The Annals of Statistics
1981, Vol. 9, No. 6, 1320-1327

ON THE ASYMPTOTIC PROBABILITY OF ERROR IN NONPARAMETRIC DISCRIMINATION

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 $R^d \times \{0, 1\}$, and let $\hat{Y}$ be the k-nearest neighbor estimate of Y from X and the (X_i, Y_i)'s. We show that for all distributions of (X, Y), the limit of $L_n = P(\hat{Y} \neq Y)$ exists and satisfies

$$\lim_{n\to\infty} L_n \leq (1 + a_k)R^*, \quad a_k = \frac{\alpha\sqrt{k}}{k - 3.25}\left(1 + \frac{\beta}{\sqrt{k - 3}}\right), \ k \text{ odd}, \ k \geq 5,$$

where R^* is the Bayes probability of error and $\alpha = 0.3399 \cdots$ and $\beta = 0.9749 \cdots$ are universal constants. This bound is shown to be best possible in a certain sense.

0. Introduction. Consider a sequence $(X, Y), (X_1, Y_1), \cdots, (X_n, Y_n)$ of independent $R^d \times \{0, 1\}$ valued random variables with a common distribution. Let μ be the probability measure of X and let

$$\eta(x) = P(Y = 1 \mid X = x), \qquad x \in R^d.$$

In discrimination problems, one considers estimates $\hat{Y}$ of Y where $\hat{Y}$ denotes a $\{0, 1\}$-valued Borel measurable function of X and $(X_1, Y_1), \cdots, (X_n, Y_n)$. For example, the k-nearest neighbor estimate $\hat{Y}$ is defined as follows (Fix and Hodges, 1951): find the k nearest neighbors of X among $X_1, \cdots, X_n$; break ties by comparing indices; take a majority vote among the Y_i's that correspond to selected X_i's; set $\hat{Y}$ equal to the chosen integer; in case of a voting tie, set $\hat{Y}$ equal to Y_i where i is the smallest index among the selected X_i's. Cover and Hart (1965) have shown that under some conditions on μ and η, if $L_n = P(\hat{Y} \neq Y)$ is the probability of error (error rate), then

$$\limsup_{n\to\infty} L_n \leq c_k R^*, \tag{1}$$

where

$$R^* = \inf_{g: R^d \to \{0, 1\}} P(g(X) \neq Y)$$

is the Bayes probability of error, and c_k is a sequence of numbers such that $c_{2k+1} = c_{2k}$, $c_k \downarrow 1$ as $k \to \infty$ and $c_1 = 2$. Stone (1977) has shown that if k varies with n in such a way that $k/n \to 0$, $k \to \infty$, then $L_n \to R^*$ as $n \to \infty$ for all distributions of (X, Y). Implicit in the same paper is the following result (see also Devroye, 1981a): for $k = 1$, and for all distributions of (X, Y),

$$\lim_{n\to\infty} L_n = E[2\eta(X)\{1 - \eta(X)\}]. \tag{2}$$

For other properties of the k-nearest neighbor estimate, see Wagner (1971), Fritz (1975), Gyorfi (1980) and Devroye (1981b, c). In this paper we will prove various results related to

Received June, 1980; revised February, 1981.

[1] This research was sponsored in part by National Research Council of Canada Grant No. A3456.
AMS 1980 *subject classifications.* Primary 62G05.

Key words and phrases. Nonparametric discrimination, pattern recognition, inequality of Cover and Hart, nearest neighbor rule, probability of error.

(1) and (2). For example, we will show that for $k \geq 5$, k odd, and for all distributions of (X, Y), (1) is valid with

$$c_k = 1 + \alpha \frac{\sqrt{k}}{k - 3.25}\left(1 + \frac{\beta}{\sqrt{k} - 3}\right), \qquad \text{some } \alpha, \beta > 0. \tag{3}$$

We will also see that this result is the best possible in the sense that

$$\lim_{k\to\infty} \frac{\sqrt{k}}{\alpha} \sup_{\text{all distributions of } (X,Y) \text{ with } R^* > 0} (\lim_{n\to\infty} L_n/R^* - 1) = 1. \tag{4}$$

In other words, the best sequence c_k in (1) must necessarily be of the form $1 + (\alpha/\sqrt{k}\,)\cdot\{1 + o(1)\}$ as $k \to \infty$. The exact values of the best possible constants are only known for a couple of integers k, e.g. $c_1 = 2$, $c_3 = (7\sqrt{7} + 17)/27 \simeq 1.3155$. They can be obtained by numerical solution of high degree polynomial equations for k greater than 3. The numbers c_k have a considerable impact on the asymptotical error rate for other estimates $\hat{Y}$ as well, and a couple of examples will be given in Section 3.

1. Definitions and lemmas. We will define a class of estimates $\hat{Y}$ that are based on a *majority voting scheme*. These estimates are completely determined by functions g_n that map $R^{d(n+1)}$ to the subsets of $\{1, \cdots, n\}$ (there are 2^n elements in the range of g_n), and we require that all g_n's be Borel measurable. To save space, we will denote $g_n(x, X_1, \cdots, X_n)$ by G_x. In general, the cardinality N_x of G_x is a random variable. For the k-nearest neighbor estimate, $N_x = k$ and G_x is the collection of those indices that correspond to the k nearest neighbors of x among $X_1, \cdots, X_n$. We say that $\hat{Y}$ is an *m.v. estimate* when $\hat{Y}$ is determined by taking a majority vote among the Y_i's, $i \in G_x$. In case of a voting tie, let $\hat{Y} = Y_i$ where i is the smallest index in G_x. If $N_x = 0$, then $\hat{Y} = 0$. We will write $\hat{Y}_x$ to make the dependence upon x explicit whenever necessary.

Let us define further

$$r_n(x) = \eta(x) P(\hat{Y}_x = 0 \mid X_1, \cdots, X_n) + \{1 - \eta(x)\} P(\hat{Y}_x = 1 \mid X_1, \cdots, X_n),$$

$$t_k(x) = \eta(x) \textstyle\sum_{0 \leq i < k/2} \binom{k}{i} \eta^i(x)\{1 - \eta(x)\}^{k-i}$$

$$+ \{1 - \eta(x)\} \textstyle\sum_{k/2 < i \leq k} \binom{k}{i} \eta^i(x)\{1 - \eta(x)\}^{k-i}, \qquad k \geq 1, k \text{ odd},$$

and $\quad t_0(x) = \eta(x)$, $t_{2k}(x) = t_{2k-1}(x)$, $\quad$ all $k \geq 1$.

LEMMA 1. *If $B_1, \cdots, B_n, B_1', \cdots, B_n'$ are independent Bernoulli random variables with probabilities $p_1, \cdots, p_n, q_1, \cdots, q_n$, then*

$$\sup_{\text{all subsets } C \text{ of } \{0,1,\cdots,n\}} |P(\textstyle\sum_{i=1}^n B_i \in C) - P(\sum_{i=1}^n B_i' \in C)| \leq \sum_{i=1}^n |p_i - q_i|.$$

PROOF. One can use the following embedding argument. Let $U_1, \cdots, U_n$ be independent uniform [0, 1] random variables, and let $A_i = I_{[U_i \leq p_i]}$ and $A_i' = I_{[U_i \leq q_i]}$ where I is the indicator function. Then $A_1, \cdots, A_n$ is distributed as $B_1, \cdots, B_n$ and $A_1', \cdots, A_n'$ is distributed as $B_1', \cdots, B_n'$. Thus, for any set C,

$$|P(\textstyle\sum_{i=1}^n A_i \in C) - P(\sum_{i=1}^n A_i' \in C)| \leq |P(\sum_{i=1}^n A_i \neq \sum_{i=1}^n A_i')| \leq \sum_{i=1}^n P(A_i \neq A_i')$$

$$= \textstyle\sum_{i=1}^n |p_i - q_i|.$$

LEMMA 2. *For any m.v. estimate,*

$$|r_n(x) - t_{N_x}(x)| \leq 3/2 \textstyle\sum_{i \in G_x} |\eta(X_i) - \eta(x)| \qquad \text{a.s., } all\ x \in R^d.$$

PROOF. $N = N_x$ is a Borel measurable function of $x, X_1, \cdots, X_n$. If $Y'_1, \cdots, Y'_N$ are independent Bernoulli random variables with probabilities all equal to $\eta(x)$, then, on $[N > 0]$,

$$t_N(x) = \eta(x)\, P\left(\sum_{i=1}^N Y'_i < \frac{N}{2} \mid N\right) + \{1 - \eta(x)\} P\left(\sum_{i=1}^N Y'_i > \frac{N}{2} \mid N\right) + \frac{1}{2} P\left(\sum_{i=1}^N Y'_i = \frac{N}{2} \mid N\right).$$

Given $X_1, \cdots, X_n$, the random variables $Y_1, \cdots, Y_n$ are independent Bernoulli with means $\eta(X_1), \cdots, \eta(X_n)$. Also, on $[N > 0]$,

$$r_n(x) = \eta(x)\, P\left(\sum_{i \in G_x} Y_i < \frac{N}{2} \mid X_1, \cdots, X_n\right) + \frac{1}{2} P\left(\sum_{i \in G_x} Y_i = \frac{N}{2} \mid X_1, \cdots, X_n\right) + \{1 - \eta(x)\}\, P\left(\sum_{i \in G_x} Y_i > \frac{N}{2} \mid X_1, \cdots, X_n\right).$$

On $[N = 0]$, we have $r_n(x) = t_0(x) = \eta(x)$. Lemma 2 now follows by a triple application of Lemma 1.

LEMMA 3. *For any m.v. estimate,*

$$|L_n - E\{t_{N_X}(X)\}| = |E\{r_n(X)\} - E\{t_{N_X}(X)\}| \le E\{|r_n(X) - t_{N_X}(X)|\} \le E\{3/2 \sum_{i \in G_X} |\eta(X_i) - \eta(X)|\}.$$

PROOF. Note that $L_n = Er_n(X)$, and apply Lemma 2.

LEMMA 4. *Consider m.v. estimates with the following properties:*

(5) $1 \le N_x \le k$, *all* $x \in R^d$, *all* n,

(6) $\sup_{i \in G_x} \|X_i - x\| \to 0$ *in probability as* $n \to \infty$, *almost all* $x(\mu)$,

(7) *there exists a constant c such that for all* [0, 1] *valued Borel measurable functions g on* R^d,

$$E\{\sum_{i \in G_X} g(X_i)\} \le cEg(X).$$

Then

$$L_n - Et_{N_X}(X) \to 0 \qquad \text{as } n \to \infty. \tag{8}$$

This conclusion remains valid if (7) is replaced by the condition that η is continuous almost everywhere (μ). Furthermore, whenever (8) holds and there is a random variable N such that $N_x \to^{\mathscr{L}} N \ge 1$, almost all $x(\mu)$, we have

$$L_n \to \sum_{j=1}^\infty P(N = j) Et_j(X) \qquad \text{as } n \to \infty. \tag{9}$$

PROOF. By Lemma 3, (8) follows if we can show that $E\{\sum_{i \in G_X} |\eta(X_i) - \eta(X)|\} \to 0$. Let x be a point of continuity of η, and let $D_x = \sup_{i \in G_x} \|X_i - x\| \to 0$ in probability. Then,

$$E\{\sum_{i \in G_x} |\eta(X_i) - \eta(x)|\} \le k\{\sup_{\|y-x\| \le r} |\eta(y) - \eta(x)| + P(D_x > r)\},$$

and this can be made arbitrarily small by choosing r small enough and then letting $n \to \infty$. By the Lebesgue dominated convergence theorem, we may conclude that (8) holds when η is continuous for almost all $x(\mu)$. For general η, we may argue as follows. For any $\epsilon > 0$, find η' bounded and continuous such that $E(|\eta(X) - \eta'(X)|) < \epsilon$. Then

$$E\{\sum_{i \in G_X} |\eta(X_i) - \eta(X)|\} \le E\{\sum_{i \in G_X} |\eta(X_i) - \eta'(X_i)|\} + E\{\sum_{i \in G_X} |\eta'(X_i) - \eta'(X)|\} + E\{\sum_{i \in G_X} |\eta(X) - \eta'(X)|\}. \tag{10}$$

By (7), the sum of the second and the fourth term in (10) is not greater than $(c + k)\epsilon$. We have already shown that the third term tends to 0 as $n \to \infty$, and thus (8) is proved. Finally, the absolute value of the difference between $E\{t_{N_X}(X)\}$ and the right-hand-side of (9) is not greater than

$$E\{\textstyle\sum_{j=1}^{\infty} | P(N_X = j | X) - P(N = j) |\} = Ea(X).$$

For almost all $x(\mu)$, we have $a(x) \to 0$ as $n \to \infty$. Also, $0 \le a(x) \le 2$, and therefore $Ea(X) \to 0$ as $n \to \infty$. This concludes the proof of (9).

LEMMA 5. *Let $\mathscr{A}$ be a class of Borel sets from R^d, and let $C_{x,r}$ be the closed sphere of R^d centered at x with radius r. If there exists $c > 0$ such that*

$$A \subseteq C_{0,1}, \qquad c\lambda(A) \ge \lambda(C_{0,1}), \qquad \text{all } A \in \mathscr{A},$$

where λ is the Lebesgue measure, and if μ is a probability measure on the Borel sets of R^d with density f, then there exists a set B such that $\mu(B) = 1$, and

$$\sup_{A\in\mathscr{A}} \left| \frac{\mu(x + rA)}{\lambda(x + rA)} - f(x) \right| \le \sup_{A\in\mathscr{A}} \int_{x+rA} | f(y) - f(x) | \, dy/\lambda(x + rA)$$

$$\le c \int_{C_{x,r}} | f(y) - f(x) | \, dy/\lambda(C_{x,r}) \to 0 \text{ as } r \to 0, \qquad \text{all } x \in B.$$

PROOF. Apply the Lebesgue density theorem. See also Wheeden and Zygmund (1977, pages 108–109).

2. Main results. From Lemma 4 we see that the quantities $Et_k(X)$ are of great importance for all m.v. estimates. In this section we study the asymptotic behavior as $k \to \infty$, uniformly over all distributions of (X, Y). We will need three universal constants related to the tail of the normal distribution. If $Q(t) = \int_t^\infty \exp(-u^2/2)\, du/\sqrt{2\pi}$ then we define

$$\alpha = \max_{t>0} 2tQ(t) = 0.3399424150\ldots,$$

and let δ be the value of t for which this maximum is attained, namely

$$\delta = 0.7517915241\ldots.$$

Furthermore, we let

$$\beta = \max_{t>0} 2t^2Q(t)/\alpha = 0.9749687445\ldots.$$

We define the sequence

$$a_k = \alpha \frac{\sqrt{k}}{k - 3.25}\left(1 + \frac{\beta}{\sqrt{k - 3}}\right).$$

The main result of this section is the following.

THEOREM 1. *Let*

$$T_k = \sup_{\text{all distributions of } (X,Y) \text{ with } R^* > 0} \frac{Et_k(X)}{R^*} - 1.$$

Then, for k odd, $k \ge 5$, $T_k \le a_k$. Also, $T_k \sim \alpha/\sqrt{k}$ as $k \to \infty$.

PROOF. Note that for $x \in R^d$ and $k \ge 1$, k odd,

$$\frac{t_k(x)}{\eta(x)} - 1 = \left\{\frac{1 - 2\eta(x)}{\eta(x)}\right\} \sum_{i>k/2} \binom{k}{i} \eta^i(x)\{1 - \eta(x)\}^{k-i}.$$

If we can show that on $A = \{x \mid \eta(x) \le ½\}$, $t_k(x)/\eta(x) - 1 \le a_k$, and that on the complement of A, A^c, $t_k(x)/\{1 - \eta(x)\} - 1 \le a_k$, then

$$\begin{aligned} Et_k(X) &= E\{t_k(X)I_A(X)\} + E\{t_k(X)I_{A^c}(X)\} \\ &\le (1 + a_k)[E\{\eta(X)I_A(X)\} + E\{(1-\eta(X))I_{A^c}(X)\}] \\ &= (1 + a_k)E[\min\{\eta(X), 1 - \eta(X)\}] \\ &= (1 + a_k)R^*. \end{aligned}$$

Let $b_i(k, p)$ be the ith term of the binomial distribution with parameters k and p. It is clear that we need only show that for k odd, $k \ge 5$,

(11) $$B_k = \sup_{0<p\le 1/2} \frac{1-2p}{p} \sum_{i>k/2} b_i(k, p) \le a_k.$$

By the relation between the binomial and the beta distribution,

(12) $$\sum_{i>k/2} b_i(k, p) = \int_0^p \{x(1-x)\}^{(k-1)/2} \frac{k!}{\left[\left\{\frac{1}{2}(k-1)\right\}!\right]^2} dx.$$

More conveniently, with

$$p = \frac{1}{2} - q,\ x = \frac{1}{2}\left(1 - \frac{z}{\sqrt{k-3}}\right),$$

this expression can be rewritten as

$$c'_k \int_{2q\sqrt{k-3}}^{\sqrt{k-3}} \left(1 - \frac{z^2}{k-3}\right)^{(k-1)/2} dz,$$

where

$$c'_k = k! \left[\left\{\left(\frac{k-1}{2}\right)!\right\}^2 2^k \sqrt{k-3}\right]^{-1}$$

Now, using the Cesaro-Buchner inequalities (Buchner, 1951; Mitrinovic, 1970, page 183),

$$\left(12k + \frac{1}{4}\right)^{-1} < \log \frac{k!}{\left(\frac{k}{e}\right)^k \sqrt{2\pi k}} < (12k)^{-1}, \qquad k \ge 2,$$

we see that

$$c'_k \le \sqrt{\frac{k}{2\pi(k-3)}} \left(\frac{k}{k-1}\right)^k \exp\left(-1 + \frac{1}{12k} - \frac{2}{6k - 23/4}\right) = c''_k.$$

Next, because $\log(1-u) \ge -u - u^2/\{2(1-u)\}$, $u > 0$, we have

$$\left(\frac{k-1}{k}\right)^k = \left(1 - \frac{1}{k}\right)^k \ge \exp\left(-1 - \frac{1}{2k-2}\right).$$

Thus,

$$c''_k \le c^*_k = \sqrt{\frac{k}{2\pi(k-3)}} \exp(\gamma_k)$$

where

$$\gamma_k = \frac{1}{12k} + \frac{1}{2k-2} - \frac{2}{6k - 23/4}.$$

Since for $z \geq 2q\sqrt{k-3}$, we have

$$2p = 1 - 2q = (1-4q^2)/(1+2q) \geq \{1 - z^2/(k-3)\}/(1+2q),$$

B_k can be estimated from above as follows:

$$B_k \leq \sup_{0\leq q<1/2} (4q)(1+2q)c_k^* \int_{2q\sqrt{k-3}}^{\sqrt{k-3}} \left(1 - \frac{z^2}{k-3}\right)^{(k-3)/2} dz$$

$$\leq \sup_{0\leq q<1/2} 2(1+2q)\frac{\sqrt{k}}{k-3}(2q\sqrt{k-3}) \int_{2q\sqrt{k-3}}^{\infty} e^{-z^2/2}\frac{1}{\sqrt{2\pi}}\, dz.$$

$$\leq \frac{\sqrt{k}}{k-3} e^{\gamma_k}\{\alpha + \sup_{u>0} 2u^2 Q(u)/\sqrt{k-3}\}$$

$$= \frac{\sqrt{k}}{k-3} e^{\gamma_k}(\alpha + \alpha\beta/\sqrt{k-3}) \leq \frac{\sqrt{k}}{k-3}\frac{\alpha}{1-\gamma_k}\left(1 + \frac{\beta}{\sqrt{k-3}}\right).$$

Now,

$$B_k \leq a_k \quad \text{for all odd} \quad k \geq 5 \text{ if } (k-3)(1-\gamma_k) \geq k - 13/4.$$

But this follows from the observation that

$$(k-3)\gamma_k = \frac{1}{12} - \frac{1}{2} + \frac{1}{3} - \frac{1}{4k} - \frac{1}{k-1} - \frac{49}{72k-69} \leq \frac{1}{4}$$

for all $k > 1$.

To prove the second half of Theorem 1, consider Y independent of X with

$$P(Y=1) = p = p(k) = \frac{1}{2}\left(1 - \frac{\delta}{\sqrt{k-1}}\right).$$

Clearly, $R^* = p$, and

$$T_k \geq \frac{1-2p}{p}\sum_{i>k/2} b_i(k,p) \sim \frac{2\delta}{\sqrt{k}}\frac{\sqrt{k}2^k}{\sqrt{2\pi}}\int_0^p \{x(1-x)\}^{(k-1)/2}\, dx \tag{13}$$

$$\sim \frac{2\delta}{\sqrt{k-1}}\int_\delta^{\sqrt{k-1}} \frac{1}{\sqrt{2\pi}}\left(1 - \frac{z^2}{k-1}\right)^{(k-1)/2} dz \sim \frac{2\delta}{\sqrt{k}} Q(\delta) = \frac{\alpha}{\sqrt{k}}.$$

Here we have used Stirling's formula to show that

$$k!\left\{\left(\frac{k-1}{2}\right)!\right\}^{-2} \sim \sqrt{k}2^k/\sqrt{2\pi}.$$

The last approximation follows from the dominated convergence theorem after noting that $\{1 - z^2/(k-1)\}^{(k-1)/2} \leq \exp(-z^2/2)$, all $z \leq \sqrt{k-1}$. Theorem 1 now follows from (13) and $T_k \leq a_k \sim \alpha/\sqrt{k}$.

REMARK 1. The proof of the theorem was based on the observation that $T_k = B_k$; see (11). The "worst" $p(k)$, i.e., the value of p for which the supremum in (11) is reached, must necessarily satisfy

$$p(k) = \frac{1}{2}\left[1 - \frac{\delta}{\sqrt{k}}\{1 + o(1)\}\right]$$

as $k \to \infty$. Notice in particular that $p(k) \to 1/2$ as $k \to \infty$.

REMARK 2. The following bound is valid for *all* $k \geq 1$:

$$Et_k(X) \leq \left(1 + \sqrt{\frac{2}{k}}\right) R^*.$$

This bound is the best possible among all the bounds of the form $\left(1 + \frac{a}{\sqrt{k}}\right) R^*$ since it is attainable for $k = 2$. Another simple bound, valid for $k \geq 3$, is

$$Et_k(X) \leq \left(1 + \frac{1}{\sqrt{k}}\right) R^*.$$

3. Examples.

The k-nearest neighbor estimate. The k-nearest neighbor estimate, mentioned in the introduction, is an m.v. estimate with $N_x = k$, all x. Also, for all $x \in S = \text{support}(\mu)$, we have $D_x = \sup_{i \in G_x} \|X_i - x\| \to 0$ a.s. as $n \to \infty$. (The notation S and D_x will be used throughout this section.) Thus, (5) and (6) are satisfied. Finally, Stone (1977) has shown that (7) holds with $c = kc_1$ where c_1 is a function of d only. We have without work the following result.

THEOREM 2. *For the k-nearest neighbor estimate,* $\lim_{n\to\infty} L_n$ *exists and is equal to* $Et_k(X)$. *Thus,*

$$\lim_{n\to\infty} L_n \leq (1 + a_k) R^*$$

and (4) *is valid.*

The sphere estimate. The sphere estimate is defined by a sequence of numbers $h = h(n)$ such that

$$h \sim \left(\frac{c}{Ln}\right)^{1/d}, \tag{14}$$

where $c > 0$ is a constant, and $L = \lambda(C_{0,1})$ is the volume of the unit sphere of R^d. We let

$$i \in G_x \quad \text{iff} \quad \|X_i - x\| \leq h.$$

Clearly, N_x is binomial $(n, \mu(C_{x,h}))$. Lemma 5 implies that $n\mu(C_{x,h}) \to cf(x)$, almost all $x(\mu)$, when μ has a density f. Therefore, for almost all x, $N_x \to^{\mathscr{L}} \mathscr{P}(cf(x))$ where $\mathscr{P}$ is the Poisson law. The condition $nh^d \to \infty$ would entail $N_x \to \infty$ in probability, almost all x. This is the classical condition required for the Bayes risk consistency of sphere estimates: Devroye and Wagner (1980) and Spiegelman and Sacks (1980) have shown that $\lim h + (nh^d)^{-1} = 0$ implies $\lim L_n = R^*$ for all distributions of (X, Y). This result remains true for the present h when μ is atomic, but it is false for (14) when μ has a density.

THEOREM 3. *Whenever X has a density* $f \in L^2(\lambda)$, *the sphere estimate with sequence h as in* (14) *satisfies*

$$\lim_{n\to\infty} L_n = E\left[\sum_{j=0}^{\infty} t_j(X) \frac{\{cf(X)\}^j e^{-cf(X)}}{j!}\right].$$

PROOF. We will first show that (8) remains valid, modifying the proof of Lemma 4 very slightly. Since $D_x \leq h \to 0$ as $n \to \infty$, (8) is valid when η is continuous and lim sup $E(N_x) < \infty$, almost all $x(\mu)$. The latter condition is satisfied in view of $E(N_x) = n\mu(C_{x,h}) \to cf(x)$, almost all x. For Borel measurable η, we use an argument as in (10). By symmetry, the sum of the second and fourth terms of (10) is

$$2E\{\textstyle\sum_{i \in G_X} |\eta(X) - \eta'(X)|\}. \tag{15}$$

The third term of (10) is $o(1)$. Thus, we should just make sure that (15) is arbitrarily small by choice of η'. Let η^* be a [0, 1]-valued Borel measurable function on R^d. Then

(16) $$E\{\sum_{i \in G_X} \eta^*(X)\} = E\{n\mu(C_{X,h})\eta^*(X)\} = (nh^d L) E\{\mu(C_{X,h})\eta^*(X)/(h^d L)\}.$$

The first factor on the right hand side of (16) tends to c as $n \to \infty$. The second factor tends to $E\{f(X)\eta^*(X)\} = \int f^2(x)\eta^*(x)\,dx$ as $h \to 0$, whenever $f \in L^2(\lambda)$. To see this, notice that

$$\mu(C_{x,h})/(Lh^d) \begin{cases} \to f(x), & \text{almost all } x(\mu), \\ \le f^*(x) = \sup_{r>0} \mu(C_{x,r})/(Lr^d), & \text{all } h > 0, \quad x \in R^d. \end{cases}$$

Since $f^* f \eta^* \le f^{*2} \in L^1(\lambda)$ whenever $f \in L^2(\lambda)$ (Wheeden and Zygmund, 1977, page 155), the Lebesgue dominated convergence theorem can be applied. But for every $\epsilon > 0$, there exists $\delta > 0$ such that $\int f(x)\eta^*(x)\,dx < \delta$ implies $\int f^2(x)\eta^*(x)\,dx < \epsilon$. Thus, since continuous functions are dense in $L^1(\mu)$, we can make (10) arbitrarily small, and (8) follows. The remainder of the proof is similar to that of Lemma 4.

REMARK 3. For the kernel estimate, let us call $L(c) = \lim L_n$. We first note that

$$\sup_{\text{all distributions of } (X,Y) \text{ with } R^* > 0} \frac{L(c)}{R^*} = \infty, \qquad \text{all fixed } c > 0.$$

Indeed, from Theorem 3 we note that $L(c) \ge E\{\eta(X)e^{-cf(X)}\}$. If we let Y be independent of X and choose $\eta \equiv p > ½$, then

$$E\{\eta(X)e^{-cf(X)}\}/R^* = E\{e^{-cf(X)}\}\frac{p}{1-p} \uparrow \infty \text{ as } p \uparrow 1.$$

Thus, distribution-free upper bounds for $L(c)$ of the type derived in Theorem 2 for the k-nearest neighbor estimate do not exist.

REFERENCES

BUCHNER, P. (1951). Bemerkungen zur die Stirlingschen Formel. *Elem. Math.* **6** 8–11.

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

DEVROYE, L. (1981a). On the inequality of Cover and Hart in nearest neighbor discrimination. *IEEE Trans. Pattern Anal. Mach. Intell.* **3** 75–78.

DEVROYE, L. (1981b). The strong convergence of empirical nearest neighbor estimates of integrals. To appear in: Proceedings of the International Symposium on Statistics and Related Topics, E. Saleh, Ed., North Holland.

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

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

FELLER, W. (1968). *An Introduction To Probability Theory And Its Applications.* Wiley, New York.

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.

FRITZ, J. (1975). Distribution-free exponential error bound for neareast neighbor pattern classification. *IEEE Trans. Inform. Theory* **21** 552–557.

GYORFI, L. (1980). Recent results on nonparametric regression estimate and multiple classification. *Problems Control Inform. Theory*, to appear.

MITRINOVIC, D. S. (1970). *Analytic Inequalities.* Springer-Verlag, Berlin.

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.

WAGNER, T. J. (1971). Convergence of the nearest neighbor rule. *IEEE Trans. Inform. Theory* **17** 566–571.

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

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