The Annals of Statistics
1981, Vol. 9, No. 6, 1310–1319

ON THE ALMOST EVERYWHERE CONVERGENCE OF NONPARAMETRIC REGRESSION FUNCTION ESTIMATES

BY LUC DEVROYE[1]

McGill University

Let (X, Y), (X_1, Y_1), $\cdots$, (X_n, Y_n) be independent identically distributed random vectors from $R^d \times R$, and let $E(|Y|^p) < \infty$ for some $p \geq 1$. We wish to estimate the regression function $m(x) = E(Y|X = x)$ by $m_n(x)$, a function of x and (X_1, Y_1), $\cdots$, (X_n, Y_n). For large classes of kernel estimates and nearest neighbor estimates, sufficient conditions are given for $E\{|m_n(x) - m(x)|^p\} \to 0$ as $n \to \infty$, almost all x. No additional conditions are imposed on the distribution of (X, Y). As a by-product, just assuming the boundedness of Y, the almost sure convergence to 0 of $E\{|m_n(X) - m(X)| \mid X_1, Y_1, \cdots, X_n, Y_n\}$ is established for the same estimates. Finally, the weak and strong Bayes risk consistency of the corresponding nonparametric discrimination rules is proved for all possible distributions of the data.

1. Introduction. Let (X, Y), (X_1, Y_1), $\cdots$, (X_n, Y_n) be independent identically distributed $R^d \times R$-valued random vectors with $E(|Y|) < \infty$. The *regression function* $m(x) = E(Y|X = x)$ for $x \in R^d$ is estimated by

$$m_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i \tag{1.1}$$

where $(W_{n1}(x), \cdots, W_{nn}(x))$ is a probability vector of weights and each $W_{ni}(x)$ is a Borel measurable function of $x, X_1, X_2, \cdots, X_n$. The *nearest neighbor estimate* is defined as follows. Rank the (X_i, Y_i), $i = 1, \cdots, n$, according to increasing values of $\|X_i - x\|$ (ties are broken by comparing indices) and obtain a vector of indices $(R_1, \cdots, R_n)$ where X_{R_i} is the ith nearest neighbor of x for all i. If $(v_{n1}, \cdots, v_{nn})$ is a given probability vector of weights, then set

$$W_{nR_i}(x) = v_{ni}; \tag{1.2}$$

see Cover (1968) for a particular choice of v_{ni}'s, and Stone (1977) for more general weight vectors. The *kernel estimate* can be obtained by putting

$$W_{ni}(x) = K((X_i - x)/h)/\textstyle\sum_{j=1}^n K((X_j - x)/h), \tag{1.3}$$

where $h = h_n$ is a positive number depending upon n only, and K is a given nonnegative function on R^d; we will treat 0/0 in (1.3) as 0. See Watson (1964), Nadaraya (1964, 1965) for the original definition, and Collomb (1976, 1977, 1981), Schuster and Yakowitz (1979), Revesz (1979), Devroye and Wagner (1978b, 1980a, 1980b), Györfi (1981) and Spiegelman and Sacks (1980) for recent developments.

Stone (1977) showed the following interesting nontrivial result. If the weight vector $\underline{v}_n = (v_{n1}, \cdots, v_{nn})$ satisfies

(1.4)
- (i) $v_{n1} \geq \cdots \geq v_{nn}$ (all n),
- (ii) $v_{n1} \to 0$ as $n \to \infty$,
- (iii) there exists a sequence of numbers $k = k_n$ such that $k/n \to 0$ and $\sum_{i=k+1}^n v_{ni} \to 0$ as $n \to \infty$,

Received August 13, 1979; revised February, 1981.

[1] This research was sponsored by National Research Council of Canada Grant No. A3456 and Air Force Grant No. AFOSR 77-3385.

AMS 1980 *subject classifications.* Primary 62G05.

Key words and phrases. Regression function, nonparametric discrimination, nearest neighbor rule, kernel estimate, universal consistency.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

then the nearest neighbor estimate is *universally consistent*, that is,

$$E(|m_n(X) - m(X)|^p) \to 0 \quad \text{as } n \to \infty \text{ whenever } E(|Y|^p) < \infty, \qquad \text{all} \quad p \geq 1. \tag{1.5}$$

Devroye and Wagner (1980b) and independently, Spiegelman and Sacks (1980), showed that the kernel estimate is also universally consistent provided that K and h satisfy:

$$\begin{aligned} &\text{(i) } h \to 0 \quad \text{and} \quad nh^d \to \infty \quad \text{as } n \to \infty, \\ &\text{(ii) there exist } r_1, r_2, c_1, c_2, \text{ all positive numbers, such that} \\ &\quad c_1 I_{(\|u\| \leq r_1)} \leq K(u) \leq c_2 I_{(\|u\| \leq r_2)} \text{ where } I \text{ is the indicator function.} \end{aligned} \tag{1.6}$$

Györfi (1981) presents universal consistency results for other estimates related to (1.1) and (1.3).

In this paper we find sufficient conditions on the W_{ni}'s that guarantee

$$\begin{aligned} &E(|m_n(x) - m(x)|^p) \to 0 \quad \text{as } n \to \infty \qquad \text{for almost all } x(\mu) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{whenever} \quad E(|Y|^p) < \infty, \qquad \text{all } p \geq 1. \end{aligned} \tag{1.7}$$

In (1.7) μ is the probability measure for X. Notice that from (1.5) we can merely conclude that $\liminf E(|m_n(x) - m(x)|^p) \to 0$ as $n \to \infty$ for almost all $x(\mu)$ by Fatou's Lemma.

In what follows, we will use the symbol S_r for the closed ball of radius r centered at x. The crucial result from real analysis that is needed here is the following (see for instance Wheeden and Zygmund, 1977, page 189):

LEMMA 1.1. *If $f \in L^1(\mu)$, that is $\int |f(x)| \mu(dx) < \infty$, then*

$$\int_{S_r} f(y)\mu(dy) \Big/ \int_{S_r} \mu(dy) \to f(x) \qquad \text{as} \quad r \to 0$$

for almost all $x(\mu)$.

REMARK 1.1. Wheeden and Zygmund prove this result for balls defined by the L_∞ norm on R^d. Their result remains valid however for the L_2 norm. To see this, it suffices to check that Besicovitch's covering Lemma (*ibid*, pages 185–186) remains valid for the L_2 norm.

The main results are stated in Section 2. From the pointwise consistency (1.7) and the dominated convergence theorem one can prove (Section 3) globally consistent behavior that comes close to (1.5). The strong pointwise consistency of both estimates is treated in Section 4 for the special case of bounded Y. In Section 5 we present analogous consistency results for the nonparametric discrimination problem.

2. Pointwise consistency.

THEOREM 2.1. *The nearest neighbor estimate satisfies* (1.7) *when there exists a sequence of integers $k = k_n$ such that*

$$\begin{aligned} &\text{(i) } k/n \to 0 \quad \textit{and} \quad k \to \infty \quad \textit{as } n \to \infty, \\ &\text{(ii) } \sup_n k \max_i v_{ni} < \infty, \\ &\text{(iii) } v_{ni} = 0 \quad \textit{when} \quad i > k. \end{aligned} \tag{2.1}$$

The kernel estimate satisfies (1.7) *when*

$$\begin{aligned} &\text{(i) } h \to 0 \quad \textit{and} \quad nh^d \to \infty \quad \textit{as } n \to \infty, \\ &\text{(ii) } \textit{there exist positive numbers } r, c_1, c_2 \textit{ such that} \\ &\quad c_1 I_{(\|u\| \leq r)} \leq K(u) \leq c_2 I_{(\|u\| \leq r)}. \end{aligned} \tag{2.2}$$

REMARK 2.1. Throughout the paper, all norms are the same: they are either all L_∞ or all L_2.

REMARK 2.2. The k-nearest neighbor estimate (defined by $v_{ni} = 1/k$, $1 \le i \le k$, and $v_{ni} = 0$, $i > k$) satisfies (2.1) when $k/n \to 0$ and $k \to \infty$ as $n \to \infty$.

The elementary result needed to prove Theorem 2.1 is:

LEMMA 2.1. *If $f \in L^p(\mu)$ for fixed $p \ge 1$ and* (2.1), (2.2) *hold, then*

$$E(\textstyle\sum_{i=1}^n W_{ni}(x) \,|\, f(X_i) - f(x) \,|^p) \to 0 \quad \text{as} \quad n \to \infty$$

for almost all $x(\mu)$ for both the nearest neighbor estimate and the kernel estimate.

PROOF OF LEMMA 2.1. Assume that $f \ge 0$. Since for $a, b \ge 0$, $p \ge 1$, $|a - b|^p \le |a^p - b^p|$, we see that for almost all $x(\mu)$,

$$\int_{S_r} |f(y) - f(x)|^p \mu(dy) \Big/ \int_{S_r} \mu(dy) \to 0 \qquad \text{as} \quad r \to 0; \tag{2.3}$$

see for example, Wheeden and Zygmund (1977, page 191, example 20). For general f, split f into its positive and negative parts, $f^+ + f^-$, note that $|f^+ + f^-|^p \le 2^{p-1}(f^{+p} + |f^-|^p)$, and apply (2.3) twice. Thus, (2.3) is valid for all $f \in L^p(\mu)$. Let A be the set of all x's for which (2.3) is true. Define further the *maximal function* corresponding to $|f|^p$ by

$$f^*(x) = \sup_{r>0} \int_{S_r} |f(y)|^p \mu(dy) \Big/ \int_{S_r} \mu(dy). \tag{2.4}$$

Fix $x \in A$, and for arbitrary $\epsilon > 0$ find $\delta > 0$ such that the expression in (2.3) is smaller than ϵ, all $r \le \delta$. Let C be the closed ball centered at x with radius $\| X_{R_{k+1}} - x \|$, and let B be the corresponding open ball. For the nearest neighbor estimate, there exist positive constants c_i such that

$$\begin{aligned}
&E\{\textstyle\sum_{i=1}^n W_{ni}(x) \,|\, f(X_i) - f(x) \,|^p\} \\
&\quad \le c_1 E\{k^{-1} \textstyle\sum_{i=1}^k |\, f(X_{R_i}) - f(x) \,|^p\} \\
&\quad \le c_1 E\left\{\max\left(\mu^{-1}(B) \int_B |\, f(y) - f(x) \,|^p \mu(dy), \mu^{-1}(C) \int_C |\, f(y) - f(x) \,|^p \mu(dy)\right)\right\} \\
&\quad \le c_1 E\left\{\sup_{0<r\le\|X_{R_{k+1}}-x\|} \mu^{-1}(S_r) \int_{S_r} |\, f(y) - f(x) \,|^p \mu(dy)\right\} \\
&\quad \le c_1 \epsilon + c_1 2^{p-1}\{f^*(x) + |\, f(x) \,|^p\} P(\| X_{R_{k+1}} - x \| \ge \delta).
\end{aligned} \tag{2.5}$$

If $x \in S = \text{support } (\mu)$, then $k/n \to 0$ implies that $P(\| X_{R_{k+1}} - x \| \ge \delta) \le c_2 \exp(-c_3 n)$ (Devroye, 1978a). Thus, the first part of Lemma 2.1 follows since $\mu(S) = 1$ (see Cover and Hart, 1967), $\mu(A) = 1$ (which we established) and $\mu(\{x: f^*(x) = \infty\}) = 0$. The last fact follows from the basic inequality for maximal functions (Wheeden and Zygmund, 1977, page 188): namely, there exists a constant $a(d) > 0$ only depending upon d such that for all $b > 0$,

$$\mu(\{x: f^*(x) > b\}) \le \{a(d)/b\} \int |\, f(y) \,|^p \mu(dy). \tag{2.6}$$

Consider now the kernel estimate, and let r, c_1, c_2 be the constants defined in (2.2). We will prove the following inequality:

$$E\{\textstyle\sum_{i=1}^n W_{ni}(x) \,|\, f(X_i) - f(x) \,|^p\} \le 7(c_2/c_1) \int_{S_{rh}} |\, f(y) - f(x) \,|^p \mu(dy) \Big/ \int_{S_{rh}} \mu(dy). \tag{2.7}$$

Lemma 2.1 then follows from (2.7) and (2.3). For $n \le 7$, (2.7) is trivially true. We fix $n > 7$, and define $U = K((X_n - x)/h)$, $u = E(U)$, $V = \sum_{i=1}^{n-1} K((X_i - x)/h)$, $Z_{n-1} = \min(1, c_2/V)$. Since $W_{nn}(x) = U/(U + V) \le Z_{n-1}$, we can estimate the left hand side of (2.7) from above by

$$nE\{|f(X_n) - f(x)|^p I_{(\|X_n - x\| \le rh)}\} E(Z_{n-1}). \tag{2.8}$$

Now, $E(Z_{n-1}) \le P(V < c) + c_2/c$ for arbitrary $c > 0$. Take $c = (n - 1)u/2$, and use Chebyshev's inequality:

$$\begin{aligned} P(V < c) &= P\{V - E(V) < -E(V)/2\} \le 4\,\mathrm{Var}(V)/E^2(V) \\ &\le 4E(U^2)/\{(n-1)u^2\} \le 4c_2/\{(n-1)u\}. \end{aligned}$$

Hence, (2.8) is not greater than

$$6\{n/(n-1)\}(c_2/u) \int_{S_{rh}} |f(y) - f(x)|^p \mu\,(dy)$$

from which (2.7) follows easily when $n > 7$.

REMARK 2.3. For the kernel estimate with $c_1 = c_2$ in (2.2), a short proof of Lemma 2.1 is possible by applying Lemma 1.1 and Lemma 1 of Spiegelman and Sacks (1980).

LEMMA 2.2. *Let $h = h_n$ be a sequence of positive numbers with $nh^d \to \infty$ as $n \to \infty$. For all $c > 0$, we have*

$$n\mu(S_{ch}) \to \infty \quad as \quad n \to \infty, \qquad almost\ all\ x(\mu).$$

PROOF OF LEMMA 2.2. We may assume that $\lim_n h = 0$. Decompose the Lebesgue measure on $R^d(\lambda)$ into its μ-absolutely continuous part (λ_1) and its μ-singular part (λ_2). By a well-known theorem on the relative differentiation of measures (see for instance Section (10.50) of Wheeden and Zygmund, 1977),

$$\lambda(S_{ch})/\mu(S_{ch}) \to \frac{d\lambda_1}{d\mu}(x), \qquad \text{almost all} \quad x(\mu),$$

where $d\lambda_1/d\mu$ is the Radon-Nikodym derivative of λ_1 with respect to μ. Thus, there exists a nonnegative function g with $g(x) < \infty$, almost all $x(\mu)$, such that

$$h^d/\mu(S_{ch}) \to g(x), \qquad \text{almost all} \quad x(\mu).$$

This concludes the proof of Lemma 2.2.

PROOF OF THEOREM 2.1. By Minkowski's inequality, for any $p \ge 1$,

$$\begin{aligned} E\{|\textstyle\sum_{i=1}^n W_{ni}(x)Y_i - m(x)|^p\}^{1/p} &\le E\{|\textstyle\sum_{i=1}^n W_{ni}(x)(Y_i - m(X_i))|^p\}^{1/p} \\ &\quad + E\{|\textstyle\sum_{i=1}^n W_{ni}(x)|m(X_i) - m(x)|^p\}^{1/p}. \end{aligned} \tag{2.9}$$

With the kernel estimate, the possibility exists that $W_{ni}(x) = 0$ for all i; thus, in that case, a third term should be added on the right hand side of (2.9), namely $|m(x)|\{P(\sum W_{ni}(x) = 0)\}^{1/p}$. Clearly, this term cannot cause any trouble because m is finite for almost all $x(\mu)$, and because by Lemma 2.2,

$$P\{\textstyle\sum W_{ni}(x) = 0\} = \{1 - \mu(S_{rh})\}^n \le \exp\{-n\mu(S_{rh})\} \to 0$$

for almost all $x(\mu)$.

The last term in (2.9) tends to 0 for almost all $x(\mu)$ and for both estimates considered here by Lemma 2.1. We will show that the first term on the right hand side of (2.9) tends to 0 for almost all $x(\mu)$ when $p \ge 2$. The case $1 \le p < 2$ is then obtained through a standard truncation argument.

Let $h(x) = E\{|Y - m(X)|^p | X = x\}$. By successive applications of inequalities of Marcinkiewicz and Zygmund (1937) (see also Petrov, 1975, pages 59–60) and Jensen, we have for some constant $a(p) > 0$ depending only upon p,

$$\begin{aligned} &E\{|\textstyle\sum_{i=1}^n W_{ni}(x)(Y_i - m(X_i))|^p\} \\ &\quad \le aE\{|\textstyle\sum_{i=1}^n W_{ni}^2(x)(Y_i - m(X_i))^2|^{p/2}\} \\ &\quad \le aE[\{\sup_i W_{ni}(x)\}^{p/2} |\textstyle\sum_{i=1}^n W_{ni}(x)(Y_i - m(X_i))^2|^{p/2}] \\ &\quad \le aE[\{\sup_i W_{ni}(x)\}^{p/2} \textstyle\sum_{i=1}^n W_{ni}(x) |Y_i - m(X_i)|^p] \\ &\quad = aE[\{\sup_i W_{ni}(x)\}^{p/2} \textstyle\sum_{i=1}^n W_{ni}(x)h(X_i)]. \end{aligned} \tag{2.10}$$

For the nearest neighbor estimate, $\sup_i W_{ni}(x) = \sup_i v_{ni} \to 0$ as $n \to \infty$. Since $h \in L^1(\mu)$, $E\{\sum_{i=1}^n W_{ni}(x)h(X_i)\}$ remains bounded for almost all $x(\mu)$ by Lemma 2.1. Thus, (2.10) tends to 0 for almost all $x(\mu)$.

For the kernel estimate, define U, u, V and Z_n as in the proof of Lemma 2.1, and estimate (2.10) from above by

$$\begin{aligned} &anE[\{\sup_i W_{ni}(x)\}^{p/2} W_{nn}(x)h(X_n)] \le anP\{V < (n-1)u/2\} \\ &\qquad \cdot E\{I_{(\|X_n - x\| \le rh)} h(X_n)\} + an\{2c_2/(n-1)u\}E\{W_{nn}(x)h(X_n)\}. \end{aligned} \tag{2.11}$$

By (2.7) and $u \ge c_1\mu(S_{rh})$ we know that the last term of (2.11) does not exceed

$$14(c_2/c_1)^2\{a/(n-1)\} \int_{S_{rh}} h(y)\mu(dy) \Big/ \left(\int_{S_{rh}} \mu(dy)\right)^2 \tag{2.12}$$

which is $o(1)$ for almost all $x(\mu)$ by Lemma 2.2. Below, we show that $P\{V < (n-1)u/2\} \le \exp\{-c_4 n\mu(S_{rh})\}$ for some $c_4 > 0$. Thus, the second term of (2.11) is not greater than

$$an \int_{S_{rh}} h(y)\mu(dy) \exp\{-c_4 n\mu(S_{rh})\} \tag{2.13}$$

which tends to 0 for almost all $x(\mu)$ in view of Lemmas 1.1 and 2.2. Thus, Theorem 2.1 is proved for $p \ge 2$.

The exponential inequality needed to obtain (2.13) follows from Bernstein's inequality for sums of bounded random variables (see Bennett, 1962 or Hoeffding, 1963):

$$\begin{aligned} P\{V < (n-1)u/2\} &= P\{V - E(V) < -(n-1)u/2\} \\ &\le \exp\{-(n-1)(u/2)^2/(2\operatorname{Var}(U) + c_2 u/2)\} \\ &\le \exp\{-(n-1)u/10c_2\} \\ &\le \exp\{-c_4 n\mu(S_{rh})\} \end{aligned} \tag{2.14}$$

where $c_4 = c_1/20c_2$, $n \ge 2$.

For $p < 2$, define for integer $t > 0$, $Y_i' = Y_i I_{(|Y_i| \le t)}$, $Y_i'' = Y_i - Y_i'$, $m'(x) = E(Y_1' | X_1 = x)$, $m''(x) = E(Y_1'' | X_1 = x)$. Thus,

$$\begin{aligned} &E\{|\textstyle\sum_{i=1}^n W_{ni}(x)(Y_i - m(X_i))|^p\} \\ &\quad \le 2^{p-1}[E\{|\textstyle\sum_{i=1}^n W_{ni}(x)(Y_i' - m'(X_i))|^p\} + E\{\textstyle\sum_{i=1}^n W_{ni}(x) |Y_i'' - m''(X_i)|^p\}]. \end{aligned} \tag{2.15}$$

The last term of (2.15) is not greater than

$$2^p E\{\textstyle\sum_{i=1}^n W_{ni}(x) |Y_i''|^p\} = 2^p E\{\textstyle\sum_{i=1}^n W_{ni}(x) g_t(X_i)\}$$

where $g_t(x) = E(|Y_1''|^p | X_1 = x)$. Let A_t be the set of all x for which the first term of (2.15) tends to 0 and $E\{\sum_{i=1}^n W_{ni}(x)g_t(X_i)\}$ tends to $g_t(x)$ as $n \to \infty$. We have already shown that

for each fixed t, $\mu(A_t) = 1$. Let B be the set of all x with $g_t(x) \to 0$ as $t \to \infty$. Clearly, $\mu(B) = 1$ because $E\{g_t(X)\} \to 0$ as $t \to \infty$ and g_t is monotone in t. For all x in $B \cap (\cap_t A_t)$, we claim that (2.15) tends to 0: first pick t large enough so that $g_t(x)$ is small, and then let n grow large. Since this set has μ-measure 1, the theorem is proved.

3. Global consistency.

THEOREM 3.1. *Let $E(|Y|^p \log^+|Y|) < \infty$ for some $p \geq 1$. If the estimate m_n satisfies the conditions of Theorem 2.1, we have*

$$E\{|m_n(X) - m(X)|^p\} \to 0 \quad as \quad n \to \infty. \tag{3.1}$$

REMARK 3.1. The condition put on Y in Theorem 3.1 is stricter than the condition $E(|Y|^p) < \infty$ needed for (1.5) in the papers of Stone (1977) and Devroye and Wagner (1980b). The conditions on the sequences of weights are not strictly nested for the nearest neighbor estimate: the monotonicity condition (1.4)(i) is absent in (2.1); but (2.1)(iii) is stricter than (1.4)(iii).

REMARK 3.2. $E(|Y|^p \log^+|Y|) < \infty$ implies $E\{|m(X)|^p \log^+|m(X)|\} < \infty$ and $E\{|Y - m(X)|^p \log^+|Y - m(X)|\} < \infty$. We say that $f \in L \log^+ L(\mu)$ when $\int |f(y)| \log^+|f(y)| \mu(dy) < \infty$.

Theorem 3.1 is an immediate consequence of the following property of maximal functions:

LEMMA 3.1. *If $f \in L \log^+ L(\mu)$, then $f^* \in L^1(\mu)$ where*

$$f^*(x) = \sup_{r>0} \int_{S_r} |f(y)| \mu(dy) \Big/ \int_{S_r} \mu(dy). \tag{3.2}$$

PROOF OF LEMMA 3.1. The proof is a slight variation of an argument of Wheeden and Zygmund (1977, pages 155–156). For $t > 0$, define $g(x) = |f(x)| I_{(|f(x)| \geq t/2)}$ and let g^* be the maximal function corresponding to g. Clearly, $|f(x)| \leq g(x) + t/2$ and $f^*(x) \leq g^*(x) + t/2$. Thus, $\{f^*(x) > t\}$ implies $\{g^*(x) > t/2\}$. So,

$$\begin{aligned}\int_{f^*(x)>t} \mu(dx) &\leq \int_{g^*(x)>t/2} \mu(dx) \leq (2a/t) \int |g(x)| \mu(dx) \\ &= (2a/t) \int_{|f(x)| \geq t/2} |f(x)| \mu(dx)\end{aligned}$$

for some $a > 0$ only depending upon d (see (2.6)). Let $t_0 = 2a \int |f(x)| \mu(dx)$. Then

$$\begin{aligned}\int f^*(x)\mu(dx) &= \int_0^\infty \mu(x: f^*(x) > t)\, dt \\ &\leq \int_{t_0}^\infty (2a/t) \int_{|f(x)| \geq t/2} |f(x)| \mu(dx)\, dt + t_0 \\ &\leq 2a \int_{2|f(x)| \geq t_0} |f(x)| \int_{t_0}^{2|f(x)|} t^{-1}\, dt \mu(dx) + t_0 \\ &\leq 2a \int |f(x)| \log^+(2|f(x)|/t_0) \mu(dx) + t_0,\end{aligned}$$

thus concluding the proof of Lemma 3.1.

PROOF OF THEOREM 3.1. The proof merely consists of exhibiting a function ϕ: $R^d \to R$ with the properties $\phi \in L^1(\mu)$ and $E\{|m_n(x) - m(x)|^p\} \le \phi(x)$. Theorem 2.1 and the dominated convergence theorem are then sufficient for (3.1).

Since $m \in L^p(\mu)$, we need only show that $E\{|m_n(x)|^p\} \le \phi(x) \in L^1(\mu)$. Let $f(x) = E(|Y|^p|X = x)$, and let f^* be the maximal function corresponding to f. We show that $E\{|m_n(x)|^p\} \le cf^*(x)$ for some constant c, and apply Lemma 3.1.

For both estimates considered here,

$$E\{|m_n(x)|^p\} \le E\{\textstyle\sum_{i=1}^n W_{ni}(x)|Y_i|^p\} = E\{\sum_{i=1}^n W_{ni}(x)f(X_i)\}. \tag{3.3}$$

Expression (3.3) is smaller than $7(c_2/c_1)f^*(x)$ for the kernel estimate (see (2.7)) and is bounded from above by $c_3 f^*(x)$ for the nearest neighbor estimate where $c_3 = \sup_n(k \max_i v_{ni})$ (see (2.5)).

4. Strong consistency. In this section we will assume that

$$|Y| \le \gamma < \infty. \tag{4.1}$$

THEOREM 4.1. *Assume that* (4.1) *holds and that $k = k_n$ is a sequence of integers such that*

$$\begin{aligned} &\text{(i) } k/n \to 0 \quad \textit{and} \quad k \to \infty \quad \textit{as} \quad n \to \infty, \\ &\text{(ii) } \sup_n k \max_i v_{ni} < \infty, \\ &\text{(iii) } \textstyle\sum_{i>k} v_{ni} \to 0 \quad \textit{as} \quad n \to \infty. \end{aligned} \tag{4.2}$$

Then, for the nearest neighbor estimate, $E\{|m_n(x) - m(x)|\} \to 0$ as $n \to \infty$ for almost all $x(\mu)$ and $E\{|m_n(X) - m(X)|\} \to 0$ as $n \to \infty$. If in addition $k/\log n \to \infty$ as $n \to \infty$, then

$$|m_n(x) - m(x)| \to 0 \quad \text{a.s. } \textit{as } n \to \infty, \qquad \textit{almost all } x(\mu), \tag{4.3}$$

and

$$E\{|m_n(X) - m(X)| \,|\, X_1, Y_1, \cdots, X_n, Y_n\} \to 0 \quad \text{a.s. } \textit{as } n \to \infty. \tag{4.4}$$

NOTE. Györfi (1981 and private communication) showed Theorem 4.1 independently of myself for the case $v_{ni} = 1/k_n$, $i \le k$, $v_{ni} = 0$, $i > k$.

PROOF OF THEOREM 4.1. When $f \in L^\infty(\mu)$, the replacement of (2.1)(iii) by (4.2)(iii) does not upset the conclusion of Lemma 2.1. In the proof of Theorem 2.1, take $p = 2$, and estimate (2.10) from above by $c \sup_i v_{ni}$ for some constant $c < \infty$. The weak convergence part of Theorem 4.1 now follows without work from (4.2) and the dominated convergence theorem.

Assertion (4.4) follows from (4.3) by a standard application of Fubini's theorem and the dominated convergence theorem (see, e.g., Glick, 1974). To prove (4.3), we have

$$|m_n(x) - m(x)| \le |\textstyle\sum_{i=1}^n W_{ni}(x)(Y_i - m(X_i))| + \sum_{i=1}^n W_{ni}(x)|m(X_i) - m(x)|, \tag{4.5}$$

and

$$P\{|\textstyle\sum_{i=1}^n W_{ni}(x)(Y_i - m(X_i))| > \epsilon \,|\, X_1, \cdots, X_n\} \le c_1 \exp\{-c_2/\sup_i W_{ni}(x)\} \quad \text{a.s.} \tag{4.6}$$

for some c_1, $c_2 > 0$ depending upon ϵ and γ only (see, e.g., Devroye, 1978a, Lemma 1). Thus, for all x, the first term on the right side of (4.5) tends to 0 a.s. as $n \to \infty$ when $(\log n)\sup_i v_{ni} \to 0$; this follows from (4.6) and the Borel-Cantelli lemma. Since $k \sup_i v_{ni} \le c_3 < \infty$ and

$$\textstyle\sum_{i=1}^n W_{ni}(x)|m(X_i) - m(x)| \le c_3 U(x) + o(1) \tag{4.7}$$

where $U(x) = k^{-1} \sum_{i=1}^{k} | m(X_{R_i}) - m(x) |$, and since $E\{U(x)\} \to 0$ as $n \to \infty$ for almost all $x(\mu)$ by Lemma 2.1, we must only check whether $U(x) - E\{U(x)\} \to 0$ a.s. as $n \to \infty$. Again by Bernstein's inequality, we have a.s.,

$$P\{| U(x) - EU(x) | > \epsilon | X_{R_{k+1}}\} \le c_4 \exp(-c_5 k) \tag{4.8}$$

for some c_4, $c_5 > 0$ depending upon ϵ and γ only. When $k/\log n \to \infty$, the right hand side of (4.8) is summable with respect to n, and $U(x) - E\{U(x)\} \to 0$ a.s. because ϵ is arbitrary.

THEOREM 4.2. *Assume that* (4.1) *and* (2.2) *are satisfied and that* $nh^d/\log n \to \infty$ *as* $n \to \infty$. *Then conclusions* (4.3) *and* (4.4) *hold for the kernel estimate.*

LEMMA 4.1. *If N is a binomial random variable with parameters n and p, then*

$$\sum_{n=1}^{\infty} E\{\exp(-sN)\} < \infty, \qquad \textit{all } s > 0,$$

whenever $np/\log n \to \infty$.

PROOF OF LEMMA 4.1. We show that $E\{\exp(-sN)\} \le 2 \exp(-s'np)$ where $s' = \min(s/2, 1/10)$. Clearly,

$$\begin{aligned} E\{\exp(-sN)\} &\le \exp(-snp/2) + P(N/n - p < -p/2) \\ &\le \exp(-s'np) + \exp\{-n(p/2)^2/(2p + p/2)\} \le 2 \exp(-s'np) \end{aligned}$$

by Bernstein's inequality (see (2.14)).

PROOF OF THEOREM 4.2. We use (4.5) and estimate the left hand side of (4.7) from above by

$$U(x) = (c_2/c_1) \sum_{i=1}^{n} | m(X_i) - m(x) | I_{A_i} / \sum_{i=1}^{n} I_{A_i}$$

where A_i is the event $(\| X_i - x \| \le rh)$. By Theorem 2.1, $E(U(x)) \to 0$ as $n \to \infty$ for almost all $x(\mu)$. Also $N = I_{A_i}$ is binomial with parameters n and $p(x)$ where for almost all $x(\mu)$, $np(x)/\log n \to \infty$ as $n \to \infty$; this follows since $nh^d/\log n \to \infty$ and $h^d/p(x) \to g(x)$, almost all $x(\mu)$, for some $g \in L^1(\mu)$, $g \ge 0$ by Lemma 2.2.
For any $\epsilon > 0$, we have a.s.,

$$P\{| U(x) - EU(x) | > \epsilon | A_1, \cdots, A_n\} \le c_3 \exp(-c_4 N)$$

where c_3, $c_4 > 0$ depend upon ϵ, γ, c_1 and c_2 only. Thus,

$$P\{| U(x) - EU(x) | > \epsilon\} \le c_3 E\{\exp(-c_4 N)\}$$

which is summable with respect to n for almost all $x(\mu)$ by Lemma 4.1. We can also estimate (4.6) from above by

$$c_5 \exp(-c_6 N)$$

because $\sup_i W_{ni}(x) \le c_2/c_1 N$. Another application of Lemma 4.1 and the Borel-Cantelli lemma shows that for almost all $x(\mu)$, $| m_n(x) - m(x) | \to 0$ a.s. as $n \to \infty$.

5. Discrimination. In discrimination, Y takes values in $\{1, \cdots, M\}$ and is estimated from X and $(X_1, Y_1), \cdots, (X_n, Y_n)$ by $g_n(X)$. This results in a *probability of error*

$$L_n = P\{g_n(X) \ne Y | X_1, Y_1, \cdots, X_n, Y_n\} \ge L^* = \inf_{g: R^d \to \{1, \cdots, M\}} P\{g(X) \ne Y\},$$

where L^* is the Bayes probability of error. Consider now functions g_n that satisfy

$$\sum_{i=1}^{n} W_{ni}(x) I_{(Y_i = g_n(x))} = \max_{1 \le k \le M} \sum_{i=1}^{n} W_{ni}(x) I_{(Y_i = k)}. \tag{5.1}$$

For particular choices of the weights, we thus obtain the nearest neighbor discrimination

rule (Cover and Hart, 1967), the k-nearest neighbor rule (Fix and Hodges, 1951) and the potential function method or kernel method. For references, see Stone (1977), Devroye (1978b) or Collomb (1981). Since (5.1) implies

$$(5.2)\qquad 0 \le L_n - L^* \le 2 \sum_{j=1}^{M} E\{|P(Y=j|X) - \sum_{i=1}^{n} W_{ni}(X) I_{(Y_i=j)} | X_1, Y_1, \cdots, X_n, Y_n\}$$

(see Stone, 1977, page 617 or Devroye, 1978b, page 3), a straightforward application of Theorems 4.1 and 4.2 gives:

THEOREM 5.1.
(i) *In* (5.1) *let the* W_{ni}'s *be nearest neighbor weights* (1.2). *If* (4.2) *holds, then* $L_n \to L^*$ *in probability as* $n \to \infty$. *If in addition* $k/\log n \to \infty$ *as* $n \to \infty$, *then* $L_n \to L^*$ a.s. *as* $n \to \infty$.
(ii) *In* (5.1) *let the* W_{ni}'s *be kernel weights* (1.3). *If* (2.2) *holds, then* $L_n \to L^*$ *in probability as* $n \to \infty$. *If in addition* $nh^d/\log n \to \infty$ *as* $n \to \infty$, *then* $L_n \to L^*$ a.s. *as* $n \to \infty$.

REMARK 5.1. In Theorem 5.1 absolutely no conditions are imposed on the distribution of (X, Y).

REMARK 5.2. Györfi (1978) has shown that (5.2) remains valid even when the coefficient "2" is deleted.

REFERENCES

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

COLLOMB, G. (1976). Estimation non paramétrique de la regression par la méthode du noyau. Ph.D. dissertation, Université Paul Sabatier, Toulouse, France.

COLLOMB, G. (1977). Quelques propriétés de la méthode du noyau pour l'estimation non paramétrique de la regression en un point fixé. *Comptes Rendus de l'Académie des Sciences de Paris* **285** 282–292.

COLLOMB, G. (1981). Estimation non paramétrique de la regression: revue bibliographique. *Internat. Statist. Review* **49** 75–93.

COVER, T. M. (1968). Estimation by the nearest neighbor rule. *IEEE Trans. Inform. Theory* **14** 50–55.

COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13** 21–27.

DEVROYE, L. (1978a). The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Trans. Inform. Theory* **24** 142–151.

DEVROYE, L. (1978b). Universal consistency in nonparametric regression and nonparametric discrimination. Technical Report SOCS-78-10, School of Computer Science, McGill University, Montreal.

DEVROYE, L. and WAGNER, T. J. (1980a). On the L1 convergence of kernel regression function estimators with applications in discrimination. *Z. Wahrsch. verw. Gebiete* **51** 15–25.

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

FIX, E. and HODGES, J. L. (1951). Discriminatory analysis, nonparametric discrimination, consistency properties. Project 21-49-004, Report No. 4, School of Aviation Medicine, Randolph Field, Texas.

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

GYÖRFI, L. (1978). On the rate of convergence of nearest neighbor rules. *IEEE Trans. Inform. Theory* IT-24, 509–512.

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. *J. Amer. Statist. Assoc.* **58** 13–30.

MARCINKIEWICZ, J. and ZYGMUND, A. (1937). Sur les fonctions indépendantes. *Fund. Math.* **29** 60–90.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Applic.* **9** 141–142.

NADARAYA, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory Probab. Applic.* **10** 186–190.

PETROV, V. V. (1975). *Sums of Independent Random Variables.* Springer, Berlin.

RÉVÉSZ, P. (1979). On the nonparametric estimation of the regression function. *Prob. Control Inform. Theory* **8** 297–302.

SCHUSTER, E. and YAKOWITZ, S. (1979). Contributions to the theory of nonparametric regression, with application to system identification. *Ann. Statist.* **7** 139–149.

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.

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

WHEEDEN, R. L. and ZYGMUND, A. (1977). *Measure and Integral.* Marcel Dekker, New York.

SCHOOL OF COMPUTER SCIENCE
MCGILL UNIVERSITY
805 SHERBROOKE STREET WEST
MONTREAL
CANADA H3A 2K6