P. R. Krishnaiah, ed., *Multivariate Analysis–V*
© North-Holland Publishing Company (1980) 59–77.

THE STRONG UNIFORM CONSISTENCY OF KERNEL DENSITY ESTIMATES

L. P. DEVROYE* and T. J. WAGNER**
School of Computer Science, McGill University,
Montreal, P.Q., Canada and
Department of Electrical Engineering,
The University of Texas, Austin, TX 78712, U.S.A.

Let $X_1,\ldots,X_n$ be independent, identically distributed random vectors taking values in $\mathbf{R}^d$ with a common probability density f. If K is a bounded probability density on $\mathbf{R}^d$ and $\{h_n\}$ is a sequence of positive numbers then $f_n(x)=\sum_1^n K((x-X_i)/h_n)/(nh_n^d)$ is the kernel estimate of f from $X_1,\ldots,X_n$. Conditions on f, K and $\{h_n\}$ are given which insure that $\sup_x|f_n(x)-f(x)|\overset{n}{\to}0$ with probability one. Additionally, conditions are discussed which allow h_n to be a function of $X_1,\ldots,X_n$ and still retain the consistency properties of f_n.

1. Introduction

Let $X_1,X_2,\ldots,X_n$ be a sequence of independent, identically distributed random vectors taking values in $\mathbf{R}^d$ with a common probability density f. The kernel estimate (Rosenblatt (1957), Parzen (1962)) is given by

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

where K, the kernel, is a bounded probability density on $\mathbf{R}^d$ and $\{h_n\}$ is a sequence of positive numbers. In this paper we are concerned mainly with conditions which insure the strong uniform consistency of f_n, that is,

$$\sup_x|f_n(x)-f(x)|\to 0 \quad \text{w.p. 1.} \tag{1.1}$$

Our results, as well as all of those that we are aware of which lead to (1.1), require that f be uniformly continuous on $\mathbf{R}^d$. The work of Schuster (1969, 1970) comes close to proving that this is a necessity for $\mathbf{R}^1$. We will be contented then just to make this assumption. If f and K are continuous on

*Research was supported in part by DOD Joint Services Electronics Program through the Air Force Office of Scientific Research (AFSC) Contract F 49620-77-C-0101.
**Research was sponsored by AFOSR Grant 77-3385.

$\mathbf{R}^d$ then $\sup_x |f_n(x) - f(x)|$ is a random variable. This remains true if f is continuous on $\mathbf{R}^d$ and if the values of K can be determined from its values on a countable dense set (e.g., for each $x \in \mathbf{R}^d$ there is a sequence $\{x_n\}$ from a countable dense set D such that $x_n \xrightarrow{n} x$ and $K(x_n) \xrightarrow{n} K(x)$). While one can easily think of kernels which do not have this property, we know of none that are interesting. Rather than explore this point further, or deal with the case that $\sup_x |f_n(x) - f(x)|$ is not a random variable, we will assume throughout this paper that it is a random variable.

Let

$$L(z) = \sup_{\|x\| > z} K(x) \qquad z \in [0, \infty)$$

and

$$L^{-1}(t) = \sup\{z : L(z) \geq t\} \qquad t \in \left[0, \sup_x K(x)\right),$$

where $\|\cdot\|$ denotes the supremum norm on $\mathbf{R}^d$. Our result may be stated as follows.

Theorem 1. *Suppose f is uniformly continuous on $\mathbf{R}^d$ and K is a bounded Riemann integrable probability density with*

$$\int_0^\infty z^{d-1} L(z)\,dz < \infty. \tag{1.2}$$

If

$$h_n \xrightarrow{n} 0,$$

then (1.1) *follows from*

$$(nh_n^d)/(L^{-1}(\varepsilon h_n^d))^d \log n \xrightarrow{n} \infty \qquad \text{for } \varepsilon > 0. \tag{1.3}$$

Remarks. For kernels with compact support the conditions (1.2) and (1.3) can be replaced by

$$nh_n^d / \log n \xrightarrow{n} \infty. \tag{1.4}$$

For kernels with

$$K(x) \leq A/\|x\|^{\alpha d} \qquad \text{for some } \alpha > 1$$

(1.2) and (1.3) can be replaced by

$$nh_n^{(1+\alpha^{-1})} / \log n \xrightarrow{n} \infty.$$

Additionally, (1.3) can always be replaced by

$$nh_n^{2d} / \log n \xrightarrow{n} \infty. \tag{1.5}$$

The condition (1.2) for K is close to the condition frequently imposed on K:

$$\|x\|^d K(x) \to 0 \quad \text{as } \|x\| \to \infty. \tag{1.6}$$

For example, (1.6) is implied by (1.2) and (1.2) follows whenever

$$\|x\|^{d+\delta} K(x) \to 0 \quad \text{as } \|x\| \to \infty$$

for some $\delta > 0$.

The above theorem does not use the restrictive assumption that K is of bounded variation on $\mathbf{R}^d$ (Nadaraya (1965), Moore and Yackel (1976) and Silverman (1978)) or that K has an integrable characteristic function (Van Ryzin (1969)). For example, the kernel which is uniform over the unit sphere satisfies neither of these assumptions and, while the kernel which is uniform over the unit cube in $\mathbf{R}^d$ has bounded variation, an orthogonal rotation of the coordinates can yield a kernel with an infinite variation while keeping $\sup_x |f_n(x) - f(x)|$ unchanged. Additionally, no moment assumptions are put on f (Deheuvels (1974), Földes and Révész (1974)) and the requirements for $\{h_n\}$, at least for kernels with compact support, are essentially the weakest possible to get (1.1) (Deheuvels (1974)).

A disadvantage of the kernel estimate is that h_n is chosen without regard to $X_1, \ldots, X_n$ (Cover (1972)). One possible remedy is to replace h_n by a function of $X_1, \ldots, X_n$, say $H_n = H_n(X_1, \ldots, X_n)$. The resulting estimate

$$\hat{f}_n(x) = \sum_{i=1}^n K((x - X_i)/H_n)/(nH_n^d)$$

has been examined by Wagner (1975), primarily for $d = 1$, where several choices for H_n are also discussed. The technique used to prove Theorem 1 also yields the following result for $\hat{f}_n$. (See also the remark at the end of Section 2.)

Theorem 2. *Suppose f is uniformly continuous on $\mathbf{R}^d$ and K is a bounded Riemann integrable probability density satisfying* (1.2). *If*

$$H_n \xrightarrow{n} 0 \qquad \text{w.p. } 1, \tag{1.7}$$

and

$$nH_n^{2d}/\log n \xrightarrow{n} \infty \qquad \text{w.p. } 1 \tag{1.8}$$

then

$$\sup_x |\hat{f}_n(x) - f(x)| \xrightarrow{n} 0 \qquad \text{w.p. } 1 \tag{1.9}$$

Additionally, if the convergence in (1.7) *is in probability and if*

$$nH_n^{2d} \xrightarrow{n} \infty \text{ in probability}, \tag{1.10}$$

then

$$\sup_x |\hat{f}_n(x) - f(x)| \xrightarrow{n} 0 \text{ in probability}. \tag{1.11}$$

2. Details

Following Nadaraya (1963) it suffices to prove that

$$\sup_x |f_n(x) - Ef_n(x)| \to 0 \qquad \text{w.p. } 1 \tag{2.1}$$

since, when f is uniformly continuous on $\mathbf{R}^d$,

$$\sup_x |Ef_n(x) - f(x)| \xrightarrow{n} 0$$

whenever $h_n \xrightarrow{n} 0$.

Consider, for the moment, the kernel which is uniform over $(0,1]^d$. Then

$$\sup_x |f_n(x) - Ef_n(x)| = h_n^{-d} \sup_{A \in \mathcal{Q}} |\mu_n(A) - \mu(A)|$$

where μ_n is the empirical measure for $X_1,\ldots,X_n$, μ is the measure on the Borel subsets of $\mathbf{R}^d$ which corresponds to f and $\mathcal{Q}$ is the class of cubes $\prod_{i=1}^d (x^i, x^i + h_n]$, $x = (x^1,\ldots,x^d) \in \mathbf{R}^d$. Thus

$$P\left\{ \sup_x |f_n(x) - Ef_n(x)| \geq \varepsilon \right\} = P\left\{ \sup_{A \in \mathcal{Q}} |\mu_n(A) - \mu(A)| \geq h_n^d \varepsilon \right\}. \tag{2.2}$$

Rather than use the inequality of Kiefer–Wolfowitz (1956) on the right-hand side of (2.2), which leads to the condition (1.5) on $\{h_n\}$, we modify a result of Vapnik–Chervonenkis (1971) allowing us to upper-bound (2.2) by $c_0 n^{2d} \exp(-cnh_n^d \varepsilon^2)$ for some c_0, $c > 0$. This now leads to condition (1.4) to get (2.1) for the kernel which is uniform over $(0,1]^d$. A careful approximation of Riemann integrable kernels by linear combinations of indicators of disjoint rectangles then yields the theorem.

We begin by proving two useful lemmas concerning upper bounds for

$$P\left\{ \sup_{A \in \mathcal{Q}} |\mu_n(A) - \mu(A)| \geq \varepsilon \right\}$$

where $\varepsilon > 0$ and $\mathcal{Q}$ is a subclass of the Borel subsets of $\mathbf{R}^d$. If μ_n and μ_n' are empirical measures for two independent samples of size n then, assuming that $\sup_{\mathcal{Q}} |\mu_n(A) - \mu_n'(A)|$ and $\sup_{\mathcal{Q}} |\mu_n(A) - \mu(A)|$ are measurable, Vapnik and Chervonenkis (1971) showed that

$$P\left\{ \sup_{\mathcal{Q}} |\mu_n(A) - \mu(A)| \geq \varepsilon \right\} \leq 4s(\mathcal{Q}, 2n) e^{-n\varepsilon^2/8}$$

where

$$s(\mathcal{Q}, n) = \max_{(x_1,\ldots,x_n)} N_{\mathcal{Q}}(x_1,\ldots,x_n)$$

and $N_{\mathcal{Q}}(x_1,\ldots,x_n)$ denotes the number of different sets in the class $\{\{x_1,\ldots,x_n\} \cap A : A \in \mathcal{Q}\}$. If, in addition to the measurability assumptions of Vapnik and Chervonenkis, we assume that $\sup_{\mathcal{Q}} \mu_n(A)$ is measurable we have the following lemma.

Lemma 1. *Let $\varepsilon > 0$ and suppose that*

$$\sup_{\mathcal{Q}} \mu(A) \leq b \leq 1/4.$$

Then

$$P\left\{\sup_{\mathscr{A}} |\mu_n(A)-\mu(A)| \geqslant \varepsilon\right\}$$

$$\leqslant 4s(\mathscr{A},2n)e^{-n\varepsilon^2/(64b+4\varepsilon)} + 2P\left\{\sup_{\mathscr{A}} \mu_{2n}(A) > 2b\right\} \tag{2.3}$$

for all $n \geqslant 8b/\varepsilon^2$.

Inequality (2.3) is useful for small b and for classes $\mathscr{A}$ for which $P\{\sup_{\mathscr{A}}\mu_{2n}(A) > 2b\}$ can be upper-bounded. The following lemma provides a useful upper bound for classes $\mathscr{A}$ whose sets have a uniform bound on their diameter. We again assume that $\sup_{\mathscr{A}}\mu_{2n}(A)$ is a random variable and, as before, $\|\cdot\|$ will denote the sup norm on $\mathbf{R}^d$. As usual, other norms could be used instead.

Lemma 2. *Let* $\mathscr{A}$ *be any class of Borel sets from* $\mathbf{R}^d$ *with*

$$\sup_{\mathscr{A}} \sup_{x,y\in A} \|y-x\| \leqslant r < \infty.$$

If $S(x,r)$ *is the closed sphere centered at* x *with radius* r *and if*

$$\sup_{x\in\mathbf{R}^d} \mu(S(x,r)) \leqslant b,$$

then

$$P\left\{\sup_{\mathscr{A}} \mu_{2n}(A) \geqslant 2b\right\} \leqslant 4ne^{-nb/10} \tag{2.4}$$

for all $n \geqslant 1/b$.

As an example, which will be used in the proof of the theorem, let $\mathscr{A}_r$ be the class of all rectangles from $\mathbf{R}^d$ with diameter not greater than r and assume that

$$\sup_{\mathscr{A}_{2r}} \mu(A) \leqslant b \leqslant 1/4.$$

Since

$$\sup_{x\in\mathbf{R}^d} \mu(S(x,r)) = \sup_{\mathscr{A}_{2r}} \mu(A)$$

and

$$s(\mathscr{A}_r, 2n) \leqslant (2n)^{2d} \qquad \text{for all } r$$

(see Cover (1965) for other calculations of this type) we have, from Lemmas 1 and 2, that

$$P\left\{\sup_{\mathscr{A}_r} |\mu_n(A)-\mu(A)| \geqslant \varepsilon\right\} \leqslant 4(2n)^{2d}e^{-n\varepsilon^2/(64b+4\varepsilon)} + 8ne^{-nb/10} \tag{2.5}$$

for $\varepsilon > 0$ and $n \geqslant \max(1/b, 8b/\varepsilon^2)$.

Proof of Lemma 1. The following arguments are variations of those of Vapnik and Chervonenkis (1971). Let $X_1,\ldots,X_{2n}$ be independent, identically distributed random vectors with a common probability measure μ. If μ_n' denotes the empirical measure for $X_{n+1},\ldots,X_{2n}$ and all unlabeled supremums below are taken over $\mathscr{A}$, then an easy modification of Lemma 1 of Vapnik and Chervonenkis yields

$$P\{\sup|\mu_n(A)-\mu(A)| \geqslant \varepsilon\} \leqslant 2P\{\sup|\mu_n(A)-\mu_n'(A)| \geqslant \varepsilon/2\} \tag{2.6}$$

where

(i) $\sup\mu(A) \leqslant b$,
(ii) $\sup|\mu_n(A)-\mu(A)|$ and $\sup|\mu_n(A)-\mu_n'(A)|$ are random variables,
(iii) $n \geqslant 8b/\varepsilon^2$.

Because

$$P\{\sup|\mu_n(A)-\mu_n'(A)| \geqslant \varepsilon/2\}$$

$$\leqslant P\{\sup|\mu_n(A)-\mu_n'(A)| \geqslant \varepsilon/2;\ \sup\mu_{2n}(A) \leqslant 2b\}$$

$$+ P\{\sup\mu_{2n}(A) > 2b\}.$$

Lemma 1 will follow from (2.6) if we can show that for any δ, $M > 0$ with $M \leqslant 1/2$

$$P\{\sup|\mu_n(A)-\mu_n'(A)| \geqslant \delta;\ \sup\mu_{2n}(A) \leqslant M\}$$

$$\leqslant 2s(\mathscr{A},2n)e^{-n\delta^2/(8M+2\delta)}. \tag{2.7}$$

The probability on the left-hand side of (2.7) equals

$$\int_{\mathbf{R}^{2nd}} \frac{1}{(2n)!} \sum I_{[\sup|\mu_n(A)-\mu_n'(A)|\geqslant\delta]} I_{[\sup \mu_{2n}(A)\leqslant M]} \,\mathrm{d}Q$$

where I_E is the indicator of a set $E \subseteq \mathbf{R}^{2nd}$, Q is the probability measure for $X_1,\ldots,X_{2n}$ defined on the Borel subsets of $\mathbf{R}^{2nd}$ and the inner summation is taken over all $(2n)!$ permutations of $x_1,\ldots,x_{2n}$. But this integral equals

$$\begin{aligned}
&\int_{\mathbf{R}^{2nd}} \frac{1}{(2n)!} \sum I_{[\sup \mu_{2n}(A)\leqslant M]} \sup I_{[|\mu_n(A)-\mu_n'(A)|\geqslant\delta]} \,\mathrm{d}Q \\
&\quad= \int_{\mathbf{R}^{2nd}} \frac{1}{(2n)!} \sum I_{[\sup \mu_{2n}(A)\leqslant M]} \sup_{\mathscr{Q}'} I_{[|\mu_n(A)-\mu_n'(A)|\geqslant\delta]} \,\mathrm{d}Q \\
&\quad= \int_{\mathbf{R}^{2nd}} \frac{1}{(2n)!} \sum I_{[\sup \mu_{2n}(A)\leqslant M]} \sup I_{[|\mu_n(A)-\mu_{2n}(A)|\geqslant\delta/2]} \,\mathrm{d}Q \\
&\quad\leqslant \int_{\mathbf{R}^{2nd}} \sum_{A\in\mathscr{Q}'} I_{[\sup \mu_{2n}(A)\leqslant M]} \left\{ \frac{1}{(2n)!} \sum I_{[|\mu_n(A)-\mu_{2n}(A)|\geqslant\delta/2]} \right\} \mathrm{d}Q
\end{aligned} \tag{2.8}$$

where $\mathscr{Q}'=\mathscr{Q}'(x_1,\ldots,x_{2n})$ is any finite subclass of $\mathscr{Q}$ which yields the same class of intersections with $\{x_1,\ldots,x_{2n}\}$ as does $\mathscr{Q}$ and where the unlabeled summations are again over all $(2n)!$ permutations of $x_1,\ldots,x_{2n}$.

If $Y_1,\ldots,Y_n$ are Bernoulli random variables with $P\{Y_1=1\}=\mathbf{p}$ then

$$P\left\{\left|\frac{1}{n}\sum_1^n Y_i-\mathbf{p}\right|\geqslant\varepsilon\right\} \leqslant 2\mathrm{e}^{-n\varepsilon((1+(b/\varepsilon))\ln(1+(b/\varepsilon))-1)} \leqslant 2\mathrm{e}^{-n\varepsilon^2/(2b+\varepsilon)} \tag{2.9}$$

provided $0\leqslant p\leqslant b\leqslant 1/2$ (Bennett (1962), Hoeffding (1963)). (The second inequality follows from $\ln(1+(a/b))\geqslant 2a/(2b+a)$ for $a,b>0$.) Hoeffding has pointed out that (2.9) remains valid if $Y_1,\ldots,Y_n$ are obtained by sampling without replacement from a sequence $y_1,\ldots,y_k$ of 0's and 1's where $k\geqslant n$ and $\sum_1^k y_i=k\mathbf{p}$. Using this last observation we have the following inequality between random variables (which holds everywhere)

$$\left\{\frac{1}{(2n)!}\sum I_{[|\mu_n(A)-\mu_{2n}(A)|\geqslant\delta/2]}\right\} \leqslant 2\mathrm{e}^{-n(\delta/2)^2/(2\mu_{2n}(A)+\delta/2)}.$$

Thus the last integral in (2.8) is upper-bounded by

$$\begin{aligned}
&\int \sum_{A\in\mathscr{Q}'} I_{[\sup \mu_{2n}(A)\leqslant M]} 2\mathrm{e}^{-n\delta^2/(8\mu_{2n}(A)+2\delta)} \,\mathrm{d}Q \\
&\quad\leqslant 2s(\mathscr{Q},2n)\mathrm{e}^{-n\delta^2/(8M+2\delta)}
\end{aligned}$$

since we can always choose $\mathscr{Q}'$ to contain no more than $s(\mathscr{Q},2n)$ sets. Inequality (2.2) and Lemma 1 now follow.

Proof of Lemma 2. If μ_{2n-1}^i is the empirical measure for $X_1,\ldots,X_{2n}$ with X_i omitted and all unlabeled supremums below are again over $\mathscr{Q}$, then

$$\begin{aligned}
P\{\sup \mu_{2n}(A)>2b\} &\leqslant P\left\{\bigcup_{i=1}^{2n}\{\mu_{2n}(S(X_i,r))>2b\}\right\} \\
&\leqslant P\left\{\bigcup_{i=1}^{2n}\{(2n-1)\mu_{2n-1}^i(S(X_i,r))>4bn-1\}\right\} \\
&\leqslant 2nP\{\mu_{2n-1}^1(S(X_1,r))>(4bn-1)/(2n-1)\}. \\
&\leqslant 2nP\{\mu_{2n-1}^1(S(X_1,r))>3b/2\} \qquad \text{if } bn\geqslant 1 \\
&\leqslant 2n \sup_x P\{\mu_{2n-1}(S(x,r))>3b/2\} \\
&\leqslant 2n \sup_x P\{\mu_{2n-1}(S(x,r))-\mu(S(x,r))>b/2\} \\
&\leqslant 4n\mathrm{e}^{-(2n-1)(b/2)^2/(2b+(b/2))} \qquad \text{(from (2.9))} \\
&\leqslant 4n\mathrm{e}^{-(2n-1)b/10} \leqslant 4n\mathrm{e}^{-nb/10},
\end{aligned}$$

which proves Lemma 2.

To prove the theorem we first approximate K by a linear combination of indicators of disjoint rectangles.

Lemma 3. *Suppose K is a nonnegative, bounded Riemann integrable function on $\mathbf{R}^d$. For each $\eta,\delta,\rho>0$ we can find a function*

$$K^*(x)=\sum_1^N \alpha_i I_{A_i}(x)$$

where

(i) $\alpha_1,\ldots,\alpha_N$ *are nonnegative real numbers,*

(ii) $A_1,\ldots,A_N$ *are disjoint rectangles contained in* $[-\rho,\rho]^d$,

(iii) $K^*(x)\leqslant \sup_x K(x)$, $x\in\mathbf{R}^d$,

(iv) $|K^*(x)-K(x)|<\eta$ *on* $[-\rho,\rho]^d$ *except on a set* D,

(v) $D\subseteq B=\bigcup_1^M B_i$ *where* $B_1,\ldots,B_M$ *are rectangles from* $[-\rho,\rho]^d$, *whose union has Lebesgue measure less than* δ.

Proof of Lemma 3. Partition $[-\rho,\rho]^d$ into disjoint rectangles in such a way that the upper and lower sums for K over the partition differ by less than $\eta\delta$ (Spivak (1965), Chapter 3). If K_1 and K_2 are, respectively, the functions corresponding to those upper and lower sums, then

$$\{x\in[-\rho,\rho]^d: K(x)-K_2(x)\geqslant\eta\}$$
$$\subseteq\{x\in[-\rho,\rho]^d: K_1(x)-K_2(x)\geqslant\eta\}.$$

The latter set is a union of disjoint rectangles with Lebesgue measure less than δ. Putting $K^*(x)=K_2(x)$ yields the lemma.

Proof of Theorem 1. We follow the notation of Lemma 3, assuming for the moment that η, δ and ρ are arbitrary positive numbers. The dependence of h_n on n will be suppressed where confusion is unlikely. First

$$\sup_x|Ef_n(x)-f_n(x)|$$
$$=\sup_x\left|h^{-d}\int K((y-x)/h)\,dF(y)-h^{-d}\int K((y-x)/h)\,dF_n(y)\right|$$
$$\leqslant\sum_{i=1}^{3}\sup_x U_i(x),$$

where

$$U_1(x)=h^{-d}\int|K((y-x)/h)-K^*((y-x)/h)|\,dF(y),$$

$$U_2(x)=h^{-d}\left|\int K^*((y-x)/h)\,dF(y)-\int K^*((y-x)/h)\,dF_n(y)\right|,$$

$$U_3(x)=h^{-d}\int|K^*((y-x)/h)-K((y-x)/h)|\,dF_n(y).$$

If $C\subseteq\mathbf{R}^d$, $x\in\mathbf{R}^d$ and $\alpha>0$, let $C(x,\alpha)=\{x+\alpha z: z\in C\}$ and let

$$C_1=S(x,\rho h)^c$$

$$C_2=S(x,\rho h)\cap D(x,h)^c$$

$$C_3=D(x,h)$$

where $(\ )^c$ denotes the complement of a set. Then

$$\sup_x U_1(x)$$
$$\leqslant\sum_1^3\sup_x\int_{C_i}h^{-d}|K^*((y-x)/h)-K((y-x)/h)|\,dF(y).$$

Recalling that K^* is zero outside of $[-\rho,\rho]^d$ we see that the first term is upper-bounded by

$$\sup_x\int_{S(x,\rho h)^c}h^{-d}K((y-x)/h)\,dF(y)$$

while the second and third terms are upper-bounded by

$$\eta M_1 h^{-d}\sup_x\lambda(S(x,\rho h)=\eta M_1 2^d\rho^d$$

and

$$2M_1M_2h^{-d}\sup_x\lambda(D(x,h))\leqslant 2M_1M_2\delta$$

respectively, where $M_1=\sup_x f(x)$, $M_2=\sup_x K(x)$ and λ denotes the Lebesgue measure on $\mathbf{R}^d$. Thus

$$\sup_x U_1(x)$$
$$\leqslant\sup_x\int_{S(x,\rho h)^c}h^{-d}K((y-x)/h)\,dF(y)+\eta M_1 2^d\rho^d+2M_1M_2\delta. \tag{2.10}$$

Next,

$$\begin{aligned}
&\sup_x U_2(x)\\
&\quad\leqslant \sup_x \left| \sum_1^N \alpha_i h^{-d}(\mu_n(A_i(x,h)-\mu(A_i(x,h)))\right|\\
&\quad\leqslant N M_2 h^{-d} \sup_{\mathcal{A}_n} |\mu_n(A)-\mu(A)| \qquad (2.11)
\end{aligned}$$

where $\mathcal{A}_n$ is the class of rectangles whose diameter does not exceed $2\rho h$. Finally,

$$\begin{aligned}
&\sup_x U_3(x)\\
&\quad\leqslant \sum_{i=1}^3 \sup_x \int_{C_i} h^{-d}|K((y-x)/h)-K^*((y-x)/h)|\,\mathrm{d}F_n(y).
\end{aligned}$$

The first term is upper-bounded by

$$\sup_x \int_{S(x,\rho h)^c} h^{-d}K((y-x)/h)\,\mathrm{d}F_n(y)$$

while the second term is upper-bounded by

$$\begin{aligned}
&\eta h^{-d}\sup_x \mu_n(S(x,\rho h))\\
&\quad\leqslant \eta h^{-d}\sup_x|\mu_n(S(x,\rho h))-\mu(S(x,\rho h))|+\eta h^{-d}\sup_x \mu(S(x,\rho h))\\
&\quad\leqslant \eta h^{-d}\sup_{\mathcal{A}_n}|\mu_n(A)-\mu(A)|+\eta M_1(2\rho)^d,
\end{aligned}$$

Recalling that $D\subseteq B=\bigcup_{i=1}^M B_i$ the third term is bounded by

$$\begin{aligned}
&M_2h^{-d}\sup_x \mu_n(B(x,h))\\
&\quad\leqslant M_2h^{-d}\sup_x|\mu_n(B(x,h))-\mu(B(x,h))|+M_2h^{-d}\sup_x \mu(B(x,h))\\
&\quad\leqslant MM_2h^{-d}\sup_{\mathcal{A}_n}|\mu_n(A)-\mu(A)|+M_1M_2\delta.
\end{aligned}$$

Thus

$$\begin{aligned}
\sup_x U_3(x) \leqslant{}& \sup_x \int_{S(x,\rho h)^c} h^{-d}K((y-x)/h)\,\mathrm{d}F_n(y)\\
&+(\eta+M_2M)h^{-d}\sup_{\mathcal{A}_n}|\mu_n(A)-\mu(A)|\\
&+\eta M_1 2^d\rho^d+M_1M_2\delta. \qquad (2.12)
\end{aligned}$$

From (2.10), (2.11) and (2.12) we see that

$$\begin{aligned}
&\sup_x |f_n(x)-Ef_n(x)|\\
&\quad\leqslant \sup_x \int_{S(x,\rho h)^c} h^{-d}K((y-x)/h)\,\mathrm{d}F_n(y)\\
&\qquad+\sup_x \int_{S(x,\rho h)^c} h^{-d}K((y-x)/h)\,\mathrm{d}F(y)\\
&\qquad+(M_2N+M_2M+\eta)h^{-d}\sup_{\mathcal{A}_n}|\mu_n(A)-\mu(A)|\\
&\qquad+3M_1M_2\delta+2\eta M_1 2^d\rho^d. \qquad (2.13)
\end{aligned}$$

The first two terms of (2.13) can be upper-bounded by

$$\begin{aligned}
&2\sup_x \int_{S(x,\rho h)^c} h^{-d}L(\|y-x\|/h)\,\mathrm{d}F(y)\\
&\quad+\sup_x\left|\int_{S(x,\rho h)^c} h^{-d}L(\|y-x\|/h)\,\mathrm{d}F_n(y)\right.\\
&\qquad\left.-\int_{S(x,\rho h)^c} h^{-d}L(\|y-x\|/h)\,\mathrm{d}F(y)\right|\\
&\leqslant 2M_1\int_\rho^\infty 2^{2d-1}t^{d-1}L(t)\,\mathrm{d}t\\
&\quad+\sup_x\left|\int_{S(x,\rho h)^c} h^{-d}L(\|y-x\|/h)\,\mathrm{d}F_n(y)\right.\\
&\qquad\left.-\int_{S(x,\rho h)^c} h^{-d}L(\|y-x\|/h)\,\mathrm{d}F(y)\right|
\end{aligned}$$

so that

$$\sup_x |f_n(x) - Ef_n(x)|$$

$$\leqslant (M_2N + M_2M + \eta)h^{-d} \sup_{\mathcal{Q}_n} |\mu_n(A) - \mu(A)|$$

$$+ \sup_x \left| \int_{S(x,\rho h)^c} h^{-d} L(\|y-x\|/h)\, \mathrm{d}F_n(y) \right.$$

$$\left. - \int_{S(x,\rho h)^c} h^{-d} L(\|y-x\|/h)\, \mathrm{d}F(y) \right|$$

$$+ 2^{2d} M_1 \int_\rho^\infty t^{d-1} L(t)\,\mathrm{d}t + 3M_1M_2\delta + 2\eta M_1 2^d \rho^d. \tag{2.14}$$

By choosing ρ sufficiently large and δ and η sufficiently small the last three terms of (2.14) can be made arbitrarily small. A straightforward application of Lemmas 1 and 2 for a fixed η, δ and ρ (e.g., in (2.5) $r = 2\rho h$, $b = 2^d M_1 \rho^d h^d$ and ε is replaced by $\varepsilon h^d/(M_2N + M_2M + \eta)$) shows that the first term of (2.14) tends to 0 with probability one if (1.3), which implies (1.4), is satisfied. The proof will be completed then if we show that

$$\sup_x \left| \int_{S(x,\rho h)^c} h^{-d} L(\|y-x\|/h)\, \mathrm{d}F_n(y) \right.$$

$$\left. - \int_{S(x,\rho h)^c} h^{-d} L(\|y-x\|/h)\, \mathrm{d}F(y) \right| \tag{2.15}$$

tends to zero with probability one for an arbitrarily large ρ.

Let $L'(t) = L(t) I_{[t \geqslant \rho]}$ so that (2.15) becomes

$$\sup_x \left| \int h^{-d} L'(\|y-x\|/h)\, \mathrm{d}F_n(y) - \int h^{-d} L'(\|y-x\|/h)\, \mathrm{d}F(y) \right|. \tag{2.16}$$

For an arbitrary integer l let

$$S_j = \left\{ x : (j-1)\frac{L(\rho)}{l} < u(\|x\|) \leqslant \frac{j}{l} L(\rho) \right\} \qquad 1 \leqslant j \leqslant l,$$

$$T_j = \left\{ x : \frac{(j-1)}{l} L(\rho) < L(\|x\|) \leqslant L(\rho) \right\} = \bigcup_{i=j}^{l} S_i,$$

and

$$L''(x) = \sum_{j=1}^{l} (j-1)\frac{L(\rho)}{l} I_{S_j}(x)$$

so that

$$|L'(\|x\|) - L''(x)| \leqslant L(\rho)/l$$

for all x. Returning to (2.16) we see that it is bounded by

$$\sup_x \int \left| h^{-d} L''((y-x)/h)\, \mathrm{d}F_n(y) - \int h^{-d} L''((y-x)/h)\, \mathrm{d}F(y) \right|$$

$$+ \sup_x \int h^{-d} |L''((y-x)/h) - L'(\|y-x\|/h)|\, \mathrm{d}F_n(y)$$

$$+ \sup_x \int h^{-d} |L''((y-x)/h) - L'(\|y-x\|/h)|\, \mathrm{d}F(y)$$

$$\leqslant \frac{2L(\rho)h^{-d}}{l} + \frac{L(\rho)h^{-d}}{l} \sup_x \left| \sum_{j=1}^{l} (j-1)\big(\mu_n(S_j(x,h)) - \mu(S_j(x,h))\big) \right|$$

$$\leqslant \frac{2L(\rho)h^{-d}}{l} + \frac{L(\rho)h^{-d}}{l} \sup_x \left| \sum_{j=2}^{l} \big(\mu_n(T_j(x,h)) - \mu(T_j(x,h))\big) \right|$$

$$\leqslant \frac{2L(\rho)h^{-d}}{l} + L(\rho)h^{-d} \sup_x \sup_{l \geqslant j \geqslant 2} |\mu_n(T_j(x,h)) - \mu(T_j(x,h))|. \tag{2.17}$$

Since T_j is the difference of two concentric rectangles of diameter at most $2L^{-1}(L(\rho)/l)$, (2.17) can be upper-bounded by

$$\frac{2L(\rho)h^{-d}}{l}+2L(\rho)h^{-d}\sup_{\mathcal{Q}_n'}|\mu_n(A)-\mu(A)| \tag{2.18}$$

where $\mathcal{Q}_n'$ is the class of rectangles with diameters at most $2hL^{-1}(L(\rho)/l)$. By taking $l=[h^{-d}]$ we can make the first term of (2.18) arbitrarily small by taking ρ large. Applying Lemmas 1 and 2 as done earlier we see that the second term of (2.18) tends to zero with probability one if (1.3) is satisfied. This completes the proof of Theorem 1.

Proof of Theorem 2. Letting

$$g_n(x)=H_n^{-d}\int K((x-y)/H_n)f(y)\,dy$$

it is straightforward to see that, with the conditions on K and f,

$$\sup_x|g_n(x)-f(x)|\xrightarrow{n}0 \quad \text{in probability} \quad (\text{w.p. } 1)$$

wherever

$$H_n\xrightarrow{n}0 \quad \text{in probability} \quad (\text{w.p. } 1)$$

(Wagner (1975)). We therefore examine the convergence of the quantity

$$\sup_x|\hat{f}_n(x)-g_n(x)|.$$

First, if $\{\mathfrak{B}_n\}$ is a sequence of positive numbers for which

$$I_{[nH_n^{2d}<\mathfrak{B}_n]}\xrightarrow{n}0 \quad \text{in probability} \quad (\text{w.p. } 1) \tag{2.19}$$

and

$$I_{[nH_n^{2d}\geq\mathfrak{B}_n]}\sup_x|\hat{f}_n(x)-g_n(x)|\xrightarrow{n}0 \quad \text{in probability} \quad (\text{w.p. } 1) \tag{2.20}$$

then

$$\sup_x|\hat{f}_n(x)-g_n(x)|\xrightarrow{n}0 \quad \text{in probability} \quad (\text{w.p. } 1).$$

Following the proof of Theorem 1, we see that

$$\sup_x|\hat{f}_n(x)-g_n(x)|I_{[nH_n^{2d}\geq\mathfrak{B}_n]}\leq\sum_{i=1}^{3}\sup_{x;\,nh^{2d}\geq\mathfrak{B}_n}U_i(x,h) \tag{2.21}$$

where the U_i are defined as before except now we make the dependence on h explicit. By following the proof of Theorem 1 we see that (2.21) can be bounded by

$$(M_2N+M_2M+\eta)(n/\mathfrak{B}_n)^{\frac{1}{2}}\sup_{\mathcal{Q}}|\mu_n(A)-\mu(A)|$$

$$+2^{2d}M_1\int_\rho^\infty t^{d-1}L(t)\,dt+3M_1M_2\delta+\eta M_1 2^d\rho^d$$

$$+\frac{2L(\rho)}{l}(n/\mathfrak{B}_n)^{\frac{1}{2}}+2L(\rho)(n/\mathfrak{B}_n)^{\frac{1}{2}}\sup_{\mathcal{Q}}|\mu_n(A)-\mu(A)| \tag{2.22}$$

where now $\mathcal{Q}$ is the class of *all* rectangles in $\mathbf{R}^d$. By taking

$$l=\left[(n/\mathfrak{B}_n)^{\frac{1}{2}}\right]$$

the middle four terms of (2.22) can be made arbitrarily small by choosing ρ large enough and η, δ small enough. The first and last terms of (2.22) can be combined to yield a term

$$c(n/\mathfrak{B}_n)^{\frac{1}{2}}\sup_{\mathcal{Q}}|\mu_n(A)-\mu(A)|. \tag{2.23}$$

Using the inequality of Kiefer–Wolfowitz (1956), we see that (2.23), and hence (2.20), tends to 0 in probability if $\mathfrak{B}_n\xrightarrow{n}\infty$, and tends to 0 w.p. 1 if

$$\sum_1^\infty e^{-\alpha\mathfrak{B}_n}<\infty\,\forall\alpha>0.$$

Using (1.10) or (1.8), it is now easy to show the existence of sequences $\{\mathfrak{B}_n\}$ which satisfy (2.19) and (2.20). This completes the proof of Theorem 2.

Remark. If f is an arbitrary density with continuity point x, Wagner (1975) has shown

$$g_n(x)\xrightarrow{n}f(x) \qquad \text{in probability} \qquad (\text{w.p. } 1)$$

whenever

$$H_n \xrightarrow{n} 0 \qquad \text{in probability} \qquad \text{(w.p. 1)}. \tag{2.24}$$

For the kernels of theorems 1 and 2, one can see, by examining the proofs of these theorems, that

$$|g_n(x) - \hat{f}_n(x)| \xrightarrow{n} 0 \qquad \text{in probability} \qquad \text{(or w.p. 1)}$$

whenever

$$nH_n^{2d} \xrightarrow{n} \infty$$

$$\left(\text{or } nH_n^{2d}/\log n \to \infty \qquad \text{w.p. 1}\right) \tag{2.25}$$

Thus, for these kernels, (2.24) and (2.25) imply

$$\hat{f}_n(x) \xrightarrow{n} f(x) \qquad \text{in probability} \qquad \text{(w.p. 1)}$$

when x is a continuity point of f.

References

[1] Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, **57**, 33–45.

[2] Cacoullos, T. (1965). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, **18**, 179–190.

[3] Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, **10**, 326–334.

[4] Cover, T. M. (1972). A hierarchy of probability density function estimates. *Frontiers of Pattern Recognition* (S. Watanabe, Editor) pp. 83–98. New York: Academic Press.

[5] Deheuvels, P. (1974). Conditions necessaires et suffisantes de convergence ponctuelle presque sure et uniforme presque sure estimateurs de la densite. *C. R. Acad. Sci. Paris Ser. A.*, **278**, 1217–1220.

[6] Földes, A. and Révész, P. (1974). A general method for density estimation. *Studia Scientiarium Mathematicarum Hungarica*, **9**, 81–92.

[7] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13–30.

[8] Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics*, **36**, 1049–1051.

[9] Moore, D. S. and Yackel, J. W. (1977). Consistency properties of nearest neighbor density function estimators. *Annals of Statistics*, **5**, 143–154.

[10] Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Probability and Its Applications*, **10**, 186–190.

[11] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, 1065–1076.

[12] Rosenblatt, M. (1957). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832–837.

[13] Schuster, E. F. (1969). Estimation of a probability density function and its derivatives. *Annals of Mathematical Statistics*, **40**, 1187–1195.

[14] Schuster, E. F. (1970). Note on the uniform convergence of density estimates. *Annals of Mathematical Statistics*, **41**, 1347–1348.

[15] Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Annals of Statistics*, **6**, 177–184.

[16] Spivak, M. (1965). *Calculus on Manifolds*. W. A. Benjamin, Inc., New York.

[17] Van Ryzin, J. (1969). On strong consistency of density estimates. *Annals of Mathematical Statistics*, **40**, 1765–1772.

[18] Vapnik, V. N. and Chervonenkis, A. Ya. (1971). On the uniform convergence of the relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, **16**, 264–280.

[19] Wagner, T. J. (1975). Nonparametric estimates of probability densities. *IEEE Transactions on Information Theory*, **IT-21**, 438–440.

Note. After this paper was in proof, we learned of the result by Bertrand–Retali (Convergence uniforme d'un estimateur de la densité par la méthode du Noyau, *Rev. Roumaine Math. Pures Appl.* **23** (1978), 361–385) which implies that Theorem 1 is true if (1.2) and (1.3) are replaced by (1.4) and

$$\int_{\mathbf{R}^d} \sup\{K(u) : \|u - x\| < 1\}\, dx < \infty.$$