Weakly universally consistent forecasting of stationary and ergodic time series *

Daniel Jones[1,†], Michael Kohler[1] and Harro Walk[2]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: jones@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

[2] *Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany, email: walk@mathematik.uni-stuttgart.de.*

March 7, 2011

Abstract

Static forecasting of stationary and ergodic time series is considered, i.e., inference of the conditional expectation of the response variable at time zero given the infinite past. It is shown that the mean squared error of a combination of suitably defined localized least squares estimates converges to zero for all distributions where the response variable is square integrable.

AMS classification: Primary 62G05; secondary 62G20.

Key words and phrases: dependent data, forecasting, time series, weak consistency, mean squared error.

1 Introduction

In this paper we study the so-called static forecasting problem. More precisely, let $((X_n, Y_n))_{n\in\mathbb{Z}}$ be a stationary and ergodic sequence of $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\left\{Y_0^2\right\} < \infty$. Given the data set

$$\mathcal{D}_{-n}^{-1} = \{(X_{-n}, Y_{-n}), \ldots, (X_{-1}, Y_{-1})\}$$

*Running title: *Weakly universally consistent forecasting*

†Corresponding author. Tel: +49-6151-16-2394, Fax: +49-6151-16-6822

and X_0, we consider the problem to construct estimates $m_n(X_0, \mathcal{D}_{-n}^{-1})$ of

$$\mathbf{E}\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\}$$

such that

$$\mathbf{E}\left\{\left|m_n(X_0, \mathcal{D}_{-n}^{-1}) - \mathbf{E}\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\}\right|^2\right\} \to 0 \quad (n \to \infty).$$

For simplification, we have introduced the notation

$$Z_k^l = (Z_k, Z_{k+1}, \ldots, Z_l), \quad k \leq l$$

for arbitrary random variables Z_n ($n \in \mathbb{Z}$). Both the static forecasting problem and the related, but more complicated dynamic forecasting problem (including the special case of autoregression, cf., e.g., Chapter 27 in Györfi et al. (2002)) have evolved from the striving for generality in the estimation of dependent series.

Most of the results in the existing literature provide consistency in some way under the assumption of more or less strong mixing conditions on the data (see, e.g., the monograph by Györfi et al. (1989) for a review). Although there exist models where these conditions are met, they are very hard to verify - no satisfactory statistical tests are known. Therefore the question arises, whether there are estimates which are consistent under considerably weaker assumptions, e.g. stationarity and ergodicity of the data. As for the dynamic forecasting problem, there are several negative findings, see for example Bailey (1976) or Ryabko (1988), a summary can be found in Györfi, Morvai and Yakowitz (1998).

Concerning static forecasting, based on works of Ornstein (1978) and Algoet (1992), Morvai, Yakowitz and Györfi (1996) proposed an estimator, a modification of which can be shown to be strongly consistent for all stationary and ergodic data in the above defined sense (see Györfi et al. (2002), Section 27.3). For more results in this respect and concerning related problems, we refer to the works of Györfi, Lugosi and Morvai (1999), Györfi and Lugosi (2000), Györfi and Ottucsák (2007), as well as Morvai, Yakowitz and Algoet (1997).

Although the idea of the estimator of Morvai, Yakowitz and Györfi (1996) is natural and although it is easy to define, it can be expected to require large amounts of data, the same applies to the estimates of Algoet and Ornstein. This drawback makes the algorithms hard to apply, to the knowledge of the authors none of them has ever been applied to any data sets yet, neither real nor simulated. This motivates to try to derive estimates which can be computed easily, like the partition estimate for example. Unfortunately, there are negative findings in a static forecasting setting similar to the one studied in this paper: Györfi, Morvai and Yakowitz (1998) showed that a partitioning estimate which is strongly universally consistent in the case of mixing assumptions fails to be consistent when the data is only stationary and ergodic. It can therefore be assumed that one cannot expect to find a "simple" estimate which is strongly universally consistent.

In many applications, it is sufficient, that weak universal consistency holds. Kohler and Walk (2010) for example derived an optimal rule for exercising an American option by estimating conditional expectations assuming only that the returns of the underlying asset are stationary and ergodic. In the definition of the estimate, they use techniques from the theory of the prediction of individual sequences (cf., e.g., Cesa-Bianchi and Lugosi (2006)), which have already been successfully applied in connection with portfolio optimization (cf., e.g., Györfi, Lugosi and Udina (2006) and Györfi, Udina and Walk (2008)). In this paper we will adapt the ideas of Kohler and Walk (2010) in order to derive an estimate which is universally consistent for all stationary and ergodic data.

One of the main tricks in the proof is an averaging of estimates of different sample sizes, which enables us to derive weak consistency results from Cesàro consistency of the original estimates. Cesàro consistency of regression estimates in case of stationary and ergodic data was already studied in Morvai and Weiss (2005), where local averaging estimates in case of a finite alphabet were analyzed. In contrast, in this article we apply a different estimation principle and use a combination of simple estimates in order to choose the smoothing parameters of our procedure.

More precisely, we use local modeling combined with techniques from the theory of the prediction of individual sequences in order to define forecasting rules applicable to arbitrary stationary and ergodic time series. We show that the resulting estimate is consistent whenever the response variable is square integrable.
We consider several function spaces for our localized least squares estimate. Piecewise constant functions will lead to an estimate similar to the well-known kernel estimate. In addition, we consider estimates based on polynomial splines.

The definition of the estimate is given in Section 2, the main results are formulated in

Section 3, the proofs are given in Section 4.

2 Definition of the estimate

First of all we choose an elementary estimate (so-called expert) for our problem, which will be a localized least squares estimate. The idea is to select via the principle of localized least squares the function, that would have performed best in the past at the task of predicting Y_i only with the knowledge of the string $X_{i-j+1}^{i}, Y_{i-j+1}^{i-1}$ and then to predict Y_0 according to this very function and the arguments $X_{-j+1}^{0}, Y_{-j+1}^{-1}$. The parameter j quantizes how far we look back for our prediction. We then define our prediction strategy as a convex combination of these experts, where the weights depend on their performance in the past: The better the performance of the expert in the past, the more reliable it seems and thus the higher (with respect to the other experts) is the weight we assign to it.
In order to be able to show consistency of the estimate, we will at some point require boundedness of the estimate and the response variable, which is why we will also use some truncation techniques.
For $j, k, r, s \in \mathbb{N}$ let $\mathcal{F}_{j,k}$ be a set of functions $f : \left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1} \to \mathbb{R}$ (with an obvious meaning in case $j = 1$). Let K be a kernel function with corresponding bandwidth h_r (which both will be specified later), and choose $0 < t < \frac{1}{2}$. Given observed data

$$d_{-n}^{-1} = \{(x_{-n}, y_{-n}), \ldots, (x_{-1}, y_{-1})\},$$

define the corresponding localized least squares estimate by

$$\begin{aligned}
&\hat{m}_{n,(j,k,r,s)}(\cdot) := \hat{m}_{n,(j,k,r,s)}(\cdot, d_{-n}^{-j}) \\
&:= \arg\min_{f \in \mathcal{F}_{jk}} \frac{\sum_{i=-n+j+s}^{-j-s-1} \left| f(x_{i-j+1}^{i}, y_{i-j+1}^{i-1}) - T_{B_k}(y_i) \right|^2 \cdot K\left(\frac{(x_{i-j-s}^{i-j}, y_{i-j-s}^{i-j}) - (x_{-j-s}^{-j}, y_{-j-s}^{-j})}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{(x_{i-j-s}^{i-j}, y_{i-j-s}^{i-j}) - (x_{-j-s}^{-j}, y_{-j-s}^{-j})}{h_r} \right)},
\end{aligned}$$

where the truncation operator is defined as

$$T_L(x) = \begin{cases} x & \text{if } |x| \leq L, \\ L \operatorname{sign}(x) & \text{else.} \end{cases}$$

This definition only makes sense, if $-n+j+s \leq -j-s-1$, so we set

$$m_{n,(j,k,r,s)}(x_0, d_{-n}^{-1}) = \begin{cases} T_{n^t}\left(\hat{m}_{n,(j,k,r,s)}(x_{-j+1}^{0}, y_{-j+1}^{-1}, d_{-n}^{-j})\right) & \text{if } n \geq 2j+2s+1, \\ 0 & \text{else} \end{cases}$$

(with $d_{-n}^{-j} = \{(x_{-n}, y_{-n}), \ldots, (x_{-j}, y_{-j})\}$). For every (j,k,r,s), this gives us an expert who guesses the outcome of the next observation with knowledge of the observations of the past and the current value of the variable X. Here j quantizes how far we look back for our prediction, k is the number of the function space considered, r describes the bandwidth which we use and s quantizes how far we look back in the localized least squares problem. The last truncation ensures that for fixed sample size all estimates are bounded by the same constant. After a certain rounds of play, we consider for $n \geq 2$ the "cumulative loss"

$$L_n(j,k,r,s) = L_n((j,k,r,s), d_{-n}^{-1}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left(m_{i,(j,k,r,s)}\left(x_{-n+i}, d_{-n}^{-n+i-1}\right) - T_{n^t}\left(y_{-n+i}\right)\right)^2.$$

The cumulative loss quantizes how well our prediction strategy performed in the past. Let $(q_{(j,k,r,s)})_{j,k,r,s\in\mathbb{N}}$ be a probability distribution such that $q_{(j,k,r,s)} > 0$ for all $j,k \in \mathbb{N}$. Set $c_n = 8n^{2t}$ and define weights (depending on the cumulative loss)

$$w_{n,(j,k,r,s)} = q_{(j,k,r,s)} \cdot e^{-(n-1)\cdot L_n(j,k,r,s)/c_n},$$

and their normalized values

$$v_{n,(j,k,r,s)} = \frac{w_{n,(j,k,r,s)}}{\sum_{\alpha,\beta,\gamma,\delta=1}^{\infty} w_{n,(\alpha,\beta,\gamma,\delta)}}.$$

Set

$$\bar{m}_n(x_0, d_{-n}^{-1}) = \sum_{j,k,r,s=1}^{\infty} v_{n,(j,k,r,s)} \cdot m_{n,(j,k,r,s)}(x_0, d_{-n}^{-1}),$$

which is a convex combination of the experts with weights $v_{n,(j,k,r,s)}$. The final estimate $\hat{m}_n$ is defined by the arithmetic mean of these convex combinations extending their "backsight" with growing index i:

$$\hat{m}_n(X_0, \mathcal{D}_{-n}^{-1}) = \frac{1}{n}\sum_{i=1}^{n} \bar{m}_i(X_0, \mathcal{D}_{-i}^{-1}).$$

3 Main Results

In order to formulate our main result, we need the notion of sup-norm covering numbers, which we introduce in the next definition.

Definition 1. *Let $\varepsilon > 0$ and let $\mathcal{G}$ be a set of functions $\mathbb{R}^d \to \mathbb{R}$. Every finite collection of functions $g_1, ..., g_N : \mathbb{R}^d \to \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there is a $j = j(g) \in \{1, ..., N\}$ such that*

$$\|g - g_j\|_\infty := \sup_z |g(z) - g_j(z)| < \varepsilon,$$

*is called an ε-cover of $\mathcal{G}$ with respect to $\|\cdot\|_\infty$. Let $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_\infty)$ be the size of the smallest ε-cover of $\mathcal{G}$ w.r.t. $\|\cdot\|_\infty$, take $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_\infty) = \infty$ if no finite ε-cover exists. Then $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_\infty)$ is called the ε-**covering number of** $\mathcal{G}$ w.r.t. $\|\cdot\|_\infty$ and will be abbreviated to $\mathcal{N}_\infty(\varepsilon, \mathcal{G})$.*

Our main theorem is valid for all (strictly) stationary and ergodic sequences $((X_j, Y_j))_{j\in\mathbb{Z}}$. Here a sequence of $\mathbb{R}^l$-valued random variables $(Z_j)_{j\in\mathbb{Z}}$ defined on the same probability space is (strictly) stationary and ergodic if for each $B \in \mathcal{B}_\mathbb{Z}$ (where $\mathcal{B}_\mathbb{Z}$ is the Borel σ-algebra in $(\mathbb{R}^l)^\mathbb{Z}$) and each $k \in \mathbb{Z}$

$$\mathbf{P}\{(Z_j)_{j\in\mathbb{Z}} \in B\} = \mathbf{P}\{(Z_{j+k})_{j\in\mathbb{Z}} \in B\},$$

and if for each $B \in \mathcal{B}_\mathbb{Z}$ with the property that the event

$$A := \{(Z_{j+k})_{j\in\mathbb{Z}} \in B\}$$

does not depend on $k \in \mathbb{Z}$ one has

$$\mathbf{P}(A) \in \{0, 1\}$$

(cf., e.g., Breiman (1968), pp. 118, 119, Doob (1954), Section X.1, or Györfi et al. (2002), p. 565).

Theorem 1. *For $j, k \in \mathbb{N}$ let $\mathcal{F}_{j,k}$ be a set of functions such that the following conditions are satisfied:*

$$\mathcal{N}_\infty(\varepsilon, \mathcal{F}_{j,k}) < \infty \text{ for all } \varepsilon > 0, \tag{1}$$

and there exists a finite ϵ-cover consisting of piecewise constant functions with respect to a finite partition.

There exist $B_k \in \mathbb{R}$ $(k \in \mathbb{N})$ with

$$\sup_{f\in\mathcal{F}_{j,k}} \|f\|_\infty \le B_k < \infty \tag{2}$$

for all j and

$$\lim_{k\to\infty} B_k = \infty. \tag{3}$$

Furthermore for all j suppose that for any probability measure μ on $\left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1}$ and for every $g \in L_2\left(\left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1}, \mu\right)$

$$\liminf_{k\to\infty} \inf_{f\in\mathcal{F}_{j,k}} \int |g-f|^2 \ d\mu = 0. \tag{4}$$

Assume that $((X_j, Y_j))_{j\in\mathbb{Z}}$ is a stationary and ergodic sequence of $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\left\{Y_0^2\right\} < \infty$. Define the estimate $\hat{m}_n$ as in Section 2, where the kernel function $K : \mathbb{R}^{(s+1)\cdot(d+1)} \to \mathbb{R}_+$ is given by

$$K(v) := H\left(\|v\|_2^{(s+1)\cdot(d+1)}\right),$$

with a nonincreasing and continuous function $H : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying

$$H(0) > 0 \quad \text{and} \quad t \cdot H(t) \to 0 \quad (t \to \infty).$$

Suppose that the bandwidth satisfies

$$\lim_{r\to\infty} h_r = 0.$$

Then

$$\mathbf{E}\left\{\left|\hat{m}_n(X_0, \mathcal{D}_{-n}^{-1}) - \mathbf{E}(Y_0|X_{-\infty}^0, Y_{-\infty}^{-1})\right|^2\right\} \to 0 \quad (n \to \infty).$$

Next we apply Theorem 1 to piecewise constant functions. In this particular case, the local modeling estimate is given by a truncated localized kernel estimate which solves the localized least squares problem in case of bounded piecewise constant functions. Application of Theorem 1 yields

Corollary 1. *For every $j \in \mathbb{N}$ let $\left\{\mathcal{P}_{j,k} = \left\{A_{j,k}^1, A_{j,k}^2, ..., A_{j,k}^{N_{j,k}}\right\}\right\}_{k\in\mathbb{N}}$ be a sequence of partitions of $\left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1}$ consisting of Borel sets $A_{j,k}^l \subseteq \left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1}$ which satisfy*

$$\lim_{k\to\infty} \sup_{1\le l\le N_{j,k}: A_{j,k}^l \cap S \neq \emptyset} diam(A_{j,k}^l) = 0, \tag{5}$$

for every sphere S centered at the origin, where $diam(A)$ denotes the diameter of $A \subseteq \mathbb{R}^t$. Set

$$\mathcal{F}_{j,k} = \left\{\sum_{l=1}^{N_{j,k}} a_l \mathbb{1}_{A_{j,k}^l} : a_l \in \mathbb{R}, |a_l| \le \beta_k\right\}.$$

Let $\beta_k > 0$ $(k \in \mathbb{N})$ be such that

$$\lim_{k\to\infty} \beta_k = \infty. \tag{6}$$

Let $((X_j, Y_j))_{j\in\mathbb{Z}}$ be a stationary and ergodic sequence of $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\left\{Y_0^2\right\} < \infty$. Define the estimate $\hat{m}_n(X_0, \mathcal{D}_{-n}^{-1})$ as in Section 2. Then

$$\mathbf{E}\left\{\left|\hat{m}_n(X_0, \mathcal{D}_{-n}^{-1}) - \mathbf{E}(Y_0|X_{-\infty}^0, Y_{-\infty}^{-1})\right|^2\right\} \to 0 \quad (n \to \infty).$$

Proof. It is easy to see that we have for the ε–supremum norm covering number

$$\mathcal{N}_\infty\left(\varepsilon, \mathcal{F}_{j,k}\right) \leq \left(\left\lceil \frac{2\beta_k}{\varepsilon} \right\rceil + 1\right)^{N_{j,k}} < \infty,$$

where $\lceil z \rceil$ denotes the smallest integer greater than or equal to z. Here the functions of the ϵ-cover can be chosen as piecewise constant with respect to $\mathcal{P}_{j,k}$. Furthermore

$$\sup_{f\in\mathcal{F}_{j,k}} \|f\|_\infty = \beta_k =: B_k,$$

with $\lim_{k\to\infty} B_k = \infty$, so (1), (2) and (3) hold. It remains to check the denseness condition. Let μ be a probability measure on $\left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1}$. As the continuous functions of bounded support are dense in $L_2(\mu)$ (cf., e.g., Theorem A.1. in Györfi et al. (2002)), it suffices to show that for each $\varepsilon > 0$ and for each continuous function g of compact support and for each $K \in \mathbb{N}$ there exists $\bar{f} \in \{f : \ f \in \mathcal{F}_{j,k}, \ k \geq K\}$ such that

$$\int \left|\bar{f}(x) - g(x)\right|^2 \mu(dx) < \varepsilon.$$

Choose k such that $\|g\|_\infty \leq \beta_k$. Set

$$\bar{f}(x) = \sum_{l=1}^{N_{j,k}} g(x_{j,k}^l) \mathbb{1}_{A_{j,k}^l}(x)$$

for some fixed $x_{j,k}^l \in A_{j,k}^l$, $l = 1, ..., N_{j,k}$. Then $\|f\|_\infty \leq \|g\|_\infty \leq \beta_k$, so $f \in \mathcal{F}_{j,k}$. Let $\varepsilon > 0$ be arbitrary and let C be the support of g. Choose a sphere S centered at the origin with $C \subseteq S$ and $\mu(S^c) \leq \frac{\varepsilon}{8\|g\|_\infty^2}$. Then

$$\begin{aligned}\int \left|\bar{f}(x) - g(x)\right|^2 \mu(dx) &\leq \int_S \left|\bar{f}(x) - g(x)\right|^2 \mu(dx) + 4\|g\|_\infty^2 \cdot \mu(S^c) \\ &\leq \sup_{1\leq l\leq N_{j,k}: A_{j,k}^l \cap S \neq \emptyset} \ \sup_{x,y\in A_{j,k}^l} |g(x) - g(y)|^2 + \frac{\varepsilon}{2}.\end{aligned}$$

By uniform continuity of g on S and by (5) we can now increase k until

$$\sup_{1\leq l\leq N_{j,k}:A^l_{j,k}\cap S\neq\emptyset}\ \sup_{x,y\in A^l_{j,k}} |g(x)-g(y)|^2 \leq \frac{\varepsilon}{2}.$$

□

The estimate above locally fits a piecewise constant function to the data. As we will see from the proof of Theorem 1, these piecewise constant functions are used as approximation of various multivariate regression functions. In case that some of these regression functions are smooth, a smooth approximation might achieve a much smaller L_2 error than a piecewise constant function. Therefore we will define next an alternative estimate based on function spaces consisting of polynomial spline functions (i.e., piecewise polynomials which are globally smooth).

Depending on some parameters $k \in \mathbb{N}$, $L_k \in \mathbb{R}$ and $M \in \mathbb{N}_0$, we will define a space of tensor product spline functions $f : \mathbb{R}^d \to \mathbb{R}$. Let $B^1_{i,M}$ be the univariate B-spline with degree M, knot sequence $\left\{-L_k + \frac{l}{k}\right\}_{l\in\mathbb{Z}}$, and support

$$\left[-L_k + \frac{i}{k}, -L_k + \frac{i+M+1}{k}\right]$$

(cf., e.g., de Boor (1978), Chapter IX or Györfi et al. (2002), Section 14.1). For $\boldsymbol{i} = (i_1, ..., i_d) \in \mathbb{Z}^d$ define the tensor product B-spline

$$B^d_{\boldsymbol{i},M}(x_1, ..., x_d) = B^d_{i_1,M}(x_1) \cdot ... \cdot B^d_{i_d,M}(x_d).$$

The tensor product spline space $S_M\left([-L_k, L_k]^d\right)$ is then defined as

$$S_M\left([-L_k, L_k]^d\right) = span\left\{B^d_{\boldsymbol{i},M} :\ supp\left(B^d_{\boldsymbol{i},M}\right) \cap [-L_k, L_k]^d \neq \emptyset\right\}.$$

We will now impose some conditions on the parameters of the tensor product space which will assure that we can apply Theorem 1 to the resulting set of functions.

Corollary 2. *For every $j \in \mathbb{N}$ put $l_j = d \cdot j + j - 1$ and consider the following set of functions on $\mathbb{R}^{l_j}$ (where we identify $\left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1}$ and $\mathbb{R}^{l_j}$):*

$$\mathcal{F}_{j,k} = \left\{ \sum_{\boldsymbol{i}\in\mathbb{Z}^{l_j}:supp\left(B^{l_j}_{\boldsymbol{i},M}\right)\cap[-L_k,L_k]^{l_j}\neq\emptyset} a_{\boldsymbol{i}}\, B^{l_j}_{\boldsymbol{i},M} : |a_{\boldsymbol{i}}| \leq \beta_k \right\},$$

where the parameters fulfil $M \leq M_{max}(k)$ for $M_{max}(k) \in \mathbb{N}_0$ and where

$$\beta_k \to \infty \quad (k \to \infty), \tag{7}$$

$$L_k \to \infty \quad (k \to \infty), \tag{8}$$

$$\frac{M_{max}(k)+1}{k} \to 0 \quad (k \to \infty). \tag{9}$$

Let $((X_j, Y_j))_{j \in \mathbb{Z}}$ be a stationary and ergodic sequence of $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\left\{Y_0^2\right\} < \infty$. Define the estimate $\hat{m}_n(X_0, \mathcal{D}_{-n}^{-1})$ as in Section 2. Then

$$\mathbf{E}\left\{\left|\hat{m}_n(X_0, \mathcal{D}_{-n}^{-1}) - \mathbf{E}(Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1})\right|^2\right\} \to 0 \quad (n \to \infty).$$

Proof. Consider the set of functions

$$\bar{\mathcal{F}}_{j,k} = \left\{ \sum_{i \in \mathbb{Z}^{l_j} : supp\left(B_{i,M}^{l_j}\right) \cap [-L_k, L_k]^{l_j} \neq \emptyset} a_{\boldsymbol{i}}\, B_{\boldsymbol{i},M}^{l_j} : a_{\boldsymbol{i}} \in \left\{-\beta_k, -\beta_k + \varepsilon, ..., -\beta_k + \left\lfloor \frac{2\beta_k}{\varepsilon} \right\rfloor \cdot \varepsilon \right\} \right\},$$

where $\lfloor z \rfloor$ denotes the largest integer less than or equal to z. Using the fact, that the B-splines are nonnegative and sum up to one (cf. de Boor (1978), p. 109, 110), it can easily be seen that $\bar{\mathcal{F}}_{j,k}$ is an ε–supremum norm cover of $\mathcal{F}_{j,k}$. Thus

$$\mathcal{N}_\infty\left(\varepsilon, \mathcal{F}_{j,k}\right) \leq \left|\bar{\mathcal{F}}_{j,k}\right| < \infty,$$

where $|\cdot|$ denotes the cardinality of a set. Furthermore $\bar{\mathcal{F}}_{j,k}$ consists of piecewise polynomials with bounded coefficients. These can be approximated arbitrarily well in supremum norm by piecewise constant functions and hence (1) holds. By construction of $\mathcal{F}_{j,k}$, (2) and (3) hold. As for the denseness condition, let μ be an arbitrary probability measure on $\mathbb{R}^{l_j}$. As in Corollary 1, it suffices to show that any continuous function g of bounded support can be approximated arbitrarily well with respect to the $\liminf$–condition (4). Because of (7) and (8), we may further assume that $\|g\|_\infty \leq \beta_k$ and that the support of g is contained in $[-L_k, L_k]^j$.
Set $I = \left\{\left(i_1, ..., i_{l_j}\right) \in \mathbb{Z}^{l_j} : supp\left(B_{(i_1,...,i_{l_j}),M}^{l_j}\right) \cap [-L_k, L_k]^{l_j} \neq \emptyset\right\}$. For $i \in I$ choose $u_i \in supp\left(B_{i,M}^{l_j}\right)$ and set

$$f = \sum_{i \in I} g(u_i) \cdot B_{i,M}^{l_j}.$$

Because of the fact that the B-splines sum up to one and are nonnegative, we have that $\|f\|_\infty \leq \|g\|_\infty$ and $f \in \mathcal{F}_{j,k}$. This implies

$$\begin{aligned}
&\int_{\mathbb{R}^{l_j}} |g(x) - f(x)|^2 \, \mu(dx) \\
&\leq 4\, \|g\|_\infty^2 \cdot \mu\left(\mathbb{R}^{l_j} \setminus [-L_k, L_k]^{l_j}\right) + \int_{[-L_k,L_k]^{l_j}} |g(x) - f(x)|^2 \, \mu(dx) \\
&\leq 4\, \|g\|_\infty^2 \cdot \mu\left(\mathbb{R}^{l_j} \setminus [-L_k, L_k]^{l_j}\right) + \sup_{x \in [-L_k,L_k]^{l_j}} |g(x) - f(x)|^2 .
\end{aligned}$$

By (8) we have that

$$\mu\left(\mathbb{R}^{l_j} \setminus [-L_k, L_k]^{l_j}\right) \to 0 \qquad (k \to \infty).$$

By using once more the fact that the B-splines sum up to one and are nonnegative we have that, for given $x \in [-L_k, L_k]^{l_j}$,

$$\begin{aligned}
|g(x) - f(x)| &\leq \sum_{i \in I} |g(u_i) - g(x)| \cdot B_{i,M}^{l_j}(x) \\
&\leq \sup_{i : x \in supp(B_{i,M}^{l_j})} |g(u_i) - g(x)| \\
&\leq \sup_{u,v \in \mathbb{R}^{l_j},\ \|u-v\|_\infty \leq (M_{max}(k)+1)/k} |g(u) - g(v)| .
\end{aligned}$$

Because of (9) and the fact that g is continuous and of bounded support we can conclude

$$\sup_{x \in [-L_k,L_k]^{l_j}} |g(x) - f(x)|^2 \to 0 \qquad (k \to \infty).$$

□

4 Proof of Theorem 1

In the proof of Theorem 1 we will apply Lemma 27.3 in Györfi et al. (2002), which we reformulate here as

Lemma 1. *For a prediction strategy g based on the sequence of decision functions $\{g_i\}_{i=1}^\infty$ with*

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

define the normalized cumulative prediction error on the string x_1^n, y_1^n *as*

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

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ *be a sequence of prediction strategies (experts), and let* $\{q_k\}_k$ *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*

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

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

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

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

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

then, for every $n \geq 1$,

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

Here $-\ln(0)$ *is treated as* ∞.

Proof. See proof of Lemma 27.3 in Györfi et al. (2002). □

Proof of Theorem 1. The proof will be divided into several steps. *In the first step of the proof* we show that the assertion follows from

$$\limsup_{n\to\infty} \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \bar{m}_i(X_i, \mathcal{D}_0^{i-1})\right|^2\right\} \leq L^*, \tag{10}$$

where

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

Because of

$$\begin{aligned}
&\mathbf{E}\left\{\left|\hat{m}_n(X_0, \mathcal{D}_{-n}^{-1}) - \mathbf{E}(Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1})\right|^2\right\} \\
&= \mathbf{E}\left\{\left|Y_0 - \hat{m}_n(X_0, \mathcal{D}_{-n}^{-1})\right|^2\right\} - \mathbf{E}\left\{\left|Y_0 - \mathbf{E}(Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1})\right|^2\right\},
\end{aligned} \tag{11}$$

the assertion of Theorem 1 follows from

$$\mathbf{E}\left\{\left|Y_0 - \hat{m}_n(X_0, \mathcal{D}_{-n}^{-1})\right|^2\right\} \to L^* \quad (n \to \infty). \tag{12}$$

By (11) we have that

$$L^* \leq \mathbf{E}\left\{\left|Y_0 - \hat{m}_n(X_0, \mathcal{D}_{-n}^{-1})\right|^2\right\}.$$

The definition of the estimate, the inequality of Jensen and the stationarity of the data imply

$$\begin{aligned}
\mathbf{E}\left\{\left|Y_0 - \hat{m}_n(X_0, \mathcal{D}_{-n}^{-1})\right|^2\right\} &= \mathbf{E}\left\{\left|Y_0 - \frac{1}{n}\sum_{i=1}^{n} \bar{m}_i(X_0, \mathcal{D}_{-i}^{-1})\right|^2\right\} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left\{\left|Y_0 - \bar{m}_i(X_0, \mathcal{D}_{-i}^{-1})\right|^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left\{\left|Y_i - \bar{m}_i(X_i, \mathcal{D}_0^{i-1})\right|^2\right\} \\
&= \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \bar{m}_i(X_i, \mathcal{D}_0^{i-1})\right|^2\right\},
\end{aligned}$$

so (12) follows indeed from (10).

In the second step of the proof we show that (10) follows in turn from

$$\inf_{j,k,r,s\in\mathbb{N}} \limsup_{n\to\infty} \mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^{0}, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - Y_0\right|^2\right\} \leq L^*. \tag{13}$$

Let $\delta > 0$ be arbitrary. By using the inequality

$$(a+b)^2 \leq (1+\delta)\, a^2 + (1+\frac{1}{\delta})\, b^2 \tag{14}$$

for arbitrary $a, b \in \mathbb{R}$, $\delta > 0$ we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \left|\bar{m}_i(X_i, \mathcal{D}_0^{i-1}) - Y_i\right|^2 \\
&\leq (1+\delta)\frac{1}{n}\sum_{i=1}^{n} \left|\bar{m}_i(X_i, \mathcal{D}_0^{i-1}) - T_{n^t}(Y_i)\right|^2 + (1+\frac{1}{\delta})\frac{1}{n}\sum_{i=1}^{n} \left|Y_i - T_{n^t}(Y_i)\right|^2.
\end{aligned}$$

From Y_0 being square integrable, Lebesgue's dominated convergence theorem and the stationarity of the sequence we conclude

$$\begin{aligned}
\limsup_{n\to\infty} (1+\frac{1}{\delta})\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} \left|Y_i - T_{n^t}(Y_i)\right|^2\right\} &= \limsup_{n\to\infty} (1+\frac{1}{\delta})\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left\{\left|Y_i - T_{n^t}(Y_i)\right|^2\right\} \\
&= (1+\frac{1}{\delta}) \limsup_{n\to\infty} \mathbf{E}\left\{\left|Y_0 - T_{n^t}(Y_0)\right|^2\right\} \\
&\leq (1+\frac{1}{\delta}) \limsup_{n\to\infty} \mathbf{E}\left\{Y_0^2\, \mathbb{1}_{\{|Y_0| > n^t\}}\right\} = 0.
\end{aligned}$$

By Lemma 1 (applied in an obviously modified version for a finite sequence of prediction strategies) we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\bar{m}_i(X_i,\mathcal{D}_0^{i-1})-T_{n^t}\left(Y_i\right)\right|^2\\
&\le \inf_{j,k,r,s\in\mathbb{N}}\left(\frac{1}{n}\sum_{i=1}^{n}\left|m_{i,(j,k,r,s)}(X_i,\mathcal{D}_0^{i-1})-T_{n^t}\left(Y_i\right)\right|^2-c_n\cdot\frac{\ln q_{j,k,r,s}}{n}\right),
\end{aligned}$$

which implies (noting $\lim_{n\to\infty}\frac{c_n}{n}=0$)

$$\begin{aligned}
&\limsup_{n\to\infty}\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\bar{m}_i(X_i,\mathcal{D}_0^{i-1})-T_{n^t}\left(Y_i\right)\right|^2\right\}\\
&\le \limsup_{n\to\infty}\mathbf{E}\left\{\inf_{j,k,r,s\in\mathbb{N}}\left(\frac{1}{n}\sum_{i=1}^{n}\left|m_{i,(j,k,r,s)}(X_i,\mathcal{D}_0^{i-1})-T_{n^t}\left(Y_i\right)\right|^2-c_n\cdot\frac{\ln q_{j,k,r,s}}{n}\right)\right\}\\
&\le \inf_{j,k,r,s\in\mathbb{N}}\limsup_{n\to\infty}\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|m_{i,(j,k,r,s)}(X_i,\mathcal{D}_0^{i-1})-T_{n^t}\left(Y_i\right)\right|^2\right\}\\
&= \inf_{j,k,r,s\in\mathbb{N}}\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left\{\left|m_{i,(j,k,r,s)}(X_i,\mathcal{D}_0^{i-1})-T_{n^t}\left(Y_i\right)\right|^2\right\}\\
&\le \inf_{j,k,r,s\in\mathbb{N}}\limsup_{n\to\infty}\mathbf{E}\left\{\left|m_{n,(j,k,r,s)}(X_n,\mathcal{D}_0^{n-1})-T_{n^t}\left(Y_n\right)\right|^2\right\}\\
&= \inf_{j,k,r,s\in\mathbb{N}}\limsup_{n\to\infty}\mathbf{E}\left\{\left|m_{n,(j,k,r,s)}(X_0,\mathcal{D}_{-n}^{-1})-T_{n^t}\left(Y_0\right)\right|^2\right\}. \qquad (15)
\end{aligned}$$

The last equality is due to the stationarity of the sequence. We observe that in the analysis of the limes superior of $m_{n,(j,k,r,s)}$ we can assume without loss of generality that $n\ge 2j+2s+1$, thus by definition

$$m_{n,(j,k,r,s)}(X_0,\mathcal{D}_{-n}^{-1})=T_{n^t}\left(\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^{0},Y_{-j+1}^{-1},\mathcal{D}_{-n}^{-j})\right).$$

Now by $|T_\beta\left(z\right)-y|\le|z-y|$ for $|y|\le\beta$ and inequality (14) it holds for arbitrary $\alpha>0$ that

$$\begin{aligned}
&\mathbf{E}\left\{\left|\left(m_{n,(j,k,r,s)}(X_0,\mathcal{D}_{-n}^{-1})\right)-T_{n^t}\left(Y_0\right)\right|^2\right\}\\
&\le \mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^{0},Y_{-j+1}^{-1},\mathcal{D}_{-n}^{-j})-T_{n^t}\left(Y_0\right)\right|^2\right\}\\
&\le (1+\alpha)\mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^{0},Y_{-j+1}^{-1},\mathcal{D}_{-n}^{-j})-Y_0\right|^2\right\}+(1+\frac{1}{\alpha})\mathbf{E}\left\{\left|Y_0-T_{n^t}\left(Y_0\right)\right|^2\right\}.
\end{aligned}$$

Similar reasoning as before yields

$$\limsup_{n\to\infty}\mathbf{E}\left\{\left|m_{n,(j,k,r,s)}(X_0,\mathcal{D}_{-n}^{-1})-T_{n^t}\left(Y_0\right)\right|^2\right\}$$

$$\leq \limsup_{n\to\infty}(1+\alpha)\mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^0, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - Y_0\right|^2\right\}$$

for arbitrary $\alpha > 0$ and hence

$$\limsup_{n\to\infty}\mathbf{E}\left\{\left|m_{n,(j,k,r,s)}(X_0, \mathcal{D}_{-n}^{-1}) - T_{n^t}\,(Y_0)\right|^2\right\}$$
$$\leq \limsup_{n\to\infty}\mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^0, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - Y_0\right|^2\right\}.$$

From this and (15) we see that (10) is indeed implied by (13).

In the following steps of the proof we will show

$$\inf_{j,k,r,s}\limsup_{n\to\infty}\mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^0, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - Y_0\right|^2\right\}$$
$$\leq \lim_{N\to\infty}\inf_{j,k\geq N}\inf_{f\in\mathcal{F}_{jk}}\mathbf{E}\left\{\left|f(X_{-j+1}^0, Y_{-j+1}^{-1}) - Y_0\right|^2\right\}. \tag{16}$$

Let $\delta > 0$ be arbitrary. By using inequality (14) we get

$$\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^0, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - Y_0\right|^2$$
$$\leq (1+\delta)\,\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^0, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - T_{B_k}\,(Y_0)\right|^2 + (1+\frac{1}{\delta})\,\left|T_{B_k}\,(Y_0) - Y_0\right|^2. \tag{17}$$

For simplification put $z_l^m := (x_l^m, y_l^m)$ for $l \leq m$ and define

$$g_f(z_{-j+1}^0) := \left|f(x_{-j+1}^0, y_{-j+1}^{-1}) - T_{B_k}\,(y_0)\right|^2$$

for $f : (\mathbb{R}^d)^j \times \mathbb{R}^{j-1} \to \mathbb{R}$. We notice that $\hat{m}_{n,(j,k,r,s)}$ depends on z_{-n}^{-j} and write in this context

$$g_{\hat{m}_{n,(j,k,r,s)}}(z_{-j+1}^0; z_{-n}^{-j}) := \left|\hat{m}_{n,(j,k,r,s)}(x_{-j+1}^0, y_{-j+1}^{-1}, d_{-n}^{-j}) - T_{B_k}\,(y_0)\right|^2.$$

For the same reason we will use below also the notation $g_{\hat{f}}(z_{-j+1}^0; z_{-n}^{-j})$.

Let $\varepsilon > 0$ be arbitrary. *In the third step of the proof* we show that for arbitrary $j, k, r, s \in \mathbb{N}$ and $\varepsilon > 0$

$$\limsup_{n\to\infty}\mathbf{E}\left\{g_{\hat{m}_{n,(j,k,r,s)}}(Z_{-j+1}^0; Z_{-n}^{-j})\right\}$$
$$\leq \limsup_{n\to\infty}\int\int g_{\hat{m}_{n,(j,k,r,s)}}(z_{-j+1}^0; z_{-n}^{-j})\, d\mathbf{P}_{Z_{-j+1}^0|Z_{-j-s}^{-j}=z_{-j-s}^{-j}}(z_{-j+1}^0)\, d\mathbf{P}_{Z_{-n}^{-j}}(z_{-n}^{-j})$$
$$+ 3\varepsilon + T_{jks\varepsilon}^{(1)}, \tag{18}$$

where

$$\limsup_{s\to\infty} T^{(1)}_{jks\varepsilon} = 0.$$

First of all we note that

$$\begin{aligned}
&\mathbf{E}\left\{g_{\hat{m}_{n,(j,k,r,s)}}(Z^0_{-j+1}; Z^{-j}_{-n})\right\} \\
&= \mathbf{E}\left\{\mathbf{E}\left\{g_{\hat{m}_{n,(j,k,r,s)}}(Z^0_{-j+1}; Z^{-j}_{-n})\Big| Z^{-j}_{-n}\right\}\right\} \\
&= \int\int g_{\hat{m}_{n,(j,k,r,s)}}(z^0_{-j+1}; z^{-j}_{-n})\, d\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-n}=z^{-j}_{-n}}(z^0_{-j+1})\, d\mathbf{P}_{Z^{-j}_{-n}}(z^{-j}_{-n}).
\end{aligned}$$

Put $\varepsilon_1 := \frac{\varepsilon}{4B_k}$ and denote by $\bar{F}^{\varepsilon_1}_{jk}$ a corresponding smallest ε_1–supremum norm cover of $\mathcal{F}_{jk}$ consisting of piecewise constant functions. Without loss of generality we can assume that $\sup_{h\in\bar{F}^{\varepsilon_1}_{jk}}\|h\|_\infty \le B_k$, so for $f\in\mathcal{F}_{j,k}$, $\bar{f}\in\bar{F}^{\varepsilon_1}_{jk}$ we have by $a^2-b^2\le|a+b|\,|a-b|$ for $a,b\in\mathbb{R}$ that

$$\left\|g_f - g_{\bar{f}}\right\|_\infty \le 4B_k\left\|f-\bar{f}\right\|_\infty.$$

Choose (depending on z^{-j}_{-n}) $\hat{f}\in\bar{F}^{\varepsilon_1}_{jk}$ with $\left\|\hat{f}-\hat{m}_{n,(j,k,r,s)}\right\|_\infty < \varepsilon_1$. Then

$$\begin{aligned}
&\int g_{\hat{m}_{n,(j,k,r,s)}}(z^0_{-j+1}; z^{-j}_{-n})\, d\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-n}=z^{-j}_{-n}}(z^0_{-j+1}) \\
&\le \int g_{\hat{f}}(z^0_{-j+1}; z^{-j}_{-n})\, d\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-n}=z^{-j}_{-n}}(z^0_{-j+1}) + \varepsilon \\
&= \int g_{\hat{f}}(z^0_{-j+1}; z^{-j}_{-n})\, d\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-j-s}=z^{-j}_{-j-s}}(z^0_{-j+1}) + \varepsilon \\
&\qquad + \int g_{\hat{f}}(z^0_{-j+1}; z^{-j}_{-n})\, d\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-n}=z^{-j}_{-n}}(z^0_{-j+1}) \\
&\quad - \int g_{\hat{f}}(z^0_{-j+1}; z^{-j}_{-n})\, d\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-j-s}=z^{-j}_{-j-s}}(z^0_{-j+1})
\end{aligned}$$

As $T_{B_k}(Y_0)$ is bounded and $\bar{F}^{\varepsilon_1}_{jk}$ consists of piecewise constant functions, there exists a finite partition $\mathcal{A}^{\varepsilon}_{jk}$ such that $g_{\hat{f}}$ can be approximated in supremum norm up to an error of at most ϵ by a function which is piecewise constant with respect to $\mathcal{A}^{\varepsilon}_{jk}$. Using this result we see that the latter difference above can be bounded in absolute value by

$$4B_k^2\cdot\sum_{A\in\mathcal{A}^{\varepsilon}_{jk}}\left|\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-n}=z^{-j}_{-n}}(A) - \mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-j-s}=z^{-j}_{-j-s}}(A)\right| + \varepsilon.$$

Here we have used that for a piecwise constant function h with respect to a finite partition $\mathcal{A}$ and measures ν, μ it holds that

$$\left|\int h d\mu - \int h d\nu\right| \le \|h\|_\infty\cdot\sum_{A\in\mathcal{A}}|\mu(A)-\nu(A)|.$$

We conclude by the martingale convergence theorem (cf., e.g., Theorem 11.10 in Klenke (2008)) and dominated convergence

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbf{E}\left\{g_{\hat{m}_{n,(j,k,r,s)}}(Z_{-j+1}^0; Z_{-n}^{-j})\right\} \\
&\leq \limsup_{n\to\infty} \int\int g_{\hat{f}}(z_{-j+1}^0; z_{-n}^{-j})\, d\mathbf{P}_{Z_{-j+1}^0|Z_{-j-s}^{-j}=z_{-j-s}^{-j}}(z_{-j+1}^0)\, d\mathbf{P}_{Z_{-n}^{-j}}(z_{-n}^{-j}) + 2\varepsilon \\
&\quad + 4B_k^2 \cdot \sum_{A\in\mathcal{A}_{jk}^{\varepsilon}} \limsup_{n\to\infty} \mathbf{E}\left\{\left|\mathbf{E}\left\{\mathbb{1}_A\left(Z_{-j+1}^0\right)\middle|Z_{-n}^{-j}\right\} - \mathbf{E}\left\{\mathbb{1}_A\left(Z_{-j+1}^0\right)\middle|Z_{-j-s}^{-j}\right\}\right|\right\} \\
&= \limsup_{n\to\infty} \int\int g_{\hat{f}}(z_{-j+1}^0; z_{-n}^{-j})\, d\mathbf{P}_{Z_{-j+1}^0|Z_{-j-s}^{-j}=z_{-j-s}^{-j}}(z_{-j+1}^0)\, d\mathbf{P}_{Z_{-n}^{-j}}(z_{-n}^{-j}) + 2\varepsilon \\
&\quad + 4B_k^2 \cdot \sum_{A\in\mathcal{A}_{jk}^{\varepsilon}} \mathbf{E}\left\{\left|\mathbf{E}\left\{\mathbb{1}_A\left(Z_{-j+1}^0\right)\middle|Z_{-\infty}^{-j}\right\} - \mathbf{E}\left\{\mathbb{1}_A\left(Z_{-j+1}^0\right)\middle|Z_{-j-s}^{-j}\right\}\right|\right\}. \\
&\leq \limsup_{n\to\infty} \int\int g_{\hat{m}_{n,(j,k,r,s)}}(z_{-j+1}^0; z_{-n}^{-j})\, d\mathbf{P}_{Z_{-j+1}^0|Z_{-j-s}^{-j}=z_{-j-s}^{-j}}(z_{-j+1}^0)\, d\mathbf{P}_{Z_{-n}^{-j}}(z_{-n}^{-j}) + 3\varepsilon \\
&\quad + 4B_k^2 \cdot \sum_{A\in\mathcal{A}_{jk}^{\varepsilon}} \mathbf{E}\left\{\left|\mathbf{E}\left\{\mathbb{1}_A\left(Z_{-j+1}^0\right)\middle|Z_{-\infty}^{-j}\right\} - \mathbf{E}\left\{\mathbb{1}_A\left(Z_{-j+1}^0\right)\middle|Z_{-j-s}^{-j}\right\}\right|\right\},
\end{aligned}$$

thus (18) holds.

In the fourth step of the proof we show that for arbitrary $j, k, r, s \in \mathbb{N}$ and $\varepsilon > 0$

$$\begin{aligned}
&\limsup_{n\to\infty} \int\int g_{\hat{m}_{n,(j,k,r,s)}}(z_{-j+1}^0; z_{-n}^{-j})\, d\mathbf{P}_{Z_{-j+1}^0|Z_{-j-s}^{-j}=z_{-j-s}^{-j}}(z_{-j+1}^0)\, d\mathbf{P}_{Z_{-n}^{-j}}(z_{-n}^{-j}) \\
&\leq \limsup_{n\to\infty} \mathbf{E}\left\{\frac{\sum_{i=-n+j+s}^{-j-s-1} g_{\hat{m}_{n,(j,k,r,s)}}(Z_{i-j+1}^i; Z_{-n}^{-j}) \cdot K\left(\frac{Z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{Z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r}\right)}\right\} + T_{jkrs\varepsilon}^{(2)} + 2\varepsilon,
\end{aligned} \tag{19}$$

with

$$\limsup_{r\to\infty} T_{jkrs\varepsilon}^{(2)} = 0.$$

We return to $\hat{f}$ at the expense of ε and proceed with

$$\begin{aligned}
&\int g_{\hat{f}}(z_{-j+1}^0; z_{-n}^{-j})\, d\mathbf{P}_{Z_{-j+1}^0|Z_{-j-s}^{-j}=z_{-j-s}^{-j}}(z_{-j+1}^0) \\
&\leq \frac{\int g_{\hat{f}}(u_{-j+1}^0; z_{-n}^{-j}) \cdot K\left(\frac{z_{-j-s}^{-j} - u_{-j-s}^{-j}}{h_r}\right) d\mathbf{P}_{Z_{-j-s}^0}(u_{-j-s}^0)}{\int K\left(\frac{z_{-j-s}^{-j} - u_{-j-s}^{-j}}{h_r}\right) d\mathbf{P}_{Z_{-j-s}^0}(u_{-j-s}^0)}
\end{aligned}$$

$$+\left|\int g_{\hat f}(z^0_{-j+1};z^{-j}_{-n})\,d\mathbf{P}_{Z^0_{-j+1}|Z^{-j}_{-j-s}=z^{-j}_{-j-s}}(z^0_{-j+1})\right.$$

$$\left.-\frac{\int g_{\hat f}(u^0_{-j+1};z^{-j}_{-n})\cdot K\left(\frac{z^{-j}_{-j-s}-u^{-j}_{-j-s}}{h_r}\right)d\mathbf{P}_{Z^0_{-j-s}}(u^0_{-j-s})}{\int K\left(\frac{z^{-j}_{-j-s}-u^{-j}_{-j-s}}{h_r}\right)d\mathbf{P}_{Z^0_{-j-s}}(u^0_{-j-s})}\right|,$$

where the latter term can be bounded by

$$\sup_{f\in\bar F^{\varepsilon_1}_{jk}}\left|\mathbf{E}\left\{g_f(Z^0_{-j+1})\,|Z^{-j}_{-j-s}=z^{-j}_{-j-s}\right\}-\frac{\mathbf{E}\left\{g_f(Z^0_{-j+1})\cdot K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}}\right|.$$

This expression tends to 0 as r tends to infinity $\mathbf{P}_{Z^{-j}_{-j-s}}$-almost surely by Lemma 24.8 in Györfi et al. (2002) as $\bar F^{\varepsilon_1}_{jk}$ is finite. Furthermore dominated convergence can be applied since we deal with bounded random variables. We continue with the analysis of the remaining term, which can be rewritten as

$$\frac{\sum_{i=-n+j+s}^{-j-s-1}g_{\hat f}(z^i_{i-j+1};z^{-j}_{-n})\cdot K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1}K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)}$$

$$+\frac{\int g_{\hat f}(u^0_{-j+1};z^{-j}_{-n})\cdot K\left(\frac{z^{-j}_{-j-s}-u^{-j}_{-j-s}}{h_r}\right)d\mathbf{P}_{Z^0_{-j-s}}(u^0_{-j-s})}{\int K\left(\frac{z^{-j}_{-j-s}-u^{-j}_{-j-s}}{h_r}\right)d\mathbf{P}_{Z^0_{-j-s}}(u^0_{-j-s})}$$

$$-\frac{\sum_{i=-n+j+s}^{-j-s-1}g_{\hat f}(z^i_{i-j+1};z^{-j}_{-n})\cdot K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1}K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)},$$

where an upper bound for the last difference is given by

$$\sup_{f\in\bar F^{\varepsilon_1}_{jk}}\left|\frac{\mathbf{E}\left\{g_f(Z^0_{-j+1})\cdot K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}}-\frac{\sum_{i=-n+j+s}^{-j-s-1}g_f(z^i_{i-j+1})\cdot K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1}K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)}\right|.$$

From this we conclude that the left-hand side of (19) is bounded from above by

$$\limsup_{n\to\infty}\mathbf{E}\left\{\frac{\sum_{i=-n+j+s}^{-j-s-1}g_{\hat m_{n,(j,k,r,s)}}(Z^i_{i-j+1};Z^{-j}_{-n})\cdot K\left(\frac{Z^{-j}_{-j-s}-Z^{i-j}_{i-j-s}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1}K\left(\frac{Z^{-j}_{-j-s}-Z^{i-j}_{i-j-s}}{h_r}\right)}\right\}+2\varepsilon$$

$$+\int \sup_{f\in\bar{F}^{\varepsilon_1}_{jk}} \Bigg| \mathbf{E}\left\{ g_f(Z^0_{-j+1}) \,|Z^{-j}_{-j-s} = z^{-j}_{-j-s} \right\} - \frac{\mathbf{E}\left\{ g_f(Z^0_{-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}}{\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}} \Bigg| d\mathbf{P}_{Z^{-j}_{-j-s}}(z^{-j}_{-j-s})$$

$$+\limsup_{n\to\infty} \int \sup_{f\in\bar{F}^{\varepsilon_1}_{jk}} \Bigg| \frac{\mathbf{E}\left\{ g_f(Z^0_{-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}}{\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}} - \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z^i_{i-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)} \Bigg| d\mathbf{P}_{Z^{-j}_{-n}}(z^{-j}_{-n}).$$

In order to complete the proof of (19) we are reduced to verifying

$$\limsup_{n\to\infty} \int \sup_{f\in\bar{F}^{\varepsilon_1}_{jk}} \Bigg| \frac{\mathbf{E}\left\{ g_f(Z^0_{-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}}{\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\}} - \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z^i_{i-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z^{-j}_{-j-s}-z^{i-j}_{i-j-s}}{h_r}\right)} \Bigg| d\mathbf{P}_{Z^{-j}_{-n}}(z^{-j}_{-n}) = 0. \quad (20)$$

Because of $K \geq c \cdot I_{S_{0,\bar{r}}}$ for suitable $c > 0$, $\bar{r} > 0$, where $S_{0,\bar{r}}$ is the ball in $(\mathbb{R}^d)^{s+1} \times \mathbb{R}^{s+1}$ centered at 0 with radius $\bar{r}$, we have

$$\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\} \geq c \cdot \mathbf{P}_{Z^{-j}_{-j-s}}\left(z^{-j}_{-j-s} + S_{0,\bar{r}\cdot h_r}\right) > 0 \quad (21)$$

$\mathbf{P}_{Z^{-j}_{-j-s}}$-almost everywhere (cf., e.g., Györfi et al. (2002), pp. 499, 500). Let ε_2 be arbitrary and set

$$S_{\varepsilon_2} = \left\{ z^{-j}_{-j-s} \in (\mathbb{R}^d \times \mathbb{R})^{s+1} \quad : \quad \mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s}-Z^{-j}_{-j-s}}{h_r}\right)\right\} > \varepsilon_2 \right\}.$$

The boundedness of the considered functions yields

$$\int \sup_{f\in \bar{F}^{\varepsilon_1}_{jk}} \left| \frac{\mathbf{E}\left\{ g_f(Z^0_{-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}}{\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}} - \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z^i_{i-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)} \right| d\mathbf{P}_{Z^{-j}_{-n}}(z^{-j}_{-n})$$

$$\leq \int_{S_{\varepsilon_2}\times(\mathbb{R}^d)^{n-j-s}\times\mathbb{R}^{n-j-s}} \sup_{f\in \bar{F}^{\varepsilon_1}_{jk}} \left| \frac{\mathbf{E}\left\{ g_f(Z^0_{-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}}{\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}} - \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z^i_{i-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)} \right| d\mathbf{P}_{Z^{-j}_{-n}}(z^{-j}_{-n})$$

$$+ c_1 \cdot \mathbf{P}_{Z^{-j}_{-j-s}}\left(S^c_{\varepsilon_2} \right)$$

for some $c_1 \in \mathbb{R}_+$. By (21) we know

$$\mathbf{P}_{Z^{-j}_{-j-s}}(S^c_{\varepsilon_2}) \to 0 \quad (\varepsilon_2 \to 0).$$

In addition it holds that

$$\int_{S_{\varepsilon_2}\times(\mathbb{R}^d)^{n-j-s}\times\mathbb{R}^{n-j-s}} \sup_{f\in \bar{F}^{\varepsilon_1}_{jk}} \left| \frac{\mathbf{E}\left\{ g_f(Z^0_{-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}}{\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}} - \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z^i_{i-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)} \right| d\mathbf{P}_{Z^{-j}_{-n}}(z^{-j}_{-n})$$

$$= \int \int_{S_{\varepsilon_2}} \sup_{f\in \bar{F}^{\varepsilon_1}_{jk}} \left| \frac{\mathbf{E}\left\{ g_f(Z^0_{-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}}{\mathbf{E}\left\{ K\left(\frac{z^{-j}_{-j-s} - Z^{-j}_{-j-s}}{h_r} \right) \right\}} - \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z^i_{i-j+1}) \cdot K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z^{-j}_{-j-s} - z^{i-j}_{i-j-s}}{h_r} \right)} \right| d\mathbf{P}_{Z^{-j}_{-j-s}|Z^{-j-s-1}_{-n}=z^{-j-s-1}_{-n}}(z^{-j}_{-j-s}) \, d\mathbf{P}_{Z^{-j-s-1}_{-n}}(z^{-j-s-1}_{-n})$$

$$
\leq \int \sup_{z_{-j-s}^{-j} \in S_{\varepsilon_2}} \sup_{f \in \bar{F}_{jk}^{\varepsilon_1}} \left| \frac{\mathbf{E}\left\{ g_f(Z_{-j+1}^0) \cdot K\left(\frac{z_{-j-s}^{-j} - Z_{-j-s}^{-j}}{h_r} \right) \right\}}{\mathbf{E}\left\{ K\left(\frac{z_{-j-s}^{-j} - Z_{-j-s}^{-j}}{h_r} \right) \right\}} \right.
$$

$$
\left. - \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z_{i-j+1}^i) \cdot K\left(\frac{z_{-j-s}^{-j} - z_{i-j-s}^{i-j}}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z_{-j-s}^{-j} - z_{i-j-s}^{i-j}}{h_r} \right)} \right| d\mathbf{P}_{Z_{-n}^{-j-s-1}}(z_{-n}^{-j-s-1})
$$

$$
= \mathbf{E}\left\{ \sup_{z_{-j-s}^{-j} \in S_{\varepsilon_2}} \sup_{f \in \bar{F}_{jk}^{\varepsilon_1}} \left| \frac{\mathbf{E}\left\{ g_f(Z_{-j+1}^0) \cdot K\left(\frac{z_{-j-s}^{-j} - Z_{-j-s}^{-j}}{h_r} \right) \right\}}{\mathbf{E}\left\{ K\left(\frac{z_{-j-s}^{-j} - Z_{-j-s}^{-j}}{h_r} \right) \right\}} \right. \right.
$$

$$
\left. \left. - \frac{\frac{1}{n-2j-2s} \sum_{i=-n+j+s}^{-j-s-1} g_f(Z_{i-j+1}^i) \cdot K\left(\frac{z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r} \right)}{\frac{1}{n-2j-2s} \sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r} \right)} \right| \right\}.
$$

Because of the fact that K is continuous and vanishes at infinity nominator and denominator of the second term above almost surely converge uniformly with respect to $z_{-j-s}^{-j} \in S_{\varepsilon_2}$ by the ergodic theorem and the circumstance that the ε-cover is finite (cf., e.g., Krengel (1985), Chapter 4, Theorem 2.1). Since the limit of the denominator is greater than ε_2 on S_{ε_2} we even have uniform convergence of the fracture. Application of the dominated convergence theorem completes the proof of (20) which in turn implies the proof of (19).

The fifth step of the proof will be to demonstrate

$$
\limsup_{n\to\infty} \mathbf{E}\left\{ \frac{\sum_{i=-n+j+s}^{-j-s-1} g_{\hat{m}_{n,(j,k,r,s)}}(Z_{i-j+1}^i; Z_{-n}^{-j}) \cdot K\left(\frac{Z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{Z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r} \right)} \right\}
$$

$$
\leq \inf_{f \in \mathcal{F}_{jk}} \mathbf{E}\left\{ \left| f(X_{-j+1}^0, Y_{-j+1}^{-1}) - T_{B_k} Y_0 \right|^2 \right\} + 2\varepsilon + T_{jkrs\varepsilon}^{(3)} \tag{22}
$$

for arbitrary $j, k, r, s \in \mathbb{N}$ and $\varepsilon > 0$ where

$$
\limsup_{r\to\infty} T_{jkrs\varepsilon}^{(3)} = 0.
$$

By definition of the estimate we have

$$
\limsup_{n\to\infty} \mathbf{E}\left\{ \frac{\sum_{i=-n+j+s}^{-j-s-1} g_{\hat{m}_{n,(j,k,r,s)}}(Z_{i-j+1}^i; Z_{-n}^{-j}) \cdot K\left(\frac{Z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r} \right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{Z_{-j-s}^{-j} - Z_{i-j-s}^{i-j}}{h_r} \right)} \right\}
$$

$$
\begin{aligned}
&= \limsup_{n\to\infty} \mathbf{E}\left\{ \inf_{f\in\mathcal{F}_{jk}} \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(Z_{i-j+1}^{i}) \cdot K\left(\frac{Z_{-j-s}^{-j}-Z_{i-j-s}^{i-j}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{Z_{-j-s}^{-j}-Z_{i-j-s}^{i-j}}{h_r}\right)} \right\} \\
&\le \inf_{f\in\mathcal{F}_{jk}} \limsup_{n\to\infty} \mathbf{E}\left\{ \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(Z_{i-j+1}^{i}) \cdot K\left(\frac{Z_{-j-s}^{-j}-Z_{i-j-s}^{i-j}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{Z_{-j-s}^{-j}-Z_{i-j-s}^{i-j}}{h_r}\right)} \right\} \\
&= \inf_{f\in\mathcal{F}_{jk}} \limsup_{n\to\infty} \int \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z_{i-j+1}^{i}) \cdot K\left(\frac{z_{-j-s}^{-j}-z_{i-j-s}^{i-j}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z_{-j-s}^{-j}-z_{i-j-s}^{i-j}}{h_r}\right)}\, d\mathbf{P}_{Z_{-n}^{-j}}(z_{-n}^{-j}) \\
&\le \inf_{f\in\mathcal{F}_{jk}} \left(\limsup_{n\to\infty} \int \left| \frac{\sum_{i=-n+j+s}^{-j-s-1} g_f(z_{i-j+1}^{i}) \cdot K\left(\frac{z_{-j-s}^{-j}-z_{i-j-s}^{i-j}}{h_r}\right)}{\sum_{i=-n+j+s}^{-j-s-1} K\left(\frac{z_{-j-s}^{-j}-z_{i-j-s}^{i-j}}{h_r}\right)} \right.\right. \\
&\qquad\qquad \left. - \frac{\mathbf{E}\left\{ g_f(Z_{-j+1}^{0}) \cdot K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}}{\mathbf{E}\left\{ K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}} \right| d\mathbf{P}_{Z_{-n}^{-j}}(z_{-n}^{-j}) \\
&\qquad\quad \left. + \int \frac{\mathbf{E}\left\{ g_f(Z_{-j+1}^{0}) \cdot K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}}{\mathbf{E}\left\{ K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}}\, d\mathbf{P}_{Z_{-j-s}^{-j}}(z_{-j-s}^{-j}) \right) \\
&= \inf_{f\in\mathcal{F}_{jk}} \int \frac{\mathbf{E}\left\{ g_f(Z_{-j+1}^{0}) \cdot K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}}{\mathbf{E}\left\{ K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}}\, d\mathbf{P}_{Z_{-j-s}^{-j}}(z_{-j-s}^{-j}),
\end{aligned}
$$

where the last equality follows from the proof of (20). With the same arguments as already used, we see that this can be bounded by

$$
\begin{aligned}
&\inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{ g_f(Z_{-j+1}^{0})\right\} + 2\varepsilon \\
&\qquad + \int \sup_{\bar f\in \bar F_{jk}^{\varepsilon_1}} \left| \mathbf{E}\left\{ g_{\bar f}(Z_{-j+1}^{0}) | Z_{-j-s}^{-j} = z_{-j-s}^{-j}\right\} \right. \\
&\qquad\qquad \left. - \frac{\mathbf{E}\left\{ g_{\bar f}(Z_{-j+1}^{0}) \cdot K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}}{\mathbf{E}\left\{ K\left(\frac{z_{-j-s}^{-j}-Z_{-j-s}^{-j}}{h_r}\right)\right\}} \right| d\mathbf{P}_{Z_{-j-s}^{-j}}(z_{-j-s}^{-j}),
\end{aligned}
$$

where the latter term tends to zero as r tends to infinity by Lemma 24.8 in Györfi et al. (2002) and dominated convergence. This completes the proof of (22).

We sum up the results of *steps three to five of the proof*: Noticing (17), we have shown that for arbitrary $\varepsilon, \delta > 0$

$$
\begin{aligned}
&\inf_{j,k,r,s\in\mathbb{N}} \limsup_{n\to\infty} \mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^{0}, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - Y_0\right|^2\right\} \\
&\le \inf_{j,k,r,s\in\mathbb{N}} \Bigg((1+\delta)\cdot\Bigg(\inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{\left|f(X_{-j+1}^{0}, Y_{-j+1}^{-1}) - T_{B_k}(Y_0)\right|^2\right\} \\
&\qquad\qquad + S_{jks\varepsilon} + T_{jkrs\varepsilon} + 7\cdot\varepsilon\Bigg) + \left(1+\frac{1}{\delta}\right)\cdot \mathbf{E}\left\{|T_{B_k}(Y_0) - Y_0|^2\right\}\Bigg) \\
&\le \inf_{j,k,r,s\in\mathbb{N}} \Bigg((1+\delta)^2\cdot\Bigg(\inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{\left|f(X_{-j+1}^{0}, Y_{-j+1}^{-1}) - Y_0\right|^2\right\} + S_{jks\varepsilon} \\
&\qquad\qquad + T_{jkrs\varepsilon} + 7\cdot\varepsilon\Bigg) + (2+\delta)\cdot\left(1+\frac{1}{\delta}\right)\cdot \mathbf{E}\left\{|T_{B_k}(Y_0) - Y_0|^2\right\}\Bigg),
\end{aligned}
$$

where

$$\limsup_{s\to\infty} S_{jks\varepsilon} = 0,$$

for arbitrary j, k,

$$\limsup_{r\to\infty} T_{jkrs\varepsilon} = 0,$$

for arbitrary j, k, s. Using

$$\inf_{j,k,r,s}(a_{jkrs} + b_k) \le \lim_{N\to\infty} \inf_{j,k\ge N} \inf_{r,s} a_{jkrs} + \lim_{N\to\infty} \sup_{k\ge N} b_k$$

we can conclude

$$
\begin{aligned}
&\inf_{j,k,r,s} \limsup_{n\to\infty} \mathbf{E}\left\{\left|\hat{m}_{n,(j,k,r,s)}(X_{-j+1}^{0}, Y_{-j+1}^{-1}, \mathcal{D}_{-n}^{-j}) - Y_0\right|^2\right\} \\
&\le (1+\delta)^2\cdot \lim_{N\to\infty}\Bigg(\inf_{j,k\ge N}\Bigg(\inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{\left|f(X_{-j+1}^{0}, Y_{-j+1}^{-1}) - Y_0\right|^2\right\} \\
&\qquad\qquad + \inf_s\left(S_{jks\varepsilon} + \inf_r T_{jkrs\varepsilon}\right) + 7\cdot\varepsilon\Bigg)\Bigg) \\
&\qquad\qquad + (2+\delta)\cdot\left(1+\frac{1}{\delta}\right)\cdot \lim_{N\to\infty}\sup_{k\ge N} \mathbf{E}\left\{|T_{B_k}(Y_0) - Y_0|^2\right\} \\
&\le (1+\delta)^2\cdot \lim_{N\to\infty} \inf_{j,k\ge N} \inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{\left|f(X_{-j+1}^{0}, Y_{-j+1}^{-1}) - Y_0\right|^2\right\} + 7\cdot(1+\delta)^2\cdot\varepsilon.
\end{aligned}
$$

The choice of ε and δ was arbitrary, hence (16) holds.

The only point remaining is to bound

$$\lim_{N\to\infty} \inf_{j,k\ge N} \inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{\left|f(X_{-j+1}^{0}, Y_{-j+1}^{-1}) - Y_0\right|^2\right\}$$

by L^* which will be *the sixth and last step of the proof.* Straightforward calculation leads to

$$\mathbf{E}\left\{\left|f(X_{-j+1}^0, Y_{-j+1}^{-1}) - Y_0\right|^2\right\} = L^* + \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\} - \mathbf{E}\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\}\right|^2\right\}$$
$$+ \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\} - f(X_{-j+1}^0, Y_{-j+1}^{-1})\right|^2\right\}.$$

Thus

$$\begin{aligned}
&\lim_{N\to\infty} \inf_{j,k\geq N} \inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{\left|f(X_{-j+1}^0, Y_{-j+1}^{-1}) - Y_0\right|^2\right\} \\
&= L^* + \lim_{N\to\infty} \inf_{j,k\geq N} \left(\inf_{f\in\mathcal{F}_{jk}} \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\} - f(X_{-j+1}^0, Y_{-j+1}^{-1})\right|^2\right\} \right. \\
&\qquad \left. + \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\} - \mathbf{E}\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\}\right|^2\right\}\right) \\
&\leq L^* + \limsup_{j\to\infty} \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\} - \mathbf{E}\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\}\right|^2\right\} \\
&\quad + \lim_{N\to\infty} \inf_{j,k\geq N} \inf_{f\in\mathcal{F}_{j,k}} \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\} - f(X_{-j+1}^0, Y_{-j+1}^{-1})\right|^2\right\}.
\end{aligned}$$

Considering the second term put

$$W_j := \mathbf{E}\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\}.$$

The sequence $(W_j)_{j\in\mathbb{N}}$ is a martingale satisfying $\sup_{j\in\mathbb{N}} \mathbf{E}\left\{|W_j|^2\right\} \leq \mathbf{E}\left\{Y_0^2\right\} < \infty$. Hence it converges almost surely and in L_2 to a square integrable random variable (cf. Theorem 11.10 in Klenke (2008)) and the limit is $\mathbf{E}\{Y_0|X_{-\infty}^0, Y_{-\infty}^{-1}\}$ (see Theorem 35.5 in Billingsley (1979)).

In a final step, set $l_j = d \cdot j + j - 1$ and

$$m_j(x,y) = \mathbf{E}\left\{Y_0|X_{-j+1}^0 = x, Y_{-j+1}^{-1} = y\right\}.$$

Then, by the inequality of Jensen, $m_j \in L_2\left(\mathbb{R}^{l_j}, \mathbf{P}_{(X_{-j+1}^0, Y_{-j+1}^{-1})}\right)$ and

$$\begin{aligned}
&\mathbf{E}\left\{\left|\mathbf{E}\left\{Y_0|X_{-j+1}^0, Y_{-j+1}^{-1}\right\} - f\left(X_{-j+1}^0, Y_{-j+1}^{-1}\right)\right|^2\right\} \\
&= \mathbf{E}\left\{\left|m_j\left(X_{-j+1}^0, Y_{-j+1}^{-1}\right) - f\left(X_{-j+1}^0, Y_{-j+1}^{-1}\right)\right|^2\right\} \\
&= \int_{\mathbb{R}^{l_j}} |m_j(x,y) - f(x,y)|^2 \, d\mathbf{P}_{(X_{-j+1}^0, Y_{-j+1}^{-1})}(x,y),
\end{aligned}$$

where we again identify $\left(\mathbb{R}^d\right)^j \times \mathbb{R}^{j-1}$ and $\mathbb{R}^{l_j}$. This allows us to conclude by (4) that

$$\lim_{N\to\infty} \inf_{j,k\geq N} \inf_{f\in\mathcal{F}_{j,k}} \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X^0_{-j+1}, Y^{-1}_{-j+1}\} - f(X^0_{-j+1}, Y^{-1}_{-j+1})\right|^2\right\}$$
$$\leq \liminf_{j\to\infty} \liminf_{k\to\infty} \inf_{f\in\mathcal{F}_{j,k}} \mathbf{E}\left\{\left|\mathbf{E}\{Y_0|X^0_{-j+1}, Y^{-1}_{-j+1}\} - f(X^0_{-j+1}, Y^{-1}_{-j+1})\right|^2\right\} = 0.$$

The proof is complete. □

Acknowledgements

The authors would like to thank two anonymous referees for various very helpful comments.

References

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

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

[3] Billingsley, P. (1979). *Probability and Measure.* Wiley, New York.

[4] de Boor, C. (1978). *A Practical Guide to Splines.* Springer, New York.

[5] Breiman, L. (1968). *Probability.* Addison-Wesley, Reading, MA.

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

[7] Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[8] Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series.* Springer Verlag, New York.

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

[10] Györfi, L., and Lugosi, G. (2000). Strategies for sequential prediction of stationary time series. *Internat. Ser. Oper. Res. Management Sci.* **46**, pp. 225–248.

[11] Györfi, L., Lugosi, G., and Udina, F. (2006). Nonparametric kernel-based sequential investment strategies. *Mathematical Finance* **16**, pp. 337-357.

[12] Györfi, L., Lugosi, G., and Morvai, G. (1999). A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Trans. Inform. Theory* **45**, pp. 2642-2650.

[13] Györfi, L., Morvai, G., and Yakowitz, S. (1998). Limits to consistent on-line forecasting for ergodic time series. *IEEE Trans. Inform. Theory* **44**, pp. 886–892.

[14] Györfi, L., and Ottucsak, G. (2007). Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory* **53**, pp. 1866–1872.

[15] Györfi, L., Udina, F., and Walk, H. (2008). Nonparametric nearest neighbor based empirical portfolio selection strategies. *Statistics & Decisions* **26**, pp. 145-157.

[16] Klenke, A. (2008). *Probability Theory: A Comprehensive Course.* Springer, London.

[17] Kohler, M., and Walk, H. (2010). On optimal exercising of American options in discrete time for stationary and ergodic data. Submitted for publication.

[18] Krengel, U. (1985). *Ergodic Theorems.* de Gruyter, Berlin, New York.

[19] Morvai, G., Yakowitz, S., and Algoet, P. (1997). Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Inform. Theory* **43**, pp. 483–498.

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

[21] Morvai, G., and Weiss, B. (2005) Forward estimation for ergodic time series. *Ann. Inst. H. Poincaré Probab. Statist.* **41**, pp. 859–870

[22] Ornstein, D. (1978). Guessing the next output of a stationary process. *Israel J. Math.* **30**, pp. 292–296.

[23] Ryabko, B. Y. (1988). Prediction of random sequences and universal coding. *Problems Inform. Transmission* **24**, pp. 87–96.