Optimal global rates of convergence for nonparametric regression with unbounded data *

Michael Kohler[1], Adam Krzyżak[2*] and Harro Walk[3]

[1] Department of Mathematics, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany, email: kohler@math.uni-sb.de

[2] Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca

[3] Department of Mathematics, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany, email: walk@mathematik.uni-stuttgart.de

June 28, 2005

Summary

Estimation of regression functions from independent and identically distributed data is considered. The L_2 error with integration with respect to the design measure is used as an error criterion. Usually in the analysis of the rate of convergence of estimates a boundedness assumption on X is made besides smoothness assumptions on the regression function and moment conditions on Y. In this article we consider the kernel estimate and show that by replacing the boundedness assumption on X by a proper moment condition the same (optimal) rate of convergence can be shown as for bounded data. This answers

*Research supported by the Alexander von Humboldt Foundation.

Running title: *Optimal global rates of convergence*

Please send correspondence and proofs to: Adam Krzyżak, Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8.

Question 1 in Stone (1982).

Key words and phrases: Regression, kernel estimate, rate of convergence.

1 Introduction

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 $\mathbf{E}\{Y^2\} < \infty$. In regression analysis we want to estimate Y after having observed X, i.e. we want to determine a function f with $f(X)$ "close" to Y. If "closeness" is measured by the mean squared error, then one wants to find a function f^* such that

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

Let $m(x) := \mathbf{E}\{Y|X=x\}$ be the regression function and denote the distribution of X by μ. The well-known relation which holds for each measurable function f

$$\mathbf{E}\{|f(X)-Y|^2\} = \mathbf{E}\{|m(X)-Y|^2\} + \int |\ f(x)-m(x)\ |^2\ \mu(dx) \tag{2}$$

implies that m is the solution of the minimization problem (1), $\mathbf{E}\{|m(X)-Y|^2\}$ is the minimum of (2) and for an arbitrary f, the L_2 error $\int |\ f(x)-m(x)\ |^2\ \mu(dx)$ is the difference between $\mathbf{E}\{|f(X)-Y|^2\}$ and $\mathbf{E}\{|m(X)-Y|^2\}$.

In the regression estimation problem the distribution of (X,Y) (and consequently m) is unknown. Given a sequence $\mathcal{D}_n = \{(X_1,Y_1),...,(X_n,Y_n)\}$ of independent observations of (X,Y), our goal is to construct an estimate $m_n(x)$=$m_n(x,\mathcal{D}_n)$ of $m(x)$ such that the L_2 error $\int |m_n(x)-m(x)|^2\mu(dx)$ is small.

It is well-known that there exist universally consistent estimates, i.e., estimates m_n with the property

$$\mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx) \to 0 \quad (n\to\infty)$$

for all distributions of (X,Y) with $\mathbf{E}\{Y^2\} < \infty$. This was first shown by Stone (1977) for the nearest neighbor estimate and later extended by numerous papers, see, e.g., Devroye et al. (1994), Greblicki, Krzyżak and Pawlak (1984), Györfi, Kohler and Walk (1998), Györfi and Walk (1996, 1997), Kohler (1999, 2002), Kohler and Krzyżak (2001), Lugosi and Zeger (1995), Nobel (1996) and Walk (2002). See also Györfi et al. (2002) and the literature cited therein.

Unfortunately, there do not exist estimates for which the expected L_2 error converges to zero with some nontrivial rate for all distributions of (X,Y), cf. Cover (1968) and Devroye (1982), or Chapter 3 in Györfi et al. (2002). So in order to derive nontrivial rates of convergence, one has to restrict the class of distributions, in particular by assuming smoothness of the regression function.

Let $\mathcal{D}$ be a class of distributions of (X,Y). In the classical minimax theory, one considers the maximal error of an estimate within the class $\mathcal{D}$ of distributions of (X,Y) and tries to construct estimates for which this maximal error is minimal, i.e., one tries to construct estimates m_n such that

$$\sup_{(X,Y)\in\mathcal{D}} \mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx) \approx \inf_{\hat{m}_n} \sup_{(X,Y)\in\mathcal{D}} \mathbf{E}\int |\hat{m}_n(x)-m(x)|^2\mu(dx). \qquad (3)$$

Here the infimum is taken over all estimates. Then the optimal minimax rate of convergence is defined as the rate of convergence at which the right-hand side of (3) converges to zero.

In Stone (1982) the optimal minimax rate of convergence for a class of distributions of (X,Y) was determined, where the regression functions is (p,C)–smooth according to the following definition.

Definition 1 *Let $p = k+\gamma$ for some $k \in \mathbb{N}_0$ and some $0 < \gamma \leq 1$. Let $C > 0$. A function $m : \mathbb{R}^d \to \mathbb{R}$ is called (p,C)–smooth if for all $k_1,\ldots,k_d \in \mathbb{N}_0$ with $k = k_1+\ldots+k_d$ the* *partial derivatives*

$$\frac{\partial^k m}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}$$

of m exist and satisfy

$$\left| \frac{\partial^k m}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(x) - \frac{\partial^k m}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(z) \right| \leq C \cdot \|x - z\|^\gamma \quad (x, z \in \mathbb{R}^d).$$

Let $\mathcal{D}^{(p,C)}$ be the class of all distributions of (X, Y) where X takes on only values in $[0, 1]^d$, X has a density with respect to the Lebesgue measure which is bounded away from zero and infinity by some constants c_1 and c_2, $\mathrm{Var}\{Y|X = x\}$ is bounded and m is (p, C)–smooth. It follows from Stone (1982), that for this class of distributions

$$\liminf_{n\to\infty} \inf_{\hat{m}_n} \sup_{(X,Y)\in\mathcal{D}^{(p,C)}} \frac{\mathbf{E} \int |\hat{m}_n(x) - m(x)|^2 dx}{C^{2d/(2p+d)} n^{-2p/(2p+d)}} > 0 \tag{4}$$

(cf. Theorem 3.2 in Györfi et al. (2002)), and that a suitably defined local polynomial kernel estimate satisfies

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}^{(p,C)}} \frac{\mathbf{E} \int |m_n(x) - m(x)|^2 dx}{C^{2d/(2p+d)} n^{-2p/(2p+d)}} < \infty. \tag{5}$$

Actually, both bounds have been proven in Stone (1982) not for the expected L_2 error but instead in probability, which is a stronger result for the lower bound and a weaker result for the upper bound. The (slightly) stronger upper bound (5) holds at least for $p \leq 1$, cf. Theorem 5.2 in Györfi et al. (2002).

Since X has a density with respect to the Lebesgue-Borel measure, which is bounded away from zero and infinity, for $(X, Y) \in \mathcal{D}^{(p,C)}$, the same result also holds for the L_2 error with integration with respect to the distribution μ of X, which is the error criterion considered in this paper. But in this case one can relax the assumption on X: It follows from Theorems 4.3, 5.2 and 6.2 in Györfi et al. (2002) (cf., Spiegelman and Sacks (1980), Györfi (1981), and Kulkarni and Posner (1995)) that suitably defined partitioning, kernel and nearest neighbor estimates satisfy

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq const \cdot C^{2d/(2p+d)} n^{-2p/(2p+d)} \tag{6}$$

provided X takes on values only in $[0,1]^d$, $\mathrm{Var}\{Y|X=x\}$ is bounded and m is (p,C)–smooth for some $p \leq 1$. In case of the nearest neighbor estimates one needs the additional condition $d > 2p$, for the other two estimates the result holds in any dimension. For smoother regression functions (i.e., (p,C)–smooth regression functions with $p > 1$), it was shown in Kohler (2000) that in case of bounded X and Y suitably defined least squares estimates also achieve the above optimal rate of convergence, regardless whether the distribution of X has a density with respect to the Lebesgue-Borel measure or not.

If one compares these rate of convergence results with the universal consistency results cited above, then one gets the impression that it should be possible to replace the boundedness assumption on X by weaker conditions like existence of some moments of $\|X\|$. This conjecture was already formulated in Stone (1982) as Question 1. In this paper we show that this conjecture is indeed true. In particular we show that for m bounded and (p,C)–smooth with $0 < p \leq 1$, $\mathrm{Var}\{Y|X=x\}$ bounded and $\mathbf{E}\|X\|^\beta < \infty$ for some $\beta > 2p$, a suitably defined kernel estimate satisfies (6). Furthermore we show that if we replace the moment condition by $\mathbf{E}\|X\|^\beta < \infty$ for some $0 < \beta < 2p$, there exists no estimate for which (6) holds for all such distributions. Similar results for partitioning and nearest neighbor regression estimates have been derived in Kohler, Krzyżak and Walk (2005).

Throughout the paper we will use the following notations: $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ are the sets of natural, real and nonnegative real numbers, respectively. The euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$. Set $S_{z,r} = \{x \in \mathbb{R}^d : ||x-z|| < r\}, z \in \mathbb{R}^d, r > 0$. 1_D denotes the indicator function of a set D. For $x \in \mathbb{R}$, $\lceil x \rceil$ is the least integer greater than or equal to x, and $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Throughout the proofs c_1, c_2, $\ldots$ denote suitable constants.

The main results are stated in Section 2 and proven in Sections 3 and 4.

2 Main results

Let m_n be the kernel estimate defined by

$$m_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n(x)}\right) \cdot Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n(x)}\right)}$$

(with $0/0 := 0$) where the kernel $K : \mathbb{R}^d \to \mathbb{R}_+$ satisfies

$$c_1 1_{S_{0,1}}(x) \le K(x) \le c_2 1_{S_{0,1}}(x) \quad (x \in \mathbb{R}^d) \tag{7}$$

for some constants $0 < c_1 \le c_2 < \infty$, and the bandwidth $h_n(x)$ depends on x. We choose $h_n(x)$ such that it will increase with $\|x\|$. More precisely, we set

$$h_n(x) = \begin{cases} h_n \cdot (1 + \|x\|)^{\beta/(2p)} & \text{if} \quad \|x\| \le \lfloor n^{\frac{2p}{(2p+d)\cdot\beta}} \rfloor, \\ \infty & \text{if} \quad \|x\| > \lfloor n^{\frac{2p}{(2p+d)\cdot\beta}} \rfloor, \end{cases} \tag{8}$$

where p and β are defined below and $h_n = C^{-2/(2p+d)} n^{-1/(2p+d)}$.

Theorem 1 *Assume that the distribution of (X, Y) satisfies the following four conditions:*

(A1) $m(x) = \mathbf{E}\{Y|X = x\}$ is bounded in absolute value by some constant $L \ge 1$.

(A2) $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)–smooth for some $0 < p \le 1$, $C \ge 1$.

(A3) The conditional variance of Y satisfies

$$\sup_{x \in \mathbb{R}^d} \mathrm{Var}\{Y|X = x\} \le \sigma_0^2$$

for some $\sigma_0 \ge 0$.

(A4) There exists a constant $\beta > 2p$ such that

$$\mathbf{E}\|X\|^\beta \le M$$

for some constant $M \ge 0$.

Define the kernel estimate m_n as above with kernel K satisfying (7) and with bandwidth $h_n(x)$ defined by (8). Then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq c_3 \cdot C^{2d/(2p+d)} \cdot n^{-2p/(2p+d)}$$

where c_3 depends only on d, p, β, L, M, σ_0, c_1 and c_2.

Theorem 1 implies the following result concerning minimax rate of convergence: Let $0 < p \leq 1, C \geq 1, \beta > 2p, L > 0$, $M \geq 0$ and $\sigma_0 > 0$. Let $\mathcal{D}(p, C, \beta, L, M, \sigma_0)$ be the class of all distributions of (X, Y) which satisfy $(A1)$, $(A2)$, $(A3)$ and $(A4)$ for these values of p, C, β, L, M and σ_0. Then

$$\sup_{(X,Y) \in \mathcal{D}(p,C,\beta,L,M,\sigma_0)} \mathbf{E} \int |m_n(x) - m(x)| \mu(dx) \leq c_3 \cdot C^{2d/(2p+d)} n^{-2p/(2p+d)}.$$

It follows from (4) that the above rate is the optimal minimax rate of convergence for the class $\mathcal{D}(p, C, \beta, L, M, \sigma_0)$ of distributions of (X, Y). Next we present a lower bound on the rate of convergence, which implies that one needs a condition on the tails of $||X||$ in order to get the above rate of convergence result.

Theorem 2 *Let $p > 0, C > 0$ and $\beta < 2p$. Then we have for M sufficiently large*

$$\liminf_{n \to \infty} n^{2p/(2p+d)} \inf_{m_n} \sup_{(X,Y) \in \mathcal{D}(p,C,\beta,C,M,1)} \mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) = \infty.$$

Remark. Let the kernel estimate m_n be defined as above with kernel K satisfying (7). By rescaling of the kernel we can assume w.l.o.g. $c_1 \geq 1$. In this case the estimate can be also defined via

$$m_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n(x)}\right) \cdot Y_i}{\max\left\{1, \sum_{i=1}^n K\left(\frac{x - X_i}{h_n(x)}\right)\right\}}. \tag{9}$$

This definition of the kernel estimate was used in Spiegelman and Sachs (1980) in connection with the anlalysis of the rate of convergence of the estimate for bounded X. By combining ideas presented there with the proof of Theorem 1, it can be shown that Theorem 1 also holds for the estimate (9) even if the kernel does not satisfy (7) but is instead bounded, has compact support and satisfies

$$K(x) \geq c^* 1_{S_{0,\delta}}(x) \quad (x \in \mathbb{R}^d)$$

for some $c^* \in \mathbb{R}$ and some $\delta > 0$.

3 Proof of Theorem 1

We have

$$\mathbf{E}\{(m_n(x) - m(x))^2 | X_1, \ldots, X_n\} = \mathbf{E}\{(m_n(x) - \hat{m}_n(x))^2 | X_1, \ldots, X_n\} + (\hat{m}_n(x) - m(x))^2$$

where

$$\hat{m}_n(x) = \mathbf{E}\{m_n(x) | X_1, \ldots, X_n\} = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n(x)}\right) \cdot m(X_i)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n(x)}\right)}.$$

Now,

$$\begin{aligned} \mathbf{E}\{(m_n(x) - \hat{m}_n(x))^2 | X_1, \ldots, X_n\} &\leq \left(\frac{c_2}{c_1}\right)^2 \cdot \frac{\sum_{i=1}^n \mathrm{Var}\{Y_i | X_i\} 1_{S_{x,h_n(x)}}(X_i)}{(\sum_{i=1}^n 1_{S_{x,h_n(x)}}(X_i))^2} \\ &\leq c_4 \sigma_0^2 \frac{1}{B(x)} \cdot 1_{\{B(x) > 0\}} \end{aligned}$$

where

$$B(x) = \sum_{i=1}^n 1_{S_{x,h_n(x)}}(X_i)$$

is binomially distributed with parameters n and $q = \mu(S_{x,h_n(x)})$.

Furthermore, by Jensen's inequality and boundedness and (p, C)–smoothness of m we get

$$\begin{aligned} &(\hat{m}_n(x) - m(x))^2 \\ &\leq \left(\frac{c_2}{c_1}\right)^2 \cdot \left(\frac{\sum_{i=1}^n |m(X_i) - m(x)| \cdot 1_{S_{x,h_n(x)}}(X_i)}{\sum_{i=1}^n 1_{S_{x,h_n(x)}}(X_i)}\right)^2 \cdot 1_{\{B(x)>0\}} + m(x)^2 1_{\{B(x)=0\}} \\ &\leq c_4 \frac{\sum_{i=1}^n (m(X_i) - m(x))^2 \cdot 1_{S_{x,h_n(x)}}(X_i)}{\sum_{i=1}^n 1_{S_{x,h_n(x)}}(X_i)} \cdot 1_{\{B(x)>0\}} + m(x)^2 1_{\{B(x)=0\}} \\ &\leq c_4 \min\left\{C^2 |h_n(x)|^{2p}, 4L^2\right\} + L^2 1_{\{B(x)=0\}}. \end{aligned}$$

Summarizing the above results we get

$$\begin{aligned}
&\mathbf{E}\{(m_n(x)-m(x))^2\}\\
&\le c_4\sigma_0^2\mathbf{E}\left\{\frac{1}{B(x)}\cdot 1_{\{B(x)>0\}}\right\}+c_4\min\left\{C^2|h_n(x)|^{2p},4L^2\right\}+L^2\mathbf{P}\{B(x)=0\}.
\end{aligned}$$

Using

$$\begin{aligned}
\mathbf{E}\left\{\frac{1}{B(x)}\cdot 1_{\{B(x)>0\}}\right\} &= \sum_{k=1}^{n}\frac{1}{k}\cdot\binom{n}{k}q^k(1-q)^{n-k}\\
&\le \sum_{k=1}^{n}\frac{2}{k+1}\cdot\binom{n}{k}q^k(1-q)^{n-k}\\
&= \frac{2}{(n+1)\cdot q}\sum_{k=1}^{n}\binom{n+1}{k+1}q^{k+1}(1-q)^{n+1-(k+1)}\\
&= \frac{2}{(n+1)\cdot q}\cdot\left(1-(1-q)^{n+1}-(n+1)\cdot q\cdot(1-q)^n\right)\\
&\le \frac{2}{(n+1)\cdot\mu(S_{x,h_n(x)})}-2\cdot\mathbf{P}\{B(x)=0\}
\end{aligned}$$

we get

$$\mathbf{E}\{(m_n(x)-m(x))^2\}\le\max\{c_4\sigma_0^2,L^2\}\cdot\frac{2}{(n+1)\cdot\mu(S_{x,h_n(x)})}+c_4\min\left\{C^2|h_n(x)|^{2p},4L^2\right\}.$$

Hence

$$\begin{aligned}
&\mathbf{E}\int|m_n(x)-m(x)|^2\mu(dx)\\
&\le\frac{2\cdot\max\{c_4\sigma_0^2,L^2\}}{n+1}\cdot\int\frac{1}{\mu(S_{x,h_n(x)})}\,dx+c_4\int\min\left\{C^2|h_n(x)|^{2p},4L^2\right\}\,\mu(dx). \quad (10)
\end{aligned}$$

Next we bound the first integral on the right-hand side of (10). Because of $\mu(S_{x,h_n(x)})=\mu(\mathbb{R}^d)=1$ for $\|x\|\ge\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor$ we have

$$\begin{aligned}
&\int\frac{1}{\mu(S_{x,h_n(x)})}\,\mu(dx)\\
&\le\int_{S_{0,2}}\frac{1}{\mu(S_{x,h_n})}\,\mu(dx)+\sum_{j=3}^{\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor}\int_{S_{0,j}\setminus S_{0,j-1}}\frac{1}{\mu(S_{x,h_n j^{\beta/(2p)}})}\,\mu(dx)\\
&\quad+\int_{\mathbb{R}^d\setminus S_{0,n^{2p/((2p+d)\cdot\beta)}-1}}1\,\mu(dx).
\end{aligned}$$

Fix $3 \le j \le \lfloor n^{2p/((2p+d)\cdot\beta)} \rfloor$ and set $r = h_n j^{\beta/(2p)}$. Then

$$r \le C^{-2/(2p+d)} n^{-1/(2p+d)} n^{1/(2p+d)} \le 1.$$

Choose $z_1, \dots, z_l$ such that the balls

$$S_{z_1,r/4}, \dots, S_{z_l,r/4} \tag{11}$$

are contained in $S_{0,j+1} \setminus S_{0,j-2}$, do not overlap and such that the number l of these balls is maximal. Then l can be bounded by

$$l \le \frac{\mathrm{Vol}(S_{0,j+1}) - \mathrm{Vol}(S_{0,j-2})}{\mathrm{Vol}(S_{0,r/4})} = \frac{(j+1)^d - (j-2)^d}{(r/4)^d} = c_5 \cdot j^{d-1-\beta\cdot d/(2p)} \cdot \frac{1}{h_n^d}$$

and $S_{z_1,r/2}, \dots, S_{z_l,r/2}$ cover $S_{0,j} \setminus S_{0,j-1}$ (because if any point $z \in S_{0,j} \setminus S_{0,j-1}$ is in none of those balls, then $S_{z,r/4}$ does not overlap with any of the balls (11) and is contained in $S_{0,j+1} \setminus S_{0,j-2}$). From this we can conclude

$$\begin{aligned}
\int_{S_{0,j} \setminus S_{0,j-1}} \frac{1}{\mu(S_{x,h_n j^{\beta/(2p)}})}\, \mu(dx) &\le \sum_{k=1}^{l} \int_{S_{z_k,r/2}} \frac{1}{\mu(S_{x,r})}\, \mu(dx) \\
&\le \sum_{k=1}^{l} \int_{S_{z_k,r/2}} \frac{1}{\mu(S_{z_k,r/2})}\, \mu(dx) = l
\end{aligned}$$

since for $x \in S_{z_k,r/2}$ we have $S_{z_k,r/2} \subseteq S_{x,r}$. Applying a similar argument to $\int_{S_{0,2}} \frac{1}{\mu(S_{x,h_n})}\, dx$ we get

$$\int \frac{1}{\mu(S_{x,h_n(x)})}\, \mu(dx) \le c_5 \cdot \frac{1}{h_n^d} \cdot \sum_{j=1}^{\infty} \left(\frac{1}{j}\right)^{1+d\cdot(\beta/(2p)-1)} + \mu\left(\mathbb{R}^d \setminus S_{0,n^{2p/((2p+d)\cdot\beta)}-1}\right).$$

By (A4) we have

$$\begin{aligned}
\sum_{j=1}^{\infty} j^{\beta} \mu\left(S_{0,j} \setminus S_{0,j-1}\right) &\le 1 + 2^{\beta} \sum_{j=1}^{\infty} (j-1)^{\beta} \mu\left(S_{0,j} \setminus S_{0,j-1}\right) \\
&\le 1 + 2^{\beta} \sum_{j=1}^{\infty} \int_{S_{0,j} \setminus S_{0,j-1}} \|x\|^{\beta} \mu(dx) \\
&\le 1 + 2^{\beta} \mathbf{E}\{\|X\|^{\beta}\} \\
&\le 1 + 2^{\beta} M < \infty,
\end{aligned} \tag{12}$$

which implies

$$N^{\beta}\mu\left(\mathrm{I\!R}^d \setminus S_{0,N}\right) \leq \sum_{j=N+1}^{\infty} j^{\beta}\mu\left(S_{0,j} \setminus S_{0,j-1}\right) \to 0 \quad (N \to \infty). \tag{13}$$

From this we can conclude

$$\int \frac{1}{\mu(S_{x,h_n(x)})}\,\mu(dx) \quad \leq \quad c_6 \cdot \left(\frac{1}{h_n^d} + n^{-2p/(2p+d)}\right)$$

for some constant c_6 depending on d, p, β and M.

Concerning the second term on the right-hand side of (10) we have

$$\begin{aligned}
&\int \min\left\{C^2|h_n(x)|^{2p}, 4L^2\right\}\,\mu(dx)\\
&\leq C^2h_n^{2p} + \sum_{j=1}^{\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor} \int_{S_{0,j}\setminus S_{0,j-1}} C^2h_n^{2p}(j+1)^{\beta}\mu(dx)\\
&\quad +4L^2\mu\left(\mathrm{I\!R}^d \setminus S_{0,n^{2p/((2p+d)\cdot\beta)}-1}\right)\\
&\leq C^2h_n^{2p}\left(1+\sum_{j=1}^{\infty}(j+1)^{\beta}\mu\left(S_{0,j}\setminus S_{0,j-1}\right)\right) + 4L^2\mu\left(\mathrm{I\!R}^d \setminus S_{0,n^{2p/((2p+d)\cdot\beta)}-1}\right)\\
&\leq c_7 \cdot \left(C^2h_n^{2p} + n^{-2p/(2p+d)}\right),
\end{aligned}$$

where the last inequality follows from (12) and (13).

Summarizing the above results we get

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2\mu(dx)\\
&\leq \frac{2\cdot\max\{c_4\sigma_0^2, L^2\}}{n+1}\cdot c_6 \cdot \left(\frac{1}{h_n^d} + n^{-2p/(2p+d)}\right) + c_4\cdot c_7\cdot\left(C^2h_n^{2p} + n^{-2p/(2p+d)}\right)\\
&\leq c_8 \cdot C^{2d/(2p+d)}n^{-2p/(2p+d)}.
\end{aligned}$$

□

4 Proof of Theorem 2.

First we define a subclass of distributions of (X,Y) contained in $\mathcal{D}(p,C,\beta,C,1)$. Assume that X has a density

$$f(x) = c_9 \cdot \frac{1}{(1+||x||)^{2p+d}} \quad (x \in \mathrm{I\!R}^d),$$

thus

$$\mathbf{E}||X||^{\beta} \leq c_9 \int \frac{1}{(1+\|x\|)^{d+(2p-\beta)}}\, dx =: M < \infty.$$

Set $g(x) = C \cdot \overline{g}(x)$ for some function $\overline{g} : \mathbb{R}^d \to \mathbb{R}$ such that $\overline{g}(x) = 0$ for $x \notin [-1/2, 1/2]^d$, $\overline{g}(x) \neq 0$, $\overline{g}(x)$ bounded in absolute value by 1, and $\overline{g}(x)$ $(p, 2^{\gamma-1})$-smooth, where $p = k+\gamma$ for some $k \in \mathbb{N}_0, 0 < \gamma \leq 1$. The class of regression functions will be indexed by a vector

$$c = (c^1_{n,1}, \ldots, c^1_{n,N_{n,1}}, \ldots, c^{j_{max}(n)}_{n,1}, \ldots, c^{j_{max}(n)}_{n,N_{n,j_{max}(n)}})$$

of $+1$ or -1 components, where $j_{max}(n)$ and $N_{n,1}, \ldots, N_{n,j_{max}(n)}$ are defined below. Denote the set of all such vectors by $\mathcal{C}_n$. For $c \in \mathcal{C}_n$ define the function

$$m^{(c)}(x) = \sum_{j=1}^{j_{max}(n)} \sum_{k=1}^{N_{n,j}} c^j_{n,k} g^j_{n,k}(x),$$

where

$$g^j_{n,k}(x) = M^{-p}_{n,j} g(M_{n,j}(x - a^j_{n,k})).$$

Set

$$M_{n,j} = \left\lceil C^{2/(2p+d)} \cdot n^{1/(2p+d)}/j \right\rceil,$$

partition $[-j, j]^d$ into $(2j)^d M^d_{n,j}$ uniform cubes $A^j_{n,k}$ of side length $h_{n,j} = 1/M_{n,j}$ and let $a^j_{n,1}, \ldots, a^j_{n,N_{n,j}}$ be the centers of those cubes $A \in \{A^j_{n,k} | k = 1, \ldots, (2j)^d M^d_{n,j}\}$ which satisfy $A \subseteq [-j, j]^d \setminus [-(j-1), j-1]^d$. (The last condition ensures that the supports of the $g^j_{n,k}$'s are disjoint.) W.l.o.g. assume that $a^j_{n,k}$ is the center of $A^j_{n,k}$. Here $N_{n,j}$ is the number of those sets $A^j_{n,k}$ which are contained in $[-j, j]^d \setminus [-(j-1), j-1]^d$. In case of

$$h_{n,j} = \frac{1}{M_{n,j}} \leq \frac{j}{C^{2/(2p+d)} n^{1/(2p+d)}} \leq \frac{1}{2},$$

(which is implied by $j < \lceil \frac{1}{2} C^{2/(2p+d)} n^{1/(2p+d)} \rceil = j_{max}(n)$) all cubes $A^j_{n,k}$ which are not contained in $[-(j-1/2), j-1/2]^d$ have this property. There are at most

$$\frac{(2j-1)^d}{h^d_{n,j}} = (2j-1)^d M^d_{n,j}$$

cubes in $[-(j-1/2), j-1/2]^d$, thus

$$N_{n,j} \geq \frac{(2j)^d}{h_{n,j}^d} - \frac{(2j-1)^d}{h_{n,j}^d} = c_{10} j^{d-1} M_{n,j}^d.$$

We can show similarly as in the proof of Theorem 3.2 in Györfi et al. (2002) that $m^{(c)}$ is (p,C)-smooth. Hence each distribution (X,Y) with $Y = m^{(c)}(X) + N$ for X, N independent, N standard normal, X having density f and $c \in \mathcal{C}_n$ is contained in $\mathcal{D}(p, C, \beta, C, 1)$. Thus it suffices to show

$$\liminf_{n\to\infty} n^{2p/(2p+d)} \inf_{m_n} \sup_{\substack{(X,Y):Y=m^{(c)}(X)+N, c\in\mathcal{C}_n,\\ X\, has\, density\, f}} \mathbf{E} \int |m_n(x) - m^{(c)}(x)|^2 \mu(dx) = \infty.$$

Let m_n be an arbitrary estimate. Since $\{g_{n,k}^j(x) : j, k\}$ is an orthogonal system in L_2, the projection $\hat{m}_n$ of m_n to $\{m^{(c)} : c \in \mathbb{R}^{\sum_{j=1}^{j_{max}(n)} N_{n,j}}\}$ is given by

$$\hat{m}_n(x) = \sum_{j=1}^{j_{max}(n)} \sum_{k=1}^{N_{n,j}} \hat{c}_{n,k}^j g_{n,k}^j(x),$$

where

$$\hat{c}_{n,k}^j = \frac{\int_{A_{n,k}^j} m_n(x) g_{n,k}^j(x) \mu(dx)}{\int_{A_{n,k}^j} (g_{n,k}^j(x))^2 \mu(dx)}.$$

Let $c \in \mathcal{C}_n$ be arbitrary. Then

$$\begin{aligned}
\int |m_n(x) - m^{(c)}(x)|^2 \mu(dx) &\geq \int |\hat{m}_n(x) - m^{(c)}(x)|^2 \mu(dx) \\
&= \sum_{j=1}^{j_{max}(n)} \sum_{k=1}^{N_{n,j}} \int (\hat{c}_{n,k}^j g_{n,k}^j(x) - c_{n,k}^j g_{n,k}^j(x))^2 \mu(dx) \\
&= \sum_{j=1}^{j_{max}(n)} \sum_{k=1}^{N_{n,j}} (\hat{c}_{n,k}^j - c_{n,k}^j)^2 \int |g_{n,k}^j(x)|^2 \mu(dx).
\end{aligned}$$

Let $\tilde{c}_{n,k}^j$ be 1 when $\hat{c}_{n,k}^j \geq 0$ and -1 otherwise. Because of

$$|\hat{c}_{n,k}^j - c_{n,k}^j| \geq 1_{\{\tilde{c}_{n,k}^j \neq c_{n,k}^j\}},$$

we have

$$\int |m_n(x) - m^{(c)}(x)|^2 \mu(dx) \geq \sum_{j=1}^{j_{max}(n)} \sum_{k=1}^{N_{n,j}} 1_{\{\tilde{c}_{n,k}^j \neq c_{n,k}^j\}} \int |g_{n,k}^j(x)|^2 \mu(dx).$$

Fix $1 \leq j \leq j_{max}(n)$ and $1 \leq k \leq N_{n,j}$. Then

$$g_{n,k}^j(x) = 0 \quad \text{for} \quad x \notin [-j, j]^d$$

so

$$\begin{aligned}
\int |g_{n,k}^j(x)|^2 \mu(dx) &= \int |g_{n,k}^j(x)|^2 f(x) dx \\
&\geq c_{11} \frac{1}{(1+\sqrt{d}j)^{2p+d}} \int |g_{n,k}^j(x)|^2 dx \\
&= c_{11} \frac{1}{(1+\sqrt{d}j)^{2p+d}} \frac{1}{M_{n,j}^{2p+d}} C^2 \int \overline{g}^2(x) dx
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m^{(c)}(x)|^2 \mu(dx) \\
&\geq \sum_{j=1}^{j_{max}(n)} \sum_{k=1}^{N_{n,j}} \frac{1}{M_{n,j}^{2p+d}} \cdot C^2 \cdot c_{11} \cdot \int \overline{g}^2(x) dx \cdot \mathbf{P}\{\tilde{c}_{n,k}^j \neq c_{n,k}^j\} \frac{1}{(1+\sqrt{d}j)^{2p+d}}.
\end{aligned}$$

Now, let us randomize c by taking a sequence $C_{n,1}^1, \ldots, C_{n,N_{n,j_{max}(n)}}^{j_{max}(n)}$ of i.i.d. random variables independent of $(X_1, N_1), (X_2, N_2), \ldots$, satisfying

$$\mathbf{P}\{C_{n,1}^1 = 1\} = \mathbf{P}\{C_{n,1}^1 = -1\} = \frac{1}{2}.$$

Then

$$\begin{aligned}
&n^{2p/(2p+d)} \inf_{m_n} \sup_{\substack{(X,Y): Y = m^{(c)}(X) + N, c \in \mathcal{C}_n, \\ X \, has \, density \, f}} \mathbf{E} \int |m_n(x) - m^{(c)}(x)|^2 \mu(dx) \\
&\geq \inf_{\tilde{c}} n^{-d/(2p+d)} \sum_{j=1}^{j_{max}(n)} \sum_{k=1}^{N_{n,j}} j^{2p+d} \cdot c_{12} \\
&\qquad \cdot \int \overline{g}^2(x) dx \frac{1}{(1+\sqrt{d}j)^{2p+d}} \mathbf{P}\{\tilde{c}_{n,k}^j \neq C_{n,k}^j\}
\end{aligned}$$

where $\tilde{c}$ is the vector of $\tilde{c}_{n,k}^j$ which can be interpreted as a decision on $C_{n,k}^j$ using the observed data. Fix $1 \leq j \leq j_{max}(n)$ and $1 \leq k \leq N_{n,j}$. Let $X_{i_1}, \ldots, X_{i_l}$ be those $X_i \in A_{n,k}^j$. Then

$$(Y_{i_1}, \ldots, Y_{i_l}) = C_{n,k}^j \cdot (g_{n,k}^j(X_{i_1}), \ldots, g_{n,k}^j(X_{i_l})) + (N_{i_1}, \ldots, N_{i_l}),$$

while

$$\{Y_1, \ldots, Y_n\} \setminus \{Y_{i_1}, \ldots, Y_{i_l}\}$$

are independent of $C^j_{n,k}$ given $X_1, \ldots, X_n$. By Lemma 3.2 in Györfi et al. (2002) we get

$$\begin{aligned} \mathbf{P}\{\tilde{c}^j_{n,k} \neq C^j_{n,k} | X_1, \ldots, X_n\} &\geq \Phi\left(-\sqrt{\sum_{r=1}^{l} (g^j_{n,k}(X_{i_r}))^2}\right) \\ &= \Phi\left(-\sqrt{\sum_{i=1}^{n} (g^j_{n,k}(X_i))^2}\right) \end{aligned}$$

where Φ is the standard normal distribution function. Since $\Phi(-\sqrt{x})$ is convex we get by Jensen's inequality

$$\mathbf{P}\{\tilde{c}^j_{n,k} \neq C^j_{n,k}\} \geq \Phi\left(-\sqrt{\mathbf{E}\left\{\sum_{i=1}^{n} (g^j_{n,k}(X_i))^2\right\}}\right) = \Phi\left(-\sqrt{n\mathbf{E}\{(g^j_{n,k}(X_1))^2\}}\right).$$

Because of

$$\begin{aligned} n\mathbf{E}\{(g^j_{n,k}(X_1))^2\} &= nM_{n,j}^{-2p} \int_{A^j_{n,k}} g^2(M_{n,j}(x - a^j_{n,k}))f(x)dx \\ &\leq nM_{n,j}^{-2p} \int_{A^j_{n,k}} g^2(M_{n,j}(x - a^j_{n,k}))\frac{c_9}{(1+(j-1))^{2p+d}}dx \\ &= nM_{n,j}^{-2p-d} C^2 \int \overline{g}^2(x)dx \cdot \frac{c_9}{j^{2p+d}} \\ &\leq \int \overline{g}^2(x)dx \cdot c_{13} < \infty \end{aligned}$$

we conclude

$$\begin{aligned} & n^{2p/(2p+d)} \inf_{m_n} \sup_{\substack{(X,Y): Y = m^{(c)}(X)+N, c \in \mathcal{C}_n, \\ X \, has \, density \, f}} \mathbf{E} \int |m_n(x) - m^{(c)}(x)|^2 \mu(dx) \\ & \geq c_{14} \cdot n^{-d/(2p+d)} \sum_{j=1}^{j_{max}(n)} N_{n,j} \cdot j^{2p+d} \cdot \frac{1}{(1+\sqrt{d}j)^{2p+d}} \\ & \geq c_{15} \cdot n^{-d/(2p+d)} \sum_{j=1}^{j_{max}(n)} j^{d-1} \cdot C^{2d/(2p+d)} \left(n^{d/(2p+d)}/j^d\right) \cdot j^{2p+d} \frac{1}{(1+\sqrt{d}j)^{2p+d}} \\ & \geq c_{16} \cdot C^{2d/(2p+d)} \sum_{j=1}^{j_{max}(n)} \frac{1}{j} \to \infty \end{aligned}$$

since $j_{max}(n) \to \infty$ as $n \to \infty$. $\square$

References

[1] Cover, T. M. (1968). Rates of convergence for nearest neighbor procedures. *Proceedings of the Hawaii International Conference on System Sciences*, pages 413–415. Honolulu, HI.

[2] Devroye, L. (1982). Any discrimination rule can have arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **4**, 154–157.

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

[4] Greblicki, W., Krzyżak, A., and Pawlak, M. (1984). Distribution-free pointwise consistency of kernel regression estimate, *Annals of Statistics*, **12**, 1570-1575.

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

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

[7] 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.

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

[9] 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.

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

[11] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, **89**, 1–23.

[12] Kohler, M. (2002). Universal consistency of local polynomial kernel regression estimates. *Annals of the Institute of Statistical Mathematics* **54**, pp. 879–899.

[13] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, 3054–3058.

[14] Kohler, M., Krzyżak, A. and Walk, H. (2005). Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. To appear in *Journal of Multivariate Analysis.*

[15] Kulkarni, S. R. and Posner, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, **41**, 1028–1039.

[16] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, 677-687.

[17] Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *Annals of Statistics*, **24**, 1084–1105.

[18] Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Annals of Statistics*, **8**, 240–246.

[19] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, 595–645.

[20] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, 1040–1053.

[21] Walk, H. (2002). Almost sure convergence properties of Nadaraya–Watson regression estimates. In *Modeling Uncertainty: An Examination of its Theory, Methods and Applications*, eds. M. Dror, P. L'Ecuyer and F. Szidarovszky, pages 201-223. Kluwer, Dordrecht.