The Annals of Probability
1982, Vol. 10, No. 2, 414–422

A LAW OF THE LOGARITHM FOR KERNEL DENSITY ESTIMATORS

BY WINFRIED STUTE[1]

University of Munich

In this paper we derive a law of the logarithm for the maximal deviation between a kernel density estimator and the true underlying density function. Extensions to higher derivatives are included. The results are applied to get optimal window-widths with respect to almost sure uniform convergence.

Introduction. Suppose that $\xi_1, \xi_2, \cdots$ are independent and identically distributed (i.i.d.) random variables on the real line, defined over some probability space $(\Omega, \mathscr{A}, \mathbb{P})$. Let F_n denote the empirical distribution function (d.f.) of the sample $\xi_1, \cdots, \xi_n$, i.e. $F_n(x)$ is the relative number of points among $\xi_1, \cdots, \xi_n$ which are less than or equal to x. It follows from the classical Glivenko-Cantelli Theorem that F_n is a consistent estimator of the true d.f. $F(x) := \mathbb{P}(\xi_1 \le x), x \in \mathbb{R}$. In many statistical procedures, however, it is more convenient to formulate hypotheses on F in terms of F^{-1} rather than F itself.
Here

$$F^{-1}(p) := \inf\{x \in \mathbb{R}: F(x) \ge p\}, \quad 0 < p < 1,$$

is the inverse or quantile function pertaining to F. One basic result in this field is the fact that $n^{1/2}(F_n^{-1}(p) - F^{-1}(p))$ is asymptotically normal (see, e.g., Wretman (1978)):

$$n^{1/2}(F_n^{-1}(p) - F^{-1}(p)) \to \mathscr{N}(0, p(1-p)/f^2(F^{-1}(p))) \quad \text{in distribution}$$

(where $\mathscr{N}(0, \sigma^2)$ is a centered normal random variable with variance σ^2). Here f is the derivative of F which is supposed to exist at $F^{-1}(p)$, being continuous and positive there. Hence to obtain proper confidence intervals for testing any hypothesis on the true value of $F^{-1}(p)$, one needs some knowledge of f. Therefore the problem arises of constructing an empirical estimate f_n of the true density f. For such an f_n the value of f_n at $F_n^{-1}(p)$ might then serve as an estimate of the variance of the normal limit distribution.

Of course, the most desired property of such an f_n would be consistency. The following concepts have been extensively studied in the literature: convergence of $f_n(t) \to f(t)$ in probability, $\mathbb{P}$-almost surely, in pth mean, and w.r.t. the mean integrated square error.

Rosenblatt (1956) proposed to study a type of density estimate, which is obtained as a convolution between F_n and a properly scaled kernel function K:

$$f_n(t) := a_n^{-1} \int K\left(\frac{t-x}{a_n}\right) F_n(dx), \quad t \in \mathbb{R}.$$

Here, $(a_n)_n$ is a sequence of (positive) "window-widths" tending to zero as $n \to \infty$. Parzen (1962) showed that under some mild smoothness conditions on K (and f), $f_n(t)$ is in any respect a consistent estimator of $f(t)$ for each $t \in \mathbb{R}$. Much attention has been given to the question whether $f_n \to f$ uniformly on some central portion of the real line. Bickel and Rosenblatt (1973) proved that for sufficiently smooth f and K the variate $\sup_t |f_n(t) - f(t)| / \sqrt{f(t)}$ when properly normalized has an extreme value limit distribution. Woodroofe (1967) investigated uniform convergence in probability.

Received December 1980; revised April 1981.

[1] This research is part of the author's Habilitationsschrift written at the Mathematical Department of the University of Munich.

AMS 1970 *subject classifications*. Primary 62G05, 60F15; secondary 62E20.

Key words and phrases. Empirical distribution function, kernel density estimator, oscillation modulus, higher derivatives, optimal window-widths.

The study of almost sure uniform convergence has been initiated by Nadaraya (1965). He showed that if K is of bounded variation and if f is uniformly continuous, then, as $n \to \infty$,

$$\sup_{-\infty<t<+\infty}|f_n(t) - f(t)| \to 0 \quad \mathbb{P}\text{—a.s.},$$

provided that $\sum_{m\geq 1} \exp[-\gamma m a_m^2] < \infty$ for each $\gamma > 0$. The convergence of this series is needed to estimate the term $\eta_n := \sup_t|f_n(t) - \mathbb{E}(f_n(t))|$, which by integrating on parts may be easily bounded by the variation of K times the maximal deviation between F_n and F (multiplied by a_n^{-1}). In this situation the convergence of the above series is needed to show that, by Borel-Cantelli and a well-known exponential inequality for $\sup_t|F_n(t) - F(t)|$ (see Dvoretzky et al. (1956)), $\eta_n \to 0$ $\mathbb{P}$—a.s.

In almost all of the subsequent papers on this topic the method of proof is essentially the same. A general defect is that possible results do not reflect the shape of f and K, respectively. This stems from the fact that the study of f_n (reflecting the local behaviour of F) is reduced to the estimation of the (global) deviation between F_n and F. Notable exceptions are the papers of Reiss (1975), Révész (1978) and Silverman (1978). See also Kolčinskiĭ (1980). In particular, in Révész (1978), the method of proof is based on a strong approximation result for empirical processes, to the effect that the choice of a_n is limited by the error of this approximation.

In this paper we obtain exact rates of convergence for a large class of kernel density estimators, depending on smoothness properties of f. It is seen that for a.s. uniform convergence, the optimal a_n is close to that obtained in the (simpler) square mean and integrated square mean analysis. The extension to higher derivatives of F is simple. As a main tool we need an (asymptotic) Hölder condition for empirical processes (cf. Stute (1982)). Actually, this property has been already used there to study histograms with random grids (e.g., nearest neighbour estimators) and nonrandom grids (classical case) as well as the "naive" kernel density estimator.

1. Limit Theorems. Recall that the kernel density estimator is defined by

$$f_n(t) = a_n^{-1} \int K\left(\frac{t-x}{a_n}\right) F_n(dx), \quad t \in \mathbb{R},$$

where F_n is the empirical d.f. of the sample $\xi_1, \cdots, \xi_n$, $(a_n)_n$ is a sequence of window-widths tending to zero, and K is a measurable function. In most cases, K is nonnegative and integrates to one. The first assumption guarantees that f_n, as an estimator of the nonnegative density f, is itself nonnegative. The condition $\int K(x)\, dx = 1$ will be needed to treat the bias $B \equiv \mathbb{E}(f_n(t)) - f(t)$. By letting K attain negative values it is always possible to reduce the bias B. This enables one to get better rates of convergence for $f_n(t) \to f(t)$, at the price that $f_n(t)$ might be negative.

In order to obtain local estimates for the empirical d.f. the sequence $(a_n)_n$ has to form a bandsequence as defined in Stute (1982), i.e.

(*) (i) $a_n \downarrow 0$ and $na_n \uparrow \infty$ (ii) $\ln a_n^{-1}/na_n \to 0$ (iii) $\ln a_n^{-1}/\ln \ln n \to \infty$.

Conditions (ii) and (iii) roughly state that a_n ranges between n^{-1} and $1/\ln n$. The deviation $f_n - f$ is measured on some central interval J of the real line. For each $\varepsilon > 0$ and $J = (a, b)$, say, put $J_\varepsilon = (a + \varepsilon, b - \varepsilon)$. To avoid trivial statements, assume $J_\varepsilon \neq \emptyset$ throughout. If $K = 1_{[-1/2,1/2)}$, i.e. when f_n is the naive estimator, equation (4.1) in Stute (1982) implies that with

$$\bar{f}_n(t) := \mathbb{E}(f_n(t)) = a_n^{-1} \int K\left(\frac{t-x}{a_n}\right) F(dx)$$

one has under some mild smoothness and boundedness assumptions on f (on J)

$$\lim_{n\to\infty} \sup_{t\in J_\varepsilon} \frac{\sqrt{na_n}\,|f_n(t) - \bar{f}_n(t)|}{\sqrt{2f(t)\ln a_n^{-1}}} = 1 \quad \mathbb{P}\text{—a.s.} \tag{1.1}$$

In the next theorem (1.1) is generalized to arbitrary step functions

$$K = \sum_{i=1}^m K(c_i)1_{[c_i, c_{i+1})}, \quad c_1 < c_2 < \cdots < c_{m+1}. \tag{1.2}$$

THEOREM 1.1 *Suppose that on* J $f = F'$ *is uniformly continuous with*

$$0 < m \le f(x) \le M < \infty \qquad \textit{for all } x \in J.$$

Then for the kernel K *from* (1.2) *and each bandsequence* $(a_n)_n$ *one has* $\mathbb{P}$—a.s.

$$\lim_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}} \sup_{t\in J_\epsilon} \frac{|f_n(t) - \bar{f}_n(t)|}{\sqrt{f(t)}} = \left(\sum_{i=1}^m K^2(c_i(c_{i+1} - c_i))\right)^{1/2}. \tag{1.3}$$

PROOF. As for the naive kernel (see Stute (1982)) the proof of (1.3) may be reduced, using a quantile transformation, to the case when $\xi_1, \xi_2, \cdots$ are uniformly distributed on [0, 1]. Then one has to prove the following limit result:

$$\lim_{n\to\infty} \sup \frac{|\sum_{i=1}^m K(c_i)[\alpha_n(u_{i+1}) - \alpha_n(u_i)]|}{(\sum_{i=1}^m K^2(c_i)(u_{i+1} - u_i))^{1/2}\sqrt{2a_n \ln a_n^{-1}}} = 1, \tag{1.4}$$

extending Theorem 2.10 in Stute (1982). Here the supremum extends over all $m + 1$-tuples $0 \le u_1 < u_2 < \cdots < u_{m+1} \le 1$ such that $\underline{c}a_n \le u_{i+1} - u_i \le \bar{c}a_n$, where $0 < \underline{c} \le \bar{c} < \infty$ are two preassigned constants and α_n is the uniform empirical process. The lim inf part in (1.4) is easily obtained from poissonization. For this we need a slight generalization of Bohman's (1963) inequality, relating the tails of a standardized sum of independent Poisson variables η_i with parameter λ_i and weights $K(c_i)$, i.e.

$$\eta = \sum_{i=1}^m K(c_i)(\eta_i - \lambda_i)/(\textstyle\sum_i K^2(c_i)\lambda_i)^{1/2}$$

to those of a standard normal one. For the lim sup part, essentially the same method applies as for Lemma 2.6 in Stute (1982). Roughly, the proof is based, using Borel-Cantelli, on a well-known exponential inequality for binomial tails and the fact that the supremum in (1.4) is practically attained on some finite set of $m + 1$-tuples with cardinality $O(a_n^{-1})$. Finally, the right-hand side of (1.3) follows from expanding the asymptotic variance of $f_n(t)$. □

In what follows we may assume w.l.o.g. that $F_n(t) = \bar{F}_n(F(t))$, $t \in \mathbb{R}$, where $\bar{F}_n$ is the empirical d.f. of a uniform sample $\bar{\xi}_1, \cdots, \bar{\xi}_n$. Let

$$\beta_n(t) := \sqrt{n}\,(F_n(t) - F(t)) = \sqrt{n}\,(\bar{F}_n(F(t)) - F(t)) \equiv \alpha_n(F(t))$$

denote the corresponding empirical process, and define

$$\bar{\omega}_n(a) := \sup_{|t-s|\le a} |\alpha_n(t) - \alpha_n(s)|, \quad a > 0,$$

the modulus function of α_n. Finally, let

$$M_0 := \sup_{x,y\in J, x\ne y} |F(x) - F(y)|/|x - y|.$$

Write $F \in \mathrm{Lip}(1, M, J)$ if $M_0 = M$ for some finite M.

THEOREM 1.2. *Let* K *be of bounded variation and suppose that* $K(x) = 0$ *outside some finite interval* $[r, s)$. *If* $F \in \mathrm{Lip}(1, M, J)$, *then for each* $\varepsilon > 0$ *and every bandsequence* $(a_n)_n$:

$$\limsup_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}} \sup_{t\in J_\epsilon} |f_n(t) - \bar{f}_n(t)| = C \quad \mathbb{P}\text{—a.s.} \tag{1.5}$$

where C *is a constant less than or equal to* $\sqrt{M(s - r)}\, V_r^s(K)$ *and* $V_r^s(K)$ *is the total variation of* K *over* $[r, s]$.

PROOF. Integration by parts yields

$$f_n(t) - \bar{f}_n(t) = a_n^{-1} \int K\left(\frac{t-x}{a_n}\right)(F_n - F)(dx)$$

$$= -a_n^{-1} \int_{(t-sa_n, t-ra_n]} [F_n(x-) - F(x-) - F_n(t - ra_n) + F(t - ra_n)]\, dK\left(\frac{t-x}{a_n}\right),$$

whence for all large $n \in \mathbb{N}$

$$\sqrt{na_n}\, \sup_{t \in J_\varepsilon} |f_n(t) - \bar{f}_n(t)| \le a_n^{-1/2} \sup_{|x-y| \le (s-r)a_n, x,y \in J} |\beta_n(x) - \beta_n(y)|\, V_r^s(K)$$

$$\le a_n^{-1/2} \bar{\omega}_n(M(s-r)a_n) V_r^s(K).$$

The finiteness of the left-hand side in (1.5) now follows from Theorem 2.14 in Stute (1982), while the existence of C is an immediate consequence of the Hewitt-Savage zero-one law. □

As may be seen from Theorem 1.1, (1.5) provides the exact rate of convergence. If F is Lipschitz on some neighbourhood of J the assertion also holds with $\varepsilon = 0$. Finally, if F is Lipschitz of order $0 < \beta < 1$, (1.5) is still valid with a_n replaced by a_n^β.

We are now going to show that for kernel functions with finite support, the constant C in (1.5) may be identified if in addition $f = F'$ exists and is sufficiently smooth.

THEOREM 1.3. *Suppose that $f = F'$ is uniformly continuous on J with $0 < m \le f(x) \le M < \infty$ for all $x \in J$. Let K be any kernel function of bounded variation with $K(x) = 0$ outside some finite interval $[r, s)$. Then with probability one*

$$\lim_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}}\, \sup_{t \in J_\varepsilon} \frac{|f_n(t) - \bar{f}_n(t)|}{\sqrt{f(t)}} = \left(\int_r^s K^2(x)\, dx\right)^{1/2}.$$

PROOF. Given $h > 0$ let $r = c_1 < c_2 < \cdots < c_{m+1} = s$ be any partition of $[r, s]$ with mesh $\max_{i=1,\dots,m}(c_{i+1} - c_i) \le h$. Define

$$K_i(y) := \begin{cases} K(y) - K(c_i), & c_i \le y < c_{i+1} \\ 0, & \text{otherwise}, i = 1, \cdots, m \end{cases}$$

$$K^1(y) := \textstyle\sum_{i=1}^m K_i(y),$$

and

$$K^0(y) := \begin{cases} K(c_i), & c_i \le y < c_{i+1} \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, $K = K^0 + K^1$. Let f_n^0 and f_n^1 denote the density estimators with respect to K^0 and K^1, so that $f_n = f_n^0 + f_n^1$. Theorem 1.2 then yields $\mathbb{P}$-almost surely

$$\limsup_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}}\, \sup_{t \in J_\varepsilon} |f_n^1(t) - \bar{f}_n^1(t)| \le \sqrt{Mh} \textstyle\sum_{i=1}^m V_{c_i}^{c_{i+1}}(K_i) \tag{1.6}$$

$$\le 2\sqrt{Mh}\, V_r^s(K).$$

For f_n^0 we obtain from Theorem 1.1 $\mathbb{P}$-almost surely

$$\lim_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}}\, \sup_{t \in J_\varepsilon} \frac{|f_n^0(t) - \bar{f}_n^0(t)|}{\sqrt{f(t)}} = \left(\textstyle\sum_{i=1}^m K^2(c_i)(c_{i+1} - c_i)\right)^{1/2}. \tag{1.7}$$

Together with (1.6) this gives

$$\left(\textstyle\sum_{i=1}^m K^2(c_i)(c_{i+1} - c_i)\right)^{1/2} - 2(Mh/m)^{1/2} V_r^s(K)$$

$$\le \liminf_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}}\, \sup_{t \in J_\varepsilon} \frac{|f_n(t) - \bar{f}_n(t)|}{\sqrt{f(t)}}$$

$$\le \limsup_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}}\, \sup_{t \in J_\varepsilon} \frac{|f_n(t) - \bar{f}_n(t)|}{\sqrt{f(t)}}$$

$$\le \left(\textstyle\sum_{i=1}^m K^2(c_i)(c_{i+1} - c_i)\right)^{1/2} + 2(Mh/m)^{1/2} V_r^s(K).$$

Since K^2 is Riemann-integrable, the first square root converges to $(\int_r^s K^2(x)\,dx)^{1/2}$ as $h \to 0$, while the second tends to zero. This proves Theorem 1.3. □

In the above results, the finiteness of $[r, s)$ was essential in order to relate the deviation between f_n and $\bar{f}_n$ to the oscillation modulus of α_n. We shall see that for kernel functions with unbounded support, the rate of convergence depends on the tail behaviour of K. For this, suppose that $R_n \uparrow \infty$ is some sequence of real numbers such that $(R_n a_n)_n$ is a bandsequence. We may then write

$$f_n(t) - \bar{f}_n(t) = a_n^{-1}\left[\int_{-R_n a_n < x-t \le R_n a_n} \mathcal{U}(x)(F_n - F)(dx) + \int_{-\infty < x-t \le -R_n a_n} \mathcal{U}(x)(F_n - F)(dx) + \int_{R_n a_n < x-t<\infty} \mathcal{U}(x)(F_n - F)(dx)\right],$$

where $\mathcal{U}(x) = K\left(\dfrac{t-x}{a_n}\right)$. Assume that K is of bounded variation with

$$K(x) \to 0 \quad \text{as } x \to \pm\infty.$$

Integration by parts yields

$$\begin{aligned} f_n(t) - \bar{f}_n(t) = a_n^{-1}n^{-1/2}\Big\{ & K(R_n)[\beta_n(t + R_n a_n) - \beta_n(t - R_n a_n)] \\ & + \int_{-R_n a_n < x-t \le R_n a_n} [\beta_n(t + R_n a_n) - \beta_n(x-)]\,\mathcal{U}(dx) + K(R_n)\beta_n(t - R_n a_n) \\ & - \int_{-\infty < x-t \le -R_n a_n} \beta_n(x-)\,\mathcal{U}(dx) - K(-R_n)\beta_n(t + R_n a_n) \\ & - \int_{R_n a_n < x-t<\infty} \beta_n(x-)\,\mathcal{U}(dx)\Big\}, \end{aligned}$$

whence (for all large n)

$$\begin{aligned} \sqrt{na_n}\, \sup_{t\in J_\varepsilon}|f_n(t) - \bar{f}_n(t)| \le a_n^{-1/2}\{ & V_{-\infty}^{\infty}(K) \sup_{|x-y|\le 2R_n a_n, x,y\in J}|\beta_n(x) - \beta_n(y)| \\ & + [V_{-\infty}^{-R_n}(K) + V_{R_n}^{+\infty}(K) + |K(R_n)| + |K(-R_n)|] \sup_{x\in\mathbb{R}}|\beta_n(x)|\}. \end{aligned} \tag{1.8}$$

Assume that for some positive x_0, K is nonincreasing on $[x_0, \infty)$ and nondecreasing on $(-\infty, -x_0]$. Then for all large n

$$V_{-\infty}^{-R_n}(K) = K(-R_n) \qquad \text{and} \qquad V_{R_n}^{\infty}(K) = K(R_n).$$

If $F \in \mathrm{Lip}(1, M, J)$, (1.8) implies

$$\begin{aligned} &\sqrt{na_n}\, \sup_{t\in J_\varepsilon}|f_n(t) - \bar{f}_n(t)| \\ &\quad \le a_n^{-1/2}\{V_{-\infty}^{+\infty}(K)\bar{\omega}_n(2MR_n a_n) + (2|K(R_n)| + 2|K(-R_n)|) \sup_{x\in\mathbb{R}}|\beta_n(x)|\}. \end{aligned}$$

Using the fact that $(R_n a_n)_n$ is a bandsequence, we get from Theorem 2.14 in Stute (1982) and the Chung-Smirnov LIL for β_n

$$\begin{aligned} &\limsup_{n\to\infty} \sqrt{\frac{na_n}{2R_n \ln (R_n a_n)^{-1}}}\, \sup_{t\in J_\varepsilon}|f_n(t) - \bar{f}_n(t)| \\ &\quad \le (2M)^{1/2}V_{-\infty}^{\infty}(K) + o(1)[|K(R_n)| + |K(-R_n)|](R_n a_n)^{-1/2}. \end{aligned} \tag{1.9}$$

The best possible rate of convergence is obtained by choosing R_n so that it is asymptotically equivalent to the implicit solution of the equation

$$|K(x)| + |K(-x)| = (xa_n)^{1/2}.$$

Under some mild growth conditions on K, this will produce a bandsequence $(R_n a_n)_n$ for which our previous estimates apply. In particular we see that kernel functions with short tails give better rates and are thus preferable. In Theorem 1.4 and 1.5 below we consider the case when $K(x)$ falls as $|x|^{-\lambda}$ or $\exp(-\lambda|x|)$, $\lambda > 0$, respectively.

THEOREM 1.4. *Let $F \in \mathrm{Lip}(1, M, J)$ and suppose that K is of bounded variation with $|K(x)| = O(|x|^{-\lambda})$ as $|x| \to \infty$ for some $\lambda > 0$. Then we have $\mathbb{P}$-almost surely*

$$\limsup_{n\to\infty} \sqrt{\frac{na_n^\beta}{\ln a_n^{-1}}}\, \sup_{t\in J_\varepsilon} |f_n(t) - \bar{f}_n(t)| = 0,$$

where $\beta = 1 + \dfrac{1}{2\lambda + 1}$.

PROOF. For any $\delta > 0$, put $R_n = \delta a_n^{-1/(2\lambda+1)}$. It is easy to see that with this choice of R_n, the sequence $(R_n a_n)_n$ is a bandsequence for which (1.9) applies. Check that $\ln(R_n a_n)^{-1} \sim \dfrac{2\lambda}{2\lambda + 1} \ln a_n^{-1}$ and use the fact that $\delta > 0$ was arbitrary to get the result. □

We do not know whether Theorem 1.4 provides the exact rate of convergence if $|K(x)| \sim |x|^{-\lambda}$, $|x| \to \infty$.

If K is of exponential degree, (1.9) implies the following theorem.

THEOREM 1.5. *Under the assumptions of Theorem 1.4, suppose that $|K(x)| = O(\exp(-\lambda|x|))$ as $|x| \to \infty$ for some $\lambda > 0$. Then $\mathbb{P}$*—a.s.

$$\limsup_{n\to\infty} \sqrt{\frac{\lambda n a_n}{\ln^2 a_n^{-1}}}\, \sup_{t\in J_\varepsilon} |f_n(t) - \bar{f}_n(t)| = C,$$

where $C \le (2M)^{1/2}\, V_{-\infty}^{\infty}(K)$.

PROOF. Put $R_n = (2\lambda)^{-1} \ln a_n^{-1}$. The assertion is now an immediate consequence of (1.9) and the Hewitt-Savage zero-one law. □

2. Optimal window-widths. To decide whether f_n is a good estimator of f, it remains to study the bias $B = \overset{*}{f}_n(t) - f(t)$. For $B \to 0$, it suffices to assume that f is continuous in t (cf. Parzen (1962)). To get further information on the rate of convergence, it is necessary to make some smoothness assumptions on f. Taylor expansion of the integrand then yields the following simple result.

LEMMA 2.1. *Suppose that $f = F'$ is $m + 1$-times continuously differentiable on J. Let K be any kernel function vanishing outside some finite interval $[r, s)$ such that*

(i) $\displaystyle\int K(x)\, dx = 1$

and

(ii) $\displaystyle\int K(x)x^i\, dx = 0$ for all $i = 1, \cdots, m$.

Then for all large enough n we have

$$\sup_{t\in J_\varepsilon} |\bar{f}_n(t) - f(t)| \le \frac{a_n^{m+1}}{(m+1)!} \sup_{x\in J} |f^{(m+1)}(x)| \int |K(y)y^{m+1}|\, dy. \tag{2.1}$$

Lemma 2.1 shows that for sufficiently smooth F's it is always possible to reduce the bias $B = \overset{*}{f}_n(t) - f(t)$ upon choosing appropriate kernels K. In the class of all nonnegative K's, condition (ii) can be achieved only for $m = 1$, thus giving a_n^2 as the best possible error rate. For better results, one has to include also those K's for which $K(y)$ may be negative. However, this is at the price that $f_n(t)$ as an estimate of $f(t)$ may be negative, too. For this reason most authors restrict themselves to nonnegative K's. For such a K the equation

$\int K(x)x\,dx = 0$ is always satisfied if the interval $[r, s)$ is symmetric about zero with $K(x) = K(-x)$ for all $x \in (r, s)$. In particular, this is true for the naive kernel $K = 1_{[-1/2,1/2)}$.

Theorem 1.3 and Lemma 2.1 imply that for small values of a_n the error $E_1 = f_n - \bar{f}_n$ becomes large, while the bias $B = \bar{f}_n - f$ decreases with a_n. In fact, B is asymptotically negligible if $a_n^2 = o((\ln a_n^{-1}/na_n)^{1/2})$, i.e. when $na_n^5/\ln a_n^{-1} \to 0$, and in this case we have with probability one

$$\lim_{n\to\infty} \sqrt{\frac{na_n}{2 \ln a_n^{-1}}}\, \sup_{t\in J_\varepsilon} \frac{|f_n(t) - f(t)|}{\sqrt{f(t)}} = \left(\int K^2(x)\,dx\right)^{1/2}. \tag{2.2}$$

If f is only once differentiable, the same argument as in the proof of Lemma 2.1 yields a failure rate $B = O(a_n)$. In this case the bias is asymptotically negligible if $na_n^3/\ln a_n^{-1} \to 0$. Under less restrictive smoothness assumptions on F, a corresponding lim sup result may be obtained from Theorem 1.2. It is also noted that $na_n^5/\ln a_n^{-1} \to 0$ automatically implies the third condition for $(a_n)_n$ to be a bandsequence (cf.(*)).

If $na_n^5/\ln a_n^{-1} \to \infty$, the stochastic behaviour of $f_n - f$ is determined by that of $\bar{f}_n - f$. To get a limit result one has to assume that f has a third continuous derivative f'''. In this case

$$\bar{f}_n(t) - f(t) = \frac{a_n^2}{2} f''(t) \int K(y)y^2\,dy + O(a_n^3) \quad \text{uniformly in } t \in J_\varepsilon \tag{2.3}$$

(if f''' and K are bounded). From Theorem 1.2 we therefore get with probability one

$$\lim_{n\to\infty} 2a_n^{-2} \sup_{t\in J_\varepsilon} |f_n(t) - f(t)| = \sup_{t\in J_\varepsilon} |f''(t)| \int K(y)y^2\,dy.$$

It is easy to check that $na_n^5/\ln a_n^{-1} \to \infty$ is stronger than (ii) in (*). For a lim sup result it is sufficient to assume the existence of f''. If $na_n^3/\ln a_n^{-1} \to \infty$, only the first derivative of f is needed.

If $f(x) \geq m > 0$ on J and f is smooth enough, the optimal a_n is obtained from (2.2) and (2.3) by minimizing the term

$$a_n^2 \sup_{t\in J_\varepsilon} \frac{|f''(t)|}{\sqrt{f(t)}} \cdot \frac{1}{2} \int K(y)y^2\,dy + \left(\frac{2 \ln a_n^{-1}}{na_n} \int K^2(y)\,dy\right)^{1/2}. \tag{2.4}$$

Differentiation w.r.t. a_n shows that one has to solve the equation

$$x^5 = \left[\frac{c_2}{4c_1}\right]^2 ((\ln x^{-1})^{-1/2} + (\ln x^{-1})^{1/2})^2 \equiv d_n((\ln x^{-1})^{-1/2} + (\ln x^{-1})^{1/2})^2, \tag{2.5}$$

where

$$c_1 = \sup_{t\in J_\varepsilon} \frac{|f''(t)|}{2\sqrt{f(t)}} \int K(y)y^2\,dy,$$

$$c_2 = \left(2 \int K^2(y)\,dy/n\right)^{1/2}.$$

Since for small values of x the right-hand side of (2.5) is asymptotically equal to $d_n \ln x^{-1}$, we try to solve the equation.

$$x^5/\ln x^{-1} = d_n. \tag{2.6}$$

If a_n denotes the solution of (2.6), $a_n \leq e^{-1}$ and $na_n \geq 1$ for all large n. Hence

$$d_n^{1/5} \leq a_n \leq (d_n \ln n)^{1/5}.$$

This implies

$$\ln a_n^{-1} \sim \frac{1}{5} \ln d_n^{-1} \sim \frac{1}{5} \ln n$$

so that

$$a_n \sim \left[\frac{\int K^2(y)\,dy}{10\,\sup_{t\in J_\varepsilon}\frac{|f''(t)|^2}{f(t)}\left(\int K(y)y^2\,dy\right)^2}\;\frac{\ln n}{n}\right]^{1/5}.$$

In most situations K is a preassigned kernel function for which the above integrals are commonly known. However, the dependence on f (and f'') makes it impossible to determine the exact order of a_n. In practice, one could try to estimate the sup in the above denominator and then use the resulting estimate as a basis for the definition of a_n. The estimation of f'', or more generally, of any derivative of F, needs no extra efforts and will be discussed below. Anyway, putting $a_n^* := (\ln n/n)^{1/5}$, this will give us the appropriate rate at which the bandsequence should converge to zero. We obtain that uniformly on J_ε, $(f_n - f)/f^{1/2}$ converges to zero as $(\ln n/n)^{2/5}$.

3. Estimation of higher derivatives. Theorem 1.3 states that, as for the naive density-estimator, the asymptotic behaviour of f_n is closely related to the local behaviour of the empirical distribution function, if K has finite support. Among these K's, some have certain advantages due to their distinguished smoothness properties. Actually, such a property plays an important role when estimating higher derivatives of F.

In the following, assume that K vanishes outside some finite interval $[r, s]$. Furthermore, suppose that K has an rth derivative of bounded variation. Define

$$f_n^{(r)}(t) = a_n^{-(r+1)} \int K^{(r)}\left(\frac{t-x}{a_n}\right) F_n\,(dx)$$

and

$$\bar{f}_n^{(r)}(t) = \mathbb{E}(f_n^{(r)}(t)) = a_n^{-(r+1)} \int K^{(r)}\left(\frac{t-x}{a_n}\right) F\,(dx).$$

If F satisfies the assumptions of Theorem 1.3 we get (with K replaced by $K^{(r)}$) $\mathbb{P}$—a.s.

$$\lim_{n\to\infty} a_n^r \sqrt{\frac{na_n}{2\ln a_n^{-1}}}\,\sup_{t\in J_\varepsilon}\frac{|f_n^{(r)}(t) - \bar{f}_n^{(r)}(t)|}{\sqrt{f(t)}} = \left(\int [K^{(r)}(x)]^2\,dx\right)^{1/2}. \tag{3.1}$$

If $f = F'$ has a derivative of order r, it is easy to see that

$$a_n^{-(r+1)}K^{(r)}\left(\frac{t-x}{a_n}\right)f(x) = a_n^{-1}K\left(\frac{t-x}{a_n}\right)f^{(r)}(x) - \frac{d}{dx}\sum_{s=0}^{r-1} a_n^{-(s+1)}K^{(s)}\left(\frac{t-x}{a_n}\right)f^{(r-s-1)}(x).$$

Since $K^{(s)}$, $s = 0, 1, \cdots, r-1$ vanish outside $[r, s]$, we get

$$\bar{f}_n^{(r)}(t) = a_n^{-1}\int K\left(\frac{t-x}{a_n}\right)f^{(r)}(x)\,dx.$$

This is the key equation by which one may be motivated to consider $f_n^{(r)}$ as an estimator of $f^{(r)}$. It was apparently introduced by Bhattacharya (1967) who obtained first results on the strong consistency of $f_n^{(r)}$. In most of the subsequent papers it was not verbally expressed by the authors that the study of $f_n^{(r)}$ presents no new difficulties compared with f_n. In fact, if $f^{(r)}$ satisfies the same assumptions as f in Lemma 2.1, we get

$$\sup_{t\in J_\varepsilon}|\bar{f}_n^{(r)}(t) - f^{(r)}(t)| \le \frac{a_n^{m+1}}{(m+1)!}\sup_{x\in J}|f^{(r+m+1)}(x)|\int |K^{(r)}(y)y^{m+1}|\,dy. \tag{3.2}$$

From (3.1) and (3.2) we see that the optimal a_n is of order $[\ln n/n]^{\frac{1}{2(r+m+1)+1}}$. For this a_n

$$\sup_{t\in J_\varepsilon}|f_n^{(r)}(t) - f^{(r)}(t)| = O([\ln n/n]^{(m+1)/(2(r+m+1)+1)})\ \mathbb{P}\text{—a.s.}$$

By putting further assumptions on the sequence $(a_n)_n$, it is straightforward to prove analogons to (2.2).

Clearly, for m fixed, the rate of convergence decreases with r. We also see that for fixed r and $\delta > 0$, it is always possible to find a K such that

$$\sup_{t\in J_\varepsilon} | f_n^{(r)}(t) - f^{(r)}(t) | = O(n^{-1/2+\delta}) \ \mathbb{P}\text{—a.s.},$$

(if f is smooth enough). However, this is only possible if one admits negative values of K. For $r \geq 1$, there is no reason to exclude such K's from consideration. In summary, this shows that one should take different kernels for f_n and $f_n^{(r)}$, if $r \geq 1$. It is also possible to define $f_n^{(r)}$ (see Singh (1977)) by

$$f_n^{(r)}(t) = a_n^{-(r+1)} \int K\left(\frac{t-x}{a_n}\right) F_n(dx),$$

where in this case K is a kernel fulfilling

$$(i!)^{-1} \int K(x) x^i \, dx = \begin{cases} 1 & \text{if } i = r \\ 0 & \text{if } i = 0, 1, \cdots, m, \ i \neq r, \end{cases} \tag{3.3}$$

$m \geq r$. In this setup, no differentiability of K is assumed. Condition (3.3) is needed in order to filter out $f^{(r)}(t)$ in the Taylor expansion of $\bar{f}_n^{(r)}(t)$. The resulting estimates are conceptually the same as before and improve on those of Singh.

REFERENCES

[1] BHATTACHARYA, P. K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā (A)* **29** 373–382.

[2] BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function etimates. *Ann. Statist.* **1** 1071–1095.

[3] BOHMAN, H. (1963). Two inequalities for Poisson distributions. *Skand. Aktuarietidskr.* **46** 47–52.

[4] DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642–669.

[5] KOLČINSKIĬ, V. I. (1980). Some limit theorems for empirical measures. *Theor. Probability and Math. Statist.* **21** 79–86.

[6] NADARAYA, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theor. Probability Appl.* **10** 186–190.

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

[8] REISS, R. D. (1975). Consistency of a certain class of empirical density functions. *Metrika* **22** 189–203.

[9] RÉVÉSZ, P. (1978). A strong law of the empirical density function. *Periodica Math. Hung.* **9** 317–324.

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

[11] SCOTT, D. W., TAPIA, R. A. and THOMPSON, J. R. (1977). Kernel density estimation revisited. *Nonlinear Anal., Theor., Math. Appl.* **1** 339–372.

[12] SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6** 177–184.

[13] SINGH, R. S. (1977). Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density. *Ann. Statist.* **5** 394–399.

[14] STUTE, W. (1982). The oscillation behaviour of empirical processes. *Ann. Probability* **10** 86–107.

[15] WOODROOFE, M. (1967). On the maximum deviation of the sample density. *Ann. Math. Statist.* **38** 475–481.

[16] WRETMAN, J. (1978). A simple derivation of the asymptotic distribution of a sample quantile. *Scand. J. Statist.* **5** 123–124.

FACHBEREICH 6
UNIVERSITY OF SIEGEN
HOELDERLINSTR. 3
D-5900 SIEGEN 21
WEST GERMANY