INFORMATION TRANSFER THROUGH CLASSIFIERS AND ITS RELATION TO PROBABILITY OF ERROR

Deniz Erdogmus, Jose C. Principe
Computational NeuroEngineering Lab (CNEL), University of Florida, Gainesville, FL 32611
[deniz,principe]@cnel.ufl.edu

ABSTRACT

Fano's bound identifies a lower bound for the classification error probability and indicates how the information transfer through classifier affects its performance. It was an important step towards linking the information theory and pattern recognition. In this paper, a family of lower bounds is derived using Renyi's entropy, which yields Fano's lower bound as a special case. Using a different set of entropy orders, Renyi's definition also allows the construction a family of upper bounds for the probability of error. This is 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 entropy order approaches to one from below. Numerical evaluations of the bounds are presented for three digital modulation schemes under AWGN channel.

1. INTRODUCTION

Fano's bound is important in that it identifies the link between information transfer through a classifier and its probability of misclassification [1]. It employs Shannon's definition of entropy to arrive at a lower bound for the error probability for a classifier. However, Fano's bound cannot be utilized in evaluating classifier performance because it is a lower bound for a quantity we wish to minimize [2]. Our purpose in this paper is to present a family of information theoretical lower and upper bounds for the error probability of classifiers that encompass Fano's inequality as a special case.

We derive our bounds using the uncommon Renyi's definition of entropy rather than the widely recognized definition of Shannon. Renyi's definition is a parametric family of entropy values with the Shannon's entropy being the limit value when the parameter alpha of the family approaches to one [3]. In fact, it turns out that the existence of this parameter becomes useful in formulating an upper bound besides a lower bound for two disjoint sets of values it can take. Fano's inequality, corresponding to Renyi's entropy with parameter equal to one, then becomes a special case in the family of lower bounds. The conditional entropy of the classifier output given the input can be regarded as the average information transfer through the classifier, thus the version of the bounds which incorporates this quantity is significant in understanding the relationship between the information transfer and misclassification probability.

Shannon's marginal entropy of a random variable is equal to the sum of the conditional entropy and mutual information [4]. The same equality is not valid for Renyi's definitions of marginal and conditional entropies and mutual information. However, it is possible to obtain lower bounds from Renyi's definitions of these information theoretical quantities by making use of Jensen's inequality. Thus, while we put the main emphasis on the lower and upper bounds, which incorporate the conditional entropy due to the above stated reasons, other versions employing mutual information and joint entropy are also introduced. We also identify the values of the parameters to obtain the tightest lower and upper bounds in among the parametric family of bounds.

In order to evaluate the usefulness of these bounds, we introduce a numerical case study, which demonstrates that the upper bound we have derived is as tight for a very wide range of classifiers as for the one with the optimal confusion matrix. Besides, we present the numerical evaluations of our lower and upper bounds as well as the commonly used union bounds for a QPSK and 16QAM digital communication scheme under AWGN channel in order to demonstrate the performance of these bounds [5].

In a recent work, we have utilized Jensen's inequality on Renyi's definition of conditional entropy, joint entropy, and mutual information to derive lower and upper bounds for the misclassification probability of the classifier under consideration [6]. It turns out that Fano's bound is still special because it is the tightest in the family of lower bounds. On the other hand, the bounds formulated from the conditional entropy reveals the relation between the amount of information transferred through the classifier and its performance. Finally, the examination of the upper bound expression reveals valuable insights to our understanding of how the probabilities in the confusion matrix of a classifier

should be distributed such that its performance is optimized.

The organization of this paper is as follows. We first give a background on the definitions of the information theoretical quantities used. Next, we revisit Fano's bound. In Section IV, we present the lower and upper bounds for probability of error using Renyi's entropy definition. Following that, we provide two numerical examples with analytical solutions, for QPSK, 4PAM, and 16QAM communication schemes. Finally, we summarize the results in the conclusions.

2. BACKGROUND DEFINITIONS

It is possible to express lower and upper bounds for the classification error probability using mutual information, conditional entropy, and joint entropy. Therefore, we first give both the Shannon's and Renyi's definitions for these quantities. Later in the following, we refer to the classes at the input and output of the classifier with the random variables M and W, respectively. 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: For a discrete random variable M, whose probability mass function (pmf) is given by $\{p(m_k)\}_{k=1}^{N_c}$, Shannon's entropy is given by [7]

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

The joint entropy, mutual information, and conditional entropy can be defined based on the entropy as [1,4,7]

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

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 the joint pmf and the conditional pmf of M and W. The following property is satisfied by Shannon's mutual information [1,4,7].

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

Renyi's Definitions: Renyi's entropy for M is [3]

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

where $\alpha>0$ is the entropy order. Accordingly, we obtain the mutual information and conditional entropy as [3]

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

where

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

In bracketing the probability of error from above and below, the entropy order will be useful as it changes the convexity of the function and allows the use of different forms of Jensen's inequality.

3. FANO'S BOUND

Fano determined a lower bound to the probability of error for the classification in discrete-symbol communication systems [1]. The symbols are selected from a discrete symbol set consisting of N_c elements with each symbol m_k having a known prior probability $p(m_k)$. The conditional probability of decision being the j^{th} symbol when k^{th} symbol was sent is $p(w_j|m_k)$. Then, Fano's lower bound for the probability of classification error can be written as

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

A common modification in the pattern recognition literature is to replace the denominator with $\log(N_c)$ to accommodate for 2-class problems. In addition, the identity in (4) is used to obtain a lower bound expressed in terms of the mutual information between the input and the output spaces [8].

4. BOUNDS WITH RENYI'S ENTROPY

In this section, we will provide the expressions for the lower and upper bounds of the probability of error in classification that employ Renyi's definitions of the entropy and mutual information. As Shannon's entropy

is the limit for Renyi's entropy when the order approaches to one, the limit of the lower bound expressions we provide is equal to Fano's bound. The detailed derivation procedure to obtain these bounds can be found in [6]. Here, it suffices to say that the derivation extensively employs the Jensen's inequality for convex and concave functions, and the order of Renyi's entropy allows us to control the convexity of the expressions. In conclusion, with some work, we obtain the following bounds for the error probability in terms of the conditional entropy of the output given the input space of the classifier.

$$\frac{H_\alpha(W|M)-H_S(e)}{\log(N_c-1)} \le P(e) \le \frac{H_\beta(W|M)-H_S(e)}{\min_k H_\beta(W|e,m_k)}, \quad \begin{matrix}\alpha \ge 1 \\ \beta < 1\end{matrix} \tag{9}$$

where

$$H_\beta(W|e,m_k) = \frac{1}{1-\beta}\log\sum_{j\neq k}\left(\frac{p(w_j|m_k)}{p(e|m_k)}\right)^\alpha \tag{10}$$

is the conditional entropy of the output given that we make an error when m_k is the actual class. The numerator of the upper bound expression is always greater than the numerator of the lower bound as a property of Renyi's entropy with varying order. The denominator is an entropy with (N_c-1) terms, hence it is smaller than or equal to log(N_c-1). Thus, the upper bound is always greater than the lower bound. The bounds incorporating the joint entropy and mutual information can be obtained by replacing the conditional entropy with $H_\gamma(W,M)-H_S(M)$, and $H_S(W)-I_\gamma(W;M)$, respectively. In addition, in the upper bound with mutual information, the denominator is evaluated using Shannon's entropy [6].

Notice that, as for a given pmf Renyi's entopy increases as order goes to one from the above, the tightest lower bound I obtained with Fano's bound. Analyzing the effect of entropy order on the upper bound is not that trivial since it appears both in the numerator and in the denominator. For that reason, we investigate a simplified example to observe the behavior of the upper bound under variations in entropy order.

Consider a three-class problem with the following confusion matrix where the ijth entropy denotes the conditional probability $p(w_i|m_j)$.

$$P_{W|M} = \begin{bmatrix} 1-p_e & p_e-\varepsilon & \varepsilon \\ \varepsilon & 1-p_e & p_e-\varepsilon \\ p_e-\varepsilon & \varepsilon & 1-p_e \end{bmatrix} \tag{11}$$

Results presented below in Fig. 1 assume equal class priors, but experiments with different prior assignments showed that the conclusion remains the same; the upper bound becomes tighter when the entropy order approaches to one. One other property of the upper bound, which is desirable, is exhibits the same level of tightness for a broad range of classifiers, whereas, the lower bound tends to be loose for these. The following plots the lower and upper bounds with conditional entropy as a function of ε in the confusion matrix. The overall probability of error is fixed to 0.2.

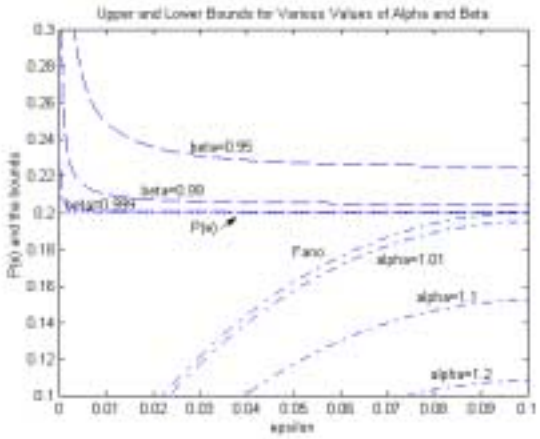

Figure 1. Bounds for different entropy orders

Although we do not present here any results using the mutual information and joint entropy bounds, experiments demonstrated that they produce very close values to those given by the conditional entropy bounds [6]. For Renyi's entropy, since the identity in (4) is not satisfied, we do not get an exact equivalence, whereas in Fano's bound, the three bounds using three different quantities are exactly equal.

Although Renyi's entropy offers a way to bracket the probability of classification error by adjusting the order of entropy, the bounds obtained are not free f problems. In some extreme cases, the lower bounds may become negative (except Fano's bound, which becomes zero), and the upper bounds may blow up if the denominator approaches to zero. Although, in most practical cases this situation will not be encountered, it is possible.

5. NUMERICAL EXAMPLES

As an example, the information theoretic bounds are evaluated for a QPSK modulation scheme over an AWGN channel. The energy per transmitted bit is E_b and the PSD for the additive white Gaussian noise is $N_0/2$. We can compute the exact expression for the confusion matrix in terms of Q-functions, with the assumption of uncorrelated noise in the in-phase and quadrature components, in this problem.

$$P_{W|M}^{QPSK} = \begin{bmatrix} (1-Q_1)^2 & Q_1*(1-Q_1) & Q_1^2 & Q_1*(1-Q_1) \\ Q_1*(1-Q_1) & (1-Q_1)^2 & Q_1*(1-Q_1) & Q_1^2 \\ Q_1*(1-Q_1) & Q_1*(1-Q_1) & (1-Q_1)^2 & Q_1*(1-Q_1) \\ Q_1^2 & Q_1^2 & Q_1*(1-Q_1) & (1-Q_1)^2 \end{bmatrix} \quad (12)$$

where $Q_x = Q\left(x\sqrt{2E_b / N_0}\right)$. The priors for symbols are assumed equal, i.e. $p(m_k) = 1/4,\ k = 1,2,3,4$. Fig. 2 shows the theoretical probability of symbol error and the lower and upper bounds for that. We note that, we could obtain an arbitrarily tight upper bound by simply making the entropy order approach arbitrarily close to one. This process will not introduce any additional computation, but numerical accuracy may become an issue.

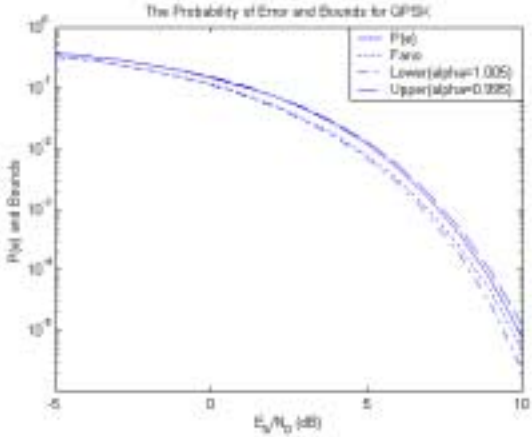

Figure 2. P_s and its bounds for QPSK

As a second example, we evaluate the bounds for a 4PAM modulation scheme over an AWGN channel. This is an example to those problematic situations that may occur. The four classes in 4PAM are located on a line, equally separated with the SNR value indicating the ratio of the distance between means of classes to the variance of the Gaussian distributions centered at these means. As the SNR increases, the denominator approaches to zero because the pmf whose entropy is evaluated approaches to a δ-distribution. In terms of the bit energy and noise power, we can write the confusion matrix as

$$P_{W|M}^{4PAM} = \begin{bmatrix} 1-Q_1 & Q_1 & Q_3 & Q_5 \\ Q_1-Q_3 & 1-2Q_1 & Q_1-Q_3 & Q_3-Q_5 \\ Q_3-Q_5 & Q_1-Q_3 & 1-2Q_1 & Q_1-Q_3 \\ Q_5 & Q_3 & Q_1 & 1-Q_1 \end{bmatrix} \quad (13)$$

The priors for symbols are again assumed equal. Fig. 3 shows the theoretical probability of symbol error and the lower and upper bounds for this case. Note that the upper bound blows for some values of SNR, whereas it is extremely tight for others.

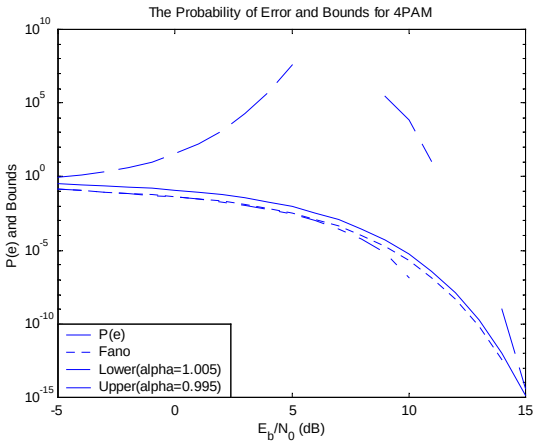

Figure 3. P_s and its bounds for 4PAM

Finally, we evaluate the bounds for a 16QAM scheme where the signal constellation consists of 16 classes uniformly distributed on a square area in two dimensions. Again, with the assumption of uncorrelated white Gaussian noise in the orthogonal directions, we can evaluate the exact confusion matrix, hence the exact probability of classification error and the bounds. The results are summarized below in Fig. 4.

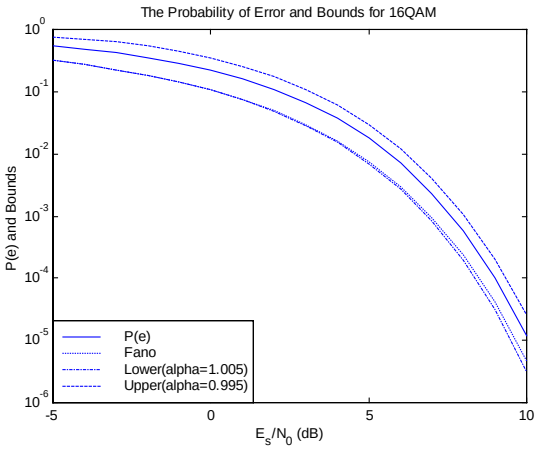

Figure 4. P_s and its bounds for 16QAM

In this section, we presented three simple examples, from digital communications, which can be framed as a pattern recognition problem, and whose analytical solutions can be easily calculated. In QPSK, two of the wrong classes are always located equidistant to the actual class, therefore, the denominator of the upper bound is the entropy of a pmf with at least two terms, hence as SNR increases, the upper bound is still tight. On the other hand, the 4PAM example, designed to illustrate that unstability of the upper bound is possible,

has 3 wrong classes for which the pmf among these may approach a δ-distribution very fast for certain SNR values. In turn, the denominator of the upper bound may become arbitrarily small causing the bound to diverge. The 16QAM example demonstrates that the upper bound may loose accuracy for the same choice of entropy order when the number of classes is high. This is due to the denominator expression, which basically depends on the probabilities of the closest classes. As the number of classes increases, the number of farther neighbors increases. Therefore, the performance degrades.

6. CONCLUSIONS

Fano's bound is a well-known result that provides an insight to how probability of classification error is linked to the information transfer through a classifier. It is derived from Shannon's definition of entropy, which is a special case of Renyi's definition, and it is only a lower bound for a quantity we wish to minimize. Inspired by the work of Fano, we have derived a family of lower and upper bounds for the probability of error starting from Renyi's entropy, where the entropy order identifies if the expression is a lower or an upper bound. Thus, we were able to exploit this property of Renyi's entropy to acquire more information about the probability of error. Interestingly enough, Fano's bound, corresponding to Renyi's entropy of order one, turned out to be the tightest of the lower bounds, and the upper bounds became arbitrarily tight as the entropy order approached one from below.

To demonstrate the performance of the bounds in action, analytical solutions for QPSK, 4PAM, and 16QAM digital communication schemes with AWGN were evaluated. The results indicated that the bounds are useful in bracketing the probability of error in realistic situations.

Although not illustrated here, it is possible to obtain estimates of the bounds by employing various nonparametric estimates for the pmfs that are required in the computation. The simplest of these estimators 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, 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.

Acknowledgments: This work was supported by the NSF grant ECS-9900394.

REFERENCES

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

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

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

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

[5] J.G. Proakis, *Digital Communications*, 3rd ed., NY: McGraw Hill, 1995.

[6] D.Edogmus, J.C. Principe, "Information Theoretical Lower and Upper Bounds for Error Probability of Classifiers," submitted to IEEE.Trans. on Pattern Analysis and Machine Intelligence, Dec. 2000.

[7] C.E. Shannon, "A Mathematical Theory of Communications," *Bell Systems Tech. J.*, vol 27, pp.379-423,623-656, 1948.

[8] K. Torkkola, W.M. Campbell, "Mutual Information in Learning Feature Transformations," *Proceedings of the International Conference on Machine Learning*, Stanford, CA, USA, 2000.