Divergence Measures Based on the Shannon Entropy

Jianhua Lin, *Member, IEEE*

***Abstract*—A new class of information-theoretic divergence measures based on the Shannon entropy is introduced. Unlike the well-known Kullback divergences, the new measures do not require the condition of absolute continuity to be satisfied by the probability distributions involved. More importantly, their close relationship with the variational distance and the probability of misclassification error are established in terms of bounds. These bounds are crucial in many applications of divergence measures. The new measures are also well characterized by the properties of nonnegativity, finiteness, semiboundedness, and boundedness.**

***Index Terms*—Divergence, dissimilarity measure, discrimination information, entropy, probability of error bounds.**

I. Introduction

Many information-theoretic divergence measures between two probability distributions have been introduced and extensively studied [2], [7], [12], [15], [17], [19], [20], [30]. The applications of these measures can be found in the analysis of contingency tables [10], in approximation of probability distributions [6], [16], [21], in signal processing [13], [14], and in pattern recognition [3]–[5]. Among the proposed measures, one of the best known is the I directed divergence [17], [19] or its symmetrized measure, the J divergence. Although the I and J measures have many useful properties, they require that the probability distributions involved satisfy the condition of absolute continuity [17]. Also, there are certain bounds that neither I nor J can provide for the variational distance and the Bayes probability of error [28], [31]. Such bounds are useful in many decisionmaking applications [3], [5], [11], [14], [31].

In this correspondence, we introduce a new directed divergence that overcomes the previous difficulties. We will show that this new measure preserves most of the desirable properties of I and is in fact closely related to I. Both the lower and upper bounds of the new divergence will also be established in terms of the variational distance. A symmetric form of the new directed divergence can be defined in a similar way as J, defined in terms of I. The behavior of I, J and the new divergences will be compared.

Based on Jensen's inequality and the Shannon entropy, an extension of the new measure, the Jensen–Shannon divergence, is derived. One of the salient features of the Jensen–Shannon divergence is that we can assign a different weight to each probability distribution. This makes it particularly suitable for the study of decision problems where the weights could be the prior probabilities. In fact, it provides both the lower and upper bounds for the Bayes probability of misclassification error.

Most measures of difference are designed for *two* probability distributions. For certain applications such as in the study of taxonomy in biology and genetics [24], [25], one is required to measure the overall difference of more than two distributions. The Jensen–Shannon divergence can be generalized to provide such a measure for any finite number of distributions. This is also useful in multiclass decisionmaking. In fact, the bounds provided by the Jensen–Shannon divergence for the two-class case can be extended to the general case.

The generalized Jensen–Shannon divergence is related to the Jensen difference proposed by Rao [23], [24] in a different context. Rao's objective was to obtain different measures of diversity [24] and the Jensen difference can be defined in terms of information measures other than the Shannon entropy function. No specific detailed discussion was provided for the Jensen difference based on the Shannon entropy.

Manuscript received October 24, 1989; revised April 20, 1990.

The author is with the Department of Computer Science, Brandeis University, Waltham, MA, 02254.

IEEE Log Number 9038865.

II. The Kullback I and J Divergence Measures

Let X be a discrete random variable and let p_1 and p_2 be two probability distributions of X. The I directed divergence [17], [19] is defined as

$$I(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)}. \tag{2.1}$$

The logarithmic base 2 is used throughout this correspondence unless otherwise stated. It is well known that $I(p_1, p_2)$ is nonnegative, additive but not symmetric [12], [17]. To obtain a symmetric measure, one can define

$$\begin{aligned} J(p_1, p_2) &= I(p_1, p_2) + I(p_2, p_1) \\ &= \sum_{x \in X} (p_1(x) - p_2(x)) \log \frac{p_1(x)}{p_2(x)}, \end{aligned} \tag{2.2}$$

which is called the J divergence [22]. Clearly, I and J divergences share most of their properties.

It should be noted that $I(p_1, p_2)$ is undefined if $p_2(x) = 0$ and $p_1(x) \neq 0$ for any $x \in X$. This means that distribution p_1 has to be *absolutely continuous* [17] with respect to distribution p_2 for $I(p_1, p_2)$ to be defined. Similarly, $J(p_1, p_2)$ requires that p_1 and p_2 be absolutely continuous with respect to each other. This is one of the problems with these divergence measures.

Effort [18], [27], [28] has been devoted to finding the relationship (in terms of bounds) between the I directed divergence and the variational distance. The variational distance between two probability distributions is defined as

$$V(p_1, p_2) = \sum_{x \in X} |p_1(x) - p_2(x)|, \tag{2.3}$$

which is a distance measure satisfying the metric properties. Several lower bounds for $I(p_1, p_2)$ in terms of $V(p_1, p_2)$ have been found, among which the *sharpest* known is given by

$$I(p_1, p_2) \geq \max\{L_1(V(p_1, p_2)), L_2(V(p_1, p_2))\}, \tag{2.4}$$

where

$$L_1(V(p_1, p_2)) = \log \frac{2 + V(p_1, p_2)}{2 - V(p_1, p_2)} - \frac{2V(p_1, p_2)}{2 + V(p_1, p_2)}, \quad 0 \leq V(p_1, p_2) \leq 2, \tag{2.5}$$

established by Vajda [28] and

$$L_2(V(p_1, p_2)) = \frac{V^2(p_1, p_2)}{2} + \frac{V^4(p_1, p_2)}{36} + \frac{V^6(p_1, p_2)}{288}, \quad 0 \leq V(p_1, p_2) \leq 2, \tag{2.6}$$

derived by Toussaint [27].

However, no general upper bound exists for either $I(p_1, p_2)$ or $J(p_1, p_2)$ in terms of the variational distance [28]. This is another difficulty in using the I directed divergence as a measure of difference between probability distributions [16], [31].

III. A New Directed Divergence Measure

In an attempt to overcome the problems of I and J divergences, we define a new directed divergence between two distri-

0018-9448/91/0100-0145$01.00 © 1991 IEEE

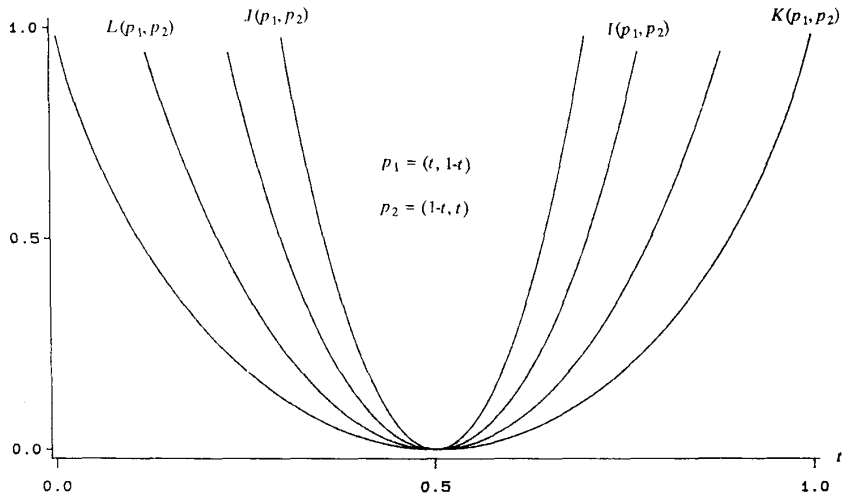

Fig. 1. Comparison of I, J, K, and L divergence measures.

butions p_1 and p_2 as

$$K(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{\frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)}. \quad (3.1)$$

This measure turns out to have numerous desirable properties. It is also closely related to I. From the Shannon inequality [1, p. 37], we know that, $K(p_1, p_2) \geq 0$ and $K(p_1, p_2) = 0$ *if and only if* $p_1 = p_2$, which is essential for a measure of difference. It is clear that $K(p_1, p_2)$ is well defined and independent of the values of $p_1(x)$ and $p_2(x)$, $x \in X$.

From both the definitions of K and I, it is easy to see that $K(p_1, p_2)$ can be described in terms of $I(p_1, p_2)$:

$$K(p_1, p_2) = I\left(p_1, \frac{1}{2}p_1 + \frac{1}{2}p_2\right). \quad (3.2)$$

The following relationship can also be established between I and K.

Theorem 1: The K directed divergence is bounded by the I divergence:

$$K(p_1, p_2) \leq \frac{1}{2} I(p_1, p_2). \quad (3.3)$$

Proof: Since $p_1(x) \geq 0$ and $p_2(x) \geq 0$ for any $x \in X$, by the inequality of the arithmetic and geometric means, we have

$$\frac{p_1(x) + p_2(x)}{2} \geq \sqrt{p_1(x) p_2(x)}, \quad x \in X.$$

Thus, it follows

$$K(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{\frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)}$$

$$\leq \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{\sqrt{p_1(x) p_2(x)}}$$

$$= \frac{1}{2} \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)} = \frac{1}{2} I(p_1, p_2). \quad \square$$

$K(p_1, p_2)$ is obviously not a symmetric measure. We can define a symmetric divergence based on K as:

$$L(p_1, p_2) = K(p_1, p_2) + K(p_2, p_1). \quad (3.4)$$

The L divergence is related to the J divergence in the same way as K is related to I. From inequality (3.3), we can easily derive the following relationship,

$$L(p_1, p_2) \leq \frac{1}{2} J(p_1, p_2). \quad (3.5)$$

A graphical comparison of I, J, K, and L divergences is shown in Fig. 1 in which we assume $p_1 = (t, 1-t)$ and $p_2 = (1-t, t)$, $0 \leq t \leq 1$. I and J have a steeper slope than K and L. It is important to note that I and J approach infinity when t approaches 0 or 1. In contrast, K and L are well defined in the entire range $0 \leq t \leq 1$.

Theorem 2: The following lower bound holds for the K directed divergence:

$$K(p_1, p_2) \geq \max\left\{ L_1\left(\frac{V(p_1, p_2)}{2}\right), L_2\left(\frac{V(p_1, p_2)}{2}\right)\right\}, \quad (3.6)$$

where L_1 and L_2 are defined by (2.5) and (2.6), respectively.

Proof: From equality (3.2) and inequality (2.4), we have

$$K(p_1, p_2) \geq \max\left\{ L_1\left(V\left(p_1, \frac{1}{2}p_1 + \frac{1}{2}p_2\right)\right), L_2\left(V\left(p_1, \frac{1}{2}p_1 + \frac{1}{2}p_2\right)\right)\right\}.$$

Since

$$V\left(p_1, \frac{1}{2}p_1 + \frac{1}{2}p_2\right) = \sum_{x \in X} \left| p_1(x) - \left(\frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)\right)\right|$$

$$= \frac{1}{2} V(p_1, p_2),$$

(3.6) follows immediately. $\square$

In contrast to situations for the I and J divergences, upper bounds also exist for the L divergence in terms of the variational distance.

Theorem 3: The variational distance and the L divergence measure satisfy the inequality:

$$L(p_1, p_2) \leq V(p_1, p_2). \quad (3.7)$$

Proof: From the definition of $L(p_1, p_2)$ given by (3.4), we have

$$L(p_1,p_2) = \sum_{x\in X} p_1(x)\log\frac{p_1(x)}{\frac{1}{2}p_1(x)+\frac{1}{2}p_2(x)} + \sum_{x\in X} p_2(x)\log\frac{p_2(x)}{\frac{1}{2}p_1(x)+\frac{1}{2}p_2(x)} \tag{3.8}$$

$$= \sum_{x\in X}(p_1(x)+p_2(x))\cdot\left(\frac{p_1(x)}{p_1(x)+p_2(x)}\log\frac{2p_1(x)}{p_1(x)+p_2(x)} + \frac{p_2(x)}{p_1(x)+p_2(x)}\log\frac{2p_2(x)}{p_1(x)+p_2(x)}\right)$$

$$= \sum_{x\in X}(p_1(x)+p_2(x))\cdot\left(1-H\left(\frac{p_1(x)}{p_1(x)+p_2(x)},\frac{p_2(x)}{p_1(x)+p_2(x)}\right)\right). \tag{3.9}$$

It has been proved in [8, p. 521] that, for any $0 \le a \le 1$,

$$H(a,1-a) \ge 2\min(a,1-a).$$

Since

$$\min(a,1-a) = \frac{1}{2}(1-|a-(1-a)|),$$

it follows that

$$1-H(a,1-a) \le |a-(1-a)|.$$

Thus, from (3.9),

$$L(p_1,p_2) \le \sum_{x\in X}(p_1(x)+p_2(x))\left|\frac{p_1(x)}{p_1(x)+p_2(x)} - \frac{p_2(x)}{p_1(x)+p_2(x)}\right| = V(p_1,p_2). \quad \square$$

Since $K(p_1, p_2)$ is clearly not greater than $L(p_1, p_2)$, from Theorem 3 we immediately obtain the following bound for the K divergence:

$$K(p_1,p_2) \le V(p_1,p_2). \tag{3.10}$$

Thus, the variational distance serves as an upper bound to both the K and L divergences.

The K and L divergences have several other desirable properties. As we mentioned earlier, both K and L are *nonnegative*, which is essential for being measures of difference. They are also *finite* and *semibounded*, that is,

$$K(p_1,p_2) < +\infty, \quad K(p_1,p_2) \ge K(p_1,p_1); \tag{3.11}$$

$$L(p_1,p_2) < +\infty, \quad L(p_1,p_2) \ge L(p_1,p_1), \tag{3.12}$$

for all probability distributions p_1 and p_2. This can easily be seen from the definition of K or L and the Shannon inequality.

Another important property of the K and L divergences is their *boundedness*, namely,

$$K(p_1,p_2) \le 1 \quad \text{and} \quad L(p_1,p_2) \le 2. \tag{3.13}$$

The second inequality can be easily derived from (3.9) and the fact that the Shannon entropy is nonnegative and the sum of two probability distributions is equal to 2. The bound for $K(p_1, p_2)$ follows directly from its definition (3.1):

$$K(p_1,p_2) = \sum_{x\in X} p_1(x)\log\frac{p_1(x)}{p_1(x)+p_2(x)} + \sum_{x\in X} p_1(x) \le 1.$$

From the equality given in (3.8), we have

$$L(p_1,p_2) = \sum_{x\in X} p_1(x)\log p_1(x) - \sum_{x\in X} p_1(x)\log\frac{p_1(x)+p_2(x)}{2} + \sum_{x\in X} p_2(x)\log p_2(x) - \sum_{x\in X} p_2(x)\log\frac{p_1(x)+p_2(x)}{2}$$

$$= 2H\left(\frac{p_1+p_2}{2}\right) - H(p_1) - H(p_2), \tag{3.14}$$

where H is the Shannon entropy function. Equation (3.14) provides one possible physical interpretation of $L(p_1, p_2)$. This entropic description also leads to a natural generalization of the L divergence.

The K divergence coincides with the f-divergence for $f(x) = x\log(2x/(1+x))$. The f-divergence is a family of measures introduced by Csiszár [7] and its many properties were studied in [29], [30]. Additional properties of the K divergence can thus be derived from the results for the f-divergence.

IV. The Jensen–Shannon Divergence Measure

Let $\pi_1, \pi_2 \ge 0$, $\pi_1 + \pi_2 = 1$, be the *weights* of the two probability distributions p_1 and p_2, respectively. The generalization of the L divergence is defined as

$$JS_\pi(p_1,p_2) = H(\pi_1 p_1 + \pi_2 p_2) - \pi_1 H(p_1) - \pi_2 H(p_2), \tag{4.1}$$

which can be termed the Jensen–Shannon divergence. Since H is a concave function, according to Jensen's inequality, $JS_\pi(p_1, p_2)$ is nonnegative and equal to zero when $p_1 = p_2$. One of the major features of the Jensen–Shannon divergence is that we can assign different weights to the distributions involved according to their importance. This is particularly useful in the study of decision problems. In fact, we will show that Jensen–Shannon divergence provides both the lower and upper bounds to the Bayes probability of error.

Let us consider a classification problem of two classes $C = \{c_1, c_2\}$ with *a priori* probabilities $p(c_1) = \pi_1$, $p(c_2) = \pi_2$ and let the corresponding conditional probability distributions be $p(x|c_1) = p_1(x)$, $p(x|c_2) = p_2(x)$. The Bayes probability of error [11] is given by

$$P_e(p_1,p_2) = \sum_{x\in X}\min(\pi_1 p_1(x), \pi_2 p_2(x)). \tag{4.2}$$

Theorem 4: The following upper bound,

$$P_e(p_1,p_2) \le \frac{1}{2}(H(\pi_1,\pi_2) - JS_\pi(p_1,p_2)), \tag{4.3}$$

holds, where $H(\pi_1, \pi_2) = -\pi_1\log\pi_1 - \pi_2\log\pi_2$.

Proof: It has been shown in [11] that

$$P_e(p_1,p_2) \le \frac{1}{2}H(C|X), \tag{4.4}$$

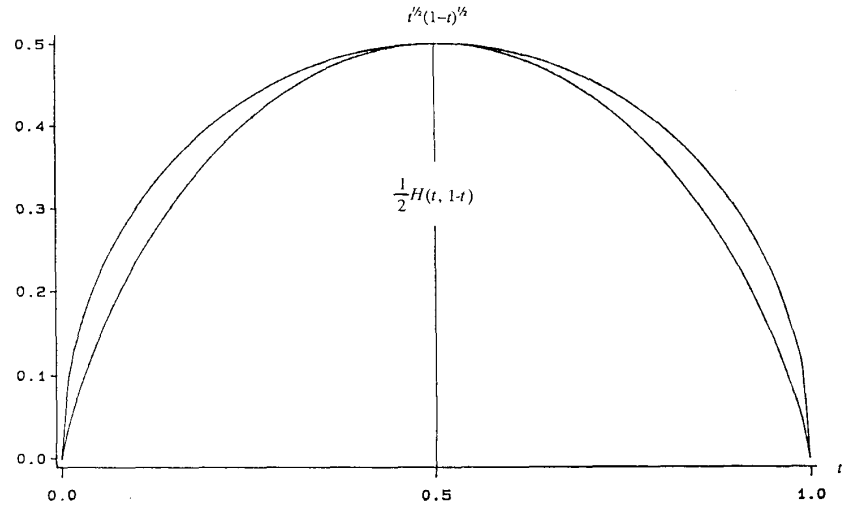

Fig. 2. Shannon entropy and geometric mean.

where

$$H(C|X) = \sum_{x \in X} p(x) H(C|x) = -\sum_{x \in X} p(x) \sum_{c \in C} p(c|x) \log p(c|x), \quad (4.5)$$

which is the equivocation or the conditional entropy [9]. It is also known that

$$H(C|X) = H(C) + H(X|C) - H(X). \quad (4.6)$$

Since there are only two classes involved, we have

$$H(C) = H(p(c_1), p(c_2)) = H(\pi_1, \pi_2), \quad (4.7)$$

and

$$H(X|C) = p(c_1) H(X|c_1) + p(c_2) H(X|c_2) = \pi_1 H(p_1) + \pi_2 H(p_2). \quad (4.8)$$

Also observing that

$$p(x) = \pi_1 p_1(x) + \pi_2 p_2(x),$$

we have

$$H(X) = H(\pi_1 p_1 + \pi_2 p_2). \quad (4.9)$$

Combining (4.7), (4.8), and (4.9) into (4.6), we obtain from inequality (4.4) that

$$P_e(p_1, p_2) \le \frac{1}{2}\left(H(\pi_1, \pi_2) + \pi_1 H(p_1) + \pi_2 H(p_2) - H(\pi_1 p_1 + \pi_2 p_2) \right) = \frac{1}{2}\left(H(\pi_1, \pi_2) - JS_\pi(p_1, p_2) \right). \quad \square$$

The previous inequality is useful because it provides an upper bound for the Bayes probability of error. In contrast, no similar bound exists in terms of either I or J divergence [31] although several lower bounds have been found [14], [26].

Theorem 5: The following lower bound also holds for the Bayes probability of error:

$$P_e(p_1, p_2) \ge \frac{1}{4}\left(H(\pi_1, \pi_2) - JS_\pi(p_1, p_2) \right)^2. \quad (4.10)$$

Proof: By the definition of $H(C|X)$ and the Cauchy inequality, we have

$$H^2(C|X) \le \left(\sum_{x \in X} p(x) \right) \cdot \left(\sum_{x \in X} p(x) \cdot H^2(C|x) \right) = \sum_{x \in X} p(x) \cdot H^2(C|x). \quad (4.11)$$

For any $0 \le t \le 1$, it can be shown that

$$\frac{1}{2} H(t, 1-t) \le \sqrt{t(1-t)}, \quad (4.12)$$

holds as depicted in Fig. 2. A rigorous proof of this inequality is given in the Appendix (Theorem 8).

Therefore, inequality (4.11) can be rewritten as

$$H^2(C|X) \le 4 \cdot \sum_{x \in X} p(x)\left(p(c_1|x) p(c_2|x) \right) \le 4 \cdot \sum_{x \in X} p(x) \min\left(p(c_1|x), p(c_2|x) \right) = 4 \cdot \sum_{x \in X} \min\left(\pi_1 p_1(x), \pi_2 p_2(x) \right) = 4 \cdot P_e(p_1, p_2).$$

From (4.6)–(4.9), inequality (4.10) follows immediately. $\square$

The Jensen–Shannon divergence $J(p_1, p_2)$ was called the increment of the Shannon entropy in [32] and used to measure the distance between random graphs. It was introduced as a criterion for the synthesis of random graphs. In the normalization process, an upper bound had to be used. Based on computer simulation, Wong and You [32] conjectured that the increment of entropy cannot be greater than 1. This conjecture can be easily verified from inequality (4.3). Since the Bayes probability of error is nonnegative, we have from (4.3) that,

$$JS_\pi(p_1, p_2) \le H(\pi_1, \pi_2) - 2P_e(p_1, p_2) \le H(\pi_1, \pi_2) \le 1.$$

This further justifies the use of this measure in [32].

V. The Generalized Jensen–Shannon Divergence Measure

Most measures of difference, including the Jensen–Shannon divergence previously discussed, are designed for two probability distributions. For certain applications such as in the study of

taxonomy in biology and genetics [24], [25], it might be necessary to measure the overall difference of more than two distributions. The Jensen–Shannon divergence can be generalized to provide such a measure for any finite number of distributions. This is useful for the study of decision problems with more than two classes involved.

Let $p_1, p_2, \cdots, p_n$ be n probability distributions with weights $\pi_1, \pi_2, \cdots, \pi_n$, respectively. The generalized Jensen–Shannon divergence can be defined as

$$JS_\pi(p_1, p_2, \cdots, p_n) = H\left(\sum_{i=1}^{n} \pi_i p_i\right) - \sum_{i=1}^{n} \pi_i H(p_i), \quad (5.1)$$

where $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$. Consider a decision problem with n classes $c_1, c_2, \cdots, c_n$ with prior probabilities $\pi_1, \pi_2, \cdots, \pi_n$. The Bayesian error for n classes can be written as

$$P(e) = \sum_{x \in X} p(x)\left(1 - \max\left(p(c_1|x), p(c_2|x), \cdots, p(c_n|x)\right)\right). \quad (5.2)$$

The relationship between the generalized Jensen–Shannon divergence and the previous Bayes probability of error is given by the following theorems.

Theorem 6:

$$P(e) \le \frac{1}{2}\left(H(\pi) - JS(p_1, p_2, \cdots, p_n)\right), \quad (5.3)$$

where

$$H(\pi) = -\sum_{i=1}^{n} \pi_i \log \pi_i \quad \text{and} \quad p_i(x) = p(x|c_i), \quad i = 1,2,\cdots,n.$$

Proof: The proof of this inequality is much the same as that of (4.3). □

Theorem 7:

$$P(e) \ge \frac{1}{4(n-1)}\left(H(\pi) - JS(p_1, p_2, \cdots, p_n)\right)^2. \quad (5.4)$$

Proof: From (4.11) and Theorem 9 in the Appendix, we have

$$H^2(C|X) \le \sum_{x \in X} p(x)\left(2\sum_{i=1}^{n-1} \sqrt{p(c_i|x)\left(1 - p(c_i|x)\right)}\right)^2. \quad (5.5)$$

By the Cauchy inequality, (5.5) becomes

$$H^2(C|X) \le 4 \sum_{x \in X} p(x)\left(\sum_{i=1}^{n-1} p(c_i|x)\right)\left(\sum_{i=1}^{n-1} \left(1 - p(c_i|x)\right)\right). \quad (5.6)$$

Assume, without loss of generality, that the $p(c_i|x)$ have been reordered in such a way that $p(c_n|x)$ is the largest. Then from

$$H^2(C|X) \le 4 \sum_{x \in X} p(x)\left(1 - \max_i \{p(c_i|x)\}\right)(n-1) = 4(n-1)P(e),$$

we immediately obtain the desired result. □

It should be pointed out that the bounds previously presented are in explicit forms and can be computed easily. Implicit lower and upper bounds for the probability of error in terms of the f-divergence can be found in [3]. It should be useful to study the relationship between these bounds but it will not be done in this correspondence.

VI. Conclusion

Based on the Shannon entropy, we were able to give a unified definition and characterization to a class of information-theoretic divergence measures. Some of these measures have appeared earlier in various applications. But their use generally suffered from a lack of theoretical justification. The results presented here not only fill this gap but provide a theoretical foundation for future applications of these measures. Some of the results such as those presented in the Appendix are related to entropy and are useful in their own right.

The unified definition is also important for further study of the measures. We are currently studying further properties of the class. Some of their key applications are also under investigation.

Acknowledgment

The author would like to thank Prof. S. K. M. Wong for his comments and suggestions on an earlier version of this correspondence. The author is also grateful to the referees, especially to the one who pointed out the connection between the divergence measures presented here and the f-divergence.

Appendix

Theorem 8: For any $0 \le x \le 1$,

$$\frac{1}{2}H(x, 1-x) \le \sqrt{x(1-x)}. \quad (A.1)$$

Proof: Consider a continuous function $f(x)$ in the closed interval $[0,1]$:

$$f(x) = 2\sqrt{x(1-x)} + x \log x + (1-x)\log(1-x).$$

$f(x)$ is twice differentiable in the open interval $(0,1)$,

$$f'(x) = \frac{1-2x}{\sqrt{x(1-x)}} + \log\frac{x}{1-x}, \quad (A.2)$$

$$f''(x) = \frac{2\sqrt{x(1-x)} - \ln 2}{2x(1-x)\sqrt{x(1-x)}\ln 2}, \quad (A.3)$$

where ln is the natural logarithm. There are two different real solutions of the equation, $f''(x) = 0$,

$$x_1 = \frac{1 - \sqrt{1 - (\ln 2)^2}}{2} \quad \text{and} \quad x_2 = \frac{1 + \sqrt{1 - (\ln 2)^2}}{2}.$$

It can be easily shown that $0 < x_1 < 1/2 < x_2 < 1$.

From (A.3), it is clear that the function $f''(x)$ is continuous in $(0,1)$ and the denominator of $f''(x)$ is nonnegative in $[0,1]$. Since

$$\lim_{x \to 0^+}\left(2\sqrt{x(1-x)} - \ln 2\right) = -\ln 2,$$

$f''(x_1) = 0$, and there exists no $x \in (0, x_1)$ such that $f''(x) = 0$, by the continuity of $f''(x)$, it follows, $f''(x) < 0$ for $0 < x < x_1$, and thus the function $f(x)$ is concave in $(0, x_1)$.

For $x = 1/2 \in (x_1, x_2)$, we obtain

$$2\sqrt{x(1-x)} - \ln 2 = 1 - \ln 2 > 0,$$

which implies $f''(1/2) > 0$. Since $f''(x_1) = f''(x_2) = 0$ and there exists no $x \in (x_1, x_2)$ such that $f''(x) = 0$, we can conclude that $f''(x) > 0$ for $x_1 < x < x_2$. $f(x)$ is therefore convex in (x_1, x_2). Similarly, from

$$\lim_{x \to 1^-}\left(2\sqrt{x(1-x)} - \ln 2\right) = -\ln 2,$$

it follows $f''(x) < 0$ for $x_2 < x < 1$. This means that the function $f(x)$ is concave in $(x_2, 1)$. In summary, the function $f(x)$ is concave in both open intervals $(0, x_1)$ and $(x_2, 1)$, and convex in (x_1, x_2). $(x_1, f(x_1))$ and $(x_2, f(x_2))$ are the points of inflections for $f(x)$.

Since $f(x)$ is continuous in $[x_1, x_2]$ and convex in (x_1, x_2), it has a unique minimum in $[x_1, x_2]$. The minimum is obtained at $x_{\min} = 1/2$ and $f(x_{\min}) = 0$. Thus, we have, for any $x \in [x_1, x_2]$,

$$f(x) \geq f(x_{\min}) = 0. \tag{A.4}$$

Also, since $f(x)$ is continuous in $[0, x_1]$ and concave in $(0, x_1)$, we have

$$f(x) \geq \min(f(0), f(x_1)) \geq \min\left(f(0), f\left(\frac{1}{2}\right)\right) = 0, \quad \text{for } x \in [0, x_1]. \tag{A.5}$$

Similarly,

$$f(x) \geq \min(f(x_2), f(1)) \geq \min\left(f\left(\frac{1}{2}\right), f(1)\right) = 0, \quad \text{for } x \in [x_2, 1]. \tag{A.6}$$

By combining (A.4), (A.5), and (A.6), we finally obtain

$$f(x) \geq 0, \quad \text{for } 0 \leq x \leq 1, \tag{A.7}$$

from which inequality (A.1) follows immediately. □

Theorem 9: Let $q = (q_1, q_2, \cdots, q_n)$, $0 \leq q_i \leq 1$, $1 \leq i \leq n$, and $\sum_{i=1}^{n} q_i = 1$. Then

$$\frac{1}{2} H(q) \leq \sum_{i=1}^{n-1} \sqrt{q_i(1-q_i)}. \tag{A.8}$$

Proof: By the recursivity of the entropy function [1, p. 30], we have

$$\begin{aligned}
H(q) &= H(q_1, q_2, \cdots, q_{n-1} + q_n) \\
&\quad + (q_{n-1} + q_n) H\left(\frac{q_{n-1}}{q_{n-1} + q_n}, \frac{q_n}{q_{n-1} + q_n}\right) \\
&= H(q_1, q_2, \cdots, q_{n-2} + q_{n-1} + q_n) \\
&\quad + (q_{n-1} + q_n) H\left(\frac{q_{n-1}}{q_{n-1} + q_n}, \frac{q_n}{q_{n-1} + q_n}\right) \\
&\quad + (q_{n-2} + q_{n-1} + q_n) H\left(\frac{q_{n-2}}{q_{n-2} + q_{n-1} + q_n}, \frac{q_{n-1} + q_n}{q_{n-2} + q_{n-1} + q_n}\right) \\
&= H(q_1, q_2 + q_3 + \cdots + q_n) \\
&\quad + (q_{n-1} + q_n) H\left(\frac{q_{n-1}}{q_{n-1} + q_n}, \frac{q_n}{q_{n-1} + q_n}\right) \\
&\quad + (q_{n-2} + q_{n-1} + q_n) H\left(\frac{q_{n-2}}{q_{n-2} + q_{n-1} + q_n}, \frac{q_{n-1} + q_n}{q_{n-2} + q_{n-1} + q_n}\right) + \cdots \\
&\quad + (q_2 + q_3 + \cdots + q_n) H\left(\frac{q_2}{q_2 + q_3 + \cdots + q_n}, \frac{q_3 + \cdots + q_n}{q_2 + q_3 + \cdots + q_n}\right).
\end{aligned}$$

By Theorem 8, we obtain

$$\begin{aligned}
\frac{1}{2} H(q) &\leq \sqrt{q_1(1-q_1)} + \sqrt{q_{n-1} q_n} + \sqrt{q_{n-2}(q_{n-1} + q_n)} + \cdots \\
&\quad + \sqrt{q_2(q_3 + q_4 + \cdots + q_n)} \\
&\leq \sqrt{q_1(1-q_1)} + \sqrt{q_2(1-q_2)} + \cdots + \sqrt{q_{n-2}(1-q_{n-2})} \\
&\quad + \sqrt{q_{n-1}(1-q_{n-1})}. \\
&= \sum_{i=1}^{n-1} \sqrt{q_i(1-q_i)}.
\end{aligned}$$

□

References

[1] J. Aczel and Z. Daroczy, *On Measures of Information and Their Characterizations*. New York: Academic, 1975.

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

[3] M. Ben Bassat, "f-entropies, probability of error, and feature selection," *Inform. Contr.*, vol. 39, pp. 227–242, 1978.

[4] C. H. Chen, *Statistical Pattern Recognition*. Rochelle Park, NJ: Hayden Book Co., 1973, Ch. 4.

[5] ____, "On information and distance measures, error bounds, and feature selection," *Inform. Sci.*, vol. 10, pp. 159–173, 1976.

[6] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 3, pp. 462–467, May 1968.

[7] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[8] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[9] S. Guiasu, *Information Theory with Applications*. New York: McGraw-Hill, 1977.

[10] D. V. Gokhale and S. Kullback, *Information in Contingency Tables*. New York: Marcel Dekker, 1978.

[11] M. E. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inform. Theory*, vol. IT-16, no. 4, pp. 368–372, July 1970.

[12] R. W. Johnson, "Axiomatic characterization of the directed divergences and their linear combinations," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 6, pp. 709–716, Nov. 1979.

[13] T. T. Kadota and L. A. Shepp, On the best finite set of linear observables for discriminating two gaussians signals," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 278–284, Apr. 1967.

[14] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions Commun. Technol.*, vol. COM-15, no. 1, pp. 52–60, Feb. 1967.

[15] J. N. Kapur, "A comparative assessment of various measures of directed divergence," *Advances Manag. Stud.*, vol. 3, no. 1, pp. 1–16, Jan. 1984.

[16] D. Kazakos and T. Cotsidas, "A decision theory approach to the approximation of discrete probability densities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, vol. 1, pp. 61–67, Jan. 1980.

[17] S. Kullback, *Information Theory and Statistics*. New York: Dover Publications, 1968.

[18] ____, "A lower bound for discrimination information in terms of variation," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 126–127, Jan. 1967.

[19] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[20] U. Kumar, V. Kumar, and J. N. Kapur, "Some normalized measures of directed divergence," *Int. J. Gen. Syst.*, vol. 13, pp. 5–16, 1986.

[21] J. Lin and S. K. M. Wong, "Approximation of discrete probability distributions based on a new divergence measure," *Congressus Numerantium*, vol. 61, pp. 75–80, 1988.

[22] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. Lon.*, Ser. A, vol. 186, 1946, pp. 453–461.

[23] C. R. Rao and T. K. Nayak, "Cross entropy, dissimilarity measures, and characterizations of quadratic entropy," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 5, pp. 589–593, Sept. 1985.

[24] C. R. Rao, "Diversity and dissimilarity coefficients: A unified approach," *Theoretical Population Biol.*, vol. 21, pp. 24–43, 1982.
[25] ____, "Diversity: Its measurement, decomposition, apportionment and analysis," *Sankhya: Indian J. Statist.*, Ser. A, vol. 44, pt. 1, pp. 1–22, Feb. 1982.
[26] G. T. Toussaint, "On some measures of information and their application to pattern recognition," in *Proc. Conf. Measures of Information and Their Applications*, Indian Inst. Technol., Bombay, Aug. 1974.
[27] ____, "Sharper lower bounds for discrimination information in terms of variation," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 1, pp. 99–100, Jan. 1975.
[28] I. Vajda, "Note on discrimination information and variation," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 771–773, Nov. 1970.
[29] ____, "On the f-divergence and singularity of probability measures," *Periodica Mathem. Hungarica*, vol. 2, pp. 223–234, 1972.
[30] ____, *Theory of Statistical Inference and Information*. Dordrecht-Boston: Kluwer, 1989.
[31] J. W. Van Ness, "Dimensionality and the classification performance with independent coordinates," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-7, pp. 560–564, July 1977.
[32] A. K. C. Wong and M. You, "Entropy and distance of random graphs with application to structural pattern recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, no. 5, pp. 599–609, Sept. 1985.

On the Error Probability of Signals in Additive White Gaussian Noise

Brian Hughes, *Member, IEEE*

***Abstract*—A new upper bound is presented to the probability of error in detecting one of M equally probable signals in additive white Gaussian noise. This bound is easy to calculate, can be applied to any signal set, and is always better than the union and minimum distance bounds.**

***Index Terms*—Error probability, Gaussian noise, signal detection, performance bounds.**

I. Introduction

Consider the classical problem of detecting one of M signals in additive white Gaussian noise: An integer message m, equally likely to be any element of $\{0, \cdots, M-1\}$, is transmitted over an N-dimensional vector channel by sending a corresponding vector $\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{M-1} \in \boldsymbol{R}^N$. When $m = i$ is transmitted, the received vector is

$$\boldsymbol{y} = \boldsymbol{x}_i + \boldsymbol{n}, \tag{1.1}$$

where the components of $\boldsymbol{n}$ are independent, identically distributed $N(0, \sigma^2)$ random variables. An estimate of m is made at the receiver using the maximum likelihood rule. From elementary communication theory, the problem of detecting one of M equally likely waveforms in additive white Gaussian noise with two-sided power spectral density σ^2 can be expressed in this form.

Manuscript received November 20, 1989; revised August 6, 1990. This work was supported by the National Science Foundation under Grant No. NCR-8804257, and by the US Army Research Office under Grant No. DAAL03-89-K-0130. This work was presented in part at the *Twenty-seventh Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 27–29, 1989.

The author is with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218.

IEEE Log Number 9040136.

Let $P_{e,0}$ be the error probability of the maximum likelihood detector given that $m = 0$ is sent. There are several simple, generally applicable, upper bounds for this error probability. Two of the most commonly used are the *union bound*

$$P_{e,0} \le \sum_{i=1}^{M-1} Q(d_i/2\sigma), \tag{1.2}$$

where

$$Q(x) \equiv \int_x^\infty (2\pi)^{-1/2} \exp(-t^2/2)\, dt \quad \text{and} \quad d_i \equiv |\boldsymbol{x}_i - \boldsymbol{x}_0|;$$

and the *minimum distance bound*

$$P_{e,0} \le \Pr\{|\boldsymbol{n}| \ge d_{\min}/2\} = \bar{\Gamma}(N/2, d_{\min}^2/8\sigma^2), \tag{1.3}$$

where $d_{\min} = \min_{i \ne 0} d_i$, and

$$\bar{\Gamma}(a, x) \equiv \frac{1}{\Gamma(a)} \int_x^\infty t^{a-1} e^{-t}\, dt \tag{1.4}$$

is a (normalized) *incomplete gamma function*. Unfortunately, (1.2) is somewhat loose for small values of $d_{\min}/2\sigma$ or large M, and (1.3) is loose for all but the smallest values of N.

In this note, we present a new upper bound to the probability of error that is also straightforward to calculate and generally applicable and improves upon the bounds in (1.2) and (1.3). In Section II, we show

$$P_{e,0} \le \sum_{i=1}^{M-1} A_N(d_i/\alpha_o, d_i/2\sigma), \tag{1.5}$$

where α_o is the unique solution of

$$\sum_{i=1}^{M-1} A_N(d_i/\alpha_o, 0) = 1, \tag{1.6}$$

and where[1]

$$A_N(y, x) \equiv \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \cdot \bar{\gamma}\left((N-1)/2, \left(y^{-2}-1\right)^+ z^2/2\right) dz, \tag{1.7}$$

$$\bar{\gamma}(a, x) \equiv 1 - \bar{\Gamma}(a, x). \tag{1.8}$$

Moreover, for virtually all signal sets of practical interest, we show that this bound reduces to

$$P_{e,0} \le N_{\min} A_N(\beta_o, d_{\min}/2\sigma), \tag{1.9}$$

where $d_{\min} = \min_{i \ne 0} d_i$, $N_{\min}$ is the number of vectors for which $d_i = d_{\min}$, and where β_o satisfies

$$N_{\min} A_N(\beta_o, 0) = 1. \tag{1.10}$$

In Section III, we present series expansions for $A_N(y, x)$ and $A_N(y, 0)$ that allow rapid calculation of (1.5) and (1.9). Finally, the bound is calculated for two examples in Section IV.

II. Derivation of the Bound

For the channel (1.1) the maximum likelihood detector is the minimum Euclidean distance rule:

$$\hat{m}(\boldsymbol{y}) \equiv \arg \min_{0 \le i \le M-1} |\boldsymbol{x}_i - \boldsymbol{y}|, \tag{2.1}$$

with ties resolved arbitrarily. For convenience, we resolve ties in favor of the *larger* index, and define *decision regions* $D_i = \{\boldsymbol{y} | \hat{m}(\boldsymbol{y}) = i\}$, $0 \le i \le M-1$.

[1] For any real number x, $x^+ \equiv \max\{0, x\}$.

0018-9448/91/0100-0151$01.00 © 1991 IEEE