Z. Wahrscheinlichkeitstheorie verw. Gebiete
51, 15–25 (1980)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

On the L_1 Convergence of Kernel Estimators of Regression Functions with Applications in Discrimination*

Luc P. Devroye[1] and T.J. Wagner[2]

[1] School of Computer Science, McGill University, 805 Sherbrooke W., Montreal, Canada H3A 2K6
[2] University of Texas, Austin, Texas 78712, U.S.A.

Summary. An estimate m_n of a regression function $m(x)=E\{Y|X=x\}$ is weakly (strongly) consistent in L_1 if $\int |m_n(x)-m(x)|\mu(dx)$ converges to 0 in probability (w.p. 1) as the sample size grows large (μ is the probability measure of X).

We show that the well-known kernel estimate (Nadaraya, Watson) and several recursive modifications of it are weakly (strongly) consistent in L_1 under no conditions on (X, Y) other than the boundedness of Y and the absolute continuity of μ. No continuity restrictions are put on the density corresponding to μ. We further notice that several kernel-type discrimination rules are weakly (strongly) Bayes risk consistent whenever X has a density.

Introduction

In nonparametric regression function estimation one is provided with a sequence $D_n=(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent $R^d \times R$-valued random vectors distributed as (X, Y) but is not given any information about the distribution of (X, Y) other than the existence of the *regression function* $m(x)=E\{Y|X=x\}$ (for this, it suffices that $E\{|Y|\}<\infty$). A regression function estimate, or simply *estimate*, is a function of $x \in R^d$ and the *data* D_n: $m_n(x)$. Criteria measuring the closeness of m_n to m include the uniform deviation,

$$U_n = \operatorname*{ess\,sup}_{(\mu)} |m_n(x)-m(x)|$$

and the distance in L_p,

$$I_{np}=\left(\int |m_n(x)-m(x)|^p \mu(dx)\right)^{1/p}.$$

Here μ is the (unknown) probability measure of X.

Nadaraya (1964, 1970) and Watson (1964) proposed the estimate

* Research of both authors was sponsored by AFOSR Grant 77-3385

0044-3719/80/0051/0015/$02.20

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

where $\{h_n\}$ is a sequence of positive numbers and $K \geqq 0$ is an integrable function on R^d. A variety of properties are known for (1). The pointwise convergence (in probability and in the mean square) of m_n to m is treated by Watson (1964), Rosenblatt (1969) and Noda (1976) for $d=1$ and by Greblicki (1974) for $d>1$. Schuster (1972) discusses the joint asymptotic normality of $m_n(x_1), \ldots, m_n(x_N)$ at fixed points $x_1, \ldots, x_N$. Nadaraya (1964) for $d=1$ gives conditions insuring that $U_n \xrightarrow{n} 0$ with probability one (w.p.1).

Greblicki (1974) and Ahmed and Lin (1976) prove some pointwise convergence results for a recursive version of (1).

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i h_i^{-d} K((X_i - x)/h_i)}{\sum_{i=1}^{n} h_i^{-d} K((X_i - x)/h_i)}. \tag{2}$$

The recursive computation of (2) can be carried out by

$$\begin{aligned} &m_0(x) = f_0(x) = 0, \\ &f_n(x) = (h_n/h_{n-1})^d f_{n-1}(x) + K((X_n - x)/h_n), \\ &m_n(x) = m_{n-1}(x) + f_n^{-1}(x)(Y_n - m_{n-1}(x))\, K((X_n - x)/h_n). \end{aligned} \tag{3}$$

A still simpler recursive estimate which the authors believe is new reads

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i K((X_i - x)/h_i)}{\sum_{i=1}^{n} K((X_i - x)/h_i)}, \tag{4}$$

or equivalently,

$$\begin{aligned} &m_0(x) = f_0(x) = 0, \\ &f_n(x) = f_{n-1}(x) + K((X_n - x)/h_n), \\ &m_n(x) = m_{n-1}(x) + f_n^{-1}(x)(Y_n - m_{n-1}(x))\, K((X_n - x)/h_n). \end{aligned} \tag{5}$$

The main theorem of this paper involves the weak and the strong convergence to 0 of

$$I_n = \int |m_n(x) - m(x)|\, \mu(dx)$$

for the estimates (1), (2) and (4) under the following standing conditions:

μ is an absolutely continuous (with respect to Lebesgue measure) probability measure, (6)

$|Y| \leqq c < \infty$ with probability one, (7)

K is a nonnegative bounded integrable function on R^d whose radial majorant ψ is integrable:

$$\psi(x) \triangleq \sup_{\|y\| \geqq \|x\|} K(y), \qquad \int \psi(x)\, dx < \infty. \tag{8}$$

Stone (1977) has shown that a large class of regression function estimates including the ones of the nearest neighbor type, satisfy $E\{I_n\} \xrightarrow{n} 0$ for *all* possible distributions of (X, Y) with $E\{|Y|\} < \infty$. For estimate (1), the same was shown by Devroye and Wagner (1979) whenever

$$h_n \xrightarrow{n} 0, \tag{9}$$

$$n h_n^d \xrightarrow{n} \infty, \tag{10}$$

and

K is a bounded nonnegative function with compact support such that for a small sphere S about the origin, $\inf_{x \in S} K(x) > 0$.

Theorem 1 below complements this result in the sense that the almost sure convergence to 0 is established for I_n under weaker conditions on K but slightly stronger conditions on the distribution of (X, Y) (see (6–7)). We note here that Theorem 1 below is *density-free:* it is valid for all random vectors X possessing a density. Also, we are not putting a continuity condition on m, and the random variable Y need not have a density at all. The condition (8) is intimately related to but not implied by the well-known condition $\|x\|^d K(x) \to 0$ as $\|x\| \to \infty$ (incidentally, this condition is equivalent to $\|x\|^d \psi(x) \to 0$ as $\|x\| \to \infty$). For example, (8) holds if K is a bounded function and either has compact support, or satisfies

$$\|x\|^{d+\varepsilon} K(x) \to 0 \quad \text{as } \|x\| \to \infty$$

or

$$\|x\|^d (\log \|x\|)^{1+\varepsilon} K(x) \to 0 \quad \text{as } \|x\| \to \infty$$

for some $\varepsilon > 0$.

Theorem 1. *Assume that* (6–9) *hold.*

(i) $I_n \xrightarrow{n} 0$ *in probability for estimate* (1) *if* (10) *holds;* $I_n \xrightarrow{n} 0$ w.p.1 *for estimate* (1) *if*

$$\sum_{n=1}^{\infty} \exp(-\alpha n h_n^d) < \infty \quad \textit{for all } \alpha > 0. \tag{11}$$

(ii) $I_n \xrightarrow{n} 0$ *in probability for estimate* (2) *if*

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{h_i^d}\xrightarrow{n}0; \tag{12}$$

$I_n \xrightarrow{n} 0$ w.p.1 *for estimate* (2) *if either*

$$\sum_{n=1}^{\infty}\frac{1}{n^2 h_n^d}<\infty \tag{13}$$

or

$$n h_n^d/\log\log n \xrightarrow{n} \infty. \tag{14}$$

(iii) $I_n \xrightarrow{n} 0$ w.p.1 *for estimate* (4) *if*

$$\sum_{n=1}^{\infty} h_n^d=\infty. \tag{15}$$

Remark 1 (Related Work with The Stochastic Approximation Method).

Révész (1973, 1977) for $d=1$ studies the integral convergence on compact sets of the recursive estimate defined by

$$m_0(x)=0; \qquad m_n(x)=m_{n-1}(x)+(n h_n^d)^{-1}(Y_n-m_{n-1}(x))\,K((X_n-x)/h_n). \tag{16}$$

His proofs of convergence are rooted in the well-known theorems of convergence for stochastic approximation methods. The approach followed in this note is more directly related to the laws of large numbers. It is curious that (5) is in form similar to (16) if $f_n(x)$ is replaced by nh_n^d. For large $n f_n(x)$ is close to $nh_n^d f(x)$ when $f(x)$ is the value of the density of X at x. Thus (5) and (16) can be expected to behave in a similar fashion for large n.

Remark 2 (Conditions on $\{h_n\}$).

The conditions for weak consistency are strictly nested: (10) implies (12) and (12) implies (15). To see this, use

$$\sum_{i=1}^{n} h_i^d \geqq \left(\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{h_i^d}\right)^{-1} \geqq \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{i h_i^d}\right)^{-1}$$

and Toeplitz's Lemma (Loève, 1963, pp. 238). Condition (11) is implied by

$$n h_n^d/\log n \xrightarrow{n} \infty, \tag{17}$$

which clearly is stronger than (14): $n h_n^d/\log\log n \xrightarrow{n} \infty$.

An Application in Discrimination

In discrimination Y is integer-valued: $Y\in\{1,\ldots,M\}$. If (X, Y) is independent of D_n, then Y is estimated from X and D_n by $g_n(X)$, which also takes values in $\{1,\ldots,M\}$. The performance with a certain mapping g_n (discrimination rule) is measured by its *probability of error*,

$$L_n=P\{g_n(X)\neq Y\,|\,D_n\}.$$

In any case, L_n can not be smaller than the Bayes probability of error

$$L^* = \inf_{g: R^d \to \{1, \dots, M\}} P\{g(X) \neq Y\}.$$

If

$$p_i(x) = E\{I_{\{Y=i\}} | X = x\} = P\{Y = i | X = x\}, \qquad 1 \leqq i \leqq M, \qquad x \in R^d,$$

(here I is the indicator function) then all discrimination rules g satisfying

$$g(x) \neq i \quad \text{whenever } p_i(x) < \max_{1 \leqq l \leqq M} p_l(x)$$

achieve L^*.
The unknown p_i can be estimated by p_{ni} using any of the above mentioned methods (1), (2) or (4), and g_n can then be picked such that

$$g_n(x) \neq i \quad \text{whenever } p_{ni}(x) < \max_{1 \leqq l \leqq M} p_{nl}(x). \tag{18}$$

Since

$$L^* = E\{1 - \max_i p_i(X)\}, \qquad L_n = E\{1 - p_{g_n(X)}(X) | D_n\}$$

and

$$p_{n g_n(x)}(x) = \max_i p_{ni}(x),$$

we have

$$\begin{aligned} 0 \leqq L_n - L^* &= E\{\max_i p_i(X) - p_{g_n(X)}(X) | D_n\} \\ &= E\{\max_i p_i(X) - \max_i p_{ni}(X) | D_n\} \\ &\quad + E\{p_{n g_n(X)}(X) - p_{g_n(X)}(X) | D_n\} \\ &\leqq 2 \sum_{i=1}^{M} E\{|p_i(X) - p_{ni}(X)| \, | D_n\}. \end{aligned} \tag{19}$$

The inequality (19) is valid for *all* discrimination rules satisfying (18). It will allow us to draw conclusions from the convergence of I_n to 0 regarding the convergence of L_n to L^*.

The rules resulting from (1) are well-known: they satisfy

$$g_n(x) \neq i$$

whenever

$$\sum_{j: Y_j = i} K((X_j - x)/h_n) < \max_{1 \leqq l \leqq M} \sum_{j: Y_j = l} K((X_j - x)/h_n). \tag{20}$$

Rules of this type are studied by Greblicki (1974, 1977, 1978), Devroye and Wagner (1976), Rejtö and Révész (1973) and Van Ryzin (1966). The starting point in all these papers is the Parzen-Rosenblatt density estimate (Parzen, 1962; Rosenblatt, 1957; Cacoullos, 1965). The rule (20) is also mentioned in the early works of Fix and Hodges (1951), Sebestyen (1962) and Meisel (1969) for special K (see also Bashkirov, Braverman and Muchnik (1964)). Pattern recognition procedures that are derived from *any* type of density estimate are discussed by Glick (1972, 1976), Greblicki (1974, 1977, 1978) and Devroye and Wagner (1976).

The rule obtained from (2) and (18) is the one first proposed by Wolverton and Wagner (1969) and later discussed by Rejtö and Révész (1973): let g_n satisfy

$$g_n(x) \neq i \tag{21}$$

whenever

$$\sum_{j: Y_j = i} h_j^{-d} K((X_j - x)/h_j) < \max_{1 \leq l \leq M} \sum_{j: Y_j = l} h_j^{-d} K((X_j - x)/h_j).$$

Combining (4) and (18) gives the very simple rule: let g_n be such that

$$g_n(x) \neq i$$

whenever

$$\sum_{j: Y_j = i} K((X_j - x)/h_j) < \max_{1 \leq l \leq M} \sum_{j: Y_j = l} K((X_j - x)/(h_j). \tag{22}$$

The following theorem, an immediate corollary of Theorem 1 and inequality (19), establishes *the weak and strong Bayes risk consistency of the rules* (20), (21) *and* (22) *under no restrictions whatsoever on the density of* X. The conditions of convergence are weaker than those reported by Wolverton and Wagner (1969), Van Ryzin (1966, 1967), Rejtö and Révész (1973), Greblicki (1974, 1977, 1978) and Devroye and Wagner (1976).

Theorem 2. *Assume that* (6), (8) *and* (9) *hold.*

(i) *For the discrimination rules* (20), $L_n \xrightarrow{n} L^*$ *in probability if* (10) *holds; also,* $L_n \xrightarrow{n} L^*$ w.p.1 *if* (11) *holds.*

(ii) *For the discrimination rules* (21), $L_n \xrightarrow{n} L^*$ *in probability if* (12) *holds; of* (13) *or* (14) *are satisfied then* $L_n \xrightarrow{n} L^*$ w.p.1.

(iii) *For the discrimination rules* (22), $L_n \xrightarrow{n} L^*$ w.p.1 *if* (15) *is satisfied.*

Proofs

Lemma 1. *Let* m *be a regression function and let* m_n *be a regression function estimate. Let further for some* $c < \infty$, $|m_n| < c$, $|m| < c$. *Then* $I_n \xrightarrow{n} 0$ *in probability* (w.p.1) *when* $m_n(x) \xrightarrow{n} m(x)$ *in probability* (w.p.1) *for almost all* $x(\mu)$.

Proof. For the weak convergence part let A be the set on which $m_n(x) \xrightarrow{n} m(x)$ in probability. By the Lebesgue dominated convergence theorem $E\{|m_n(x) - m(x)|\} \xrightarrow{n} 0$ on A. By another application of the Lebesgue dominated convergence theorem,

$$E\{I_n\} = \int E\{|m_n(x) - m(x)|\} \mu(dx) \xrightarrow{n} 0,$$

from which Lemma 1 follows by Markov's inequality.

For the strong convergence part of Lemma 1, we let A be the set on which $m_n(x) \xrightarrow{n} m(x)$ w.p.1, and let $(\Omega, \mathfrak{J}, P)$ be the probability space of (X_1, Y_1), $(X_2, Y_2), \ldots$. We write ω for the probability element of Ω. By Fubini's theorem

$$P\{\omega: m_n(x) \not\rightarrow m(x)\} = 0 \qquad \text{for almost all } x(\mu),$$

if and only if

$$\{(\omega, x): m_n(x) \not\rightarrow m(x)\} \qquad \text{has measure } (P x \mu) \text{ zero},$$

if and only if

$$\mu(\{x: m_n(x) \not\rightarrow m(x)\}) = 0 \qquad \text{for almost all } \omega(P).$$

Let Ω' be this set of $\omega \in \Omega$. But for every $\omega \in \Omega'$, $I_n = \int |m_n(x) - m(x)| \mu(dx \xrightarrow{n} 0$ by the Lebesgue dominated convergence theorem. Since $P\{\Omega'\} = 1$, Lemma 1 is proved. Q.E.D.

Proof of Theorem 1. If c is the constant of (7) then (1), (2) and (4) satisfy $|m_n| \leqq c$, $|m| \leqq c$. By Lemma 1 we need only show that $m_n(x) \xrightarrow{n} m(x)$ in probability (w.p.1) for almost all $x(\mu)$. Let us call

$$Z_n^1(x) = \frac{1}{n h_n^d} \sum_{i=1}^{n} Y_i K((X_i - x)/h_n),$$

$$Z_n^2(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i^d} Y_i K((X_i - x)/h_i),$$

$$Z_n^3(x) = \left(\sum_{i=1}^{n} h_i^d\right)^{-1} \sum_{i=1}^{n} Y_i K((X_i - x)/h_i).$$

If we can show that under the conditions of Theorem 1 $Z_n^i(x) \xrightarrow{n} m(x) f(x)$ in probability (w.p.1) for almost all $x(\mu)$, then we have also shown that $W_n^i(x) \xrightarrow{n} f(x)$ in probability (w.p.1) for almost all $x(\mu)$ where $W_n^i(x)$ is defined as $Z_n^i(x)$ with $Y_i \equiv 1$ (and thus $m \equiv 1$). Since in all three cases $m_n(x) = Z_n^i(x)/W_n^i(x)$ and since for almost all $x(\mu)$ $f(x) > 0$ we can conclude that $m_n(x) \xrightarrow{n} m(x)$ in probability (w.p.1) for almost all $x(\mu)$.

We can split the proof into two parts in view of

$$|Z_n^i(x) - m(x) f(x)| \leqq |Z_n^i(x) - E\{Z_n^i(x)\}| + |E\{Z_n^i(x)\} - m(x) f(x)|, \qquad i = 1, 2, 3.$$

First we use a theorem of Stein (1970, pp. 62–63) regarding the following function:

$$\begin{aligned}\phi(x,s)&=E\{s^{-d}Y_1K((X_1-x)/s)\}\\&=E\{s^{-d}m(X_1)K((X_1-x)/s)\}\\&=\int s^{-d}K((y-x)/s)\,m(y)\,f(y)\,dy.\end{aligned}$$

If $mf\geqq 0$, $\int m(x)f(x)\,dx<\infty$, $K\geqq 0$, $\int K(x)\,dx=1$ and $\int\psi(x)\,dx<\infty$ (here ψ is the radial majorant of K) then $\phi(x,s)\to m(x)f(x)$ as $s\to 0$ for almost all (Lebesgue measure) x. Let us collect these x in a set A. Since f is a density, we obviously have that $\mu(A)=1$. Notice that we can always take K such that $\int K(x)\,dx=1$ since K appears in both the denominator and numerator of m_n.

For the estimate (1), $E\{Z_n^1(x)\}=\phi(x,h_n)\xrightarrow{n}m(x)f(x)$ as $h_n\xrightarrow{n}0$, $x\in A$. It follows that for $x\in A$,

$$E\{Z_n^2(x)\}=n^{-1}\sum_{i=1}^{n}\phi(x,h_i)\xrightarrow{n}m(x)f(x)\qquad\text{as } h_n\xrightarrow{n}0$$

and

$$E\{Z_n^3(x)\}=\left(\sum_{i=1}^{n}h_i^d\right)^{-1}\sum_{i=1}^{n}h_i^d\,\phi(x,h_i)\xrightarrow{n}m(x)f(x)\qquad\text{as } h_n\xrightarrow{n}0$$

whenever $\sum_{i=1}^{\infty}h_i^d=\infty$.

Theorem 1 is thus proved if we can show that for all $x\in A$ $Z_n^i(x)-E\{Z_n^i(x)\}\xrightarrow{n}0$ in probability (w.p.1), $i=1,2,3$. Let us consider Z_n^1 first. Clearly,

$$Z_n^1(x)-E\{Z_n^1(x)\}=n^{-1}\sum_{i=1}^{n}(T_i(h_n)-E\{T_i(h_n)\})$$

where $T_i(s)=s^{-d}Y_iK((X_i-x)/s)$. If $b=\sup_x K(x)$, $c'=\sup_n\phi(x,h_n)$ (c' is finite if $x\in A$ but should depend on x), then we notice that

$$|T_i(h_n)|\leqq bc/h_n^d\qquad\text{w.p.1}$$

and

$$E\{T_i^2(h_n)\}\leqq(bc/h_n^d)\,E\{T_i(h_n)\}\leqq bcc'/h_n^d.$$

If $U_1,\dots,U_n$ are independent random variables with $|U_i|\leqq b$, $E\{U_i\}=0$, $E\{U_i^2\}\leqq\sigma_i^2$, then an inequality due to Bennett (1962, pp. 39) (see also Hoeffding (1963, pp. 16)) states that

$$\begin{aligned}&P\{|n^{-1}\sum_{i=1}^{n}U_i|\geqq\varepsilon\}\\&\quad\leqq 2\exp\{-n(\varepsilon/2b)((1+\sigma^2/2b\varepsilon)\log(1+2b\varepsilon/\sigma^2)-1)\}\\&\quad\leqq 2\exp\{-n\varepsilon^2/2(\sigma^2+b\varepsilon)\}\end{aligned}$$

where $\sigma^2 = n^{-1} \sum_{i=1}^{n} \sigma_i^2$ and the latter inequality follows from $\log(1+u) > 2u/(2+u)$ for all $u>0$. Applying this inequality yields

$$\begin{aligned} &P\{|Z_n^1(x) - E\{Z_n^1(x)\}| \leqq \varepsilon\} \\ &\quad \leqq 2\exp\{-n\varepsilon^2/(2bcc'/h_n^d + 4bc\varepsilon/h_n^d)\} \\ &\quad = 2\exp\{-\alpha n h_n^d\} \qquad \text{where } \alpha = \varepsilon^2/(2bcc' + 4bc\varepsilon). \end{aligned} \tag{23}$$

Condition (10) implies that (23) goes to 0 as $n \to \infty$ for all $\varepsilon > 0$; (11) and the Borel-Cantelli lemma are sufficient in order to be able to conclude that $Z_n^1(x) - E\{Z_n^1(x)\} \xrightarrow{n} 0$ w.p.1 for all $x \in A$.

Let us turn our attention to

$$Z_n^2(x) - E\{Z_n^2(x)\} = n^{-1} \sum_{i=1}^{n} (T_i(h_i) - E\{T_i(h_i)\}),$$

assuming that (12) holds. Since

$$\begin{aligned} E\{(Z_n^2(x) - E\{Z_n^2(x)\})^2\} &\leqq n^{-2} \sum_{i=1}^{n} E\{T_i^2(h_i)\} \\ &\leqq n^{-2} \sum_{i=1}^{n} bcc'/h_i^d \xrightarrow{n} 0, \end{aligned}$$

we conclude by Čebyšev's inequality that $Z_n^2(x) - E\{Z_n^2(x)\} \xrightarrow{n} 0$ in probability. Assume now that (13) holds. By Kolmogorov's second moment version of the strong law of large numbers (Loeve, 1963, pp. 253) we know that $Z_n^2(x) - E\{Z_n^2(x)\} \xrightarrow{n} 0$ w.p.1 if

$$\sum_{n=1}^{\infty} E\{T_n^2(h_n)\}/n^2 < \infty.$$

But in view of $E\{T_n^2(h_n)\} \leqq bcc'/h_n^d$ this condition reduces to (13). It is a bit harder to show that the same conclusion can be drawn if just (14) holds. From Loeve (1963, pp. 253) we conclude that $Z_n^2(x) - E\{Z_n^2(x)\} \xrightarrow{n} 0$ w.p.1 if $|T_n(h_n)| \leqq Ln$ for all n and some $L < \infty$ (which is the case here since $|T_n(h_n)| \leqq bc/h_n^d$ and $nh_n^d/\log\log n \xrightarrow{n} \infty$) provided that

$$\sum_{k=1}^{\infty} P\left\{\left|2^{-k} \sum_{i=2^k+1}^{2^{k+1}} (T_i(h_i) - E\{T_i(h_i)\})\right| \geqq \varepsilon\right\} < \infty, \qquad \text{all } \varepsilon > 0. \tag{24}$$

If we define $\bar{h}_n = \inf_{i \leqq n} h_i$, then by another application of Bennett's inequality we see that condition (24) is satisfied if

$$\sum_{k=0}^{\infty} 2\exp\{-\alpha 2^k \bar{h}_{2^k}^d\} < \infty \qquad \text{for all } \alpha > 0. \tag{25}$$

This in turn follows from $2^k \bar{h}_{2^k}^d/\log k \xrightarrow{k} \infty$, which itself follows whenever $n\bar{h}_n^d/\log\log n \xrightarrow{n} \infty$. We now show that this is true if $nh_n^d/\log\log n \xrightarrow{n} \infty$. Indeed,

$$n\bar{h}_n^d/\log\log n \geqq \min\{\inf_{i>N} ih_i^d/\log\log i;\ n\bar{h}_N^d/\log\log n\}. \tag{26}$$

The right hand side of (26) can be made aritrarily large by first picking N large enough (here we use (14)) and then letting n grow unbounded.

We finally show that under condition (15) $Z_n^3(x) - E\{Z_n^3(x)\} \xrightarrow{n} 0$ w.p.1.

$$Z_n^3(x) - E\{Z_n^3(x)\} = \left(\sum_{i=1}^n h_i^d\right)^{-1} \sum_{i=1}^n (h_i^d T_i(h_i) - h_i^d E\{T_i(h_i)\})$$

tends to 0 w.p.1 if

$$\sum_{n=1}^\infty E\{(h_n^d T_n(h_n))^2\}/\left(\sum_{i=1}^n h_i^d\right)^2 < \infty \tag{27}$$

(Loeve, 1963, pp. 253). Since $E\{(h_n^d T_n(h_n))^2\} \leqq bcc' h_n^d$, (27) is satisfied if

$$\sum_{n=1}^\infty h_n^d/\left(\sum_{i=1}^n h_i^d\right)^2 < \infty.$$

Assume that $h_1 > 0$. Then from (15) we deduce the following inequality:

$$\begin{aligned}
&\sum_{n=1}^\infty h_n^d/\left(\sum_{i=1}^n h_i^d\right)^2 \\
&\quad \leqq 1/h_1^d + \sum_{n=2}^\infty h_n^d \left(\sum_{i=1}^n h_i^d\right)^{-1} \left(\sum_{i=1}^{n-1} h_i^d\right)^{-1} \\
&\quad = 1/h_1^d + \sum_{n=2}^\infty \left(\left(\sum_{i=1}^{n-1} h_i^d\right)^{-1} - \left(\sum_{i=1}^n h_i^d\right)^{-1}\right) \\
&\quad = 2/h_1^d < \infty. \quad \text{Q.E.D.}
\end{aligned}$$

References

1. Ahmad, I.A., Lin, P.: Nonparametric sequential estimation of a multiple regression function. Bull. Math. Statist. **17**, 63–75 (1976)
2. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control **25**, 917–936 (1964a)
3. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I.: The probability problem of pattern recognition learing and the method of potential functions. Automat. Remote Control **25**, 1307–1323 (1964b)
4. Bashkirov, O.A., Braverman, E.M., Muchnik, I.E.: Potential function algorithms for pattern recognition learning machines. Automat. Remote Control **25**, 692–695 (1964)
5. Bennett, G.: Probability inequalities for the sum of independent random variables. J. Amer. Statist. Assoc. **57**, 33–45 (1962)
6. Cacoullos, T.: Estimation of a multivariate density. Ann. Inst. Statist. Math. **18**, 179–190 (1965)
7. Devroye, L.P., Wagner, T.J.: Nonparametric discrimination and density estimation. Electronics Research Center, Univ. of Texas. Technical Report 183. 1976.

8. Devroye, L., Wagner, T.J.: Distribution free consistency results in nonparametric discrimination and regression function estimation. [To appear in Ann. Statist. 1979]
9. Fix, E., Hodges, J.L.: Discriminatory analysis. Nonparametric discrimination: consistency properties. USAF School of Aviation Medicine, Randolph Field, Texas. Report 4. Project No. 21-49-004. 1951
10. Glick, N.: Sample-based classification procedures derived from density estimators. J. Amer. Statist. Assoc. **67**, 116–122 (1972)
11 Glick, N.: Sample-based classification procedures related to empiric distributions. IEEE Trans. Information Theory IT-**22**, 454–461 (1976)
12. Greblicki, W.: Asymptotically optimal probabilistic algorithms for pattern recognition and identification. Monografie No. 3. Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wroclawsjiej Nr. 18. Wroclaw Poland 1974.
13. Greblicki, W.: Pattern recognition procedures with nonparametric density estimates. Submitted to the IEE Trans. Systems, Man and Cybernetics (1977)
14. Greblicki, W.: Asymptotically optimal pattern recognition procedures with density estimates. To appear in the IEEE Trans. Information Theory (1978)
15. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. **58**, 13–30 (1963)
16. Loève, M.: Probability theory. Princeton, New Jersey: Van Nostrand 1963
17. Meisel, W.: Potential functions in mathematical pattern recognition. IEEE Trans. Computers **18**, 911–918 (1969)
18. Nadaraya, E.A.: On estimating regression. Theor. Probab. Appl. **9**, 141–142 (1964)
19. Nadaraya, E.A.: Remarks on nonparametric estimates for density functions and regression curves. Theor. Probab. Appl. **15**, 134–137 (1970)
20. Noda, K.: Estimation of a regression function by the Parzen kernel-type density estimators. Ann. Inst. Statist. Math. **28**, 221–234 (1976)
21. Parzen, E.: On the estimation of a probability density function and the mode. Ann. Math. Statist. **33**, 1065–1076 (1962)
22. Rejtö, L., Révész, P.: Density estimation and pattern classification. Problems of Control and Information Theory **2**, 67–80 (1973)
23. Révész, P.: Robbins-Monroe procedure in a Hilbert space and its application in the theory of learning processes. I. Studia Sci. Math. Hungar. **8**, 391–398 (1973)
24. Révész, P.: How to apply the method of stochastic approximation in the nonparametric estimation of a regression function. Math. Operationsforsch. Statist. Ser. Statist. **8**, 119–126 (1977)
25. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. **27**, 832–837 (1957)
26. Rosenblatt, M.: Conditional probability density and regression estimators. Multivariate Analysis II. pp. 25–31. P.R. Krishnaiah Ed. New York: Academic Press 1969
27. Schuster, E.F.: Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. Ann. Math. Statist. **43**, 84–88 (1972)
28. Sebestyen, G.: Decision making processes in pattern recognition. New York: Macmillan 1962
29. Stein, E.M.: Singular integrals and differentiability properties of functions. Princeton, New Jersey: Princeton Univ. Press 1970
30. Stone, C.J.: Consistent nonparametric regression. Annals of Statistics **5**, 595–645 (1977)
31. Van Ryzin, J.: Bayes risk consistency of classification procedures using density estimation. Sankhyā Ser. A **28**, 161–170 (1966)
32. Van Ryzin, J.: A stochastic a posteriori updating algorithm for pattern recognition. J. Math. Anal. Appl. **20**, 359–379 (1967)
33. Watson, G.S.: Smooth regression analysis. Sankhyā Ser. A **26**, 359–372 (1964)
34. Wolverton, C.T., Wagner, T.J.: Asymptotically optimal discriminant functions for pattern classification. IEEE Trans. Information Theory IT-**15**, 258–265 (1969)

Received April 20, 1978; in revised form March 15, 1979