The Annals of Statistics
1983, Vol. 11, No. 3, 896–904

THE EQUIVALENCE OF WEAK, STRONG AND COMPLETE CONVERGENCE IN L_1 FOR KERNEL DENSITY ESTIMATES[1]

By Luc Devroye

McGill University

Let f be a density on R^d, and let f_n be the kernel estimate of f,

$$f_n(x) = (nh^d)^{-1} \sum_{i=1}^n K((x - X_i)/h)$$

where $h = h_n$ is a sequence of positive numbers, and K is an absolutely integrable function with $\int K(x)\,dx = 1$. Let $J_n = \int |f_n(x) - f(x)|\,dx$. We show that when $\lim_n h = 0$ and $\lim_n nh^d = \infty$, then for every $\varepsilon > 0$ there exist constants $r, n_0 > 0$ such that $P(J_n \geq \varepsilon) \leq \exp(-rn)$, $n \geq n_0$. Also, when $J_n \to 0$ in probability as $n \to \infty$ and K is a density, then $\lim_n h = 0$ and $\lim_n nh^d = \infty$.

1. Introduction. The purpose of this paper is to point out that for the celebrated Parzen-Rosenblatt density estimate (Parzen, 1962; Rosenblatt, 1956) all types of L_1 consistency are equivalent. We consider a sample $X_1, \cdots, X_n$ of independent R^d-valued random vectors with common density f, and estimate $f(x)$ by

$$f_n(x) = (nh^d)^{-1} \sum_{i=1}^n K((x - X_i)/h)$$

where $h = h_n$ is a sequence of positive numbers and K is a Borel measurable function satisfying $k \geq 0$, $\int K = 1$. The natural measure of the closeness of f_n to f is its L_1 distance,

$$J_n = \int |f_n(x) - f(x)|\,dx.$$

Our main result is:

THEOREM 1. *Let K be a nonnegative Borel measurable function on R^d with $\int K(x)\,dx = 1$. Then the following conditions are equivalent:* (i) *$J_n \to 0$ in probability as $n \to \infty$, some f;* (ii) *$J_n \to 0$ in probability as $n \to \infty$, all f;* (iii) *$J_n \to 0$ almost surely as $n \to \infty$, all f;* (iv) *$J_n \to 0$ exponentially as $n \to \infty$ (i.e. for all $\varepsilon > 0$, there exist $r, n_0 > 0$ such that $P(J_n \geq \varepsilon) \leq e^{-rn}$, $n \geq n_0$), all f;* (v) *$\lim_n h = 0$ and $\lim_n nh^d = \infty$. Also,* (v) *implies* (iv) *when K is merely absolutely integrable and $\int K(x)\,dx = 1$.* □

A weak analogue of Theorem 1 for histogram estimates was obtained by Abou-Jaoude (1976a, 1976b, 1976c). Theorem 1 improves Devroye and Wagner (1979), where L_1 convergence results are obtained from pointwise convergence results (such as Deheuvels, 1974) and Scheffé's Theorem (Scheffé, 1947; see also Glick, 1974 and Devroye, 1979).

2. Proof of Theorem 1. We will try to extract the key facts needed in the proof of Theorem 1. They are condensed in several lemmas of independent interest. Lemmas 1 and 2 are integral and pointwise versions of the Lebesgue density theorem. Lemma 3 contains a crucial inequality for the multinomial distribution, and in Lemma 4 we prove that (v) $\Rightarrow$ (iv). Lemma 5 is an L_1 version of the non-existence of unbiased kernel density estimates.

Received August 1981; revised August 1982

[1] This paper was written during the summer of 1981 while the author visited Applied Research Laboratories and the University of Texas at Austin. The sponsor was the Office of Naval Research under contract N00014-81-K-0145.

AMS 1970 *subject classifications.* Primary, 60F15; secondary, 62G05.

Key words and phrases. Nonparametric density estimation, L_1 convergence, kernel estimate, strong consistency.

The implication (i) ⇒ (v) is established in Lemma 6. Since (iv) ⇒ (iii) ⇒ (ii) ⇒ (i), this would then complete the proof of Theorem 1.

LEMMA 1. (*L_1 version of Bochner's theorem*). *Let K be an absolutely integrable function on R^d with $\int K(x)\,dx = 1$, and let $h = h_n$ be a sequence of positive numbers satisfying* $\lim_n h = 0$. *For each density f, we have* $\lim_n \int |g_h(x) - f(x)|\,dx = 0$, *where* $g_h(x) = h^{-d}\int K((x-y)/h)f(y)\,dy$.

PROOF OF LEMMA 1. The proof is based on a technique of Kantorovich and Akilov (1964). I am grateful to Laszlo Györfi for pointing this reference out to me. We let $C = \int |K(x)|\,dx$, and note that by a change of integral, for any function f,

$$\int |g_h(x)|\,dx \le \iint h^{-d}|K((x-y)/h)||f(y)|\,dy\,dx = C\int |f(y)|\,dy. \tag{1}$$

For each $\varepsilon > 0$ there exists a continuous function f^* vanishing outside a compact set, say S_{0R}, where S_{xr} is the closed sphere of radius r centered at x, such that $\int |f(x) - f^*(x)|\,dx < \varepsilon$. Thus, if we write $g_h(f, x)$ to make the dependence upon f explicit, then

$$\begin{aligned}
\int |g_h(f,x) - f(x)|\,dx & \\
&\le \int |g_h(f-f^*,x)|\,dx + \int |g_h(f^*,x) - f^*(x)|\,dx + \int |f^*(x) - f(x)|\,dx \\
&\le (C+1)\int |f^*(x) - f(x)|\,dx + \int |g_h(f^*,x) - f^*(x)|\,dx \\
&\le (C+1)\varepsilon + \int |g_h(f^*,x) - f^*(x)|\,dx.
\end{aligned}$$

Thus, we need only show the Lemma for all functions f^*. For each $\varepsilon > 0$, find $\delta(\varepsilon) > 0$ such that $\|x - y\| < \delta(\varepsilon)$ implies $|f^*(x) - f^*(y)| < \varepsilon$. Thus, if $f^* = 0$ outside S_{0R}, then

$$\begin{aligned}
\int |g_h(f^*,x) - f^*(x)|\,dx &= \int \left| \int_{\|x\|\le R, \|y\|\le R} h^{-d}K\left(\frac{(x-y)}{h}\right)\{f^*(y) - f^*(x)\}\,dy \right| dx \\
&\le \int_{\|x\|\le R} \left| \int_{\|y\|\le R, \|x-y\|\le\delta(\varepsilon)} + \int_{\|y\|\le R, \|x-y\|>\delta(\varepsilon)} \right| dx \\
&\le \int_{\|x\|\le R} \left(C\varepsilon + C_1 \int_{\|y\|\le R, \|x-y\|>\delta(\varepsilon)} h^{-d}\left| K\left(\frac{(x-y)}{h}\right)\right| dy \right) dx \\
&\le C\varepsilon(2R)^d + C_1(2R)^d \int_{\|hy\|>\delta(\varepsilon)} |K(y)|\,dy \\
&= C\varepsilon(2R)^d + o(1),
\end{aligned}$$

where $C_1 = \sup_x f^*(x)$. This concludes the proof of Lemma 1.

LEMMA 2. (*Lebesgue density theorem*). *If f is a density on R^d and B is a compact set of R^d with $\lambda(B) > 0$, then*

$$\lim_{h\downarrow 0}\lambda^{-1}(hB)\int_{x+hB} f(y)\,dy = f(x), \quad \textit{almost all} \quad x.$$

PROOF OF LEMMA 2. We know that

$$\lim_{h\downarrow 0}\lambda^{-1}(S_{xh})\int_{S_{xh}}|f(y)-f(x)|\,dy=0$$

for almost all x, by the classical version of the Lebesgue density theorem; see for example, Stein (1970, pages 62–63) or Wheeden and Zygmund (1977, pages 100–109). If S_{0R} is the smallest sphere containing B, then for almost all x,

$$\lambda^{-1}(x+hB)\int_{x+hB}|f(y)-f(x)|\,dy\le(\lambda(S_{0R})/\lambda(B))\lambda^{-1}(x+hS_{0R})\int_{x+hS_{0R}}|f(y)-f(x)|\,dy$$

which tends to zero as $h\downarrow 0$.

LEMMA 3. *(A multinomial distribution inequality). Let $(X_1,\cdots,X_k)$ be a multinomial $(n,p_1,\cdots,p_k)$ random vector. For all $\varepsilon\in(0,1)$ and all k satisfying $k/n\le\varepsilon^2/20$, we have*

$$P(\textstyle\sum_{i=1}^k|X_i-E(X_i)|>n\varepsilon)\le 3\exp(-n\varepsilon^2/25).$$

PROOF OF LEMMA 3. The proof is based upon a Poissonization. Let N be a Poisson(n) random variable independent of $U_1, U_2, \cdots$, which is a sequence of independent $\{1,\cdots,k\}$-valued variables distributed according to $P(U_1=i)=p_i$, $1\le i\le k$. Let X_i be the number of occurrences of the value i among $U_1,\cdots,U_n$, and let X_i' be the number of occurrences of the value i among $U_1,\cdots,U_N$. It is clear that $X_1',\cdots,X_k'$ are independent Poisson random variables with means $np_1,\cdots,np_k$, and that $X_1,\cdots,X_k$ is a multinomial $(n,p_1,\cdots,p_k)$ random vector. Since $E(X_i)=np_i$, we have

$$\sum_{i=1}^k\frac{1}{n}|X_i-np_i|\le\sum_{i=1}^k\frac{1}{n}|X_i-X_i'|+\sum_{i=1}^k\frac{1}{n}|X_i'-np_i|. \tag{2}$$

Now, when U is Poisson(λ), then for $t>0$,

$$E(e^{t|U-\lambda|})\le E\{e^{t(U-\lambda)}+e^{t(\lambda-U)}\}=e^{\lambda(e^t-1)-t\lambda}+e^{\lambda(e^{-t}-1)+t\lambda}\le 2e^{\lambda(e^t-1-t)},$$

because $e^{-t}+t\le e^t-t$. Thus,

$$\begin{aligned}P(|U-\lambda|\ge\lambda\varepsilon)&\le E(e^{t|U-\lambda|-t\lambda\varepsilon})\le 2e^{-t\lambda\varepsilon}e^{\lambda(e^t-1-t)}\\&=2e^{\lambda\{\varepsilon-(1+\varepsilon)\ln(1+\varepsilon)\}}\le 2e^{-\lambda\varepsilon^2/2(1+\varepsilon)}\le 2e^{-\lambda\varepsilon^2/4},\end{aligned}\tag{3}$$

where we took $t=\ln(1+\varepsilon)$. By a repetition of the previous argument, using (3) and making the substitution $t=\ln(1+3\varepsilon/5)$, we have

$$\begin{aligned}P\left(\sum_{i=1}^k\frac{1}{n}|X_i-np_i|\ge\varepsilon\right)&\le P\left(|N-n|\ge n\frac{2\varepsilon}{5}\right)+P\left(\sum_{i=1}^k\frac{1}{n}|X_i'-np_i|\ge n\frac{3\varepsilon}{5}\right)\\&\le 2e^{-n(2\varepsilon/5)^2/4}+e^{-tn(3\varepsilon/5)}\textstyle\prod_{i=1}^k\{2e^{np_i(e^t-1-t)}\}\\&\le 2e^{-n\varepsilon^21/25}+2^ke^{n(e^t-1-t-3\varepsilon t/5)}\\&\le 2e^{-n\varepsilon^21/25}+e^{k-n(3\varepsilon/5)^2/4}\\&\le 3e^{-n\varepsilon^21/25}\quad\text{when}\quad k\le n\varepsilon^2/20.\end{aligned}\tag{4}$$

REMARK 1. The original manuscript had the bound $1134/(n^2\varepsilon^8)$, valid for $k\le n\varepsilon^2/9$. I am grateful to Laszlo Györfi for suggesting the exponential inequality of Lemma 3.

LEMMA 4. *For any density f on R^d, and any absolutely integrable function K with $\int K(x)\,dx=1$, $J_n\to 0$ completely as $n\to\infty$ whenever $\lim_n h=0$ and $\lim_n nh^d=\infty$.*

PROOF OF LEMMA 4. Let g_h be defined as in the statement of Lemma 3. By Lemma 3, it suffices to show that $\int |f_n(x) - g_h(x)| \, dx \to 0$ completely as $n \to \infty$. Let μ_n be the empirical probability measure for $X_1, \cdots, X_n$, and note that

$$f_n(x) = h^{-d} \int K\left(\frac{(x-y)}{h}\right) \mu_n(dy).$$

For given $\varepsilon > 0$, find finite constants $M, L, N, a_1, \cdots, a_N$ and disjoint finite rectangles $A_1, \cdots, A_N$ in R^d such that the function

$$K^*(x) = \sum_{i=1}^N a_i I_{A_i}(x)$$

satisfies: $|K^*| \leq M$, $K^* = 0$ outside $[-L, L]^d$, and $\int |K(x) - K^*(x)| \, dx < \varepsilon$. Define g_h^* and f_n^* as g_h and f_n with K^* instead of K. Then

$$\begin{aligned}\int |f_n(x) - g_h(x)| \, dx &\leq \int |f_n(x) - f_n^*(x)| \, dx \\ &\quad + \int |f_n^*(x) - g_h^*(x)| \, dx + \int |g_h^*(x) - g_h(x)| \, dx \\ &\leq \int h^{-d} \int |K^*((x-y)/h) - K((x-y)/h)| f(y) \, dy \, dx \\ &\quad + \int h^{-d} \int |K^*((x-y)/h) - K((x-y)h)| \mu_n(dy) \, dx \\ &\quad + \int |f_n^*(x) - g_h^*(x)| \, dx \\ &\leq 2\varepsilon + \int |f_n^*(x) - g_h^*(x)| \, dx\end{aligned}$$

by a double change of integral. But if μ is the probability measure for f, then

$$\begin{aligned}\int |f_n^*(x) - g_h^*(x)| \, dx &\leq \sum_{i=1}^N |a_i| \int \left| h^{-d} \int_{x+hA_i} f(y) \, dy - h^{-d} \int_{x+hA_i} \mu_n(dy) \right| dx \\ &\leq M h^{-d} \sum_{i=1}^N \int |\mu(x + hA_i) - \mu_n(x + hA_i)| \, dx.\end{aligned}$$

Lemma 4 follows if we can show that for all finite rectangles A of R^d, $h^{-d} \int |\mu(x + hA) - \mu_n(x + hA)| \, dx \to 0$ exponentially as $n \to \infty$. Choose an A, and let $\varepsilon > 0$ be arbitrary. Consider the partition of R^d into sets B that are d-fold products of intervals of the form $[(i-1)h/N, ih/N)$, where i is an integer, and N is a fixed constant to be chosen later. Call the partition Ψ. Let $A = \prod_{i=1}^d [x_i, x_i + a_i)$, $\min_i a_i \geq 2/N$ and $A^* = \prod_{i=1}^d [x_i + 1/N, x_i + a_i - 1/N)$. Define

$$C_x = x + hA - \cup_{B \in \Psi, B \subseteq x+hA} B \subseteq x + h(A - A^*) = C_x^*.$$

Clearly,

$$\begin{aligned}(5) \quad &\int |\mu(x + hA) - \mu_n(x + hA)| \, dx \\ &\qquad \leq \int \sum_{B \in \Psi, B \subseteq x+hA} |\mu(B) - \mu_n(B)| \, dx + \int \{\mu(C_x) + \mu_n(C_x)\} \, dx.\end{aligned}$$

The last term in (5) equals

$$2\lambda(h(A-A^*)) = 2h^d\lambda(A-A^*) = 2h^d(\textstyle\prod_{i=1}^d a_i - \prod_{i=1}^d (a_i - 2/N))$$
$$= 2h^d\lambda(A)(1-\textstyle\prod_{i=1}^d (1-2/(Na_i))) \le 4h^d\lambda(A)\sum_{i=1}^d a_i^{-1}/N \le \varepsilon h^d$$

by choice of N. We used the fact that for any set C, and any probability measure ν on the Borel sets of R^d, $\int \nu(x+hC)\,dx = \lambda(hC)$. For any finite constant $R>0$, we can bound the first term in (5) from above by

$$(6)\qquad \textstyle\sum_{B\in\Psi, B\cap S_{0R}\ne\phi} |\mu_n(B)-\mu(B)| \displaystyle\int_{B\subseteq x+hA} dx + \int_{B\subseteq x+hA} dx\{\mu_n(S_{0R}^c) - \mu(S_{0R}^c) + 2\mu(S_{0R}^c)\}.$$

Here $(\cdot)^c$ denotes the complement of a set. Clearly, $h^{-d}\int_{B\subseteq x+hA} dx \le \lambda(A)$, and $\mu(S_{0R}^c) < \varepsilon$ by our choice of R. Also,

$$P\{\mu_n(S_{0R}^c) - \mu(S_{0R}^c) > \varepsilon\} \le e^{-2n\varepsilon^2}$$

by Hoeffding's inequality for binomial random variables (Hoeffding, 1963). Finally, since the collection of sets $B\in\Psi$ with $B\cap S_{0R}\ne\phi$ has at most $(2RN/h+2)^d = o(n)$ elements, we see that by Lemma 3, for all n large enough,

$$P(\textstyle\sum_{B\in\Psi, B\cap S_{0R}\ne\phi} |\mu_n(B)-\mu(B)| > \varepsilon) \le 3e^{-1n\varepsilon^2/25}.$$

Now collect bounds. This concludes the proof of Lemma 4.

LEMMA 5. (*Nonexistence of unbiased kernel density estimates*). *Let K and f be arbitrary densities on R^d, and let g_h be defined as in Lemma* 1. *Then* $\int |f(x)-g_a(x)|\,dx > 0$ *for all* $a>0$. *Also, when a_n is a positive number sequence,* $\lim_n \int |f(x)-g_{a_n}(x)|\,dx = 0$ *implies that* $\lim_n a_n = 0$.

PROOF OF LEMMA 5. Let ϕ and ψ be the characteristic functions of f and K respectively. Clearly, $g_a(x) = E\{f_n(x)\}$ has characteristic function $\psi(at)\phi(t)$. Now, $\int |f(x)-g_a(x)|\,dx = 0$ implies $f=g_a$ for almost all x, and thus $\phi(t) = \phi(t)\psi(at)$ for all $t\in R^d$. For $\phi(t)\ne 0$, i.e. at least in a neighborhood of the origin, $\psi(at) = 1$. But since $a\ne 0$, this implies that ψ cannot be the characteristic function of a density on R^d, and we have a contradiction. Thus, $\int |f(x)-g_a(x)|\,dx = 0$ implies $a=0$.

To prove the second statement of the Lemma, we assume first that $\lim_n a_n = \infty$. By Fatou's Lemma, $\int |f(x)-g_{a_n}(x)|\,dx \to 0$ implies $\lim\inf_n |f(x)-g_{a_n}(x)| = 0$, almost all x. But since $g_a(x)\to 0$ for almost all x, we have $f(x) = 0$ for almost all x, and this is impossible. Assume next that $\lim_n a_n = c\in(0,\infty)$. Now, $\int |f(x)-g_{a_n}(x)|\,dx \ge \int |f(x)-g_c(x)|\,dx - \int |g_c(x)-g_{a_n}(x)|\,dx$. By the first part of this Lemma, it suffices to show that $\int |g_c(x)-g_{a_n}(x)|\,dx \to 0$ to reach a contradiction, thereby concluding the proof of Lemma 5. Let $K_a(x) = a^{-d}K(x/a)$. For every $\varepsilon>0$ we can find a continuous bounded function K^* with compact support such that $\int |K^*(x)-K(x)|\,dx < \varepsilon$. Now, by (1),

$$\begin{aligned}\int |g_c(x)-g_{a_n}(x)|\,dx &\le \int |K_c(x)-K_{a_n}(x)|\,dx \le \int |K_c(x)-K_c^*(x)|\,dx \\ &\quad + \int |K_c^*(x)-K_{a_n}^*(x)|\,dx + \int |K_{a_n}^*(x)-K_{a_n}(x)|\,dx \\ &= 2\int |K^*(x)-K(x)|\,dx + \int |K_c^*(x)-K_{a_n}^*(x)|\,dx \le 2\varepsilon + o(1)\end{aligned}$$

where for the $o(1)$ part we used the Lebesgue dominated convergence theorem.

LEMMA 6. *Let K and f be densities on R^d. If $J_n\to 0$ in probability as $n\to\infty$, then* $\lim_n h = 0$ *and* $\lim_n nh^d = \infty$.

PROOF OF LEMMA 6. Since $J_n\le 2$ for all n, $J_n\to 0$ in probability if and only if

$\lim_n E(J_n) = 0$. Define g_h as in Lemma 1. Then

$$E(J_n) = E\left(\int |f_n(x) - f(x)|\,dx\right) \ge \int |E(f_n(x)) - f(x)|\,dx = \int |g_h(x) - f(x)|\,dx.$$

Apply Lemma 5, and conclude that $\lim_n h = 0$. This will be assumed for the remainder of the proof. For the second part, we note that by Lemma 1, $\lim_n E(\int |f_n(x) - g_h(x)|\,dx) = 0$. Let M be a large number, and let $K^*(x)$ be defined as $K(x)I_{K(x)\le M}$. Define f_n^* and g_h^* as f_n, g_h with K^* instead of K. By (1),

$$\begin{aligned}
&\int |f_n(x) - g_h(x)|\,dx \\
&\ge \int |f_n^*(x) - g_h^*(x)|\,dx - \int |f_n(x) - f_n^*(x)|\,dx - \int |g_h(x) - g_h^*(x)|\,dx \qquad (7)\\
&= \int |f_n^*(x) - g_h^*(x)|\,dx - 2\int |K(x) - K^*(x)|\,dx.
\end{aligned}$$

Let us introduce some more notation: L is another large number, A is the event that no X_i, $1 \le i \le n$, belongs to S_{xhL}, $K' = K^* I_{S_{0L}}$, $K'' = K^* - K'$, and f_n' and f_n'' are defined as f_n after replacement of K by K' and K'' in the definition. Clearly,

$$\begin{aligned}
\int E(|f_n^*(x) - g_h^*(x)|\,dx) &\ge \int E(|f_n^*(x) - g_h^*(x)|\,I_A)\,dx \\
&\ge \int g_h^*(x)P(A)\,dx - \int E(f_n''(x)I_A)\,dx = U_n - V_n. \qquad (8)
\end{aligned}$$

We will need the following facts, all corollaries of Lemma 2 (see also Devroye and Wagner, 1979): for bounded K^* with compact support, $g_h^*(x) \to f(x)\int K^*(x)\,dx$, almost all x, and $\mu(S_{y+hzhL})/\lambda(S_{y+hzhL}) \to f(y)$ for all $z \in R^d$ and almost all $y \in R^d$. Let C be the volume of S_{01}, and assume that $\lim_n nh^d = r \in [0, \infty)$. By Fatou's Lemma, we have

$$\begin{aligned}
\liminf_n U_n &\ge \int \liminf_n g_h'(x) \liminf_n P(A)\,dx \\
&= \int f(x) \liminf_n \{1 - \mu(S_{xhL})\}^n\,dx \int K'(z)\,dz \\
&\ge \int f(x)\exp(-\limsup_n [n\mu(S_{xhL})/\{1 - \mu(S_{xhL})\}])\,dx \int K'(z)\,dz \qquad (9)\\
&= \int f(x)\exp\{-rCL^d f(x)\}\,dx \int_{S_{0L}} K^*(z)\,dz.
\end{aligned}$$

Also,

$$\begin{aligned}
V_n &\le \int E\left\{\frac{1}{n}\textstyle\sum_{i=1}^n h^{-d}K''((x - X_i)/h)I_A\right\}dx \\
&= \int\int h^{-d}K''((x - y)/h)\,I_{y\notin S_{xhL}} f(y)\,dy\{1 - \mu(S_{xhL})\}^{n-1}\,dx \\
&= \int f(y)\int_{x\notin S_{yhL}} h^{-d}K''((x - y)/h)\{1 - \mu(S_{xhL})\}^{n-1}\,dx\,dy \qquad (10)\\
&\le \int f(y)\int_{z\notin S_{0L}} K''(z)\exp\{-(n - 1)\mu(S_{y+hzhL})\}\,dz\,dy.
\end{aligned}$$

The integrand of the inner integral of (10) is bounded by an integrable function, K''. Thus, by the Lebesgue dominated convergence theorem and an earlier remark, we can conclude that

$$\begin{aligned}\limsup_n V_n &\le \int f(y) \int_{z \notin S_{0L}} K^*(z) \exp\{-rCL^d f(y)\}\, dz\, dy \\ &= \int f(y) \exp\{-rCL^d f(y)\}\, dy \int_{z \notin S_{0L}} K^*(z)\, dz.\end{aligned} \tag{11}$$

Combining (7), (8), (9) and (11) gives

$$\begin{aligned}\liminf_n \int E(|f_n(x) - g_h(x)|)\, dx + 2 \int |K(x) - K^*(x)|\, dx \\ \ge \int f(x) \exp\{-rCL^d f(x)\}\, dx \left\{ 2 \int_{S_{0L}} K^*(z)\, dz - 1 \right\}.\end{aligned} \tag{12}$$

Keeping L fixed, and letting M grow large shows that the right-hand-side of (12) is ≤ 0, with K instead of K^* in the last integral. Now, choose any finite L for which $\int_{S_{0L}} K(z)\, dz > \frac{1}{2}$. Then, (12) can only be 0 when $r = \infty$, and this is a contradiction. Thus, no subsequence of nh^d can tend to a finite limit r, and therefore, we must have $\lim_n nh^d = \infty$.

3. Discrimination. We would like to point out one important application of Theorem 1. In the discrimination problem, we are given a sequence $(X_1, Y_1), \cdots, (X_n, Y_n)$ of independent $R^d \times \{1, \cdots, M\}$-valued random vectors distributed as (X, Y) but independent of (X, Y). We construct an estimate Y from X and the data sequence, say, $Y = g_n(X)$. The probability of error for the given estimate and data sequence is $L_n = P\{g_n(X) \ne Y \mid X_1, Y_1, \cdots, X_n, Y_n\}$, and this is always at least equal to the Bayes probability of error

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

If X has a density f, and if we construct the density estimates

$$f_{ni}(x) = (nh^d)^{-1} \textstyle\sum_{j=1}^n K((x - X_j)/h) I_{Y_j = i}, \quad 1 \le i \le M, \tag{13}$$

and if we define $g_n(x)$ as the first integer i for which $f_{ni}(x) = \max_{1 \le k \le M} f_{nk}(x)$, then how is L_n related to L^*? In other words, in what senses does L_n converge to L^*? The simple rule mentioned here can be found under the name "potential function method" in the Russian literature (see e.g. Bashkirov, Braverman and Muchnik, 1964). Its properties were subsequently studied by Van Ryzin (1966), Rejtö and Révész (1973), Glick (1972, 1976), Greblicki (1978), Devroye and Wagner (1980a, 1980b) and Spiegelman and Sacks (1980). In this note, we can offer the following result:

THEOREM 2. *Let K be an absolutely integrable function with positive integral over R^d, and let X have a density f. Then the discrimination rule defined by* (13) *satisfies*

$$\textstyle\sum_{n=1}^\infty n^q P(L_n - L^* > \varepsilon) < \infty, \quad \text{all} \quad q, \varepsilon > 0,$$

whenever

$$\lim_n h = 0, \quad \text{and} \quad \lim_n nh^d = \infty.$$

REMARK 2. Theorem 2 contains all previously known consistency results for the discrimination rule (13) that are based on the assumption that X has a density f. With additional conditions on K (i.e., $c_1 I_{S_{0r_1}} \ge K \ge c_2 I_{S_{0r_2}}$ for some $c_1, c_2, r_1, r_2 > 0$), we know that

$L_n \to L^*$ in probability for all distributions of (X, Y) (Devroye and Wagner, 1980; Spiegelman and Sacks, 1980). If we also ask that $r_1 = r_2$ and $nh^d/\log n \to \infty$, then $L_n \to L^*$ almost surely for all distributions of (X, Y). From our Theorem, it is clear that the condition $nh^d/\log n \to \infty$ is not needed whenever X has a density.

PROOF OF THEOREM 2. We introduce some new notation: $p_i = P(Y = i)$, $p_{ni} = (1/n) \sum_{j=1}^{n} I_{Y_j = i}$, f_i is the density of X given that $Y = i$, and $f_{n0} = \sum_{i=1}^{M} f_{ni}$. Then, by (12) of Devroye and Wagner (1980b), and defining 0/0 by 0,

$$\begin{aligned} L_n - L^* &\le \textstyle\sum_{i=1}^{M} \displaystyle\int \left| \frac{f_{ni}(x)}{f_{n0}(x)} - \frac{p_i f_i(x)}{f(x)} \right| f(x)\, dx \\ &\le \textstyle\sum_{i=1}^{M} \displaystyle\int | p_i f_i(x) - f_{ni}(x) |\, dx + \textstyle\sum_{i=1}^{M} \displaystyle\int f_{ni}(x) \left| \frac{f(x)}{f_{n0}(x)} - 1 \right| dx \\ &\le \textstyle\sum_{i=1}^{M} p_{ni} \displaystyle\int \left| f_i(x) - \frac{f_{ni}(x)}{p_{ni}} \right| dx + \int | f(x) - f_{n0}(x) |\, dx + \textstyle\sum_{i=1}^{M} | p_i - p_{ni} | \\ &\le 2 \textstyle\sum_{i=1}^{M} p_i \displaystyle\int \left| f_i(x) - \frac{f_{ni}(x)}{p_{ni}} \right| dx + \textstyle\sum_{i=1}^{M} | p_i - p_{ni} |. \end{aligned}$$

Let us look at $i = 1$ only. By Hoeffding's inequality (Hoeffding, 1963), $P(|p_1 - p_{n1}| > \varepsilon) \le 2 \exp(-2n\varepsilon^2)$, all $\varepsilon > 0$. Assume that $p_1 > 0$, and let $N = np_{n1}$. Note next that $E\{f_{n1}(x)/p_{n1} \mid N\} = g_h(x)$, which is defined as in Lemma 1 when f is replaced by f_1. Thus,

$$\begin{aligned} &\int \left| f_1(x) - \frac{f_{n1}(x)}{p_{n1}} \right| dx \\ (14) \qquad &\quad \le \int | f_1(x) - g_h(x) |\, dx I_{N>0} + \int \left| g_h(x) - \frac{f_{n1}(x)}{p_{n1}} \right| dx I_{N>0} + 2 I_{N=0}. \end{aligned}$$

The first term on the right-hand-side of the inequality tends to 0 as $h \to 0$ by Lemma 1. Conditional on N, the second term is distributed as $\int | E\{f_N(x)\} - f_N(x) |\, dx I_{N>0}$, where

$$f_N(x) = (Nh^d)^{-1} \textstyle\sum_{i=1}^{N} K((x - X_i)/h)$$

and $X_1, \cdots, X_N$ are independent random vectors with common density f_1. In the proof of Theorem 1, we have seen that for every $\varepsilon > 0$ there exist positive constants c_i only depending upon ε, K and f_1 such that $P(\int | E(f_N(x)) - f_N(x) |\, dx > \varepsilon \mid N) \le c_1/N^q$, valid when $(c_2/h + 1)^d < c_3 N$. Thus

$$P\left(\int \left| g_h(x) - \frac{f_{n1}(x)}{p_{n1}} \right| dx\, I_{N>0} > \varepsilon \right) \le P\left(N < \frac{np_1}{2} \right) + c_1 \left(\frac{np_1}{2} \right)^{-q},$$

valid when $(c_2/h + 1)^d < \frac{1}{2} np_1 c_3$.

Since $nh^d \to \infty$, the last inequality is valid for all n large enough. The term $P(N < np_1/2)$ does not exceed $\exp(-np_1^2/2)$ by Hoeffding's inequality, and the last term of (14) is treated similarly. Theorem 2 now follows by the arbitrariness of ε and q.

Acknowledgments. The author wishes to thank Clark Penrod and Charles Baker for their constant help, and Laszlo Györfi for pointing out a crucial improvement.

REFERENCES

ABOU-JAOUDE, S. (1976a). Sur une condition nécessaire et suffisante de L_1-convergence presque complète de l'estimateur de la partition fixe pour une densité. *C. R. Acad. Sci. Paris, Sér. A* **283** 1107–1110.

ABOU-JAOUDE, S. (1976b). Sur la convergence L_1 et L_∞ de l'estimateur de la partition aléatoire pour une densité. *Ann. Inst. Henri Poincaré* **12** 299–317.

ABOU-JAOUDE, S. (1976c). Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. *Ann. Inst. Henri Poincaré* **12** 213–231.

BASHKIROV, O. A., BRAVERMAN, E. M. and MUCHNIK, I. E. (1964). Potential function algorithms for pattern recognition learning machines. *Automat. Remote Control* **25** 692–695.

DEHEUVELS, P. (1974). Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. *C. R. Acad. Sci. Paris, Sér. A* **278** 1217–1220.

DEVROYE, L. (1979). On the pointwise and the integral convergence of recursive kernel estimates of probability densities. *Utilitas Math.* **15** 113–128.

DEVROYE, L. and WAGNER, T. J. (1979). The $L1$ convergence of kernel density estimates. *Ann. Statist.* **7** 1136–1139.

DEVROYE, L. and WAGNER, T. J. (1980a). On the $L1$ convergence of kernel estimators of regression functions 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.

GLICK, N. (1972). Sample-based classification procedures derived from density estimators. *J. Amer. Statist. Assoc.* **67** 116–122.

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

GLICK, N. (1976). Sample-based classification procedures related to empiric distributions. *IEEE Trans. Inform. Theory* **IT-22** 454–461.

GREBLICKI, W. (1978). Asymptotically optimal procedures with density estimates. *IEEE Trans. Inform. Theory* **IT-24** 250–251.

HOEFFDING, W. (1963). Probability inequalities for the sum of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KANTOROVICH, L. V. and AKILOV, G. P. (1964). *Functional Analysis in Normed Spaces.* Pergamon Press, Oxford, England.

PARZEN, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33** 1065–1076.

REJTÖ, L. and RÉVÉSZ, P. (1973). Density estimation and pattern classification. *Problems of Control and Information Theory* **2** 67–80.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

SCHEFFÉ, H. (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* **18** 434–458.

SLUD, E. V. (1977). Distribution inequalities for the binomial law. *Ann. Probability* **5** 404–412.

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

STEIN, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions.* Princeton University Press, Princeton, New Jersey.

VAN RYZIN, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhya Ser. A* **28** 161–170.

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