The Annals of Statistics
1984, Vol. 12, No. 4, 1231–1249

THE CONSISTENCY OF AUTOMATIC KERNEL DENSITY ESTIMATES[1]

BY LUC DEVROYE AND CLARK S. PENROD

McGill University and The University of Texas at Austin

We consider the Parzen-Rosenblatt kernel density estimate on $\mathbb{R}^d$ with data-dependent smoothing factor. Sufficient conditions on the asymptotic behavior of the smoothing factor are given under which the estimate is pointwise consistent almost everywhere for all densities f to be estimated. When the smoothing factor is a function only of the sample size n, it is shown that these conditions are also necessary, a generalization of results by Deheuvels. The consistency of various automatic kernel density estimates is a simple consequence of these theorems.

1. Introduction. The recent trend towards automatization of the kernel density estimate has led to the development of many estimates that are not known to be consistent. In this paper our primary goal is to give a consistency theorem of sufficient generality for deriving the consistency of most automatic kernel density estimates.

The *kernel estimate* on R^d is

$$f_n(x) = (nh_n^d)^{-1} \sum_{i=1}^n K((x - X_i)/h_n) \tag{1}$$

where $X_1, \cdots, X_n$ is an independent sample drawn from a density f on R^d, K is a given density (kernel), and h_n is a positive number depending upon n only (the *smoothing factor*, or *window width*) (Parzen, 1962, Rosenblatt, 1956). In an *automatic kernel estimate*, h_n is a measurable function of n, $X_1, \cdots, X_n$. The function h_n does not depend upon x however, since this would in general lead to an estimate f_n that is no longer a density on R^d. Ideally, h_n does not depend upon parameters that have to be chosen by the user. In Section 2 we will give several examples of automatic kernel estimates. In this section, we present our main results, all based on the behavior of

$$D_n(x) = \sup_{H_n} |f_{nh}(x) - f(x)| \tag{2}$$

where $f_{nh}(x) = (nh^d)^{-1} \sum_{i=1}^n K((x - X_i)/h)$, $h > 0$, and the supremum is taken over all values of h in an interval $H_n = [h_n', h_n'']$, where $0 < h_n' \leq h_n'' < \infty$ only depend upon n.

Received October 1982; revised December 1983.

[1] Research of both authors was sponsored by the Office of Naval Research under Contract N00014-81-K-0145.

AMS 1980 *subject classifications.* Primary 60F15; secondary 62G05.

Key words and phrases. Nonparametric density estimation, kernel density estimates, automatic kernel estimates, consistency.

THEOREM 1. *Let K be a bounded Riemann integrable density with compact support, and let $h_n'' = o(1)$.*

A. *If $\lim_{n\to\infty} nh_n'^d = \infty$, then $D_n(x) \to 0$ in probability, almost all x.*

B. *If h_n' varies regularly with coefficient $r \leq 0$ (i.e. $h_{nt}'/h_n' \to t^r$, all $t > 0$), and $\lim_{n\to\infty} nh_n'^d/\log\log n = \infty$, then $D_n(x) \to 0$ almost surely, almost all x.*

C. *If $\lim_{n\to\infty} nh_n'^d/\log n = \infty$, then $D_n(x) \to 0$ completely, almost all x (i.e. $\sum_{n=1}^{\infty} n^q P(D_n(x) > \varepsilon) < \infty$, all q, $\varepsilon > 0$).*

The main theorem of this paper can be deduced without much effort from Theorem 1:

THEOREM 2. *Let K be a bounded Riemann integrable density with compact support, and let f_n be an automatic kernel estimate with smoothing factor $h_n = h_n(X_1, \cdots, X_n)$. Let f be a fixed but arbitrary density on R^d.*

A. *If $h_n \to 0$ and $nh_n^d \to \infty$ in probability, then $f_n(x) \to f(x)$ in probability, almost all x, and $\int |f_n(x) - f(x)|\, dx \to 0$ in probability.*

B. *If $h_n \to 0$ and $nh_n^d/\log\log n \to \infty$ almost surely, then $f_n(x) \to f(x)$ almost surely, almost all x, and $\int |f_n(x) - f(x)|\, dx \to 0$ almost surely.*

C. *If $h_n \to 0$ and $nh_n^d/\log n \to \infty$ completely, then $f_n(x) \to f(x)$ completely, almost all x.*

The proofs are given in Section 3. We point out that there are no conditions whatsoever on the density f, and that the conditions on K are weak enough to cover all interesting kernels except possibly the normal kernel. The qualification "almost all x" refers to all Lebesgue points of f. The conditions on h_n can essentially not be improved. To see this, we take h_n as a function of n only, and note that the conditions in A, B and C are necessary. The necessity of these conditions (and in particular of B) was first proved by Deheuvels (1974) under various regularity conditions on K, f and h_n. For the sake of completeness, we give here a generalization of Deheuvels' theorem, stripped of most regularity conditions, together with a different, shorter proof.

DEFINITION. A sequence of positive numbers a_n is called *semimonotone* if there exists a constant $c > 0$ such that $a_{n+m} \geq ca_n$ for all m, $n \geq 1$. (Note that this implies that either $\liminf_{n\to\infty} a_n = \infty$ or $\sup_n a_n < \infty$.)

THEOREM 3. *Let f_n be the kernel estimate* (1) *defined on R^d, and let K be a bounded density with compact support.*

1. [*Weak version.*] *The following statements are equivalent*:

A. *$f_n(x) \to f(x)$ in probability, almost all x, some f.*
B. *$f_n(x) \to f(x)$ in probability, almost all x, all f.*
C. *$\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d = \infty$.*

D. $\int |f_n(x) - f(x)|\, dx \to 0$ *in probability, some* f.
E. $\int |f_n(x) - f(x)|\, dx \to 0$ *completely, all* f.

2. [*Strong version.*] *Let* K *also be Riemann integrable, and let the sequence* $nh_n^d/\log\log n$ *be semimonotone. Then the following are equivalent*:

A. $f_n(x) \to f(x)$ *almost surely, almost all* x, *some* f.
B. $f_n(x) \to f(x)$ *almost surely, almost all* x, *all* f.
C. $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d/\log\log n \to \infty$.

The Riemann integrability of K *is only used in the proof of* $C \Rightarrow B$. *The semimonotonicity of* $nh_n^d/\log\log n$ *is only used in the proof of* $A \Rightarrow C$.

3. [*Complete version.*] *Let* $nh_n^d/\log n$ *be semimonotone. Then the following are equivalent*:

A. $f_n(x) \to f(x)$ *completely, almost all* x, *some* f.
B. $f_n(x) \to f(x)$ *completely, almost all* x, *all* f.
C. $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d/\log n = \infty$.

The semimonotonicity of $nh_n^d/\log n$ *is only used in the proof of* $A \Rightarrow C$.

The equivalence of 1C, 1D and 1E is due to Devroye (1983). Another by-product of Theorem 3 is that the kernel estimate is pointwise convergent for almost all x (in one of the senses given) for *all* f simultaneously, or for no f. There is no intermediate situation.

In 1975, Wagner proved a theorem for the case $d = 1$ that is contained in Theorem 2. He showed that when K has bounded variation, $\lim_{|x|\to\infty} |x|\, K(x) = 0$, and $h_n \to 0$ and $nh_n^2 \to \infty$ in probability, then $f_n(x) \to f(x)$ in probability at continuity points of f. He remarked that "in probability" can be replaced by "almost surely" if also $h_n \to 0$ and $n^a h_n^2 \to \infty$ almost surely for some $0 < a < 1$. By quick inspection of his proof, we see that the last condition can be replaced by $nh_n^2/\log\log n \to \infty$ almost surely (use Kiefer, 1961, Theorem 2), but that no further improvements can be made without major changes in the proof. Because of its relevance in this paper, we reproduce here a uniform convergence theorem similar to Theorem 2:

THEOREM 4. (Devroye and Wagner, 1980). *Let* K *be a bounded Riemann integrable density with compact support, and let* f *be a uniformly continuous density on* R^d. *If* $h_n \to 0$ *and* $nh_n^{2d}/\log n \to \infty$ *almost surely, then the automatic kernel density estimate defined by* K *and* h_n *satisfies*

$$\sup_x |f_n(x) - f(x)| \to 0 \quad \textit{almost surely.}$$

The thrust of this paper is the replacement of Wagner's suboptimal conditions on h_n for pointwise convergence by optimal ones. His argument, based upon tight bounds for the empirical distribution function, is replaced by a finer argument. In Section 2, we apply Theorem 2 to several automatic kernel estimates. Because of the generality of the theorem, we can only discuss the consistency of these

estimates and not the rate of convergence to 0 of some global measure of deviation such as $\int |f_n(x) - f(x)|\, dx$.

2. Applications. The fundamental result underlying most choices of h_n is due to Rosenblatt (1956, 1971): when $d = 1$, K is a bounded symmetric density with $\int x^2K(x)\, dx < \infty$, f is bounded and has two continuous derivatives, and $f, f'' \in L_2$, then the kernel estimate (1) satisfies:

$$E\left(\int (f_n(x) - f(x))^2\, dx \right) \sim (nh_n)^{-1} \int K^2(x)\, dx + \frac{1}{4} h_n^4 \left(\int x^2K(x)\, dx \right)^2 \int f''^2(x)\, dx \tag{3}$$

when $h_n \to 0$ and $nh_n \to \infty$. From (3), it appears that the best value for h_n is given by

$$h_n = \left[A \Big/ \left(n \int f''^2(x)\, dx \right) \right]^{1/5} \tag{4}$$

where $A = \int K^2(x)\, dx / (\int x^2K(x)\, dx)^2$ is a factor depending upon K only. Unfortunately, (4) depends on the unknown density f. There have been many attempts at replacing (4) by a data dependent estimate. To cite a few:

1. *The semi-parametric estimate.* The statistician assumes that f can be roughly estimated by some density in a family of densities g_θ parametrized by θ. The parameter vector θ is estimated from $X_1, \cdots, X_n$ by standard parametric techniques (maximum likelihood, method of moments, etc.). The unknown value $\int f''^2(x)\, dx$ in (4) is then replaced by the known value $\int g''^2_{\hat\theta}(x)\, dx$ where $\hat\theta$ is the estimate of θ. This approach allows us to use a priori information about f. Its first in-depth development is due to Deheuvels (1977) who in particular considered the case of a normal parametric family in R^1 with mean μ and variance σ^2. For the normal (μ, σ^2) density g, $\int g''^2(x)\, dx = 3/(8\sqrt{\pi}\sigma^5)$, so that our estimate of (4) for nearly-normal densities f becomes

$$h_n = [A\ 8\ \sqrt{\pi}/(3n)]^{1/5}\sigma_n \tag{5}$$

where σ_n^2 is the sample variance. See also Deheuvels and Hominal (1980) for further discussions.

Invoking the strong law of large numbers, we deduce without further work that for all densities f for which $\int x^2f(x)\, dx < \infty$, and for all kernels of Theorem 2, the automatic kernel estimate (1)(5) satisfies $f_n(x) \to f(x)$ almost surely, almost all x, and $\int |f_n(x) - f(x)|\, dx \to 0$ almost surely. Estimate (5) can be made more robust by using sample quantile based estimates for σ.

2. *Iterative estimation.* Scott, Tapia and Thompson (1977) give a nonparametric estimate $C_n(h)$ of $\int f''^2(x)\, dx$ for fixed smoothing factor h. With this

value of $C_n(h)$, a new value of h can be obtained by setting

$$h_{\text{new}} = [A/(nC_n(h))]^{1/5},$$

and this process can be repeated. The authors report that their algorithm does not always converge. Theorem 2 is thus not directly applicable here. See also Scott and Factor (1981) for experimental results.

3. *Direct nonparametric estimation.* In the hope of achieving the optimal MISE rate as determined by (3)(4), Nadaraya (1974) proposed the following scheme, valid for all even bounded K in R^1 for which $\int x^2K(x)\,dx < \infty$, K'' exists and is continuous, $|K'(x)| + K(x) \to 0$ as $|x| \to \infty$ and $\int x^2|K''(x)|\,dx < \infty$:

> Choose any sequence $t_n \to 0$, $t_n n^{1/50} \to \infty$, and any sequence $b_n \to 0$ such that $nb_n \geq c > 0$ for some constant c. Compute $C_n(t_n) = \int f''^2_{nt_n}(x)\,dx$ where f_{nt_n} is defined as in (1) with smoothing factor t_n. Estimate (4) by $h_n^* = [A/(n(b_n + C_n(t_n)))]^{1/5}$.

Nadaraya has shown that $E(|h_n^* - h_n|) = O(n^{-2/5})$ when f is twice continuously differentiable with $f, f'' \in L_2$. Thus, $h_n^*/h_n \to 1$ in probability. By Theorem 2, we note that $f_n(x) \to f(x)$ in probability, almost all x, and $\int |f_n(x) - f(x)|\,dx \to 0$ in probability. Of course, this result is overshadowed by the finer result of Nadaraya's that estimate (1) with h_n^* has MISE asymptotic to the optimal MISE (3)(4) under some additional conditions on K. See also Woodroofe (1970), Bretagnolle and Huber (1979) and Scott and Factor (1981).

For an excellent discussion of the MISE of density estimates, see Tapia and Thompson (1978). The heaviness of the tail of f has little influence on (3). Yet, because large tails are allowed within the class of densities for which (3) is valid, we are faced with the curious phenomenon that within this class of densities, any slow rate of convergence to 0 for $E(\int |f_n(x) - f(x)|\,dx)$ can be achieved, for *any* density estimate (Devroye, 1983). Thus, we should perhaps look for a smoothing factor that minimizes the average L_1 error. Rosenblatt (1979) showed that under the conditions for (3) and the additional condition that x^2f, x^2f' and x^2f'' are absolutely bounded,

$$E\left(\int |f_n(x) - f(x)|\,dx\right) \leq \left[\int K^2(x)\,\frac{dx}{nh_n}\right]^{1/2} \int \sqrt{f(x)}\,dx + \frac{1}{2}h_n^2 \int x^2K(x)\,dx \int |f''(x)|\,dx + o(h_n^2 + (nh_n)^{-1/2}).$$

This suggests the choice

$$h_n = \left[A\left(\int \sqrt{f(x)}\,dx\right)^2 \Big/ \left(4n\left(\int |f''(x)|\,dx\right)^2\right)\right]^{1/5}$$

where A is as in (4). Here we notice a dependence upon $\int \sqrt{f(x)}\,dx$, which is a

measure of the heaviness of the tail of f. We are not aware of any automatic kernel estimates in which this new value of h_n is estimated from the data. For more research along these lines, and for a wealth of inequalities linking $E(\int |f_n(x) - f(x)|^p \, dx)$ $(p \geq 1)$ and functionals of f and its derivatives, we refer the reader to Bretagnolle and Huber (1979).

Expression (3) is only valid for densities with continuous second derivative in L_2. This requirement is often unrealistic. Choosing h_n by *maximum likelihood principles* effectively avoids this drawback. The ground-breaking work in this area is due to Duin (1976, paper submitted in 1973) and Habbema et al. (1974). They suggest that h be chosen so as to maximize the likelihood

$$L(h) = \prod_{i=1}^{n} f_{ni}^{*}(X_i) \tag{6}$$

where

$$f_{ni}^{*}(x) = (nh^d)^{-1} \sum_{j=1, j\neq i}^{n} K((x - X_j)/h).$$

This cross-validated kernel density estimate seems to work well in most, but not all, situations. For several years, it was not even known whether this estimate was consistent or not. Schuster and Gregory (1981) proved that the solution h_n maximizing (6) in the case $d = 1$ satisfies $h_n \not\to 0$ in probability when $\lim_{x\downarrow-\infty} F(x)/f(x) > 0$, and $h_n \to \infty$ in probability when $\lim_{x\downarrow-\infty} F(x)/f(x) = \infty$, where F is the distribution function corresponding to f. In the latter case, we have the disturbing result that $\sup_x f_n(x) \to 0$ in probability. This happens, for example, when $f(x) \sim c/|x|^a$ as $|x| \to \infty$ for some $a > 1$. Chow, Geman and Wu (1983) showed that if we choose h_n such that $L(h_n) \geq a \sup_{h>0} L(h)$ for some fixed a in (0, 1), then $h_n \to 0$ almost surely, and $nh_n/\log n < \varepsilon$ finitely often almost surely for some $\varepsilon > 0$ (see their Lemma 1.1) under the following assumptions on f and K: f is bounded and has compact support; K is bounded, has compact support, is nondecreasing on $(-\infty, 0]$, nonincreasing on $[0, \infty)$, and stays bounded away from 0 on $[-\delta, \delta]$ for some $\delta > 0$. By our Theorem 2, we conclude that under these conditions, $f_n(x) \to f(x)$ almost surely, almost all x, and $\int |f_n(x) - f(x)| \, dx \to 0$ almost surely. Theorem 1 of Chow, Geman and Wu follows from this, because we can choose h_n within the allowed range of values so as to maximize the L_1 error; thus,

$$\sup_{h_n: L(h_n) \geq a \sup_{h>0} L(h)} \int |f_n(x) - f(x)| \, dx \to 0 \quad \text{almost surely.}$$

For other details, see also Geman (1981). We stress the fact that f is only required to be bounded and to have compact support. On the negative side, Hall (1982a, 1982b) gives evidence that the cross-validation method yields h_n of magnitude $n^{-1/3}$ when f is concave on [0, 1]: these are necessarily suboptimal in certain cases.

Schuster and Gregory (1978) determine h_n by maximizing

$$\prod_{i=1}^{n/2} f_{n/2}(X_i)$$

where $f_{n/2}(x) = (2/n) \sum_{j=n/2+1}^{n} h^{-d} K((x - X_j)/h)$. The sample is artificially cut

into two parts in order to preserve independence between h_n and half of the original sample. To correct for the nonconsistency of the cross-validated kernel estimate in the case of medium- or long-tailed f, Schuster and Gregory (1981) modify the cross-validation estimate slightly by including the variable kernel estimate (Breiman, Meisel and Purcell, 1977) in the class of densities over which the maximization is carried out. Strictly speaking, this estimate is no longer an automatic kernel estimate (as defined in the introduction).

We found other interesting ideas in the literature, e.g. Silverman (1978) and Wagner (1975). For example, Wagner (1975) computes $D_{n1}, \cdots, D_{nn}$, the distances between $X_1, \cdots, X_n$ and their respective kth nearest neighbors where $k = \lfloor n^a \rfloor$, $0 < a < 1$. He suggests many schemes for determining h_n such as (i) h_n is chosen at random from $D_{n1}, \cdots, D_{nn}$; (ii) $h_n = \sum D_{ni}/n$; (iii) $h_n = \max(D_{ni})$; (iv) $h_n = \min(D_{ni})$. The number of possibilities is nearly unlimited. For (i) he has shown that for *all* f, $h_n \to 0$ and $n^b h_n^{2d} \to \infty$ almost surely for all $b > 1 - a$. Thus, by our Theorem 2, for the kernels considered there and for *all* f, $f_n(x) \to f(x)$ almost surely for almost all x, and $\int |f_n(x) - f(x)|\, dx \to 0$ almost surely.

Finally, we note that there are many authors who take

$$h_n = h_n(x, X_1, \cdots, X_n).$$

Such estimates are disregarded in this paper although they may be good pointwise estimates. See for example Sacks and Ylvisaker (1981) and Krieger and Pickands (1981).

3. Proofs. Throughout this section, K is a density bounded by K^* and vanishes outside $[-c, c]^d$. h is sometimes a real number and sometimes a random variable (this will be clear from the context). Finally, we will often write f_n and $E(f_n)$ instead of f_{nh} and $E(f_{nh})$. In particular, when h is random, $E(f_{nh})$ and $E(f_n)$ are thus both functions of h, and should be thought of as convolutions $\int f(x - y)h^{-d}K(y/h)\, dy$.

LEMMA 1. (Lebesgue density theorem.) *Let f be a density on R^d. There exists a set $B \subseteq R^d$ such that almost all x belong to B, and for all bounded sets A of positive Lebesgue measure $\lambda(A) > 0$:*

$$\lim_{h \downarrow 0} \int_{x+hA} f(y)\, dy \Big/ \int_{x+hA} dy = f(x), \quad x \in B.$$

PROOF. See Wheeden and Zygmund (1977).

LEMMA 2. (Convergence of the bias.) *Let h_n'' be a sequence of positive numbers tending to 0 as $n \to \infty$. For all densities f, and for the kernel estimate* (1) *with smoothing factor h,*

$$\lim_{n \to \infty} \sup_{0 < h \le h_n''} |E(f_n(x)) - f(x)| = 0, \quad \textit{almost all } x.$$

PROOF.

$$\sup_{0<h\le h_n''} |E(f_n(x)) - f(x)|$$

$$\le \sup_{0<h\le h_n''} \int_{S_{0,ch}} |f(x-y) - f(x)| h^{-d} K\left(\frac{y}{h}\right) dy$$

$$\le K^* \sup_{h\le h_n''} \left(\int_{S_{0,ch}} |f(x-y) - f(x)| \, dy \Big/ \int_{S_{0,ch}} dy \right) \cdot \int_{S_{0,c}} dy$$

$$\to 0, \quad \text{almost all } x$$

by Lemma 1. Here $S_{x,r}$ denotes the closed sphere of radius r centered at x.

LEMMA 3. *For every nonnegative Riemann integrable function K bounded by K^* on $[0, 1]^d$, and for every $\varepsilon > 0$, there exists an integer N and nonnegative numbers $a_i \in [0, K^*]$, $1 \le i \le N^d$, such that the function*

$$K_1(x) = \sum_{i=1}^{N^d} a_i I_{A_i}(x), \quad x \in [0, 1]^d,$$

in which the A_i's are the rectangles formed by the products of intervals of the form $[(j-1)/N, j/N)$, $1 \le j < N$, or $[(N-1)/N, 1]$, satisfies:

(i) $|K_1(x) - K(x)| < \varepsilon$, *all $x \notin A$ = union of some A_i's;*
(ii) $0 \le K_1(x) \le K^*$, *all x;*
(iii) $\lambda(A) < \varepsilon$.

LEMMA 4. (Fundamental inequalities for the uniform deviation.) *Let $\varepsilon > 0$ be an arbitrary number, let x be a Lebesgue point for f (i.e., $x \in B$ as defined in Lemma 1), and let h_n' and h_n'' be two positive number sequences satisfying $0 < h_n' \le h_n'' \downarrow 0$. Let f_n be the estimate (1) with smoothing factor h. Then*

$$\sup_{h_n'\le h\le h_n''} P(|f_n(x) - E(f_n(x))| \ge \varepsilon) \le 2 \exp(-bnh_n'^d)$$

where b can be taken to be $\varepsilon^2/(2K^*(f(x) + o(1) + \varepsilon))$.

If K is Riemann integrable, then also

$$P(\sup_{h_n'\le h\le h_n''} |f_n(x) - E(f_n(x))| \ge \varepsilon) \le a \exp(-bnh_n'^d)/(1 - \exp(-b'nh_n'^d))$$

for some positive constants a, b, b' not depending upon n.

PROOF. Bennett (1962) has shown that for independent identically distributed zero mean random variables Z_i with $|Z_i| \le t$, and for all $\varepsilon > 0$,

$$P\left(\left| \frac{1}{n} \sum_{i=1}^n Z_i \right| > \varepsilon \right) \le 2 \exp\left(-\frac{n}{2t} \left(\left(1 + \frac{\sigma^2}{2t\varepsilon}\right) \log\left(1 + \frac{2t\varepsilon}{\sigma^2}\right) - 1 \right) \right)$$

$$\le 2 \exp\left(-\frac{n\varepsilon^2}{2(\sigma^2 + t\varepsilon)} \right)$$

where $\sigma^2 = E(Z_1^2)$. The last inequality follows from $\log(1 + u) \ge 2u/(2 + u)$, valid for all $u > 0$.

Our first inequality follows by replacing Z_i by $h^{-d}(K((x - X_i)/h) - E(K((x - X_i)/h)))$, which is bounded in absolute value by $t = K^*/h^d$, and has variance $\sigma^2 \le K^*E(f_n(x))/h^d = K^*(f(x) + o(1))/h^d$ uniformly on $[0, h_n'']$ (by Lemma 2).

For the second inequality, we take a positive number δ (to be specified later), and $h_{ni} = h_n'(1 + \delta)^i$, $i \ge 0$. Let i_0 be such that $h_{ni_0-1} \le h_n'' < h_{ni_0}$. We have

$$
\begin{aligned}
&\sup_{h_n' \le h \le h_n''} |f_n(x) - E(f_n(x))| \\
&\qquad \le \sup_{0<i\le i_0}[|f_{nh_{ni-1}}(x) - E(f_{nh_{ni-1}}(x))| \\
&\qquad\qquad + \sup_{h_{ni-1}\le h,h'\le h_{ni}} |E(f_{nh}(x)) - E(f_{nh'}(x))| \\
&\qquad\qquad + \sup_{h_{ni-1}\le h,h'\le h_{ni}} |f_{nh}(x) - f_{nh'}(x)|] \\
&\qquad = \sup_{0<i\le i_0}[U_i + V_i + W_i].
\end{aligned} \tag{7}
$$

By the first part of Lemma 4, for $\varepsilon > 0$,

$$P(U_i \ge \varepsilon) \le 2 \exp(-nh_{ni-1}\varepsilon^2/(2K^*(f(x) + \varepsilon + o(1))))$$

where the $o(1)$ term does not depend upon i (since $0 < i \le i_0$). By Lemma 2,

$$\sup_{0<i\le i_0} V_i \le 2 \sup_{h\le h_{ni_0}} |E(f_{nh}(x)) - f(x)| \to 0. \tag{8}$$

For fixed $\varepsilon > 0$, find K_1, N, $a_1, \cdots, a_{N^d}$, and sets A_i as in Lemma 3 (after having replaced $[0, 1]^d$ by $[-c, c]^d$). The set A also keeps its meaning from Lemma 3. We introduce the notation μ and μ_n for the measure induced by f, and the empirical measure defined by $X_1, \cdots, X_n$ respectively. Also, Δ is the difference operator between sets.

Without loss of generality, we can assume that all sets A_i are strictly contained in one quadrant, such as $[0, c]^d$. We need a few geometrical facts now. Let h, h' be numbers in the interval $[h_{ni-1}, h_{ni}]$, and let A_j be fixed, e.g. $A_j = [a_1, a_1'] \times \cdots \times [a_d, a_d']$. Then, $(x + hA_j)\Delta(x + h'A_j) \subseteq (x + h_{ni}B_j)$ where B_j is a set of fixed form and dimensions determined by A_j, d and δ only. Also, $\lambda(B_j) \le 2c^d\delta$.

To prove this first geometrical fact, we need only show that $uA_j \,\Delta u'A_j \subseteq B_j$ for all $u, u' \in [1/(1 + \delta), 1]$. First, take

$$
\begin{aligned}
B_j = [a_1, a_1'] \times \cdots \times [a_d, a_d'] - \left[a_1, \frac{a_1'}{1+\delta}\right] \times \cdots \times \left[a_d, \frac{a_d'}{1+\delta}\right] \\
+ \left[\frac{a_1}{1+\delta}, a_1'\right] \times \cdots \times \left[\frac{a_d}{1+\delta}, a_d'\right] - [a_1, a_1'] \times \cdots \times [a_d, a_d'],
\end{aligned}
$$

where the $-$ operators are considered before the union operator $+$. We note that B_j is contained in $[-c, c]^d$. Also,

$$
\begin{aligned}
\lambda(B_j) &\le \left(a_1'\left(1 - \frac{1}{1+\delta}\right) + a_1\left(1 - \frac{1}{1+\delta}\right)\right)a_2'a_3' \cdots a_d' \\
&\le 2\, a_1'a_2' \cdots a_d'\delta \le 2c^d\delta.
\end{aligned}
$$

Next, let A be the set of Lemma 3, i.e. it is the union of M disjoint rectangles A_j, and let B be the set of all points contained in uB where $u \in [1/(1+\delta), 1]$. Then, $B \subseteq [-c, c]^d$, and by the previous derivation for a single rectangle, $\lambda(B) \le \lambda(A) + \sum_{j=1}^{M} \lambda(B_j) \le \lambda(A) + 2Mc^d\delta$. We can now obtain the following crucial upper bound for W_i:

$$
\begin{aligned}
W_i &\le \sup_{h_{ni-1} \le h, h' \le h_{ni}} \sum_{j=1}^{N^d} \int a_j \left| I_{x+hA_j}(y) - I_{x+h'A_j}(y) \right| \mu_n(dy)/h_{ni-1}^d \\
&\quad + 2 \sup_{h_{ni-1} \le h \le h_{ni}} \sum_{j=1}^{N^d} \int_{x+hA_j} \left| K((x-y)/h) - K_1((x-y)/h) \right| \mu_n(dy)/h_{ni-1}^d \\
&\le \sum_{j=1}^{N^d} a_j \mu_n(x + h_{ni}A_j \,\Delta\, x + h_{ni-1}A_j)/h_{ni-1}^d \\
&\quad + 2\varepsilon\mu_n(x + [-c, c]^d h_{ni})/h_{ni-1}^d + 2K^*\mu_n(x + h_{ni}B)/h_{ni-1}^d \qquad (9) \\
&\le h_{ni-1}^{-d} \left(\sum_{j=1}^{N^d} K^*\mu_n(x + h_{ni}B_j) + 2\varepsilon\mu_n(x + [-c, c]^d h_{ni}) \right. \\
&\quad \left. + 2K^*\mu_n(x + h_{ni}B) \right) \\
&= \sum_{j=1}^{N^d} W_{ij} + W_i' + W_i''
\end{aligned}
$$

where μ_n is the empirical measure for $X_1, \cdots, X_n$.

For a given $\eta > 0$, we find $\varepsilon, \delta > 0$ such that the expected value of each W_{ij} does not exceed $\eta/(3N^d)$, and the expected value of W_i' and of W_i'' does not exceed $\eta/3$. This corresponds to the requirement that

$$
\begin{aligned}
(f(x) + o(1))(1+\delta)^d K^* 2c^d\delta &< \eta/(3N^d); \\
(f(x) + o(1))(1+\delta)^d 2\varepsilon(2c)^d &< \eta/3; \\
(f(x) + o(1))(1+\delta)^d 2K^*(\varepsilon + 2Mc^d\delta) &< \eta/3.
\end{aligned}
$$

Once again, the $o(1)$ terms do not depend upon i, so that all three inequalities can be satisfied for all n large enough, uniformly in i. A small technical note is in order here: it seems necessary to choose ε first under the assumption that δ does not exceed 2. This fixes N and M, so that in a second step we can choose δ.

For each i, we have by simple bounding techniques,

$$
\begin{aligned}
P(W_i > 2\eta) &\le \sum_{j=1}^{N^d} P(W_{ij} - E(W_{ij}) > \eta/(3N^d)) \\
&\quad + P(W_i' - E(W_i') > \eta/3) + P(W_i'' - E(W_i'') > \eta/3). \qquad (10)
\end{aligned}
$$

Uniformly in i and j, we know that for all $n \ge n_0$, all expected values are smaller than $\eta/3$. Also, each of the W_{ij}'s, W_i''s and W_i'''s can be written as $(1/n)\sum_{m=1}^{n} Y_m$ where the Y_m's are independent bounded nonnegative random variables with absolute value not exceeding r/h_{ni}^d where

$$
r = \max(2K^*, 2\varepsilon)(1+\delta)^d.
$$

Thus, by another application of Bennett's inequality, we see that each probability

on the right-hand side of (10) does not exceed

$$(11)\quad 2\exp\left(-n(\eta/(3N^d))^2 \Big/ \left(2\left(\frac{\eta}{3}\frac{r}{h_{ni}^d} + \frac{r}{h_{ni}^d}\cdot\frac{\eta}{3N^d}\right)\right)\right) = 2\exp(-bnh_{ni}^d)$$

by definition of b. A combination of all the bounds derived above shows us that for all n greater than some n_1,

$$(12)\quad \begin{aligned} &P(\sup_{h_n'\le h\le h_n''}|f_{nh}(x) - f^*K_h(x)| > 4\eta) \\ &\le \sum_{i=1}^{i_0} 2\exp(-snh_n'^d(1+\delta)^{d(i-1)}) + (N^d+2)2\exp(-bnh_n'^d(1+\delta)^{di}) \end{aligned}$$

where $s = \eta^2/(4K^*(f(x)+\eta))$. Here, $*$ is the convolution operator, and $K_h(x) = h^{-d}K(x/h)$ (thus, $f^*K_h(x) = E(f_{nh}(x))$). The right-hand side of (12) is again bounded from above, albeit very crudely, by

$$(13)\quad \begin{aligned} &\sum_{i=0}^{\infty} b'\exp(-b''nh_n'^d(1+\delta)^i) \\ &\le \sum_{i=0}^{\infty} b'\exp(-b''nh_n'^d(1+\delta i)) \\ &= b'\exp(-b''nh_n'^d)/(1-\exp(-b''\delta nh_n'^d)) \end{aligned}$$

for some positive constants b', b''. This concludes the proof of Lemma 4.

LEMMA 5. (A binomial tail inequality.) *Let Z be a binomial (n, p) random variable, with $p = p(n) \in (0, 1)$ varying in such a way that $np^2 = o(1)$ but $\lim_{n\to\infty} np = \infty$. Then, for constant $\delta > 0$,*

$$P(Z - np \ge \delta np) \ge \frac{1+o(1)}{(2\pi(1+\delta)^3 np)^{1/2}}\exp(-npH(\delta))$$

where $0 < H(\delta) = (1+\delta)\log(1+\delta) - \delta \to 0$ as $\delta \downarrow 0$.

PROOF. *Let k be $\lceil np(1+\delta)\rceil$.* Then,

$$\begin{aligned} P(Z - np \ge \delta np) &\ge \binom{n}{k}p^k(1-p)^{n-k} \\ &\ge \frac{(n-k+1)^k}{k!}p^k(1-p)^n e^{pk} \ge \frac{(np)^k}{k!}(1-p)^n e^{pk}\left(1 - \frac{k^2}{n}\right). \end{aligned}$$

Since $k^2 = o(n)$, $pk = o(1)$ and $k! \sim (k/e)^k\sqrt{2\pi k}$, the lower bound is

$$\begin{aligned} (1+o(1))\left(\frac{npe}{k}\right)^k e^{-np} \Big/ \sqrt{2\pi k} &= (1+o(1))e^{k-np}(1+\delta)^{-k}/\sqrt{2\pi k} \\ &\ge (1+o(1))e^{\delta np - k\log(1+\delta)}/\sqrt{2\pi k} \\ &\ge (1+o(1))e^{-npH(\delta)}/((1+\delta)\sqrt{2\pi k}), \end{aligned}$$

from which the sought inequality follows.

LEMMA 6. (Exponential lower bounds for large deviations.) *Let f be an*

arbitrary density on R^d, and let x be a Lebesgue point of f with $f(x) > 0$. Let $\varepsilon > 0$ be a constant, and let $h = h_n$ be a sequence of positive numbers satisfying $nh^{2d} = o(1)$, $\lim_{n\to\infty} nh^d = \infty$. Let $H(\cdot)$ be defined as in Lemma 5, and let $\delta = 2\varepsilon/f(x)$. Then, for the kernel estimate (1),

$$P(f_n(x) - E(f_n(x)) \geq \varepsilon)$$

$$\geq \frac{1 + o(1)}{(2\pi nh^d f(x)(2c)^d(1+\delta)^3)^{1/2}} \exp(-nh^d H(\delta)(f(x) + o(1))(2c)^d).$$

PROOF. Let Y be a random vector defined as X restricted to $x + [-c, c]^d h$. Define

$$g_n(x) = (1/n) \sum_{i=1}^n h^{-d} K((x - Y_i)/h)$$

where $Y_1, Y_2, \cdots$ are independent and distributed as Y. It is clear that $f_n(x)$ is distributed as $(N/n)g_N(x)$ where N is independent of the Y_i's, and distributed as the number of X_i's in $x + [-c, c]^d h$. Also, $E(f_n(x)) = pE(g_n(x))$ where $p = P(X_1 \in A = x + [-c, c]^d h) = (2c)^d h^d (f(x) + o(1))$. We have the following inclusion, valid for all n large enough:

$$P(f_n(x) \geq E(f_n(x)) + \varepsilon)$$

$$\geq P(N \geq np(1+\delta)) \inf_{k \geq np(1+\delta)} P\left(g_k(x) \geq E(g_k(x)) - \frac{\varepsilon}{2p(1+\delta)}\right). \tag{14}$$

Indeed, on a rich enough probability space, we can think of $f_n(x)$ as being equal to $(N/n)g_N(x)$ where $Y_1, \cdots, Y_N$ is the subset of $X_1, \cdots, X_n$ that falls in A. If $N \geq np(1+\delta)$ and $g_N(x) \geq E(g_N(x)) - \varepsilon/(2p(1+\delta))$, then

$$f_n(x) = \frac{N}{n} g_N(x) \geq \frac{np(1+\delta)}{n}\left(E(g_N(x)) - \frac{\varepsilon}{2p(1+\delta)}\right)$$

$$= p(1+\delta)\left(E(f_n(x))/p - \frac{\varepsilon}{2p(1+\delta)}\right) = E(f_n(x)) + \delta E(f_n(x)) - \frac{\varepsilon}{2}$$

$$\geq E(f_n(x)) + \varepsilon, \quad n \text{ large enough.}$$

This explains (14). By Chebyshev's inequality and the fact that $\mathrm{Var}(g_k(x)) \leq K^*(f(x) + o(1))/(kh^d p)$, we see that (14) is at least equal to

$$P(N - np \geq \delta np) \inf_{k \geq np(1+\delta)} \left(1 - \left(\frac{2p(1+\delta)}{\varepsilon}\right)^2 \mathrm{Var}(g_k(x))\right)$$

$$\geq P(N - np \geq \delta np)\left(1 - \frac{K^*(f(x) + o(1))(2p)^2(1+\delta)^2}{np(1+\delta)\varepsilon^2 h^d p}\right)$$

$$= P(N - np \geq \delta np)(1 - o(1)),$$

to which Lemma 5 can be applied since N is binomial (n, p) with $np^2 = o(1)$ and $\lim_{n\to\infty} np = \infty$. This concludes the proof of Lemma 6.

PROOF OF THEOREM 1. Parts A and C follow directly from Lemmas 2 and 4 and the trivial inequality

$$D_n(x) \le \sup_{H_n} |f_{nh}(x) - E(f_{nh}(x))| + \sup_{H_n} |E(f_{nh}(x)) - f(x)|.$$

To prove statement B, we fix a small $\delta > 0$, and define a subsequence $n_i = \lfloor (1+\delta)^i \rfloor$, $i = 0, 1, 2, \cdots$. Let

$$E_i = \sup_{n_i \le n < n_{i+1}} \sup_{h \in H_i^*} |f_{nh}(x) - E(f_{nh}(x))|$$

where $H_i^* = [\inf_{n_i \le n < n_{i+1}} h_n', \sup_{n_i \le n < n_{i+1}} h_n''] = [h_i^*, h_i^{**}]$. By Lemma 3, it is clear that

$$\sup_{n_i \le n < n_{i+1}} D_n(x) \le E_i + o(1) \quad \text{as} \quad i \to \infty, \quad \text{all Lebesgue points } x.$$

Thus, to show that $D_n(x) \to 0$ almost surely for almost all x, it suffices to show that for all Lebesgue points, all $\varepsilon > 0$ and some $\delta(\varepsilon) > 0$,

$$\sum_{i=0}^{\infty} P(E_i > \varepsilon) < \infty$$

(by the Borel-Cantelli lemma). A simple bounding argument yields for all $n_i \le n < n_{i+1}$, and fixed h, writing f^*K_h instead of $E(f_{nh}(x))$:

$$\begin{aligned}
&|f_{nh} - f^*K_h| \\
&\quad \le |f_{nh} - f_{n_ih}| + |f_{n_ih} - f^*K_h| \\
&\quad \le \left(\frac{1}{n_i} - \frac{1}{n_{i+1}}\right) \sum_{j=1}^{n_i} K_h(x - X_j) \\
&\quad\quad + \frac{1}{n_i} \sum_{j=n_i+1}^{n_{i+1}} K_h(x - X_j) + |f_{n_ih} - f^*K_h| \qquad (15) \\
&\quad \le (\delta + o(1))(f_{n_ih} + \tilde{f}_{n_{i+1}-n_ih}) + |f_{n_ih} - f^*K_h| \\
&\quad \le (1 + \delta + o(1))|f_{n_ih} - f^*K_h| + (\delta + o(1))|\tilde{f}_{n_{i+1}-n_ih} - f^*K_h| \\
&\quad\quad + (\delta + o(1))\, 2 f^*K_h.
\end{aligned}$$

Here $\tilde{f}_{nh}$ is an estimate independent of f_{nh} but distributed as f_{nh}. It is clear that E_i is not greater than the right-hand side of (15), preceded by $\sup_{h \in H_i^*}$. Since $h_i^{**} \to 0$ as $i \to \infty$, the last term in the upper bound is $2\delta f(x) + o(1)$ (Lemma 2). Now, for fixed $\varepsilon > 0$, let us choose δ so small that $\delta \le \frac{1}{2}$, $2\delta f(x) < \varepsilon/4$, and i so large that all the $o(1)$ terms in (15) do not exceed $\frac{1}{2}$ and the $o(1)$ term in $2\delta f(x) + o(1)$ does not exceed $\varepsilon/12$ (thus, the entire term does not exceed $\varepsilon/3$). For such large i, we have

$$E_i \le 2 \sup_{h \in H_i^*} |f_{n_ih} - f^*K_h| + \sup_{h \in H_i^*} |\tilde{f}_{n_{i+1}-n_ih} - f^*K_h| + \varepsilon/3. \qquad (16)$$

By Lemma 4 there exist positive constants a, a', a'', b, b', b'' such that the probabilities that the first and second terms on the right-hand side of (17) exceed

$\varepsilon/3$ do not exceed

$$(17)\qquad \begin{aligned} &a\exp(-a'n_ih_i^{*d})/(1-\exp(-a''n_ih_i^{*d}))\quad\text{and}\\ &b\exp(-b'(n_{i+1}-n_i)h_i^{*d})/(1-\exp(-b''(n_{i+1}-n_i)h_i^{*d})) \end{aligned}$$

respectively. The constants do not depend upon i.

For every $M>0$, we can find i large enough such that $j\geq i$ implies $n_jh_j^{*d}\geq M\log\log n_j\geq M\log(j\log(1+\delta))$. For $j\geq i$, the bounds in (17) are smaller than

$$(18)\qquad \frac{a+o(1)}{(j\log(1+\delta))^{Ma'}}\quad\text{and}\quad\frac{b+o(1)}{(j\log(1+\delta))^{Mb'\delta}}$$

respectively. But both expressions in (18) are summable in j when $Ma'>1$ and $Mb'\delta>1$. This shows that $\delta(\varepsilon)>0$ can be found such that

$$\sum_{i=0}^{\infty}P(E_i>\varepsilon)<\infty,\quad\text{all }\varepsilon>0,\quad\text{all Lebesgue points of }f.$$

This concludes the proof of Theorem 1.

PROOF OF THEOREM 2. Theorem 2 is based upon the inequality

$$(19)\qquad |f_n(x)-f(x)|\leq \sup_{H_n}|f_{nh}(x)-f(x)|+\infty.\,I_{[h_n\notin H_n]}$$

where I is the indicator function of an event, and $\infty.0$ is 0. The integral versions follow from the pointwise versions (statements A and B) after noting that f_n is a density on R^d for each n, and that weak and strong extensions of Scheffé's theorem are applicable (Glick, 1974, Devroye and Wagner, 1979). The proofs of the pointwise parts proceed by construction of a proper sequence $H_n=[h_n',h_n'']$. They are based upon increasing subsequences of the integers, n_k' and n_k'' respectively. In all cases (A, B and C), we have $n_1'=n_1''=1$. Also, $h_n''=1/k$ on $[n_k'',n_{k+1}'')\to 0$ as $k\to\infty$. Finally, h_n' and h_n'' and arbitrarily defined on $[n_1',n_2')$ and $[n_1'',n_2'')$ respectively.

Part A. Let

$$\begin{aligned} n_k'' &= \inf(n:\ n>n_{k-1}'',\ \sup_{m\geq n}P(h_m\geq 1/k)\leq 1/k),\quad k\geq 2,\\ n_k' &= \inf(n:\ n>n_{k-1}',\ \sup_{m\geq n}P(mh_m^d\leq k)\leq 1/k),\quad k\geq 2,\\ h_n' &= (k/n)^{1/d}\text{ on }[n_k',n_{k+1}'),\quad k\geq 2. \end{aligned}$$

Clearly, $nh_n'^d\to\infty$. Also, on $[n_k'',n_{k+1}'')$, $P(h_n\geq h_n'')=P(h_n\geq 1/k)\leq 1/k\to 0$ as $k\to\infty$. Similarly, $P(nh_n^d\leq nh_n'^d)=P(nh_n^d\leq k)\leq 1/k$ on $[n_k',n_{k+1}')$, and this tends to 0 as $k\to\infty$. This completes the proof of part A.

Part C. Let

$$\begin{aligned} n_k'' &= \inf(n:\ n>n_{k-1}'',\ \textstyle\sum_{m\geq n}m^kP(h_m\geq 1/k)\leq 2^{-k}),\quad k\geq 2,\\ n_k' &= \inf(n:\ n>n_{k-1}',\ \textstyle\sum_{m\geq n}m^kP(mh_m^d/\log m\leq k)\leq 2^{-k}),\quad k\geq 2,\\ h_n' &= (k\log n/n)^{1/d}\quad\text{on}\quad[n_k',n_{k+1}'),\quad k\geq 2. \end{aligned}$$

Clearly, $nh_n'^d/\log n \to \infty$. Also, for all $q \geq 0$, if $s = \max(2, \text{ceiling of } q)$,

$$\sum_{n=1}^{\infty} n^q P(h_n \geq h_n'') \leq n_s''^{q+1} + \sum_{k \geq s} \sum_{n=n_k''}^{n_{k+1}''-1} n^k P(h_n \geq 1/k)$$
$$\leq n_s''^{q+1} + \sum_{k \geq s} 2^{-k} < \infty.$$

By an identical argument,

$$\sum_{n=1}^{\infty} n^q P(h_n \leq h_n') \leq n_s'^{q+1} + \sum_{k \geq s} 2^{-k} < \infty.$$

Thus, $\sum_n n^q P(h_n \notin H_n) < \infty$, all $q \geq 0$, and therefore, the right-hand side of (19) tends to 0 completely in view of Theorem 1.

Part B. Let

$$n_k'' = \inf(n: n > n_{k-1}'', P(U_{m \geq n}[mh_m^d/\log\log m \leq k]) \leq 2^{-k}), \quad k \geq 2,$$

$$n_k' = \inf(n: n > n_{k-1}', P(U_{m \geq n}[h_m \geq 1/k]) \leq 2^{-k}), \quad k \geq 2,$$

$$h_n' = (k \log\log n/n)^{1/d} \text{on } [n_k', n_{k+1}'), \quad k \geq 2.$$

Check that $nh_n'^d/\log\log n \to \infty$, and that $h_n \geq h_n''$ finitely often almost surely because on $[n_k'', n_{k+1}'')$,

$$P(U_{m \geq n}[h_m \geq h_m'']) \leq \sum_{j=k}^{\infty} P(U_{m=n_j''}^{n_{j+1}''}[h_m \geq 1/j])$$
$$\leq \sum_{j=k}^{\infty} 2^{-j} = 2^{-k+1} \to 0 \quad \text{as} \quad k \to \infty.$$

In a similar way, it can be checked that $h_n \leq h_n'$ finitely often almost surely. Part B will be complete if we can find a sequence of positive numbers $h_n^* \leq h_n'$ such that $nh_n^{*d}/\log\log n \to \infty$ and that h_n^* is *regularly varying*. Theorem 1 and (19) will then complete the proof. The sequence $\phi(n) = nh_n'^d/\log\log n$ is nondecreasing by construction, and it tends to ∞. Define $\phi(t)$ on the real line by linear interpolation from $\phi(n)$. We will attempt to find a function $\psi(t)$ with $0 \leq \psi \leq \phi$, $\psi(t) \uparrow \infty$ as $t \uparrow \infty$, and $t\psi'(t)/\psi(t) \to 0$ as $t \to \infty$. This function ψ is thus slowly varying (Seneta, 1976, pages 6–7). Then, we define $h_n^* = (\psi(n)\log\log n/n)^{1/d}$, and note that it satisfies all our requirements.

The function ψ that we suggest is continuous and piecewise linear with knots at $t_1 < t_2 < \cdots$, where $t_k \to \infty$. Let $t_1 = 1$, and set $\psi(t) = \phi(t)$ on $[0, 1]$. Given t_k and $\psi(t_k)$ we define t_{k+1} and $\psi(t_{k+1})$ as follows:

$$\psi(t_{k+1}) = \min(\phi(t_k), \psi(t_k)(1 + 1/(2 \log k))),$$

$$t_{k+1} = \inf(t: t \geq t_k + 1, t/t_k \geq \psi(t_{k+1})/\psi(t_k),$$
$$t - t_k \geq (\psi(t_{k+1}) - \psi(t_k))t \log k/\psi(t_{k+1})).$$

Note that $t_k \geq k \to \infty$ as $k \to \infty$, that $\psi(t)/t \downarrow$, and that on $[t_k, t_{k+1})$, $\psi'(t) \leq \psi(t_{k+1})/(t_{k+1}\log k) \leq (\psi(t)/t)/\log k$. The existence of t_{k+1} follows from the fact that we can always find $t \geq t_k + 1$ such that $t \geq t_k/(1 - \log k(1 - \psi(t_k)/(t_{k+1})))$, because the denominator in the last expression is always at least ½ (in other words, $t \geq 2t_k$ will always satisfy the given condition). Finally, $0 \leq \psi \leq \phi$ and

$\psi(t) \uparrow \infty$ because $\phi(t) \to \infty$ and

$$\prod_{k=2}^{\infty} (1 + 1/(2 \log k)) = \infty.$$

PROOF OF THEOREM 3. Part 1 requires no new proof. The equivalence of C, D and E is established in Devroye (1983). Obviously, $C \Rightarrow B \Rightarrow A$ (see, for example, Devroye and Wagner, 1979). Finally, $A \Rightarrow D$ by Glick's extension of Scheffé's theorem (Glick, 1974).

Part 3 is partially shown in Devroye and Wagner (1979) (i.e. $C \Rightarrow B \Rightarrow A$). To show $A \Rightarrow C$, we note that the necessity of $h_n = o(1)$ follows from part 1, and that the necessity of $nh^d/\log n \to \infty$ follows from Lemma 6: indeed,

$$\sum_{n=1}^{\infty} n^q P(f_n(x) - E(f_n(x)) > \varepsilon) < \infty, \quad \text{all} \quad q \geq 0, \quad \varepsilon > 0,$$

$$\text{almost all} \quad x,$$

$h_n = o(1)$ and $nh_n^d \to \infty$ (both consequences of part 1 of this theorem) imply that

$$\sum_{n=1}^{\infty} \min(1, (nh_n^d)^{-1/2} \exp(-anh_n^d)) < \infty, \quad \text{all} \quad a > 0 \tag{20}$$

by Lemma 6, since we can restrict ourselves to Lebesgue points for f, with $f(x) > 0$. If $nh_n^d/\log n$ is bounded by M, then the sum in (20) is at least equal to

$$\sum_{n \geq e^{1/M}} (M \log n)^{-1/2} n^{-aM},$$

which is not summable for $a \leq 1/M$. But if $nh_n^d/\log n$ cannot remain bounded, then $\lim_{n \to \infty} nh_n^d/\log n = \infty$ by its semimonotonicity. Hence $A \Rightarrow C$.

Part 2 is the only nontrivial part of the Theorem. Clearly, $B \Rightarrow A$. Also, $C \Rightarrow B$ by Theorem 2 when K is Riemann integrable. We will now show that Lemma 6 suffices to prove that $A \Rightarrow C$. Fix a constant $a > 0$, and define the subsequence n_i by $\exp(ai \log i)$, $i \geq 1$. Notice that $(n_{i+1} - n_i)/n_i \sim (ei)^a$. Assume that we can show that whenever $nh_n^d/\log \log n \leq M < \infty$, $h_n \to 0$, $nh_n^d \to \infty$, and x is a Lebesgue point of f with $f(x) > 0$, then

$$P(|f_{n_i}(x) - E(f_{n_i}(x))| > \varepsilon \text{ infinitely often}) = 1 \tag{21}$$

for ε small enough. By the semimonotonicity of $nh_n^d/\log \log n$, we must have that $\lim_{n \to \infty} nh_n^d/\log \log n = \infty$, to avoid a contradiction. The necessity of $h_n = o(1)$ follows from part 1 of this Theorem, as does the necessity of $\lim_{n \to \infty} nh_n^d = \infty$. We will thus show (21) under the stated conditions. We have

$$\begin{aligned} &[|f_{n_i}(x) - E(f_{n_i}(x))| > \varepsilon \text{ i.o.}] \\ &\subseteq [|\tilde{f}_i(x) - E(\tilde{f}_i(x))| > 2\varepsilon \text{ i.o.}] \\ &\cap [(n_i/n_{i+1}) |f_i^*(x) - E(f_i^*(x))| > \varepsilon \text{ f.o.}] \end{aligned} \tag{22}$$

where

$$\tilde{f}_i(x) = (n_{i+1} - n_i)^{-1} \sum_{j=n_i+1}^{n_{i+1}} K((x - X_j)/h_{n_{i+1}})/h_{n_{i+1}}^d,$$

$$f_i^*(x) = n_i^{-1} \sum_{j=1}^{n_i} K((x - X_j)/h_{n_{i+1}})/h_{n_{i+1}}^d.$$

Implication (22) follows from the inequality

$$|f_{n_{i+1}}(x) - E(f_{n_{i+1}}(x))|$$
$$\geq \frac{n_{i+1} - n_i}{n_{i+1}} |\tilde{f}_i(x) - E(\tilde{f}_i(x))| - \frac{n_i}{n_{i+1}} |f_i^*(x) - E(f_i^*(x))|.$$

By Lemma 4,

$$P\left(\frac{n_i}{n_{i+1}} |f_i^*(x) - E(f_i^*(x))| > \varepsilon\right)$$
$$\leq 2 \exp\left(- n_i h_{n_{i+1}}^d \frac{\varepsilon^2 (n_{i+1}/n_i)^2}{2K^*(f(x) + \varepsilon + o(1))}\right)$$
$$= 2 \exp\left(- n_{i+1} h_{n_{i+1}}^d \frac{(\varepsilon^2 + o(1))(ei)^a}{2K^*(f(x) + \varepsilon + o(1))}\right)$$

which is summable in i for all a, $\varepsilon > 0$ (since $nh_n^d \to \infty$), so that by the Borel-Cantelli lemma, the last event in (22) has probability 1. By the independence of its component events, the middle event in (22) occurs with probability one if and only if

$$\sum_{i=1}^{\infty} P(|\tilde{f}_i(x) - E(\tilde{f}_i(x))| > 2\varepsilon) = \infty. \tag{23}$$

A lower bound for the ith probability in (23) is given in Lemma 6 if we replace n and h there by $n_{i+1} - n_i$ and $h_{n_{i+1}}$ respectively. By our assumptions, $h_{n_{i+1}} = o(1)$, $(n_{i+1} - n_i)h_n^{2d} = o(1)$ and $(n_{i+1} - n_i)h_{n_{i+1}}^d \to \infty$, so that Lemma 6 indeed applies. The lower bound for the ith term is of the form

$$c_1(n_{i+1}h_{n_{i+1}}^d)^{-1/2} \exp(-c_2 n_{i+1} h_{n_{i+1}}^d), \quad i \text{ large enough}, \tag{24}$$

where c_1, c_2 are positive constants for all $\varepsilon > 0$, $\liminf_{\varepsilon \downarrow 0} c_1 > 0$ and $\liminf_{\varepsilon \downarrow 0} c_2 = 0$. Clearly, (24) is at least equal to

$$c_1(M \log \log n_{i+1})^{-1/2} \exp(-c_2 M \log \log n_{i+1})$$
$$\sim c_1(M \log i)^{-1/2}(ai \log i)^{c_2 M},$$

and no tail sum is finite when $c_2 < 1/M$ (i.e. when ε is small enough). This concludes the proof of (23), (21) and Theorem 3.

REFERENCES

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

BREIMAN, L., MEISEL, W. and PURCELL, E. (1977). Variable kernel estimates of multivariate densities and their calibration. *Technometrics* **19** 135–141.

BRETAGNOLLE, C. and HUBER, C. (1979). Estimation des densités: risque minimax. *Z. Wahrsch. verw. Gebiete* **47** 119–137.

CHOW, Y. S., GEMAN, S. and WU, L. D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11** 25–38.

DEHEUVELS, P. and HOMINAL, P. (1980). Estimation automatique de la densité. *Revue de Statist. Appl.* **28** 25–55.

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 Ser. A* **278** 1217–1220.

DEHEUVELS, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statist. Appl.* **25** 5–42.

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. (1980). The strong uniform consistency of kernel density estimates. In *Multivariate Analysis V* 59–77. P. R. Krishnaiah Ed. North-Holland, New York.

DEVROYE, L. (1983). The equivalence of weak, strong and complete convergence in $L1$ for kernel density estimates. *Ann. Statist.* **11** 896–904.

DEVROYE, L. (1983). On arbitrarily slow rates of global convergence in density estimation. *Z. Wahrsch. verw. Gebiete* **62** 475–483.

DUIN, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.* **C-25** 1175–1179.

GEMAN, S. (1981). Sieves for nonparametric estimation of densities and regressions. Reports in Pattern Analysis No. 99, Division of Applied Mathematics, Brown University, Providence, Rhode Island.

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

HABBEMA, J. D. F., HERMANS, J. and VANDENBROEK, K. (1974). A stepwise discriminant analysis program using density estimation. *Compstat 1974* 101–110. G. Bruckmann Ed. Physica Verlag, Wien. 1974.

HALL, P. (1982a). Cross-validation in density estimation. *Biometrika* **69** 383–390.

HALL, P. (1982b). Limit theorems for stochastic measures of the accuracy of nonparametric density estimators. *Stochastic Process. Appl.* **13** 11–25.

KIEFER, J. (1961). On large deviations of the empiric D. F. of vector chance variables and a law of the iterated logarithm. *Pacific J. Math.* **11** 649–660.

KRIEGER, A. M., and PICKANDS, J. (1981). Weak convergence and efficient density estimation at a point. *Ann. Statist.* **9** 1066–1078.

NADARAYA, E. A. (1974). On the integral mean square error of some nonparametric estimates for the density function. *Theory Probab. Appl.* **19** 133–141.

PARZEN, E. (1962). On estimates of a probability density function. *Ann. Math. Statist.* **40** 854–864.

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

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

ROSENBLATT, M. (1979). Global measures of deviation for kernel and nearest neighbor density estimates. In *Smoothing Techniques for Curve Estimation.* Th. Gasser and M. Rosenblatt, eds. *Lecture Notes in Mathematics* **757** 181–190. Springer-Verlag, Berlin.

SACKS, J. and YLVISAKER, D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.* **9** 334–346.

SCHUSTER, E. F. and GREGORY, G. G. (1978). Choosing the shape factor(s) when estimating a density. Technical Report, Department of Mathematics, University of Texas at El Paso.

SCHUSTER, E. F. and GREGORY, G. G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* 295–298. W. F. Eddy Ed. Springer-Verlag, New York.

SCOTT, D. W., TAPIA, R. A. and THOMPSON, J. R. (1977). Kernel density estimation revisited. *J. Nonlinear Anal. Theory, Methods Appl.* **1** 339–372.

SCOTT, D. W. and FACTOR, L. E. (1981). Monte Carlo study of three data-based nonparametric probability density estimators. *J. Amer. Statist. Assoc.* **76** 9–15.

SENETA, E. (1976). Regularly Varying Functions. *Lecture Notes in Math.* **508**. Springer-Verlag, Heidelberg.

SILVERMAN, B. W. (1978). Choosing the window width when estimating a density. *Biometrika* **65** 1–11.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation.* The Johns Hopkins University Press, Baltimore, Maryland.

WAGNER, T. J. (1975). Nonparametric estimates of probability densities. *IEEE Trans. Inform. Theory* **IT-21** 438–440.

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

WOODROOFE, M. (1970). On choosing a delta-sequence. *Ann. Math. Statist.* **41** 1665–1671.

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

APPLIED RESEARCH LABORATORIES
THE UNIVERSITY OF TEXAS
P.O. BOX 8029
AUSTIN, TEXAS 78712