Probability of Error, Equivocation, and the Chernoff Bound

MARTIN E. HELLMAN, MEMBER, IEEE, AND JOSEF RAVIV, MEMBER, IEEE

Abstract—Relationships between the probability of error, the equivocation, and the Chernoff bound are examined for the two-hypothesis decision problem. The effect of rejections on these bounds is derived. Finally, the results are extended to the case of any finite number of hypotheses.

I. Introduction

LET US consider the usual decision-theory problem of classifying an observation X as coming from one of m possible classes (hypotheses) $C_1, C_2, \cdots, C_m$. Let $\pi_1, \cdots, \pi_m$ denote the a priori probabilities on the hypotheses, and let $p_1(x), \cdots, p_m(x)$ denote the conditional probability density functions given the true hypothesis. Let us assume that these are known. Then it is well known that the decision rule that minimizes the probability of error $P(e)$ is the Bayes decision rule; i.e., choose the hypothesis with the largest a posteriori probability. Although $P(e)$ can theoretically be calculated, this computation is often impractical [1]. In such cases bounds on $P(e)$ that are easy to calculate are desirable, and several bounds have been presented in the literature [3], [4] for the two-class decision problem ($m = 2$).

In this paper the Bhattacharyya bound [1] and the more general Chernoff bound [2], [4] are examined. Section II gives simple derivations of these bounds for the two-class problem. Section III explores the connection between $P(e)$ and the equivocation I. In particular, it is shown that $P(e) \leq (\frac{1}{2})I$ for the two-class problem. Furthermore, using Chernoff-type bounds on I, we obtain an alternative proof of the Chernoff bound on $P(e)$. In Section IV this method of proof yields tighter bounds on $P(e)$ when rejections (erasures) are allowed. Finally in Section V these results are extended to any finite number of hypotheses. In the Appendix we explore a fine point that relates to the Chernoff bound on I. It is shown that the bound need not hold, unless the $p_i(x)$ have the same support.

II. Bounds for the Two-Class Problem

If only two classes are involved ($m = 2$), the Bhattacharyya bound states that

$$P(e) \leq \sqrt{\pi_1\pi_2}\,\rho \tag{1}$$

where[1]

$$\rho = \int_X \sqrt{p_1(x)p_2(x)}\,dx. \tag{2}$$

The following theorem, which gives the Chernoff bound on $P(e)$, is seen to include (1) as a special case. The proof is similar to that contained in [11].

Theorem 1

For any $\alpha \in [0, 1]$

$$P(e) \leq \pi_1^{\alpha}\pi_2^{1-\alpha} \int [p_1(x)]^{\alpha}[p_2(x)]^{1-\alpha}\,dx. \tag{3}$$

Proof: If x is observed, the posterior probability of class i is

$$\Pr(C_i \mid x) = \frac{\pi_i p_i(x)}{\pi_1 p_1(x) + \pi_2 p_2(x)} \qquad i = 1, 2. \tag{4}$$

To minimize $P(e)$ choose the class with the larger posterior probability. Therefore,

$$P(e \mid x) = \min\left\{\frac{\pi_1 p_1(x)}{\pi_1 p_1(x) + \pi_2 p_2(x)}, \frac{\pi_2 p_2(x)}{\pi_1 p_1(x) + \pi_2 p_2(x)}\right\} \tag{5}$$

and

$$P(e) = E_x[P(e \mid x)] = \int_X \min\{\pi_1 p_1(x), \pi_2 p_2(x)\}\,dx, \tag{6}$$

where E_x denotes expectation with respect to x. Since for $0 \leq \alpha \leq 1$, $a \geq 0$, $b \geq 0$,

$$\min\{a, b\} \leq a^{\alpha}b^{1-\alpha}, \tag{7}$$

it follows that

$$P(e) \leq \int_X [\pi_1 p_1(x)]^{\alpha}[\pi_2 p_2(x)]^{1-\alpha}\,dx. \tag{8}$$

Q.E.D.

Define

$$\rho^* = \inf_{0\leq\alpha\leq 1} \int [p_1(x)]^{\alpha}[p_2(x)]^{1-\alpha}\,dx \tag{9}$$

and

$$K_1 = \pi_1^{\alpha^*}\pi_2^{1-\alpha^*} \tag{10}$$

where α^* is the limiting value of α that minimizes (9). Then

$$P(e) \leq K_1\rho^*. \tag{11}$$

Manuscript received September 29, 1969; revised December 11, 1969. Portions of this paper were presented at the 1969 IEEE International Conference on Communications, Boulder, Colo., June 9–11, 1969.

M. E. Hellman was with the IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y. 10598. He is now with the Department of Electrical Engineering, Massachusettes Institute of Technology, Cambridge, Mass.

J. Raviv is with the IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y. 10598.

[1] If x is a discrete random variable this and future integrals should be interpreted as sums.

Note that (11) does not follow directly from (3) since ρ^* is an infimum not a minimum. The Appendix has a rigorous proof of (11).

Let $X = (X_1, \cdots, X_n)$ where the X_i are conditionally independent and identically distributed (i.i.d.) as $p_i(x)$. It is easily shown that $P_n(e)$, the probability of error of the Bayes decision based on X, is bounded by

$$P_n(e) \le K_1(\rho^*)^n. \tag{12}$$

This form is analogous to

$$P_n(e) \le \sqrt{\pi_1\pi_2}\, \rho^n \tag{13}$$

for the usual Bhattacharyya bound. Since $\rho^* \le \rho$, (12) is exponentially tighter, except when $\alpha^* = \frac{1}{2}$, in which case the two bounds are identical.

III. Equivocation

If x is observed, then

$$H(C \mid x) = -\sum_{i=1}^{m} \Pr\{C_i \mid x\} \log \Pr\{C_i \mid x\} \tag{14}$$

is the conditional entropy [9], where the logarithm is to the base 2. The equivocation I is defined by

$$I = E_x[H(C \mid x)]. \tag{15}$$

Rényi [5] has shown[2] that the missing information after n i.i.d. observations $I(n)$ obeys a bound similar to (9),

$$I(n) \le K_2(\rho^*)^n. \tag{16}$$

He also showed that

$$P_n(e) \le I(n) \tag{17}$$

so that by combining (16) and (17) one can obtain the Chernoff bound (12) within a multiplicative constant.

Let us note that the inequality (17) can be tightened by a factor of 2, i.e.,

$$P(e) \le I/2. \tag{18}$$

Proof: Since only two classes are involved

$$p(X) \equiv \Pr\{C_1 \mid X\} \tag{19}$$

completely determines the posterior distribution. Also $H(C \mid X)$ can be expressed as

$$H(p) = -p \log p - (1-p) \log(1-p), \tag{20}$$

where $p = p(X)$.

Since X is a random variable, so is $p = p(X)$. Therefore, (15) can be rewritten as an expectation over p:

$$I = E_p[H(p)]. \tag{21}$$

Similarly, it is seen that

$$P(e) = E_p[\min\{p, 1-p\}] \tag{22}$$

and, since

$$\min\{p, 1-p\} \le (\tfrac{1}{2})H(p), \tag{23}$$

as can be seen from Fig. 1, (18) follows immediately from (21)–(23).

IV. Bounds With Rejection Option

Thus far only two actions, decide C_1 or decide C_2, have been allowed. Forcing decisions on certain observations could result in a large $P(e)$. Therefore, it is desirable to have the option of making no decision at all, that is, rejecting the observation. Let

$$w(x) = \max\{p(x), 1 - p(x)\}, \tag{24}$$

where $p(x)$ is defined by (19). As shown by Chow [8], the optimal rejection criterion is to choose a threshold $t \ge \frac{1}{2}$ and reject whenever $w(x) < t$; i.e., reject if $P(e \mid x) > 1-t$. (In communication theory rejections are often called erasures [12].) The optimal value of t is determined by the relative costs of error and rejection [8].

Let $R(t)$ denote the rejection rate (probability of rejection), and let $f(w)$ denote the density of the random variable $w(x)$. Then

$$R(t) = \int_{1/2}^{t} f(w)\, dw \tag{25}$$

and

$$P(e \mid t) = \int_t^1 (1-w) f(w)\, dw. \tag{26}$$

Note that $R(t)$ is increasing in t and $P(e \mid t)$ is decreasing in t. Equation (26) may be rewritten as

$$P(e \mid t) = E_p\{g(p \mid t)\}, \tag{27}$$

where

$$g(p \mid t) = \begin{cases} \min\{p, 1-p\} & p \le 1-t \text{ or } p \ge t \\ 0 & 1-t < p < t. \end{cases} \tag{28}$$

It is easily seen that

$$g(p \mid t) \le C_t H(p), \tag{29}$$

where

$$C_t = (1-t)/[H(1-t)]. \tag{30}$$

Fig. 2 shows this pictorially.

Then

$$P(e \mid t) \le C_t I, \tag{31}$$

which follows from (21), (27), and (29).

The importance of (31) is that $C_t \le \frac{1}{2}$ for all t and if $t \to 1$, then $C_t \to 0$, as can be seen from the fact that $H(p)$ has infinite slope at the origin. Thus, for values of t near 1 (31) yields a tighter bound on $P(e \mid t)$.

Using this development, it is also possible to show that the rejection rate after n observations $R_n(t)$ decays exponentially as $(\rho^*)^n$. This can be seen by observing that

$$R_n(t) \le \frac{P_n(e)}{1-t}, \tag{32}$$

[2] The proof and result given by Rényi are not correct in certain cases. The Appendix elaborates on this point.

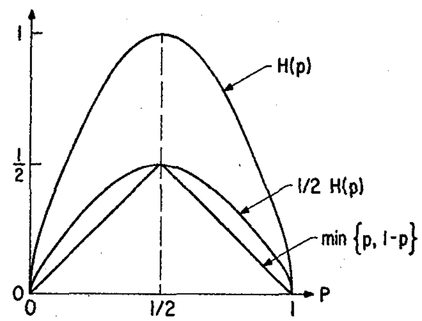

Fig. 1. Entropy and error of posterior distribution.

where $P_n(e)$ is the probability of error without rejections; i.e., $t = \frac{1}{2}$.

To derive (32) note that (dropping the subscript n)

$$R(t) = E_p[\mathcal{I}_{(1-t,t)}(p)], \tag{33}$$

where $\mathcal{I}_{(1-t,t)}(p)$ is the indicator function for the open interval $(1 - t, t)$. Now

$$P(e) = E_p[\min \{p, 1 - p\}] \tag{22}$$

implies

$$P(e) \geq E_p[\mathcal{I}_{(1-t,t)}(p) \min \{p, 1 - p\}]$$

$$\geq (1 - t)E_p[\mathcal{I}_{(1-t,t)}(p)] \tag{34}$$

or

$$P(e) \geq (1 - t)R(t). \tag{35}$$

But (35) is just a restatement of (32).

V. Extension to More Than Two Classes

The exponential bound (16) on $I(n)$ has been extended by Rényi [7] to the case where $m > 2$. He shows that if

$$\rho_{ij} \equiv \int_X \sqrt{p_i(x)p_j(x)}\, dx \tag{36}$$

and

$$q \equiv \max_{i \neq j} \rho_{ij}, \tag{37}$$

then

$$I(n) \leq K_3 q^n. \tag{38}$$

He also proves that $P_n(e)$ is bounded above by a constant times q^n.

Following are some extensions of Rényi's results. First, except for pathological cases like that discussed in the Appendix, (38) can be tightened to

$$I(n) \leq K_4(\rho^{**})^n, \tag{39}$$

where ρ^{**} is defined by

$$\rho_{ij}^* = \inf_{0 \leq \alpha \leq 1} \int_X [p_i(x)]^\alpha [p_j(x)]^{1-\alpha}\, dx \tag{40a}$$

$$\rho^{**} = \max_{i \neq j} \rho_{ij}^*. \tag{40b}$$

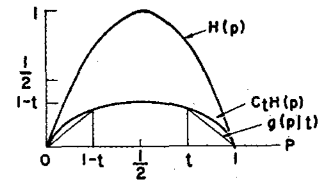

Fig 2. Entropy and error with rejection option.

The proof of (39) is much the same as the proof of (38) contained in [7], merely substituting p_i^α for $p_i^{1/2}$. Furthermore, in [7] it is not shown that $P(e)$ is bounded above by a constant times I (for $m > 2$). Such a proof exists in the literature [10], [13], but an alternative proof is given here.

That is, we will show that

$$P(e) \leq \tfrac{1}{2} I \tag{41}$$

for all integers $m \geq 2$, thereby proving that $I(n)$ decreases exponentially as $(\rho^{**})^n$.

The proof will depend on the fact that if X results in posterior distribution $\Pr \{C_i \mid x\} \equiv p(i \mid x)$ for $i = 1, \cdots, m$, then

$$P(e \mid x) = 1 - \max_i \{p(i \mid x)\}. \tag{42}$$

Since

$$P(e) = E_x[1 - \max_i \{p(i \mid x\}] \tag{43}$$

and

$$I = E_x[H(p(1 \mid x), \cdots, p(m \mid x))], \tag{44}$$

it is sufficient to prove that if $\sum_{i=1}^m a_i = 1$ and $a_i \geq 0$, then

$$1 - \max_i \{a_i\} \leq (\tfrac{1}{2})H(a_1, \cdots, a_m) \tag{45}$$

in order to prove (41).

The proof of (45) will proceed by induction. For $m = 2$ (45) is equivalent to (23), which has already been established. Therefore, under the assumption that (45) is true for m, we must show it to be true for $m + 1$ to complete the proof.

Let $(a_1, a_2, \cdots, a_m, a_{m+1})$ be a probability distribution on $m + 1$ classes. Assume, without loss of generality, that the a_i have been reordered in such a way that a_{m+1} is the largest. Now consider the m vector $(a_1, a_2, \cdots, a_m + a_{m+1})$. From the assumption that (45) is true for m

$$1 - (a_m + a_{m+1})$$

$$\leq \tfrac{1}{2}H(a_1, a_2, \cdots, a_{m-1}, a_m + a_{m+1}). \tag{46}$$

Using the grouping axiom [9]

$$H(a_1, a_2, \cdots, a_{m-1}, a_m, a_{m+1})$$

$$= H(a_1, a_2, \cdots, a_{m-1}, a_m + a_{m+1})$$

$$+ (a_m + a_{m+1})H\left(\frac{a_m}{a_m + a_{m+1}}, \frac{a_{m+1}}{a_m + a_{m+1}}\right). \tag{47}$$

Further, since (45) is true for $m = 2$ (or by (23))

$$\frac{a_m}{a_m + a_{m+1}} \leq (\tfrac{1}{2})H\left(\frac{a_m}{a_m + a_{m+1}}, \frac{a_{m+1}}{a_m + a_{m+1}}\right). \tag{48}$$

Combining (46)–(48) yields

$$1 - a_{m+1} \leq (\tfrac{1}{2})H(a_1, a_2, \cdots, a_m, a_{m+1}), \tag{49}$$

completing the proof of (41).

It should be noted that an alternative proof of the ρ^{**} bound on $P(e)$ is possible using the union bound [14] and the bound (11) on $P(e)$ for $m = 2$. From the alternative proof it also follows that, for any $\epsilon > 0$ there exists an $N(\epsilon)$ such that, for $n > N(\epsilon)$

$$P_n(e) \leq (2 + \epsilon)(\rho^{**})^n \tag{50}$$

as long as ρ^{**} is achieved by only one pair i, j. Note that this proof is valid even for the pathological cases excluded from (39), and discussed in the Appendix.

Appendix

Extreme-Point Infima

Consider the two-class problem in which $\pi_1 = \pi_2 = \frac{1}{2}$ and

$$p_1(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases};$$
$$p_2(x) = \begin{cases} \frac{1}{2} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}. \tag{51}$$

Then for $0 < \alpha < 1$

$$\rho(\alpha) \equiv \int_X p_1^\alpha(x) p_2^{1-\alpha}(x)\, dx = (\tfrac{1}{2})^{1-\alpha} \tag{52}$$

and

$$\rho^* = \inf_{0<\alpha<1} \rho(\alpha) = \tfrac{1}{2}.$$

Thus (12) and (16) would predict the existence of finite constants K_1 and K_2 such that

$$P_n(e) \leq K_1(\tfrac{1}{2})^n \tag{53}$$

and

$$I(n) \leq K_2(\tfrac{1}{2})^n. \tag{54}$$

Evaluation of $P_n(e)$ results in

$$P_n(e) = (\tfrac{1}{2})^{n+1} \tag{55}$$

so that (53) is a valid bound with $K_1 = \frac{1}{2}$. However, evaluation of $p(\boldsymbol{x})$ as defined by (19) yields only two possible values for $p(\boldsymbol{x})$. If any of $x_1, x_2, \cdots, x_n$ are greater than 1, then the $\boldsymbol{x}$ must, with probability 1, be drawn according to $p_2(x)$. In this case $p(\boldsymbol{x}) = 0$ and evaluation shows that

$$\Pr\{p(\boldsymbol{x}) = 0\} = \pi_2[1 - (\tfrac{1}{2})^n] = (\tfrac{1}{2})[1 - (\tfrac{1}{2})^n]. \tag{56}$$

If, on the other hand, all of the observations are less than 1, $p(\boldsymbol{x}) = \pi_1/[\pi_1 + \pi_2(\frac{1}{2})^n] \equiv \gamma_n$. In this problem $\gamma_n = 2^n/(1 + 2^n)$ and $\Pr\{p(\boldsymbol{x}) = \gamma_n\} = \pi_1 + \pi_2(\frac{1}{2})^n = (\frac{1}{2})[1 + (\frac{1}{2})^n]$. From (21)

$$\begin{aligned} I(n) &= (\tfrac{1}{2})[1 - (\tfrac{1}{2})^n]H(0) \\ &\quad + (\tfrac{1}{2})[1 + (\tfrac{1}{2})^n]H(\gamma_n) \\ &= (\tfrac{1}{2})[1 + (\tfrac{1}{2})^n]H(\gamma_n) \\ &\geq (\tfrac{1}{2})(1 - \gamma_n)\log[1/(1 - \gamma_n)] \\ &= (\tfrac{1}{2})[1/(1 + 2^n)]\log(1 + 2^n) \\ &\geq (\tfrac{1}{2})^{n+2}\log 2^n \\ &= n(\tfrac{1}{2})^{n+2} \end{aligned} \tag{57}$$

or in summary

$$I(n) \geq (n/4)(\tfrac{1}{2})^n. \tag{58}$$

But (58) shows explicitly that (54) is not true.

The reason for this is that, although $I(n) \leq K_\alpha[\rho(\alpha)]^n$ for any $0 < \alpha < 1$, in this problem there is no α for which $\rho(\alpha) = \frac{1}{2}$. The value $\rho^* = \frac{1}{2}$ is only approached in the limit as $\alpha \to 0$. However, if $\alpha = 0$, then obviously $\rho(\alpha) \equiv 1$. The problem with Rényi's reasoning that led him to (54) is that $\rho(\alpha)$ need not be continuous at $\alpha = 0$ and $\alpha = 1$, and so ρ^* need not be achieved by any $\alpha^*(0 \leq \alpha^* \leq 1)$.

At this point one might ask why $P_n(e)$ does obey (53), since all that has been shown so far is that $P_n(e) \leq K_\alpha[\rho(\alpha)]^n$, the same as had been done for $I(n)$. The difference is that for $a \geq 0$, $b \geq 0$

$$\min\{a, b\} \leq \lim_{\alpha\to 0} a^\alpha b^{1-\alpha}$$

even if $a = 0$ or $b = 0$. Thus

$$\begin{aligned} P(e) &= \int_X \min\{\pi_1 p_1(x), \pi_2 p_2(x)\}\, dx \\ &\leq \int_X \lim_{\alpha\to 0} \pi_1^\alpha p_1^\alpha(x) \pi_2^{1-\alpha} p_2^{1-\alpha}(x)\, dx. \end{aligned} \tag{59}$$

Then, since for all $0 < \alpha < 1$

$$\begin{aligned} \pi_1^\alpha p_1^\alpha(x)\pi_2^{1-\alpha}p_2^{1-\alpha}(x) &\leq \max\{\pi_1 p_1(x), \pi_2 p_2(x)\} \\ &\leq \pi_1 p_1(x) + \pi_2 p_2(x) \end{aligned} \tag{60}$$

and $p_1(x)$ and $p_2(x)$ are integrable, the Lebesgue convergence theorem [15] allows the integral and limit to be interchanged in (59), yielding

$$P(e) \leq \pi_2 \lim_{\alpha\to 0} \int_X p_1^\alpha(x) p_2^{1-\alpha}(x)\, dx$$

or

$$P(e) \leq \pi_2 \lim_{\alpha\to 0} \rho(\alpha). \tag{61}$$

A similar proof exists for $\alpha \to 1$.

It should be noted that the equivocation does obey (54) provided that $p_1(x) = 0$ if and only if $p_2(x) = 0$ (a.e.). That this is so can be deduced from the following.

Theorem 2

For $0 < \alpha < 1$, $\rho(\alpha)$ is continuous.

Proof: For $0 < \alpha < 1$ and all x

$$\lim_{\Delta \to 0} p_1^{\alpha+\Delta}(x) p_2^{1-\alpha-\Delta}(x) = p_1^{\alpha}(x) p_2^{1-\alpha}(x). \tag{62}$$

Then since for $0 < \alpha + \Delta < 1$

$$p_1^{\alpha+\Delta}(x) p_2^{1-\alpha-\Delta}(x) \leq p_1(x) + p_2(x) \tag{63}$$

and $p_1(x) + p_2(x)$ is integrable, application of the Lebesgue convergence theorem yields

$$\lim_{\Delta \to 0} \rho(\alpha + \Delta) = \rho(\alpha). \tag{64}$$

Q.E.D.

Theorem 3

The function $\rho(\alpha)$ is convex $\cup$.

Remark: Since $\rho(\alpha)$ need not be continuous at $\alpha = 0$ and $\alpha = 1$, we cannot proceed merely by proving $\partial^2\rho(\alpha)/\partial\alpha^2 \geq 0$. Therefore, let us use an equivalent condition for convexity. Given $0 \leq \alpha_1 < \alpha_2 < \alpha_3 \leq 1$ so that

$$\alpha_2 = k\alpha_1 + (1 - k)\alpha_3, \tag{65}$$

where $0 < k < 1$, then we show

$$\rho(\alpha_2) \leq k\rho(\alpha_1) + (1 - k)\rho(\alpha_3). \tag{66}$$

Proof: For all x

$$p_1^{\alpha_2}(x) p_2^{1-\alpha_2}(x) \leq k p_1^{\alpha_1}(x) p_2^{1-\alpha_1}(x) + (1 - k) p_1^{\alpha_3}(x) p_2^{1-\alpha_3}(x). \tag{67}$$

Integrating (67) yields (66). Q.E.D.

Theorem 4

If $p_1(x) = 0$ when and only when $p_2(x) = 0$, then $\rho(\alpha)$ is continuous at $\alpha = 0$ and $\alpha = 1$.

Proof: Obviously $\rho(0) = \rho(1) \equiv 1$. Further, since $p_1(x)$ and $p_2(x)$ have the same support. if we define $X^1 \subseteq X$ as

$$X^1 = \{x \in X : p_1(x) > 0 \text{ and } p_2(x) > 0\} \tag{68}$$

then

$$\rho(\alpha) = \int_{X^1} p_1^{\alpha}(x) p_2^{1-\alpha}(x)\, dx. \tag{69}$$

Now for all $x \in X^1$

$$\lim_{\alpha \to 0} p_1^{\alpha}(x) p_2^{1-\alpha}(x) = p_2(x). \tag{70}$$

So using the Lebesgue convergence theorem yields

$$\lim_{\alpha \to 0} \rho(\alpha) = \int_{X^1} p_2(x) = 1.$$

A similar proof holds for $\alpha \to 1$. Q.E.D.

Taken together, Theorems 2–4, tell us:

1) for $0 \leq \alpha \leq 1$, $\rho(\alpha) \leq 1$;
2) if $p_1(x) \neq p_2(x)$ a.e., then for $0 < \alpha < 1$, $\rho(\alpha) < 1$;
3) if $p_1(x)$ and $p_2(x)$ have the same support, then

$$\inf_{0 \leq \alpha \leq 1} \rho(\alpha) = \min_{0 < \alpha < 1} \rho(\alpha). \tag{71}$$

That is, there exists $0 < \alpha^* < 1$ such that for all $0 \leq \alpha \leq 1$

$$\rho(\alpha^*) \leq \rho(\alpha). \tag{72}$$

Thus, (16) is valid in this case. However, the example given shows that (16) need not be valid otherwise.

References

[1] T. Kailath, "The divergence and Bhattacharyya distance in signal selection," *IEEE Trans. Communications Technology*, vol. COM-15, pp. 52–60, February 1967.

[2] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, vol. 23, pp. 493–507, 1952.

[3] H. Kobayashi and J. Thomas, "Distance measures and related criteria," *Proc. 5th Ann. Allerton Conf. on Circuits and Systems*, (University of Illinois, Urbana), pp. 491–500, 1967.

[4] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Stat.*, vol. 28, pp. 131–142, 1966.

[5] A. Rényi, "On the amount of missing information and the Neyman–Pearson lemma," in *Festschrift for J. Neyman*. New York: Wiley, 1966, pp. 281–288.

[6] ——, "On some basic problems of statistics from the point of view of information theory," *Proc. 5th Berkeley Symp. on Math. Stat. and Probability*, vol. 1. Berkeley, Calif: University of California Press, 1967, pp. 531–543.

[7] ——, "On the amount of information concerning an unknown parameter in a sequence of observations," *Magyar Tud. Akad. Mat. Kutató Int. Köyl*, vol. 9, pp. 617–625, 1965.

[8] a. C. K. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electronic Computers*, vol. EC-6, pp. 247–254, December 1957.
b. ——, "On optimum recognition error and reject tradeoff," IBM Res. RC 2408 (11712); also *IEEE Trans. Information Theory*, vol. IT-16, pp. 41–46, January 1970.

[9] R. Ash, *Information Theory*. New York: Interscience, 1965, pp. 8, 24.

[10] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968, pp. 520–521.

[11] C. Shannon, R. Gallagher, and E. Berlekamp, "Lower bounds to error probabilities for coding on discrete memoryless channels—I," *Inform. and Control*, vol. 10, pp. 65–103, 1967.

[12] G. D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Information Theory*, vol. IT-14, pp. 206–220, March 1968.

[13] J. Chu and J. Chueh, "Inequalities between information measures and error probability," *J. Franklin Inst.*, vol. 282, pp. 121–125, August 1966.

[14] J. Wozencraft and I. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965, p. 264.

[15] H. Royden, *Real Analysis*. New York: Macmillan, 1963, p. 76.