STATISTICS AND RELATED TOPICS
M. Csörgö, D.A. Dawson, J.N.K. Rao, A.K.Md.E. Saleh (eds.)

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 μ on the Borel sets of 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 $f\in L^1(\mu)$, then it is known that $\sum_{i=1}^{n} f(X_i)\ \mu(A_{ni}) \to \int f(x)\ \mu(dx)$ in probability as $n\to\infty$ (Stone, 1977). We show that "in probability" can be replaced by "almost surely" whenever $f\in L^\infty(\mu)$. No conditions are placed upon μ.

1. INTRODUCTION

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

$$\int f(x)\ \mu_n(dx) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

and $I_n \to I=\int f(x)\ \mu(dx)$ a.s. as $n\to\infty$ for all $f\in L^1(\mu)$. We are interested here in the same type of result for the empirical nearest neighbor measure μ_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 μ 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.

$$\sum_{i=1}^{n} f(X_i)\ \mu(A_{ni})$$

where $A_{ni} = \{x|x\in R^1,\ X_i$ 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

$$\text{either } |x-X_i| < |x-X_j|$$

$$\text{or } |x-X_i| = |x-X_j|\ ,\ X_i < X_j$$

$$\text{or } X_i = X_j\ ,\ i < j\ .$$

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

Stone (1977) has shown that when $f\in L^P(\mu)$, $p \ge 1$,

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

for all probability measures μ. 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 $|f| \le c < \infty$ be a Borel measurable function and let μ be a probability measure on the Borel sets of R^1. Then the empirical nearest neighbor estimate I_n satisfies

$$|I_n-I| \le \sum_{i=1}^{n} \int_{A_{ni}} |f(X_i)-f(x)|\ \mu(dx) \to 0 \text{ a.s. as } n \to \infty\ . \tag{1}$$

2. PROOFS

Lemma 1. Theorem 1 is true whenever μ is nonatomic.

Proof of Lemma 1.

Replace all X_i's by $F(X_i)$'s where F is the distribution function corresponding to μ. Let $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ 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

$$\sum_{i=1}^{n} \int_{A_{ni}} |f(X_i)-f(x)|\mu(dx)$$

$$\le \sum_{i=1}^{n} \int_{B_{ni}\cup B_{n\,i+1}} |g(X_{(i)}-g(x)|\ dx \tag{2}$$

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

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

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

which can be split up into two sums $V'_{2n}+V''_{2n} = \sum_{i\text{ even}} + \sum_{i\text{ odd}}$. It suffices to show that $V'_{2n} \to 0$ a.s. as $n \to \infty$. Lemma 1 then follows by symmetry. Let us define $C_i = B_{2n\ 2i}\cup B_{2n\ 2i+1}$, $1 \le i \le n$, and $D_n=(X_{(1)},X_{(3)},\ldots,X_{(2n-1)})$. We will show that

$$V'_{2n} - E(V'_{2n}|D_n) \to 0 \text{ a.s. as } n \to \infty\ , \tag{3}$$

and

$$E(V'_{2n}|D_n) \to 0 \text{ a.s. as } n \to \infty\ . \tag{4}$$

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

$$A_i = \int_{C_i} |g(X_{(2i)})-g(x)|\ dx\ ,\ 1 \le i \le n\ ,$$

$$E(|V'_{2n}-E(V'_{2n}|D_n)|^r) = E(|\sum_{i=1}^{n} (Z_i-E(Z_i|D_n))|^r)$$

$$\le a_r\, n^{\frac{r}{2}-1} \sum_{i=1}^{n} E(|Z_i-E(Z_i|D_n)|^r)$$

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

$$\le (4c)^r a_r n^{\frac{r}{2}-1} \sum_{i=1}^{n} E(\mu^r(C_i))$$

$$= (4c)^r a_r n^{\frac{r}{2}-1}\ E(\mu^r(C_1))$$

$$= O(n^{-\frac{r}{2}})$$

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

Let us next define

$$\rho^+(x,b) = b^{-1} \int_{y > x,\ ||y-x|| \le b} |f(y)-f(x)|dy$$

for $x\in[0,1]$ and $b > 0$. Define $\rho^-(x,b)$ similarly on the set $y<x, ||y-x||\le b$, and let

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

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

$$E(V'_{2n}|D_n) \le \sum_{i=1}^{n} \int_{C_i} \rho(x,U_n)\,dx = \int_0^1 \rho(x,U_n)\,dx \qquad (5)$$

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

When μ has atoms, we consider the decomposition of μ into its atomic part (μ_1) and its nonatomic part (μ_2) : $\mu=\mu_1+\mu_2$. Let A be the set of of atoms of μ. If A is empty, theorem 1 follows from lemma 1. If $\mu(A)=1$, theorem 1 is almost trivial. For the other cases, the following lemma will be useful.

Lemma 2. $W_n = \sum_{i:X_i\in A} \mu_2(A_{ni}) \to 0$ 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 $Y_1,\ldots,Y_{M_n}$ be the collection of X_i's, $1\le i\le n$, belonging to A^c, and let $Z_1,\ldots,Z_{N_n}$ be the corresponding collection for A. Y_i's and Z_i's are added in order of their appearance in the X_i sequence. Clearly, $M_n+N_n=n$ for all n. If $F(x)=\mu_2((-\infty,x])$, then

$$W_n \le \sum_{i=0}^{M_n} (F(Y_{(i+1)})-F(Y_{(i)}))\, T_{ni} = W^*_n$$

where $Y_{(1)} \le \cdots \le Y_{(M_n)}$ are the order statistics corresponding to $Y_1,\ldots,Y_{M_n}$; $Y_{(0)}=-\infty$; $Y_{(M_n+1)}=+\infty$; T_{ni} is the indicator function of the event that at least one Z_j, $1\le j\le N_n$, belongs to $(Y_{(i)},Y_{(i+1)}]$. Assume that $\mu_2(R)=q > 0$.

If $E_0,\ldots,E_{M_n}$ are independent identically distributed exponential random variables with sum S_{M_n}, then W^*_n is distributed as

$$\sum_{i=1}^{M_n} E_i T_{ni}/S_{M_n} \le S_{M_n}^{-1} \Big[\sum_{i=0}^{M_n} E_i^2\, Q_n \Big]^{1/2}$$

where Q_n is the number of different values among $Z_1,\ldots,Z_{N_n}$ and T_{ni} is the indicator function of the event that one or more of the $F(Z_j)$'s belongs to $(S_{M_n}^{-1}(E_0+\ldots+E_{i-1}),S_{M_n}^{-1}(E_0+\ldots+E_i)]$, with $E_{-1}=0$. To conclude that $W^*_n \to 0$ a.s. as $n\to\infty$, it suffices to show that

(i) $Q_n/n \to 0$ a.s. as $n\to\infty$,

(ii) $M_n/n \to q$ a.s. as $n\to\infty$,

and

(iii) $\sum_{n=1}^{\infty} P(|(n+1)^{-1}\sum_{i=0}^{n} E_i^2 - 2| > \varepsilon) < \infty$, all $\varepsilon > 0$.

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

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

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

$$\begin{aligned} E(Q_n) &\le \sum_{a_i<c_2/n} na_i + \sum_{a_i\ge c_2/n} 1 \\ &\le n \sum_{a_i<c_2/n} a_i + 1 + \frac{n}{c_2} \\ &= n\,o(1) + 1 + \frac{n}{c_2} \\ &\le \frac{2n}{c_2}\ , \quad n \text{ large enough.} \end{aligned}$$

Since $c_2 > 0$ was arbitrary, we conclude that $E(Q_n)/n \to 0$ as $n\to\infty$. If $Q_{(k,\ell]}$ is the number of different values among $Z_{N_k+1},\ldots,Z_{N_\ell}$, then obviously

$$0 \le Q_{(k,\ell]} \le Q_{(k,s]}+Q_{(s,\ell]} \ , \text{ all } k < s < \ell \ ,$$

and the distribution of $Q_{(k,\ell]}$ only depends upon $\ell-k$. By the subadditive ergodic theorem (Kingman, 1968, 1973), we may conclude that $Q_{(0,n]}/n \to c_1 = \lim_{n\to\infty} E(Q_{(0,n]})/n$ a.s., $n\to\infty$. Since $Q_{(0,n]}=Q_n$, we have $c_1=0$ and $Q_n/n \to 0$ a.s. as $n\to\infty$.

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_{ni}, $1 \le i \le M_n$, be the nearest neighbor partition of R^1 corresponding to $Y_1,\ldots,Y_{M_n}$. Let C_n be $\{x|\ x\in A\ ,\ X_i \ne x\ ,\ \text{all}\ i \le n\}$. Then,

$$\sum_{i=1}^{n} \int_{A_{ni}} |f(X_i)-f(x)|\ \mu(dx)$$
$$\le \sum_{i=1}^{M_n} \int_{B_{ni}} |f(Y_i)-f(x)|\ \mu_2(dx) + \sum_{i:X_i\in A} 2c\mu_2(A_{ni}) + 2c\mu_1(C_n) . \tag{6}$$

Since $M_n \to \infty$ a.s. 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 $\downarrow$, and

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

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), $(X_1,Y_1),\ldots,(X_n,Y_n),\ldots$ be independent identically distributed $R^d\times\{0,1\}$-valued random vectors, and let Y be estimated by $\hat{Y}_n = Y_i$ when $X \in A_{ni}$. 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

$$L_n = P(\hat{Y}_n \ne Y | X_1,Y_1,\ldots,X_n,Y_n)$$

and

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

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, $L_n \ge L^*$, all n. Under some restrictions on the distribution of (X,Y), Cover and Hart have shown that

$$\lim_{n\to\infty} E(L_n) = 2E(\eta(X)(1-\eta(X))) \le 2L^*(1-L^*) \tag{7}$$

where

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

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

$$L_n = \frac{1}{3} I_{[Y_1=0]} + \frac{2}{3} I_{[Y_1=1]} ,$$

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

$$L_n \to 2E(\eta(X)(1-\eta(X))) \quad \text{in probability as} \quad n \to \infty$$

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

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

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

Proof of Theorem 2.

Clearly,

$$L_n = \sum_i I_{[Y_i=1]} \int_{A_{ni}} (1-\eta(x))\mu(dx)$$
$$+ \sum_i I_{[Y_i=0]} \int_{A_{ni}} \eta(x)\ \mu(dx)$$
$$= L_{n1}+L_{n0} .$$

We will show that $L_{n1} \to E(\eta(X)(1-\eta(X)))$ 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

$$C_{ni} = \int_{A_{ni}} (1-\eta(x))\ \mu(dx) ,$$

$$Z_i = I_{[Y_i=1]} - \eta(X_i) .$$

Then,

$$\left|L_{n1} - \int\eta(x)(1-\eta(x))\mu(dx)\right|$$
$$\le \sum_i C_{ni}Z_i + \sum_i \int_{A_{ni}} |\eta(X_i)-\eta(x)|\mu(dx) . \tag{8}$$

The last term of (8) tends to 0 a.s. for all probability measures μ (Theorem 1). Check that $\sum C_{ni} \le 1$, $C_{ni} \ge 0$, $E(C_{ni}|X_1,\ldots,X_n)=C_{ni}$ a.s., $|Z_i| \le 1$ and $E(Z_i|X_1,\ldots,X_n)=0$ 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

$$E(|\sum_i c_{ni}Z_i|^{2r}) = E(E|\sum_i c_{ni}Z_i|^{2r}|X_1,\ldots,X_n))$$

$$\le E(an^{r-1}\sum_i E(|c_{ni}Z_i|^{2r}|X_1,\ldots,X_n))$$

$$= a\,n^r\,E(|c_{n1}Z_1|^{2r})$$

$$\le a\,n^r\,E(\mu^{2r}(A_{n1}))$$

$$\le a\,n^r\,E(W^{2r})$$ (where W is the second largest of a sample of n independent identically distributed uniform (0,1) random variables)

$$\le a\,n^r\int_0^1 x^{2r}\,n(n-1)(1-x)^{n-2}x\,dx$$

$$= a\,n^r\,n(n-1)\,\frac{\Gamma(n-1)\Gamma(2r+2)}{\Gamma(n+2r+1)}$$

$$= O(n^{-r})\ .$$

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 of the empirical nearest neighbor estimate I_n when f is bounded and μ 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 μ is nonatomic.

Proof of Lemma 3.

By Theorem 1, (1) is valid for the function $f(x)I_{[|f(x)|\le M]}$ and any constant M. Let us fix a constant $M>0$, and define $g(x) = f(x)I_{[|f(x)|>M]}$. We have,

$$|I_n-I| \le \sum_{i=1}^n \int_{A_{ni}} |g(X_i)-g(x)|\,\mu(dx) + o(1) \text{ a.s. as } n\to\infty\ .$$

Also,

$$\sum_{i=1}^n \int_{A_{ni}} |g(X_i)-g(x)|\mu(dx)$$

$$\le \int|g(x)|\mu(dx) + \sum_{i=1}^n |g(X_i)|\mu(A_{ni})\ . \tag{9}$$

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

$$[\sum_{i=1}^n |g(X_i)|^p]^{\frac{1}{p}}\ [\sum_{i=1}^n \mu^q(A_{ni})]^{\frac{1}{q}} \tag{10}$$

where $p,q>1$, $\frac{1}{p}+\frac{1}{q}=1$. Now, $\sum_{i=1}^n |g(X_i)|^p \sim n\,E(|g(X_1)|^p)$ a.s. as $n\to\infty$. Let S_{ni}, $0\le i\le n$, be the spacings of n i.i.d. uniform (0,1) random variables. Clearly, for all n and all $u>0$,

$$P(\sum_{i=1}^n \mu^q(A_{ni}) > u) \le P(2^{q-1}\sum_{i=0}^n S_{ni}^q > u)\ . \tag{11}$$

Now, we will show that $\frac{1}{n}\sum_{i=0}^n (nS_{ni})^q$ converges completely to $\Gamma(q+1)$ as $n\to\infty$. This in turn implies that for all $\varepsilon>0$, $\sum_{i=1}^n \mu^q(A_{ni}) \le 2^{q-1}\Gamma(q+1)\,(1+\varepsilon)/n^{q-1}$ except possibly for finitely many n, a.s. Thus, almost surely, (10) is smaller than

$$E(|g(X_1)|^p)^{\frac{1}{p}}\ 2\ \Gamma(q+1)^{\frac{1}{q}}\ ,$$

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 $E_0/S, E_1/S,\ldots,E_n/S$ where $E_0,E_1,\ldots,E_n$ are i.i.d. exponential random variables and $S=E_0+\ldots+E_n$. Thus, Thus, $\frac{1}{n}\sum_{i=0}^n (nS_{ni})^q \to \Gamma(q+1)$ completely when

(i) $\frac{1}{n}\sum_{i=0}^n E_i^q \to \Gamma(q+1)$ completely,

(ii) $\frac{1}{n}\sum_{i=0}^n E_i \to 1$ completely.

The latter two results follow from the fact that $E(E_0^q) = \Gamma(q+1)$, all $q\ge 0$, and that $E(E_0^{3q}) < \infty$ (apply for example, Theorem 28, p. 286 of Petrov (1975) which states that under the said conditions $P(|\frac{1}{n}\sum_{i=0}^n (E_i^q-\Gamma(q+1))|>\varepsilon) = o(n^{-2})$ for all $\varepsilon>0$).

5. REFERENCES

[1] Baum, L.E., Katz, M. and Read, R.R., "Exponential convergence rates in the law of large numbers", Transactions of the American Mathematical Society 120, 108-123, 1962.

[2] Cover, T.M. and Hart, P.E., "Nearest neighbor pattern classification", IEEE Transactions on Information Theory 13, 21-27, 1967.

[3] Devroye, L., "On the inequality of Cover and Hart in nearest neighbor discrimination", to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980.

[4] Dharmadhikari, S.W. and Jogdeo, K., "Bounds on moments of certain random variables", Annals of Mathematical Statistics 40, 1506-1508, 1969.

[5] Fix, E. and Hodges, J.L., "Discriminatory analysis. Nonparametric discrimination, consistency properties", Project 21-49-004, Report No. 4, School of Aviation Medicine, Randolph Field, Texas, 1951.

[6] Glick, N., "Consistency conditions for probability estimators and integrals of density estimators", Utilitas Mathematica 6, 61-74, 1974.

[7] Kingman, J.F.C., "The ergodic theory of subadditive stochastic processes", Journal of the Royal Statistical Society Series B 30, 499-510, 1968.

[8] Kingman, J.F.C., "Subadditive ergodic theory", Annals of Probability 1, pp. 883-909, 1973.

[9] Petrov, V.V., Sums of Independent Random Variables, Springer-Verlag, Berlin, 1975.

[10] Slud, E., "Entropy and maximal spacings for random partitions", Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete 41, 341-352, 1978.

[11] Stein, E.M., Singular Integrals and Differentiability Properties of Functions, Princeton University Press, Princeton, New Jersey, 1970.

[12] Stone, C.J., "Consistent nonparametric regression", Annals of Statistics 5, 595-645, 1977.

[13] Yakowitz, S., Computational Probability and Simulation, Addison-Wesley, Mass., 1977.