Ann. Inst. Statist. Math.
Vol. 54, No. 4, 879–899 (2002)

UNIVERSAL CONSISTENCY OF LOCAL POLYNOMIAL KERNEL REGRESSION ESTIMATES*

MICHAEL KOHLER

Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany,
e-mail: kohler@mathematik.uni-stuttgart.de

(Received December 18, 2000; revised May 25, 2001)

Abstract. Regression function estimation 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. It is shown that suitably defined local polynomial kernel estimates are weakly and strongly universally consistent, i.e., it is shown that the L_2 errors of these estimates converge to zero almost surely and in L_1 for all distributions.

Key words and phrases: Local polynomial kernel estimates, regression estimates, weak and strong universal consistency.

1. Introduction

1.1 *Nonparametric regression function estimation*

Let (X,Y), (X_1,Y_1), $(X_2,Y_2),\dots$ 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), and for an arbitrary f, L_2 error $\int |f(x) - m(x)|^2\mu(dx)$ is the difference between $\boldsymbol{E}\{|f(X) - Y|^2\}$ and $\boldsymbol{E}\{|m(X) - Y|^2\}$—the minimum of (1.2).

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

*Research supported by DAAD.

1.2 *Universal consistency*

A sequence of estimators $(m_n)_{n\in\mathbb{N}}$ is called **weakly universally consistent** if $\mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx)\to 0$ $(n\to\infty)$ for all distributions of (X,Y) with $\mathbf{E}Y^2<\infty$. It is called **strongly universally consistent** if $\int |m_n(x)-m(x)|^2\mu(dx)\to 0$ $(n\to\infty)$ a.s. for all distributions of (X,Y) with $\mathbf{E}Y^2<\infty$.

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

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

is one if X_i is among the k_n-nearest neighbors of x in $\{X_1,\ldots,X_n\}$ 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. *Kernel estimates* belong to the class of these estimates. There

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

$(0/0=0$ by definition) for some kernel function $K:\mathbb{R}^d\to\mathbb{R}_+$ and bandwidth $h_n>0$. Another example of local averaging estimates are *partitioning estimates*, which depend on a partition $\mathcal{P}_n=\{A_{n,1},A_{n,2},\ldots\}$ of $\mathbb{R}^d$. There the weights (1.4) are defined by

$$W_{n,i}(x)=\frac{I_{A_n(x)}(X_i)}{\sum_{j=1}^n I_{A_n(x)}(X_j)},$$

where $A_n(x)=A_{n,j}$ if $x\in A_{n,j}$ and $I_{A_{n,j}}$ denotes the indicator function of $A_{n,j}$.

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 corresponding result for partitioning estimates has been obtained by Györfi (1991). The strong universal consistency of kernel and partitioning estimates for suitably defined kernels, sequences of bandwidths and sequences of partitions has been shown by Walk (2002). Various results concerning consistency of variants of kernel and partitioning estimates can be found in Devroye and Krzyżak (1989), Nobel (1996), Györfi and Walk (1996, 1997) and Györfi *et al.* (1998).

It is easy to see that the partitioning estimate minimizes the so-called empirical L_2 risk

$$\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 \tag{1.5}$$

over the class of all real-valued functions f which are piecewise constant with respect to $\mathcal{P}_n$. *Least squares estimates* are defined by minimizing the empirical L_2 risk over general classes of functions (consisting e.g. of piecewise polynomials). The weak and strong universal consistency of various least squares estimates has been shown in Lugosi and Zeger (1995) and Kohler (1997, 1999).

Instead of minimizing the empirical L_2 risk (1.5) over some small class of functions one can also add a penalty term to (1.5) which penalizes the roughness of a function (e.g. a constant times the squared integral of the second derivative of f) and minimize the resulting sum over basically all functions (see Eubank (1988) or Wahba (1990) for details). The strong universal consistency of such *smoothing spline estimates* has been shown in Kohler and Krzyżak (2001).

1.3 *Local polynomial kernel estimates*

It is easy to see that the kernel estimate

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

satisfies for each $x \in \mathbb{R}^d$

$$\frac{1}{n}\sum_{i=1}^n |m_n(x) - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right) = \min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |a - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right).$$

Instead of fitting locally a constant to the data, the *local polynomial kernel estimate* fits locally a polynomial of some fixed degree M to the data, i.e., it is defined by

$$m_n(x) = \hat{p}_x(x) \tag{1.6}$$

where

$$\begin{aligned} \hat{p}_x(\cdot) &= \hat{p}_x(\cdot, \mathcal{D}_n) \in \mathcal{F}_M \\ &= \left\{ \sum_{0\le j_1,\dots,j_d\le M} a_{j_1,\dots,j_d} \cdot (x^{(1)})^{j_1} \cdot \dots \cdot (x^{(d)})^{j_d} : a_{j_1,\dots,j_d} \in \mathbb{R} \right\} \end{aligned} \tag{1.7}$$

satisfies

$$\frac{1}{n}\sum_{i=1}^n |\hat{p}_x(X_i) - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right) = \min_{p\in\mathcal{F}_M} \frac{1}{n}\sum_{i=1}^n |p(X_i) - Y_i|^2 K\left(\frac{x-X_i}{h_n}\right). \tag{1.8}$$

Local polynomial kernel estimates have been considered by many authors, see e.g. the monographs Härdle (1990), Korostelev and Tsybakov (1993) and Fan and Gijbels (1996) and the literature cited therein.

1.4 *Main results*

As defined in the previous subsection, local polynomial kernel estimates are in general not weakly consistent, even if the regression function is smooth and the distribution of X is nice (Devroye (1998), personal communication): Let X be uniformly distributed on $[0,1]$, Y be uniformly distributed on $\{-1,1\}$ and assume that X and Y are independent. Then it can been shown that the local linear estimate m_n defined by (1.6)–(1.8) with $M = 1$ and $K = I_{[-1,1]}$ satisfies $\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) = \infty$ for all n and all $h_n > 0$. The proof of this fact uses that if an interval of length h_n contains exactly two of the X_i's, if the corresponding Y_i's are different and if all other X_j's are more than h_n away from this interval, then the estimate will be on this interval equal to the line which interpolates the two data points with x-values in this interval. This line can have an arbitrary large slope and therefore also the estimate can take arbitrary large values on this interval.

In this paper we modify the definition (1.6)–(1.8). We minimize in (1.8) only over those polynomials whose coefficients are bounded in absolute value by some constant which depends on n and tends to infinity for n tending to infinity. We show that this modified local polynomial kernel estimate is, under some mild conditions on the kernel and the bandwidths, weakly and strongly consistent for all distributions of (X, Y) with X bounded and Y square integrable. Furthermore we show, that if we set this estimate to zero outside of some cube which depends on n and tends to $\mathbb{R}^d$ for n tending to infinity, then the resulting estimate is weakly and strongly universally consistent.

1.5 *Main idea in the proof*

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a square integrable function. Under some regularity conditions on the kernel the generalized Lebesgue density theorem implies that for μ-almost all x the pointwise error $|g(x) - m(x)|^2$ can be approximated for sufficiently small $h > 0$ by

$$\frac{\int |g(z) - m(z)|^2 \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \mu(dz)}{\int \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \mu(dz)}.$$

The nominator in the above integral is equal to

$$\begin{aligned} &\mathbf{E}\left\{|g(X) - m(X)|^2 \frac{1}{h^d} K\left(\frac{x-X}{h}\right)\right\} \\ &= \mathbf{E}\left\{|Y - g(X)|^2 \frac{1}{h^d} K\left(\frac{x-X}{h}\right)\right\} - \mathbf{E}\left\{|Y - m(X)|^2 \frac{1}{h^d} K\left(\frac{x-X}{h}\right)\right\}. \end{aligned}$$

By the strong law of large numbers this term is close to

$$\frac{1}{n} \sum_{i=1}^{n} (|Y_i - g(X_i)|^2 - |Y_i - m(X_i)|^2) \frac{1}{h^d} K\left(\frac{x-X_i}{h}\right),$$

if n is large. In the definition of the local polynomial kernel estimate the function g is chosen such that the last term is small.

The main difficulty in the proof is to show that the previous approximations also hold if g is chosen in some data-dependent way from some fixed set of polynomials.

To prove that in this case the Lebesgue density theorem still holds we use that in the definition of the estimate we consider only polynomials, whose coefficients are bounded by some data independent constant. This implies that these polynomials satisfy some Lipschitz condition for some constant, which doesn't depend on the data.

To prove that in this case also something similar to the strong law of large numbers holds, we use techniques from empirical process theory.

1.6 *Notation*

$\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ are the sets of natural, real and nonnegative real numbers, respectively. I_A denotes the indicator function, $card(A)$ the cardinality of a set A. The natural logarithm is denoted by $\log(\cdot)$.

The euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$, the components of x are denoted by $x^{(1)}, \ldots, x^{(d)}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$ set

$$\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)| \quad \text{and} \quad \|f\|^2 = \int_{\mathbb{R}^d} |f(x)|^2 \mu(dx).$$

For $h > 0$, $z \in \mathbb{R}^d$ and $K : \mathbb{R}^d \to \mathbb{R}$ define

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

$C_0^\infty(\mathbb{R}^d)$ is the set of all real-valued functions on $\mathbb{R}^d$ which are infinitely often differentiable and have compact support, $supp(X)$ is the support of the distribution of the random variable X.

1.7 *Outline*

The main results are stated in Section 2 and proven in Section 3. In the appendix a list of some results of empirical process theory, which are used in the proofs, is given.

2. Main results

Let $M \in \mathbb{N}_0$ and β_n, $h_n > 0$. Set

$$\mathcal{F}_M(\beta_n) = \left\{ \sum_{0 \le j_1, \ldots, j_d \le M} a_{j_1,\ldots,j_d} \cdot (x^{(1)})^{j_1} \cdot \ldots \cdot (x^{(d)})^{j_d} : |a_{j_1,\ldots,j_d}| \le \beta_n \right\}.$$

For given data $\mathcal{D}_n$ and $x \in \mathbb{R}^d$ choose

$$\hat{p}_x(\cdot) = \hat{p}_x(\cdot, \mathcal{D}_n) \in \mathcal{F}_M(\beta_n) \tag{2.1}$$

such that

$$\frac{1}{n}\sum_{i=1}^n |Y_i - \hat{p}_x(X_i)|^2 K_{h_n}(x - X_i) \le \inf_{p \in \mathcal{F}_M(\beta_n)} \left(\frac{1}{n}\sum_{i=1}^n |Y_i - p(X_i)|^2 K_{h_n}(x - X_i) + \frac{1}{n} \right), \tag{2.2}$$

and set

$$m_n(x) = \hat{p}_x(x). \tag{2.3}$$

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

$$b \cdot I_{[0,r^2]}(v) \le \tilde{K}(v) \le B \cdot I_{[0,R^2]}(v) \qquad (v \in \mathbb{R}_+)$$

for some $0 < r \leq R < \infty$, $0 < b \leq B < \infty$. Define the kernel $K : \mathbb{R}^d \to \mathbb{R}$ by

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

Let $M \in \mathbb{N}_0$. For $n \in \mathbb{N}$ choose β_n, $h_n > 0$ such that

$$\beta_n \to \infty \qquad (n \to \infty), \tag{2.4}$$

$$h_n \cdot \beta_n^2 \to 0 \qquad (n \to \infty) \tag{2.5}$$

and

$$\frac{n \cdot h_n^d}{\beta_n^2 \cdot \log(n)} \to \infty \qquad (n \to \infty). \tag{2.6}$$

Let the estimate m_n be defined by (2.1)–(2.3). Then

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

and

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

for every distribution of (X, Y) with $\|X\|$ bounded a.s. and $\mathbf{E}Y^2 < \infty$.

In Theorem 2.1 we need boundedness of $\|X\|$ to ensure that the estimate is weakly and strongly consistent. This assumption can be avoided, if we set the estimate to zero outside of a cube which depends on the sample size n and tends to $\mathbb{R}^d$ for n tending to infinity:

THEOREM 2.2. *Let $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotone decreasing and left-continuous function which satisfies*

$$b \cdot I_{[0,r^2]}(v) \leq \tilde{K}(v) \leq B \cdot I_{[0,R^2]}(v) \qquad (v \in \mathbb{R}_+)$$

for some $0 < r \leq R < \infty$, $0 < b \leq B < \infty$. Define the kernel $K : \mathbb{R}^d \to \mathbb{R}$ by

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

Let $M \in \mathbb{N}_0$. For $n \in \mathbb{N}$ choose A_n, β_n, $h_n > 0$ such that

$$A_n \to \infty \qquad (n \to \infty), \tag{2.7}$$

$$\beta_n \to \infty \qquad (n \to \infty), \tag{2.8}$$

$$h_n \cdot \beta_n^2 \cdot A_n^{2M \cdot d} \to 0 \quad (n \to \infty) \tag{2.9}$$

and

$$\frac{n \cdot h_n^d}{A_n^d \cdot \beta_n^2 \cdot \log(n)} \to \infty \qquad (n \to \infty). \tag{2.10}$$

Define m_n by (2.1)–(2.3) and set $\bar{m}_n(x) = m_n(x) \cdot I_{[-A_n, A_n]^d}(x)$. Then

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

and

$$E\left\{\int |\bar{m}_n(x) - m(x)|^2 \mu(dx)\right\} \to 0 \quad (n \to \infty)$$

for every distribution of (X, Y) *with* $EY^2 < \infty$, *i.e.,* $\bar{m}_n$ *is weakly and strongly universally consistent.*

Remark 1. We want to stress that in Theorem 2.2 there is no assumption on the underlying distribution of (X, Y) besides $EY^2 < \infty$. 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 2. It is well–known that one cannot derive a non–trivial rate of convergence result for the L_2 error of any estimate without restricting the class of distributions considered, e.g. by assuming some smoothness property on m (see, e.g., Theorem 7.2 in Devroye *et al.* (1996) and Section 3 in Devroye and Wagner (1980)). Stone (1982) showed that local polynomial kernel estimates achieve, in probability, the optimal rate of convergence if the regression function is k-times continuously differentiable, $M \geq k$ and and the distribution of X has a density with respect to the Lebesgue-Borel measure which is bounded away from zero and infinity.

Remark 3. It follows from the proofs given below that Theorems 2.1 and 2.2 also hold if the bandwidth h of the estimate is chosen in an arbitrary data-driven way from some deterministic interval $[h_{min}(n), h_{max}(n)]$, where $h_{min}(n), h_{max}(n) \in \mathbb{R}_+$ satisfy (2.5) and (2.9) with h_n replaced by $h_{max}(n)$ and (2.6) and (2.10) with h_n replaced by $h_{min}(n)$.

Remark 4. Let $M = 0$. Then the kernel estimate satisfies (1.6)–(1.8). It is easy to see that if one truncates the kernel estimates at height $\pm\beta_n$, then this truncated kernel estimate satisfies (2.1)–(2.3). Hence Theorem 2.2 implies that a modified kernel estimate, which is truncated at height $\pm\beta_n$ and is set equal to zero outside of some cube tending to $\mathbb{R}^d$ for n tending to infinity, is weakly and strongly universally consistent. It follows from Devroye and Wagner (1980) and Spiegelman and Sachs (1980) that these modifications are not necessary in order to get weak universal consistency. Walk (2001) shows that under suitable assumptions on the kernel and the bandwidth (including the assumption that the bandwidth doesn't change for every n) these modifications are also not necessary to prove strong universal consistency.

3. Proofs

In the proof of Theorems 2.1 and 2.2 we will apply the following lemma.

LEMMA 3.1. *Assume that the kernel* K *satisfies the assumptions of Theorem* 2.1. *Then there exists a constant* $c_1 \in \mathbb{R}_+$ *such that for all* $h > 0$ *and all distributions* μ *of* X *the following three inequalities are valid:*

a) *For all* $z \in \mathbb{R}^d$:

$$\int \frac{K_h(x-z)}{E\{K_h(x-X)\}} \mu(dx) \leq c_1.$$

b) *For all* $A \geq 1$:

$$\int_{[-A,A]^d} \frac{1}{E\{K_h(x-X)\}} \mu(dx) \leq c_1 \cdot A^d.$$

c) *For all* $f : \mathbb{R}^d \to \mathbb{R}_+$:

$$\int \frac{\mathbf{E}\{f(X)K_h(x-X)\}}{\mathbf{E}\{K_h(x-X)\}}\mu(dx) \le c_1 \cdot \int f(x)\mu(dx).$$

PROOF. a) follows from Lemma 1 in Devroye and Wagner (1980). In order to prove b) choose $z_1, \dots, z_K \in \mathbb{R}^d$ such that the union of all balls $S_{r\cdot h}(z_i)$ of radius $r \cdot h$ around z_i cover $[-A, A]^d$ and $K \le c \cdot A^d \cdot h^{-d}$ for some constant c which depends only on d. Then

$$\begin{aligned}\int_{[-A,A]^d} \frac{1}{\mathbf{E}\{K_h(x-X)\}}\mu(dx) &\le \sum_{i=1}^{K} \int_{S_{r\cdot h}(z_i)} \frac{1}{\mathbf{E}\{K_h(x-X)\}}\mu(dx) \\ &\le \frac{1}{b} \cdot h^d \sum_{i=1}^{K} \int_{S_{r\cdot h}(z_i)} \frac{K_h(x-z_i)}{\mathbf{E}\{K_h(x-X)\}}\mu(dx).\end{aligned}$$

This together with a) implies the assertion of b). c) follows from a) and

$$\int \frac{\mathbf{E}\{f(X)K_h(x-X)\}}{\mathbf{E}\{K_h(x-X)\}}\mu(dx) = \int f(z) \int \frac{K_h(x-z)}{\mathbf{E}\{K_h(x-X)\}}\mu(dx)\mu(dz). \qquad \Box$$

PROOF OF THEOREM 2.1. Choose $A \in \mathbb{R}_+$, $A > 1$ such that $supp(X) \subseteq [-A, A]^d$. Let L, $\epsilon > 0$ be arbitrary. Then there exists $\bar{m}_\epsilon \in C_0^\infty(\mathbb{R}^d)$ such that $\int |\bar{m}_\epsilon(x) - m(x)|^2\mu(dx) < \epsilon$. For $z \in \mathbb{R}$ set

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

Set $Y_L = T_L Y$ and $Y_{i,L} = T_L Y_i$ $(i = 1, \dots, n)$. Without loss of generality we assume that n is so large that $\|\bar{m}_\epsilon\|_\infty \le \beta_n$ and $L \le \beta_n$.

In the first step of the proof we show

$$\begin{aligned}(3.1) \int &|m_n(x) - m(x)|^2\mu(dx) \\ &\le 4 \cdot \int \frac{\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) + c_2 \cdot (\epsilon + A^{2M\cdot d}\beta_n^2 \cdot h_n)\end{aligned}$$

for some constant c_2 which depends only on M and d.

We have

$$\begin{aligned}(3.2) \int &|m_n(x) - m(x)|^2\mu(dx) \\ &\le 2\int |\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2\mu(dx) + 2\int |\bar{m}_\epsilon(x) - m(x)|^2\mu(dx) \\ &\le 2\epsilon + 2\int \left(|\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2 - \frac{\mathbf{E}\{|\hat{p}_x(X) - \bar{m}_\epsilon(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}} \right)\mu(dx) \\ &\quad +4\int \frac{\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \\ &\quad +4\int \frac{\mathbf{E}\{|m(X) - \bar{m}_\epsilon(X)|^2 K_{h_n}(x-X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx).\end{aligned}$$

By Lemma 3.1 c) the last integral is bounded by $c_1 \cdot \int |m(x) - \bar{m}_\epsilon(x)|^2 \mu(dx) \le c_1\epsilon$. In order to bound the first integral on the right-hand side of (3.2) observe that the first derivative of any $f \in \mathcal{F}_M(\beta_n)$ is on the cube $[-A, A]^d$ bounded (with respect to the euclidean norm) by $d \cdot M \cdot (M+1)^d A^{M\cdot d}\beta_n$. Hence by mean value theorem $|f(x) - f(u)| \le c_3 \cdot A^{M\cdot d}\beta_n \cdot \|x - u\|$ for all $f \in \mathcal{F}_M(\beta_n)$ and all $x, u \in [-A, A]^d$. Here c_3 is a constant which depends only on M and d. Furthermore by definiton of $\mathcal{F}_M(\beta_n)$

$$\sup_{x\in[-A,A]^d} |f(x)| \le (M+1)^d \cdot A^{M\cdot d} \cdot \beta_n \qquad (f \in \mathcal{F}_M(\beta_n)).$$

Because of $\bar{m}_\epsilon \in C_0^\infty(\mathbb{R}^d)$ we can assume without loss of generality that these two relations also hold for $f = \bar{m}_\epsilon$. We conclude that for all $x, u \in [-A, A]^d$ with $\|x - u\| \le R \cdot h_n$ and all $f \in \mathcal{F}_M(\beta_n)$

$$\begin{aligned}
&||f(x) - \bar{m}_\epsilon(x)|^2 - |f(u) - \bar{m}_\epsilon(u)|^2| \\
&\quad = |(f(x) - f(u)) + (\bar{m}_\epsilon(u) - \bar{m}_\epsilon(x))| \cdot |f(x) + f(u) - \bar{m}_\epsilon(u) - \bar{m}_\epsilon(x)| \\
&\quad \le 2 \cdot c_3 \cdot A^{M\cdot d}\beta_n \cdot \|x - u\| \cdot 4 \cdot (M+1)^d \cdot A^{M\cdot d} \cdot \beta_n \\
&\quad \le c_4 \cdot A^{2M\cdot d}\beta_n^2 \cdot h_n.
\end{aligned}$$

From this, together with $K_{h_n}(x - u) = 0$ for $\|x - u\| > R \cdot h_n$, we get

$$\begin{aligned}
&\int \left(|\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2 - \frac{\mathbf{E}\{|\hat{p}_x(X) - \bar{m}_\epsilon(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x - X)\}} \right) \mu(dx) \\
&\quad = \int \frac{\mathbf{E}\{(|\hat{p}_x(x) - \bar{m}_\epsilon(x)|^2 - |\hat{p}_x(X) - \bar{m}_\epsilon(X)|^2) K_{h_n}(x - X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x - X)\}} \mu(dx) \\
&\quad \le c_4 A^{2M\cdot d} \cdot \beta_n^2 \cdot h_n \cdot \int \frac{\mathbf{E}\{K_{h_n}(x - X) \mid \mathcal{D}_n\}}{\mathbf{E}\{K_{h_n}(x - X)\}} \mu(dx) \\
&\quad = c_4 A^{2M\cdot d} \cdot \beta_n^2 \cdot h_n.
\end{aligned}$$

This proves (3.1).

In the second step of the proof we bound $\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\}$ by a sum of several terms. For $x \in \mathbb{R}^d$ define $\bar{p}_x \in \mathcal{F}_M(\beta_n)$ by $\bar{p}_x(u) = \bar{m}_\epsilon(x)$ $(u \in \mathbb{R}^d)$. Then

$$\begin{aligned}
&\mathbf{E}\{|\hat{p}_x(X) - m(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\} \\
&\quad = \mathbf{E}\{|Y - \hat{p}_x(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\} - \mathbf{E}\{|Y - m(X)|^2 K_{h_n}(x - X)\} \\
&\quad = \mathbf{E}\{|Y - \hat{p}_x(X)|^2 K_{h_n}(x - X) \mid \mathcal{D}_n\} - (1+\epsilon)^3 \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{p}_x(X_i)|^2 K_{h_n}(x - X_i) \\
&\qquad + (1+\epsilon)^3 \cdot \left(\frac{1}{n} \sum_{i=1}^n |Y_i - \hat{p}_x(X_i)|^2 K_{h_n}(x - X_i) - \frac{1}{n} \sum_{i=1}^n |Y_i - \bar{p}_x(X_i)|^2 K_{h_n}(x - X_i) \right) \\
&\qquad + (1+\epsilon)^3 \frac{1}{n} \sum_{i=1}^n |Y_i - \bar{p}_x(X_i)|^2 K_{h_n}(x - X_i) - (1+\epsilon)^5 \mathbf{E}\{|Y - m(X)|^2 K_{h_n}(x - X)\} \\
&\qquad + ((1+\epsilon)^5 - 1) \mathbf{E}\{|Y - m(X)|^2 K_{h_n}(x - X)\} \\
&\quad = \sum_{j=1}^4 T_{j,n}(x).
\end{aligned}$$

In the next steps we give an upper bound for

$$\int \frac{T_{j,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \tag{3.3}$$

$(j \in \{1,2,3,4\})$.

In the third step of the proof we show

$$\int \frac{T_{4,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \leq \left((1+\epsilon)^5 - 1\right) \cdot c_1 \boldsymbol{E}\left\{|Y-m(X)|^2\right\}. \tag{3.4}$$

By Lemma 3.1 c) we get

$$\begin{aligned}
&\int \frac{T_{4,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad = ((1+\epsilon)^5-1)\int \frac{\boldsymbol{E}\{\boldsymbol{E}\{|Y-m(X)|^2 \mid X\}K_{h_n}(x-X)\}}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad \leq ((1+\epsilon)^5-1)\cdot c_1 \int \boldsymbol{E}\{|Y-m(X)|^2 \mid X=x\}\mu(dx) \\
&\quad = ((1+\epsilon)^5-1)\cdot c_1 \boldsymbol{E}\{|Y-m(X)|^2\},
\end{aligned}$$

which proves (3.4).

In the fourth step of the proof we show

$$\begin{aligned}
&\limsup_{n\to\infty} \int \frac{T_{3,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad \leq 2c_1\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\boldsymbol{E}\{|Y-Y_L|^2\} + c_1(1+\epsilon)^5\epsilon \quad \text{a.s.}
\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}
&\limsup_{n\to\infty} \boldsymbol{E} \int \frac{T_{3,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\quad \leq 2c_1\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\boldsymbol{E}\left\{|Y-Y_L|^2\right\} + c_1(1+\epsilon)^5\epsilon.
\end{aligned} \tag{3.6}$$

We use the decomposition

$$\begin{aligned}
&T_{3,n}(x) \\
&= (1+\epsilon)^3\frac{1}{n}\sum_{i=1}^n |Y_i - \bar{m}_\epsilon(x)|^2 K_{h_n}(x-X_i) - (1+\epsilon)^5\boldsymbol{E}\{|Y-m(X)|^2K_{h_n}(x-X)\} \\
&= (1+\epsilon)^3\Bigg(\tfrac{1}{n}\sum_{i=1}^n |Y_i - \bar{m}_\epsilon(x)|^2K_{h_n}(x-X_i) - (1+\epsilon)\cdot\tfrac{1}{n}\sum_{i=1}^n |Y_{i,L} - \bar{m}_\epsilon(x)|^2K_{h_n}(x-X_i)\Bigg) \\
&\quad + (1+\epsilon)^4\Bigg(\tfrac{1}{n}\sum_{i=1}^n |Y_{i,L} - \bar{m}_\epsilon(x)|^2K_{h_n}(x-X_i) - \tfrac{1}{n}\sum_{i=1}^n |Y_{i,L} - \bar{m}_\epsilon(X_i)|^2K_{h_n}(x-X_i)\Bigg) \\
&\quad + (1+\epsilon)^4\left(\frac{1}{n}\sum_{i=1}^n |Y_{i,L} - \bar{m}_\epsilon(X_i)|^2K_{h_n}(x-X_i) - \boldsymbol{E}\{|Y_L - \bar{m}_\epsilon(X)|^2K_{h_n}(x-X)\}\right) \\
&\quad + (1+\epsilon)^4(\boldsymbol{E}\{|Y_L - \bar{m}_\epsilon(X)|^2K_{h_n}(x-X)\} - (1+\epsilon)\boldsymbol{E}\{|Y - \bar{m}_\epsilon(X)|^2K_{h_n}(x-X)\})
\end{aligned}$$

$$+(1+\epsilon)^5(\boldsymbol{E}\{|Y-\bar{m}_\epsilon(X)|^2K_{h_n}(x-X)\}-\boldsymbol{E}\{|Y-m(X)|^2K_{h_n}(x-X)\})$$
$$=\sum_{j=5}^{9}T_{j,n}.$$

Using $(a+b)^2\le(1+\frac{1}{\epsilon})a^2+(1+\epsilon)b^2$ $(a,b\in\mathbb{R})$ we get

$$T_{5,n}(x)\le\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^3\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_{i,L}|^2K_{h_n}(x-X_i)$$

and

$$T_{8,n}(x)\le\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\boldsymbol{E}\left\{|Y_L-Y|^2K_{h_n}(x-X)\right\}.$$

Hence by Lemma 3.1 a)

$$\text{(3.7)}\qquad\int\frac{T_{5,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^3\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_{i,L}|^2\int\frac{K_{h_n}(x-X_i)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^3c_1\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_{i,L}|^2,$$

and by Lemma 3.1 c)

$$\text{(3.8)}\qquad\int\frac{T_{8,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4\int\frac{\boldsymbol{E}\left\{\boldsymbol{E}\{|Y_L-Y|^2\mid X\}K_{h_n}(x-X)\right\}}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le\left(1+\frac{1}{\epsilon}\right)\cdot(1+\epsilon)^4c_1\boldsymbol{E}\left\{|Y-Y_L|^2\right\}.$$

Furthermore

$$T_{6,n}(x)$$
$$=(1+\epsilon)^4\frac{1}{n}\sum_{i=1}^{n}(Y_{i,L}-\bar{m}_\epsilon(x)-(Y_{i,L}-\bar{m}_\epsilon(X_i)))$$
$$\cdot(Y_{i,L}-\bar{m}_\epsilon(x)+Y_{i,L}-\bar{m}_\epsilon(X_i))\,K_{h_n}(x-X_i)$$
$$\le(1+\epsilon)^4(2L+2\|\bar{m}_\epsilon\|_\infty)\cdot\sup_{\|u-v\|\le R\cdot h_n}|\bar{m}_\epsilon(u)-\bar{m}_\epsilon(v)|\cdot\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(x-X_i),$$

which together with Lemma 3.1 a) implies

$$\int\frac{T_{6,n}(x)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx)\le(1+\epsilon)^4(2L+2\|\bar{m}_\epsilon\|_\infty)\cdot\sup_{\|u-v\|\le R\cdot h_n}|\bar{m}_\epsilon(u)-\bar{m}_\epsilon(v)|\cdot c_1.$$

Because of $\bar{m}_\epsilon \in C_0^\infty(\mathbb{R}^d)$ this together with $h_n \to 0$ $(n \to \infty)$ implies

$$\limsup_{n\to\infty} \int \frac{T_{6,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \le 0 \tag{3.9}$$

and

$$\limsup_{n\to\infty} \mathbf{E} \int \frac{T_{6,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \le 0. \tag{3.10}$$

Next, we observe

$$\begin{aligned}
&\int \frac{T_{7,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&= (1+\epsilon)^4 \left(\frac{1}{n}\sum_{i=1}^n |Y_{i,L} - \bar{m}_\epsilon(X_i)|^2 \cdot \int \frac{K_{h_n}(x-X_i)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \right. \\
&\qquad \left. - \mathbf{E}\left\{ |Y_L - \bar{m}_\epsilon(X)|^2 \cdot \int \frac{K_{h_n}(x-X)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \right\} \right) \\
&= (1+\epsilon)^4 \left(\frac{1}{n}\sum_{i=1}^n Z_{i,n} - \mathbf{E}\{Z_{1,n}\} \right).
\end{aligned}$$

The random variables $Z_{1,n}, \ldots, Z_{n,n}$ are independent and identically distributed. It follows from Lemma 3.1 a) that they take, with probability one, only values in an interval of length $c_1(2L^2 + 2\|\bar{m}_\epsilon\|_\infty^2)$. Hence Hoeffding's inequality together with Borel-Cantelli lemma imply

$$\frac{1}{n}\sum_{i=1}^n Z_{i,n} - \mathbf{E}\{Z_{1,n}\} \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

This proves

$$\limsup_{n\to\infty} \int \frac{T_{7,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) = 0 \quad \text{a.s.} \tag{3.11}$$

Furthermore, independence and identical distribution of $Z_{1,n}, \ldots, Z_{n,n}$ imply

$$\mathbf{E} \int \frac{T_{7,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) = 0 \quad (n \in \mathbb{N}). \tag{3.12}$$

Finally by Lemma 3.1 c) and definition of $\bar{m}_\epsilon$ we get

$$\begin{aligned}
&\int \frac{T_{9,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&= (1+\epsilon)^5 \int \frac{\mathbf{E}\{|\bar{m}_\epsilon(X) - m(X)|^2 K_{h_n}(x-X)\}}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \\
&\le (1+\epsilon)^5 c_1 \int |\bar{m}_\epsilon(x) - m(x)|^2 \mu(dx) \\
&\le (1+\epsilon)^5 c_1 \epsilon.
\end{aligned}$$

This together with (3.7)–(3.12) and the strong law of large numbers implies (3.5) and (3.6).

In the fifth step of the proof we show

$$\limsup_{n\to\infty} \int \frac{T_{2,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \le 0 \quad \text{and} \tag{3.13}$$
$$\limsup_{n\to\infty} \mathbf{E} \int \frac{T_{2,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \le 0.$$

By definition of $\hat{p}_x$

$$T_{2,n}(x) \le (1+\epsilon)^3 \frac{1}{n}.$$

This together with Lemma 3.1 b) and $supp(X) \subseteq [-A,A]^d$ implies

$$\int \frac{T_{2,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \le (1+\epsilon)^3 \frac{1}{n} \cdot c_1 A^d,$$

which in turn implies (3.13).

In the sixth step of the proof we show

$$\int \frac{T_{1,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx) \tag{3.14}$$
$$\le c_1\left(1+\frac{1}{\epsilon}\right)\mathbf{E}\{|Y-Y_L|^2\} + c_1(1+\epsilon)^2\left(1+\frac{1}{\epsilon}\right)\frac{1}{n}\sum_{i=1}^n |Y_{i,L}-Y_i|^2$$
$$+c_1(1+\epsilon)A^d T_{10,n},$$

where

$$T_{10,n} = \sup_{f\in\mathcal{F}_M(\beta_n), z\in\mathbb{R}^d}\Bigg(\mathbf{E}\{|Y_L-f(X)|^2K_{h_n}(z-X)\} \tag{3.15}$$
$$-(1+\epsilon)\cdot\frac{1}{n}\sum_{i=1}^n |Y_{i,L}-f(X_i)|^2K_{h_n}(z-X_i)\Bigg).$$

We use the decomposition

$$T_{1,n}(x)$$
$$= \mathbf{E}\{|Y-\hat{p}_x(X)|^2K_{h_n}(x-X) \mid \mathcal{D}_n\} - (1+\epsilon)\mathbf{E}\{|Y_L-\hat{p}_x(X)|^2K_{h_n}(x-X) \mid \mathcal{D}_n\}$$
$$+(1+\epsilon)\mathbf{E}\{|Y_L-\hat{p}_x(X)|^2K_{h_n}(x-X) \mid \mathcal{D}_n\} - (1+\epsilon)^2\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-\hat{p}_x(X_i)|^2K_{h_n}(x-X_i)$$
$$+(1+\epsilon)^2\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-\hat{p}_x(X_i)|^2K_{h_n}(x-X_i) - (1+\epsilon)^3\frac{1}{n}\sum_{i=1}^n|Y_i-\hat{p}_x(X_i)|^2K_{h_n}(x-X_i).$$

Bounding the first and third terms on the right hand side as in the fourth step (cf. proof of (3.7) and (3.8)) we get

$$\int \frac{T_{1,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}}\mu(dx)$$
$$\le c_1\left(1+\frac{1}{\epsilon}\right)\mathbf{E}\{|Y-Y_L|^2\} + c_1(1+\epsilon)^2\left(1+\frac{1}{\epsilon}\right)\frac{1}{n}\sum_{i=1}^n |Y_{i,L}-Y_i|^2$$

$$+\int\left(\frac{(1+\epsilon)\boldsymbol{E}\{|Y_L-\hat{p}_x(X)|^2K_{h_n}(x-X)\mid\mathcal{D}_n\}}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\right.$$
$$\left.-\frac{(1+\epsilon)^2\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-\hat{p}_x(X_i)|^2K_{h_n}(x-X_i)}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\right)\mu(dx).$$

The difference of the nominators in the integral above is bounded by $(1+\epsilon)$ times $T_{10,n}$. $T_{10,n}$ doesn't depend on x, hence the whole integral can be bounded by $T_{10,n}$ times

$$(1+\epsilon)\cdot\int\frac{1}{\boldsymbol{E}\{K_{h_n}(x-X)\}}\mu(dx).$$

Applying Lemma 3.1 b) to the last term yields (3.14).

In the seventh step of the proof we show

$$\limsup_{n\to\infty}T_{10,n}\le 0\quad\text{a.s.}\qquad\text{and}\qquad\limsup_{n\to\infty}\boldsymbol{E}T_{10,n}\le 0.\tag{3.16}$$

To this end let $t>0$ be arbitray. Then

$$\boldsymbol{P}\{T_{10,n}>t\}$$
$$=\boldsymbol{P}\Bigg\{\exists f\in\mathcal{F}_M(\beta_n),z\in\mathbb{R}^d:$$
$$\boldsymbol{E}\{|Y_L-f(X)|^2K_{h_n}(z-X)\}-(1+\epsilon)\cdot\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-f(X_i)|^2K_{h_n}(z-X_i)>t\Bigg\}$$
$$\le\boldsymbol{P}\Bigg\{\exists f\in\mathcal{F}_M(\beta_n),z\in\mathbb{R}^d:$$
$$\frac{\boldsymbol{E}\{|Y_L-f(X)|^2K_{h_n}(z-X)\}-\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-f(X_i)|^2K_{h_n}(z-X_i)}{t+\epsilon\cdot\boldsymbol{E}\{|Y_L-f(X)|^2K_{h_n}(z-X)\}}>\frac{1}{1+\epsilon}\Bigg\}$$
$$=\boldsymbol{P}\Bigg\{\exists f\in\mathcal{F}_M(\beta_n),z\in\mathbb{R}^d:$$
$$\frac{\boldsymbol{E}\left\{|Y_L-f(X)|^2K\left(\frac{z-X}{h_n}\right)\right\}-\frac{1}{n}\sum_{i=1}^n|Y_{i,L}-f(X_i)|^2K\left(\frac{z-X_i}{h_n}\right)}{\frac{t\cdot h_n^d}{\epsilon}+\boldsymbol{E}\left\{|Y_L-f(X)|^2K\left(\frac{z-X}{h_n}\right)\right\}}>\frac{\epsilon}{1+\epsilon}\Bigg\}.$$

By Lemma A.1 in the Appendix, which uses the notion of covering numbers introduced in Definition A.1 in the Appendix, the last probability is bounded by

$$4 \cdot \boldsymbol{E}\mathcal{N}_1\left(\frac{t \cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \cdot \exp\left(-\frac{n \cdot \frac{t \cdot h_n^d}{\epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon}\right)^2}{64B\beta_n^2}\right),$$

where

$$\mathcal{G} = \left\{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = |T_L y - f(x)|^2 K\left(\frac{u-x}{h_n}\right) \; ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \right.$$
$$\left. \text{for some } u \in \mathbb{R}^d, f \in \mathcal{F}_M(\beta_n)\right\}.$$

We will show in the eighth step of the proof that

$$\mathcal{N}_1\left(\frac{t \cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \le \left(c_5 \frac{(1+\epsilon) \cdot \beta_n^2 A^{2M \cdot d}}{t \cdot h_n^d}\right)^{c_6} \tag{3.17}$$

for some constants c_5 and c_6 which depend only on M, B and d. This implies

(3.18) $\boldsymbol{P}\{T_{10,n} > t\}$

$$\le 4\left(c_5 \frac{(1+\epsilon) \cdot \beta_n^2 A^{2M \cdot d}}{t \cdot h_n^d}\right)^{c_6} \exp\left(-\frac{n \cdot \frac{t \cdot h_n^d}{\epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon}\right)^2}{64B\beta_n^2}\right)$$

$$= 4 \cdot \exp\left(-\log(n^2) \cdot \frac{nh_n^d}{\beta_n^2 2\log(n)} \cdot \left(\frac{t \cdot \epsilon}{64B(1+\epsilon)^2} - \frac{c_6 \log\left(c_5(1+\epsilon)\frac{\beta_n^2 A^{2Md}}{t \cdot h_n^d}\right)\beta_n^2}{nh_n^d}\right)\right).$$

The assumptions of Theorem 2.1 imply

$$\frac{nh_n^d}{\beta_n^2 2\log(n)} \to \infty \quad (n \to \infty)$$

and for n sufficiently large

$$\frac{c_6 \log\left(c_5(1+\epsilon)\frac{\beta_n^2 A^{2Md}}{t \cdot h_n^d}\right)\beta_n^2}{nh_n^d} \le \frac{c_6 \log(n)\, \beta_n^2}{nh_n^d} \to 0 \quad (n \to \infty).$$

It follows that the right-hand side of (3.18) is summable for each $t > 0$, hence the Borel-Cantelli lemma yields the first part of (3.16). In order to prove the second part, let $\delta > 0$ be arbitrary. Then

$$\boldsymbol{E}T_{10,n} \le \int_0^\infty \boldsymbol{P}\{T_{10,n} > t\}dt$$

$$
\begin{aligned}
&\leq \delta + \int_\delta^\infty 4\left(c_5 \frac{(1+\epsilon)\cdot\beta_n^2 A^{2M\cdot d}}{\delta \cdot h_n^d}\right)^{c_6} \exp\left(- \frac{n \cdot \frac{t\cdot h_n^d}{\epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon}\right)^2}{64B\beta_n^2} \right) dt \\
&= \delta + 4\left(c_5 \frac{(1+\epsilon)\cdot\beta_n^2 A^{2M\cdot d}}{\delta \cdot h_n^d}\right)^{c_6} \cdot \frac{64B\beta_n^2(1+\epsilon)^2}{n\cdot h_n^d \epsilon} \exp\left(- \frac{n \cdot h_n^d}{\beta_n^2} \cdot \frac{\delta\cdot\epsilon}{64B(1+\epsilon)^2}\right) \\
&\to \delta \quad (n \to \infty)
\end{aligned}
$$

by the assumptions of Theorem 2.1. With $\delta \to 0$ the second part of (3.16) follows.

In the eighth step of the proof we show (3.17). Therefore we use arguments from the proof of Theorem 2 in Krzyżak *et al.* (1996). We have $\mathcal{G} = \{g_1 \cdot g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$, where

$$\mathcal{G}_1 = \{g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} : g(x,y) = |T_L y - f(x)|^2 ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \text{ for some } f \in \mathcal{F}_M(\beta_n)\}$$

and

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

The functions in $\mathcal{G}_1$ and $\mathcal{G}_2$ are bounded on $[-A,A]^d \times \mathbb{R}$ in absolute value by

$$(2L^2 + 2(\beta_n (M+1)^d A^{M\cdot d})^2) \leq 4\beta_n^2 (M+1)^{2d} A^{2M\cdot d}$$

and B, respectively. Hence by Lemma A.2 in the Appendix we get

$$
\begin{aligned}
\mathcal{N}_1\left(\frac{t\cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \leq\ & \mathcal{N}_1\left(\frac{t\cdot h_n^d}{16(1+\epsilon)B}, \mathcal{G}_1, (X,Y)_1^n\right) \\
& \cdot \mathcal{N}_1\left(\frac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}, \mathcal{G}_2, (X,Y)_1^n\right).
\end{aligned}
$$

If $h_i(x,y) = |f_i(x) - T_L y|^2$ for some $f_i : [-A,A]^d \to \mathbb{R}$ bounded in absolute value by $\beta_n (M+1)^d A^{M\cdot d}$, then

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n |h_1(X_i,Y_i) - h_2(X_i,Y_i)|^2 \\
&\quad = \frac{1}{n}\sum_{i=1}^n |f_1(X_i) - T_L Y_i + f_2(X_i) - T_L Y_i| \cdot |f_1(X_i) - f_2(X_i)| \\
&\quad \leq (2L + 2\beta_n (M+1)^d A^{M\cdot d}) \cdot \frac{1}{n}\sum_{i=1}^n |f_1(X_i) - f_2(X_i)|
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\mathcal{N}_1\left(\frac{t\cdot h_n^d}{16(1+\epsilon)B}, \mathcal{G}_1, (X,Y)_1^n\right) \\
&\quad \leq \mathcal{N}_1\left(\frac{t\cdot h_n^d}{16(1+\epsilon)B\,(2L + 2\beta_n(M+1)^d A^{M\cdot d})}, \mathcal{F}_M(\beta_n), X_1^n\right).
\end{aligned}
$$

Next we need the notion of VC dimension, which is introduced in Defintion A.2 in the Appendix. $\mathcal{F}_M(\beta_n)$ is a subset of a linear vector space of dimension $(M+1)^d$, hence by Lemma A.4 in the Appendix

$$V_{\mathcal{F}_M(\beta_n)^+} \le (M+1)^d + 1 \le (M+2)^d.$$

This together with Lemma A.3 in the Appendix implies

$$\begin{aligned}\mathcal{N}_1\left(\frac{t\cdot h_n^d}{16(1+\epsilon)B}, \mathcal{G}_1, (X,Y)_1^n\right) &\le 2\left(\frac{4e(M+1)^d\beta_n A^{M\cdot d}}{\dfrac{t\cdot h_n^d}{16(1+\epsilon)B\left(2L+2\beta_n(M+1)^d A^{M\cdot d}\right)}}\right)^{2(M+2)^d}\\ &\le \left(c_7\cdot\frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t\cdot h_n^d}\right)^{2(M+2)^d},\end{aligned}$$

where c_7 is a constant which depends only on M, B and d.

Next we bound

$$\mathcal{N}_1\left(\frac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}, \mathcal{G}_2, (X,Y)_1^n\right).$$

By Lemma A.3 in the Appendix we get

$$\begin{aligned}&\mathcal{N}_1\left(\frac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}, \mathcal{G}_2, (X,Y)_1^n\right)\\ &\quad\le 2\left(\frac{4eB}{\dfrac{t\cdot h_n^d}{64(1+\epsilon)\beta_n^2(M+1)^{2d}A^{2M\cdot d}}}\right)^{2V_{\mathcal{G}_2^+}}\\ &\quad\le \left(c_8\cdot\frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t\cdot h_n^d}\right)^{2V_{\mathcal{G}_2^+}},\end{aligned}$$

where c_8 is a constant which depends only on M, B and d. Hence it suffices to derive a bound on the VC dimension of the class of all subgraphs of

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

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

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

where $\phi(t)=\sup\{z:\tilde{K}(z)\ge t\}$. Equivalently, (x,y,t) must satisfy

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

Consider now the set of real functions

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

If for a given collection of points $\{(x_i, y_i, t_i)\}_{i=1,\dots,n}$ a set $\{(x,y,t) : g(x,y) \geq t\}$, $g \in \mathcal{G}_2$ picks out the points $\{(x_{i_1}, y_{i_1}, t_{i_1}), \dots, (x_{i_l}, y_{i_l}, t_{i_l})\}$ then there exist α, β, γ, δ such that $\{(x,y,s) : g_{\alpha,\beta,\gamma,\delta}(x,y,s) \geq 0\}$ 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 $V_{\mathcal{G}_2^+} \leq V_{\{\{(x,y,s):g(x,y,s)\geq 0\}:g\in\mathcal{G}_3\}}$. $\mathcal{G}_3$ is a linear vector space of dimension $d+3$, hence we can conclude from Lemma A.4 in the Appendix $V_{\mathcal{G}_2^+} \leq d+3$. Summarizing the above results we get

$$\begin{aligned}
&\mathcal{N}_1\left(\frac{t \cdot h_n^d}{8(1+\epsilon)}, \mathcal{G}, (X,Y)_1^n\right) \\
&\leq \left(c_7 \cdot \frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t \cdot h_n^d}\right)^{2(M+2)^d} \cdot \left(c_8 \cdot \frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t \cdot h_n^d}\right)^{2(d+3)} \\
&\leq \left(c_5 \cdot \frac{(1+\epsilon)\beta_n^2 A^{2M\cdot d}}{t \cdot h_n^d}\right)^{c_6}
\end{aligned}$$

for constants c_5 and c_6 which depend only on M, B and d.

In the ninth and last step of the proof we finish the proof by summarizing the above results. By the results of the first and second step we have

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mu(dx) \\
&\leq c_2(\epsilon + A^{2M\cdot d}\beta_n^2 h_n) + 4\sum_{j=1}^{4} \int \frac{T_{j,n}(x)}{\mathbf{E}\{K_{h_n}(x-X)\}} \mu(dx).
\end{aligned}$$

Using the results of steps three to seven and $\beta_n^2 h_n \to 0$ ($n \to \infty$) one gets

$$\begin{aligned}
&\limsup_{n\to\infty} \int |m_n(x) - m(x)|^2 \mu(dx) \\
&\leq c_2\epsilon + 4((1+\epsilon)^5 - 1)c_1 \mathbf{E}|Y - m(X)|^2 \\
&\quad + 8c_1(1+1/\epsilon)(1+\epsilon)^4 \mathbf{E}|Y - Y_L|^2 + 4c_1(1+\epsilon)^5\epsilon \\
&\quad + 4c_1(1+1/\epsilon)\mathbf{E}|Y - Y_L|^2 + 4c_1(1+\epsilon)^2(1+1/\epsilon)\mathbf{E}|Y - Y_L|^2 \quad \text{a.s.}
\end{aligned}$$

With $L \to \infty$ and $\epsilon \to 0$ this implies $\int |m_n(x) - m(x)|^2 \mu(dx) \to 0$ ($n \to \infty$) a.s. The proof of $\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \to 0$ ($n \to \infty$) follows in an analogous way from the previous results. □

Proof of Theorem 2.2. By definition of $\bar{m}_n$

$$\begin{aligned}
&\int |\bar{m}_n(x) - m(x)|^2 \mu(dx) \\
&= \int_{[-A_n, A_n]^d} |m_n(x) - m(x)|^2 \mu(dx) + \int_{\mathbb{R}^d \setminus [-A_n, A_n]^d} |m(x)|^2 \mu(dx).
\end{aligned}$$

Because of $A_n \to \infty$ $(n \to \infty)$ and $\int |m(x)|^2 \mu(dx) < \infty$ we have

$$\int_{\mathbb{R}^d \setminus [-A_n, A_n]^d} |m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty).$$

Hence it suffices to show

$$\int_{[-A_n, A_n]^d} |m_n(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty)$$

a.s. and *in* L_1. This can be done by replacing in the proof of Theorem 2.1 A by A_n and $\int \dots$ by $\int_{[-A_n, A_n]^d} \dots$ Then one has to show in the seventh step

$$\limsup_{n \to \infty} A_n^d \cdot T_{10,n} \leq 0 \quad \text{a.s.} \qquad \text{and} \qquad \limsup_{n \to \infty} A_n^d \cdot \boldsymbol{E} T_{10,n} \leq 0.$$

To this end one uses

$$\begin{aligned} &\boldsymbol{P}\{A_n^d \cdot T_{10,n} > t\} \\ &\quad = \boldsymbol{P}\left\{T_{10,n} > \frac{t}{A_n^d}\right\} \\ &\quad \leq 4 \left(c_5 \frac{(1+\epsilon) \cdot \beta_n^2 A_n^{2M \cdot d}}{(t/A_n^d) \cdot h_n^d} \right)^{c_6} \exp\left(- \frac{n \cdot \frac{t \cdot h_n^d}{A_n^d \epsilon} \cdot \left(\frac{\epsilon}{1+\epsilon} \right)^2}{64 B \beta_n^2} \right) \end{aligned}$$

and proceeds otherwise as before. □

Acknowledgements

The author thanks two anonymous referees for several helpful comments.

Appendix

A. Some results of empirical process theory

In this section we list the definitions and results of empirical process theory which we have used in Section 3. 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

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

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

If $Z_1^n = (Z_1, \ldots, Z_n)$ 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), Th. 2)) *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [0, B]$, and let $Z_1^n = (Z_1, \ldots, Z_n)$ be $\mathbb{R}^d$-valued i.i.d. random variables. Then for any α, $\epsilon > 0$*

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

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

LEMMA A.2 (Devroye *et al.* (1996), Th. 29.7) *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two families of real functions on $\mathbb{R}^d$ with $|g_1(z)| \leq B_1$ and $|g_2(z)| \leq B_2$ 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*

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

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 following lemma, which uses the notation $V_{\mathcal{F}^+}$ for the VC dimension of the set

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

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

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

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

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.*

References

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

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.

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

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

Györfi, L. (1991). Universal consistency of a regression estimate for unbounded regression functions, *Nonparametric Functional Estimation and Related Topics* (ed. G. Roussas), NATO ASI Series, 329–338, Kluwer, Dordrecht.

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.

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

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

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

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

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

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image ReConstruction*, Springer, Berlin.

Krzyżak, A., Linder, T. and Lugosi, G. (1996). Nonparametric estimation and classification using radial basis function nets and empirical risk minimization, *IEEE Transactions on Neural Networks*, **7**, 475–487.

Lugosi, G. and 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. and 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.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia, Pennsylvania.

Walk, H. (2002). Almost sure convergence properties of Nadaraya–Watson regression estimates, *Essays on Uncertainty-S. Yakowitz Memorial Volume* (eds. M. Dror, P. L'Ecuyer and F. Szidarovszky), 201–223, Kluwer, Dordrecht.