6.3 Stone's Consistency Theorem

A stunning result regarding the universal consistency of such a weighted average estimator of $\eta(x)$ was proved by Stone, C., in his paper 'Consistent nonparametric regression' (1977).

Theorem 19 (Stone's consistency theorem)**.** *Assume that for any distribution of X, the weights satisfy the following three conditions:*

(i) *There is a constant C such that for every non-negative measurable function f satisfying $E\,f(X) < \infty$,*

$$E\left\{\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right\} \leq C\,E\,f(X),$$

(ii) *For all $a > 0$, $\lim_{n\to\infty} E\{\sum_{i=1}^{n} W_{ni}(X) I(\|X_i - X\| > a)\} = 0$, and*

(iii) $\lim_{n\to\infty} E\left\{\max_{1\leq i\leq n} W_{ni}(X)\right\} = 0.$

Then, $f_n(X) = I(\eta_n(X) > 1/2)$ is ***universally consistent****. In other words,*

$$E_{\mathcal{D}_n} R(f_n) \to R^* \qquad \text{as } n \to \infty.$$

Remark 12. Condition (ii) says that the expected total weight of X_i outside $B_a(X)$ must go to zero. So only those points in a shrinking neighborhood of X should be counted for the averaging. Condition (iii) means that no single X_i has too large a contribution to the estimate.

Proof. The expectations in the above three conditions of the theorem are w.r.t $\mathcal{D}_n$ and the new test point (X, Y). Previously, we have proved that for a plug-in decision rule

$$R(f_n) - R^* \leq 2\,E_X|\eta_n(X) - \eta(X)|.$$

By taking expectation and using Jensen's inequality,

$$E_{\mathcal{D}_n} R(f_n) - R^* \leq 2\,E_{X,\mathcal{D}_n}|\eta_n(X) - \eta(X)| \leq 2\sqrt{E_{X,\mathcal{D}_n}(\eta_n(X) - \eta(X))^2}.$$

Thus, for the consistency of f_n, it is sufficient to prove that

$$E_{X,\mathcal{D}_n}(\eta_n(X) - \eta(X))^2 \to 0.$$

Let $\hat{\eta}_n(x) = \sum_{i=1}^{n} \eta(x_i) W_{ni}(x)$. Then, using the inequality $(a+b)^2 \leq 2(a^2+b^2)$,

$$\begin{aligned} E\left(\eta_n(X) - \eta(X)\right)^2 &= E\left(\eta_n(X) - \hat{\eta}_n(X) + \hat{\eta}_n(X) - \eta(X)\right)^2 \\ &\leq 2\left[E(\eta_n(X) - \hat{\eta}_n(X))^2 + E(\hat{\eta}_n(X) - \eta(X))^2\right]. \end{aligned}$$

Let $T_1 := (\eta_n(X) - \hat{\eta}_n(X))^2$ and $T_2 := (\hat{\eta}_n(X) - \eta(X))^2$. Then

$$\begin{aligned} T_2 &= \left(\sum_{i=1}^{n} W_{ni}(X)\eta(X_i) - \sum_{i=1}^{n} W_{ni}(X)\eta(X)\right)^2 \\ &= \left(\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta(X_i))\right)^2 \\ &\leq \sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta(X_i))^2 \qquad \text{by Jensen's inequality.} \end{aligned}$$

For $\eta(x)$, we can always find $\eta^*(x)$ which is continuous and has a bounded support such that

$$E\left(\eta(X) - \eta^*(X)\right)^2 \leq \epsilon \tag{6.4}$$

as the set of continuous functions with a bounded support is dense in $L_2(dP_X)$. Then, $\eta^*(X)$ is uniformly continuous, i.e., for any $\epsilon > 0, \exists\, a > 0$ such that

$$\|X_1 - X\| \leq a \Rightarrow (\eta^*(X) - \eta^*(X_1))^2 < \epsilon. \tag{6.5}$$

$$\begin{aligned}
&E\left(\sum_{i=1}^n W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2\right)\\
&= E\left(\sum_{i=1}^n W_{ni}(X) I(\|X_i - X\| > a)(\eta^*(X) - \eta^*(X_i))^2\right)\\
&\qquad + E\left(\sum_{i=1}^n W_{ni}(X) I(\|X_i - X\| \leq a)(\eta^*(X) - \eta^*(X_i))^2\right)\\
&\leq E\left(\sum W_{ni}(X) I(\|X_i - X\| > a)\right) + \epsilon\, E\left(\sum W_{ni}\right)\\
&\qquad\qquad \text{using (6.5) and } |\eta^*(X) - \eta^*(X_i)| \leq 1\\
&= E\left(\sum W_{ni}(X) I(\|X_i - X\| > a)\right) + \epsilon\\
&\to \epsilon \;\text{ as } n \to \infty \qquad \text{by the assumption (ii).}
\end{aligned} \tag{6.6}$$

Therefore,

$$\begin{aligned}
&E\left(\sum W_{ni}(X)(\eta(X) - \eta(X_i))^2\right)\\
&= E\left(\sum W_{ni}(X)(\eta(X) - \eta^*(X) + \eta^*(X) - \eta^*(X_i) + \eta^*(X_i) - \eta(X_i))^2\right)\\
&\leq 3\, E\left(\sum W_{ni}(X)\left[(\eta(X) - \eta^*(X))^2 + (\eta^*(X) - \eta^*(X_i))^2 + (\eta^*(X_i) - \eta(X_i))^2\right]\right)\\
&= 3\, E\left(\left[\sum W_{ni}(X)\right](\eta(X) - \eta^*(X))^2\right) + 3\, E\left(\sum W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2\right)\\
&\qquad + 3\, E\left(\sum W_{ni}(X)(\eta^*(X_i) - \eta(X_i))^2\right).
\end{aligned} \tag{6.7}$$

Considering each term individually,

$$E\left(\left[\sum W_{ni}(X)\right](\eta(X) - \eta^*(X))^2\right) < \epsilon \qquad \text{using (6.4),} \tag{6.8}$$

$$\lim_{n\to\infty} E\left(\sum W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2\right) \leq \epsilon \qquad \text{using (6.6),} \tag{6.9}$$

and

$$\begin{aligned}
E\left(\sum W_{ni}(X)(\eta^*(X_i) - \eta(X_i))^2\right) &\leq C\, E(\eta^*(X) - \eta(X))^2\\
&< C\epsilon \qquad \text{using the assumption (i) and (6.4).}
\end{aligned} \tag{6.10}$$

Combining (6.8), (6.9), and (6.10) into (6.7), we get

$$\lim_{n\to\infty} ET_2 \le \lim_{n\to\infty} E\left(\sum W_{ni}(X)(\eta(X)-\eta(X_i))^2\right) \le 3(2+C)\epsilon. \tag{6.11}$$

For T_1, we have

$$\begin{aligned}
ET_1 &= E\left(\hat{\eta}_n(X)-\eta_n(X)\right)^2 = E\left(\sum W_{ni}(X)(\eta(X_i)-Y_i)\right)^2 \\
&= \sum_{i=1}^n\sum_{j=1}^n E\left(W_{ni}(X)W_{nj}(X)(\eta(X_i)-Y_i)(\eta(X_j)-Y_j)\right) \\
&= \sum\sum E\left(W_{ni}(X)W_{nj}(X)\cdot E\left\{(\eta(X_i)-Y_i)(\eta(X_j)-Y_j)|X,X_1,\ldots,X_n\right\}\right) \\
&= \sum\sum_{i\neq j} E\left(W_{ni}(X)W_{nj}(X)\cdot E\left\{(\eta(X_i)-Y_i)|X_i\right\}\cdot E\left\{(\eta(X_j)-Y_j)|X_j\right\}\right) \\
&\qquad\qquad + \sum_i E\left(W_{ni}^2(X)\cdot E\left\{(\eta(X_i)-Y_i)^2|X_i\right\}\right) \\
&= 0 + \sum_{i=1}^n E\left(W_{ni}^2(X)\cdot E\left\{(\eta(X_i)-Y_i)^2|X_i\right\}\right) \\
&\le \sum_{i=1}^n EW_{ni}^2(X) = E\left(\sum_{i=1}^n W_{ni}^2(X)\right) \\
&\le E\left(\max_{1\le i\le n} W_{ni}(X)\cdot\sum W_{ni}(X)\right) \\
&= E\left(\max_{1\le i\le n} W_{ni}(X)\right) \to 0 \qquad \text{by the assumption (iii).}
\end{aligned} \tag{6.12}$$

Combining (6.11) and (6.12) completes the proof. □

Universal consistency of versions of the k_n-NN rules immediately follows as a result of the consistency theorem.

Corollary 2. *Let f_n be the k_n-NN rules as defined in* (6.1)*. If $k_n \to \infty$ and $k_n/n \to 0$, then for all distributions, $E_{\mathcal{D}_n} R(f_n) \to E\, R^*$ as $n \to \infty$.*

Proof. For the k-NN rule, recall that the weights are given by

$$W_{ni}(x) = \begin{cases} \frac{1}{k} & \text{if } x_i \in N_k(x) \\ 0 & \text{otherwise.} \end{cases}$$

Since $k_n \to \infty$, condition (iii) of Stone's theorem is trivially satisfied by the weights,

$$E\left\{\max_{1\le i\le n} W_{ni}(X)\right\} = \frac{1}{k_n} \to 0.$$

For the condition (ii), we need to verify that for all $a > 0$,

$$\lim_{n\to\infty} E\left\{\sum_{i=1}^n W_{ni}(X) I(\|X_i - X\| > a)\right\} = 0.$$

It is equivalent to prove that $\sum_{i=1}^n W_{ni}(X) I(\|X_i - X\| > a) \xrightarrow{P} 0$. That is, for $\epsilon > 0$, $P(\sum_{i=1}^n W_{ni}(X) I(\|X_i - X\| > a) > \epsilon) \to 0$. Given $\epsilon > 0$ find a k_n such that $1/k_n \leq \epsilon$. Then $\sum_{i=1}^n W_{ni}(X) I(\|X_i - X\| > a) > \epsilon \geq 1/k_n$ means that there exists at least one $X_i \in N_{k_n}(X)$ such that $\|X_i - X\| > a$, which implies $\|X_{(k_n)}(X) - X\| > a$. Therefore,

$$P\left(\sum_{i=1}^n W_{ni}(X) I(\|X_i - X\| > a) > \epsilon\right) \leq P(\|X_{(k_n)}(X) - X\| > a).$$

However, by Lemma 6, the k_nth NN of X converges to X w.p.1. So the right-hand side probability must go to 0 as $n \to \infty$.

Now to prove the first condition of Stone's Theorem, we need to show that there exists a constant C such that for every non-negative integrable f,

$$E\{\sum_{i=1}^n W_{ni}(X) f(X_i)\} \leq C\, Ef(X).$$

For the k-NN rule,

$$E\left\{\sum_{i=1}^n W_{ni}(X) f(X_i)\right\} = \frac{1}{k_n} \sum_{i=1}^{k_n} Ef(X_{(i)}(X)).$$

The condition is verified by Stone's Lemma given below. □

Lemma 7 (Stone's Lemma (1977))**.** *For every non-negative integrable f,*

$$\frac{1}{k} \sum_{i=1}^{k} Ef(X_{(i)}(X)) \leq \gamma_d Ef(X)$$

where γ_d is a constant that depends only on the dimension d of X.

In order to prove the lemma, we need a new concept.

Definition 8 (Cones)**.** For $\theta \in (0, \pi/2)$, define a cone by

$$C(x, \theta) = \{y \in \mathbb{R}^d : \text{ angle between } x \text{ and } y \leq \theta\}.$$

See Figure 8(a) for illustration. Let $z + C(x, \theta)$ be the translation of the cone $C(x, \theta)$ by z.

Lemma 8. *For a fixed $\theta \in (0, \pi/2)$ there exists a set $\{z_1, \ldots, z_{\gamma_d}\} \subset \mathbb{R}^d$ such that*

$$\mathbb{R}^d = \bigcup_{i=1}^{\gamma_d} C(z_i, \theta).$$

For $\theta = \pi/6$, $\gamma_d \leq \left(1 + \frac{2}{\sqrt{2 - \sqrt{3}}}\right)^d - 1$.

The above lemma will be used for proof of Stone's lemma, but not proved here. For illustration, Figure 8(b) depicts an example with $d = 2$ and $\theta = \pi/6$.

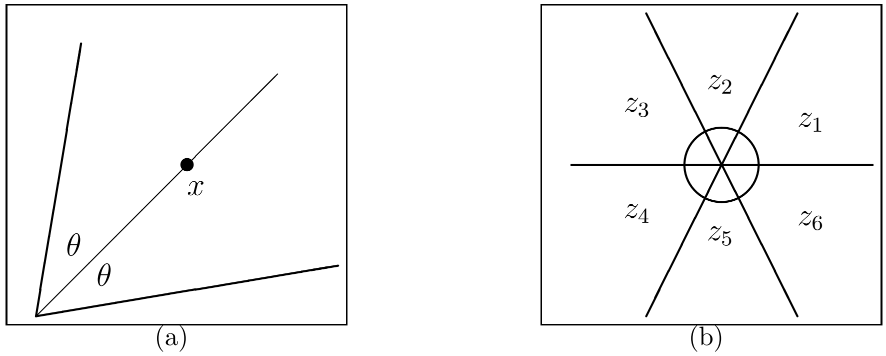

Figure 8: (a) a cone, $C(x,\theta)$, and (b) $\pi/6$ cones that cover $\mathbb{R}^2$

Lemma 9. *If $y, z \in C(x, \pi/6)$ and $\|y\| < \|z\|$, then $\|y-z\|^2 < \|z\|$.*

Proof. Note that if the angle between two vectors y and z is θ then, $y'z = \|y\|\,\|z\|\,cos(\theta)$. For $y, z \in C(x, \pi/6)$, since $\|y\| < \|z\|$,

$$\begin{aligned}\|y-z\|^2 &= \|y\|^2 + \|z\|^2 - 2y'z \le \|y\|^2 + \|z\|^2 - 2\|y\|\,\|z\|\,cos(\pi/3)\\ &= \|y\|^2 + \|z\|^2 - \|y\|\,\|z\| < \|z\|^2.\end{aligned}$$

□

Proof of Stone's Lemma (Lemma 7). Given X, cover $\mathbb{R}^d$ by γ_d cones $\{X + C(z_l, \pi/6) : \ l = 1, \dots, \gamma_d\}$. Mark in each cone the X_i that is nearest to X if such an X_i exists.

If $X_i \in \{X + C(z_l + \pi/6)\}$ but it is not marked then X cannot be the NN of X_i in $X_{(-i)} := \{X, \dots, X_{i-1}, X, X_{i+1}, \dots, X_n\}$. To see this, note that if X_i is not marked then there exists $X_j \in \{X + C(z_l + \pi/6)\}$ such that $\|X_j - X\| < \|X_i - X\|$. Then by Lemma 9, $\|X_i - X_j\| < \|X_i - X\|$. Similarly, mark all the k-NNs of X in each cone. If there are less than k points in a cone then mark all of them. Thus the following claim, If $X_i \in \{X + C(z_l + \pi/6)\}$ but it is not marked then X cannot be among the k-NNs of X_i in $X_{(-i)}$.

$$\begin{aligned}\sum_{i=1}^{k} Ef(X_{(i)}(X)) &= E\left(\sum_{i=1}^{n} I(X_i \in N_k(X)) f(X_i)\right)\\ &= E\left(\sum_{i=1}^{n} I(X \text{ is among the } k\text{-NNs of } X_i \text{ in } X_{(-i)}) f(X)\right)\\ &\le E\left(f(X)\sum_{i=1}^{n} I(X_i \text{ is marked})\right)\\ &= E\left(f(X)\sum_{j=1}^{\gamma_d}\sum_{i=1}^{n} I(X_i \text{ is marked and } X_i \in \{X + C(z_j, \pi/6)\})\right)\\ &\le E\left(f(X)\sum_{j=1}^{\gamma_d} k\right) = k\gamma_d Ef(X).\end{aligned}$$

The second equality holds because $X, X_1, \dots, X_n$ are iid and thus they are exchangeable. In particular, X and X_i are interchanged for the equality. □