Journal of Multivariate Analysis **68**, 138–164 (1999)
Article ID jmva.1998.1782, available online at http://www.idealibrary.com on IDEAL®

Universally Consistent Regression Function Estimation Using Hierarchial B-Splines

Michael Kohler

Universität Stuttgart, Stuttgart, Germany
E-mail: kohler@mathematik.uni-stuttgart.de

Received September 25, 1996; revised July 1, 1998

Estimation of multivariate regression functions from i.i.d. data is considered. We construct estimates by empirical L_2-error minimization over data-dependent spaces of polynomial spline functions. For univariate regression function estimation these spaces are spline spaces with data-dependent knot sequences. In the multivariate case, we use so-called hierarchical spline spaces which are defined as linear span of tensor product B-splines with nested knot sequences. The knot sequences of the chosen B-splines depend locally on the data.

We show the strong L_2-consistency of the estimators without any condition on the underlying distribution.

The estimators are similar to histogram regression estimators using data-dependent partitions and partitioning regression estimators based on local polynomial fits. The main difference is that the estimators considered here are smooth functions, which seems to be desirable especially in the case that the regression function to be estimated is smooth. © 1999 Academic Press

AMS 1991 subject classifications: 62G07, 62G20.

Key words and phrases: data-dependent partitions; integrated squared error; least squares estimate; polynomial splines; regression estimate; universal consistency.

1. INTRODUCTION

Let $(\mathbf{X}, Y)$, $(\mathbf{X}_1, Y_1)$, $(\mathbf{X}_2, Y_2)$, ... be independent identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random vectors with $EY^2 < \infty$. In the regression analysis one wants to estimate Y after having observed $\mathbf{X}$; i.e., one wishes to find a function f with $f(\mathbf{X})$ "close" to Y. If the main goal of the analysis is the minimization of the mean squared error, then one wants to find a function m^* with

$$E(m^*(\mathbf{X}) - Y)^2 = \min_f E(f(\mathbf{X}) - Y)^2. \tag{1}$$

0047-259X/99 $30.00

Let $m(\mathbf{x}) := E(Y \mid \mathbf{X} = \mathbf{x})$ be the regression function. It is well known that we have for each measurable function f,

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

Here μ stands for the distribution of $\mathbf{X}$. Therefore m is the solution of the minimization problem (1) and for an arbitrary f the so-called *excess error* $\int_{\mathbb{R}^d} |m(\mathbf{x}) - f(\mathbf{x})|^2 \, \mu(d\mathbf{x})$ is the difference between $E(f(\mathbf{X}) - Y)^2$ and the optimal value $E(m(\mathbf{X}) - Y)^2$.

For the regression estimation problem the distribution of $(\mathbf{X}, Y)$ (and therefore also m) is unknown. Given only a sample $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)$ of the distribution of $(\mathbf{X}, Y)$ one has to construct an estimator $m_n(\mathbf{x}) = m_n(\mathbf{x}, (\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n))$ of $m(\mathbf{x})$.

A sequence of estimators $(m_n)_{n \in \mathbb{N}}$ is called **weakly (strongly) universally consistent** if

$$\int_{\mathbb{R}^d} |m_n(\mathbf{x}, (\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)) - m(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \to 0 \qquad \text{in } L_1 \quad \text{(a.s.)}$$

for all distributions of $(\mathbf{X}, Y)$ with $EY^2 < \infty$.

Stone (1977) first pointed out that there exist weakly universally consistent estimators. Since then various results about weak and strong universal consistency of special estimators, e.g., kernel estimators, nearest neighbor estimators, histogram estimators, and series estimators, have been published. See Devroye *et al.* (1994) for a list of papers on universal consistency and, in addition, Györfi and Walk (1996, 1997), Györfi *et al.* (1998), Kohler (1997), and Walk (1997).

In this paper we examine estimators which are defined by the following three steps: In the first step, one uses the sample to construct a space $\mathcal{F}_n = \mathcal{F}_n((\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n))$ of functions $f\colon \mathbb{R}^d \to \mathbb{R}$. In the second step, one chooses a function

$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, (\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)) \in \mathcal{F}_n \tag{3}$$

such that

$$\frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(\mathbf{X}_i) - Y_i|^2 = \inf_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2, \tag{4}$$

i.e., $\tilde{m}_n$ minimizes the empirical risk over $\mathcal{F}_n$. In the third step the estimator is truncated at β_n and $-\beta_n$, i.e., m_n is defined by

$$m_n(x) = (\tilde{m}_n(x) \vee (-\beta_n)) \vee \beta_n \qquad (x \in \mathbb{R}^d). \tag{5}$$

Here we use the notation $a \wedge b := \min\{a, b\}$, $a \vee b := \max\{a, b\}$ $(a, b \in \mathbb{R})$ with a constant $\beta_n \in \mathbb{R}_+ \cup \{\infty\}$ depending only on n, $\beta_n \to \infty$ $(n \to \infty)$.

Next we want to motivate the data-dependent choice of $\mathcal{F}_n$. We are interested in a small $L_2(\mu)$-error $\int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \, \mu(d\mathbf{x})$ between the estimator m_n and the true regression function m. The influence of the pointwise error $|m_n(\mathbf{x}) - m(\mathbf{x})|^2$ in some area of $\mathbb{R}^d$ on this $L_2(\mu)$-error depends on μ. Thus to get a small $L_2(\mu)$-error the estimator should approximate m in some areas of $\mathbb{R}^d$ much better than in others, where these areas depend on the distribution μ of $\mathbf{X}$. The data-dependent choice of $\mathcal{F}_n$ allows us first to use the sample to estimate μ and then to choose $\mathcal{F}_n$ such that m can be approximated especially well in areas where the pointwise error has a high influence on the $L_2(\mu)$-error.

We now give two examples for possible choices of $\mathcal{F}_n$.

Example 1 (Histogram Regression Estimators Using Data-Dependent Partitions). Use the sample to construct a finite or countably infinite partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \ldots\}$ of Borel-measurable subsets of $\mathbb{R}^d$, e.g., such that each $A_{n,i}$ contains the same number of the $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Define $\mathcal{F}_n$ as the space of all functions which are constant on each set of this partition. Set $A_n(x) := A_{n,j}$ if $x \in A_{n,j}$. Then it is easy to see that for

$$\tilde{m}_n(x) := \frac{\sum_{1 \leqslant i \leqslant n,\, \mathbf{X}_i \in A_n(x)} Y_i}{\sum_{1 \leqslant i \leqslant n,\, \mathbf{X}_i \in A_n(x)} 1}$$

($\frac{0}{0}$ is 0 by definition) (3) and (4) hold. In the case $\beta_n = \infty$, $m_n = \tilde{m}_n$ is the so-called data-dependent histogram regression estimator, see, e.g., Nobel (1996) and references therein.

Example 2 (Partitioning Regression Estimators Based on Local Polynomial Fits). A natural extension of Example 1 is to fit a polynomial of fixed degree greater than zero (instead of a constant) to the data in each set of the partition. In order to do this, one defines $\mathcal{F}_n$ as the set of all functions which are equal to a polynomial of fixed degree on each set of a data-dependent partition.

In both examples the data are used to construct a partition of $\mathbb{R}^d$ and then the estimator is defined locally in each set of the partition independently of the data not contained in this set. As a result the estimators as functions of $\mathbf{x}$ are generally not continuous. This seems to be unpleasant especially in the case that the regression function to be estimated is smooth. In order to avoid this we use polynomial spline functions, i.e., piecewise polynomial functions with global smoothness.

For univariate X we define $\mathcal{F}_n$ as polynomial spline space with data-dependent knots. We give general conditions for the data-dependent location and number of these knots which imply strong universal consistency of the resulting estimator.

For multivariate X we need spaces of multivariate spline functions. For these one often uses tensor products of univariate spline spaces. In the case d = 2 this means that one chooses knots $u_0 \leqslant u_1 \leqslant \cdots \leqslant u_{K_x}$ and $v_0 \leqslant v_1 \leqslant \cdots \leqslant v_{K_y}$ and defines tensor product splines as functions which are equal to a polynomial (of fixed degree in x and y) in each rectangle $[u_i, u_{i+1}) \times [v_j, v_{j+1})$ and which satisfy some global smoothness condition (e.g. one requires that the functions are continuous). The main drawback of this is that a local refinement of one of these rectangles is not possible: If one wants to refine the rectangle $[u_i, u_{i+1}) \times [v_j, v_{j+1})$ one must insert a new knot u between u_i and u_{i+1} (or a new knot v between v_j and v_{j+1}) which leads to a refinement of all rectangles $[u_i, u_{i+1}) \times [v_k, v_{k+1})$ $(0 \leqslant k \leqslant K_y)$ (or all rectangles $[u_k, u_{k+1}) \times [v_j, v_{j+1})$ $(0 \leqslant k \leqslant K_x)$). Therefore we don't simply use tensor product spline spaces with fixed knot sequences.

In the multivariate case, we define the data-dependent spaces $\mathcal{F}_n$ as linear span of tensor product B-splines with nested knot-sequences, where the knot-sequences of the chosen B-splines depend locally on the data. These so-called hierarchical spline spaces were used for surface approximation in computational geometry by Forsey and Bartels (1988), and their vector space dimension and local approximation properties were examined in Kraft (1994, 1997). By the aid of a local approximation property of such spline spaces we show strong universal consistency of the resulting estimators.

1.1. *Discussion of Related Work*

Empirical L_2-error minimization over data-independent spaces of functions was used in Lugosi and Zeger (1995) to construct strongly universally consistent series and neural network estimators. In Kohler (1997) this principle was used to show the strong universal consistency of a modification of the classical least squares spline estimator.

In this paper we consider estimators which are defined by empirical L_2-error minimization over *data-dependent* spaces of functions. For such estimators it is possible first to use the sample to derive properties of the distribution μ of $\mathbf{X}$ and the regression function m and then to choose $\mathcal{F}_n$ such that m can ben approximated well by functions of $\mathcal{F}_n$ in $L_2(\mu)$. For example, if $\mathbf{X}_1, ..., \mathbf{X}_n$ are contained in a hyperplane then one can choose $\mathcal{F}_n$ such that functions defined on this hyperplane can be approximated especially well.

Another difference from Lugosi and Zeger (1995) and Kohler (1997) is that we don't restrict the values of the functions in $\mathcal{F}_n$. In Lugosi and Zeger (1995) and also in Kohler (1997) all functions in $\mathcal{F}_n$ are bounded in absolute value by a constant depending on the size of the sample. The drawback of this is that for such spaces $\mathcal{F}_n$ we have no knowledge of any fast algorithm to compute a function $\tilde{m}_n$ which satisfies (4) for a given sample. In this paper we set

$$\mathcal{F}_n = \left\{ \sum_{j=1}^{K} a_j f_{j,n} \,\middle|\, a_j \in \mathbb{R} \right\},$$

with $K \in \mathbb{N}$ and functions $f_{1,n}, ..., f_{K,n}: \mathbb{R}^d \to \mathbb{R}$ depending on the sample. In this case (3) and (4) are equivalent to

$$\tilde{m}_n = \sum_{j=1}^{K} a_j f_{j,n} \tag{6}$$

for some $\mathbf{a} = (a_j)_j \in \mathbb{R}^K$ depending on the sample which satisfies

$$\mathbf{A}^T \mathbf{A} \mathbf{a} = \mathbf{A}^T \mathbf{Y}, \tag{7}$$

where

$$\mathbf{A} = (f_{j,n}(\mathbf{X}_i))_{1 \leqslant i \leqslant n,\, 1 \leqslant j \leqslant K} \qquad \text{and} \qquad \mathbf{Y} = (Y_1, ..., Y_n)^T.$$

Thus to compute $\tilde{m}_n$ one simply has to solve the linear equation system (7).

Estimators which are similar to the estimators defined in this paper are data-dependent histogram regression estimators and partitioning regression estimators based on data-dependent partitions (see Examples 1 and 2). General sufficient conditions for strong universal consistency in the case of bounded Y of these estimators can be found in Nobel (1996) and Lugosi and Nobel (1996). These results differ in two ways from the results in this paper: First, Nobel (1996) and Lugosi and Nobel (1996) consider only the case of bounded Y. Second, Nobel (1996) and Lugosi and Nobel (1996) use non-smooth function spaces $\mathcal{F}_n$. As a consequence, their estimators are generally not continuous. In this paper we use spaces of smooth functions; therefore the estimators yield always smooth functions as estimates of regression functions, which seems to be desirable especially in the case that the regression function to be estimated is smooth.

1.2. *Organization of the Paper*

The main results are formulated in Section 2 for univariate splines and in Section 3 for (multivariate) hierarchical splines. In Section 4 some preliminary results are presented. We use these results in Sections 5 and 6 to prove the results of Section 2 and 3, resp. In the Appendix we prove a local approximation property of hierarchical B-splines and we give a list of some results of the so-called Vapnik–Chervonenkis theory which are used in Sections 4, 5, and 6.

2. UNIVARIATE POLYNOMIAL SPLINE FUNCTIONS

In this section we use spaces of univariate polynomial spline functions, i.e., piecewise polynomials with global smoothness, in the general definition of the estimator of Section 1.

Let $M \in \mathbb{N}_0$, $\tilde{K} \in \mathbb{N}$, $t_0, ..., t_{\tilde{K}} \in \mathbb{R}$ with $t_0 < t_1 < \cdots < t_{\tilde{K}}$ and let $v_1, ..., v_{\tilde{K}-1} \in \{1, ..., M+1\}$. The spline space $S_{\mathbf{t}, v, M}$ with knots $t_0, ..., t_{\tilde{K}}$, knot multiplicities $v_1, ..., v_{\tilde{K}-1}$ and degree M is defined as the set of all functions $f = [t_0, t_{\tilde{K}}) \to \mathbb{R}$ which satisfy

(I) For each $i \in \{1, ..., \tilde{K}\}$ f is equal to a polynomial of degree $\leqslant M$ on $[t_{i-1}, t_i)$,

(II) If $i \in \{1, ..., \tilde{K}-1\}$ and $M - v_i \geqslant 0$ then f is $M - v_i$ times continuously differentiable at t_i.

EXAMPLE 3 For $M = 1$ and $v_1 = \cdots = v_{\tilde{K}-1} = 1$ the functions in $S_{\mathbf{t}, v, M}$ are continuous and piecewise linear.

Clearly, $S_{\mathbf{t}, v, M}$ is a linear space with finite dimension. In order to handle such functions on a computer one can represent them as linear combinations of chosen basis functions. For computational purposes it is convenient to use basis functions with support as small as possible. These are the so-called normalized B-splines, which will be introduced next.

Let $u_{-M} \leqslant u_{-M+1} \leqslant \cdots \leqslant u_{K+M}$ be real values such that $K = 1 + \sum_{i=1}^{\tilde{K}-1} v_i$, $u_0 = t_0$, $u_K = t_{\tilde{K}}$ and each t_i is contained exactly v_i times among the $u_1, ..., u_{K-1}$ $(i = 1, ..., \tilde{K}-1)$. Then the normalized B-splines $B_{j, l, \mathbf{u}}$ can be defined recursively by

$$B_{j, 0, \mathbf{u}}(x) := \begin{cases} 1 & \text{if} \quad u_j \leqslant x < u_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

for $j = -M, -M+1, ..., K+M-1$ and

$$B_{j, l+1, \mathbf{u}}(x) := \frac{x - u_j}{u_{j+l+1} - u_j} \cdot B_{j, l, \mathbf{u}}(x) + \frac{u_{j+l+2} - x}{u_{j+l+2} - u_{j+1}} \cdot B_{j+1, l, \mathbf{u}}(x) \tag{9}$$

for $j = -M, ..., K+M-2-l$, $l = 0, ..., M-1$.

It is well known that $\{B_{j,M,\mathbf{u}} \mid j=-M, ..., K-1\}$ is a basis for $S_{\mathbf{u},M} := S_{\mathbf{t},\nu,M}$ (see de Boor (1978), pp. 113, 114, 131). The basic properties of this B-spine basis are

$$B_{j,M,\mathbf{u}}(x) \geqslant 0 \tag{10}$$

for all $x \in \mathbb{R}$, $j=-M, ..., K-1$,

$$supp\ B_{j,M,\mathbf{u}} = [u_j, u_{j+M+1}] \tag{11}$$

for $j=-M, ..., K-1$, and

$$\sum_{j=-M}^{K-1} B_{j,M,\mathbf{u}}(x) = 1 \tag{12}$$

for all $x \in [u_0, u_K)$ (see de Boor (1978), pp. 109, 110). Here we have used the notation *supp f* for the support of a function f: $\mathbb{R}^d \to \mathbb{R}$.

Example 3 (continued). For $M=1$ and $u_{-1} < u_0 < \cdots < u_K < u_{K+1}\, B_{j,M,\mathbf{u}}$ is one at u_{j+1}, zero at $u_{-1}, ..., u_j, u_{j+2}, ..., u_{K+1}$ and piecewise linear between the knots (hat-function).

Observe that it is possible to define the B-splines $B_{j,M,\mathbf{u}}$ on whole $\mathbb{R}$ by (8) and (9). We will do this in the sequel. Then $S_{\mathbf{u},M}$ is a subset of the set of all functions which are equal to a polynomial on each of the sets $(-\infty, u_{-M})$, $[u_{-M}, u_{-M+1})$, ..., $[u_{K+M}, \infty)$.

In order to define the space $S_{\mathbf{u},M}$ one simply has to choose the knot sequence $\mathbf{u}$ and the degree M. In the next theorem we show how to choose the knots of $\mathscr{F}_n = S_{\mathbf{u},M}$ in dependence of the sample in order to define a strongly universally consistent estimator via (3), (4), and (5).

Let μ_n be the empirical measure to $X_1, ..., X_n$, i.e., $\mu_n(A) := 1/n \sum_{i=1}^{n} I_A(X_i)$ $(A \subseteq \mathbb{R})$.

Theorem 1. *Let* $M \in \mathbb{N}_0$, $K_n \in \mathbb{N}$, $\beta_n \in \mathbb{R}_+$ $(n \in \mathbb{N})$. *Depending on the sample* $(X_1, Y_1), ..., (X_n, Y_n)$ *choose* $K \in \mathbb{N}$ *and knots* $u_{-M}, ..., u_{K+M} \in \mathbb{R}$ *such that* $K \leqslant K_n$, $u_{-M} \leqslant u_{-M+1} \leqslant \cdots \leqslant u_{K+M}$ *and* $u_i < u_{i+M+1}$ *for all* $-M \leqslant i < K-1$. *Assume that*

$$\beta_n \to \infty \qquad (n \to \infty), \tag{13}$$

$$\frac{K_n \cdot \beta_n^4 \cdot \ln n}{n} \to 0 \qquad (n \to \infty) \tag{14}$$

and

$$\frac{\beta_n^4}{n^{1-\delta}} \to 0 \qquad (n \to \infty) \tag{15}$$

for some $\delta > 0$. If, in addition, for every distribution μ of X

$$\mu\left(\left\{(-\infty, u_0) \cup \bigcup_{\substack{k=1, ..., K, \\ u_k - u_{k-M-1} > \gamma}} [u_{k-1}, u_k) \cup [u_K, \infty)\right\} \cap [-L, L]\right)$$
$$\xrightarrow{(n \to \infty)} 0 \qquad \text{a.s.} \tag{16}$$

for every $L, \gamma > 0$, or

$$\mu_n\left(\left\{(-\infty, u_0) \cup \bigcup_{\substack{k=1, ..., K, \\ u_k - u_{k-M-1} > \gamma}} [u_{k-1}, u_k) \cup [u_K, \infty)\right\} \cap [-L, L]\right)$$
$$\xrightarrow{(n \to \infty)} 0 \qquad \text{a.s.} \tag{17}$$

for every $L, \gamma > 0$, then every sequence $(m_n)_{n \in \mathbb{N}}$ of estimators which satisfy (3), (4), *and* (5) *with $\mathcal{F}_n = S_{\mathbf{u}, M}$ is strongly universally consistent.*

Remark 1. The left-hand sides of (16) and (17) are random variables because the knots $u_{-M}, ..., u_{K+M}$ are random variables depending on the sample $(X_1, Y_1), ..., (X_n, Y_n)$.

We now give examples which show that it is easy to choose the knots such that (16) or (17) hold. In the first example we consider data-independent knots.

Example 4. Let $L_n, R_n \in \mathbb{R}$ $(n \in \mathbb{N})$ such that

$$L_n \to -\infty, \qquad R_n \to \infty \qquad (n \to \infty) \tag{18}$$

and

$$\frac{R_n - L_n}{K_n} \to 0 \qquad (n \to \infty). \tag{19}$$

Set $K = K_n$ and

$$u_k := L_n + k \cdot \frac{R_n - L_n}{K_n} \qquad (k = -M, ..., K_n + M). \tag{20}$$

Then (16) holds because for fixed $L, \gamma > 0$

$$\left\{(-\infty, u_0) \cup \bigcup_{k=1, ..., K,\, u_k - u_{k-M-1} > \gamma} [u_{k-1}, u_k) \cup [u_K, \infty)\right\} \cap [-L, L] = \varnothing$$

for n big enough.

In the next example we consider data-dependent knots.

Example 5. Let $C_n \in \mathbb{N}$, $\delta_n \geqslant 0$ $(n \in \mathbb{N})$ such that

$$\delta_n \to 0 \qquad (n \to \infty) \tag{21}$$

and

$$\frac{C_n}{n} \to 0 \qquad (n \to \infty). \tag{22}$$

Set $K = K_n$ and choose the knots such that there are less than C_n of the $X_1, ..., X_n$ in each of the intervals $(-\infty, u_0)$ and $[u_{K_n}, \infty)$ and such that for every $k \in \{1, ..., K_n\}$ with $u_k - u_{k-M-1} > \delta_n$ there are less than C_n of the $X_1, ..., X_n$ in $[u_{k-1}, u_k)$.

We now show that in this case (17) holds. Let $L, \gamma > 0$. Because of (21) we can assume w.l.o.g. that $\delta_n < \gamma$. Then $u_k - u_{k-M-1} > \gamma$ implies $\mu_n([u_{k-1}, u_k)) \leqslant C_n/n$; thus

$$\mu_n\left(\left\{(-\infty, u_0) \cup \bigcup_{\substack{k=1, ..., K_n \\ u_k - u_{k-M-1} > \gamma}} [u_{k-1}, u_k) \cup [u_{K_n}, \infty)\right\} \cap [-L, L]\right)$$
$$\leqslant \mu_n((-\infty, u_0)) + \sum_{\substack{u_k - u_{k-M-1} > \gamma \\ [u_{k-1}, u_k) \cap [-L, L] \neq \varnothing}} \mu_n([u_{k-1}, u_k)) + \mu_n([u_{K_n}, \infty))$$
$$\leqslant 2\,\frac{C_n}{n} + (M+1)\left(\frac{2L}{\gamma} + 2\right)\frac{C_n}{n} \to 0 \qquad (n \to 0)$$

because of (22), and thus (17) is proved.

Example 6. Choose each $\lceil n/K_n \rceil$th order statistic of $X_1, ..., X_n$ as a knot. Then each sequence $(m_n)_{n \in \mathbb{N}}$ of estimators which satisfies (3), (4), and (5) with $\mathscr{F}_n = S_{\mathbf{u}, M}$ is strongly consistent for every distribution of (X, Y) with X non-atomic and $EY^2 < \infty$, provided that (13)–(15) and

$$K_n \to \infty \qquad (n \to \infty) \tag{23}$$

hold.

This follows immediately from Theorem 1 and Example 5 by setting $C_n = \lceil n/K_n \rceil + 1$ and $\delta_n = 0$.

3. MULTIVARIATE POLYNOMIAL SPLINE FUNCTIONS—HIERARCHICAL B-SPLINES

In this section we introduce the so-called hierarchical B-splines and use them for the estimator of Section 1.

For notational convenience we restrict our considerations to the case $d=2$. The general case can be handled in an analogous way.

We define multivariate spline functions, i.e., multivariate piecewise polynomials with global smoothness, by tensor products of univariate spline functions.

Choose degrees $M_x, M_y \in \mathbb{N}_0$, $K_x, K_y \in \mathbb{N}$ and knot sequences $\mathbf{u}=(u_{-M_x}, ..., u_{K_x+M_x})$, $\mathbf{v}=(v_{-M_y}, ..., v_{K_y+M_y})$ and set

$$S_{\mathbf{u}, M_x, \mathbf{v}, M_y} := \left\{ f\colon \mathbb{R}^2 \to \mathbb{R} \;\middle|\; f(x, y) = \sum_{i=1}^{l} a_i g_i(x)\, h_i(y) \right.$$
$$\forall x, y \in \mathbb{R} \text{ for some } l \in \mathbb{N},$$
$$\left. a_1, ..., a_l \in \mathbb{R}, g_1, ..., g_l \in S_{\mathbf{u}, M_x} \text{ and } h_1, ..., h_l \in S_{\mathbf{v}, M_y} \right\}.$$

As the basis of $S_{\mathbf{u}, M_x, \mathbf{v}, M_y}$ we use tensor products of the B-spline basis of $S_{\mathbf{u}, M_x}$ and $S_{\mathbf{v}, M_y}$: Set

$$J := \{(j, l) \in \mathbb{Z}^2 \mid -M_x \leqslant j < K_x \text{ and } -M_y \leqslant l < K_y\}$$

and define

$$B_{(j,l)}(x, y) := B_{(j,l), M_x, \mathbf{u}, M_y, \mathbf{v}}(x, y) := B_{j, M_x, \mathbf{u}}(x) \cdot B_{l, M_y, \mathbf{v}}(y)$$

for $(j, l) \in J$. Then $\{B_{\mathbf{j}} \mid \mathbf{j} \in J\}$ is a basis for $S_{\mathbf{u}, M_x, \mathbf{v}, M_y}$.

Next we define hierarchical spline spaces as linear span of tensor product B-splines with nested knot sequences. Choose $M_x, M_y \in \mathbb{N}_0$, $u_0, v_0 \in \mathbb{R}$ and $h_x^0, h_y^0 > 0$. Set

$$h_x^p := \frac{h_x^0}{\max\{2, M_x\}^p}, \qquad h_y^p := \frac{h_y^0}{\max\{2, M_y\}^p},$$
$$u_i^p := u_0 + i \cdot h_x^p, \qquad v_i^p := v_0 + i \cdot h_y^p$$

for $p \in \mathbb{N}_0$, $i \in \mathbb{Z}$. The B-spline $B_{\mathbf{j}}^p$ of *level* p is the tensor product B-spline to the knot sequences $\{u_i^p\}_i$ and $\{v_l^p\}_l$ with support $[u_{j_1}^p, u_{j_1+M_x+1}^p] \times [v_{j_2}^p, v_{j_2+M_x+1}^p]$. Let

$$\mathbb{R}^2 \supseteq D_1 \supseteq D_2 \supseteq \cdots \supseteq D_p \supseteq D_{p+1} := \varnothing$$

be a finite sequence of nested domains such that

$$D_l = \bigcup_{\mathbf{j} \in J_l} \mathit{supp}\, B_{\mathbf{j}}^l \qquad \text{for some finite set} \quad J_l \subseteq \mathbb{Z}^2 \tag{24}$$

for $1 \leqslant l \leqslant p$. Then the hierarchical spline space $S_{\mathbf{D}}$ is defined as linear span of all B-splines of level l with support contained in D_l $(1 \leqslant l \leqslant p)$, i.e., with

$$I_l := \{\mathbf{j} \in \mathbb{Z}^2 \mid supp\ B^l_{\mathbf{j}} \subseteq D_l\}$$

we define the hierarchical spline space by

$$S_{\mathbf{D}} := span\{B^l_{\mathbf{j}} \mid 1 \leqslant l \leqslant p \text{ and } \mathbf{j} \in I_l\}.$$

Remark 2. If one sets

$$J_l := \{\mathbf{j} \in \mathbb{Z}^2 \mid supp\ B^l_{\mathbf{j}} \subseteq D_l \text{ and } supp\ B^l_{\mathbf{j}} \nsubseteq D_{l+1}\}$$

for $1 \leqslant l \leqslant p$, then $\{B^l_{\mathbf{j}} \mid 1 \leqslant l \leqslant p \text{ and } \mathbf{j} \in J_l\}$ is a basis for $S_{\mathbf{D}}$ (see Kraft (1997)).

Next we investigate how well smooth functions can be approximated by functions of $S_{\mathbf{D}}$. Let $C^{(\alpha, \beta)}$ be the set of all functions $f: \mathbb{R}^2 \to \mathbb{R}$ which are α (β) times continuously differentiable in x (y). It is well known that the error of approximating a function $f \in C^{(\alpha, \beta)}$ by linear combinations of B-splines of level l is of order $(h^l_x)^\alpha + (h^l_y)^\beta$ (see Schumaker (1981), Theorem 12.7). It follows that for each $1 \leqslant l \leqslant p$ and each $f \in C^{(\alpha, \beta)}$ there exists a $g_l \in S_{\mathbf{D}}$ for which the error of approximating f in

$$D^0_l := \{\mathbf{x} \in \mathbb{R}^2 \mid \forall \mathbf{j} \in \mathbb{Z}^2 \colon B^l_{\mathbf{j}}(\mathbf{x}) \neq 0 \Rightarrow supp\ B^l_{\mathbf{j}} \subseteq D_l\} = D_l \Big\backslash \bigcup_{\mathbf{j}:\ supp\ B^l_{\mathbf{j}} \nsubseteq D_l} supp\ B^l_{\mathbf{j}}$$

is of order $(h^l_x)^\alpha + (h^l_y)^\beta$. It is shown in Kraft (1994) that there even exists a $g \in S_{\mathbf{D}}$ which has this approximation property *simultaneously* for all $1 \leqslant l \leqslant p$. This is the content of the next lemma, where we use the notation $f^{(\alpha, \beta)}$ for the (α, β) partial derivative of a function $f: \mathbb{R}^2 \to \mathbb{R}$.

LEMMA 1. *Let* $1 \leqslant \alpha \leqslant M_x + 1$, $1 \leqslant \beta \leqslant M_y + 1$ *and* $f \in C^{(\alpha, \beta)}$. *Set*

$$p_0 = \min\{l \in \mathbb{N} \mid h^l_x \leqslant 1 \text{ and } h^l_y \leqslant 1\}.$$

Then there exist $Qf \in S_{\mathbf{D}}$ *and constants* $c_1, c_2 \in \mathbb{R}$ *independent of* f, h^0_x, h^0_y *such that*

$$|f(\mathbf{x}) - (Qf)(\mathbf{x})| \leqslant c_1 \cdot (\|f^{(\alpha, 0)}\|_\infty \cdot (h^l_x)^\alpha + \|f^{(0, \beta)}\|_\infty \cdot (h^l_y)^\beta) \qquad (25)$$

for all $\mathbf{x} \in \bigcap_{i=p_0}^l D^0_i$, $p_0 \leqslant l \leqslant p$ *and*

$$\|Qf\|_\infty \leqslant c_2 \cdot (\|f\|_\infty + \|f^{(1, 0)}\|_\infty + \|f^{(0, 1)}\|_\infty). \qquad (26)$$

This lemma is shown in Kraft (1994) under some additional conditions on the sequences of domains. We show in the appendix that it is also valid under the conditions given here.

Remark 3. The conditions $p_0 \leqslant l \leqslant p$ and $h_x^{p_0} \leqslant 1$, $h_y^{p_0} \leqslant 1$ are used to show that c_2 in (26) is independent of h_x^0 and h_y^0.

Next we give general conditions on data-dependent sequences of nested domains which imply the strong universal consistency of the resulting estimator. Therefore we use data-independent parameters $L_n, R_n \in \mathbb{R}$, $\beta_n \in \mathbb{R}_+$, $K_n, C_n, p_n \in \mathbb{N}$ $(n \in \mathbb{N})$ and the abbreviations

$$[u_{\mathbf{i}}^l, u_{\mathbf{i}+1}^l) := [u_{i_1}^l, u_{i_1+1}^l) \times [v_{i_2}^l, v_{i_2+1}^l) \qquad \text{and}$$
$$[u_{\mathbf{i}-1}^l, u_{\mathbf{i}+2}^l) := [u_{i_1-1}^l, u_{i_1+2}^l) \times [v_{i_2-1}^l, v_{i_2+2}^l)$$

for $\mathbf{i} = (i_1, i_2) \in \mathbb{Z}^2$, $l \in \mathbb{N}_0$. According to the next theorem the sequence of nested domains should satisfy the following conditions:

First, all domains should be contained in a data-independent rectangle $[L_n, R_n]^2$ (see (28)), where this rectangle converges to $\mathbb{R}^2$ $(n \to \infty)$ (see (18)). Second, the vector space dimension of the resulting hierarchical spline space (see Remark 2) should be not greater than a data-independent number K_n (see (30)), which converges not too fast to infinity (see (14)). Third, each rectangle $[u_{\mathbf{i}}^l, u_{\mathbf{i}+1}^l)$ which satisfies $[u_{\mathbf{i}-1}^l, u_{\mathbf{i}+2}^l) \subseteq D_l$ and contains more than a data-independent number C_n of the $\mathbf{X}_1, ..., \mathbf{X}_n$ should be contained in D_{l+1}^0 (see (31)), where C_n converges not too fast to infinity (see (22)). This is required for all levels $0 \leqslant l < p_n$, where p_n is data-independent and converges to infinity (see (27)).

THEOREM 2. *Let* $M_x, M_y \in \mathbb{N}_0$. *For* $n \in \mathbb{N}$ *let* $L_n, R_n \in \mathbb{R}$, $\beta_n \in \mathbb{R}_+$ *and* $K_n, C_n, p_n \in \mathbb{N}$ *such that* (13)–(15), (18), (22) *and, in addition,*

$$p_n \to \infty \qquad (n \to \infty) \tag{27}$$

hold. Depending on the sample choose $p = p((\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)) \in \mathbb{N}$ *and a sequence* $\mathbf{D} = (D_l)_{l=1, ..., p}$ *of nested domains of the data-dependent hierarchical spline space such that*

$$[L_n, R_n]^2 =: D_0 \supseteq D_1 \supseteq D_2 \supseteq \cdots \supseteq D_p \supseteq D_{p+1} := D_{p+2} :\cdots := \emptyset, \tag{28}$$

$$D_l = \bigcup_{\mathbf{j} \in J_l} supp\ B_{\mathbf{j}}^l \tag{29}$$

for some data-dependent $J_l \subseteq \mathbb{Z}^2$ $(1 \leqslant l \leqslant p)$,

$$\sum_{l=1}^{p} |\{\mathbf{j} \in \mathbb{Z}^2 : supp\ B_{\mathbf{j}}^l \subseteq D_l \text{ and } supp\ B_{\mathbf{j}}^l \nsubseteq D_{l+1}\}| \leqslant K_n \tag{30}$$

and

$$\bigcup_{\substack{\mathbf{j}:\, [u^l_{\mathbf{j}} - 1,\, u^l_{\mathbf{j}} + 2] \subseteq D_l, \\ \mu_n([u^l_{\mathbf{j}},\, u^l_{\mathbf{j}+1}]) \geqslant C_n/n}} [u^l_{\mathbf{j}-1}, u^l_{\mathbf{j}+2}] \subseteq D_{l+1} \tag{31}$$

for all $0 \leqslant l < p_n$.

Then each sequence $(m_n)_{n \in \mathbb{N}}$ *of estimators which satisfy* (3), (4), *and* (5) *with* $\mathscr{F}_n = S_{\mathbf{D}}$ *is strongly universally consistent.*

Example 7 (28), (29), and (31) are satisfied if one defines D_{l+1} by the left-hand side of (31) and sets $p := p_n$. (30) is then implied by

$$\sum_{l=1}^{p_n} \max\{2, M_x\}^l \max\{2, M_y\}^l \frac{(R_n - L_n)^2}{h^0_x h^0_y} \leqslant K_n. \tag{32}$$

4. PRELIMINARIES TO THE PROOFS OF THEOREM 1 AND THEOREM 2

In this section we will use the following notations: For $L > 0$ and $z \in \mathbb{R}$ set $T_L z := T_L(z) := (z \vee (-L)) \wedge L$. For $f = \mathbb{R}^d \to \mathbb{R}$ define $T_L f \colon \mathbb{R}^d \to \mathbb{R}$ by $(T_L f)(\mathbf{x}) := T_L(f(\mathbf{x}))$ $(\mathbf{x} \in \mathbb{R}^d)$. Let

$$\mathscr{B}_n \mathscr{F}_n := \{ f \in \mathscr{F}_n \mid \forall \mathbf{x} \in \mathbb{R}^d \colon |f(\mathbf{x})| \leqslant \beta_n \}$$

be the set of functions in $\mathscr{F}_n$ which are bounded in absolute value by β_n and let

$$\mathscr{T}_n \mathscr{F}_n := \{ T_{\beta_n} f \mid f \in \mathscr{F}_n \}$$

be the set of truncated functions from $\mathscr{F}_n$.

Our first lemma is a modification of Theorem 1 of Lugosi and Zeger (1995).

Lemma 2. *Assume that*

$$\beta_n \to \infty \qquad (n \to \infty), \tag{33}$$

$$\sup_{f \in \mathscr{T}_n \mathscr{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{X}_i) - Y_i)^2 - E(f(\mathbf{X}) - Y)^2 \right| \to 0 \qquad \text{a.s.} \tag{34}$$

for every distribution of $(\mathbf{X}, Y)$ *with* $|Y| \leqslant L < \infty$ *for some* $L \in \mathbb{R}$ *and*

$$\inf_{f \in \mathscr{B}_n \mathscr{F}_n} \int_{\mathbb{R}^d} |f(\mathbf{x}) - m(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \to 0 \qquad \text{a.s.} \tag{35}$$

for every distribution of $(\mathbf{X}, Y)$ *with* $EY^2 < \infty$. *Then each sequence* $(m_n)_{n \in \mathbb{N}}$ *of estimators which satisfy* (3), (4), *and* (5) *is strongly consistent.*

Remark 4. To ensure measurability of the supremum in (34) it is necessary to impose regularity conditions on uncountable collections $\mathscr{F}_n$ of functions. For the spline spaces in Sections 2 and 3 one can use that every spline function is a pointwise limit of a sequence contained in the countable set of all spline functions with rational knots and rational B-spline coefficients (see (12), Schumaker (1981), Theorem 4.26 and Pollard (1984), p. 38).

Proof of Lemma 2. Let $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots$ be i.i.d. random variables with $EY^2 < \infty$. Set

$$\mathscr{D}_n := \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n\}.$$

Because of

$$\int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \, \mu(d\mathbf{x}) = E(|m_n(\mathbf{X}) - Y|^2 \mid \mathscr{D}_n) - E\,|m(\mathbf{X}) - Y|^2$$

it suffices to show

$$\{E(|m_n(\mathbf{X}) - Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \{E\,|m(\mathbf{X}) - Y|^2\}^{1/2} \to 0 \qquad \text{a.s.} \tag{36}$$

We use the decomposition

$$\begin{aligned} 0 &\leqslant \{E(|m_n(\mathbf{X}) - Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \{E\,|m(\mathbf{X}) - Y|^2\}^{1/2} \\ &= (\{E(|m_n(\mathbf{X}) - Y^2 \mid \mathscr{D}_n)\}^{1/2} - \inf_{f \in \mathscr{B}_n \mathscr{F}_n} \{E\,|f(\mathbf{X}) - Y|^2\}^{1/2}) \\ &\quad + (\inf_{f \in \mathscr{B}_n \mathscr{F}_n} \{E\,|f(\mathbf{X}) - Y|^2\}^{1/2} - \{E\,|m(\mathbf{X}) - Y|^2\}^{1/2}). \end{aligned} \tag{37}$$

It follows from (35) by the aid of the triangle inequality that the second term of (37) converges to zero a.s. Therefore for (36) it suffices to show

$$\limsup_{n \to \infty} (\{E(|m_n(\mathbf{X}) - Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \inf_{f \in \mathscr{B}_n \mathscr{F}_n} \{E\,|f(\mathbf{X}) - Y|^2\}^{1/2}) \leqslant 0 \qquad \text{a.s.} \tag{38}$$

In order to show (38) let $L>0$ be arbitrary. Set $Y_{,L}:=T_L Y$ and $Y_{i,L}:=T_L Y_i$ $(i=1, ..., n)$. Because of (33) we can assume w.l.o.g. that $\beta_n > L$. Then

$$
\begin{aligned}
&\{E(|m_n(\mathbf{X})-Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \inf_{f\in\mathscr{B}_n\mathscr{F}_n} \{E\,|f(\mathbf{X})-Y|^2\}^{1/2} \\
&\quad = \sup_{f\in\mathscr{B}_n\mathscr{F}_n} \{E(|m_n(\mathbf{X})-Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \{E\,|f(\mathbf{X})-Y|^2\}^{1/2} \\
&\quad \leqslant \sup_{f\in\mathscr{B}_n\mathscr{F}_n} \Bigg\{ \{E(|m_n(\mathbf{X})-Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \{E(|m_n(\mathbf{X})-Y_{,L}|^2 \mid \mathscr{D}_n)\}^{1/2} \\
&\qquad + \{E(|m_n(\mathbf{X})-Y_{,L}|^2 \mid \mathscr{D}_n)\}^{1/2} - \left\{\frac{1}{n}\sum_{i=1}^n |m_n(\mathbf{X}_i)-Y_{i,L}|^2\right\}^{1/2} \\
&\qquad + \left\{\frac{1}{n}\sum_{i=1}^n |m_n(\mathbf{X}_i)-Y_{i,L}|^2\right\}^{1/2} - \left\{\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(\mathbf{X}_i)-Y_{i,L}|^2\right\}^{1/2} \\
&\qquad + \left\{\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(\mathbf{X}_i)-Y_{i,L}|^2\right\}^{1/2} - \left\{\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(\mathbf{X}_i)-Y_i|^2\right\}^{1/2} \\
&\qquad + \left\{\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(\mathbf{X}_i)-Y_i|^2\right\}^{1/2} - \left\{\frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i)-Y_i|^2\right\}^{1/2} \\
&\qquad + \left\{\frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i)-Y_i|^2\right\}^{1/2} - \left\{\frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i)-Y_{i,L}|^2\right\}^{1/2} \\
&\qquad + \left\{\frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i)-Y_{i,L}|^2\right\}^{1/2} - \{E\,|f(\mathbf{X})-Y_{,L}|^2\}^{1/2} \\
&\qquad + \{E\,|f(\mathbf{X})-Y_{,L}|^2\}^{1/2} - \{E\,|f(\mathbf{X})-Y|^2\}^{1/2} \Bigg\}.
\end{aligned}
$$

Now we give upper bounds for the terms in each row of the right-hand side of the last inequality: The second and seventh term are bounded above by

$$
\sup_{f\in\mathscr{T}_n\mathscr{F}_n} \left| \left\{\frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i)-Y_{i,L}|^2\right\}^{1/2} - \{E\,|f(\mathbf{X})-Y_{,L}|^2\}^{1/2} \right|
$$

(observe $m_n \in \mathscr{T}_n\mathscr{F}_n$ and $\mathscr{B}_n\mathscr{F}_n \subset \mathscr{T}_n\mathscr{F}_n$). Because of (4) and $f\in\mathscr{B}_n\mathscr{F}_n\subset\mathscr{F}_n$ the fifth term is bounded above by 0. For the third term observe that if $x, y\in\mathbb{R}$ with $|y|\leqslant\beta_n$ and $z:=(x\vee(-\beta_n))\wedge\beta_n$, then $|z-y|\leqslant|x-y|$. Therefore the third term is also not greater than zero.

Using these upper bounds and the triangle inequality for the remaining terms one gets

$$\begin{aligned}&\{E(|m_n(\mathbf{X})-Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \inf_{f \in \mathscr{B}_n \mathscr{F}_n} \{E\,|f(\mathbf{X})-Y|^2\}^{1/2}\\ &\quad \leqslant 2 \cdot \{E\,|Y-Y_{,L}|^2\}^{1/2} + 2 \cdot \left\{\frac{1}{n} \sum_{i=1}^{n} |Y_i - Y_{i,L}|^2\right\}^{1/2}\\ &\qquad + 2 \cdot \sup_{f \in \mathscr{T}_n \mathscr{F}_n} \left| \left\{\frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_{i,L}|^2\right\}^{1/2} - \{E\,|f(\mathbf{X}) - Y_{,L}|^2\}^{1/2} \right|.\end{aligned}$$

Because of (34) and the strong law of large number this implies

$$\begin{aligned}&\limsup_{n\to\infty} (\{E(|m_n(\mathbf{X})-Y|^2 \mid \mathscr{D}_n)\}^{1/2} - \inf_{f \in \mathscr{B}_n \mathscr{F}_n} \{E\,|f(\mathbf{X})-Y|^2\}^{1/2})\\ &\quad \leqslant 4 \cdot \{E\,|Y-Y_{,L}|^2\}^{1/2} \qquad \text{a.s.}\end{aligned}$$

With $L \to \infty$ one gets the assertion. ∎

Because of Lemma 2 the assertion of Theorem 1 and Theorem 2 follows from (34) and (35) with $\mathscr{F}_n = S_{\mathbf{u},M}$ and $\mathscr{F}_n = S_{\mathbf{D}}$, resp. To show (35) we will use the following lemma.

LEMMA 3. *Let $\mathscr{F}_n = \mathscr{F}_n((\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n))$ be a data-dependent set of functions $f\colon \mathbb{R}^d \to \mathbb{R}$ $(n \in \mathbb{N})$. Let $h_{\mathscr{F}_n}\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a data-dependent function and $c\colon C_0^\infty(\mathbb{R}^d) \to \mathbb{R}$ be a data-independent function such that for every $m \in C_0^\infty(\mathbb{R}^d)$ there is a function $g \in \mathscr{F}_n$ with*

$$|g(\mathbf{x})| \leqslant c(m) \qquad \text{and} \qquad |m(\mathbf{x}) - g(\mathbf{x})| \leqslant c(m) \cdot h_{\mathscr{F}_n}(\mathbf{x}) \qquad (\mathbf{x} \in \mathbb{R}^d). \tag{39}$$

Then (33) and

$$\mu(\{\mathbf{x} \in \mathbb{R}^d \mid h_{\mathscr{F}_n}(\mathbf{x}) > \gamma\} \cap [-L, L]^d) \to 0 \qquad \text{a.s.} \tag{40}$$

for every $L, \gamma > 0$ imply

$$\inf_{f \in \mathscr{B}_n \mathscr{F}_n} \int_{\mathbb{R}^d} |f(\mathbf{x}) - m(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \to 0 \qquad \text{a.s.}$$

for every $m \in L_2(\mu)$

Before we will prove this lemma we give an example for possible functions $h_{\mathscr{F}_n}$ and c.

EXAMPLE 8. Let $\mathscr{F}_n$ be the set of all functions $f\colon \mathbb{R}^2 \to \mathbb{R}$ which are constant on each set of a data-dependent partition $\mathscr{P}_n = \{A_{n,1}, A_{n,2}, ...\}$ of $\mathbb{R}^2$. Then it is easy to see that (39) holds with

$$h_{\mathscr{F}_n}(\mathbf{x}) = diam(A_{n,i}) := \sup_{\mathbf{u}, \mathbf{v} \in A_{n,i}} |\mathbf{u} - \mathbf{v}| \qquad (\mathbf{x} \in A_{n,i})$$

and

$$c(m) := \max\{\sqrt{2} \cdot \|m^{(1,0)}\|_\infty, \sqrt{2} \cdot \|m^{(0,1)}\|_\infty, \|m\|_\infty\}.$$

In this case one can choose fixed $\mathbf{x}_i \in A_{n,i}$ $(i \in \mathbb{N})$ and can define

$$g(\cdot) = \sum_i I_{A_{n,i}}(\cdot)\, m(\mathbf{x}_i).$$

Proof of Lemma 3. Because of $m \in L_2(\mu)$ and $C_0^\infty(\mathbb{R}^d)$ dense in $L_2(\mu)$ we can assume w.l.o.g. that $m \in C_0^\infty(\mathbb{R}^d)$.

Choose $g \in \mathscr{F}_n$ with $|g(\mathbf{x})| \leqslant c(m)$ and $|m(\mathbf{x}) - g(\mathbf{x})| \leqslant c(m) \cdot h_{\mathscr{F}_n}(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^d$. Because of (33) and $c(m) < \infty$ we can assume w.l.o.g. $g \in \mathscr{B}_n \mathscr{F}_n$.

Let $L, \gamma > 0$. Then

$$\begin{aligned}
\inf_{f \in \mathscr{B}_n \mathscr{F}_n} \int_{\mathbb{R}^d} |f(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) &\leqslant \int_{\mathbb{R}^d} |g(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) \\
&\leqslant \int_{\mathbb{R}^d \setminus [-L, L]^d} |g(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) \\
&\quad + \int_{\{\mathbf{x} \in \mathbb{R}^d \mid h_{\mathscr{F}_n}(\mathbf{x}) > \gamma\} \cap [-L, L]^d} |g(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) \\
&\quad + \int_{\{\mathbf{x} \in \mathbb{R}^d \mid h_{\mathscr{F}_n}(\mathbf{x}) \leqslant \gamma\} \cap [-L, L]^d} |g(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) \\
&\leqslant 2 \cdot (c(m)^2 + \|m\|_\infty^2) \\
&\quad \cdot \{\mu(\mathbb{R}^d \setminus [-L, L]^d) + \mu(\{\mathbf{x} \in \mathbb{R}^d \mid h_{\mathscr{F}_n}(\mathbf{x}) > \gamma\} \cap [-L, L]^d)\} \\
&\quad + \gamma^2 \cdot c(m)^2 \\
&\to 2 \cdot (c(m)^2 + \|m\|_\infty^2) \cdot \mu(\mathbb{R}^d \setminus [-L, L]^d) + \gamma^2 \cdot c(m)^2 \qquad \text{a.s.}
\end{aligned}$$

because of (40). With $L \to \infty$ and $\gamma \to 0$ one gets the assertion. ∎

Next we explain how we will show (34). In Theorem 1 and in Theorem 2 $\mathcal{F}_n$, and therefore also $\mathcal{T}_n\mathcal{F}_n$, are spaces of piecewise defined functions which depend on the data. We will construct spaces of piecewise defined functions which do not depend on the data and which contain $\mathcal{T}_n\mathcal{F}_n$ for every sample, and we will show (34) for these spaces.

To describe such sets of piecewise defined functions we use the following notation: Let $\Pi = (\pi_j)_j$ be a family of finite or countably infinite partitions $\pi_j = \{A_{j,1}, A_{j,2}, ...\}$ of Borel-measurable subsets of $\mathbb{R}^d$ and let $\mathcal{G}$ be a fixed set of functions $g: \mathbb{R}^d \to \mathbb{R}$. Then

$$\mathcal{G} \circ \Pi := \left\{ f = \sum_{A_j \in \pi} g_j I_{A_j} \,\middle|\, \pi = \{A_j\} \in \Pi,\ g_j \in \mathcal{G} \right\}$$

is the set of all functions which are obtained by applying a different function of $\mathcal{G}$ within each cell of a selected partition of Π.

In order to show (34) for such spaces of functions, we will use the following lemma, whose formulation needs the notions of VC dimension and partitioning numbers (Definitions 2 and 3 in the appendix).

Lemma 4. *Let $N \in \mathbb{N}$ and let Π be a family of partitions of $\mathbb{R}^d$ such that no partition of Π consists of more than N sets. Let $\mathcal{G}$ be a set of functions $g: \mathbb{R}^d \to \mathbb{R}$. Assume $|Y| \leqslant \beta_n$ a.s. and let $t > 0$ be arbitrary. Then the following inequality holds*:

$$P\left(\sup_{f \in \mathcal{T}_n \mathcal{G} \circ \Pi} \left| \frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2 - E\,|f(\mathbf{X}) - Y|^2 \right| > t \right)$$
$$\leqslant 8 \cdot \Delta_n(\Pi) \cdot 2^N \cdot \left(\frac{128 e \beta_n^2}{t} \log\left(\frac{128 e \beta_n^2}{t} \right) \right)^{V_{\mathcal{G}^+} \cdot N} \cdot \exp\left(-\frac{n t^2}{2048 \beta_n^4} \right).$$

We will apply the Borel–Cantelli lemma in order to obtain (34) from the above inequality.

Proof of Lemma 4. Set

$$\mathcal{H}_n := \{ h: \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}: h(\mathbf{x}, y) = |f(\mathbf{x}) - T_{\beta_n} y|^2\ ((\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R})$$
$$\text{for some } f \in \mathcal{T}_n \mathcal{G} \circ \Pi \}.$$

For $h \in \mathcal{H}_n$ one has $0 \leqslant h(\mathbf{x}, y) \leqslant 4\beta_n^2$ $((\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R})$. Using the notion of covering numbers (Definition 1 in the appendix) and Lemma 5 one concludes

$$P\left[\sup_{f\in\mathcal{T}_n\mathcal{G}\circ\Pi}\left|\frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{X}_i)-Y_i|^2-E\,|f(\mathbf{X})-Y|^2\right|>t\right]$$

$$=P\left[\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}h(\mathbf{X}_i,Y_i)-Eh(\mathbf{X},Y)\right|>t\right]$$

$$\leqslant 8E\left(\mathcal{N}\left(\frac{t}{8},\mathcal{H}_n,(\mathbf{X},Y)_1^n\right)\right)\exp\left(-\frac{nt^2}{2048\beta_n^4}\right). \tag{41}$$

Next we bound the covering number in (41). Observe first, that if $h_j(\mathbf{x},y)=|f_j(\mathbf{x})-T_{\beta_n}y|^2$ $((\mathbf{x},y)\in\mathbb{R}^d\times\mathbb{R})$ for some functions f_j bounded in absolute value by β_n $(j=1,2)$, then

$$\tfrac{1}{n}\sum_{i=1}^{n}|h_1(\mathbf{X}_i,Y_i)-h_2(\mathbf{X}_i,Y_i)|$$

$$=\frac{1}{n}\sum_{i=1}^{n}|f_1(\mathbf{X}_i)-T_{\beta_n}Y_i+f_2(\mathbf{X}_i)-T_{\beta_n}Y_i|\cdot|f_1(\mathbf{X}_i)-f_2(\mathbf{X}_i)|$$

$$\leqslant 4\beta_n\frac{1}{n}\sum_{i=1}^{n}|f_1(\mathbf{X}_i)-f_2(\mathbf{X}_i)|.$$

Thus

$$\mathcal{N}\left(\frac{t}{8},\mathcal{H}_n,(\mathbf{X},Y)_1^n\right)\leqslant\mathcal{N}\left(\frac{t}{32\beta_n},\mathcal{T}_n\mathcal{G}\circ\Pi,\mathbf{X}_1^n\right). \tag{42}$$

Using the notions of VC dimension and partitioning numbers (Definitions 2 and 3 in the appendix), Lemma 8 and Lemma 6, one gets

$$\mathcal{N}\left(\frac{t}{32\beta_n},\mathcal{T}_n\mathcal{G}\circ\Pi,\mathbf{X}_1^n\right)\leqslant\Delta_n(\Pi)\left\{\sup_{\mathbf{z}_1,\dots,\mathbf{z}_m\in\mathbf{X}_1^n}\mathcal{N}\left(\frac{t}{32\beta_n},\mathcal{T}_n\mathcal{G},\mathbf{z}_1^m\right)\right\}^N$$

$$\leqslant\Delta_n(\Pi)\left(2\left(\frac{4e\beta_n}{t/32\beta_n}\log\left(\frac{4e\beta_n}{t/32\beta_n}\right)\right)^{V_{\mathcal{T}_n\mathcal{G}^+}}\right)^N$$

$$\leqslant\Delta_n(\Pi)\,2^N\left(\frac{128e\beta_n^2}{t}\log\left(\frac{128e\beta_n^2}{t}\right)\right)^{V_{\mathcal{G}^+}\cdot N} \tag{43}$$

where we have used the relation $V_{\mathcal{T}_n\mathcal{G}^+}\leqslant V_{\mathcal{G}^+}$, which follows directly from the definition of the VC dimension. (41)–(43) imply the assertion. ∎

5. PROOF OF THEOREM 1

Because of Lemma 2 it suffices to show (34) and (35).

Proof of (34). Let Π_n be the family of all partitions of $\mathbb{R}$ consisting of $K_n + 2M + 2$ or less intervals and let $\mathcal{G}$ be the set of all polynomials of degree not greater than M. Then $\mathcal{F}_n \subset \mathcal{G} \circ \Pi_n$ and therefore it suffices to show (34) with $\mathcal{F}_n$ replaced by $\mathcal{G} \circ \Pi_n$. $\mathcal{G}$ is a linear space of functions of dimension $M+1$, thus $V_{\mathcal{G}^+} \leqslant M+2$ (see Lemma 7). It follows from Example 1 in Nobel (1996) that

$$\Delta_n(\Pi_n) \leqslant \binom{n + K_n + 2M + 1}{n} \leqslant (n + K_n + 2M + 1)^{K_n + 2M + 1}.$$

Now the assertion follows from this and Lemma 4 by an easy application of the Borel–Cantelli lemma with the help of conditions (14) and (15).

Proof of (35). Because of Lemma 3 it suffices to show (39) and (40). For $m \in C_0^\infty(\mathbb{R})$ define $Qm \in \mathcal{F}_n$ by $Qm := \sum_{j=-M}^{K_n - 1} m(u_j) \cdot B_{j, M, \mathbf{u}}$. Then (10) and (12) imply

$$|Qm(x)| \leqslant \max_{j=-M, \ldots, K_n - 1} |m(u_j)| \cdot \sum_{j=-1}^{K_n - 1} B_{j, M, \mathbf{u}}(x) \leqslant \|m\|_\infty$$

for $x \in \mathbb{R}$. Let $x \in [u_i, u_{i+1})$ for some $0 \leqslant i \leqslant K_n - 1$. Using (12), (11), and (10) one gets

$$\begin{aligned}
|m(x) - Qm(x)| &= \left| \sum_{j=-M}^{K_n - 1} (m(x) - m(u_j)) \cdot B_{j, M, \mathbf{u}}(x) \right| \\
&= \left| \sum_{j=i-M}^{i} (m(x) - m(u_j)) \cdot B_{j, M, \mathbf{u}}(x) \right| \\
&\leqslant \max_{j=i-M, \ldots, i} |m(x) - m(u_j)| \cdot \sum_{j=i-M}^{i} B_{j, M, \mathbf{u}}(x) \\
&\leqslant |u_{i+1} - u_{i-M}| \cdot \|m'\|_\infty.
\end{aligned}$$

Therefore (39) is satisfied with

$$h_{\mathcal{F}_n}(x) := \begin{cases} |u_{i+1} - u_{i-M}| & \text{if } \ x \in [u_i, u_{i+1}) \ \text{ and } \ 0 \leqslant i \leqslant K_n - 1 \\ \infty & \text{if } \ x < u_0 \ \text{ or } \ x \geqslant u_{K_n}, \end{cases}$$

and $c(m) := \max\{\|m\|_\infty, \|m'\|_\infty\}$.

Let $L, \gamma > 0$. Then

$$\begin{aligned}
&\mu(\{x \in \mathbb{R} \mid h_{\mathscr{F}_n}(x) > \gamma\} \cap [-L, L]) \\
&\quad = \mu\Bigg(\Bigg\{(-\infty, u_0) \cup \bigcup_{\substack{j=0, ..., K_n - 1, \\ u_{j+1} - u_{j-M} > \gamma}} [u_j, u_{j+1}) \cup [u_{K_n}, \infty)\Bigg\} \cap [-L, L]\Bigg) \\
&\quad \leqslant \mu_n\Bigg(\Bigg\{(-\infty, u_0) \cup \bigcup_{\substack{j=0, ..., K_n - 1, \\ u_{j+1} - u_{j-M} > \gamma}} [u_j, u_{j+1}) \cup [u_{K_n}, \infty)\Bigg\} \cap [-L, L]\Bigg) \\
&\qquad + \sup_{f \in \mathscr{G}_0 \circ \Pi_n} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - Ef(X) \right|,
\end{aligned}$$

with $\mathscr{G}_0$ consisting of two functions which are constant 0 and constant 1 (resp.).

It follows from (34) that

$$\begin{aligned}
&\sup_{f \in \mathscr{G}_0 \circ \Pi_n} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - Ef(X) \right| \\
&\quad = \sup_{f \in \mathscr{G}_0 \circ \Pi_n} \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - 0)^2 - E(f(X) - 0)^2 \right| \to 0 \qquad \text{a.s.} \quad (n \to \infty).
\end{aligned}$$

Thus (40) holds provided that (16) or (17) hold. ▌

6. PROOF OF THEOREM 2

Because of Lemma 2 it suffices to show (34) and (35).

Proof of (34). Let $\mathscr{G}$ be the set of all polynomials of degree less than or equal to M_x (M_y) in x (y), resp. Each function in $S_{\mathbf{D}}$ is on each set $[u^l_{\mathbf{i}}, u^l_{\mathbf{i}+1}) \subseteq D_l \backslash D_{l+1}$ equal to such a polynomial ($l \in \{1, ..., p\}$). From this, (29) and (30) one obtains that there exists a partition $\pi_{\mathbf{D}}$ consisting of at most $K_n(M_x + 1)(M_y + 1)$ rectangles and one additional set such that each function in $S_{\mathbf{D}}$ is on each cell of $\pi_{\mathbf{D}}$ equal to some function contained in $\mathscr{G}$. Let Π_n be the family of all such partitions. Then $S_{\mathbf{D}} \subseteq \mathscr{G} \circ \Pi_n$ and therefore it suffices to show (34) with $\mathscr{F}_n$ replaced by $\mathscr{G} \circ \Pi_n$.

It is easy to see that in $\mathbb{R}^2$ rectangles can partition n points in at most $(n+1)^4$ ways, hence

$$\Delta_n(\Pi_n) \leqslant (n+1)^{4K_n(M_x+1)(M_y+1)}.$$

From this one obtains the assertion as in the proof of Theorem 1.

Proof of (35). Because of Lemma 3 it suffices to show (39) and (40). It follows from Lemma 1 that (39) is satisfied with

$$h_{\mathscr{F}_n}(x) := \begin{cases} \dfrac{h_x^0}{\max\{2, M_x\}^l} + \dfrac{h_y^0}{\max\{2, M_y\}^l} & \text{if } \; x \in \bigcap_{i=p_0}^{l} D_i^0 \Big\backslash \bigcap_{i=p_0}^{l-1} D_i^0 \\ & \quad \text{for some } \; p_0 \leqslant l \leqslant p, \\ \infty & \text{otherwise} \end{cases}$$

and $c(m) := \max\{c_1, c_2\} \cdot (\|m\|_\infty + \|m^{(1,0)}\|_\infty + \|m^{(0,1)}\|_\infty)$. As in the proof of Theorem 1 (40) follows from (34) and

$$\mu_n(\{\mathbf{x} \in \mathbb{R}^2 \mid h_{S_\mathbf{D}}(\mathbf{x}) > \gamma\} \cap [-L, L]^2) \to 0 \qquad \text{a.s.} \tag{44}$$

for every $L, \gamma > 0$.

Let $L, \gamma > 0$ be arbitrary. Choose $q \in \mathbb{N}$ minimal such that $h_x^0/\max\{2, M_x\}^q + h_y^0/\max\{2, M_y\}^q \leqslant \gamma$. Because of (18) and (27) we can assume w.l.o.g. that

$$[-L, L]^2 \subseteq [L_n + 2h_x^0, R_n - 2h_x^0] \times [L_n + 2h_y^0, R_n - 2h_y^0]$$

(thus $[u_\mathbf{j}^0, u_{\mathbf{j}+1}^0] \cap [-L, L]^2 \neq \varnothing$ implies $[u_{\mathbf{j}-1}^0, u_{\mathbf{j}+2}^0] \subseteq [L_n, R_n]^2$) and $q \leqslant p_n$.

If $[u_\mathbf{j}^q, u_{\mathbf{j}+1}^q] \cap [-L, L]^2 \neq \varnothing$ and $[u_\mathbf{j}^q, u_{\mathbf{j}+1}^q)$ contains more than C_n of the $\mathbf{X}_1, ..., \mathbf{X}_n$, then (31) implies $[u_\mathbf{j}^q, u_{\mathbf{j}+1}^q) \subseteq D_l^0$ for $1 \leqslant l \leqslant q$ and therefore

$$\{\mathbf{x} \in \mathbb{R}^2 \mid h_{S_\mathbf{D}}(\mathbf{x}) > \gamma\} \cap [u_\mathbf{j}^q, u_{\mathbf{j}+1}^q) = \varnothing.$$

Using this and (22) one gets

$$\begin{aligned} &\mu_n(\{\mathbf{x} \in \mathbb{R}^2 \mid h_{S_\mathbf{D}}(\mathbf{x}) > \gamma\} \cap [-L, L]^2) \\ &\quad \leqslant \sum_{[u_\mathbf{j}^q, u_{\mathbf{j}+1}^q) \cap [-L, L]^2 \neq \varnothing, \, \mu_n([u_\mathbf{j}^q, u_{\mathbf{j}+1}^q)) \leqslant C_n/n} \mu_n([u_\mathbf{j}^q, u_{\mathbf{j}+1}^q)) \\ &\quad \leqslant \frac{C_n}{n} \cdot \left(\max\{2, M_x\}^q \cdot \left\lceil \frac{2L}{h_x^0} \right\rceil + 2 \right) \\ &\qquad \cdot \left(\max\{2, M_y\}^q \cdot \left\lceil \frac{2L}{h_y^0} \right\rceil + 2 \right) \to 0 \qquad (n \to \infty). \quad \blacksquare \end{aligned}$$

APPENDIX

A. Proof of Lemma 1

For $l \in \mathbb{N}_0$ and $f \in C(\mathbb{R}^2)$ define

$$Q^l f := \sum_{\mathbf{j} \in \mathbb{Z}^2} Q^l_{\mathbf{j}} f \cdot B^l_{\mathbf{j}}$$

for some linear functionals $Q^l_{\mathbf{j}} \colon C(\mathbb{R}^2) \to \mathbb{R}$. It is possible to choose $Q^l_{\mathbf{j}}$ such that

$$|Q^l_{\mathbf{j}} f| \leqslant c_3 \cdot \|f\|_\infty \tag{45}$$

for all $f \in C(\mathbb{R}^2)$, $\mathbf{j} \in \mathbb{Z}^2$,

$$\|Q^l f - f\|_\infty \leqslant c_4 \cdot (\|f^{(\alpha, 0)}\|_\infty \cdot (h^l_x)^\alpha + \|f^{(0, \beta)}\|_\infty \cdot (h^l_y)^\beta) \tag{46}$$

for all $\alpha \leqslant M_x + 1$, $\beta \leqslant M_y + 1$, $f \in C^{(\alpha, \beta)}$, and

$$Q^{l+1}(Q^l f) = Q^l f \tag{47}$$

for all $l \in \mathbb{N}_0$, $f \in C(\mathbb{R}^2)$. Here c_3, c_4 are constants independent of f, l, h^0_x, h^0_y (see Schumaker (1981), Theorems 12.5–12.7).

Define

$$\tilde{Q}^l f := \sum_{\mathbf{j} \in \mathbb{Z}^2,\, supp\, B^l_{\mathbf{j}} \subseteq D_l} (Q^l_{\mathbf{j}} f) \cdot B^l_{\mathbf{j}}$$

and

$$Q^l_{S_{\mathbf{D}}} f := \tilde{Q}^{p_0} f + \sum_{i=p_0}^{l-1} \tilde{Q}^{i+1}(f - Q^i f)$$

for $p_0 \leqslant l \leqslant p$.

It follows directly from the definition of D^0_l that

$$\tilde{Q}^l f_{|D^0_l} = Q^l f_{|D^0_0}$$

for all $l \in \mathbb{N}_0$. Using this and (47) one gets

$$\begin{aligned} Q^l_{S_{\mathbf{D}}} f_{|\bigcap_{i=p_0}^l D^0_i} &= \tilde{Q}^{p_0} f_{|\bigcap_{i=p_0}^l D^0_i} + \sum_{i=p_0}^{l-1} \tilde{Q}^{i+1}(f - Q^i f)_{|\bigcap_{i=p_0}^l D^0_i} \\ &= Q^{p_0} f_{|\bigcap_{i=p_0}^l D^0_i} + \sum_{i=p_0}^{l-1} (Q^{i+1} f_{|\bigcap_{i=p_0}^l D^0_i} - Q^{i+1}(Q^i f)_{|\bigcap_{i=p_0}^l D^0_i}\bigg) \\ &= Q^l f_{|\bigcap_{i=p_0}^l D^0_i}. \end{aligned} \tag{48}$$

Now we prove (25). Let $p_0 \leqslant l \leqslant p$ and $\mathbf{x} \in \bigcap_{i=p_0}^{l} D_i^0$. Then (48), (45), (12), and (46) imply

$$\begin{aligned}
|f(\mathbf{x}) - Q_{S_\mathbf{D}}^{p} f(\mathbf{x})| &\leqslant |f(\mathbf{x}) - Q_{S_\mathbf{D}}^{l} f(\mathbf{x})| + \left| \sum_{i=l}^{p-1} \tilde{Q}^{i+1}(f - Q^i f)(\mathbf{x}) \right| \\
&\leqslant |f(\mathbf{x}) - Q^l f(\mathbf{x})| + c_3 \cdot \sum_{i=l}^{p-1} \|f - Q^i f\|_\infty \\
&\leqslant c_4 \cdot \{ \|f^{(\alpha, 0)}\|_\infty \cdot (h_x^l)^\alpha + \|f^{(0, \beta)}\|_\infty \cdot (h_y^l)^\beta \} \\
&\quad + c_3 \cdot \sum_{i=l}^{p-1} c_4 \cdot \left\{ \|f^{(\alpha, 0)}\|_\infty \cdot \left(\frac{h_x^l}{\max\{2, M_x\}^{i-l}} \right)^\alpha \right. \\
&\quad \left. + \|f^{(0, \beta)}\|_\infty \cdot \left(\frac{h_y^l}{\max\{2, M_y\}^{i-l}} \right)^\beta \right\} \\
&\leqslant (c_4 + 2 \cdot c_3 \cdot c_4) \cdot \{ \|f^{(\alpha, 0)}\|_\infty \cdot (h_x^l)^\alpha + \|f^{(0, \beta)}\|_\infty \cdot (h_y^l)^\beta \};
\end{aligned}$$

thus (25) is proved.

Next we prove (26). Let $\mathbf{x} \in \mathbb{R}^2$ be arbitrary. Then (45), (12), (46), and the definition of p_0 imply

$$\begin{aligned}
|(Q_{S_\mathbf{D}}^{p} f)(\mathbf{x})| &\leqslant c_3 \cdot \|f\|_\infty + \sum_{i=p_0}^{p} c_3 \cdot \|f - Q^i f\|_\infty \\
&\leqslant c_3 \|f\|_\infty + c_3 c_4 \sum_{i=p_0}^{p} \left\{ \|f^{(1, 0)}\|_\infty \cdot \frac{h_x^{p_0}}{\max\{2, M_x\}^{i-p_0}} \right. \\
&\quad \left. + \|f^{(0, 1)}\|_\infty \cdot \frac{h_y^{p_0}}{\max\{2, M_y\}^{i-p_0}} \right\} \\
&\leqslant c_3 \cdot \|f\|_\infty + c_3 \cdot c_4 \cdot \{ \|f^{(1, 0)}\|_\infty + \|f^{(0, 1)}\|_\infty \}. \quad \blacksquare
\end{aligned}$$

B. Some Results of the Vapnik–Chervonenkis Theory

In this section we list the definitions and results of the Vapnik–Chervonenkis theory which we have use in Sections 4, 5, and 6. An excellent introduction to most of these results can be found in Devroye *et al.* (1996).

We start with the definition of covering numbers of classes of functions.

Definition 1. Let $\mathscr{F}_n$ be a class of functions $f: \mathbb{R}^d \to \mathbb{R}$. The covering number $\mathscr{N}(\varepsilon, \mathscr{F}_n, \mathbf{z}_1^n)$ is defined for any $\varepsilon > 0$ and $z_1^n = (\mathbf{z}_1, ..., \mathbf{z}_n) \in \mathbb{R}^{d \cdot n}$

as the smallest integer k such that there exist functions $g_1, ..., g_k\colon \mathbb{R}^d \to \mathbb{R}$ with

$$\min_{1 \leqslant i \leqslant k} \frac{1}{n} \sum_{j=1}^{n} |f(\mathbf{z}_j) - g_i(\mathbf{z}_j)| \leqslant \varepsilon$$

for each $f \in \mathscr{F}_n$.

If $\mathbf{Z}_1^n = (\mathbf{Z}_1, ..., \mathbf{Z}_n)$ is a sequence of $\mathbb{R}^d$-valued random variables, then $\mathscr{N}(\varepsilon, \mathscr{F}_n, \mathbf{Z}_1^n)$ is a random variable with expected value $E\mathscr{N}(\varepsilon, \mathscr{F}_n, \mathbf{Z}_1^n)$. The next result due to Pollard is the main tool in the proof of (34).

Lemma 5 (Pollard (1984), Section II.5, Th. 24). *Let $\mathscr{F}_n$ be a class of functions $f\colon \mathbb{R}^d \to [0, B]$, and let $\mathbf{Z}_1^n = (\mathbf{Z}_1, ..., \mathbf{Z}_n)$ be $\mathbb{R}^d$-valued i.i.d. random variables. Then for any $\varepsilon > 0$*

$$P\left[\sup_{f \in \mathscr{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{Z}_i) - Ef(\mathbf{Z}_1) \right| > \varepsilon\right]$$
$$\leqslant 8E\left(\mathscr{N}\left(\frac{\varepsilon}{8}, \mathscr{F}_n, \mathbf{Z}_1^n\right)\right) \exp\left(-\frac{n\varepsilon^2}{128B^2}\right).$$

To bound covering numbers we use the following definition of the VC dimension.

Definition 2. Let $\mathscr{D}$ be a class of subsets of $\mathbb{R}^d$ and let $F \subseteq \mathbb{R}^d$. One says that $\mathscr{D}$ shatters F if each subset of F has the form $D \cap F$ for some D in $\mathscr{D}$. The VC dimension $V_{\mathscr{D}}$ of $\mathscr{D}$ is defined as the largest integer k for which a set of cardinality k exists which is shattered by $\mathscr{D}$.

A connection between covering numbers and VC dimensions is given by the following lemma, which uses the notation $V_{\mathscr{F}_n^+}$ for the VC dimension of the set

$$\mathscr{F}_n^+ := \{\{(\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R} : t \leqslant f(\mathbf{x})\} : f \in \mathscr{F}_n\}$$

of all graphs of functions of $\mathscr{F}_n$.

Lemma 6 (Haussler (1992), Th. 6). *Let $\mathscr{F}_n$ be a class of functions $f\colon \mathbb{R}^d \to [-B, B]$. Then one has for any $\mathbf{z}_1^n \in \mathbb{R}^{d \cdot n}$ and any $\varepsilon > 0$*

$$\mathscr{N}(\varepsilon, \mathscr{F}_n, \mathbf{z}_1^n) \leqslant 2\left(\frac{4eB}{\varepsilon} \log\left(\frac{4eB}{\varepsilon}\right)\right)^{V_{\mathscr{F}_n^+}}.$$

The following result is often useful for bounding the VC dimension.

LEMMA 7 (Dudley (1978)). *Let $\mathscr{F}_n$ be a k-dimensional vector space of functions $f\colon \mathbb{R}^d \to \mathbb{R}$. Then the class of sets of the form $\{\mathbf{x} \in \mathbb{R}^d\colon f(\mathbf{x}) \geqslant 0\}$, $f \in \mathscr{F}_n$, has VC dimension less than or equal to k.*

We apply the above results to sets of piecewise defined functions. Let $\Pi = (\pi_j)_j$ be a family of partitions of $\mathbb{R}^d$ and let $\mathscr{G}$ be a fixed set of functions $g\colon \mathbb{R}^d \to \mathbb{R}$. Set

$$\mathscr{G} \circ \Pi := \left\{ f = \sum_{A_j \in \pi} g_j I_{A_j} \,\middle|\, \pi = \{A_j\} \in \Pi,\ g_j \in \mathscr{G} \right\}.$$

In order to bound covering numbers of such sets of functions we need the following definition, which is due to Lugosi and Nobel (1996).

DEFINITION 3. Let Π be a family of partitions of $\mathbb{R}^d$. Let $\mathbf{x}_1^n = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subseteq \mathbb{R}^d$. Every element $\pi = \{A_j\colon j\} \in \Pi$ induces a partition $\{A_j \cap \mathbf{x}_1^n\colon j\}$ of $\mathbf{x}_1^n$. Let $\Delta(\mathbf{x}_1^n, \Pi)$ be the maximal number of distinct partitions of $\mathbf{x}_1^n$ induced by elements of Π (without regarding the order of appearence of the individual sets) and define the partitioning number $\Delta_n(\Pi)$ by

$$\Delta_n(\Pi) = \max\{\Delta(\mathbf{x}_1^n, \pi)\colon \mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d\}.$$

LEMMA 8 (Nobel (1996)). *Let $N \in \mathbb{N}$ and let Π be a family of partitions of $\mathbb{R}^d$ such that no partition of Π consist of more than N sets. Let $\mathscr{G}$ be a class of functions $g\colon \mathbb{R}^d \to \mathbb{R}$. Then one has for each $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ and each $\varepsilon > 0$*

$$\mathscr{N}(\varepsilon, \mathscr{G} \circ \Pi, \mathbf{x}_1^n) \leqslant \Delta_n(\Pi)\{\sup_{\mathbf{z}_1, ..., \mathbf{z}_m \in \mathbf{x}_1^n,\, m \leqslant n} \mathscr{N}(\varepsilon, \mathscr{G}, \mathbf{z}_1^m)\}^N.$$

ACKNOWLEDGMENTS

The author thanks J. Dippon, L. Györfi, and H. Walk for various stimulating discussions and an anonymous referee and an editor for many detailed and helpful comments.

REFERENCES

1. P. Chaudhuri, M. C. Huang, W.-Y. Loh and R. Yao, Piecewise polynomial regression trees, *Statist. Sinica* **4** (1994), 143–167.
2. C. De Boor, "A Practical Guide to Splines," Springer, New York, 1978.
3. L. Devroye, L. Györfi, A. Krzyżak, and G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates, *Ann. Statist.* **22** (1994), 1371–1385.

4. L. Devroye, L. Györfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," Springer, Berlin, 1996.
5. R. Dudley, Central limit theorems for empirical measures, *Ann. Probab.* **6** (1978), 899–929.
6. D. Forsey and R. Bartels, Hierarchical B-Spline Refinement, *Computer Graphics* **4** (1988), 205–212.
7. L. Györfi, M. Kohler, and H. Walk, Weak and strong universal consistency of semi-recursive kernel and partitioning regression estimates, *Statist. Decisions* **16** (1998), 1–18.
8. L. Györfi and H. Walk, On the strong universal consistency of a series type regression estimate, *Math. Methods Statist.* **5** (1996), 332–342.
9. L. Györfi and H. Walk, On the strong universal consistency of a recursive regression estimate by Pál Révész, *Statist. Probab. Lett.* **31** (1997), 177–183.
10. D. Haussler, Decision theoretic generalizations of the PAC model for neutral net and other learning applications, *Inform. Comput.* **100** (1992), 78–150.
11. M. Kohler, On the universal consistency of a least squares spline regression estimator, *Math. Methods Statist.* **6** (1997), 349–364.
12. R. Kraft, Hierarchical B-splines, Preprint 94-14, Mathematisches Institut A, Universität Stuttgart, (1994).
13. R. Kraft, Adaptive and linearly independent multilevel B-splines, *in* "Surface Fitting and Multiresolution Methods" (A. Le Méhauté, C. Rabut, and L. L. Schumaker, Eds.), pp. 209–216, Vanderbilt Univ. Press, Nashville, TN, 1997.
14. G. Lugosi and K. Zeger, Nonparametric estimation via empirical risk minimization, *IEEE Trans. Inform. Theory* **41** (1995), 677–687.
15. G. Lugosi and A. Nobel, Consistency of data-driven histogram methods for density estimation and classification, *Ann. Statist.* **24** (1996), 687–706.
16. A. Nobel, Histogram regression estimation using data-dependent partitions, *Ann. Statist.* **24** (1996), 1084–1105.
17. D. Pollard, "Convergence of Stochastic Processes," Springer-Verlag, New York, 1984.
18. L. Schumaker, "Spline Functions: Basic Theory," Wiley, New York, 1981.
19. C. J. Stone, Consistent nonparametric regression, *Ann. Statist.* **5** (1977), 595–645.
20. C. J. Stone, The use of polynomial splines and their tensor products in multivariate function estimation, *Ann. Statist.* **22** (1994), 118–184.
21. H. Walk, Strong universal consistency of kernel and partitioning regression estimates, Preprint 97-1, Mathematisches Institut A, Universität Stuttgart, (1997).

Printed in Belgium