JOURNAL OF MULTIVARIATE ANALYSIS 10, 539–550 (1980)

Consistency of a Recursive Nearest Neighbor Regression Function Estimate

LUC DEVROYE

McGill University, Montreal, Canada

AND

GARY L. WISE

University of Texas, Austin, Texas 78712

Communicated by M. Rosenblatt

Let (X, Y) be an $\mathbb{R}^d \times \mathbb{R}$-valued random vector and let $(X_1, Y_1),\ldots, (X_N, Y_N)$ be a random sample drawn from its distribution. Divide the data sequence into disjoint blocks of length $l_1,\ldots, l_n$, find the nearest neighbor to X in each block and call the corresponding couple (X_i^*, Y_i^*). It is shown that the estimate $m_n(X) = \sum_{i=1}^n w_{ni} Y_i^* / \sum_{i=1}^n w_{ni}$ of $m(X) = E\{Y|X\}$ satisfies $E\{|m_n(X) - m(X)|^p\} \to 0$ $(p \geqslant 1)$ whenever $E\{|Y|^p\} < \infty$, $l_n \to \infty$, and the triangular array of positive weights $\{w_{ni}\}$ satisfies $\sup_{l<n} w_{nl} / \sum_{i=1}^n w_{ni} \to 0$. No other restrictions are put on the distribution of (X, Y). Also, some distribution-free results for the strong convergence of $E\{|m_n(X) - m(X)|^p | X_1, Y_1,\ldots, X_N, Y_N\}$ to zero are included. Finally, an application to the discrimination problem is considered, and a discrimination rule is exhibited and shown to be strongly Bayes risk consistent for all distributions.

1. INTRODUCTION AND SUMMARY

It is reasonable to expect that with a large amount of empirical data we could achieve a good estimate of a regression function. However, with a large amount of data, we may be faced with computational burdens in processing them. Therefore, a recursive method of estimation may seem attractive. In this paper we present distribution-free consistency results for the recursive nonparametric regression function estimation problem.

Received August 23, 1978; revised November 5, 1979.

AMS 1970 Subject Classifications: 6205, 62H30.

Key words and phrases: Consistency, recursive estimation, regression function, nearest neighbors, weak convergence, nonparametric estimation.

0047-259X/80/040539–12$02.00/0

Assume that (X, Y), (X_1, Y_1),..., (X_N, Y_N) are independent identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random vectors with $E\{|Y|\} < \infty$. Consider estimating the regression function

$$m(x) = E\{Y|X = x\}$$

from the *data*, (X_1, Y_1),..., (X_N, Y_N). We propose the following estimate. Break the data up into disjoint blocks of length $l_1, l_2, \ldots, l_n$, and among all X_i in the jth block, find the one that is closest to x using the 1_q norm $\|\cdot\|$ on $\mathbb{R}^d$ (in case of a tie, pick the X_i with the lowest index i; another more efficient way of handling ties will be mentioned later). Let us call the corresponding $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X_j^*, Y_j^*). The dependency on x is suppressed for the sake of brevity.

If $\{\{w_{ni}, \ldots, w_{nn}\}, n \geqslant 1\}$ is a triangular array of positive weights, then we propose to estimate $m(x)$ by

$$m_n(x) = \frac{\sum_{j=1}^n w_{nj} Y_j^*}{\sum_{j=1}^n w_{nj}}, \tag{1}$$

when $N = l_1 + \cdots + l_n$ *observations* (X_i, Y_i) are available. Note that when $w_{ni} = v_i$ for all n, i, then the computation in (1) can be performed recursively. That is, there is no need to store all the observations (X_i, Y_i), and if we are not satisfied with m_n we can collect more observations and update our estimate. Also, (1) retains the flavor of the nearest neighbor estimates (Royall, 1966; Cover, 1968; Stone, 1977), but the processing burden arising from the ranking procedure is less. The conditions which we put upon l_n and w_{ni} are weak:

$$l_n \xrightarrow{n} \infty, \tag{2}$$

$$\sup_{1 \leqslant i \leqslant n} w_{ni} \Big/ \sum_{j=1}^n w_{nj} \xrightarrow{n} 0. \tag{3}$$

We wish to investigate which consistency properties of m_n hold without additional restrictions on the joint distribution of (X, Y).

The classical nearest neighbor estimate is defined as (1) except that (X_1^*, Y_1^*),..., (X_n^*, Y_n^*) is a reordering of (X_1, Y_1),..., (X_n, Y_n) according to increasing values of $\|X_i - x\|$. For this estimate, Stone (1977) gives conditions on w_{ni} insuring that for all distributions of (X, Y) with $E\{|Y|^p\} < \infty$ $(p \geqslant 1)$, $E\{I_{np}\} \to 0$, where

$$I_{np} = \int |m_n(x) - m(x)|^p \, \mu(dx),$$

and μ is the probability measure of X. Devroye and Wagner (1980) have shown the same result for the kernel estimate

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

under some conditions on the positive number sequence h_n and the functions K: $\mathbb{R}^d \to \mathbb{R}$. For the convergence to 0 of ess $\sup_{(\mu)} |m_n(x) - m(x)|$, we refer to Watson (1964), Nadaraya (1964, 1965, 1970), Rosenblatt (1969), Schuster (1972), Greblicki (1974), Noda (1976), and Devroye (1978b) for the kernel estimate, and to Devroye (1978a) for the nearest neighbor estimate.

In this paper, (1) is shown to satisfy $E\{I_{np}\} \xrightarrow{n} 0$ wherever $E\{|Y|^p\} < \infty$ $(p \geqslant 1)$ and $I_{np} \xrightarrow{n} 0$ with probability one (*wp*1) when Y is almost surely (a.s.) bounded. Also, the necessity of condition (3) is investigated. Some brief comments on the discrimination problem are given and a discrimination rule is exhibited and shown to be strongly Bayes risk consistent.

2. Convergence in L_p

THEOREM 1. *If* $E\{|Y|^p\} < \infty$ $(p \geqslant 1)$ *and if* (2) *and* (3) *hold, then* $E\{I_{np}\} \to 0$.

Remarks. (i) No restriction is put on the joint distribution of (X, Y).

(ii) If $w_{ni} = v_i$, all n, i, then (3) reduces to

$$\sup_{i \leqslant n} v_i \Big/ \sum_{i=1}^{n} v_i \xrightarrow{n} 0.$$

This is equivalent to

$$v_n \Big/ \sum_{i=1}^{n} v_i \xrightarrow{n} 0, \qquad \sum_{n=1}^{\infty} v_n = \infty..$$

For the nontrivial implication, note that

$$\sup_{i \leqslant n} v_i \Big/ \sum_{i=1}^{n} v_i \leqslant \sup_{i \leqslant A} v_i \Big/ \sum_{i=1}^{n} v_i + \sup_{A < i} v_i \Big/ \sum_{j=1}^{i} v_j,$$

which can be made arbitrarily small by first picking A large enough and then letting n grow large.

(iii) The question of the selection of $\{l_n\}$ and $\{w_{ni}\}$ is not treated in this paper. Note that with $w_{ni} = v_i = i^\alpha$, $\alpha \geqslant -1$, (3) is satisfied. Thus, the sequence of weights can be increasing or slowly decreasing.

(iv) Consider the case that Y is $\{1,..., M\}$-valued and that Y must be estimated from X and the data (the discrimination problem), by, say, $g_n(X)$, where g_n is a Borel measurable function g_n: $\mathbb{R}^d \times (\mathbb{R}^d \times \{1,..., M\})^N \to \{1,..., M\}$. For each state (class) j, add up the weights w_{ni} that correspond to $Y_i^* = j$, and let

$$g_n(X) \neq j \quad \text{if} \quad \sum_{i=1}^{n} w_{ni} I_{[j]}(Y_i^*) < \max_{1 \leqslant k \leqslant M} \sum_{i=1}^{n} w_{ni} I_{[k]}(Y_i^*), \tag{4}$$

where I denotes the indicator function. The probability of error with such rules is

$$L_n = P\{g_n(X) \neq Y | X_1, Y_1,..., X_N, Y_N\}.$$

Let L^* be the Bayes probability of error:

$$L^* = \inf P\{g(X) \neq Y\},$$

where the infimum is taken over all Borel measurable functions g: $\mathbb{R}^d \to \{1,..., M\}$. It follows from Theorem 4 of Stone (1977) (see also Devroye and Wagner, 1980, expression (12)) that $E\{L_n\} \to L^*$ whenever $E\{I_{n1}\} \xrightarrow{n} 0$ under the condition that Y is bounded. Thus, from Theorem 1 we obtain

THEOREM 2. *All discrimination rules g_n satisfying* (4) *are Bayes risk consistent* ($E\{L_n\} \xrightarrow{n} L^*$) *if* (2) *and* (3) *hold. No restriction is put on the probability measure μ.*

(v) Theorems 1 and 2 remain valid if ties are handled differently. Assume that in the jth block, $X_{i_1},..., X_{i_K}$ are all equally close to x and closer to x than all the other X_i in the jth block. Replace Y_j^* in (1) by

$$\frac{1}{K} \sum_{k=1}^{K} Y_{i_k}.$$

To see this, check first that Lemma 1 remains valid by Propositions 11, 12 of Stone (1977). The remainder of the proof is not affected at all.

(vi) A slight improvement over (1) may result if instead of considering the nearest neighbor X_i^* in the ith block, one considers the s nearest neighbors $X_i^*(1),..., X_i^*(s)$ and replaces Y_i^* in (1) by

$$a_1 Y_i^*(1) + \cdots + a_s Y_i^*(s),$$

where $(a_1,..., a_s)$ is a fixed probability vector. Since $X_i^*(1)$ is closer to x than $X_i^*(2)$, etc., it is not unreasonable to expect the best performance when $a_1 \geqslant \cdots \geqslant a_s$. Of course, Theorems 1 and 2 remain valid for this generalization.

Proof of Theorem 1. The proof of Theorem 1 follows the lines of Stone's proof for nearest neighbor estimates (Stone, 1977). We indicate along the way where changes are needed.

LEMMA 1. *For any sequence of nonnegative constants* $w_{n1},..., w_{nn}$, *and for any* $l_1,..., l_n$,

$$E\left\{\left[\sum_{i=1}^n w_{ni}|Y_i^*|\Big/\sum_{i=1}^n w_{ni}\right]^p\right\} \leqslant E\left\{\sum_{i=1}^n w_{ni}|Y_i^*|^p\Big/\sum_{i=1}^n w_{ni}\right\} \leqslant \alpha(d)\, E\{|Y|^p\},$$

where $p \geqslant 1$ *and* $\alpha(d)$ *is a constant only depending upon* d *and* q.

Proof of Lemma 1. The first inequality is an application of Jensen's Theorem. The second inequality follows if we can show that for all i, $E\{|Y_i^*|^p\} \leqslant \alpha(d)\, E\{|Y|^p\}$. By the definition of Y_i^*, this is a corollary of Stone (1977, Propositions 11, 12). Q.E.D.

Next, notice that

$$\int |m_n(x) - m(x)|^p\, \mu(dx) \leqslant 4^{p-1} \sum_{j=1}^4 U_j,$$

where

$$U_1 = \int |g(x) - m(x)|^p\, \mu(dx),$$

$$U_2 = \int \left[\sum_{i=1}^n w_{ni}|g(X_i^*) - g(x)|^p\Big/\sum_{i=1}^n w_{ni}\right] \mu(dx),$$

$$U_3 = \int \left[\sum_{i=1}^n w_{ni}|g(X_i^*) - m(X_i^*)|^p\Big/\sum_{i=1}^n w_{ni}\right] \mu(dx),$$

$$U_4 = \int \left|\sum_{i=1}^n w_{ni}[Y_i^* - m(X_i^*)]\Big/\sum_{i=1}^n w_{ni}\right|^p \mu(dx),$$

(X_i^*, Y_i^*) is the ith block couple of observations singled out with respect to x, and g is an arbitrary Borel measurable function on $\mathbb{R}^d$. Since m is in $L_p(\mu)$, for every $\varepsilon > 0$ one can find a function g that is bounded, continuous, and

zero outside a compact set such that $U_1 < \varepsilon$ (Dunford and Schwartz, 1957, p. 298). By Lemma 1,

$$E\{U_3\} \leqslant \alpha(d)U_1 < \alpha(d)\varepsilon.$$

Thus, we need only show that $E\{U_2\} \not\to 0$ and $E\{U_4\} \not\to 0$ after selection of g.

Next, if $c = \sup_x |g(x)|$ and if δ is so small that $\|x-y\| < \delta$ implies that $|g(y)-g(x)| < \varepsilon$ for all x, y, then

$$E\{U_2\} \leqslant \sum_{i=1}^{n} w_{ni}(c^p P\{\|X_i^* - X\| \geqslant \delta\} + \varepsilon^p) \Big/ \sum_{i=1}^{n} w_{ni}$$
$$\leqslant 2\varepsilon^p \qquad \text{for large } n$$

provided that $P\{\|X_n^* - X\| \geqslant \delta\} \not\to 0$. (Use Toeplitz's lemma (Loeve, 1963, p. 238) and (3)).

Clearly,

$$P\{\|X_n^* - X\| \geqslant \delta\} = E\{(P\{\|X_1 - X\| \geqslant \delta | X\})^{l_n}\}$$
$$\leqslant E\{\exp(-l_n P\{\|X_1 - X\| < \delta | X\})\}$$
$$\leqslant P\{X \notin B_\beta\} + \exp(-\beta l_n),$$

where $B_\beta = \{x: P\{\|X_1 - X\| < \delta | X = x\} \geqslant \beta\}$. For almost all $x(\mu)$, we know that $P\{\|X_1 - X\| < \delta | X = x\} > 0$. Thus, by the Lebesgue Dominated Convergence Theorem, $P\{X \notin B_\beta\} \to 0$ as $\beta \to 0$. First, find β small enough and then l_n large enough so that the last expression is small.

Finally, assume that $|Y| \leqslant \gamma$ a.s. so that $|Y - m(X)| \leqslant 2\gamma$ a.s. Since for each x, $Y_1^* - m(X_1^*), \ldots, Y_n^* - m(X_n^*)$ are independent and since

$$Y_i^* - m(X_i^*) = \sum_{j=1}^{l_i} C_{ij}[Y_{ij} - m(X_{ij})],$$

where $\{(X_{ij}, Y_{ij}),\ 1 \leqslant j \leqslant l_i\}$ is the ith block of observations and

$$C_{ij} = \begin{cases} 1 \text{ if } X_{ij} \text{ is the nearest neighbor to } x \text{ in the } i\text{th block} \\ 0 \text{ otherwise,} \end{cases}$$

and since the $Y_{ij} - m(X_{ij})$ are independent zero mean random variables, we may write for $1 \leqslant p \leqslant 2$,

$$(E\{U_4\})^{2/p}$$
$$\leqslant \int E\left\{\left(\sum_{i=1}^{n} w_{ni}[Y_i^* - m(X_i^*)] \Big/ \sum_{i=1}^{n} w_{ni}\right)^2\right\} \mu(dx)$$

$$= \int E\left\{ \sum_{i=1}^{n} \sum_{j=1}^{l_i} w_{ni}^2 C_{ij}^2 E\{[Y_{ij} - m(X_{ij})]^2 | X_{ij}\} \Big/ \left(\sum_{i=1}^{n} w_{ni} \right)^2 \right\} \mu(dx)$$

$$\leqslant \left[\sup_{1 \leqslant i \leqslant n} w_{ni} \Big/ \sum_{j=1}^{n} w_{nj} \right]$$

$$\times \int E\left\{ \sum_{i=1}^{n} w_{ni} E\{[Y_i^* - m(X_i^*)]^2 | X_i^*\} \Big/ \sum_{i=1}^{n} w_{ni} \right\} \mu(dx)$$

$$\leqslant \alpha(d) \left[\sup_{1 \leqslant i \leqslant n} w_{ni} \Big/ \sum_{j=1}^{n} w_{ni} \right] E\{[Y - m(X)]^2\}$$

$$\leqslant \alpha(d) \left[\sup_{1 \leqslant i \leqslant n} w_{ni} \Big/ \sum_{j=1}^{n} w_{nj} \right] 4\gamma^2 \xrightarrow{n} 0.$$

For $p > 2$, use the fact that

$$E\{U_4\} \leqslant (2\gamma)^{p-2} \int E\left\{ \left(\sum_{i=1}^{n} w_{ni} [Y_i^* - m(X_i^*)] \Big/ \sum_{i=1}^{n} w_{ni} \right)^2 \right\} \mu(dx)$$

$$\xrightarrow{n} 0.$$

To complete the proof of Theorem 1, we only have to show that $E\{U_4\}$ can be made arbitrarily small even if Y is not a.s. bounded. Let $Y_i = Y_i' + Y_i''$, where

$$Y_i' = Y_i I_{[-\gamma,\gamma]}(Y_i) \qquad \text{and} \qquad Y_i'' = Y_i I_{[-\gamma,\gamma]^c}(Y_i).$$

Further, let $m'(x) = E\{Y'_1 | X_1 = x\}$, $m''(x) = E\{Y_1'' | X_1 = x\}$, and note that $m(x) = m'(x) + m''(x)$ for almost all $x(\mu)$. Now, it is straightforward that

$$E\{U_4\} \leqslant 2^{p-1} \int E\left\{ \left| \sum_{i=1}^{n} w_{ni} [Y_i'^* - m'(X_i^*)] \Big/ \sum_{i=1}^{n} w_{ni} \right|^p \right\} \mu(dx)$$

$$+ 2^{p-1} \int E\left\{ \left| \sum_{i=1}^{n} w_{ni} [Y_i''^* - m''(X_i^*)] \Big/ \sum_{i=1}^{n} w_{ni} \right|^p \right\} \mu(dx).$$

The first term tends to zero as $n \to \infty$ for all finite γ. The last term is not more than

$$2^{p-1} \alpha(d) E\{|Y_1'' - m''(X_1)|^p\}$$

$$\leqslant 2^{2p-2} \alpha(d) \left(E\{|Y_1|^p I_{[-\gamma,\gamma]^c}(Y_1)\} + E\{|m''(X_1)|^p\} \right)$$

$$\leqslant 2^{2p-1} \alpha(d) E\{|Y_1|^p I_{[-\gamma,\gamma]^c}(Y_1)\} \to 0 \qquad \text{as} \quad \gamma \to \infty,$$

by the finiteness of $E\{|Y_1|^p\}$. Q.E.D.

3. Necessary Conditions for Convergence in $\mathbb{L}_p$

We have shown that the sequence of weights $\{w_{nj}/\sum_{i=1}^n w_{ni},\ 1 \leqslant j \leqslant n, n = 1, 2,...\}$ is universally consistent in Stone's sense (Stone, 1977, p. 598) (that is, $E\{I_{np}\} \xrightarrow{n} 0$ whenever $p \geqslant 1$ and $E\{|Y|^p\} < \infty$). Conversely, if $E\{I_{np}\} \xrightarrow{n} 0$ for all distributions of (X, Y) with $E\{|Y|^p\} < \infty$ $(p \geqslant 1)$, then

$$\sup_{1 \leqslant i \leqslant n} w_{ni} \Big/ \sum_{j=1}^n w_{nj} \xrightarrow{n} 0,$$

$$E\left\{ \sum_{i=1}^n w_{ni} |f(X_i^*)| \Big/ \sum_{i=1}^n w_{ni} \right\} \leqslant \alpha E\{|f(X)|\}$$

for all Borel functions f on $\mathbb{R}^d$ (α is a constant not depending upon n or f or μ), and

$$\sum_{i=1}^{\infty} w_{ni} I_{(a,\infty)}(\|X_i^* - X\|) \Big/ \sum_{i=1}^n w_{ni} \xrightarrow{n} 0$$

in probability for all $a > 0$ (Stone, 1977, Corollary 1). Thus, if $l_n \xrightarrow{n} \infty$, *the sequence of weights* $\{w_{nj}/\sum_{i=1}^n w_{ni},\ 1 \leqslant j \leqslant n, n \geqslant 1\}$ *is universally consistent if and only if* (3) *holds.*

4. Strong Convergence

In this section we study the strong convergence of I_{np} under the condition

$$|Y| \leqslant c < \infty \qquad \text{wpl} \tag{5}$$

but with no other restriction on the joint distribution of (X, Y). We prove

Theorem 3. *Let* (5–6) *hold*:

$$\sum_{n=1}^{\infty} \exp\{-\alpha l_n\} < \infty, \qquad \textit{all} \quad \alpha > 0. \tag{6}$$

Assume that the positive weights w_{ni} *satisfy*

$$\sum_{n=1}^{\infty} \exp\left\{-\alpha \left(\sum_{j=1}^n w_{nj}\right)^2 \Big/ \left(\sum_{j=1}^n w_{nj}^2\right)\right\} < \infty, \qquad \textit{all} \quad \alpha > 0, \tag{7}$$

then $I_{np} \xrightarrow{n} 0$ wpl *for all* $p \geqslant 1$. *If* $w_{ni} = v_i > 0$ *for all* n, i, *then* $I_{np} \xrightarrow{n} 0$ wpl *for all* $p \geqslant 1$ *if one of the following conditions is satisfied*:

$$\text{(i)} \quad \sum_{n=1}^{\infty} \left(v_n \Big/ \sum_{i=1}^{n} v_i \right)^2 < \infty, \quad \sum_{n=1}^{\infty} v_n = \infty,$$

$$\text{(ii)} \quad \left(\sup_{i \leqslant n} v_i \Big/ \sum_{i=1}^{n} v_i \right) \log \log n \xrightarrow{n} 0. \tag{8}$$

In discrimination, the weak convergence of L_n to L^* only guarantees that for fixed large n, "most" data sequences $X_1, Y_1, \ldots, X_N, Y_N$ have a probability of error associated with them that is close to L^*. In practice, only one data sequence $X_1, Y_1, \ldots, X_N, Y_N, \ldots$ is available, and we would really want to know if for *this* sequence $L_n \to L^*$ as $n \to \infty$. We will see that for the rules given here, this is the case for almost all sequences (that is, $L_n \xrightarrow{n} L^*$ a.s.), regardless of the joint distribution of (X_1, Y_1).

From Theorem 3 and Stone's Theorem 4 (1977) we can deduce *Theorem* 4. All discrimination rules g_n satisfying (4) are *strongly Bayes risk consistent* ($L_n \xrightarrow{n} L^*$ wpl) if (6) holds and one of the conditions (7), (8(i)), (8(ii)) is satisfied. No restriction is put on the probability measure.

Remarks. (i) Condition (6) holds if $l_n / \log n \xrightarrow{n} \infty$.

(ii) For sequences $w_{ni} = v_i$ (all n, i), (7) is satisfied when

$$\left(\sum_{j=1}^{n} v_j^2 \Big/ \left(\sum_{j=1}^{n} v_j \right)^2 \right) \log n \xrightarrow{n} 0,$$

which in turn follows from

$$\left((v_n + 1) \Big/ \sum_{j=1}^{n} v_j \right) \log n \xrightarrow{n} 0,$$

if $\log n / \sum_{j=1}^{n} v_j$ is eventually monotone. When $v_n / \sum_{i=1}^{n} v_i$ is monotone, then (8(i)) implies $n^{1/2} v_n / (\sum_{i=1}^{n} v_i) \xrightarrow{n} 0$. This should be compared with (8(ii)).

(iii) All sequences $v_n = n^a$, $-1 \leqslant a$, satisfy (8(i)) and (8(ii)).

(iv) When $\{\log \log n / \sum_{j=1}^{n} v_j\}$ is eventually monotone, then (8(ii)) is equivalent to $((1 + v_n) / \sum_{j=1}^{n} v_j) \log \log n \xrightarrow{n} 0$. The proof follows the lines of Remark (ii) of Section 2. The condition (5) is too strong in general. For the limited scope of this paper and for applications such as discrimination, it is all that is needed. Theorem 4 is the first distribution-free strong Bayes risk consistency result that the authors are aware of in the literature.

Proof of Theorem 3. The notation of Theorem 1 is inherited, but instead of U_j, we will use $U_j(n)$ to make the dependency on n explicit. By choice of

g, $U_1(n) \equiv U_1(1)$ can be made arbitrarily small. Also, $E\{U_3(n)\} < \varepsilon$ by choice of g. Thus,

$$P\left\{\bigcup_{n \geqslant N} [U_3(n) > 2\varepsilon]\right\} \leqslant P\left\{\bigcup_{n \geqslant N} [U_3(n) - E\{U_3(n)\} > \varepsilon]\right\}.$$

However, note that $U_3(n) - E\{U_3(n)\}$ can be written as

$$\frac{\sum_{i=1}^{n} w_{ni} Z_i}{\sum_{i=1}^{n} w_{ni}}, \tag{9}$$

where

$$Z_i = \int |g(X_i^*) - m(X_i^*)|^p \, \mu(dx) - E\left\{\int |g(X_i^*) - m(X_i^*)|^p \, \mu(dx)\right\}.$$

Thus, Z_1, Z_2,..., Z_n are independent zero mean uniformly bounded random variables.

Therefore, by a Theorem of Chow (1966),

$$\sum_{i=1}^{n} w_{ni} Z_i \Big/ \sum_{i=1}^{n} w_{ni} \xrightarrow{n} 0 \qquad \text{wpl}$$

whenever (7) holds. If $w_{ni} = v_i$, all n and i, then

$$\sum_{i=1}^{n} v_i Z_i \Big/ \sum_{i=1}^{n} v_i \xrightarrow{n} 0 \qquad \text{wpl}$$

from the boundedness of Z_1, Z_2,...,

$$\sum_{n=1}^{\infty} \left(v_n \Big/ \sum_{i=1}^{n} v_i\right)^2 < \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} v_n = \infty$$

by Kolmogorov's strong law of large numbers (see Loeve, 1963, pp. 238). We now show that the same is true if just

$$\sup_{i \leqslant n} v_i \log \log n \Big/ \sum_{i=1}^{n} v_i \xrightarrow{n} 0.$$

From Loeve (1963, pp. 253) we know that (10) follows from $\sum_{n=1}^{\infty} v_n = \infty$, $|v_n Z_n| \leqslant \sum_{i=1}^{n} v_i$ for all n large enough (which is the case here since $|Z_n| \leqslant b \leqslant \infty$ and (8(ii)) is assumed), and

$$\sum_{k=0}^{\infty} P\left\{\left|\frac{1}{2^k} \sum_{i=2^k+1}^{2^{k+1}} v_i Z_i\right| \geqslant \frac{\varepsilon}{2^k} \sum_{i=1}^{2^{k+1}} v_i\right\} < \infty, \qquad \text{all} \quad \varepsilon > 0. \tag{11}$$

In view of $E\{Z_n\}=0$, $|Z_n|\leqslant b$, $E\{Z_n^2\}\leqslant b^2$ and an inequality of Bennett (1962, p. 39) the kth term of (11) is not more than

$$2\exp\left\{-2^k\left(\frac{\varepsilon}{2^k}\sum_{i=1}^{2^{k+1}}v_i\right)^2\Big/2\left(\frac{b^2}{2^k}\sum_{i=2^k+1}^{2^{k+1}}v_i^2+b\varepsilon\,\frac{1}{2^k}\sum_{i=1}^{2^{k+1}}v_i\sup_{i\leqslant 2^{k+1}}v_i\right)\right\}$$
$$\leqslant 2\exp\left\{-a\sum_{i=1}^{2^{k+1}}v_i\Big/\sup_{i\leqslant 2^{k+1}}v_i\right\}$$

for some $a>0$. These terms are summable with respect to k for all $a>0$ when

$$\left(\sum_{i=1}^{2^k}v_i\Big/\sup_{i\leqslant 2^k}v_i\right)\Big/\log k\xrightarrow{k}\infty,$$

or when

$$\left(\sup_{i\leqslant n}v_i\Big/\sum_{i=1}^{n}v_i\right)\log\log n\xrightarrow{n}0.$$

The random variables $U_2(n)$ and $U_4(n)$ are treated differently. Note that $U_2(n)=\int U_2(n,x)\,\mu(dx)$ and $U_4(n)=\int U_4(n,x)\,\mu(dx)$, where $U_2(n,x)$ and $U_4(n,x)$ are random variables and Borel measurable functions of x. Both are uniformly bounded by b, say. If for almost all x (μ), $U_2(n,x)\to 0$ wpl, it follows that $U_2(n)\xrightarrow{n}0$ wpl. That is, let $(\Omega,\mathcal{F},P)$ be the probability space of $(X_1,Y_1,X_2,Y_2,\ldots)$ with probability element $\omega\in\Omega$. By Fubini's theorem, $P\{\omega\colon U_2(n,x)\not\to 0\}=0$ for almost all $x(\mu)$ if and only if the set $\{(\omega,x)\colon U_2(n,x)\not\to 0\}$ has $P\times\mu$ measure zero, and this is true if and only if $\mu(x\colon U_2(n,x)\not\to 0\})=0$ for almost all $\omega(P)$, say $\omega\in\Omega'$. For every $\omega\in\Omega'$ $U_2(n)\xrightarrow{n}0$ by the Lebesgue Dominated Convergence Theorem, and since $P(\Omega')=1$, the claim follows. Thus we need only show that for all $x\in$ support (μ), $U_2(n,x)\xrightarrow{n}0$ wpl and $U_4(n,x)\to 0$ wpl.

First, $U_2(n,x)\xrightarrow{n}0$ wpl if $|g(X_n^*)-g(x)|\xrightarrow{n}0$ wpl in view of (3) and Toeplitz's Lemma. It is easy to check that (3) holds if either (7) or (8(i)) or (8(ii)) hold. Now,

$$\begin{aligned}&P\{|g(X_n^*)-g(x)|>\varepsilon\}\\ &\quad\leqslant P\{\|X_n^*-x\|\geqslant\delta\}\qquad\text{(for some }\delta>0)\\ &\quad\leqslant\exp(-\beta l_n),\end{aligned}$$

where $\beta=P\{\|X_n^*-x\|<\delta\}$ and $\beta>0$ since $x\in$ support (μ). Thus, $|g(X_n^*)-g(x)|\xrightarrow{n}0$ wpl by the Borel–Cantelli Lemma.

Next, $U_4(n,x)$ can be written in the form (9) with $Z_i=Y_i^*-m(X_i^*)$. Since $Y_1^*-m(X_1^*)\cdots$, $Y_n^*-m(X_n^*)$ are independent zero mean uniformly bounded

random variables, conditions (7), (8(i)) or (8(ii)) imply that $U_4(n, x) \xrightarrow{n} 0$ wpl for all $x \in$ support (μ).

This concludes the proof of Theorem 3.

ACKNOWLEDGMENTS

L. Devroye was supported by the Department of Defense Joint Services Electronics Program under Contract F49620-77-C-0101 and by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant AFOSR-76-3062. G. L. Wise was supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant AFOSR-76-3062.

REFERENCES

[1] CHOW, Y. S. (1966). Some convergence theorems for independent random variables. *Ann. Math. Statist.* **35** 1482–1493.

[2] BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57** 33–45.

[3] COVER, T. M. (1968). Estimation by the nearest neighbor rule. *IEEE Trans. Information Theory* **IT-14** 50–55.

[4] DEVROYE, L. P. (1978a). The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Trans. Information Theory* **IT-24** 142–151.

[5] DEVROYE, L. P. (1978b). The uniform convergence of the Nadaraya–Watson regression function estimate. *Canad. J. Statist.* **6** 179–191.

[6] DEVROYE, L. P., AND WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation, *Ann. Statist.* **8** 231–239.

[7] DUNFORD, N., AND SCHWARTZ, J. T. (1957). *Linear Operators, Part I: General Theory.* Interscience, New York.

[8] GREBLICKI, W. (1974). Asymptotically optimal probabilistic algorithms for pattern recognition and identification. In *Monografie No.* 3 *Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wroclawskiej No. 18.* Wroclaw, Poland.

[9] LOEVE, M. (1963). *Probability Theory.* Van Nostrand, Princeton, N.J.

[10] NADARAYA, E. A. (1964). On estimating regression. *Theor. Probability Appl.* **9** 141–142.

[11] NADARAYA, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theor. Probability Appl.* **10** 186–190.

[12] NADARAYA, E. A. (1970). Remarks on some nonparametric estimates for density functions and regression curves. *Theor. Probability Appl.* **15** 134–137.

[13] NODA, K. (1976). Estimation of a regression function by the Parzen kernel-type density estimators. *Ann. Inst. Statist. Math.* **28** 221–234.

[14] ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis-II* (P. R. Krishnaiah, Ed.), pp. 25–31. Academic Press, New York.

[15] ROYALL, R. M. (1966). *A class of nonparametric estimators of a smooth regression function.* Ph.D. dissertation, Department of Statistics, Stanford University.

[16] SCHUSTER, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Ann. Math. Statist.* **43** 84–88.

[17] STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

[18] WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.