cn^{-2\beta}
$$

with $C < \infty$, then

$$
n^{\alpha}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \to 0
$$

a.s., where $0 < \alpha < \beta \wedge (1/2)$. Thus, (22) satisfies the condition in Theorem 2.1 of Ninness [29], which implies $n^{\alpha}|r_n(k) - r(k)| \to 0$ a.s. for each k. Hence, by the union bound, the lemma is true for any $a_n \le n^{\alpha}/k_n$. □

We slightly modify the coefficient vector as follows: introduce the notations $\tilde{R}_n^{(k)} = R_n^{(k)} + \frac{1}{\ln n} I$ and $\tilde{C}_n^{(k)} = (\tilde{R}_n^{(k)})^{-1} M_n^{(k)}$. Moreover, put $\tilde{c}^{(k)} = (\tilde{R}^{(k)})^{-1} M^{(k)}$, where $\tilde{R}^{(k)} = \mathbf{E}\{\tilde{R}_n^{(k)}\}$.

Proposition 4. *Under the conditions of Lemma 2 and for the choice $k_n = (\ln n)^{\gamma}$ with $\gamma > 0$,*

$$
\sum_{j=1}^{k_n} (\tilde{C}_{n,j}^{(k_n)} - \tilde{c}_j^{(k_n)}) Y_{n-j} \to 0
$$

a.s.

Proof. Because of the Cauchy-Schwarz inequality

$$
\begin{aligned}
|J_n| &= \left|\sum_{j=1}^{k_n} (\tilde{c}_j^{(k_n)} - \tilde{C}_{n,j}^{(k_n)}) Y_{n-j}\right| \\
&\le \sqrt{\sum_{j=1}^{k_n} (\tilde{c}_j^{(k_n)} - \tilde{C}_{n,j}^{(k_n)})^2 \sum_{j=1}^{k_n} Y_{n-j}^2}
\end{aligned}
$$

$$\leq \sqrt{\sum_{j=1}^{k_n} (\tilde{c}_j^{(k_n)} - \tilde{C}_{n,j}^{(k_n)})^2 k_n \max_{1\leq i\leq n} Y_i^2}.$$

Pisier [30] proved the following: let $Z_1,\ldots,Z_n$ be zero-mean Gaussian random variables with $\mathbf{E}\{Z_i^2\} = \sigma^2$, $i = 1,\ldots,n$. Then

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

and for each $u > 0$,

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

This implies, by taking $u = 2\sigma\sqrt{2\ln(2n)}$,

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

and therefore

$$\sum_{n=1}^{\infty} \mathbf{P}\{\max_{1\leq i\leq n} |Y_i| > 3\sigma\sqrt{2\ln(2n)}\} < \infty,$$

and so the Borel-Cantelli lemma implies that

$$\limsup_{n\to\infty} \frac{\max_{1\leq i\leq n} |Y_i|}{\sqrt{\ln n}} \leq \infty$$

a.s. Thus, we have to show that for the choice of $k_n = (\ln n)^\gamma$ we get that

$$k_n \ln n \sum_{j=1}^{k_n} (\tilde{c}_j^{(k_n)} - \tilde{C}_{n,j}^{(k_n)})^2 \to 0$$

a.s. Let $\|\cdot\|$ denote the Euclidean norm and the norm of a matrix. Then

$$\begin{aligned}
&\sum_{j=1}^{k_n} (\tilde{c}_j^{(k_n)} - \tilde{C}_{n,j}^{(k_n)})^2 \\
&= \|\tilde{c}^{(k_n)} - \tilde{C}_n^{(k_n)}\|^2 \\
&= \|(\tilde{R}^{(k)})^{-1} M^{(k)} - (\tilde{R}_n^{(k)})^{-1} M_n^{(k)}\|^2 \\
&\leq 2\|(\tilde{R}_n^{(k)})^{-1}(M^{(k)} - M_n^{(k)})\|^2 + 2\|((\tilde{R}^{(k)})^{-1} - (\tilde{R}_n^{(k)})^{-1}) M^{(k)}\|^2
\end{aligned}$$

Concerning the first term of the right hand side, we have that

$$\begin{aligned}\|(\tilde{R}_n^{(k)})^{-1}(M^{(k)}-M_n^{(k)})\|^2 &\le \|(\tilde{R}_n^{(k)})^{-1}\|^2\|M^{(k)}-M_n^{(k)}\|^2\\ &\le (\ln n)^2\sum_{i=1}^{k_n}(r(i)-r_n(i))^2\\ &\le (\ln n)^2 k_n \max_{1\le i\le k_n}(r(i)-r_n(i))^2.\end{aligned}$$

The derivation for the second term of the right hand side is similar:

$$\begin{aligned}\|((\tilde{R}^{(k)})^{-1}-(\tilde{R}_n^{(k)})^{-1})M^{(k)}\|^2 &\le \|(\tilde{R}^{(k)})^{-1}-(\tilde{R}_n^{(k)})^{-1}\|^2\|M^{(k)}\|^2\\ &\le \|(\tilde{R}^{(k)})^{-1}\|^2\|(\tilde{R}_n^{(k)})^{-1}\|^2\|\tilde{R}^{(k)}-\tilde{R}_n^{(k)}\|^2\|M^{(k)}\|^2\\ &\le (\ln n)^4\|R^{(k)}-R_n^{(k)}\|^2\sum_{i=1}^{k_n}r(i)^2\\ &\le (\ln n)^4\sum_{i=1}^{k_n}\sum_{j=1}^{k_n}(r(i-j)-r_n(i-j))^2\sum_{i=1}^{\infty}r(i)^2\\ &\le c_1(\ln n)^4 2k_n\sum_{i=1}^{k_n}(r(i)-r_n(i))^2\\ &\le c_1(\ln n)^4 2k_n^2\max_{1\le i\le k_n}(r(i)-r_n(i))^2.\end{aligned}$$

For the choice $k_n=(\ln n)^\gamma$, summarizing these inequalities we get that

$$\begin{aligned}k_n\ln n\sum_{j=1}^{k_n}(\tilde{c}_j^{(k_n)}-\tilde{C}_{n,j}^{(k_n)})^2 &\le c_2(k_n^2(\ln n)^3+k_n^3(\ln n)^5)\max_{1\le i\le k_n}(r(i)-r_n(i))^2)\\ &\le c_2(\ln n)^{5+3\gamma}\max_{1\le i\le k_n}(r(i)-r_n(i))^2)\\ &\to 0\end{aligned}$$

a.s., where we used Lemma 2 with $a_n=(\ln n)^{5+3\gamma}$. □

Remark 6. In this section we considered deterministic choices of k_n. One can introduce data driven choices of K_n, for example, via complexity regularization or via boosting. In principle, it is possible, that there is a data driven choice, for which the corresponding prediction is strongly consistent without any condition on the process. We conjecture the contrary: for any data driven sequence K_n, there is a stationary, ergodic Gaussian process such that the prediction error

$$\mathbf{E}\{Y_n\mid Y_1^{n-1}\}-\sum_{j=1}^{K_n}C_{n,j}^{(K_n)}Y_{n-j}$$

does not converge to 0 a.s.

4 Aggregation of elementary predictors

After n time instants, the (*normalized*) *cumulative squared prediction error* on the strings Y_1^n is

$$L_n(g) = \frac{1}{n}\sum_{i=1}^{n}\left(g_i(Y_1^{i-1}) - Y_i\right)^2.$$

There is a fundamental limit for the predictability of the sequence, which is determined by a result of Algoet [1]: for any prediction strategy g and stationary ergodic process $\{Y_n\}_{-\infty}^{\infty}$ with $\mathbf{E}\{Y_0^2\} < \infty$,

$$\liminf_{n\to\infty} L_n(g) \geq L^* \quad \text{almost surely}, \tag{23}$$

where

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

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past observation sequences $Y_{-\infty}^{-1} = (\ldots, Y_{-2}, Y_{-1})$. A prediction strategy g is called *universally consistent* with respect to a class $\mathscr{C}$ of stationary and ergodic processes $\{Y_n\}_{-\infty}^{\infty}$ if for each process in the class,

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

There are universally consistent prediction strategies for the class of stationary and ergodic processes with $\mathbf{E}\{Y^4\} < \infty$, (cf. Györfi and Ottucsák [19], and Bleakley et al. [5]).

With respect to the combination of elementary experts $\tilde{h}^{(k)}$, Györfi and Lugosi applied in [17] the so-called "doubling-trick", which means that the time axis is segmented into exponentially increasing epochs and at the beginning of each epoch the forecaster is reset.

Bleakley et al. [5] proposed a much simpler procedure which avoids in particular the doubling-trick. Set

$$h_n^{(k)}(Y_1^{n-1}) = T_{\min\{n^\delta, k\}}\left(\tilde{h}_n^{(k)}(Y_1^{n-1})\right),$$

where the truncation function T_a was introduced in Section 3 and $0 < \delta < \frac{1}{8}$.

Combine these experts as follows. Let $\{q_k\}$ be an arbitrarily probability distribution over the positive integers such that for all k, $q_k > 0$, and define the weights

$$w_{k,n} = q_k e^{-(n-1)L_{n-1}(h_n^{(k)})/\sqrt{n}} = q_k e^{-\sum_{i=1}^{n-1}(h_i^{(k)}(Y_1^{i-1}) - Y_i)^2/\sqrt{n}}$$

($k = 1, \ldots, n-2$) and their normalized values

$$p_{k,n} = \frac{w_{k,n}}{\sum_{i=1}^{n-2} w_{i,n}}.$$

The prediction strategy g at time n is defined by

$$g_n(Y_1^{n-1}) = \sum_{k=1}^{n-2} p_{k,n} h_n^{(k)}(Y_1^{n-1}), \qquad n = 1, 2, \ldots$$

Bleakley et al. [5] proved that the prediction strategy g defined above is universally consistent with respect to the class of all stationary and ergodic zero-mean Gaussian processes, i.e.,

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

which implies that

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

(cf. Györfi and Lugosi [17], and Györfi and Ottucsák [19]).

This later convergence is expressed in terms of an almost sure Cesáro consistency. We guess that even the almost sure consistency (1) holds. In order to support this conjecture mention that

$$g_n(Y_1^{n-1}) = \sum_{k=1}^{n-2} p_{k,n} h_n^{(k)}(Y_1^{n-1}) \approx \sum_{k=1}^{n-2} p_{k,n} \tilde{h}_n^{(k)}(Y_1^{n-1}) = \sum_{k=1}^{n-2} p_{k,n} \sum_{j=1}^{k} c_{n,j}^{(k)} Y_{n-j},$$

and so

$$g_n(Y_1^{n-1}) = \sum_{j=1}^{n-2} c_{n,j} Y_{n-j},$$

where

$$c_{n,j} = \sum_{k=j}^{n-2} p_{k,n} c_{n,j}^{(k)}.$$

References

1. Algoet, P.: The strong law of large numbers for sequential decisions under uncertainty. IEEE Transactions on Information Theory **40**, 609–634, (1994).
2. Alquier, P., Wintenberger, O.: Model selection for weakly dependent time series forecasting. Bernoulli **18**, 883–913 (2012).
3. Bailey, D.H.: Sequential schemes for classifying and predicting ergodic processes. PhD thesis, Stanford University (1976).
4. Bierens, H.J.: The Wold decomposition. Manuscript (2012). http://econ.la.psu.edu/ hbierens/WOLD.PDF
5. Bleakley, K., Biau, G., Györfi, L., Ottucs, G.: Nonparametric sequential prediction of time series. Journal of Nonparametric Statistics **22**, 297–317 (2010).
6. Brockwell, P., Davis, R.A.: Time Series: Theory and Methods, 2nd edn. Springer-Verlag, New York (1991)

7. Buldygin, V.V., Donchenko, V.S.: The convergence to zero of Gaussian sequences. Matematicheskie Zametki **21**, 531–538 (1977)
8. Cornfeld, I., Fomin, S., Sinai, Y.G.: Ergodic Theory. Springer-Verlag, New York (1982)
9. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York (1996)
10. Doob, J.L.: The elementary Gaussian processes. Annals of Mathematical Statistics **15**, 229–282 (1944)
11. Feller, W. An Introduction to Probability and its Applications, Vol. I. Wiley, New York (1957)
12. Gerencsér, L.: $AR(\infty)$ estimation and nonparametric stochastic complexity. IEEE Transactions on Information Theory **38**, 1768–1779 (1992)
13. Gerencsér, L.: On Rissanen's predictive stochastic complexity for stationary ARMA processes. J. of Statistical Planning and Inference **41**, 303–325 (1994)
14. Gerencsér, L., Rissanen, J.: A prediction bound for Gaussian ARMA processes. Proc. of the 25th Conference on Decision and Control, 1487–1490 (1986)
15. Grenander, U.: Stochastic processes and statistical inference. Arkiv Math. **1**, 195–277 (1950)
16. Györfi, L.: Adaptive linear procedures under general conditions. IEEE Transactions on Information Theory **30**, 262–267 (1984)
17. Györfi, L. and Lugosi, G.: Strategies for sequential prediction of stationary time series. In: Dror, M., L'Ecuyer, P., Szidarovszky, F. (eds.) Modelling Uncertainty: An Examination of its Theory, Methods and Applications, pp. 225–248. Kluweer Acad. Publ., Boston (2001)
18. Györfi, L., Morvai, G., Yakowitz, S.: Limits to consistent on-line forecasting for ergodic time series. IEEE Trans. Information Theory **44**, 886–892 (1998)
19. Györfi, L., Ottucsák, G.: Sequential prediction of unbounded time series. IEEE Trans. Inform. Theory **53**, 1866–1872 (2007)
20. Inoue, A.: AR and MA representations of partial autocorrelation functions, with applications. Probabilty Theory and Related Fields **140**, 523–551 (2008)
21. Isserlis, L.: On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika **12**, 134–39 (1918)
22. Karlin, S., Taylor, H.M.: A First Course in Stochastic Processes. Academic Press, New York (1975)
23. Klimo, L.A. Nelson, P.I.: On conditional least squares estimation for stochastic processes. Annals of Statistics **6**, 629–642 (1978).
24. Lindgren, G.: Lectures on Stationary Stochastic Processes. Lund University (2002). http://www.maths.lu.se/matstat/staff/georg/Publications/lecture2002.pdf
25. Maruyama, G.: The harmonic analysis of stationary stochastic processes. Mem. Fac. Sci. Kyusyu Univ. **A4**, 45–106 (1949)
26. McDonald, D.J., Shalizi, C.R., Schervish, M.: Generalization error bounds for stationary autoregressive models (2012) http://arxiv.org/abs/1103.0942
27. Meir, R.: Nonparametric time series prediction through adaptive model selection. Machine Learning **39**, 5–34 (2000)
28. Morvai, G., Yakowitz, S., Györfi, L.: Nonparametric inference for ergodic, stationary time series. Annals of Statistics **24**, 370–379 (1996)
29. Ninness, B.: Strong laws of large numbers under weak assumptions with application. IEEE Trans. Automatic Control **45**, 2117–2122 (2000)
30. Pisier, G.: Probabilistic methods in the geometry of Banach spaces. In Probability and Analysis. Lecture Notes in Mathematics **1206**, pp. 167–241. Springer, New York (1986)
31. Ryabko, B.Y.: Prediction of random sequences and universal coding. Problems of Information Transmission **24**, 87–96 (1988)
32. Schäfer, D.: Strongly consistent online forecasting of centered Gaussian processes. IEEE Trans. Inform. Theory **48**, 791–799 (2002)
33. Singer, A., Feder, M.: Universal linear prediction by model order weighting. IEEE Transactions on Signal Processing **47**, 2685–2699 (1999)