The Annals of Statistics
1980, Vol. 8, No. 6, 1348–1360

OPTIMAL RATES OF CONVERGENCE FOR NONPARAMETRIC ESTIMATORS[1]

BY CHARLES J. STONE

University of California, Los Angeles

Let d denote a positive integer, $\|x\| = (x_1^2 + \cdots + x_d^2)^{1/2}$ the Euclidean norm of $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, k a nonnegative integer, $\mathscr{C}_k$ the collection of k times continuously differentiable functions on $\mathbb{R}^d$, and g_k the Taylor polynomial of degree k about the origin corresponding to $g \in \mathscr{C}_k$. Let M and $p > k$ denote positive constants and let U be an open neighborhood of the origin of $\mathbb{R}^d$. Let $\mathscr{G}$ denote the collection of functions $g \in \mathscr{C}_k$ such that $|g(x) - g_k(x)| \leq M\|x\|^p$ for $x \in U$. Let $m \leq k$ be a nonnegative integer, let $\theta_0 \in \mathscr{C}_m$ and set $\Theta = \{\theta_0 + g : g \in \mathscr{G}\}$. Let L be a linear differential operator of order m on $\mathscr{C}_m$ and set $T(\theta) = L\theta(0)$ for $\theta \in \Theta$. Let (X, Y) be a pair of random variables such that X is $\mathbb{R}^d$ valued and Y is real valued. It is assumed that the distribution of X is absolutely continuous and that its density is bounded away from zero and infinity on U. The conditional distribution of Y given X is assumed to be (say) normal, with a conditional variance which is bounded away from zero and infinity on U. The regression function of Y on X is assumed to belong to Θ. It is shown that $r = (p - m)/(2p + d)$ is the optimal (uniform) rate of convergence for a sequence $\{\hat{T}_n\}$ of estimators of $T(\theta)$ such that $\hat{T}_n$ is based on a random sample of size n from the distribution of (X, Y). An analogous result is obtained for nonparametric estimators of a density function.

1. Introduction. Let Θ denote a collection of functions on a fixed subset of $\mathbb{R}^d$. Let $T(\theta)$, $\theta \in \Theta$, be a real valued functional on Θ. Consider an unknown distribution which depends on $\theta \in \Theta$. Let $\{\hat{T}_n\}$ denote a sequence of estimators of $T(\theta)$ such that $\hat{T}_n$ is based on a random sample of size n from the unknown distribution. Let r denote a positive number. Then r is called an *upper bound to the rate of convergence* if for every sequence $\{\hat{T}_n\}$ of estimators

$$\text{(1.1)} \qquad \liminf_n \sup_{\theta \in \Theta} P_\theta(|\hat{T}_n - T(\theta)| > cn^{-r}) > 0 \qquad \text{for all } c > 0$$

and

$$\text{(1.2)} \qquad \lim_{c \to 0} \liminf_n \sup_{\theta \in \Theta} P_\theta(|\hat{T}_n - T(\theta)| > cn^{-r}) = 1.$$

Also r is called an *achievable rate of convergence* if there is a sequence $\{\hat{T}_n\}$ of estimators such that

$$\text{(1.3)} \qquad \lim_{c \to \infty} \limsup_n \sup_{\theta \in \Theta} P_\theta(|\hat{T}_n - T(\theta)| > cn^{-r}) = 0;$$

r is called the *optimal rate of convergence* if it is both an upper bound to the rate of convergence and an achievable rate of convergence. (Note that if r is an upper bound to the rate of convergence and s is an achievable rate of convergence, then $s \leq r$.) If Θ is a collection of functions on a finite subset of $\mathbb{R}^d$, then under appropriate regularity conditions, $r = ½$ is the well-known optimal rate of convergence. From now on Θ will denote a collection of functions on all of $\mathbb{R}^d$.

Let $\alpha = (\alpha_1, \cdots, \alpha_d)$ denote a d-tuple of nonnegative integers and set $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $\alpha! = \alpha_1! \cdot \cdots \cdot \alpha_d!$. For $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$ set $x^\alpha = x_1^{\alpha_1} \cdot \cdots \cdot x_d^{\alpha_d}$. Let D^α denote the differential operator defined by

Received June 1978; revised June 1979.

[1] Research was supported by NSF Grant MCS 77-02121.

AMS 1980 *subject classifications.* Primary 62G20, secondary 62G05.

Key words and phrases. Optimal rate of convergence, nonparametric estimator, regression function, density function.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.*
www.jstor.org

$$D^\alpha = \frac{\partial^{\alpha_1 + \cdots + \alpha_d}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Let k denote a nonnegative integer and let $\mathscr{C}_k$ denote the collection of k times continuously differentiable real valued functions g on $\mathbb{R}^d$. Given $g \in \mathscr{C}_k$, its Taylor polynomial g_k of degree k about the origin is defined by

$$g_k(x) = \sum_{|\alpha| \le k} \frac{1}{\alpha!} D^\alpha g(0) x^\alpha.$$

Let $p > k$ and M denote positive constants and let U denote an open neighborhood of the origin of $\mathbb{R}^d$. Let $\mathscr{G}$ denote the collection of functions $g \in \mathscr{C}_k$ such that

$$|g(x) - g_k(x)| \le M \|x\|^p, \qquad x \in U, \tag{1.4}$$

where $\|x\| = (x_1^2 + \cdots + x_d^2)^{1/2}$. (If U is convex, p is a positive integer and $k = p - 1$, then (1.4) is implied by an appropriate boundedness condition on the restriction to U of the pth derivative of g.)

Let $m \le k$ denote a nonnegative integer and let

$$L = \sum_{|\alpha| \le m} C_\alpha D^\alpha$$

denote a linear differential operator of order m on $\mathscr{C}_m$; i.e., such that $C_\alpha \ne 0$ for some α with $|\alpha| = m$, where C_α is a real constant for $|\alpha| \le m$. Let Θ denote a collection of functions $g \in \mathscr{C}_m$. Define the functional T on Θ by $T(\theta) = L\theta(0)$. The functional $T(\theta) = \theta(0)$ corresponds to a differential operator of order $m = 0$, while

$$T(\theta) = \frac{\partial \theta}{\partial x_1}(0)$$

corresponds to a differential operator of order $m = 1$.

Model 1 (*Unknown regression function*). Let θ_0 be a fixed function in $\mathscr{C}_m$ and set $\Theta = \{\theta_0 + g : g \in \mathscr{G}\}$. Let (X, Y) be a pair of random variables such that X is $\mathbb{R}^d$ valued and Y is real valued. It is assumed that the distribution of X is absolutely continuous and that its density f is bounded away from zero and infinity on U. The regression function $\theta(x) = E(Y \mid X = x)$, $x \in \mathbb{R}^d$, is assumed to be an unknown member of Θ. The conditional variance $\sigma^2(x) = \operatorname{Var}(Y \mid X = x)$, $x \in \mathbb{R}^d$, is assumed to be bounded away from zero and infinity on U.

The conditional distribution of Y given $X = x \in U$ is assumed to be of the form $f(y \mid x, \theta(x))\varphi(dy)$, where φ is a measure on $\mathbb{R}$. It is assumed that $f(y \mid x, t)$ is strictly positive and jointly measurable in x, y and t as x varies over U, y varies over the support of φ and t varies over an open interval containing $\{\theta(x) : \theta \in \Theta \text{ and } x \in U\}$; and that

$$\int y f(y \mid x, t) \varphi(dy) = t$$

for t in the indicated open interval. It is further assumed that $f(y \mid x, t)$ is twice continuously differentiable on the indicated domain and that the equation

$$\int f(y \mid x, t) \varphi(dy) = 1$$

can be differentiated to yield

$$\int f'(y \mid x, t) \varphi(dy) = 0$$

and

$$\int f''(y \mid x, t) \varphi(dy) = 0,$$

where $'$ and $''$ denote differentiation with respect to t. Set $l(y\,|\,x, t) = \log f(y\,|\,x, t)$. It is finally assumed that there are positive constants ϵ_0 and C and there is a measurable function $M(y\,|\,x, t)$ such that on the indicated domain

$$|\,l''(y\,|\,x, t+\epsilon)\,| \le M(y\,|\,x, t), \qquad \epsilon \le \epsilon_0,$$

and

$$\int M(y\,|\,x, t) f(y\,|\,x, t)\varphi(dy) \le C.$$

These conditions on $f(y\,|\,x, t)$ are satisfied in each of the following five examples. In the last four examples $f(y\,|\,t) = f(y\,|\,x, t)$ is independent of x, but the conditional distribution $f(y\,|\,\theta(x))\varphi(dy)$ of Y given $X = x$ still depends on x.

Example 1 (*Normal*). Let

$$f(y\,|\,x, t) = \frac{1}{\sigma(x)(2\pi)^{1/2}}\, e^{-(y-t)^2/2\sigma^2(x)}, \qquad t \in \mathbb{R},$$

where φ is Lebesgue measure on $\mathbb{R}$ and σ is positive and bounded away from zero and infinity on U.

Example 2 (*Exponential*). Let

$$f(y\,|\,t) = \frac{1}{t}\, e^{-y/t},$$

where φ is Lebesgue measure on $[0, \infty)$ and t ranges over a relatively compact open subinterval of $(0, \infty)$ (i.e., an open interval I whose closure is a compact subset of $(0, \infty)$, $\{\theta(x) : \theta \in \Theta$ and $x \in U\}$ being required to be contained in I.)

Example 3 (*Poisson*). Let

$$f(y\,|\,t) = \frac{t^y e^{-t}}{y!},$$

where φ is counting measure on the set $\mathbb{Z}^+$ of nonnegative integers and t ranges over a relatively compact open subinterval of $(0, \infty)$.

Example 4 (*Geometric*). Let

$$f(y\,|\,t) = \left(\frac{1}{1+t}\right)\left(\frac{t}{1+t}\right)^y,$$

where φ is counting measure on $\mathbb{Z}^+$ and t ranges over a relatively compact open subinterval of $(0, \infty)$.

Example 5 (*Bernoulli*). Let

$$f(y\,|\,t) = t^y(1-t)^{1-y},$$

where φ is counting measure on $\{0, 1\}$ and t ranges over a relatively compact open subinterval of $(0, 1)$.

In the context of Model 1, $\hat{T}_n$ is an estimator of $T(\theta)$ based on $(X_1, Y_1), \cdots, (X_n, Y_n)$, where $(X_1, Y_1), (X_2, Y_2), \cdots$ are independent pairs of random variables each having the same distribution as (X, Y).

Model 2 (*Unknown density function*). Let θ_0 be a fixed probability density function in $\mathscr{C}_m$ such that $\theta_0(0) > 0$. Set $\Theta = \{\theta_0(1+g) : g \in \mathscr{G}, |\,g\,| \le 1$ on $\mathbb{R}^d$ and $\int \theta_0 g\, dx = 0\}$. Let X be an $\mathbb{R}^d$ valued random variable having unknown density $\theta \in \Theta$. In the context of this model, $\hat{T}_n$

is an estimator of $T(\theta)$ based on $X_1, \cdots, X_n$, where $X_1, X_2, \cdots$ are independent random variables each having the same distribution as X.

THEOREM. *Let Model 1 or Model 2 hold. Then $r = (p - m)/(2p + d)$ is the optimal rate of convergence.*

A number of observations concerning this theorem are in order, starting with Model 1. The proof is given in Section 2. The estimator that is used to show that the indicated rate r is achievable will now be described. Set $\gamma = 1/(2p + d)$. Let $\{\epsilon_n\}$ be a sequence of positive numbers satisfying either of the following two conditions:

(i) ϵ_n is nonrandom and $0 < \lim_n n^\gamma \epsilon_n < \infty$;

(ii) ϵ_n is the N_nth smallest value among $\| X_1 \|, \cdots, \| X_n \|$, where $\{N_n\}$ is a sequence of nonrandom positive integers such that $0 < \lim_n n^{-2\gamma p} N_n < \infty$.

Set $I_n = \{i: 1 \leq i \leq n \text{ and } \| X_i \| \leq \epsilon_n\}$. Let $\hat{\theta}_{kn}$ denote the polynomial on $\mathbb{R}^d$ of degree k which minimizes

$$\sum_{I_n} [(Y_i - \hat{\theta}_{kn}(X_i))^2]$$

and set $\hat{T}_n = L\hat{\theta}_{kn}(0)$. Estimators of this type have been considered by Stone (1975), (1977) and Cleveland (1979). The proof that this estimator has rate of convergence r does not depend on the assumptions stated above on the conditional distribution of Y given X. It does not depend on the assumption that σ^2 is bounded away from zero on U. It does depend on the assumption that σ^2 is bounded on U, but if σ^2 approaches infinity at the origin, r is probably not achievable. The proof depends on the assumption that the marginal density f of X is bounded away from zero and infinity on U, but if f approaches zero at the origin r is probably not achievable. (If f approaches infinity at the origin, r is probably not an upper bound to the rate of convergence.)

The proof that r is an upper bound to the rate of convergence does not depend on the assumption that σ^2 is bounded on U. It does depend on the assumption that σ^2 is bounded away from zero on U, but if σ^2 approaches zero at the origin, r is probably not an upper bound to the rate of convergence. The proof does not depend on the assumption that f is bounded away from zero on U. It does depend on the (apparently necessary as noted above) assumption that f is bounded on U. The proof depends on the assumptions regarding the conditional distribution of Y given X. These assumptions can obviously be dropped, however, if the conditional distribution of Y given X is regarded as unknown but possibly, say, normal and (1.1) and (1.2) are only required to hold for some choice of the unknown conditional distribution. Alternatively the consequence

$$\liminf_n \sup_{\theta \in \Theta} n^{2r} E_\theta(\hat{T}_n - T(\theta))^2 > 0 \tag{1.5}$$

of (1.1) holds without any assumption on the conditional distribution of Y given X if $\hat{T}_n$ is required to be linear in $Y_1, \cdots Y_n$. For then

$$E_\theta(\hat{T}_n - T(\theta))^2 = E_\theta E_\theta((\hat{T}_n - T(\theta))^2 \mid X_1, \cdots, X_n)$$

depends on the conditional distribution of Y given X only through the conditional mean and variance. Thus $E_\theta(\hat{T}_n - T(\theta))^2$ is unaltered if this conditional distribution is replaced by the normal distribution with the same mean and variance. Consequently (1.5) remains valid as desired. (It is not hard to give a direct proof of this result.)

The theorem is proven for Model 2 in Section 3. Previous results on upper bounds to local and global rates of convergence for nonparametric estimators of a density function have been obtained by Farrell (1972), Chentsov (1972), Wahba (1975), Samarov (1976), Meyer (1977), Khasminskii (1978), Boyd and Steele (1978) and Bretagnolle and Huber (1979).

The literature on asymptotic properties of various estimators of regression functions, density functions and their derivatives is too numerous to list here. Some of these results show that r is achievable in various contexts. The asymptotic results on estimating regression functions and their derivatives typically assume as much smoothness on the marginal density f of X as on the regression function. The above theorem shows that such smoothness assumptions on f are unnecessary.

2. Estimation of a regression function. Assume that Model 1 holds. It will first be shown that r, as defined in the theorem, is an upper bound to the rate of convergence. Now $L = \sum_{0 \le j \le m} L_j$, where $L_j = \sum_{|\alpha| = j} C_\alpha D^\alpha$. Let ψ be an infinitely differentiable function with compact support such that $L_m \psi(0) > 0$ and

$$|\psi(x) - \psi_k(x)| \le M \|x\|^p, \qquad x \in \mathbb{R}^d.$$

Choose $N > 0$ and $\delta \in (0, 1]$ and recall that $\gamma = 1/(2p + d)$. Define g_n on $\mathbb{R}^d$ by

$$g_n(x) = \delta N^p n^{-\gamma p} \psi(N^{-1} n^\gamma x).$$

Then $g_n \in \mathscr{G}$, so $\theta_n = \theta_0 + g_n \in \Theta$. Since the density of X is bounded above near the origin,

$$\limsup_n nEg_n^2(X) < \infty. \tag{2.1}$$

Let μ_n and ν_n denote the joint distribution of $(X_1, Y_1), \cdots, (X_n, Y_n)$ under P_{θ_0} and P_{θ_n} respectively, let L_n denote the Radon-Nikodym derivative $d\nu_n/d\mu_n$ and set $l_n = \log_e L_n$. It will now be shown that

$$\limsup_n E_{\theta_0} |l_n| < \infty \tag{2.2}$$

and

$$\lim_{\delta \to 0} \limsup_n E_{\theta_0} |l_n| = 0. \tag{2.3}$$

To this end choose n sufficiently large so that $|g_n| \le \epsilon_0$ and $g_n = 0$ on U^c. Observe that by Taylor's theorem with remainder

$$\begin{aligned} l_n &= \textstyle\sum_1^n [l(Y_i | X_i, \theta_0(X_i) + g_n(X_i)) - l(Y_i | X_i, \theta_0(X_i)); X_i \in U] \\ &= \textstyle\sum_1^n g_n(X_i) l'(Y_i | X_i, \theta_0(X_i)) + Z_n, \end{aligned}$$

where

$$Z_n = \tfrac{1}{2} \textstyle\sum_1^n g_n^2(X_i) l''(Y_i | X_i, \theta_0(X_i) + \epsilon_i)$$

for some ϵ_i, $1 \le i \le n$, satisfying $|\epsilon_i| \le \epsilon_0$. Thus

$$|Z_n| \le \tfrac{1}{2} \textstyle\sum_1^n g_n^2(X_i) M(Y_i | X_i, \theta_0(X_i))$$

and hence

$$E_{\theta_0} |Z_n| \le \frac{Cn}{2} Eg_n^2(X). \tag{2.4}$$

Also

$$\int l'(y | x, t) f(y | x, t) \varphi(dy) = \int f'(y | x, t) \varphi(dy) = 0, .$$

so

$$E_{\theta_0} g_n(X) l'(Y | X, \theta_0(X)) = 0$$

and hence

$$E_{\theta_0} \textstyle\sum_1^n g_n(X_i) l'(Y_i | X_i, \theta_0(X_i)) = 0.$$

Moreover

$$\begin{aligned}
&\int (l'(y \mid x, t))^2 f(y \mid x, t)\varphi(dy) \\
&= \int f''(y \mid x, t)\varphi(dy) - \int l''(y \mid x, t) f(y \mid x, t)\varphi(dy) \\
&\leq \int M(y \mid x, t) f(y \mid x, t)\varphi(dy) \leq C,
\end{aligned}$$

so

$$E_{\theta_0}\left[\left(\sum_1^n g_n(X_i) l'(Y_i \mid X_i, \theta_0(X_i))\right)^2\right] \leq CnEg_n^2(X).$$

Therefore

$$E_{\theta_0} \left| \sum_1^n g_n(X_i) l'(Y_i \mid X_i, \theta_0(X_i)) \right| \leq (CnEg_n^2(X))^{1/2}. \tag{2.5}$$

By (2.4) and (2.5)

$$E_{\theta_0} \mid l_n \mid \leq (CnEg_n^2(X))^{1/2} + \frac{Cn}{2} Eg_n^2(X).$$

This and (2.1) together yield (2.2) and (2.3).

By (2.2) there is a finite positive constant M such that

$$\limsup_n E_{\theta_0} \mid \log_e L_n \mid < M.$$

Choose $\epsilon > 0$ such that if $L_n > (1 - \epsilon)/\epsilon$ or $L_n < \epsilon/(1 - \epsilon)$, then $\mid \log_e L_n \mid \geq 2M$. By the Markov inequality

$$\liminf_n \mu_n \left(\frac{\epsilon}{1 - \epsilon} \leq L_n \leq \frac{1 - \epsilon}{\epsilon}\right) > \frac{1}{2}.$$

Let n be sufficiently large so that

$$\mu_n \left(\frac{\epsilon}{1 - \epsilon} \leq L_n \leq \frac{1 - \epsilon}{\epsilon}\right) > \frac{1}{2}.$$

Put prior probabilities ½ each on θ_0 and θ_n. Then

$$P(\theta = \theta_n \mid (X_1, Y_1), \cdots, (X_n, Y_n)) = \frac{L_n/2}{L_n/2 + ½} = \frac{L_n}{L_n + 1}$$

and hence

$$\begin{aligned}
&P(\epsilon \leq P(\theta = \theta_n \mid (X_1, Y_1), \cdots, (X_n, Y_n)) \leq 1 - \epsilon) \\
&= P\left(\epsilon \leq \frac{L_n}{L_n + 1} \leq 1 - \epsilon\right) = P\left(\frac{\epsilon}{1 - \epsilon} \leq L_n \leq \frac{1 - \epsilon}{\epsilon}\right) \\
&\geq \frac{1}{2} \mu_n \left(\frac{\epsilon}{1 - \epsilon} \leq L_n \leq \frac{1 - \epsilon}{\epsilon}\right) \geq \frac{1}{4}.
\end{aligned}$$

Therefore any method of deciding between θ_0 and θ_n based on $(X_1, Y_1), \cdots, (X_n, Y_n)$ must have overall error probability at least $\epsilon/4$. Apply this result to the classifier $\bar{\theta}_n$ defined by

$$\begin{aligned}
\bar{\theta}_n &= \theta_0 \quad \text{if } \hat{T}_n \leq \frac{T(\theta_0) + T(\theta_n)}{2}, \\
&= \theta_n \quad \text{if } \hat{T}_n > \frac{T(\theta_0) + T(\theta_n)}{2}.
\end{aligned}$$

It follows that

$$\frac{1}{2}P_{\theta_0}\left(|\hat{T}_n - T(\theta_0)| \ge \frac{T(\theta_n) - T(\theta_0)}{2}\right) + \frac{1}{2}P_{\theta_n}\left(|\hat{T}_n - T(\theta_n)| > \frac{T(\theta_n) - T(\theta_0)}{2}\right) \ge \frac{\epsilon}{4}.$$

Consequently

$$\sup_{\theta\in\Theta} P_\theta\left(|\hat{T}_n - T(\theta)| \ge \frac{T(\theta_n) - T(\theta_0)}{2}\right) \ge \frac{\epsilon}{4}.$$

In particular

$$\liminf_n \sup_{\theta\in\Theta} P_\theta\left(|\hat{T}_n - T(\theta)| \ge \frac{T(\theta_n) - T(\theta_0)}{2}\right) > 0.$$

Now

$$\begin{aligned}\frac{T(\theta_n) - T(\theta_0)}{2} &= \tfrac{1}{2}\, T(g_n) = \tfrac{1}{2} \textstyle\sum_{j=0}^m L_j g_n(0) \\ &= \frac{\delta N^p n^{-\gamma p}}{2} \textstyle\sum_{j=0}^m (N^{-1}n^\gamma)^j L_j \Psi(0) \\ &\ge \frac{\delta N^{p-m} L_m \Psi(0)}{4} n^{-r}\end{aligned}$$

for n sufficiently large. Since N can be arbitrarily large, (1.1) holds.

The proof of (1.2) is very similar. Choose a positive integer $i_0 \ge 2$ and put prior probability i_0^{-1} on each of the i_0 points

$$\theta_{ni} = \theta_0 + \frac{i-1}{i_0 - 1}(\theta_n - \theta_0), \qquad 1 \le i \le i_0.$$

Equation (2.3) can be used to show that there is a $\delta > 0$ such that for n sufficiently large any method of classifying $\theta \in \{\theta_{n1}, \cdots, \theta_{ni_0}\}$ based on $(X_1, Y_1), \cdots, (X_n, Y_n)$ must have overall error probability at least $1-2/i_0$, which can be made arbitrarily close to 1 by choosing i_0 sufficiently large. Equation (1.2) follows easily from this observation. This completes the proof that r is an upper bound to the rate of convergence.

It remains to construct a sequence $\{\hat{T}_n\}$ of estimators of $T(\theta)$ such that (1.3) holds. Without loss of generality it can be assumed that $\theta_0 = 0$. Choose $\delta_0 > 0$ such that U contains the ball $B_{\delta_0} = \{x \in \mathbb{R}^d : \|x\| \le \delta_0\}$, $\delta_0 \le f \le \delta_0^{-1}$ on B_{δ_0} and $\sigma^2 \le \delta_0^{-1}$ on B_{δ_0}. Given $0 < \delta \le \delta_0$, let f_δ be the probability density function on $\mathbb{R}^d$ defined by

$$\begin{aligned} f_\delta(x) &= f(\delta x) \Big/ \int_{\|y\|\le 1} f(\delta y)\, dy, && \|x\| \le 1, \\ &= 0, && \|x\| > 1. \end{aligned}$$

Then $c_d \delta_0^2 \le f_\delta \le c_d \delta_0^{-2}$ on $\mathbb{R}^d$, where c_d^{-1} denotes the volume of a unit ball in $\mathbb{R}^d$.

Let A denote the collection of d-tuples α of nonnegative integers such that $|\alpha| \le k$ and let $|A|$ denote the cardinality of A. For $0 < \delta \le \delta_0$ let $\mathscr{A}_\delta = (\mathscr{A}_{\delta\alpha\beta})$ denote the positive definite symmetric $|A| \times |A|$ matrix defined by

$$\mathscr{A}_{\delta\alpha\beta} = \int_{\|x\|\le 1} x^\alpha x^\beta f_\delta(x)\, dx, \qquad \alpha, \beta \in A.$$

Then

$$\inf_{0<\delta\le\delta_0} \det \mathscr{A}_\delta > 0. \tag{2.6}$$

For suppose otherwise and let λ_δ denote the smallest eigenvalue of $\mathscr{A}_\delta$. Then $\inf_{0<\delta\le\delta_0} \lambda_\delta = 0$. Let $p_\delta = (p_{\delta\alpha})$ be an eigenvector of $\mathscr{A}_\delta$ corresponding to the eigenvalue λ_δ and such that $\sum_\alpha p_{\delta\alpha}^2 = 1$. Let P_δ be the polynomial on $\mathbb{R}^d$ defined by $P_\delta(x) = \sum_\alpha p_{\delta\alpha} x^\alpha$. Then

$$\lambda_\delta = \sum_{\alpha,\beta} \mathscr{A}_{\delta\alpha\beta} p_{\delta\alpha} p_{\delta\beta} = \int_{\|x\| \le 1} P_\delta^2(x) f_\delta(x) dx \ge c_d \delta_0^2 \int_{\|x\| \le 1} P_\delta^2(x)\, dx,$$

so

$$\inf_{0<\delta\le\delta_0} \int_{\|x\|\le 1} P_\delta^2(x) dx = 0.$$

Consequently, by a compactness argument, there is a nonzero polynomial P on $\mathbb{R}^d$ such that

$$\int_{\|x\|\le 1} P^2(x)\, dx = 0.$$

By continuity $P = 0$ on $\{x \in \mathbb{R}^d : \|x\| \le 1\}$, which is impossible. Therefore (2.6) holds as desired.

Set $\gamma = 1/(2p + d)$. Let $\{\epsilon_n\}$ be a sequence of positive numbers satisfying either of the following two conditions: (i) ϵ_n is nonrandom and $0 < \lim_n n^\lambda \epsilon_n < \infty$; (ii) ϵ_n is the N_nth smallest value among $\|X_1\|, \cdots, \|X_n\|$, where $\{N_n\}$ is a sequence of nonrandom positive integers such that $N_n \le n$ and $0 < \lim_n n^{-2\gamma p} N_n < \infty$. Set

$$I_n = \{i : 1 \le i \le n \quad \text{and} \quad \|X_i\| \le \epsilon_n\}$$

and

$$\delta_n = \max[\|X_i\| : i \in I_n].$$

Note that under (ii), $\delta_n = \epsilon_n$ and I_n has cardinality N_n. Let N_n also denote the cardinality of I_n under (i). Since f is bounded away from zero and infinity near the origin, $n\epsilon_n^d/N_n$ is bounded in probability away from zero and infinity and hence so are $n^\gamma \epsilon_n$ and $n^{-2\gamma p} N_n$; also δ_n/ϵ_n converges to one in probability. Consequently (note that $r = (p - m)\gamma$)

$$\delta_n^{p-m} = n^{-r} O_p(1) \tag{2.7}$$

and

$$N_n^{-1} \delta_n^{-2m} = n^{-2r} O_p(1). \tag{2.8}$$

Clearly $\lim_n P(0 < \delta_n \le \delta_0) = 1$. In the definitions below it is assumed that $0 < \delta_n \le \delta_0$. Arbitrary definitions can be employed on the complementary event of vanishingly small probability.

Let $\mathscr{X}_n = (\mathscr{X}_{ni\alpha})$ denote the $N_n \times |A|$ matrix defined by

$$\mathscr{X}_{ni\alpha} = \frac{X_i^\alpha}{\delta_n^{|\alpha|}}, \qquad i \in I_n \quad \text{and} \quad \alpha \in A,$$

and let $\mathscr{X}_x'$ denote the transpose of $\mathscr{X}_n$. Then $\mathscr{X}_n' \mathscr{X}_n$ is the $|A| \times |A|$ matrix determined by

$$(\mathscr{X}_n' \mathscr{X}_n)_{\alpha\beta} = \sum_{i \in I_n} \frac{X_i^\alpha X_i^\beta}{\delta_n^{|\alpha|} \delta_n^{|\beta|}}, \qquad \alpha, \beta \in A.$$

It is clear that the elements of $N_n^{-1} \mathscr{X}_n' \mathscr{X}_n - \mathscr{A}_{\delta_n}$ converge to zero in probability. Thus by (2.6), $N_n(\mathscr{X}_n' \mathscr{X}_n)^{-1}$ is bounded in probability; that is

$$N_n(\mathscr{X}_n' \mathscr{X}_n)^{-1} = O_p(1). \tag{2.9}$$

Define the N_n-dimensional vectors $\mathscr{T}_n = (\mathscr{T}_{ni})$, $\mathscr{T}_{kn} = (\mathscr{T}_{kni})$ and $\mathscr{Y}_n = (\mathscr{Y}_{ni})$ by

$$\mathscr{T}_{ni} = \theta(X_i), \qquad i \in I_n,$$

$$\mathscr{T}_{kni} = \theta_k(X_i), \qquad i \in I_n,$$

θ_k denoting kth degree Taylor polynomial approximation to θ, and

$$\mathscr{Y}_{ni} = Y_i, \qquad i \in I_n.$$

Set $\mathscr{Z}_n = \mathscr{Y}_n - \mathscr{T}_n$. Define the $|A|$-dimensional vector $\mathscr{L}_n = (\mathscr{L}_{n\alpha})$ by

$$\mathscr{L}_{n\alpha} = \frac{\alpha! C_\alpha}{\delta_n^{|\alpha|}} \qquad \alpha \in A,$$

where $C_\alpha = 0$ for $m < |\alpha| \le k$. Now

$$\mathscr{T}_{kni} = \sum\nolimits_{|\alpha| \le k} \left(\frac{X_i^\alpha}{\delta_n^{|\alpha|}} \right) \frac{\delta_n^{|\alpha|} D^\alpha \theta(0)}{\alpha!} \qquad \text{for } i \in I_n,$$

so

$$((\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{T}_{kn})_\alpha = \frac{\delta_n^{|\alpha|}}{\alpha!} D^\alpha \theta(0)$$

and hence

$$T(\theta) = \mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{T}_{kn}.$$

Let $\hat{\theta}_{kn}$ denote the polynomial on $\mathbb{R}^d$ of degree k which minimizes

$$\sum\nolimits_{i \in I_n} (Y_i - \hat{\theta}_{kn}(X_i))^2$$

and set $\hat{T}_n = L\hat{\theta}_{kn}(0)$. It follows from the normal equations for least squares estimators that

$$\hat{T}_n = \mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{Y}_n.$$

(Write $\hat{\theta}_{kn}(x) = \sum_{|\alpha| \le k} b_\alpha \frac{x^\alpha}{\delta_n^{|\alpha|}}$ and note that $b_\alpha = ((\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{Y}_n)_\alpha$.) Consequently

$$\hat{T}_n - T(\theta) = \mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{Z}_n + \mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' (\mathscr{T}_n - \mathscr{T}_{kn}). \tag{2.10}$$

Now

$$|\mathscr{T}_{ni} - \mathscr{T}_{nki}| \le M \| X_i \|^p \le M \delta_n^p, \qquad i \in I_n.$$

Thus

$$|(\mathscr{X}_n' (\mathscr{T}_n - \mathscr{T}_{nk}))_\alpha| \le M \sum\nolimits_{i \in I_n} \frac{\| X_i \|^{|\alpha|}}{\delta_n^{|\alpha|}} \delta_n^p \le M N_n \delta_n^p, \qquad \alpha \in A,$$

and hence by (2.7) and (2.9)

$$\mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' (\mathscr{T}_n - \mathscr{T}_{kn}) = n^{-r} O_p(1). \tag{2.11}$$

Observe also that

$$E(\mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{Z}_n \mid \mathscr{X}_n) = 0$$

and by (2.8) and (2.9)

$$\operatorname{Var}(\mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{Z}_n \mid \mathscr{X}_n) \le \delta_0^{-1} \mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{L}_n = n^{-2r} O_p(1).$$

It follows that

$$\mathscr{L}_n' (\mathscr{X}_n' \mathscr{X}_n)^{-1} \mathscr{X}_n' \mathscr{Z}_n = n^{-r} O_p(1). \tag{2.12}$$

By (2.10) – (2.12)

$$\hat{T}_n - T(\theta) = n^{-r} O_p(1). \tag{2.13}$$

It is easily seen by examining the proof of (2.13) that it actually holds uniformly in θ, so that (1.3) holds. This completes the proof that r is an achievable rate of convergence. Therefore the theorem is valid for Model 1.

3. Estimation of a density function.

Assume that Model 2 holds. It will first be shown that r is an upper bound to the rate of convergence. Let x_0 be a point in $\mathbb{R}^d$ other than the origin.

Let ψ be a nonnegative infinitely differentiable function with compact support such that $L_m\psi(0) > 0$, ψ vanishes on a neighborhood of x_0,

$$|\psi(x) - \psi_k(x)| \leq M\|x\|^p, \qquad x \in \mathbb{R}^d,$$

and

$$|\psi(x + x_0)| \leq M\|x\|^p, \qquad x \in \mathbb{R}^d.$$

Let γ still be $1/(2p + d)$, let δ and N be positive constants, and define g_n on $\mathbb{R}^d$ for n sufficiently large by

$$g_n(x) = \delta N^p n^{-\gamma p}(\psi(N^{-1}n^\gamma x) - b_n\psi(N^{-1}n^\gamma x + x_0)),$$

where b_n is chosen so that $\int g_n\theta_0\, dx = 0$. Since θ_0 is bounded away from zero and infinity on some neighborhood of the origin there is a positive constant B independent of δ such that $0 \leq b_n \leq B$ for n sufficiently large. Now

$$|g_n(x) - (g_n)_k(x)| \leq \delta(1 + b_n)M\|x\|^p \leq \delta(1 + B)M\|x\|^p \leq M\|x\|^p, \qquad x \in U,$$

for n sufficiently large, where δ is now chosen so that $\delta(1 + B) \leq 1$. Thus $g_n \in \mathscr{G}$ for n sufficiently large. Since $\lim_n \max_x |g_n(x)| = 0$, $\theta_n = \theta_0(1 + g_n) \in \Theta$ for n sufficiently large. Since θ_0 is bounded on a neighborhood of the origin,

$$\text{(3.1)} \qquad \limsup_n nE_{\theta_0}g_n^2(X) < \infty.$$

Let μ_n and ν_n denote the joint distribution of $X_1, \cdots, X_n$ under P_{θ_0} and P_{θ_n} respectively, let L_n denote the Radon-Nikodym derivative $d\nu_n/d\mu_n$ and set $l_n = \log_e L_n$. Then

$$\text{(3.2)} \qquad \limsup_n E_{\theta_0}|l_n| < \infty$$

and

$$\text{(3.3)} \qquad \lim_{\delta\to 0}\limsup_n E_{\theta_0}|l_n| = 0.$$

To see this note that (for n sufficiently large)

$$|g_n(x)| \leq ½, \qquad x \in \mathbb{R}^d;$$

also

$$L_n = \prod_{i=1}^n (1 + g_n(x_i))$$

and hence

$$l_n = \sum_1^n \log(1 + g_n(X_i)).$$

Thus

$$|l_n - \sum_1^n g_n(X_i)| \leq \sum_1^n g_n^2(X_i)$$

and therefore

$$|l_n| \leq \left|\sum_1^n g_n(X_i)\right| + \sum_1^n g_n^2(X_i).$$

Since $E_{\theta_0} g_n(X) = \int g_n\theta_0\, dx = 0$,

$$E_{\theta_0}(\sum_1^n g_n(X_i))^2 = nE_{\theta_0}g_n^2(X).$$

Schwarz's inequality now implies that

$$E_{\theta_0}|l_n| \leq (nE_{\theta_0}g_n^2(X))^{1/2} + nE_{\theta_0}g_n^2(X).$$

The last inequality and (3.1) together yield (3.2) and (3.3).

The argument required to conclude from (3.2) and (3.3) that r is an upper bound to the rate of convergence is essentially the same as the argument required to prove the same conclusion in Section 2 based on (2.2) and (2.3). (Note that

$$\begin{aligned} T(\theta_n) - T(\theta_0) &= L(\theta_0 g_n)(0) = \textstyle\sum_{j=0}^{m} L_j(\theta_0 g_n)(0) \\ &= \theta_0(0) L_m g_n(0) + o(n^{-r}) \\ &= \delta N^{p-m} \theta_0(0) L_m \psi(0) n^{-r} + o(n^{-r}).) \end{aligned}$$

It remains to construct a sequence $\{\hat{T}_n\}$ of estimators of $T(\theta)$ such that (1.3) holds. It suffices to verify that

$$\limsup_n n^{2r} \sup_{\theta\in\Theta} E_\theta(\hat{T}_n - T(\theta))^2 < \infty. \tag{3.4}$$

To this end write $\theta = \theta_0 h$, where $h = 1 + g$. Then

$$T(\theta) = L\theta(0) = L(\theta_0 h)(0).$$

Since θ_0 is fixed, $L(\theta_0 h)(0)$ is clearly a linear combination of $D^\alpha h(0)$, $|\alpha| \le m$. Consequently

$$L(\theta_0 h)(0) = \bar{L}h(0) = \bar{L}h_k(0),$$

where $\bar{L}$ is a fixed linear differential operator on $\mathscr{C}_k$ of order m. Thus

$$T(\theta) = \bar{L}h_k(0).$$

Let δ be a positive number such that $\theta_0(x) \ge \delta$ for $\|x\| \le \delta$. Let K be an infinitely differentiable function with compact support such that $\int K(y)\,dy = 1$ and

$$\int y^\alpha K(y)\,dy = 0 \qquad \text{for} \quad 1 \le |\alpha| \le k.$$

Let ϵ_n be a sequence of positive numbers such that

$$0 < \lim_n n^\gamma \epsilon_n < \infty.$$

Let K_n be the function on $\mathbb{R}^d$ defined by

$$K_n(x) = \epsilon_n^{-d} K(\epsilon_n^{-1} x).$$

It can be assumed that

$$K_n(x) = 0 \qquad \text{for } n \ge 1 \text{ and } \|x\| \ge \delta.$$

Set $I_n = \{i: 1 \le i \le n \text{ and } \|X_i\| \le \delta\}$.

Let $\hat{h}_n$ denote the estimator of h defined by

$$\hat{h}_n(x) = \frac{1}{n} \sum_{i\in I_n} \frac{K_n(x - X_i)}{\theta_0(X_i)}.$$

Set

$$\hat{T}_n = \bar{L}\hat{h}_n(0) = \frac{1}{n} \sum_{i\in I_n} \frac{\bar{L}K_n(-X_i)}{\theta_0(X_i)}.$$

Then

$$E_\theta \hat{T}_n = \int_{\|y\|\le\delta} \frac{\bar{L}K_n(-y)}{\theta_0(y)} \theta_0(y) h(y)\,dy = \int \bar{L}K_n(-y) h(y)\,dy.$$

Next it will be shown that

$$E_\theta \hat{T}_n = T(\theta) + \int \bar{L}K_n(-y)(h(y) - h_k(y))\,dy. \tag{3.5}$$

To see this note that

$$\int K_n(x-y)h_k(y)\,dy = \int K_n(y)h_k(x-y)\,dy = h_k(x)$$

and hence that

$$\int K_n(x-y)h(y)\,dy = h_k(x) + \int K_n(x-y)(h(y)-h_k(y))\,dy.$$

Consequently

$$\int \bar{L}K_n(x-y)h(y)\,dy = \bar{L}h_k(x) + \int \bar{L}K_n(x-y)(h(y)-h_k(y))\,dy,$$

from which (3.5) follows.

By (3.5)

$$|E_\theta \hat{T}_n - T(\theta)| \le M \int |\bar{L}K_n(-y)|\,\|y\|^p\,dy,$$

so that

$$\limsup_n \epsilon_n^{-(p-m)} \sup_{\theta\in\Theta} |E_\theta \hat{T}_n - T(\theta)| < \infty.$$

Therefore

$$\limsup_n n^r \sup_{\theta\in\Theta} |E_\theta \hat{T}_n - T(\theta)| < \infty. \tag{3.6}$$

Observe next that

$$\begin{aligned}\mathrm{Var}_\theta \hat{T}_n &\le \frac{1}{n}\int_{\|y\|\le\delta} \left(\frac{\bar{L}K_n(-y)}{\theta_0(y)}\right)^2 \theta_0(y)h(y)\,dy\\ &= \frac{1}{n}\int_{\|y\|\le\delta} \frac{(\bar{L}K_n(-y))^2}{\theta_0(y)} h(y)\,dy\\ &\le \frac{2}{n\delta}\int (\bar{L}K_n(-y))^2\,dy.\end{aligned}$$

Consequently

$$\limsup_n n\,\epsilon_n^{2m+d} \sup_{\theta\in\Theta} \mathrm{Var}_\theta \hat{T}_n < \infty$$

and hence

$$\limsup_n n^{2r} \sup_{\theta\in\Theta} \mathrm{Var}_\theta \hat{T}_n < \infty. \tag{3.7}$$

Now (3.4) follows from (3.6) and (3.7), so r is an achievable rate of convergence. Therefore the theorem is valid for Model 2.

REFERENCES

[1] BOYD, D. W. and STEELE, J. M. (1978). Lower bounds for nonparametric density estimation rates. *Ann. Statist.* **6** 932–934.

[2] BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densities: risque minimax. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **47** 119–137.

[3] CHENTSOV, N. N. (1972). *Statistical Decision Functions and Optimal Inference.* In Russian. Nauka, Moscow.

[4] CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

[5] FARRELL, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43** 170–180.

[6] KHASMINSKII, R. Z. (1978). A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theor. Probability Appl.* **23** 794–798.

[7] MEYER, T. G. (1977). Bounds for estimation of density functions and their derivatives. *Ann. Statist.* **5** 136–142.

[8] SAMAROV, A. M. (1976). Minimax bound on the risk of nonparametric density estimates. *Problems of Information Transmission* **12** 242–244.

[9] STONE, C. J. (1975). Nearest neighbor estimators of a nonlinear regression function. *Proc. Computer Sci. Statist. 8th Ann. Symp. Interface.* 413–418. Health Sciences Computer Facility, UCLA.

[10] STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

[11] WAHBA, C. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–29.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, LOS ANGELES
LOS ANGELES, CALIFORNIA 90024