Optimal Global Rates of Convergence for Nonparametric Regression
Author(s): Charles J. Stone
Source: *The Annals of Statistics*, Vol. 10, No. 4 (Dec., 1982), pp. 1040-1053
Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/2240707
Accessed: 10/11/2008 19:14

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=ims.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

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

http://www.jstor.org

The Annals of Statistics
1982, Vol. 10, No. 4, 1040–1053

Special Invited Paper

OPTIMAL GLOBAL RATES OF CONVERGENCE FOR NONPARAMETRIC REGRESSION[1]

BY CHARLES J. STONE

University of California, Berkeley

Consider a p-times differentiable unknown regression function θ of a d-dimensional measurement variable. Let $T(\theta)$ denote a derivative of θ of order m and set $r = (p - m)/(2p + d)$. Let $\hat{T}_n$ denote an estimator of $T(\theta)$ based on a training sample of size n, and let $\| \hat{T}_n - T(\theta) \|_q$ be the usual L^q norm of the restriction of $\hat{T}_n - T(\theta)$ to a fixed compact set. Under appropriate regularity conditions, it is shown that the optimal rate of convergence for $\| \hat{T}_n - T(\theta) \|_q$ is n^{-r} if $0 < q < \infty$; while $(n^{-1} \log n)^r$ is the optimal rate if $q = \infty$.

1. Discussion. Let (X, Y) be a pair of random variables which are respectively d and 1 dimensional, and let θ denote the regression function of the response Y on the measurement variable X, so that $E(Y \mid X) = \theta(X)$. Let $\hat{\theta}_n$, $n \geq 1$, denote estimators of θ, $\hat{\theta}_n$ being based on a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ of size n from the distribution of (X, Y). The estimators $\hat{\theta}_n$, $n \geq 1$, are said to be *parametric* if $\hat{\theta}_n \in \Theta$ for all $n \geq 1$, where Θ is a collection of functions which are defined in terms of a finite number of unknown parameters. Otherwise the estimators $\hat{\theta}_n$, $n \geq 1$, are said to be *nonparametric*.

In Stone (1977) a consistency theorem was obtained for a large class of nonparametric regression estimators and used to establish the consistency of nearest neighbor estimators. Since then, consistency has been established for kernel estimators by Devroye and Wagner (1980) and Spiegelman and Sacks (1980) and for partition estimators by Gordon and Olshen (1980) and Breiman, et al. (1983). Stone (1980) and the present paper are devoted to optimal rates of convergence for nonparametric regression.

Nearest neighbor, kernel, and partition methods of nonparametric regression, as usually defined, are based on local averages. In Stone (1975, 1977) the suggestion was made that nonparametric regression based on locally linear fits should also be considered. This suggestion, and its extension to local polynomial fits, can be given theoretical justification in terms of optimal rates of convergence.

To see this in the simplest possible setting, suppose that $d = 1$ and let U be an open interval containing $[0, 1]$. Suppose that the distribution of X is absolutely continuous and that its density f is bounded away from zero on U; that the conditional variance of Y given X is both bounded and bounded away from zero on U; and that the regression function θ is p-times continuously differentiable on U, where p is a positive integer. The estimators $\hat{\theta}_n$, $n \geq 1$, are said (temporarily) to be *asymptotically optimal* if

$$n^{-2p/(2p+1)} \int_0^1 \{\hat{\theta}_n(x) - \theta(x)\}^2 \, dx$$

is bounded in probability as $n \to \infty$. (This definition is reasonable in light of Theorem 1 below.)

Received July 1981; revised May 1982.

[1] This paper is based on a Special Invited Lecture presented at the Eastern Regional Meeting of IMS held in Philadelphia, May 1981. The research was supported in part by National Science Foundation Grant MCS 80-02732.

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

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

Given $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, set $|x| = (x_1^2 + \cdots + x_d^2)^{1/2}$. Also let $\#(A)$ denote the number of elements in a set A. Let δ_n, $n \geq 1$, be positive constants which tend to zero as $n \to \infty$. Set

$$\mathcal{N}_n(x) = \{i : 1 \leq i \leq n \text{ and } |X_i - x| \leq \delta_n\}$$

and $N_n(x) = \#(\mathcal{N}_n(x))$. Let $\hat{\theta}_n$ denote the kernel estimator of θ defined by the local average

$$\hat{\theta}_n = \frac{1}{N_n(x)} \sum_{\mathcal{N}_n(x)} Y_i.$$

If $p = 1$ and $\delta_n = n^{-1/3}$, then $\hat{\theta}_n$ is asymptotically optimal; see Theorem 2 of Spiegelman and Sacks (1980). Suppose instead that $p = 2$. If f is absolutely continuous and f' is square integrable on U, then δ_n can be chosen (e.g., $\delta_n = n^{-1/5}$) to make $\hat{\theta}_n$ asymptotically optimal; but the conclusion is not generally true without the corresponding smoothness assumption on f. For any positive integer p, the following approach yields an asymptotically optimal estimator without any smoothness assumption of f. Given $x \in [0, 1]$, choose $\hat{P}_n(\cdot; x)$ to be the polynomial of degree $p - 1$ (or less) which minimizes

$$\sum_{\mathcal{N}_n(x)} \{Y_i - \hat{P}_n(X_i; x)\}^2$$

and set $\hat{\theta}_n(x) = \hat{P}_n(x; x)$. Then $\hat{\theta}_n$ is asymptotically optimal provided that f is bounded (as well as bounded away from zero) on U. The requirement that f be bounded can be dropped if $\hat{P}_n(\cdot; x)$ is chosen to minimize

$$\sum_{\mathcal{N}_n(x)} \frac{\{Y_i - \hat{P}_n(X_i; x)\}^2}{W_n(X_i; x)},$$

where $W_n(\cdot; x)$ is an appropriate positive weight function. This approach, involving local weighted polynomial regression, is easily generalized to handle multidimensional measurement variables ($d > 1$). The details will be given in Section 3, along with extensions to estimators $\hat{\theta}_n^{(m)}$ of the mth derivative $\theta^{(m)}$ of θ and modifications involving linear interpolation to achieve the optimal rate of convergence for the maximum of $|\hat{\theta}_n^{(m)}(x) - \theta^{(m)}(x)|$ as x ranges over a fixed compact set.

Halász (1978) used another approach, also based on $(p - 1)$th degree polynomials, to obtain an asymptotically optimal estimator of the regression function. It is indispensable to his approach that $d = 1$.

The theory to be developed below can handle deterministic as well as random measurement variables. It is convenient to describe the precise results in terms of the following model. The distribution of a response Y depends on the value $x \in \mathbb{R}^d$ of a measurement variable; it has the specific form $h(y \mid x, t)\phi(dy)$, where ϕ is a measure on $\mathbb{R}$, t is an unknown real-valued parameter which belongs to an open interval J, and t is the mean of the distribution; that is,

$$\int y h(y \mid x, t)\phi(dy) = t \quad \text{for} \quad x \in \mathbb{R}^d \quad \text{and} \quad t \in J.$$

The parameter t is allowed to vary with x according to $t = \theta(x)$, where the unknown regression function θ is assumed to belong to a collection Θ of suitably smooth functions on $\mathbb{R}^d$. It is assumed that $\theta(x) \in J$ for $\theta \in \Theta$ and $x \in \mathbb{R}^d$.

For each given $n \geq 1$, consider a training sample $X_1, Y_1, \cdots, X_n, Y_n$ (i.e., $X_1^{(n)}, Y_1^{(n)}, \cdots, X_n^{(n)}, Y_n^{(n)}$), where $X_1, \cdots, X_n$ are $\mathbb{R}^d$-valued and may be random or nonrandom. Conditioned on $X_1 = x_1, \cdots, X_n = x_n$, the random variables $Y_1, \cdots, Y_n$ are assumed to be independent and distributed according to the model described in the previous paragraph, so that Y_i has distribution $h(y \mid x_i, \theta(x_i))\phi(dy)$. Let P_θ denote the dependence of various probabilities on θ.

Let $T(\theta) = T(\cdot; \theta)$ denote an arbitrary finite linear combination (with constant coefficients) of the derivatives of θ; two examples are $T(x; \theta) = \theta(x)$ and $T(x; \theta) = \partial^2\theta/\partial x_1^2 + \cdots + \partial^2\theta/\partial x_d^2$. Let $\hat{T}_n$ denote an arbitrary (measurable) estimator of $T(\theta)$ based on the

training sample $X_1, Y_1, \cdots, X_n, Y_n$.

Let C be a compact subset of $\mathbb{R}^d$ having a nonempty interior and let $q \in (0, \infty]$. Define the L^q norm $\| \ \|_q$ by $\| g \|_q = \sup_{x \in C} | g(x) |$ if $q = \infty$ and $\| g \|_q = (\int_C | g(x) |^q \, dx)^{1/q}$ if $0 < q < \infty$. Let $\{b_n\}$ be a sequence of (eventually) positive constants. It is called a *lower rate of convergence* if there is a $c > 0$ such that

$$\lim_n \inf_{\hat{T}_n} \sup_\Theta P_\theta(\| \hat{T}_n - T(\theta) \|_q \geq c b_n) = 1;$$

here $\inf_{\hat{T}_n}$ denotes the infimum over all possible estimators $\hat{T}_n$. The sequence is said to be an *achievable rate of convergence* if there is a sequence $\{\hat{T}_n\}$ of estimators and a $c > 0$ such that

$$\lim_n \sup_\Theta P_\theta(\| \hat{T}_n - T(\theta) \|_q \geq c b_n) = 0. \tag{1.1}$$

It is called an *optimal rate of convergence* if it is both a lower and an achievable rate of convergence. If $\{b_n\}$ is a lower rate of convergence and $\{b'_n\}$ is an achievable rate of convergence, there are positive constants c and n_0 such that $b'_n \geq c b_n$ for $n \geq n_0$. If $\{b_n\}$ and $\{b'_n\}$ are both optimal rates of convergence, there are positive constants c and n_0 such that $c b_n \leq b'_n \leq c^{-1} b_n$ for $n \geq n_0$; so it is reasonable to refer to any optimal rate of convergence as *the* optimal rate of convergence. If $\{b_n\}$ is the optimal rate of convergence and $\{\hat{T}_n\}$ satisfies (1.1), the estimators $\hat{T}_n$, $n \geq 1$, are said to be *asymptotically optimal.*

The assumption that C have a nonempty interior is required to show that an appropriate sequence $\{b_n\}$ is a lower rate of convergence, and the assumption that C be compact is required to show that $\{b_n\}$ is achievable. In order to obtain precise results, conditions must also be imposed on the function h appearing in the basic model, on the collection Θ of possible regression functions, and on the asymptotic distribution of $X_1, \cdots, X_n$. Let U denote an open subset of $\mathbb{R}^d$ containing C. The first condition is needed to verify that $\{b_n\}$ is a lower convergence sequence.

CONDITION 1. *Let x and t respectively range over U and J. As a function of t, h is strictly positive and continuously differentiable; the equation*

$$\int h(y \mid x, t) \phi(dy) = 1$$

can be differentiated with respect to t to yield

$$\int h'(y \mid x, t) \phi(dy) = 0$$

and

$$\int h''(y \mid x, t) \phi(dy) = 0.$$

Set $\ell(y \mid x, t) = \log h(y \mid x, t)$. There are positive constants ε_0 and K_1 and there is a function $M(y \mid x, t)$ such that on the indicated domain

$$| \ell''(y \mid x, t + \varepsilon) | \leq M(y \mid x, t) \quad \text{for} \quad | \varepsilon | \leq \varepsilon_0$$

and

$$\int M(y \mid x, t) h(y \mid x, t) \phi(dy) \leq K_1.$$

The next condition is required to verify that certain rates of convergence are achievable and certain estimators are asymptotically optimal.

CONDITION 2. *For some* $s > 0$

$$\int e^{s|y-t|} h(y \mid x, t)\phi(dy)$$

is bounded for $x \in U$ *and* $t \in J$.

Conditions 1 and 2 are satisfied in the following two examples (for several other examples, see page 1350 of Stone, 1980).

EXAMPLE 1 *Normal.* Let ϕ denote Lebesgue measure on $\mathbb{R}$, let $J = \mathbb{R}$ and let

$$h(y \mid x, t) = \frac{1}{\sigma(x)(2\pi)^{1/2}} \exp\{-(y-t)^2/2\sigma^2(x)\},$$

where $\sigma(\cdot)$ is bounded as well as bounded away from zero on U.

EXAMPLE 2 *Bernoulli.* Let ϕ be a counting measure on $\{0, 1\}$, let J be a relatively compact open subinterval of $(0, 1)$ and let

$$h(y \mid t) = h(y \mid x, t) = t^y(1-t)^{1-y}.$$

A condition on the asymptotic distribution of $X_1, \cdots, X_n$ is required to guarantee achievability and asymptotic optimality.

CONDITION 3. *For every* $\lambda \in (0, 1/d)$ *and* $c > 0$ *there is a* $c' > 0$ *such that*

$$\lim_n P(\#\{i : 1 \le i \le n \text{ and } |X_i - x| \le cn^{-\lambda}\} \ge c'n^{1-\lambda d} \text{ for all } x \in U) = 1.$$

If U is, say, a polyhedron, this condition is implied by the following one (e.g., as a consequence of Lemma 1 in Section 2).

CONDITION 3′. *The random variables* $X_1, \cdots, X_n$ *are the first* n *terms of an* i.i.d. *sequence of random variables each having distribution* F, *the density of whose absolutely continuous component is bounded away from zero on* U.

Let $\alpha = (\alpha_1, \cdots, \alpha_d)$ denote a d-tuple of nonnegative integers and set $[\alpha] = \alpha_1 + \cdots + \alpha_d$. Let D^α denote the differential operator defined by

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

Let k be a nonnegative integer, let $0 < \beta \le 1$ and let $0 < K_2 < \infty$. Let Θ denote the collection of k-times continuously differentiable functions θ on $\mathbb{R}^d$ such that $\theta(x) \in J$ for $x \in \mathbb{R}^d$ and

$$|D^\alpha\theta(x) - D^\alpha\theta(x_0)| \le K_2|x - x_0|^\beta \quad \text{for} \quad x_0, x \in U \quad \text{and} \quad [\alpha] = k. \tag{1.2}$$

Let $T(\theta)$ be a linear combination with constant coefficients of $D^\alpha\theta$, $[\alpha] \le k$; so that $T(\theta) = Q\theta$, where

$$Q = \textstyle\sum_{[\alpha]\le k} q_\alpha D^\alpha,$$

q_α being real constants for $[\alpha] \le k$. Let m denote the *order* of Q, defined by

$$m = \max([\alpha] : [\alpha] \le k \text{ and } q_\alpha \ne 0).$$

(If $T(\theta) = \theta$, then $m = 0$; while if $T(\theta) = \partial^2\theta/\partial x_1^2 + \cdots + \partial^2\theta/\partial x_d^2$, then $m = 2$.)

Think of $p = k + \beta$ as a measure of the smoothness of the functions in Θ and set $r = (p - m)/(2p + d)$.

THEOREM 1. *Suppose that Conditions* 1–3 *hold. If* $0 < q < \infty$, *then* $\{n^{-r}\}$ *is the optimal rate of convergence*; *while if* $q = \infty$, *then* $\{(n^{-1} \log n)^r\}$ *is the optimal rate of convergence.*

This theorem generalizes in various ways previous results for $d = 1$ and $m = 0$ by Halász (1978) and Ibragimov and Hásminskii (1980). According to Stone (1980), $\{n^{-r}\}$ is also the optimal rate of convergence if $\| \hat{T}_n - T(\theta) \|_q$ is replaced by $| \hat{T}_n(x_0) - T(x_0; \theta) |$, where $x_0 \in U$ is fixed. The results on optimal rates of convergence for nonparametric density estimates are surprisingly similar to those for nonparametric regression; see Stone (1980, 1982) and the references cited therein.

The proof that the indicated sequences are lower rates of convergence will be given in Section 2. The proof that the indicated rates are achievable and that certain estimators are asymptotically optimal will be given in Section 3. Each proof is a refinement of the corresponding proof in Stone (1980).

There are several interesting open questions related to Theorem 1. Suppose, say, that $q = 2$. The desire to eliminate the restriction that C be compact leads to the first question.

QUESTION 1. Let Condition 3′ hold and let $\| \hat{T}_n - T(\theta) \|_2$ be replaced by

$$\left(\int_{\mathbb{R}^d} | \hat{T}_n(x) - T(x; \theta) |^2 F(dx) \right)^{1/2}.$$

Under which additional conditions (e.g., on F) is $\{n^{-r}\}$ an achievable rate of convergence?

The next question is suggested by the practical success of projection pursuit regression (see Friedman and Stuetzle, 1981).

QUESTION 2. Let $d \geq 2$ and let $\mathscr{A}$ denote either the collection of functions θ on $\mathbb{R}^d$ which are additive, i.e., of the form

$$\theta(x_1, \cdots, x_d) = \theta_1(x_1) + \cdots + \theta_d(x_d)$$

or the collection of the form

$$\theta(x_1, \cdots, x_d) = \psi(\beta_1 x_1 + \cdots + \beta_d x_d).$$

Let Θ be replaced by $\mathscr{A} \cap \Theta$ and set $r_1 = (p - m)/(2p + 1)$. Is $\{n^{-r_1}\}$ an achievable rate of convergence?

The dependence on Θ of the asymptotically optimal estimators constructed in Section 3 suggests the following:

QUESTION 3. Let Θ_r denote the dependence of Θ, as defined above, on r. Are there positive constants c_r, $r > 0$, and estimators $\hat{T}_n$, defined independently of r, such that

$$\lim_n \sup_{\Theta_r} P_\theta(\| \hat{T}_n - T(\theta) \|_2 \geq c_r n^{-r}) = 0 \quad \text{for all} \quad r > 0?$$

The desire to make nonparametric regression robust leads to the following:

QUESTION 4. Suppose that t is the median of the distribution of $h(y \mid x, t)\phi(dy)$ instead of its mean. Is $\{n^{-r}\}$ still an achievable rate of convergence?

2. Lower rates of convergence. (The reader should study the first half of Section 2 of Stone, 1980, before reading the related but more complicated argument in this section.) First an elementary inequality regarding sums of independent Bernoulli random variables will be recorded.

LEMMA 1. *Let I_v, $v \in V$, be independent Bernoulli random variables such that $I = \sum_v I_v$ has finite mean M. Then $P(I \le M/2) \le (2/e)^{M/2}$.*

PROOF. Clearly

$$E\left[\left(\frac{1}{2}\right)^I\right] = \Pi_v\left(1 - \frac{EI_v}{2}\right) \le e^{-\sum_v EI_v/2} = e^{-M/2},$$

so

$$P(I \le M/2) \le 2^{M/2}E[(1/2)^I] \le (2/e)^{M/2}$$

as desired.

In proving that the indicated sequences are lower rates of convergence it can, without loss of generality, be assumed that C is the cube $[0, 1]^d$ in $\mathbb{R}^d$ consisting of those points all of whose coordinates lie in the interval $[0, 1]$. Let M_n denote a positive integer. Write C as the disjoint union of M_n^d cubes C_{nv} having center x_{nv} and length M_n^{-1}, where $v \in V_n = \{1, \cdots, M_n^d\}$.

Now $Q = \sum_{0 \le j \le m} Q_j$, where $Q_j = \sum_{[\alpha]=j} q_\alpha D^\alpha$ and $Q_m \not\equiv 0$. Let ψ be an infinitely differentiable function which vanishes outside $(-1/2, 1/2)^d$ and is such that $Q_m\psi(0) > 0$ and $|D^\alpha\psi(x) - D^\alpha\psi(x_0)| \le K_2 2^{\beta-1}|x - x_0|^\beta$ for $x_0, x \in \mathbb{R}^d$ and $[\alpha] = k$ (recall (1.2)). Define g_{nv} on U for $v \in V_n$ by $g_{nv}(x) = M_n^{-p}\psi(M_n(x - x_{nv}))$. Given a $\{0, 1\}$-valued sequence $\tau_n = \{\tau_{nv}\}_{v \in V_n}$, set $g_n = \sum_{V_n} \tau_{nv} g_{nv}$. Clearly g_n is an infinitely differentiable function on $\mathbb{R}^d$ which vanishes outside C. It is also easily seen that g_n satisfies (1.2). (First suppose that x and x_0 lie in a common C_{nv}. Next, to handle the general case, consider the straight line from x_0 to x, note that only the two end boxes yield a relevant contribution, and verify that $a^\beta + b^\beta \le 2^{1-\beta}(a + b)^\beta$ for $a, b \ge 0$.)

Let θ_0 be a constant function in θ. Let Θ_n denote the collection of all functions of the form $\theta = \theta_0 + g_n$ as τ_n ranges over the $2^{M_n^d}$ possible sequences. Then $\Theta_n \subset \Theta$ for n sufficiently large. In the argument to follow, it is convenient to think of J_n as a function of θ on Θ_n,

For $v \in V_n$ set $\mathcal{I}_{nv} = \{i : 1 \le i \le n \text{ and } X_i \in C_{nv}\}$, $N_{nv} = \#(\mathcal{I}_{nv})$,

$$\ell_{nv} = \sum_{\mathcal{I}_{nv}} [\ell(Y_i \mid X_i, \theta_0(X_i) + g_{nv}(X_i)) - \ell(Y_i \mid X_i, \theta_0(X_i))],$$

$L_{nv} = \exp(\ell_{nv})$ (the likelihood ratio or Radon-Nikodym derivative of $P_{\theta_0+q_n}$ with respect to P_{θ_0}), and

$$\pi_{nv} = \min\left(\frac{L_{nv}}{L_{nv} + 1}, \frac{1}{L_{nv} + 1}\right).$$

Also set

$$\bar{\tau}_{nv} = \begin{cases} 1 & \text{if } L_{nv} \ge 1, \\ 0 & \text{if } L_{nv} < 1. \end{cases}$$

Let the uniform prior probability distribution be placed on the $2^{M_n^d}$ possible choices of the sequence τ_n. Then the posterior distribution of τ_{nv}, $v \in V_n$, given the data $X_1, Y_1, \cdots, X_n, Y_n$ is that of M_n^d independent Bernoulli random variables with

$$P(\tau_{nv} = 1 \mid X_1, Y_1, \cdots, X_n, Y_n) = \frac{L_{nv}}{L_{nv} + 1}, \quad v \in V_n.$$

Consequently, conditioned on the data, $|\bar{\tau}_{nv} - \tau_{nv}|$ are independent Bernoulli random

variables with

$$P(|\bar{\tau}_{nv} - \tau_{nv}| = 1 \mid X_1, Y_1, \cdots, X_n, Y_n) = \pi_{nv}, \quad v \in V_n.$$

Thus by Lemma 1

$$P(\textstyle\sum_{V_n} |\bar{\tau}_{nv} - \tau_{nv}| \geq \frac{1}{2}\sum_{V_n} \tau_{nv} \mid X_1, Y_1, \cdots, X_n, Y_n) \geq 1 - \left(\frac{2}{e}\right)^{\sum_{V_n} \pi_{nv}/2}. \tag{2.1}$$

Conditioned on N_{nv}, $v \in V_n$, the random variables π_{nv}, $v \in V_n$, are independent and the conditional distribution of π_{nv} given $N_{n\cdot}$ depends only on N_{nv}. Let $0 < q_n < \frac{1}{2}$. Then

$$\begin{aligned} P(\pi_{nv} \geq q_n \mid N_{nv}) &\geq \frac{1}{2} P(\pi_{nv} \geq q_n \mid N_{nv}, \tau_{nv} = 0) \\ &= \frac{1}{2} P\left\{ |\ell_{nv}| \leq \log\left(\frac{1 - q_n}{q_n}\right) \middle| N_{nv}, \tau_{nv} = 0 \right\}, \end{aligned}$$

so

$$P(\pi_{nv} \geq q_n \mid N_{n\cdot}) \geq \frac{1}{2}\left\{1 - \frac{E(|\ell_{nv}| \mid N_{nv}, \tau_{nv} = 0)}{\log\left(\frac{1 - q_n}{q_n}\right)}\right\}. \tag{2.2}$$

Set $\gamma = 1/(2p + d)$ and $W_n = \{v \in V_n : N_{nv} \leq 2nM_n^{-d}\}$. Then $\#(W_n) \geq M_n^d/2$. It follows from Condition 1, by computations similar to those on pages 1352–1353 of Stone (1980), that there is a positive integer n_0 such that for $n \geq n_0$ and $v \in W_n$,

$$E(|\ell_{nv}| \mid N_{nv}, \tau_{nv} = 0) \leq K_1 \max \psi^2 n M_n^{-1/\gamma} + (2K_1 \max \psi^2 n M_n^{-1/\gamma})^{1/2}.$$

Suppose that, for some positive constant K_3,

$$M_n \leq K_3 n^{\gamma}, \quad n \geq n_0. \tag{2.3}$$

Then

$$E(|\ell_{nv}| \mid N_{nv}, \tau_{nv} = 0) \leq K_4 n M_n^{-1/\gamma}, \quad n \geq n_0 \quad \text{and} \quad v \in W_n, \tag{2.4}$$

where

$$K_4 = K_1 \max \psi^2 \{1 + (2K_3^{1/\gamma}/K_1 \max \psi^2)^{1/2}\}.$$

Define q_n implicitly by

$$\log\{(1 - q_n)/q_n\} = 2K_4 n M_n^{-1/\gamma}, \quad n \geq 1. \tag{2.5}$$

Then by (2.2) and (2.4),

$$P(\pi_{nv} \geq q_n \mid N_{n\cdot}) \geq \tfrac{1}{4}, \quad n \geq n_0 \quad \text{and} \quad v \in W_n,$$

so by Lemma 1

$$P\left(\#\{v \in W_n : \pi_{nv} \geq q_n\} \leq \frac{M_n^d}{16} \,\middle|\, N_{n\cdot}\right) \leq \left(\frac{2}{e}\right)^{M_n^d/16}, \quad n \geq n_0,$$

and hence

$$P\left(\textstyle\sum_{V_n} \pi_{nv} \leq \frac{q_n M_n^d}{16} \,\middle|\, N_{n\cdot}\right) \leq \left(\frac{2}{e}\right)^{M_n^d/16}, \quad n \geq n_0.$$

Suppose that

$$\lim_n M_n = \infty. \tag{2.6}$$

Then

$$\lim_n P\left(\sum_{V_n} \pi_{nv} \geq \frac{q_n M_n^d}{16}\right) = 1. \tag{2.7}$$

Suppose that also

$$\lim_n q_n M_n^d = \infty. \tag{2.8}$$

Then by (2.1) and (2.7),

$$\lim_n P\left(\sum_{V_n} | \bar{\tau}_{nv} - \tau_{nv} | \geq \frac{q_n M_n^d}{32}\right) = 1. \tag{2.9}$$

Let $\hat{\tau}_{nv} \in \{0, 1\}$ denote an arbitrary estimator of τ_{nv}, $v \in V_n$, based on $X_1, Y_1, \cdots, X_n, Y_n$.

LEMMA 2. *Suppose that* (2.3), (2.5), (2.6) *and* (2.8) *hold. Then*

$$\lim_n \inf_{\hat{\tau}_n} \sup_{\Theta_n} P_\theta\left(\sum_{V_n} | \hat{\tau}_{nv} - \tau_{nv} | \geq \frac{q_n M_n^d}{32}\right) = 1.$$

PROOF. Consider the loss $L_n(\hat{\tau}_n, \tau_n)$ defined by

$$L_n(\hat{\tau}_n, \tau_n) = \begin{cases} 1 & \text{if } \sum_{V_n} | \hat{\tau}_{nv} - \tau_{nv} | \geq \dfrac{q_n M_n^d}{32}, \\ 0 & \text{otherwise.} \end{cases}$$

The Bayes risk of $\hat{\tau}_n$ is

$$P\left(\sum_{V_n} | \hat{\tau}_{nv} - \tau_{nv} | \geq \frac{q_n M_n^d}{32}\right).$$

It is easily seen that $\bar{\tau}_n$ is a Bayes rule. (Observe that if I_v, $v \in V$, are independent Bernoulli random variables with $P(I_v = 1) = p_v$, then $P(\Sigma_{v \in V} I_v \geq s)$ is a nondecreasing function of p_v, $v \in V$.) The desired result now follows immediately from (2.9).

Suppose $q = \infty$. It will now be shown that $\{(n^{-1} \log n)^r\}$ is a lower rate of convergence. To this end choose $K_5 > 0$ and let M_n be chosen so that

$$M_n = [(K_5 n / \log n)^\gamma]$$

for n sufficiently large, where [ ] denotes the greatest integer function. Then (2.3) holds for, say, $K_3 = 1$ and n_0 sufficiently large, and (2.6) also holds. Furthermore (2.8) holds provided that K_5 is chosen sufficiently large, in which case by Lemma 2

$$\lim_n \inf_{\hat{\tau}_n} \sup_{\Theta_n} P_\theta(\textstyle\sum_{V_n} | \hat{\tau}_{nv} - \tau_{nv} | \geq 1) = 1. \tag{2.10}$$

Let $\hat{T}_n$ be an estimator of $T(\theta)$ based on $X_1, Y_1, \cdots, X_n, Y_n$. Suppose that $\theta = \theta_0 + \sum_{V_n} \tau_{nv} g_{nv} \in \Theta_n$. Define $\hat{\tau}_n$ in terms of $\hat{T}_n$ by

$$\hat{\tau}_{nv} = \begin{cases} 1 & \text{if } | \hat{T}_n(x_{nv}) - T(x_{nv}; \theta_0 + g_{nv})| \leq | \hat{T}_n(x_{nv}) - T(x_{nv}; \theta_0)|, \\ 0 & \text{otherwise.} \end{cases}$$

If $\hat{\tau}_{nv} \neq \tau_{nv}$, then

$$| \hat{T}_n(x_{nv}) - T(x_{nv}; \theta)| \geq \frac{| Q g_{nv}(x_{nv})|}{2}.$$

Now

$$Q g_{nv}(x_{nv}) = \textstyle\sum_0^m M_n^{j-p} Q_j \psi(0)$$

and

$$\lim_n M_n^{p-m} \textstyle\sum_0^m M_n^{j-p} Q_j \psi(0) = Q_m \psi(0) > 0.$$

Since

$$M_n^{m-p} \sim (n^{-1}\log n/K_5)^r,$$

there is a positive constant c and a positive integer n_0 such that

$$\frac{Qg_{nv}(x_{nv})}{2} \geq c\left(\frac{\log n}{n}\right)^r, \quad n \geq n_0 \quad \text{and} \quad v \in V_n.$$

Consequently, if $n \geq n_0$ and $\hat{\tau}_{nv} \neq \tau_{nv}$ for some $v \in V_n$, then $\| \hat{T}_n - T(\theta)\|_\infty \geq c(n^{-1}\log n)^r$. Therefore, by (2.10)

$$\lim_n \inf_{\hat{T}_n} \sup_{\Theta_n} P_\theta\left(\| \hat{T}_n - T(\theta)\|_\infty \geq c\left(\frac{\log n}{n}\right)^r\right) = 1,$$

which implies that $\{(n^{-1}\log n)^r\}$ is a lower rate of convergence.

Suppose instead that $0 < q < \infty$. It will now be shown that $\{n^{-r}\}$ is a lower rate of convergence. Choose $K_6 > 0$ and let M_n be chosen so that $M_n = [(K_6 n)^\gamma]$ for n sufficiently large. Then (2.3), (2.6) and (2.8) hold so it follows from Lemma 2 that for some $\varepsilon > 0$

$$\lim_n \inf_{\hat{\tau}_n} \sup_{\Theta_n} P_\theta(\textstyle\sum_{V_n} | \hat{\tau}_{nv} - \tau_{nv} | \geq \varepsilon M_n^d) = 1. \tag{2.11}$$

There is a positive integer n_0 and $\delta \in (0, \frac{1}{2})$ such that

$$\frac{Qg_{nv}(x)}{2} \geq \delta n^{-r} \quad \text{for} \quad n \geq n_0, v \in V_n \quad \text{and} \quad | x - x_{nv} | \leq \delta M_n^{-1}.$$

Set $D_{nv} = \{x \in C_{nv} : | x - x_{nv}| \leq \delta M_n^{-1}\}$. Then $| D_{nv}| = \rho M_n^{-d}$ for $n \geq 1$ and $v \in V_n$, where ρ is a positive constant and $| A |$ denotes the Lebesgue measure of a Borel subset A of $\mathbb{R}^d$.

Let $\hat{T}_n$ be an estimator of $T(\theta)$ based on $X_1, Y_1, \cdots, X_n, Y_n$. Suppose that $\theta = \theta_0 + \sum_{V_n} \tau_{nv} g_{nv} \in \Theta_n$. Define $\hat{\tau}_n$ in terms of $\hat{T}_n$ by

$$\hat{\tau}_{nv} = \begin{cases} 1 & \text{if } |\{x \in D_{nv} : | \hat{T}_n(x) - T(x; \theta_0 + g_{nv})| < | \hat{T}_n(x) - T(x; \theta_0)|\}| \geq | D_{nv} |/2, \\ 0 & \text{otherwise.} \end{cases}$$

If $n \geq n_0$, $v \in V_n$ and $\hat{\tau}_{nv} \neq \tau_{nv}$, then

$$| \{x \in D_{nv} : | \hat{T}_n(x) - T(x; \theta)| \geq \delta n^{-r}\}| \geq \frac{\rho}{2} M_n^{-d}.$$

Thus by (2.11)

$$\lim_n \inf_{\hat{T}_n} \sup_{\Theta_n} P_\theta\left(|\{x \in C : | \hat{T}_n(x) - T(x; \theta)| \geq \delta n^{-r}\}| \geq \frac{\rho\varepsilon}{2}\right) = 1$$

and hence

$$\lim_n \inf_{\hat{T}_n} \sup_{\Theta_n} P_\theta(\| \hat{T}_n - T(\theta)\|_q \geq cn^{-r}) = 1,$$

where $c = \delta(\rho\varepsilon/2)^{1/q} > 0$. Therefore, $\{n^{-r}\}$ is a lower rate of convergence as desired.

3. Achievability. (The reader should review the discussion of local weighted polynomial regression in the introduction and also study the achievability proof in Section 2 of Stone, 1980, before reading the present section.) Without loss of generality it can be assumed that C is contained in the interior of the cube $C_0 = [-1, 1]^d \subset U$. Let $\{M_n\}$ denote an increasing sequence of positive integers which tends to infinity, but sufficiently slowly so that

$$\lim_n M_n n^{-\lambda} = 0 \quad \text{for some } \lambda \in (0, 1/d). \tag{3.1}$$

Write C_0 as the disjoint union of M_n^d cubes C_{nv} of length $2M_n^{-1}$ where $1 \leq v \leq M_n^d$. For $1 \leq v \leq M_n^d$ set $\mathcal{I}_{nv} = \{i : 1 \leq i \leq n \text{ and } X_i \in C_{nv}\}$ and $N_{nv} = \#(\mathcal{I}_{nv})$. By (3.1) and Condition 3 there is a positive constant K_3 such that

$$\lim_n P(N_{nv} \geq K_3 n M_n^{-d} \text{ for } 1 \leq v \leq M_n^d) = 1. \tag{3.2}$$

Let $\{\delta_n\}$ denote a sequence of positive constants which tend to zero, but sufficiently slowly so that $\delta_n M_n$ tends to infinity. For $x \in C$ set

$$V_n(x) = \{v : 1 \le v \le M_n^d \text{ and } |z - x| \le \delta_n \text{ for all } z \in C_{nv}\}.$$

There are positive constants n_0 and K_4 such that

$$\#(V_n(x)) \ge K_4(\delta_n M_n)^d \quad \text{for} \quad n \ge n_0 \quad \text{and} \quad x \in C. \tag{3.3}$$

Let $\hat{\theta}_n(\cdot; x)$ denote the polynomial on $\mathbb{R}^d$ of degree k which minimizes

$$\sum_{V_n(x)} \frac{1}{N_{nv}} \sum_{\mathscr{I}_{nv}} \{Y_i - \hat{\theta}_n(X_i; x)\}^2.$$

(Assume temporarily that there is a unique such minimizing function.) Write

$$\hat{\theta}_n(z; x) = \sum_A \hat{b}_{n\alpha}(x) \frac{(z - x)^\alpha}{\delta_n^{[\alpha]}},$$

where A denotes the collection of all d-tuples α of nonnegative integers such that $[\alpha] \le k$. Then

$$\hat{b}_{n\alpha}(x) = (\mathscr{A}_n^{-1}(x)\, \mathscr{X}_n'(x)\, \mathscr{Y}_n(x))_\alpha, \tag{3.4}$$

where $\mathscr{Y}_n(x)$, $\mathscr{X}_n(x)$ and $\mathscr{A}_n(x)$ are defined as follows: $\mathscr{Y}_n(x)$ is the n-dimensional (column) vector defined by

$$\mathscr{Y}_{ni}(x) = \begin{cases} Y_i, & i \in \mathscr{I}_n(x) = \cup_{V_n(x)} \mathscr{I}_{nv}, \\ 0, & i \notin \mathscr{I}_n(x). \end{cases}$$

$\mathscr{X}_n(x)$ is the $n \times \#(A)$ matrix defined by

$$\mathscr{X}_{ni\alpha}(x) = \begin{cases} \dfrac{1}{\#(V_n(x))N_{nv}} \dfrac{(X_i - x)^\alpha}{\delta_n^{[\alpha]}}, & i \in \mathscr{I}_n(x) \quad \text{and} \quad \alpha \in A, \\ 0, & i \notin \mathscr{I}_n(x) \quad \text{and} \quad \alpha \in A, \end{cases}$$

where for $i \in \mathscr{I}_n(x)$, v is the unique member of $V_n(x)$ such that $i \in \mathscr{I}_{nv}$. $\mathscr{A}_n(x)$ is the $\#(A) \times \#(A)$ matrix defined by

$$\begin{aligned} \mathscr{A}_{n\alpha\beta}(x) &= \frac{1}{\#(V_n(x))} \sum_{V_n(x)} \frac{1}{N_{nv}} \sum_{\mathscr{I}_{nv}} \frac{(X_i - x)^\alpha (X_i - x)^\beta}{\delta_n^{[\alpha]+[\beta]}} \\ &= \#(V_n(x)) \sum_{V_n(x)} N_{nv} \sum_{\mathscr{I}_{nv}} \mathscr{X}_{ni\alpha}(x)\, \mathscr{X}_{ni\beta}(x). \end{aligned}$$

The indicated minimization problem has a unique solution if and only if $\det \mathscr{A}_n(x) > 0$.

Define the $\#(A) \times \#(A)$ matrix $\mathscr{A}$ by

$$\mathscr{A}_{\alpha\beta} = \int_{|z| \le 1} z^\alpha z^\beta \, dz \Big/ \int_{|z| \le 1} dz.$$

Then $\det \mathscr{A} > 0$ (see pages 1354–1355 of Stone, 1980). By (3.2) for every $\varepsilon > 0$,

$$\lim_n P(|\mathscr{A}_{n\alpha\beta}(x) - \mathscr{A}_{\alpha\beta}| \le \varepsilon \text{ for all } x \in C \text{ and } \alpha, \beta \in A) = 1.$$

Consequently for every $\varepsilon > 0$,

$$\lim_n P(|(\mathscr{A}_n^{-1}(x))_{\alpha\beta} - (\mathscr{A}^{-1})_{\alpha\beta}| \le \varepsilon \text{ for all } x \in C \text{ and } \alpha, \beta \in A) = 1. \tag{3.5}$$

Let $\hat{T}_n$ denote the estimator of $T(\theta) = Q\theta$ defined by $\hat{T}_n(x) = Q\hat{\theta}_n(x; x)$. Define the vector $\mathscr{Q}_n \in \mathbb{R}^{\#(\alpha)}$ by

$$\mathscr{Q}_{n\alpha} = \frac{\alpha! q_\alpha}{\delta_n^{[\alpha]}}, \qquad \alpha \in A,$$

where $q_\alpha = 0$ for $m < [\alpha] \leq k$. By (3.4)

$$\hat{T}_n(x) = \mathscr{Q}'_n \mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x)\, \mathscr{Y}_n(x). \tag{3.6}$$

Let $\theta_k(\cdot; x)$ denote the Taylor polynomial approximation to θ of degree k about x, defined by

$$\theta_k(z; x) = \sum_{[\alpha]\leq k} \frac{D^\alpha \theta(z)}{\alpha!} (z - x)^\alpha,$$

where $z^\alpha = z_1^{\alpha_1} \cdot \ldots \cdot z_d^{\alpha_d}$ and $\alpha! = \alpha_1! \cdot \ldots \cdot \alpha_d!$. Define the vectors $\mathscr{J}_n(x)$ and $\mathscr{J}_{kn}(x)$ in $\mathbb{R}^n$ by

$$\mathscr{J}_{ni}(x) = \theta(X_i), \qquad i \in \mathscr{I}_n(x),$$

$$\mathscr{J}_{kni}(x) = \theta_k(X_i; x), \qquad i \in \mathscr{I}_n(x),$$

and $\mathscr{J}_{ni}(x) = \mathscr{J}_{kni}(x) = 0$ for $i \notin \mathscr{I}_n(x)$. Now

$$\mathscr{J}_{kni}(x) = \sum_A \frac{(X_i - x)^\alpha}{\delta_n^{[\alpha]}} \frac{\delta_n^{[\alpha]} D^\alpha \theta(x)}{\alpha!}, \qquad i \in \mathscr{I}_n(x),$$

from which it follows easily that

$$(\mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x)\, \mathscr{J}_{kn}(x))_\alpha = \frac{\delta_n^{[\alpha]} D^\alpha \theta(x)}{\alpha!}$$

and therefore that

$$T(x; \theta) = \mathscr{Q}'_n \mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x)\, \mathscr{J}_{kn}(x). \tag{3.7}$$

By (3.6) and (3.7)

$$\begin{aligned} \hat{T}_n(x) - T(x; \theta) = {} & \mathscr{Q}'_n \mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x) \{\mathscr{Y}_n(x) - \mathscr{J}_n(x)\} \\ & + \mathscr{Q}'_n \mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x) \{\mathscr{J}_n(x) - \mathscr{J}_{kn}(x)\}. \end{aligned} \tag{3.8}$$

There is a positive constant K_5 such that for n sufficiently large

$$|\, \mathscr{J}_{ni}(x) - \mathscr{J}_{kni}(x)\,| \leq K_5 |\, X_i - x\,|^p \leq K_5 \delta_n^p, \qquad x \in C \quad \text{and} \quad i \in \mathscr{I}_n(x).$$

Thus by (3.5) there is a positive constant K_6 such that

$$\lim_n \sup_\Theta P_\theta(|\, \mathscr{Q}'_n \mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x) \{\mathscr{J}_n(x) - \mathscr{J}_{kn}(x)\}\,| \geq K_6 \delta_n^{p-m} \text{ for some } x \in C) = 0. \tag{3.9}$$

Observe that

$$\mathscr{Q}'_n \mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x) \{\mathscr{Y}_n(x) - \mathscr{J}_n(x)\} = \sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x) \{Y_i - \theta(X_i)\}, \qquad x \in C, \tag{3.10}$$

where

$$\mathscr{W}_n(x) = (\mathscr{W}_{ni}(x)) = \mathscr{Q}'_n \mathscr{A}_n^{-1}(x)\, \mathscr{X}'_n(x), \qquad x \in C.$$

By (3.2), (3.3) and (3.5) there is a positive constant K_7 such that

$$\lim_n P(|\, \mathscr{W}_n(x)\,|^2 \leq K_7 n^{-1} \delta_n^{-(2m+d)} \text{ for all } x \in C) = 1;$$

in other words $\lim_n P(\Omega_n) = 1$, where Ω_n is the event that $|\, \mathscr{W}_n(x)\,|^2 \leq K_7 n^{-1} \delta_n^{-(2m+d)}$ for $x \in C$.

Suppose now that $0 < q < \infty$. It will be shown that $\{n^{-r}\}$ is an achievable rate of convergence. To this end observe first that by Condition 2 there is a positive constant K_8 (which can depend on q) such that

$$E_\theta(|\, \textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x) \{Y_i - \theta(X_i)\}\,|^q \,|\, X_1, \cdots, X_n) \leq K_8 \{n^{-1} \delta_n^{-(2m+d)}\}^{q/2}$$

for $x \in C$ on Ω_n. (It is enough to prove this result for q an even integer.) Consequently,

there is a positive constant K_9 such that

$$E_\theta\left[\int_C |\textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\}|^q\, dx \,|\, X_1, \cdots, X_n\right] \le K_9\{n^{-1}\delta_n^{-(2m+d)}\}^{q/2} \quad \text{on} \quad \Omega_n. \tag{3.11}$$

Now

$$\mathrm{Var}_\theta\left[\int_C |\textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\}|^q\, dx \,|\, X_1, \cdots, X_n\right]$$

$$= \int\int_{x_1,x_2\in C, |x_1-x_2|\le 2\delta_n} \mathrm{cov}_\theta[|\textstyle\sum_{\mathscr{I}_n(x_1)} \mathscr{W}_{ni}(x_1)\{Y_i - \theta(X_i)\}|^q,$$

$$|\textstyle\sum_{\mathscr{I}_n(x_2)} \mathscr{W}_{ni}(x_2)\{Y_i - \theta(X_i)\}|^q \,|\, X_1, \cdots, X_n]\, dx_1\, dx_2$$

on Ω_n, so there exist positive constants κ_n tending to zero such that

$$\mathrm{Var}_\theta\left[\int_C |\textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\}|^q\, dx \,|\, X_1, \cdots, X_n\right] \le \kappa_n\{n^{-1}\delta_n^{-(2m+d)}\}^q \quad \text{on} \quad \Omega_n. \tag{3.12}$$

By (3.11) and (3.12) there is a positive constant c_1 and there are positive constants λ_n tending to zero such that

$$P_\theta\left[\int_C |\textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\}|^q\, dx \ge \{c_1^2 n^{-1}\delta_n^{-(2m+d)}\}^{q/2} \,|\, X_1, \cdots, X_n\right] \le \lambda_n \text{ on } \Omega_n.$$

Since $\lim_n P(\Omega_n) = 1$,

$$\lim_n \sup_\Theta P_\theta\left[\int_C |\textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\}|^q\, dx \ge (c_1^2 n^{-1}\delta_n^{-(2m+d)})^{q/2}\right] = 0. \tag{3.13}$$

By (3.8)–(3.10) and (3.13), there is a positive constant c such that

$$\lim_n \sup_\Theta P_\theta\left[\| \hat{T}_n - T(\theta)\|_q \ge \frac{c}{2}\,\delta_n^{-m}\{\delta_n^p + (n\delta_n^d)^{-1/2}\}\right] = 0. \tag{3.14}$$

Choose δ_n so that $\delta_n^p = (n\delta_n^d)^{-1/2}$ or, equivalently, $\delta_n = n^{-1/(2p+d)} = n^{-\gamma}$. Then

$$\frac{c}{2}\,\delta_n^{-m}\{\delta_n^p + (n\delta_n^d)^{-1/2}\} = cn^{-r},$$

so it follows from (3.14) that $\{n^{-r}\}$ is an achievable rate of convergence.

Suppose instead that $q = \infty$. It will be shown that $\{(n^{-1}\log n)^r\}$ is an achievable rate of convergence. Without loss of generality it can be assumed that $C = [-\frac{1}{2}, \frac{1}{2}]^d$. It follows from Condition 2 that there are positive constants K_{10} and s_0 such that

$$\int e^{s(y-t)} h(y \,|\, x, t)\phi(dy) \le \exp(K_{10}s^2), \qquad |s| \le s_0,$$

for $x \in U$ and $t \in \mathscr{J}$. (Recall that t is the mean of $h(y \,|\, x, t)\phi(dy)$.) By (3.2), (3.3) and (3.5) there is a positive constant K_{11} such that $\lim_n P(\Omega_n') = 1$, where Ω_n' is the intersection of Ω_n and the event that

$$|\mathscr{W}_{ni}(x)| \le K_{11} n^{-1}\delta_n^{-(m+d)} \quad \text{for} \quad x \in C \quad \text{and} \quad i \in \mathscr{I}_n(x).$$

Set $K_{12} = K_7K_{10}$. Then

$$E_\theta(\exp[s_n \textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\}] \mid X_1, \cdots, X_n) \le \exp\{K_{12}s_n^2 n^{-1}\delta_n^{-(2m+d)}\}$$

for $x \in C$ on Ω_n' provided that

$$K_{11} \mid s_n \mid n^{-1}\delta_n^{-(m+d)} \le s_0. \tag{3.15}$$

Therefore

$$P_\theta[\mid \textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\} \mid \ge t_n \mid X_1, \cdots, X_n] \le 2\exp\{K_{12}s_n^2 n^{-1}\delta_n^{-(2m+d)} - s_n t_n\}$$

for $x \in C$ on Ω_n' provided that s_n, $t_n \in (0, \infty)$ and (3.15) holds. It follows by (optimally) choosing $s_n = nt_n\delta_n^{2m+d}/2K_{12}$ that

$$P_\theta[\mid \textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\} \mid \ge t_n \mid X_1, \cdots, X_n] \le 2\exp\{-nt_n^2\delta_n^{2m+d}/(4K_{12})\}$$

for $x \in C$ on Ω_n' provided that $t_n > 0$ and

$$K_{11}t_n\delta_n^m \le 2K_{12}s_0. \tag{3.16}$$

Let C_n be a finite subset of C such that $\#(C_n) \le n^{K_{13}}$ for some fixed positive constant K_{13}. Then

$$P_\theta[\max_{C_n} \mid \textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\} \mid \ge t_n \mid X_1, \cdots, X_n] \le 2n^{K_{13}}\exp\{-nt_n^2\delta_n^{2m+d}/(4K_{12})\}$$

on Ω_n' provided that $t_n > 0$ and (3.16) holds. Consequently, for some $c_1 > 0$

$$\lim_n \sup_\Theta P_\theta[\max_{C_n} \mid \textstyle\sum_{\mathscr{I}_n(x)} \mathscr{W}_{ni}(x)\{Y_i - \theta(X_i)\} \mid \ge c_1(n^{-1}\delta_n^{-2m-d}\log n)^{1/2}] = 0 \tag{3.17}$$

provided that

$$\lim_n n^{-1}\delta_n^{-d}\log n = 0. \tag{3.18}$$

By (3.8)–(3.10) and (3.17), if (3.18) holds, then for some $c > 0$

$$\lim_n \sup_\Theta P_\theta\left[\max_{C_n} \mid \hat{T}_n(x) - T(x;\theta) \mid \ge \frac{c}{4}\{\delta_n^{p-m} + (n^{-1}\delta_n^{-2m-d}\log n)^{1/2}\}\right] = 0. \tag{3.19}$$

Choose $\delta_n > 0$ so that

$$\delta_n^{p-m} = (n^{-1}\delta_n^{-2m-d}\log n)^{1/2}$$

or equivalently

$$\delta_n = \left(\frac{\log n}{n}\right)^{1/(2p+d)} = \left(\frac{\log n}{n}\right)^{\gamma},$$

which satisfies (3.18). Then by (3.19)

$$\lim_n \sup_\Theta P_\theta\left\{\max_{C_n} \mid \hat{T}_n(x) - T(x;\theta) \mid \ge \frac{c}{2}\left(\frac{\log n}{n}\right)^{r}\right\} = 0. \tag{3.20}$$

In order to complete the proof of the theorem it is convenient to restrict $\hat{T}_n$ to a grid $C_n \subset C$ and define a new estimator $\bar{T}_n$ on all of C by linear interpolation. This modification was proposed from a practical viewpoint in Stone (1975). When $d = 2$ it is particularly useful if contour plots of the estimates are desired.

Set $L_n = [n^{K_{13}}]$ for some constant $K_{13} > r/\min(p, 1)$. Let C_n be the collection of $(2L_n + 1)^d$ points in C each of whose coordinates is of the form $j/(2L_n)$ for some integer j such that $|j| \le L_n$. Correspondingly C can be written as the union of $(2L_n)^d$ subcubes, each having length $(2L_n)^{-1}$ and all of its vertices in C_n. Each $x \in C$ can be written as a convex combination $\sum \lambda_w w$ of the vertices of one of these subcubes. Set $\bar{T}(x;\theta) = \sum \lambda_w T(w;$

θ). There is a positive constant K_{14} such that

$$\max_{x\in C}|\bar{T}(x;\theta) - T(x;\theta)| \le \frac{K_{14}}{L_n^{\min(p,1)}} = o\left(\left(\frac{\log n}{n}\right)^r\right).$$

Set $\bar{T}_n(x) = \sum \lambda_w \hat{T}_n(w)$, where $x = \sum \lambda_w w$. Then

$$\begin{aligned}|\bar{T}_n(x) - T(x;\theta)| &\le \sum \lambda_w |\hat{T}_n(w) - T(w;\theta)| + |\bar{T}(x;\theta) - T(x;\theta)| \\ &\le \max_{C_n}|\hat{T}_n(x) - T(x;\theta)| + o\left(\left(\frac{\log n}{n}\right)^r\right),\end{aligned}$$

so by (3.20)

$$\lim_n \sup_\Theta P_\theta\left\{\|\bar{T}_n - T(\theta)\|_\infty \ge c\left(\frac{\log n}{n}\right)^r\right\} = 0$$

and hence $\{(n^{-1}\log n)^r\}$ is an achievable rate of convergence. This completes the proof of Theorem 1.

Acknowledgment. The author wishes to thank Richard Olshen and David Hinkley, who each made several very helpful suggestions for improving the readability of the paper.

REFERENCES

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1983). *Tree-structured Methods for Classification and Regression.* Wadsworth, Belmont, to appear.

DEVROYE, L. P. and WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GORDON, L. and OLSHEN, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **10** 611–627.

HALÁSZ, G. (1978). Statistical Interpolation. *Fourier Analysis and Approximation Theory*, Vol. 1, 403–410, G. Alexits and P. Turán (eds.) North Holland, Amsterdam.

IBRAGIMOV, I. A. and HAŚMINSKII, R. Z. (1980). On nonparametric estimation of regression. *Soviet Math. Dokl.* **21** 810–814.

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

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.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 549–645.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. *Recent Advances in Statistics: Papers Presented in Honor of Herman Chernoff's Sixtieth Birthday*, M. H. Rizvi, J. S. Rustagi, and D. Siegmund, (eds.)Academic Press, New York, to appear.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720