LOWER AND UPPER BOUNDS FOR MISCLASSIFICATION PROBABILITY BASED ON RENYI'S INFORMATION

Deniz Erdogmus, Jose C. Principe

Computational NeuroEngineering Laboratory, NEB 454, University of Florida, Gainesville, FL 32611

[deniz,principe]@cnel.ufl.edu

ABSTRACT

Fano's inequality has proven to be one important result in Shannon's information theory having found applications in innumerous proofs of convergence. It also provides us with a lower bound on the symbol error probability in a communication channel, in terms of Shannon's definitions of entropy and mutual information. This result is also significant in that it suggests insights on how the classification performance is influenced by the amount of information transferred through the classifier. We have previously extended Fano's lower bound on the probability of error to a family of lower and upper bounds based on Renyi's definitions of entropy and mutual information. These new bounds however, despite their theoretical appeal, were practically incomputable. In this paper, we present some modifications to these bounds that will allow us to utilize them in practical situations. The significance of these new bounds is threefold: Illustrating a theoretical use of Renyi's definition of information, extending Fano's result to include an upper bound for probability of classification error, and providing insights on how the information transfer through a classifier affects its performance. The performance of the modified bounds is investigated in various numerical examples, including applications to digital communication channels that are designed to point out the major conclusions.

Keywords: Classifier performance, Error bounds, Renyi's entropy, Fano's inequality

Symbols

M, W : Discrete random variables with probability mass functions $\{p(m_k)\}_{k=1}^{N_c}$ and $\{p(w_k)\}_{k=1}^{N_c}$ (representing actual classes and decision classes)

$p(w_j, m_k)$: Joint probability mass function of W and M

$p(w_j \mid m_k)$: Conditional probability mass function of W given M (entries of the confusion matrix)

E : Discrete random variable with Bernoulli distribution $\{p_e, 1\text{-}p_e\}$ for values e and c (representing erroneous and correct classification)

log : Throughout this paper, this function represents the base-2 logarithm

1. Introduction

Fano's bound is a well-known inequality in the information theory literature. It is essential to prove the converse to Shannon's second theorem [1]. Applied to a classifier, by providing a lower bound for classification error probability, it is useful in terms of giving an indication of attainable performance. In addition, it provides some insights as to how the process of information transfer progresses in this setting, linking classification performance with information theory. In fact, this is one of the outstanding advantages of information theory, the abstract level of investigation and analysis. Linsker's infomax principle progresses along similar lines. As a principle for self-organization, Infomax states that an optimal system must transfer as much information as possible from its input to its output, i.e. maximize the mutual information between its input and output [2]. Fano's bound entails similar conclusions about the structure of optimal classifiers; these must maximize the mutual information between actual and decision classes to minimize probability of error [3].

The question of determining optimal features has been one of the major focal points in pattern recognition research, and information theory has played a central role in this quest [4,5]. It has been established that information is not preserved in subspace projections, yet maximization of information across the mapping is essential in this process [6]. Fisher and Torkkola have recently utilized this approach to train neural networks directly from samples for optimal feature extraction using a nonparametric estimator based on Renyi's entropy [3,7]. In all of these, Fano's bound appears as the central-piece because it relates classification error to conditional entropy. Although Fano's lower bound for the probability of error in classification is a valuable indicator of attainable performance, the goal in statistical pattern recognition and machine learning is to minimize the probability of error [8], or possibly an upper bound for the error probability as in structural risk minimization [9]. Therefore, a family of lower and upper bounds would encompass the advantages of both; identify the limitations and indicate the possible generalization performance simultaneously.

Fano's inequality is derived utilizing Shannon's entropy definition, as it was the only one available to him [10]. Motivated by Shannon's brilliant work [11], researchers concentrated their efforts on information theory. Renyi was able to also formulate the theory of information starting from four basic postulates [12]. His definitions of information theoretic quantities like entropy and mutual information encompassed Shannon's definitions as special cases. Inspired by Fano's bound, many researchers have also proposed modifications, generalizations, or alternative information theoretic inequalities, mainly with applications to communication theory [13,14,15,16,17]. The recent work of Feder and Merhav is especially important as it provides a lower and an upper bound for the minimal probability of error in estimating the value of a discrete random variable [15]. Their bounds demonstrate the association between the probability of value-prediction error and Shannon's entropy, and Renyi's entropy of order infinity as well. Han and Verdu's generalization to Fano's bound, again using Renyi's entropy of order infinity is theoretically appealing and also useful in proving a generalized source-channel separation theorem [16]. Yet, the bounds presented in these works do not explicitly consider the classification process, thus do not make use of the confusion matrix of the classifier under consideration. Nevertheless, motivated by these works that extend on classical results utilizing Renyi's alternative definition of information, we have developed a family of lower and upper bounds, using Renyi's definitions of information theoretic quantities. For this, the free parameter in Renyi's definitions was exploited along with Jensen's inequality for convex and concave functions.

The organization of this paper is as follows: We will first, briefly review Shannon's and Renyi's definitions of information theoretic quantities like entropy and mutual information. Then we will present Fano's bound and the generalized family of lower and upper bounds obtained using Renyi's definitions. Next, we will introduce some modifications to these new bounds to eliminate unwanted terms, followed by a discussion of some properties of the bounds and numerical case studies each designed to illustrate a different aspect of the proposed bounds. In the conclusions, we summarize the main results and important lessons learned.

2. Shannon's and Renyi's Definitions of Entropy and Mutual Information

In the development of the aforementioned bounds, we will utilize several information theoretic quantities as defined by Shannon and Renyi. These are the joint entropy, (average) conditional entropy, and (average) mutual information. We use the random variable *M* to denote the actual class (input space) and *W* to denote the decided class (output sample) when applying these arguments to classifiers with a known confusion matrix and priors. The random variable *E*, which takes the values *e* or *c*, is used to denote the events of wrong and correct classification with probabilities $\{p_e, 1\text{-p}_e\}$

Shannon's Definitions: For a discrete random variable *M*, whose probability mass function (pmf) is $\{p(m_k)\}_{k=1}^{N_c}$, Shannon's entropy is given by [11]

$$H_S(M) = -\sum_{k=1}^{N_c} p(m_k) \log p(m_k) \tag{1}$$

Based on this definition, the joint entropy, mutual information, and conditional entropy are defined as

$$\begin{aligned} H_S(M,W) &= -\sum_{k=1}^{N_c}\sum_{j=1}^{N_c} p(m_k, w_j) \log p(m_k, w_j) \\ I_S(M,W) &= \sum_{k=1}^{N_c}\sum_{j=1}^{N_c} p(m_k, w_j) \log \frac{p(m_k, w_j)}{p(m_k)\, p(w_j)} \\ H_S(M \mid W) &= \sum_{j=1}^{N_c} H_S(M \mid w_j) p(w_j) \end{aligned} \tag{2}$$

where

$$H_S(M \mid w_j) = -\sum_{k=1}^{N_c} p(m_k \mid w_j) \log p(m_k \mid w_j) \tag{3}$$

and $p(m_k, w_j)$ and $p(m_k|w_j)$ are respectively the joint probability mass function and the conditional probability mass function of *M* given *W*, respectively. Shannon's mutual information is equal to the Kullback-Leibler divergence [18] between the joint distribution and the product of m1arginal distributions and it satisfies the following property [10].

$$I_S(M,W) = H_S(W) - H_S(W \mid M) \tag{4}$$

Renyi's Definitions: Renyi's entropy for *M* is given by [12]

$$H_a(M) = \frac{1}{1-a} \log \sum_{k=1}^{N_c} p^a(m_k) \tag{5}$$

where a is a real positive constant different from 1. The (average) mutual information and (average) conditional entropy are consequently

$$\begin{aligned} H_a(M,W) &= \frac{1}{1-a} \log \sum_{k=1}^{N_c} \sum_{j=1}^{N_c} p^a(m_k, w_j) \\ I_a(M,W) &= \frac{1}{a-1} \log \sum_{k=1}^{N_c} \sum_{j=1}^{N_c} \frac{p^a(m_k, w_j)}{p^{a-1}(m_k) p^{a-1}(w_j)} \\ H_a(W \mid M) &= \sum_{k=1}^{N_c} p(m_k) H_a(W \mid m_k) \end{aligned} \tag{6}$$

where

$$H_a(W \mid m_k) = \frac{1}{1-a} \log \sum_{j=1}^{N_c} p^a(w_j \mid m_k) \tag{7}$$

The free parameter a of Renyi's definitions will be helpful in the following sections, when we apply Jensen's inequality to obtain the lower and upper bounds for the probability of error. In order to perceive the effect of a on the value of entropy, consider the following fact: Renyi's entropy is a monotonically decreasing function of a whose values range from $\log N_c$ to $-\log(\max_k p(m_k))$ as it is varied from zero to infinity. Although Renyi's definitions have a discontinuity at a =1, using L'Hopital's rule, it is easily seen that the limit of Renyi's entropy as a goes to one is Shannon's entropy. In fact this statement is true for all Renyi's definitions.

3. Fano's Bound on Probability of Error

Fano's inequality determines a lower bound for the probability of classification error in terms of the information transferred through the classifier. More specifically, consider a classifier for which the actual classes, denoted by *M*, have prior probabilities $\{p(m_k)\}_{k=1}^{N_c}$ and the decided classes, denoted by *W*, have the conditional probabilities $p(w_j \mid m_k)$. Fano's bound for the probability of classification error, in accordance with the definitions of the previous section, is then given by [10]

$$p_e \geq \frac{H_S(W\,|\,M) - h_S(p_e)}{\log(N_c - 1)} \tag{8}$$

where the special notation $h_S(p_e)$ = $-p_e\log p_e - (1-p_e)\log(1-p_e)$ is used for the binary Shannon's entropy. Notice that this original bound, as it appears in Fano's derivation has the probability, has the probability of error appearing on both sides of the inequality. Also the denominator prevents the application of this bound to two-class situations. To account for these problems, the binary entropy of p_e is replaced by its maximum possible value, log2=1, and the denominator is replaced with the larger $\log N_c$. In addition, the conditional entropy is replaced by the sum of a marginal entropy and a mutual information term in accordance with (4). After all these modifications, the commonly presented version of Fano's bound in the literature is [3]

$$p_e \geq \frac{H_S(W) - I_S(M;W) - 1}{\log N_c} \tag{9}$$

4. Bounds Using Renyi's Entropy and Mutual Information

We have recently applied Jensen's inequality on Renyi's definition of conditional entropy, joint entropy and mutual information to obtain the following lower and upper bounds for the probability of error [19]. Since Renyi's mutual information and conditional entropy do not share the identity in (4), these bounds had to be separately derived, starting from their corresponding basic definitions. The existence of entropy order as a free parameter in Renyi's definitions allows us to utilize two different forms of Jensen's inequality: For the two regions a < 1 and a > 1, inequalities for convex and concave functions are used separately, resulting in two different directions of the inequality and yielding an upper bound and a lower bound. For convenience, we provide the derivation for the bound that uses the conditional entropy in the appendix.

$$L = \frac{H_a(W\,|\,M) - h_S(p_e)}{\log(N_c - 1)} \leq p_e \leq \frac{H_b(W\,|\,M) - h_S(p_e)}{\min\limits_k H_b(W\,|\,e, m_k)} = U, \quad \begin{matrix} a \geq 1 \\ b < 1 \end{matrix} \tag{10}$$

$$\frac{H_a(W,M) - H_S(M) - h_S(p_e)}{\log(N_c - 1)} \leq p_e \leq \frac{H_b(W,M) - H_S(M) - h_S(p_e)}{\min\limits_k H_b(W\,|\,e, m_k)}, \quad \begin{matrix} a \geq 1 \\ b < 1 \end{matrix} \tag{11}$$

$$\frac{H_S(W) - I_a(W;M) - h_S(p_e)}{\log(N_c - 1)} \leq p_e \leq \frac{H_S(W) - I_b(W;M) - h_S(p_e)}{\min\limits_k H_S(W\,|\,e, m_k)}, \quad \begin{matrix} a \geq 1 \\ b < 1 \end{matrix} \tag{12}$$

Notice that in all three cases, the lower bounds for $\alpha = 1$ corresponds to Fano's bound through equality (4). The term in the denominator of the upper bound is the entropy of the conditional distribution given the actual class and that the classifier makes an error.

From a theoretical point of view, these bounds are interesting as they indicate how the information transfer through the classifier relates to its performance. Since the family parameter of Renyi's definition does not affect the location of minimum and maximum points of the entropy and mutual information, it is safely concluded, for example, from (12) that, as the mutual information between the input and the output of the classifier is increased its probability of error decreases. Consequently, this result also provides a theoretical basis for utilizing mutual information for feature extraction.

The denominators of the upper bounds also offer an interesting insight about the success of the classification process. As these entropy terms are maximized, the upper bounds become tighter. This happens when the corresponding distribution is uniform; that is when the distribution of probabilities over the erroneous classes is uniform. This conclusion conforms to the observations of Feder and Merhav [15]. They have also noted that in a prediction process, their upper bound is tightest when the probabilities are distributed uniformly over the *wrong* values.

It was previously demonstrated that the tightest lower and upper bounds are obtained when α and β both approached to 1 from above and below, respectively. Furthermore, it was shown that all three bounds were almost equally tight [19]. A problem in these bounds, however, as in Fano's original bound, is the fact that the probability of error appears on both sides of the inequalities. In the next sections, we will propose some modifications to eliminate this inconvenience.

5. Obtaining the Modified Bounds

As in Fano's bound, we wish to get rid of the binary entropy term in the numerator of the bounds and also we would like to modify the bounds to account for two-class situations. We will only consider the bounds using the conditional entropy in (10), and since we already know that $\alpha = \beta = 1$ provides the tightest bounds in both cases [19], we will focus on these values of the family parameters. Nevertheless, even if these parameters are not fixed to the value above, the obtained modified bounds would still be valid, but looser.

Lower Bound: In the numerator, we would like to substitute a tight approximation for the binary entropy whenever possible. Consider the following argument. Suppose in a classification problem, the priors of the classes are given by $\{p(m_k)\}_{k=1}^{N_c}$. Even without designing any classifier, by simply always picking the class with the largest prior, it is possible to obtain an average error probability of $p_e^* = 1 - \max_k p(m_k)$. We call this value of the probability of error as the (trivial) upper bound for a non-trivial classifier. This means, any classifier performing worse than this may simply be thrown away and any classifier design that utilizes a posteriori information about the problem will perform better than this value. Assuming that the classifier under consideration is a non-trivial classifier, its error probability will be smaller than this value, therefore the binary entropy for its probability of error will be smaller than $h_S(\min[1/2, p_e^*])$, which is either equal to log2=1 when $p_e^* \geq 1/2$ or equal to $h_S(p_e^*)$ when $p_e^* < 1/2$. With this substitution, the new lower bound becomes

$$L \geq L^* = \frac{H_S(W \mid M) - h_S(\min[1/2, p_e^*])}{\log(N_c - 1)} \tag{13}$$

It is possible to calculate the worst-case difference between the actual bound and the modified bound. The maximum difference between L and L^* is given by (14). The final expression is obtained by assuming zero probability of error, which results in zero binary entropy for p_e and zero conditional entropy.

$$\begin{aligned} \max_{p_e} L - L^* &= \max_{p_e} \frac{H_S(W \mid M) - h_S(p_e)}{\log(N_c - 1)} - \frac{H_S(W \mid M) - h_S(\min[1/2, p_e^*])}{\log(N_c - 1)} \\ &= \max_{p_e} \frac{h_S(\min[1/2, p_e^*])}{\log(N_c - 1)} - \frac{h_S(p_e)}{\log(N_c - 1)} \\ &= \frac{h_S(\min[1/2, p_e^*])}{\log(N_c - 1)} \end{aligned} \tag{14}$$

We could further modify L^* to account for two-class situations by replacing the numerator with $\log N_c$ as is commonly done when utilizing Fano's bound, in the literature.

Upper Bound: In modifying the upper bound we will again make use of p_e^*. In order to get rid of the binary entropy of the probability of error in the upper bound, we need to replace it with something smaller. Define $\overline{L} = \max[0, L^*]$. We know that the probability of error will satisfy $\overline{L} \leq p_e \leq p_e^*$. Since $h_S(p_e)$ is a concave function, it is greater than or equal to a line passing through the points ($\bar{L}$,h_S($\bar{L}$)) and (p_e^* ,h_S(p_e^*)). This inequality is given by

$$h_S(p_e) \geq \left[\frac{h_S(p_e^*) - h_S(\bar{L})}{p_e^* - \bar{L}}\right] \cdot p_e + \left[\frac{h_S(\bar{L}) \cdot p_e^* - h_S(p_e^*) \cdot \bar{L}}{p_e^* - \bar{L}}\right] \tag{15}$$

Substituting this in the upper bound and rearranging terms to collect all p_e's on one side, we obtain the following modified upper bound expression for the probability of error.

$$U \leq U^* = \frac{H_S(W|M) - \left[\dfrac{h_S(\bar{L}) \cdot p_e^* - h_S(p_e^*) \cdot \bar{L}}{p_e^* - \bar{L}}\right]}{\min\limits_k H_S(W|e,m_k) + \left[\dfrac{h_S(p_e^*) - h_S(\bar{L})}{p_e^* - \bar{L}}\right]} \tag{16}$$

The maximum deviation of U^* from U occurs at the value of p_e where its binary entropy is the furthest away from the line in (15). This value of p_e is easily computed by taking the derivative of the difference and equating it to zero. Denoting the slope of the line in (15) with a, the value of p_e where maximum deviation occurs is found to be $\bar{p}_e = (2^a + 1)^{-1}$. Assuming $\bar{L} = 0$, at this value of error probability, the deviation between the two bounds can be calculated to be (this is an upper bound to the case where $\bar{L} \geq 0$).

$$\begin{aligned} U^* - U &= \frac{H_S(W|M)}{\min\limits_k H_S(W|e,m_k) + h_S(p_e^*)/p_e^*} - \frac{H_S(W|M) - h_S(\bar{p}_e)}{\min\limits_k H_S(W|e,m_k)} \\ &= \frac{h_S(\bar{p}_e) \cdot \min\limits_k H_S(W|e,m_k) + h_S(\bar{p}_e) \cdot h_S(p_e^*)/p_e^* - H_S(W|M) \cdot h_S(p_e^*)/p_e^*}{\left(\min\limits_k H_S(W|e,m_k) + h_S(p_e^*)/p_e^*\right) \cdot \min\limits_k H_S(W|e,m_k)} \end{aligned} \tag{17}$$

This expression is very difficult to simplify further, however, a looser but simpler upper bound could be obtained using the fact that the minimum value of the term $\min\limits_k H_S(W|e,m_k)$ is zero, but this bound will most likely be extremely loose.

In order to get an idea of how much accuracy is lost when performing these approximations, consider the following situations. For the lower bound, we vary p_e^* in the interval [0,1] and evaluate the worst-case deviation of the lower bound given in (14) as a function of it. For the upper bound, while varying p_e^* in the same interval, we set the confusion matrix of a hypothetical classifier such that its diagonal entries are equal to $1 - \bar{p}_e$ and all off-diagonal entries are equal to each other (note that in a confusion matrix each column, corresponding to the conditional distribution given a specific class, must add to one). Under these assumptions we can evaluate $\min_k H_S(W \mid e, m_k)$ as log(N_c-1) and the conditional entropy can also be evaluated from the confusion matrix depending on the value of p_e^*. Fig. 1 depicts the results of these two case studies. The upper panel shows the maximum deviation in the lower bound for three different values of N_c and the lower panel shows the same quantity for the upper bound as a function of p_e^*. From the worst-case results presented here, we observe that the modified bounds perform better as the number of classes increase. In addition, for small values of the probability of error, the loss in the bound due to the introduced modifications increases approximately linearly with the probability of error. Note that these worst-case values are obtained using the confusion matrix and prior probability configurations that maximize the loss by the modifications. Therefore, in general, the modified bounds will be more accurate than as shown in Fig. 1.

The substitution we have used in the above upper bound for the binary entropy of the probability of error is, in fact, at least as tight as the bound provided in Feder and Merhav for general discrete random variables [15]. This general inequality can be summarized as follows.

Fact: Given a random variable X with pmf $\{p(m_k)\}$, k=1,…,N_c. Define p =1-max$_k p(m_k)$. Then the Shannon's entropy of X is bounded by $h_S(p) \geq H_S(X) \geq 2p$.

We are specifically interested for the case where X is a Bernoulli distributed variable with $\{p_e, 1\text{-}p_e\}$ and p=p_e<1/2. In this case, $h_S(p) = H_S(X) \geq 2p$. Noticing that when we assume $p_e^* = 1/2$, we have $h_S(p_e^*)/p_e^* = 2$, this inequality reads $h_S(p_e) \geq p_e \cdot h_S(p_e^*)/p_e^*$, which is, in general, looser than the lower bound line we have utilized in (15).

6. Discussions

In the preceding sections, we have considered the extensions to Fano's bound for the classification probability of error derived using Renyi's definition of entropy and mutual information. The free family parameter of Renyi's definitions were useful in generating these lower and upper bounds by allowing the application of both concave and convex versions of Jensen's inequality. These original inequalities in (10)-(12) are theoretically intriguing as they elucidate the relationship between the classifier performance and its information preservation properties. Investigating these bounds, we can argue that, in order to improve the performance of a classifier one needs to maximize the mutual information between actual and decision classes generated by the classifier (according to (12)), while minimizing the average output uncertainty, i.e. $H(W)$. If we were to examine the inequalities in (10), then we conclude that, in order to increase performance, one needs to minimize the average output uncertainty given inputs from a specific class (averaged over all input classes in accordance with their prior probabilities).

Observing that the probability of error appears on both sides of the inequalities, we then introduce further approximations into the lower and upper bounds. Investigation of the modified bounds and comparisons of their values with the original bounds have revealed that whenever the classifier under consideration has a small probability of error the loss resulting from these modifications is also small.

It is also imperative at this point to discuss the practical evaluation of these bounds in realistic situations. Generally, analytical expressions for probability distributions will not be available to the designer and the quantities in the bounds need to be estimated from the samples. In order to estimate the bounds, the following probabilities need to be estimated: $\{p(w_j \mid m_k)\}_{j,k=1}^{N_c}$ and $\{p(m_k)\}_{k=1}^{N_c}$.

One simple way to estimate these probabilities is to count an divide the appropriate samples in the training set with the appropriate numbers of samples. For this approach to work, however, there must be a sufficiently large number of samples to cover for all possible combinations of actual input classes and decisions. If the classifier is a multiplayer perceptron with as many outputs as the number of classes that is trained to generate a large value only at the output corresponding to the class of the presented input, the output values may be used as estimations of the conditional probabilities [20]. These estimates may be averaged over the training samples of the same class to yield more accurate values. The advantage of this method is that it may provide accurate estimates of the conditional probabilities with relatively fewer training samples.

It must be noted, however, that the common drawback in information theoretic bounds is that the required information to evaluate the bounds readily suffices to evaluate the probability of error itself. On the other hand, the evaluated bound can still be utilized as a validation value and can provide a level of confidence about the estimated probability of error. In order to see this, consider the case where both an upper bound and the probability of error is evaluated using the conditional and prior probabilities necessarily all estimated by counting appropriately a number of samples. Since the counting method is the maximum likelihood method to predict probability masses and since an asymptotically maximum likelihood estimation tends to have a Gaussian distribution, we can assume that the estimated p_e and U values will both be normally distributed. As we will demonstrate in the numerical case studies, the variances of the error and bound estimation using this procedure is approximately equal. Let P denote the probability that the estimated probability of error is greater than the estimated bound; then $P{=}Prob[(U{-}p_e){+}n_1{-}n_2 < 0]$, where n_1 and n_2 are the (assumed to be) *zero-mean* normal distributed estimation errors with equal variance s^2. Letting $\Delta{=}U{-}p_e$, we find $P = Q(\Delta/(s\sqrt{2}))$. Finally, we can say that the confidence in the evaluated probability of error being smaller than the estimated upper bound is $1{-}P$. Since we have a parametric family of upper bounds in (10)-(12), we can evaluate the bound using a couple of different entropy order values and obtain the confidence levels for a number of upper bound values. The specific details of this idea, however, is not fully developed yet. In order to proceed with such a formulization, it is necessary to utilize well-known results from statistics that relate the actual value of a probability mass function to its approximation obtained from frequency counts. Then, it is possible todetermine the probability density of the value of the bound given its value estimated from frequency counts (over the training samples of the classifier for example). These possibilities will be pursued as future work by the authors.

7. Numerical Case Studies

First, consider a simple example to demonstrate the performance of the bounds in a toy problem. This example has three classes and a hypothetical classifier with the following confusion matrix. For this specific confusion matrix, the priors have no effect on the bounds, due to the structural symmetry.

$$P_{W|M} = \begin{bmatrix} 1-p_e & p_e-e & e \\ e & 1-p_e & p_e-e \\ p_e-e & e & 1-p_e \end{bmatrix} \tag{18}$$

The average probability of error is determined by p_e, and e controls the distribution of the probability among the erroneous classes. By changing e in the interval $[0, p_e/2]$, we can study the performance of the bounds in terms of tightness. Fig. 2 shows the lower and upper bounds of (19) for this example as a function of e. The original bounds use the expressions in (10) with $a = 1$ and $b = 0.995$. For comparison, we also include Feder and Merhav's lower and upper bounds for the same situations. For detailed explanation of how these bounds are evaluated please refer to the appendix. Investigating these results, we observe that the original upper bounds are tighter than the modified bounds, as expected. Nonetheless, the modified upper bounds compared equally to the Feder & Merhav bound.

In the above examples, the modified lower bounds are not depicted because they turn out to be negative. Notice that Feder and Merhav's lower bound is tighter than that of Fano's, however, our original upper bound, which uses Renyi's conditional entropy, is tighter than their upper bound. In fact, if we increase b towards one, we can get the upper bound tighter.

As a second example, we evaluate our information theoretic bounds for a QPSK digital communication scheme over an AWGN channel. This scheme can be considered a four-class problem where each class has a two-dimensional Gaussian distribution (due to channel noise) centered at the class-mean (given by the symbols). The energy per transmitted bit is E_b and the PSD for the additive white Gaussian noise is $N_0/2$. In this problem, it is possible to evaluate the exact values for average bit error rate, p_e, and all the probability distributions required for the evaluation of the bounds in terms of Q-functions. Under these assumptions, the confusion matrix for a correlation receiver is given by

$$P_{W|M}^{QPSK} = \begin{bmatrix} (1-Q_1)^2 & Q_1 * (1-Q_1) & Q_1^2 & Q_1 * (1-Q_1) \\ Q_1 * (1-Q_1) & (1-Q_1)^2 & Q_1 * (1-Q_1) & Q_1^2 \\ Q_1 * (1-Q_1) & Q_1 * (1-Q_1) & (1-Q_1)^2 & Q_1 * (1-Q_1) \\ Q_1^2 & Q_1^2 & Q_1 * (1-Q_1) & (1-Q_1)^2 \end{bmatrix} \tag{19}$$

where $Q_1 = Q\left(\sqrt{2E_b / N_0}\right)$. The priors for symbols were assumed equal, i.e. $p(m_k) = 1/4, \quad k = 1,2,3,4$. Fig. 3 shows the probability of error and the original lower and upper bounds for this scenario.

Finally, we demonstrate on the QPSK example the estimation accuracy for the bounds when samples are used to estimate the confusion matrix and the prior probabilities instead of the ideal Q-function expressions. Using only a small number of samples, it is possible to get highly accurate estimations of the original bounds. In the example below, the average of 1000 Monte Carlo runs each with a total of 500 randomly selected samples (approximately 125 from each symbol) is presented. As expected, as the bit-energy-to noise-power-ratio increases the estimates become more accurate.

8. Conclusions

Fano’s bound is a widely appreciated inequality that has applications in the proofs of many key theorems in information theory and its applications to communication theory. Besides, it is useful for it offers an explanation and insights as to how information theory melds with pattern recognition. In order to design more efficient, more robust, and more robust pattern recognition systems, it is imperative to understand how the information propagation through classifiers and feature extractors affect the overall performance of the pattern recognition system in terms of the final product, the probability of classification error. Fano’s lower bound is significant because it provides the designer the limits of attainable performance. An upper bound, on the other hand is necessary to guarantee that in the system optimization phase worst-case performance of the final product is improved within known bounds.

In this paper, inspired by the work of Fano and utilizing Renyi’s definitions of entropy and mutual information, we have derived a family of lower and upper bounds for the probability of classification error in which the free parameter of Renyi’s definitions identifies which specific bound in this family is selected. It has been determined previously that In the family of lower bounds, which encompass Fano’s as a special case, Fano’s bound is the tightest. In the family of upper bounds, similarly and interestingly, tighter results are obtained when the entropy parameter approaches one, which corresponds to Shannon’s entropy.

In the theoretical development section, we have also provided some modifications to the originally proposed bounds in order to replace the term dependent on the probability of error. The modification we have proposed for the lower bounds was determined to be extremely loose, sometimes resulting in negative values, which is ineffective. This is not a problem, however, because Fano’s bound and especially Feder and Merhav’s lower bound proved to be extremely tight, providing a strong alternative. The modifications proposed for the upper bounds, although causing some performance loss with respect to the original bounds, have compared competitively with the Feder and Merhav upper bound in the numerical case studies The original upper bounds have proved themselves to be the tightest.

In the numerical studies, we have also demonstrated how to apply the proposed bounds to realistic problems, exemplified in a QPSK digital communication scheme. Evaluations of the bounds using theoretical and sample-estimated versions (Monte Carlo simulations were performed in this scheme) of the confusion matrix and the prior probabilities have confirmed that the bounds are useful in bracketing the probability of classification error from above and below accurately and tightly.

Although we have presented alternative bounds using conditional entropy, joint entropy and mutual information, only the conditional entropy bound was utilized in all considerations and simulations as it was shown before that the values of the bounds offered by these alternative quantities were similar to each other, therefore, either one of these bounds can be utilized in practice interchangeably.

The key theoretical conclusion is that, by training classifiers to maximize the mutual information between its input and output spaces, its probability of error must decrease. Similar conclusions apply to optimal feature extraction that will result in minimal classification error probability; it is essential to consider the progression of information through the designed systems in order to assure their optimality.

We have finally pointed out that these kinds of information theoretic bounds always require an amount of information that is usually sufficient to get an estimate of the probability of error itself. These bounds, however, can be helpful in determining confidence intervals for this probability of error if one can determine the estimation variance in these quantities arising from the finite number of samples. This line of further research stands as an interesting and worthwhile subject to focus future effects onto.

Acknowledgments: This work is partially supported by NSF grant ECS-9900394.

References

[1] T. Cover, J. Thomas, *Elements of Information Theory*, New York: John Wiley, 1991.

[2] R. Linsker, Towards an Organizing Principle for a Layered Perceptual Network, *Neural Information Processing Systems*, 1988, pp.485-494.

[3] K. Torkkola, “Visualizing Class Structure in Data Using Mutual Information,” *Proceedings of Neural Networks for Signal Processing X*, Sydney, Australia, Dec 2000, pp.376-385.

[4] K. Fu, “Statistical Pattern Recognition,” in *Adaptive, Learning and Pattern Recognition* Systems, (Ed. Mendel and Fu), New York: Academic Press, 1970, pp.35-76.

[5] K. Fukunaga, *An Introduction to Statistical Pattern Recognition*, New York : Academic Press, 1972.

[6] G. Deco, D. Obradovic, *An Information Theoretic Approach to Neural Computing*, New York: Springer, 1996.

[7] J. Fisher, *Nonlinear Extensions to the MACE Filter*, Ph.D. Dissertation, University of Florida, 1997.

[8] B. Ripley, *Pattern Recognition and Neural Networks*, New York: Cambridge University Press, 1996.

[9] V. Vapnik, *The Nature of Statistical Learning Theory*, New York : Springer Verlag, 1995.

[10] R.M. Fano, *Transmission of Information: A Statistical Theory of Communications*, New York: MIT Press and John Wiley & Sons Inc, 1961.

[11] C.E. Shannon, “A Mathematical Theory of Communications,” *Bell Systems Technical Journal*, vol 27, 1948, pp.379-423,623-656.

[12] A. Renyi, *Probability Theory*, New York: American Elsevier Publishing Company Inc, 1970.

[13] R.G. Gallager, *Information Theory and Reliable Communication*, New York: John Wiley & Sons Inc, 1968.

[14] M.B. Bassat, J. Raviv, “Renyi’s Entropy and the Prbability of Error,” IEEE Transactions on Information Theory, vol. 24, no. 3, 1978, pp.324-330.

[15] M. Feder, N. Merhav, “Relations Between Entropy and Error Probability,” IEEE Transactions on Information Theory, vol. 40, no. 1, 1994, pp.259-266.

[16] T.S. Han, S. Verdu, “Generalizing the Fano Inequality,” IEEE Transactions on Information Theory, vol. 40, no. 4, 1994, pp.1247-1251.

[17] H.V. Poor, S. Verdu, “A Lower Bound on the Probability of Error in Multihypothesis Testing,” IEEE Transactions on Information Theory, vol. 41, no. 6, 1995, pp.1992-1994.

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

[19] D. Erdogmus, J.C. Principe, “Information Transfer Through Classifiers and its Relation to Probability of Error,” Proceedings of International Joint Conference on Neural Networks (IJCNN'01), Washington, DC, 2001, pp. 50-54.

[20] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press, 1995.

Appendix

Derivation of the Conditional-Entropy Bound: In this derivation, we will employ the well-known Jensen's inequality. This inequality is

Jensen's Inequality: Assume $g(x)$ is convex (if concave reverse inequality), $x \in [a,b]$, then for $\sum_k w_k = 1,\ w_k > 0$, we have $g\left(\sum_k w_k x_k\right) \le \sum_k w_k g(x_k)$.

We also write the conditional probability of error given a specific input class as

$$\begin{aligned} p(e \mid m_k) &= \sum_{j \neq k} p(w_j \mid m_k) \\ 1 - p(e \mid m_k) &= p(w_k \mid m_k) \end{aligned} \tag{A.1}$$

Consider Renyi's conditional entropy of W given m_k.

$$\begin{aligned} H_a(W \mid m_k) &= \frac{1}{1-a} \log \sum_j p^a(w_j \mid m_k) \\ &= \frac{1}{1-a} \log \left[\sum_{j \neq k} p^a(w_j \mid m_k) + p^a(w_k \mid m_k) \right] \\ &= \frac{1}{1-a} \log \left[p^a(e \mid m_k) \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^a + (1 - p(e \mid m_k))^a \right] \end{aligned} \tag{A.2}$$

Using Jensen's inequality, and (A.1), we obtain two inequalities for $a > 1$ and $a < 1$ cases.

$$\begin{aligned} H_a(W \mid m_k) &\underset{a<1}{\overset{a>1}{\underset{\ge}{\le}}} p(e \mid m_k) \frac{1}{1-a} \log p^{a-1}(e \mid m_k) \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^a \\ &\qquad + (1 - p(e \mid m_k)) \frac{1}{1-a} \log (1 - p(e \mid m_k))^{a-1} \\ &= H_S(e \mid m_k) + p(e \mid m_k) \frac{1}{1-a} \log \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^a \end{aligned} \tag{A.3}$$

Recall that for an (N_c-*1*)-point entropy we have

$$\frac{1}{1-a} \log \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^a \le \log(N_c - 1) \tag{A.4}$$

equality being achieved for a uniform distribution. Hence, for $a > 1$, from (A.3) and (A.4) we obtain

$$H_a(W \mid m_k) \le H_S(e \mid m_k) + p(e \mid m_k) \log(N_c - 1) \tag{A.5}$$

Finally, using Bayes' rule on the conditional distributions and entropies we get the lower bound for p_e.

$$H_a(W \mid M) \le H_S(e) + p_e \log(N_c - 1) \tag{A.6}$$

For $a < 1$, from (A.3) we have

$$\begin{aligned} H_a(W \mid m_k) &\ge H_S(e \mid m_k) + p(e \mid m_k) H_a(W \mid e, m_k) \\ &\ge H_S(e \mid m_k) + p(e \mid m_k)[\min_k H_a(W \mid e, m_k)] \end{aligned} \tag{A.7}$$

where the 'conditional entropy given an error is made in classification and actual class was m_k' is

$$H_a(W \mid e, m_k) = \frac{1}{1-a} \log \sum_{j \ne k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^a \tag{A.8}$$

Finally, combining these results and fusing Fano's special case into the lower bound, we obtain the following interval for classification error probability.

$$\frac{H_a(W \mid M) - H_S(e)}{\log(N_c - 1)} \le p_e \le \frac{H_b(W \mid M) - H_S(e)}{\min_k H_b(W \mid e, m_k)}, \quad \begin{array}{l} a \ge 1 \\ b < 1 \end{array} \tag{A.9}$$

Applying Feder &Merhav Bound to Misclassification Probability: Feder & Merhav inequality is originally designed for minimal prediction error probability in estimating the value of a discrete random variable. It is derived in [Feder&Merhav] as follows:

Feder & Merhav Inequality: Suppose a discrete random variable X has the pmf $\{p(x_1),\ldots, p(x_M)\}$. Define the minimal prediction error probability as $p = 1\text{-}\max_k p(x_k)$ and let H be the Shannon's entropy for X. Then it can be shown that $f_*^{-1}(H) \ge p \ge \Phi^{-1}(H)$, where $f_*(.)$ and $\Phi(.)$ are convex and concave functions respectively.

In order to apply this to a classifier, consider the following substitutions. $X_k \sim (W \mid m_k)$ with pmf $\{p(w_1|m_k),\ldots, p(w_N|m_k)\}$ and assuming that $p(w_k|m_k)$ is the maximum component, define $p_k = 1\text{-}\ p(w_k|m_k)$. Let H_k be the Shannon's entropy for X_k. Notice that $p(e|m_k) = p_k$, therefore $p_e = \sum_k p(m_k) p_k$. Now we can apply Feder & Merhav's inequality to each of these variables X_k for $k = 1,\ldots,N$. Taking the weighted average of these inequalities in accordance with their corresponding prior probabilities, we get the following inequality for the probability of error.

$$\sum_{k=1}^{N} p(m_k) f_*^{-1}(H_k) \ge p_e \ge \sum_{k=1}^{N} p(m_k) \Phi_*^{-1}(H_k) \tag{A.10}$$

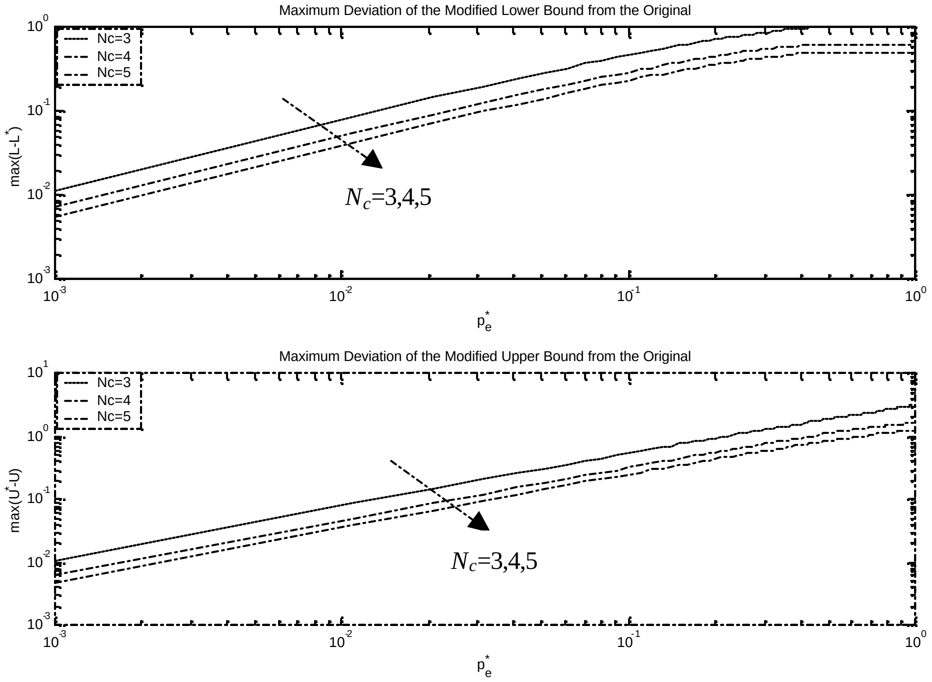


Fig. 1. Maximum deviation of (a) lower and (b) upper bounds from the original bounds after the modifications, as a function of p_e^* for various values of N_c.

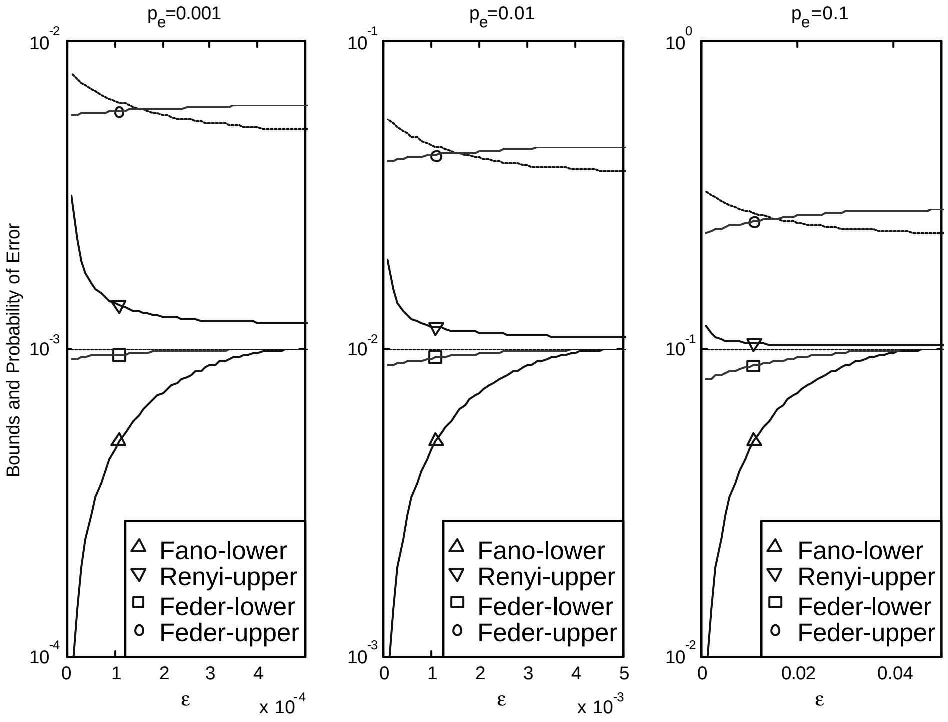


Fig. 2. Probability of error – constant with respect to e (dotted), original Fano's lower (Δ) and Reny's upper (∇) bounds, Feder and Merhav's upper (o) and lower ( ) bounds, and the modified Reny's upper bound (dotted) vs e.

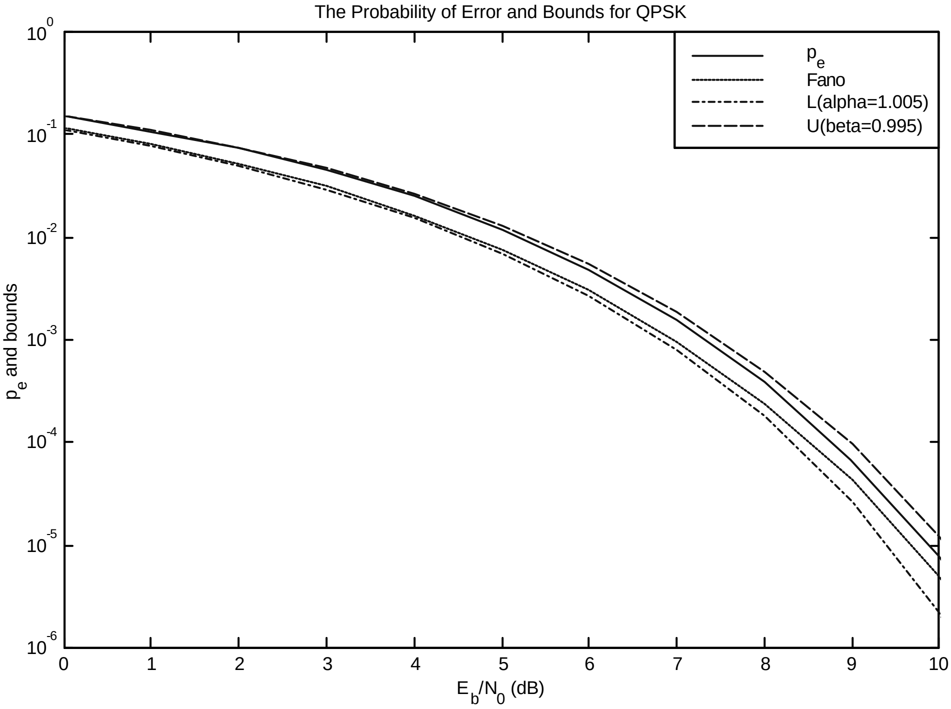


Fig. 3. Probability of error and its bounds for QPSK.

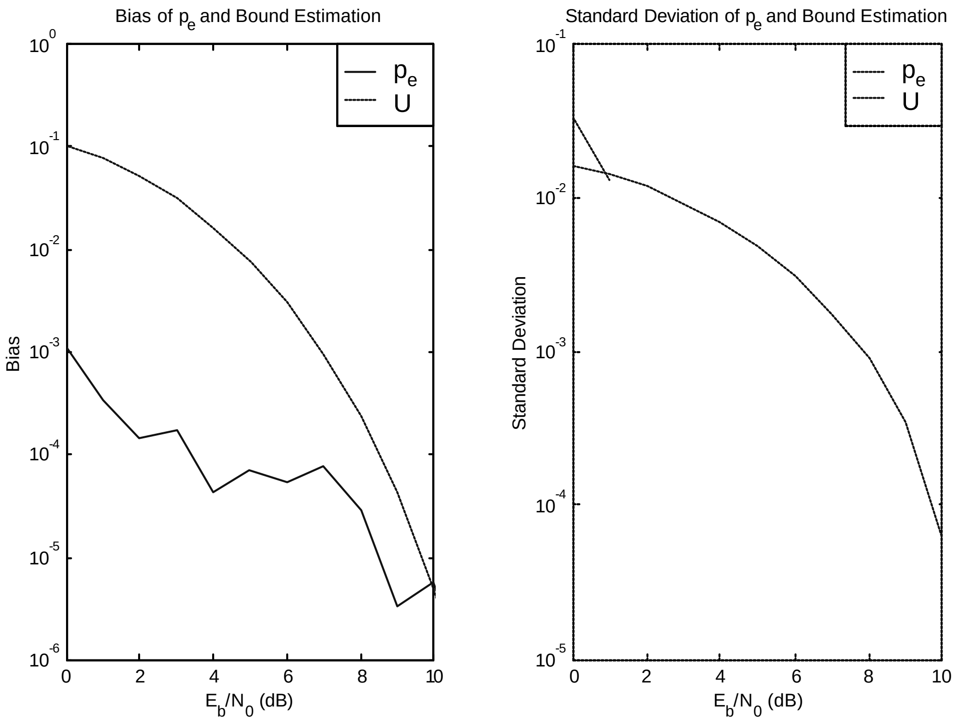


Fig. 4. Bias and standard deviation in estimating the upper bound in the previous QPSK example using 500 symbols as a function of bit-energy-to-noise-power-ratio.

Figure Captions

Fig. 1. Maximum deviation of (a) lower and (b) upper bounds from the original bounds after the modifications, as a function of p_e^* for various values of N_c.

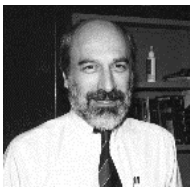
Fig. 2. Probability of error – constant with respect to e (dotted), original Fano's lower (Δ) and Reny's upper (∇) bounds, Feder and Merhav's upper (o) and lower ( ) bounds, and the modified Reny's upper bound (dotted) vs e.

Fig. 3. Probability of error and its bounds for QPSK.

Fig. 4. Bias and standard deviation in estimating the upper bound in the previous QPSK example using 500 symbols as a function of bit-energy-to-noise-power-ratio.

Author Biographies

Deniz Erdogmus received his B.S. in electrical & electronics engineering and B.S. in mathematics in 1997, and his M.S. in electrical & electronics engineering, with emphasis on systems and control, in 1999, all from the Middle East Technical University, Ankara, Turkey. From 1997 to 1999, he worked as a research engineer at the Defense Industries Research and Development Institute (SAGE) under The Scientific and Technical Research Council of Turkey (TUBITAK). He has received his Ph.D. degree from the Electrical and Computer Engineering Department of the University of Florida, under the supervision of Jose C. Principe, in May 2002. His current research interests include information theory, its applications to adaptive systems, and adaptive systems for signal processing, communications and control. He is a member of Tau Beta Pi and Eta Kappa Nu, and IEEE.



Jose C. Principe is Professor of Electrical and Computer Engineering and Biomedical Engineering at the University of Florida where he teaches advanced signal processing, machine learning and artificial neural networks (ANNs) modeling. He is BellSouth Professor and the Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). His primary area of interest is processing of time varying signals with adaptive neural models. The CNEL Lab has been studying signal and pattern recognition principles based on information theoretic criteria (entropy and mutual information).

Dr. Principe is an IEEE Fellow. He is the Chair of the Technical Committee on Neural Networks of the IEEE Signal Processing Society, Member of the Board of Governors of the International Neural Network Society, and Editor in Chief of the IEEE Transactions on Biomedical Engineering. He is a member of the Advisory Board of the University of Florida Brain Institute. Dr. Principe has more than 70 publications in refereed journals, 10 book chapters, and 160 conference papers. He directed 35 Ph.D. dissertations and 45 Master theses. He recently wrote an interactive electronic book entitled “Neural and Adaptive Systems: Fundamentals Through Simulation” published by John Wiley and Sons.