ASYMPTOTIC PROPERTIES OF KERNEL ESTIMATORS BASED ON LOCAL MEDIANS

by

Young K. Truong
Department of Biostatistics
University of North Carolina, Chapel Hill

Institute of Statistics Mimeo Series #1831

October 1987

ASYMPTOTIC PROPERTIES OF KERNEL ESTIMATORS BASED ON LOCAL MEDIANS

By YOUNG K. TRUONG
Department of Biostatistics
University of North Carolina, Chapel Hill

July 1987

Abstract. Let $(\mathbf{X}, Y)$ be a pair of random variables that are respectively d and 1 dimensional and set $r = (2+d)^{-1}$. Consider $\theta(\mathbf{X}) = \text{Median}(Y \mid \mathbf{X})$ and let $\hat{\theta}_n(\cdot)$ be an estimator of $\theta(\cdot)$ based on a training sample of size n. Under some regularity conditions, $\hat{\theta}_n(\cdot)$ can be chosen to achieve the optimal rate of convergence n^{-r} both pointwise and in L^q norms $(1 \leq q < \infty)$ restricted to compacts. Futhermore, it also achieves the optimal rate of convergence $(n^{-1}\log n)^r$ in L^∞ norm restricted to compacts. For this class of nonparametric estimators, the results presented in this paper constitute an answer to one of the open questions of Stone (1982, Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053).

Keywords. Kernel estimators, local averages, local medians, rates of convergence, nonparametric regression.

1. INTRODUCTION

Let $(\mathbf{X}, Y)$ be a pair of random variables that are respectively d and 1 dimensional; the random variable Y is called the response and the random vector $\mathbf{X}$ is refered to as the predictor variable. One of the important problems in statistics is to construct a function $\theta(\cdot)$ in order to (i) study the relationship between the response and the explanatory variable or (ii) obtain the predictor $\theta(\mathbf{X})$ of Y based on $\mathbf{X}$.

The simplest and most widely used measure of accuracy of $\theta(\mathbf{X})$ as a predictor of Y is the *Mean Square Error*, $E|Y - \theta(\mathbf{X})|^2$. The function $\theta(\cdot)$ which minimizes this measure of accuracy is the regression function of Y on $\mathbf{X}$, defined by $\theta(\mathbf{X}) = E(Y \,|\, \mathbf{X})$.

Recently, there has been an increasing interest in adopting the *Mean Absolute Deviation* $E|Y - \theta(\mathbf{X})|$ as a measure of accuracy, especially when outliers may be present (Bloomfield and Steiger, 1983). The optimal function $\theta(\cdot)$ is now defined so that $\theta(\mathbf{X})$ is the conditional median, Median$(Y \,|\, \mathbf{X})$, of Y given $\mathbf{X}$. Note that this function is not necessary uniquely defined.

In practice, it is necessary to construct estimators of these functions based on a training sample from the distribution of $(\mathbf{X}, Y)$.

Parametric Approach vs Nonparametric Approach

To estimate these predictors, the ***parametric*** approach starts with specific assumptions about the relationship between the response and the explanatory variables and about the variation in the response that may or may not be accounted for by the explanatory variables. For instance, the standard regression method starts with an a priori model for the regression function $\theta(\cdot)$ which, by assumption or prior knowledge, is a linear function that contains finitely many unknown parameters. Under the assumption that the joint distribution is Gaussian, it is an optimal prediction rule; if the distribution is non-Gaussian, it is not generally posible to determine the funtion $\theta(\cdot)$; so one might settle for the *best* linear predictor. By constrast, in the *nonparametric* approach, the regression function will be estimated directly without assuming such an a priori model for $\theta(\cdot)$. As pointed out

in Stone (1985), the nonparametric approach is more *flexible* than the standard regression method; *flexibility* means the ability of the model to provide accurate fits in a wide variety of realistic situations, inaccuracy here leading to *bias* in estimation.

The present approach deals with the asymptotic properties (in terms of rates of convergence) of a class of nonparametric estimators constructed by kernel methods based on local medians. It is hoped that the results obtained here serve as a starting point for further development and understanding of the sampling properties of more complicated nonparametric procedures involving robustification, local polynomial fits, additive regression, and spline approximation.

Some previous work on the nonparametric regression will now be surveyed. In Stone (1977) a consistency theorem was obtained for a large class of nonparametric regression estimators and used to establish the consistency of nearnest neighbor estimators. Since then, consistency has been established for kernel estimators by Devroye and Wagner (1980a, 1980b) and Spiegelman and Sacks (1980) and for partition estimators by Gordon and Olshen (1980) and Breiman, et al. (1984).

Nearest neighbor, kernel, and recursive 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 together with theoretical justification in terms of optimal rates of convergence are given in Stone (1980, 1982).

Recently, there has been an interest in robustifying the nonparametric regression by estimating the conditional median. In Stone (1977) a consistency theorem was obtained for nearest neighbor estimators. Härdle and Luckhaus (1984) considered the L^∞ rate of convergence for a class of robust nonparametric estimators including an estimator of the conditional median. But the problem of L^q $(1 \leq q < \infty)$ rates of convergence was still unsolved.

Both the pointwise (local) and the L^q $(1 \leq q \leq \infty)$ (global) rates of convergence for kernel estimators based on local medians will be described in Section 2. For this class of nonparametric estimators, the results presented there constitute an answer to one of

the open questions of Stone (1982). Proofs of these results are given in Section 4, which include a different and more intuitive proof (compared to Härdle and Luckhaus, 1984) of the uniform rate of convergence.

2. NONPARAMETRIC ESTIMATION OF THE CONDITIONAL MEDIAN

Results on the local and global rates of convergence of nonparametric estimators of the conditional median, based on a random sample from the distribution of $(\mathbf{X}, Y)$, will be treated in this section. Let d denote a positive integer and let U be a nonempty bounded open neighborhood of the origin of $\mathbf{R}^d$. Let $(\mathbf{X}, Y)$ denote a pair of random variables such that $\mathbf{X} \in \mathbf{R}^d$, $Y \in \mathbf{R}$ and define $\theta(\cdot)$ by

$$\theta(\mathbf{x}) = \text{Median}(Y \mid \mathbf{X} = \mathbf{x}) = \text{Median}(Y \mid X_1 = x_1, \ldots, X_d = x_d).$$

where $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbf{R}^d$ and $\mathbf{X} = (X_1, \ldots, X_d)$. Given $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbf{R}^d$, set $\|\mathbf{x}\| = (x_1^2 + \cdots + x_d^2)^{1/2}$. The rate of convergence of the estimator treated here depends on the the following smoothness assumption on $\theta(\cdot)$.

Assumption 1. *There is a positive constant M_0 such that*

$$|\theta(\mathbf{x}) - \theta(\mathbf{x}')| \leq M_0 \|\mathbf{x} - \mathbf{x}'\| \quad \text{for } \mathbf{x}, \mathbf{x}' \in U.$$

(If U is convex, the above condition is implied by an appropriate boundedness condition on the restriction to U of the first derivative of θ.)

A condition on the distribution of the explanatory variables is required to guarantee the achievability of the desired rate of convergence.

Assumption 2. *The distribution of $\mathbf{x} = (x_1, \ldots, x_d)$ is absolutely continuous and its density $f(\cdot)$ is bounded away from zero and infinity on U; that is, there is a positive constant M_1 such that $M_1^{-1} < f(\cdot) < M_1$ for $\mathbf{x} \in U$.*

A condition on the conditional distribution of Y given $\mathbf{X}$ is required to guarantee the uniqueness of the conditional median (uniqueness will ensure consistency) and also the

achievability of the desired rate of convergence. If the conditional density is not bounded away from zero around the median the desired rate of convergence will not be achievable. (The same condition is required in order to obtain the usual asymptotic result about the sample median in the univariate case.)

Assumption 3. *The conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ is absolutely continuous and its density $h(y \mid \mathbf{x}, \theta)$ is bounded away from zero and infinity over a neighborhood of the median; that is, there is a positive constant ϵ_0 such that $M_1^{-1} \leq h(y \mid \mathbf{x}, \theta) \leq M_1$ for $\mathbf{x} \in U$ and $y \in (\theta(\mathbf{x}) - \epsilon_0, \theta(\mathbf{x}) + \epsilon_0)$.*

The kernel estimator $\hat{\theta}_n(\cdot)$ of $\theta(\cdot)$ will now be described. Given $n \geq 1$, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ denote a random sample of size n from the distribution of $(\mathbf{X}, Y)$. Let δ_n, $n \geq 1$, be positive numbers that tend to zero as n tends to infinity. For $\mathbf{x} \in C$, set $I_n(\mathbf{x}) = 1 \leq i \leq n$ and $\|\mathbf{X}_i - \mathbf{x}\| \leq \delta_n\}$, $N_n(\mathbf{x}) = \#I_n(\mathbf{x})$ and, $\hat{\theta}_n(\mathbf{x}) = \text{Median}\{Y_i : i \in I_n(\mathbf{x})\}$ (use the average of the two middle ordered statistics if $N_n(\mathbf{x})$ is even).

Given positive numbers a_n and b_n, $n \geq 1$, let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity. Given random variables $V_n, n \geq 1$, let $V_n = O_{pr}(b_n)$ mean that the random variables $b_n^{-1}V_n, n \geq 1$ are bounded in probability or, equivalently, that

$$\lim_{c \to \infty} \liminf_n P(|V_n| > cb_n) = 0.$$

Set $r = (2 + d)^{-1}$.

Theorem 1. *Suppose that Assumptions 1–3 hold and that $\delta_n \sim n^{-r}$. Then*

$$|\hat{\theta}_n(\mathbf{0}) - \theta(\mathbf{0})| = O_{pr}(n^{-r}).$$

Let C be a fixed compact subset of U having nonempty interior and let $g(\cdot)$ be a real-valued function on U. Set

$$\|g\|_q = \left\{ \int_C |g(\mathbf{x})|^q \, d\mathbf{x} \right\}^{\frac{1}{q}}, \qquad 1 \leq q < \infty;$$
$$\|g\|_\infty = \sup_{\mathbf{x} \in C} |g(\mathbf{x})|.$$

Theorem 2. *Suppose that Assumptions 1–3 hold and that* $\delta_n \sim (n^{-1}\log n)^r$. *Then there exists a* $c > 0$ *such that*

$$\lim_n P\left(\|\hat{\theta}_n - \theta\|_\infty \geq c(n^{-1}\log n)^r\right) = 0.$$

Theorem 3. *Suppose that Assumptions 1–3 hold and that* $\delta_n \sim n^{-r}$. *Then there exists a* $c > 0$ *such that*

$$\lim_n P(\|\hat{\theta}_n - \theta\|_q \geq cn^{-r}) = 0 \quad \text{for } 1 \leq q < \infty.$$

Proofs of these theorems are given in Section 4. The proof of Theorem 2 is simpler and more intuitive than the corresponding proof given by Härdle and Luckhaus (1984) (only the calculation of binomial probabilities is required).

With a simple modification of Assumption 3, Theorems 1–3 are easily extended to yield rates of convergence for nonparametric estimators of other conditional quantiles.

3. DISCUSSION

For $n \geq 1$, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a random sample of size n from the distribution of $(\mathbf{X}, Y)$ and let k denote a non-negative integer. Let $\theta(\cdot)$ be the regression function of Y on $\mathbf{X}$ and suppose that $\theta(\cdot)$ has bounded $(k+1)$th derivative. Set $r = p/(2p+d)$ where $p = k+1$. Stone (1980, 1982) showed that if $1 \leq q < \infty$, then n^{-r} is the optimal rate of convergence in both pointwise and L^q norms; while $(n^{-1}\log n)^{-r}$ is the optimal rate of convergence in L^∞ norm. To find an estimator of $\theta(\cdot)$ that achieves these optimal rates of concergence, given $\mathbf{x}$, let $\hat{P}_n(\cdot;\mathbf{x})$ be the polynomial on $\mathbf{R}^d$ of degree k that minimizes

$$\sum_{I_n(\mathbf{x})} [Y_i - \hat{P}_n(\mathbf{X}_i;\mathbf{x})]^2$$

and set $\hat{\theta}_n(\mathbf{x}) = \hat{P}_n(\mathbf{x};\mathbf{x})$ (if $q = \infty$, define $\hat{\theta}_n$ as above over a finite subset of C and then extend it to all of C by suitable interpolation). Note that this estimator can be easily obtained by solving the corresponding normal equation.

Based on results presented in the previous sections, the following generalization to the case of conditional median seems plausible. Suppose that the conditional median $\theta(\cdot)$ has

bounded pth derivative. To find an estimator that achieves the above L^q $(1 \leq q \leq \infty)$ rates of convergence, given $\mathbf{x}$, let $\hat{P}_n(\cdot;\mathbf{x})$ be a polynomial on $\mathbf{R}^d$ of degree k which minimizes

$$\sum_{I_n(\mathbf{x})} |Y_i - \hat{P}_n(\mathbf{X}_i;\mathbf{x})|$$

and set $\hat{\theta}_n(\mathbf{x}) = \hat{P}_n(\mathbf{x};\mathbf{x})$. Though there may not be a unique solution, this numerical optimization problem is readily solved by the simplex method (see, for example, Bloomfield and Steiger (1983)). It is an interesting open question to determine whether the asymptotic properties described above (for $p > 2$) still hold in this context.

One drawback that the nonparametric approach has is the high *dimensionality*, which can be thought of in terms of the *variance* in estimation. In other words: A *huge* data set may be required for nonparametric estimation of a function of many variables; otherwise the variance of the estimator may be unacceptably large.

A possible solution would be to use *additivity* as in Stone (1985) to alleviate *curse of dimensionality*. More formally, let $\theta(\cdot)$ be the regression function defined on $\mathbf{R}^d$ and suppose that θ is additive; that is, that there is smooth functions $\theta_1(\cdot),\ldots,\theta_d(\cdot)$ defined on $\mathbf{R}^1$ such that

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

where $\mu = E(Y)$. Using B-splines, an estimator of $\theta(\cdot)$ can be constructed to achieve the optimal rates of convergence n^{-r}, where r now is equal to $p/(2p+1)$. The rates of convergence here do not depend on the dimensional parameter d. Another nice feature about this estimator is that it is smoother and is as flexible as ordinary nonparametric procedures constructed by the kernel method.

The corresponding methodology is generalized immediately to the estimation of conditional medians, and it is an interesting open problem to determine whether the asymptotic properties described above (with r independent of d) also hold in this context.

4. PROOF OF THEOREMS OF SECTION 2

Proof of Theorem 1. By symmetry, it suffices to show that

$$\lim_{c\to\infty} \limsup_n P(\hat{\theta}_n(\mathbf{0}) > \theta(\mathbf{0}) + cn^{-r}) = 0. \tag{4.1}$$

Set $N_n = N_n(\mathbf{x})$ and $I_n = I_n(\mathbf{x})$ for $\mathbf{x} \in U$. The proof of (4.1) depends on the following lemma, whose proof uses the Bernstein and Hoeffding's inequalities. Let ϵ_n denote a sequence of positive numbers tending to zero as $n \to \infty$.

Lemma 1. *Suppose that Assumptions 1–3 hold and that c is a positive constant greater than M_0. Then there are positive constants c_1 and c_2 such that*

$$P\left(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i \geq \theta(\mathbf{x})+c\delta_n\}} \geq \frac{1}{2} - \epsilon_n\delta_n\right) \leq \exp\left(-(c-M_0)^2 c_1 n\delta_n^{d+2}\right) + \exp(-c_2 n\delta_n^d).$$

Proof. Let $K_n = K_n(\mathbf{x})$ and $K_{ni} = K_{ni}(\mathbf{x})$ denote respectively the events $\{\|\mathbf{X}-\mathbf{x}\| \leq \delta_n\}$ and $\{\|\mathbf{X}_i - \mathbf{x}\| \leq \delta_n\}$. According to Assumtion 2.1, $\theta(\mathbf{X}) \leq \theta(\mathbf{x}) + M_0\delta_n$ whenever $\|\mathbf{X}-\mathbf{x}\| \leq \delta_n$. Thus

$$\tfrac{1}{2} - P(Y \geq \theta(\mathbf{x}) + c\delta_n \,|\, K_n) = P(M_0\delta_n \leq Y - \theta(\mathbf{x}) \leq c\delta_n \,|\, K_n).$$

Hence by Assumption 3, there is a positive constant k_0 such that (note that $\epsilon_n \to 0$ as $n \to \infty$)

$$\tfrac{1}{2} - \epsilon_n\delta_n - P(Y \geq \theta(\mathbf{x}) + c\delta_n \,|\, K_n) \geq (c - M_0)k_0\delta_n \qquad \text{for } c > M_0. \tag{4.2}$$

Set $p_n \equiv p_n(\mathbf{x}) = P(\|\mathbf{X}-\mathbf{x}\| \leq \delta_n)$, $R_i = \frac{1}{2} - \epsilon_n\delta_n - P(Y_i \geq \theta(\mathbf{x}) + c\delta_n \,|\, K_{ni})$ and $Z_i = 1_{\{Y_i \geq \theta(\mathbf{x})+c\delta_n\}} - P(Y_i \geq \theta(\mathbf{x}) + c\delta_n \,|\, K_{ni})$ for $i \in I_n$. Then by (4.2),

$$N_n^{-1}\sum_{I_n} R_i \geq (c-M_0)k_0\delta_n \qquad \text{for } c > M_0.$$

Thus

$$\begin{aligned} P(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i \geq \theta(\mathbf{x})+c\delta_n\}} \geq \frac{1}{2} - \epsilon_n\delta_n) &\leq E\left[P^X(N_n^{-1}\textstyle\sum_{I_n} Z_i \geq N_n^{-1}\sum_{I_n} R_i)\right] \\ &\leq E\left[P^X(N_n^{-1}\textstyle\sum_{I_n} Z_i \geq (c-M_0)k_0\delta_n)\right], \end{aligned} \tag{4.4}$$

where $P^X(\cdot) = P(\cdot \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n)$. It follows from $\sum_{I_n} P^X(Y_i \geq \theta(\mathbf{x}) + c\delta_n) = \sum_{I_n} P(Y_i \geq \theta(\mathbf{x}) + c\delta_n \,|\, K_{ni})$ (since $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent) that $E(\sum_{I_n} Z_i \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) = 0$. Consequently, by Hoeffding's inequality (see Theorem 2 of Hoeffding, 1963)

$$P^X(N_n^{-1}\textstyle\sum_{I_n} Z_i \geq (c-M_0)k_0\delta_n) \leq \exp(-2N_n[(c-M_0)k_0\delta_n]^2). \tag{4.5}$$

According to Bernstein's inequality (see Theorem 3 of Hoeffding, 1963)

$$P(N_n < \tfrac{1}{2}np_n) \le \exp\left(-\frac{n(\frac{1}{2}p_n)^2}{2p_n + p_n}\right). \tag{4.6}$$

By (4.4)–(4.6),

$$\begin{aligned} P(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i \ge \theta(\mathbf{x}) + c\delta_n\}} \ge \tfrac{1}{2}) &\le P(N_n < \tfrac{1}{2}np_n) \\ &\quad + E\left[\exp(-2[(c - M_0)M_1^{-1}]^2 N_n\delta_n^2)1_{\{N_n \ge \frac{1}{2}np_n\}}\right] \\ &\le \exp(-\tfrac{1}{12}np_n) + \exp(-[(c - M_0)M_1^{-1}]^2 np_n\delta_n^2). \end{aligned} \tag{4.7}$$

By Assumption 2, $p_n \sim \delta_n^d$. The conclusion of Lemma 1 follows from (4.7).

The proof of (4.1) will now be given. Note that the event $\{\hat{\theta}_n(\mathbf{0}) \ge \theta(\mathbf{0}) + c\delta_n\}$ is contained in the event $\{N_n^{-1}\sum_{I_n} 1_{\{Y_i \ge \theta(\mathbf{x}) + c\delta_n\}} \ge \frac{1}{2}\}$. It follows from Lemma 1 and $\delta_n \sim n^{-r}$ or, equivalently, $n\delta_n^{d+2} \sim 1$ that

$$\begin{aligned} P(\hat{\theta}_n(\mathbf{0}) \ge \theta(\mathbf{0}) + c\delta_n) &\le P(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i \ge \theta(\mathbf{x}) + c\delta_n\}} \ge \tfrac{1}{2}) \\ &\le \exp(-c_2 n\delta_n^d) + \exp(-[(c - M_0)]^2 c_1 n\delta_n^{d+2}) = o(1) \end{aligned}$$

as $n, c \to \infty$. This completes the proof of Theorem 1.

Proof of Theorem 2. Without loss of generality it can be assumed that $C = [-\frac{1}{2}, \frac{1}{2}]^d$. Choose $s > 1$ and let $\{L_n\}$ denote a sequence of positive integers such that $L_n \sim n^s$. Let W_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 $|j| \le L_n$. Then C can be written as the union of $(2L_n)^d$ subcubes, each having length $2\lambda_n = (2L_n)^{-1}$ and all of its vertices in W_n. For each $\mathbf{x} \in C$ there is a subcube Q_w with center w such that $\mathbf{x} \in Q_w$. Let C_n denote the collection of the centers of these subcubes. Then

$$P\left(\sup_{\mathbf{x}\in C} |\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| \ge c(n^{-1}\log n)^r\right) = P\left(\max_{C_n} \sup_{\mathbf{x}\in Q_w} |\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| \ge c(n^{-1}\log n)^r\right).$$

It follows from $\lambda_n \sim n^s = o(c(n^{-1}\log n)^r)$ and Assumption 1 that $|\theta(\mathbf{x}) - \theta(w)| \le M_0\|\mathbf{x} - w\| \le M_0\delta_n$ for $\mathbf{x} \in Q_w$, $w \in C_n$ (for n sufficiently large). Therefore, to prove the theorem, it is sufficient to show that there is a positive constant c such that

$$\lim_n P\left(\max_{w\in C_n} \sup_{\mathbf{x}\in Q_w} |\hat{\theta}_n(\mathbf{x}) - \theta(w)| \ge c(n^{-1}\log n)^r\right) = 0. \tag{4.8}$$

To prove (4.8), let $\eta \equiv \sqrt{d}$, $\mathbf{x} \in Q_w$ and $N_n' \equiv N_n'(w) = \#\{i : \|\mathbf{X}_i - w\| \le \delta_n - \eta\lambda_n\}$. Now $N_n \equiv N_n(\mathbf{x}) = \#\{i : \|\mathbf{X}_i - \mathbf{x}\| \le \delta_n\} \ge N_n'$ for $\mathbf{x} \in Q_w$, hence $\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \ge c\delta_n\} \subseteq \{N_n^{-1}\sum_{I_n} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \frac{1}{2}\} \subseteq \{\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \frac{1}{2}N_n'\}$, where $I_n^* \equiv I_n^*(w) = \{i : 1 \le i \le n \text{ and } \|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n\}$. Thus

$$\cup_{Q_w}\left\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \ge c\delta_n\right\} \subseteq \left\{\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \tfrac{1}{2}N_n'\right\}. \tag{4.9}$$

Set $N_n^* = N_n^*(w) = \#I_n^*(w)$. By Assumption 2 and Theorem 12.2 of Breiman et al. (1984, p.334), there are positive constants d_1 and k_1 such that

$$\lim_n P(\Psi_n) = 1, \tag{4.10}$$

where $\Psi_n = \left\{N_n^*(w) - N_n'(w) \le d_1 \text{ and } N_n^*(w) \ge k_1 n\delta_n^d \text{ for all } w \in C_n\right\}$. (note that $N_n^* - N_n' = \#\{i : \delta_n - \eta\lambda_n \le \|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n\}$ is Binomial random variable with parameters n and p where $p \sim ((\delta_n + \eta\lambda_n)^d - (\delta_n - \eta\lambda_n)^d) \sim \lambda_n\delta_n^{d-1}$ for n sufficient large.

Note that $n^r N_n^{*-1} \le (k_1 n\delta_n^d)^{-1} \sim \delta_n^2/\log n$ on Ψ_n. According to (4.9) and (4.10), there is a squence of positive constants ϵ_n tending to zero such that

$$\begin{aligned}
&P\left(\max_{C_n}\sup_{x\in Q_w}|\hat{\theta}_n(\mathbf{x}) - \theta(w)| \ge c\delta_n\right)\\
&\le P\left(\cup_{C_n}\cup_{Q_w}\left\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \ge c\delta_n\right\}\right)\\
&\le P\left(\cup_{C_n}\left\{\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \tfrac{1}{2}N_n'\right\}\right)\\
&\le P\left(\cup_{C_n}\left\{\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \tfrac{1}{2}N_n^* - \tfrac{1}{2}d_1\right\}\cap\Psi_n\right) + P(\Psi_n^c)\\
&\le P\left(\cup_{C_n}\left\{N_n^{*-1}\textstyle\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \frac{1}{2} - \epsilon_n\delta_n\right\}\right) + P(\Psi_n^c).
\end{aligned} \tag{4.11}$$

According to Assumption 2, $P(\|\mathbf{X} - \mathbf{w}\| \le \delta_n + \eta\lambda_n) \sim \delta_n^d$ for $\mathbf{w} \in C_n$. Thus by Lemma 1, there are positive constants c, M_3 and M_4 such that

$$\begin{aligned}
&P\left(\cup_{C_n}\left\{N_n^{*-1}\textstyle\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \frac{1}{2} - \epsilon_n\delta_n\right\}\right)\\
&\le n^{sd}\max_{C_n} P\left(N_n^{*-1}\textstyle\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+c\delta_n\}} \ge \frac{1}{2} - \epsilon_n\delta_n\right)\\
&\le n^{sd}\exp(-c^2 M_3 n\delta_n^{d+2}) + n^{sd}\exp(-M_4 n\delta_n^d) = o(1),
\end{aligned} \tag{4.12}$$

for δ_n is chosen so that $\delta_n \sim (n^{-1}\log n)^r$ or, equivalently, $n\delta_n^{d+2} \sim \log n$. Hence, by (4.10)–(4.12) there is a positive constant c such that

$$\lim_n P(\max_{C_n} \sup_{\mathbf{x}\in Q_w} [\,\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})\,] \geq c(n^{-1}\log n)^r) = 0. \tag{4.13}$$

Similarly,

$$\lim_n P(\max_{C_n} \sup_{\mathbf{x}\in Q_w} [\,\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})\,] \leq -c(n^{-1}\log n)^r) = 0. \tag{4.14}$$

It follows from (4.13) and (4.14) that (4.8) is valid. This completes the proof of Theorem 2.

Proof of Theorem 3. By Assumption 3 and the argument presented in the proof of Lemma 1, there are positive constants c_1 and c_2 such that

$$\begin{aligned} &P\left(\cup_{C_n}\left\{N_n^{*-1}\textstyle\sum_{I_n^*} 1_{\{Y_i \geq \theta(w)+c\delta_n\}} \geq \frac{1}{2} - \epsilon_n\delta_n\right\}\right) \\ &\leq n^{sd}\exp(-c_1[\epsilon_0 \wedge c]^2 n\delta_n^d) + n^{sd}\exp(-c_2 n\delta_n^d) = o(1), \quad \text{for } c > 0 \end{aligned} \tag{4.15}$$

since δ_n is chosen so that $n\delta_n^d \sim \delta_n^{-2} \sim n^{2r}$. It follows from (4.9)–(4.11), (4.15) (with $c\delta_n$ replaced by c in (4.9), (4.11)) and the boundedness of $\theta(\cdot)$ on C that there is a positive constant $T \geq 1$ such that

$$\lim_n P(\Pi_n) = 1 \tag{4.16}$$

where $\Pi_n \equiv \{\|\hat{\theta}_n\|_\infty \leq T\}$. For $i = 1, \ldots, n$, set

$$Y_i' = \begin{cases} -T & \text{if } Y_i \leq -T; \\ Y_i & \text{if } |Y_i| \leq T; \\ T & \text{if } Y_i \geq T. \end{cases}$$

Set $\bar{\theta}_n(\mathbf{x}) \equiv \text{Median}\{Y_i' : i \in I_n(\mathbf{x})\}$. Note that $\bar{\theta}_n(\mathbf{x}) = \hat{\theta}_n(\mathbf{x})$ except on Π_n^c for $\mathbf{x} \in C$. Thus by (4.24), in order to prove the theorem, it is sufficient to show

$$\lim_n P(\|\bar{\theta}_n - \theta\|_q \geq cn^{-r}) = 0. \tag{4.17}$$

To prove (4.17), we may assume that C is contained in the interior of the cube $C_0 = [-\frac{1}{2}, \frac{1}{2}]^d \subset U$. Put $E^X(\cdot) = E(\cdot\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Then

$$E^X(\|\bar{\theta}_n - \theta\|_q^q) = \textstyle\int_C E^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q)\,d\mathbf{x},$$

and

$$E^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q) = \int_0^\infty qt^{q-1} P^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| > t)\, dt$$

$$= \int_0^{2M_0\delta_n} qt^{q-1} P^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| > t)\, dt$$

$$+ \int_{2M_0\delta_n}^T qt^{q-1} P^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| > t)\, dt$$

$$\leq (2M_0\delta_n)^q + \int_{2M_0\delta_n}^T qt^{q-1} P^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| > t)\, dt. \quad (4.18)$$

Recall that $N_n(\mathbf{x}) = \#(I_n(\mathbf{x}))$. By Assumption 2 and Theorem 12.2 of Breiman, et al. (1984, p.334), there is a positive constant k_1 such that

$$\lim_n P(\Omega_n) = 1, \quad (4.19)$$

where $\Omega_n = \{N_n(\mathbf{x}) \geq k_1 n\delta_n^d \quad \text{for all } \mathbf{x} \in C\}$. By Assumption 3, there is a positive constant k_2 such that

$$\int_{2M_0\delta_n}^T qt^{q-1} P^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| > t)\, dt \leq k_2 T^q [N_n(\mathbf{x})]^{-\frac{q}{2}} \quad \text{for } \mathbf{x} \in C \text{ and } T \geq 1. \quad (4.20)$$

(The proof of (4.20) will be given at the end of this section.) It follows from (4.18)–(4.20) that there is a positive constant k_3 such that

$$E^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q) \leq k_3[N_n(\mathbf{x})^{-q/2} + \delta_n^q] \qquad \text{for } \mathbf{x} \in C.$$

Thus there is a positive constant k_4 such that

$$E^X(\|\bar{\theta}_n - \theta\|_q^q) \leq k_4(n\delta_n^d)^{-q/2} \quad \text{on } \Omega_n. \quad (4.21)$$

Put $\mathrm{Var}^X(\cdot) = \mathrm{Var}(\cdot \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $\mathrm{Cov}^X(\cdot,\cdot) = \mathrm{Cov}((\cdot,\cdot) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n)$. Then

$$\mathrm{Var}^X(\|\bar{\theta}_n - \theta\|_q^q) = \mathrm{Var}^X(\textstyle\int_C |\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q\, d\mathbf{x})$$

$$= \iint_D \mathrm{Cov}^X\left\{|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q, |\bar{\theta}_n(\mathbf{x}') - \theta(\mathbf{x}')|^q\right\} d\mathbf{x}\, d\mathbf{x}'$$

$$\leq \iint_D \left\{E^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^{2q}) E^X(|\bar{\theta}_n(\mathbf{x}') - \theta(\mathbf{x}')|^{2q})\right\}^{\frac{1}{2}} d\mathbf{x}\, d\mathbf{x}',$$

where $D = \{\mathbf{x}, \mathbf{x}' \in C : \|\mathbf{x} - \mathbf{x}'\| \leq 2\delta_n\}$. It now follows from (4.21) that there is a positive constant k_5 and a sequence of positive numbers κ_n tending to 0 such that

$$\mathrm{Var}^X(\|\bar{\theta}_n - \theta\|_q^q) \leq k_5 \kappa_n (n\delta_n^d)^{-q} \qquad \text{on } \Omega_n. \tag{4.22}$$

Note that $\delta_n \sim n^{-r}$. By (4.21) and (4.22), there is a sequence of positive numbers λ_n tending to zero such that

$$P^X(\|\bar{\theta}_n - \theta\|_q^q \geq (cn^{-r})^q) \leq \frac{\mathrm{Var}^X(\|\bar{\theta}_n - \theta\|_q^q)}{[(c - k_4)n^{-r}]^{2q}} \leq \lambda_n \qquad \text{on } \Omega_n. \tag{4.23}$$

It follows from (4.19) and (4.23) that there is a positive constant c such that (4.16) holds, as desired.

Proof of (4.20). Suppose that $\mathbf{x} \in C$, $\|\mathbf{X} - \mathbf{x}\| \leq \delta_n$ and $T \geq 1$. By Assumption 3 (for n sufficiently large)

$$\int_{\theta(\mathbf{x})+M_0\delta_n}^{\theta(\mathbf{x})+t} h(y \mid \mathbf{X}, \theta(\mathbf{X}))\, dy \geq M_1^{-1}(t - M_0\delta_n)$$

$$\geq M_1^{-1}T^{-1}(t - M_0\delta_n) \quad \text{for } M_0\delta_n \leq t \leq \epsilon_0$$

and

$$\int_{\theta(\mathbf{x})+M_0\delta_n}^{\theta(\mathbf{x})+t} h(y \mid \mathbf{X}, \theta(\mathbf{X}))\, dy \geq M_1^{-1}(\epsilon_0 - M_0\delta_n)T^{-1}(t - M_0\delta_n) \quad \text{for } \epsilon_0 \leq t \leq T.$$

Thus there is a positive constant c_0 such that for $\mathbf{x} \in C$, $\|\mathbf{X} - \mathbf{x}\| \leq \delta_n$, $T \geq 1$ and $t \in [M_0, T]$,

$$\int_{\theta(\mathbf{x})+M_0\delta_n}^{\theta(\mathbf{x})+t} h(y \mid \mathbf{X}, \theta(\mathbf{X}))\, dy \geq c_0 T^{-1}(t - M_0\delta_n). \tag{4.24}$$

Let $K_{ni} = K_{ni}(\mathbf{x})$ be the event that $\|\mathbf{X}_i - \mathbf{x}\| \leq \delta_n$. Set $Z_i = 1_{\{Y_i > \theta(\mathbf{x})+t\}} - P(Y_i > \theta(\mathbf{x}) + t \mid K_{ni})$ and $N_n = N_n(\mathbf{x})$. Since $\{Y_i' > \theta(\mathbf{x}) + t\} \subset \{Y_i > \theta(\mathbf{x}) + t\}$, it follows from (4.24) that

$$P^X(\bar{\theta}_n(\mathbf{x}) > \theta(\mathbf{x}) + t) \leq P^X(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i' > \theta(\mathbf{x})+t\}} \geq \frac{1}{2})$$

$$\leq P^X(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i > \theta(\mathbf{x})+t\}} \geq \frac{1}{2})$$

$$\leq P^X\left(N_n^{-1}\textstyle\sum_{I_n} Z_i \geq N_n^{-1}\sum_{I_n} \int_{\theta(\mathbf{x})+M_0\delta_n}^{\theta(\mathbf{x})+t} h(y \mid \mathbf{X}_i, \theta(\mathbf{X}_i))\, dy \right)$$

$$\leq P^X(N_n^{-1}\textstyle\sum_{I_n} Z_i \geq c_0 T^{-1}(t - M_0\delta_n)).$$

By Hoeffding's inequality (see the proof of Lemma 1), there is a positive constant k_6 such that for $M\delta_n \leq t \leq T$ and $\mathbf{x} \in C$

$$P^X(\bar{\theta}_n(\mathbf{x}) > \theta(\mathbf{x}) + t) \leq \exp[-k_6 N_n T^{-2}(t - M_0\delta_n)^2] \tag{4.25}$$

and, similarly,

$$P^X(\bar{\theta}_n(\mathbf{x}) < \theta(\mathbf{x}) - t) \leq \exp[-k_6 N_n T^{-2}(t - M_0\delta_n)^2] \tag{4.26}$$

It follows from (4.25) and (4.26) that there is a positive constants k_7 such that for $\mathbf{x} \in C$,

$$\int_{2M_0\delta_n}^{T} t^{q-1} P^X(|\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| > t)\, dt \leq \int_{2M_0\delta_n}^{T} t^{q-1} \exp[-k_6 N_n T^{-2}(t - M_0\delta_n)^2]\, dt$$

$$\leq 2^{q-1} \int_0^{\infty} t^{q-1} \exp[-k_6 N_n T^{-2}(t - M_0\delta_n)^2]\, dt$$

$$= k_7 T^q N_n^{-\frac{q}{2}},$$

as desired.

Acknowledgment. I would like to express my deepest gratitude to my thesis advisor, Chuck Stone, for suggesting this problem and for his generous guidance. Also, I would like to thank my wife, Psyche Lee, for her patience and inspiration.

REFERENCES

BLOOMFIELD, P. and STEIGER, (1983) *Least Absolute Deviations.* Birkhäuser, Boston.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984) *Classification and Regression Trees.* Wadsworth, Belmont.

DEVROYE, L. P. and WAGNER, T. J. (1980a) On the L_1 convergence of kernel estimators of regression function with application in discrimination. *Z. Wahrsch. verw. Gebiete* **51**, 15–25.

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

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

HÄRDLE, W. and LUCKHAUS, S. (1984) Uniform consistency of a class of regression function estimators. *Ann. Statist.* **12**, 612–623.

HOEFFDING, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

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

STONE, C. J. (1977) Consistent nonparametric regression. *Ann. Statist.* **5**, 595–645.

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

STONE, C. J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

STONE, C. J. (1985) Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689–705.