THE STRONG CONVERGENCE OF EMPIRICAL NEAREST NEIGHBOR ESTIMATES OF INTEGRALS

Luc Devroye
McGill University

SUMMARY

Let $X_1, \ldots, X_n$ be independent random variables with common probability measure $\mu$ on the Borel sets of $\mathbb{R}^1$, and let $A_{n1}, \ldots, A_{nn}$ be the nearest neighbor partition of the real line obtained from $X_1, \ldots, X_n$. When $\text{fcl}^1(\mu)$, then it is known that $\frac{1}{n} \sum_{i=1}^{n} f(X_i) \mu(A_{ni}) \rightarrow \int f(x) \mu(dx)$ in probability as $n \to \infty$ (Stone, 1977). We show that "in probability" can be replaced by "almost surely" whenever $\text{fcl}^\infty(\mu)$. No conditions are placed upon $\mu$.

1. INTRODUCTION

We consider the problem of the approximation of $\int f(x) \mu(dx)$ by $I_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$, where $\mu$ is an arbitrary probability measure on the Borel sets of $\mathbb{R}^1$, $f$ is a Borel measurable function and $\mu_n$ is an empirical probability measure. We assume throughout that $X_1, \ldots, X_n$ are independent identically distributed random variables with probability measure $\mu$. When $\mu$ is the classical empirical measure, then

$$
\text{fcl}^1(\mu) \rightarrow I = \int f(x) \mu(dx) \quad \text{a.s. as } n \to \infty.
$$

and $I_n \to I = \int f(x) \mu(dx)$ a.s. as $n \to \infty$ for all $\text{fcl}^1(\mu)$. We are interested here in the same type of result for the empirical nearest neighbor measure $\mu_n$. In section 3 we will highlight the impact of this result on the study of the nearest neighbor method in discrimination. Yakowitz (1977) has suggested the use of the empirical nearest neighbor estimate $I_n$ in Monte Carlo integration, and he has given evidence showing that $I_n$ converges faster to $I$ when $\mu$ satisfies some regularity conditions.

The empirical nearest neighbor estimate $I_n$ is defined by

---

*This research was sponsored by U.S. Air Force Grant AFOSR 76-3062 while the author was at the University of Texas during the summer of 1979.*
where $A_{ni} = \{x | e_1 x_i \}, X_i$ is the nearest neighbor of $x$ among $X_1, \ldots, X_n$.

We say that $X_i$ is closer to $x$ than $X_j$ when

- $|x - X_i| < |x - X_j|$
- $|x - X_i| = |x - X_j|$, $X_i < X_j$
- $X_i = X_j$, $i < j$.

Thus, $X_i = X_j$, $i < j$, implies that $A_{nj}$ is empty.

Stone (1977) has shown that when $f \in L_p(\mu), p \geq 1$,

$$E(\|I_n - I_p\|) \to 0 \quad \text{as} \quad n \to \infty,$$

for all probability measures $\mu$. In particular, when $f \in L^1(\mu)$, it is true $I_n \to I$ in probability as $n \to \infty$. The almost sure convergence of $I_n$ to $I$ cannot be established by the methods employed by Stone. The main result of this paper is the following Theorem:

**Theorem 1.** Let $|f| \leq c < \infty$ be a Borel measurable function and let $\mu$ be a probability measure on the Borel sets of $\mathbb{R}^d$. Then the empirical nearest neighbor estimate $I_n$ satisfies

$$\|I_n - I_p\| \leq \frac{n}{\binom{n}{2}} \int \int |f(x) - f(x')| \, \mu(dx) \, \mu(dx') \to 0 \quad \text{as} \quad n \to \infty. \quad (1)$$

**2. PROOFS**

**Lemma 1.** Theorem 1 is true whenever $\mu$ is nonatomic.

**Proof of Lemma 1.**

Replace all $X_i$'s by $F(X_i)$'s where $F$ is the distribution function corresponding to $\mu$. Let $X^{(1)} \leq X^{(2)} \leq \ldots \leq X^{(n)}$ be the order statistics of $F(X_1), \ldots, F(X_n)$ and let $B_{ni} = (X^{(i-1)} - X^{(i)})$ where $X^{(0)} = 0, X^{(n+1)} = 1$. Clearly

$$\frac{n}{\binom{n}{2}} \int \int |f(x) - f(x')| \, \mu(dx) \, \mu(dx')$$

$$\leq \frac{n}{\binom{n}{2}} \int \int |g(X^{(i)}) - g(x)| \, \mu(dx)$$

where $g(u) = f(F^{-1}(u))$ and $F^{-1}(u) = \inf\{y | f(y) = u\}$, and $0 \leq u \leq 1$.

Consider a sample of size $2n$, and define

$$V_{2n} = \frac{1}{2n} \sum_{i=1}^{2n} \int_{B_{2n} i}^{B_{2n} i+1} |g(x) - g(x')| \, dx$$

which can be split up into two sums $V_1 + V_2 = \sum_{i=1}^{2n} \int_{B_{2n} i}^{B_{2n} i+1} \int_{B_{2n} i}^{B_{2n} i+1} |g(x) - g(x')| \, dx$. It suffices to show that $V_1 + V_2 \to 0$ a.s. as $n \to \infty$. Lemma 1 then follows by symmetry. Let us define $C_i = B_{2n} i$ for $1 \leq i \leq n$, and $D_n = (X^{(1)}, X^{(2)}, \ldots, X^{(2n-1)})$. We will show that

$$V_1 \to E(V_1 | D_n) \to 0 \quad \text{as} \quad n \to \infty,$$

and

$$E(V_1 | D_n) \to 0 \quad \text{as} \quad n \to \infty. \quad (3)$$

Since $C_1, \ldots, C_n$ are determined by $D_n$, we have for all $r \geq 2$ and some constant $a_r < \infty$, by a result of Dharmadhikari and Jogdeo (1969), after defining

$$A_i = \int |g(X^{(i)}) - g(x)| \, dx, \quad 1 \leq i \leq n,$$

$$E(V_1 = E(V_1 | D_n) | F) = E(\sum_{i=1}^{n} (Z_i - E(Z_i | D_n)) | F)$$

$$\leq a_r \sum_{i=1}^{n} E(\|Z_i - E(Z_i | D_n)\|)$$

$$\leq 2^r a_r \sum_{i=1}^{n} E(\|Z_i\|)$$

$$\leq 2^{r-1} a_r \sum_{i=1}^{n} E(\|Z_i\|)$$

$$\leq (4c)^r a_r^{n - 1} E(\|Z_1\|)$$

$$\leq (4c)^r a_r^{n - 1} E(\|Z_1\|)$$

$$= (4c)^r a_r^{n - 1} E(\|Z_1\|)$$

which is summable in $n$ for $r > 2$; (3) then follows by the Borel-Cantelli lemma.

Let us next define

$$\rho^+(x, b) = b^{-1} \int_{y > x, |y-x| > sb} |f(y) - f(x)| \, dy$$
for $x \in [0,1]$ and $b > 0$. Define $\rho^-(x,b)$ similarly on the set $y < x, |y-x| < b$, and let

$$\rho(x,a) = \sup_{0 < b < a} (\rho^+(x,b), \rho^-(x,b)),$$  

$a > 0$.

By the Lebesgue density theorem (Stein, 1970), we know that $\rho(x,a) = 0$ as $a \uparrow 0$ for almost all $x$. Also, $\rho(x,a)$ is nonincreasing as $a \uparrow 0$. Now,

$$E(V_{2n}^n | D_n) \leq \sum_{i=1}^{n} E \rho(x,U_i) dx \leq \int \rho(x,U) dx$$

$$= \int_{C_1}^1 \rho(x) dx = 0$$

where $U = \sup_{n \to \infty} |X_{2(2i+1)} - X_{2i}^n|$. Because $U = 0$ a.s. as $n \to \infty$ (Slud, 1978) and $0 < \rho < 2c$, we have by a generalization of the dominated convergence theorem (Glick, 1974), $\rho(x,U) dx \to 0$ a.s. as $n \to \infty$.

When $u$ has atoms, we consider the decomposition of $u$ into its atomic part $(u_1^a)$ and its nonatomic part $(u_2) : u = u_1 + u_2$. Let $A$ be the set of atoms of $u$. If $A$ is empty, theorem 1 follows from lemma 1. If $\mu(A) = 1$, theorem 1 is almost trivial. For other cases, the following lemma will be useful.

**Lemma 2.** $W_n = \sum_{i \in A} u_2(n_i) = 0$ a.s. as $n \to \infty$.

**Proof of Lemma 2.**

Consider the two subsequences of random variables from $X_1, X_2, \ldots$, defined by membership in $A$. Let $Y_1, \ldots, Y_n$ be the collection of $X_i$'s, $1 \leq i \leq n$, belonging to $A^c$, and let $Z_1, \ldots, Z_n$ be the corresponding collection for $A$. $Y_1$'s and $Z_1$'s are added in order of their appearance in the $X_1$ sequence. Clearly, $M_n \to M = n$ for all $n$. If $F(x) = u_2(-x,x)$, then

$$E(Y_{i+1}) - E(Y_1) \leq \sum_{i=0}^{M-1} (F(Y_{i+1}) - F(Y_i)) T_n = W_n$$

where $Y_1, \ldots, Y_n$ are the order statistics corresponding to $Y_1, \ldots, Y_n$. $Y(0) \to \infty$; $Y(M+1) \to \infty$; $T_n$ is the indicator function of the event that at least one $Z_j, 1 \leq j \leq n$, belongs to $(Y_1, Y_{i+1})$. Assume that $u_2(0) = q > 0$.

If $E_0, \ldots, E_n$ are independent identically distributed exponential random variables with sum $S_n$, then $W_n$ is distributed as

$$W_n \leq \sum_{i=1}^{n} E_i T_i / S_n \leq S_n^{-1} \sum_{i=0}^{n} E_i^2 Q_0^{-1 / 2}$$

where $Q_0$ is the number of different values among $Z_1, \ldots, Z_n$, and $T_n$ is the indicator function of the event that one or more of the $F(Z_j)$'s belongs to $(S_0^{-1} F_{+} + \cdots + E_{-1}^{-1}, S_0^{-1} F_{+} + \cdots + E_{-1}^{-1})$, with $E_{-1} = 0$. To conclude that $W_n \to 0$ a.s. as $n \to \infty$, it suffices to show that

$$E \left[ \sum_{i=0}^{n} \int \left| \sum_{i=0}^{n} F(Z_j) - E_{-1}^{-1} \right|^2 \right] = 0$$

$$E \left[ \sum_{i=0}^{n} \int \left| \sum_{i=0}^{n} F(Z_j) - E_{-1}^{-1} \right|^2 \right] = 0$$

here $E_0, E_1, \ldots, E_n$ are i.i.d. exponential random variables with sum $S_n$.

Statement (ii) is true by the strong law of large numbers, and statement (iii) can be proved without difficulty by using exponential inequalities for sums of independent random variables (e.g., see Baum, Katz, and Read, 1962). Lemma 2 will follow if we can show (i).

It is clear that

$$E(Q_n) \leq \sum_{i=1}^{n} (1 - (1 - a_i)^{n_i})$$

for some sequence of $a_i$'s with $\sum a_i = 1$, $a_i \geq 0$. Thus, for $c_2 > 0$,

$$E(Q_n) \leq \sum_{a_i < c_2} a_i + \sum_{a_i > c_2} a_i c_2$$

$$\leq \sum_{a_i < c_2} a_i + 1 + n \sum_{a_i > c_2} a_i$$

$$\leq n o(1) + 1 + n \sum_{a_i > c_2} a_i$$

since $c_2 > 0$ was arbitrary, we conclude that $E(Q_n)/n + 0$ a.s. as $n \to \infty$. If $Q(k, \varepsilon)$ is the number of different values among $Z_{k+1}, \ldots, Z_k$, then obviously

$$0 \leq Q(k, \varepsilon) \leq Q(k, s)^{Q(a, \varepsilon)}$$

all $k \leq k < k'$, and the distribution of $Q(k, \varepsilon)$ only depends upon $k - k'$. By the subadditive ergodic theorem (Kingman, 1968, 1973), we may conclude that $Q(k, s)^{Q(a, s)} = \lim E(Q(k, s))^{Q(a, s)} a.s., n \to \infty$. Since $Q(k, s)^{Q(a, s)}$, we have $c_1 = 0$ and $Q_n / n + 0$ a.s. as $n \to \infty$. 

Nearest Neighbor Estimates

247
Proof of Theorem 1.

Let $M_n$, $N_n$, $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$ be as in the proof of Lemma 2, and let $B_{nl}$, $1 \leq i \leq M_n$, be the nearest neighbor partition of $R^d$ corresponding to $Y_1, \ldots, Y_n$. Let $C_n$ be $\{x| x \notin A, X_i \notin x, \text{ all } i \leq n\}$. Then,

$$
\sum_{i=1}^{M_n} \sum_{x \in A} \int_{B_{nl}} |f(X_i) - f(x)| \mu(dx) \leq \sum_{i=1}^{M_n} \int_{B_{nl}} |f(Y_i) - f(x)| \mu(dx) + \sum_{1 \leq i < \ell \leq n} \frac{2c_{nl} \mu_1(A_{nl}) + 2c_{nl} \mu_1(C_n)}{1 - \mu_1((x_i))}.
$$

Since $M_n \to \infty$ as $n \to \infty$, the first term on the right-hand side of (6) tends to 0 a.s. as $n \to \infty$ (Lemma 1). The second term tends to 0 a.s. as $n \to \infty$ by Lemma 2. Finally, $\mu_1(C_n)$ is monotone and

$$
E(\mu_1(C_n)) = \int_{x \in A} \mu_1(x) (1 - \mu_1((x))) \to 0 \text{ as } n \to \infty,
$$

so that $\mu_1(C_n) \to 0$ a.s. as $n \to \infty$. This concludes the proof of Theorem 1.

3. THE NEAREST NEIGHBOR RULE

We will now consider the implications of Theorem 1 in nonparametric discrimination. Let $(X,Y), (X_1,Y_1), \ldots, (X_n,Y_n)$ be independent identically distributed $R^d \times (0,1)$-valued random vectors, and let $Y$ be estimated by $\hat{Y}_n = \hat{Y}_n$, $1 \leq n \leq N$ when $X \in A_{nl}$. Thus, $\hat{Y}_n$ depends upon $X$, and $(X_1,Y_1), \ldots, (X_n,Y_n)$. It is called the nearest neighbor estimate of $Y$ (Fix and Hodges, 1951; Cover and Hart, 1967). We define

$$
L_n = P(\hat{Y}_n \neq Y_1, \ldots, \hat{Y}_n \neq Y_n)
$$

and

$$
L^* = \inf_{g:R^d \to (0,1)} P(g(X) \neq Y).
$$

$L^*$ is called the Bayes probability of error, and $L_n$ is the probability of error for the nearest neighbor estimate and the given data. Clearly, $L_n \geq L^*$, all $n$.

Under some restrictions on the distribution of $(X,Y)$, Cover and Hart have shown that

$$
\lim_{n \to \infty} E(L_n) = 2E(\eta(X)(1-\eta(X))) \leq 2L^*(1-L^*)
$$

where

$$
\eta(x) = P(Y=1|X=x), \ x \in R^d.
$$

Stone (1977) and Devroye (1980) showed that (7) remains valid for all distributions of $(X,Y)$. In general, $L_n$ does not converge to a constant in probability. For example, when $X=0$ a.s., and $\eta(0) = \frac{1}{3}$, then

$$
L_n = \frac{1}{3} \int_{|y|=0} \mu(dx) + \frac{2}{3} \int_{|y|=1} \mu(dx)
$$

so that convergence to a constant is excluded. Devroye (1980) has shown recently that

$$
\lim_{n \to \infty} 2E(\eta(X)(1-\eta(X))) = 2L^*(1-L^*)
$$

whenever the probability measure $\mu$ of $X$ is nonatomic. This result can be strengthened now for $d=1$:

Theorem 2. For $d=1$ and all nonatomic probability measures $\mu$, we have, a.s.,

$$
\lim_{n \to \infty} L_n = 2E(\eta(X)(1-\eta(X))) \leq 2L^*(1-L^*)
$$

Proof of Theorem 2.

Clearly,

$$
L_n = \int_{|y|=0} \mu(dx) + \int_{|y|=1} \mu(dx)
$$

and

$$
L^* = \inf_{g:R^d \to (0,1)} P(g(X) \neq Y)
$$

We will show that $L_n \to E(\eta(X)(1-\eta(X)))$ a.s. as $n \to \infty$. By symmetry, we can then conclude that $L_n \to 2E(\eta(X)(1-\eta(X)))$ a.s. as $n \to \infty$. The inequality in Theorem 2 is a simple consequence of Jensen's inequality when one notices that $L^* = E(\min(\eta(X), 1-\eta(X)))$.

Let

$$
\eta(x) = P(Y=1|X=x), \ x \in R^d
$$

and

$$
E(\eta(X)(1-\eta(X))) \leq 2L^*(1-L^*)
$$

Then,

$$
\int_{|y|=0} \eta(x)(1-\eta(x)) \mu(dx)
$$

and

$$
\int_{|y|=1} \eta(x)(1-\eta(x)) \mu(dx)
$$

The last term of (8) tends to 0 a.s. for all probability measures $\mu$ (Theorem 1). Check that $\sum_{i=1}^{M_n} \mu_1(A_{nl}) \leq 1$, $\mu_1(A_{nl}) \geq 0$, $E(C_n, \xi_1, \ldots, \xi_n) = C_n$ a.s., $|Z_1| \leq 1$ and $E(Z_1) = 0$ a.s. Thus, by an inequality of Dharmadhikari and Jogdeo (1969), for all $r>1$, there exists a constant $a=a(r)>0$ such that
For all \( g \in L^1(\mu) \), the first term on the right-hand-side of (9) is small by the choice of \( M \). By Hölder’s inequality, the last term of (9) is not greater than

\[
E\left( \sum_{i=1}^{n} |g(x_i)|^{2r} \right)^{\frac{1}{r}} E\left( \sum_{i=1}^{n} \left| g(x_i) \right|^{q} \right)^{\frac{1}{q}} \leq \frac{1}{n^{\frac{r}{q}}} \int \left| g(x) \right|^{q} \mu(dx) \leq \frac{1}{n^{\frac{r}{q}}} \int \left| g(x) \right|^{q} \mu(dx)
\]

(10)

where \( p, q > 1, \frac{1}{p} + \frac{1}{q} = 1 \). Now, \( \frac{1}{n^{\frac{r}{q}}} \int \left| g(x_i) \right|^{q} \mu(dx) \leq n E\left( |g(X_i)|^{q} \right) \) a.s. as \( n \to \infty \).

Let \( S_{ni} \), \( 0 \leq i \leq n \), be the spacings of i.i.d. uniform (0,1) random variables. Clearly, for all \( n \) and all \( u > 0 \),

\[
P\left( \sum_{i=1}^{n} \mu(A_{ni}) > u \right) \leq P\left( \sum_{i=1}^{n} \mu(A_{ni}) > u \right) \leq P\left( \sum_{i=1}^{n} \mu(A_{ni}) > u \right).
\]

(11)

Now, we will show that \( \frac{1}{n} \sum_{i=0}^{n} \left( nS_{ni} \right)^{q} \) converges completely to \( \Gamma(q+1) \) as \( n \to \infty \).

This in turn implies that for all \( \varepsilon > 0 \), \( \sum_{i=1}^{n} \mu(A_{ni}) \leq 2^{q-1}\Gamma(q+1) (1+\varepsilon)^{q-1} n^{-1} \)

except possibly for finitely many \( n \), a.s. Thus, almost surely, (10) is smaller than

\[
\frac{1}{n^{\frac{r}{q}}} \int \left| g(x) \right|^{q} \mu(dx) \leq \frac{1}{n^{\frac{r}{q}}} \int \left| g(x) \right|^{q} \mu(dx)
\]

except possibly for finitely many \( n \). This too can be made arbitrarily small by choosing \( M \) large enough. This would complete the proof of Lemma 3.

It is known that \( S_{n0}, \ldots, S_{nn} \) are distributed as \( E_{0}/S_{0}, E_{1}/S_{1}, \ldots, E_{n}/S_{n} \) where \( E_{0}, E_{1}, \ldots, E_{n} \) are i.i.d. exponential random variables and \( S = E_{0} + \cdots + E_{n} \). Thus,

\[
\frac{1}{n} \sum_{i=0}^{n} \left( nS_{ni} \right)^{q} \Gamma(q+1) \text{ completely when}
\]

\[\begin{align*}
&\text{(i)} \quad \frac{1}{n} \sum_{i=0}^{n} E_{i}^{q} \Gamma(q+1) \text{ completely,} \\
&\text{(ii)} \quad \frac{1}{n} \sum_{i=0}^{n} E_{i} + 1 \text{ completely.}
\end{align*}
\]

The latter two results follow from the fact that \( E(E_{i}^{q}) = \Gamma(q+1) \), all \( q \geq 0 \), and that \( E(E_{i}^{q}) < \infty \) (apply for example, Theorem 28, p. 286 of Petrov (1975) which states that under the said conditions \( P\left( \left( \sum_{i=0}^{n} \left( E_{i}^{q} \right) \right) / \varepsilon \right) = o(n^{-2}) \) for all \( \varepsilon > 0 \)).

Thus, by the Borel-Cantelli lemma, (8) tends to 0 a.s. as \( n \to \infty \).

4. **Refinements**

In Lemma 1, we established the strong convergence to \( I_{n} \) of the empirical nearest neighbor estimate \( I_{n} \) when \( f \) is bounded and \( \mu \) is nonatomic. The condition that \( f \) is bounded can be dropped without much trouble.

**Lemma 3.** (1) is valid whenever \( f \in L^{p}(\mu) \) for some \( p > 1 \), and \( \mu \) is nonatomic.

**Proof of Lemma 3.**

By Theorem 1, (1) is valid for the function \( f(x)I_{\left( \mid f(x) \mid > M \right)} \) and any constant \( M \). Let us fix a constant \( M > 0 \), and define \( g(x) = f(x)I_{\left( \mid f(x) \mid > M \right)} \).

We have,

\[
|I_{n} - I_{x}| \leq \sum_{i=1}^{n} g(x_i) - g(x)|u(dx)| + o(1) \text{ a.s. as } n \to \infty.
\]

Also,

\[
\frac{1}{n} \sum_{i=1}^{n} g(x_i) - g(x)|u(dx)| \leq \frac{1}{n} \sum_{i=1}^{n} |g(x_i)||u(A_{ni})|.
\]

(9)
5. REFERENCES


