Journal of Statistical Planning and Inference 23 (1989) 71–82
North-Holland

AN EQUIVALENCE THEOREM FOR L_1 CONVERGENCE OF THE KERNEL REGRESSION ESTIMATE*

Luc DEVROYE

School of Computer Science and Department of Mathematics and Statistics, McGill University, Montreal, Canada H3A 2A7

Adam KRZYŻAK

Department of Computer Science, Concordia University, Montreal, Canada H3G 1M8

Received 29 February 1988; revised manuscript received 5 July 1988
Recommended by M. Rosenblatt

Abstract: We show that all modes of convergence in L_1 (in probability, almost surely, complete) for the standard kernel regression estimate are equivalent.

AMS Subject Classification: Primary 62G05.

Key words and phrases: Regression function; nonparametric estimation; consistency; strong convergence; convergence; kernel estimate.

1. Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent observations of an $R^d \times [-M, M]$-valued random vector (X, Y). The regression function $m(x) = E(Y \mid X = x)$ can be estimated by the *kernel estimate*

$$m_n(x) = \sum_{i=1}^{n} Y_i K_h(x - X_i) \Big/ \sum_{i=1}^{n} K_h(x - X_i),$$

where $h > 0$ is a smoothing factor depending upon n, K is an absolutely integrable function (the kernel), and $K_h(x) = K(x/h)$ (Nadaraya (1964, 1970), Watson (1964)). For a survey of other estimates, see e.g. Collomb (1981) or Györfi (1981).

We are concerned with the L_1 convergence of m_n to m as measured by $J_n \triangleq \int |m_n(x) - m(x)| \mu(\mathrm{d}x)$ where μ is the (unknown) probability measure for X. This quantity is particularly important in discrimination based on the kernel rule (see Devroye and Wagner (1980) or Stone (1977)). Stone (1977) first pointed out that

* Research of the authors was supported by NSERC grants A3456 and A0270, and FCAR grants EQ-1679 and EQ-2904.

there exist estimators for which $J_n \to 0$ in probability for all distributions of (X, Y) with $E(|Y|) < \infty$. This included the nearest neighbor and histogram estimates. In 1980, Devroye and Wagner, and independently Spiegelman and Sacks, showed that this is also the case for the kernel estimate provided that K is a bounded nonnegative function with compact support such that for a small fixed sphere S centered at the origin, $\inf_{x \in S} K(x) > 0$, and that

$$\lim_{n \to \infty} h = 0, \quad \lim_{n \to \infty} nh^d = \infty. \tag{1}$$

Older proofs of the strong convergence of J_n to 0 proceed from pointwise convergence (i.e., $m_n - m \to 0$ almost surely at almost all x (μ)), and move on to L_1 convergence via a result of Glick's (1974) from which it is possible to conclude that $J_n \to 0$ almost surely. This route is indirect and brings with it additional conditions on h (notably, $nh^d/\log(n) \to \infty$ for complete convergence, and at least $nh^d/\operatorname{loglog}(n) \to \infty$ for strong convergence) and K (Devroye and Wagner (1980), Devroye (1981), Krzyżak and Pawlak (1984), Greblicki, Krzyżak and Pawlak (1984) and Krzyżak (1986) discuss possible conditions on K). Greblicki et al. (1984) for example required that, in addition to the condition for weak convergence mentioned earlier, $K(x)/H(\|x\|) \in [a, b]$ for some $0 < a \le b < \infty$, where H is a bounded nonincreasing Borel function with $t^d H(t) \to 0$ as $t \to \infty$.

In 1983, Devroye showed that all modes of convergence in L_1 are equivalent for the kernel density estimate, which brings up the question whether a similar result is not valid for the kernel regression estimate. As shown in Chapter 10 of Devroye and Györfi (1985), this is indeed possible whenever X has a density and Y is bounded. Difficulties arise with the regression estimate when we no longer assume that the X-variable has a density. This follows from the fact that the estimate is a ratio and that the identification of a limit for the numerator or denominator is no longer obvious. The technical hurdles can be overcome without much trouble for the histogram regression estimate (Devroye and Györfi (1983) obtained the equivalence for all distributions of (X, Y) with $|Y| \le M < \infty$). In a personal communication, Györfi has pointed out that $J_n \to 0$ almost surely for all partitioning estimates whenever $E|Y| < \infty$, provided that a bin width condition similar to (1) is satisfied. The purpose of the present paper is to explain a simple technique based upon exponential martingale inequalities for proving the equivalence of all modes of convergence of J_n for the kernel estimate under no conditions on the distribution of (X, Y) other than the boundedness of Y.

We will say that a kernel K is *regular* if $K(x) \ge B l_{S_r}$ for some positive constants B and r, where S_r is the ball of radius r centered at the origin, and if

$$\int_{y \in x + S_r} K(y) \mathrm{d}x < \infty.$$

The latter condition is for example satisfied if K is merely Riemann integrable.

Theorem (The main result). *Assume that K is a regular kernel. Let m_n be the kernel regression estimate with kernel K, and let $0/0$ be defined as 0. Then the following statements are equivalent:*

(A) *For every distribution of (X, Y) with $|Y| \le M < \infty$, and for every $\varepsilon > 0$, there exist constants c and n_0 such that for all $n \ge n_0$,*

$$P(J_n > \varepsilon) \le e^{-cn}.$$

(B) *For every distribution of (X, Y) with $|Y| \le M < \infty$,*

$$J_n \to 0 \text{ almost surely as } n \to \infty.$$

(C) *For every distribution of (X, Y) with $|Y| \le M < \infty$,*

$$J_n \to 0 \text{ in probability as } n \to \infty.$$

(D)

$$\lim_{n\to\infty} h = 0, \qquad \lim_{n\to\infty} nh^d = \infty.$$

Remark 1 (A curiosity). It is interesting that we can find sequences h for which $J_n \to 0$ almost surely for all distributions of (X, Y) with bounded Y, yet m_n does not tend to m in the almost sure pointwise sense (take $h^d = \log\log\log(n)/n$, and note that $nh^d/\log\log(n) \to \infty$ is necessary for the almost sure pointwise convergence of the kernel estimate whenever X has a density and m is twice continuously differentiable with $m'' \neq 0$).

Remark 2 (Unbounded Y). The present proof needs to be generalized in order to replace the condition $|Y| \le M < \infty$ by the natural condition $E(|Y|) < \infty$. In an unpublished document, Györfi has obtained a result similar to the Theorem for the histogram estimate, with the relaxed condition on Y. The boundedness of Y can be relaxed to the condition $E|Y|^{2+\delta} < \infty$. See the Appendix for the proof.

Remark 3 (Necessity of the conditions). It is not true that when $J_n \to 0$ in probability for one distribution of (X, Y), the conditions (1) on h follow: just consider the case that $Y = 0$ with probability one. Of course, the implication is true for 'most' distributions of (X, Y). See for example the distribution used in the proof of the Theorem.

2. Proof of the theorem

Clearly, (A) implies (B), which implies (C). To see that (C) implies (D), we need only construct one prototype distribution for which the implication holds. One can for example consider a triple (X, U, Y) on $[0, 1]^d \times [0, 1] \times \{0, 1\}$ where X and U are independent and uniformly distributed on their supports, and $Y = I_{[U \le m(X)]}$ where m is an absolutely continuous $[0, 1]$-valued function on the real line, with an ab-

solutely continuous derivative m', zero outside the hypercube, and $\int m(x)\,dx = \frac{1}{2}$. The details of the proof are left as a simple exercise, but basically, the X_i's with $Y_i = 1$ have density $2m(x)$ on the hypercube, so that estimating m consistently is equivalent to estimating a density consistently. One can then invoke Devroye (1983), where it is shown that estimating one (any!) density consistently implies (D).

The novelty in this paper is the proof that condition (D) implies the exponential inequality (A). We begin with

$$|m_n(x) - m(x)| = \left| \frac{\sum_{i=1}^n (Y_i - m(x))K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)} \right|$$

$$= \left| \frac{\sum_{i=1}^n (Y_i - m(x))K_h(x - X_i)/nEK_h(x - X_1)}{\sum_{i=1}^n K_h(x - X_i)/nEK_h(x - X_1)} \right|$$

$$\triangleq \left| \frac{U_n(x)}{L_n(x)} \right|.$$

We will need several constants; so as to be able to optimize constants after the proof has been completed, it is to our advantage to give each constant a name. Small constants are called ε_i's with

$$\varepsilon_0 = \frac{\varepsilon}{4}, \quad \varepsilon_1 = \frac{\varepsilon}{2}, \quad \varepsilon_2 = \frac{\varepsilon_1}{M} = \frac{\varepsilon}{2M}, \quad \varepsilon_3 = \frac{\varepsilon_0 \varepsilon_2}{2} = \frac{\varepsilon^2}{16M},$$

$$\varepsilon_4 = \frac{\varepsilon_2}{4} = \frac{\varepsilon}{8M}, \quad \varepsilon_5 = \frac{\varepsilon_3}{2} = \frac{\varepsilon^2}{32M}, \quad \varepsilon_6 = \frac{\varepsilon_4}{2} = \frac{\varepsilon}{16M}.$$

Clearly,

$$P\left(\int \left| \frac{U_n}{L_n} \right| d\mu > \varepsilon \right) = P\left(\int \left| \frac{U_n}{L_n} \right| (I_{A(x)} + I_{A^c(x)})\, d\mu > \varepsilon \right)$$

(where $A(x) = \{|U_n(x)| < \varepsilon_0, |L_n(x) - 1| < \frac{1}{2}\}$)

$$\le P\left(2\varepsilon_0 + \sup_x \left| \frac{U_n(x)}{L_n(x)} \right| \int I_{A^c(x)} \mu(dx) > \varepsilon \right)$$

$$\le P\left(P(|U_n(X)| \ge \varepsilon_0 | D_n) + P(|L_n(X) - 1| \ge \tfrac{1}{2} | D_n) > \frac{\varepsilon_1}{M} \right)$$

(since $\varepsilon - 2\varepsilon_0 = \varepsilon_1 = M\varepsilon_2$).

Here X is a random variable distributed as X_1 and independent of the *data* $D_n = (X_1, Y_1, \ldots, Y_n)$. We now apply Markov's inequality to the last expression, and bound it from above by

$$P\left(\int \frac{|U_n(x)|}{\varepsilon_0} \mu(dx) \ge \frac{\varepsilon_2}{2} \right) + P\left(\int \frac{|L_n(x) - 1|}{\frac{1}{2}} \mu(dx) \ge \frac{\varepsilon_2}{2} \right)$$

$$= P\left(\int |U_n(x)|\, \mu(dx) \ge \varepsilon_3 \right) + P\left(\int |L_n(x) - 1|\, \mu(dx) \ge \varepsilon_4 \right).$$

We are now at a crucial junction in the proof, having bounded a probability of an event involving a ratio into a sum of probabilities of events involving no ratios at all. Each term in the last expression will now be treated separately. We need to rewrite U_n and L_n so that we can apply some Hoeffding-like inequalities for martingale difference sequences. Define

$$Z_i = E\left(\int |U_n| \,|\, D_i\right) - E\left(\int |U_n| \,|\, D_{i-1}\right)$$

and

$$\sum_{i=1}^{n} Z_i = E\left(\int |U_n| \,|\, D_n\right) - E\left(\int |U_n| \,|\, D_0\right) = \int |U_n| - E\left(\int |U_n|\right),$$

where D_0 stands for 'no data', or the trivial sigma algebra. Observe that $E(Z_i \,|\, D_{i-1}) = 0$, so that the Z_i's form a martingale difference sequence. In order to be able to apply the exponential inequality used below, it is necessary that each Z_i be bounded from above uniformly over all i. With the convenient notation

$$W_{i,k} = \frac{1}{n} \sum_{i \le j \le k} \frac{(Y_j - m(x))K_h(x - X_j)}{EK_h(x - X_1)},$$

we have (see Devroye, 1988)

$$\begin{aligned} |Z_i| &\le \int |E(|W_{1,i-1} + W_{i,i} + W_{i+1,n}| \,|\, D_i) - E(|W_{1,i-1} + W_{i,i} + W_{i+1,n}| \,|\, D_{i-1})| \\ &\le \int \sup_{a \in R} |E(|a + W_{i,i}|) - |a + W_{i,i}|| \\ &\le \int |W_{i,i} - EW_{i,i}| + \int E(|W_{i,i} - EW_{i,i}|). \end{aligned}$$

Now,

$$\begin{aligned} \int |W_{i,i}| &= \int \left| \frac{1}{n} \frac{(Y_i - m(x))K_h(x - X_i)}{EK_h(x - X_1)} \right| \mu(\mathrm{d}x) \\ &\le \frac{2M}{n} \int \frac{K_h(x - X_i)}{EK_h(x - X_1)} \mu(\mathrm{d}x) \\ &\le \frac{2M}{n} \sup_y \frac{K_h(x - y)}{EK_h(x - X_1)} \mu(\mathrm{d}x) \\ &\le \frac{2M\varrho}{n}, \end{aligned}$$

where ϱ is the finite constant of the covering lemma (Lemma 1) described in the next

section. We also have $\int |EW_{i,i}| \le 2M\varrho/n$ and

$$|Z_i| \le \frac{1}{n}(4M + 4M\varrho) \triangleq \frac{C}{n}.$$

We have

$$\begin{aligned} P\left(\int |U_n| > \varepsilon_3\right) &\le P\left(\left|\int |U_n| - E\int |U_n|\right| > \tfrac{1}{2}\varepsilon_3\right) + P\left(E\int |U_n| > \tfrac{1}{2}\varepsilon_3\right) \\ &\le 2e^{-\varepsilon_5^2/2n(C/n)^2} + P\left(E\int |U_n| > \tfrac{1}{2}\varepsilon_3\right) \\ &= 2e^{-n\varepsilon_5^2/2C^2} + P\left(E\int |U_n| > \varepsilon_5\right) \end{aligned}$$

where $\varepsilon_5 = \frac{1}{2}\varepsilon_3$. We have used an exponential martingale inequality due to Azuma (1967) (see e.g. Stout (1974)) which states that for Z_i's as defined here,

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > \varepsilon\right) \le 2e^{-\varepsilon^2/2\sum_{i=1}^{n}\|Z_i\|_\infty^2}$$

where $\|Z_i\|_\infty$ is the essential supremum of $|Z_i|$. The probability $P(\int |L_n - 1| > \varepsilon_4)$ can be treated similarly. The only difference is that we have to define

$$W_{i,i} = \frac{1}{n}\left(\frac{K_h(x - X_i)}{EK_h(x - X_1)} - 1\right)$$

and note that $EW_{i,i} = 0$. By the covering argument used above, we have $\int |W_{i,i}| \le (\varrho + 1)/n$, and thus, if in the definition of Z_i, we replace U_n by $L_n - 1$, $|Z_i| \le 2(\varrho + 1)/n \triangleq C^*/n$. Again, applying the exponential martingale inequality, we obtain, with $\varepsilon_6 = \frac{1}{2}\varepsilon_4$,

$$\begin{aligned} &P\left(\int |L_n - 1| > \varepsilon_4\right) \\ &\le P\left(\left|\int |L_n - 1| - E\int |L_n - 1|\right| > \tfrac{1}{2}\varepsilon_4\right) + P\left(E\int |L_n - 1| > \tfrac{1}{2}\varepsilon_4\right) \\ &\le 2e^{-\varepsilon_6^2/2n(C^*/n)^2} + P\left(E\int |L_n - 1| > \varepsilon_6\right) \\ &= 2e^{-n\varepsilon_6^2/2C^2} + P\left(E\int |L_n - 1| > \varepsilon_6\right). \end{aligned}$$

We can conclude that

$$P\left(\int |m_n(x) - m(x)|\,\mu(dx) > \varepsilon\right) \le 4e^{-n\min^2(\varepsilon_5/C,\, \varepsilon_6/C^*)},$$

when $E\int |U_n| \le \varepsilon_4$ and $E\int |L_n - 1| \le \varepsilon_6$. The latter condition holds for all n large enough because $E\int |U_n| \to 0$ and $E\int |L_n - 1| \to 0$ as $n \to \infty$ (see Lemma 2 below). □

3. A covering lemma

Lemma 1. *Let K be a regular kernel, and let μ be an arbitrary probability measure on the Borel sets of R^d. Then there exists a finite constant $\varrho = \varrho(K)$ only depending upon K such that*

$$\sup_{y \in R^d, h>0} \int \frac{K_h(x-y)}{EK_h(x-X)} \mu(\mathrm{d}x) \leq \varrho.$$

Also, for any $\delta, \varepsilon > 0$, there exists $h_0 > 0$ such that

$$\sup_{y \in R^d, h \leq h_0} \int \frac{K_h(x-y) I_{(\|x-y\| \geq \delta)}}{EK_h(x-X)} \mu(\mathrm{d}x) \leq \varepsilon.$$

Proof. First we find a bounded overlap cover of R^d with translates of $S_{r/2}$. This cover has an infinite number of member balls, but every x gets covered at most k_1 times where k_1 depends upon d only. The centers of the balls are called x_i, $i = 1, 2, \ldots$. The integral condition on K implies that

$$\sum_{i=1}^{\infty} \sup_{x \in x_i + S_{r/2}} K(x) \leq \frac{k_1}{\int_{S_{r/2}} \mathrm{d}x} \int \sup_{y \in x + S_r} K(y) \, \mathrm{d}x \leq k_2$$

for another finite constant k_2. Clearly,

$$K_h(x-y) \leq \sum_{i=1}^{\infty} \sup_{x \in y + hx_i + S_{rh/2}} K_h(x-y) I_{[x \in y + hx_i + S_{rh/2}]},$$

and

$$EK_h(x-X) \geq B\mu(y + hx_i + S_{rh/2}) \quad (x \in y + hx_i + S_{rh/2})$$

so that combining both inequalities,

$$\begin{aligned} \int \frac{K_h(x-y)}{EK_h(x-X)} \mu(\mathrm{d}x) &\leq \sum_{i=1}^{\infty} \int_{x \in y + hx_i + S_{rh/2}} \frac{\sup_{z \in hx_i + S_{rh/2}} K_h(z)}{B\mu(y + hx_i + S_{rh/2})} \mu(\mathrm{d}x) \\ &= \sum_{i=1}^{\infty} \frac{\mu(y + hx_i + S_{rh/2}) \sup_{z \in hx_i + S_{rh/2}} K_h(z)}{B\mu(y + hx_i + S_{rh/2})} \\ &= \frac{1}{B} \sum_{i=1}^{\infty} \sup_{z \in x_i + S_{r/2}} K(z) \leq \frac{k_2}{B}. \end{aligned}$$

To obtain the second inequality in Lemma 1, substitute $K_h(z)$ above by $K_h(z) I_{(\|z\| \geq \delta)}$ and notice that

$$\begin{aligned} \int \frac{K_h(x-y) I_{(\|x-y\| \geq \delta)}}{EK_h(x-X)} \mu(\mathrm{d}x) &\leq \sum_{i=1}^{\infty} \sup_{z \in x_i + S_{r/2}} K(z) I_{(\|z\| \geq \delta/h)} \\ &\to 0 \quad \text{as } h \downarrow 0. \end{aligned}$$

□

4. A technical lemma

Lemma 2. *Assume that K is a regular kernel, and that h is such that $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$. Let m_n be the kernel regression estimate with kernel K, and let $0/0$ be defined as 0. Then, for every distribution of (X, Y) with $|Y| \le M < \infty$,*

$$\lim_{n\to\infty} E\left(\int |U_n(x)|\, \mu(\mathrm{d}x) + \int |L_n(x) - 1|\, \mu(\mathrm{d}x) \right) = 0,$$

where U_n and L_n are as defined in the proof of the theorem.

Proof. The proof is rather standard, combining a covering argument and exploiting the fact that continuous functions are dense in the space $L_1(\mu)$. Fix an arbitrary $\varepsilon > 0$ and find a continuous function $g \in L_1(\mu)$ with compact support such that $\int |m(x) - g(x)|\, \mu(\mathrm{d}x) < \varepsilon$. We will begin with a good look at $E\int |U_n|$. Clearly,

$$\begin{aligned}
E\int |U_n| &= E\int \left| \frac{\sum_{i=1}^n (Y_i - m(x)) K_h(x - X_i)}{nEK_h(x - X_1)} \right| \mu(\mathrm{d}x) \\
&\le E\int \left| \frac{\sum_{i=1}^n (Y_i - m(X_i)) K_h(x - X_i)}{nEK_h(x - X_1)} \right| \mu(\mathrm{d}x) \\
&\quad + E\int \left| \frac{\sum_{i=1}^n (g(X_i) - m(X_i)) K_h(x - X_i)}{nEK_h(x - X_1)} \right| \mu(\mathrm{d}x) \\
&\quad + E\int \left| \frac{\sum_{i=1}^n (g(X_i) - g(x)) K_h(x - X_i)}{nEK_h(x - X_1)} \right| \mu(\mathrm{d}x) \\
&\quad + E\int \left| \frac{\sum_{i=1}^n (m(x) - g(x)) K_h(x - X_i)}{nEK_h(x - X_1)} \right| \mu(\mathrm{d}x) \\
&\triangleq \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned}$$

We have for arbitrary $\delta > 0$, by our choice of g,

$$\begin{aligned}
\mathrm{IV} &\le \int |g(x) - m(x)|\, \mu(\mathrm{d}x) < \varepsilon, \\
\mathrm{III} &\le \sum_{i=1}^n \int E\left(\frac{|g(X_i) - g(x)|\, K_h(x - X_i)}{nEK_h(x - X_1)} \right) \mu(\mathrm{d}x) \\
&= \sum_{i=1}^n \int E\left(\frac{|g(X_i) - g(x)|\, K_h(x - X_i)}{nEK_h(x - X_1)} I_{(\|x - X_i\| < \delta)} \right) \mu(\mathrm{d}x) \\
&\quad + \sum_{i=1}^n \int E\left(\frac{|g(X_i) - g(x)|\, K_h(x - X_i)}{nEK_h(x - X_1)} I_{(\|x - X_i\| \ge \delta)} \right) \mu(\mathrm{d}x) \\
&\le \sup_x \sup_{y \in x + \delta h} |g(y) - g(x)|
\end{aligned}$$

$$+2\sup|g(x)|\sum_{i=1}^{n}\int\frac{E(K_h(x-X_i)I_{(\|x-X_i\|\geq\delta)})}{nEK_h(x-X_1)}\mu(\mathrm{d}x)$$

$$<\varepsilon$$

for h small enough. Furthermore,

$$\mathrm{II}\leq E\int\frac{\sum_{i=1}^{n}|g(X_i)-m(X_i)|\,K_h(x-X_i)}{nEK_h(x-X_1)}\mu(\mathrm{d}x)$$

$$=\int\frac{E(|g(X_1)-m(X_1)|\,K_h(x-X_1))}{EK_h(x-X_1)}\mu(\mathrm{d}x)$$

$$\leq\sup_y\int\frac{K_h(x-y)}{EK_h(x-X)}\mu(\mathrm{d}x)\times E\,|g(X)-m(X)|\leq\varrho\varepsilon$$

where ϱ is the constant of Lemma 1. To treat I, observe that it can be written as $\int\psi(x)\,\mu(\mathrm{d}x)$ for some function ψ taking values on $[0,2M]$. Thus, we can split the integral into an integral over a huge bounded cube Q, and its complement, so that $2M\mu(Q^c)<\varepsilon$. This leaves us with the task of bounding $\int_Q\psi(x)\mu(\mathrm{d}x)$. By the Cauchy–Schwarz inequality,

$$\mathrm{I}\leq\varepsilon+\int_Q\psi(x)\mu(\mathrm{d}x)$$

$$\leq\varepsilon+E\int_Q E^{1/2}\left(\left(\frac{\sum_{i=1}^{n}(Y_i-m(X_i))K_h(x-X_i)}{nEK_h(x-X_1)}\right)^2\,\middle|\,X_1,\ldots,X_n\right)\mu(\mathrm{d}x)$$

$$=\varepsilon+E\int_Q\left\{\frac{E((Y_1-m(X_1))^2K_h^2(x-X_1)\,|\,X_1)}{nE^2K_h(x-X)}\right\}^{1/2}\mu(\mathrm{d}x)$$

$$\leq\varepsilon+\frac{2M}{\sqrt{n}}\int_Q\left\{\frac{EK_h^2(x-X)}{E^2K_h(x-X)}\right\}^{1/2}\mu(\mathrm{d}x)$$

$$\leq\varepsilon+\frac{2M}{\sqrt{nh^d}}\left\{\int_Q\frac{h^dEK_h^2(x-X)}{E^2K_h(x-X)}\mu(\mathrm{d}x)\right\}^{1/2}\quad\text{(Jensen's inequality)}$$

$$\leq\varepsilon+\frac{2M}{\sqrt{nh^d}}\left\{\int_Q\frac{ch^d}{EK_h(x-X)}\mu(\mathrm{d}x)\right\}^{1/2},$$

where c is an upper bound on K (regularity of K implies that c is finite). The second term in the last expression tends to 0 as $n\to\infty$ because $nh^d\to\infty$ and the integral in the term is uniformly bounded over all $h\geq1$. To see that, note that $EK_h(x-X)\geq B\mu(x+S_{rh})$, that Q can be covered by at most $c_1+c_2h^{-d}$ translates of $S_{rh/2}$ for some constants c_1,c_2 depending upon d only, and that the integral over one of these translates is not larger than ch^d/B. Therefore, the integral over Q does not exceed $(c/B)(c_1h^d+c_2)$. Collecting all this shows that

$$E\int|U_n|\leq\varepsilon(3+\varrho)+\mathrm{o}(1).$$

We can use a similar argument for $E\int |L_n - 1|$. Indeed, using the same large hypercube Q as before, chosen so as to insure that $2\mu(Q^c) < \varepsilon$, we have

$$\begin{aligned} E\int |L_n - 1| &= E\int \left| \frac{\sum_{i=1}^n (K_h(x-X_i) - EK_h(x-X_1))}{nEK_h(x-X_1)} \right| \mu(\mathrm{d}x) \\ &\le 2\mu(Q^c) + E\int_Q \left| \frac{\sum_{i=1}^n (K_h(x-X_i) - EK_h(x-X_1))}{nEK_h(x-X_1)} \right| \mu(\mathrm{d}x) \\ &\le \varepsilon + \left(\int \frac{E\sum_{i=1}^n K_h^2(x-X_i)}{n^2 E^2 K_h(x-X)} \mu(\mathrm{d}x) \right)^{1/2} \\ &\le \varepsilon + \left\{ \frac{c}{nh^d} \int \frac{h^d}{EK_h(x-X)} \mu(\mathrm{d}x) \right\}^{1/2} \\ &= \varepsilon + o(1). \qquad \square \end{aligned}$$

5. Appendix: Proof of Remark 2

We will use standard truncation techniques (see Devroye and Wagner (1980) or Krzyżak (1986)). Following the proof of the Theorem, we get

$$\begin{aligned} \int |W_{i,i}| &\le \int \left| \frac{1}{n}(Y_i - \bar{Y}_i) \frac{K_h(x-X_i)}{EK_h(x-X_1)} \right| \mu(\mathrm{d}x) \\ &\quad + \int \left| \frac{1}{n}(\bar{Y}_i - g(x)) \frac{K_h(x-X_i)}{EK_h(x-X_1)} \right| \mu(\mathrm{d}x) \\ &\quad + \int \left| \frac{1}{n}(g(x) - m(x)) \frac{K_h(x-X_i)}{EK_h(x-X_1)} \right| \mu(\mathrm{d}x) \\ &\triangleq \mathrm{I} + \mathrm{II} + \mathrm{III}, \end{aligned}$$

where $g(x)$ is defined as in the proof of Lemma 2, and

$$\bar{Y} = YI_{(|Y| \le n^{1/(2+\delta)})},$$

and $\delta > 0$. For n large enough, $\mathrm{I} = 0$ almost surely. If ϱ is the constant of Lemma 1, then by Hölder's inequality, for n large enough,

$$\mathrm{II} \le 2\varrho n^{-(1+\delta)/(2+\delta)}.$$

And provided that $\int |g - m| < \varepsilon$, $\mathrm{III} < \varepsilon\varrho/n$. Hence,

$$\int |W_{i,i}| \le (2+\varepsilon)\varrho n^{-(1+\delta)/(2+\delta)} \triangleq c_1 n^{-(1+\delta)/(2+\delta)}$$

and, still in the notation of the Theorem,

$$P\left(\int |U_n| > \varepsilon_3 \right) \le 2\exp\left(-\frac{\varepsilon_5^2}{2c_1^2} n^{\delta/(2+\delta)} \right) + P\left(E\int |U_n| > \tfrac{1}{2}\varepsilon_2 \right).$$

It remains to show that $E\int |U_n| \to 0$ as $n \to \infty$. To do this, we will use a truncation technique, in which t denotes a positive constant. Let $Y = Y' + Y''$, where $Y' = YI_{(|Y| \le t)}$ and $Y'' = YI_{(|Y| > t)}$. By Lemma 1 and the proof of Lemma 2 we have

$$E\int |U_n| \le E\int |U_n'| + E\int |U_n''| \le O(t/\sqrt{nh^d}) + 2\varrho E|Y''| + \varepsilon(2+\varrho),$$

where

$$U_n' = \sum_{i=1}^{n} (Y_i' - m'(x))K_h(x - X_i)/(nEK_h(x - X_1)),$$

$$U_n'' = \sum_{i=1}^{n} (Y_i'' - m''(x))K_h(x - X_i)/(nEK_h(x - X_1)),$$

$m'(x) = E(Y' \mid X = x)$ and $m''(x) = E(Y'' \mid X = x)$. First choose t large enough so that the second term on the r.h.s. of the above inequality becomes small, and then let n grow large to make the first term small.

References

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Math. J.* **37**, 357-367.

Collomb, G. (1981). Estimation non parametrique de la regression: revue bibliographique. *Internat. Statist. Rev.* **49**, 75-93.

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

Devroye, L. (1983). The equivalence of weak, strong and complete convergence in $L1$ for kernel density estimates. *Ann. Statist.* **11**, 896-904.

Devroye, L. (1988). The kernel estimate is relatively stable. *Probab. Theory Related Fields* **77**, 521-536.

Devroye, L. and L. Györfi (1983). Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function. In: F. Konecny, J. Mogyorodi, W. Wertz, Eds., *Proceedings of the Fourth Panonian Symposium on Mathematical Statistics.* Akademiai Kiado, Budapest, Hungary, 67-76.

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

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

Glick, N. (1974). Consistency conditions for probability estimators and integrals of density estimators. *Utilitas Math.* **6**, 61-74.

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

Györfi, L. (1981). Recent results on nonparametric regression estimate and multiple classification. *Problems Control Inform. Theory* **10**, 43-52.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**, 13-30.

Krzyżak, A. (1986). The rates of convergence of kernel regression estimates and classification rules. *IEEE Trans. Inform. Theory.* **32**, 668-679.

Krzyżak, A. and M. Pawlak (1984). Distribution-free consistency of a nonparametric kernel regression estimate and classification. *IEEE Trans. Inform. Theory* **30**, 78-81.

Nadaraya, E.A. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141-142.

Nadaraya, E.A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory Probab. Appl.* **15**, 134–137.

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

Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* **8**, 1348–1360.

Stout, W.F. (1974). *Almost Sure Convergence.* Academic Press, New York.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26**, 359–372.