The Annals of Statistics
1980, Vol. 8, No. 2, 240–246

CONSISTENT WINDOW ESTIMATION IN NONPARAMETRIC REGRESSION

BY C. SPIEGELMAN[1] AND J. SACKS[2, 3]

Florida State University and Northwestern University

Stone proved that nearest neighbor estimates of a nonparametric regression function are "universally" consistent. We show that the same holds for window estimates, and we obtain a rate of convergence under some restrictions.

1. Introduction. Let (θ, X) be a random pair such that $E\theta^2 < \infty$. For any estimate $\delta(X)$ of θ the statistician incurs loss $L(\theta, \delta) = (\theta - \delta)^2$. The risk to the statistician is defined by $R(\delta) = EL(\theta, \delta)$. It is well known that the Bayes estimator $\delta^*(X) = E[\theta|X]$ has the property that $R(\delta^*) = \min_\delta R(\delta)$.

The form of the joint distribution of (θ, X) is often not known, even approximately. However, from prior experience, independent and identically distributed random variables (θ_i, X_i), $i = 1, \cdots, n$, having the same distribution as (θ, X) are available. The usual maximum likelihood procedures for estimating δ^* are inappropriate when the joint distribution of (δ, X) does not have a finite number of unknown parameters. Similar comments apply to ordinary least square procedures. A number of authors show that certain nonparametric estimates, δ_n, of δ^* have the property (L_2-consistency) that $E(\delta_n - \delta^*)^2 \to 0$ as $n \to \infty$. The most general work is by Stone (1977) and references to other work may be found therein. In particular, Stone shows that, under minimal assumptions, nearest neighbor estimates have this L_2-consistency property and raises the question about whether kernel estimates behave similarly. We show here that window estimates do behave similarly (Theorem 1 in Section 2) and note that the same is true for a class of kernel estimates.

More generally, Stone treats the class of estimates $\delta_n = \sum_{i=1} W_{ni}\theta_i$, where $W_{n1}, \cdots, W_{nn}$ are termed "weights" and gives, in Theorem 1 of his paper, conditions on $\{W_{ni}\}$ (these are restated below in (2.1)) under which δ_n is L_r-consistent for all $r \geqslant 1$. (The sequence of estimates $\{\delta_n\}$ is L_r-consistent if, whenever $E|\theta|^r < \infty$, $\lim_{n\to\infty} E(|\delta_n - \delta^*|^r) = 0$.) If $\{\delta_n\}$ is L_r-consistent for all $r \geqslant 1$ regardless of the distribution of X (the conditions stated in (2.1) depend on the distribution of X) then $\{\delta_n\}$ or $\{W_{ni}\}$ is said (by Stone) to be universally consistent. Stone

Received July 1977; revised May 1979.

[1]Presently at the National Bureau of Standards, Gaithersburg, MD.

[2]Research partially supported by NSF Grant MCS 77-01657.

[3]Presently at the Department of Statistics, Rutgers University.

AMS 1970 *subject classifications.* Primary 6215; secondary 6240.
Key words and phrases. Empirical Bayes, nonlinear regression.

shows that the weights defining nearest neighbor estimates are universally consistent. If $\{b_n\}$ is a sequence of positive numbers with $b_n \to 0$ and if

$$K_{ni} = 1, \quad \text{if } \|X_i - X\| \leqslant b_n$$

$$= 0, \quad \text{otherwise,} \tag{1.1}$$

$$K_n^* = \max\{0, \Sigma_1^n K_{ni}\},$$

where $\|\cdot\|$ is a norm on R^d, then $W_{ni} = K_{ni}/K_n^*$ defines the weights for a window estimate of δ^*. We will show, under a natural restriction on b_n, that K_{ni}/K_n^* satisfies the conditions of Stone's Theorem 1 for any probability distribution on X, thereby establishing the universal consistency of window estimates (see Theorem 1 in Section 2 below). In a comment at the end of Section 2 we point out how this can be extended to a class of kernel estimates. In Theorem 2 a rate of convergence is provided when δ^* satisfies a Lipschitz condition and the distribution of X has compact support.

When this paper was originally submitted, a theorem more special than Theorem 1 was stated. During revision the improvement to the present Theorem 1 was obtained. Independently, Devroye and Wagner (1980) obtained the same result for a class of kernel estimates. Upon learning of the Devroye-Wagner result we were able to extend Theorem 1, as indicated in the Remark at the end of Section 2, to the same class of kernel estimates. There is some similarity between our proof and that in Devroye-Wagner. By relying on Stone's result our proof is somewhat more concise. Both proofs depend on the covering number of a sphere which we define as follows:

DEFINITION. If $\|\cdot\|$ is a norm on R^d and S is a compact set, let $C(S, \rho)$ be the minimum number of closed balls of radius ρ needed to cover S. If S is a closed ball of radius 1 then we use the shorter notation $C(\rho)$ to denote the covering number of S.

Similar covering numbers play a crucial role in Stone's verification of his conditions for nearest neighbor estimates.

2. Results. Let $\delta_n(X) = \Sigma_1^n W_{ni}\theta_i$ where $W_{ni} = K_{ni}/K_n^*$ with K_{ni} and K_n^* defined in (1.1). Our goal is to show that the conditions of Theorem 1 of Stone (1977) are satisfied by W_{ni}. These conditions are:

(a) For some C_1, $E[\Sigma_{i=1}^n W_{ni} g(X_i)] \leqslant C_1 Eg(X)$ for all nonnegative g.

(b) For some D, $P(\Sigma_1^n |W_{ni}| \leqslant D) = 1$, for all n.

(2.1) (c) $\Sigma |W_{ni}| I_{\{\|X_i - X\| > a\}} \to 0$ in probability for each $a > 0$ (I_B = indicator of the set B.)

(d) $\Sigma W_{ni} \to 1$ in probability.

(e) $\sup_i |W_{ni}| \to 0$ in probability.

It is trivial that (b) and (c) are satisfied. It remains to establish (a), (d) and (e).

LEMMA 1. *Let* $\varphi(x, y) \geqslant 0$ *for* $x, y \in R^d$. *Let*

$$D_n(x) = E[\varphi(X_1, x)| \, |X_1 - x| \leqslant b_n].$$

Then,

$$E[\Sigma_1^n \varphi(X_i, X) W_{ni}] \leqslant E[D_n(X)].$$

PROOF. Since $X, X_1, \cdots, X_n$ are i.i.d., and W_{ni} depends on X_i only through K_{ni}, we obtain

$$E[\varphi(X_i, X) W_{ni} | X, X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n, K_{ni}]$$

$$= 0, \quad \text{if } K_{ni} = 0 \tag{2.2}$$

$$= W_{ni} E[\varphi(X_i, x) | K_{ni} = 1, X = x], \quad \text{if } K_{ni} = 1, X = x.$$

Since $W_{ni} = 0$ if $K_{ni} = 0$ and $E[\varphi(X_i, x)|K_{ni} = 1, X = x]$ doesn't depend on i, the right side of (2.2) equals $W_{ni} D_n(x)$ if $X = x$. Thus

$$E[\varphi(X_i, X) W_{ni}] = E[W_{ni} D_n(X)],$$

and summing produces

$$E[\Sigma \varphi(X_i, X) W_{ni}] = E[\Sigma W_{ni} D_n(X)]$$

$$\leqslant E D_n(x).$$

LEMMA 2. *Let* μ *be a probability measure on* R^d (μ *is the distribution of* X). *Let* $\| \cdot \|$ *be a norm on* R^d *such that all closed bounded balls are compact. Let* $S(a, b)$ *be the ball with center at* a *and radius* b. *Then*

$$\sup_{\mu, y \in R^d, b>0} \int_{\|u-y\| \leqslant b} \frac{1}{\mu[S(u, b)]} \mu(du) \leqslant C\left(\tfrac{1}{2}\right)$$

where $C(1/2)$ *is defined in Section* 1.

PROOF. It is easy to see that, if $S = S(a, b)$, then $C(S, b/2) = C(\frac{1}{2})$. Let $S(x_1, b/2), \cdots, S(x_c, b/2)$ ($c = C(\frac{1}{2})$) be balls which cover $S(y, b)$. If $u \in S(x_i, b/2)$ then $S(x_i, b/2) \subset S(u, b)$. Hence

$$\int_{S(y, b)} \frac{1}{\mu[S(u, b)]} \mu(du) \leqslant \sum_{i=1}^{C(\frac{1}{2})} \int_{S(x_i, b/2)} \frac{\mu(du)}{\mu[S(u, b)]}$$

$$\leqslant \sum_{i=1}^{C(\frac{1}{2})} \int_{S(x_i, b/2)} \frac{\mu(du)}{\mu[S(x_i, b/2)]}$$

$$= C\left(\tfrac{1}{2}\right)$$

and Lemma 2 is proved.

REMARK. If $\|z\|$ is $\sup_{1 \leqslant i \leqslant d} |z_i|$ where $z = (z_1, \cdots, z_d)$ (so balls are rectangles in d-space) then $C(\frac{1}{2}) = 2^d$.

LEMMA 3. (a) *If* $\| \cdot \|$ *is equivalent to the Euclidean norm then* $C(b_n) = 0(b_n^{-d})$. (b) *In general,*

$$C(b_n) = 0\left(b_n^{-\log C(\frac{1}{2})/\log 2}\right).$$

PROOF. By dealing with rectangles (a) is easy to get. In the general case, use $C(2^{-k}) \leqslant [C(\frac{1}{2})]^k$.

THEOREM 1. *If*

$$\lim_{n\to\infty} \frac{C(b_n)}{n} = 0 \tag{2.3}$$

then $\{K_{ni}/K_n^*\}$ *is universally consistent. Thus, by Lemma* 3, *if* $\|\cdot\|$ *is equivalent to the Euclidean norm, then* $nb_n^d \to \infty$ *implies that the window weights are universally consistent.*

PROOF. As noted after (2.1), we have to show that (a), (d) and (e) of (2.1) hold. To obtain (a) apply Lemma 1 to $\varphi(x_i, x) = g(x_i)$ and obtain, with μ denoting the distribution of X,

$$\begin{aligned} E\left[\Sigma_1^n W_{ni} g(X_i)\right] &\leqslant E\left[D_n(X)\right] \\ &= \int \mu(dx) \int_{S(x, b_n)} \frac{g(u)\mu(du)}{\mu\left[S(x, b_n)\right]}, \\ &= \int g(u)\mu(du) \int_{S(u, b_n)} \frac{\mu(dx)}{\mu\left[S(x, b_n)\right]}. \end{aligned}$$

Now use Lemma 2 and bound the last expression by $C(\frac{1}{2})Eg(X)$ which establishes (a) of (2.1).

To show that (d) and (e) hold it is enough to show that $K_n^* \to \infty$ in probability. Let $\pi_n(x) = P[|X_1 - x| \leqslant b_n]$ and note that, given $X = x$, K_n^* is binomial with parameters n, $\pi_n(x)$. Then, if B is a positive number $\leqslant \frac{1}{2} n\pi_n(x)$, use Chebyshev's inequality and get

$$\begin{aligned} P\left[K_n^* \leqslant B | X = x\right] &= P\left[K_n^* - n\pi_n(x) \leqslant B - n\pi_n(x) | X = x\right] \\ &\leqslant \frac{1}{n\pi_n(x)}. \end{aligned} \tag{2.4}$$

Let A be a compact subset of the support of μ with $\mu(A) > 0$. Then

$$\begin{aligned} P\left[K_n^* \leqslant B, X \in A\right] &\leqslant \int_{A \cap \{n\pi_n(x)/2 > B\}} \frac{1}{n\pi_n(x)} \mu(dx) + \mu\left[A \cap \{n\pi_n(X) < 2B\}\right] \\ &\leqslant \frac{1}{2B} + \int_{A \cap \{n\pi_n(x) < 2B\}} \frac{B}{2n\pi_n(x)} \mu(dx) \\ &\leqslant \frac{1}{2B} + \frac{B}{2n} \int_A \frac{1}{\pi_n(x)} \mu(dx). \end{aligned} \tag{2.5}$$

Let $x_1 \in R^d$. Then, from Lemma 2,

$$\int_{S(x_1, b_n)} \frac{1}{\pi_n(x)} \mu(dx) \leqslant C\left(\tfrac{1}{2}\right).$$

Since A is compact, A can be covered by $C(A, b_n)$ balls of the type $S(x_1, b_n)$ with $x_1 \in R^d$. So the last term on the right side of (2.5) is bounded by

$$\frac{B}{2n} C(A, b_n) C\left(\tfrac{1}{2}\right).$$

Since A is compact, $C(A, b_n) = 0(C(b_n))$, and therefore

$$P[K_n^* \leqslant B, X \in A] \leqslant \frac{1}{2B} + \gamma(A) B \frac{C(b_n)}{n}. \tag{2.6}$$

Since $C(b_n)/n$ is assumed to go to 0 we get

$$\limsup_{n\to\infty} P[K_n^* \leqslant B, X \in A] \leqslant \frac{1}{2B} \tag{2.7}$$

and this is enough to show that $K_n^* \to \infty$ in probability. The theorem is proved.

THEOREM 2. *Let* $\delta(x) = E[\theta|X)]$. *Assume*
(a) $|\delta(x) - \delta(z)| \leqslant B_1\|x - z\|$ *where* $\|\cdot\|$ *is equivalent to the Euclidean norm,*
(b) μ *has compact support,*
(c) $b_n = n^{-1/(2+d)}$,
(d) $E[|\theta - \delta(X)|^2|X] \leqslant \sigma^2 < \infty$.
Let $\delta_n = \Sigma_{i=1}^n (K_{ni}/K_n^*)\theta_i$. *Then*

$$E[\delta_n - \delta(X)]^2 = 0(n^{-2/(2+d)}).$$

PROOF. It is straightforward to obtain

$$\begin{aligned} E[\delta_n - \delta(X)]^2 = {} & E[\Sigma(\theta_i - \delta(X_i))W_{ni}]^2 \\ & + E[\Sigma(\delta(X_i) - \delta(X))W_{ni}]^2 \\ & + E\delta^2(X)(1 - \Sigma W_{ni}). \end{aligned} \tag{2.8}$$

From (a) and (b) we conclude that δ is bounded on the support of μ so that

$$\begin{aligned} E\delta^2(X)(1 - \Sigma_1^n W_{ni}) &= 0(1)P[K_n^* = 1] \\ &= 0(1)E\left[\frac{1}{K_n^*}\right]. \end{aligned} \tag{2.9}$$

From (a) and Cauchy-Schwarz, we get

$$\begin{aligned} E(\Sigma(\delta(X_i) - \delta(X))W_{ni})^2 &\leqslant E[\Sigma(\delta(X_i) - \delta(X))^2 W_{ni}] \\ &\leqslant B_1 E[\Sigma_1^n \|X_i - X\|^2 W_{ni}]. \end{aligned}$$

Now use Lemmas 1 and 2 and bound the last term by

$$B_1 \int \mu(dx_1) \int_{S(x_1, b_n)} \frac{\|x_1 - x\|^2}{\pi_n(x)} \mu(dx)$$

$$= 0(b_n^2) \int \mu(dx_1) \int_{S(x_1, b_n)} \frac{\mu(dx)}{\pi_n(x)} \tag{2.10}$$

$$= 0(b_n^2) = 0(n^{-2/2+d}).$$

Next we note, with the help of (d), that

$$E(\Sigma(\theta_i - \delta(X_i)) W_{ni})^2 = E\left[\Sigma(\theta_i - \delta(X_i))^2 W_{ni}^2\right]$$

$$\leqslant \sigma^2 E\left[\Sigma_1^n W_{ni}^2\right] \leqslant \sigma^2 E\left[\frac{1}{K_n^*}\right].$$

Let $b(j, \pi_n(x), n)$ be the binomial probability of j heads in n tosses with probability of heads $= \pi_n(x)$ and get

$$E\left[\frac{1}{K_n^*} \middle| X = x\right] = b(0, \pi_n(x), n) + \Sigma_{j=1}^n \frac{1}{j} b(j, \pi_n(x), n) \tag{2.11}$$

$$\leqslant \Sigma_{j < n\pi_n(x)/2} b(j, \pi_n(x), n) + \frac{2}{n\pi_n(x)}.$$

The first term on the right side of (2.11) is bounded by $1/n\pi_n(x)$ as shown at (2.4) (take $B = n\pi_n(x)/2$). Therefore

$$E\frac{1}{K_n^*} \leqslant \frac{3}{n} \int \frac{\mu(dx)}{\pi_n(x)}. \tag{2.12}$$

If μ has compact support then the argument following (2.5) shows that the second term on the right side of (2.12) is

$$0(C(b_n)/n) = 0(n^{-1} b_n^{-d}) = 0(n^{-2/(2+d)}).$$

Then (2.9), (2.10) and this last result establish Theorem 2.

REMARK. Theorems 1 and 2 can be extended to some kernel weights. Let J be a function on $[0, \infty)$ with $0 \leqslant J(t) \leqslant B$, $J = 0$ outside a compact set, $J(t) \geqslant \alpha > 0$ on $[0, t_0]$. For simplicity assume $B = 1$ and the compact set is $[0, 1]$. Define $J_{ni} = J(\|x_i - x\|/b_n)$, $J_n^* = \max(1, \Sigma J_{ni})$. The only difficulty in extending the argument to the weights J_{ni}/J_n^* lies in confirming (a) of (2.1). To do this, note, as in Lemma 1, that

(2.13)

$$E\left[\frac{J_{ni}}{J_n^*} g(X_i) \middle| X, X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n, W_{ni}\right] \leqslant \frac{K_{ni}}{K_{ni} + \Sigma_{s \neq i} J_{ns}} D_n(X).$$

Also note that

$$(2.14)\quad E\left[K_{ni}\frac{1}{K_{ni}+\Sigma_{s\neq i}J_{ns}}\middle|X=x\right]\leqslant E[K_{ni}|X=x]E\left[\frac{1}{\Sigma_{s\neq i}J_{ns}}\middle|X=x\right].$$

Let $q_n(x)=P[|X_1-x|\leqslant t_0b_n]$. As before let $E[K_{ni}|X=x]=\pi_n(x)$. Argue as in (2.11) to conclude that

$$(2.15)\quad E\left[\frac{1}{\Sigma_{s\neq i}J_{ns}}\middle|X=x\right]=0\left(\frac{1}{nq_n(x)}\right).$$

Combine (2.13), (2.14), and (2.15) and get

$$(2.16)\quad E\left[\frac{J_{ni}}{J_n^*}g(X_i)\middle|X=x\right]=0(1)D_n(x)\frac{\pi_n(x)}{nq_n(x)}.$$

Add up and obtain, as at the beginning of the proof of Theorem 1,

$$(2.17)\quad E\left[\Sigma\frac{J_{ni}}{J_n^*}g(X_i)\right]=0(1)E\left[D_n(X)\frac{\pi_n(X)}{q_n(X)}\right]$$

$$=0(1)\int g(u)\mu(du)\int_{S(u,\,b_n)}\frac{\pi_n(x)}{q_n(x)}\frac{\mu(dx)}{\pi_n(x)}.$$

Since $S(u, b_n)$ can be covered by $C(t_0)$ spheres of radius t_0b_n the inner integral in (2.17) is bounded by $C(t_0)C(\frac{1}{2})Eg(X)0(1)$, where, as in the proof of Theorem 1, Lemma 2 is used. This confirms (a) of (2.1).

REFERENCES

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

[2] STONE, C. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.

DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306

DEPARTMENT OF MATHEMATICS
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60201