Ann. Inst. Statist. Math.
Vol. 55, No. 2, 287–308 (2003)
©2003 The Institute of Statistical Mathematics

STRONG CONSISTENCY OF AUTOMATIC KERNEL REGRESSION ESTIMATES*

MICHAEL KOHLER[1], ADAM KRZYŻAK[2]** AND HARRO WALK[1]

[1] *Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany,* e-mail: kohler@mathematik.uni-stuttgart.de; walk@mathematik.uni-stuttgart.de
[2] *Department of Computer Science, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8,* e-mail: krzyzak@cs.concordia.ca

(Received January 4, 2002; revised August 19, 2002)

Abstract. Regression function estimation from independent and identically distributed bounded data is considered. The L_2 error with integration with respect to the design measure is used as an error criterion. It is shown that the kernel regression estimate with an arbitrary random bandwidth is weakly and strongly consistent for *all* distributions whenever the random bandwidth is chosen from some deterministic interval whose upper and lower bounds satisfy the usual conditions used to prove consistency of the kernel estimate for deterministic bandwidths. Choosing discrete bandwidths by cross-validation allows to weaken the conditions on the bandwidths.

Key words and phrases: Automatic kernel estimates, cross-validation, regression estimates, strong consistency.

1. Introduction

1.1 *Nonparametric regression function estimation*

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 $\boldsymbol{E}Y^2 < \infty$. 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

$$\boldsymbol{E}\{|f^*(X)-Y|^2\} = \min_f \boldsymbol{E}\{|f(X)-Y|^2\}. \tag{1.1}$$

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

$$\boldsymbol{E}\{|f(X)-Y|^2\} = \boldsymbol{E}\{|m(X)-Y|^2\} + \int |f(x)-m(x)|^2\mu(dx) \tag{1.2}$$

implies that m is the solution of the minimization problem (1.1), $\boldsymbol{E}\{|m(X)-Y|^2\}$ is the minimum of (1.2) and for an arbitrary f, the L_2 error $\int |f(x)-m(x)|^2\mu(dx)$ is the difference between $\boldsymbol{E}\{|f(X)-Y|^2\}$ and $\boldsymbol{E}\{|m(X)-Y|^2\}$.

*Research supported by DAAD, NSERC and Alexander von Humboldt Foundation.

**The research of the second author was completed during his stay at the Technical University of Szczecin, Poland.

In the regression estimation problem the distribution of (X, Y) (and consequently m) is unknown. Given a sequence $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ 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.

1.2 *Universal consistency*

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

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

for all distributions of (X, Y) with $\mathbf{E}Y^2 < \infty$. It is called **strongly universally consistent** if

$$\int |m_n(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

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

C. J. Stone (1977) first pointed out that there exist weakly universally consistent estimators. He considered k_n*-nearest neighbor estimates*

$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x) \cdot Y_i \tag{1.3}$$

where

$$W_{n,i}(x) = W_{n,i}(x, X_1, \dots, X_n) \tag{1.4}$$

is one if X_i is among the k_n-nearest neighbors of x in $\{X_1, \dots, X_n\}$ and zero otherwise, and where $k_n \to \infty$ and $k_n/n \to 0$ $(n \to \infty)$. The strong universal consistency of nearest neighbor estimates has been shown in Devroye *et al.* (1994).

Estimates of the form (1.3) with weight functions (1.4) are called local averaging estimates. The most popular examples of local averaging estimates are *kernel estimates*, where

$$W_{n,i}(x) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}$$

$(0/0 = 0$ by definition) for some function $K : \mathbb{R}^d \to \mathbb{R}_+$ (called kernel) and some $h_n > 0$ (called bandwidth).

The weak universal consistency of kernel estimates has been shown under certain conditions on h_n and K independently by Devroye and Wagner (1980) and Spiegelman and Sachs (1980). The strong universal consistency of kernel estimates for suitably defined kernels and sequences of bandwidths has been shown by Walk (2002*b*). Various results concerning consistency of variants of kernel estimates can be found in Devroye and Krzyżak (1989), Györfi and Walk (1996, 1997) and Györfi *et al.* (1998).

1.3 *Automatic kernel estimates*

The consistency results mentioned above were proven for sequences of bandwidths which do not depend on the data. In order to achieve the optimal rate of convergence one has to choose these sequences in dependence of the smoothness of the regression function. Usually this is not possible in applications because there the smoothness of the regression function is unknown.

Therefore one often uses the data to generate a random bandwith $H = H(\mathcal{D}_n) \in \mathbb{R}_+$ and considers the kernel regression estimate

$$m_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{H}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x - X_i}{H}\right)} \tag{1.5}$$

which we shall call *automatic kernel regression estimate*. A list of various methods to choose a random bandwidth $H = H(\mathcal{D}_n) \in \mathbb{R}_+$ can be found e.g. in Chapter 4 of Fan and Gijbels (1996).

The aim of choosing a random bandwidth is to automatically adapt to the (smoothness of the) regression function. Usually one tries to show that one has achieved this goal by proving that the estimate achieves the corresponding optimal rate of convergence under various regularity assumptions (e.g. on the smoothness of the regression function). In the case that, unknown to the statistician, m is Lipschitz continuous with Lipschitz constant C and X and Y are bounded, choice of H from a finite, but growing set especially by splitting the data or by cross-validation, yields the convergence rate $O(C^{2d/(d+2)}n^{-2/(d+2)})$, see Hamers and Kohler (2003), Walk (2002*a*), and Györfi *et al.* (2002), Theorem 8.1. This rate is optimal according to Stone (1982). Unfortunately the results on optimal convergence rate of automatic regression estimates do not imply that the estimates are consistent if these conditions are not satisfied.

In Theorem 2.1 below we prove a general result which states that if one restricts the choice of the random bandwidths to some fixed deterministic intervals which satisfy some mild conditions then one gets for *all possible* choices of H a kernel estimate which is weakly and strongly consistent for all distributions with bounded Y. In the case that the bandwidths are chosen by cross-validation from (large) discrete sets, we show in Theorem 2.2 below that these conditions can be weakened.

The proof of Theorem 2.1 is based on techniques introduced in Kohler (2002) in the context of the proof of universal consistency of local polynomial kernel estimates. As mentioned in Remark 3 there, these results can be extended to data-dependent bandwidths. In this paper we show how the arguments there can be simplified in the case of kernel estimates.

The concept of cross-validation in statistics was introduced by Lunts and Brailovsky (1967), Allen (1974) and M. Stone (1974). Strong consistency of cross-validated kernel regression estimates and weak consistency of cross-validated nearest neighbor regression estimates were obtained by Wong (1983) and Li (1984), respectively, for fixed design and continuous regression function. As to further literature on cross-validation we refer to Härdle (1990) and Simonoff (1996).

Related results concerning partitioning, least squares and penalized least squares estimates can be found in Nobel (1996), Kohler (1999) and Kohler and Krzyżak (2001), resp. In the context of density estimation and classification corresponding results have been shown by Devroye and Györfi (1985) and in Chapter 25 of Devroye *et al.* (1996), resp. Various methods for the automatical choice of parameters of density estimates are described in Devroye and Lugosi (2001).

1.4 *Notation*

$\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ are the sets of natural, real and nonnegative real numbers, respectively. For $L > 0$ and $z \in \mathbb{R}$ set

$$T_L z = \begin{cases} L & \text{if } \; z > L, \\ z & \text{if } \; -L \leq z \leq L, \\ -L & \text{if } \; z < -L. \end{cases}$$

The natural logarithm is denoted by $\log(\cdot)$, the euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$. Set $S_{0,r} = \{x \in \mathbb{R}^d : \|x\| < r\}$, $x \in \mathbb{R}^d$, $r > 0$. For $h > 0$, $z \in \mathbb{R}^d$ and $K : \mathbb{R}^d \to \mathbb{R}$ define

$$K_h(z) = \frac{1}{h^d} K\left(\frac{z}{h}\right).$$

1.5 *Outline*

The main results are stated in Section 2 and proven in Sections 3 and 4. In the Appendix a list of some results of empirical process theory, which are used in the proofs, is given, together with specialized McDiarmid and Hoeffding inequalities.

2. Main results

THEOREM 2.1. *Let $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotone decreasing and left-continuous function satisfying*

$$\tilde{K}(+0) > 0,$$
$$t^d \tilde{K}(t^2) \to 0 \qquad \textit{as} \quad t \to \infty.$$

Define the kernel $K : \mathbb{R}^d \to \mathbb{R}$ by

$$K(u) = \tilde{K}(\|u\|^2) \qquad (u \in \mathbb{R}^d).$$

Let m_n be the kernel estimate defined by (1.5) *with the data-dependent bandwidth $H = H(\mathcal{D}_n)$ satisfying*

$$H \in [h_{\min}(n), h_{\max}(n)] \tag{2.1}$$

for some nonnegative numbers $h_{\min}(n), h_{\max}(n)$ such that

$$h_{\max}(n) \to 0 \qquad (n \to \infty) \tag{2.2}$$

and

$$\frac{n h_{\min}^d(n)}{\log n} \to \infty \qquad (n \to \infty). \tag{2.3}$$

Then

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

and

$$\int |m_n(x) - m(x)|^2 \mu(dx) \to 0 \qquad (n \to \infty) \qquad \textit{a.s.}$$

for all distributions of (X, Y) *with* $|Y| \leq L < \infty$ *a.s. for some* $L > 0$.

The proof of Theorem 2.1 is given in Section 3. The main idea is as follows: In the first part of the proof we approximate $|m_n(x) - m(x)|^2$ via a generalization of the classical Lebesgue density theorem by

$$\frac{\int |m_n(x) - m(z)|^2 K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)}$$
$$= \frac{\boldsymbol{E}\{|Y - m_n(x)|^2 K_H(x-X) \mid \mathcal{D}_n\} - \boldsymbol{E}\{|Y - m(X)|^2 K_H(x-X) \mid \mathcal{D}_n\}}{\int K_H(x-z)\mu(dz)}.$$

In the remaining parts we use that the kernel estimate minimizes an empirical version of the latter term, which enables us to analyze the kernel estimate similarly to least squares estimates.

In Theorem 2.1 above we do not assume anything about the method used to define the data dependent bandwidths H besides $H \in [h_{\min}(n), h_{\max}(n)]$. In particular, in Theorem 2.1 it is allowed that the method chooses the worst bandwidths in that interval. Therefore it is clear that condition (2.2) and (up to logarithmic factor) (2.3) are necessary. We show next that in case of cross-validation the conditions can be weakened to the assumption that the set of possible bandwidth sequences contains a sequence (h_n^*) satisfying the minimal condition $h_n^* \to 0$, $nh_n^{*d} \to \infty$ $(n \to \infty)$ for strong consistency. Due to technical difficulties in the proof we assume that the sets of possible bandwidths are finite. In applications minimization over a finite set of possible bandwidths is usual.

Let $\mathcal{Q}_n \subset (0, \infty)$ be a finite set of possible bandwidths, and for $h \in \mathcal{Q}_n$ set

$$m_n^{(h)}(x) = \frac{\sum_{i=1}^n K_h(x - X_i)Y_i}{\sum_{i=1}^n K_h(x - X_i)},$$
$$m_{n,i}^{(h)}(x) = \frac{\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} K_h(x - X_j)Y_j}{\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} K_h(x - X_j)} \qquad (i = 1, \ldots, n).$$

The cross-validated regression estimate is defined by $m_n^{(H_n)}(x)$ with data-dependent bandwidth

$$H_n = \arg\min_{h\in\mathcal{Q}_n} \sum_{i=1}^n (m_{n,i}^{(h)}(X_i) - Y_i)^2.$$

THEOREM 2.2. *Let* $K : \mathbb{R}^d \to \mathbb{R}_+$ *be a boxed kernel, that is, assume that* K *is measurable and satisfies*

$$bI_{S_{0,r}} \leq K \leq pI_{S_{0,R}}$$

for some $0 < r < R < \infty$, $0 < b < p < \infty$. *Let* $m_n^{(H_n)}$ *be the cross-validated regression estimate with*

$$(2.4) \qquad |\mathcal{Q}_n| = O(n^\tau) \quad \textit{for some} \quad \tau > 0,$$

$$(2.5) \qquad h_n^* \to 0, \ nh_n^{*d} \to \infty \quad (n \to \infty) \quad \textit{for some} \quad h_n^* \in \mathcal{Q}_n \quad (n \in \mathbb{N}).$$

Then

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

and

$$\int |m_n^{(H_n)}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s. \tag{2.6}$$

for all distributions of (X,Y) *with* $|Y| \le L < \infty$ *a.s. for some* $L > 0$.

In the proof of Theorem 2.2, which is given in Section 4, (2.6) is shown by an investigation of the left-hand side of (2.6) centered by a conditional expectation term and of this expectation term. For the first part we use a variant of the Efron-Stein inequality, for the second part we use the optimality of H_n, the Hoeffding and McDiarmid inequalities, and the well-known weak universal consistency of kernel estimates with deterministic bandwidths.

Remark 1. We briefly compare Theorems 2.1 and 2.2. The conditions on the kernel are slightly different in both theorems. Theorem 2.2 deals with a special data-dependent bandwidth device (cross-validation) and has a restriction for the numbers of possible bandwidth values, but requires with (2.5) the weakest possible condition.

Remark 2. We want to stress that in Theorems 2.1 and 2.2 there is no assumption on the underlying distribution of (X,Y) besides $|Y| \le L < \infty$. In particular it is not required that X have a density with respect to the Lebesgue-Borel measure or that m be (in some sense) smooth.

Remark 3. Theorem 2.1 still holds if one replaces (2.1), (2.2) and (2.3) by

$$H \to 0 \text{ a.s.} \quad \text{and} \quad \frac{n \cdot H^d}{\log n} \to \infty \text{ a.s.} \tag{2.7}$$

Indeed, proceeding as on pages 158–159 in Devroye and Györfi (1985) one can conclude from (2.7) that there exists $h_{\min}(n), h_{\max}(n) \in \mathbb{R}_+$ which satisfy (2.2), (2.3) and

$$I_{\{H \notin [h_{\min}(n), h_{\max}(n)]\}} \to 0 \quad \text{a.s.}$$

From this and

$$\int |m_n(x) - m(x)|^2 \mu(dx)$$
$$\le 4L^2 I_{\{H \notin [h_{\min}(n), h_{\max}(n)]\}} + \int |m_n(x) - m(x)|^2 \mu(dx) \cdot I_{\{H \in [h_{\min}(n), h_{\max}(n)]\}}$$

one gets the assertion as in the proof of Theorem 2.1.

3. Proof of Theorem 2.1

In the proof we use the following lemma of Greblicki *et al.* (1984)

LEMMA 3.1. *Assume*

$$\begin{aligned} &c_1 G(\|x\|) \le K(x) \le c_2 G(\|x\|), \qquad c_1, c_2 > 0 \\ &G(+0) > 0 \\ &t^d G(t) \to 0 \quad \textit{as} \quad t \to \infty \end{aligned}$$

where G is a nonincreasing Borel function on $[0,\infty)$. Then for all μ-integrable functions f

$$\lim_{h\to 0}\frac{\int K((x-z)/h)f(z)\mu(dz)}{\int K((x-z)/h)\mu(dz)} = f(x) \text{ mod } \mu.$$

We will also need the following bound on the covering number of the class of functions

$$\mathcal{G} = \{g_1 \cdot g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\},$$

where

$$\mathcal{G}_1 = \{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = |T_L y - a|^2((x,y) \in \mathbb{R}^d \times \mathbb{R}) \text{ for some } a \in [-L,L]\}$$

and

$$\mathcal{G}_2 = \left\{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = K\left(\frac{u-x}{h}\right)((x,y) \in \mathbb{R}^d \times \mathbb{R}) \right.$$
$$\left. \text{for some } u \in \mathbb{R}^d, h \in \mathbb{R}_+\right\}$$

(see Definition A.1 in the Appendix for the definition of the covering number).

LEMMA 3.2. *Set $(X,Y)_1^n = ((X_1,Y_1),\dots,(X_n,Y_n))$. Let $L > 0$ and let $\mathcal{G}$ be defined as above. Then for all $0 < \epsilon < 2L^2K(0)$*

$$\mathcal{N}_1(\epsilon, \mathcal{G}, (X,Y)_1^n) \le \left(\frac{c_3}{\epsilon}\right)^{c_4},$$

where c_3 and c_4 are constants which depend only on d, L and $K(0)$.

PROOF OF LEMMA 3.2. The functions in $\mathcal{G}_1$ and $\mathcal{G}_2$ are bounded in absolute value by $4L^2$ and $K(0)$, respectively. Hence by Lemma A.2 in the Appendix we get

$$\mathcal{N}_1(\epsilon,\mathcal{G},(X,Y)_1^n) \le \mathcal{N}_1\left(\frac{\epsilon}{2K(0)},\mathcal{G}_1,(X,Y)_1^n\right)\cdot \mathcal{N}_1\left(\frac{\epsilon}{8L^2},\mathcal{G}_2,(X,Y)_1^n\right).$$

If $h_i(x,y) = |a_i - T_L y|^2$ for some $a_i \in [-L,L]$ $(i = 1,2)$, then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n |h_1(X_i,Y_i) - h_2(X_i,Y_i)| &= \frac{1}{n}\sum_{i=1}^n |a_1 - T_LY_i + a_2 - T_LY_i|\cdot|a_1-a_2| \\ &\le 4L\cdot|a_1-a_2|\end{aligned}$$

which implies

$$\begin{aligned}\mathcal{N}_1\left(\frac{\epsilon}{2K(0)},\mathcal{G}_1,(X,Y)_1^n\right) &\le \mathcal{N}_1\left(\frac{\epsilon}{8K(0)L},\{a:|a|\le L\},X_1^n\right) \\ &\le \frac{2L}{\epsilon/(8K(0)L)} = \frac{16K(0)L^2}{\epsilon}.\end{aligned}$$

Next we bound

$$\mathcal{N}_1\left(\frac{\epsilon}{8L^2},\mathcal{G}_2,(X,Y)_1^n\right).$$

By Lemma A.3 in the Appendix, which uses the notion of VC dimension introduced in Definition A.2 in the Appendix, we get

$$\mathcal{N}_1\left(\frac{\epsilon}{8L^2}, \mathcal{G}_2, (X,Y)_1^n\right) \le 2\left(\frac{4eK(0)}{\frac{\epsilon}{8L^2}}\right)^{2V_{\mathcal{G}_2^+}}.$$

Hence it suffices to derive a bound on the VC dimension of the class of all subgraphs of

$$\mathcal{G}_2 = \left\{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = \tilde{K}\left(\frac{\|u-x\|^2}{h^2}\right)\ ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \right.$$
$$\left. \text{for some } u \in \mathbb{R}^d, h \in \mathbb{R}_+ \right\}.$$

Since $\tilde{K}$ is left continuous and monotone decreasing we have

$$\tilde{K}\left(\frac{\|u-x\|^2}{h^2}\right) \ge t \quad \text{if and only if} \quad \frac{\|u-x\|^2}{h^2} \le \phi(t)$$

where

$$\phi(t) = \sup\{z : \tilde{K}(z) \ge t\}.$$

Equivalently, (x,y,t) must satisfy

$$x^Tx - 2u^Tx + u^Tu - h^2\phi(t) \le 0.$$

Consider now the set of real functions

$$\mathcal{G}_3 = \{g_{\alpha,\beta,\gamma,\delta} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} : g_{\alpha,\beta,\gamma,\delta}(x,y,s) = \alpha x^Tx + \beta^Tx + \gamma s + \delta$$
$$((x,y,s) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}) \text{ for some } \alpha,\gamma,\delta \in \mathbb{R}, \beta \in \mathbb{R}^d\}.$$

If for a given collection of points $\{(x_i,y_i,t_i)\}_{i=1,\dots,n}$ a set $\{(x,y,t) : g(x,y) \ge t\}$, $g \in \mathcal{G}_2$, picks out the points $\{(x_{i_1},y_{i_1},t_{i_1}),\dots,(x_{i_l},y_{i_l},t_{i_l})\}$, i.e.,

$$\{(x,y,t) : g(x,y) \ge t\} \cap \{(x_i,y_i,t_i)\}_{i=1,\dots,n} = \{(x_{i_1},y_{i_1},t_{i_1}),\dots,(x_{i_l},y_{i_l},t_{i_l})\},$$

then there exist $\alpha,\beta,\gamma,\delta$ such that $\{(x,y,s) : g_{\alpha,\beta,\gamma,\delta}(x,y,s) \ge 0\}$ picks out exactly $\{(x_{i_1},y_{i_1},\phi(t_{i_1})),\dots,(x_{i_l},y_{i_l},\phi(t_{i_l}))\}$ from $\{(x_1,y_1,\phi(t_1)),\dots,(x_n,y_n,\phi(t_n))\}$. This shows

$$V_{\mathcal{G}_2^+} \le V_{\{\{(x,y,s):g(x,y,s)\ge 0\}:g\in\mathcal{G}_3\}}.$$

$\mathcal{G}_3$ is a linear vector space of dimension $d+3$, hence we can conclude from Lemma A.4 in the Appendix

$$V_{\mathcal{G}_2^+} \le d+3.$$

Summarizing the above results we get

$$\mathcal{N}_1(\epsilon,\mathcal{G},(X,Y)_1^n) \le \frac{16K(0)L^2}{\epsilon} \cdot 2\left(\frac{4eK(0)}{\frac{\epsilon}{8L^2}}\right)^{2(d+3)} = \left(\frac{c_3}{\epsilon}\right)^{c_4}$$

for constants c_3 and c_4 which depend only on d, L and $K(0)$. □

PROOF OF THEOREM 2.1. It suffices to show that

$$m_n(x) \to m(x) \qquad (n \to \infty) \qquad \text{a.s. mod } \mu.$$

We have

$$\begin{aligned}
&|m_n(x) - m(x)|^2 - \frac{\int |m_n(x) - m(z)|^2 K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)} \\
&= |m_n(x)|^2 - |m_n(x)|^2 \frac{\int K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)} \\
&\quad - 2m_n(x)\left(m(x) - \frac{\int m(z) K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)}\right) \\
&\quad + |m(x)|^2 - \frac{\int m(z)^2 K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)} \\
&= 0 - 2m_n(x)\left(m(x) - \frac{\int m(z) K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)}\right) \\
&\quad + |m(x)|^2 - \frac{\int m(z)^2 K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)} \\
&\to 0 \qquad \text{a.s. mod } \mu
\end{aligned}$$

as $H \to 0$ a.s. by Lemma 3.1 (the conditions there are satisfied for $c_1 = c_2 = 1$ and $G(z) = \tilde{K}(z^2)$). Hence it suffices to show

$$\frac{\int |m_n(x) - m(z)|^2 K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)} \to 0 \qquad \text{a.s. mod } \mu. \tag{3.1}$$

Let $\epsilon > 0$. We have

$$\begin{aligned}
&\int |m_n(x) - m(z)|^2 K_H(x-z)\mu(dz) \\
&\quad = \boldsymbol{E}\{|Y - m_n(x)|^2 K_H(x-X) \mid \mathcal{D}_n\} - \boldsymbol{E}\{|Y - m(X)|^2 K_H(x-X) \mid \mathcal{D}_n\} \\
&\quad = \sum_{j=1}^{5} T_{j,n}
\end{aligned}$$

where

$$T_{1,n} = \boldsymbol{E}\{|Y - m_n(x)|^2 K_H(x-X) \mid \mathcal{D}_n\} - (1+\epsilon) \cdot \frac{1}{n}\sum_{i=1}^{n} |Y_i - m_n(x)|^2 K_H(x - X_i),$$

$$T_{2,n} = (1+\epsilon)\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i - m_n(x)|^2 K_H(x-X_i) - \frac{1}{n}\sum_{i=1}^{n} |Y_i - m(x)|^2 K_H(x-X_i)\right),$$

$$\begin{aligned}
T_{3,n} = {} & (1+\epsilon) \cdot \frac{1}{n}\sum_{i=1}^{n} |Y_i - m(x)|^2 K_H(x-X_i) - (1+\epsilon)^2 \\
& \cdot \boldsymbol{E}\{|Y - m(x)|^2 K_H(x-X) \mid \mathcal{D}_n\},
\end{aligned}$$

$$
\begin{aligned}
T_{4,n} &= (1+\epsilon)^2 \cdot (\mathbf{E}\{|Y-m(x)|^2 K_H(x-X) \mid \mathcal{D}_n\} \\
&\qquad\qquad - \mathbf{E}\{|Y-m(X)|^2 K_H(x-X) \mid \mathcal{D}_n\}), \\
T_{5,n} &= ((1+\epsilon)^2 - 1) \cdot \mathbf{E}\{|Y-m(X)|^2 K_H(x-X) \mid \mathcal{D}_n\}.
\end{aligned}
$$

In the remainder of the proof we bound

$$\frac{T_{j,n}}{\int K_H(x-z)\mu(dz)}, \quad j=1,\dots,5.$$

We start with showing that

$$\limsup_{n\to\infty} \frac{T_{1,n}}{\int K_H(x-z)\mu(dz)} \le 0 \quad \text{a.s. mod } \mu.$$

It follows from (2.2) and the proof of Lemma 2.2. in Devroye (1981) that there exists a function g which satisfies $g(x)>0$ mod μ and

$$\int K_H(x-z)\mu(dz) \to g(x) \quad \text{a.s. mod } \mu,$$

and thus we only need to show that

$$\limsup_{n\to\infty} T_{1,n} \le 0 \quad \text{a.s. mod } \mu. \tag{3.2}$$

To prove (3.2), fix $t>0$. Then by Lemma A.1 in the Appendix and by Lemma 3.2 we have for n sufficiently large

$$
\begin{aligned}
&\mathbf{P}\{T_{1,n} > t \mid \mathcal{D}_n\} \\
&= \mathbf{P}\left\{ \frac{\mathbf{E}\{|Y-m_n(x)|^2 K_H(x-X) \mid \mathcal{D}_n\} - \frac{1}{n}\sum_{i=1}^n |Y_i - m_n(x)|^2 K_H(x-X_i)}{t + \epsilon \cdot \mathbf{E}\{|Y-m_n(x)|^2 K_H(x-X) \mid \mathcal{D}_n\}} \right. \\
&\qquad\qquad \left. > \frac{1}{1+\epsilon} \,\middle|\, \mathcal{D}_n \right\} \\
&\le \mathbf{P}\left\{ \exists a \in [-L,L], \exists h \in [h_{\min}(n), h_{\max}(n)] : \right. \\
&\qquad \left. \frac{\mathbf{E}\left\{|Y-a|^2 K\left(\frac{x-X}{h}\right)\right\} - \frac{1}{n}\sum_{i=1}^n |Y_i - a|^2 K\left(\frac{x-X_i}{h}\right)}{\frac{h_{\min}^d(n)\cdot t}{\epsilon} + \cdot \mathbf{E}\left\{|Y-a|^2 K\left(\frac{x-X}{h}\right)\right\}} > \frac{\epsilon}{1+\epsilon} \,\middle|\, \mathcal{D}_n \right\} \\
&\le 4 \cdot \mathbf{E}\mathcal{N}_1\left(\frac{h_{\min}^d(n)\cdot t}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \cdot \exp\left(- \frac{n \cdot \frac{t\cdot h_{\min}^d(n)}{\epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon}\right)^2}{64 K(0) L^2} \right) \\
&\le 4 \left(\frac{8c_3(1+\epsilon)}{h_{\min}^d(n)\cdot t}\right)^{c_4} \cdot \exp\left(-\frac{n h_{\min}^d(n)\cdot t \cdot \epsilon}{64 L^2 K(0)(1+\epsilon)^2}\right).
\end{aligned}
$$

This together with (2.3) and the Borel-Cantelli lemma implies (3.2).

It is easy to see that the kernel estimate satisfies for all $x \in \mathbb{R}^d$

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(x)-Y_i|^2 K_H(x-X_i)=\min_{a\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}|a-Y_i|^2 K_H(x-X_i),$$

which implies

$$T_{2,n}\le 0 \quad \text{a.s.}$$

Furthermore, using a similar argument as for $T_{1,n}$ we get

$$\limsup\frac{T_{3,n}}{\int K_H(x-z)\mu(dz)}\le 0 \quad \text{a.s. mod } \mu.$$

Next by Lemma 3.1

$$\begin{aligned}\frac{1}{(1+\epsilon^2)}\frac{T_{4,n}}{\int K_H(x-z)\mu(dz)}&=\frac{\int|m(z)-m(x)|^2K_H(x-z)\mu(dz)}{\int K_H(x-z)\mu(dz)}\\ &\to |m(x)-m(x)|^2=0 \quad \text{a.s. mod } \mu.\end{aligned}$$

Finally,

$$\frac{T_{5,n}}{\int K_H(x-z)\mu(dz)}\le 4L^2((1+\epsilon^2)-1)\to 0$$

as $\epsilon\to 0$. Thus (3.1) has been shown. □

4. Proof of Theorem 2.2

In the proof of Theorem 2.2 we need the following two lemmas.

LEMMA 4.1. a) *There is a constant $c>0$ depending only on the boxed kernel $K:\mathbb{R}^d\to\mathbb{R}_+$ (in Theorem 2.2) such that*

$$\sum_{i=1}^{n-1}\frac{K\left(\frac{x_i-x_n}{h}\right)}{\sum_{j\in\{1,\dots,n\}\setminus\{i\}}K\left(\frac{x_i-x_j}{h}\right)}\le c \tag{4.1}$$

for each $n\ge 2$, $(x_1,\dots,x_n)\in\mathbb{R}^{dn}$, $h>0$.

b) *For each $q\in\mathbb{N}$, there is a constant $c>0$ depending only on q and on the boxed kernel $K:\mathbb{R}^d\to\mathbb{R}_+$ such that*

$$\boldsymbol{E}\left[\int\frac{K\left(\frac{x-X_n}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}\mu(dx)\right]^{2q}\le\frac{c}{n^{2q}} \tag{4.2}$$

for each $n\ge 1$, $h>0$.

PROOF OF LEMMA 4.1. We use a known covering argument from kernel regression estimation, see Devroye and Wagner (1980), Spiegelman and Sacks (1980), Devroye and

Krzyżak (1989). Let $z_1+S_{0,r/2},\dots,z_M+S_{0,r/2}$ be a finite cover of $S_{0,R}$. Assume without loss of generality that $K\le 1$.

a) The left-hand side of (4.1) is bounded above by

$$\sum_{k=1}^{M}\sum_{i=1}^{n-1}\frac{I_{z_k+S_{0,r/2}}\left(\frac{x_i-x_n}{h}\right)K\left(\frac{x_i-x_n}{h}\right)}{b\sum_{j\in\{1,\dots,n-1\}\setminus\{i\}}I_{S_{0,r}}\left(\frac{x_i-x_j}{h}\right)+K\left(\frac{x_i-x_n}{h}\right)}$$

$$\le\sum_{k=1}^{M}\sum_{i=1}^{n-1}\frac{I_{x_n+hz_k+S_{0,rh/2}}(x_i)}{1+b\sum_{j\in\{1,\dots,n-1\}\setminus\{i\}}I_{x_i+S_{0,rh}}(x_j)}$$

(because $z/(z+w)\le 1/(1+w)$ for $0\le z\le 1, w\ge 0$)

$$\le\sum_{k=1}^{M}\sum_{i=1}^{n-1}\frac{I_{x_n+hz_k+S_{0,rh/2}}(x_i)}{1+b\sum_{j\in\{1,\dots,n-1\}\setminus\{i\}}I_{x_n+hz_k+S_{0,rh/2}}(x_j)}$$

(because $x_i\in x_n+hz_k+S_{0,rh/2}$ implies $x_n+hz_k+S_{0,rh/2}\subset x_i+S_{0,rh}$)

$$\le\frac{M}{b}.$$

b) Let $n\ge 2q+2$, $0<b<\frac{1}{2q}$ without loss of generality. Using similar arguments as in a) we can bound the left-hand side of (4.2) above by

$$\mathbf{E}\left[\sum_{k=1}^{M}\int\frac{I_{z_k+S_{0,r/2}}\left(\frac{x-X_n}{h}\right)}{1+b\sum_{j=2q+1}^{n-1}I_{S_{0,r}}\left(\frac{x-X_j}{h}\right)}\mu(dx)\right]^{2q}$$

$$=\mathbf{E}\left\{\sum_{k_1,\dots,k_{2q}\in\{1,\dots,M\}}\int\frac{I_{z_{k_1}+S_{0,r/2}}\left(\frac{x-X_n}{h}\right)}{1+b\sum_{j=2q+1}^{n-1}I_{S_{0,r}}\left(\frac{x-X_j}{h}\right)}\mu(dx)\dots\int\frac{I_{z_{k_{2q}}+S_{0,r/2}}\left(\frac{x-X_n}{h}\right)}{1+b\sum_{j=2q+1}^{n-1}I_{S_{0,r}}\left(\frac{x-X_j}{h}\right)}\mu(dx)\right\}$$

$$=\mathbf{E}\sum_{k_1,\dots,k_{2q}\in\{1,\dots,M\}}\frac{I_{z_{k_1}+S_{0,r/2}}\left(\frac{X_1-X_n}{h}\right)\cdots I_{z_{k_{2q}}+S_{0,r/2}}\left(\frac{X_{2q}-X_n}{h}\right)}{\left[1+b\sum_{j=2q+1}^{n-1}I_{S_{0,r}}\left(\frac{X_1-X_j}{h}\right)\right]\cdots\left[1+b\sum_{j=2q+1}^{n-1}I_{S_{0,r}}\left(\frac{X_{2q}-X_j}{h}\right)\right]}$$

(by independence of $X_1,\dots,X_{2q},X_{2q+1},\dots,X_n$)

$$\le\frac{1}{b^{2q}}\mathbf{E}\sum_{k_1,\dots,k_{2q}\in\{1,\dots,M\}}\frac{I_{z_{k_1}+S_{0,r/2}}\left(\frac{X_1-X_n}{h}\right)\cdots I_{z_{k_{2q}}+S_{0,r/2}}\left(\frac{X_{2q}-X_n}{h}\right)}{\sum_{j=1}^{n-1}I_{S_{0,r}}\left(\frac{X_1-X_j}{h}\right)\cdots\sum_{j=1}^{n-1}I_{S_{0,r}}\left(\frac{X_{2q}-X_j}{h}\right)}$$

$$\left(\text{because } 1 > 2qb, 2q > \sum_{j=1}^{2q} I_{S_{0,r}}\left(\frac{X_l - X_j}{h}\right) \quad (l = 1, \dots, 2q)\right)$$

$$= \frac{1}{b^{2q}(n-1)\dots(n-2q)} \boldsymbol{E} \sum_{k_1,\dots,k_{2q}\in\{1,\dots,M\}} \sum_{\substack{l_1,\dots,l_{2q}\in\{1,\dots,n-1\}\\ l_i\neq l_{i'}}}$$

$$\frac{I_{z_{k_1}+S_{0,r/2}}\left(\frac{X_{l_1}-X_n}{h}\right)\cdots I_{z_{k_{2q}}+S_{0,r/2}}\left(\frac{X_{l_{2q}}-X_n}{h}\right)}{\sum_{j=1}^{n-1} I_{S_{0,r}}\left(\frac{X_{l_1}-X_j}{h}\right)\cdots\sum_{j=1}^{n-1} I_{S_{0,r}}\left(\frac{X_{l_{2q}}-X_n}{h}\right)}$$

(by exchangeability of $X_1, \dots, X_{n-1}$)

$$\leq \frac{1}{b^{2q}(n-1)\cdots(n-2q)} \boldsymbol{E} \sum_{k_1,\dots,k_{2q}\in\{1,\dots,M\}} \sum_{i=1}^{n-1} \frac{I_{z_{k_1}+S_{0,r/2}}\left(\frac{X_i-X_n}{h}\right)}{\sum_{j=1}^{n-1} I_{S_{0,r}}\left(\frac{X_i-X_j}{h}\right)} \cdots$$

$$\sum_{i=1}^{n-1} \frac{I_{z_{k_{2q}}+S_{0,r/2}}\left(\frac{X_i-X_n}{h}\right)}{\sum_{j=1}^{n-1} I_{S_{0,r}}\left(\frac{X_i-X_j}{h}\right)}$$

$$\leq \frac{M^{2q}}{b^{2q}(n-1)\cdots(n-2q)}. \qquad \square$$

The following lemma contains an inequality of Efron-Stein (1981) and Steele (1986) type for higher central moments in the case of identically distributed random variables and symmetric statistics.

LEMMA 4.2. *Let $Z_1, \dots, Z_n, \tilde{Z}_n$ be independent identically distributed random variables with values in some Borel set $A \subset \mathbb{R}^m$, and let the functions $f = f_n : A^n \to \mathbb{R}$ be measurable and symmetric (i.e., the function values are not changed by a permutation of the arguments). Let $q \in \mathbb{N}$. If $f(Z_1, \dots, Z_n) \in \mathcal{L}_{2q}$, then a constant $C \in \mathbb{R}_+$ (dependent only on q, but not on n or f) exists such that*

$$\begin{aligned}(4.3)\qquad &\boldsymbol{E}\{|f(Z_1, \dots, Z_n) - \boldsymbol{E}f(Z_1, \dots, Z_n)|^{2q}\} \\ &\leq Cn^q \boldsymbol{E}\{|f(Z_1, \dots, Z_n) - f(Z_1, \dots, Z_{n-1}, \tilde{Z}_n)|^{2q}\}.\end{aligned}$$

PROOF OF LEMMA 4.2. We use arguments from Devroye *et al.* (1996), pp. 136, 137. First we notice existence of a constant $C = C(q) > 1$ such that

$$(4.4)\qquad \sum_{j=2}^{2q} \binom{2q}{j} C^{(2q-j)/2q} n^{(2q-j)/2} + Cn^q \leq C(n+1)^q$$

for all $n \in \mathbb{N}$, since this relation holds for $C > (2q-1)^q$, if n is sufficiently large, say $n > n_0(q)$, and for each $n \in \{1, \dots, n_0(q)\}$, if C is sufficiently large. Now let $n \in \mathbb{N}$ be fixed. Set

$$\begin{aligned}
&V^{(n)} = f_n(Z_1,\dots,Z_n) - \mathbf{E}f_n(Z_1,\dots,Z_n),\\
&V_1^{(n)} = \mathbf{E}\{V^{(n)}|Z_1\}\\
&V_k^{(n)} = \mathbf{E}\{V^{(n)}|Z_1,\dots,Z_k\} - \mathbf{E}\{V^{(n)}|Z_1,\dots,Z_{k-1}\}, \qquad k \in \{2,\dots,n\},\\
&d_n = \mathbf{E}\{|f_n(Z_1,\dots,Z_n) - f_n(Z_1,\dots,Z_{n-1},\tilde{Z}_n)|^{2q}\}.
\end{aligned}$$

Then $V^{(n)} = V_1^{(n)} + \dots + V_n^{(n)}$ and $V_1^{(n)},\dots,V_n^{(n)}$ form a martingale difference sequence with respect to $Z_1,\dots,Z_n$. We prove (4.3) for the above C by induction on n. (4.3) holds for $n = 1$, because

$$\begin{aligned}
&\mathbf{E}\{|f_1(Z_1) - \mathbf{E}f_1(Z_1)|^{2q}\}\\
&\quad = \int |u - \mathbf{E}f_1(Z_1)|^{2q}\mathbf{P}_{f_1(\tilde{Z}_1)}(du)\\
&\quad \le \int \mathbf{E}\{|u - f_1(Z_1)|^{2q}\}\mathbf{P}_{f_1(\tilde{Z}_1)}(du) \qquad \text{(by Jensen's inequality)}\\
&\quad = \mathbf{E}\{|f_1(\tilde{Z}_1) - f_1(Z_1)|^{2q}\}.
\end{aligned}$$

Assume (4.3) for fixed n. To show (4.3) for $n+1$ rather than for n we use the martingale property and Hölder's inequality and obtain

$$\begin{aligned}
&\mathbf{E}\{|V^{(n+1)}|^{2q}\}\\
&\quad = \mathbf{E}\left\{\left|V_{n+1}^{(n+1)} + \sum_{i=1}^{n} V_i^{(n+1)}\right|^{2q}\right\}\\
&\quad \le \sum_{j=2}^{2q}\binom{2q}{j}\mathbf{E}\left\{\left|\sum_{i=1}^{n} V_i^{(n+1)}\right|^{2q-j}|V_{n+1}^{(n+1)}|^j\right\} + \mathbf{E}\left\{\left|\sum_{i=1}^{n} V_i^{(n+1)}\right|^{2q}\right\}\\
&\quad \le \sum_{j=2}^{2q}\binom{2q}{j}\left(\mathbf{E}\left\{\left|\sum_{i=1}^{n} V_i^{(n+1)}\right|^{2q}\right\}\right)^{(2q-j)/2q}(\mathbf{E}\{|V_{n+1}^{(n+1)}|^{2q}\})^{j/2q}\\
&\quad\quad + \mathbf{E}\left\{\left|\sum_{i=1}^{n} V_i^{(n+1)}\right|\right\}^{2q}.
\end{aligned}$$

We notice

$$\begin{aligned}
&\mathbf{E}\left\{\left|\sum_{i=1}^{n} V_i^{(n+1)}\right|^{2q}\right\}\\
&\quad = \mathbf{E}\{|\mathbf{E}\{f_{n+1}(Z_1,\dots,Z_{n+1})|Z_1,\dots,Z_n\}\\
&\qquad\quad - \mathbf{E}\mathbf{E}\{f_{n+1}(Z_1,\dots,Z_{n+1})|Z_1,\dots,Z_n\}|^{2q}\}\\
&\quad \le Cn^q\mathbf{E}\{|\mathbf{E}\{f_{n+1}(Z_1,\dots,Z_{n+1})|Z_1,\dots,Z_n\}\\
&\qquad\quad - \mathbf{E}\{f_{n+1}(Z_1,\dots,\tilde{Z}_n,Z_{n+1})|Z_1,\dots,Z_{n-1},\tilde{Z}_n\}|^{2q}\}
\end{aligned}$$

(by the induction assumption applied to

$$\begin{aligned}
&g(z_1,\ldots,z_n) = \boldsymbol{E}\{f_{n+1}(Z_1,\ldots,Z_{n+1}) \mid Z_1 = z_1,\ldots,Z_n = z_n\})\\
&= Cn^q\boldsymbol{E}\left\{\left|\int f_{n+1}(Z_1,\ldots,Z_n,z) - f_{n+1}(Z_1,\ldots,Z_{n-1},\tilde{Z}_n,z)dP_{Z_n}(z)\right|^{2q}\right\}\\
&\le Cn^q\boldsymbol{E}\int |f_{n+1}(Z_1,\ldots,Z_n,z) - f_{n+1}(Z_1,\ldots,Z_{n-1},\tilde{Z}_n,z)|^{2q}dP_{Z_n}(z)\\
&\quad\text{(by Jensen's inequality)}\\
&= Cn^q\boldsymbol{E}\{|f_{n+1}(Z_1,\ldots,Z_{n+1}) - f_{n+1}(Z_1,\ldots,Z_{n-1},\tilde{Z}_n,Z_{n+1})|^{2q}\}\\
&= Cn^q d_{n+1},
\end{aligned}$$

further

$$\begin{aligned}
&\boldsymbol{E}\{|V_{n+1}^{(n+1)}|^{2q}\}\\
&\quad= \boldsymbol{E}\{|f_{n+1}(Z_1,\ldots,Z_{n+1}) - \boldsymbol{E}\{f_{n+1}(Z_1,\ldots,Z_{n+1}) \mid Z_1,\ldots,Z_n\}|^{2q}\}\\
&\quad= \boldsymbol{E}\boldsymbol{E}\{|f_{n+1}(Z_1,\ldots,Z_{n+1})\\
&\qquad\qquad - \boldsymbol{E}\{f_{n+1}(Z_1,\ldots,Z_{n+1}) \mid Z_1,\ldots,Z_n\}|^{2q} \mid Z_1,\ldots,Z_n\}\\
&\quad\le \boldsymbol{E}\boldsymbol{E}\{|f_{n+1}(Z_1,\ldots,Z_{n+1}) - f_{n+1}(Z_1,\ldots,Z_n,\tilde{Z}_{n+1})|^{2q} \mid Z_1,\ldots,Z_n\}\\
&\qquad\text{(see above proof of (4.1) for } n = 1)\\
&\quad= d_{n+1}.
\end{aligned}$$

These results together with (4.4) yield

$$\begin{aligned}
\boldsymbol{E}\{|V^{(n+1)}|^{2q}\} &\le \left(\sum_{j=2}^{2q}\binom{2q}{j}(Cn^q)^{(2q-j)/(2q)} + Cn^q\right)d_{n+1}\\
&\le C(n+1)^q d_{n+1},
\end{aligned}$$

i.e., (4.3) for $n+1$. □

PROOF OF THEOREM 2.2. It suffices to show (2.6). Set

$$F_n^{(h)} = \int |m_n^{(h)}(x) - m(x)|^2\mu(dx) - \boldsymbol{E}\int |m_n^{(h)}(x) - m(x)|^2\mu(dx)$$

and

$$\triangle_n^{(h)} = \boldsymbol{E}\int |m(x) - m_n^{(h)}(x)|^2\mu(dx), \qquad h > 0.$$

Now it remains to show

(4.5) $$F_n^{(H_n)} \to 0 \quad (n \to \infty) \quad \text{a.s.},$$

and

(4.6) $$\triangle_n^{(H_n)} \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

In order to show (4.5) we choose $q \in \mathbb{N}$ satysfying $q > \tau + 1$ with τ from (2.4). To obtain an upper bound for $\boldsymbol{E}\{|F_n^{(h)}|^{2q}\}$, $h \in \mathcal{Q}_n$, we use Lemma 4.2. Let $\tilde{m}_n^{(h)}$ be obtained from $m_n^{(h)}$ via replacing (X_n, Y_n) by $(\tilde{X}_n, \tilde{Y}_n)$, where $(X_1, Y_1), \ldots, (X_n, Y_n), (\tilde{X}_n, \tilde{Y}_n)$

are independent identically distributed random vectors. A straightforward computation shows

$$(4.7)\quad |m_n^{(h)}(x) - \tilde m_n^{(h)}(x)| \le 2L \frac{K\left(\frac{x-X_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} + 2L \frac{K\left(\frac{x-\tilde X_n}{h}\right)}{\sum_{i=1}^{n-1} K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x-\tilde X_n}{h}\right)}.$$

Thus we obtain by Lemma 4.2

$$\begin{aligned} &\boldsymbol{E}\{|F_n^{(h)}|^{2q}\} \\ &\le c^* n^q \boldsymbol{E}\left\{\left|\int [|m_n^{(h)}(x) - m(x)|^2 - |\tilde m_n^{(h)}(x) - m(x)|^2]\mu(dx)\right|^{2q}\right\} \\ &\le (4L)^{2q} c^* n^q \boldsymbol{E}\left\{\left[\int |m_n^{(h)}(x) - \tilde m_n^{(h)}(x)|\mu(dx)\right]^{2q}\right\} \\ &\le (4L)^{4q} c^* n^q \boldsymbol{E}\left\{\left[\int \frac{K\left(\frac{x-X_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}\mu(dx)\right]^{2q}\right\} \\ &\le c^{**}\frac{1}{n^q} \end{aligned}$$

with suitable constants $c^*, c^{**} \in \mathbb{R}_+$, where the latter inequality follows from Lemma 4.1b). Now for some $c \in \mathbb{R}_+$ and for all $\epsilon' > 0$ we obtain

$$\begin{aligned} \sum_{n=1}^\infty \boldsymbol{P}[F_n^{(H_n)} \ge \epsilon'] &\le \sum_{n=1}^\infty \sum_{h \in \mathcal{Q}_n} \boldsymbol{P}[F_n^{(h)} \ge \varepsilon'] \\ &\le \sum_{n=1}^\infty \sum_{h \in \mathcal{Q}_n} \frac{1}{\varepsilon'^{2q}} \boldsymbol{E}\{|F_n^{(h)}|^{2q}\} \\ &\le c^{**}\frac{1}{\epsilon'^{2q}} \sum_{n=1}^\infty |\mathcal{Q}_n| \frac{1}{n^q} < \infty \end{aligned}$$

(the latter via (2.4)), which yields (4.5).

For the proof of (4.6) choose $h_n^* \in \mathcal{Q}_n$ such that $h_n^* \to 0$, $nh_n^{*^d} \to \infty$ $(n \to \infty)$. We first show

$$(4.8)\qquad \limsup_{n\to\infty} (\Delta_{n-1}^{(H_n)} - \Delta_{n-1}^{(h_{n-1}^*)}) \le 0 \quad \text{a.s.}$$

For arbitrary fixed $\varepsilon' > 0$, we notice

$$\boldsymbol{P}[\Delta_{n-1}^{(H_n)} - \Delta_{n-1}^{(h_{n-1}^*)} > \varepsilon']$$

$$\leq \boldsymbol{P}\left[\triangle_{n-1}^{(H_n)} - \frac{1}{n}\sum_{i=1}^{n}((m_{n,i}^{(H_n)}(X_i) - Y_i)^2 - (m(X_i) - Y_i)^2) + \frac{1}{n}\sum_{i=1}^{n}((m_{n,i}^{(h_{n-1}^*)}(X_i) - Y_i)^2 - (m(X_i) - Y_i)^2) - \triangle_{n-1}^{(h_{n-1}^*)} > \varepsilon'\right],$$

because of the optimality property of H_n, and thus

$$\begin{aligned} &\boldsymbol{P}[\triangle_{n-1}^{(H_n)} - \triangle_{n-1}^{(h_{n-1}^*)} > \varepsilon'] \\ &\leq \boldsymbol{P}\left[2\max_{h\in\mathcal{Q}_n}\left|\frac{1}{n}\sum_{i=1}^{n}((m_{n,i}^{(h)}(X_i) - Y_i)^2 - (m(X_i) - Y_i)^2) - \boldsymbol{E}((m_{n,i}^{(h)}(X_i) - Y_i)^2 - (m(X_i) - Y_i)^2)\right| > \varepsilon'\right] \\ &\leq \sum_{h\in\mathcal{Q}_n}\boldsymbol{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}((m_{n,i}^{(h)}(X_i) - Y_i)^2 - (m(X_i) - Y_i)^2) - \boldsymbol{E}((m_{n,i}^{(h)}(X_i) - Y_i)^2 - (m(X_i) - Y_i)^2)\right| > \frac{\varepsilon'}{2}\right] \\ &\leq \sum_{h\in\mathcal{Q}_n}\boldsymbol{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}((m(X_i) - Y_i)^2 - \boldsymbol{E}(m(X_i) - Y_i)^2)\right| > \frac{\varepsilon'}{4}\right] \\ &\quad + \sum_{h\in\mathcal{Q}_n}\boldsymbol{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}((m_{n,i}^{(h)}(X_i) - Y_i)^2 - \boldsymbol{E}(m_{n,i}^{(h)}(X_i) - Y_i)^2)\right| > \frac{\varepsilon'}{4}\right] \\ &= C_n + D_n. \end{aligned}$$

By Hoeffding's inequality (Lemma A.5 in the Appendix)

$$C_n \leq 2|\mathcal{Q}_n|e^{-n\varepsilon^2/(128L^4)}.$$

To bound the summands of D_n we use McDiarmid's inequality (Lemma A.5 in the Appendix). Let $\tilde{m}_{n,i}^{(h)}$ be obtained from $m_{n,i}^{(h)}$ via replacing (X_n, Y_n) by $(\tilde{X}_n, \tilde{Y}_n)$, where $(X_1, Y_1), \dots, (X_n, Y_n), (\tilde{X}_n, \tilde{Y}_n)$ are independent identically distributed random vectors $(i = 1, \dots, n-1)$. We set

$$V_n = \sum_{i=1}^{n}(m_{n,i}^{(h)}(X_i) - Y_i)^2 - \left[\sum_{i=1}^{n-1}(\tilde{m}_{n,i}^{(h)}(X_i) - Y_i)^2 + (m_{n,n}(\tilde{X}_n) - \tilde{Y}_n)^2\right].$$

In order to be able to apply Lemma A.5, we show next

$$|V_n| \leq c$$

for some $c > 0$ independent of n. Let

$$U_i = m_{n,i}^{(h)}(X_i) - Y_i$$

$$= \sum_{l\in\{1,\dots,n\}\setminus\{i\}} \frac{Y_l K\left(\frac{X_i - X_l}{h}\right)}{\sum_{j\in\{1,\dots,n\}\setminus\{i\}} K\left(\frac{X_i - X_j}{h}\right)} - Y_i,$$

$$W_i = \tilde{m}_{n,i}^{(h)}(X_i) - Y_i$$

$$= \sum_{l\in\{1,\dots,n-1\}\setminus\{i\}} \frac{Y_l K\left(\frac{X_i - X_l}{h}\right)}{\sum_{j\in\{1,\dots,n-1\}\setminus\{i\}} K\left(\frac{X_i - X_j}{h}\right) + K\left(\frac{X_i - \tilde{X}_n}{h}\right)}$$

$$+ \frac{\tilde{Y}_n K\left(\frac{X_i - \tilde{X}_n}{h}\right)}{\sum_{j\in\{1,\dots,n-1\}\setminus\{i\}} K\left(\frac{X_i - X_j}{h}\right) + K\left(\frac{X_i - \tilde{X}_n}{h}\right)} - Y_i$$

for $i = 1, \dots, n-1$. Thus

$$V_n = \sum_{i=1}^{n-1} (U_i^2 - W_i^2) + |m_{n,n}^{(h)}(X_n) - Y_n|^2 - |m_{n,n}^{(h)}(\tilde{X}_n) - \tilde{Y}_n|^2,$$

$$V_n^2 \leq 3\left|\sum_{i=1}^{n-1}(U_i^2 - W_i^2)\right|^2 + 3|m_{n,n}^{(h)}(X_n) - Y_n|^4 + 3|m_{n,n}^{(h)}(\tilde{X}_n) - \tilde{Y}_n|^4.$$

We obtain

$$|U_i| \leq 2L, \qquad |W_i| \leq 2L,$$

$$|U_i - W_i| \leq 2L \frac{K\left(\frac{X_i - X_n}{h}\right)}{\sum_{j\in\{1,\dots,n\}\setminus\{i\}} K\left(\frac{X_i - X_j}{h}\right)}$$

$$+ 2L \frac{K\left(\frac{X_i - \tilde{X}_n}{h}\right)}{\sum_{j\in\{1,\dots,n-1\}\setminus\{i\}} K\left(\frac{X_i - X_j}{h}\right) + K\left(\frac{X_i - \tilde{X}_n}{h}\right)}$$

$(i = 1, \dots, n-1)$ (for the latter compare (4.7)), thus, by Lemma 4.1a),

$$\sum_{i=1}^{n-1} |U_i - W_i| \leq c^*$$

for some $c^* > 0$, therefore

$$\left|\sum_{i=1}^{n-1}(U_i^2 - W_i^2)\right|^2 = \left|\sum_{i=1}^{n-1}(U_i - W_i)(U_i + W_i)\right|^2$$

$$\leq 16L^2 {c^*}^2$$

which implies

$$V_n^2 \le c^2 = 48L^2 {c^*}^2 + 96L^4.$$

Now by Lemma A.5

$$D_n \le 2|\mathcal{Q}_n| e^{-n\varepsilon'^2/(8c^2)}.$$

Thus

$$\sum_{n=1}^{\infty} \boldsymbol{P}[\triangle_{n-1}^{(H_n)} - \triangle_{n-1}^{(h_{n-1}^*)} > \varepsilon'] \le \sum_{n=1}^{\infty} (C_n + D_n) < \infty$$

for each $\varepsilon' > 0$, which yields (4.8). Now by the weak universal consistency result of Devroye and Wagner (1980) we have

$$\triangle_n^{(h_n^*)} \to 0. \tag{4.9}$$

We notice

$$\triangle_n^{(H_n)} - \triangle_{n-1}^{(H_n)} \to 0 \qquad (n \to \infty), \tag{4.10}$$

which follows from

$$\begin{aligned}
| \triangle_n^{(h)} - \triangle_{n-1}^{(h)} | &\le 4L\boldsymbol{E} \int |m_n^{(h)}(x) - m_{n-1}^{(h)}(x)| \mu(dx) \\
&\le 8L^2 \int \boldsymbol{E} \frac{K_h(x - X_n)}{\sum_{i=1}^n K_h(x - X_i)} \mu(dx) \\
&\quad \text{(by a straightforward computation as in (4.7))} \\
&= 8L^2 \frac{1}{n} \sum_{j=1}^n \int \boldsymbol{E} \frac{K_h(x - X_j)}{\sum_{i=1}^n K_h(x - X_i)} \mu(dx) \\
&\quad \text{(by exchangeability of } X_1, \ldots, X_n) \\
&\le \frac{8L^2}{n} \qquad (h > 0, n \in \mathbb{N}).
\end{aligned}$$

From (4.8), (4.9), (4.10) we obtain (4.7). □

Acknowledgements

The authors wish to thank two anonymous referees and an associate editor for several helpful comments.

Appendix

A. Some results on empirical process theory

In this section we list the definitions and results of empirical process theory which we have used in Section 3. We also formulate specialized versions of the inequalities of Hoeffding and McDiarmid used in Section 4. An excellent introduction to most of these results can be found in Devroye *et al.* (1996).

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

Definition A.1. Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$. The covering number $\mathcal{N}_1(\epsilon, \mathcal{F}, z_1^n)$ is defined for any $\epsilon > 0$ and $z_1^n = (z_1, \dots, z_n) \in \mathbb{R}^{d \cdot n}$ as the smallest integer k such that there exist functions $g_1, \dots, g_k : \mathbb{R}^d \to \mathbb{R}$ with

$$\min_{1 \le i \le k} \frac{1}{n} \sum_{j=1}^{n} |f(z_j) - g_i(z_j)| \le \epsilon$$

for each $f \in \mathcal{F}$.

If $Z_1^n = (Z_1, \dots, Z_n)$ is a sequence of $\mathbb{R}^d$-valued random variables, then $\mathcal{N}_1(\epsilon, \mathcal{F}, Z_1^n)$ is a random variable with expected value $\boldsymbol{E}\mathcal{N}_1(\epsilon, \mathcal{F}, Z_1^n)$.

Lemma A.1. (Haussler (1992), Theorem 2) *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [0, B]$, and let $Z_1^n = (Z_1, \dots, Z_n)$ be $\mathbb{R}^d$-valued i.i.d. random variables. Then for any $\alpha, \epsilon > 0$*

$$\boldsymbol{P}\left[\sup_{f \in \mathcal{F}} \frac{\left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \boldsymbol{E}f(Z_1)\right|}{\alpha + \boldsymbol{E}f(Z_1)} > \epsilon\right] \le 4\boldsymbol{E}\left(\mathcal{N}_1\left(\frac{\alpha\epsilon}{8}, \mathcal{F}, Z_1^n\right)\right) \exp\left(-\frac{n\alpha\epsilon^2}{16B}\right).$$

The following lemma is useful for bounding covering numbers of products of functions.

Lemma A.2. (Devroye *et al.* (1996), Theorem 29.7) *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two families of real functions on $\mathbb{R}^d$ with $|g_1(z)| \le B_1$ and $|g_2(z)| \le B_2$ for all $z \in \mathbb{R}^d$, $g_1 \in \mathcal{G}_1$ and $g_2 \in \mathcal{G}_2$. Then for any $z_1^n \in \mathbb{R}^{d \cdot n}$ and $\epsilon > 0$ we have*

$$\mathcal{N}_1(\epsilon, \{g_1 \cdot g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}, z_1^n) \le \mathcal{N}_1\left(\frac{\epsilon}{2B_2}, \mathcal{G}_1, z_1^n\right) \cdot \mathcal{N}_1\left(\frac{\epsilon}{2B_1}, \mathcal{G}_2, z_1^n\right).$$

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

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

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

$$\mathcal{F}^+ := \{\{(x, t) \in \mathbb{R}^d \times \mathbb{R} : t \le f(x)\} : f \in \mathcal{F}\}$$

of all subgraphs of functions of $\mathcal{F}$.

Lemma A.3. (Haussler (1992), Theorem 6) *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [0, B]$. Then one has for any $z_1^n \in \mathbb{R}^{d \cdot n}$ and any $0 < \epsilon < B/4$*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, z_1^n) \le 2\left(\frac{4eB}{\epsilon} \log\left(\frac{4eB}{\epsilon}\right)\right)^{V_{\mathcal{F}^+}}.$$

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

Lemma A.4. (Dudley (1978)) *Let $\mathcal{F}$ be a k-dimensional vector space of functions $f : \mathbb{R}^d \to \mathbb{R}$. Then the class of sets of the form $\{x \in \mathbb{R}^d : f(x) \geq 0\}$, $f \in \mathcal{F}$, has VC dimension less than or equal to k.*

The following lemma states the inequalities of McDiarmid (1989) and Hoeffding (1963) in the special case of identically distributed random variables and symmetric statistic.

Lemma A.5. *Let $Z_1, \ldots, Z_n, \tilde{Z}_n$ be independent identically distributed random variables with values in some Borel set $A \subset \mathbb{R}^m$, and let the functions $f = f_n : A^n \to \mathbb{R}$ be measurable and symmetric.*

If $|f(Z_1, \ldots, Z_n) - f(Z_1, \ldots, Z_{n-1}, \tilde{Z}_n)| \leq c < \infty$, then for each $\varepsilon > 0$

$$\boldsymbol{P}[|f(Z_1, \ldots, Z_n) - \boldsymbol{E} f(Z_1, \ldots, Z_n)| \geq \varepsilon] \leq 2e^{-2\varepsilon^2/(nc^2)}.$$

If especially the Z_i's are real-valued with $|Z_i| \leq \frac{c}{2}$, then for each $\varepsilon > 0$

$$\boldsymbol{P}\left[\left|\sum_{i=1}^{n} Z_i - \boldsymbol{E}\sum_{i=1}^{n} Z_i\right| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2/(nc^2)}.$$

References

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, **16**, 125–127.

Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates, *Ann. Statist.*, **9**, 1310–1319.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*, Wiley, New York.

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate, *J. Statisti. Plann. Inference*, **23**, 71–82.

Devroye, L. and Lugosi, G. (2001). *Combinatorical Methods in Density Estimation*, Springer, New York.

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

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.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

Dudley, R. (1978). Central limit theorems for empirical measures, *Ann. Probab.*, **6**, 899–929.

Efron, B. and Stein, C. (1980). The jackknife estimate of variance, *Ann. Statist.*, **9**, 586–596.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

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

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

Györfi, L. and 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. and 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. and Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression*, *Springer Ser. Statist.*, Springer, New York.

Hamers, M. and Kohler, M. (2003). A bound on the expected maximal deviations of sample averages from their means, *Statist. Probab. Lett.*, **62**, 137–144.

Härdle, H. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. and Comput.*, **100**, 78–150.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.*, **58**, 13–30.

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

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

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

Li, K. C. (1984). Consistency for cross-validated nearest neighbor estimates in nonparametric regression, *Ann. Statist.*, **12**, 230–240.

Lunts, A. and Brailovsky, V. (1967). Evaluation of attributes obtained in statistical decision rules, *Engineering Cybernetics*, **3**, 98–103.

McDiarmid, C. (1989). On the method of bounded differences, *Surveys in Combinatorics*, 148–188, Cambridge University Press, Cambridge.

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

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

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

Steele, J. (1986). An Efron-Stein inequality for nonsymmetric statistics, *Ann. Statist.*, **14**, 753–758.

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.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. Ser. B*, **36**, 111–147.

Walk, H. (2002*a*). On cross-validation in kernel and partitioning regression estimation, *Statist. Probab. Lett.*, **59**, 113–123.

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

Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression, *Ann. Statist.*, **11**, 1136–1141.