INSIGHTS ON THE RELATIONSHIP BETWEEN PROBABILITY OF MISCLASSIFICATION AND INFORMATION TRANSFER THROUGH CLASSIFIERS

D Erdogmus, JC Principe

Computational NeuroEngineering Laboratory, University of Florida, Gainesville, USA
deniz@cnel.ufl.edu, principe@cnel.ufl.edu

Received: 21/09/2001
Accepted in the final form: 21/01/2002

Abstract. Using Shannon's definition of entropy, Fano's bound identifies an inequality that lower bounds the error probability in a communication channel. The expression is important in communication theory, and it offers insights on how classification performance is affected by the information transfer through linear or nonlinear transformations. In this paper, we derive a family of lower bounds using Renyi's entropy, which yield Fano's lower bound as a special case. Using various values for the free parameter in Renyi's definition of entropy, it also becomes possible to construct a *family of upper bounds* for the probability of error, which was impossible using Shannon's definition of entropy. Further analysis to obtain the tightest lower and upper bounds revealed the fact that Fano's bound is indeed the tightest lower bound, and the upper bounds become tighter as the free parameter approaches one from below. Numerical evaluations of the bounds are provided for two cases, including baseband QPSK communication with AWGN.

Keywords: Classifier Performance, Error Bounds, Renyi's Entropy, Fano's Inequality

Symbols:

M, W: Discrete random variables with probability mass $\{p(m_k)\}_{k=1}^{N_c}$ and $\{p(w_k)\}_{k=1}^{N_c}$
$p(w_j, m_k)$: Joint probability mass function of M and W
$p(w_j|m_k)$: Conditional probability mass function of W given M
e: Discrete random variable with distribution $\{p_e, 1-p_e\}$

1 Introduction

In the information theory literature, Fano's bound is a well-known inequality that is essential to prove the converse to Shannon's second theorem [1]. As it is well known, Shannon's capacity theorem states the rate conditions to transfer information through a channel with arbitrary low probability of error. Shannon's brilliant insight was to model the communication channel as a system that includes uncertainty in the transmission and as such its effect can be modeled by conditional entropy, but in reality we use channels close to the Shannon's bound so our messages have a non-vanishing probability

of being corrupted by noise. In this case it is important to quantify the probability of error. Fano's bound is exactly the result we are seeking, since it expresses the probability of error as a function of the conditional entropy.

One of the great assets of information theory is the abstract level of the analysis. Hence, there are many other domains where the same type of ideas and formulation applies. We are particularly interested in adaptive systems' learning, so we would like to mention Linker's infomax principle [2] and feature extraction in statistical pattern recognition [3]. Linsker's infomax is a principle of self-organization for multilayer systems. The basic idea is to require that an optimal sub-subsystem should transfer as much information as possible from its input to its output (maximize the mutual information between its input and output). The analogy with the communication channel is obvious.

We know that when signals are mapped to subspaces, information is not preserved [4], but even in these cases the goal of the system designer should be to transfer as much information as possible from the input to the output. A good example of this type of subspace mappings is feature extraction. The issue of choosing optimal features has been central to pattern recognition, and concepts of information theory have been utilized since the early 60's by Fu [5] and others. Feature extraction can be formulated as a process that projects data from a high to a smaller dimensional space, preserving the discriminability among the classes. Conversely, feature extraction can be formulated as maximizing the mutual information between the output of the mapper and the desired (classification) output. We have recently utilized this concept to train neural networks directly from samples for optimal feature extraction using a nonparametric estimator based on Renyi's entropy [6]. In all of these cases, Fano's bound appears as the central-piece because it relates classification error to conditional entropy [7, 8].

Unfortunately for statistical pattern recognition and machine learning, Fano's bound is not of practical use to determine the probability of error (p_e) since it provides a lower bound to p_e, although it is important because it points out the best possible performance that can be attained. The goal in classification is to minimize p_e, hence a lower bound is not as useful as an upper bound. Unfortunately, upper bounds for p_e are not easy to compute nor that tight [9, 10].

Fano's inequality is based on Shannon's definition of entropy, as this definition was the only one available to him at the time [11]. Inspired by Shannon's exemplary work [12], many researchers concentrated their efforts on information theory. One of them was Alfred Renyi, who was able to formulate the theory of information starting from basic postulates [13]. Renyi was able to establish a profound mathematical theory for the concept of information and he devised alternative expressions for quantities like entropy and mutual information for which Shannon's definitions became special cases.

Motivated by these facts, we have developed a family of lower bounds, using Renyi's definitions of information theoretic quantities that are counterparts of Fano's bound. Exploiting the advantage of having a free parameter in Renyi's definitions, we were also able to formulate a *family of upper bounds*, which was impossible to achieve with Shannon's definitions. As a result, we are able to bracket the classification error probability by utilizing different values of the mentioned parameter.

The organization of this paper is as follows. First, we provide the definitions of all information theoretic quantities, both Shannon's and Renyi's versions, that will be necessary in formulating the bounds. Second, we will review Fano's bound. Next we present the derivation of the lower and upper bounds for probability of error using the conditional entropy, and state the equivalent expressions when joint entropy and mutual information are utilized. We then investigate the effect of the free parameter in Renyi's entropy on the value of entropy and making use of these results, we put forward several modifications that can be applied on the lower and upper bound expressions. Next, we study the optimal values for the free parameter that yield the tightest lower and upper bounds, and show that Fano's bound happens to be the tightest in the family of lower bounds. Finally, we present numerical evaluations of the bounds in a number of case studies, to provide a better understanding of their performances.

2 Definitions of Information Theoretical Quantities

Several information theoretical quantities are of interest for the development of the lower and upper bounds of the classification error probability. These are joint entropy, average conditional entropy, and average mutual information. We will drop the term 'average' in these from now on. In the application of the below arguments to classifiers, we use the random variable M to denote the actual class of a sample (called the input class), and W to denote the decided class of a sample (called the output class). The random variable e is used to denote the events of erroneous and correct classification with probabilities $\{p_e, 1-p_e\}$.

Shannon's Definitions: Shannon's entropy definition is the inspiration for all the quantities mentioned. For discrete random variables M and W, whose probability mass functions (pmf) are given by $\{p(m_k)\}_{k=1}^{N_c}$ and $\{p(w_j)\}_{j=1}^{N_c}$, Shannon's entropy is given by [12]

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

Based on this definition of entropy, the joint entropy, mutual information, and conditional entropy are defined as [11]

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

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 $p(m_k \mid w_j)$ are the joint probability mass function and the conditional probability mass function of the event in M and W, respectively. Shannon's mutual information is equal to the Kullback-Leibler divergence [14] between the joint distribution and the product of marginal distributions and it satisfies the following desirable property [11].

$$I_s(M,W) = H_s(M) - H_s(M \mid W) \tag{4}$$

Renyi's Definitions: Renyi's entropy for M with $\{p(m_k)\}_{k=1}^{N_c}$ is given by [13]

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

where α is a real positive constant different from 1. Accordingly, we get the average mutual information and average conditional entropy expressions as [13]

$$H_\alpha(M,W) = \frac{1}{1-\alpha} \log \sum_{k=1}^{N_c}\sum_{j=1}^{N_c} p^\alpha(m_k,w_j)$$

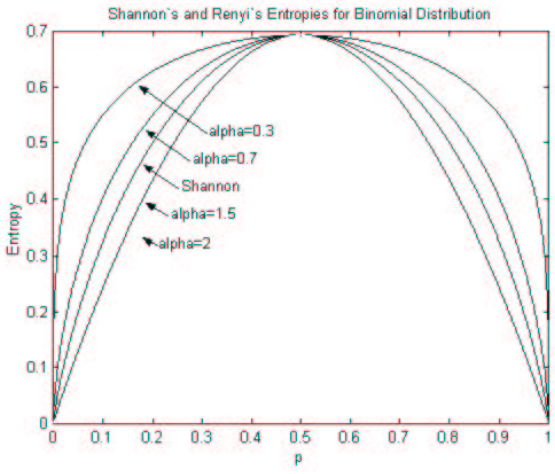

Figure 1: Shannon's and Renyi's entropies for a binomial distribution

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

where

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

We will see that the free parameter α of the Renyi's definitions is helpful in bracketing the probability of error from above and below. From a mathematical point of view, the parameter α acts as a variable that adjusts the value of entropy assigned to a given distribution without changing the ordering among distributions. From an engineering point of view, on the other hand, the role of α as a free parameter is to provide flexibility to the designer in determining an entropy definition that will best fit the problem at hand. In order to demonstrate the effect of α on the value of entropy, consider the Bernoulli distribution $\{p, 1-p\}$. In Figure 1, Shannon's and Renyi's entropies of this distribution as a function of p are shown for various values of α. It is worth noting that in the limit as $\alpha \to 1$ Renyi's definitions approach those of Shannon's (this is easily seen using L'Hopital's rule). This fact is also evident from Figure 1.

3 Fano's Bound

While working on the applications of information theory to digital communications, Fano determined a lower bound to the probability of error for the classification in discrete-symbol communication systems [11]. In this scheme, the symbols are selected from a discrete symbol set consisting of N_c elements. Each symbol m_k in the set has a known prior probability $p(m_k)$. At the receiver side, the signal space is partitioned into M mutually exclusive sets w_k where the decision is made according to which set the received signal falls into. The conditional probability of 'k-th symbol was sent while the decision was j-th symbol' is denoted by $p(m_k \mid w_j)$. Then, Fano's lower bound for the probability of classification error is written as

$$
p_e \geq \frac{H_S(M \mid W) - H_S(e)}{\log(N_c - 1)} \tag{8}
$$

This original inequality had found its way to the classification literature with slight modifications that arise from the relationship between Shannon's conditional entropy and mutual information given in (4). Substituting the appropriate terms for the conditional entropy in (8), assuming base-2 log to replace $H_S(e)$ with its maximum possible value, which is 1, and in addition, replacing $(N_c - 1)$ with N_c to accommodate for 2-class problems Fano's inequality becomes [8]

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

4 Bounds Using Renyi's Conditional Entropy

In this section, we will derive the lower and upper bound expressions using the conditional entropy. We will need Jensen's inequality.

Jensen's Inequality: Assume $g(x)$ is convex, $x \in [a, b]$, then

$$g\left(\sum_k w_k x_k\right) \leq \sum_k w_k g(x_k)$$

where $\sum_k w_k = 1$, and $w_k > 0$. If $g(x)$ is concave, then the inequality is reversed.

The following identities on the conditional error probabilities will be useful in the derivation of the bounds.

$$\begin{aligned} p(e \mid m_k) &= \sum_{j \neq k} p(w_j \mid m_k) \\ 1 - p(e \mid m_k) &= p(w_k \mid m_k) \end{aligned} \tag{10}$$

Consider Renyi's conditional entropy of W given m_k.

$$\begin{aligned} H_\alpha(W \mid m_k) &= \frac{1}{1-\alpha} \log \sum_j p^\alpha(w_j \mid m_k) \\ &= \frac{1}{1-\alpha} \log \left[\sum_{j \neq k} p^\alpha(w_j \mid m_k) + p^\alpha(w_k \mid m_k) \right] \\ &= \frac{1}{1-\alpha} \log \left[p^\alpha(e \mid m_k) \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^\alpha + (1 - p(e \mid m_k))^\alpha \right] \end{aligned} \tag{11}$$

Using Jensen's inequality, and (9), we obtain two inequalities for $\alpha > 1$ and $\alpha < 1$ cases.

$$\begin{aligned} H_\alpha(W \mid m_k) \underset{\alpha<1}{\overset{\alpha>1}{\lessgtr}} & p(e \mid m_k) \frac{1}{1-\alpha} \log p^{\alpha-1}(e \mid m_k) \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^\alpha \\ & + (1 - p(e \mid m_k)) \frac{1}{1-\alpha} \log (1 - p(e \mid m_k))^{\alpha-1} \\ & = H_s(e \mid m_k) + p(e \mid m_k) \frac{1}{1-\alpha} \log \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^\alpha \end{aligned} \tag{12}$$

Notice that the following inequality holds for an $(N_c - 1)$-point entropy, and the equality is achieved when the distribution is uniform.

$$\frac{1}{1-\alpha} \log \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^\alpha \leq \log(N_c - 1) \tag{13}$$

Hence, for $\alpha > 1$, from (12) and (13) we obtain

$$H_\alpha(W \mid m_k) \leq H_S(e \mid m_k) + p(e \mid m_k)\log(N_c - 1) \tag{14}$$

Using Baye's rule on the conditional distributions and entropies we get the lower bound for p_e.

$$H_\alpha(W \mid M) \leq H_S(e) + p_e \log(N_c - 1) \tag{15}$$

For $\alpha < 1$, from (12) we have

$$\begin{aligned} H_\alpha(W \mid m_k) &\geq H_S(e \mid m_k) + p(e \mid m_k) H_\alpha(W \mid e, m_k) \\ &\geq H_S(e \mid m_k) + p(e \mid m_k)[\min_k H_\alpha(W \mid e, m_k)] \end{aligned} \tag{16}$$

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

$$H_\alpha(W \mid e, m_k) = \frac{1}{1-\alpha}\log \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^\alpha \tag{17}$$

At this step, one can also obtain a tighter upper bound by using the average instead of the minimum of these entropies (see Appendix A). Again using Baye's rule, we obtain the upper bound for the probability of classification error from (16) as

$$H_\alpha(W \mid M) \geq H_S(e) + p_e[\min_k H_\alpha(W \mid e, m_k)] \tag{18}$$

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

$$\frac{H_\alpha(W \mid M) - H_S(e)}{\log(N_C - 1)} \leq p_e \leq \frac{H_\beta(W \mid M) - H_S(e)}{\min_k H_\beta(W \mid e, m_k)}, \quad \begin{matrix} \alpha \geq 1 \\ \beta < 1 \end{matrix} \tag{19}$$

It is interesting to note here that, the free parameter of Renyi's entropy definition allows us to obtain lower, and upper bounds, thus bracket the error probability. Observe that the denominator of the upper bound is always smaller than that of the lower bound. Also, we will see in the next section that the numerator of the upper bound is always greater than that of the lower bound. Due to these facts, the upper bound is always larger than the lower bound expression.

Going through a similar process, one can obtain lower and upper bounds for p_e using Renyi's joint entropy and mutual information definitions. The derivations are provided in Appendix B. Here, we summarize the results.

$$\frac{H_\alpha(W, M) - H_S(M) - H_S(e)}{\log(N_C - 1)} \leq p_e \leq \frac{H_\beta(W, M) - H_S(M) - H_S(e)}{\min_k H_\beta(W \mid e, m_k)}, \quad \begin{matrix} \alpha \geq 1 \\ \beta < 1 \end{matrix} \tag{20}$$

$$\frac{H_S(W) - I_\alpha(W; M) - H_S(e)}{\log(N_C - 1)} \leq p_e \leq \frac{H_S(W) - I_\beta(W; M) - H_S(e)}{\min_k H_S(W \mid e, m_k)}, \quad \begin{matrix} \alpha \geq 1 \\ \beta < 1 \end{matrix} \tag{21}$$

Although the additivity of information in the manner given by (4) does not hold for Renyi's definitions of the corresponding quantities, we observe that similar relations arise in the above bounds when the joint entropy or the mutual information replaces Renyi's conditional entropy. Also note that, in these expressions if α is chosen as 1, then the lower bound expressions become equivalent to Fano's bound (where Renyi's definitions with $\alpha = 1$ are assumed to be evaluated using Shannon's definitions).

These bounds provide an understanding of how a classifier, optimal in the sense of minimum error probability, should behave. By observing the numerators of lower and upper bounds in (21), we see that, the probability of error can be minimized, by making the mutual information $I_\alpha(W; M)$ larger.

We can regard this quantity as the transmitted information through the classifier as it represents the shared information between the output classes of the classifier and the input classes. In addition, the denominator of the upper bound in all the given forms suggests that we increase the entropy of the probability distribution over the wrong classes at the output, given that a classification error is made for a given input. Since entropy is maximized when the subjected distribution is uniform, this term suggests that the probability distribution among the erroneous decisions is made uniform. The intuition behind this is to make the wrong classes equally probable so that none of them stands out as a compelling distracter beside the correct decision. It must be noted, however, that the overall performance is a composite function of both the numerators and the denominators, thus it is imperative to find the *right* balance between maximizing the denominator of the upper bound and minimizing the numerator when trying to force the error probability to smaller values. This insight induces a possible new training criterion and learning rule for classifiers, one that has never been considered before. We can regard this phenomenon in analogy with the concept of *learning from mistakes*. It is possible to improve one's (the classifier's in this context) knowledge even from erroneous decisions, yet in regard of the requirement of the minimization of the numerator, this rule has to be utilized in conjunction with a learning rule that maximizes the mutual information between the input and the output. The details of how such an algorithm would be and how successful it would be is beyond the scope of this paper, since our main motivation here is to provide a theoretical understanding between the classifier performance and the information transfer through it. On the other hand, the possibility of a methodology that maximizes the *information transfer* while *extracting useful information from mistakes* is tempting to pursue further research towards the development of such training algorithms for classifiers.

5 The Effect of α on Entropy

It is important to investigate the effect of the parameter α on the value of entropy. This analysis will aid us understand how to choose this parameter so that we obtain the tightest bounds. In addition, it will also clarify how choosing different values of α result in a lower or upper bound.

Fact 1 *Given a discrete random variable X, if $0 < \alpha < 1$, and $1 < \beta$ then $H_\alpha(X) \geq H_S(X) \geq H_\beta(X)$ [6].*

Proof: We start from the definition of Renyi's entropy and apply Jensen's inequality.

$$\begin{aligned} H_\alpha(X) &= \frac{1}{(1-\alpha)} \log \sum_j p^\alpha(x_j) \\ &\underset{\alpha<1}{\overset{\alpha>1}{\lessgtr}} \sum_j p(x_j) \frac{1}{(1-\alpha)} \log p^{\alpha-1}(x_j) \\ &= -\sum_j p(x_j) \log p(x_j) = H_S(X) \end{aligned} \tag{22}$$

This fact will be useful in showing that the upper bound is, in fact, always greater than the lower bound.

Fact 2 *Given a discrete random variable X, if $1 < \alpha < \beta$, then $H_\alpha(X) \geq H_\beta(X)$.*

Proof: Since $1 < \alpha < \beta$. We have $p^\alpha(x_j) \geq p^\beta(x_j)$, and we can write the following inequality.

$$H_\alpha(X) = \frac{1}{(1-\alpha)} \log \sum_j p^\alpha(x_j) \geq \frac{1}{(1-\alpha)} \log \sum_j p^\beta(x_j)$$

$$\geq \frac{\alpha-1}{\beta-1}\frac{1}{1-\alpha}\log\sum_j p^\beta(x_j) = \frac{1}{1-\beta}\log\sum_j p^\beta(x_j)$$
$$= H_\beta(X) \tag{23}$$

This fact will be useful in proving that Fano's bound is the tightest lower bound.

6 Obtaining Looser Bounds

Now that we know the relationship between the Shannon's entropy and Renyi's entropy for choices of α, we can construct lower and upper bound expressions that involve only Renyi's entropy, by replacing the Shannon's entropy by appropriate terms. This substitution, however, may result in looser bounds. From Fact 1 we know that $H_{\overline{\alpha}}(e) \geq H_S(e) \geq H_{\overline{\beta}}(e)$ when $0 < \overline{\alpha} < 1$ and $1 < \overline{\beta}$. Thus, we can replace $H_S(e)$ with these two quantities to obtain the following bounds.

$$\frac{H_\alpha(W \mid M) - H_{\overline{\alpha}}(e)}{\log(N_c - 1)} \leq p_e \leq \frac{H_\beta(W \mid M) - H_{\overline{\beta}}(e)}{\min_k H_\beta(W \mid e, m_k)}, \quad \begin{matrix}\alpha \geq 1, & \overline{\alpha} < 1 \\ \beta < 1, & \overline{\beta} > 1\end{matrix} \tag{24}$$

It is possible to make similar substitutions in the bounds involving joint entropy and mutual information for the terms with Shannon's entropy.

In the lower bounds, it is also possible to make more drastic substitutions for the term $H_S(e)$ and the denominator as in Fano's bound. In the upper bounds, $H_S(e)$ can be replaced with 0. With these substitutions, the inequality reads

$$\frac{H_\alpha(W \mid M) - \log 2}{\log N_c} \leq p_e \leq \frac{H_\beta(W \mid M)}{\min_k H_\beta(W \mid e, m_k)}, \quad \begin{matrix}\alpha \geq 1 \\ \beta < 1\end{matrix} \tag{25}$$

It must be noted, however, that these replacements may severely degrade the tightness of the bounds, therefore may not be practically useful.

7 Finding the Tightest Bounds

Now that we have a family of lower and upper bounds (19, 20, 21), where α and β are the parameters for lower and upper bounds respectively, and the relationship between the entropy values for various values of these parameters, we are ready to address the problem of determining the tightest bounds. Consider the lower bound given in terms of Renyi's conditional entropy of W given M, given in (19). This lower bound is maximized when $H_\alpha(W \mid M)$ takes its maximum value with respect to α. Since $\alpha \geq 1$, choosing a smaller value for α, due to Fact 2, will maximize the conditional entropy, and eventually Shannon's entropy ($\alpha = 1$) will yield the maximum value for this quantity, due to Fact 1. Hence, the corresponding value gives the tightest lower bound for probability of error (i.e. Fano's bound). Although this tightest lower bound is not exactly equal to Fano's bound in (8), it could have been obtained using Fano's original proof presented in [11], with a slight modification of variable names. Therefore, we call this tightest lower bound, the Fano's bound.

Computing the best value of $\beta < 1$ to yield the tightest upper bound, however, is not as simple. In the upper bound expression in (19), two terms, one in the numerator, the other in the denominator, depends on the value of β. Therefore, the optimal value must be a balance between minimizing the conditional entropy in the numerator, and maximizing the conditional entropy in the denominator. Intuitively, we conjecture that for a broad range of classifiers, the upper bound will assume its tightest value when the parameter β approaches to 1, from below and infinitesimally close. In fact, we observe this behavior in the case studies that are to be presented in the following section. Thus, an interesting property of these information theoretical bounds arises: Shannon's entropy acts as a threshold point at which the transition from the lower bound domain to upper bound domain is made.

8 Practical Evaluation of Bounds from Samples

The lower and upper bound expressions obtained in the previous sections are significant theoretical results that enable us to identify the effect of information transfer through a classifier and its classification error probability. However, in practice, one has to estimate the required quantities from the samples. For this, there are plenty of approaches one can follow. 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 $\{p(m_k)\}_{k=1}^{N_c}$.

The first and the trivial approach is to estimate these probabilities by counting appropriate samples and then dividing by the total number of samples. Clearly, if the number of samples is large, this method will result in accurate estimations. A possible problem with this approach occurs in the small sample case, i.e. the number of samples is not large enough to estimate all the required probabilities accurately due to the sparsity of the samples.

The second approach, which may overcome this problem, is to use Parzen windowing, shifted histogram or similar methods to obtain continuous distribution estimates and then to integrate them over the appropriate regions to obtain the distributions for the discrete variables W and M. We have recently proposed a method to overcome the integration when we use Parzen windowing and Gaussian kernels [6, 7]. It is possible to formulate the windowing problem such that the same procedure can be applied to avoid the integrations.

The third approach is suitable for classifiers that are trained in the minimum MSE sense. We know that, in this configuration, the decision functions of the classifier tend to approximate the conditional probabilities $p(w_j \mid m_k), j = 1, \ldots, N_c$ when a sample from class m_k is introduced (especially in the case of multilayer perceptrons the outputs for each class are representations of these probabilities) [15]. One can exploit this property and obtain estimates for these conditional probabilities, and hence evaluate the lower and upper bounds. It should be noted that the accuracy of such an estimate would depend heavily on the classifier's ability to represent these conditional probability functions.

In any case, we stress here that it is not the accuracy of these bounds in estimating the probability of error, but it is the theoretical insights that they provide the designer towards understanding the relationship between the final performance and the amount of information residing in the data. The more important implications of these bounds are rather the possibilities they offer us in designing better classifiers in the sense of maximum *information transfer*.

9 Numerical Evaluation of the Bounds

In order to fully understand the behavior of these bounds, and to observe their performance in different situations we studied two cases. The first is a simple example to evaluate the lower and upper bounds for various values of α and β, and the second one is the evaluation of lower and upper bounds for a QPSK communication scheme over an AWGN channel (we name this example as the QPSK because this four-class scheme occurs in practice under the stated assumptions about the communication system). The case study is constructed as follows. Given three classes and a classifier with the following conditional probability matrix, where the ij-th entry denotes the probability of classifier decision is class-i, when the actual class was class-j.

$$P_{W|M} = \begin{bmatrix} 1-p_e & p_e-\varepsilon & \varepsilon \\ \varepsilon & 1-p_e & p_e-\varepsilon \\ p_e-\varepsilon & \varepsilon & 1-p_e \end{bmatrix} \tag{26}$$

Each column represents the distribution of probabilities among the output classes of the classifier, and the diagonal entries correspond to the probability of correct classification given the actual class. Hence, with these conditional probabilities, the bounds are indifferent to the actual class and their prior probabilities. In addition, the probability of error is fixed at 0.2 and is equal to the conditional probability of error given any actual class label. This way, the analysis is simplified to the investigation of the effect of probability distribution among the wrong classes. By varying ε in the interval $[0, p_e/2]$,

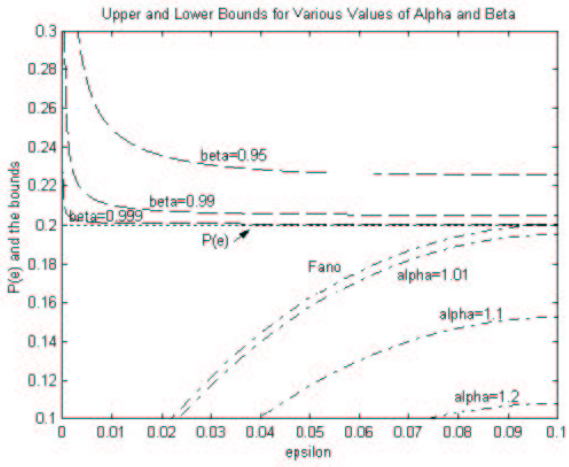

Figure 2: Family of lower and upper bounds for probability of error

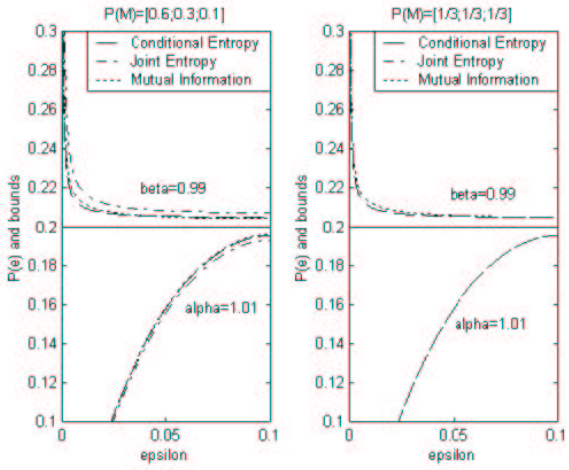

Figure 3: Bounds using the three schemes for different priors

we can study the performance of the bounds in terms of tightness. Figure 2 shows the lower and upper bounds of (19) for this example as a function of ε.

We observe that the family of lower bounds become tighter as α is decreased, eventually attaining Fano's bound as the tightest. Similarly, we observe that, the upper bounds become tighter as β is increased to approach 1. One other interesting observation is that the upper bounds remain virtually flat over a wide range of ε, suggesting that it practically provides a bound that is as tight for a broad variety of classifiers as the optimal classifier where the probability distribution among wrong classes is uniform ($\varepsilon = p_e/2$ in the three class case).

The following case study examines the differences between the three alternatives in (19, 20, 21), namely, conditional entropy, joint entropy, and mutual information bounds. Also, the effect of prior class probabilities on these three bounds is investigated. Figure 3 summarizes the results for two choices for prior probabilities, one of them uniform.

In this case study, we observe that that three expressions for the upper and lower bounds using alternative information theoretic quantities yield almost the same value, but their closeness is related to how the priors are distributed. We observe that if the prior distribution is close to uniform, then all three bounds are practically the same, and as the prior distribution diverges from uniform, the three bounds exhibit slight deviations from each other.

As a second example, we evaluate the information theoretic bounds for a baseband QPSK digital

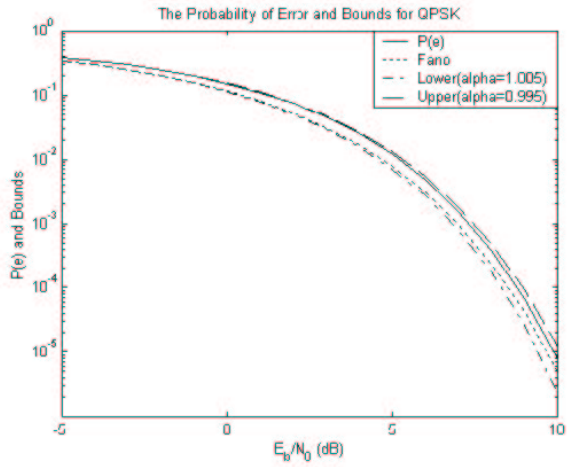

Figure 4: Probability of error and its bounds for QPSK

communication scheme over an AWGN channel. 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 and is given in (27).

$$P_{W|M}^{QPSK} = \begin{bmatrix} (1-Q_1)^2 & Q_1*(1-Q_1) & Q_1^2 & Q_1*(1-Q_1) \\ Q_1*(1-Q_1) & (1-Q_1)^2 & Q_1*(1-Q_1) & Q_1^2 \\ Q_1*(1-Q_1) & Q_1*(1-Q_1) & (1-Q_1)^2 & Q_1*(1-Q_1) \\ Q_1^2 & Q_1^2 & Q_1*(1-Q_1) & (1-Q_1)^2 \end{bmatrix} \tag{27}$$

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

As a last example, consider a 16-class problem where the input classes are centered as shown in Figure 5 with circularly symmetric two-dimensional Gaussian distributions located at these centers (this situation occurs in the 16-QAM modulation scheme with AWGN channel model in digital communications). The confusion matrix of this classification problem is 16x16 and similar to (27) in structure. We omit this matrix for the sake of saving space. Since the number of classes increased in this case, compared to the QPSK example, we expect the bounds to be looser, as the inequalities led to both denominators (in the lower and upper bounds) will be stronger. However, this phenomenon does not change the main conclusion of this work, which is about the relationship between the classification error probability and the amount of information transferred through the classifier.

10 Conclusions

Fano's bound is a widely recognized inequality in the information theory literature, and it provides an understanding of how probability of error in classification is related to the information transfer through a classifier from its input space to its output space. Fano's work, however, is based on Shannon's definition of entropy, which is a special case of Renyi's definition. In this paper, inspired by the work of Fano, we have derived a family of lower and upper bounds for the probability of error in which the free parameter of Renyi's entropy identifies which specific bound in this family is selected. An interesting result arising from these inequalities was that while the lower bounds employed Renyi's entropy with parameter greater than or equal to one (latter is the Fano's bound), the upper bounds utilized Renyi's entropy with parameter less than one. Thus, we were able to exploit this property of Renyi's entropy to acquire more information about the probability of error.

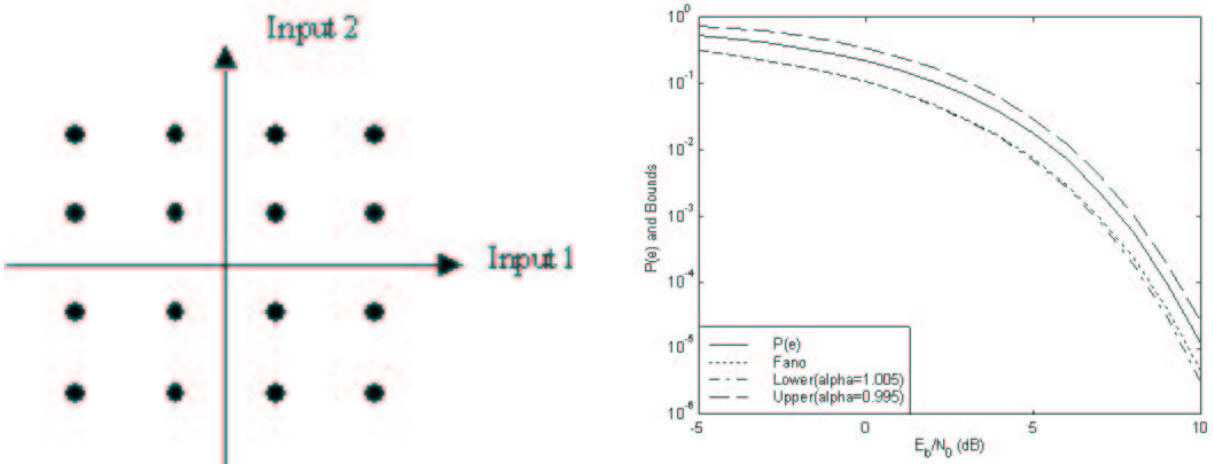

Figure 5: a) 16-QAM constellation; centers of classes in two dimensional input space b) Probability of error and its bounds for the 16-class case

The effect of the parameter of Renyi's entropy, on the tightness of the bounds, was also examined and it was proven that in the family of lower bounds, Fano's bound offers the tightest lower bound. As for the tightest upper bound, it was conjectured that, as the parameter approached to one (from below) the expression provided a tighter bound. This conjecture was supported by numerical evaluations, and in these evaluations, it was noted that the tightness of the upper bound was the same for a wide range of classifiers. Numerical evaluations for comparing the performance of bounds incorporating different information theoretic quantities, namely conditional entropy, joint entropy, and mutual information, revealed that there was practically little or no deviation among them. Although small, it was observed that the prior probabilities of the classes in the input space had an effect on the values of the lower and upper bounds.

In addition, the bounds for a QPSK communication scheme with AWG noise were evaluated as to demonstrate how these bounds would be applicable to real-life problems and it was shown that by appropriate choice of the family parameters, it is possible to obtain extremely tight bounds for the average bit error probability in this realistic problem.

Although not illustrated here, we mentioned briefly that it is possible to obtain estimates of the bounds by employing various nonparametric estimates for the probability mass functions that are required in the computation. The simplest of these estimators we have mentioned is the sample-count method. Our simulations have showed that with a reasonably small number of samples (around 500), the bounds for QPSK can be estimated with a small variance. Alternatively, neural networks can be trained to produce estimates of the desired conditional probabilities or nonparametric pdf estimation methods like Parzen windowing can be employed to obtain pdf estimates, which can then be integrated over the appropriate regions in the output space to yield estimates of the required conditional probabilities. As a final remark on this, in practice, it is possible to obtain an estimate of the probability of error with the information that is required to obtain an estimate of the bounds. Nevertheless, the bounds can still be informative and may be used as confirmation parameters for these estimates.

The key conclusion from all these bounds on misclassification probability is that, by training classifiers to maximize mutual information between its input and output vectors, its probability of error is forced to decrease. Similarly, for optimal feature extraction that will result in minimal classification error probability, one needs to consider the amount of information transferred, by the feature selection mechanism, from the raw data to the selected features. The bounds involving the conditional entropy, on the other hand, assert that to improve performance, the uncertainty of the output distribution (variation of the decisions given a sample) must be minimized; this is a validation of common sense.

Acknowledgments: This work is partially supported by NSF grant ECS-9900394. Special thanks to AP Engelbrecht, editor-in-chief of The International Journal of Computers, Systems and Signals (IJCSS), for encouraging the authors to submit this extended version of the paper given as [16] in the reference list. The authors also would like to thank to the anonymous reviewers whose comments helped improve the quality of the manuscript.

References

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

[2] R Linsker, "Towards an organizing principle for a layered perceptual network," *Neural Infor. Proc. Systems*, pp. 485–494, 1988.

[3] K Fukunaga, "An introduction to statistical pattern recognition," *Academic Press*, NY, 1972.

[4] G Deco, D Obradovic, *An Information Theoretic Approach to Neural Computing*, Springer, NY, 1996.

[5] K Fu, "Statistical pattern recognition," *Adaptive, Learning and Pattern Recognition Systems*, JM Mendel, KS Fu (Eds.), Academic Press, NY, pp. 35–76, 1970.

[6] J Principe, D Xu, J Fisher, "Information theoretic learning," *Unsupervised Adaptive Filtering*, S Haykin (Ed.), John Wiley, NY, pp. 265–319, 2000.

[7] J Fisher, *Nonlinear Extensions to the MACE Filter*, PhD Thesis, University of Florida, 1997.

[8] K Torkkola, WM Campbell, "Mutual information in learning feature transformations," *Proc. of Int. Conf. on Machine Learning*, Stanford, CA, USA, 2000.

[9] B Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, NY, 1996.

[10] V Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, NY, 1995.

[11] RM Fano, *Transmission of Information: A Statistical Theory of Communications*, MIT Press and John Wiley & Sons, NY, 1961.

[12] CE Shannon, "A mathematical theory of communications," *Bell Systems Tech. J.*, vol. 27, pp. 379–423, 623-656, 1948.

[13] A Renyi, *Probability Theory*, American Elsevier Publishing Company Inc., NY, 1970.

[14] S Kullback, *Information Theory and Statistics*, Dover Publications, NY, 1968.

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

[16] D Erdogmus, JC Principe, "Information Transfer Through Classifiers and its Relation to Probability of Error," *Proc. of Int. Joint Conf. on Neural Networks (IJCNN'01)*, Washington, DC, 2001.

Appendix A

Instead of applying the drastic minimum operator in (16), we could have obtained another bound as follows. We have the following inequality for the conditional error probability from (16).

$$p(e \mid m_k) \leq \frac{H_\alpha(W \mid m_k) - H_S(e \mid m_k)}{H_\alpha(W \mid e, m_k)} \tag{28}$$

We now multiply both sides with $p(m_k)$ and sum over all k to obtain an upper bound for error probability.

$$p(e) = \sum_k p(m_k)p(e \mid m_k) \leq \sum_k p(m_k) \left(\frac{H_\alpha(W \mid m_k) - H_S(e \mid m_k)}{H_\alpha(W \mid e, m_k)} \right) \tag{29}$$

Similar upper bounds with joint entropy and mutual information may be derived with the same argument. It is also possible to obtain tighter lower bounds in the same manner by taking the average of the individual bounds for the conditional probabilities instead of substituting $\log(N_c - 1)$ for the multiplier of $p(e \mid m_k)$ in (12).

Appendix B

Derivation of bounds using joint entropy and mutual information. Consider Renyi's joint entropy. Staring from the definition, and applying Bayes' rule and Jensen's inequality, for different values of α, we obtain two inequalities.

$$\begin{aligned}
H_\alpha(W, M) &= \frac{1}{1-\alpha} \log \sum_k \sum_j p^\alpha(w_j, m_k) = \frac{1}{1-\alpha} \log \sum_k \sum_j p^\alpha(w_j \mid m_k) p^\alpha(m_k) \\
&\underset{\alpha<1}{\overset{\alpha>1}{\gtrless}} \sum_k p(m_k) \frac{1}{1-\alpha} \log \sum_j p^\alpha(w_j \mid m_k) \\
&= \sum_k p(m_k) \left[-\log p(m_k) + \frac{1}{1-\alpha} \log \left[\sum_{j \neq k} p^\alpha(w_j \mid m_k) + p^\alpha(w_k \mid m_k) \right] \right] \\
&= H_S(M) + \sum_k p(m_k) \frac{1}{1-\alpha} \log \left[\sum_{j \neq k} p^\alpha(w_j \mid m_k) + p^\alpha(w_k \mid m_k) \right] \\
&\underset{\alpha<1}{\overset{\alpha>1}{\gtrless}} H_S(M) + \sum_k p(m_k) \left[H_S(e \mid m_k) + p(e \mid m_k) \frac{1}{1-\alpha} \log \sum_{j \neq k} \left(\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \right)^\alpha \right]
\end{aligned} \tag{30}$$

Hence, rearranging the terms, we obtain the following inequality

$$\frac{H_\alpha(W, M) - H_S(M) - H_S(e)}{\log(N_c - 1)} \leq p_e \leq \frac{H_\beta(W, M) - H_S(M) - H_S(e)}{\min_k H_\beta(W \mid e, m_k)}, \quad \begin{matrix} \alpha \geq 1 \\ \beta < 1 \end{matrix} \tag{31}$$

Now consider Renyi's mutual information. Once again applying Jensen's inequality in two steps, we can obtain the lower and upper bounds for error probability.

$$\begin{aligned}
I(M; W) &= \frac{1}{1-\alpha} \log \sum_k \sum_j \frac{p^\alpha(w_j, m_k)}{p^{\alpha-1}(w_j) p^{\alpha-1}(m_k)} = \frac{1}{1-\alpha} \log \sum_k \sum_j \frac{p^\alpha(w_j \mid m_k) p(m_k)}{p^{\alpha-1}(w_j)} \\
&\underset{\alpha<1}{\overset{\alpha>1}{\gtrless}} \sum_k p(m_k) \frac{1}{\alpha-1} \log \sum_j p^\alpha(w_j \mid m_k) p^{1-\alpha}(w_j) \\
&= \sum_k p(m_k) \frac{1}{\alpha-1} \log \left[\sum_{j \neq k} p^\alpha(w_j \mid m_k) p^{1-\alpha}(w_j) + p^\alpha(w_k \mid m_k) p^{1-\alpha}(w_k) \right]
\end{aligned}$$

$$\overset{\alpha>1}{\underset{\alpha<1}{\gtrless}} \sum_k p(m_k)\left[p(e \mid m_k)\frac{1}{\alpha-1}\log\left(p^{\alpha-1}(e \mid m_k)\sum_{j\neq k}\frac{p^{\alpha}(w_j \mid m_k)p^{1-\alpha}(w_j)}{p^{\alpha}(e \mid m_k)}\right)\right.$$
$$\left.+(1-p(e \mid m_k))\frac{1}{\alpha-1}\log(1-p(e \mid m_k))^{\alpha-1}p^{1-\alpha}(w_k)\right] \tag{32}$$

Now, rearranging the terms, and applying Jensen's inequality,

$$I_\alpha(M;W)\overset{\alpha>1}{\underset{\alpha<1}{\lessgtr}}\sum_k p(m_k)\left[-H_S(e \mid m_k)+p(e \mid m_k)\frac{1}{1-\alpha}\log\sum_{j\neq k}\frac{p^{\alpha}(w_j \mid m_k)p^{1-\alpha}(w_j)}{p^{\alpha}(e \mid m_k)}\right.$$
$$\left.-(1-p(e \mid m_k))\log p(w_k)\right]$$
$$=\sum_k p(m_k)\left[-H_S(e \mid m_k)+p(e \mid m_k)\log p(w_k)\right.$$
$$\left.-\log p(w_k)+p(e \mid m_k)\frac{1}{1-\alpha}\log\sum_{j\neq k}\frac{p^{\alpha}(w_j \mid m_k)p^{1-\alpha}(w_j)}{p^{\alpha}(e \mid m_k)}\right]$$
$$\overset{\alpha>1}{\underset{\alpha<1}{\lessgtr}}\sum_k p(m_k)\left[-H_S(e \mid m_k)+p(e \mid m_k)\log p(w_k)-\log p(w_k)\right.$$
$$\left.+p(e \mid m_k)\sum_{j\neq k}\frac{p(w_j \mid m_k)}{p(e \mid m_k)}\frac{1}{1-\alpha}\log\frac{p^{\alpha-1}(w_j \mid m_k)p^{1-\alpha}(w_j)}{p^{\alpha-1}(e \mid m_k)}\right]$$
$$=\sum_k p(m_k)\left[-H_S(e \mid m_k)+p(e \mid m_k)\log p(w_k)-\log p(w_k)\right.$$
$$\left.+p(e \mid m_k)\sum_{j\neq k}\frac{p(w_j \mid m_k)}{p(e \mid m_k)}\left[\log\frac{p(w_j \mid m_k)}{p(e \mid m_k)}-\log p(w_j)\right]\right]$$
$$=-H_S(e)+\sum_k p(m_k)\left[-p(w_k \mid m_k)\log p(w_k)+p(e \mid m_k)\sum_{j\neq k}\frac{p(w_j \mid m_k)}{p(e \mid m_k)}\log\frac{p(w_j \mid m_k)}{p(e \mid m_k)}\right.$$
$$\left.-\sum_{j\neq k}p(w_j \mid m_k)\log p(w_j)\right]$$
$$=-H_S(e)+H_S(W)+\sum_k p(m_k)p(e \mid m_k)\sum_{j\neq k}\frac{p(w_j \mid m_k)}{p(e \mid m_k)}\log\frac{p(w_j \mid m_k)}{p(e \mid m_k)} \tag{33}$$

Finally, rearranging the terms and substituting appropriate extreme values for multiplier of p_e, we obtain the following inequality on error probability in terms of Renyi's mutual information.

$$\frac{H_S(W)-I_\alpha(W;M)-H_S(e)}{\log(N_c-1)}\leq p_e\leq\frac{H_S(W)-I_\beta(W;M)-H_S(e)}{\min_k H_S(W \mid e,m_k)},\quad\begin{matrix}\alpha\geq 1\\ \beta<1\end{matrix} \tag{34}$$