ON OPTIMAL GLOBAL RATES OF CONVERGENCE FOR NONPARAMETRIC REGRESSION WITH RANDOM DESIGN *

BY MICHAEL KOHLER

Universität Stuttgart

Estimation of univariate regression functions from bounded i.i.d. data is considered. The L_2–error with integration with respect to the design measure is used as an error criterion.

The aim of this paper is to show that it is possible to achieve for k–smooth regression functions the well–known optimal global rate of convergence $n^{-\frac{2k}{2k+1}}$ without assuming anything on the underlying distribution besides boundedness and without knowing the smoothness of the regression function in advance.

The estimates for which this result is shown are smoothing spline estimates. Motivated by Vapnik's principle of structural risk minimization a data-driven method for the choice of the smoothing parameter is defined. Empirical process theory is used to analyze the estimates.

March 16, 2000

1 Introduction

1.1 Scope of this article. Nonparametric regression with random design is a field with many applications, e.g. many questions in Data Mining can be solved by applying its techniques (see, e.g., Chapter 10 in Mattison (1997), Chapter 5 in Krahl, Windheuser and Zwick (1998) or Chapter 1 in John (1997)). Most corresponding theoretical results in the literature require regularity assumptions on the distribution of the design like the existence of a density with respect to the Lebesgue-Borel measure. But in applications typically some components of the design are discrete, which implies that such regularity assumptions are not satisfied.

*Research supported by DFG.

AMS 1991 subject classifications. Primary 62G07; secondary 62G20.

Key words and phrases. Penalized least squares, smoothing splines, regression estimate, rate of convergence, structural risk minimization.

This article contributes to the methodology in nonparametric regression. It is shown that one can get the usual optimal rates of convergence even without the above regularity assumptions on the design. The author hopes that this result will push the research in nonparametric regression with random design towards results which need less assumptions and hence are closer to applications.

1.2 Nonparametric regression with random design. Let (X,Y), (X_1,Y_1), (X_2,Y_2),... be independent identically distributed $\mathbb{R}^d \times \mathbb{R}$ - valued random vectors with $EY^2 < \infty$. The distributions of the components of X may be e.g. discrete, or absolutely continuous or a mixture of a discrete and an absolutely continuous distribution. Hence, we do not assume anything on the distribution of X. In regression analysis one wants to estimate Y after having observed X, i.e., one wishes to find a function f with $f(X)$ "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 f^* with

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

Let $m(x) := E(Y|X=x)$ be the regression function and denote the distribution of X by μ. It is well known that for each measurable function f the relation

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

holds. Therefore m is the solution of the minimization problem (1).

For the regression estimation problem the distribution of (X,Y) (and therefore also m) is unknown. Given only a sample (X_1,Y_1),...,(X_n,Y_n) of the distribution of (X,Y) one has to construct an estimate $m_n(x)$=$m_n(x,(X_1,Y_1),...,(X_n,Y_n))$ of $m(x)$.

1.3 How to measure the error? In the literature the error of a regression estimate is usually measured either *pointwise* by

$$m_n(x) - m(x)$$

for some fixed $x \in \mathbb{R}^d$, or in the *supremum norm* by

$$\sup_{x\in K} |m_n(x) - m(x)|,$$

or in the L_2 *norm with integration with respect to the Lebesgue–Borel measure* by

$$\int_K |m_n(x) - m(x)^2 dx.$$

Here K is some compact subset of $\mathbb{R}^d$ and one assumes $X \in K$ *a.s.*

In this article we use none of the above error criteria. Motivated by (2), which implies that $E(f(X) - Y)^2$ is close to the optimal value $E(m(X) - Y)^2$ if and only if $\int_{\mathbb{R}^d} \mid f(x) - m(x) \mid^2 \mu(dx)$ is small, we measure the error by the so-called L_2 *error*

$$\int_{\mathbb{R}^d} \mid m_n(x) - m(x) \mid^2 \mu(dx). \tag{3}$$

There are two advantages of using (3) to measure the error: Firstly, a function with small L_2 error is always nearly optimal for predicting Y given X (with respect to the mean squared error). Secondly, as we will see in the sequel, it is possible to derive asymptotic results for the L_2 error without strong assumptions on the distribution of the design.

1.4 White-noise-models. Sometimes nonparametric regression is studied in the so-called white–noise–model

$$Y_\epsilon(dt) = f(t)dt + \epsilon W(dt) \quad t \in [0,1],$$

where $\epsilon > 0$ is a small parameter, f is an unknown function which one would like to recover, and $W(t)$ is a standard Wiener process. As it is shown in Brown and Low (1996) this model is equivalent to nonparametric regression with random design in the sense that if one shows suitable error bounds in the white–noise–model for an estimate, then one can explicitely construct estimates which achieve a corresponding bound in the regression model described above (and vice versa). However, this equivalence requires regularity assumptions on the underlying distribution, in particular that the distribution of the design has a density with respect to the Lebesgue–Borel measure (see (4.3) in Brown and Low (1996)). Therefore it is not possible to derive the results which we want to get in this article from results derived in white–noise–models.

1.5 Universal consistency. A sequence of estimates m_n is called *weakly universally consistent* if $E \int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx) \to 0$ $(n \to \infty)$ for every distribution of (X, Y) with $EY^2 < \infty$. It is called *strongly universally consistent* if $\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx) \to 0$ $(n \to \infty)$ *a.s.* for every distribution of (X, Y) with $EY^2 < \infty$.

In Stone (1977) it was shown for the first time, that there exists weakly universally consistent estimates. Later, this result, which was proven for nearest neighbor estimates, was extended by many additional results concerning weak and strong universal consistency

of various estimates: see Devroye et al. (1994) for a list of references, and, in addition, Györfi and Walk (1996, 1997), Györfi, Kohler and Walk (1998), Walk (1997), Kohler (1997, 1999) and Kohler and Krzyżak (2000).

1.6 Slow rate. Universal consistency implies that the L_2 error of the estimate converges to zero for all distributions as the sample size tends to infinity. But it says nothing about the rate of convergence of the L_2 error towards zero. In view of applications, where one has a finite sample size, it would be very interesting to have results which imply that the L_2 error converges to zero with some given rate of convergence for all distributions.

Unfortunately, such results do not exist. Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) and Section 3 in Devroye and Wagner (1980) imply the following slow rate of convergence result: Let $\{a_n\}$ be a sequence of positive numbers converging to zero with $1/64 \geq a_1 \geq a_2 \geq \cdots$. For every sequence of regression estimates, there exists a distribution of (X, Y), such that X is uniformly distributed, $Y = m(X)$ and $E \int |m_n(x) - m(x)|^2 \mu(dx) \geq a_n$ for all n.

1.7 Rate of convergence. As we have seen above, one has to restrict the class of distributions which one considers in order to obtain non–trivial results for the rate of convergence. This was done in Stone (1982) under some smoothness conditions on the regression function. Let $p = k + \beta$ for some $k \in \mathbb{N}$ and $\beta \in [0, 1]$, and let $C > 0$. Call $f : [0, 1]^d \to \mathbb{R}$ (p, C)*–smooth* if

$$|D^\alpha f(x) - D^\alpha f(z)| \leq C|x - z|^\beta,$$

where D^α dentotes the partial derivatives for $\alpha = (\alpha_1, \ldots, \alpha_d)$, with α_i nonnegative integer, $\sum_{j=1}^d \alpha_j = k$. It was shown in Stone (1982) that the optimal rate of convergence for estimation of (p, C)–smooth regression functions is $n^{-2p/(2p+d)}$. This result was later extended by many other results (see, e.g., van de Geer (1990), Shen and Wong (1994), Birgé and Massart (1998) and Huang (1998)). All these results impose some regularity assumptions on the distribution of the design. (One should note that most of these results measure the error in the L_2 norm with integration with respect to the Lebesgue–Borel measure, in which case the analysis needs regularity conditions on the design.)

1.8 Adaptation. The results mentioned above require that parameters of the estimates are chosen depending on the smoothness of the regression function. This is not possible

in an application, because there the regression function (and also its smoothness) are unknown. Therefore it is important to construct estimates which choose their estimates in a data–driven way and nevertheless achieve under various smoothness assumptions on the regression function the corresponding optimal rate of convergence.

Such estimates were constructed for the first time in white–noise–models by Efroimovich and Pinsker (1984). Later this result was extended by many other results: A list of references for results concerning white–noise–models can be found in Tsybakov (1998). Concerning regression models similar to the one in this paper see e.g. Härdle and Marron (1985), Baraud (1997) and Barron, Birgé and Massart (1999). An extensive list of refrences is given in Section 5 in Barron, Birgé and Massart (1999).

1.9 Further results. All the results mentioned above concerning rate of convergence and adaption require some regularity assumptions on the distribution of the designs. If one considers the fact, that it is possible to derive consistency of the estimates without assuming anything on the underlying distribution, then it is natural to conjecture that one should be able to derive rate of convergence results without regularity assumptions on X. Next we comment on known corresponding results.

Under no assumption on the distribution of the design the optimal rate of convergence has been obtained for (p, C)–smooth regression functions in the case $0 < p \leq 1$ for k nearest neighbor regression estimates by Györfi (1981) and for kernel regression estimates by Krzyżak (1986). In both papers the estimates are non-adaptive in the sense that their parameters have to be chosen depending on the smoothness of the regression function.

In Kohler (2000), which was written after the original version of this manuscript, adaptive least squares spline estimates are defined and under no assumptions on the distribution of the design it is shown that the estimates achieve for bounded data and (p, C)–smooth regression functions $(0 < p < \infty)$ the optimal rate of convergence.

1.10 Main result in this article. In this articel we investigate smoothing spline estimates. Motivated by Vapnik's principle of structural risk minimization a data-driven method for the choice of the smoothing parameter is defined. It shown that the resulting estimates m_n satisfy

$$E \int |m_n(x) - m(x)|^2 \mu(dx) = O\left((\int_0^1 |m^{(k)}(x)|^2 dx)^{1/(2k+1)} \cdot n^{-2k/(2k+1)} \right)$$

for all distributions of (X, Y) with $X \in [0,1]$ $a.s.$, $|Y|$ bounded $a.s.$ and $1 \leq \int_0^1 |m^{(k)}(x)|^2 dx < \infty$ and all $k \in \{1, \ldots, K_{max}\}$, where $K_{max} \in \mathbb{N}$ is a parameter of the estimate.

Hence for bounded bivariate data and under no assumptions on the distribution of the design the estimates achieve the optimal rate of convergence. They automatically adapt to the smoothness of the regression function measured by k and $\int_0^1 |m^{(k)}(x)|^2 dx$.

1.11 What is new in this article? If one ignores the assumptions then all results which are presented in the sequel are well–known. Special in this articel is that the result holds for all distributions of X with X bounded $a.s.$ In particular it is not required that X has a density with respect to the Lebesgue–Borel measure. The difference to the result in Kohler (2000) is that in this paper the adaptation is done with respect to a more precise measure of smoothness (which is based on the Sobolev norm instead of Hölder norm).

1.12 Notation. $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ denote the sets of natural, real and nonnegative real numbers, resp. For x in $\mathbb{R}$ let $\lceil x \rceil$ be the smallest integer greater than or equal to x. I_A denotes the indicator function, $card(A)$ the cardinality of a set A. The natural logarithm is denoted by $\log(\cdot)$.

Let $L_2 = L_2([0,1])$ be the set of all measurable square integrable functions $f : [0,1] \to \mathbb{R}$ and for $k \in \mathbb{N}$ let $W_k = W_k([0,1])$ be the set of all functions $g : [0,1] \to \mathbb{R}$ whose derivatives up to order $k-1$ are absolutely continuous and whose kth derivative $f^{(k)}$ satisfies $f^{(k)} \in L_2$. For $f \in W_k$ set $J_k^2(f) = \int_0^1 |f(x)|^2 dx$.

For $L \in \mathbb{R}_+$ and a function $f : [0,1] \to \mathbb{R}$ define $T_L f : [0,1] \to \mathbb{R}$ by

$$(T_L f)(x) = \begin{cases} L & \text{if} \quad f(x) > L, \\ f(x) & \text{if} \quad -L \leq f(x) \leq L, \\ -L & \text{if} \quad f(x) < -L, \end{cases}$$

and set

$$\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|, \ \|f\|^2 = \int_{\mathbb{R}} |f(x)|^2 \mu(dx) \text{ and } \|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n |f(X_i)|^2.$$

1.13 Outline. The exact definition of the estimates is given in Section 2, the main result is formulated in Section 3. It is proven by the aid of three lemmas in Section 4, and the proofs of these lemmas are given in Sections 5, 6 and 7. An auxiliary result is presented in the appendix.

2 Definition of the estimate

In the sequel we will assume $X \in [0,1]$ *a.s.* and $|Y| \leq L$ *a.s.* for some $L \in \mathbb{R}_+$.

Let $k \in \mathbb{N}$ and $\lambda \in \mathbb{R}_+$. Depending on the data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ choose

$$\tilde{m}_{n,(k,\lambda)}(\cdot) = \tilde{m}_{n,(k,\lambda)}(\cdot, \mathcal{D}_n) \in W_k \tag{4}$$

such that

$$\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}_{n,(k,\lambda)}(X_i) - Y_i|^2 + \lambda^2 J_k^2(\tilde{m}_{n,(k,\lambda)}) = \min_{f \in W_k} \left(\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 + \lambda^2 J_k^2(f) \right), \tag{5}$$

where

$$J_k^2(f) = \int_0^1 |f^{(k)}(x)|^2 dx \tag{6}$$

is a penalty term which penalizes the roughness of the function f.

It is well–known that the minimum in (5) is achieved by some polynomial spline function of degree $2k-1$ with simple knots at the $\{X_1, \ldots, X_n\}$ (i.e., by a function which is equal to a polynomial of degree $2k-1$ (or less) between adjacent values of the X_i's and which is $(2k-2)$–times continuously differentiable on $\mathbb{R}$), and that one can compute this function easily (see e.g. de Boor (1978), chapter XIV). More informations about these so–called smoothing spline estimates can be found in the monographs Eubank (1988), Wahba (1990) and Green and Silverman (1994).

It follows from the proof of Theorem 1 below, that for a suitable choice of k and $\lambda = \lambda_{n,k}$ a truncated version of the estimate $\tilde{m}_{n,(k,\lambda_{n,k})}$ achieves the optimal rate of convergence. But this choice of k and λ depends on the smoothness of the regression function, which is unknown in an application. Therefore it is important to determine the values of the parameters k and λ in a data–dependent way.

In Kohler, Krzyżak and Schäfer (2000) a data–driven method for choosing k and λ was proposed, which is related to Vapnik's principle of structural risk minimization. The basic idea there is to derive an upper bound on the L_2 error of (a truncated version of) the estimate $\tilde{m}_{n,(k,\lambda)}$ and to choose (k, λ) such that this upper bound is minimial.

More precisely, based on a covering result for truncated functions of a Sobolev space (Lemma 3 in Kohler, Krzyżak and Schäfer (2000)), the following bound was shown:

Lemma 1 *(Kohler, Krzyżak and Schäfer (2000), Lemma 1). Let $1 \leq L < \infty$, $\lambda \in \mathbb{R}_+$ and $0 < \eta \leq 1$. Then for n sufficiently large one has with probability greater than or equal to $1 - \eta$*

$$\begin{aligned}
&\int |T_L \tilde{m}_{n,(k,\lambda)}(x) - m(x)|^2 \mu(dx) \\
&\leq L^4 \frac{\log(n)}{n} + \frac{2L^5 \log(n)^2}{n \cdot \lambda^{1/k}} + 2\lambda^2 J_k^2(\tilde{m}_{n,(k,\lambda)}) \\
&\quad +2 \left(\frac{1}{n} \sum_{i=1}^n |T_L \tilde{m}_{n,(k,\lambda)}(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right)
\end{aligned}$$

for every distribution of (X, Y) with $X \in [0, 1]$ a.s. and $|Y| \leq L$ a.s.

Then (k, λ) was chosen such that the right–hand side of the above inequality is minimal, i.e., it was chosen such that

$$\frac{1}{n} \sum_{i=1}^n |T_L \tilde{m}_{n,(k,\lambda)}(X_i) - Y_i|^2 + p\bar{e}n_n(\tilde{m}_{n,(k,\lambda)}, k, \lambda)$$

is minimal, where $p\bar{e}n_n(\tilde{m}_{n,(k,\lambda)}, k, \lambda)$ is a penalty term for the complexity of $\tilde{m}_{n,(k,\lambda)}$ defined by

$$p\bar{e}n_n(\tilde{m}_{n,(k,\lambda)}, k, \lambda) = \lambda^2 J_k^2(\tilde{m}_{n,(k,\lambda)}) + \frac{L^5 \log(n)^2}{n \cdot \lambda^{1/k}}.$$

It was shown that this estimates automatically adapts to the unknown smoothness of the regression function and that it achieves up to a logarithmic factor the optimal rate of convergence.

In the sequel we describe a modification of this estimate and show that the modified estimate achieves the optimal rate of convergence. The modification is motivated by a covering lemma of Birman and Solomjak (1967) for bounded functions of a Sobolev space (Lemma 5 below). There the complexity of a function f is measured by $J_k(f) + \|f\|_\infty$. This result motivates to choose (k, λ) by minimizing

$$\frac{1}{n} \sum_{i=1}^n |T_L \tilde{m}_{n,(k,\lambda)}(X_i) - Y_i|^2 + pen_n(\tilde{m}_{n,(k,\lambda)}, k, \lambda) \tag{7}$$

where

$$pen_n(\tilde{m}_{n,(k,\lambda)}, k, \lambda) = \lambda^2 \left(1 + J_k^2(\tilde{m}_{n,(k,\lambda)}) + \|\tilde{m}_{n,(k,\lambda)}\|_\infty^2\right) + c_1 (n\lambda^{1/k})^{-1}. \tag{8}$$

Here c_1 is a constant which is specified in Theorem 1 below.

More precisely, set

$$\mathcal{P}_n = \{1, \ldots, K_{max}\} \times \left\{2^{l/2} \quad : \quad l = -\lceil \log(n) \rceil, \ldots, \lceil \log(n) \rceil\right\}.$$

For $(k, \lambda) \in \mathcal{P}_n$ define $\tilde{m}_{n,(k,\lambda)}$ as above and set

$$m_{n,(k,\lambda)}(\cdot) = T_L \tilde{m}_{n,(k,\lambda)}(\cdot).$$

Depending on the data we choose that estimate of the family

$$\left\{m_{n,(k,\lambda)} \quad : \quad (k, \lambda) \in \mathcal{P}_n\right\}$$

which minimizes (7). More precisely, we choose

$$(k^*, \lambda^*) = (k^*(\mathcal{D}_n), \lambda^*(\mathcal{D}_n)) \in \mathcal{P}_n \tag{9}$$

such that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} |m_{n,(k^*,\lambda^*)}(X_i) - Y_i|^2 + pen_n(\tilde{m}_{n,(k^*,\lambda^*)}, k^*, \lambda^*) \\ &= \min_{(k,\lambda)\in\mathcal{P}_n} \left(\frac{1}{n}\sum_{i=1}^{n} |m_{n,(k,\lambda)}(X_i) - Y_i|^2 + pen_n(\tilde{m}_{n,(k,\lambda)}, k, \lambda)\right), \end{aligned} \tag{10}$$

and define our adaptive estimate m_n by

$$m_n(x) = m_n(x, \mathcal{D}_n) = m_{n,(k^*(\mathcal{D}_n),\lambda^*(\mathcal{D}_n))}(x, \mathcal{D}_n).$$

3 Main result

The main result of this article is the following theorem.

Theorem 1 *Let $L \in \mathbb{R}_+$ and $K_{max} \in \mathbb{N}$ be arbitrary. Let the estimate m_n be defined as in Section 2. Choose the constant $c_1 \in \mathbb{R}_+$ in (8) such that $c_1 \geq \max\{85 \cdot c_5 L^2, 512 \cdot 4^{K_{max}} t_0^2\}$ where c_5 and t_0 are defined in Lemma 5 and Lemma 6 below. Then*

$$E\int_{\mathbb{R}} |m_n(x) - m(x)|^2 \mu(dx) = O\left(J_k(m)^{2/(2k+1)} n^{-2k/(2k+1)}\right)$$

for every $k \in \{1, \ldots, K_{max}\}$ and for every distribution of (X, Y) with $X \in [0, 1]$ a.s., $|Y| \leq L$ a.s. and $1 \leq J_k(m) < \infty$.

Remark 1. To compute the estimate one has to compute the supremum norm of the spline functions $\tilde{m}_{n,(k,\lambda)}$. Up to some constant factors the supremum norm of a spline function can be bounded from below and from above by the maximum of the absolute values of its B–spline coefficients (see de Boor (1978), Corollary XI.3). Using this it follows from the proof of Theorem 1 that the result also holds if one replaces $\|\tilde{m}_{n,(k,\lambda)}\|_\infty$ by the maximum of the absolute values of the B–spline coefficients of $\tilde{m}_{n,(k,\lambda)}$, which can be computed easily.

Remark 2. The estimate m_n achieves the optimal rate of convergence although its definition does not depend on the smoothness of the regression function measured by k and $J_k(m)$. Hence it adapts automatically to the smoothness of the regression function.

Remark 3. It is possible to replace the assumption that Y is bounded $a.s.$ in Theorem 1 by some weaker assumption on the existence of exponential moments of Y. But in this case the constant used in the penalty term (8) will depend on this exponential moment, which is usually unknown in an application.

Remark 4. We want to stress that in Theorem 1 there is no assumption on the distribution of X besides $X \in [0,1]$ $a.s.$, in particular it is not required that X has a density with respect to the Lebesgue-Borel measure. To the knowledge of the author it is an open problem wether one can show a similar result without assuming that X is bounded $a.s.$ (cf. Question 1 in Stone (1982)).

4 Proof of Theorem 1

Recall that

$$\|f\|^2 = \int_{\mathbb{R}} |f(x)|^2 \mu(dx) \text{ and } \|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n |f(X_i)|^2$$

for functions $f : [0,1] \to \mathbb{R}$.

Because of $\|m_n - m\|^2 \leq 4L^2$ it suffices to show that for some $c_2 \in \mathbb{R}_+$

$$P\left[\|m_n - m\|^2 > c_2 J_k(m)^{2/(2k+1)} n^{-2k/(2k+1)}\right] \to 0 \quad (n \to \infty). \tag{11}$$

First we consider the case $card(supp(X)) > K_{max}$. In this case we use the error decomposition

$$\|m_n - m\|^2 = T_{1,n} + 8T_{2,n}$$

where

$$T_{1,n} = \|m_n - m\|^2 - 8T_{2,n}$$

and

$$T_{2,n} = \|m_n - m\|_n^2 + (\lambda^*)^2 \left(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) + \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}.$$

To bound $T_{1,n}$ we will show

Lemma 2 *Let m_n be defined as in Section 2. Assume that the constant $c_1 \in \mathbb{R}_+$ in (8) is chosen such such that $c_1 \geq 85 \cdot c_5 L^2$ where c_5 is defined in Lemma 5 below. Then*

$$P\left[T_{1,n} > \frac{\ln(n)}{n}\right] \to 0 \quad (n \to \infty)$$

for every distribution of (X, Y) with $X \in [0,1]$ a.s. and $|Y| \leq L$ a.s.

The proof of Lemma 2 is given in Section 5. It uses standard arguments concerning the equivalence of the empirical norm $\|\cdot\|_n$ and the theoretical norm $\|\cdot\|$ together with a result about supremum norm covering of balls in Sobolev spaces derived in Birman and Solomjak (1967).

To bound $T_{2,n}$ we will show

Lemma 3 *Let m_n be defined as in Section 2. Assume that the constant $c_1 \in \mathbb{R}_+$ in (8) is chosen such such that $c_1 \geq 521 \cdot 4^{K_{max}} t_0^2$ where t_0 is defined in Lemma 6 below. Then*

$$P\left[T_{2,n} > c_2\, J_k(m)^{2/(2k+1)} n^{-2k/(2k+1)}\right] \to 0 \quad (n \to \infty)$$

for every distribution of (X, Y) with $X \in [0,1]$ a.s., $card(supp(X)) > K_{max}$, $|Y| \leq L$ a.s. and $1 \leq J_k(m) < \infty$. Here $c_2 \in \mathbb{R}_+$ is a constant which depends only on k, L and the distribution of X.

To prove Lemma 3 we will first use the definition of m_n to show

$$\begin{aligned}
(12)\quad & \|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + (\lambda^*)^2 \left(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) + \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1} \\
& \leq \lambda_{n,k}^2 \left(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2\right) + c_1(n\lambda_{n,k}^{1/k})^{-1} \\
& \quad + \frac{2}{n}\sum_{i=1}^n \left(\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i)\right)(Y_i - m(X_i)) - \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1},
\end{aligned}$$

where $\lambda_{n,k} \approx (J_k^2(m) \cdot n)^{-k/(2k+1)}$. Then we will apply results of van de Geer (1990) to bound the term

$$\frac{2}{n} \sum_{i=1}^{n} \left(\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i) \right) (Y_i - m(X_i)) \tag{13}$$

by some constant times

$$n^{-\frac{1}{2}} \|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^{1-\frac{1}{2k^*}} \left(1 + J_{k^*}(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty \right)^{\frac{1}{2k^*}}$$

and use this and (12) to show that the term (13) is less than $\frac{1}{2} c_1 (n(\lambda^*)^{1/k^*})^{-1}$, from which the result follows by (12) together with a bound on the supremum norm of the estimate $\tilde{m}_{n,(k,\lambda_{n,k})}$. The complete proof of Lemma 3 is given in Section 6.

From Lemma 2 and Lemma 3 one obtains the assertion in the case $card(supp(X)) > K_{max}$. Next we consider the case $card(supp(X)) \leq K_{max}$. On the random event $[k^* < card(supp(X))]$ the estimate m_n is equal to an estimate which is defined like m_n but with K_{max} replaced by $\bar{K}_{max} = card(supp(X)) - 1$. But for the latter estimate $card(supp(X)) > \bar{K}_{max}$, therefore one can apply the results of the first case to this estimate which implies

$$P\left[\|m_n - m\|^2 > c_2 \, J_k(m)^{2/(2k+1)} n^{-2k/(2k+1)}, k^* < card(supp(X)) \right] \to 0 \quad (n \to \infty).$$

Hence the assertion follows from

Lemma 4 *Let m_n be defined as in Section 2. Then*

$$E\left[\int |m_n(x) - m(x)|^2 \mu(dx) \cdot I_{[card(supp(X)) \leq k^*]} \right] = O\left(\frac{1}{n}\right)$$

for every distribution of (X, Y) with $X \in [0,1]$ a.s., $card(supp(X)) \leq K_{max}$ and $|Y| \leq L$ a.s.

The main trick in the proof of Lemma 4, which is given in Section 7, is the observation that $k \geq card\{X_1, \ldots, X_n\}$ implies that for any $\lambda \in \mathbb{R}_+$ $m_{n,(k,\lambda)}(x)$ is equal to the average value of those Y_i where $X_i = x$ $(x \in \{X_1, \ldots, X_n\})$.

This completes the proof. □

5 Proof of Lemma 2

To prove Lemma 2 we will need the following notation: Let $(M, \|\cdot\|')$ be a normed space, let $A \subseteq M$ and let $\delta > 0$. For $z \in M$ we denote the $\|\cdot\|'$ ball of radius δ around z by $B_\delta(z)$, i.e. we set $B_\delta(z) = \{x \in M \ : \ \|x - z\|' < \delta\}$. We call $\{z_1, \ldots, z_k\} \subseteq M$ a δ–cover of A if $A \subseteq \bigcup_{i=1,\ldots,k} B_\delta(z_i)$. Note that if $\{z_1, \ldots, z_k\} \subseteq M$ is a δ–cover of A, then there exists a 2δ–cover of A of the same size which is contained in A. The δ–covering number $\mathcal{N}(\delta, A, \|\cdot\|')$ is the minimal size of a δ–cover of A. Next we present a result concerning covers with respect to the norm $\|\cdot\|_\infty$. (Recall that $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$ for a function $f : [0,1] \to \mathbb{R}$.)

Lemma 5 *Let $k \in \mathbb{N}$ and $c_4 > 0$. Then there exists a constant $c_5 \in \mathbb{R}_+$ (independent of k and c_4) such that*

$$\mathcal{N}\left(\delta, \{f \in W_k \ : \ \|f\|_\infty + J_k(f) \le c_4\}, \|\cdot\|_\infty\right) \le \exp\left(c_5 \left(\frac{c_4}{\delta}\right)^{\frac{1}{k}}\right).$$

Proof. See Birman and Solomjak (1967), Theorem 5.2. □

We start now with the proof of Lemma 2. We have

$$\begin{aligned}
&P\left[T_{1,n} > \frac{\ln(n)}{n}\right] \\
&\le \sum_{(k,\lambda)\in\mathcal{P}_n} P\Big[\|m_{n,(k,\lambda)} - m\|^2 - 8\|m_{n,(k,\lambda)} - m\|_n^2 > \\
&\qquad\qquad 8\lambda^2\left(1 + J_k^2(\tilde{m}_{n,(k,\lambda)}) + \|\tilde{m}_{n,(k,\lambda)}\|_\infty^2\right) + 4c_1(n\lambda^{1/k})^{-1} + \frac{\ln(n)}{n}\Big] \\
&\le \sum_{(k,\lambda)\in\mathcal{P}_n} \Bigg(P\Big[\exists f \in W_k : \ J_k^2(f) + \|f\|_\infty^2 \le 2, \\
&\qquad\qquad \|T_L f - m\|^2 - 8\|T_L f - m\|_n^2 > 8\lambda^2 + 4c_1(n\lambda^{1/k})^{-1} + \frac{\ln(n)}{n}\Big] \\
&\qquad + \sum_{l=2}^{\infty} P\Big[\exists f \in W_k : \ 2^{l-1} < J_k^2(f) + \|f\|_\infty^2 \le 2^l, \\
&\qquad\qquad \|T_L f - m\|^2 - 8\|T_L f - m\|_n^2 > 8\cdot 2^{l-1}\lambda^2 + 4c_1(n\lambda^{1/k})^{-1} + \frac{\ln(n)}{n}\Big]\Bigg) \\
&= \sum_{(k,\lambda)\in\mathcal{P}_n} \sum_{l=1}^{\infty} P\Big[\exists f \in W_k : \ J_k^2(f) + \|f\|_\infty^2 \le 2^l, \\
&\qquad\qquad \|T_L f - m\|^2 - 8\|T_L f - m\|_n^2 > 2^{l+2}\lambda^2 + 4c_1(n\lambda^{1/k})^{-1} + \frac{\ln(n)}{n}\Big],
\end{aligned}$$

where we have used $\tilde{m}_{n,(k,\lambda)} \in W_k$.

In the sequel we will use some standard techniques concerning the equivalence of theoretical norm $\|f\|$ and empirical norm $\|f\|_n$ to bound the above probabilities. (Related arguments can be found e.g. in van de Geer (1988), Section 6.3)

Fix $(k,\lambda) \in \mathcal{P}_n$, $l \in \mathbb{N}$ and set $\delta = \sqrt{2^{l+2}\lambda^2/20}$. Because of Lemma 5 there exists a δ–cover $\mathcal{F}_\delta \subseteq \{f \in W_k \ : \ J_k^2(f) + \|f\|_\infty^2 \leq 2^l\}$ of the set $\{f \in W_k \ : \ J_k^2(f) + \|f\|_\infty^2 \leq 2^l\}$, whose cardinality $card(\mathcal{F}_\delta)$ satisfies

$$card(\mathcal{F}_\delta) \leq \exp\left(c_5 \left(\frac{2\sqrt{2^l}}{\delta}\right)^{\frac{1}{k}}\right) \leq \exp\left(c_5 \cdot 2 \cdot \sqrt{5} \cdot \frac{1}{\lambda^{1/k}}\right)$$

We will bound the above probability by a probability involving only functions which are contained in $\mathcal{F}_\delta$. To do this, we observe that for $f, g \in W_k$ with $\|f-g\|_\infty \leq \delta$ one has

$$\|f-m\|^2 \leq 2\|g-m\|^2 + 2\delta^2$$

and

$$2\|f-m\|_n^2 \geq \|g-m\|_n^2 - 2\delta^2.$$

Thus

$$\begin{aligned}
&P\Big[\exists f \in W_k : \ J_k^2(f) + \|f\|_\infty^2 \leq 2^l, \\
&\qquad\qquad \|T_L f - m\|^2 - 8\|T_L f - m\|_n^2 > 2^{l+2}\lambda^2 + 4c_1(n\lambda^{1/k})^{-1} + \frac{\ln(n)}{n}\Big] \\
&\leq P\Big[\exists g \in \mathcal{F}_\delta : 2\|T_L g - m\|^2 + 2\delta^2 - 4\|T_L g - m\|_n^2 + 8\delta^2 > \\
&\qquad\qquad 2^{l+2}\lambda^2 + 4c_1(n\lambda^{1/k})^{-1} + \frac{\ln(n)}{n}\Big] \\
&\leq card(\mathcal{F}_\delta) \cdot \max_{g \in \mathcal{F}_\delta} P\Big[\|T_L g - m\|^2 - \|T_L g - m\|_n^2 > \\
&\qquad\qquad \frac{1}{4}(2\|T_L g - m\|^2 + 2^{l+1}\lambda^2 + 4c_1(n\lambda^{1/k})^{-1} + \frac{\ln(n)}{n})\Big],
\end{aligned}$$

where we have used $10\delta^2 = 2^{l+1}\lambda^2$.

Fix $g \in \mathcal{F}_\delta$ and set $Z_i = |T_L g(X_i) - m(X_i)|^2$ $(i = 1, \ldots, n)$. Then $\|T_L g - m\|_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i$, $\|T_L g - m\|^2 = EZ_1$, $Z_1, \ldots, Z_n$ are i.i.d. and $0 \leq Z_1 \leq 4L^2$ $a.s.$ Using the inequality of Bernstein (see e.g. Devroye, Györfi and Lugosi (1996), Theorem 8.2) and

$$Var(Z_1) \leq EZ_1^2 \leq 4L^2 EZ_1 = 4L^2\|T_L g - m\|^2$$

one obtains:

$$P\Big[\|T_Lg-m\|^2-\|T_Lg-m\|_n^2>\frac{1}{4}(2\|T_Lg-m\|^2+2^{l+1}\lambda^2+4c_1(n\lambda^{1/k})^{-1}+\frac{\ln(n)}{n})\Big]$$

$$\leq \exp\left(-\frac{n\cdot(\frac{1}{2}\|T_Lg-m\|^2+2^{l-1}\lambda^2+c_1(n\lambda^{1/k})^{-1}+\frac{\ln(n)}{4\cdot n})^2}{2Var(Z_1)+\frac{2}{3}4L^2(\frac{1}{2}\|T_Lg-m\|^2+2^{l-1}\lambda^2+c_1(n\lambda^{1/k})^{-1}+\frac{\ln(n)}{4\cdot n})}\right)$$

$$\leq \exp\left(-\frac{n\cdot(\frac{1}{2}\|T_Lg-m\|^2+2^{l-1}\lambda^2+c_1(n\lambda^{1/k})^{-1}+\frac{\ln(n)}{4\cdot n})}{(16+\frac{8}{3})L^2}\right)$$

$$\leq \exp\left(-\frac{c_1}{19L^2}\cdot\lambda^{-1/k}\right)\cdot\exp\left(-\frac{\ln(n)}{76L^2}\right)\cdot\exp\left(-2^{l-1}\frac{n\lambda^2}{19L^2}\right).$$

Because of $c_1\geq c_5\cdot 2\sqrt{5}\cdot 19L^2$ one gets

$$P\Big[\exists f\in W_k:\ J_k^2(f)+\|f\|_\infty^2\leq 2^l,$$
$$\|T_Lf-m\|^2-8\|T_Lf-m\|_n^2>2^{l+2}\lambda^2+4c_1(n\lambda^{1/k})^{-1}+\frac{\ln(n)}{n}\Big]$$

$$\leq \exp\left(-\frac{\ln(n)}{76L^2}\right)\cdot\exp\left(-2^{l-1}\frac{n\lambda^2}{19L^2}\right)$$

$$\leq \exp\left(-\frac{\ln(n)}{76L^2}\right)\cdot\exp\left(-\frac{2^l}{76L^2}\right)$$

because of $n\lambda^2\geq\frac{1}{2}$ for $(k,\lambda)\in\mathcal{P}_n$. Hence

$$P\left[T_{1,n}>\frac{\ln(n)}{n}\right]\leq\sum_{(k,\lambda)\in\mathcal{P}_n}\left(\sum_{l=1}^{\infty}\exp\left(-\frac{2^l}{76L^2}\right)\right)\exp\left(-\frac{\ln(n)}{76L^2}\right)$$

$$\leq card(\mathcal{P}_n)\cdot\frac{\exp(-1/(76L^2))}{1-\exp(-1/(76L^2))}\cdot\exp\left(-\frac{\ln(n)}{76L^2}\right)\to 0\quad(n\to\infty).$$

□

6 Proof of Lemma 3

The main tool in the proof of Lemma 3 is the following lemma.

Lemma 6 *There exist constants $c_6,t_0\in\mathbb{R}_+$ depending only on L such that for all $k\in\mathbb{N}$, $t,\sigma>0$ with $t>2^k\sigma^{\frac{1}{2k}}t_0$ and with $t\geq 2^kt_0$ and all distributions of (X,Y) with $|Y|\leq L$ a.s.*

$$P\left[\sup_{f:\|f-m\|_n\leq\sigma}\frac{\frac{1}{n}\sum_{i=1}^n(f(X_i)-m(X_i))(Y_i-m(X_i))}{n^{-\frac{1}{2}}\|f-m\|_n^{1-\frac{1}{2k}}(1+J_k(f)+\|f\|_\infty)^{\frac{1}{2k}}}>t\right]\leq 2\exp\left(-\frac{c_6\,t^2}{4^k\sigma^{\frac{1}{k}}}\right).$$

Proof. The proof is a modification of the proof of Lemma 6.1 in van de Geer (1990). Therefore we give only the outline of the proof.

$$P\left[\sup_{f:\|f-m\|_n\le\sigma}\frac{\frac{1}{n}\sum_{i=1}^n(f(X_i)-m(X_i))(Y_i-m(X_i))}{n^{-\frac{1}{2}}\|f-m\|_n^{1-\frac{1}{2k}}(1+J_k(f)+\|f\|_\infty)^{\frac{1}{2k}}}>t\right]$$

$$\le\sum_{j=0}^\infty\left\{P\left[\sup_{\substack{f:\|f-m\|_n\le\frac{\sigma}{2^{j\cdot k}},\ \|f-m\|_n>\frac{\sigma}{2^{(j+1)k}},\\ J_k(f)+\|f\|_\infty\le 1}}\ldots>t\right]\right.$$

$$\left.+\sum_{l=1}^\infty P\left[\sup_{\substack{f:\|f-m\|_n\le\frac{\sigma}{2^{j\cdot k}},\ \|f-m\|_n>\frac{\sigma}{2^{(j+1)k}},\\ 2^{(l-1)k}<J_k(f)+\|f\|_\infty\le 2^{l\cdot k}}}\ldots>t\right]\right\}$$

$$\le\sum_{j=0}^\infty\sum_{l=0}^\infty P\left[\sup_{\substack{f:\|f-m\|_n\le\frac{\sigma}{2^{j\cdot k}},\\ J_k(f)+\|f\|_\infty\le 2^{l\cdot k}}}\frac{1}{n^{\frac{1}{2}}}\sum_{i=1}^n(f(X_i)-m(X_i))(Y_i-m(X_i))\right.$$

$$\left.>t\left(\frac{\sigma}{2^{j\cdot k}}\right)^{1-\frac{1}{2k}}\left(2^{l\cdot k}\right)^{\frac{1}{2k}}2^{-k}\right].$$

It follows from Theorem 3.3 in van de Geer (1990) and Lemma 5 above that the last probability is bounded from above by

$$2\exp\left(-\frac{\alpha\left(t\left(\frac{\sigma}{2^{j\cdot k}}\right)^{1-\frac{1}{2k}}\left(2^{l\cdot k}\right)^{\frac{1}{2k}}2^{-k}\right)^2}{\left(\frac{\sigma}{2^{j\cdot k}}\right)^2}\right)=2\exp\left(-\frac{\alpha t^2 2^{j+l}}{4^k\sigma^{\frac{1}{k}}}\right)$$

for some constant $\alpha>0$ (depending only on L). From this one easily obtains the assertion.

□

We start now with the proof of Lemma 3. Because of $\|m\|_\infty\le L$ we have $\|m_n-m\|_n^2\le\|\tilde{m}_n-m\|_n^2$, hence it suffices to show

$$P\Big[\|\tilde{m}_{n,(k^*,\lambda^*)}-m\|_n^2+(\lambda^*)^2(1+J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)})+\|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2)+\frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}$$
$$>c_2J_k(m)^{2/(2k+1)}n^{-2k/(2k+1)}\Big]\to 0\quad(n\to\infty).$$

Choose $\lambda_{n,k}$ such that $(k,\lambda_{n,k})\in\mathcal{P}_n$ and that for n sufficiently large

$$\frac{1}{2}\left(J_k^2(m)\cdot n\right)^{-k/(2k+1)}\le\lambda_{n,k}\le\left(J_k^2(m)\cdot n\right)^{-k/(2k+1)}.$$

We will first show

$$(14)\quad \|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + (\lambda^*)^2 \left(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) + \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}$$
$$\leq \lambda_{n,k}^2 \left(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2\right) + c_1(n\lambda_{n,k}^{1/k})^{-1}$$
$$+\frac{2}{n}\sum_{i=1}^n \left(\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i)\right)(Y_i - m(X_i)) - \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}.$$

To do this we use

$$\frac{1}{n}\sum_{i=1}^n |\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - Y_i|^2$$
$$= \|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 - \frac{2}{n}\sum_{i=1}^n (\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i))(Y_i - m(X_i)) + \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2,$$

(10) and (5) and conclude

$$\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + (\lambda^*)^2 \left(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) + \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}$$
$$= \frac{1}{n}\sum_{i=1}^n |\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - Y_i|^2 + (\lambda^*)^2 \left(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) + c_1(n(\lambda^*)^{1/k^*})^{-1}$$
$$-\frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2 + \frac{2}{n}\sum_{i=1}^n (\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i))(Y_i - m(X_i)) - \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}$$
$$\leq \frac{1}{n}\sum_{i=1}^n |\tilde{m}_{n,(k,\lambda_{n,k})}(X_i) - Y_i|^2 + \lambda_{n,k}^2 \left(1 + J_k^2(\tilde{m}_{n,(k,\lambda_{n,k})}) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2\right) + c_1(n\lambda_{n,k}^{1/k})^{-1}$$
$$-\frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2 + \frac{2}{n}\sum_{i=1}^n (\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i))(Y_i - m(X_i)) - \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}$$
$$\leq \lambda_{n,k}^2 \left(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2\right) + c_1(n\lambda_{n,k}^{1/k})^{-1}$$
$$+\frac{2}{n}\sum_{i=1}^n (\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i))(Y_i - m(X_i)) - \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}.$$

This implies (14).

It follows from (5) that with probability one

$$\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 \leq 2\|\tilde{m}_{n,(k^*,\lambda^*)}\|_n^2 + 2\|m\|_n^2$$
$$\leq 4\left(\frac{1}{n}\sum_{i=1}^n |Y_i - \tilde{m}_{n,(k^*,\lambda^*)}(X_i)|^2 + (\lambda^*)^2 J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)})\right) + 4\frac{1}{n}\sum_{i=1}^n |Y_i|^2 + 2\frac{1}{n}\sum_{i=1}^n m(X_i)^2$$
$$\leq 4\left(\frac{1}{n}\sum_{i=1}^n |Y_i - 0|^2 + (\lambda^*)^2 \cdot 0\right) + 4\frac{1}{n}\sum_{i=1}^n |Y_i|^2 + 2\frac{1}{n}\sum_{i=1}^n m(X_i)^2 \leq 10L^2.$$

Thus applying Lemma 6 with $\sigma = \sqrt{10}L$ and $t = \frac{\ln(n)}{8}$ yields that the probability of the event

$$\left[\frac{2}{n}\sum_{i=1}^{n}(\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i))(Y_i - m(X_i)) > \right.$$
$$\left. \frac{\ln(n)}{4n^{\frac{1}{2}}}\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^{1-\frac{1}{2k^*}}(1 + J_{k^*}(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty)^{\frac{1}{2k^*}}\right]$$

is bounded above by

$$\sum_{(k,\lambda)\in\mathcal{P}_n} 2\exp\left(-\frac{c_6\ \ln(n)^2}{64\cdot 4^{K_{max}}(\sqrt{10}L)^{\frac{1}{k}}}\right)$$

and thus tends to zero ($n \to \infty$). Outside of this event it follows from (14) that for n sufficiently large

$$\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + \lambda_{n,k^*}^2\left(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) + \frac{1}{2}c_1(n(\lambda^*)^{1/k^*})^{-1}$$
$$\leq \lambda_{n,k}^2\left(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2\right) + c_1(n\lambda_{n,k}^{1/k})^{-1}$$
$$+\frac{\ln(n)}{4n^{\frac{1}{2}}}\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^{1-\frac{1}{2k^*}}\left(1 + J_{k^*}(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty\right)^{\frac{1}{2k^*}}$$
$$\leq n^{-\frac{1}{3}}(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2)$$
$$+\frac{\ln(n)}{n^{\frac{1}{2}}}\left(\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + (\lambda^*)^2(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2)\right)^{\frac{1}{2}}(\lambda^*)^{-\frac{1}{2k^*}}.$$

If the left-hand side of the above inequality is larger than $2n^{-\frac{1}{3}}(1+J_k^2(m)+\|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2)$, then the above inequality implies

$$\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n \leq 2\frac{\ln(n)}{n^{\frac{1}{2}}}(\lambda^*)^{-\frac{1}{2k^*}} \leq 2\frac{\ln(n)}{n^{\frac{1}{2}}}(1/(2n))^{-1/4} = 4\ln(n)\cdot n^{-1/4}.$$

If the left-hand side of the above inequality is less than or equal to $2n^{-\frac{1}{3}}(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2)$, then

$$\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n \leq \sqrt{2}n^{-1/6}(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda)}\|_\infty^2).$$

It follows from the proof of Lemma 3.1 in Mammen and van de Geer (1997), that there exists a constant c_8 depending only on k, L and the distribution of X such that outside of an event whose probability tends to zero ($n \to \infty$) one has

$$\|\tilde{m}_{n,(k\lambda)}\|_\infty \leq c_8 J_k^2(m). \tag{15}$$

For the sake of completeness we sketch the proof of this result in appendix A.

Hence we have shown

$$\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n \leq \ln(n) n^{-\frac{1}{6}}$$

outside of an event whose probability tends to zero ($n \to \infty$).

Now we apply again Lemma 6, but this time with $\sigma = \sigma_n = \ln(n) n^{-\frac{1}{6}}$ and $t = t_1 = 2^{K_{max}} t_0$. It follows that the probability of the event

$$\left[\frac{2}{n}\sum_{i=1}^{n}(\tilde{m}_{n,(k^*,\lambda^*)}(X_i) - m(X_i))(Y_i - m(X_i)) > \frac{2t_1}{n^{\frac{1}{2}}}\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^{1-\frac{1}{2k^*}}(1 + J_{k^*}(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty)^{\frac{1}{2k^*}}\right]$$

tends to zero. Outside of this event, (14) implies

$$\begin{aligned}
&\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + (\lambda^*)^2\left(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) + \frac{1}{2}c_1(n(\lambda^*)^{-1/k^*}) \\
&\leq \lambda_{n,k}^2\left(1 + J_k^2(m) + \|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty^2\right) + c_1(n\lambda_{n,k}^{-1/k}) \\
&\quad + \frac{8t_1}{n^{\frac{1}{2}}}\left(\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + (\lambda^*)^2(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2)\right)^{\frac{1}{2}}(\lambda^*)^{-\frac{1}{2k^*}} \\
&\quad - \frac{1}{2}c_1(n(\lambda^*)^{-1/k^*}). \qquad (16)
\end{aligned}$$

If the first line on the right–hand side of (16) is greater than the second line, then (16) implies

$$\begin{aligned}
&\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + \lambda_{n,k^*}^2\left(\frac{1}{2} + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2\right) \\
&\leq 2\lambda_{n,k}^2\left(1 + J_k^2(m) + \|m_{n,(k,\lambda_{n,k})}\|_\infty^2\right). \qquad (17)
\end{aligned}$$

If this is not the case, then (16) implies

$$\begin{aligned}
&\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + \lambda^*(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2) \\
&\leq \frac{16t_1}{n^{1/2}}\left(\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + \lambda^*(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2)\right)^{1/2}(\lambda^*)^{-1/(2k^*)}
\end{aligned}$$

which is equivalent to

$$\left(\|\tilde{m}_{n,(k^*,\lambda^*)} - m\|_n^2 + \lambda^*(1 + J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + \|\tilde{m}_{n,(k^*,\lambda^*)}\|_\infty^2)\right)^{1/2} \leq \frac{16t_1}{n^{1/2}}(\lambda^*)^{-1/(2k^*)}.$$

From this, (16) and

$$c_1 \geq 512t_1^2 = 512 \cdot 4^{K_{max}} t_0^2$$

one can again conclude (17).

Summarizing the above results we get, that (17) holds outside of an event whose probability tends to zero ($n \to \infty$). Together with (15) this implies the assertion. □

7 Proof of Lemma 4

Assume $k^* \geq card(supp(X))$. Then there exists a polynomial f of degree $k^* - 1$ (or less) such that

$$f(x) = \frac{\sum_{i=1}^n Y_i \cdot I_{[X_i=x]}}{\sum_{i=1}^n I_{[X_i=x]}} \quad \text{for all } x \in \{X_1, \ldots, X_n\}. \tag{18}$$

It is easy to see that an arbitrary function $f : [0,1] \to \mathbb{R}$ minimizes

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

with respect to all other functions if and only if f satisfies (18). This together with the definition of m_n, $|Y| \leq L$ *a.s.* and the fact that $J_{k^*}(f) = 0$ for every polynomial f of degree $k^* - 1$ (or less) implies that with probability one

$$m_n(x) = \frac{\sum_{i=1}^n Y_i \cdot I_{[X_i=x]}}{\sum_{i=1}^n I_{[X_i=x]}} \quad \text{for all } x \in \{X_1, \ldots, X_n\}. \tag{19}$$

Furthermore, by definition of m_n,

$$|m_n(x)| \leq L \quad \text{for all } x \in \mathbb{R}. \tag{20}$$

Hence it suffices to show

$$E \int |m_n(x) - m(x)|^2 \mu(dx) = O\left(\frac{1}{n}\right) \tag{21}$$

for any estimate m_n which satisfies (19) and (20).

To this end, let $supp(X) = \{x_1, \ldots, x_K\}$ and set $p_k = P[X = k] = \mu(\{x_k\})$. Then

$$E \int |m_n(x) - m(x)|^2 \mu(dx) = \sum_{k=1}^K E\left[|m_n(x_k) - m(x_k)|^2\right] \cdot p_k$$

$$= \sum_{k=1}^K E\left[|m_n(x_k) - m(x_k)|^2 \cdot I_{[\sum_{i=1}^n I_{[X_i=x_k]}=0]}\right] \cdot p_k$$

$$+ \sum_{k=1}^K E\left[|m_n(x_k) - m(x_k)|^2 \cdot I_{[\sum_{i=1}^n I_{[X_i=x_k]}>0]}\right] \cdot p_k$$

$$\leq 4L^2 \cdot \sum_{k=1}^K (1-p_k)^n \cdot p_k + \sum_{k=1}^K E\left[\left|\frac{\sum_{i=1}^n (Y_i - m(x_k)) \cdot I_{[X_i=x_k]}}{\sum_{i=1}^n I_{[X_i=x_k]}}\right|^2 \cdot I_{[\sum_{i=1}^n I_{[X_i=x_k]}>0]}\right] \cdot p_k.$$

By independence of $(X_1, Y_1), \ldots, (X_n, Y_n)$ we have for $i \neq j$

$$\begin{aligned} &E\left[(Y_i - m(x_k)) \cdot I_{[X_i = x_k]} \cdot (Y_j - m(x_k)) \cdot I_{[X_j = x_k]} \middle| X_1, \ldots, X_n\right] \\ &= (m(X_i) - m(x_k)) \cdot I_{[X_i = x_k]} \cdot (m(X_j) - m(x_k)) \cdot I_{[X_j = x_k]} = 0. \end{aligned}$$

From this together with $|Y| \leq L$ *a.s.* one concludes

$$\begin{aligned} &E\left[\left|\frac{\sum_{i=1}^n (Y_i - m(x_k)) \cdot I_{[X_i = x_k]}}{\sum_{i=1}^n I_{[X_i = x_k]}}\right|^2 \cdot I_{[\sum_{i=1}^n I_{[X_i = x_k]} > 0]}\right] \\ &= E\left[\frac{\sum_{i=1}^n E\left[|Y_i - m(x_k)|^2 | X_1, \ldots, X_n\right] \cdot I_{[X_i = x_k]}}{|\sum_{i=1}^n I_{[X_i = x_k]}|^2} \cdot I_{[\sum_{i=1}^n I_{[X_i = x_k]} > 0]}\right] \\ &\leq 4L^2 E\left[\frac{1}{\sum_{i=1}^n I_{[X_i = x_k]}} I_{[\sum_{i=1}^n I_{[X_i = x_k]} > 0]}\right]. \end{aligned}$$

An easy computation shows that if Z is a binomially distributed random variable with parameters n and p, then

$$E\left[\frac{1}{Z} \cdot I_{[Z>0]}\right] \leq \frac{2}{(n+1) \cdot p}.$$

Hence

$$E\left[\left|\frac{\sum_{i=1}^n (Y_i - m(x_k)) \cdot I_{[X_i = x_k]}}{\sum_{i=1}^n I_{[X_i = x_k]}}\right|^2 \cdot I_{[\sum_{i=1}^n I_{[X_i = x_k]} > 0]}\right] \leq \frac{8L^2}{(n+1) \cdot p_k}.$$

Summarizing the above results one gets

$$E \int |m_n(x) - m(x)|^2 \mu(dx) \leq 4L^2 \cdot \sum_{k=1}^K (1 - p_k)^n p_k + 8L^2 \frac{K}{n+1}$$

which implies (21). □

Appendix

A A bound on the supremum norm of $\tilde{m}_{n,(k,\lambda_{n,k})}$

In this section we use arguments of Mammen and van de Geer (1997) to show

Lemma 7 *Let $k \in \mathbb{N}$ and $1 \leq L < \infty$. Then*

$$P\left[\|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty > c_8 J_k^2(m)\right] \to 0 \quad (n \to \infty)$$

for every distribution of (X,Y) *with* $X \in [0,1]$ *a.s.,* $card(supp(X)) > k$, $|Y| \leq L$ *and* $1 \leq J_k(m) < \infty$. *Here* $c_8 \in \mathbb{R}_+$ *is a constant which depends only on* k, L *and the distribution of* X.

Proof.

We will first show

$$P\left[\|\tilde{m}_{n,(k,\lambda_{n,k})}\|_\infty > 5kc_9L(1+J_k(\tilde{m}_{n,(k,\lambda_{n,k})}))\right] \to 0 \quad (n \to \infty) \tag{22}$$

for some constant c_9 depending only on the distribution of X.

By the Sobolev–embedding theorem, one can write

$$\tilde{m}_{n,(k,\lambda_{n,k})}(x) = \tilde{m}_{n,(k,\lambda_{n,k}),1}(x) + \tilde{m}_{n,(k,\lambda_{n,k}),2}(x)$$

with

$$\tilde{m}_{n,(k,\lambda_{n,k}),1}(x) = \sum_{j=0}^{k-1} \beta_j x^j \text{ and } |\tilde{m}_{n,(k,\lambda_{n,k}),2}(x)| \leq J_k(\tilde{m}_{n,(k,\lambda_{n,k})}) \quad (x \in [0,1])$$

(see, e.g., Oden and Reddy (1976)).

It follows from (5) that

$$\begin{aligned}
\|\tilde{m}_{n,(k,\lambda_{n,k})}\|_n^2 &\leq 2\left(\frac{1}{n}\sum_{i=1}^n |Y_i - \tilde{m}_{n,(k,\lambda_{n,k})}(X_i)|^2 + \lambda_{n,k}^2 J_k^2(\tilde{m}_{n,(k,\lambda_{n,k})})\right) + 2\frac{1}{n}\sum_{i=1}^n |Y_i|^2 \\
&\leq 2\left(\frac{1}{n}\sum_{i=1}^n |Y_i - 0|^2 + \lambda_{n,k}^2 \cdot 0\right) + 2\frac{1}{n}\sum_{i=1}^n |Y_i|^2 \leq 4L^2.
\end{aligned}$$

Hence

$$\begin{aligned}
&\|\tilde{m}_{n,(k,\lambda_{n,k}),1}\|_n^2 \leq 2\|\tilde{m}_{n,(k,\lambda_{n,k})}\|_n^2 + 2\|\tilde{m}_{n,(k,\lambda_{n,k}),2}\|_n^2 \\
&\leq 8L^2 + 2J_k^2(\tilde{m}_{n,(k,\lambda_{n,k})}) \leq 8L^2(1+J_k(\tilde{m}_{n,(k,\lambda_{n,k})}))^2.
\end{aligned}$$

On the other hand

$$\|\tilde{m}_{n,(k,\lambda_{n,k}),1}\|_n^2 = \beta^T \left(\frac{1}{n}\sum_{i=1}^n X_i^{j+l}\right)_{0 \leq j,l \leq k-1} \beta,$$

where $\beta^T = (\beta_0, \ldots, \beta_{k-1})$.

The matrix $(EX^{j+l})_{0 \leq j,l \leq k-1}$ is positive definite. Indeed, if $\gamma = (\gamma_0, \ldots, \gamma_{k-1})^T \in \mathbb{R}^k$ such that

$$\gamma^T (EX^{j+l})_{0 \leq j,l \leq k-1} \gamma = 0,$$

then

$$E\left|\sum_{i=0}^{k-1}\gamma_i X^i\right|^2 = \gamma^T (EX^{j+l})_{0\le j,l\le k-1}\gamma = 0,$$

which implies

$$\sum_{i=0}^{k-1}\gamma_i x^i \quad \text{for } P_X \text{almost all } x\in[0,1].$$

Because of $card(supp(X)) > k$ this in turn implies $\gamma = 0$.

Choose $c_9 > 0$ such that

$$\gamma^T (EX^{j+l})_{0\le j,l\le k-1}\gamma \ge \frac{1}{c_9}\gamma^T\gamma$$

for all $\gamma \in \mathbb{R}^k$. By Hoeffding inequality (see e.g. Devroye, Györfi and Lugosi (1996), Theorem 8.1)

$$P\left[\sup_{0\le j,l\le k-1}\left|\frac{1}{n}\sum_{i=1}^n X_i^{j+l} - EX^{j+l}\right| > \frac{1}{2\,k\,c_9}\right] \to 0 \quad (n\to\infty).$$

Thus outside of an event whose probability tends to zero ($n\to\infty$)

$$\begin{aligned}
&\beta^T\left(\frac{1}{n}\sum_{i=1}^n X_i^{j+l}\right)_{0\le j,l\le k-1}\beta \\
&= \beta^T\left(EX^{j+l}\right)_{0\le j,l\le k-1}\beta + \beta^T\left(\frac{1}{n}\sum_{i=1}^n X_i^{j+l} - EX^{j+l}\right)_{0\le j,l\le k-1}\beta \\
&\ge \frac{1}{c_9}\beta^T\beta - \frac{1}{2kc_9}k\beta^T\beta = \frac{1}{2c_9}\beta^T\beta.
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\tilde m_{n,(k,\lambda_{n,k}),1}\|_\infty &\le k\left(\beta^T\beta\right)^{\frac12} \le k\left(2c_9\beta^T\left(\frac{1}{n}\sum_{i=1}^n X_i^{j+l}\right)_{0\le j,l\le k-1}\beta\right)^{\frac12} = k\left(2c_9\|\tilde m_{n,(k,\lambda_{n,k}),1}\|_n^2\right)^{\frac12} \\
&\le 4kc_9L(1+J_k(\tilde m_{n,(k,\lambda_{n,k})}))
\end{aligned}$$

outside of an event whose probability tends to zero ($n\to\infty$), which implies

$$\|\tilde m_{n,(k,\lambda_{n,k})}\|_\infty \le \|\tilde m_{n,(k,\lambda_{n,k}),1}\|_\infty + \|\tilde m_{n,(k,\lambda_{n,k}),2}\|_\infty \le 5kCL(1+J_k(\tilde m_{n,(k,\lambda_{n,k})}))$$

outside of an event whose probability tends to zero ($n\to\infty$). This proves (22).

Next we bound $J_k(\tilde m_{n,(k,\lambda_{n,k})})$. Similarly to (14) one obtains

$$\begin{aligned}
&\|\tilde m_{n,(k,\lambda_{n,k})} - m\|_n^2 + \lambda_{n,k}^2(1+J_k^2(\tilde m_{n,(k,\lambda_{n,k})})) \\
&\le \lambda_{n,k}^2(1+J_k^2(m)) + \frac{2}{n}\sum_{i=1}^n(\tilde m_{n,(k,\lambda_{n,k})}(X_i) - m(X_i))(Y_i - m(X_i)). \qquad (23)
\end{aligned}$$

Using (22), Lemma 6 and $\|\tilde{m}_{n,(k,\lambda_{n,k})} - m\|_n^2 \leq 10L^2$ one obtains that the probability of the event

$$\left[\frac{2}{n}\sum_{i=1}^{n}(\tilde{m}_{n,(k,\lambda_{n,k})}(X_i) - m(X_i))(Y_i - m(X_i)) \right.$$
$$\left. > \frac{\ln(n)}{4n^{\frac{1}{2}}}\|\tilde{m}_{n,(k,\lambda_{n,k})} - m\|_n^{1-\frac{1}{2k}}(1 + J_k(\tilde{m}_{n,(k,\lambda_{n,k})}))^{\frac{1}{2k}}\right]$$

converges to zero ($n \to \infty$). Outside of this event (23) implies

$$\|\tilde{m}_{n,(k,\lambda_{n,k})} - m\|_n^2 + \lambda_{n,k}^2(1 + J_k^2(\tilde{m}_{n,(k,\lambda_{n,k})})) \leq \ln^4(n)n^{-\frac{2k}{2k+1}}.$$

Applying now again Lemma 6, but this time with $\sigma = \sigma_n = \ln^2(n)n^{-\frac{k}{2k+1}}$ and $t = 2^k t_0$ implies as in the proof of Lemma 3 that

$$\|\tilde{m}_{n,(k,\lambda_{n,k})} - m\|_n^2 + \lambda_{n,k}^2(1 + J_k^2(\tilde{m}_{n,(k,\lambda_{n,k})})) \leq c_1 0(J_k(m))^{1/(2k+1)}n^{-\frac{2k}{2k+1}}$$

outside of an event whose probability tends to zero ($n \to \infty$). This together with (22) yields the assertion. □

Acknowledgement

The author wishes to thank Y. Baraud and H. Walk for various stimulating discussions, and two referees and an associate editor for many helpful comments.

References

[1] Baraud, Y. (1997). Model selection for regression on random design. Submitted for publication.

[2] Barron, A. R. , Birgé, L. and Massart,P. (1999). Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields* **113** 301–413.

[3] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

[4] Birman, M.S. and Solomjak, M.Z. (1967). Piecewise polynomial approximations of functions of the class W_p^α. *Mathematics of the USSR-Sbornik*, Translation of Mat. Sbornik, Vol. 73, pp. 331-355. Vol. 2, pp. 295-316.

[5] Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **6** 2384–2398.

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

[7] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.* **22** 1371–1385.

[8] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, Berlin.

[9] Devroye, L. P. and Wagner, T. J. (1980). Distribution–free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

[10] Efroimovich, S. and Pinsker, M. (1984). Learning algorithm for nonparametric filtering. *Automat. Remote Control* **11** 1434–1440.

[11] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

[12] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman & Hall, London.

[13] Györfi, L. (1981). The rate of convergence of $k_n - NN$ regression estimates and classification rules. *IEEE Trans. Information Theory* **27** 362–364.

[14] Györfi, L., Kohler, M. and Walk, H. (1998). Weak and strong universal consistency of semi-recursive partitioning and kernel regression estimates. *Statistics & Decisions* **16** 1–18.

[15] Györfi, L. and Walk, H. (1996). On the strong universal consistency of a series type regression estimate. *Mathematical Methods of Statistics* **5** 332–342.

[16] Györfi, L. and Walk, H. (1997). On the strong universal consistency of a recursive regression estimate by Pál Révész. *Statistics and Probability Letters* **31** 177–183.

[17] Härdle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

[18] Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *Ann. Statist.* **26** 242–272.

[19] John, G. H. (1997). *Enhancements to the Data Mining Process.* PhD–thesis, Satnford University.

[20] Kohler, M. (1997). On the universal consistency of a least squares spline regression estimator. *Math. Methods of Statistics* **6** 349–364.

[21] Kohler, M. (1999). Universally consistent regression function estimation using hierarchical B-splines. *Journal of Multivariate Analysis* **67** 138–164.

[22] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. To appear in *Journal of statistical Planning and Inference.*

[23] Kohler, M. and Krzyżak, A. (2000). Nonparametric regression via penalized least squares. Submitted for publication.

[24] Kohler, M., Krzyżak, A. and Schäfer, D. (2000). Application of structural risk minimization to multivariate smoothing spline estimates. Submitted for publication.

[25] Krahl, D., Windheuser, U. and Zick, F.-K. (1998). *Data Mining. Einsatz in der Praxis.* Addison–Wesley, Bonn.

[26] Krzyżak, A. (1986). The rates of convergence of kernel regression estimates and classification rules. *IEEE Trans. Information Theory* **32** 668–679.

[27] Krzyżak, A. and Linder, T. (1996). Radial basis function networks and complexity regularization in function learning. *IEEE Transactions on Nueral Networks* **9**, 247–256.

[28] Mammen, E. and van de Geer, S. (1997). Penalized quasi–likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014–1035.

[29] Mattison, R. (1997). *Data Warehousing and Data Mining for Telecommunications.* Artech House.

[30] Oden, J. T. and Reddy, J. N. (1976). *An Introduction to the Mathematical Theory of Finite Elements.* Wiley, New York.

[31] Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615.

[32] Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

[33] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

[34] Tsybakov, A. B. (1998). Pointwise and sup–norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** 2420–2469.

[35] van de Geer, S. (1988). *Regression Analysis and Empirical Processes.* CWI-tract 45, Centre for Mathematics and Computer Science, Amsterdam.

[36] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **25** 1014–1035.

[37] van de Geer, S. (2000). *Empirical Process in M-estimation.* Cambridge University Press.

[38] Vapnik, V. N. and Chervonenkis, A. Y. (1974). *Theory of Pattern Recognition* (in Russisch), Nauka, Moskau.

[39] Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley & Sons, New York.

[40] Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

[41] Walk, H. (1997). Strong universal consistency of kernel and partitioning regression estimates. Preprint 97–1, Mathematisches Institut A, Universität Stuttgart.

MATHEMATISCHES INSTITUT A
UNIVERSITÄT STUTTGART
PFAFFENWALDRING 57

D-70569 STUTTGART
GERMANY
E-MAIL: kohler@mathematik.uni-stuttgart.de