Non-parametric sequential prediction of time series

Gérard BIAU [a,*], Kevin BLEAKLEY [b,c,d],
László GYÖRFI [e] and György OTTUCSÁK [e]

[a] LSTA & LPMA
Université Pierre et Marie Curie – Paris VI
Boîte 158, 175 rue du Chevaleret
75013 Paris, France
gerard.biau@upmc.fr

[b] Institut Curie, Centre de Recherche, 75248 Paris, France
[c] INSERM, U900, 75248 Paris, France
[d] Centre for Computational Biology, Ecole des Mines de Paris
35 rue St Honoré, 77305 Fontainebleau, France
kevbleakley@gmail.com

[e] Department of Computer Science and Information Theory
Budapest University of Technology and Economics
H-1117 Magyar Tudósok krt. 2, Budapest, Hungary
{gyorfi,oti}@szit.bme.hu

Abstract

Time series prediction covers a vast field of every-day statistical applications in medical, environmental and economic domains. In this paper we develop non-parametric prediction strategies based on the combination of a set of "experts" and show the universal consistency of these strategies under a minimum of conditions. We perform an in-depth analysis of real-world data sets and show that these non-parametric strategies are more flexible, faster and generally outperform ARMA methods in terms of normalized cumulative prediction error.

Index Terms — Time series, sequential prediction, universal consistency, kernel estimation, nearest neighbor estimation, generalized linear estimates.

*Corresponding author.

1 Introduction

The problem of time series analysis and prediction has a long and rich history, probably dating back to the pioneering work of Yule in 1927 [32]. The application scope is vast, as time series modeling is routinely employed across the entire and diverse range of applied statistics, including problems in genetics, medical diagnoses, air pollution forecasting, machine condition monitoring, financial investments, marketing and econometrics. Most of the research activity until the 1970s was concerned with parametric approaches to the problem whereby a simple, usually linear model is fitted to the data (for a comprehensive account we refer the reader to the monograph of Brockwell and Davies [5]). While many appealing mathematical properties of the parametric paradigm have been established, it has become clear over the years that the limitations of the approach may be rather severe, essentially due to overly rigid constraints which are imposed on the processes. One of the more promising solutions to overcome this problem has been the extension of classic non-parametric methods to the time series framework (see for example Györfi, Härdle, Sarda and Vieu [16] and Bosq [3] for a review and references).

Interestingly, related schemes have been proposed in the context of sequential investment strategies for financial markets. Sequential investment strategies are allowed to use information about the market collected from the past and determine at the beginning of a training period a portfolio, that is, a way to distribute the current capital among the available assets. Here, the goal of the investor is to maximize their wealth in the long run, without knowing the underlying distribution generating the stock prices. For more information on this subject, we refer the reader to Algoet [1], Györfi and Schäfer [21], Györfi, Lugosi and Udina [19], and Györfi, Udina and Walk [22].

The present paper is devoted to the non-parametric problem of sequential prediction of real valued sequences which we do not require to necessarily satisfy the classical statistical assumptions for bounded, autoregressive or Markovian processes. Indeed, our goal is to show powerful consistency results under a strict minimum of conditions. To fix the context, we suppose that at each time instant $n = 1, 2, \ldots$, the statistician (also called the *predictor* hereafter) *is asked to guess the next outcome* y_n of a sequence of real numbers $y_1, y_2, \ldots$ with knowledge of the past $y_1^{n-1} = (y_1, \ldots, y_{n-1})$ (where y_1^0 denotes the empty string) and the side information vectors $x_1^n = (x_1, \ldots, x_n)$, where $x_n \in \mathbb{R}^d$. In other words, adopting the perspective of on-line learning, the elements $y_0, y_1, y_2, \ldots$ and $x_1, x_2, \ldots$ are revealed one at a time, in order, beginning with $(x_1, y_0), (x_2, y_1), \ldots$, and the predictor's estimate of y_n at time n is based on the strings y_1^{n-1} and x_1^n. Formally, the strategy of the

predictor is a sequence $g = \{g_n\}_{n=1}^{\infty}$ of forecasting functions

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

and the prediction formed at time n is just $g_n(x_1^n, y_1^{n-1})$.

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

After n time instants, the (*normalized*) *cumulative squared prediction error* on the strings X_1^n and Y_1^n is

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

Ideally, the goal is to make $L_n(g)$ small. There is, however, a fundamental limit for the predictability of the sequence, which is determined by a result of Algoet [2]: for any prediction strategy g and jointly stationary ergodic process $\{(X_n, Y_n)\}_{-\infty}^{\infty}$,

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

where

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

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past observation sequences $Y_{-\infty}^{-1} = (\ldots, Y_{-2}, Y_{-1})$ and $X_{-\infty}^0 = (\ldots, X_{-2}, X_{-1})$. Generally, we cannot hope to design a strategy whose prediction error exactly achieves the lower bound L^*. Rather, we require that $L_n(g)$ gets arbitrarily close to L^* as n grows. This gives sense to the following definition:

Definition 1.1 *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,*

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

Thus, universally consistent strategies asymptotically achieve the best possible loss for all processes in the class. Algoet [1] and Morvai, Yakowitz and Györfi [26] proved that there exist universally consistent strategies with respect to the class $\mathcal{C}$ of all bounded, stationary and ergodic processes. However, the prediction algorithms discussed in these papers are either very

complex or have an unreasonably slow rate of convergence, even for well-behaved processes. Building on the methodology developed in recent years for prediction of individual sequences (see Cesa-Bianchi and Lugosi [8] for a survey and references), Györfi and Lugosi introduced in [18] a histogram-based prediction strategy which is "simple" and yet universally consistent with respect to the class $\mathcal{C}$. A similar result was also derived independently by Nobel [27]. Roughly speaking, both methods consider several partitioning estimates (called *experts* in this context) and combine them at time n according to their past performance. For this, a probability distribution on the set of experts is generated, where a "good" expert has relatively large weight, and the average of all experts' predictions is taken with respect to this distribution.

The purpose of this paper is to further investigate non-parametric expert-oriented strategies for unbounded time series prediction. With this aim in mind, in Section 2.1 we briefly recall the histogram-based prediction strategy initiated in [18], which was recently extended to unbounded processes by Györfi and Ottucsák [20]. In Section 2.2 and 2.3 we offer two "more flexible" strategies, called respectively *kernel* and *nearest neighbor-based* prediction strategies, and state their universal consistency with respect to the class of all (non-necessarily bounded) stationary and ergodic processes with finite fourth moment. In Section 2.4 we consider as an alternative a prediction strategy based on combining generalized linear estimates. In Section 2.5 we use the techniques of the previous section to give a simpler prediction strategy for stationary Gaussian ergodic processes. Extensive experimental results based on real-life data sets are discussed in Section 3, and proofs of the main results are given in Section 4.

2 Universally consistent prediction strategies

2.1 Histogram-based prediction strategy

In this section, we briefly describe the histogram-based prediction scheme due to Györfi and Ottucsák [20] for *unbounded* stationary and ergodic sequences. The strategy is defined at each time instant as a convex combination of *elementary predictors* (the so-called *experts*), where the weighting coefficients depend on the past performance of each elementary predictor. To be more precise, we first define an infinite array of experts $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ as follows. Let $\mathcal{P}_\ell = \{A_{\ell,j}, j = 1, 2, \ldots, m_\ell\}$ be a sequence of finite partitions of $\mathbb{R}^d$, and let $\mathcal{Q}_\ell = \{B_{\ell,j}, j = 1, 2, \ldots, m'_\ell\}$ be a sequence of finite partitions of

$\mathbb{R}$. Introduce the corresponding quantizers:

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

and

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

To lighten notation a bit, for any n and $x_1^n \in (\mathbb{R}^d)^n$, we write $F_\ell(x_1^n)$ for the sequence $F_\ell(x_1), \ldots, F_\ell(x_n)$ and similarly, for $y_1^n \in \mathbb{R}^n$ we write $G_\ell(y_1^n)$ for the sequence $G_\ell(y_1), \ldots, G_\ell(y_n)$.

The sequence of experts $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ is defined as follows. Let $J_n^{(k,\ell)}$ be the locations of the matches of the last seen strings x_{n-k}^n of length $k+1$ and y_{n-k}^{n-1} of length k in the past according to the quantizer with parameters k and ℓ:

$$J_n^{(k,\ell)} = \left\{ k < t < n : F_\ell(x_{t-k}^t) = F_\ell(x_{n-k}^n), G_\ell(y_{t-k}^{t-1}) = G_\ell(y_{n-k}^{n-1}) \right\},$$

and introduce the truncation function

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

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

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = T_{n^\delta}\left(\frac{1}{|J_n^{(k,\ell)}|} \sum_{\{t \in J_n^{(k,\ell)}\}} y_t \right), \qquad n > k+1,$$

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

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

Here and throughout, for any finite set J, the notation $|J|$ stands for the size of J. We note that the expert $h_n^{(k,\ell)}$ can be interpreted as a (truncated) histogram regression function estimate drawn in $(\mathbb{R}^d)^{k+1} \times \mathbb{R}^k$ (Györfi, Kohler, Krzyżak and Walk [17]).

The proposed prediction algorithm proceeds with an exponential weighting average method. Formally, let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k, ℓ) of positive integers such that for all k and ℓ, $q_{k,\ell} > 0$. Fix a learning parameter $\eta_n > 0$, and define the weights

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

and their normalized values

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

The prediction strategy g at time n is defined by

$$g_n(x_1^n, y_1^{n-1}) = \sum_{k,\ell=1}^{\infty} p_{k,\ell,n} h_n^{(k,\ell)}(x_1^n, y_1^{n-1}), \qquad n = 1, 2, \ldots$$

It is proved in [20] that this scheme is universally consistent with respect to the class of all (non-necessarily bounded) stationary and ergodic processes with finite fourth moment, as stated in the following theorem. Here and throughout the document, $\|\cdot\|$ denotes the Euclidean norm.

Theorem 2.1 (Györfi and Ottucsák [20]) *Assume that*

(a) The sequence of partitions $\mathcal{P}_\ell$ is nested, that is, any cell of $\mathcal{P}_{\ell+1}$ is a subset of a cell of $\mathcal{P}_\ell$, $\ell = 1, 2, \ldots$;

(b) The sequence of partitions $\mathcal{Q}_\ell$ is nested;

(c) The sequence of partitions $\mathcal{P}_\ell$ is asymptotically fine, i.e., if

$$\mathrm{diam}(A) = \sup_{x,y \in A} \|x - y\|$$

denotes the diameter of a set, then for each sphere S centered at the origin

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

(d) The sequence of partitions $\mathcal{Q}_\ell$ is asymptotically fine.

Then, if we choose the learning parameter η_n of the algorithm as

$$\eta_n = \frac{1}{\sqrt{n}},$$

the histogram-based prediction scheme g defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes such that

$$\mathbb{E}\{Y_0^4\} < \infty.$$

The idea of combining a collection of concurrent estimates was originally developed in a non-stochastic context for on-line sequential prediction from deterministic sequences (see Cesa-Bianchi and Lugosi [8] for a comprehensive introduction). Following the terminology of the prediction literature, the combination of different procedures is sometimes termed *aggregation* in the stochastic context. The overall goal is always the same: use aggregation to improve prediction. For a recent review and an updated list of references, see Bunea and Nobel [6] and Bunea, Tsybakov and Wegkamp [7].

2.2 Kernel-based prediction strategies

We introduce in this section a class of *kernel-based* prediction strategies for (non-necessarily bounded) stationary and ergodic sequences. The main advantage of this approach in contrast to the histogram-based strategy is that it replaces the rigid discretization of the past appearances by more flexible rules. This also often leads to faster algorithms in practical applications.

To simplify the notation, we start with the simple "moving-window" scheme, corresponding to a uniform kernel function, and treat the general case briefly later. Just like before, we define an array of experts $h^{(k,\ell)}$, where k and ℓ are positive integers. We associate to each pair (k,ℓ) two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$ such that, for any fixed k

$$\lim_{\ell\to\infty} r_{k,\ell} = 0, \tag{2}$$

and

$$\lim_{\ell\to\infty} r'_{k,\ell} = 0. \tag{3}$$

Finally, let the location of the matches be

$$J_n^{(k,\ell)} = \left\{k < t < n : \|x_{t-k}^t - x_{n-k}^n\| \le r_{k,\ell},\ \|y_{t-k}^{t-1} - y_{n-k}^{n-1}\| \le r'_{k,\ell}\right\} .$$

Then the elementary expert $h_n^{(k,\ell)}$ at time n is defined by

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = T_{\min\{n^\delta,\ell\}}\left(\frac{\sum_{\{t\in J_n^{(k,\ell)}\}} y_t}{|J_n^{(k,\ell)}|}\right), \qquad n > k+1, \tag{4}$$

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

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

The pool of experts is mixed the same way as in the case of the histogram-based strategy. That is, letting $\{q_{k,\ell}\}$ be a probability distribution over the

set of all pairs (k, ℓ) of positive integers such that $q_{k,\ell} > 0$ for all k and ℓ, for $\eta_n > 0$, we define the weights

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

together with their normalized values

$$p_{k,\ell,n} = \frac{w_{k,\ell,n}}{\sum_{i,j=1}^{\infty} w_{i,j,n}}. \tag{5}$$

The general prediction scheme g_n at time n is then defined by weighting the experts according to their past performance and the initial distribution $\{q_{k,\ell}\}$:

$$g_n(x_1^n, y_1^{n-1}) = \sum_{k,\ell=1}^{\infty} p_{k,\ell,n} h_n^{(k,\ell)}(x_1^n, y_1^{n-1}), \qquad n = 1, 2, \ldots$$

Theorem 2.2 *Denote by $\mathcal{C}$ the class of all jointly stationary and ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ such that $\mathbb{E}\{Y_0^4\} < \infty$. Choose the learning parameter η_n of the algorithm as*

$$\eta_n = \frac{1}{\sqrt{n}},$$

and suppose that (2) and (3) are verified. Then the moving-window-based prediction strategy defined above is universally consistent with respect to the class $\mathcal{C}$.

The proof of Theorem 2.2 is in Section 4. This theorem may be extended to a more general class of kernel-based strategies, as introduced in the next remark.

Remark 2.1 (General kernel function) *Define a kernel function as any map $K : \mathbb{R}_+ \to \mathbb{R}_+$. The kernel-based strategy parallels the moving-window scheme defined above, with the only difference that in definition (4) of the elementary strategy, the regression function estimate is replaced by*

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = T_{\min\{n^\delta, \ell\}} \left(\frac{\sum_{\{k<t<n\}} K\left(\|x_{t-k}^t - x_{n-k}^n\|/r_{k,\ell}\right) K\left(\|y_{t-k}^{t-1} - y_{n-k}^{n-1}\|/r'_{k,\ell}\right) y_t}{\sum_{\{k<t<n\}} K\left(\|x_{t-k}^t - x_{n-k}^n\|/r_{k,\ell}\right) K\left(\|y_{t-k}^{t-1} - y_{n-k}^{n-1}\|/r'_{k,\ell}\right)} \right).$$

Observe that if K is the naive kernel $K(x) = \mathbf{1}_{\{x \le 1\}}$ (where $\mathbf{1}$ denotes the indicator function and $x \in \mathbb{R}_+$), we recover the moving-window strategy discussed above. Typical nonuniform kernels assign a smaller weight to the

observations x_{t-k}^t and y_{t-k}^{t-1} whose distance from x_{n-k}^n and y_{n-k}^{n-1} is larger. It can be shown that kernel K can be any regular kernel (cf. Chapter 23 in Györfi, Kohler, Krzyżak and Walk [17]). For example, if K is a continuous density function with compact support and $K(0) > 0$, then it is a regular kernel. Such kernels promise a better prediction of the local structure of the conditional distribution.

2.3 Nearest neighbor-based prediction strategy

This strategy is yet more robust with respect to the kernel strategy and thus also with respect to the histogram strategy. This is because it does not suffer from the scaling problems of histogram and kernel-based strategies where the quantizer and the radius have to be carefully chosen to obtain "good" performance.

To introduce the strategy, we start again by defining an infinite array of experts $h^{(k,\ell)}$, where k and ℓ are positive integers. Just like before, k is the length of the past observation vectors being scanned by the elementary expert and, for each ℓ, choose $p_\ell \in (0,1)$ such that

$$\lim_{\ell\to\infty} p_\ell = 0\,, \tag{6}$$

and set

$$\bar{\ell} = \lfloor p_\ell n \rfloor$$

(where $\lfloor . \rfloor$ is the floor function). At time n, for fixed k and ℓ ($n > k+\bar{\ell}+1$), the expert searches for the $\bar{\ell}$ nearest neighbors (NN) of the last seen observation x_{n-k}^n and y_{n-k}^{n-1} in the past and predicts accordingly. More precisely, let

$$J_n^{(k,\ell)} = \big\{\, k < t < n : (x_{t-k}^t, y_{t-k}^{t-1}) \text{ is among the } \bar{\ell} \text{ NN of } (x_{n-k}^n, y_{n-k}^{n-1}) \text{ in } (x_1^{k+1}, y_1^k), \ldots, (x_{n-k-1}^{n-1}, y_{n-k-1}^{n-2}) \big\}$$

and introduce the elementary predictor

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = T_{\min\{n^\delta,\ell\}} \left(\frac{\sum_{\{t\in J_n^{(k,\ell)}\}} y_t}{|J_n^{(k,\ell)}|} \right)$$

if the sum is non void, and 0 otherwise. Next, set

$$0 < \delta < \frac{1}{8}.$$

Finally, the experts are mixed as before: starting from an initial probability distribution $\{q_{k,\ell}\}$, the aggregation scheme is

$$g_n(x_1^n, y_1^{n-1}) = \sum_{k,\ell=1}^{\infty} p_{k,\ell,n} h_n^{(k,\ell)}(x_1^n, y_1^{n-1}), \qquad n = 1, 2, \ldots,$$

where the probabilities $p_{k,\ell,n}$ are the same as in (5).

We say that a tie occurs with probability zero if, for any vector $\mathbf{s}$, the random variable

$$\|(X_1^{k+1}, Y_1^k) - \mathbf{s}\|$$

has a continuous distribution function.

Theorem 2.3 *Assume that a tie occurs with probability zero. Denote by $\mathcal{C}$ the class of all jointly stationary and ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ such that $\mathbb{E}\{Y_0^4\} < \infty$. Choose the parameter η_n of the algorithm as*

$$\eta_n = \frac{1}{\sqrt{n}}\,,$$

and suppose that (6) is verified. Then the nearest neighbor prediction strategy defined above is universally consistent with respect to the class $\mathcal{C}$.

A discussion on how to deal with ties that may appear in some cases can be found in [23], in the related context of portfolio selection strategies. It turns out that one can guarantee that the tie condition is satisfied if an additional dummy variable with density is included to the vectors X_n's.

The proof of Theorem 2.3 is a combination of the proof of Theorem 2.2 and the technique used in [22].

Remark 2.2

1. *The truncation index T in the definition (4) of the elementary expert $h_n^{(k,\ell)}$ is merely a technical choice that avoids having to assume that $|Y_0|$ is almost surely bounded. On the practical side, it has little influence on results for relatively short time series.*

2. *The choice of the learning parameter η_n as $1/\sqrt{n}$ ensures consistency of the method, but may be suboptimal in the finite-sample case and significantly affect the performance of the algorithm for real data. Though an optimal theoretical calibration of η_n is beyond the purpose of the present paper, a closer look at the proofs reveals that η_n and $\delta > 0$ should satisfy the conditions $n\eta_{n+1} \to \infty$ and $n^{4\delta}\eta_n \to 0$ as $n \to \infty$. A thorough discussion on the practical choice of η_n can be found in Section 3.*

   Note also that if the time series are bounded then the squared loss is not only convex but is also exp-concave. In this case (if we know the

bounds of the series in advance), we can choose an appropriate constant η that offers the fast rate $O(1/n)$ (see e.g. Cesa-Bianchi and Lugosi [8]). However, in the unbounded case, this technique does not work in plug-in style. Some new results show that, in some special cases, it is possible to obtain fast rates (Moon and Weissman [25]). However, for regression, it is still an open problem to find an algorithm achieving such rates.

2.4 Generalized linear prediction strategy

This section is devoted to an alternative way of defining a universal predictor for stationary and ergodic processes. It is in effect an extension of the approach presented in Györfi and Lugosi [18] to non-necessarily bounded processes. Once again, we apply the method described in the previous sections to combine elementary predictors, but now we use elementary predictors which are generalized linear predictors. More precisely, we define an infinite array of elementary experts $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ as follows. Let $\{\phi_j^{(k)}\}_{j=1}^{\ell}$ be real-valued functions defined on $(\mathbb{R}^d)^{(k+1)} \times \mathbb{R}^k$. The elementary predictor $h_n^{(k,\ell)}$ generates a prediction of form

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = T_{\min\{n^\delta, \ell\}} \left(\sum_{j=1}^{\ell} c_{n,j} \phi_j^{(k)}(x_{n-k}^n, y_{n-k}^{n-1}) \right) ,$$

where the coefficients $c_{n,j}$ are calculated according to the past observations x_1^n, y_1^{n-1}, and

$$0 < \delta < \frac{1}{8}.$$

Formally, the coefficients $c_{n,j}$ are defined as the real numbers which minimize the criterion

$$\sum_{t=k+1}^{n-1} \left(\sum_{j=1}^{\ell} c_j \phi_j^{(k)}(x_{t-k}^t, y_{t-k}^{t-1}) - y_t \right)^2 \tag{7}$$

if $n > k + 1$, and the all-zero vector otherwise. It can be shown using a recursive technique (see e.g., Tsypkin [31], Györfi [15], Singer and Feder [29], and Györfi and Lugosi [18]) that the $c_{n,j}$ can be calculated with small computational complexity.

The experts are mixed via an exponential weighting, which is defined the same way as earlier. Thus, the aggregated prediction scheme is

$$g_n(x_1^n, y_1^{n-1}) = \sum_{k,\ell=1}^{\infty} p_{k,\ell,n} h_n^{(k,\ell)}(x_1^n, y_1^{n-1}), \qquad n = 1, 2, \ldots,$$

where the $p_{k,\ell,n}$ are calculated according to (5).

Combining the proof of Theorem 2.2 and the proof of Theorem 2 in [18] leads to the following result:

Theorem 2.4 *Suppose that* $|\phi_j^{(k)}| \leq 1$ *and, for any fixed* k*, suppose that the set*

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

is dense in the set of continuous functions of $d(k+1)+k$ *variables. Then the generalized linear prediction strategy defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes such that*

$$\mathbb{E}\{Y_0^4\} < \infty.$$

We give a sketch of the proof of Theorem 2.4 in Section 4.

2.5 Prediction of Gaussian processes

We consider in this section the classical problem of Gaussian time series prediction (cf. Brockwell and Davis [5]). In this context, parametric models based on distribution 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 [13], Gerencsér [11, 12], Goldenshluger and Zeevi [14]). However, in the spirit of modern non-parametric inference, we try to avoid such restrictions on the process structure. Thus, we only assume that we observe a 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. Note that there are no side information vectors x_1^n in this purely time series prediction framework.

It is well known for Gaussian time series that the best predictor is a linear function of the past:

$$\mathbb{E}\{Y_n \mid Y_{n-1}, Y_{n-2}, \ldots\} = \sum_{j=1}^{\infty} c_j^* Y_{n-j},$$

where the c_j^* minimize the criterion

$$\mathbb{E}\left\{ \left(\sum_{j=1}^{\infty} c_j Y_{n-j} - Y_n \right)^2 \right\}.$$

Following Györfi and Lugosi [18], we extend the principle of generalized linear estimates to the prediction of Gaussian time series by considering the special case

$$\phi_j^{(k)}(y_{n-k}^{n-1}) = y_{n-j}\mathbf{1}_{\{1\le j\le k\}},$$

i.e.,

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

Once again, the coefficients $c_{n,j}$ are calculated according to the past observations y_1^{n-1} by minimizing the criterion:

$$\sum_{t=k+1}^{n-1}\left(\sum_{j=1}^{k} c_j y_{t-j} - y_t\right)^2$$

if $n > k$, and the all-zero vector otherwise.

With respect to the combination of elementary experts $\tilde{h}^{(k)}$, Györfi and Lugosi applied in [18] 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.

In this section we propose a much simpler procedure which avoids in particular the doubling-trick. To begin, we set

$$h_n^{(k)}(y_1^{n-1}) = T_{\min\{n^\delta,k\}}\left(\tilde{h}_n^{(k)}(y_1^{n-1})\right),$$

where

$$0 < \delta < \frac{1}{8},$$

and combine these experts as before. Precisely, let $\{q_k\}$ be an arbitrarily probability distribution over the positive integers such that for all k, $q_k > 0$, and for $\eta_n > 0$, define the weights

$$w_{k,n} = q_k e^{-\eta_n (n-1) L_{n-1}(h_n^{(k)})}$$

and their normalized values

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

The prediction strategy g at time n is defined by

$$g_n(y_1^{n-1}) = \sum_{k=1}^{\infty} p_{k,n} h_n^{(k)}(y_1^{n-1}), \qquad n = 1, 2, \ldots$$

By combining the proof of Theorem 2.2 and Theorem 3 in [18], we obtain the following result:

Theorem 2.5 *The linear prediction strategy g defined above is universally consistent with respect to the class of all jointly stationary and ergodic zero-mean Gaussian processes.*

The following corollary shows that the strategy g provides asymptotically a good estimate of the regression function in the following sense:

Corollary 2.1 (Györfi and Ottucsák [20]) *Under the conditions of Theorem 2.5,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\{Y_t \mid Y_1^{t-1}\} - g(Y_1^{t-1}) \right)^2 = 0 \quad \textit{almost surely.}$$

Corollary 2.1 is expressed in terms of an almost sure Cesáro consistency. It is an open problem to know whether there exists a prediction rule g such that

$$\lim_{n\to\infty} \left(\mathbb{E}\{Y_n | Y_1^{n-1}\} - g(Y_1^{n-1}) \right) = 0 \quad \text{almost surely} \tag{8}$$

for all stationary and ergodic Gaussian processes. Schäfer [28] proved that, under some conditions on the time series, the consistency (8) holds.

3 Experimental results and analyses

We evaluated the performance of the histogram, moving-window kernel, NN and Gaussian process strategies on two real world data sets. Furthermore, we compared these performances to those of the standard ARMA family of methods on the same data sets. We show in particular that the four methods presented in this paper usually perform better than the best ARMA results, with respect to three different criteria.

The two real-world time series we investigated were the monthly USA unemployment rate from January 1948 until March 2007 (710 points) and daily USA federal funds interest rate from 12 January 2003 until 21 March 2007 (1200 points) respectively, extracted from the website *economagic.com*. In order to remove first-order trends, we transformed these time series into time series of *percentage change* compared to the previous month or day, respectively. The resulting time series are shown in Figs 1 and 2.

Before testing the four methods of the present paper alongside the ARMA methods, we tested whether the resulting time series were trend/level stationary using two standard tests, the KPSS test [24] and the PP test [10]. For both series using the KPSS test, we did not reject the null hypothesis of level stationarity at $p = .01, .05$ and $.1$ respectively, and for both series using

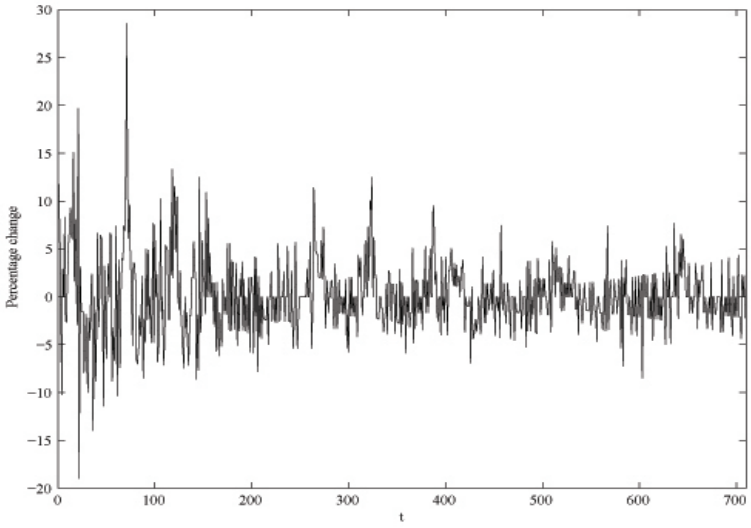

Figure 1: Monthly percentage change in USA unemployment rate from January 1948 to March 2007.

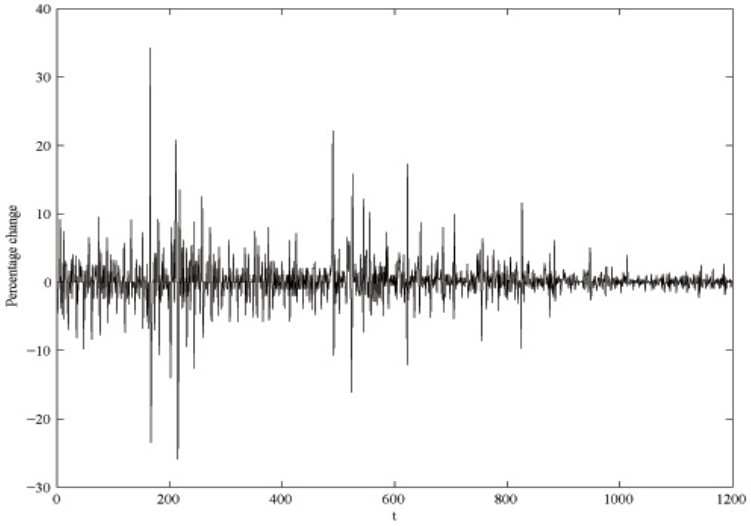

Figure 2: Daily percentage change in USA federal funds interest rate from 12 January 2003 to 21 March 2007.

the PP test (which has for null hypothesis the existence of a unit root and for alternative hypothesis, level stationarity), the null hypothesis was rejected at $p = .01, .05$ and $.1$.

We remark that this means the ARIMA(p, d, q) family of models, richer than ARMA(p, q) is unnecessary, or equivalently, we need only to consider the ARIMA family ARIMA$(p, 0, q)$. As well as this, the Gaussian process method requires the normality of the data. Since the original data in both data sets is discretized (and not very finely), this meant that the data, when transformed into percentage changes only took a small number of fixed values. This had the consequence that directly applying standard normality tests gave curious results even when histograms of the data appeared to have near-perfect Gaussian forms; however adding small amounts of random noise to the data allowed us to not systematically reject the hypothesis of normality.

Given each method and each time series $(y_1, \ldots, y_m)$ (here, $m = 710$ or 1200), we used the data $(y_1, \ldots, y_n), n < m$, to predict the value of y_{n+1}. We used three criteria to measure the quality of the overall set of predictions. First, as described in the present paper, we calculated the normalized cumulative prediction squared error L_m. Secondly, we calculated L_m^{50}, the normalized cumulative prediction error over only the last 50 predictions of the time series in order to see how the method was working after having learned nearly the whole time series. Thirdly, since in practical situations we may want to predict only the *direction* of change, we compared the direction (positive or negative) of the last 50 predicted points with respect to each previous, known point, to the 50 real directions. This gave us the criteria A^{50}: the *percentage of the direction of the last 50 points correctly predicted.*

The histogram, kernel and NN strategies were implemented in Matlab and the ARMA and Gaussian strategies in **R** using the *arima* function with d set to 0. As in [19] and [22], for practical reasons we initially chose a finite grid of experts: $k = 1, \ldots, K$ and $\ell = 1, \ldots, L$ for the histogram, kernel and NN strategies, fixing $K = 5$ and $L = 10$. For the histogram strategy we partitioned the space into each of $\{2^2, 2^3, \ldots, 2^{11}\}$ equally sized intervals, for the kernel strategy we let the radius $r'_{k,\ell}$ take the values $r'_{k,\ell} \in \{.001, .005, .01, .05, .1, .5, 1, 5, 10, 50\}$ and for the NN strategy we set $\bar{\ell} = \ell$. We initially fixed η as constants $\eta = 1/6540$ and $\eta = 1/10000$ respectively for the unemployment and interest rate data. Furthermore, we fixed the probability distribution $\{q_{k,\ell}\}$ as the uniform distribution over the $K \times L$ experts. For the Gaussian process method, we simply let $K = 5$ and fixed the probability distribution $\{q_k\}$ as the uniform distribution over the K experts. Running times on the unemployment rate data were respectively around 3 mins, 4 secs and 4 secs for the histogram, kernel and NN meth-

ods, around 4 mins for the Gaussian method and from 15 secs to 2 mins for the various ARMA methods. Running times for the interest rate data were approximately double the above figures.

Used to compare standard methods with the present non-parametric strategies, the ARMA(p, q) algorithm was run for all pairs $(p, q) \in \{0, 1, 2, 3, 4, 5\}^2$, i.e., 36 independent trials. In the ARMA experiments, at each time instant we simply used ARMA to output the real-valued prediction of the next point, and continued until we reached the end of the time series, calculating the cumulative error in exactly the same way as for the other methods. Tables 1 and 2 show the histogram, kernel, NN, Gaussian process and ARMA results for the unemployment and interest rate time series respectively. The three ARMA results shown in each table are those which had the best L_m, L_m^{50} and A^{50} respectively (sometimes two or more had the same A^{50}, in which case we chose one of these randomly). The best results with respect to each of the three criteria are shown in bold.

	L_m	L_m^{50}	A^{50}
histogram	15.66	**4.82**	68
kernel	15.44	4.99	68
NN	**15.40**	4.97	70
Gaussian	16.35	5.02	76
ARMA$(1, 1)$	16.26	5.31	72
ARMA$(0, 0)$	16.68	4.86	**78**
ARMA$(2, 0)$	16.46	5.12	**78**

Table 1: Results for histogram, kernel, NN, Gaussian process and ARMA prediction methods on the monthly percentage change in USA unemployment rate from January 1948 until March 2007. The three ARMA results are those which performed the best in terms of the L_m, L_m^{50} and A^{50} criteria respectively.

We see via Tables 1 and 2 that the histogram, kernel and NN strategies presented here outperform all 36 possible ARMA(p, q) models $(0 \leq p, q \leq 5)$ in terms of normalized cumulative prediction error L_m, and that the Gaussian process method performs similarly to the best ARMA method. In terms of the L_m^{50} and A^{50} criteria, all of the present methods and the best ARMA method provide broadly similar results.

Using the NN algorithm and the unemployment rate data as a test case, we also performed a critical analysis of the influence of the choices of η, K and L on experimental results. First, for $K = 5$ and $L = 10$ fixed, we repeated the same experiment for a variety of fixed, constant values of η. Results with respect to the L_m and L_m^{50} criteria are shown in Fig. 3(A) and (B). We see that

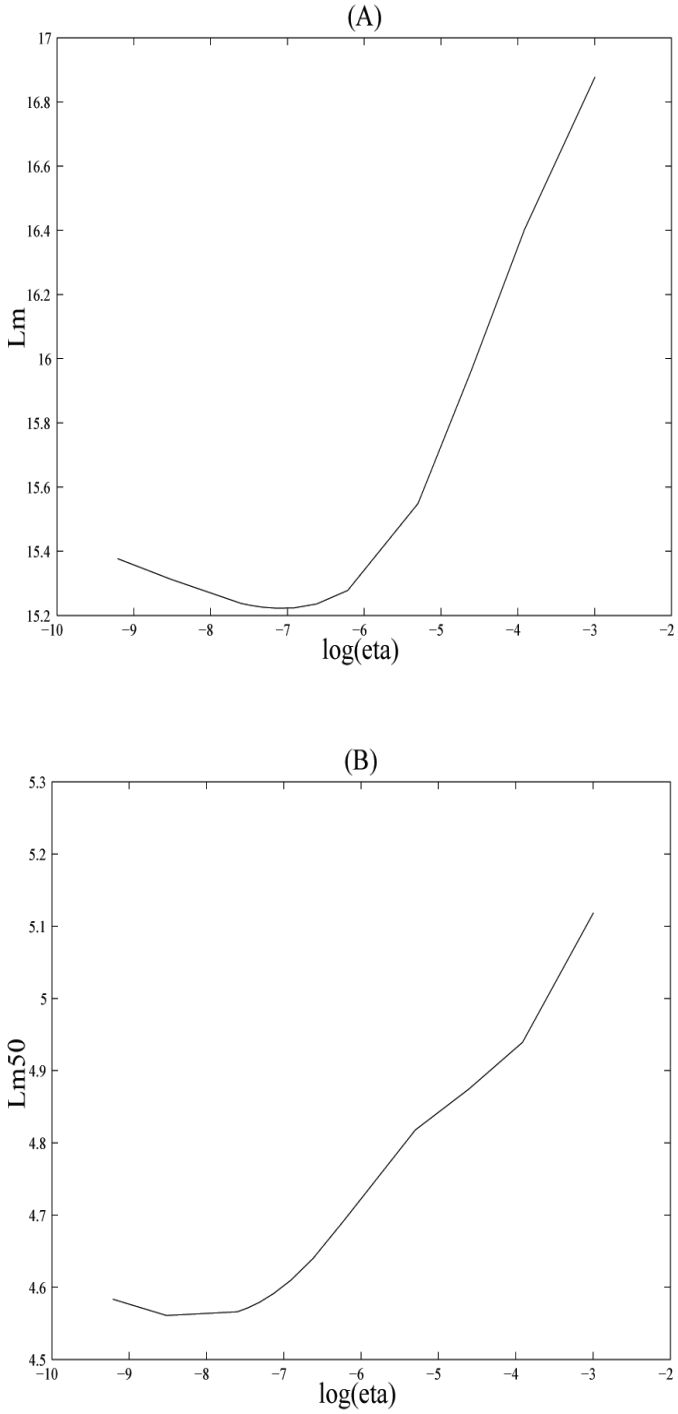

Figure 3: Effect of the choice of η on the values of (A) L_m and (B) L_m^{50}.

	L_m	L_m^{50}	A^{50}
histogram	9.78	**0.52**	**88**
kernel	**9.77**	0.57	86
NN	9.86	0.79	80
Gaussian	9.98	0.62	82
ARMA$(1,1)$	9.90	0.78	70
ARMA$(0,1)$	10.30	0.60	82
ARMA$(3,0)$	10.12	0.63	**88**

Table 2: Results for histogram, kernel, NN, Gaussian process and ARMA prediction methods on the daily percentage change in the USA federal funds interest rate from 12 January 2003 until 21 March 2007. The three ARMA results are those which performed the best in terms of the L_m, L_m^{50} and A^{50} criteria respectively.

the minimum for L_m (≈ 15.23) is found somewhere slightly below $\eta = 1/1000$ (or $\log 1/1000 = -6.91$), and close to $\eta = 1/5000$ ($\log 1/5000 = -8.52$) with respect to L_m^{50} (≈ 5.56). The theoretical results of the present article set $\eta_n = 1/\sqrt{n}$, i.e., varying with n. In the present experiment, this gave an L_m result of 17.21 and an L_m^{50} result of 5.05, suboptimal with respect to the best choice of fixed η. The reason for this is that the numerical efficiency of the algorithm is highly dependent on the relative sizes of $\eta(n-1)$ and the cumulative squared error L_{n-1} in the negative exponential used to calculate the weights of experts. If they are both too large, the computer can calculate the weights to be extremely close to zero (or even numerically zero) if we end up taking the negative of too large an exponential value, leading either to a slight instability in the calculation of weights, as was the case for the choice $\eta_n = 1/\sqrt{n}$, or divide by zero errors in extreme cases.

In practice, choosing a fixed "small enough" η should be sufficient to keep the sum of weights different to zero as the algorithm performs its first iterations. It only needs to be chosen to be small enough to balance the maximum imagineable cumulative squared error in the first iterations, which is feasible even if we have no prior knowledge of the data. For example, fixing η to balance the highly unlikely case of a 50% unemployment rate change should cover all reasonable scenarios. If in practice the time series being used are much longer than the ones considered here, after a certain number of iterations it may be necessary to let η start to tend to zero with $1/\sqrt{n}$ to avoid the same kind of numerical difficulties mentioned above.

Second, we fixed $\eta = 1/2000$ and varied (K, L) over $(i, j), 1 \leq i, j \leq 15$. We found that as long as both $K > 3$ and $L > 3$, the final values of L_m and L_m^{50} were relatively stable across the grid, though both very slowly decreasing

(i.e., improving) to minima at $(K, L) = (15, 15)$. The conclusion that the larger K and L, the better, is nevertheless misleading. For each K and L, the algorithm has to be able to initialize itself, and for this it (the NN algorithm) needs the first $K + L - 1$ data points. Thus, for larger K and L, the algorithm runs on the same dataset but does not perform quite as many iterations as for smaller K and L, and in particular avoids making any big mistakes which could be made within the first $K + L - 1$ points by choices of smaller K and/or L. As the considered series are relatively short, this difference of a few iterations at the start of the learning algorithm actually has an effect on the final values of L_m and L_m^{50}. In practice, we suggest making an ad hoc choice for K and L of at least 4, and as large as possible, given running time constraints and knowledge of the numerical stability of the algorithm being used.

4 Proofs

4.1 Proof of Theorem 2.2

The proof of Theorem 2.2 strongly relies on the following two lemmas. The first one is known as Breiman's generalized ergodic theorem.

Lemma 4.1 (Breiman [4]) *Let $Z = \{Z_n\}_{-\infty}^{\infty}$ be a stationary and ergodic process. For each positive integer t, let T^t denote the left shift operator, shifting any sequence $\{\ldots, z_{-1}, z_0, z_1, \ldots\}$ by t digits to the left. Let $\{f_t\}_{t\geq 1}$ be a sequence of real-valued functions such that $\lim_{t\to\infty} f_t(Z) = f(Z)$ almost surely for some function f. Suppose that $\mathbb{E}\sup_t |f_t(Z)| < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} f_t(T^t Z) = \mathbb{E}\{f(Z)\} \quad \textit{almost surely.}$$

Lemma 4.2 (Györfi and Ottucsák [20]) *Let $h^{(1)}, h^{(2)}, \ldots$ 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 any expert $h = \{h_n\}_{n=1}^{\infty}$ by*

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

where the loss function $\mathcal{L}$ is convex in its first argument h_t. Define

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

where $\eta_n > 0$ is monotonically decreasing, and set

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

If the prediction strategy $g = \{g_n\}_{n=1}^{\infty}$ is defined by

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

then, for every $n \geq 1$,

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

Proof of Theorem 2.2. Because of (1) it is enough to show that

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

With this in mind, we introduce the following notation:

$$\widehat{E}_n^{(k,\ell)}(X_1^n, Y_1^{n-1}, \mathbf{z}, \mathbf{s}) = \frac{\sum_{\left\{k<t<n:\ \|x_{t-k}^t - \mathbf{z}\| \leq r_{k,\ell},\ \|y_{t-k}^{t-1} - \mathbf{s}\| \leq r'_{k,\ell}\right\}} Y_t}{\left|\left\{k < t < n :\ \|x_{t-k}^t - \mathbf{z}\| \leq r_{k,\ell},\ \|y_{t-k}^{t-1} - \mathbf{s}\| \leq r'_{k,\ell}\right\}\right|}$$

for all $n > k+1$, where $0/0$ is defined to be 0, $\mathbf{z} \in (\mathbb{R}^d)^{k+1}$ and $\mathbf{s} \in \mathbb{R}^k$. Thus, for any $h^{(k,\ell)}$, we can write

$$h_n^{(k,\ell)}(X_1^n, Y_1^{n-1}) = T_{\min\{n^\delta, \ell\}} \left(\widehat{E}_n^{(k,\ell)}(X_1^n, Y_1^{n-1}, X_{n-k}^n, Y_{n-k}^{n-1}) \right).$$

By a double application of the ergodic theorem, as $n \to \infty$, almost surely, for a fixed $\mathbf{z} \in (\mathbb{R}^d)^{k+1}$ and $\mathbf{s} \in \mathbb{R}^k$, we may write

$$\begin{aligned}
\widehat{E}_n^{(k,\ell)}(X_1^n, Y_1^{n-1}, \mathbf{z}, \mathbf{s}) &= \frac{\frac{1}{n}\sum_{\left\{k<t<n:\ \|X_{t-k}^t - \mathbf{z}\| \leq r_{k,\ell},\ \|Y_{t-k}^{t-1} - \mathbf{s}\| \leq r'_{k,\ell}\right\}} Y_t}{\frac{1}{n}\left|\left\{k < t < n :\ \|X_{t-k}^t - \mathbf{z}\| \leq r_{k,\ell},\ \|Y_{t-k}^{t-1} - \mathbf{s}\| \leq r'_{k,\ell}\right\}\right|} \\
&\to \frac{\mathbb{E}\{Y_0 \mathbf{1}_{\{\|X_{-k}^0 - \mathbf{z}\| \leq r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \leq r'_{k,\ell}\}}\}}{\mathbb{P}\left\{\|X_{-k}^0 - \mathbf{z}\| \leq r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \leq r'_{k,\ell}\right\}} \\
&= \mathbb{E}\{Y_0 \,|\, \|X_{-k}^0 - \mathbf{z}\| \leq r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \leq r'_{k,\ell}\}.
\end{aligned}$$

Therefore, for all $\mathbf{z}$ and $\mathbf{s}$,

$$\begin{aligned}
&\lim_{n\to\infty} T_{\min\{n^\delta, \ell\}} \left(\widehat{E}_n^{(k,\ell)}(X_1^n, Y_1^{n-1}, \mathbf{z}, \mathbf{s}) \right) \\
&\quad = \quad T_\ell \left(\mathbb{E}\{Y_0 \,|\, \|X_{-k}^0 - \mathbf{z}\| \leq r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \leq r'_{k,\ell}\} \right) \\
&\quad \stackrel{\text{def}}{=} \quad \varphi_{k,\ell}(\mathbf{z}, \mathbf{s}).
\end{aligned}$$

Thus, by Lemma 4.1, as $n \to \infty$, almost surely,

$$\begin{aligned} L_n(h^{(k,\ell)}) &= \frac{1}{n}\sum_{t=1}^{n}(h_t^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t)^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}\left(T_{\min\{t^\delta,\ell\}}\left(\widehat{E}_t^{(k,\ell)}(X_1^t, Y_1^{t-1}, X_{t-k}^t, Y_{t-k}^{t-1})\right) - Y_t\right)^2 \\ &\to \mathbb{E}\left\{(\varphi_{k,\ell}(X_{-k}^0, Y_{-k}^{-1}) - Y_0)^2\right\} \\ &\stackrel{\text{def}}{=} \varepsilon_{k,\ell}\ . \end{aligned}$$

Denote, for Borel sets $A \subset (\mathbb{R}^d)^{k+1}$ and $B \subset \mathbb{R}^k$,

$$\mu_k(A, B) \stackrel{\text{def}}{=} \mathbb{P}\{X_{-k}^0 \in A, Y_{-k}^{-1} \in B\},$$

and set

$$\psi_k(\mathbf{z}, \mathbf{s}) \stackrel{\text{def}}{=} \mathbb{E}\{Y_0 \,|\, X_{-k}^0 = \mathbf{z}, Y_{-k}^{-1} = \mathbf{s}\}.$$

Next, let $S_{\mathbf{s},r}$ denote the closed ball with center $\mathbf{s}$ and radius r. Let

$$\tilde{\varphi}_{k,\ell}(\mathbf{z}, \mathbf{s}) \stackrel{\text{def}}{=} \mathbb{E}\{Y_0 \,|\, \|X_{-k}^0 - \mathbf{z}\| \le r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \le r'_{k,\ell}\},$$

then for any $\mathbf{z}$ and $\mathbf{s}$ which are in the support of μ_k, we have

$$\begin{aligned} \varphi_{k,\ell}(\mathbf{z}, \mathbf{s}) &= T_\ell\left(\mathbb{E}\{Y_0 \,|\, \|X_{-k}^0 - \mathbf{z}\| \le r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \le r'_{k,\ell}\}\right) \\ &= T_\ell\left(\frac{\mathbb{E}\{Y_0 \mathbf{1}_{\{\|X_{-k}^0 - \mathbf{z}\| \le r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \le r'_{k,\ell}\}}\}}{\mathbb{P}\left\{\|X_{-k}^0 - \mathbf{z}\| \le r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \le r'_{k,\ell}\right\}}\right) \\ &= T_\ell\left(\frac{1}{\mu_k(S_{\mathbf{z},r_{k,\ell}}, S_{\mathbf{s},r'_{k,\ell}})}\int_{x \in S_{\mathbf{z},r_{k,\ell}},\ y \in S_{\mathbf{s},r'_{k,\ell}}} \tilde{\varphi}_{k,\ell}(x,y)\,\mu_k(\mathrm{d}x, \mathrm{d}y)\right) \\ &\to \psi_k(\mathbf{z}, \mathbf{s}), \end{aligned}$$

as $\ell \to \infty$ and for μ_k-almost all $\mathbf{s}$ and $\mathbf{z}$ by the Lebesgue density theorem (see Györfi, Kohler, Krzyżak and Walk [17], Lemma 24.5). Therefore,

$$\lim_{\ell\to\infty} \varphi_{k,\ell}(X_{-k}^0, Y_{-k}^{-1}) = \psi_k(X_{-k}^0, Y_{-k}^{-1}) \quad \text{almost surely.}$$

Observe that

$$\begin{aligned} \varphi_{k,\ell}^2(\mathbf{z}, \mathbf{s}) &= \left[T_\ell\left(\mathbb{E}\{Y_0 \,|\, \|X_{-k}^0 - \mathbf{z}\| \le r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \le r'_{k,\ell}\}\right)\right]^2 \\ &\le \left(\mathbb{E}\{Y_0 \,|\, \|X_{-k}^0 - \mathbf{z}\| \le r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \le r'_{k,\ell}\}\right)^2 \\ &\qquad \text{(since } |T_\ell(z)| \le |z|\text{)} \\ &\le \mathbb{E}\{Y_0^2 \,|\, \|X_{-k}^0 - \mathbf{z}\| \le r_{k,\ell},\ \|Y_{-k}^{-1} - \mathbf{s}\| \le r'_{k,\ell}\} \\ &\qquad \text{(by Jensen's inequality).} \end{aligned}$$

Consequently,

$$\sup_{\ell \geq 1} \mathbb{E}\{\varphi_{k,\ell}^2(X_{-k}^0, Y_{-k}^{-1})\} \leq \mathbb{E}Y_0^2 < \infty,$$

due to the assumptions of the theorem. Therefore, for fixed k the sequence of random variables $\{\varphi_{k,\ell}(X_{-k}^0, Y_{-k}^{-1})\}_{\ell=1}^{\infty}$ is uniformly integrable and by using the dominated convergence theorem we obtain

$$\begin{aligned}
\lim_{\ell \to \infty} \varepsilon_{k,\ell} &= \lim_{\ell \to \infty} \mathbb{E}\left\{(\varphi_{k,\ell}(X_{-k}^0, Y_{-k}^{-1}) - Y_0)^2\right\} \\
&= \mathbb{E}\left\{\left(\mathbb{E}\{Y_0 | X_{-k}^0, Y_{-k}^{-1}\} - Y_0\right)^2\right\} \\
&\stackrel{\text{def}}{=} \varepsilon_k.
\end{aligned}$$

Invoking the martingale convergence theorem (see, e.g., Stout [30]), we then have

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

and consequently,

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

We next apply Lemma 4.2 with the choice $\eta_n = 1/\sqrt{n}$ and the squared loss

$$\mathcal{L}(h_t, Y_t) = (h_t - Y_t)^2.$$

We obtain

$$\begin{aligned}
L_n(g) \leq \inf_{k,\ell} \Bigg(& L_n(h^{(k,\ell)}) - \frac{2 \ln q_{k,\ell}}{\sqrt{n}} \Bigg) \\
& + \frac{1}{2n} \sum_{t=1}^{n} \frac{1}{\sqrt{t}} \sum_{k,\ell=1}^{\infty} p_{k,\ell,t} \left(h_t^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t \right)^4.
\end{aligned}$$

On one hand, almost surely,

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

On the other hand,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell=1}^{\infty}p_{k,\ell,t}(h_t^{(k,\ell)}(X_1^t,Y_1^{t-1})-Y_t)^4\\
&\quad\le\frac{8}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell=1}^{\infty}p_{k,\ell,t}\left(h_t^{(k,\ell)}(X_1^t,Y_1^{t-1})^4+Y_t^4\right)\\
&\quad\le\frac{8}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k=1}^{\infty}\left(\sum_{\ell=1}^{\lfloor t^\delta\rfloor}p_{k,\ell,t}\ell^4+\sum_{\ell=\lceil t^\delta\rceil}^{\infty}p_{k,\ell,t}t^{4\delta}+\sum_{\ell=1}^{\infty}p_{k,\ell,t}Y_t^4\right)\\
&\quad\le\frac{8}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell=1}^{\infty}p_{k,\ell,t}\left(t^{4\delta}+Y_t^4\right)\\
&\quad=\frac{8}{n}\sum_{t=1}^{n}\frac{t^{4\delta}+Y_t^4}{\sqrt{t}}.
\end{aligned}$$

Therefore, almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell=1}^{\infty}p_{k,\ell,t}(h_t^{(k,\ell)}(X_1^t,Y_1^{t-1})-Y_t)^4\\
&\quad\le\quad\limsup_{n\to\infty}\frac{8}{n}\sum_{t=1}^{n}\frac{Y_t^4}{\sqrt{t}}\\
&\quad=0\\
&\qquad\text{(since }\delta<1/8\text{ and }\mathbb{E}\{Y_0^4\}<\infty\text{)}.
\end{aligned}$$

Summarizing these bounds, we get that, almost surely,

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

and the theorem is proved. □

4.2 Sketch of the proof of Theorem 2.4

For fixed k and ℓ, let

$$(c_1^*,\ldots,c_\ell^*)\in\arg\min_{(c_1,\ldots,c_\ell)}\mathbb{E}\left\{\left(\sum_{j=1}^{\ell}c_j\phi_j^{(k)}(X_{-k}^0,Y_{-k}^{-1})-Y_0\right)^2\right\}.$$

Then, following the proof of Theorem 2 in [18] one can show that for all $j\in\{1,\ldots,\ell\}$,

$$\lim_{n\to\infty}c_{n,j}=c_j^*\quad\text{almost surely,}\tag{9}$$

where the $c_{n,j}$ are defined in (7). Using equality (9) and Lemma 4.1, for any fixed k and ℓ we obtain that, almost surely,

$$\begin{aligned}
\lim_{n\to\infty} L_n(h^{(k,\ell)}) &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=k+1}^{n} \left(h_t^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t \right)^2 \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{t=k+1}^{n} \left(T_{\min\{t^\delta,\ell\}} \left(\sum_{j=1}^{\ell} c_{t,j} \phi_j^{(k)}(X_{t-k}^t, Y_{t-k}^{t-1}) \right) - Y_t \right)^2 \\
&= \mathbb{E}\left\{ \left(T_\ell \left(\sum_{j=1}^{\ell} c_j^* \phi_j^{(k)}(X_{-k}^0, Y_{-k}^{-1}) \right) - Y_0 \right)^2 \right\} \\
&\stackrel{\text{def}}{=} \varepsilon_{k,\ell}.
\end{aligned}$$

Then, with similar arguments to Theorem 2 in [18], it can be shown that

$$\lim_{k,\ell\to\infty} \varepsilon_{k,\ell} \leq L^*.$$

Finally, by using Lemma 4.2, the assumptions $\delta < 1/8$ and $\mathbb{E}\{Y_0^4\} < \infty$, and repeating the arguments of the proof of Theorem 2.2, we obtain

$$\limsup_{n\to\infty} L_n(g) \leq \inf_{k,\ell} \varepsilon_{k,\ell} \leq L^*,$$

as desired. □

Acknowledgements. We thank two referees for valuable comments and insightful suggestions.

References

[1] Algoet, P. Universal schemes for prediction, gambling and portfolio selection, *Ann. Probab.*, Vol. 20, pp. 901–941, 1992.

[2] Algoet, P. The strong law of large numbers for sequential decisions under uncertainty, *IEEE Trans. Inform. Theory*, Vol. 40, pp. 609–633, 1994.

[3] Bosq, D. *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*, Lecture Notes in Statistics, 110, Springer-Verlag, New York, 1996.

[4] Breiman, L. The individual ergodic theorem of information theory, *Ann. Math. Statist.*, Vol. 28, pp. 809–811, 1957. Correction. *Ann. Math. Statist.*, Vol. 31, pp. 809–810, 1960.

[5] Brockwell, P. and Davis, R. A. *Time Series: Theory and Methods*, Second edition, Springer-Verlag, New York, 1991.

[6] Bunea, F. and Nobel, A. Sequential procedures for aggregating arbitrary estimators of a conditional mean, *IEEE Trans. Inform. Theory*, Vol. 54, pp. 1725–1735, 2008.

[7] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. Aggregation for Gaussian regression, *Ann. Statist.*, Vol. 35, pp. 1674–1697, 2007.

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

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

[10] Durlauf, S. N. and Phillips, P. C. B. Trends versus random walks in time series analysis, *Econometrica*, Vol. 56, pp. 1333–1354, 1988.

[11] Gerencsér, L. $AR(\infty)$ estimation and nonparametric stochastic complexity, *IEEE Trans. Inform. Theory*, Vol. 38, pp. 1768–1779, 1992.

[12] Gerencsér, L. On Rissanen's predictive stochastic complexity for stationary ARMA processes, *J. Statist. Plann. Inference*, Vol. 41, pp. 303–325, 1994.

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

[14] Goldenshluger, A. and Zeevi, A. Nonasymptotic bounds for autoregressive time series modeling, *Ann. Statist.*, Vol. 29, pp. 417–444, 2001.

[15] Györfi, L. Adaptive linear procedures under general conditions, *IEEE Trans. Inform. Theory*, Vol. 30, pp. 262–267, 1984.

[16] Györfi, L., Härdle, W., Sarda, P. and Vieu, P. *Nonparametric Curve Estimation from Time Series*, Lecture Notes in Statistics, 60, Springer-Verlag, Berlin, 1989.

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

[18] Györfi, L. and Lugosi, G. Strategies for sequential prediction of stationary time series, in *Modeling Uncertainty*, Internat. Ser. Oper. Res. Management Sci., Vol. 46, pp. 225–248, Kluweer Acad. Publ., Boston, 2001.

[19] Györfi, L., Lugosi, G. and Udina, F. Nonparametric kernel-based sequential investment strategies, *Math. Finance*, Vol. 16, pp. 337–357, 2006.

[20] Györfi, L. and Ottucsák, G. Sequential prediction of unbounded time series, *IEEE Trans. Inform. Theory*, Vol. 53, pp. 1866–1872, 2007.

[21] Györfi, L. and Schäfer, D. Nonparametric prediction, in J. A. K. Suykens, G. Horváth, S. Basu, C. Micchelli and J. Vandevalle, editors, *Advances in Learning Theory: Methods, Models and Applications*, pp. 339–354, IOS Press, NATO Science Series, 2003.

[22] Györfi, L., Udina, F. and Walk, H. Nonparametric nearest neighbor based empirical portfolio selection strategies, *Statist. Decis.*, Vol. 26, pp. 145–157, 2008.

[23] Györfi, L., Udina, F. and Walk, H. Experiments on universal portfolio selection using data from real markets, *Technical Report*, 2008. http://tukey.upf.es/papers/NNexp.pdf

[24] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root, *J. Econometrics*, Vol. 54, pp. 159–178, 1992.

[25] Moon, T. and Weissman, T. Universal FIR MMSE filtering, *Technical Report*, 2008. http://www.stanford.edu/~tsmoon/J3.pdf

[26] Morvai, G., Yakowitz, S. and Györfi, L. Nonparametric inference for ergodic, stationary time series, *Ann. Statist.*, Vol. 24, pp. 370–379, 1996.

[27] Nobel, A. On optimal sequential prediction for general processes, *IEEE Trans. Inform. Theory*, Vol. 49, pp. 83–98, 2003.

[28] Schäfer, D. Strongly consistent online forecasting of centered Gaussian processes, *IEEE Trans. Inform. Theory*, Vol. 48, pp. 791–799, 2002.

[29] Singer, A. C. and Feder, M. Universal linear least-squares prediction, *Proceedings of the 2000 International Symposium on Information Theory*, Sorrento, Italy, June 25-30, 2000. http://www.ifp.uiuc.edu/~singer/confpapers/singer_2000.pdf

[30] Stout, W. F. *Almost Sure Convergence*, Academic Press, New York, 1974.

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

[32] Yule, U. On a method of investigating periodicities in disturbed series, with special reference to Wölfer's sunspot numbers, *Philos. Trans. Roy. Soc.*, Vol. A 226, pp. 267–298, 1927.