On the Inequality of Cover and Hart in Nearest Neighbor Discrimination

LUC DEVROYE

Abstract—When \((X_1, \theta_1), \ldots, (X_n, \theta_n)\) are independent identically distributed random vectors from \(\mathbb{R}^d \times \{0, 1\}\) distributed as \((X, \theta)\), and when \(\theta\) is estimated by its nearest neighbor estimate \(\hat{\theta}(x)\), then Cover and Hart have shown that
\[ E[P(\theta \neq \hat{\theta}(x)) \leq \frac{2}{n} 2E[\eta(X)(1 - \eta(X))^{1/\kappa}] \leq 2R^*(1 - R^*) \]

where \(R^* = E[\min(\eta(X), 1 - \eta(X))]\) is the Bayes probability of error. They require, however, that \(X\) have a density \(f\) and that \(f\) and \(\eta\) are almost everywhere continuous. It should be noted, however, that the proof in Cover and Hart holds for \(X\) taking values in a separable metric space. Stone [4] has implicitly shown that (1) is true for all distributions of \((X, \theta)\). The purpose of this note is to give two short proofs of (1) and to obtain additional results on the convergence of \(E[L_n]\).

I. INTRODUCTION

Let \((X, \theta), (X_1, \theta_1), \ldots, (X_n, \theta_n)\) be independent identically distributed \(\mathbb{R}^d \times \{0, 1\}\)-valued random vectors and estimate \(\theta\) from \(X\) and the \((X_i, \theta_i)\)'s by \(\hat{\theta}(x)\), the nearest neighbor estimate that is obtained by reordering the \((X_i, \theta_i)\) according to increasing values for \(\|X_i - X\|\) and taking \(\hat{\theta}(x)\) from the nearest neighbor \(X_\hat{\theta}(x)\) (ties are broken by comparing original indices).

Cover and Hart [1] have shown the following inequality. When
\[ E[L_n] = P(\theta \neq \hat{\theta}) \leq \frac{2}{n} 2E[\eta(X)(1 - \eta(X))] \]
and
\[ \eta(x) = P(\theta = 1 | X = x), \]
then
\[ E[L_n] = P(\theta \neq \hat{\theta}) \leq \frac{2}{n} 2E[\eta(X)(1 - \eta(X))] \]
\[ \leq 2R^*(1 - R^*) \]

where
\[ R^* = E[\min(\eta(X), 1 - \eta(X))] \]
is the Bayes probability of error. They require, however, that \(X\) have a density \(f\) and that \(f\) and \(\eta\) are almost everywhere continuous. It should be noted, however, that the proof in Cover and Hart holds for \(X\) taking values in a separable metric space. Stone [4] has implicitly shown that (1) is true for all distributions of \((X, \theta)\). The purpose of this note is to give two short proofs of (1) and to obtain additional results on the convergence of \(E[L_n]\).

II. THE BASIC THEOREM

Theorem 1:
\[ E[L_n] \leq 2R^*(1 - R^*) \]

Proof: Fix a version \(\eta\) of \(P(\theta = 1 | X = x)\), and let \(X_\eta(x)\) be the nearest neighbor of \(x\) while \(X_\hat{\theta}(x)\) is the nearest neighbor of the random variable \(X\). Further, let
\[ \xi(x) = E[\eta(X_\eta(x))] \]
and
\[ r_n(x) = \xi(x)(1 - \eta(x)) + (1 - \xi(x)) \eta(x). \]
The inequality in Theorem 1 follows from \(\eta(x)(1 - \eta(x)) = \)
\[
\min (\eta(x), 1 - \eta(x)) \times (1 - \min (\eta(x), 1 - \eta(x))) \text{ and Jensen's inequality.}
\]

Next,
\[
|r_n(x) - 2\eta(x) (1 - \eta(x))| \\
\leq E \left[ \left| \eta(X_{\bar{0}}) - \eta(x) \right| \right] \\
\leq E \left[ \left| \eta(X_{\bar{0}}) - \eta(x) \right| \right],
\]

for almost all \( x(\mu) \) \((\mu \text{ is the probability measure for } X)\), that (2) tends to 0 as \( n \to \infty \). By the dominated convergence theorem, we may then certainly conclude that
\[
E \left[ \left| r_n(X) - 2\eta(x) (1 - \eta(x)) \right| \right] \to 0.
\]

We will show for almost all \( x(\mu) \) \((\mu \text{ is the probability measure for } X)\), that (2) tends to 0 as \( n \to \infty \). By the dominated convergence theorem, we may then certainly conclude that
\[
E \left[ \left| r_n(X) - 2\eta(x) (1 - \eta(x)) \right| \right] \to 0.
\]

Theorem then follows because
\[
E \left[ r_n(X) \right] = E \left[ \eta(X) (1 - \eta(x)) + (1 - \eta(X)) \eta(x) \right] \\
= E \left[ \eta(X_{\bar{0}}) (1 - \eta(x)) + (1 - \eta(X_{\bar{0}})) \eta(x) \right] \\
= E \left[ L_{11} \right].
\]

When \( I \) is the indicator function and \( a > 0 \) is a constant, we have
\[
E \left[ \left| \eta(X_{\bar{0}}) - \eta(x) \right| \right] \\
\leq P \left( \left| X_{\bar{0}} - x \right| > a \right) \\
\leq \sup_{S_{X_{\bar{0}}}, b} \int_{S_{X_{\bar{0}}}} \eta(y) - \eta(x) \, \mu(dy) \\
\leq \sup_{S_{X_{\bar{0}}}, b} \int_{S_{X_{\bar{0}}}} \eta(y) - \eta(x) \, \mu(dy),
\]

where \( a > 0 \) is arbitrary and \( S_{X_{\bar{0}}} \) is the closed sphere centered at \( x \) and radius \( r \). The last term in (4) tends to 0 as \( a \to 0 \) for almost all \( x(\mu) \) by a theorem on the relative differentiation of measures (Wheeden and Zygmund [6, pp. 185-190]). The first term on the right-hand side of (4) tends to 0 for all \( a > 0 \) whenever \( x \in \text{support}(\mu) \). But \( \mu(\text{support}(\mu)) = 1 \) (see [1]) and the theorem is proved.

We will sketch a second proof that is essentially due to Stone [4]. Again, we will show that
\[
E \left[ \left| \eta(X_{\bar{0}}) - \eta(x) \right| \right] \to 0.
\]

For fixed \( \epsilon > 0 \), find \( g : \mathbb{R}^d \to [0, 1] \), \( g \) continuous, such that \( E \left[ \left| g(X) - \eta(x) \right| \right] < \epsilon \) (see [6, p. 149]). Estimate (5) from above by
\[
E \left[ \left| \eta(X_{\bar{0}}) - g(X_{\bar{0}}) \right| \right] \\
+ E \left[ \left| g(x) - \eta(x) \right| \right] \\
+ E \left[ \left| g(X) - \eta(x) \right| \right].
\]

Stone [4, p. 613] has shown that for any function \( f \in L^1(\mu) \),
\[
E \left[ \left| f(X_{\bar{0}}) \right| \right] \leq a(d) E \left[ \left| f(X) \right| \right]
\]

where \( a(d) > 0 \) is a constant depending upon \( d \) only. Thus, the first and third terms of (6), summed together, are not greater than \( a(d) + 1 \) \( \epsilon \).

For all \( x \in \text{support}(\mu) \), we have \( X_{\bar{0}} \to X \) a.s., and thus
\[
g(X_{\bar{0}}) \to g(x) \text{ a.s., so that the second term of (6) tends to 0 by the dominated convergence theorem.}
\]

Since \( \epsilon > 0 \) was arbitrary, we may conclude the proof of the theorem.

Remark 1: We have in fact shown that
\[
r_n(x) \to 2\eta(x) (1 - \eta(x))
\]

for almost all \( x(\mu) \).

Remark 2: If \( g \) and \( h_n \) are uniformly bounded Borel measurable functions of their arguments, then it is true that
\[
E \left[ \left| g(X_{\bar{0}}) - g(x) \right| \right] \to 0
\]

for almost all \( x(\mu) \), and
\[
E \left[ \left| g(X_{\bar{0}}) - g(x) \right| \right] \to 0
\]

for all distributions of \( X \).

Remark 3: The proof given above work for IR\(^d\), but it is not clear how they can be generalized to separable metric spaces.

III. THE CONDITIONAL PROBABILITY OF ERROR

For general \( \mu \), \( L_{11} \) does not converge to a constant in probability. For example, take \( \mu(\{0\}) = 1, \eta(0) = \frac{1}{2} \). Clearly,
\[
L_{11} = \frac{1}{2} I_{[a, -a]} + \frac{1}{2} I_{[a, 1]}
\]

and convergence to a constant is thus excluded. Nevertheless, we have the following.

Theorem 2: If \( \mu \) has no atoms, then
\[
L_{11} \to 2E \left[ \eta(X) (1 - \eta(x)) \right]
\]

in probability.

Note: Wagner [5] has shown Theorem 2 for the special case that \( \mu \) has a density \( f \) and that \( \eta \) and \( f \) are almost everywhere continuous. For \( d = 1 \), he has shown that
\[
L_{11} \to 2E \left[ \eta(X) (1 - \eta(x)) \right] \text{ a.s.,}
\]

under the same assumptions. Fritz [2] proved the a.s. convergence of \( L_{11} \) to \( 2E \left[ \eta(X) (1 - \eta(x)) \right] \) when \( \mu \) has no atoms and \( \eta \) is almost everywhere continuous (\( \mu \)). Our Theorem 2, in contrast, holds for all nonatomic measures \( \mu \) and all \( \eta \).

The proof of Theorem 2 will be postponed until Theorem 3.

To take care of the atomic part of \( \mu \), Stone [4] proposed replacing \( \theta(l) \) by \( \hat{\theta} \), where \( \hat{\theta} \) is defined as follows.

Recorder the \( (X_i, \theta_i) \) to obtain \( (X_{\bar{0}}, \theta_{\bar{0}}), 1 \leq i \leq \infty \). If
\[
\|X_{\bar{0}} - X\| = \cdots = \|X_{\bar{0}} - X\|
\]

then let \( \hat{\theta} \) be the integer most frequently occurring among \( \theta_{l_1}, \cdots, \theta_{l_k} \) (in case of a tie, \( \hat{\theta} \) is taken arbitrarily among the integers involved in the tie). Define
\[
L_n = P \left( \hat{\theta} = \theta_{l_i} \right) \quad 1 \leq i \leq \infty, \quad A = \text{set of atoms of } \mu
\]

and for general Borel sets \( B \) from \( \mathbb{R}^d \),
\[
R^*(B) = E I_{(X_{\bar{0}} \in B)} \min (\eta(X), 1 - \eta(X))
\]

and the Bayes probability of error is \( R^* = R^*(\mathbb{R}^d) \). In fact, \( R^*(\cdot) \) and \( L(\cdot) \) can be considered as finite measures on the
Boyle sets of IR^d, but this matter will not be pursued any further.

Theorem 3:

\[ L_n \xrightarrow{n \to \infty} R^A + L(A^c) \]

in probability and

\[ E\{L_n\} \xrightarrow{n \to \infty} R^A + L(A^c) \]

where A^c is the complement of A.

Remark 4: R^A(A) is the portion of the Bayes probability of error due to the atomic part of \( \mu; L(A^c) \) is the portion of the asymptotic nearest neighbor probability of error due to the nonatomic part of \( \mu \). Clearly,

\[ R^A(\mathbb{R}^d) \leq R^A(A) + L(A^c) \leq L(\mathbb{R}^d) = E(2\eta(X)(1 - \eta(X))) \]

(by the asymptotic nearest neighbor probability of error for \( \theta_{(1)} \)) so that, in a sense, \( \hat{\theta} \) is always better than \( \theta_{(1)} \).

Remark 5: If \( \mu \) is nonatomic, then \( L_n \approx L_{\eta/2} \) a.s. because \( \hat{\theta} = \theta_{(1)} \) a.s. Therefore, Theorem 2 is a corollary of Theorem 3.

Remark 6: If \( \mu \) is atomic, then \( L_n \to R^A \) a.s. by Lemma 4 below.

IV. LEMMAS NEEDED TO PROVE THEOREM 3

In this section, we give some lemmas, all of a measure-theoretical nature, that will be used further on. The proofs can be found in the Appendix.

Lemma 1 is an extension of the dominated convergence theorem.

Lemma 1 [3]: Let \( (f_n) \leq c < \infty \) be a sequence of Borel measurable functions of \( x, x_1, x_2, \ldots, x_n, \theta_n \), and let

\[ f_n(x, x_1, \theta_1, \ldots, x_n, \theta_n) \xrightarrow{n \to \infty} f(x) \text{ a.s.} \]

for almost all \( x \in \mathbb{R}^d \), then

\[ E\{f_n(X, X_1, \theta_1, \ldots, X_n, \theta_n) - f(X) \mid X_1, \theta_1, \ldots, X_n, \theta_n\} \xrightarrow{n \to \infty} 0 \text{ a.s.} \]

The sample \( X_1, \ldots, X_n \) partitions \( \mathbb{R}^d \) up into at most \( n \) sets \( A_1, \ldots, A_n \), where \( A_n \) is the collection of all \( x \in \mathbb{R}^d \) for which \( X_1 \) is the nearest neighbor among \( X_1, \ldots, X_n \). Lemma 2 below states that for nonatomic measures, the \( \mu \)-measure of these sets tends to 0 a.s. uniformly in \( x \) as \( n \to \infty \).

Lemma 2 [5]: If \( \mu \) is a nonatomic finite measure, then

\[ \sup_{1 \leq i \leq n} \mu(A_{In}) \xrightarrow{n \to \infty} 0 \text{ a.s.} \]

and

\[ E\{\sup_{1 \leq i \leq n} \mu(A_{In})\} \xrightarrow{n \to \infty} 0. \]

Remark 7: For any finite measure \( \mu \) we thus have

\[ \sup_{1 \leq i \leq n} \mu(A_{In} \cap A^c) \xrightarrow{n \to \infty} 0 \text{ a.s.} \]

and

\[ E\{\sup_{1 \leq i \leq n} \mu(A_{In} \cap A^c)\} \xrightarrow{n \to \infty} 0. \]

We will also need some result on the "separation" of the atomic and the nonatomic parts of \( \mu \). Lemma 3 below to some degree qualifies the statement that almost all "nonatomic" \( x \)'s have "nonatomic" nearest neighbors \( X_{(1)}^x \) with probability tending to 1 as \( n \to \infty \).

Lemma 3:

\[ P\{\|X_{(1)}^x - X\| = \|X_{(1)}^x - X\|, \ x \in A^c\} \xrightarrow{n \to \infty} 0. \]

Lemma 4:

\[ P\{\hat{\theta} \neq \theta, \ X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n\} \xrightarrow{n \to \infty} R^A(\hat{\theta}) \text{ a.s.} \]

Lemma 5:

\[ P\{\hat{\theta} \neq \theta, \ X \in A^c \mid X_1, \theta_1, \ldots, X_n, \theta_n\} \xrightarrow{n \to \infty} L(A^c) \]

in probability when

\[ P\{\theta_{(1)} \neq \theta, \ X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n\} \xrightarrow{n \to \infty} L(A^c) \]

in probability.

We can now handle the atomic and nonatomic parts of the probability of error separately. The basic results are that for \( X \in A \), \( \hat{\theta} \) is asymptotically Bayes (Lemma 4), and that for \( X \not\in A \), \( \hat{\theta} \) and \( \theta_{(1)} \) are asymptotically equivalent (Lemma 5).

Appendix

Proof of Lemma 2: We first recall that for any compact set \( K \subseteq \mathbb{R}^d \),

\[ V_n(K) = \sup_{K \subseteq \text{support}(\mu)} \|X_{(1)}^x - x\| \xrightarrow{n \to \infty} 0 \text{ a.s.} \] [5].

Let \( W_n = \sup_{x \in K} \mu(A_{In}) \). Arguing again as in [5], we have for arbitrary \( e > 0 \), for all \( n \geq 1 \) and for arbitrary compact \( K \),

\[ [V_n(K) < e] \subseteq [W_n < \mu(K) + \sup_K \mu(S_{x,e})] \]

where \( S_{x,e} \) is a closed sphere centered at \( x \) with radius \( e \). Choose \( K \) such that \( \mu(K) < \delta/2 \) where \( \delta > 0 \) is given. Since \( \sup_K \mu(S_{x,e}) \to 0 \) as \( e \to 0 \), it is clear that by choosing \( e \) sufficiently small we can ensure

\[ [V_n(K) < e] \subseteq [W_n < \delta]. \]

Hence, \( W_n \xrightarrow{n \to \infty} 0 \) a.s. and \( E\{W_n\} \xrightarrow{n \to \infty} 0. \)

Proof of Lemma 3: We will show that for almost all \( x \in A^c \),

\[ P\{\|X_{(1)}^x - x\| = \|X_{(1)}^x - x\| \xrightarrow{n \to \infty} 0, \text{ and Lemma 3} \}

then follows by the dominated convergence theorem.

Without loss of generality we can assume that \( x \in \text{support}(\mu) \) because \( \mu(\text{support}(\mu)) = 1 \). Let \( \nu \) be the measure on \([0, \infty)\) corresponding to \( X_1 - x \), and let \( A^* \) be the set of atoms of \( \nu \). Clearly,

\[ P\{\|X_{(1)}^x - x\| < \|X_{(1)}^x - x\| \xrightarrow{n \to \infty} 0, \text{ and Lemma 3} \}

\[ \leq P\{\|X_{(1)} - x\| > a\} + \sup_{0 < a} \mu([0, b]) \]

for arbitrary \( a \). The last term in this expression is arbitrarily small by choice of \( a \) [6, Corollary 10.50] and the first term tends to 0 as \( n \to \infty \) because \( x \in \text{support}(\mu) \). This concludes the proof of Lemma 3.

Proof of Lemma 4: When \( \eta(x) < 1 - \eta(x), \mu(\{x\}) > 0 \), we have

\[ P\{\hat{\theta} = 0, X = x \mid X_1, \theta_1, \ldots, X_n, \theta_n\} \xrightarrow{n \to \infty} \mu(\{x\}) \text{ a.s.} \]
by the strong law of large numbers. If $\eta(x) > 1 - \eta(x)$, then it tends to 0 a.s.

Thus,
\[ P \{ \hat{\theta} \neq 0, X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n \} \]
\[ = \sum_{x \in A} (\eta(x) P \{ \hat{\theta} = 0, X = x \mid X_1, \theta_1, \ldots, X_n, \theta_n \} + (1 - \eta(x)) P \{ \hat{\theta} = 1, X = x \mid X_1, \theta_1, \ldots, X_n, \theta_n \}) \]
\[ \xrightarrow{n \to \infty} \sum_{x \in A} \mu(x) \min(\eta(x), 1 - \eta(x)) \text{ a.s.} \]

by Lemma 1.

Proof of Lemma 5:
\[ |P \{ \hat{\theta} \neq 0, X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n \} - P \{ \theta_0 \neq 0, X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n \}| \]
\[ \leq P \{ \| X_0^X - X \| = \| X_0^X \|, \}
\[ X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n \} \]
\[ \xrightarrow{n \to \infty} 0 \]

in probability (Lemma 3).

Proof of Theorem 3: Let us define $L_n(1), L_n(2), L_2^*(2)$ as follows:
\[ L_n = P \{ \hat{\theta} \neq 0, X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n \} \]
\[ + P \{ \theta_0 \neq 0, X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n \} \]
\[ = L_n(1) + L_2^*(2) \]
(11)
and $L_n(2) = P \{ \theta_0 \neq 0, X \in A \mid X_1, \theta_1, \ldots, X_n, \theta_n \}$. We have seen that $L_n(1) \to R^*(A)$ a.s. (Lemma 4). $L_n(2)$ can be rewritten as
\[ E[I_{X \in A} \eta(X) I_{\theta_0 = 1} \mid X_1, \theta_1, \ldots, X_n, \theta_n] \]
\[ + E[I_{X \in A} \eta(X) (1 - \eta(X)) I_{\theta_0 = 1} \mid X_1, \theta_1, \ldots, X_n, \theta_n] \]
(12)
and each term of (12) converges in probability to $\frac{1}{2} L(A^*)$ as we will see below. Because $|L_n(2) - L_2^*(2)| \xrightarrow{n \to \infty} 0$ in probability (Lemma 5), we will have shown Theorem 3.

Consider the first term of (12):
\[ E[I_{X \in A} \eta(X) I_{\theta_0 = 1} \mid X_1, \theta_1, \ldots, X_n, \theta_n] \]
\[ - E[I_{X \in A} \eta(X) (1 - \eta(X)) \mid X_1, \theta_1, \ldots, X_n, \theta_n] \]
\[ \leq E[|\eta(X_1) - \eta(X)| \mid X_1, \theta_1, \ldots, X_n, \theta_n] \]
\[ + |E[I_{X \in A} \eta(X) I_{\theta_0 = 1} \mid X_1, \theta_1, \ldots, X_n, \theta_n] - E[I_{X \in A} \eta(X) (1 - \eta(X)) \mid X_1, \theta_1, \ldots, X_n, \theta_n]| \]
\[ = U_n(1) + U_n(2). \]
(13)

Cleary, $E \{ U_n(1) \} \xrightarrow{n \to \infty} 0$ by Remark 2. If $A_{in}$ is defined as in Lemma 2, and
\[ \nu(B) = \int_B \eta(x) \mu(dx), \quad \text{for } B \text{ a Borel set of } \mathbb{R}^d, \]
then
\[ U_n(2) = \sum_{i=1}^n (I_{[\eta_i - 1]} - \eta(X_i)) \nu(A_{in}) \]
and
\[ E(U_n^2(2)) = E[E(U_n^2(2) \mid X_1, X_2, \ldots, X_n)] \]
\[ = E\left[ \sum_{i=1}^n \eta(X_i) (1 - \eta(X_i)) \nu^2(A_{in}) \right] \]
\[ < E\left[ \sup_{1 \leq i \leq n} \nu(A_{in}) \right] \xrightarrow{n \to \infty} 0 \quad \text{(Lemma 2).} \]

Thus, $L_n(2) \xrightarrow{n \to \infty} L(A^*)$ in probability, thereby completing the proof of Theorem 3.

REFERENCES


Luc Devroye was born in Tienen, Belgium, on August 6, 1946. He received the Ph.D. degree from the University of Texas at Austin in 1976. In 1977 he became an Assistant Professor at the School of Computer Science, McGill University, Montreal, P.O., Canada. He is interested in various applications of probability theory and mathematical statistics such as nonparametric estimation, probabilistic algorithms, the computer generation of random numbers, and the strong convergence of random processes.