*The Annals of Statistics*
1977, Vol. 5, No. 4, 595–645

# CONSISTENT NONPARAMETRIC REGRESSION[1]

BY CHARLES J. STONE

*University of California, Los Angeles*

Let $(X, Y)$ be a pair of random variables such that $X$ is $\mathbb{R}^d$-valued and $Y$ is $\mathbb{R}^{d'}$-valued. Given a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ from the distribution of $(X, Y)$, the conditional distribution $P^Y(\cdot \mid X)$ of $Y$ given $X$ can be estimated nonparametrically by $\hat{P}_n^Y(A \mid X) = \sum_1^n W_{ni}(X) I_A(Y_i)$, where the weight function $W_n$ is of the form $W_{ni}(X) = W_{ni}(X, X_1, \cdots, X_n)$, $1 \leq i \leq n$. The weight function $W_n$ is called a probability weight function if it is nonnegative and $\sum_1^n W_{ni}(X) = 1$. Associated with $\hat{P}_n^Y(\cdot \mid X)$ in a natural way are nonparametric estimators of conditional expectations, variances, covariances, standard deviations, correlations and quantiles and nonparametric approximate Bayes rules in prediction and multiple classification problems. Consistency of a sequence $\{W_n\}$ of weight functions is defined and sufficient conditions for consistency are obtained. When applied to sequences of probability weight functions, these conditions are both necessary and sufficient. Consistent sequences of probability weight functions defined in terms of nearest neighbors are constructed. The results are applied to verify the consistency of the estimators of the various quantities discussed above and the consistency in Bayes risk of the approximate Bayes rules.

**1. Introduction.** Let $(X, Y)$ be a pair of random variables such that $X$ is $\mathbb{R}^d$-valued and $Y$ is $\mathbb{R}^{d'}$-valued. An important concept in probability and statistics is that of the conditional distribution $P^Y(\cdot \mid X)$ of $Y$ given $X$ and quantities defined in terms of this conditional distribution—conditional expectations, variances, standard deviations, covariances, correlations and quantiles.

There are simple formulas for these conditional quantities if the joint distribution $P^{X,Y}$ of $(X, Y)$ is a multivariate Gaussian distribution $\mathscr{N}(\mu, \Sigma)$ with known mean $\mu$ and covariance matrix $\Sigma$. Typically in practice $P^{X,Y}$ is not known exactly but a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ from $P^{X,Y}$ is available. In the Gaussian case estimators $\hat{\mu}$ and $\hat{\Sigma}$ of $\mu$ and $\Sigma$ based on this data can be obtained and $P^{X,Y}$ can be estimated as $\hat{P}_n^{X,Y} = \mathscr{N}(\hat{\mu}, \hat{\Sigma})$. Then $P^Y(\cdot \mid X)$ can be estimated by $\hat{P}_n^Y(\cdot \mid X)$, defined to be the conditional distribution of $Y$ given $X$ corresponding to the joint distribution $\hat{P}_n^{X,Y}$. The various conditional quantities defined in terms of $P^Y(\cdot \mid X)$ can in turn be estimated by the corresponding quantities defined in terms of $\hat{P}_n^Y(\cdot \mid X)$.

This paper is concerned with the problem of estimating the conditional

Received October 1974; revised September 1976.

[1] Research was supported by NSF Grant No. MPS 72-04591 and, through the Health Sciences Computer Facility at UCLA, by NIH special research grant RR-3.

*AMS* 1970 *subject classifications.* Primary 62G05; Secondary 62H30.

*Key words and phrases.* Regression function, conditional quantities, prediction, multiple classification, consistency in Bayes risk, approximate Bayes rules, nonparametric estimators, nearest neighbor rules.

distribution of $Y$ given $X$ and the various conditional quantities related to it when the joint distribution of $X$ and $Y$ is not assumed to be Gaussian or, in fact, to belong to any prespecified parametric family of distributions. In this context nonparametric methods of estimation are appropriate.

If a number of the $X_i$'s in the random sample are exactly equal to $X$, which can happen if $X$ is a discrete random variable, $P^Y(\cdot \mid X)$ can be estimated by the empirical distribution of the $Y_i$'s corresponding to $X_i$'s equal to $X$. If few or none of the $X_i$'s are exactly equal to $X$, it is necessary to use $Y_i$'s corresponding to $X_i$'s near $X$. This leads to estimators $\hat{P}_n^Y(\cdot \mid X)$ of the form

$$\hat{P}_n^Y(A \mid X) = \sum_{i=1}^n W_{ni}(X) I_A(Y_i) ,$$

where $W_{ni}(X) = W_{ni}(X, X_1, \cdots, X_n)$, $1 \leqq i \leqq n$, weights those values of $i$ for which $X_i$ is close to $X$ more heavily than those values of $i$ for which $X_i$ is far from $X$. Set $W_{ni}(X) = 0$ for $i > n$. The *weight function* $W_n$ is said to be *normal* if $\sum_i W_{ni}(X) = 1$, *nonnegative* if $W_n \geqq 0$, and a *probability weight function* if it is both normal and nonnegative. In the last case $\hat{P}_n^Y(\cdot \mid X)$ is a probability distribution on $\mathbb{R}^{d'}$.

Let $g$ be a Borel function on $\mathbb{R}^{d'}$ such that $E|g(Y)| < \infty$ and let $E(g(Y) \mid X)$ denote the conditional expectation of $g(Y)$ given $X$. Corresponding to $W_n$ is the estimator $\hat{E}_n(g(Y) \mid X)$ of $E(g(Y) \mid X)$ defined by

$$\hat{E}_n(g(Y) \mid X) = \int g(y) \hat{P}_n^Y(dy \mid X) = \sum_i W_{ni}(X) g(Y_i) .$$

Note that if $A$ is a Borel set in $\mathbb{R}^{d'}$, then

$$\hat{P}_n^Y(A \mid X) = \hat{E}_n(I_A(Y) \mid X) .$$

Other conditional quantities defined in terms of $P^Y(\cdot \mid X)$ can again be estimated by the corresponding quantities defined in terms of $\hat{P}_n^Y(\cdot \mid X)$.

Observe that the estimators considered here are estimators of function values at specified points of the domain, not estimators of parameters of the function. Suppose, for example, that $Y$ is real valued and $E|Y| < \infty$. The value $E(Y \mid X = x) = \int y P^Y(dy \mid X = x)$ of the regression function of $Y$ on $X$ at the point $x$ is estimated by

$$\hat{E}_n(Y \mid X = x) = \int y \hat{P}_n^Y(dy \mid X = x) = \sum_i W_{ni}(x) Y_i .$$

This setup differs from that of nonparametric linear regression models. There the regression function is assumed to belong to the parametric family of linear functions on $\mathbb{R}^d$, but no parametric form is assumed for the distribution of the residuals. For such models the goal is to obtain robust estimators of the regression coefficients (see Adichie (1967), Jurečková (1971), Jaeckel (1972) and Bickel (1973)).

Let $X, X_1, X_2, \cdots$ be a fixed sequence of independent and identically distributed (i.i.d.) $\mathbb{R}^d$-valued random variables on a probability space $\Omega$. It is assumed that there is a sequence of independent standard normal random variables on $\Omega$ which is independent of $(X, X_1, X_2, \cdots)$ (which fact is required to obtain the necessity of (5) in Theorem 1 below). A sequence $\{W_n\}$ of weights is said to be *consistent* if whenever $(X, Y), (X_1, Y_1), (X_2, Y_2), \cdots$ are i.i.d., $Y$ is real valued, $r > 1$, and $E|Y|^r < \infty$; then $\hat{E}_n(Y \mid X) \to E(Y \mid X)$ in $L^r$ ($Z_n \to Z$ in $L^r$ means that $E|Z_n - Z|^r \to 0$).

In Theorem 1 of Section 2 sufficient conditions for $\{W_n\}$ to be consistent are stated. If $\{W_n\}$ is a sequence of probability weights, then, as noted in Corollary 1, the conditions simplify and becomes both necessary and sufficient for consistency.

The conditions in Theorem 1 and Corollary 1 involve the unknown underlying distribution of $X$. A sequence $\{W_n\}$ of weights is said to be *universally consistent* if it is consistent regardless of the distribution of $X$. Theorem 2 of Section 3 shows how to obtain universally consistent sequences of weights defined in terms of the ranks of the distances from $X_1, \cdots, X_n$ to $X$. The proof of Theorem 2 depends crucially on an inequality stated as Proposition 11 in Section 11. This inequality, which is interesting in itself, is the key to the truly nonparametric (distribution free) aspect of this paper—i.e., to the fact that results are obtained which are completely free of regularity conditions on the distribution of $X$ or the joint distribution of $(X, Y)$.

In Section 4 a method is discussed for modifying a consistent sequence of weights to obtain another consistent sequence which hopefully yields more accurate estimators. Section 5 discusses "trend removal," which can take advantage of a fairly accurate linear approximation to $E(Y \mid X)$. By definition a consistent sequence of weights yields consistent estimators of conditional expectations. Sections 6, 7 and 8 show respectively how to obtain consistent estimators of conditional second order quantities, conditional quantiles, and Bayes rules. The results in Sections 4–8 are all based on starting out with a consistent sequence of weights. They become truly nonparametric if the weights are assumed to be universally consistent. Related papers in the literature are briefly reviewed in Section 9. The results from Sections 2–8 are proved in Sections 10–13.

An experimental packaged program is currently being developed in cooperation with the Health Sciences Computer Facility at UCLA, which should make it easy to determine the performance of the estimators discussed in this paper on real and simulated data sets. Preliminary experience in using this program on simulated data sets shows the effectiveness of the modifications discussed in Sections 4 and 5.

**2. Consistent sequences of weights.** Let $\mathbb{R}^d$ denote $d$-dimensional Euclidean space with the usual inner product $x \cdot y$ and norm $||x||$. For $x$ and $y$ in $\mathbb{R}$ let $x \vee y$ and $x \wedge y$ denote respectively the maximum and minimum of $x$ and $y$. For $x \in \mathbb{R}$ set $x^+ = x \vee 0$, $x^- = -(x \wedge 0)$ and $\operatorname{sign}(x) = -1, 0$, or $1$ according as $x < 0$, $x = 0$, or $x > 0$. Given any set $A$, let $\#(A)$ denote the number of elements in $A$.

All random variables considered in this paper are assumed to be defined on the probability space $\Omega$. Let $Z_n$, $n \geqq 1$, and $Z$ be real valued random variables. Then $Z_n \to Z$ in probability if $\lim_n P(|Z_n - Z| > \varepsilon) = 0$ for all $\varepsilon > 0$. For $r \geqq 1$, $Z_n \to Z$ in $L^r$ if $\lim_n E|Z_n - Z|^r = 0$. Note that $Z_n \to Z$ in $L^r$ implies that $Z_n \to Z$ in probability. Finally $Z_n$ is bounded in probability if $\lim_{M\to\infty} \limsup_n P(|Z_n| \geqq M) = 0$.

The following result will be proven in Section 10.

THEOREM 1. *Let $\{W_n\}$ be a sequence of weights. Suppose the following five conditions are satisfied: there is a $C \geqq 1$ such that for every nonnegative Borel function $f$ on $\mathbb{R}^d$*

$$E \textstyle\sum_i |W_{ni}(X)| f(X_i) \leqq CEf(X) \qquad \text{for all} \quad n \geqq 1\,; \tag{1}$$

*there is a $D \geqq 1$ such that*

$$P(\textstyle\sum_i |W_{ni}(X)| \leqq D) = 1 \qquad \text{for all} \quad n \geqq 1\,; \tag{2}$$

$$\textstyle\sum_i |W_{ni}(X)| I_{\{\|X_i - X\| > a\}} \to 0 \qquad \text{in probability for all} \quad a > 0\,; \tag{3}$$

$$\textstyle\sum_i W_{ni}(X) \to 1 \quad \text{in probability; and} \tag{4}$$

$$\max_i |W_{ni}(X)| \to 0 \quad \text{in probability.} \tag{5}$$

*Then $\{W_n\}$ is consistent.*

*Suppose, conversely, that $\{W_n\}$ is consistent. Then* (4) *and* (5) *hold. If $W_n \geqq 0$ for all $n \geqq 1$, then* (3) *holds, and if $W_n \geqq 0$ for all $n \geqq 1$ and* (2) *holds, then* (1) *holds.*

If $\{W_n\}$ is a sequence of probability weights, then (2) and (4) hold automatically and the three remaining conditions are necessary and sufficient for consistency. This result is summarized in the following corollary.

COROLLARY 1. *Let $\{W_n\}$ be a sequence of probability weights. It is consistent if and only if the following three conditions hold: there is a $C \geqq 1$ such that, for every nonnegative Borel function $f$ on $\mathbb{R}^d$, $E \sum_i W_{ni}(X) f(X_i) \leqq CEf(X)$ for all $n \geqq 1$; $\sum_i W_{ni}(X) I_{\{\|X_i - X\| > a\}} \to 0$ in probability for all $a > 0$; and* $\max_i W_{ni}(X) \to 0$ *in probability.*

The following consequence of Theorem 1 will be used in Section 4.

COROLLARY 2. *Let $\{U_n\}$ be a consistent sequence of probability weights, let $\{W_n\}$ be a sequence of normal weights, and suppose that there is an $M \geqq 1$ such that $|W_n| \leqq MU_n$ for all $n \geqq 1$. Then $\{W_n\}$ is consistent.*

**3. Nearest neighbor weights.** In this section consistent sequences of probability weights will be constructed. The weights will depend on the distances from $X$ to $X_1, \cdots, X_n$ in terms of a suitable metric on $\mathbb{R}^d$.

The obvious metric on $\mathbb{R}^d$ to use is the Euclidean metric. This metric may well be appropriate if the various coordinates of $X$ are measured in the same units, but it is most likely inappropriate otherwise.

When the individual coordinates are measured in dissimilar units, e.g., grams, centimeters, and seconds, it makes sense to transform them to be unit free before applying the Euclidean metric. Let $s_n$ be a *scale* based on $X_1, \cdots, X_n$, that is a nonnegative function of the form $s_{nj} = s_{nj}(X, X_1, \cdots, X_n)$, $1 \leqq j \leqq d$. The random (pseudo) metric $\rho_n$ corresponding to this scale is defined by

$$\rho_n(u, v) = \left(\sum_j \left(\frac{u_j - v_j}{s_{nj}}\right)^2\right)^{\frac{1}{2}}, \tag{6}$$

where $u = (u_1, \cdots, u_d)$, $v = (v_1, \cdots, v_d)$, and the sum extends over all $j$, $1 \leqq j \leqq d$, such that $s_{nj} > 0$.

Let $\{s_n\}$ be a sequence of scales and let $\{\rho_n\}$ be the corresponding sequence of metrics determined by (6). In order to obtain a consistent sequence of weights, a number of assumptions need to be imposed on $\{s_n\}$. First, it is assumed that if $1 \leqq j \leqq d$ and the $j$th coordinate of $X$ has a nondegenerate distribution, then $\lim_n P(s_{nj} > 0) = 1$. Secondly, it is assumed that if $1 \leqq j, l \leqq d$ and the $j$th and $l$th coordinates of $X$ both have nondegenerate distributions, then $s_{nj}/s_{nl}$ is bounded in probability. Finally it is assumed that there are positive constants $a$ and $b \geqq a$ independent of $n$ such that whenever $n \geqq 1$, $1 \leqq i \leqq n$, $1 \leqq j \leqq d$, and the $j$th coordinates of $X_1, \cdots, X_n$ do not coincide, then

$$\begin{aligned} as_{nj}(X_i, X_1, \cdots, X, \cdots, X_n) &\leqq s_{nj}(X, X_1, \cdots, X_n) \\ &\leqq bs_{nj}(X_i, X_1, \cdots, X, \cdots, X_n)\,. \end{aligned} \tag{7}$$

Here $(X_i, X_1, \cdots, X, \cdots, X_n)$ denotes the sequence $(X, X_1, \cdots, X_n)$ with $X$ and $X_i$ interchanged. The last condition is obviously satisfied with $a = b = 1$ if $s_{nj}(X, X_1, \cdots, X_n)$ is a symmetric function of $X, X_1, \cdots, X_n$. The condition allows for a certain amount of asymmetry. If $\{s_n\}$ satisfies these assumptions it is said to be *regular*.

If $s_n \equiv 1$ for all $n \geqq 1$, then $\{s_n\}$ is obviously regular. If $E||X||^2 < \infty$ and $s_{nj}$ is the sample standard deviation of the $j$th coordinate of $X, X_1, \cdots, X_n$, then $\{s_n\}$ is regular. From now on it is assumed that $\{s_j\}$ is a regular sequence of scales and that $\{\rho_n\}$ is the corresponding sequence of metrics.

For $1 \leqq k \leqq n$ let $I_{nk}(X)$ denote the collection of all indices $i$, $1 \leqq i \leqq n$, such that fewer than $k$ of the points $X_1, \cdots, X_n$ are strictly closer to $X$ in the metric $\rho_n$ than is $X_i$. Suppose, for example, that $n = 4$ and that

$$\rho_4(X_3, X) < \rho_4(X_2, X) = \rho_4(X_4, X) < \rho_4(X_1, X)\,.$$

Then $I_{41}(X) = \{3\}$, $I_{42}(X) = I_{43}(X) = \{2, 3, 4\}$ and $I_{44}(X) = \{1, 2, 3, 4\}$. Clearly $\#(I_{nk}(X)) \geqq k$, and $\#(I_{nk}(X)) = k$ for $1 \leqq k \leqq n$ if and only if the $n$ numbers $\rho_n(X_1, X), \cdots, \rho_n(X_n, X)$ are distinct. The points $i$ in $I_{nk}(X)$ are called the $k$ *nearest neighbors* of $X$. If $W_n$ is a weight function such that $W_{ni}(X) = 0$ for $i \notin I_{nk}(X)$, it is called a $k$ *nearest neighbor* ($k$-NN) *weight function*.

Let $c_{ni}$, $i \geqq 1$, be such that $c_{n1} \geqq \cdots \geqq c_{nn} \geqq 0$, $c_{ni} = 0$ for $i > n$, and $c_{n1} + \cdots + c_{nn} = 1$. Associated with $c_n$ is the probability weight function $W_n$

defined as follows: for $1 \leqq i \leqq n$

$$W_{ni}(X) = \frac{c_{n\nu} + \cdots + c_{n,\nu+\lambda-1}}{\lambda} \tag{8}$$

where

$$\nu = 1 + \#(\{l: 1 \leqq l \leqq n, l \neq i, \text{ and } \rho_n(X_l, X) < \rho_n(X_i, X)\})$$

and

$$\lambda = 1 + \#(\{l: 1 \leqq l \leqq n, l \neq i, \text{ and } \rho_n(X_l, X) = \rho_n(X_i, X)\}) .$$

In particular, if $X_i$ is the unique $\nu$th closest point among $X_1, \cdots, X_n$ to $X$ in the metric $\rho_n$, then $\lambda = 1$ and hence $W_{ni}(X) = c_{n\nu}$. Since $W_n \geqq 0$ and $\sum_i W_{ni}(X) = \sum_i c_{ni} = 1$, $W_n$ is indeed a probability weight function.

EXAMPLE 1 (uniform $k$-NN weight function). $c_{ni} = 1/k$ for $1 \leqq i \leqq k$ and $c_{ni} = 0$ for $i > k$.

EXAMPLE 2 (triangular $k$-NN weight function). $c_{ni} = (k - i + 1)/b_k$ for $1 \leqq i \leqq k$ and $c_{ni} = 0$ for $i > k$. Here $b_k = k(k + 1)/2$.

EXAMPLE 3 (quadratic $k$-NN weight function). $c_{ni} = (k^2 - (i - 1)^2)/b_k$ for $1 \leqq i \leqq k$ and $c_{ni} = 0$ for $i > k$. Here $b_k = k(k + 1)(4k - 1)/6$.

One expects $\hat{E}_n(g(Y) \mid X)$ to be a smoother function of $X$ for triangular and quadratic $k$-NN weight functions than for uniform $k$-NN weight functions.

The next result will be proven in Section 11.

THEOREM 2. *For $n \geqq 1$ let $W_n$ be the probability weight function corresponding to $c_n$. If $\lim_n \sum_{i>\alpha n} c_{ni} = 0$ for all $\alpha > 0$ and $\lim_n c_{n1} = 0$, then $\{W_n\}$ is consistent.*

COROLLARY 3. *For $n \geqq 1$, let $W_n$ be the uniform, triangular, or quadratic $k_n$-NN probability weight function. If $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, then $\{W_n\}$ is consistent.*

**4. Local linear weights.** Assume through Section 9 that $(X, Y), (X_1, Y_1), (X_2, Y_2), \cdots$ is an i.i.d. sequence, where $X$ is $\mathbb{R}^d$-valued and $Y$ is $\mathbb{R}^{d'}$-valued. In this section it is assumed that $d' = 1$, so that $Y$ is real valued.

Let $U_n$ be a probability weight function. A related weight function, corresponding to a different method for estimating $E(Y \mid X)$, will now be constructed.

Choose $\hat{a}_n \in \mathbb{R}$ and $\hat{b}_n \in \mathbb{R}^d$ to be values of $a$ and $b$ which minimize

$$\textstyle\sum_i U_{ni}(X)(Y_i - a - b \cdot X_i)^2$$

and set $\hat{E}_n(Y \mid X) = \hat{a}_n + \hat{b}_n \cdot X$, where $\cdot$ denotes the usual inner product on $\mathbb{R}^d$. This local linear regression estimator, in effect, uses weighted least squares, with the $i$th case having weight $U_{ni}(X)$, to fit a linear regression function to the data and then evaluates this function at $X$. It can be written in the form

$$\hat{E}_n(Y \mid X) = \textstyle\sum_i V_{ni}(X) Y_i ,$$

where

$$V_{ni}(X) = U_{ni}(X)(1 + (X - \bar{X}) \cdot C^{-1}(X)(X_i - \bar{X})) .$$

Here

$$\bar{X} = \sum_i U_{ni}(X)X_i ,$$
$$C_{lm}(X) = \sum_i U_{ni}(X)(X_{il} - \bar{X}_l)(X_{im} - \bar{X}_m) , \qquad 1 \leqq l, m \leqq d ,$$

and, for simplicity, the matrix $(C_{lm}(X))$ is assumed to be nonsingular with probability one (in implementing this procedure, a "tolerance" is used to avoid pivoting on small elements). The weight function $V_n$ is called the (untrimmed) *local linear weight function* corresponding to $U_n$. It is normal but generally not a probability weight function.

Let $\{U_n\}$ be a consistent sequence of probability weights. The corresponding sequence $\{V_n\}$ of local linear weights is not necessarily consistent. It will now be shown how to trim $V_n$ to obtain consistency.

Choose $A \leqq 1$ and $B \geqq 1$ and set $W_n^{(1)} = (V_n \vee AU_n) \wedge BU_n$. Then $AU_n \leqq W_n^{(1)} \leqq BU_n$. Now $W_n^{(1)}$ is not necessarily normal. To guarantee normality one more trimming is necessary: if $\sum_i W_{ni}^{(1)}(X) < 1$, set

$$W_{ni}(X) = W_{ni}^{(1)}(X) \vee (A_n(X)U_{ni}(X)) \qquad \text{for} \quad i \geqq 1 ,$$

where $A_n(X) \in (A, 1]$ is chosen so that $\sum_i W_{ni}(X) = 1$; if $\sum_i W_{ni}^{(1)}(X) > 1$, set

$$W_{ni}(X) = W_{ni}^{(1)}(X) \wedge (B_n(X)U_{ni}(X)) \qquad \text{for} \quad i \geqq 1 ,$$

where $B_n(X) \in [1, B)$ is chosen so that $\sum_i W_{ni}(X) = 1$; and if $\sum_i W_{ni}^{(1)}(X) = 1$, set $W_{ni}(X) = W_{ni}^{(1)}(X)$ for $i \geqq 1$. The weight function $W_n$ so defined is called the *trimmed local linear weight function* corresponding to $U_n$ and the parameters $A$ and $B$. By construction, $W_n$ is normal and $AU_n \leqq W_n \leqq BU_n$. If $A \geqq 0$, then $W_n$ is a probability weight function. If $U_n$ is a $k_n$-NN weight function, then so is $W_n$.

The following result follows immediately from Corollary 2.

COROLLARY 4. *Let $\{U_n\}$ be a consistent sequence of probability weights, let $A \leqq 1 \leqq B$ and, for $n \geqq 1$, let $W_n$ be the trimmed local linear weight function corresponding to $U_n$ and the parameters $A$ and $B$. Then $\{W_n\}$ is consistent.*

**5. Trend removal.** In this section it is assumed that $Y$ is real valued. The untrimmed local linear weight function defined in the previous section yields an estimator of $E(Y \mid X)$ which, in effect, extrapolates a local linear trend in each direction out to infinity (this is most easily seen when $d = 1$). It may be more reliable to extrapolate a global linear trend out to infinity. This can be done by first removing the global linear trend, then applying a suitable estimator $\hat{E}_n(\cdot \mid X)$ to the residuals, and finally adding back the global linear trend.

Specifically suppose that $E||X||^2 < \infty$, $EY^2 < \infty$, and that the covariance matrix of $X$ is nonsingular. Let $a_0 \in \mathbb{R}$ and $b_0 \in \mathbb{R}^d$ be the values of $a$ and $b$ which minimize $E(Y - a - b \cdot X)^2$. Let $\hat{a}_n \in \mathbb{R}$ and $\hat{b}_n \in \mathbb{R}^d$ be the values of $a$ and $b$ which minimize

$$\sum_{i=1}^n (Y_i - a - b \cdot X_i)^2 .$$

It follows easily from the normal equations corresponding to this minimization problem that $\hat{a}_n \to a_0$ and $\hat{b}_n \to b_0$ in probability and hence that

$$\hat{a}_n + \hat{b}_n \cdot X \to a_0 + b_0 \cdot X \quad \text{in probability.}$$

Let $\{W_n\}$ be a consistent sequence of weights. The estimator $\tilde{E}_n(Y \mid X)$ corresponding to $W_n$ obtained by *trend removal* is given by

$$\tilde{E}_n(Y \mid X) = \hat{a}_n + \hat{b}_n \cdot X + \sum_i W_{ni}(X)(Y_i - \hat{a}_n - \hat{b}_n \cdot X_i) .$$

Since $\sum_i W_{ni}(X) \to 1$ in probability and each coordinate of $\sum_i W_{ni}(X)X_i$ converges in $L^2$ to the corresponding coordinate of $X$, it follows that

$$\sum_i W_{ni}(X)(\hat{a}_n + \hat{b}_n \cdot X_i) \to a_0 + b_0 \cdot X \quad \text{in probability.}$$

It also follows from the consistency of $\{W_n\}$ that

$$\sum_i W_{ni}(X) Y_i \to E(Y \mid X) \quad \text{in probability.}$$

By the above four displayed results

$$\tilde{E}_n(Y \mid X) \to E(Y \mid X) \quad \text{in probability.}$$

Thus trend removal results in estimators of $E(Y \mid X)$ which are consistent in probability. It is not such an easy matter to determine when these estimators are consistent in $L^2$ or even to determine when the usual linear regression estimator $\hat{a}_n + \hat{b}_n \cdot X$ converges to $a_0 + b_0 \cdot X$ in $L^2$.

**6. Estimation of conditional second order quantities.** Let $g$ and $h$ be Borel functions on $\mathbb{R}^{d'}$ such that $Eg^2(Y) < \infty$ and $Eh^2(Y) < \infty$. For example, $g(Y)$ and $h(Y)$ could be two of the $d'$ coordinates of $Y$. The conditional covariance $\operatorname{Cov}(g(Y), h(Y) \mid X)$ of $g(Y)$ and $h(Y)$ given $X$ is defined as

$$\operatorname{Cov}(g(Y), h(Y) \mid X) = E(g(Y)h(Y) \mid X) - E(g(Y) \mid X)E(h(Y) \mid X) .$$

The conditional variance $\operatorname{Var}(g(Y) \mid X)$ of $g(Y)$ given $X$ is defined as

$$\operatorname{Var}(g(Y) \mid X) = \operatorname{Cov}(g(Y), g(Y) \mid X) .$$

The conditional standard deviation $\operatorname{Std}(g(Y) \mid X)$ of $g(Y)$ given $X$ is defined as

$$\operatorname{Std}(g(Y) \mid X) = (\operatorname{Var}(g(Y) \mid X))^{\frac{1}{2}} .$$

The conditional correlation $\operatorname{Cor}(g(Y)h(Y) \mid X)$ of $g(Y)$ and $h(Y)$ given $X$ is defined as

$$\operatorname{Cor}(g(Y), h(Y) \mid X) = \frac{\operatorname{Cov}(g(Y), h(Y) \mid X)}{\operatorname{Std}(g(Y) \mid X) \operatorname{Std}(h(Y) \mid X)} ,$$

if the denominator of the right-hand side is positive and by $\operatorname{Cor}(g(Y), h(Y) \mid X) = 0$ otherwise.

Let $W_n$ be a weight function and let $\hat{E}_n(\cdot \mid X)$ be the corresponding estimator of $E(\cdot \mid X)$. The above conditional second order quantities can be estimated as follows:

$$\widehat{\mathrm{Cov}}_n\,(g(Y), h(Y)\,|\,X) = \hat{E}_n(g(Y)h(Y)\,|\,X) - \hat{E}_n(g(Y)\,|\,X)\hat{E}_n(h(Y)\,|\,X)\,;$$
$$\widehat{\mathrm{Var}}_n\,(g(Y)\,|\,X) = \widehat{\mathrm{Cov}}_n^{+}\,(g(Y), g(Y)\,|\,X)\,;$$
$$\widehat{\mathrm{Std}}_n\,(g(Y)\,|\,X) = (\widehat{\mathrm{Var}}_n\,(g(Y)\,|\,X))^{\frac{1}{2}}\,;$$
$$\widehat{\mathrm{Cor}}_n\,(g(Y), h(Y)\,|\,X) = \frac{\widehat{\mathrm{Cov}}_n\,(g(Y), h(Y)\,|\,X)}{\widehat{\mathrm{Std}}_n\,(g(Y)\,|\,X)\,\widehat{\mathrm{Std}}_n\,(h(Y)\,|\,X)}\,,$$

if the right-hand side is well defined and lies in $[-1, 1]$, and

$$\widehat{\mathrm{Cor}}_n\,(g(Y), h(Y)\,|\,X) = \mathrm{sign}\,(\widehat{\mathrm{Cov}}_n\,(g(Y), h(Y)\,|\,X))$$

otherwise. Suppose $W_n$ is a probability weight function. Then these estimators equal the corresponding second order quantities of the probability distribution $\hat{P}_n^Y(\cdot\,|\,X)$. Consequently $\widehat{\mathrm{Cov}}_n\,(g(Y), g(Y)\,|\,X) \geqq 0$ and Schwarz's inequality

$$(\widehat{\mathrm{Cov}}_n\,(g(Y), h(Y)\,|\,X)^2 \leqq \widehat{\mathrm{Var}}_n\,(g(Y)\,|\,X)\,\widehat{\mathrm{Var}}_n\,(h(Y)\,|\,X)$$

holds.

Suppose now that $\{W_n\}$ is consistent. Then

$$\widehat{\mathrm{Cov}}_n\,(g(Y), h(Y)\,|\,X) \to \mathrm{Cov}\,(g(Y), h(Y)\,|\,X) \quad \text{in} \quad L^1\,,$$
$$\widehat{\mathrm{Var}}_n\,(g(Y)\,|\,X) \to \mathrm{Var}\,(g(Y)\,|\,X) \quad \text{in} \quad L^1\,,$$

and

$$\widehat{\mathrm{Std}}_n\,(g(Y)\,|\,X) \to \mathrm{Std}\,(g(Y)\,|\,X) \quad \text{in} \quad L^1\,.$$

If $\mathrm{Std}\,(g(Y)\,|\,X)\,\mathrm{Std}\,(h(Y)\,|\,X) > 0$ with probability one, then

$$\widehat{\mathrm{Cor}}_n\,(g(Y), h(Y)\,|\,X) \to \mathrm{Cor}\,(g(Y), h(Y)\,|\,X) \quad \text{in probability}$$

and hence in $L^r$ for all $r \geqq 1$.

**7. Estimation of conditional quantiles.** In this section $Y$ is real valued. The conditional distribution function $F^Y(\cdot\,|\,X)$ is defined by $F^Y(y\,|\,X) = P^Y((-\infty, y]\,|\,X)$. Let $0 < p < 1$. *The lower* $p$th *quantile* $L^Y(p\,|\,X)$, *upper* $p$th *quantile* $U^Y(p\,|\,X)$, *and* $p$th *quantile* $Q^Y(p\,|\,X)$ of $F^Y(\cdot\,|\,X)$ are defined by

$$L^Y(p\,|\,X) = \inf\,[y: F^Y(y\,|\,X) \geqq p]\,,$$
$$U^Y(p\,|\,X) = \sup\,[y: F^Y(y\,|\,X) \leqq p]\,,$$

and

$$Q^Y(p\,|\,X) = (L^Y(p\,|\,X) + U^Y(p\,|\,X))/2\,.$$

Let $W_n$ be a weight function. The above conditional quantities can be estimated by

$$\hat{F}_n^Y(y\,|\,X) = \hat{P}_n^Y((-\infty, y]\,|\,X) = \textstyle\sum_i W_{ni}(X)I_{\{Y_i \leqq y\}}\,,$$
$$\hat{L}_n^Y(p\,|\,X) = \inf\,[y: \hat{F}_n^Y(y\,|\,X) \geqq p]\,,$$
$$\hat{U}_n^Y(p\,|\,X) = \sup\,[y: \hat{F}_n^Y(y\,|\,X) \leqq p]\,,$$

and

$$\hat{Q}_n^Y(p\,|\,X) = (\hat{L}_n^Y(p\,|\,X) + \hat{U}_n^Y(p\,|\,X))/2\,.$$

The next result will be proven in Section 12.

THEOREM 3. *Let $\{W_n\}$ be a consistent sequence of probability weights and let $0 < p < 1$. Then*

$$(\hat{L}_n^Y(p \mid X) - L^Y(p \mid X))^- \to 0 \quad \textit{in probability} \tag{9}$$

*and*

$$(\hat{U}_n^Y(p \mid X) - U^Y(p \mid X))^+ \to 0 \quad \textit{in probability.} \tag{10}$$

*If $r \geqq 1$ and $E|Y|^r < \infty$, then in (9) and (10) convergence in probability can be replaced by convergence in $L^r$.*

COROLLARY 5. *Let $\{W_n\}$ be a consistent sequence of probability weights, let $0 < p < 1$, and suppose that $L^Y(p \mid X) = U^Y(p \mid X)$ with probability one. Then*

$$\hat{Q}_n^Y(p \mid X) \to Q^Y(p \mid X) \quad \textit{in probability.} \tag{11}$$

*If $r \geqq 1$ and $E|Y|^r < \infty$, then in (11) convergence in probability can be replaced by convergence in $L^r$.*

COROLLARY 6. *Let $\{W_n\}$ be a consistent sequence of probability weights, let $0 < p_1 < p_2 < 1$, and let $J(p)$, $p_1 \leqq p \leqq p_2$, be a continuous function. Then*

$$\int_{p_1}^{p_2} J(p)\hat{Q}_n^Y(p \mid X)\, dp \to \int_{p_1}^{p_2} J(p)Q^Y(p \mid X)\, dp \quad \textit{in probability.} \tag{12}$$

*If $r \geqq 1$ and $E|Y|^r < \infty$, then in (12) convergence in probability can be replaced by convergence in $L^r$.*

**8. Approximate Bayes rules.** In this section $Y$ is real valued. Let $\mathscr{A}$ be a measurable space of "actions" and let $\mathscr{L}: \mathbb{R} \times \mathscr{A} \to \mathbb{R}$ be a jointly measurable nonnegative loss function. In each model considered in this section $E\mathscr{L}(Y, a) < \infty$ for all $a \in \mathscr{A}$.

Let $d: \mathbb{R}^d \to \mathscr{A}$ be a (measurable) decision rule for choosing $a \in \mathscr{A}$ after having observed $X$ but before having observed $Y$. The *Bayes risk* associated with such a rule is $E\mathscr{L}(Y, d(X)) = EE(\mathscr{L}(Y, d(X)) \mid X)$. In the specific models discussed below there will be a *minimum Bayes risk* $R$ associated with a (not necessarily uniquely determined) *Bayes rule* $\delta$ which satisfies

$$E(\mathscr{L}(Y, \delta(X)) \mid X) = \inf_{a \in \mathscr{A}} E(\mathscr{L}(Y, a) \mid X)\,.$$

Then

$$R = E\mathscr{L}(Y, \delta(X)) \leqq E\mathscr{L}(Y, d(X))$$

for all decision rules $d$.

The Bayes rule is defined in terms of $E(\mathscr{L}(Y, a) \mid X)$ for $a \in \mathscr{A}$. If this is unknown it can be estimated by

$$\hat{E}_n(\mathscr{L}(Y, a) \mid X) = \sum_i W_{ni}(X)\mathscr{L}(Y_i, a)\,,$$

where $W_n$ is a weight function. The Bayes rule $\delta$ can in turn be approximated by $\hat{\delta}_n$ chosen so that

$$\hat{E}_n(\mathscr{L}(Y, \hat{\delta}_n(X) X) = \inf_{a \in \mathscr{A}} \hat{E}_n(\mathscr{L}(Y, a) \mid X)\,.$$

Such $a$ (not necessarily uniquely determined) decision rule $\hat{\delta}_n$ is called an *approximate Bayes rule.* A sequence $\{\hat{\delta}_n\}$ of such rules is said to be *consistent in Bayes risk* if

$$\lim_n E\mathscr{L}(Y, \hat{\delta}_n(X)) = E\mathscr{L}(Y, \delta(X)) = R .$$

Consistency in Bayes risk will be obtained in three important models.

MODEL 1 (prediction with squared error loss). $\mathscr{A} = \mathbb{R}$, $EY^2 < \infty$, and $\mathscr{L}(y, a) = (y - a)^2$. In this model $\delta(X) = E(Y \mid X)$ and $R = E(Y - E(Y \mid X))^2$. Let $W_n$ be a normal weight function. Then $\hat{\delta}_n(X) = \hat{E}_n(Y \mid X)$. The Bayes risk of $\hat{\delta}_n$ is given by

$$E(Y - \hat{E}_n(Y \mid X))^2 = E(Y - E(Y \mid X))^2 + E(\hat{E}_n(Y \mid X) - E(Y \mid X))^2 .$$

MODEL 2 (prediction with weighted absolute error loss). $\mathscr{A} = \mathbb{R}$, $E|Y| < \infty$, and $\mathscr{L}(y, a) = c(p(y - a)^+ + (1 - p)(y - a)^-)$ for some constants $c > 0$ and $p \in (0, 1)$. In this model $\delta(X)$ is any value in $[L^Y(p \mid X), U^Y(p \mid X)]$, e.g., $\delta(X) = Q^Y(p \mid X)$ (see Problem 3 on page 51 of Ferguson (1967)). Let $W_n$ be a probability weight function. Then $\hat{\delta}_n(X)$ is any value in $[\hat{L}_n{}^Y(p \mid X), \hat{U}_n{}^Y(p \mid X)]$, e.g., $\hat{\delta}_n(X) = \hat{Q}_n{}^Y(p \mid X)$. This model can be applied to prediction problems with absolute value loss by setting $c = 2$ and $p = .5$, so that $\mathscr{L}(y, a) = |y - a|$. In this case $\delta(X) = Q^Y(.5 \mid X)$ is the conditional median of $Y$ given $X$ and $\hat{\delta}_n(X) = \hat{Q}_n{}^Y(.5 \mid X)$ is an estimate of this conditional median.

MODEL 3 (multiple classification). $\mathscr{A} = \mathbb{R}$ and $\mathscr{L}(y, a) = 0$ or $1$ according as $y = a$ or $y \neq a$. In this model $\delta(X)$ is any value of $y \in \mathbb{R}$ such that

$$P^Y(\{\delta(X)\} \mid X) = \max_y P^Y(\{y\} \mid X) .$$

Let $W_n$ be a weight function. Then $\hat{\delta}_n(X)$ is any value $y \in \mathbb{R}$ such that

$$\hat{P}_n{}^Y(\{\delta(X)\} \mid X) = \max_y \hat{P}_n{}^Y(\{y\} \mid X) .$$

This model is applicable to multiple classification problems. Here $Y$ takes values from some finite set and $\delta(X)$ and $\hat{\delta}_n(X)$ take values from this set.

The following result will be proven in Section 14.

THEOREM 4. *Let Model* 1, 2 *or* 3 *hold. Let* $\{W_n\}$ *be a consistent sequence of weights which are normal if Model* 1 *holds and probability weights if Model* 2 *holds. Then* $\{\hat{\delta}_n\}$ *is consistent in Bayes risk.*

**9. Related work.** Nearest neighbor procedures were first studied in the context of nonparametric classification by Fix and Hodges (1951). They verified the consistency in Bayes risk of $\{\hat{\delta}_n\}$ in the simple classification problem under some regularity conditions when $W_n$ is the uniform $k_n$-NN weight function, $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$.

A probability weight function $W_n$ is called a *unit weight function* if there is a function $i_n(X) = i_n(X, X_1, \cdots, X_n)$ ranging over $\{1, \cdots, n\}$ such that $W_{n, i_n(X)}(X) = 1$ with probability one. Let $\hat{\delta}_n{}^{(1)}(X)$ be the approximate Bayes rule corresponding to such a unit weight function. If $P(i_n(X) \in I_{n1}(X)) = 1$, $\hat{\delta}_n{}^{(1)}(X)$ is called the *nearest neighbor* (NN) rule. In Models 1–3, $\hat{\delta}_n(X) = Y_{i_n}(X)$.

Cover and Hart (1967) and Cover (1968) studied NN rules and rules corresponding to uniform $k$-NN weights under some regularity conditions. The first paper considered Model 1 and the second paper considered Models 2 and 3. Their results for NN rules can be extended to the level of generality of the present paper as follows (the proofs combine their arguments with the results of this paper in a straightforward manner): let $\{W_n\}$ be a sequence of unit weights satisfying the first two conditions of Corollary 1 (according to the proof of Theorem 2 this allows $W_n$ to be the sequence of 1-NN weights if $P(\#(I_{n1}(X)) = 1) = 1$ for all $n \geqq 1$). Let $\hat{\delta}_n^{(1)}$ be the NN rule corresponding to $W_n$. In Model 1

$$\lim_n E\mathscr{L}(Y, \hat{\delta}_n^{(1)}(X)) = 2R\,;$$

in Model 2

$$\limsup_n E\mathscr{L}(Y, \hat{\delta}_n^{(1)}(X)) \leqq 2R\,;$$

and in Model 3

$$\lim_n E\mathscr{L}(Y, \hat{\delta}_n^{(1)}(X)) = 1 - E \textstyle\sum_y (P(Y = y \mid X))^2 \leqq R(2 - \alpha R) \leqq 2R\,,$$

where $\alpha = M/(M-1)$ if the support of the distribution of $Y$ is a finite set having $M$ points and $\alpha = 1$ otherwise. Thus in Models 1–3 there is no decision rule which for large $n$ has a Bayes risk noticeably less than one-half that of $\hat{\delta}_n^{(1)}$. This point was emphasized in Cover and Hart (1967) and Cover (1968). Fritz (1975) studied the NN rule for Model 3. The geometrical arguments used to prove Proposition 12 below are similar in part to those used by Fritz to prove his Lemma 3.

Liggett (1976) has obtained an extension of the Erdös–Ko–Rado combinatorial theorem and applied it to obtain the following nontrivial result: let $Y, Y_1, Y_2, \cdots, Y_n$ be i.i.d. and take on only finitely many values, let $w_1, \cdots, w_n$ be nonnegative numbers adding up to one and let the estimator $\hat{Y}$ of $Y$ be chosen randomly from among the values of $y$ which maximize $\sum_1^n w_i I_{\{y\}}(Y_i)$. Then $P(\hat{Y} = Y) \geqq P(Y_1 = Y)$. Liggett's result can be used to show that in Model 3 if $\{W_n\}$ is a sequence of probability weights satisfying the first two conditions of Corollary 1 and $\{\hat{\delta}_n\}$ is the corresponding sequence of approximate Bayes rules, then under a variety of mild additional conditions

$$\limsup_n E\mathscr{L}(Y, \hat{\delta}_n(X)) \leqq \lim_n E\mathscr{L}(Y, \hat{\delta}_n^{(1)}(X))\,.$$

Thus these rules all do asymptotically at least as well as the rule $\hat{\delta}_n^{(1)}$.

Watson (1964) mentioned the possibility of estimating $E(Y \mid X)$ using uniform $k$-NN weights. Royall (1966) obtained some asymptotic results for estimators $\hat{E}_n(Y \mid X = x)$ of $E(Y \mid X = x)$ for fixed $x$ determined by weights $W_n$ corresponding to $c_n$ as in Section 3. Stone (1975) discussed results of applying nearest neighbor estimators to some simulated data.

Kernel weights are of the form

$$W_{ni}(X) = \frac{K_n(\rho_n(X_i, X))}{\sum_i K_n(\rho_n(X_i, X))}\,,$$

where $K_n$ is a positive nonincreasing function on $[0, \infty)$ and $\rho_n$ is an appropriate metric on $\mathbb{R}^d$. It follows from Proposition 11 below that if $\rho_n$ is given by (6) and $f$ is a nonnegative Borel function on $\mathbb{R}^d$, then

$$E \sum_i W_{ni}(X)f(X_i) \leqq \beta\left(d, \frac{a}{b}\right) \sum_{i=1}^n \frac{1}{i} Ef(X) .$$

Since $\sum_1^\infty i^{-1} = \infty$, this result does not quite show that the first condition of Corollary 1 holds. Thus it is not clear when a sequence of kernel weights is consistent. Of course one can always trim kernel weights (as was done for local linear weights in Section 4) and use Corollary 2 to obtain a consistent sequence of weights. For work on kernel estimators see Watson (1964), Nadaraya (1964), (1970), Schuster (1968), (1972), Rosenblatt (1969), Benedetti (1974), (1975), and Butler (1975). For a related method based on Fourier series expansions see Raman (1971). For other related methods see Priestley and Chao (1972) and Major (1973). These methods were suggested by work of Rosenblatt (1956), Parzen (1962) and others on kernel methods of nonparametric density estimation.

Some other approaches to nonparametric regression are potential functions (Aizerman, Braverman and Rozonoer (1970) and the references cited therein, Yakowitz and Fisher (1975), Fisher and Yakowitz (1976)); stochastic approximation (Révész (1973)); splines (Wold (1974) and the references cited therein, Wahba and Wold (1975)); AID (Morgan and Sonquist (1963), Sonquist and Morgan (1964)); SMOFIT (Beaton and Tukey (1974)); and random piecewise linear functions (Breiman and Meisel (1976)).

In the multiple classification problem Van Ryzin (1966) obtained rules which are consistent in Bayes risk under various regularity conditions. Recently, after the original version of this paper was written, Gordon and Olshen (1975) showed that a variation of a procedure of Friedman (1976) yields rules which are consistent in Bayes risk under no regularity conditions.

Beaton and Tukey (1974) used "running medians," which is a special case of the estimator $\hat{Q}_n^Y(.5 \mid X)$ discussed in Section 7. Trend removal, discussed in Section 5, was suggested by similar techniques used in their paper. Estimators closely related to those analyzed in Corollary 6 of Section 7 were used successfully by Cleveland and Kleiner (1975). The numerical example they considered shows the need for handling ties properly, as was done in (8).

**10. Proof of Theorem 1.** In this section $X, X_1, X_2, \cdots$ are i.i.d. $\mathbb{R}^d$-valued random variables and $\{W_n\}$ is a sequence of weights.

PROPOSITION 1. *Suppose that* (1)—(3) *hold. Let* $r \geqq 1$ *and let* $f$ *be a Borel function on* $\mathbb{R}^d$ *such that* $E|f(X)|^r < \infty$, *Then*

$$\lim_n E \sum_i |W_{ni}(X)||f(X_i) - f(X)|^r = 0 .$$

PROOF. Choose $\varepsilon > 0$. Let $h$ be a continuous function on $\mathbb{R}^d$ having compact support and such that $E|f(X) - h(X)|^r \leqq \varepsilon$. By (1)

$$E \sum_i |W_{ni}(X)||f(X_i) - h(X_i)|^r \leqq CE|f(X) - h(X)|^r \leqq C\varepsilon .$$

It follows from (2) that

$$E \sum_i |W_{ni}(X)||f(X) - h(X)|^r \leqq DE|f(X) - h(X)|^r \leqq D\varepsilon .$$

Thus to prove that the conclusion of Proposition 1 holds for $f$, it suffices to prove that the conclusion holds with $f$ replaced by $h$. In other words, without loss of generality it can be assumed that $f$ itself is continuous and has compact support. Let this be the case and let $M$ be an upper bound to $|f|$. Choose $\varepsilon > 0$. There is an $a > 0$ such that $|f(x_1) - f(x)|^r \leqq \varepsilon$ if $x \in \mathbb{R}^d$, $x_1 \in \mathbb{R}^d$, and $||x_1 - x|| \leqq a$. Then by (2)

$$E \sum_i |W_{ni}(X)||f(X_i) - f(X)|^r \leqq (2M)^r E \sum_i |W_{ni}(X)| I_{\{||X_i - X|| > a\}} + D\varepsilon .$$

It follows from (2) and (3) that

$$\lim_n E \sum_i |W_{ni}(X)| I_{\{||X_i - X|| > a\}} = 0 .$$

Thus

$$\limsup_n E \sum_i |W_{ni}(X)||f(X_i) - f(X)|^r \leqq D\varepsilon .$$

Since $\varepsilon$ can be made arbitrarily small, the conclusion of Proposition 1 holds.

PROPOSITION 2. *Let $\{W_n\}$ be a sequence of nonnegative weights. Suppose that (1)—(3) hold and that there are sequences $\{M_n\}$ and $\{N_n\}$ of nonnegative constants such that*

$$\lim_n P(M_n \leqq \sum_i W_{ni}(X) \leqq N_n) = 1 .$$

*Let $f$ be a nonnegative Borel function on $\mathbb{R}^d$ such that $Ef(X) < \infty$. Then*

$$\liminf_n E \sum_i W_{ni}(X) f(X_i) \geqq (\liminf_n M_n) Ef(X)$$

*and*

$$\limsup_n E \sum_i W_{ni}(X) f(X_i) \leqq (\limsup_n N_n) Ef(X) .$$

PROOF. Set $A_n = \{M_n \leqq \sum_i W_{ni}(X) \leqq N_n\}$. Without loss of generality it can be assumed that $M_n \leqq D$ for all $n \geqq 1$. Then

$$M_n - DI_{A_n^c} \leqq \sum_i W_{ni}(X) \leqq N_n + DI_{A_n^c}$$

and hence

$$M_n Ef(X) - DEI_{A_n^c} f(X) \leqq E \sum_i W_{ni}(X) f(X) \leqq N_n Ef(X) + DEI_{A_n^c} f(X) .$$

Now $\lim_n P(A_n^c) = 0$ and hence $\lim_n EI_{A_n^c} f(X) = 0$. Consequently

$$\liminf_n E \sum_i W_{ni}(X) f(X) \geqq (\liminf_n M_n) Ef(X)$$

and

$$\limsup_n E \sum_i W_{ni}(X) f(X) \leqq (\limsup_n N_n) Ef(X) .$$

Since by Proposition 1

$$\lim_n (E \sum_i W_{ni}(X) f(X_i) - E \sum_i W_{ni}(X) f(X)) = 0 ,$$

the desired conclusion holds.

PROPOSITION 3. *Suppose that* (1)—(3) *hold and that there are sequences* $\{M_n\}$ *and* $\{N_n\}$ *of nonnegative constants such that*

$$\lim_n P(M_n \leqq \sum_i W_{ni}^2(X) \leqq N_n) = 1 .$$

*Let $f$ be a nonnegative Borel function on* $\mathbb{R}^d$ *such that* $Ef(X) < \infty$. *Then*

$$\liminf_n E \sum_i W_{ni}^2(X) f(X_i) \geqq (\liminf_n M_n) Ef(X)$$

*and*

$$\limsup_n E \sum_i W_{ni}^2(X) f(X_i) \leqq (\limsup_n N_n) Ef(X) .$$

PROOF. This result follows by applying Proposition 2 directly to $\{W_n{}^2\}$, noting that this sequence of weights satisfies (1)—(3) if $C$ and $D$ are replaced by $CD$ and $D^2$ respectively.

PROPOSITION 4. *Suppose that* (1)—(3) *hold and let $f$ be a Borel function on* $\mathbb{R}^d$. *Then for every* $\varepsilon > 0$

$$\sum_i |W_{ni}(X)| I_{\{|f(X_i) - f(X)| > \varepsilon\}} \to 0 \quad \textit{in probability.}$$

PROOF. Let $\varepsilon > 0$ be given. Choose $M > 0$. Set $h = (f \wedge M) \vee (-M)$, so that $|h| \leqq M$ and $h(x) = f(x)$ whenever $|f(x)| \leqq M$. It follows from Proposition 1 that

$$\lim_n E \sum_i |W_{ni}(X)| |h(X_i) - h(X)| = 0$$

and hence that

$$\sum_i |W_{ni}(X)| I_{\{|h(X_i) - h(X)| > \varepsilon\}} \to 0 \quad \text{in probability.}$$

Since $\{h(X_i) \neq f(X_i)\} \subset \{|f(X_i)| > M\}$, it follows from (1) that

$$E \sum_i |W_{ni}(X)| I_{\{h(X_i) \neq f(X_i)\}} \leqq CP(|f(X)| > M) .$$

By (2)

$$E \sum_i |W_{ni}(X)| I_{\{h(X) \neq f(X)\}} \leqq DP(|f(X)| > M) .$$

Since $P(|f(X)| > M)$ can be made arbitrarily small by choosing $M$ sufficiently large, the conclusion of the proposition follows from the last three displayed equations.

PROPOSITION 5. *Suppose* $\{W_n\}$ *satisfies* (1)—(4). *If* $r \geqq 1$ *and $f$ is a Borel function on* $\mathbb{R}^d$ *such that* $E|f(X)|^r < \infty$, *then* $\sum_i W_{ni}(X) f(X_i) \to f(X)$ *in* $L^r$.

PROOF. It follows from (2) that $|\sum_i W_{ni}(X) - 1|^r \leqq (1 + D)^r$. Thus by (4)

$$\lim_n E|(\sum_i W_{ni}(X) - 1) f(X)|^r = 0 .$$

It follows from (2) and Proposition 1, together with Hölder's inequality for $r > 1$, that

$$\lim_n E|\sum_i W_{ni}(X)(f(X_i) - f(X))|^r = 0 .$$

The conclusion of the proposition follows easily from the last two displayed results.

PROPOSITION 6. *Suppose that* $\{W_n\}$ *is a sequence of nonnegative weights and that* *for every bounded and continuous function $f$ on $\mathbb{R}^d$*

$$\textstyle\sum_i W_{ni}(X)f(X_i) \to f(X) \quad \text{in probability.}$$

*Then $\{W_n\}$ satisfies* (3).

PROOF. Let $a > 0$ be given. Choose $x_0 \in \mathbb{R}^d$ and let $f$ be a bounded and continuous nonnegative function on $\mathbb{R}^d$ such that $f(x) = 0$ for $||x - x_0|| \leqq a/3$ and $f(x) = 1$ for $||x - x_0|| \geqq 2a/3$. Then on $\{||X - x_0|| \leqq a/3\}$, $f(X) = 0$ and

$$\textstyle\sum_i W_{ni}(X)f(X_i) \geqq \sum_i W_{ni}(X)I_{\{||X_i - X|| > a\}} \,.$$

Consequently

$$I_{\{||X - x_0|| \leqq a/3\}} \textstyle\sum_i W_{ni}(X)I_{\{||X_i - X|| > a\}} \to 0 \quad \text{in probability.}$$

Thus for every compact subset $B$ of $\mathbb{R}^d$

$$I_B(X) \textstyle\sum_i W_{ni}(X)I_{\{||X_i - X|| > a\}} \to 0 \quad \text{in probability.}$$

Therefore (3) holds as desired.

PROPOSITION 7. *Let $\{W_n\}$ be a sequence of nonnegative weights satisfying the following property: for every nonnegative Borel function $f$ on $\mathbb{R}^d$ such that $Ef(X) < \infty$,* $\limsup_n E \sum_i W_{ni}(X)f(X_i) < \infty$. *Then there is a positive integer $n_0$ and a positive constant $C$ such that for every nonnegative Borel function $f$ on $\mathbb{R}^d$*

$$E \textstyle\sum_i W_{ni}(X)f(X_i) \leqq CEf(X) \qquad \text{for all} \quad n \geqq n_0 \,.$$

PROOF. Suppose the conclusion of the proposition is false. Then there is an increasing sequence $\{n_\nu\}$ of positive integers and a sequence $\{f_\nu\}$ of nonnegative Borel functions on $\mathbb{R}^d$ such that $Ef_\nu(X) = 2^{-\nu}$ and

$$E \textstyle\sum_i W_{n_\nu i}(X)f_\nu(X_i) \geqq \nu \,.$$

Set $f = \sum_{\nu=1} f_\nu$. Then $f$ is a nonnegative Borel function on $\mathbb{R}^d$, $Ef(X) = 1 < \infty$, and

$$E \textstyle\sum_i W_{n_\nu i}(X)f(X_i) \geqq E \sum_i W_{n_\nu i}(X)f_\nu(X_i) \geqq \nu \,.$$

Thus $\limsup_n E \sum_i W_{ni}(X)f(X_i) = \infty$ and hence the hypothesis of the proposition is false. Thus the proposition is valid.

PROPOSITION 8. *Let $\{W_n\}$ be a sequence of weights satisfying the following property: there is a sequence $\{Y_i\}$ of independent standard normal real valued random variables such that $\{Y_i\}$ is independent of $(X, X_1, X_2, \cdots)$ and $\sum_i W_{ni}(X)Y_i \to 0$ in probability. Then $\sum_i W_{ni}^2(X) \to 0$ in probability.*

PROOF. The conditional distribution of $\sum_i W_{ni}(X)Y_i$ given $X, X_1, X_2, \cdots, X_n$ is normal with mean zero and variance $\sum_i W_{ni}^2(X)$. Thus for $\varepsilon > 0$

$$P(|\textstyle\sum_i W_{ni}(X)Y_i| > \varepsilon) \geqq \left(2 \int_1^\infty \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-y^2/2}\, dy\right) P(\sum_i W_{ni}^2(X) > \varepsilon^2)$$

and hence $\lim_n P(\sum_i W_{ni}^2(X) > \varepsilon^2) = 0$. The conclusion of the proposition now follows easily.

Theorom 1 will now be proven. The various necessity results follow easily from Propositions 6, 7, and 8. To complete the proof of Theorem 1 it suffices to show that if $\{W_n\}$ is a sequence of weights satisfying (1)—(5), $r \geqq 1$, $(X, Y)$, $(X_1, Y_1)$, $(X_2, Y_2)$, $\cdots$ are i.i.d., $Y$ is real valued, and $E|Y|^r < \infty$, then

$$\lim_n E|\textstyle\sum_i W_{ni}(X)Y_i - E(Y\,|\,X)|^r = 0 . \tag{13}$$

Consider first the case $r = 2$. Set $Z = Y - E(Y\,|\,X)$, $Z_i = Y_i - E(Y_i\,|\,X_i)$, $f(X) = E(Y\,|\,X)$, and $h(X) = E(Z^2\,|\,X)$. Then $E(Z_i\,|\,X_i) = 0$, $Ef^2(X) < \infty$, and $Eh(X) = E(Y - E(Y\,|\,X))^2 \leqq EY^2 < \infty$. Write

$$\textstyle\sum_i W_{ni}(X)Y_i - E(Y\,|\,X) = (\sum_i W_{ni}(X)f(X_i) - f(X)) + \sum_i W_{ni}(X)Z_i .$$

By Proposition 5, $\sum_i W_{ni}(X)f(X_i) \rightarrow f(X)$ in $L^2$. Now

$$\begin{aligned} E(\textstyle\sum_i W_{ni}(x)Z_i)^2 &= EE((\textstyle\sum_i W_{ni}(x)Z_i)^2\,|\,X_1, \cdots, X_n) \\ &= E\textstyle\sum_i W_{ni}^2(x)E(Z_i^2\,|\,X_i) \\ &= E\textstyle\sum_i W_{ni}^2(x)h(X_i) . \end{aligned}$$

Thus

$$E(\textstyle\sum_i W_{ni}(X)Z_i)^2 = E\sum_i W_{ni}^2(X)h(X_i) .$$

By (2) and (5) $\sum_i W_{ni}^2(X) \rightarrow 0$ in probability. Proposition 3 now implies that

$$\lim_n E(\textstyle\sum_i W_{ni}(X)Z_i)^2 = 0$$

and hence that (13) holds for $r = 2$.

Consider now the general case $r \geqq 1$. Given a positive number $M$ set $Y^{(M)} = (Y \wedge M) \vee (-M)$ and $Y_i^{(M)} = (Y_i \wedge M) \vee (-M)$. Then $\lim_{M\rightarrow\infty} E|Y - Y^{(M)}|^r = 0$. It now follows from (1) and (2) (and Hölder's inequality for $r > 1$) that

$$\lim_{M\rightarrow\infty} E|\textstyle\sum W_{ni}(X)(Y_i - Y_i^{(M)})|^r = 0 \qquad \text{uniformly in} \quad n .$$

Observe also that

$$E|E(Y\,|\,X) - E(Y^{(M)}\,|\,X)|^r = E|E(Y - Y^{(M)})\,|\,X)|^r \leqq E|Y - Y^{(M)}|^r ,$$

which approaches zero as $M \rightarrow \infty$. Thus to prove that (13) holds for $Y$, it is enough to show that it holds for $Y^{(M)}$. In other words, without loss of generality it can be assumed that $Y$ is bounded. But if $Y$ is bounded, then to prove that (13) holds for all $r \geqq 1$, it is enough to show that it holds for $r = 2$. Since this has already been done, the proof of Theorem 1 is complete.

**11. Proof of Theorem 2.** In this section the notation and terminology from Section 3 is used. In particular $s_{nj}$, $1 \leqq j \leqq d$, is the scale based on $X_1, \cdots, X_n$. Also set $I_{n0}(X) = \phi$ and $I_{nt}(X) = I_{nk}(X)$ for $t > 0$ and $k \leqq t < k + 1$.

PROPOSITION 9. *For every $a > 0$*

$$\lim_{\alpha\downarrow 0} (\limsup_n P(\max_{i \in I_{n,\alpha n}(X)} ||X_i - X|| > a)) = 0 .$$

PROOF. Choose $a > 0$ and $\varepsilon > 0$. It suffices to show that there is an $\alpha \in (0, 1)$ such that

$$\limsup_n P(\max_{i \in I_{n,\alpha n}(X)} ||X_i - X|| > a) \leqq \varepsilon . \tag{14}$$

In proving this result it can be assumed, without loss of generality, that each coordinate of $X$ has a nondegenerate distribution. It can also be assumed that $s_{n1} = 1$ on the set where $s_{n1} > 0$; for dividing all the numbers $s_{nj}$ by a positive random variable ($s_{n1}$ if $s_{n1} > 0$ and 1 otherwise) does not affect $I_{nk}(X)$ or the regularity of $\{s_n\}$. It now follows from the definition of regularity that there are positive numbers $t$ and $T$ such that for $n$ sufficiently large

$$P\left(\frac{1}{T} \leqq s_{nj} \leqq \frac{1}{t} \text{ for } 1 \leqq j \leqq d\right) \geqq 1 - \frac{\varepsilon}{2};$$

and hence for $n$ sufficiently large, the random metric $\rho_n$ satisfies

$$P(t\|u - v\| \leqq \rho_n(u, v) \leqq T\|u - v\| \text{ for all } u, v \in \mathbb{R}^d) \geqq 1 - \frac{\varepsilon}{2}. \tag{15}$$

Let $S$ denote the *support* of the distribution of $X$, that is, the set of all $x \in \mathbb{R}^d$ such that $P(|X - x| < \delta) > 0$ for every $\delta > 0$. Then $S$ is a closed subset of $\mathbb{R}^d$ and $P(X \in S) = 1$. For $x \in \mathbb{R}^d$ let $N_n(x)$ denote the number of points $X_i$, $1 \leqq i \leqq n$, such that $\|X_i - x\| \leqq at/T$. If $x \in S$, then $P(\lim_n N_n(x)/n > 0) = 1$ by the strong law of large numbers. Therefore $P(\lim_n N_n(X)/n > 0) = 1$ and hence there is an $\alpha \in (0, 1)$ such that for $n$ sufficiently large

$$P(N_n(X) \geqq \alpha n) \geqq 1 - \frac{\varepsilon}{2}. \tag{16}$$

Suppose that $N_n(X) \geqq \alpha n$ and that $t\|u - v\| \leqq \rho_n(u, v) \leqq T\|u - v\|$ for all $u, v \in \mathbb{R}^d$. If $\|X_i - X\| \leqq at/T$, then $\rho_n(X_i, X) \leqq at$. Thus there are at least $\alpha n$ values of $i$ such that $\rho_n(X_i, X) \leqq at$ and hence $\rho_n(X_i, X) \leqq at$ for all $i \in I_{n,\alpha n}(X)$. Therefore

$$\|X_i - X\| \leqq \frac{1}{t}\rho_n(X_i, X) \leqq a \qquad \text{for all} \quad i \in I_{n,\alpha n}(X)$$

and hence $\max_{i \in I_{n,\alpha n}(X)} \|X_i - X\| \leqq a$. It now follows from (15) and (16) that for $n$ sufficiently large

$$P(\max_{i \in I_{n,\alpha n}(X)} \|X_i - X\| \leqq a) \geqq 1 - \varepsilon.$$

Thus (14) holds as desired.

For $0 < c \leqq 1$ let $\mathscr{V}(d, c)$ denote the collection of all subsets $V$ of $\mathbb{R}^d$ such that if $u$ and $v$ are two nonzero elements of $V$, the cosine of the angle between them is greater than $(1 - c^2/2)$, i.e.,

$$u \cdot v > \left(1 - \frac{c^2}{2}\right)\|u\|\,\|v\|.$$

Since the unit sphere in $\mathbb{R}^d$ is compact, $\mathbb{R}^d$ can be covered by a finite subcollection of $\mathscr{V}(d, c)$ (only cones in $\mathscr{V}(d, c)$ need be considered). Let $\beta(d, c)$ denote the minimum cardinality of subcollections of $\mathscr{V}(d, c)$ which cover $\mathbb{R}^d$. Then $\beta(d, c)$ is a positive integer valued function which is nondecreasing in $d$ and nonincreasing in $c$. It is easily seen that $\beta(1, c) = 2$ and that $\beta(2, 1) = 6$. The explicit determination of $\beta(d, c)$ for $d \geqq 3$ is a difficult combinatorial geometry problem. Fortunately it is not necessary in the present context.

PROPOSITION 10. *Let $0 < a \leqq b_j \leqq b$ for $1 \leqq j \leqq d$. Let $V \in \mathscr{V}(d, a/b)$ and let $u = (u_1, \cdots, u_d)$ and $v = (v_1, \cdots, v_d)$ be in $V$ and such that $0 < ||u|| \leqq ||v||$. Let $\tilde{u}$ and $\tilde{v}$ be determined by $\tilde{u}_j = b_j u_j$ and $\tilde{v}_j = b_j v_j$ for $1 \leqq j \leqq d$. Then $||\tilde{v}|| > ||\tilde{v} - \tilde{u}||$.*

PROOF. Suppose first that $||u|| = ||v||$. Then

$$\frac{u \cdot v}{||u||^2} = \frac{u \cdot v}{||u||\,||v||} > \frac{2b^2 - a^2}{2b^2}$$

and hence

$$\begin{aligned} a^2||v||^2 - b^2||v - u||^2 &= (a^2 - b^2)||v||^2 - b^2||u||^2 + 2b^2 u \cdot v \\ &= (a^2 - 2b^2)||u||^2 + 2b^2 u \cdot v > 0 . \end{aligned}$$

Consequently $||\tilde{v}|| \geqq a||v|| > b||v - u|| \geqq ||\tilde{v} - \tilde{u}||$.

Consider now the general case. Set $t = ||v||/||u|| \geqq 1$, $v_0 = t^{-1}v$ and $\tilde{v}_0 = t^{-1}\tilde{v}$. Then $||u|| = ||v_0||$, $v = tv_0$ and $\tilde{v} = t\tilde{v}_0$. Now $||\tilde{v}_0|| > ||\tilde{v}_0 - \tilde{u}||$ by what has already been shown. It follows easily from the formula $||u - v||^2 = ||u||^2 - 2u \cdot v + ||v||^2$ that

$$\begin{aligned} ||\tilde{v}||^2 - ||\tilde{v} - \tilde{u}||^2 &= ||t\tilde{v}_0||^2 - ||t\tilde{v}_0 - \tilde{u}||^2 \\ &= t(||\tilde{v}_0||^2 - ||\tilde{v}_0 - \tilde{u}||^2) + (t - 1)||\tilde{u}||^2 > 0 \end{aligned}$$

and hence that $||\tilde{v}|| > ||\tilde{v} - \tilde{u}||$ as desired.

PROPOSITION 11. *Let $W_n$ be the probability weight function corresponding to $c_n$ and let $a$ and $b \geqq a$ satisfy* (7). *If $f$ is a nonnegative Borel function on $\mathbb{R}^d$ such that $Ef(X) < \infty$, then*

$$E \textstyle\sum_i W_{ni}(X) f(X_i) \leqq \beta\left(d, \frac{a}{b}\right) Ef(X) .$$

PROOF. Now $W_{ni}(X) = W_{ni}(X, X_1, \cdots, X_n)$, where $X, X_1, \cdots, X_n$ are i.i.d. Thus $X$ and $X_i$ can be interchanged to obtain

$$E(W_{ni}(X, X_1, \cdots, X_n) f(X_i)) = E(W_{ni}(X_i, X_1, \cdots, X, \cdots, X_n) f(X)) \qquad \text{for} \quad 1 \leqq i \leqq n .$$

Set $U_{ni}(X) = U_{ni}(X, X_1, \cdots, X_n) = W_{ni}(X_i, X_1, \cdots, X, \cdots, X_n)$ for $1 \leqq i \leqq n$ and $U_{ni}(X) = 0$ for $i > n$. Then $E(W_{ni}(X) f(X_i)) = E(U_{ni}(X) f(X))$ and hence

$$E \textstyle\sum_i W_{ni}(X) f(X_i) = E(f(X) \sum_i U_{ni}(X)) . \tag{17}$$

Proposition 11 follows immediately from (17) and the next result.

PROPOSITION 12. $\sum_i U_{ni}(X) \leqq \beta(d, a/b)$.

PROOF. Think of $X, X_1, \cdots, X_n$ as fixed points in $\mathbb{R}^d$. Write $\rho_n = \rho_{n,X,X_1,\cdots,X_n}$

and set

$$\rho_{ni} = \rho_{n,X_i,X_1,\cdots,X,\cdots,X_n} \qquad \text{for} \quad 1 \leqq i \leqq n\,.$$

It follows from the definitions of $W_n$ and $U_n$ that $U_{ni}(X) = (c_{n\nu} + \cdots + c_{n,\nu+\lambda-1})/\lambda$, where

$$\nu = 1 + \#\{l\colon 1 \leqq l \leqq n, l \neq i, \text{ and } \rho_{ni}(X_l, X_i) < \rho_{ni}(X, X_i)\}\}$$

and

$$\lambda = 1 + \#\{l\colon 1 \leqq l \leqq n, l \neq i, \text{ and } \rho_{ni}(X_l, X_i) = \rho_{ni}(X, X_i)\}\}\,.$$

Assume first that (7) holds and that $s_{nj} > 0$ for $1 \leqq j \leqq d$. Set $I_0 = \{i\colon 1 \leqq i \leqq n \text{ and } X_i = X\}$ and $t = \#(I_0)$. If $i \in I_0$, then $\nu = 1$ and $\lambda = t$, so that $U_{ni}(X) = (c_{n1} + \cdots + c_{nt})/t$. Thus

$$\sum_{i \in I_0} U_{ni}(X) = \sum_{i=1}^{t} c_{ni}\,. \tag{18}$$

For $1 \leqq i \leqq n$ and $1 \leqq j \leqq d$ define $s_{nji}$ by

$$s_{nji} = s_{nji}(X, X_1, \cdots, X_n) = s_{nj}(X_i, X_1, \cdots, X, \cdots, X_n)\,.$$

Consider the transformations $T, T_1, \cdots, T_n$ from $\mathbb{R}^d$ to itself defined as follows:

$$(Tu)_j = \frac{u_j}{s_{nj}} \qquad \text{and} \qquad (T_i u)_j = \frac{u_j}{s_{nji}} \quad \text{for} \quad 1 \leqq j \leqq d\,,$$

where $u = (u_1, \cdots, u_d)$. Observe that $\rho_n(u, v) = ||Tu - Tv||$ and $\rho_{ni}(u, v) = ||T_i u - T_i v||$. Observe also that $(T_i u)_j = b_{ji}(Tu)_j$, where $b_{ji} = s_{nj}/s_{nji}$. It follows from (7) that $a \leqq b_{ji} \leqq b$ for $1 \leqq j \leqq d$.

Choose $V \in \mathscr{V}(d, a/b)$. Set

$$I = \{i\colon 1 \leqq i \leqq n, X_i \neq X, \text{ and } TX_i - TX \in V\}$$

and $p = \#(I)$. Then $I = \{i_1, \cdots, i_p\}$, where $0 < ||TX_{i_1} - TX|| \leqq \cdots \leqq ||TX_{i_p} - TX||$. Let $1 \leqq q < r \leqq p$. Then $TX_{i_q} - TX \in V$, $TX_{i_r} - TX \in V$, and $0 < ||TX_{i_q} - TX|| \leqq ||TX_{i_r} - TX||$. It now follows from Proposition 10 that $||T_{i_r}X_{i_r} - T_{i_r}X_{i_q}|| < ||T_{i_r}X_{i_r} - T_{i_r}X||$ or equivalently that $\rho_{ni_r}(X_{i_q}, X_{i_r}) < \rho_{ni_r}(X, X_{i_r})$. Thus $U_{ni_r}(X) = (c_{n\nu} + \cdots + c_{n,\nu+\lambda-1})/\lambda$, where $\nu \geqq r$ and $\lambda \geqq t + 1$. Since $c_{n1} \geqq \cdots \geqq c_{nn}$, it follows that $U_{ni_r}(X) \leqq (c_{nr} + \cdots + c_{n,r+t})/(t+1)$ and hence that

$$\sum_{i \in I} U_{ni}(X) \leqq \frac{1}{t+1} \sum_{r=1}^{n} \sum_{m=r}^{r+t} c_{nm}\,.$$

Since $\mathbb{R}^d$ can be covered by $\beta(d, a/b)$ elements $V \in \mathscr{V}(d, a/b)$, it now follows that

$$\sum_{i \notin I_0} U_{ni}(X) \leqq \frac{\beta(d, a/b)}{t+1} \sum_{r=1}^{n} \sum_{m=r}^{r+t} c_{nm}\,. \tag{19}$$

It follows from (18) and (19) and elementary algebra that

$$\sum_i U_{ni}(X) \leqq \sum_{i=1}^{t} c_{ni} + \beta\left(d, \frac{a}{b}\right)\left(\frac{1}{t+1} \sum_{i=1}^{t} i c_{ni} + \sum_{i=t+1}^{n} c_{ni}\right).$$

To verify the inequality of Proposition 12, it is necessary to show that the right side of the above inequality is bounded above by $\beta(d, a/b)$. By elementary algebra and the formula $\sum_{i=1}^{n} c_{ni} = 1$, this reduces to showing that

$$\sum_{i=1}^{t}\left[\beta\left(d, \frac{a}{b}\right)(t+1-i) - (t+1)\right]c_{ni} \geqq 0 .$$

The last inequality follows easily from the observation that $c_{n1} \geqq \cdots \geqq c_{nn} \geqq 0$, $\beta(d, a/b) \geqq 2$, and $\sum_{i=1}^{t}[2(t+1-i)-(t+1)] = 0$. This shows that the inequality of Proposition 12 is valid whenever (7) holds and $s_{nj} > 0$ for $1 \leqq j \leqq d$.

Consider now the general case. Let $J$ denote the collection of all $j$ such that $1 \leqq j \leqq d$, $s_{nj} > 0$, and the $j$th coordinates of $X_1, \cdots, X_n$ do not coincide. Set $\tilde{d} = \#(J)$.

Suppose first that $\tilde{d} = 0$. Then $\rho_{ni}(X_l, X_i) = 0$ for $1 \leqq i, l \leqq d$. It follows easily that $U_{ni}(X) = c_{nn}$ if $\rho_{ni}(X, X_i) > 0$ and $U_{ni}(X) = 1/n$ if $\rho_{ni}(X, X_i) = 0$. In any case $U_{ni}(X) \leqq 1/n$ for $1 \leqq i \leqq n$ and hence $\sum_i U_{ni}(X) \leqq 1 < \beta(d, a/b)$.

Suppose next that $\tilde{d} > 0$. Let $\tilde{\rho}_{ni}$ be the pseudometric obtained by setting

$$\tilde{\rho}_{ni}^2(u, v) = \sum_{j \in J}\left(\frac{u_j - v_j}{s_{nji}}\right)^2 ,$$

where $u = (u_1, \cdots, u_d)$ and $v = (v_1, \cdots, v_d)$. Note that $\tilde{\rho}_{ni}(X_l, X_i) = \rho_{ni}(X_l, X_i)$ and $\tilde{\rho}_{ni}(X, X_i) \leqq \rho_{ni}(X, X_i)$ for $1 \leqq i, l \leqq n$. Set

$$\tilde{\nu} = 1 + \#(\{l\colon 1 \leqq l \leqq n, l \neq i, \text{ and } \tilde{\rho}_{ni}(X_l, X_i) < \tilde{\rho}_{ni}(X, X_i)\})$$

and

$$\tilde{\lambda} = 1 + \#(\{l\colon 1 \leqq l \leqq n, l \neq i, \text{ and } \tilde{\rho}_{ni}(X_l, X_i) = \tilde{\rho}_{ni}(X, X_i)\}) .$$

Then $\tilde{\nu} \leqq \nu$ and $\tilde{\nu} + \tilde{\lambda} \leqq \nu + \lambda$. Set $\tilde{U}_{ni}(X) = (c_{n\tilde{\nu}} + \cdots + c_{n,\tilde{\nu}+\tilde{\lambda}-1})/\tilde{\lambda}$. Then $U_{ni}(X) \leqq \tilde{U}_{ni}(X)$ for $1 \leqq i \leqq n$. Thus

$$\sum_i U_{ni}(X) \leqq \sum_i \tilde{U}_{ni}(X) \leqq \beta\left(\tilde{d}, \frac{a}{b}\right) \leqq \beta\left(d, \frac{a}{b}\right).$$

Thus Proposition 12 holds in general, and hence Proposition 11 is valid.

Proof of Theorem 2. Let $W_n$ be the probability weight function corresponding to $c_n$ and suppose that $\lim_n \sum_{i>\alpha n} c_{ni} = 0$ for all $\alpha > 0$. Proposition 11 implies that the first condition of Corollary 1 holds. To show that the second condition of Corollary 1 holds, choose $a > 0$ and $\varepsilon > 0$. For given $\alpha > 0$ let $A_n$ denote the event that

$$\max_{i \in I_{n,\alpha n}(X)} ||X_i - X|| > a .$$

It follows from Proposition 9 that $\alpha$ can be chosen so that $\limsup_n P(A_n) \leqq \varepsilon$. Now

$$\sum_i W_{ni}(X) I_{\{||X_i - X|| > a\}} \leqq \sum_{i>\alpha n} c_{ni} + I_{A_n} .$$

Since $\varepsilon$ can be made arbitrarily small, the second condition of Corollary 1 is valid. Suppose also that $c_{n1} \to 0$. Since $\max_i W_{ni}(X) \leqq c_{n1}$, the third condition of Corollary 1 holds and hence $\{W_n\}$ is consistent.

**12. Proof of Theorem 3.** In this section Theorem 3 of Section 7 will be proven using the notation of that section. Also, in various proofs, the abbreviated notations $L(X)$ for $L^Y(p \mid X)$, $\hat{L}_n(X)$ for $\hat{L}_n{}^Y(p \mid X)$, etc., will be used.

PROPOSITION 13. *Let $\{W_n\}$ be a consistent sequence of probability weights and let $0 < p < 1$. Then for every $\varepsilon > 0$*

$$\lim_n P(\hat{L}_n{}^Y(p \mid X) \geqq L^Y(p \mid X) - \varepsilon) = 1$$

*and*

$$\lim_n P(\hat{U}_n{}^Y(p \mid X) \leqq U^Y(p \mid X) + \varepsilon) = 1 .$$

PROOF. Only the first result will be proven, the proof of the second result being similar. Define the function $f$ on $\mathbb{R}^d$ by

$$f(X) = P\left(Y \leqq L(X) - \frac{\varepsilon}{2} \,\middle|\, X\right).$$

Then $0 \leqq f < p$. It follows from the consistency of $\{W_n\}$ that

$$\lim_n E|\textstyle\sum_i W_{ni}(X) I_{\{Y_i \leqq L(X_i) - \varepsilon/2\}} - f(X)| = 0$$

and hence that

$$\lim_n P\left(\textstyle\sum_i W_{ni}(X) I_{\{Y_i \leqq L(X_i) - \varepsilon/2\}} \geqq \dfrac{p + f(X)}{2}\right) = 0 . \tag{20}$$

It follows from Proposition 4 that

$$\textstyle\sum_i W_{ni}(X) I_{\{L(X_i) \leqq L(X) - \varepsilon/2\}} \to 0 \quad \text{in probability}$$

and hence that

$$\lim_n P\left(\textstyle\sum_i W_{ni}(X) I_{\{L(X_i) \leqq L(X) - \varepsilon/2\}} \geqq \dfrac{p - f(X)}{2}\right) = 0 . \tag{21}$$

Equations (20) and (21) together imply that

$$\lim_n P(\textstyle\sum_i W_{ni}(X) I_{\{Y_i \leqq L(X) - \varepsilon\}} < p) = 1$$

and hence that $\lim_n P(\hat{L}_n(X) \geqq L(X) - \varepsilon) = 1$. Thus the first result of Proposition 13 is valid, as desired.

PROPOSITION 14. *Let $0 < p < 1$ and $r > 1$. Then*

$$E|L^Y(p \mid X)|^r \leqq \frac{E|Y|^r}{p \wedge (1 - p)}$$

*and*

$$E|U^Y(p \mid X)|^r \leqq \frac{E|Y|^r}{p \wedge (1 - p)} .$$

PROOF. Only the first result will be proven, the proof of the second result being similar. If $L(X) \leqq 0$, then $E(|Y|^r \mid X) \geqq p|L(X)|^r$. If $L(X) \geqq 0$, then $E(|Y|^r \mid X) \geqq (1 - p)|L(X)|^r$. Thus in general $E(|Y|^r \mid X) \geqq p \wedge (1 - p)|L(X)|^r$ and hence

$$E|L(X)|^r \leqq \frac{EE(|Y|^r \mid X)}{p \wedge (1 - p)} = \frac{E|Y|^r}{p \wedge (1 - p)}$$

as desired.

PROPOSITION 15. *Let $W_n$ be a probability weight function satisfying* (1) *and let* $0 < p < 1$ *and* $M > 0$. *Then*

$$E|\hat{L}_n^Y(p \mid X)|^r I_{\{|\hat{L}_n^Y(p|X)| \geqq M\}} \leqq \frac{C}{p \wedge (1-p)} E|Y|^r I_{\{|Y| \geqq M\}}$$

*and the same inequality holds with* $\hat{L}_n^Y(p \mid X)$ *replaced by* $\hat{U}_n^Y(p \mid X)$.

PROOF. It is easily seen that

$$|\hat{L}_n(X)|^r I_{\{|\hat{L}_n(X)| \geqq M\}} \leqq \frac{1}{p \wedge (1-p)} \sum_i W_{ni}(X)|Y_i|^r I_{\{|Y_i| \geqq M\}} .$$

Thus by (1)

$$E|\hat{L}_n(X)|^r I_{\{|\hat{L}_n(X)| \geqq M\}} \leqq \frac{C}{p \wedge (1-p)} E|Y|^r I_{\{|Y| \geqq M\}} .$$

The same argument works if $\hat{L}_n(X)$ is replaced by $\hat{U}_n(X)$.

PROOF OF THEOREM 3. Let $\{W_n\}$ be a consistent sequence of probability weights and let $0 < p < 1$. It follows from Proposition 13 that (9) and (10) hold. It now follows from Propositions 14 and 15 that if $r \geqq 1$ and $E|Y|^r < \infty$, then in (9) and (10) convergence in probability can be replaced by convergence in $L^r$. This completes the proof of Theorem 3.

**13. Proof of Theorem 4.** When applied to Model 1, Theorem 4 follows immediately from the consistency of $\{W_n\}$ and the formula for the Bayes risk of $\hat{\delta}_n$ given in the discussion of Model 1. When applied to Model 2, Theorem 4 follows immediately from Theorem 3 and the inequality

$$\begin{aligned} &|\mathscr{L}(Y, \hat{\delta}_n(X)) - \mathscr{L}(Y, \delta(X))| \\ &\qquad \leqq c(1-p)(\hat{L}_n^Y(p \mid X) - L^Y(p \mid X))^- + cp(\hat{U}_n^Y(p \mid X) - U^Y(p \mid X))^+ . \end{aligned}$$

Consider now Model 3 and let $\{W_n\}$ be a consistent sequence of weights. Set

$$\begin{aligned} e_n(X) &= \max_y |\hat{E}_n(\mathscr{L}(Y, y) \mid X) - E(\mathscr{L}(Y, y) \mid X)| \\ &= \max_y |\hat{P}_n^Y(\{y\} \mid X) - P^Y(\{y\} \mid X)| . \end{aligned}$$

It will be shown that

$$\lim_n Ee_n(X) = 0 . \tag{22}$$

Observe that

$$\begin{aligned} E(\mathscr{L}(Y, \hat{\delta}_n(X)) \mid X) &\leqq \hat{E}_n(\mathscr{L}(Y, \hat{\delta}_n(X)) \mid X) + e_n(X) \\ &\leqq \hat{E}_n(\mathscr{L}(Y, \delta(X)) \mid X) + e_n(X) \\ &\leqq E(\mathscr{L}(Y, \delta(X)) \mid X) + 2e_n(X) \end{aligned}$$

and hence that

$$E\mathscr{L}(Y, \hat{\delta}_n(X)) \leqq R + 2Ee_n(X) .$$

Thus it follows from (22) that $\{\hat{\delta}_n\}$ is consistent in Bayes risk.

In the important special case that the distribution of $Y$ has finite support, (22) follows immediately from the consistency of $\{W_n\}$. To prove the result in general set $A_1 = \{y: P(Y = y) = 0\}$. Then with probability one, no value of $y \in A_1$ occurs more than once among $Y_1, Y_2, \cdots$ and hence

$$\max_{y \in A_1} |\hat{P}_n^Y(\{y\} \mid X) \leqq \max_i |W_{ni}(X)| .$$

It now follows from (2) and (5) that

$$\lim_n E \max_{y \in A_1} |\hat{P}_n^Y\{y\} \mid X) - P^Y(\{y\} \mid X)| = 0 .$$

Set $A_2 = \{y: P(Y = y) > 0\}$. Choose $\varepsilon > 0$ and let $A_3$ and $A_4$ be disjoint sets whose union is $A_2$ and such that $A_3$ is finite and $P(Y \in A_4) \leqq \varepsilon$. It follows from the consistency of $\{W_n\}$ that

$$\lim_n E \max_{y \in A_3} |\hat{P}_n(\{y\} \mid X) - P^Y(\{y\} \mid X)| = 0 .$$

Clearly

$$E \max_{y \in A_4} P^Y(\{y\} \mid X) \leqq \varepsilon .$$

It follows from (1) that

$$E \max_{y \in A_4} |\hat{P}_n(\{y\} \mid X)| \leqq C\varepsilon .$$

Since $\varepsilon$ can be made arbitrarily small, (22) follows from the last four displayed results. This completes the proof of Theorem 4.

## REFERENCES


[1] Adichie, J. N. (1967). Estimates of regression parameters based on rank tests. *Ann. Math. Statist.* **38** 894–904.

[2] Aizerman, M. A., Braverman, E. and Rozonoer, L. (1970). Extrapolative problems in automatic control and the method of potential functions. *Amer. Math. Soc. Transl.* **87** 281–303.

[3] Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16** 147–192.

[4] Benedetti, J. (1974). Kernel estimation of regression functions. Ph. D. dissertation, Univ. of Washington.

[5] Benedetti, J. (1975). Kernel estimation of regression functions. *Proc. of Computer Science and Statistics*: 8th *Annual Symposium on the Interface* 405–408. Health Science Computing Facility, UCLA.

[6] Bickel, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *Ann. Statist.* **1** 597–616.

[7] Breiman, L. and Meisel, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71** 301–307.

[8] Butler, G. A. (1975). Heuristic regression for large commercial problems. *Proc. of Computer Science and Statistics*: 8th *Annual Symposium on the Interface* 398–404. Health Science Computing Facility, UCLA.

[9] Cleveland, W. S. and Kleiner, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. *Technometrics* **17** 447–454.

[10] Cover, T. M. (1968). Estimation by the nearest neighbor rule. *IEEE Trans. Information Theory* **IT-14** 50–55.

[11] Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Information Theory* **IT-13** 21–27.

[12] Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York.

[13] Fisher, L. and Yakowitz, S. (1976). Uniform convergence of the potential function algorithm. *SIAM J. Control* **14** 95–103.

[14] Fix, E. and Hodges, J. L., Jr. (1951). Discriminatory analysis, nonparametric discrimination, consistency properties. Randolph Field, Texas, Project 21-49-004, Report No. 4.

[15] Friedman, J. H. (1976). A variable metric decision rule for nonparametric classification. *IEEE Trans. Comput.*, to appear.

[16] Fritz, J. (1975). Distribution-free exponential error bound for nearest neighbor pattern classification. *IEEE Trans. Information Theory* **IT-21** 552–557.

[17] Gordon, L. and Olshen, R. A. (1975). Asymptotically efficient, computationally feasible solutions to the classification problem. Unpublished manuscript.

[18] Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43** 1449–1458.

[19] Jurečková, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42** 1328–1338.

[20] Liggett, T. M. (1977). Extensions of the Erdös-Ko-Rado theorem and a statistical application. *J. Combinatorial Theory* (A) **22**, No. 3.

[21] Major, P. (1973). On non-parametric estimation of the regression function. *Studia Sci. Math. Hungar.* **8** 347-361.

[22] Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* **58** 415–434.

[23] Nadaraya, E. A. (1964). On estimating regression. *Theor. Probability Appl.* **9** 141–142.

[24] Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theor. Probability Appl.* **15** 134–137.

[25] Parzen, E. (1962). On estimation of a probability density and mode. *Ann. Math. Statist.* **35** 1065–1076.

[26] Priestley, M. B. and Chao, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392.

[27] Raman, S. (1971). Contribution to the theory of Fourier estimation of multivariate probability density functions with application to data on bone age determinations. Ph. D. dissertation, Univ. of California, Berkeley.

[28] Révész, P. (1973). Robbins–Munro procedure in a Hilbert space and its application in the theory of learning processes. I. *Studia Sci. Math. Hungar.* **8** 391–398.

[29] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 642–669.

[30] Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis* II 25–31, Academic Press, New York.

[31] Royall, R. M. (1966). A class of nonparametric estimators of a smooth regression function. Ph. D. dissertation, Stanford Univ.

[32] Schuster, E. F. (1968). Estimation of a probability density function with applications in statistical inference. Ph. D. dissertation, Univ. of Arizona.

[33] Schuster, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Ann. Math. Statist.* **43** 84–88.

[34] Sonquist, J. A. and Morgan, J. N. (1964). *The Detection of Interaction Effects.* Survey Research Center Monograph No. 35, Institute for Social Research, Univ. of Michigan, Ann Arbor.

[35] Stone, C. J. (1975). Nearest neighbor estimators of a nonlinear regression function. *Proc. of Computer Science and Statistics*: 8th *Annual Symposium on the Interface* 413–418. Health Sciences Computer Facility, UCLA.

[36] Van Ryzin, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhyā Ser. A* **28** 261–270.

[37] Wahba, G. and Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Comm. Statist.* **6** 1–17.

[38] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

[39] Wold, S. (1974). Spline functions in data analysis. *Technometrics* **16** 1–11.

[40] YAKOWITZ, S. and FISHER, L. (1975). Experiments and developments on the method of potential functions. *Proc. of Computer Science and Statistics*: 8th *Annual Symposium on the Interface* 419–423. Health Science Computer Facility, UCLA.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90024

## DISCUSSION

PETER J. BICKEL
*University of California at Berkeley*

As Professor Stone has pointed out, over the years a large variety of methods have been proposed for the estimation of various features of the conditional distributions of $Y$ given $X$ on the basis of a sample $(X_1, Y_1), \cdots, (X_n, Y_n)$. The asymptotic consistency of these methods has always been subject to a load of regularity conditions. In this elegant paper, Professor Stone has given a unified treatment of consistency under what seem to be natural necessary as well as sufficient conditions.

His work really reveals the essentials of the problem. He has been able to do this by defining the notion of consistency properly from a mathematical point of view in terms of $L_r$ convergence. However, the notions of convergence that would seem most interesting practically are pointwise notions. An example is uniform convergence on $(x, y)$ compacts of the conditional density of $Y$ given $X = x$. The study of this convergence necessarily involves more regularity conditions. At the very least there must be a natural, unique choice of the conditional density. However, such a study and its successors, studies of speed of asymptotic convergence, asymptotic normality of the estimates of the density at a point, asymptotic behavior of the maximum deviation of the estimated density from its limit (see [1] for the marginal case), etc., would seem necessary to me and to Professor Stone too! (He informed me, when I raised this question at a lecture he recently gave in Berkeley, that a student of his had started work on such questions.)

One important question that could be approached by such a study is, how much is lost by using a nonparametric method over an efficient parametric one? If density estimation is a guide, the efficiency would be 0 at the parametric model for any of the nonparametric methods surveyed by Professor Stone. However, even if this is the case, it seems clear that one can construct methods which are asymptotically efficient under any given parametric model and are generally consistent in Stone's sense. This could be done by forming a convex combination of the best parametric and a nonparametric estimate, with weights