Chapter 5

Nonparametric Sequential Prediction of Stationary Time Series

László Györfi and György Ottucsák

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

We present simple procedures for the prediction of a real valued time series with side information. For squared loss (regression problem), survey the basic principles of universally consistent estimates. The prediction algorithms are based on a combination of several simple predictors. We show that if the sequence is a realization of a stationary and ergodic random process then the average of squared errors converges, almost surely, to that of the optimum, given by the Bayes predictor. We offer an analog result for the prediction of stationary gaussian processes. These prediction strategies have some consequences for $0-1$ loss (pattern recognition problem).

5.1. Introduction

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

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

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

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

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

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

Our aim to achieve small $L_n(g)$ when n is large.

For this prediction problem, an example can be the forecasting daily relative prices y_t of an asset, while the side information vector x_t may contain some information on other assets in the past days or the trading volume in the previous day or some news related to the actual assets, etc. This is a widely investigated research problem. However, in the vast majority of the corresponding literature the side information is not included in the model, moreover, a parametric model (AR, MA, ARMA, ARIMA, ARCH, GARCH, etc.) is fitted to the stochastic process $\{Y_t\}$, its parameters are estimated, and a prediction is derived from the parameter estimates. (cf. [Tsay (2002)]). Formally, this approach means that there is a parameter θ such that the best predictor has the form

$$\mathbb{E}\{Y_t \mid Y_1^{t-1}\} = g_t(\theta, Y_1^{t-1}),$$

for a function g_t. The parameter θ is estimated from the past data Y_1^{t-1}, and the estimate is denoted by $\hat{\theta}$. Then the data-driven predictor is

$$g_t(\hat{\theta}, Y_1^{t-1}).$$

Here we don't assume any parametric model, so our results are fully nonparametric. This modelling is important for financial data when the process is only approximately governed by stochastic differential equations, so the parametric modelling can be weak, moreover the error criterion of the parameter estimate (usually the maximum likelihood estimate) has no relation to the mean square error of the prediction derived. The main aim of this research is to construct predictors, called universally consistent predictors, which are consistent for all stationary time series. Such universal feature can be proven using the recent principles of nonparametric statistics and machine learning algorithms.

The results below are given in an autoregressive framework, that is, the value Y_t is predicted based on X_1^t and Y_1^{t-1}. The fundamental limit for the predictability of the sequence can be determined based on a result of [Algoet (1994)], who showed that for any prediction strategy g and stationary ergodic process $\{(X_n, Y_n)\}_{-\infty}^{\infty}$,

$$\liminf_{n\to\infty} L_n(g) \geq L^* \quad \text{almost surely,} \tag{5.1}$$

where

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

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past $X_{-\infty}^0, Y_{-\infty}^{-1}$. Note that it follows by stationarity and the martingale convergence theorem (see, e.g., [Stout (1974)]) that

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

This lower bound gives sense to the following definition:

Definition 5.1. A prediction strategy g is called *universally consistent with respect to a class $\mathcal{C}$ of stationary and ergodic processes* $\{(X_n, Y_n)\}_{-\infty}^{\infty}$, if for each process in the class,

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

Universally consistent strategies asymptotically achieve the best possible squared loss for all ergodic processes in the class. [Algoet (1992)] and [Morvai *et al.* (1996)] proved that there exists a prediction strategy universal with respect to the class of all bounded ergodic processes. However, the prediction strategies exhibited in these papers are either very complex or have an unreasonably slow rate of convergence even for well-behaved processes.

Next we introduce several simple prediction strategies which, apart from having the above mentioned universal property of [Algoet (1992)] and [Morvai *et al.* (1996)], promise much improved performance for “nice” processes. The algorithms build on a methodology worked out in recent years for prediction of individual sequences, see [Vovk (1990)], [Feder *et al.* (1992)], [Littlestone and Warmuth (1994)], [Cesa-Bianchi *et al.* (1997)], [Kivinen and Warmuth (1999)], [Singer and Feder (1999)], [Merhav and Feder (1998)], [Cesa-Bianchi and Lugosi (2006)] for a survey.

An approach similar to the one of this paper was adopted by [Györfi *et al.* (1999)], where prediction of stationary binary sequences was addressed. There they introduced a simple randomized predictor which predicts asymptotically as well as the optimal predictor for all binary ergodic processes. The present setup and results differ in several important points from those of [Györfi *et al.* (1999)]. On the one hand, special properties of the squared loss function considered here allow us to avoid randomization of the predictor, and to define a significantly simpler prediction scheme. On the other hand, possible unboundedness of a real-valued process requires special care, which we demonstrate on the example of gaussian processes. We refer to [Nobel (2003)], [Singer and Feder (1999, 2000)], [Yang (2000)] to recent closely related work.

In Section 5.2 we survey the basic principles of nonparametric regression estimates. In Section 5.3 introduce universally consistent strategies for bounded ergodic processes which are based on a combination of partitioning or kernel or nearest neighbor or generalized linear estimates. In Section 5.4 consider the prediction of unbounded sequences including the ergodic gaussian process. In Section 5.5 study the classification problem of time series.

5.2. Nonparametric regression estimation

5.2.1. *The regression problem*

For the prediction of time series, an important source of the basic principles is the nonparametric regression. In regression analysis one considers a random vector (X, Y), where X is $\mathbb{R}^d$-valued and Y is $\mathbb{R}$-valued, and one is interested how the value of the so-called response variable Y depends on the value of the observation vector X. This means that one wants to find a function $f : \mathbb{R}^d \to \mathbb{R}$, such that $f(X)$ is a "good approximation of Y," that is, $f(X)$ should be close to Y in some sense, which is equivalent to making $|f(X)-Y|$ "small." Since X and Y are random vectors, $|f(X)-Y|$ is random as well, therefore it is not clear what "small $|f(X)-Y|$" means. We can resolve this problem by introducing the so-called L_2 *risk* or *mean squared error* of f,

$$\mathbb{E}|f(X) - Y|^2,$$

and requiring it to be as small as possible.

So we are interested in a function $m^* : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathbb{E}|m^*(X) - Y|^2 = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}|f(X) - Y|^2.$$

Such a function can be obtained explicitly as follows. Let

$$m(x) = \mathbb{E}\{Y|X = x\}$$

be the *regression function.* We will show that the regression function minimizes the L_2 risk. Indeed, for an arbitrary $f : \mathbb{R}^d \to \mathbb{R}$, a version of the

Steiner theorem implies that

$$\begin{aligned}\mathbb{E}|f(X)-Y|^2 &= \mathbb{E}|f(X)-m(X)+m(X)-Y|^2\\ &= \mathbb{E}|f(X)-m(X)|^2+\mathbb{E}|m(X)-Y|^2,\end{aligned}$$

where we have used

$$\begin{aligned}&\mathbb{E}\left\{(f(X)-m(X))(m(X)-Y)\right\}\\ &=\mathbb{E}\left\{\mathbb{E}\left\{(f(X)-m(X))(m(X)-Y)\big|X\right\}\right\}\\ &=\mathbb{E}\left\{(f(X)-m(X))\mathbb{E}\{m(X)-Y|X\}\right\}\\ &=\mathbb{E}\left\{(f(X)-m(X))(m(X)-m(X))\right\}\\ &=0.\end{aligned}$$

Hence,

$$\mathbb{E}|f(X)-Y|^2=\int_{\mathbb{R}^d}|f(x)-m(x)|^2\mu(dx)+\mathbb{E}|m(X)-Y|^2, \tag{5.2}$$

where μ denotes the distribution of X. The first term is called the L_2 error of f. It is always nonnegative and is zero if $f(x)=m(x)$. Therefore, $m^*(x)=m(x)$, i.e., the optimal approximation (with respect to the L_2 risk) of Y by a function of X is given by $m(X)$.

5.2.2. *Regression function estimation and L_2 error*

In applications the distribution of (X,Y) (and hence also the regression function) is usually unknown. Therefore it is impossible to predict Y using $m(X)$. But it is often possible to observe data according to the distribution of (X,Y) and to estimate the regression function from these data.

To be more precise, denote by (X,Y), (X_1,Y_1), $(X_2,Y_2),\dots$ independent and identically distributed (i.i.d.) random variables with $\mathbb{E}Y^2<\infty$. Let $\mathcal{D}_n$ be the set of *data* defined by

$$\mathcal{D}_n=\left\{(X_1,Y_1),\dots,(X_n,Y_n)\right\}.$$

In the regression function estimation problem one wants to use the data $\mathcal{D}_n$ in order to construct an estimate $m_n:\mathbb{R}^d\to\mathbb{R}$ of the regression function m. Here $m_n(x)=m_n(x,\mathcal{D}_n)$ is a measurable function of x and the data. For simplicity, we will suppress $\mathcal{D}_n$ in the notation and write $m_n(x)$ instead of $m_n(x,\mathcal{D}_n)$.

In general, estimates will not be equal to the regression function. To compare different estimates, we need an error criterion which measures the difference between the regression function and an arbitrary estimate

m_n. One of the key points we would like to make is that the motivation for introducing the regression function leads naturally to an L_2 error criterion for measuring the performance of the regression function estimate. Recall that the main goal was to find a function f such that the L_2 risk $\mathbb{E}|f(X)-Y|^2$ is small. The minimal value of this L_2 risk is $\mathbb{E}|m(X)-Y|^2$, and it is achieved by the regression function m. Similarly to (5.2), one can show that the L_2 risk $\mathbb{E}\{|m_n(X)-Y|^2|\mathcal{D}_n\}$ of an estimate m_n satisfies

$$\mathbb{E}\left\{|m_n(X)-Y|^2|\mathcal{D}_n\right\} = \int_{\mathbb{R}^d} |m_n(x)-m(x)|^2\mu(dx)+\mathbb{E}|m(X)-Y|^2. \quad (5.3)$$

Thus the L_2 risk of an estimate m_n is close to the optimal value if and only if the L_2 error

$$\int_{\mathbb{R}^d} |m_n(x)-m(x)|^2\mu(dx) \quad (5.4)$$

is close to zero. Therefore we will use the L_2 error (5.4) in order to measure the quality of an estimate and we will study estimates for which this L_2 error is small.

In this section we describe the basic principles of nonparametric regression estimation: *local averaging*, *local modelling*, *global modelling* (or *least squares estimation*), and *penalized modelling*. (Concerning the details see [Györfi *et al.* (2002)].)

Recall that the data can be written as

$$Y_i = m(X_i) + \epsilon_i,$$

where $\epsilon_i = Y_i - m(X_i)$ satisfies $\mathbb{E}(\epsilon_i|X_i) = 0$. Thus Y_i can be considered as the sum of the value of the regression function at X_i and some error ϵ_i, where the expected value of the error is zero. This motivates the construction of the estimates by *local averaging*, i.e., estimation of $m(x)$ by the average of those Y_i where X_i is "close" to x. Such an estimate can be written as

$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x)\cdot Y_i,$$

where the weights $W_{n,i}(x) = W_{n,i}(x, X_1, \dots, X_n) \in \mathbb{R}$ depend on $X_1, \dots, X_n$. Usually the weights are nonnegative and $W_{n,i}(x)$ is "small" if X_i is "far" from x.

5.2.3. *Partitioning estimate*

An example of such an estimate is the *partitioning estimate*. Here one chooses a finite or countably infinite partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \dots\}$ of $\mathbb{R}^d$ consisting of cells $A_{n,j} \subseteq \mathbb{R}^d$ and defines, for $x \in A_{n,j}$, the estimate by averaging Y_i's with the corresponding X_i's in $A_{n,j}$, i.e.,

$$m_n(x) = \frac{\sum_{i=1}^n I_{\{X_i \in A_{n,j}\}} Y_i}{\sum_{i=1}^n I_{\{X_i \in A_{n,j}\}}} \quad \text{for } x \in A_{n,j}, \tag{5.5}$$

where I_A denotes the indicator function of set A, so

$$W_{n,i}(x) = \frac{I_{\{X_i \in A_{n,j}\}}}{\sum_{l=1}^n I_{\{X_l \in A_{n,j}\}}} \quad \text{for } x \in A_{n,j}.$$

Here and in the following we use the convention $\frac{0}{0} = 0$. In order to have consistency, on the one hand we need that the cells $A_{n,j}$ should be "small", and on the other hand the number of non-zero terms in the denominator of (5.5) should be "large". These requirements can be satisfied if the sequences of partition $\mathcal{P}_n$ is asymptotically fine, i.e., if

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

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

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

and

$$\lim_{n\to\infty} \frac{|\{j \,:\, A_{n,j} \cap S \neq \emptyset\}|}{n} = 0.$$

For the partition $\mathcal{P}_n$, the most important example is when the cells $A_{n,j}$ are cubes of volume h_n^d. For cubic partition, the consistency conditions above mean that

$$\lim_{n\to\infty} h_n = 0 \quad \text{and} \quad \lim_{n\to\infty} n h_n^d = \infty. \tag{5.6}$$

Next we bound the rate of convergence of $\mathbb{E}\|m_n - m\|^2$ for cubic partitions and regression functions which are Lipschitz continuous.

Proposition 5.1. *For a cubic partition with side length h_n assume that*

$$\mathbf{Var}(Y|X = x) \leq \sigma^2,\ x \in \mathbb{R}^d,$$

$$|m(x) - m(z)| \leq C\|x - z\|,\ x, z \in \mathbb{R}^d, \tag{5.7}$$

and that X has a compact support S. Then

$$\mathbb{E}\|m_n - m\|^2 \leq \frac{c_1}{n \cdot h_n^d} + d \cdot C^2 \cdot h_n^2,$$

thus for

$$h_n = c_2 n^{-\frac{1}{d+2}}$$

we get

$$\mathbb{E}\|m_n - m\|^2 \leq c_3 n^{-2/(d+2)}.$$

In order to prove Proposition 5.1 we need the following technical lemma. An integer-valued random variable $B(n,p)$ is said to be binomially distributed with parameters n and $0 \leq p \leq 1$ if

$$\mathbb{P}\{B(n,p) = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Lemma 5.1. *Let the random variable $B(n,p)$ be binomially distributed with parameters n and p. Then:*

(i)

$$\mathbb{E}\left\{\frac{1}{1 + B(n,p)}\right\} \leq \frac{1}{(n+1)p},$$

(ii)

$$\mathbb{E}\left\{\frac{1}{B(n,p)} I_{\{B(n,p)>0\}}\right\} \leq \frac{2}{(n+1)p}.$$

Proof. Part (i) follows from the following simple calculation:

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{1 + B(n,p)}\right\} &= \sum_{k=0}^{n} \frac{1}{k+1} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{1}{(n+1)p} \sum_{k=0}^{n} \binom{n+1}{k+1} p^{k+1} (1-p)^{n-k} \\
&\leq \frac{1}{(n+1)p} \sum_{k=0}^{n+1} \binom{n+1}{k} p^k (1-p)^{n-k+1} \\
&= \frac{1}{(n+1)p} \left(p + (1-p)\right)^{n+1} \\
&= \frac{1}{(n+1)p}.
\end{aligned}$$

For (ii) we have

$$\mathbb{E}\left\{\frac{1}{B(n,p)}I_{\{B(n,p)>0\}}\right\} \le \mathbb{E}\left\{\frac{2}{1+B(n,p)}\right\} \le \frac{2}{(n+1)p}$$

by (i). □

Proof of Proposition 5.1. Set

$$\hat{m}_n(x) = \mathbb{E}\{m_n(x)|X_1,\dots,X_n\} = \frac{\sum_{i=1}^n m(X_i)I_{\{X_i\in A_n(x)\}}}{n\mu_n(A_n(x))},$$

where μ_n denotes the empirical distribution for $X_1,\dots,X_n$. Then

$$\begin{aligned}&\mathbb{E}\{(m_n(x)-m(x))^2|X_1,\dots,X_n\}\\ &= \mathbb{E}\{(m_n(x)-\hat{m}_n(x))^2|X_1,\dots,X_n\} + (\hat{m}_n(x)-m(x))^2. \qquad (5.8)\end{aligned}$$

We have

$$\begin{aligned}&\mathbb{E}\{(m_n(x)-\hat{m}_n(x))^2|X_1,\dots,X_n\}\\ &= \mathbb{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i-m(X_i))I_{\{X_i\in A_n(x)\}}}{n\mu_n(A_n(x))}\right)^2 \Bigg| X_1,\dots,X_n\right\}\\ &= \frac{\sum_{i=1}^n \mathbf{Var}(Y_i|X_i)I_{\{X_i\in A_n(x)\}}}{(n\mu_n(A_n(x)))^2}\\ &\le \frac{\sigma^2}{n\mu_n(A_n(x))}I_{\{n\mu_n(A_n(x))>0\}}.\end{aligned}$$

By Jensen's inequality

$$\begin{aligned}(\hat{m}_n(x)-m(x))^2 &= \left(\frac{\sum_{i=1}^n (m(X_i)-m(x))I_{\{X_i\in A_n(x)\}}}{n\mu_n(A_n(x))}\right)^2 I_{\{n\mu_n(A_n(x))>0\}}\\ &\quad + m(x)^2 I_{\{n\mu_n(A_n(x))=0\}}\\ &\le \frac{\sum_{i=1}^n (m(X_i)-m(x))^2 I_{\{X_i\in A_n(x)\}}}{n\mu_n(A_n(x))} I_{\{n\mu_n(A_n(x))>0\}}\\ &\quad + m(x)^2 I_{\{n\mu_n(A_n(x))=0\}}\\ &\le d\cdot C^2 h_n^2 I_{\{n\mu_n(A_n(x))>0\}} + m(x)^2 I_{\{n\mu_n(A_n(x))=0\}}\\ &\qquad \text{(by (5.7) and } \max_{z\in A_n(x)}\|x-z\| \le d\cdot h_n^2)\\ &\le d\cdot C^2 h_n^2 + m(x)^2 I_{\{n\mu_n(A_n(x))=0\}}.\end{aligned}$$

Without loss of generality assume that S is a cube and the union of $A_{n,1},\dots,A_{n,l_n}$ is S. Then

$$l_n \le \frac{\tilde{c}}{h_n^d}$$

for some constant $\tilde{c}$ proportional to the volume of S and, by Lemma 5.1 and (5.8),

$$
\begin{aligned}
&\mathbb{E}\left\{\int (m_n(x)-m(x))^2\mu(dx)\right\}\\
&=\mathbb{E}\left\{\int (m_n(x)-\hat{m}_n(x))^2\mu(dx)\right\}+\mathbb{E}\left\{\int (\hat{m}_n(x)-m(x))^2\mu(dx)\right\}\\
&=\sum_{j=1}^{l_n}\mathbb{E}\left\{\int_{A_{n,j}} (m_n(x)-\hat{m}_n(x))^2\mu(dx)\right\}\\
&\quad+\sum_{j=1}^{l_n}\mathbb{E}\left\{\int_{A_{n,j}} (\hat{m}_n(x)-m(x))^2\mu(dx)\right\}\\
&\le\sum_{j=1}^{l_n}\mathbb{E}\left\{\frac{\sigma^2\mu(A_{n,j})}{n\mu_n(A_{n,j})}I_{\{\mu_n(A_{n,j})>0\}}\right\}+dC^2h_n^2\\
&\quad+\sum_{j=1}^{l_n}\mathbb{E}\left\{\int_{A_{n,j}} m(x)^2\mu(dx)I_{\{\mu_n(A_{n,j})=0\}}\right\}\\
&\le\sum_{j=1}^{l_n}\frac{2\sigma^2\mu(A_{n,j})}{n\mu(A_{n,j})}+dC^2h_n^2+\sum_{j=1}^{l_n}\int_{A_{n,j}} m(x)^2\mu(dx)\mathbb{P}\{\mu_n(A_{n,j})=0\}\\
&\le l_n\frac{2\sigma^2}{n}+dC^2h_n^2+\sup_{z\in S}\left\{m(z)^2\right\}\sum_{j=1}^{l_n}\mu(A_{n,j})(1-\mu(A_{n,j}))^n\\
&\le l_n\frac{2\sigma^2}{n}+dC^2h_n^2+l_n\frac{\sup_{z\in S}m(z)^2}{n}\sup_j n\mu(A_{n,j})e^{-n\mu(A_{n,j})}\\
&\le l_n\frac{2\sigma^2}{n}+dC^2h_n^2+l_n\frac{\sup_{z\in S}m(z)^2e^{-1}}{n}\\
&\qquad\qquad(\text{since } \sup_z ze^{-z}=e^{-1})\\
&\le\frac{(2\sigma^2+\sup_{z\in S}m(z)^2e^{-1})\tilde{c}}{nh_n^d}+dC^2h_n^2.
\end{aligned}
$$

□

5.2.4. *Kernel estimate*

The second example of a local averaging estimate is the *Nadaraya–Watson kernel estimate*. Let $K:\mathbb{R}^d\to\mathbb{R}_+$ be a function called the kernel function, and let $h > 0$ be a bandwidth. The kernel estimate is defined by

$$m_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}, \tag{5.9}$$

so

$$W_{n,i}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)}.$$

Here the estimate is a weighted average of the Y_i, where the weight of Y_i (i.e., the influence of Y_i on the value of the estimate at x) depends on the distance between X_i and x. For the bandwidth $h = h_n$, the consistency conditions are (5.6). If one uses the so-called naïve kernel (or window kernel) $K(x) = I_{\{\|x\| \le 1\}}$, then

$$m_n(x) = \frac{\sum_{i=1}^n I_{\{\|x-X_i\| \le h\}} Y_i}{\sum_{i=1}^n I_{\{\|x-X_i\| \le h\}}},$$

i.e., one estimates $m(x)$ by averaging Y_i's such that the distance between X_i and x is not greater than h.

In the sequel we bound the rate of convergence of $\mathbb{E}\|m_n - m\|^2$ for a naïve kernel and a Lipschitz continuous regression function.

Proposition 5.2. *For a kernel estimate with a naïve kernel assume that*

$$\mathbf{Var}(Y|X = x) \le \sigma^2,\ x \in \mathbb{R}^d,$$

and

$$|m(x) - m(z)| \le C\|x - z\|,\ x, z \in \mathbb{R}^d,$$

and X has a compact support S^*. Then

$$\mathbb{E}\|m_n - m\|^2 \le \frac{c_1}{n \cdot h_n^d} + C^2 h_n^2,$$

thus for

$$h_n = c_2 n^{-\frac{1}{d+2}}$$

we have

$$\mathbb{E}\|m_n - m\|^2 \le c_3 n^{-2/(d+2)}.$$

Proof. We proceed similarly to Proposition 5.1. Put

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n m(X_i) I_{\{X_i \in S_{x,h_n}\}}}{n\mu_n(S_{x,h_n})},$$

then we have the decomposition (5.8). If $B_n(x) = \{n\mu_n(S_{x,h_n}) > 0\}$, then

$$\begin{aligned}
&\mathbb{E}\{(m_n(x) - \hat{m}_n(x))^2 | X_1, \ldots, X_n\} \\
&= \mathbb{E}\left\{ \left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{X_i \in S_{x,h_n}\}}}{n\mu_n(S_{x,h_n})} \right)^2 | X_1, \ldots, X_n \right\} \\
&= \frac{\sum_{i=1}^n \mathbf{Var}(Y_i | X_i) I_{\{X_i \in S_{x,h_n}\}}}{(n\mu_n(S_{x,h_n}))^2} \\
&\le \frac{\sigma^2}{n\mu_n(S_{x,h_n})} I_{B_n(x)}.
\end{aligned}$$

By Jensen's inequality and the Lipschitz property of m,

$$\begin{aligned}
&(\hat{m}_n(x) - m(x))^2 \\
&= \left(\frac{\sum_{i=1}^n (m(X_i) - m(x)) I_{\{X_i \in S_{x,h_n}\}}}{n\mu_n(S_{x,h_n})} \right)^2 I_{B_n(x)} + m(x)^2 I_{B_n(x)^c} \\
&\le \frac{\sum_{i=1}^n (m(X_i) - m(x))^2 I_{\{X_i \in S_{x,h_n}\}}}{n\mu_n(S_{x,h_n})} I_{B_n(x)} + m(x)^2 I_{B_n(x)^c} \\
&\le C^2 h_n^2 I_{B_n(x)} + m(x)^2 I_{B_n(x)^c} \\
&\le C^2 h_n^2 + m(x)^2 I_{B_n(x)^c}.
\end{aligned}$$

Using this, together with Lemma 5.1,

$$\begin{aligned}
&\mathbb{E}\left\{ \int (m_n(x) - m(x))^2 \mu(dx) \right\} \\
&= \mathbb{E}\left\{ \int (m_n(x) - \hat{m}_n(x))^2 \mu(dx) \right\} + \mathbb{E}\left\{ \int (\hat{m}_n(x) - m(x))^2 \mu(dx) \right\} \\
&\le \int_{S^*} \mathbb{E}\left\{ \frac{\sigma^2}{n\mu_n(S_{x,h_n})} I_{\{\mu_n(S_{x,h_n}) > 0\}} \right\} \mu(dx) + C^2 h_n^2 \\
&\quad + \int_{S^*} \mathbb{E}\left\{ m(x)^2 I_{\{\mu_n(S_{x,h_n}) = 0\}} \right\} \mu(dx) \\
&\le \int_{S^*} \frac{2\sigma^2}{n\mu(S_{x,h_n})} \mu(dx) + C^2 h_n^2 + \int_{S^*} m(x)^2 (1 - \mu(S_{x,h_n}))^n \mu(dx) \\
&\le \int_{S^*} \frac{2\sigma^2}{n\mu(S_{x,h_n})} \mu(dx) + C^2 h_n^2 + \sup_{z \in S^*} m(z)^2 \int_{S^*} e^{-n\mu(S_{x,h_n})} \mu(dx) \\
&\le 2\sigma^2 \int_{S^*} \frac{1}{n\mu(S_{x,h_n})} \mu(dx) + C^2 h_n^2 \\
&\quad + \sup_{z \in S^*} m(z)^2 \max_u u e^{-u} \int_{S^*} \frac{1}{n\mu(S_{x,h_n})} \mu(dx).
\end{aligned}$$

We can find $z_1, \ldots, z_{M_n}$ such that the union of $S_{z_1, rh_n/2}, \ldots, S_{z_{M_n}, rh_n/2}$ covers S^*, and

$$M_n \leq \frac{\tilde{c}}{h_n^d}.$$

Then

$$\begin{aligned}
\int_{S^*} \frac{1}{n\mu(S_{x,rh_n})} \mu(dx) &\leq \sum_{j=1}^{M_n} \int \frac{I_{\{x \in S_{z_j, rh_n/2}\}}}{n\mu(S_{x,rh_n})} \mu(dx) \\
&\leq \sum_{j=1}^{M_n} \int \frac{I_{\{x \in S_{z_j, rh_n/2}\}}}{n\mu(S_{z_j, rh_n/2})} \mu(dx) \\
&\leq \frac{M_n}{n} \\
&\leq \frac{\tilde{c}}{nh_n^d}.
\end{aligned}$$

Combining these inequalities the proof is complete. □

5.2.5. *Nearest neighbor estimate*

Our final example of local averaging estimates is the *k-nearest neighbor* (*k*-NN) *estimate*. Here one determines the k nearest X_i's to x in terms of distance $\|x - X_i\|$ and estimates $m(x)$ by the average of the corresponding Y_i's. More precisely, for $x \in \mathbb{R}^d$, let

$$(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$$

be a permutation of

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

such that

$$\|x - X_{(1)}(x)\| \leq \cdots \leq \|x - X_{(n)}(x)\|.$$

The k-NN estimate is defined by

$$m_n(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(x). \tag{5.10}$$

Here the weight $W_{ni}(x)$ equals $1/k$ if X_i is among the k nearest neighbors of x, and equals 0 otherwise. If $k = k_n \to \infty$ such that $k_n/n \to 0$ then the k-nearest-neighbor regression estimate is consistent.

Next we bound the rate of convergence of $\mathbb{E}\|m_n - m\|^2$ for a k_n-nearest neighbor estimate.

Proposition 5.3. *Assume that X is bounded,*

$$\sigma^2(x) = \mathbf{Var}(Y|X = x) \le \sigma^2 \quad (x \in \mathbb{R}^d)$$

and

$$|m(x) - m(z)| \le C\|x - z\| \quad (x, z \in \mathbb{R}^d).$$

Assume that $d \ge 3$. Let m_n be the k_n-NN estimate. Then

$$\mathbb{E}\|m_n - m\|^2 \le \frac{\sigma^2}{k_n} + c_1 \left(\frac{k_n}{n}\right)^{2/d},$$

thus for $k_n = c_2 n^{\frac{2}{d+2}}$,

$$\mathbb{E}\|m_n - m\|^2 \le c_3 n^{-\frac{2}{d+2}}.$$

For the proof of Proposition 5.3 we need the rate of convergence of nearest neighbor distances.

Lemma 5.2. *Assume that X is bounded. If $d \ge 3$, then*

$$\mathbb{E}\{\|X_{(1,n)}(X) - X\|^2\} \le \frac{\tilde{c}}{n^{2/d}}.$$

Proof. For fixed $\epsilon > 0$,

$$\mathbb{P}\{\|X_{(1,n)}(X) - X\| > \epsilon\} = \mathbb{E}\{(1 - \mu(S_{X,\epsilon}))^n\}.$$

Let $A_1, \dots, A_{N(\epsilon)}$ be a cubic partition of the bounded support of μ such that the A_j's have diameter ϵ and

$$N(\epsilon) \le \frac{c}{\epsilon^d}.$$

If $x \in A_j$, then $A_j \subset S_{x,\epsilon}$, therefore

$$\begin{aligned}
\mathbb{E}\{(1 - \mu(S_{X,\epsilon}))^n\} &= \sum_{j=1}^{N(\epsilon)} \int_{A_j} (1 - \mu(S_{x,\epsilon}))^n \mu(dx) \\
&\le \sum_{j=1}^{N(\epsilon)} \int_{A_j} (1 - \mu(A_j))^n \mu(dx) \\
&= \sum_{j=1}^{N(\epsilon)} \mu(A_j)(1 - \mu(A_j))^n.
\end{aligned}$$

Obviously,

$$\begin{aligned}\sum_{j=1}^{N(\epsilon)} \mu(A_j)(1-\mu(A_j))^n &\leq \sum_{j=1}^{N(\epsilon)} \max_z z(1-z)^n \\ &\leq \sum_{j=1}^{N(\epsilon)} \max_z ze^{-nz} \\ &= \frac{e^{-1}N(\epsilon)}{n}.\end{aligned}$$

If L stands for the diameter of the support of μ, then

$$\begin{aligned}\mathbb{E}\{\|X_{(1,n)}(X)-X\|^2\} &= \int_0^\infty \mathbb{P}\{\|X_{(1,n)}(X)-X\|^2 > \epsilon\}\, d\epsilon \\ &= \int_0^{L^2} \mathbb{P}\{\|X_{(1,n)}(X)-X\| > \sqrt{\epsilon}\}\, d\epsilon \\ &\leq \int_0^{L^2} \min\left\{1, \frac{e^{-1}N(\sqrt{\epsilon})}{n}\right\} d\epsilon \\ &\leq \int_0^{L^2} \min\left\{1, \frac{c}{en}\epsilon^{-d/2}\right\} d\epsilon \\ &= \int_0^{(c/(en))^{2/d}} 1\, d\epsilon + \frac{c}{en}\int_{(c/(en))^{2/d}}^{L^2} \epsilon^{-d/2} d\epsilon \\ &\leq \frac{\tilde{c}}{n^{2/d}}\end{aligned}$$

for $d \geq 3$. □

Proof of Proposition 5.3. We have the decomposition

$$\begin{aligned}\mathbb{E}\{(m_n(x)-m(x))^2\} &= \mathbb{E}\{(m_n(x)-\mathbb{E}\{m_n(x)|X_1,\dots,X_n\})^2\} \\ &\quad + \mathbb{E}\{(\mathbb{E}\{m_n(x)|X_1,\dots,X_n\}-m(x))^2\} \\ &= I_1(x)+I_2(x).\end{aligned}$$

The first term is easier:

$$\begin{aligned}I_1(x) &= \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\big(Y_{(i,n)}(x)-m(X_{(i,n)}(x))\big)\right)^2\right\} \\ &= \mathbb{E}\left\{\frac{1}{k_n^2}\sum_{i=1}^{k_n}\sigma^2(X_{(i,n)}(x))\right\} \\ &\leq \frac{\sigma^2}{k_n}.\end{aligned}$$

For the second term

$$
\begin{aligned}
I_2(x) &= \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}(m(X_{(i,n)}(x)) - m(x))\right)^2\right\}\\
&\leq \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}|m(X_{(i,n)}(x)) - m(x)|\right)^2\right\}\\
&\leq \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}C\|X_{(i,n)}(x) - x\|\right)^2\right\}.
\end{aligned}
$$

Put $N = k_n\lfloor\frac{n}{k_n}\rfloor$. Split the data $X_1, \dots, X_n$ into $k_n + 1$ segments such that the first k_n segments have length $\lfloor\frac{n}{k_n}\rfloor$, and let $\tilde{X}_j^x$ be the first nearest neighbor of x from the jth segment. Then $\tilde{X}_1^x, \dots, \tilde{X}_{k_n}^x$ are k_n different elements of $\{X_1, \dots, X_n\}$, which implies

$$
\sum_{i=1}^{k_n}\|X_{(i,n)}(x) - x\| \leq \sum_{j=1}^{k_n}\|\tilde{X}_j^x - x\|,
$$

therefore, by Jensen's inequality,

$$
\begin{aligned}
I_2(x) &\leq C^2\mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{j=1}^{k_n}\|\tilde{X}_j^x - x\|\right)^2\right\}\\
&\leq C^2\frac{1}{k_n}\sum_{j=1}^{k_n}\mathbb{E}\left\{\|\tilde{X}_j^x - x\|^2\right\}\\
&= C^2\mathbb{E}\left\{\|\tilde{X}_1^x - x\|^2\right\}\\
&= C^2\mathbb{E}\left\{\|X_{(1,\lfloor\frac{n}{k_n}\rfloor)}(x) - x\|^2\right\}.
\end{aligned}
$$

Thus, by Lemma 5.2,

$$
\begin{aligned}
\frac{1}{C^2}\left\lfloor\frac{n}{k_n}\right\rfloor^{2/d}\int I_2(x)\mu(dx) &\leq \left\lfloor\frac{n}{k_n}\right\rfloor^{2/d}\mathbb{E}\left\{\|X_{(1,\lfloor\frac{n}{k_n}\rfloor)}(X) - X\|^2\right\}\\
&\leq const.
\end{aligned}
$$

□

5.2.6. *Empirical error minimization*

A generalization of the partitioning estimate leads to *global modelling* or *least squares estimates*. Let $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \dots\}$ be a partition of $\mathbb{R}^d$ and let $\mathcal{F}_n$ be the set of all piecewise constant functions with respect to that partition, i.e.,

$$\mathcal{F}_n = \left\{ \sum_j a_j I_{A_{n,j}} \;:\; a_j \in \mathbb{R} \right\}. \tag{5.11}$$

Then it is easy to see that the partitioning estimate (5.5) satisfies

$$m_n(\cdot) = \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 \right\}. \tag{5.12}$$

Hence it minimizes the empirical L_2 risk

$$\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 \tag{5.13}$$

over $\mathcal{F}_n$. Least squares estimates are defined by minimizing the empirical L_2 risk over a general set of functions $\mathcal{F}_n$ (instead of (5.11)). Observe that it doesn't make sense to minimize (5.13) over all functions f, because this may lead to a function which interpolates the data and hence is not a reasonable estimate. Thus one has to restrict the set of functions over which one minimizes the empirical L_2 risk. Examples of possible choices of the set $\mathcal{F}_n$ are sets of piecewise polynomials with respect to a partition $\mathcal{P}_n$, or sets of smooth piecewise polynomials (splines). The use of spline spaces ensures that the estimate is a smooth function. An important member of least squares estimates is the generalized linear estimates. Let $\{\phi_j\}_{j=1}^{\infty}$ be real-valued functions defined on $\mathbb{R}^d$ and let $\mathcal{F}_n$ be defined by

$$\mathcal{F}_n = \left\{ f;\; f = \sum_{j=1}^{\ell_n} c_j \phi_j \right\}.$$

Then the generalized linear estimate is defined by

$$\begin{aligned} m_n(\cdot) &= \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \right\} \\ &= \operatorname*{arg\,min}_{c_1,\ldots,c_{\ell_n}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{\ell_n} c_j \phi_j(X_i) - Y_i \right)^2 \right\}. \end{aligned}$$

If the set

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

is dense in the set of continuous functions of d variables, $\ell_n \to \infty$ and $\ell_n/n \to 0$ then the generalized linear regression estimate defined above is consistent. For least squares estimates, other example can be the neural networks or radial basis functions or orthogonal series estimates.

Next we bound the rate of convergence of empirical error minimization estimates.

Condition (sG). The error $\varepsilon := Y - m(X)$ is subGaussian random variable, that is, there exist constants $\lambda > 0$ and $\Lambda < \infty$ with

$$\mathbb{E}\left\{ \exp(\lambda\varepsilon^2) \middle| X \right\} < \Lambda$$

a.s. Furthermore, define $\sigma^2 := \mathbb{E}\{\varepsilon^2\}$ and set $\lambda_0 = 4\Lambda/\lambda$.

Condition (C). The class $\mathcal{F}_n$ is totally bounded with respect to the supremum norm. For each $\delta > 0$, let $M(\delta)$ denote the δ-covering number of $\mathcal{F}$. This means that for every $\delta > 0$, there is a δ-cover $f_1, \ldots, f_M$ with $M = M(\delta)$ such that

$$\min_{1 \le i \le M} \sup_x |f_i(x) - f(x)| \le \delta$$

for all $f \in \mathcal{F}_n$. In addition, assume that $\mathcal{F}_n$ is uniformly bounded by L, that is,

$$|f(x)| \le L < \infty$$

for all $x \in \mathbb{R}$ and $f \in \mathcal{F}_n$.

Proposition 5.4. *Assume that conditions (C) and (sG) hold and*

$$|m(x)| \le L < \infty.$$

Then, for the estimate m_n defined by (5.5) and for all $\delta_n > 0$, $n \ge 1$,

$$\begin{aligned}
&\mathbb{E}\left\{ (m_n(X) - m(X))^2 \right\} \\
&\le 2 \inf_{f \in \mathcal{F}_n} \mathbb{E}\{(f(X) - m(X))^2\} \\
&\quad + (16L + 4\sigma)\delta_n + \left(16L^2 + 4\max\left\{ L\sqrt{2\lambda_0}, 8\lambda_0 \right\} \right) \frac{\log M(\delta_n)}{n}.
\end{aligned}$$

In the proof of this proposition we use the following lemma:

Lemma 5.3 (Wegkamp, 1999). *Let Z be a random variable with*

$$\mathbb{E}\{Z\} = 0 \text{ and } \mathbb{E}\left\{ \exp(\lambda Z^2) \right\} \le A$$

for some constants $\lambda > 0$ and $A \geq 1$. Then

$$\mathbb{E}\left\{\exp(\beta Z)\right\} \leq \exp\left(\frac{2A\beta^2}{\lambda}\right)$$

holds for every $\beta \in \mathbb{R}$.

Proof. Since for all $t > 0$, $\mathbb{P}\{|Z| > t\} \leq A\exp(-\lambda t^2)$ holds, we have for all integers $m \geq 2$,

$$\mathbb{E}\left\{|Z|^m\right\} = \int_0^\infty \mathbb{P}\big\{|Z|^m > t\big\}\, dt \leq A\int_0^\infty \exp\left(-\lambda t^{2/m}\right)\, dt = A\lambda^{-m/2}\Gamma\left(\frac{m}{2}+1\right).$$

Note that $\Gamma^2(\frac{m}{2}+1) \leq \Gamma(m+1)$ by Cauchy-Schwarz. The following inequalities are now self-evident.

$$\begin{aligned}
\mathbb{E}\left\{\exp\left(\beta Z\right)\right\} &= 1 + \sum_{m=2}^{\infty}\frac{1}{m!}\mathbb{E}(\beta Z)^m \\
&\leq 1 + \sum_{m=2}^{\infty}\frac{1}{m!}|\beta|^m\mathbb{E}|Z|^m \\
&\leq 1 + A\sum_{m=2}^{\infty}\lambda^{-m/2}|\beta|^m\frac{\Gamma\left(\frac{m}{2}+1\right)}{\Gamma\left(m+1\right)} \\
&\leq 1 + A\sum_{m=2}^{\infty}\lambda^{-m/2}|\beta|^m\frac{1}{\Gamma\left(\frac{m}{2}+1\right)} \\
&= 1 + A\sum_{m=1}^{\infty}\left(\frac{\beta^2}{\lambda}\right)^m\frac{1}{\Gamma\left(m+1\right)} \\
&\quad + A\sum_{m=1}^{\infty}\left(\frac{\beta^2}{\lambda}\right)^{m+\frac{1}{2}}\frac{1}{\Gamma\left(m+\frac{3}{2}\right)} \\
&\leq 1 + A\sum_{m=1}^{\infty}\left(\frac{\beta^2}{\lambda}\right)^m\left(1+\left(\frac{\beta^2}{\lambda}\right)^{\frac{1}{2}}\right)\frac{1}{\Gamma\left(m+1\right)}.
\end{aligned}$$

Finally, invoke the inequality $1+(1+\sqrt{x})(\exp(x)-1) \leq \exp(2x)$ for $x > 0$, to obtain the result. □

Lemma 5.4 (Antos *et al.*, 2005). *Let X_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots M$ be random variables such that for each fixed j, $X_{1j},\ldots,X_{nj}$ are independent and identically distributed such that for each $s_0 \geq s > 0$*

$$\mathbb{E}\{e^{sX_{ij}}\} \leq e^{s^2\sigma_j^2}.$$

For $\delta_j > 0$, put

$$\vartheta = \min_{j \le M} \frac{\delta_j}{\sigma_j^2}.$$

Then

$$\mathbb{E}\left\{\max_{j \le M}\left(\frac{1}{n}\sum_{i=1}^{n} X_{ij} - \delta_j\right)\right\} \le \frac{\log M}{\min\{\vartheta, s_0\} n}. \tag{5.14}$$

If

$$\mathbb{E}\{X_{ij}\} = 0$$

and

$$|X_{ij}| \le K,$$

then

$$\mathbb{E}\left\{\max_{j \le M}\left(\frac{1}{n}\sum_{i=1}^{n} X_{ij} - \delta_j\right)\right\} \le \max\{1/\vartheta^*, K\}\frac{\log M}{n}, \tag{5.15}$$

where

$$\vartheta^* = \min_{j \le M} \frac{\delta_j}{\mathbf{Var}(X_{ij})}.$$

Proof. For the notation

$$Y_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} - \delta_j$$

we have that for any $s_0 \ge s > 0$

$$\begin{aligned}
\mathbb{E}\{e^{snY_j}\} &= \mathbb{E}\{e^{sn\left(\frac{1}{n}\sum_{i=1}^{n} X_{ij} - \delta_j\right)}\} \\
&= e^{-sn\delta_j}\left(\mathbb{E}\{e^{sX_{1j}}\}\right)^n \\
&\le e^{-sn\delta_j} e^{ns^2\sigma_j^2} \\
&\le e^{-sn\alpha\sigma_j^2 + s^2 n\sigma_j^2}.
\end{aligned}$$

Thus

$$\begin{aligned}
e^{sn\mathbb{E}\{\max_{j \le M} Y_j\}} &\le \mathbb{E}\{e^{sn\max_{j \le M} Y_j}\} \\
&= \mathbb{E}\{\max_{j \le M} e^{snY_j}\} \\
&\le \sum_{j \le M} \mathbb{E}\{e^{snY_j}\} \\
&\le \sum_{j \le M} e^{-sn\sigma_j^2(\alpha - s)}.
\end{aligned}$$

For $s = \min\{\alpha, s_0\}$ it implies that

$$\mathbb{E}\{\max_{j\le M} Y_j\} \le \frac{1}{sn} \log\left(\sum_{j\le M} e^{-sn\sigma_j^2(\alpha-s)}\right) \le \frac{\log M}{\min\{\alpha, s_0\}n}.$$

In order to prove the second half of the lemma, notice that, for any $L > 0$ and $|x| \le L$ we have the inequality

$$\begin{aligned} e^x &= 1 + x + x^2 \sum_{i=2}^{\infty} \frac{x^{i-2}}{i!} \\ &\le 1 + x + x^2 \sum_{i=2}^{\infty} \frac{L^{i-2}}{i!} \\ &= 1 + x + x^2 \frac{e^L - 1 - L}{L^2}, \end{aligned}$$

therefore $0 < s \le s_0 = L/K$ implies that $s|X_{ij}| \le L$, so

$$e^{sX_{ij}} \le 1 + sX_{ij} + (sX_{ij})^2 \frac{e^L - 1 - L}{L^2}.$$

Thus,

$$\mathbb{E}\{e^{sX_{ij}}\} \le 1 + s^2 \mathbf{Var}(X_{ij}) \frac{e^L - 1 - L}{L^2} \le e^{s^2 \mathbf{Var}(X_{ij}) \frac{e^L - 1 - L}{L^2}},$$

so (5.15) follows from (5.14). □

Proof of Proposition 5.4. This proof is due to [Györfi and Wegkamp (2008)]. Set

$$D(f) = \mathbb{E}\{(f(X) - Y)^2\}$$

and

$$\widehat{D}(f) = \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

and

$$\Delta_f(x) = (m(x) - f(x))^2$$

and define

$$R(\mathcal{F}_n) := \sup_{f\in\mathcal{F}_n} \left[D(f) - 2\widehat{D}(f)\right] \le R_1(\mathcal{F}_n) + R_2(\mathcal{F}_n),$$

where

$$R_1(\mathcal{F}_n) := \sup_{f\in\mathcal{F}_n} \left[\frac{2}{n} \sum_{i=1}^{n} \{\mathbb{E}\Delta_f(X_i) - \Delta_f(X_i)\} - \frac{1}{2}\mathbb{E}\{\Delta_f(X)\}\right]$$

and

$$R_2(\mathcal{F}_n) := \sup_{f\in\mathcal{F}_n}\Big[\frac{4}{n}\sum_{i=1}^n \varepsilon_i(f(X_i)-m(X_i)) - \frac{1}{2}\mathbb{E}\{\Delta_f(X)\}\Big],$$

with $\varepsilon_i := Y_i - m(X_i)$. By the definition of $R(\mathcal{F}_n)$ and m_n, we have for all $f\in\mathcal{F}_n$

$$\begin{aligned}\mathbb{E}\left\{(m_n(X)-m(X))^2 \mid \mathcal{D}_n\right\} &= \mathbb{E}\left\{D(m_n)\mid\mathcal{D}_n\right\} - D(m)\\ &\le 2\{\widehat{D}(m_n)-\widehat{D}(m)\} + R(\mathcal{F}_n)\\ &\le 2\{\widehat{D}(f)-\widehat{D}(m)\} + R(\mathcal{F}_n)\ .\end{aligned}$$

After taking expectations on both sides, we obtain

$$\mathbb{E}\left\{(m_n(X)-m(X))^2\right\} \le 2\mathbb{E}\left\{(f(X)-m(X))^2\right\} + \mathbb{E}\{R(\mathcal{F}_n)\}.$$

Let $\mathcal{F}_n'$ be a finite δ_n-covering net (with respect to the sup-norm) of $\mathcal{F}_n$ with $M(\delta_n)=|\mathcal{F}_n'|$. It means that for any $f\in\mathcal{F}_n$ there is an $f'\in\mathcal{F}_n'$ such that

$$\sup_x |f(x)-f'(x)| \le \delta_n,$$

which implies that

$$\begin{aligned}&|\ (m(X_i)-f(X_i))^2 - (m(X_i)-f'(X_i))^2\ |\\ &\le |f(X_i)-f'(X_i)|\cdot\big(|m(X_i)-f(X_i)| + |m(X_i)-f'(X_i)|\big)\\ &\le 4L|f(X_i)-f'(X_i)|\\ &\le 4L\delta_n,\end{aligned}$$

and, by Cauchy-Schwarz inequality,

$$\begin{aligned}&\mathbb{E}\{|\varepsilon_i(m(X_i)-f(X_i)) - \varepsilon_i(m(X_i)-f'(X_i))|\}\\ &\le \sqrt{\mathbb{E}\{\varepsilon_i^2\}}\sqrt{\mathbb{E}\{(f(X_i)-f'(X_i))^2\}}\\ &\le \sigma\delta_n.\end{aligned}$$

Thus,

$$\mathbb{E}\{R(\mathcal{F}_n)\} \le 2\delta_n(4L+\sigma) + \mathbb{E}\{R(\mathcal{F}_n')\},$$

and therefore

$$\begin{aligned}&\mathbb{E}\left\{(m_n(X)-m(X))^2\right\}\\ &\le 2\mathbb{E}\left\{(f(X)-m(X))^2\right\} + \mathbb{E}\{R(\mathcal{F}_n)\}\\ &\le 2\mathbb{E}\left\{(f(X)-m(X))^2\right\} + (16L+4\sigma)\delta_n + \mathbb{E}\left\{R(\mathcal{F}_n')\right\}\\ &\le 2\mathbb{E}\left\{(f(X)-m(X))^2\right\} + (16L+4\sigma)\delta_n + \mathbb{E}\left\{R_1(\mathcal{F}_n')\right\} + \mathbb{E}\left\{R_2(\mathcal{F}_n')\right\}.\end{aligned}$$

Define, for all $f \in \mathcal{F}_n$ with $D(f) > D(m)$,

$$\tilde{\rho}(f) := \frac{\mathbb{E}\left\{(m(X) - f(X))^4\right\}}{\mathbb{E}\left\{(m(X) - f(X))^2\right\}} .$$

Since $|m(x)| \leq 1$ and $|f(x)| \leq 1$, we have that

$$\tilde{\rho}(f) \leq 4L^2 .$$

Invoke the second part of Lemma 5.4 below to obtain

$$\begin{aligned}\mathbb{E}\left\{R_1(\mathcal{F}'_n)\right\} &\leq \max\left(8L^2, 4L^2 \sup_{f\in\mathcal{F}'_n} \tilde{\rho}(f)\right) \frac{\log M(\delta_n)}{n} \\ &\leq \max\left(8L^2, 16L^2\right) \frac{\log M(\delta_n)}{n} \\ &= 16L^2 \frac{\log M(\delta_n)}{n}.\end{aligned}$$

By Condition (sG) and Lemma 5.3, we have for *all* $s > 0$,

$$\mathbb{E}\left\{\exp\left(s\varepsilon(f(X) - m(X))\right)\mid X\right\} \leq \exp(\lambda_0 s^2 (m(X) - f(X))^2/2).$$

For $|z| \leq 1$, apply the inequality $e^z \leq 1 + 2z$. Choose

$$s_0 = \frac{1}{L\sqrt{2\lambda_0}},$$

then

$$\frac{1}{2}\lambda_0 s^2 (f(X) - m(X))^2 \leq 1,$$

therefore, for $0 < s \leq s_0$,

$$\begin{aligned}\mathbb{E}\left\{\exp\left(s\varepsilon(f(X) - m(X))\right)\right\} &\leq \mathbb{E}\left\{\exp\left(\frac{1}{2}\lambda_0 s^2 (f(X) - m(X))^2\right)\right\} \\ &\leq 1 + \lambda_0 s^2 \mathbb{E}\left\{(f(X) - m(X))^2\right\} \\ &\leq \exp\left(\lambda_0 s^2 \mathbb{E}\left\{(f(X) - m(X))^2\right\}\right).\end{aligned}$$

Next we invoke the first part of Lemma 5.4. We find that the value ϑ in Lemma 5.4 becomes

$$1/\vartheta = 8 \sup_{f\in\mathcal{F}'_n} \frac{\lambda_0 \mathbb{E}\{(f(X) - m(X))^2\}}{\mathbb{E}\{\Delta_f(X)\}} \leq 8\lambda_0,$$

and we get

$$\mathbb{E}\left\{R_2(\mathcal{F}'_n)\right\} \leq 4\frac{\log M(\delta_n)}{n} \max\left(L\sqrt{2\lambda_0}, 8\lambda_0\right),$$

and this completes the proof of Proposition 5.4. □

Instead of restricting the set of functions over which one minimizes, one can also add a penalty term to the functional to be minimized. Let $J_n(f) \geq 0$ be a penalty term penalizing the "roughness" of a function f. The *penalized modelling* or *penalized least squares estimate* m_n is defined by

$$m_n = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 + J_n(f) \right\}, \tag{5.16}$$

where one minimizes over all measurable functions f. Again we do not require that the minimum in (5.16) be unique. In the case it is not unique, we randomly select one function which achieves the minimum.

A popular choice for $J_n(f)$ in the case $d = 1$ is

$$J_n(f) = \lambda_n \int |f''(t)|^2 dt, \tag{5.17}$$

where f'' denotes the second derivative of f and λ_n is some positive constant. One can show that for this penalty term the minimum in (5.16) is achieved by a cubic spline with knots at the X_i's, i.e., by a twice differentiable function which is equal to a polynomial of degree 3 (or less) between adjacent values of the X_i's (a so-called smoothing spline).

5.3. Universally consistent predictions: bounded Y

5.3.1. *Partition-based prediction strategies*

In this section we introduce our first prediction strategy for bounded ergodic processes. We assume throughout the section that $|Y_0|$ is bounded by a constant $B > 0$, with probability one, and the bound B is known.

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

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

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

and

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

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

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

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

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

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

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

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

In contrast to the nonparametric regression estimation problem from i.i.d. data, for ergodic observations, it is impossible to choose $k = k_n$ and $\ell = \ell_n$ such that the corresponding predictor is universally consistent for the class of bounded ergodic processes. The very important new principle is the combination or aggregation of elementary predictors (cf. [Cesa-Bianchi and Lugosi (2006)]). The proposed prediction algorithm proceeds as follows: let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k, ℓ) of positive integers such that for all k, ℓ, $q_{k,\ell} > 0$. Put $c = 8B^2$, and define the weights

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/c} \tag{5.18}$$

and their normalized values

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{W_t} \ , \tag{5.19}$$

where

$$W_t = \sum_{i,j=1}^{\infty} w_{t,i,j} \ . \tag{5.20}$$

The prediction strategy g is defined by

$$g_t(x_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(x_1^t, y_1^{t-1}) \ , \qquad t = 1, 2, \dots \tag{5.21}$$

i.e., the prediction g_t is the convex linear combination of the elementary predictors such that an elementary predictor has non-negligible weight in the combination if it has good performance until time $t-1$.

Theorem 5.1 (Györfi and Lugosi, 2001). *Assume that*
(a) the sequences of partition $\mathcal{P}_\ell$ is nested, that is, any cell of $\mathcal{P}_{\ell+1}$ is a subset of a cell of $\mathcal{P}_\ell$, $\ell = 1, 2, \dots$;
(b) the sequences of partition $\mathcal{Q}_\ell$ is nested;
(c) the sequences of partition $\mathcal{P}_\ell$ is asymptotically fine;
(d) the sequences of partition $\mathcal{Q}_\ell$ is asymptotically fine;
Then the prediction scheme g defined above is universal with respect to the class of all jointly stationary and ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ such that $|Y_0| \le B$.

One of the main ingredients of the proof is the following lemma, whose proof is a straightforward extension of standard arguments in the prediction theory of individual sequences, see, for example, [Kivinen and Warmuth (1999)], [Singer and Feder (2000)].

Lemma 5.5. *Let $\tilde{h}_1, \tilde{h}_2, \dots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $\tilde{h}_i(x_1^n, y_1^{n-1}) \in [-B, B]$ and $y_1^n \in [-B, B]^n$. Define*

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

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

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

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

$$\tilde{g}_t(x_1^n, y_1^{t-1}) = \sum_{k=1}^{\infty} v_{t,k} \tilde{h}_k(x_1^n, y_1^{t-1}) \qquad t = 1, 2, \dots$$

then for every $n \ge 1$,

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

Here $-\ln 0$ is treated as ∞.

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

$$\left[\sum_{k=1}^{\infty} v_{t,k}\left(y_t - \tilde{h}_k(x_1^n, y_1^{t-1})\right)\right]^2 \le -c \ln \frac{W_{t+1}}{W_t} \ . \tag{5.22}$$

Note that

$$W_{t+1} = \sum_{k=1}^{\infty} w_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^n, y_1^{t-1})\right)^2/c} = W_t \sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^n, y_1^{t-1})\right)^2/c},$$

so that

$$-c \ln \frac{W_{t+1}}{W_t} = -c \ln \left(\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^n, y_1^{t-1})\right)^2/c}\right).$$

Therefore, (5.22) becomes

$$\exp\left(\frac{-1}{c}\left[\sum_{k=1}^{\infty} v_{t,k}\left(y_t - \tilde{h}_k(x_1^n, y_1^{t-1})\right)\right]^2\right) \ge \sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^n, y_1^{t-1})\right)^2/c},$$

which is implied by Jensen's inequality and the concavity of the function $F_t(z) = e^{-(y_t - z)^2/c}$ for $c \ge 8B^2$. Thus, (5.22) implies that

$$\begin{aligned} nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^n, y_1^{t-1})\right)^2 \\ &= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k}\left(y_t - \tilde{h}_k(x_1^n, y_1^{t-1})\right)\right]^2 \\ &\le -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\ &= -c \ln W_{n+1} \end{aligned}$$

and therefore

$$\begin{aligned} nL_n(\tilde{g}) &\le -c \ln \left(\sum_{k=1}^{\infty} w_{n+1,k}\right) \\ &= -c \ln \left(\sum_{k=1}^{\infty} q_k e^{-nL_n(\tilde{h}_k)/c}\right) \\ &\le -c \ln \left(\sup_k q_k e^{-nL_n(\tilde{h}_k)/c}\right) \\ &= \inf_k \left(-c \ln q_k + nL_n(\tilde{h}_k)\right) \ , \end{aligned}$$

which concludes the proof. □

Another main ingredient of the proof of Theorem 5.1 is known as Breiman's generalized ergodic theorem [Breiman (1957)], see also [Algoet (1994)] and [Györfi *et al.* (2002)].

Lemma 5.6 (Breiman, 1957). *Let $Z = \{Z_i\}_{-\infty}^{\infty}$ be a stationary and ergodic process. Let T denote the left shift operator. Let f_i be a sequence of real-valued functions such that for some function f, $f_i(Z) \to f(Z)$ almost surely. Assume that $\mathbb{E}\{\sup_i |f_i(Z)|\} < \infty$. Then*

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

Proof of Theorem 5.1. Because of (5.1), it is enough to show that

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

By a double application of the ergodic theorem, as $n \to \infty$, almost surely,

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

and therefore

$$\lim_{n\to\infty} \sup_z \sup_s |\widehat{E}_n^{(k,\ell)}(X_1^n, Y_1^{n-1}, z, s) - \mathbb{E}\{Y_0 | G_\ell(X_{-k}^0) = z,\, F_\ell(Y_{-k}^{-1}) = s\}| = 0$$

almost surely. Thus, by Lemma 5.6, as $n \to \infty$, almost surely,

$$\begin{aligned} L_n(h^{(k,\ell)}) &= \frac{1}{n}\sum_{i=1}^{n} (h^{(k,\ell)}(X_1^i, Y_1^{i-1}) - Y_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} (\widehat{E}_n^{(k,\ell)}(X_1^i, Y_1^{i-1}, G_\ell(X_{i-k}^i), F_\ell(Y_{i-k}^{i-1})) - Y_i)^2 \\ &\to \mathbb{E}\{(Y_0 - \mathbb{E}\{Y_0 | G_\ell(X_{-k}^0),\, F_\ell(Y_{-k}^{-1})\})^2\} \\ &\overset{\text{def}}{=} \epsilon_{k,\ell}. \end{aligned}$$

Since the partitions $\mathcal{P}_\ell$ and $\mathcal{Q}_\ell$ are nested, $\mathbb{E}\left\{Y_0 | G_\ell(X_{-k}^0),\, F_\ell(Y_{-k}^{-1})\right\}$ is a martingale indexed by the pair (k, ℓ). Thus, the martingale convergence theorem (see, e.g., [Stout (1974)]) and assumption (c) and (d) for the sequence of partitions implies that

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

Now by Lemma 5.5,

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c\ln q_{k,\ell}}{n} \right), \tag{5.23}$$

and therefore, almost surely,

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

and the proof of the theorem is finished. □

Theorem 5.1 shows that asymptotically, the predictor g_t defined by (5.21) predicts as well as the optimal predictor given by the regression function $\mathbb{E}\{Y_t|Y_{-\infty}^{t-1}\}$. In fact, g_t gives a good estimate of the regression function in the following (Cesáro) sense:

Corollary 5.1. *Under the conditions of Theorem 5.1*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}\{Y_i|X_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(X_1^i, Y_1^{i-1}) \right)^2 = 0 \qquad \textit{almost surely.}$$

Proof. By Theorem 5.1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g_i(X_1^i, Y_1^{i-1}) \right)^2 = L^* \qquad \text{almost surely.}$$

Consider the following decomposition:

$$\begin{aligned}
&\left(Y_i - g_i(X_1^i, Y_1^{i-1})\right)^2 \\
&= \left(Y_i - \mathbb{E}\{Y_i|X_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)^2 \\
&\quad +2\left(Y_i - \mathbb{E}\{Y_i|X_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|X_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(X_1^i, Y_1^{i-1})\right) \\
&\quad + \left(\mathbb{E}\{Y_i|X_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(X_1^i, Y_1^{i-1})\right)^2.
\end{aligned}$$

Then the ergodic theorem implies that

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

It remains to show that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbb{E}\{Y_i|X_{-\infty}^{i}, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|Y_{-\infty}^{i-1}\} - g_i(X_1^i, Y_1^{i-1})\right) = 0. \tag{5.24}$$

almost surely. But this is a straightforward consequence of Kolmogorov's classical strong law of large numbers for martingale differences due to [Chow (1965)] (see also Theorem 3.3.1 in [Stout (1974)]). It states that if $\{Z_i\}$ is a martingale difference sequence with

$$\sum_{n=1}^{\infty}\frac{\mathbb{E}Z_n^2}{n^2} < \infty, \tag{5.25}$$

then

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

Thus, (5.24) is implied by Chow's theorem since the martingale differences $Z_i = \left(Y_i - \mathbb{E}\{Y_i|X_{-\infty}^{i}, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|X_{-\infty}^{i}, Y_{-\infty}^{i-1}\} - g_i(X_1^i, Y_1^{i-1})\right)$ are bounded by $4B^2$. (To see that the Z_i's indeed form a martingale difference sequence just note that $\mathbb{E}\{Z_i|X_{-\infty}^{i}, Y_{-\infty}^{i-1}\} = 0$ for all i.) □

Remark 5.1 (Choice of $q_{k,\ell}$). . *Theorem 5.1 is true independently of the choice of the $q_{k,\ell}$'s as long as these values are strictly positive for all k and ℓ. In practice, however, the choice of $q_{k,\ell}$ may have an impact on the performance of the predictor. For example, if the distribution $\{q_{k,\ell}\}$ has a very rapidly decreasing tail, then the term $-\ln q_{k,\ell}/n$ will be large for moderately large values of k and ℓ, and the performance of g will be determined by the best of just a few of the elementary predictors $h^{(k,\ell)}$. Thus, it may be advantageous to choose $\{q_{k,\ell}\}$ to be a large-tailed distribution. For example, $q_{k,\ell} = c_0 k^{-2}\ell^{-2}$ is a safe choice, where c_0 is an appropriate normalizing constant.*

5.3.2. *Kernel-based prediction strategies*

We introduce in this section a class of *kernel-based* prediction strategies for stationary and ergodic sequences. The main advantage of this approach in contrast to the partition-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

$$\lim_{\ell\to\infty} r_{k,\ell} = 0, \tag{5.26}$$

and

$$\lim_{\ell\to\infty} r'_{k,\ell} = 0. \tag{5.27}$$

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

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{\{t\in J_n^{(k,\ell)}\}} y_t}{|J_n^{(k,\ell)}|}, \qquad n > k+1, \tag{5.28}$$

where $0/0$ is defined to be 0. The pool of experts is mixed the same way as in the case of the partition-based strategy (cf. (5.18), (5.19), (5.20) and (5.21)).

Theorem 5.2. *Suppose that (5.26) and (5.27) are verified. Then the kernel-based strategy defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes* $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ *such that* $|Y_0| \le B$.

Remark 5.2. This theorem may be extended to a more general class of kernel-based strategies. 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 (5.28) of the elementary strategy, the regression function estimate is replaced by

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

Observe that if K is the naïve kernel $K(x) = I_{\{x\le 1\}}$, 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. Such kernels promise a better prediction of the local structure of the conditional distribution.

5.3.3. *Nearest neighbor-based prediction strategy*

This strategy is yet more robust with respect to the kernel strategy and thus also with respect to the partition strategy. Since it does not suffer from scaling problem as partition and kernel-based strategies where the quantizer and the radius has to be carefully chosen to obtain "good" performance. As well as this, in practical applications it runs extremely fast compared with the kernel and partition schemes as it is much less likely to get bogged down in calculations for certain experts.

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

$$\lim_{\ell\to\infty} p_\ell = 0\,, \tag{5.29}$$

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

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

and introduce the elementary predictor

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

if the sum is nonvoid, and 0 otherwise. Finally, the experts are mixed as before (cf. (5.18), (5.19), (5.20) and (5.21)).

Theorem 5.3. *Suppose that (5.29) is verified and that for each vector* $\mathbf{s}$ *the random variable*

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

has a continuous distribution function. Then the nearest neighbor strategy defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ *such that* $|Y_0| \leq B$.

5.3.4. *Generalized linear estimates*

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 (2001)]. 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

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

where the coefficients $c_{n,j}$ are calculated according to the past observations x_1^n, y_1^{n-1}. More precisely, the coefficients $c_{n,j}$ are defined as the real numbers which minimize the criterion

$$\sum_{t=k+1}^{n-1} \left(\sum_{j=1}^{\ell} c_j \phi_j^{(k)}(x_{t-k}^t, y_{t-k}^{t-1}) - y_t \right)^2 \tag{5.30}$$

if $n > k + 1$, and the all-zero vector otherwise. It can be shown using a recursive technique (see e.g., [Tsypkin (1971)], [Györfi (1984)], [Singer and Feder (2000)], and [Györfi and Lugosi (2001)]) 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 (cf. (5.18), (5.19), (5.20) and (5.21)).

Theorem 5.4 (Györfi and Lugosi, 2001). *Suppose that* $|\phi_j^{(k)}| \le 1$ *and, for any fixed* k*, suppose that the set*

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

is dense in the set of continuous functions of $d(k+1)+k$ *variables. Then the generalized linear strategy defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes* $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ *such that* $|Y_0| \le B$*.*

5.4. Universally consistent predictions: unbounded Y

5.4.1. *Partition-based prediction strategies*

Let $\widehat{E}_n^{(k,\ell)}(x_1^n, y_1^{n-1}, z, s)$ be defined as in Section 5.3.1. Introduce the truncation function

$$T_m(z) = \begin{cases} m & \text{if } z > m \\ z & \text{if } |z| < m \\ -m & \text{if } z < -m, \end{cases}$$

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

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

where

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

for $n = 1, 2, \dots$. That is, $h_n^{(k,\ell)}$ is the truncation of the elementary predictor introduced in Section 5.3.1.

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

$$w_{t,k,\ell} = q_{k,\ell} e^{-\eta_t (t-1) L_{t-1}(h^{(k,\ell)})} \tag{5.31}$$

and their normalized values

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{W_t}\ , \tag{5.32}$$

where

$$W_t = \sum_{i,j=1}^{\infty} w_{t,i,j} \ . \tag{5.33}$$

The prediction strategy g is defined by

$$g_t(x_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(x_1^t, y_1^{t-1}) \ , \qquad t = 1, 2, \dots \tag{5.34}$$

Theorem 5.5 (Györfi and Ottucsák, 2007). *Assume that the conditions (a), (b), (c) and (d) of Theorem 5.1 are satisfied. Choose $\eta_t = 1/\sqrt{t}$. Then the prediction scheme g defined above is universally consistent with respect to the class of all ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ such that*

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

Here we describe a result, which is used in the analysis. This lemma is a modification of the analysis of *et al.* [Auer *et al.* (2002)], which allows of the handling the case when the learning parameter of the algorithm (η_t) is time-dependent and the number of the elementary predictors is infinite.

Lemma 5.7 (Györfi and Ottucsák, 2007). *Let $h^{(1)}, h^{(2)}, \dots$ be a sequence of prediction strategies (experts). Let $\{q_k\}$ be a probability distribution on the set of positive integers. Denote the normalized loss of the expert $h = (h_1, h_2, \dots)$ by*

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

where

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

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

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

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

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

where

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

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

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

then for every $n \geq 1$,

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

Proof. Introduce some notations:

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

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

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

We start the proof with the following chain of bounds:

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

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

$$\begin{aligned}
&\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} \\
&\le \frac{1}{\eta_t} \ln \left(1 - \eta_t \sum_{k=1}^{\infty} p_{t,k} \ell_t(h^{(k)}) + \frac{\eta_t^2}{2} \sum_{k=1}^{\infty} p_{t,k} \ell_t^2(h^{(k)}) \right) \\
&\le - \sum_{k=1}^{\infty} p_{t,k} \ell_t(h^{(k)}) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \ell_t^2(h^{(k)}) \qquad (5.35) \\
&= - \sum_{k=1}^{\infty} p_{t,k} \ell(h_t^{(k)}, Y_t) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \ell_t^2(h^{(k)}) \\
&\le -\ell \left(\sum_{k=1}^{\infty} p_{t,k} h_t^{(k)}, Y_t \right) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \ell_t^2(h^{(k)}) \qquad (5.36) \\
&= -\ell_t(g) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \ell_t^2(h^{(k)}) \qquad (5.37)
\end{aligned}$$

where (5.35) follows from the fact that $\ln(1+x) \le x$ for all $x > -1$ and in (5.36) we used the convexity of the loss $\ell(h, y)$ in its first argument h. From (5.37) after rearranging we obtain

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

Then write a telescope formula:

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

We have that

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

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

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

Thus,

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

We can summarize the bounds:

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

□

Proof of Theorem 5.5. Because of (5.1), it is enough to show that

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

Because of the proof of Theorem 5.1, as $n \to \infty$, a.s.,

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

and therefore for all z and s

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

By Lemma 5.6, as $n \to \infty$, almost surely,

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

In the same way as in the proof of Theorem 5.1, we get that

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

Apply Lemma 5.7 with choice $\eta_t = \frac{1}{\sqrt{t}}$ and for the squared loss $\ell_t(h) = (h_t - Y_t)^2$, then the square loss is convex in its first argument h, so

$$\begin{aligned}
L_n(g) \le \inf_{k,\ell}&\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right)\\
&+ \frac{1}{2n}\sum_{t=1}^n \frac{1}{\sqrt{t}} \sum_{k,\ell=1}^\infty p_{t,k,\ell}\big(h^{(k,\ell)}(X_1^t, Y_1^{t-1}) - Y_t\big)^4\,. \qquad (5.38)
\end{aligned}$$

On the one hand, almost surely,

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

On the other hand,

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

therefore, almost surely,

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

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

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

and the proof of the theorem is finished. □

Corollary 5.2 (Györfi and Ottucsák, 2007). *Under the conditions of Theorem 5.5,*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1})\right)^2 = 0 \qquad a.s. \tag{5.39}$$

Proof. By Theorem 5.5,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 = L^* \qquad \text{a.s.} \tag{5.40}$$

and by the ergodic theorem we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left\{\left(Y_t - \mathbb{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\right\} = L^* \tag{5.41}$$

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

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1})\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\{\left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&\quad - \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\{\left(Y_t - \mathbb{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\{\left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&\quad - \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - g_t(X_1^t, Y_1^{t-1})\right)^2 + o(1) \qquad\qquad (5.42) \\
&= 2\frac{1}{n}\sum_{t=1}^{n} g_t(X_1^t, Y_1^{t-1})(Y_t - \mathbb{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}) \\
&\quad - \frac{1}{n}\sum_{t=1}^{n}\left(Y_t^2 - \mathbb{E}\{Y_t^2 \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\}\right) + o(1) \qquad \text{a.s.}
\end{aligned}$$

where (5.42) holds because of (5.40) and (5.41). The second sum is

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

by the ergodic theorem. Put

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

In order to finish the proof it suffices to show

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} Z_t = 0\ . \tag{5.43}$$

Then

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

for all t, so the Z_t's form a martingale difference sequence. By the strong law of large numbers for martingale differences due to [Chow (1965)], one has to verify (5.25). By the construction of g_n,

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

therefore (5.25) is verified, (5.43) is proved and the proof of the corollary is finished. □

5.4.2. *Kernel-based prediction strategies*

Apply the notations of Section 5.3.2. Then the elementary expert $h_n^{(k,\ell)}$ at time n is defined by

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

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 partition-based strategy (cf. (5.31), (5.32), (5.33) and (5.34)).

Theorem 5.6 (Biau *et al.*, 2010). *Choose $\eta_t = 1/\sqrt{t}$ and suppose that (5.26) and (5.27) are verified. Then the kernel-based strategy defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ such that*

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

5.4.3. *Nearest neighbor-based prediction strategy*

Apply the notations of Section 5.3.3. Then the elementary expert $h_n^{(k,\ell)}$ at time n is defined by

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

if the sum is nonvoid, and 0 otherwise and $0 < \delta < 1/8$. The pool of experts is mixed the same way as in the case of the histogram-based strategy (cf. (5.31), (5.32), (5.33) and (5.34)).

Theorem 5.7 (Biau *et al.*, 2010). *Choose $\eta_t = 1/\sqrt{t}$ and suppose that (5.29) is verified. Suppose also that for each vector* $\mathbf{s}$ *the random variable*

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

has a continuous distribution function. Then the nearest neighbor strategy defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ such that

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

5.4.4. *Generalized linear estimates*

Apply the notations of Section 5.3.4 . The elementary predictor $h_n^{(k,\ell)}$ generates a prediction of form

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

with $0 < \delta < 1/8$. The pool of experts is mixed the same way as in the case of the histogram-based strategy (cf. (5.31), (5.32), (5.33) and (5.34)).

Theorem 5.8 (Biau *et al.*, 2010). *Choose $\eta_t = 1/\sqrt{t}$ and suppose that $|\phi_j^{(k)}| \le 1$ 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 strategy defined above is universally consistent with respect to the class of all jointly stationary and ergodic processes $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ such that

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

5.4.5. *Prediction of Gaussian processes*

We consider in this section the classical problem of Gaussian time series prediction. 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 (1986)], [Gerencsér (1992, 1994)]). 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. Note that there is 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 (2001)], we extend the principle of generalized linear estimates to the prediction of Gaussian time series by considering the special case

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

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 $\tilde{h}^{(k)}$, applied in [Györfi and Lugosi (2001)] 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

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

Theorem 5.9 (Biau *et al.*, 2010). *Choose $\eta_t = 1/\sqrt{t}$. Then the prediction strategy g defined above is universally consistent with respect to the class of all jointly stationary and ergodic zero-mean Gaussian processes $\{Y_n\}_{-\infty}^\infty$.*

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

Corollary 5.3 (Biau *et al.*, 2010). *Under the conditions of Theorem 5.9,*

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

Corollary 5.3 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

$$\lim_{n\to\infty} \left(\mathbb{E}\{Y_n | Y_1^{n-1}\} - g(Y_1^{n-1})\right) = 0 \quad \text{almost surely} \tag{5.44}$$

for all stationary and ergodic Gaussian processes. [Schäfer (2002)] proved that, under some additional mild conditions on the Gaussian time series, the consistency (5.44) holds.

5.5. Pattern recognition for time series

In this section we apply the same ideas to the seemingly more difficult classification (or pattern recognition) problem. The setup is the following: let $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ be a stationary and ergodic sequence of pairs taking values in $\mathbb{R}^d \times \{0,1\}$. The problem is to predict the value of Y_n given the data (X_1^n, Y_1^{n-1}).

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

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

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

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

In this case there is a fundamental limit for the predictability of the sequence, i.e., [Algoet (1994)] proved that for any classification strategy f and stationary ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\liminf_{n\to\infty} R_n(f) \geq R^* \quad \text{a.s.}, \tag{5.45}$$

where

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

therefore the following definition is meaningful:

Definition 5.2. A classification strategy f is called *universally consistent* if for all stationary and ergodic processes $\{X_n, Y_n\}_{-\infty}^{\infty}$,

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

Therefore, universally consistent strategies asymptotically achieve the best possible loss for all ergodic processes. The first question is, of course, if such a strategy exists. [Ornstein (1978)] and [Bailey (1976)] proved the existence of universally consistent predictors. This was later generalized by [Algoet (1992)]. A simpler estimator with the same convergence property was introduced by [Morvai *et al.* (1996)]. Motivated by the need of a practical estimator, [Morvai *et al.* (1997)] introduced an even simpler algorithm. However, it is not known whether their predictor is universally consistent.

[Györfi *et al.* (1999)] introduced a simple randomized universally consistent procedure with a practical appeal. Their idea was to combine the decisions of a small number of simple experts in an appropriate way.

The same idea was used in [Weissman and Merhav (2004)]. They studied the consistency in noisy environment. In their model the past of Y_t is not available for the predictor, it has only access to the noisy past X_1^{t-1}. X_t is a noisy function of Y_t, that is, $X_t = u(Y_t, N_t)$, where $u : \{0,1\} \times \mathbb{R} \to \mathbb{R}$ is a function and $\{N_t\}$ is some noise process. A general loss function $\ell(f_t'(X_1^{t-1}), Y_t)$ is considered, where $f_t' : \mathbb{R}^{t-1} \to \mathbb{R}$ and $f_t'(X_1^{t-1})$ is the estimate of Y_t. They used an algorithm based on [Vovk (1998)] to combine the simple experts and used doubling trick to fit the algorithm to infinite time horizon. In case of $0-1$ loss, one may easily modify the results in the sequel such that, they can be applied for the problem of [Weissman and Merhav (2004)].

5.5.1. *Pattern recognition*

In pattern recognition, the label Y takes only finitely many values. For simplicity assume that Y takes two values, say 0 and 1. The aim is to predict the value of Y given the value of feature vector X (e.g., to predict whether a patient has a special disease or not, given some measurements of the patient like body temperature, blood pressure, etc.). The goal is to find a function $g^* : \mathbb{R}^d \to \{0,1\}$ which minimizes the probability of $g^*(X) \neq Y$, i.e., to find a function g^* such that

$$\mathbb{P}\{g^*(X) \neq Y\} = \min_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}\{g(X) \neq Y\}, \tag{5.46}$$

where g^* is called the Bayes decision function, and $\mathbb{P}\{g(X) \neq Y)$ is the probability of misclassification. (Concerning the details see [Devroye *et al.* (1996)].)

The Bayes decision function can be obtained explicitly.

Lemma 5.8.

$$g^*(x) = \begin{cases} 1 \ \textit{if} & \mathbb{P}\{Y = 1|X = x\} \geq 1/2, \\ 0 \ \textit{if} & \mathbb{P}\{Y = 1|X = x\} < 1/2, \end{cases}$$

is the Bayes decision function, i.e., g^* satisfies (5.46).

Proof. Let $g : \mathbb{R}^d \to \{0,1\}$ be an arbitrary (measurable) function. Fix

$x \in \mathbb{R}^d$. Then

$$\begin{aligned}\mathbb{P}\{g(X) \neq Y | X = x\} &= 1 - \mathbb{P}\{g(X) = Y | X = x\} \\ &= 1 - \mathbb{P}\{Y = g(x) | X = x\}.\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{P}\{g(X) \neq Y | X = x\} - \mathbb{P}\{g^*(X) \neq Y | X = x\} \\ &= \mathbb{P}\{Y = g^*(x) | X = x\} - \mathbb{P}\{Y = g(x) | X = x\} \geq 0,\end{aligned}$$

because

$$\mathbb{P}\{Y = g^*(x) | X = x\} = \max\left\{\mathbb{P}\{Y = 0 | X = x\}, \mathbb{P}\{Y = 1 | X = x\}\right\}$$

by the definition of g^*. This proves

$$\mathbb{P}\{g^*(X) \neq Y | X = x\} \leq \mathbb{P}\{g(X) \neq Y | X = x\}$$

for all $x \in \mathbb{R}^d$, which implies

$$\begin{aligned}\mathbb{P}\{g^*(X) \neq Y\} &= \int \mathbb{P}\{g^*(X) \neq Y | X = x\}\mu(dx) \\ &\leq \int \mathbb{P}\{g(X) \neq Y | X = x\}\mu(dx) \\ &= \mathbb{P}\{g(X) \neq Y\}.\end{aligned}$$

□

$\mathbb{P}\{Y = 1 | X = x\}$ and $\mathbb{P}\{Y = 0 | X = x\}$ are the so-called a posteriori probabilities. Observe that

$$\mathbb{P}\{Y = 1 | X = x\} = \mathbb{E}\{Y | X = x\} = m(x).$$

A natural approach is to estimate the regression function m by an estimate m_n using data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and then to use a so-called plug-in estimate

$$g_n(x) = \begin{cases} 1 \text{ if } m_n(x) \geq 1/2, \\ 0 \text{ if } m_n(x) < 1/2, \end{cases}$$

to estimate g^*. The next lemma implies that if m_n is close to the real regression function m, then the error probability of decision g_n is near to the error probability of the optimal decision g^*.

Lemma 5.9. *Let $\hat{m} : \mathbb{R}^d \to \mathbb{R}$ be a fixed function and define the plug-in decision $\hat{g}$ by*

$$\hat{g}(x) = \begin{cases} 1 \text{ if } \hat{m}(x) \geq 1/2, \\ 0 \text{ if } \hat{m}(x) < 1/2. \end{cases}$$

Then

$$\begin{aligned}0 &\le \mathbb{P}\{\hat{g}(X) \neq Y\} - \mathbb{P}\{g^*(X) \neq Y\}\\ &\le 2\int_{\mathbb{R}^d} |\hat{m}(x) - m(x)|\mu(dx)\\ &\le 2\left(\int_{\mathbb{R}^d} |\hat{m}(x) - m(x)|^2\mu(dx)\right)^{\frac{1}{2}}.\end{aligned}$$

Proof. It follows from the proof of Lemma 5.8 that, for arbitrary $x \in \mathbb{R}^d$,

$$\begin{aligned}&\mathbb{P}\{\hat{g}(X) \neq Y|X = x\} - \mathbb{P}\{g^*(X) \neq Y|X = x\}\\ &= \mathbb{P}\{Y = g^*(x)|X = x\} - \mathbb{P}\{Y = \hat{g}(x)|X = x\}\\ &= I_{\{g^*(x)=1\}}m(x) + I_{\{g^*(x)=0\}}(1 - m(x))\\ &\qquad - \left(I_{\{\hat{g}(x)=1\}}m(x) + I_{\{\hat{g}(x)=0\}}(1 - m(x))\right)\\ &= I_{\{g^*(x)=1\}}m(x) + I_{\{g^*(x)=0\}}(1 - m(x))\\ &\qquad - \left(I_{\{g^*(x)=1\}}\hat{m}(x) + I_{\{g^*(x)=0\}}(1 - \hat{m}(x))\right)\\ &\quad + \left(I_{\{g^*(x)=1\}}\hat{m}(x) + I_{\{g^*(x)=0\}}(1 - \hat{m}(x))\right)\\ &\qquad - \left(I_{\{\hat{g}(x)=1\}}\hat{m}(x) + I_{\{\hat{g}(x)=0\}}(1 - \hat{m}(x))\right)\\ &\quad + \left(I_{\{\hat{g}(x)=1\}}\hat{m}(x) + I_{\{\hat{g}(x)=0\}}(1 - \hat{m}(x))\right)\\ &\qquad - \left(I_{\{\hat{g}(x)=1\}}m(x) + I_{\{\hat{g}(x)=0\}}(1 - m(x))\right)\\ &\le I_{\{g^*(x)=1\}}(m(x) - \hat{m}(x)) + I_{\{g^*(x)=0\}}(\hat{m}(x) - m(x))\\ &\quad + I_{\{\hat{g}(x)=1\}}(\hat{m}(x) - m(x)) + I_{\{\hat{g}(x)=0\}}(m(x) - \hat{m}(x))\\ &\qquad \text{(because of}\\ &\qquad\quad I_{\{\hat{g}(x)=1\}}\hat{m}(x) + I_{\{\hat{g}(x)=0\}}(1 - \hat{m}(x)) = \max\{\hat{m}(x), 1 - \hat{m}(x)\}\\ &\qquad \text{by definition of } \hat{g})\\ &\le 2|\hat{m}(x) - m(x)|.\end{aligned}$$

Hence

$$\begin{aligned}0 &\le \mathbb{P}\{\hat{g}(X) \neq Y\} - \mathbb{P}\{g^*(X) \neq Y\}\\ &= \int \left(\mathbb{P}\{\hat{g}(X) \neq Y|X = x\} - \mathbb{P}\{g^*(X) \neq Y|X = x\}\right)\mu(dx)\\ &\le 2\int |\hat{m}(x) - m(x)|\mu(dx).\end{aligned}$$

The second assertion follows from the Cauchy-Schwarz inequality. □

It follows from Lemma 5.9 that the error probability of the plug-in decision g_n defined above satisfies

$$\begin{aligned} 0 &\le \mathbb{P}\{g_n(X) \neq Y | \mathcal{D}_n\} - \mathbb{P}\{g^*(X) \neq Y\} \\ &\le 2 \int_{\mathbb{R}^d} |m_n(x) - m(x)| \mu(dx) \\ &\le 2 \left(\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx) \right)^{\frac{1}{2}} . \end{aligned}$$

Thus estimates m_n with small L_2 error automatically lead to estimates g_n with small misclassification probability.

This can be generalized to the case where Y takes $M \ge 2$ distinct values, without loss of generality (w.l.o.g.) 1, …, M (e.g., depending on whether a patient has a special type of disease or no disease). The goal is to find a function $g^* : \mathbb{R}^d \to \{1, \ldots, M\}$ such that

$$\mathbb{P}\{g^*(X) \neq Y\} = \min_{g:\mathbb{R}^d \to \{1,\ldots,M\}} \mathbb{P}\{g(X) \neq Y\},$$

where g^* is called the Bayes decision function. It can be computed using the a posteriori probabilities $\mathbb{P}\{Y = k | X = x\}$ $(k \in \{1, \ldots, M\})$:

$$g^*(x) = \arg \max_{1 \le k \le M} \mathbb{P}\{Y = k | X = x\}.$$

The a posteriori probabilities are the regression functions

$$\mathbb{P}\{Y = k | X = x\} = \mathbb{E}\{I_{\{Y=k\}} | X = x\} = m^{(k)}(x).$$

Given data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, estimates $m_n^{(k)}$ of $m^{(k)}$ can be constructed from the data set

$$\mathcal{D}_n^{(k)} = \{(X_1, I_{\{Y_1=k\}}), \ldots, (X_n, I_{\{Y_n=k\}})\},$$

and one can use a plug-in estimate

$$g_n(x) = \arg \max_{1 \le k \le M} m_n^{(k)}(x)$$

to estimate g^*. If the estimates $m_n^{(k)}$ are close to the a posteriori probabilities, then again the error of the plug-in estimate is close to the optimal error.

5.5.2. *Prediction for binary labels*

In this section we present a simple (non-randomized) on-line classification strategy, and prove its universal consistency. Consider the prediction scheme $g_t(X_1^t, Y_1^{t-1})$ introduced in Sections 5.3.1 or 5.3.2 or 5.3.3 or 5.3.4, and then introduce the corresponding classification scheme:

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

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

Theorem 5.10 (Györfi and Ottucsák, 2007). *Assume that the conditions of Theorems 5.1 or 5.2 or 5.3 or 5.4. Then the classification scheme f defined above satisfies*

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

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

Proof. Because of (5.45) we have to show that

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

By Corollary 5.1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\{Y_t \mid X_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(X_1^t, Y_1^{t-1})\right)^2 = 0 \qquad \text{a.s.} \tag{5.47}$$

Introduce the Bayes classification scheme using the infinite past:

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

and its normalized cumulative $0-1$ loss:

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

Put

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

and

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

Then

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

and

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

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

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

so we have to show that

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

Lemma 5.9 implies that

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

where in the last step we applied (5.47). □

References

Algoet, P. (1992). Universal schemes for prediction, gambling, and portfolio selection, *Annals of Probability* **20**, pp. 901–941.

Algoet, P. (1994). The strong law of large numbers for sequential decisions under uncertainity, *IEEE Transactions on Information Theory* **40**, pp. 609–634.

Auer, P., Cesa-Bianchi, N. and Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms, *Journal of Computer and System Sciences* **64**, 1, pp. 48–75, a preliminary version has appeared in *Proc. 13th Ann. Conf. Computational Learning Theory.*

Bailey, D. H. (1976). *Sequential schemes for classifying and predicting ergodic processes*, Ph.D. thesis, Stanford University.

Breiman, L. (1957). The individual ergodic theorem of information theory, *Annals of Mathematical Statistics* **28**, pp. 809–811, correction. *Annals of Mathematical Statistics*, 31:809–810, 1960.

Cesa-Bianchi, N., Freund, Y., Helmbold, D. P., Haussler, D., Schapire, R. and Warmuth, M. K. (1997). How to use expert advice, *Journal of the ACM* **44**, 3, pp. 427–485.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games* (Cambridge University Press, Cambridge).

Chow, Y. S. (1965). Local convergence of martingales and the law of large numbers, *Annals of Mathematical Statistics* **36**, pp. 552–558.

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

Feder, M., Merhav, N. and Gutman, M. (1992). Universal prediction of individual sequences, *IEEE Trans. Inform. Theory* **IT-38**, pp. 1258–1270.

Gerencsér, L. (1992). $ar(\infty)$ estimation and nonparametric stochastic complexity, *IEEE Transactions on Information Theory* **38**, pp. 1768–1779.

Gerencsér, L. (1994). On rissanen's predictive stochastic complexity for stationary arma processes, *Journal of Statistical Planning and Inference* **41**, pp. 303–325.

Gerencsér, L. and Rissanen, J. (1986). A prediction bound for gaussian arma processes, in *Proc. of the 25th Conference on Decision and Control*, pp. 1487–1490.

Györfi, L. (1984). Adaptive linear procedures under general conditions, *IEEE Transactions on Information Theory* **30**, pp. 262–267.

Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression* (Springer, New York).

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

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

Györfi, L. and Wegkamp, M. (2008). Quantization for nonparametric regression, *IEEE Transactions on Information Theory* **54**, pp. 867–874.

Kivinen, J. and Warmuth, M. K. (1999). Averaging expert predictions, in P. F. H. U. Simon (ed.), *Computational Learning Theory: Proceedings of the Fourth European Conference, EuroCOLT'99*, no. 1572 in Lecture Notes in Artificial Intelligence (Springer-Verlag, Berlin), pp. 153–167.

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm, *Information and Computation* **108**, pp. 212–261.

Merhav, N. and Feder, M. (1998). Universal prediction, *IEEE Transactions on Information Theory* **IT-44**, pp. 2124–2147.

Morvai, G., Yakowitz, S. and Algoet, P. (1997). Weakly convergent stationary time series, *IEEE Transactions on Information Theory* **43**, pp. 483–498.

Morvai, G., Yakowitz, S. and Györfi, L. (1996). Nonparametric inference for ergodic, stationary time series, *Annals of Statistics* **24**, pp. 370–379.

Nobel, A. (2003). On optimal sequential prediction for general processes, *IEEE Transactions on Information Theory* **49**, pp. 83–98.

Ornstein, D. S. (1978). Guessing the next output of a stationary process, *Israel Journal of Mathematics* **30**, pp. 292–296.

Schäfer, D. (2002). Strongly consistent online forecasting of centered Gaussian processes, *IEEE Transactions on Information Theory* **48**, pp. 791–799.

Singer, A. C. and Feder, M. (1999). Universal linear prediction by model order weighting, *IEEE Transactions on Signal Processing* **47**, pp. 2685–2699.

Singer, A. C. and Feder, M. (2000). Universal linear least-squares prediction, in *Proceedings of the IEEE International Symposium on Information Theory.*

Stout, W. F. (1974). *Almost sure convergence* (Academic Press, New York).

Tsay, R. S. (2002). *Analysis of Financial Time Series* (Wiley, New York).

Tsypkin, Y. Z. (1971). *Adaptation and Learning in Automatic Systems* (Academic Press, New York).

Vovk, V. (1990). Aggregating strategies, in *Proceedings of the Third Annual Workshop on Computational Learning Theory* (Morgan Kaufmann, Rochester, NY), pp. 372–383.

Vovk, V. (1998). A game of prediction with expert advice, *Journal of Computer and System Sciences* **56**, pp. 153–173.

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

Yang, Y. (2000). Combining different procedures for adaptive regression, *Journal of Multivariate Analysis* **74**, pp. 135–161.