Journal of Statistical Planning and Inference 139 (2009) 1286 – 1296

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimal global rates of convergence for nonparametric regression with unbounded data[☆]

Michael Kohler[a], Adam Krzyżak[b,*], Harro Walk[c]

[a]Department of Mathematics, Technische Universität Darmstadt, Schlossgartenstr. 7, D-64289 Darmstadt, Germany
[b]Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8
[c]Department of Mathematics, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 1 July 2005
Received in revised form
12 September 2007
Accepted 2 July 2008
Available online 19 August 2008

Keywords:
Regression
Kernel estimate
Rate of convergence

ABSTRACT

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 the explanatory variable X is made besides smoothness assumptions on the regression function and moment conditions on the response variable 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 Question 1 in Stone [1982. Optimal global rates of convergence for nonparametric regression. Ann. Statist., 10, 1040–1053].

1. Introduction

Let $(X,Y),(X_1,Y_1),(X_2,Y_2),\ldots$ 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 the so-called response variable Y after having observed the value of the so-called explanatory variable 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

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

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

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

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 $\int|f(x)-m(x)|^2\mu(\mathrm{d}x)$ 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 $\mathscr{D}_n=\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ of independent observations of (X,Y), our goal is to construct an estimate $m_n(x)=m_n(x,\mathscr{D}_n)$ of $m(x)$ such that the L_2 error $\int|m_n(x)-m(x)|^2\mu(\mathrm{d}x)$ is small.

[☆] Research supported by the Alexander von Humboldt Foundation.

[*] Corresponding author.

E-mail addresses: kohler@mathematik.tu-darmstadt.de (M. Kohler), krzyzak@cs.concordia.ca (A. Krzyżak), walk@mathematik.uni-stuttgart.de (H. Walk).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.07.012

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 et al. (1984), Györfi et al. (1998), Györfi and Walk (1996, 1997), Kohler (1999, 2002), Kohler and Krzyżak (2001), Lugosi and Zeger (1995), Nobel (1996) and Walk (2005, 2008). 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 Györfi et al. (2002, Chapter 3). 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 $\mathscr{D}$ be a class of distributions of (X,Y). In the classical minimax theory, one considers the maximal error of an estimate within the class $\mathscr{D}$ 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

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

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 are (p,C)-smooth according to the following definition.

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

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

of m exist and satisfy

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

Let $\mathscr{D}^{(p,C)}$ be the class of all distributions of (X,Y) where X takes on 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

$$\liminf_{n\to\infty}\inf_{\hat{m}_n}\sup_{(X,Y)\in\mathscr{D}^{(p,C)}}\frac{\mathbf{E}\int|\hat{m}_n(x)-m(x)|^2\,dx}{C^{2d/(2p+d)}n^{-2p/(2p+d)}}>C_1>0 \tag{4}$$

for some constant C_1 independent of C (cf. Györfi et al., 2002, Theorem 3.2), and that a suitably defined local polynomial kernel estimate satisfies

$$\limsup_{n\to\infty}\sup_{(X,Y)\in\mathscr{D}^{(p,C)}}\frac{\mathbf{E}\int|m_n(x)-m(x)|^2\,dx}{C^{2d/(2p+d)}n^{-2p/(2p+d)}}<C_2<\infty \tag{5}$$

for some constant C_2 independent of C. 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 $p\leqslant 1$, cf. Györfi et al. (2002, Theorem 5.2).

Since X has a density with respect to the Lebesgue–Borel measure, which is bounded away from zero and infinity, for $(X,Y)\in\mathscr{D}^{(p,C)}$, 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; Kulkarni and Posner, 1995) that suitably defined partitioning, kernel and nearest neighbor estimates satisfy

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

provided X takes on values in $[0,1]^d$, $\mathrm{Var}\{Y|X=x\}$ is bounded and m is (p,C)-smooth for some $p \leqslant 1$. In case of the nearest neighbor estimates one needs the additional condition $d \geqslant 2p$, for the other two estimates the result holds in any dimension (cf., Györfi et al., 2002, Lemma 6.4; Liitiäinen et al., 2007, Proposition 2.3). 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 the 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 the conjecture is indeed true. This is interesting in applications because it implies that one can get reasonable results for distributions with bounded support even in case of large values of $\|X\|$. In particular we show that for m bounded and (p,C)-smooth with $0<p\leqslant 1$, $\mathrm{Var}\{Y|X=x\}$ 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 et al. (2006).

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 measurable kernel $K:\mathbb{R}^d\to\mathbb{R}_+$ satisfies

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

for some constants $0<c_1\leqslant c_2<\infty$, 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

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

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 $L\geqslant 1$.*
(A2) *$m(x)=\mathbf{E}\{Y|X=x\}$ is (p,C)-smooth for some $0<p\leqslant 1$, $C\geqslant 1$.*
(A3) *The conditional variance of Y satisfies*

$$\mathrm{Var}\{Y|X=x\}\leqslant\sigma_0^2$$

for some $\sigma_0\geqslant 0$.
(A4) *There exists a constant $\beta>2p$ such that*

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

for some constant $M\geqslant 0$.

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

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

where c_3 depends only on $d, p, \beta, L, M, \sigma_0$, c_1 and c_2.

Theorem 1 implies the following result concerning minimax rate of convergence: Let $0<p\leqslant 1, C\geqslant 1, \beta>2p, L>0, M\geqslant 0$ 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)–(A4) for these values of p,C,β,L,M and σ_0. Then

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

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 1. 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. that $c_1\geqslant 1$. In this case the estimate can be also defined via

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

This definition of the kernel estimate was used in Spiegelman and Sacks (1980) in connection with the analysis 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

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

for some $c^*>0$ and some $\delta>0$.

Remark 2. Theorem 1 above extends well-known optimal rate of convergence results to unbounded distributions of the explanatory variable X in case of (p,C)-smooth regression functions where p satisfies $p\leqslant 1$. In L_2 regression it is known that kernel estimates (like other local averaging estimates) are not able to achieve the optimal rate of convergence for $p>1.5$ even in case of bounded $\|X\|$ (cf., Eqs. (5.2) and (5.4) in Györfi et al., 2002). Therefore it is not possible to extend Theorem 1, which considers the kernel estimate, in such a way that we get the optimal rate of convergence for arbitrarily smooth regression functions.

Remark 3. The kernel estimate above is easy to compute in practice, but as pointed out in the previous remark it does not achieve the optimal L_2 rate of convergence for very smooth regression functions. We next show that under stronger conditions on Y than in Theorem 1 and with a least squares estimate, which is very hard to compute in practice, it is possible to get the optimal L_2 rate of convergence also for very smooth regression functions and X with unbounded support.

To see this, define a partition $\mathcal{P}_n$ of $\mathbb{R}^d$ depending on $C,p,\beta>0$ and $n\in\mathbb{N}$ as follows: For $j\in\mathbb{N}$ set

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

and let $A^j_{n,1},\dots,A^j_{n,(2j)^dM^d_{n,j}}$ be the uniform partition of $[-j,j]^d$ consisting of $(2j)^d\cdot M^d_{n,j}$ cubes of the side length $h_{n,j}=1/M_{n,j}$. Set

$$j_{\max}(n)=\lceil n^{2p/((2p+d)\cdot\beta)}\rceil$$

and define $\mathcal{P}_n$ by

$$\mathcal{P}_n=\{\mathbb{R}^d\backslash[-j_{\max}(n),j_{\max}(n)]^d\}\cup\{A^1_{n,k}\,:\,k=1,\dots,2^dM^d_{n,1}\}\cup\bigcup_{j=2}^{j_{\max}(n)}\{A^j_{n,k}\backslash[-(j-1),j-1]^d\,:\,A^j_{n,k}\backslash[-(j-1),j-1]^d\neq\emptyset\}.$$

Let $\mathcal{F}_n$ be the set of all piecewise polynomials of degree M with respect to $\mathcal{P}_n$, which are bounded in absolute value by $(L+1)$. Let m_n be the corresponding least squares estimate, i.e.,

$$m_n(\cdot)=\operatorname*{argmin}_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^n|f(X_i)-Y_i|^2.$$

Assume that

$$(\mathrm{A3}')\quad K^2\cdot\mathbf{E}\{e^{(Y-m(X))^2/K^2}-1|X\}\leqslant\sigma_0\quad\text{a.s.}$$

for some $K,\sigma_0>0$, and, in addition, (A1), (A2) (with $p>0$ arbitrary) and (A4) hold.

Then

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

so the least squares estimate achieves the optimal rate of convergence also for unbounded support of X and $p>1$.

We can derive this result using the proof of Corollary 1 in Kohler (2006) which implies

$$\mathbf{E}\int|m_n(x)-m(x)|^2\mu(dx)\leqslant\text{const}\cdot\left(\frac{(M+1)\cdot|\mathscr{P}_n|}{n}+\inf_{f\in\mathscr{F}_n}\int|f(x)-m(x)|^2\mu(dx)\right).$$

(Actually, this result is proven only in probability in Kohler, 2006. Nevertheless this implies the above result, since the probability in Kohler, 2006, converges to zero exponentially fast and since the L_2 error is bounded.)

Now it is easy to see that

$$\frac{(M+1)\cdot|\mathscr{P}_n|}{n}\leqslant\text{const}\cdot C^{2d/(2p+d)}\cdot n^{-2p/(2p+d)}.$$

Furthermore, by Lemma 11.1 in Györfi et al. (2002), we can bound the approximation error by

$$\inf_{f\in\mathscr{F}_n}\int|f(x)-m(x)|^2\mu(dx)\leqslant\text{const}\cdot\sum_{A\in\mathscr{P}_n,A\subseteq[-j_{\max}(n),j_{\max}(n)]^d}C^2\cdot|A|^{2p}\cdot\mu(A)+L^2\cdot\mu(\mathbb{R}^d\setminus[-j_{\max}(n),j_{\max}(n)]^d).$$

By the definition of $\mathscr{P}_n$ we can bound the right-hand side above by

$$\text{const}\cdot C^2\cdot\int_{[-j_{\max}(n),j_{\max}(n)]^d}[h_n\cdot(1+\|x\|)^{\beta/(2p)}]\mu(dx)+L^2\cdot\mu(\mathbb{R}^d\setminus[-j_{\max}(n),j_{\max}(n)]^d),$$

and by bounding this as in the proof of Theorem 1 below one gets the desired result.

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,

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

where

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

is binomially distributed with parameters n and $q=\mu(S_{x,h_n(x)})$ and $c_4=(c_2/c_1)^2$.

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

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

Gathering the above results we get

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

Using

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

we get

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

Hence

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

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 $\|x\|>\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor$ we have

$$\int\frac{1}{\mu(S_{x,h_n(x)})}\mu(\mathrm{d}x)\leqslant\int_{S_{0,2}}\frac{1}{\mu(S_{x,h_n})}\mu(\mathrm{d}x)+\sum_{j=3}^{\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor}\int_{S_{0,j}\setminus S_{0,j-1}}\frac{1}{\mu(S_{x,h_nj^{\beta/(2p)}})}\mu(\mathrm{d}x)+\int_{\mathbb{R}^d\setminus S_{0,\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor}}1\,\mu(\mathrm{d}x).$$

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

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

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

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

where $c_5 = 2d8^d$ and $S_{z_1,r/2},\dots,S_{z_l,r/2}$ cover $S_{0,j}\backslash S_{0,j-1}$ (because if any point $z \in S_{0,j}\backslash S_{0,j-1}$ is in none of those balls, then $S_{z,r/4}$ does not overlap with any of the balls (11) and is contained in $S_{0,j+1}\backslash S_{0,j-2}$). From this we can conclude

$$\int_{S_{0,j}\backslash S_{0,j-1}} \frac{1}{\mu(S_{x,h_n j^{\beta/(2p)}})}\,\mu(\mathrm{d}x) \leqslant \sum_{k=1}^{l} \int_{S_{z_k,r/2}} \frac{1}{\mu(S_{x,r})}\,\mu(\mathrm{d}x)$$
$$\leqslant \sum_{k=1}^{l} \int_{S_{z_k,r/2}} \frac{1}{\mu(S_{z_k,r/2})}\,\mu(\mathrm{d}x) = l$$

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 $\int_{S_{0,2}} \frac{1}{\mu(S_{x,h_n})}\,\mu(\mathrm{d}x)$ we get

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

The Markov inequality implies

$$\mu(\mathbb{R}^d \backslash S_{0,\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor}) \leqslant \frac{\mathbf{E}\|X\|^\beta}{(\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor)^\beta} \leqslant \mathbf{E}\|X\|^\beta \cdot 2^\beta n^{-2p/(2p+d)}. \tag{12}$$

From this we can conclude

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

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

$$\int \min\{C^2 |h_n(x)|^{2p}, 4L^2\}\,\mu(\mathrm{d}x)$$
$$\leqslant C^2 \cdot \int (h_n \cdot (1+\|x\|)^{\beta/(2p)})^{2p} \mu(\mathrm{d}x) + 4L^2 \mu(\mathbb{R}^d \backslash S_{0,\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor})$$
$$\leqslant C^2 h_n^{2p} \cdot 2^\beta \cdot (1 + \mathbf{E}\|X\|^\beta) + 4L^2 \mu(\mathbb{R}^d \backslash S_{0,\lfloor n^{2p/((2p+d)\cdot\beta)}\rfloor})$$
$$\leqslant c_7 \cdot (C^2 h_n^{2p} + n^{-2p/(2p+d)}),$$

where the last inequality follows from (12) and $c_7 = \max\{2^\beta(1+\mathbf{E}\|X\|^\beta), 4L^2\mathbf{E}\|X\|^\beta 2^\beta\}$.

Putting together the above results we get

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x)$$
$$\leqslant \frac{2 \cdot \max\{c_4\sigma_0^2, L^2\}}{n+1} \cdot c_6 \cdot \left(\frac{1}{h_n^d} + n^{-2p/(2p+d)}\right) + c_4 \cdot c_7 \cdot (C^2 h_n^{2p} + n^{-2p/(2p+d)})$$
$$\leqslant c_3 \cdot C^{2d/(2p+d)} n^{-2p/(2p+d)}$$

where $c_3 = 3\max\{c_4\sigma_0^2, L^2\}c_6 + 2c_4c_7$. □

4. Proof of Theorem 2

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

$$f(x) = c_8 \cdot \frac{1}{(1+||x||)^{2p+d}} \quad (x \in \mathbb{R}^d),$$

thus

$$\mathbf{E}||X||^\beta \leqslant c_8 \int \frac{1}{(1+\|x\|)^{d+(2p-\beta)}}\,\mathrm{d}x =: M < \infty.$$

Set $g(x)=C\cdot\bar{g}(x)$ for some function $\bar{g}:\mathbb{R}^d\to\mathbb{R}$ such that $\bar{g}(x)=0$ for $x\notin[-1/2,1/2]^d, \bar{g}(x)\neq 0$ elsewhere, $\bar{g}$ bounded in absolute value by 1, and $\bar{g}$ $(p,2^{\gamma-1})$-smooth, where $p=k+\gamma$ for some $k\in\mathbb{N}_0, 0<\gamma\leqslant 1$. 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 $\mathscr{C}_n$. For $c\in\mathscr{C}_n$ 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

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

Set

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

partition $[-j,j]^d$ into $(2j)^dM^d_{n,j}$ 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 $A\in\{A^j_{n,k}|k=1,\ldots,(2j)^dM^d_{n,j}\}$ which satisfy $A\subseteq[-j,j]^d\backslash[-(j-1),j-1]^d$. (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 $[-j,j]^d\backslash[-(j-1),j-1]^d$. In case of

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

(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

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

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

$$N_{n,j}\geqslant\frac{(2j)^d}{h^d_{n,j}}-\frac{(2j-1)^d}{h^d_{n,j}}=c_9j^{d-1}M^d_{n,j},$$

where $c_9=2^{d-1}$. 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 $c\in\mathscr{C}_n$ is contained in $\mathscr{D}(p,C,\beta,C,M,1)$. Thus it suffices to show

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

Let m_n be an arbitrary estimate. Since $\{g^j_{n,k}(x):j,k\}$ 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

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

where

$$\hat{c}^j_{n,k}=\frac{\int_{A^j_{n,k}}m_n(x)g^j_{n,k}(x)\mu(\mathrm{d}x)}{\int_{A^j_{n,k}}(g^j_{n,k}(x))^2\mu(\mathrm{d}x)}.$$

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

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

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

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

we have

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

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

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

so

$$\begin{aligned}\int |g_{n,k}^j(x)|^2\mu(dx) &= \int |g_{n,k}^j(x)|^2 f(x)\,dx\\ &\geqslant c_8\frac{1}{(1+\sqrt{d}j)^{2p+d}}\int |g_{n,k}^j(x)|^2\,dx\\ &= c_8\frac{1}{(1+\sqrt{d}j)^{2p+d}}\frac{1}{M_{n,j}^{2p+d}}C^2\int \bar{g}^2(x)\,dx\end{aligned}$$

which implies

$$\mathbf{E}\int |m_n(x)-m^{(c)}(x)|^2\mu(dx) \geqslant \sum_{j=1}^{j_{\max}(n)}\sum_{k=1}^{N_{n,j}}\frac{1}{M_{n,j}^{2p+d}}\cdot C^2\cdot c_8\cdot\int \bar{g}^2(x)\,dx\cdot\mathbf{P}\{\tilde{c}_{n,k}^j \neq c_{n,k}^j\}\frac{1}{(1+\sqrt{d}j)^{2p+d}}.$$

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

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

Then

$$\begin{aligned}&n^{2p/(2p+d)}\inf_{m_n}\sup_{\substack{(X,Y):Y=m^{(c)}(X)+N,c\in\mathscr{C}_n,\\ X \text{ has density } f}}\mathbf{E}\int |m_n(x)-m^{(c)}(x)|^2\mu(dx)\\ &\geqslant \inf_{\tilde{c}} n^{-d/(2p+d)}\sum_{j=1}^{j_{\max}(n)}\sum_{k=1}^{N_{n,j}} j^{2p+d}\cdot c_8\cdot\int \bar{g}^2(x)\,dx\frac{1}{(1+\sqrt{d}j)^{2p+d}}\mathbf{P}\{\tilde{c}_{n,k}^j \neq C_{n,k}^j\},\end{aligned}$$

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

$$\{Y_1,\dots,Y_n\}\backslash\{Y_{i_1},\dots,Y_{i_l}\}$$

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

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

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

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

Because of

$$\begin{aligned}n\mathbf{E}\{(g^j_{n,k}(X_1))^2\}&=nM_{n,j}^{-2p}\int_{A^j_{n,k}}g^2(M_{n,j}(x-a^j_{n,k}))f(x)\,\mathrm{d}x\\&\leqslant nM_{n,j}^{-2p}\int_{A^j_{n,k}}g^2(M_{n,j}(x-a^j_{n,k}))\frac{c_8}{(1+(j-1))^{2p+d}}\,\mathrm{d}x\\&=nM_{n,j}^{-2p-d}C^2\int\bar{g}^2(x)\,\mathrm{d}x\cdot\frac{c_8}{j^{2p+d}}\\&=c_8\cdot\int\bar{g}^2(x)\,\mathrm{d}x<\infty\end{aligned}$$

we conclude

$$\begin{aligned}&n^{2p/(2p+d)}\inf_{m_n}\sup_{\substack{(X,Y):Y=m^{(c)}(X)+N,c\in\mathscr{C}_n,\\ X\text{ has density }f}}\mathbf{E}\int|m_n(x)-m^{(c)}(x)|^2\mu(\mathrm{d}x)\\&\geqslant c_{10}\cdot n^{-d/(2p+d)}\sum_{j=1}^{j_{\max}(n)}N_{n,j}\cdot j^{2p+d}\cdot\frac{1}{(1+\sqrt{d}j)^{2p+d}}\\&\geqslant c_{11}\cdot n^{-d/(2p+d)}\sum_{j=1}^{j_{\max}(n)}j^{d-1}\cdot C^{2d/(2p+d)}(n^{d/(2p+d)}/j^d)\cdot j^{2p+d}\frac{1}{(1+\sqrt{d}j)^{2p+d}}\\&\geqslant c_{12}\cdot C^{2d/(2p+d)}\sum_{j=1}^{j_{\max}(n)}\frac{1}{j}\to\infty\end{aligned}$$

since $j_{\max}(n)\to\infty$ as $n\to\infty$, where c_{12} depends on C, d, p and c_8. □

Acknowledgment

The authors wish to thank an anonymous referee and an Associate Editor for several helpful comments.

References

Cover, T.M., 1968. Rates of convergence for nearest neighbor procedures. Proceedings of the Hawaii International Conference on System Sciences, Honolulu, HI, pp. 413–415.

Devroye, L., 1982. Any discrimination rule can have arbitrarily bad probability of error for finite sample size. IEEE Trans. Pattern Anal. Mach. Intell. 4, 154–157.

Devroye, L., Györfi, L., Krzyżak, A., Lugosi, G., 1994. On the strong universal consistency of nearest neighbor regression function estimates. Ann. Statist. 22, 1371–1385.

Greblicki, W., Krzyżak, A., Pawlak, M., 1984. Distribution-free pointwise consistency of kernel regression estimate. Ann. Statist. 12, 1570–1575.

Györfi, L., 1981. The rate of convergence of k_n-NN regression estimates and classification rules. IEEE Trans. Inform. Theory 27, 362–364.

Györfi, L., Walk, H., 1996. On the strong universal consistency of a series type regression estimate. Math. Methods Statist. 5, 332–342.

Györfi, L., Walk, H., 1997. On the strong universal consistency of a recursive regression estimate by Pál Révész. Statist. Probab. Lett. 31, 177–183.

Györfi, L., Kohler, M., Walk, H., 1998. Weak and strong universal consistency of semi-recursive partitioning and kernel regression estimates. Statist. Decisions 16, 1–18.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H., 2002. A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics. Springer, New York.

Kohler, M., 1999. Universally consistent regression function estimation using hierarchical B-splines. J. Multivariate Anal. 67, 138–164.

Kohler, M., 2000. Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. J. Statist. Plann. Inference 89, 1–23.

Kohler, M., 2002. Universal consistency of local polynomial kernel regression estimates. Ann. Inst. Statist. Math. 54, 879–899.

Kohler, M., 2006. Nonparametric regression with additional measurement errors in the dependent variable. J. Statist. Plann. Inference 136, 3339–3361.

Kohler, M., Krzyżak, A., 2001. Nonparametric regression estimation using penalized least squares. IEEE Trans. Inform. Theory 47, 3054–3058.

Kohler, M., Krzyżak, A., Walk, H., 2006. Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. J. Multivariate Anal. 97, 311–323.

Kulkarni, S.R., Posner, S.E., 1995. Rates of convergence of nearest neighbor estimation under arbitrary sampling. IEEE Trans. Inform. Theory 41, 1028–1039.

Liitiäinen, E., Corona, F., Lendasse, A., 2007. Nearest neighbor distributions and noise variance estimation. ESANN'2007 Proceedings—European Symposium on Artificial Neural Networks, Bruge, Belgium, pp. 67–72.

Lugosi, G., Zeger, K., 1995. Nonparametric estimation via empirical risk minimization. IEEE Trans. Inform. Theory 41, 677–687.

Nobel, A., 1996. Histogram regression estimation using data-dependent partitions. Ann. Statist. 24, 1084–1105.

Spiegelman, C., Sacks, J., 1980. Consistent window estimation in nonparametric regression. Ann. Statist. 8, 240–246.

Stone, C.J., 1977. Consistent nonparametric regression. Ann. Statist. 5, 595–645.

Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Ann. Statist. 10, 1040–1053.

Walk, H., 2005. Strong universal consistency of smooth kernel regression estimates. Ann. Inst. Statist. Math. 57, 665–685.

Walk, H., 2008. A universal strong law of large numbers for conditional expectations via nearest neighbors. J. Multivariate Anal. 99, 1035–1050.