Journal of Machine Learning Research 14 (2013) 1033-1090 Submitted 8/12; Revised 1/13; Published 4/13

Beyond Fano's Inequality: Bounds on the Optimal F-Score, BER, and Cost-Sensitive Risk and Their Implications

Ming-Jie Zhao[*] MZHAO@CS.MANCHESTER.AC.UK
Narayanan Edakunni EDAKUNNI@CS.MANCHESTER.AC.UK
Adam Pocock ADAM.POCOCK@CS.MANCHESTER.AC.UK
Gavin Brown GAVIN.BROWN@CS.MANCHESTER.AC.UK
School of Computer Science
University of Manchester
Manchester M13 9PL, UK

Editor: Charles Elkan

Abstract

Fano's inequality lower bounds the probability of transmission error through a communication channel. Applied to classification problems, it provides a lower bound on the Bayes error rate and motivates the widely used Infomax principle. In modern machine learning, we are often interested in more than just the error rate. In medical diagnosis, different errors incur different cost; hence, the overall risk is cost-sensitive. Two other popular criteria are balanced error rate (BER) and F-score. In this work, we focus on the two-class problem and use a general definition of conditional entropy (including Shannon's as a special case) to derive upper/lower bounds on the optimal F-score, BER and cost-sensitive risk, extending Fano's result. As a consequence, we show that *Infomax is not suitable for optimizing F-score or cost-sensitive risk*, in that it can potentially lead to low F-score and high risk. For cost-sensitive risk, we propose a new conditional entropy formulation which avoids this inconsistency. In addition, we consider the common practice of using a threshold on the posterior probability to tune performance of a classifier. As is widely known, a threshold of 0.5, where the posteriors cross, minimizes error rate—we derive similar optimal thresholds for F-score and BER.

Keywords: balanced error rate, F-score (F_β-measure), cost-sensitive risk, conditional entropy, lower/upper bound

1. Introduction

In the information theory literature, Fano's inequality (Fano, 1961) is a well known result linking the transmission error probability of a noisy communication channel to standard information theoretic quantities such as conditional entropy and mutual information (Shannon, 1948). From a machine learning perspective, we can treat a classification *problem* as a noisy channel; then the inequality provides us with a lower bound on the *Bayes error rate*, that is, the minimum error rate attainable by any classifier, for that problem. A few years later, several upper bounds were also reported, of which the simplest one is as follows: the Bayes error rate of a multi-class problem cannot exceed half of the Shannon conditional entropy (of the class label given the feature vector). This relationship was first obtained by Tebbe and Dwyer III (1968)[1]—see Equation (7) therein, and later by Hellman

[*] The corresponding author

[1] We thank an anonymous reviewer for bringing this to our attention.

and Raviv (1970) using a different argument from Tebbe's. It will be nevertheless referred to as Hellman's bound or Hellman's inequality in the paper, as Tebbe's result is actually stronger than the one we have just stated. See Appendix A (Figure 10) for more detail.

In practice, information measures are often easier than the error probability to evaluate and manipulate (Kailath, 1967). Consequently, both Fano's and Hellman's bounds are useful since they give, from the respective side, some indication of the minimum achievable error rate for a given classification task. More importantly, as shown by Figure 1, the two bounds are both increasing functions of the conditional entropy. Therefore, minimizing the conditional entropy of a system is roughly equivalent to minimizing its probability of error or Bayes error rate. This justifies a general learning principle proposed in the late 1980's, called the *Infomax* or *maximum information preservation* principle:[2]

> **The Infomax Principle** (Linsker, 1989, p. 186): The principle applies to a layer L of cells that provides input to a next layer M. The mapping of the input signal vector L onto an output signal vector M, $f : L \rightarrow M$, is characterized by a conditional probability density function ("pdf") $P(M|L)$. The set S of allowed mappings f is specified. The input pdf $P_L(L)$ is also given. The infomax principle states that a mapping f should be chosen for which the Shannon information rate [*the authors: that is, the mutual information* $I(L;M)$] is a maximum (over all f in the set S).
>
> **The Infomax Principle** (Linsker, 1988, p. 486): An equivalent statement of this principle is: The L-to-M transformation is chosen so as to minimize the amount of information that would be conveyed by the input values L to someone who already knows the output values M. [*The authors: that is, the Shannon conditional entropy* $H(M|L)$ *is the quantity to be minimized.*]

As an optimization principle, Infomax has been employed to devise learning algorithms for a wide range of applications. For instance, Linsker (1989) used it to identify independent input signals fed into a linear system from the system's output. His work was later extended by Bell and Sejnowski (1995) to nonlinear systems, yielding an *independent component analysis* algorithm that is capable of successfully separating unknown mixtures of up to ten speakers.

Another important example is the family of information theoretic filtering methods for feature selection and extraction (Guyon and Elisseeff, 2003; Torkkola, 2003; Duch, 2006). In feature selection (for classification problems), the input signal could be any subset of features, X_θ, where, following Brown et al. (2012), θ is a binary vector with a 1 indicating the corresponding feature is selected and a 0 indicating it is discarded. The output signal is the class label Y. The Infomax principle in this context can thus be stated as:

> *A subset of features* X_θ *should be chosen so that the mutual information* $I(X_\theta;Y)$ *is maximized, or, equivalently, the conditional entropy* $H(Y|X_\theta)$ *is minimized.*

Indeed, this is well justified by Fano's inequality and the monotonically increasing relationship between Fano's bound on error probability and conditional entropy (see Figure 1). As shown by Brown et al. (2012), most of the mutual-information-based feature selection filters in the literature are in fact heuristic approximations of the above Infomax principle, under different independence

[2] While Linsker directly introduced it as a heuristic principle, we highlight the close relationship between Infomax and the error rate minimization principle; and regard the former as a "derived" principle of the latter.

assumptions on features. It seems that people have been taking Infomax for granted: many believe that choosing those features sharing the maximum mutual information with the class label will best facilitate the subsequent classification procedure. In this paper, however, we will show that *this is* **not** *necessarily the case when F-score or cost-sensitive risk is concerned*, via both analytical analysis and numerical examples.

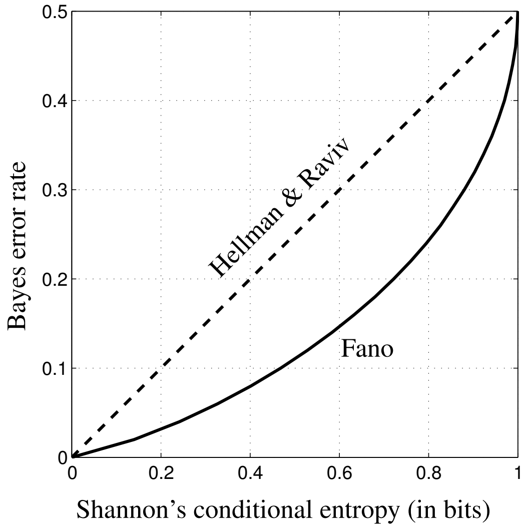


Figure 1: The lower (Fano) and upper (Hellman & Raviv) bounds on the Bayes error rate in terms of Shannon's conditional entropy, for the two-class problem. As both bounds are increasing functions of the conditional entropy, minimizing the conditional entropy would implicitly minimize the Bayes error rate.

Inspired by Fano's result and its widespread utility in machine learning, over the past several decades many researchers have focused on deriving new lower/upper bounds of the Bayes error rate, using various definitions of conditional entropy. We have already mentioned the works of Tebbe and Dwyer III (1968) and Hellman and Raviv (1970). Besides that, Ben-Bassat (1978) derived the lower and upper bounds by means of f-entropies, following the lines originally proposed by Kovalevsky (1968). Using the same method, Golic (1987) discussed the lower and upper bounds based on what he called *concave measures* and *information measures*. Later, Feder and Merhav (1994) re-derived the same upper bound in terms of Shannon's conditional entropy. More recently, Erdogmus and Principe (2004) proposed a family of lower/upper bounds in terms of the Rényi entropy (Rényi, 1961). These bounds are all *increasing* functions of the concerned entropy. This extends our understanding of the Infomax principle, since there are dozens of definitions of entropy in the literature (Taneja, 2001) of which most can be used as the objective function. For instance, Hild II et al. (2006) proposed a mutual information measure based on Rényi's quadratic entropy; and use it to perform feature extraction in the Infomax framework.

All the above bounds are on the Bayes error rate; and to date no *analytical* investigation has been reported on the relationship between conditional entropy and other performance criteria of classifiers such as F-score and balanced error rate.[3] On the other hand, both balanced error rate and F-score are widely employed in practice; and under certain circumstances they are of more interest than the error rate. Indeed, F-score is used widely in the field of *information retrieval* (Manning et al., 2008); whereas balanced error rate is suitable for situations where the distribution of objects is biased among classes. Another situation for which the error rate alone is of little interest is that when different decision errors incur different penalties. In this case, the cost difference between different kinds of errors should be taken into account; the resulting performance measure is called *cost-sensitive risk* in this paper.

As we have discussed above, it is the monotonicity of Fano's and Hellman's bounds that justifies Infomax as an optimization principle for minimizing the error rate. An important question arises: "*is it still a rational principle when the ultimate goal is to minimize the balanced error rate, to maximize the F-score, or to minimize the cost-sensitive risk?*" In this work, we provide an answer to this question, by first deriving the *tight* lower/upper bounds on the *minimum* balanced error rate, the *maximum* F-score and the *minimum* cost-sensitive risk, as functions of conditional entropy; and then examining the monotonicity of these bounds.

1.1 Paper Outline

For binary classification problems Fano's and Hellman's inequalities provide respectively the tight lower and upper bounds on the Bayes error rate (the *minimal* achievable error rate), in terms of the Shannon conditional entropy. Analogously, in this paper we concentrate on the two-class problem and aim to derive the *tight* lower and upper bounds on the *minimum* balanced error rate, on the *maximum* F-score, and on the *minimum* cost-sensitive risk. We however shall do this using a general definition of conditional entropy that includes Shannon's as a special case, in three steps:

1. Derive the analytical expressions of balanced error rate, F-score and cost-sensitive risk for a given classifier applied to a given classification task. These three quantities will be denoted as BER, FSC and CSR, respectively. See Table 1 for a list of the notations consistently employed in this paper.

2. Compute the optimum values of BER, FSC and CSR over *all classifiers*. The resulting quantities are denoted as $\underline{\text{BER}}$, $\overline{\text{FSC}}$ and $\underline{\text{CSR}}$, respectively. Note here that we use the underline (overline) to indicate that a quantity has been minimized (maximized).

3. Derive the tight lower and upper bounds on $\underline{\text{BER}}$, on $\overline{\text{FSC}}$ and on $\underline{\text{CSR}}$, by means of the conditional entropy (of the considered problem) as given by Definition 2 (page 1043).

Notice that while the values of BER, FSC and CSR depend on both the given task and the concerned classifier, their optimum values $\underline{\text{BER}}$, $\overline{\text{FSC}}$ and $\underline{\text{CSR}}$ are classifier-independent. In other words, here we emphasize that the paper is mainly concerned with *problems*, rather than *classifiers* or *algorithms*. More precisely, one main target of this paper is to establish the *universal* relationship between the conditional entropy and one of the three optimum performance measures: $\underline{\text{BER}}$, $\overline{\text{FSC}}$ and $\underline{\text{CSR}}$. Here the word "universal" refers to that our results hold for *any* classification task instead of a particular one. To make this point even clearer, a formal expression unifying the main results of this paper will be highlighted at the end of Section 3, after we have set forth the necessary notions.

[3] That being said, we should point out that *empirical* analysis and comparison of different performance criteria does exist in the literature. See Caruana and Niculescu-Mizil (2004) for example.

Symbol	Meaning
$\mathcal{X}$	space of feature vectors (or objects)
x, y; x, y	feature vector of an object and its *true* class label, the sans-serif font is used when they are treated as random variables
$\hat{y}(\cdot)$; $\hat{y}(x)$, $\hat{y}(\mathsf{x})$	classifiers, or the *predicted* class label for a given object
$\mathcal{X}_0$, $\mathcal{X}_1$	decision region corresponding to class 0 and class 1, see Equation (1)
μ	marginal distribution of the feature vector x, see Equation (2)
$\eta(x)$, $\eta(\mathsf{x})$	posterior probability of class 1 given the object x, see Equation (2); the symbol $\eta(\mathsf{x})$ is used when it is seen as a random variable
(μ, η)	the pair (μ, η) is called a (*classification*) *task*, see page 1038
$\tilde{t}$ for $t \in [0,1]$	shorthand of $1-t$, for example, $\tilde{\eta}(x) = \Pr\{\mathsf{y} = 0 \mid \mathsf{x} = x\}$ (cf. Equation (2))
t^+ for $t \in \mathbb{R}$	shorthand of $\max\{0,t\}$, used in Equation (47) and thereafter
π, $\tilde{\pi}$	prior probability of class 1 and 0, see Equations (4) and (5)
TP, FP, TN, FN	*proportion* of true positive, false positive, true negative, false negative; see also Table 2
c_0, c_1	the cost of false positive and false negative, see Equation (15)
PREC, REC, SPEC	precision, recall and specificity of classifiers, see page 1039
ERR, BER, CSR, FSC	error rate, balanced error rate, F-score and cost-sensitive risk of a given classifier $\hat{y}(\cdot)$; see Equations (11), (13), (15) and (17)
$\underline{\text{ERR}}$, $\underline{\text{BER}}$, $\underline{\text{CSR}}$, $\overline{\text{FSC}}$	the optimum value of ERR, BER, CSR or FSC in a given task
$h_{\text{bin}}(\eta)$, $\eta \in [0,1]$	binary entropy function, see Equation (19) for its definition

Table 1: List of symbols consistently used in the paper and their meaning

The rest of the paper is organized as follows. Section 2 explains some terminologies and notations to be used in this paper; these include the asymptotic expressions of (balanced) error rate (Section 2.2), cost-sensitive risk (Section 2.3) and F-score (Section 2.4). In Section 3, after briefly introducing Fano's and Hellman's inequalities, we present a novel geometric derivation of the two for the case where the conditional entropy is defined by a concave function. We then derive the analytical expression of the minimum cost-sensitive risk, as well as its tight lower and upper bounds in Section 4. The expression of minimum balanced error rate and its lower/upper bounds are given in Section 5. While Section 6 is devoted to computing the maximum F-score, in Section 7 we examine the relationship between the maximum F-score and conditional entropy. In Section 8, we show that *minimizing conditional entropy does not necessarily maximize the F-score or minimize the cost-sensitive risk*. Consequently, standard mutual information is *not* a proper criterion for learning if the final target is to minimize the cost-sensitive risk or maximize the F-score of the subsequent classification process. A proper information measure for cost-sensitive risk, called *cost-sensitive conditional entropy*, is proposed in Section 8.2. Finally, Section 9 concludes the paper with a summary of the main contributions and some possible extensions of this work.

2. Background

In this section we introduce the necessary background and establish the appropriate formal notions to frame the contributions of the paper.

2.1 Classification Tasks and Binary Classifiers

In this paper, we denote by $\mathcal{X}$ the space of feature vectors; and identify each object with its feature vector $x \in \mathcal{X}$. In the binary classification problem, each object x is assumed to belong to one of two classes which are labeled as $y = 0$ (negative) and $y = 1$ (positive), respectively. A classifier can then be described as a binary-valued function, $\hat{y} : \mathcal{X} \to \{0, 1\}$, that maps each object $x \in \mathcal{X}$ to its predicted class label $\hat{y}(x)$.[4] Each such classifier $\hat{y}(\cdot)$ induces naturally a partition of the feature space $\mathcal{X}$ into two *decision regions*, $\mathcal{X}_0$ and $\mathcal{X}_1$, as defined respectively by

$$\mathcal{X}_0 = \{x \in \mathcal{X} \mid \hat{y}(x) = 0\}, \qquad \mathcal{X}_1 = \{x \in \mathcal{X} \mid \hat{y}(x) = 1\}. \tag{1}$$

By definition, it is obvious that $\mathcal{X}_0 \cup \mathcal{X}_1 = \mathcal{X}$ and $\mathcal{X}_0 \cap \mathcal{X}_1 = \varnothing$ for any classifier. Conversely, any pair $(\mathcal{X}_0, \mathcal{X}_1)$ satisfying the two conditions defines a binary classifier $\hat{y}(x)$ that takes the value 0 for $x \in \mathcal{X}_0$ and 1 for $x \in \mathcal{X}_1$. In this paper we shall use the two representations of classifiers interchangeably.

In the traditional probabilistic framework, both the feature vector and the class label are seen as random variables. For the sake of clarity, we shall use the sans-serif font for random variables; so $\mathsf{x} \in \mathcal{X}$ represents the feature vector of an object and $\mathsf{y} \in \{0, 1\}$ the corresponding class label. To specify the joint distribution of x and y, we denote by μ the (marginal) distribution of x and by $\eta(x) \in [0, 1]$ the conditional probability of class 1 given that $\mathsf{x} = x$—the two symbols are borrowed from Devroye et al. (1996, Chapter 2). Formally, for any measurable subset A of $\mathcal{X}$ and any feature vector $x \in \mathcal{X}$, we write

$$\mu(A) := \Pr\{\mathsf{x} \in A\}, \qquad \eta(x) := \Pr\{\mathsf{y} = 1 \mid \mathsf{x} = x\}. \tag{2}$$

Furthermore, for any $t \in [0, 1]$, we define $\tilde{t} := 1 - t$. Then $\tilde{\eta}(x) = \Pr\{\mathsf{y} = 0 \mid \mathsf{x} = x\}$ for any $x \in \mathcal{X}$; and the joint distribution of (x, y) can be written as

$$\Pr\{\mathsf{x} \in A, \mathsf{y} = 1\} = \textstyle\int_A \eta(x) \mathrm{d}\mu, \qquad \Pr\{\mathsf{x} \in A, \mathsf{y} = 0\} = \textstyle\int_A \tilde{\eta}(x) \mathrm{d}\mu. \tag{3}$$

We shall call (μ, η) a *classification task*, or simply a *task*, as it completely describes the problem in the sense that other quantities can all be computed from the pair. For instance, putting $A = \mathcal{X}$ in the two equations of Equation (3), we get the (marginal) probability of the two classes, which will be denoted as π and $\tilde{\pi}$, respectively:

$$\pi := \Pr\{\mathsf{y} = 1\} = \textstyle\int_{\mathcal{X}} \eta(x) \mathrm{d}\mu, \tag{4}$$

$$\tilde{\pi} \ = \Pr\{\mathsf{y} = 0\} = \textstyle\int_{\mathcal{X}} \tilde{\eta}(x) \mathrm{d}\mu. \tag{5}$$

[4] Such classifiers are sometimes called *deterministic* in the literature; the other type being *probabilisitc*, which produce a vector of estimated class probabilities instead of a class label for each given object (Garg and Roth, 2001). A more general variant of the latter is a *discriminant function*, which outputs vectors of continuous scores (often bearing no probabilistic interpretations). See Steinwart (2007) and Tewari and Bartlett (2007) for instance. In this paper we consider only deterministic classifiers.

2.2 Error Rate and Balanced Error Rate of a Classifier

The *error rate* of a classifier is the proportion of misclassified examples in a test data set; and the *balanced error rate* is the arithmetic mean of the misclassification rate in each class. So the value of (balanced) error rate depends not only on the classifier, but also on the test data set selected. To remove finite sample effects, we consider a data set of infinite size and hence the *asymptotic* expressions of error rate and balanced error rate. In particular, for the two-class problem, these can be defined based on the notions of *true positive*, *true negative*, *false positive* and *false negative*.

Let $\hat{y}(\cdot)$ be a classifier applied to the task (μ, η); and $\{(x_i, y_i)\}_{i=1}^{n}$ a set of test examples independently drawn from the distribution (3). According to the value of true class labels y_i and their predictions $\hat{y}_i = \hat{y}(x_i)$, $i = 1, \ldots, n$, the n examples fall into four categories, as shown in Table 2. Denote by TP, FP, FN and TN the *proportion*[5] of examples in the four types, then, as these are also the frequency of the respective events, when $n \to \infty$ they tend to

$$\text{TP} \to \Pr\{\hat{y}(\mathrm{x}) = 1, \mathrm{y} = 1\} = \Pr\{\mathrm{x} \in \mathcal{X}_1, \mathrm{y} = 1\} = \textstyle\int_{\mathcal{X}_1} \eta(x) d\mu, \tag{6}$$

$$\text{FP} \to \Pr\{\hat{y}(\mathrm{x}) = 1, \mathrm{y} = 0\} = \Pr\{\mathrm{x} \in \mathcal{X}_1, \mathrm{y} = 0\} = \textstyle\int_{\mathcal{X}_1} \tilde{\eta}(x) d\mu, \tag{7}$$

$$\text{FN} \to \Pr\{\hat{y}(\mathrm{x}) = 0, \mathrm{y} = 1\} = \Pr\{\mathrm{x} \in \mathcal{X}_0, \mathrm{y} = 1\} = \textstyle\int_{\mathcal{X}_0} \eta(x) d\mu, \tag{8}$$

$$\text{TN} \to \Pr\{\hat{y}(\mathrm{x}) = 0, \mathrm{y} = 0\} = \Pr\{\mathrm{x} \in \mathcal{X}_0, \mathrm{y} = 0\} = \textstyle\int_{\mathcal{X}_0} \tilde{\eta}(x) d\mu, \tag{9}$$

respectively, where the subsets $\mathcal{X}_0$ and $\mathcal{X}_1$ are defined by Equation (1); and the last equality in each equation follows from Equation (3).

	$y = 1$	$y = 0$
$\hat{y} = 1$	true positive (TP); cost: c_{11}	false positive (FP); cost: c_{10}
$\hat{y} = 0$	false negative (FN); cost: c_{01}	true negative (TN); cost: c_{00}

Table 2: Confusion matrix for two possible outcomes and the associated cost matrix

We now define some commonly known performance criteria of binary classifiers for later use. As shown in Nguyen et al. (2009), these can all be written as functions of the above four quantities.

- The *error rate* of a classifier is denoted as ERR in this paper, which is the proportion of misclassified objects, that is, $\text{ERR} := \Pr\{\hat{y}(\mathrm{x}) \neq \mathrm{y}\} = \text{FN} + \text{FP}$.
- The *precision*, PREC, is the proportion of predicted positives ($\hat{y} = 1$) which are actual positive ($\hat{y} = y = 1$), that is, $\text{PREC} := \Pr\{\mathrm{y} = 1 \mid \hat{y}(\mathrm{x}) = 1\} = \text{TP}/(\text{TP} + \text{FP})$.
- The *recall*, denoted REC, is the proportion of actual positives ($y = 1$) which are predicted positive ($y = \hat{y} = 1$), that is, $\text{REC} := \Pr\{\hat{y}(\mathrm{x}) = 1 \mid \mathrm{y} = 1\} = \text{TP}/(\text{TP} + \text{FN})$.
- Finally, the *balanced error rate* is defined as the arithmetic mean of the error rate within the two classes 0 and 1, that is,

$$\begin{aligned} \text{BER} :=& \tfrac{1}{2} \Pr\{\hat{y}(\mathrm{x}) = 1 \mid \mathrm{y} = 0\} + \tfrac{1}{2} \Pr\{\hat{y}(\mathrm{x}) = 0 \mid \mathrm{y} = 1\} \\ =& \tfrac{1}{2} \left\{ \text{FP}/(\text{TN} + \text{FP}) + \text{FN}/(\text{TP} + \text{FN}) \right\} . \end{aligned} \tag{10}$$

[5] Typically, the four quantities refer to the *number* of examples; by "rescaling" them to the *proportion* we are able to discuss the case where the test set contains infinitely many examples, that is, $n \to \infty$.

We now derive the analytical expressions of error rate and balanced error rate. By the asymptotic expressions of FP and FN, Equations (7) and (8), we immediately obtain

$$\text{ERR} = \text{FN} + \text{FP} = \textstyle\int_{\mathcal{X}_0} \eta(x)\mathrm{d}\mu + \int_{\mathcal{X}_1} \tilde{\eta}(x)\mathrm{d}\mu. \tag{11}$$

Furthermore, by Equations (6)–(9) and the facts that $\mathcal{X}_1 \cap \mathcal{X}_0 = \varnothing$ and $\mathcal{X}_1 \cup \mathcal{X}_0 = \mathcal{X}$, we know

$$\text{TP} + \text{FN} = \textstyle\int_{\mathcal{X}} \eta(x)\mathrm{d}\mu = \pi, \qquad \text{TN} + \text{FP} = \int_{\mathcal{X}} \tilde{\eta}(x)\mathrm{d}\mu = \tilde{\pi}. \tag{12}$$

It then follows that

$$\frac{\text{FN}}{\text{TP} + \text{FN}} = \pi^{-1} \cdot \textstyle\int_{\mathcal{X}_0} \eta(x)\mathrm{d}\mu, \qquad \displaystyle\frac{\text{FP}}{\text{TN} + \text{FP}} = \tilde{\pi}^{-1} \cdot \textstyle\int_{\mathcal{X}_1} \tilde{\eta}(x)\mathrm{d}\mu.$$

Therefore, by Equation (10),

$$\text{BER} = \tfrac{1}{2}\left(\pi^{-1} \textstyle\int_{\mathcal{X}_0} \eta(x)\mathrm{d}\mu + \tilde{\pi}^{-1} \int_{\mathcal{X}_1} \tilde{\eta}(x)\mathrm{d}\mu\right). \tag{13}$$

2.3 Cost-Sensitive Risk

According to Table 2, when an object gets misclassified, it can be either a false positive or a false negative. In the criterion of error rate, the two types of errors are treated equally. In some applications, however, the two kinds of errors may have significantly different consequences. In medical testing, for instance, a false negative (i.e., a mistaken diagnosis that a disease is absent, when it is actually present) is typically more serious than a false positive.

One common way to capture the different effects of false positive and false negative is to assign a (different) *cost* to each of the four outcomes in Table 2. Following the convention of Elkan (2001), we denote by $c_{\hat{y}y}$ the cost of classifying an object to the class $\hat{y}$, when it is actually from the class y. For the two-class problem, this gives rise to a 2×2 matrix called the *cost matrix*, which is presented also in Table 2. The expected cost of a given classifier $\hat{y}(\cdot)$ is called the *cost-sensitive risk* and denoted CSR in the paper—here the modifier "cost-sensitive" is borrowed from Elkan (2001). From Table 2, we see that

$$\text{CSR} = c_{11} \cdot \text{TP} + c_{10} \cdot \text{FP} + c_{01} \cdot \text{FN} + c_{00} \cdot \text{TN}. \tag{14}$$

As has been pointed out by Elkan (2001), for a "reasonable" cost matrix, the cost of labeling an example incorrectly should always be greater than the cost of labeling it correctly. In our notation, this is equivalent to requiring that $c_{10} > c_{00}$ and $c_{01} > c_{11}$. In this paper, we further assume that $c_{11} = c_{00} = 0$;[6] and, to simplify our notations, write $c_0 = c_{10}$ and $c_1 = c_{01}$—that is, the first subscript (which is $\hat{y}$) is dropped; so c_y ($y = 0, 1$) is the cost incurred when an object in the class y is misclassified. Using these notations and Equations (7), (8), we can rewrite Equation (14) as

$$\text{CSR} = c_{01} \cdot \text{FN} + c_{10} \cdot \text{FP} = c_1 \cdot \textstyle\int_{\mathcal{X}_0} \eta(x)\mathrm{d}\mu + c_0 \cdot \int_{\mathcal{X}_1} \tilde{\eta}(x)\mathrm{d}\mu. \tag{15}$$

Obviously, the above expression degenerates into Equation (11) when $c_1 = c_0 = 1$. This confirms that the error rate ERR is in fact a special case of the family of cost-sensitive risks.

[6] This condition can actually be weakened to $c_{11} = c_{00}$; in other words, the cost of labeling an object correctly is a constant, regardless of the true class of that object. In this case, we have $\text{CSR} = c_{00} + (c_{10} - c_{00}) \cdot \text{FP} + (c_{01} - c_{00}) \cdot \text{FN}$; so by subtracting the constant c_{00} from CSR, we obtain essentially the same expression as in Equation (15).

The relationship between BER and CSR is little more subtle. At first glance one may think of BER also as a special case of CSR, since we can get Equation (13) by putting

$$c_1 = \tfrac{1}{2}\pi^{-1}, \qquad c_0 = \tfrac{1}{2}\tilde{\pi}^{-1} \tag{16}$$

in Equation (15). But a closer look at the expressions of BER and CSR reveals that they are both functionals of the task (μ, η) and the classifier $(\mathcal{X}_0, \mathcal{X}_1)$ under consideration. Moreover, the value of CSR depends on the two costs c_0 and c_1, whereas BER does not—the two coefficients in Equation (16) are computed from (μ, η). Hence the two quantities should be written, in a more formal way, as $\text{BER}(\mu, \eta, \hat{y})$ and $\text{CSR}(\mu, \eta, \hat{y}; c_0, c_1)$, respectively. It is now clear that in general we cannot treat BER as a special CSR, because there is no uniform setting of c_0 and c_1 such that $\text{BER}(\mu, \eta, \hat{y}) = \text{CSR}(\mu, \eta, \hat{y}; c_0, c_1)$. On the other hand, most machine learning papers are about learning algorithms, with the underlying distribution (μ, η) assumed to be fixed. In that case, or, more generally, as far as only the tasks (μ, η) with fixed priors π and $\tilde{\pi}$ are concerned, BER can be regarded as the cost-sensitive risk as defined by Equations (15) and (16). We will discuss this problem further in Section 5 when we derive bounds on the minimum BER.

2.4 Information Retrieval and F-Score

Manning et al. (2008, p. 1) defines information retrieval as:

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

As an illustrative example, let us consider a typical document retrieval system which accepts a query from the user and returns a subset of "matched" documents retrieved from a huge collection. To evaluate the performance of the system, we assume that each document is known to be either relevant or non-relevant to a particular query. This has formulated the process as a two-class problem in which the positive class consists of those relevant documents; and the negative class corresponds to the set of irrelevant ones. Accordingly, the retrieval system acts as a classifier: the retrieved documents are (seen as) predicted positive. Therefore, we can rewrite, for example,

$$\text{precision as:} \qquad \text{PREC} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|},$$

$$\text{recall as:} \qquad \text{REC} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}.$$

From the above two expressions, we see that precision can be seen as the probability that a retrieved document is truly relevant to the query. Therefore, a high value of precision can be obtained by only returning those documents that are relevant with high confidence. In this way, however, we probably will miss lots of relevant documents. Similarly, the recall can be viewed as the probability that a relevant document is retrieved for the query. So it is trivial to achieve recall of 100% by returning all documents in response to any query. In conclusion, neither precision nor recall alone is enough to serve as a performance measure of information retrieval systems; and we need to take the two into account simultaneously.

A well-known criterion in the community of information retrieval is *F-score*, defined as the harmonic mean of precision and recall,

$$\text{FSC} := \frac{2 \times \text{PREC} \times \text{REC}}{\text{PREC} + \text{REC}} = \frac{2 \times \text{TP}}{(\text{TP} + \text{FN}) + (\text{TP} + \text{FP})} = \frac{2 \cdot \int_{\mathcal{X}_1} \eta(x) \mathrm{d}\mu}{\pi + \mu(\mathcal{X}_1)} . \tag{17}$$

In the above computation, we have used the first equality of Equation (12) and the identity

$$\text{TP} + \text{FP} = \int_{\mathcal{X}_1} \eta(x) \mathrm{d}\mu + \int_{\mathcal{X}_1} \tilde{\eta}(x) \mathrm{d}\mu = \int_{\mathcal{X}_1} 1 \, \mathrm{d}\mu = \mu(\mathcal{X}_1) .$$

F-score is also known as F_1 measure. It is a member of a broader family of performance measures called F_β, where β varies the emphasis on precision versus recall. In this paper we shall focus on the case of $\beta = 1$, and consistently use the term “F-score”.

3. Extending Fano’s and Hellman’s Bounds for the Two-Class Problem

For a given classification problem, the minimum achievable error rate by any classifier is called its *Bayes error rate*, and denoted as $\underline{\text{ERR}}$ in this paper. In the introduction section, we have already mentioned the main results in the literature that are related to our work. They are all about bounding the quantity $\underline{\text{ERR}}$ by means of different conditional entropies, to which a unifying introduction will be given shortly. Although these known bounds hold for the multi-class problem in general, here we shall review only the binary case, leaving a brief introduction to the multi-class case to Appendix A. More precisely, in this section we will present a novel geometric derivation of Fano’s and Hellman’s inequalities for the two-class case and extend it to a broad family of conditional entropies. We do this because the same technique will be used throughout the paper to derive bounds on other optimal performance criteria.

For the two-class problem, it is well known that the classifier which predicts all objects x with the posterior $\eta(x) = \Pr\{\mathsf{y} = 1 \mid \mathsf{x} = x\} > 0.5$ as positive (and others as negative) minimizes the error rate; and the minimum error rate is[7]

$$\underline{\text{ERR}} = \int_{\mathcal{X}} \min\{\eta(x), \tilde{\eta}(x)\} \mathrm{d}\mu . \tag{18}$$

It is also well known that for a binary random variable with the distribution $(\eta, 1-\eta) = (\eta, \tilde{\eta})$, its Shannon entropy is defined by the *binary entropy function*

$$h_{\text{bin}}(\eta) := -\eta \cdot \log \eta - \tilde{\eta} \cdot \log \tilde{\eta} , \qquad \eta \in [0,1] . \tag{19}$$

For binary classification, the value of η depends on the input object $x \in \mathcal{X}$, as given by Equation (2). The expectation of the above function with respect to the object distribution, $\mathsf{x} \sim \mu$, is the *Shannon conditional entropy* (of the class y given the object x):

$$H_{\text{s}}(\mathsf{y}|\mathsf{x}) := \mathbb{E}_{\mathsf{x} \sim \mu}[h_{\text{bin}}(\eta(\mathsf{x}))] = \int_{\mathcal{X}} h_{\text{bin}}(\eta(x)) \mathrm{d}\mu , \tag{20}$$

where the subscript s stands for “Shannon”.

Fano’s inequality connects the Shannon conditional entropy, $H_{\text{s}}(\mathsf{y}|\mathsf{x})$, to the Bayes error rate, $\underline{\text{ERR}}$, by $h_{\text{bin}}(\underline{\text{ERR}}) \geqslant H_{\text{s}}(\mathsf{y}|\mathsf{x})$. As $\underline{\text{ERR}} \leqslant 0.5$ and the function $h_{\text{bin}}(\eta)$ is monotonically increasing

[7] These facts will become clear after we have derived the expression of the minimum cost-sensitive risk in Section 4.

for $0 \leqslant \eta \leqslant 0.5$, this actually provides a lower bound on $\underline{\mathrm{ERR}}$ in terms of $H_{\mathrm{s}}(\mathsf{y}|\mathsf{x})$. The upper bound is defined by Hellman's inequality, which can be written in our notation as $\underline{\mathrm{ERR}} \leqslant \frac{1}{2}H_{\mathrm{s}}(\mathsf{y}|\mathsf{x})$. The two bounds had been graphically shown in Figure 1.

In the literature, the two inequalities were proven using different methods; see, for example, Cover and Thomas (2006, Section 2.10) and Hellman and Raviv (1970). Here we propose a novel geometric proof that they can be obtained simultaneously, based upon an "obvious" fact which we state as a theorem (because of its fundamental importance in the paper).

Theorem 1 *The expectation of a random vector (assume it exists) in the Euclidean space $\mathbb{R}^m$ lies in the convex hull of the range of that random vector.*

This proposition, probably well known and intuitively clear—since the expectation of a random vector is essentially the convex combination of the vectors in its range, is in fact nontrivial. To the best of our knowledge (and to our surprise), there is no proof to Theorem 1 in the literature (we thought it should be in some textbooks on probability theory, but we cannot find one). We hence provide one of ourselves in Appendix B.4.

Theorem 1 gives rise to a *general geometric strategy* for deriving/proving inequalities like Fano's, as outlined in Scheme 3, where the derivation of the lower and upper bounds on $H_{\mathrm{s}}(\mathsf{y}|\mathsf{x})$ has been used as a demonstration. One should have no difficulty to see that this geometric method can be extended, straightforwardly, to the family of concave and symmetric functions $h(\eta)$, instead of the particular function $h_{\mathrm{bin}}(\eta)$. In fact, we even can go one step further, by dropping the requirement that $h(\eta)$ be symmetric. We hence introduce the following general definition of conditional entropy.

Definition 2 *Let $h(\eta)$ with $\eta \in [0,1]$ be a concave function satisfying*[8] $h(0) = h(1) = 0$. *The conditional entropy of a given classification task (μ, η) is defined as*

$$H(\mathsf{y}|\mathsf{x}) := \mathbb{E}_{\mathsf{x}\sim\mu}[h(\eta(\mathsf{x}))] = \int_{\mathcal{X}} h(\eta(x))\mathrm{d}\mu. \tag{21}$$

This definition of conditional entropy is general enough to include most of entropies in the literature. For example, the Shannon entropy is obtained by setting $h(\eta) = h_{\mathrm{bin}}(\eta)$; and letting $h(\eta) = -\log(\eta^2 + \tilde{\eta}^2)$, we get Rényi's quadratic entropy (Principe and Xu, 1999). Another example is *weighted entropy* (Guiasu, 1971), which for the binary case is defined by the function $h(\eta) = -w_1\eta \cdot \log\eta - w_0\tilde{\eta} \cdot \log\tilde{\eta}$, where $w_0, w_1 > 0$ are two weights. Note that this function is asymmetric when $w_0 \neq w_1$.

Scheme 3 A general geometric strategy for deriving tight lower and upper bounds on one performance/information measure in terms of another measure

Assume we want to derive the tight lower/upper bounds on $H_{\mathrm{s}}(\mathsf{y}|\mathsf{x})$ in terms of $\underline{\mathrm{ERR}}$:

1. The first step is to find a random vector with expectation $[\underline{\mathrm{ERR}}, H_{\mathrm{s}}(\mathsf{y}|\mathsf{x})]$ (the vector comprising the concerned quantities). In fact, by Equations (20) and (18) we easily see that $[\underline{\mathrm{ERR}}, H_{\mathrm{s}}(\mathsf{y}|\mathsf{x})] = \mathbb{E}_{\mathsf{x}\sim\mu}[e(\eta(\mathsf{x})), h_{\mathrm{bin}}(\eta(\mathsf{x}))]$, where the function $e(\cdot)$ is defined by

$$e(\eta) := \min\{\eta, \tilde{\eta}\}, \qquad \eta \in [0,1]. \tag{22}$$

So the random vector $[e(\eta(\mathsf{x})), h_{\mathrm{bin}}(\eta(\mathsf{x}))]$ is what we want.

[8] As subtracting a linear function (of η) from a concave function still gives a concave function, imposing the condition $h(0) = h(1) = 0$ on the definition will result in no loss of generality.

2. Next, we need to find the range of $[e(\eta(\mathsf{x})), h_{\text{bin}}(\eta(\mathsf{x}))]$, the random vector obtained in Step 1. Apparently, this is the curve $\ell := \{[e(\eta), h_{\text{bin}}(\eta)] \mid \eta \in [0,1]\}$.[9] But $h_{\text{bin}}(\eta)$ is a symmetric function, that is, $h_{\text{bin}}(\eta) = h_{\text{bin}}(\tilde{\eta})$, by the definition of $e(\eta)$ we know the curve ℓ is in fact the left half of $h_{\text{bin}}(\eta)$, as depicted in Figure 2-a.

3. We then construct the convex hull of the curve ℓ, which for this example is the bow shape OABCO bounded by the curve OCB (i.e., $h = h(e)$) from above and by the line segment OAB (i.e., $h = 2e$) from below—see Appendix B.2 and B.3 for a rigorous discussion on the convex hull of a given curve or subset.

4. Now Theorem 1 shows the point $[\underline{\text{ERR}}, H_{\text{s}}(\mathsf{y}|\mathsf{x})]$ is in the area OABCO. We can thus *directly "read"*, for any given value of $\underline{\text{ERR}}$, the lower and upper bounds of $H_{\text{s}}(\mathsf{y}|\mathsf{x})$ from the convex hull of ℓ, in an obvious way and *simultaneously*. The correctness of the bounds so obtained is *guaranteed* by Theorem 1. For this example, the two bounds are $2e|_{e=\underline{\text{ERR}}} \leqslant H_{\text{s}}(\mathsf{y}|\mathsf{x}) \leqslant h(e)|_{e=\underline{\text{ERR}}}$, that is, $2\underline{\text{ERR}} \leqslant H_{\text{s}}(\mathsf{y}|\mathsf{x}) \leqslant h(\underline{\text{ERR}})$, which are exactly Fano's and Hellman's results.

5. Last but not least, it is easy to show that each point in the convex hull of ℓ can be attained by some classification task (see the proof to Theorem 5). Thus, the bounds obtained as above are tight.

To illustrate the generality of the proposed geometric scheme, we apply it to a general concave function $h(\eta)$ which might be *asymmetric*. For this, only the second and third steps in Scheme 3 need to be adapted slightly, as follows.

2. For an asymmetric function $h(\eta)$, the curve $\ell = \{[e(\eta), h(\eta)] \mid \eta \in [0,1]\}$ consists of two parts which can be expressed as $h = h(e)$—as $e(\eta) = \eta$ for $\eta \in [0, 0.5]$, and $h = h(1-e)$—as $e(\eta) = \tilde{\eta} = 1 - \eta$ for $\eta \in [0.5, 1]$. Graphically, this means that ℓ is the left half of the curve $h = h(\eta)$ plus its right half flipped along the vertical line $\eta = 0.5$, as is shown in Figure 2-b.

3. The convex hull of ℓ, denoted $\operatorname{co}\ell$, can then be expressed as[10] (recall that $\tilde{e} = 1 - e$)

$$\operatorname{co}\ell = \{(e,h) \mid e \in [0,0.5],\ [\min\{h(e), h(\tilde{e})\}]_{\smile} \leqslant h \leqslant [\max\{h(e), h(\tilde{e})\}]^{\frown}\}, \tag{23}$$

where, for any real-valued function $f(\cdot)$ defined on a convex set, $f_{\smile}$ denotes the *convex hull* of f, that is, the greatest convex function with the same domain as f that does not exceed f; and $f^{\frown}$ is the *concave hull* of f, the smallest concave function that is larger than or equal to f at each point in the domain of f.

We are now ready to "read" the lower and upper bounds of $H(\mathsf{y}|\mathsf{x})$ from the set $\operatorname{co}\ell$, as Theorem 1 has already told us that $[\underline{\text{ERR}}, H(\mathsf{y}|\mathsf{x})] \in \operatorname{co}\ell$. But before that, we would first simplify the two bounds $[\ldots]_{\smile}$ and $[\ldots]^{\frown}$ in Equation (23), to get a cleaner result. The function $h(\cdot)$ is concave, so is

[9] Strictly speaking, this should be $\ell := \{[e(\eta(x)), h_{\text{bin}}(\eta(x))] \mid x \in \mathcal{X}\}$. But as we are investigating the universal relationship, the "wildest" case where the range of $\eta(x)$ is $[0,1]$ should be considered.

[10] See Appendix B.3 for a proof for this.

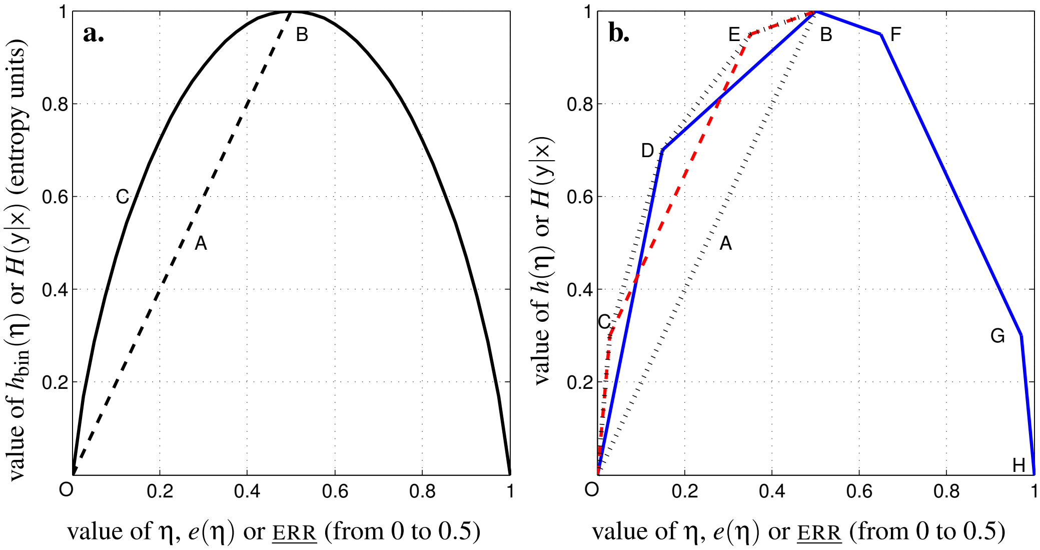


Figure 2: **a.** The graph of the function $h_{\text{bin}}(\eta)$ (solid line), with η as the x-axis and h_{bin} as the y-axis. As this is a symmetric function, its left part BCO represents the curve $\ell = \{[e(\eta), h_{\text{bin}}(\eta)] \mid \eta \in [0,1]\}$—now the x-axis stands for $e(\eta)$. The convex hull of ℓ is hence the region bounded by the solid curve (left half) and the dashed line OAB—now we have $\underline{\text{ERR}}$ for the x-axis and $H(\text{y}|\text{x})$ for the y-axis. By flipping this bow shape along the diagonal line through the points $[0,0]$ and $[1,1]$, we get exactly Figure 1, that is, Fano's and Hellman's bounds.
b. The graph of an asymmetric function $h(\eta)$—the broken line ODBFGH, and the curve $\ell = \{[e(\eta), h(\eta)] \mid \eta \in [0,1]\}$, which consists of the broken lines ODB and OCEB (obtained from HGFB through a right-to-left flipping). The convex hull of ℓ is then the polygon OBEDCO (dotted line). As in the symmetric case, here the upper/lower bounds on $\underline{\text{ERR}}$ can be obtained by flipping this polygon along the diagonal line.

$\min\{h(e), h(1-e)\}$ as a function of $e \in [0, 0.5]$, as it is the minimum of two concave functions. But the convex hull of a concave function is an affine function through its two endpoints. Therefore,

$$[\min\{h(e), h(1-e)\}]_{\smile} = 2 \cdot h(0.5) \cdot e\,.$$

Here we have used the assumption that $h(0) = h(1) = 0$. Moreover, if $h(\cdot)$ is symmetric, that is, $h(e) = h(\tilde{e})$, then $\max\{h(e), h(\tilde{e})\} = h(e)$ is a concave function; and its concave hull is itself: $[\max\{h(e), h(\tilde{e})\}]^{\frown} = h(e)$.

Putting the above discussion together, we obtain the following theorem that extends Fano's and Hellman's results to that using an arbitrary concave function $h : [0,1] \to \mathbb{R}$ in the definition of the conditional entropy.

Theorem 4 (extension of Fano's and Hellman's inequalities) *Let $h : [0,1] \to \mathbb{R}$ be a concave function with $h(0) = h(1) = 0$. Then for any classification task (μ, η) we have*

$$2 \cdot h(0.5) \cdot \underline{\text{ERR}} \leqslant H(\text{y}|\text{x}) \leqslant [\max\{h(\underline{\text{ERR}}), h(1 - \underline{\text{ERR}})\}]^{\frown}\,. \tag{24}$$

In particular, for symmetric functions $h(\cdot)$ the above inequality can be simplified to

$$2 \cdot h(0.5) \cdot \underline{\text{ERR}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant h(\underline{\text{ERR}})\,.$$

In next section we shall extend the above theorem further to a relationship between the conditional entropy $H(\mathsf{y}|\mathsf{x})$ and the minimum cost-sensitive risk $\underline{\text{CSR}}$—of which $\underline{\text{ERR}}$ is a special case.

Furthermore, from Theorem 1 and the definition of the convex hull of a set, we can easily see the two bounds of $H(\mathsf{y}|\mathsf{x})$ given by Equation (24) are tight, in the sense that for *any concave function* $h(\cdot)$ and *any given value of* $\underline{\text{ERR}}$, both bounds are reachable by some task (μ,η). In fact, we have an even stronger result, for which a short proof is presented as the "template" for other similar tightness proofs in the paper.

Theorem 5 *For any concave function $h : [0,1] \to \mathbb{R}$ with $h(0) = h(1) = 0$, and any point $[e_0,h_0]$ inside the convex set* $\operatorname{co}\ell$ *as given by Equation* (23)*, there exists a task (μ,η) for which it holds that* $\underline{\text{ERR}} = e_0$ *and* $H(\mathsf{y}|\mathsf{x}) = h_0$.

Proof Since the point $[e_0,h_0]$ lies in the convex hull of $\ell = \{[e(\eta),h(\eta)] \mid \eta \in [0,1]\}$, there are n points on the curve ℓ, say $\{[e(\eta_i),h(\eta_i)]\}_{i=1,\ldots,n}$, such that $[e_0,h_0]$ is the convex combination of the n points with the coefficients $\{\beta_i\}_{i=1,\ldots,n}$, that is,

$$e_0 = \textstyle\sum_{i=1}^{n} \beta_i e(\eta_i)\,, \qquad h_0 = \sum_{i=1}^{n} \beta_i h(\eta_i)\,,$$

where $\beta_i \geqslant 0$ satisfy $\sum_{i=1}^{n} \beta_i = 1$. We can thus construct a classification task in which the feature space $\mathcal{X}$ consists exactly of n points, $\{x^{(1)},\ldots,x^{(n)}\}$, with the probability mass $\mu(x^{(i)})$ and the posterior $\eta(x^{(i)})$ given by

$$\mu(x^{(i)}) = \Pr\{\mathsf{x} = x^{(i)}\} = \beta_i\,, \qquad \eta(x^{(i)}) = \Pr\{\mathsf{y} = 1 \mid \mathsf{x} = x^{(i)}\} = \eta_i\,.$$

Clearly, for this task (μ,η) we have $\underline{\text{ERR}} = e_0$ and $H(\mathsf{y}|\mathsf{x}) = h_0$. ■

Corollary 6 *In Theorem 4 (Equation* (24)*), the two bounds on $H(\mathsf{y}|\mathsf{x})$ are tight. That is, given any concave function $h : [0,1] \to \mathbb{R}$ with $h(0) = h(1) = 1$ and any value of* $\underline{\text{ERR}} \in [0,0.5]$*, there are two (different) tasks for which the two inequalities in Equation* (24) *become equalities, respectively.*

Proof Apply Theorem 5 to the point $[e_0,h_0] = [\underline{\text{ERR}},\ 2 \cdot h(0.5) \cdot \underline{\text{ERR}}]$ and to the point $[e_0,h_0] = [\underline{\text{ERR}},\ [\max\{h(\underline{\text{ERR}}),h(1-\underline{\text{ERR}})\}]^{\frown}]$. ■

To summarize, in this fundamental section we proposed a general geometric approach to deriving/proving inequalities that links the conditional entropy of classification tasks with an optimal performance measure, for example, the Bayes error rate $\underline{\text{ERR}}$. By Theorem 1, Theorem 5 and Corollary 6, the inequalities obtained in this way are guaranteed to be *correct* and *sharp*. They are also *universal* in that Theorem 4 holds for any task (μ,η).

Following the discussion at the end of Section 2.3, here we would emphasize again that the two quantities in the inequality, $\underline{\text{ERR}}$ and $H(\mathsf{x}|\mathsf{y})$, are actually functionals of tasks; and should be written respectively as $\underline{\text{ERR}}(\mu,\eta)$ and $H_h(\mu,\eta)$ in a more formal way. Here the subscript $_h$ is used to stress the role of the function $h(\cdot)$ in the definition of conditional entropy. In accordance, Equation (24) should be written as

$$2 \cdot h(0.5) \cdot \underline{\text{ERR}}(\mu,\eta) \leqslant H_h(\mu,\eta) \leqslant [\max\{h(\underline{\text{ERR}}(\mu,\eta)), h(1-\underline{\text{ERR}}(\mu,\eta))\}]^{\frown};$$

and it holds for *any* concave function $h:[0,1]\to\mathbb{R}$ satisfying $h(0)=h(1)=0$ and *any* classification task (μ,η), as has been asserted by Theorem 4.

In the next four sections we will derive the similar inequalities for the quantities $\underline{\text{CSR}}$, $\underline{\text{BER}}$ and $\overline{\text{FSC}}$, which have the following uniform form:

$$f(\text{XX}(\mu,\eta)) \leqslant H_h(\mu,\eta) \leqslant g(\text{XX}(\mu,\eta)), \qquad \text{XX stands for } \underline{\text{CSR}}, \underline{\text{BER}} \text{ or } \overline{\text{FSC}}, \tag{25}$$

where $f(\cdot)$ is a proper *convex* function and $g(\cdot)$ a proper *concave* function. Like Equation (24), for $\underline{\text{CSR}}$ and $\overline{\text{FSC}}$ the corresponding inequality holds for *any* task (μ,η). The quantity $\underline{\text{BER}}$ is special, for which the two "bounding" functions $f(\cdot)$ and $g(\cdot)$ involve an extra parameter: the positive prior π, which presents also in the expression of $\underline{\text{BER}}$—see Equation (32) in page 1052. Consequently, the result holds only for the tasks (μ,η) with *fixed* class priors, namely, $\Pr\{\mathsf{y}=1\}=\pi$ and $\Pr\{\mathsf{y}=0\}=\tilde{\pi}$. But when π is also seen as a functional of (μ,η), then the inequality (25)—which now links $\underline{\text{BER}}$, H_h and π, becomes universal.

4. Bounds on the Minimum Cost-Sensitive Risk

We now study the relationship between the conditional entropy $H(\mathsf{y}|\mathsf{x})$ and the minimum cost-sensitive risk $\underline{\text{CSR}}$, using the same geometric strategy as given in the preceeding section. To this end, we need first to derive the expression of $\underline{\text{CSR}}$.

We have already derived in Section 2.3 the analytical expression of the cost-sensitive risk for a given classifier, which, as shown in Equation (15), is the sum of two integrals of the functions $c_1\eta(x)$ and $c_0\tilde{\eta}(x)$ over the disjoint subsets $\mathcal{X}_0$ and $\mathcal{X}_1$ of the space $\mathcal{X}$, respectively. To obtain its minimum (over all possible classifiers), we use that both $c_1\eta(x)$ and $c_0\tilde{\eta}(x)$ are larger than or equal to the minimum of the two. It thus follows that

$$\begin{aligned} \text{CSR} &= \textstyle\int_{\mathcal{X}_0} c_1\eta(x)\mathrm{d}\mu + \int_{\mathcal{X}_1} c_0\tilde{\eta}(x)\mathrm{d}\mu \\ &\geqslant \textstyle\int_{\mathcal{X}_0} \min\{c_1\eta(x), c_0\tilde{\eta}(x)\}\mathrm{d}\mu + \int_{\mathcal{X}_1} \min\{c_1\eta(x), c_0\tilde{\eta}(x)\}\mathrm{d}\mu. \end{aligned}$$

But as $\mathcal{X}_0\cap\mathcal{X}_1=\varnothing$ and $\mathcal{X}_0\cup\mathcal{X}_1=\mathcal{X}$, the above two integrals $\int_{\mathcal{X}_i}\min\{\ldots\}\mathrm{d}\mu$, $i=0,1$, can be combined into one (over the whole space $\mathcal{X}$), yielding $\text{CSR}\geqslant\int_{\mathcal{X}}\min\{c_1\eta(x), c_0\tilde{\eta}(x)\}\mathrm{d}\mu$. Moreover, this inequality becomes equality when (and only when) the condition:

$$\begin{aligned} c_1\eta(x) &= \min\{c_1\eta(x), c_0\tilde{\eta}(x)\} \quad \text{on } \mathcal{X}_0\text{; and} \\ c_0\tilde{\eta}(x) &= \min\{c_1\eta(x), c_0\tilde{\eta}(x)\} \quad \text{on } \mathcal{X}_1 \end{aligned}$$

is fulfilled. This is equivalent to requiring that $c_1\eta(x)\leqslant c_0\tilde{\eta}(x)$, that is, $\eta(x)\leqslant\frac{c_0}{c_0+c_1}$ for (and only for) all $x\in\mathcal{X}_0$. Therefore, the minimum cost-sensitive risk is

$$\underline{\text{CSR}} = \textstyle\int_{\mathcal{X}} \min\{c_1\eta(x), c_0\tilde{\eta}(x)\}\mathrm{d}\mu; \tag{26}$$

and this minimum is achieved by the classifier $\hat{y}(x)=[\![\eta(x)>\frac{c_0}{c_0+c_1}]\!]$, where $[\![\cdot]\!]$ denotes the indicator function which takes value 1 if the bracketed statement is true and 0 otherwise. This result is well known in Bayesian decision theory; see, for example, Duda et al. (2001, page 26).

Note that the error rate can be seen as a special cost-sensitive risk with $c_0 = c_1 = 1$, so its minimum can be obtained from Equation (26) by putting $c_0 = c_1 = 1$. This gives us exactly the expression Equation (32), and the corresponding optimal classifier is $\hat{y}(x) = [\![\eta(x) > \frac{c_0}{c_0+c_1}]\!] = [\![\eta(x) > 0.5]\!]$, which have been stated in Section 3 as well established in the literature.

We now derive, in terms of $\underline{\text{CSR}}$, the lower and upper bounds on the conditional entropy $H(\mathsf{y}|\mathsf{x})$ as given by Definition 2. Following the geometric lines in Scheme 3, we define the function $e(\eta)$ as (again, this is reduced to Equation (22) for $c_0 = c_1 = 1$)

$$e(\eta) := \min\{c_1\eta, c_0\tilde{\eta}\}, \qquad \eta \in [0,1]. \tag{27}$$

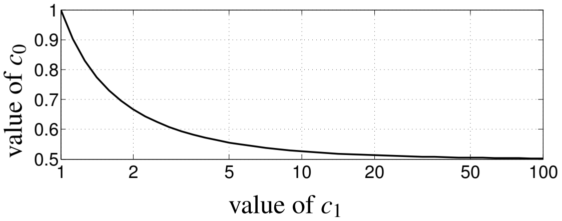
Then, Equations (21) and (26) can rewritten as the mathematical expectations of $h(\eta(\mathsf{x}))$ and $e(\eta(\mathsf{x}))$, respectively:

$$[\underline{\text{CSR}}, H(\mathsf{y}|\mathsf{x})] = \mathbb{E}_{\mathsf{x}\sim\mu}[e(\eta(\mathsf{x})), h(\eta(\mathsf{x}))]. \tag{28}$$

We thus have accomplished the first step in Scheme 3. By Theorem 1, in the e-h plane the point $[\underline{\text{CSR}}, H(\mathsf{y}|\mathsf{x})]$ lies in the convex hull of the curve $\ell := \{[e(\eta), h(\eta)] \mid \eta \in [0,1]\}$. The problem then amounts to finding the convex hull of ℓ which we shall discuss shortly.

By the definition of $e(\eta)$, Equation (27), one easily sees that $e = c_1\eta$ when $\eta \leqslant \frac{c_0}{c_0+c_1}$ and that $e = c_0\tilde{\eta} = c_0 - c_0\eta$ when $\eta \geqslant \frac{c_0}{c_0+c_1}$. It then follows that $0 \leqslant e(\eta) \leqslant \frac{c_0c_1}{c_0+c_1}$, with the minimum value 0 attained at $\eta = 0$ or $\eta = 1$; and the maximum $\frac{c_0c_1}{c_0+c_1}$ obtained at $\eta = \frac{c_0}{c_0+c_1}$. At this point, we find it most convenient to normalize the two costs c_0 and c_1 (by multiplying them by a common factor) to such that $\frac{c_0c_1}{c_0+c_1} = 0.5$, that is, $c_0^{-1} + c_1^{-1} = 2$. Then the range of e is always $[0, 0.5]$.

To simplify the derivation procedure we further assume, without loss of generality, that $c_1 \geqslant c_0$. In Section 5, this assumption will be used to obtain bounds on $\underline{\text{BER}}$ from that on $\underline{\text{CSR}}$, see the proof to Corollary 8. The inequality $c_1 \geqslant c_0$ is equivalent to $c_1^{-1} \leqslant c_0^{-1}$, which together with $c_0^{-1} + c_1^{-1} = 2$ implies that $1 \leqslant c_0^{-1} < 2$ and $0 < c_1^{-1} \leqslant 1$. We thus get $c_0 \in (0.5, 1]$, $c_1 \in [1, \infty)$ and $\frac{c_0}{c_0+c_1} = \frac{1}{2c_1} \leqslant \frac{1}{2}$. Furthermore, from the equality $c_0^{-1} + c_1^{-1} = 2$ we know $c_0 = \frac{c_1}{2c_1-1}$. So the cost matrix is now characterized by a single parameter $c_1 \in [1, \infty)$, as shown in Figure 3.



	$y=1$	$y=0$
$\hat{y}=1$	0	$c_0 = \frac{c_1}{2c_1-1}$
$\hat{y}=0$	c_1	0

Figure 3: The relationship between the values of c_0 and c_1 (left) and the cost-matrix as characterized by the cost c_1 (right).

We now study the curve $\ell = \{[e(\eta), h(\eta)] \mid \eta \in [0,1]\}$ (Step 2 in Scheme 3). Based on the above assumptions, we see that when η changes from 0 to $\frac{1}{2c_1}$, $e = c_1\eta$ changes from 0 to 0.5; and when η changes from $\frac{1}{2c_1}$ *further* to 1, $e = c_0 - c_0\eta$ changes from 0.5 *back* to 0, both in a linear manner. Therefore, the curve $\ell = \{[e(\eta), h(\eta)] \mid \eta \in [0,1]\}$ consists of two parts, namely $h = h(c_1^{-1}e)$, $e \in [0, 0.5]$ (corresponding to $\eta \in [0, \frac{1}{2c_1}]$) and $h = h(1 - c_0^{-1}e)$, $e \in [0, 0.5]$ (corresponding to $\eta \in [\frac{1}{2c_1}, 1]$). The first part $h = h(c_1^{-1}e)$ is obtained from the graph of $h(\eta)$, $\eta \in [0, \frac{1}{2c_1}]$ by linearly lengthening it from the interval $[0, \frac{1}{2c_1}]$ to that on the interval $[0, 0.5]$. The second part

$h = h(1 - c_0^{-1}e)$ is obtained from the graph of $h(\eta)$, $\eta \in [\frac{1}{2c_1}, 1]$ by first linearly shrinking it from the interval $[\frac{1}{2c_1}, 1]$ to over the interval $[\frac{1}{2}, 1]$, and then flipping the resulting curve along the vertical line at $\eta = 0.5$.

The above dynamical procedure is demonstrated in Figure 4-a for the settings $c_1 = 2.5$, $c_0 = 0.625$ and $h(\eta) = -\eta \cdot \log \eta - (1-\eta)\log(1-\eta)$ (Shannon). In Figure 4-a, we start with the graph of $h = h(\eta)$, the curve OABF. This curve is divided into two parts by the point A whose coordinate is $(\frac{1}{2c_1}, h(\frac{1}{2c_1}))$. To obtain the curve ℓ, we first move horizontally A to the point C which has the coordinate of $(0.5, h(\frac{1}{2c_1}))$. The other points on the curve OABF are moved linearly, with the two endpoints O and F being fixed. This gives us the curve OCDF, whose left part OC represents the function $h = h(c_1^{-1}e)$, $e \in [0, 0.5]$; and its right half CDF is described by the function $h = h(1 - c_0^{-1}(1-e))$, $e \in [0.5, 1]$. Next, we flip the right part CDF along the vertical line at $\eta = 0.5$, yielding the curve OHEC which is the graph of $h = h(1 - c_0^{-1}e)$, $e \in [0, 0.5]$. The curve ℓ is then the union of the curve OHEC and the curve OC, that is, the closed curve OHECO.

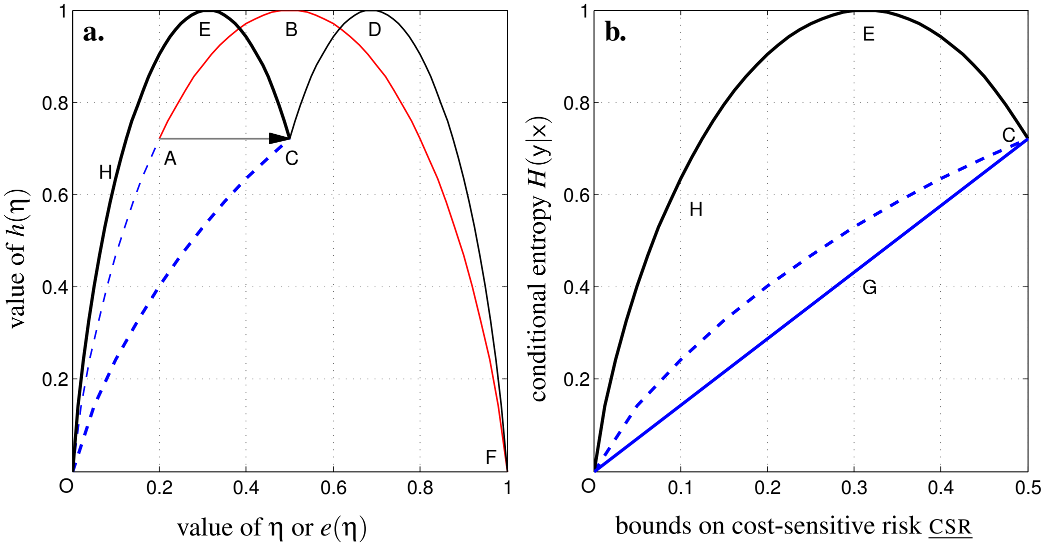


Figure 4: **a.** The procedure for getting the curve $\ell = \{[e(\eta), h(\eta)] \mid \eta \in [0,1]\}$ (OHECO) from the graph of $h(\eta)$ (the curve OABF), for the settings of $c_0 = 0.625$, $c_1 = 2.5$ and $h(\eta) = -\eta \log \eta - \tilde{\eta} \log \tilde{\eta}$. See the text for more explanation.
b. The tight lower (the line OGC) and upper (the curve OHEG) bounds on $H(y|x)$ as functions of $\underline{\text{CSR}}$. Stated in the other way, if the conditional entropy $H(y|x)$ is known, then the upper bound of $\underline{\text{CSR}}$ is determined by the curve OGCE; and its lower bound is given by the curve OHE.

The next step is to find the convex hull of ℓ. By Definition 2, $h(\eta)$ is a concave function. So both $h = h(c_1^{-1}e)$ and $h = h(1 - c_0^{-1}e)$ are concave functions (of e). Moreover, for any symmetric function $h(\eta)$, from Figure 4-a we see that the curve OHEC is above the curve OC. Mathematically, this can be expressed as $h(c_1^{-1}e) \leqslant h(1 - c_0^{-1}e)$, which is true for all *symmetric* concave functions

$h(\eta)$.[11] Therefore, the convex hull of the curve OHECO can be obtained by simply connecting the points O and C with a straight line. This is plotted in Figure 4-b, where the *region* OHECGO is the convex hull of the curve OHECO; it also represents the reachable region of the point $[\underline{\text{CSR}}, H(\mathsf{y}|\mathsf{x})]$.

It is now clear that the value of $H(\mathsf{y}|\mathsf{x})$ is lower bounded by the straight line OC, which is the graph of the function $h = 2h(\frac{1}{2c_1})e$, $e \in [0, 0.5]$, and upper bounded by the curve OHEC, which is described by the function $h = h(1 - c_0^{-1}e) = h(c_0^{-1}e)$, $e \in [0, 0.5]$—since $h(\cdot)$ has been assumed to be symmetric here. We thus obtain

$$2 \cdot h(\tfrac{1}{2c_1}) \cdot \underline{\text{CSR}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant h(c_0^{-1} \cdot \underline{\text{CSR}})\,. \tag{29}$$

By a similar discussion to that above Theorem 4, we can extend Equation (29) to asymmetric functions $h(\cdot)$. In this case it is not necessarily that $h(c_1^{-1}e) \leqslant h(1 - c_0^{-1}e)$ for $e \in [0, 0.5]$. In Figure 4-b, this means that the dashed curve OC, that is, $h = h(c_1^{-1}e)$, could intersect with the curve OHEC, that is, $h = h(1 - c_0^{-1}e)$, at points other than O and C. Consequently, the right hand side of Equation (29) should now be replaced by the concave hull function of the maximum of $h(c_1^{-1}e)$ and $h(1 - c_0^{-1}e)$. That is,

$$2 \cdot h(\tfrac{1}{2c_1}) \cdot \underline{\text{CSR}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant [\max\{h(c_1^{-1} \cdot \underline{\text{CSR}}),\, h(1 - c_0^{-1} \cdot \underline{\text{CSR}})\}]^{\frown}\,.$$

Moreover, analogous to Theorem 5 and Corollary 6, we can prove the above two bounds on $H(\mathsf{y}|\mathsf{x})$ are both tight. These results are summarized in the following theorem.

Theorem 7 (tight bounds on $H(\mathsf{y}|\mathsf{x})$ as functions of $\underline{\text{CSR}}$) *Let $h : [0, 1] \to \mathbb{R}$ be a concave function that satisfies $h(0) = h(1) = 0$. Then for any binary classification problem (μ, η) we have*

$$2 \cdot h(\tfrac{1}{2c_1}) \cdot \underline{\text{CSR}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant [\max\{h(c_1^{-1} \cdot \underline{\text{CSR}}),\, h(1 - c_0^{-1} \cdot \underline{\text{CSR}})\}]^{\frown}\,, \tag{30}$$

where $H(\mathsf{y}|\mathsf{x})$ is defined as in Definition 2, and $\underline{\text{CSR}}$ given by Equation (26). In particular, when $h(\cdot)$ is a symmetric function, it holds that $2 \cdot h(\frac{1}{2c_1}) \cdot \underline{\text{CSR}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant h(c_0^{-1} \cdot \underline{\text{CSR}})$.

One observes that when $c_0 = c_1 = 1$ the above theorem is reduced to Theorem 4, so it is the extension of Fano's and Hellman's results to the cost-sensitive case.

To simplify the discussion, let us return to the symmetric case. To get the lower/upper bounds on $\underline{\text{CSR}}$ in terms of $H(\mathsf{y}|\mathsf{x})$ from Theorem 7, we notice that the function $h(\eta)$ is symmetric on the interval $[0, 1]$ and hence monotonically non-decreasing on $[0, 0.5]$. It thus follows from the inequality $H(\mathsf{y}|\mathsf{x}) \leqslant h(c_0^{-1} \cdot \underline{\text{CSR}})$ that

$$h^{-1}_{[0,0.5]}(H(\mathsf{y}|\mathsf{x})) \leqslant c_0^{-1} \cdot \underline{\text{CSR}} \leqslant 1 - h^{-1}_{[0,0.5]}(H(\mathsf{y}|\mathsf{x}))\,,$$

where $h^{-1}_{[0,0.5]}$ denotes the inverse of the function $h(\eta)$ restricted on $[0, 0.5]$. This inequality together with $2h(\frac{1}{2c_1}) \cdot \underline{\text{CSR}} \leqslant H(\mathsf{y}|\mathsf{x})$ implies

$$c_0 \cdot h^{-1}_{[0,0.5]}(H(\mathsf{y}|\mathsf{x})) \leqslant \underline{\text{CSR}} \leqslant \min\left\{c_0 - c_0 \cdot h^{-1}_{[0,0.5]}(H(\mathsf{y}|\mathsf{x})),\, \tfrac{1}{2} \cdot [h(\tfrac{1}{2c_1})]^{-1} \cdot H(\mathsf{y}|\mathsf{x})\right\}\,. \tag{31}$$

In Figure 4-b, the above lower bound $c_0 \cdot h^{-1}_{[0,0.5]}(H(\mathsf{y}|\mathsf{x}))$ corresponds to the curve OHE; and the upper bound $\min\{\ldots\}$ corresponds to the curve OGCE (see the description of Figure 4-b).

[11] Here is a short proof. As $h(\eta)$ is symmetric and concave, it attains its maximum at $\eta = 0.5$; and it is monotonically non-decreasing on the interval $[0, 0.5]$ and monotonically non-increasing on $[0.5, 1]$. So to prove $h(c_1^{-1}e) \leqslant h(1 - c_0^{-1}e)$ it suffices to show that $c_1^{-1}e \leqslant 1 - c_0^{-1}e \leqslant 1 - c_1^{-1}e$, of which the first inequality follows from the facts $e \in [0, 0.5]$ and $c_0^{-1} + c_1^{-1} = 2$; and the second inequality is clear from the assumption $c_0 \leqslant c_1$.

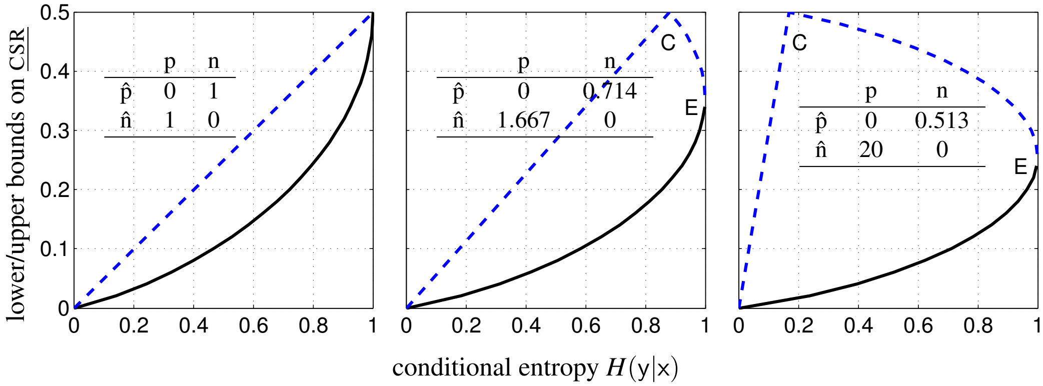


Figure 5: The lower (solid line) and upper (dashed line) bounds on the minimum cost-sensitive risk, $\underline{\text{CSR}}$, in terms of Shannon's conditional entropy. From left to right the cost of false negative is set to be $c_1 = 1, \frac{5}{3}$ and 20, respectively; and the cost of false positive is determined by the condition $c_0^{-1} + c_1^{-1} = 2$. Attached with each graph is the corresponding cost matrix, where p (n) refers to the real positive (negative); and p̂ (n̂) is the predicted positive (negative). Note that the left figure reproduces Fano's and Hellman's bounds (see Figure 1).

To give the reader an intuitive feeling about how the two bounds on the minimum cost-sensitive risk as shown in Equation (31) vary in accordance with different settings of c_0 and c_1, we plotted in Figure 5 curves of these lower/upper bounds for $c_1 = 1$, $c_1 = \frac{5}{3}$ and $c_1 = 20$, with the corresponding cost matrix attached for each subfigure. From the figure we see that when the parameter c_1 increases from 1 to ∞, the peak point C (at which $\underline{\text{CSR}}$ takes the maximum value 0.5) moves left from the top right corner $[1, 0.5]$ to the top left corner $[0, 0.5]$; whereas another extreme point E (at which $H(\mathsf{y}|\mathsf{x})$ takes the maximum value 1) moves down from the point $[1, 0.5]$ to the point $[1, 0.25]$.

A fact one should notice is that here the (tight) upper bound on $\underline{\text{CSR}}$ is no longer monotonically increasing with $H(\mathsf{y}|\mathsf{x})$, especially when c_1 is large, that is, the positive class is regarded as much more important than the negative class. This implies a non-intuitive situation. Usually with classification problems, as we decrease the conditional entropy, we would expect the worst case classification scenario to improve. With cost-sensitive risk, however, as we decrease entropy, in the worst case the cost-sensitive risk could possibly become larger—compare the points C and E in Figure 5. This important observation was *not* noted before in the literature. We will discuss it in more detail in Section 8.

5. Bounding the Minimum Balanced Error Rate by Conditional Entropy

The main goal in this section is to derive the *tight* lower and upper bounds on minimum balanced error rate, $\underline{\text{BER}}$, for binary classification problems with given conditional entropy $H(\mathsf{y}|\mathsf{x})$ and (prior) class probabilities, $\pi = \Pr\{\mathsf{y} = 1\}$ and $\tilde{\pi} = \Pr\{\mathsf{y} = 0\}$. To do so, we need first to derive the expression of $\underline{\text{BER}}$.

As we have already mentioned at the end of Section 2.3, for any given task (μ,η) the balanced error rate BER as a functional of classifiers $\hat{y}(\cdot)$ can be seen as a special cost-sensitive risk with $c_1 = \frac{1}{2}\pi^{-1}$ and $c_0 = \frac{1}{2}\tilde{\pi}^{-1}$. This observation allows us to re-use the analysis presented in Section 4 to obtain the expression of the minimum balanced error rate, $\underline{\text{BER}}$. In fact, putting $c_1 = \frac{1}{2}\pi^{-1}$ and $c_0 = \frac{1}{2}\tilde{\pi}^{-1}$ in Equation (26)—note that the derivation of Equation (26) does not rely on the conditions $c_0^{-1} + c_1^{-1} = 2$ and $c_0 \leqslant c_1$, we at once get

$$\underline{\text{BER}} = \tfrac{1}{2}\textstyle\int_X \min\{\pi^{-1}\eta(x), \tilde{\pi}^{-1}\tilde{\eta}(x)\}\mathrm{d}\mu; \tag{32}$$

and the corresponding optimal classifier is $\hat{y}(x) = [\![\eta(x) > \frac{c_0}{c_0+c_1}]\!] = [\![\eta(x) > \pi]\!]$.

We have also pointed out that, as far as only those tasks (μ,η) with constant class probabilities $\pi = \Pr\{\mathsf{y} = 1\}$ and $\tilde{\pi} = \Pr\{\mathsf{y} = 0\}$ are concerned, the quantity BER can still be regarded as a special case of CSR. Accordingly, $\underline{\text{BER}}$ as a functional of tasks (μ,η) is a special case of $\underline{\text{CSR}}$, that is, $\underline{\text{BER}}(\mu,\eta) = \underline{\text{CSR}}(\mu,\eta;c_0,c_1)$ with $c_0 = \frac{1}{2}\tilde{\pi}^{-1}$ and $c_1 = \frac{1}{2}\pi^{-1}$. (Conceptually, however, the two are totally different, as we will see soon.) We thus get from Theorem 7 the following corollary.

Corollary 8 (bounds on $H(\mathsf{y}|\mathsf{x})$ as functions of $\underline{\text{BER}}$) *Let $h(\eta)$, $\eta \in [0,1]$, be a concave function satisfying $h(0) = h(1) = 0$ and $h(\eta) = h(1-\eta)$ (symmetric). Then for any binary classification task (μ,η) with $\Pr\{\mathsf{y} = 1\} = \pi$ it holds that*

$$2 \cdot h(\pi) \cdot \underline{\text{BER}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant \begin{cases} h(2\tilde{\pi} \cdot \underline{\text{BER}}) & \textit{if } \pi \leqslant 0.5 \\ h(2\pi \cdot \underline{\text{BER}}) & \textit{if } \pi > 0.5 \end{cases}, \tag{33}$$

where $H(\mathsf{y}|\mathsf{x})$ is defined as in Definition 2, and $\underline{\text{BER}}$ is given by Equation (32).

Proof If $\pi \leqslant \frac{1}{2}$, the settings $c_1 = \frac{1}{2}\pi^{-1}$ and $c_0 = \frac{1}{2}\tilde{\pi}^{-1}$ satisfy the two conditions $c_0 \leqslant c_1$ and $c_0^{-1} + c_1^{-1} = 2$. So from Equation (29) we get $h(\pi) \cdot 2\underline{\text{BER}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant h(2\tilde{\pi} \cdot \underline{\text{BER}})$. For the case of $\pi \geqslant \frac{1}{2}$, we still set $c_1 = \frac{1}{2}\pi^{-1}$ and $c_0 = \frac{1}{2}\tilde{\pi}^{-1}$ but interchange the roles of c_0 and c_1 in Equation (29). This gives us $h(\pi) \cdot 2\underline{\text{BER}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant h(2\pi \cdot \underline{\text{BER}})$. Here we have used the fact that $h(\pi) = h(\tilde{\pi})$—notice that $h(\cdot)$ is a symmetric function. ■

As we have just said, balanced error rate and cost-sensitive risk are two different concepts. The main difference is that in the expression of CSR, Equation (15), the two coefficients c_0 and c_1 depend on neither μ nor η; whereas for BER, Equation (13), the values of c_0 and c_1 depend on the concerned problem (μ,η). This difference results in that the two bounds on $H(\mathsf{y}|\mathsf{x})$ as claimed by Corollary 8 are *not* necessarily tight any more. The reason is that, in terms of $\underline{\text{CSR}}$, the lower/upper bounds on $H(\mathsf{y}|\mathsf{x})$ are obtained over all tasks with a fixed value of $\underline{\text{CSR}}$; whereas in terms of $\underline{\text{BER}}$, we implicitly imposed an additional condition on the tasks, namely, the class probabilities should also be fixed as π and $\tilde{\pi}$. Consequently, the development presented in Section 4 can not be directly applied here to obtain the *tight* bounds.

We now use Scheme 3 to derive the tight bounds on $H(\mathsf{y}|\mathsf{x})$ in terms of $\underline{\text{BER}}$ and π. The resulting bounds are presented in Theorem 11.—This is only for theoretical convenience: in practice we are often more interested in using $H(\mathsf{y}|\mathsf{x})$ to bound $\underline{\text{BER}}$, which can be obtained from Theorem 11 by simply interchanging the axes of $H(\mathsf{y}|\mathsf{x})$ and $\underline{\text{BER}}$, as has been done in Figure 6-b. In fact, we have already used this technique to study the bounds on $\underline{\text{CSR}}$ in terms of $H(\mathsf{y}|\mathsf{x})$, see the derivation of Equation (31) in page 1050.

In Equation (21), we have already expressed $H(\mathsf{y}|\mathsf{x})$ as the mathematical expectation of some function of $\eta(\mathsf{x})$. The other two quantities involved, π and $\underline{\mathrm{BER}}$, can also be written in this way. In fact, by Equations (32) and (4), one easily sees that $\underline{\mathrm{BER}} = \frac{1}{2}\mathbb{E}_{\mathsf{x}\sim\mu}[r(\eta(\mathsf{x}))]$ and $\pi = \mathbb{E}_{\mathsf{x}\sim\mu}[\eta(\mathsf{x})]$, where the function $r(\eta)$ is defined as

$$r(\eta) := \min\{\pi^{-1}\eta, \tilde{\pi}^{-1}\tilde{\eta}\}, \qquad \eta \in [0,1].$$

So Equations (4), (32) and (21) can be rewritten together in vector notation as

$$[\pi, 2\underline{\mathrm{BER}}, H(\mathsf{y}|\mathsf{x})] = \mathbb{E}_{\mathsf{x}\sim\mu}[\eta(\mathsf{x}), r(\eta(\mathsf{x})), h(\eta(\mathsf{x}))]. \tag{34}$$

It then follows from Theorem 1 that the point $[\pi, 2\underline{\mathrm{BER}}, H(\mathsf{y}|\mathsf{x})]$ is in the convex hull of the curve $\ell = \{[\eta, r(\eta), h(\eta)] \mid \eta \in [0,1]\}$ in the *three* dimensional η-r-h space.

Next we would implement the second and third steps in Scheme 3, for which we use the graph of ℓ and its convex hull with $\pi = 0.3$ and $h(\eta) = -\eta\log\eta - \tilde{\eta}\log\tilde{\eta}$ (Shannon) as the example (see Figure 6-a). For any given value of π, the function $r(\eta)$ is piecewise linear: it equals to $\pi^{-1}\eta$ if $\eta \leqslant \pi$ and $\tilde{\pi}^{-1}\tilde{\eta}$ otherwise. The curve ℓ is hence divided into two parts by the point $[\pi, 1, h(\pi)]$—the point A in Figure 6-a. The part with $\eta \leqslant \pi$ is in the plane $\eta = \pi r$ (plane AOD); and the part with $\eta \geqslant \pi$ is in the plane $\tilde{\eta} = \tilde{\pi} r$ (plane ABD). For such a curve simple geometry tells us that its convex hull is bounded by the triangle OAB, the two bow shapes OAO and ABA, and the minimal "concave" curved surface OAB bordered by the curve ℓ and the line segment OB—see Appendix B.3 for more detail.

Finally, we want to compute the *tight* lower and upper bounds on $H(\mathsf{y}|\mathsf{x})$ from the convex hull of ℓ. For any fixed value of π, this is equivalent to seeking the intersection of the plane $\eta = \pi$ and the convex hull of ℓ. From Figure 6-a, it is easy to see that $H(\mathsf{y}|\mathsf{x})$ is lower bounded by the line segment AC, the intersection of the planes ADC (the plain $\eta = \pi$) and OAB (the "tight lower bound" of the convex hull of ℓ). It is also obvious that the two endpoints of AC have the coordinates $\mathrm{A}(\pi, 1, h(\pi))$ and $\mathrm{C}(\pi, 0, 0)$. Therefore,

$$h(\pi) \cdot 2\underline{\mathrm{BER}} \leqslant H(\mathsf{y}|\mathsf{x}). \tag{35}$$

This inequality is same as the first inequality in Equation (33); they are nevertheless obtained by different methods. The most important difference is that here we can safely claim the sharpness of the inequality (by an argument similar to that in Theorem 5 and Corollary 6), which is not clear from Corollary 8.

Analogously, the tight upper bound on $H(\mathsf{y}|\mathsf{x})$ is determined by the intersection curve of plane $\eta = \pi$ (plane ADC) and the aforementioned curved surface OAB. Therefore, to compute this tight upper bound we need to find the maximal value of h such that the point $[\pi, 2\underline{\mathrm{BER}}, h]$ is in the convex hull of ℓ. By the definition of convex hull, this can be done as follows. Pick any two points, say M and N (not plotted in Figure 6-a), on the curve ℓ or the line segment OB, so that the line segment MN meets the vertical line defined by $\eta = \pi$ and $r = 2\underline{\mathrm{BER}}$, say the line EF in the figure, at some point K. By definition, we know point K is in the convex hull. So its h-coordinate, K_h, is no more than the maximal value of $H(\mathsf{y}|\mathsf{x})$; and the maximum of K_h (over all possible pairs M and N) is exactly the *tight* upper bound of $H(\mathsf{y}|\mathsf{x})$.

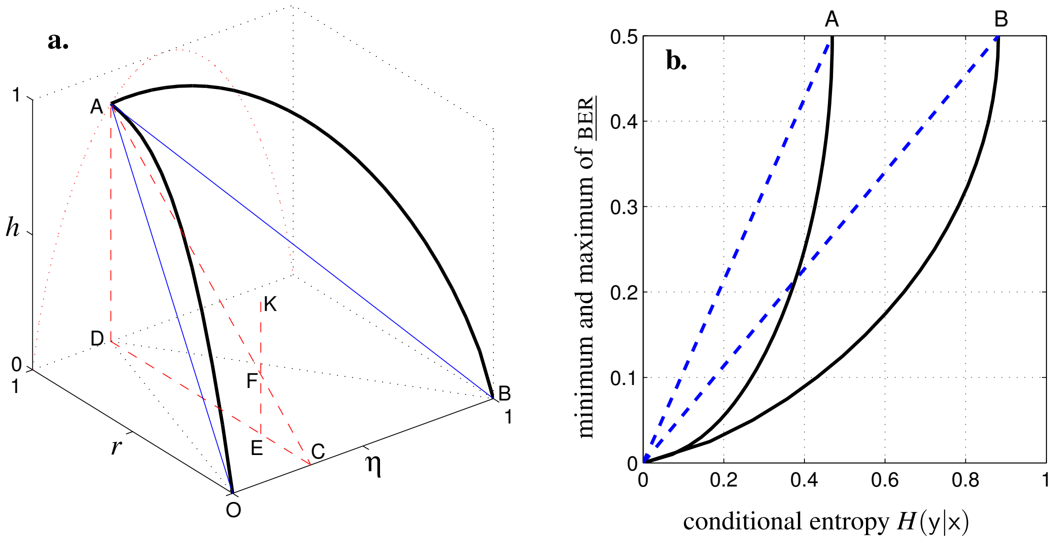


Figure 6: **a.** The curve $\ell = \{[\eta, r(\eta), h(\eta)] \mid \eta \in [0,1]\}$ and its convex hull in the η-r-h space. **b.** The tight lower (solid lines) and upper (dashed lines) bounds on $\underline{\mathrm{BER}}$ versus the Shannon conditional entropy for $\pi = 0.1$ (A) and $\pi = 0.3$ (B). Note the difference between the bow shape (B) here and that in the middle graph of Figure 5—the two use the same parameters: $\pi = 0.3$, that is, $c_1 = \frac{5}{3}$.

To compute the maximal value of K_h, let M, N be as above and write $\rho := 2\underline{\mathrm{BER}}$. Then, as K is on the line segment MN, there exists a unique $t \in (0,1)$ such that

$$\mathrm{K}_\eta = \tilde{t} \cdot \mathrm{M}_\eta + t \cdot \mathrm{N}_\eta = \pi\,, \tag{36}$$

$$\mathrm{K}_r = \tilde{t} \cdot \mathrm{M}_r + t \cdot \mathrm{N}_r = \rho\,, \tag{37}$$

$$\mathrm{K}_h = \tilde{t} \cdot \mathrm{M}_h + t \cdot \mathrm{N}_h\,, \tag{38}$$

where the subscript η, r or h denotes the corresponding coordinate of the concerned point. We shall discuss two different cases separately.

Case 1: One of M *and* N, *say* M, *is on line segment* OB. That is, $\mathrm{M}_r = \mathrm{M}_h = 0$ and $0 \leqslant \mathrm{M}_\eta \leqslant 1$. If $0 \leqslant \mathrm{M}_\eta \leqslant \pi$, by Equation (36), $\mathrm{N}_\eta \geqslant \pi$ and so $\mathrm{N}_r = \tilde{\pi}^{-1}(1-\mathrm{N}_\eta)$. This equation, together with Equations (36) and (37), implies that $\mathrm{N}_\eta = 1 - t^{-1}\tilde{\pi}\rho$ and $\mathrm{M}_\eta = 1 - \tilde{t}^{-1}\tilde{\pi}\tilde{\rho}$. So from Equation (38) we know $\mathrm{K}_h = t \cdot h(1 - t^{-1}\tilde{\pi}\rho)$, where the range of t is determined by the condition $0 \leqslant \mathrm{M}_\eta \leqslant \pi$, from which we obtain $t \in [\rho, \rho + \pi\tilde{\rho}]$. Similarly, for $\pi \leqslant \mathrm{M}_\eta \leqslant 1$ it holds that $\mathrm{N}_\eta \leqslant \pi$ and $\mathrm{N}_r = \pi^{-1}\mathrm{N}_\eta$; and by solving the three equations (36)–(38) we get $\mathrm{N}_\eta = t^{-1}\pi\rho$, $\mathrm{M}_\eta = \tilde{t}^{-1}\pi\tilde{\rho}$, and $\mathrm{K}_h = t \cdot h(t^{-1}\pi\rho)$, with $t \in [\rho, \rho + \tilde{\pi}\tilde{\rho}]$.

Case 2: Both M *and* N *are on the curve* ℓ. Without loss of generality, assume $\mathrm{M}_\eta \leqslant \mathrm{N}_\eta$. Then by Equation (36) we know $\mathrm{M}_\eta \leqslant \pi \leqslant \mathrm{N}_\eta$; and hence $\mathrm{M}_r = \pi^{-1}\mathrm{M}_\eta$ and $\mathrm{N}_r = \tilde{\pi}^{-1}(1-\mathrm{N}_\eta)$. Substituting the two equations into Equation (37) and solving the resulting linear equations (36) and (37), we arrive at $\mathrm{M}_\eta = \pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}$ and $\mathrm{N}_\eta = \pi + t^{-1}\pi\tilde{\pi}\tilde{\rho}$. It then follows from Equation (38) that $\mathrm{K}_h = \tilde{t} \cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) + t \cdot h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho})$, where $t \in [\pi\tilde{\rho}, \rho + \pi\tilde{\rho}]$, as is determined by the conditions $\mathrm{M}_\eta \geqslant 0$ and $\mathrm{N}_\eta \leqslant 1$.

Summing up the above discussion, we conclude that the tight upper bound on $H(\mathsf{y}|\mathsf{x})$ is the maximum of the three maxima:

$$\mathrm{K}_h^{(1)} = \max\{f_1(t) := t \cdot h(1 - t^{-1}\tilde{\pi}\rho) \mid t \in [\rho, \rho + \pi\tilde{\rho}]\}\,, \tag{39}$$

$$\mathrm{K}_h^{(2)} = \max\{f_2(t) := t \cdot h(t^{-1}\pi\rho) \mid t \in [\rho, \rho + \tilde{\pi}\tilde{\rho}]\}\,, \tag{40}$$

$$\mathrm{K}_h^{(3)} = \max\{f_3(t) := \tilde{t} \cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) + t \cdot h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho}) \mid t \in [\pi\tilde{\rho}, \rho + \pi\tilde{\rho}]\}\,. \tag{41}$$

Lemma 9 *Let $h : [0,1] \to \mathbb{R}$ be a concave function. Let $u, v, w \in [0,1]$ and $\alpha, \beta, \gamma \geqslant 0$ be such that $\alpha + \beta = \gamma$ and $\alpha u + \beta v = \gamma w$. Then*

$$\alpha \cdot h(u) + \beta \cdot h(v) \leqslant \gamma \cdot h(w)\,. \tag{42}$$

Proof The case of $\alpha = \beta = \gamma = 0$ is trivial. If at least one of the three is nonzero, then $\gamma > 0$ since $\gamma = \alpha + \beta$. Equation (42) is then just a reformulation of the characterizing (defining) inequality of concave functions, $t \cdot h(u) + \tilde{t} \cdot h(v) \leqslant h(t \cdot u + \tilde{t} \cdot v)$, with $t = \gamma^{-1}\alpha$. ■

Lemma 10 *In Equations* (39)–(41)*,* (a) *the functions $f_1(t)$ and $f_2(t)$ are monotonically non-decreasing, so $\mathrm{K}_h^{(1)} = f_1(\rho + \pi\tilde{\rho})$ and $\mathrm{K}_h^{(2)} = f_2(\rho + \tilde{\pi}\tilde{\rho})$;* (b) *the function $f_3(t)$ is concave, and its value at the two endpoints are $f_3(\rho + \pi\tilde{\rho}) = \mathrm{K}_h^{(1)}$ and $f_3(\pi\tilde{\rho}) = \mathrm{K}_h^{(2)}$, respectively.*

Proof (a) For $t_1, t_2 \in [\rho, \rho + \pi\tilde{\rho}]$ satisfying $t_1 \leqslant t_2$, we need to show that $f_1(t_1) \leqslant f_1(t_2)$, that is, $t_1 \cdot h(1 - t_1^{-1}\tilde{\pi}\rho) \leqslant t_2 \cdot h(1 - t_2^{-1}\tilde{\pi}\rho)$. This can be obtained by substituting $(\alpha, \beta, \gamma) = (t_1, t_2 - t_1, t_2)$ and $(u, v, w) = (1 - t_1^{-1}\tilde{\pi}\rho, 1, 1 - t_2^{-1}\tilde{\pi}\rho)$ into Equation (42) and using the fact that $h(1) = 0$.

Similarly, the inequality $f_2(t_1) \leqslant f_2(t_2)$, that is, $t_1 \cdot h(t_1^{-1}\pi\rho) \leqslant t_2 \cdot h(t_2^{-1}\pi\rho)$ can be proven using the settings $(\alpha, \beta, \gamma) = (t_1, t_2 - t_1, t_2)$ and $(u, v, w) = (t_1^{-1}\pi\rho, 0, t_2^{-1}\pi\rho)$ for Equation (42), as well as the fact that $h(0) = 0$.

(b) For any $t_1, t_2 \in [\pi\tilde{\rho}, \rho + \pi\tilde{\rho}]$ and $\alpha \in (0,1)$, write $t = \alpha \cdot t_1 + \tilde{\alpha} \cdot t_2$. We want to prove that $f_3(t) \geqslant \alpha \cdot f_3(t_1) + \tilde{\alpha} \cdot f_3(t_2)$. But Lemma 9 implies that

$$\tilde{t} \cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) \geqslant \alpha\tilde{t}_1 \cdot h(\pi - \tilde{t}_1^{-1}\pi\tilde{\pi}\tilde{\rho}) + \tilde{\alpha}\tilde{t}_2 \cdot h(\pi - \tilde{t}_2^{-1}\pi\tilde{\pi}\tilde{\rho})\,,$$
$$t \cdot h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho}) \geqslant \alpha t_1 \cdot h(\pi + t_1^{-1}\pi\tilde{\pi}\tilde{\rho}) + \tilde{\alpha} t_2 \cdot h(\pi + t_2^{-1}\pi\tilde{\pi}\tilde{\rho})\,.$$

The sum of the above two inequalities is exactly what we want: $f_3(t) \geqslant \alpha \cdot f_3(t_1) + \tilde{\alpha} \cdot f_3(t_2)$. Finally, the two identities $f_3(\rho + \pi\tilde{\rho}) = f_1(\rho + \pi\tilde{\rho}) = \mathrm{K}_h^{(1)}$ and $f_3(\pi\tilde{\rho}) = f_2(\rho + \tilde{\pi}\tilde{\rho}) = \mathrm{K}_h^{(2)}$ can be verified by direct computation. ■

As a consequence of the above lemma, we see that $H(\mathsf{y}|\mathsf{x})$ is actually upper bounded by the quantity $\mathrm{K}_h^{(3)}$ as defined in Equation (41), that is,

$$H(\mathsf{y}|\mathsf{x}) \leqslant \max\left\{f_3(t) = \tilde{t} \cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) + t \cdot h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho}) \mid t \in [\pi\tilde{\rho}, \rho + \pi\tilde{\rho}]\right\}\,, \tag{43}$$

where $\rho = 2\underline{\mathrm{BER}}$ is two times of the minimum balanced error rate and $\pi = \Pr\{\mathsf{y}=1\}$ the prior probability of the positive class. Furthermore, using an argument similar to that for Theorem 5 and Corollary 6, we can prove the above upper bound on $H(\mathsf{y}|\mathsf{x})$ is tight.

Combining Equation (35) and Equation (43), we get

Theorem 11 (tight bounds on $H(\mathsf{y}|\mathsf{x})$ in terms of $\underline{\mathrm{BER}}$ and π) *Let $h(\eta)$, $\eta \in [0,1]$, be a concave function satisfying $h(0) = h(1) = 0$. Then for any binary classification task (μ, η) with $\Pr\{\mathsf{y}=1\} = \pi$ it holds that*

$$2 \cdot h(\pi) \cdot \underline{\mathrm{BER}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant \max\{f_3(t) \mid t \in [\pi\tilde{\rho}, \rho + \pi\tilde{\rho}]\} = \mathrm{K}_h^{(3)}\,, \tag{44}$$

where $\rho = 2\underline{\mathrm{BER}}$ and the function $f_3(t)$ is defined by Equation (43).

Notice that Theorem 11 does not require that $h(\cdot)$ be symmetric. Furthermore, as has been pointed out earlier, there are two ways to understand this theorem. The first way is to see π as a given parameter, then Equation (44) describes the relationship between the functionals $H(\mathsf{y}|\mathsf{x})$ and $\underline{\mathrm{BER}}$; and it holds for any task with $\Pr\{\mathsf{y}=1\} = \pi$. We can also regard π as a functional of tasks, then Equation (44) connects the three quantities: π, $\underline{\mathrm{BER}}$ and $H(\mathsf{y}|\mathsf{x})$; and holds for any classification task (μ, η).

In Theorem 11, the tight upper bound on $H(\mathsf{y}|\mathsf{x})$ has been written as the maximum of a concave function $f_3(t) = \tilde{t} \cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) + t \cdot h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho})$ over the interval $[\pi\tilde{\rho}, \rho + \pi\tilde{\rho}]$. This maximum has *no* closed-form expression in general, we therefore resort to numerical methods. If the function $h(\cdot)$ is differentiable, so is $f_3(t)$—the derivative of $f_3(t)$ is

$$\begin{aligned} f_3'(t) = &- h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho} \cdot h'(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) \\ &+ h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho}) - t^{-1}\pi\tilde{\pi}\tilde{\rho} \cdot h'(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho})\,. \end{aligned}$$

In this case, the maximum of $f_3(t)$ can be obtained by checking the values of its derivative $f_3'(t)$. First at the two endpoints $\pi\tilde{\rho}$ and $\rho + \pi\tilde{\rho}$: if $f_3'(\pi\tilde{\rho}) \leqslant 0$, then $\mathrm{K}_h^{(3)} = f_3(\pi\tilde{\rho})$; if $f_3'(\rho + \pi\tilde{\rho}) \geqslant 0$, then $\mathrm{K}_h^{(3)} = f_3(\rho + \pi\tilde{\rho})$. Otherwise we need to calculate the unique solution $t_0 \in (\pi\tilde{\rho}, \rho + \pi\tilde{\rho})$ to the equation $f_3'(t) = 0$ and obtain $\mathrm{K}_h^{(3)} = f_3(t_0)$. This can be done very efficiently by simple numerical methods such as bisection, since $f_3'(t)$ is a non-increasing function of t. If $h(\cdot)$ is not differentiable, one still can use simple numerical methods such as the Fibonacci search and the golden section search (Brent, 1973, p. 68) to locate the maximum of $f_3(t)$, since $f_3(t)$ is a unimodal function.

For the Shannon conditional entropy, $h(\eta) = -\eta \log p - \tilde{\eta} \log \tilde{\eta}$ and the value of $f_3'(t)$ at the two endpoints are $f_3'(\rho + \pi\tilde{\rho}) = -\infty$ and $f_3'(\pi\tilde{\rho}) = \infty$, respectively. The problem is thus reduced to solving the equation $f_3'(t) = 0$, which can be simplified to

$$\pi \log(1 - \tilde{t}^{-1}\tilde{\pi}\tilde{\rho}) + \tilde{\pi} \log(1 + \tilde{t}^{-1}\pi\tilde{\rho}) = \pi \log(1 + t^{-1}\tilde{\pi}\tilde{\rho}) + \tilde{\pi} \log(1 - t^{-1}\pi\tilde{\rho})\,.$$

Its solution t_0 is then substituted into the expression of $f_3(t)$, yielding the tight upper bound of $H(\mathsf{y}|\mathsf{x})$. In Figure 6-b the lower and upper bounds on $H(\mathsf{y}|\mathsf{x})$ are plotted versus minimum balanced error rate $\underline{\mathrm{BER}} = \frac{1}{2}\rho$ for $\pi = 0.1$ and 0.3—corresponding to $c_1 = \frac{1}{2}\pi^{-1} = 5$ and $\frac{5}{3}$. The graph has used the x-axis for $H(\mathsf{y}|\mathsf{x})$ and the y-axis for $\underline{\mathrm{BER}}$, so that one can easily check the bounds on $\underline{\mathrm{BER}}$ for given values of $H(\mathsf{y}|\mathsf{x})$. One may compare the middle graph of Figure 5 with Figure 6-b, to confirm that the upper bound on $H(\mathsf{y}|\mathsf{x})$ stated by Corollary 8 is indeed untight (for those tasks with $0 < \underline{\mathrm{BER}} < 0.5$). In Appendix C, we will show that *the upper bound of $H(\mathsf{y}|\mathsf{x})$ given by Theorem 11 is never looser than that in Corollary 8.*

To conclude, we have used the proposed geometric method to derive the *tight* lower and upper bounds on $H(\mathsf{y}|\mathsf{x})$ in terms of $\underline{\text{BER}}$ and π. By flipping the curve of these bounds along the diagonal line, we can also get the *tight* lower and upper bounds on $\underline{\text{BER}}$ as functions of $H(\mathsf{y}|\mathsf{x})$ and π, as we have done for the quantity $\underline{\text{CSR}}$—see Equation (31). As shown by Corollary 8, these tight bounds are not obtained by simply taking the balanced error rate as a special cost-sensitive risk, even though here the value of π is assumed to be a constant. This confirms that balanced error rate and cost-sensitive risk are two *essentially* different performance measures.

6. Maximum F-score in the Binary Classification Problem

We now consider the relationship between F-score and conditional entropy. As before, we shall first derive the maximum value of F-score for a given classification problem (μ, η). By Equation (17), this amounts to maximizing the set function

$$\Gamma(\mathcal{X}_1) := \tfrac{1}{2}\text{FSC} = [\pi + \mu(\mathcal{X}_1)]^{-1} \cdot \textstyle\int_{\mathcal{X}_1} \eta(x)\mathrm{d}\mu, \tag{45}$$

under the assumption that the object distribution μ and the conditional class probability $\eta(x)$ are constants (so the class probability π is also a constant).

Lemma 12 *Let the set function* $\Gamma(\mathcal{X}_1)$ *be as in Equation* (45)*. For any measurable subset* $\mathcal{X}_1$ *of* $\mathcal{X}$*, let* $\mathcal{X}_1' = \{x \in \mathcal{X} \mid \eta(x) > \Gamma(\mathcal{X}_1)\}$*. Then* $\Gamma(\mathcal{X}_1') \geqslant \Gamma(\mathcal{X}_1)$*.*

Proof Write $\theta = \Gamma(\mathcal{X}_1)$. Let $A = \{x \in \mathcal{X}_1 \mid \eta(x) \leqslant \theta\}$ and $B = \{x \notin \mathcal{X}_1 \mid \eta(x) > \theta\}$. Then $A \subseteq \mathcal{X}_1$, $B \cap \mathcal{X}_1 = \varnothing$ and $\mathcal{X}_1' = (\mathcal{X}_1 \setminus A) \cup B$. Thus, by Equation (45),

$$\Gamma(\mathcal{X}_1') = \frac{\int_{\mathcal{X}_1'} \eta(x)\mathrm{d}\mu}{\pi + \mu(\mathcal{X}_1')} = \frac{\int_{\mathcal{X}_1} \eta(x)\mathrm{d}\mu + \int_B \eta(x)\mathrm{d}\mu - \int_A \eta(x)\mathrm{d}\mu}{\pi + \mu(\mathcal{X}_1) + \mu(B) - \mu(A)}.$$

As $\eta(x) > \theta$ on B, $\int_B \eta(x)\mathrm{d}\mu \geqslant \theta\mu(B)$. Similarly, $\int_A \eta(x)\mathrm{d}\mu \leqslant \theta\mu(A)$. Furthermore, by the definition of $\Gamma(\mathcal{X}_1)$, we have $\int_{\mathcal{X}_1} \eta(x)\mathrm{d}\mu = \theta[\pi + \mu(\mathcal{X}_1)]$. All these three facts together imply that $\Gamma(\mathcal{X}_1') \geqslant \theta = \Gamma(\mathcal{X}_1)$. ■

This theorem allows us to consider only classifiers of the form $\hat{y}(x) = [\![\eta(x) > \theta]\!]$ when maximizing the F-score, where $\theta \in [0,1]$ is a threshold. To determine the optimal threshold θ so that the F-score, or, equivalently, the function $\Gamma(\mathcal{X}_1)$ is maximized, where the set $\mathcal{X}_1$ is defined via θ as $\mathcal{X}_1(\theta) = \{x \in \mathcal{X} \mid \eta(x) > \theta\}$, we rewrite Equation (45) as a function of θ:

$$\Gamma(\theta) = \frac{\int_{\mathcal{X}_1} \eta(x)\mathrm{d}\mu}{\pi + \mu(\mathcal{X}_1)} = \frac{\theta\mu(\mathcal{X}_1) + \int_{\mathcal{X}_1}[\eta(x) - \theta]\mathrm{d}\mu}{\pi + \mu(\mathcal{X}_1)} = \theta + \frac{\int_{\mathcal{X}_1}[\eta(x) - \theta]\mathrm{d}\mu - \pi\theta}{\pi + \mu(\mathcal{X}_1)}. \tag{46}$$

For any $r \in \mathbb{R}$, write $r^+ := \max\{0, r\}$. By the definition of $\mathcal{X}_1$, $\eta(x) - \theta > 0$ iff $x \in \mathcal{X}_1$. So for $x \in \mathcal{X}_1$, $[\eta(x) - \theta]^+ = \eta(x) - \theta$; and for $x \notin \mathcal{X}_1$, $[\eta(x) - \theta]^+ = 0$. It therefore follows that $\int_{\mathcal{X}_1}[\eta(x) - \theta]\mathrm{d}\mu = \int_{\mathcal{X}}[\eta(x) - \theta]^+\mathrm{d}\mu$. Substituting this into Equation (46), we obtain

$$\Gamma(\theta) = \frac{\theta\mu(\mathcal{X}_1) + \int_{\mathcal{X}}[\eta(x) - \theta]^+\mathrm{d}\mu}{\pi + \mu(\mathcal{X}_1)} = \theta + \frac{\int_{\mathcal{X}}[\eta(x) - \theta]^+\mathrm{d}\mu - \pi\theta}{\pi + \mu(\mathcal{X}_1)}. \tag{47}$$

Lemma 13 *The function* $g(\theta) := \int_{\mathcal{X}} [\eta(x) - \theta]^+ \mathrm{d}\mu - \pi\theta$, *where* $\theta \in [0,1]$, *is continuous and strictly decreasing, with* $g(0) = \pi$ *and* $g(1) = -\pi$.

Proof Since $|(\eta - \theta_1)^+ - (\eta - \theta_2)^+| \leqslant |\theta_1 - \theta_2|$, we have $|g(\theta_1) - g(\theta_2)| \leqslant (1+\pi)|\theta_1 - \theta_2|$, so $g(\theta)$ is continuous. It is strictly decreasing because its first term is non-increasing (with respect to θ) and its second term, $-\pi\theta$, is strictly decreasing. Finally, as $0 \leqslant \eta(x) \leqslant 1$, we know $[\eta(x)]^+ = \eta(x)$ and $[\eta(x) - 1]^+ = 0$; so $g(0) = \pi$ and $g(1) = -\pi$. ■

By this lemma, we know there exists a unique $\theta^* \in (0,1)$ such that $g(\theta^*) = 0$, that is,

$$\int_{\mathcal{X}} [\eta(x) - \theta^*]^+ \mathrm{d}\mu - \pi\theta^* = 0. \tag{48}$$

We now prove it is this θ^* that maximizes the function $\Gamma(\theta)$; and the maximum value is $\Gamma(\theta^*) = \theta^*$, as can be easily seen from Equation (47) and Equation (48).

Lemma 14 *The function* $\Gamma(\theta)$ *as given by Equation* (46) *is maximized at* θ^*.

Proof We shall use the first expression of $\Gamma(\theta)$ from Equation (46), in which the subset $\mathcal{X}_1$ is defined as $\mathcal{X}_1 = \{x \in \mathcal{X} \mid \eta(x) > \theta\}$. If $\theta < \theta^*$, define $A := \{x \in \mathcal{X} \mid \eta(x) > \theta^*\}$ and $B := \{x \in \mathcal{X} \mid \theta^* \geqslant \eta(x) > \theta\}$. Then it is clear that $A \cap B = \varnothing$ and $A \cup B = \mathcal{X}_1$. Thus,

$$\Gamma(\theta) = \frac{\int_A \eta(x)\mathrm{d}\mu + \int_B \eta(x)\mathrm{d}\mu}{\pi + \mu(A) + \mu(B)}.$$

Now, as $\theta^* = \Gamma(\theta^*) = [\pi + \mu(A)]^{-1} \cdot \int_A \eta(x)\mathrm{d}\mu$, we have $\int_A \eta(x)\mathrm{d}\mu = \theta^* \cdot [\pi + \mu(A)]$. Moreover, $\int_B \eta(x)\mathrm{d}\mu \leqslant \theta^* \mu(B)$ since $\eta(x) \leqslant \theta^*$ on B. Therefore, $\Gamma(\theta) \leqslant \theta^* = \Gamma(\theta^*)$.

If $\theta > \theta^*$, define A as before and $B := \{x \in \mathcal{X} \mid \theta \geqslant \eta(x) > \theta^*\}$. Then $B \subseteq A$ and $\mathcal{X}_1 = A \setminus B$. Thus,

$$\Gamma(\theta) = \frac{\int_A \eta(x)\mathrm{d}\mu - \int_B \eta(x)\mathrm{d}\mu}{\pi + \mu(A) - \mu(B)}.$$

Since $\eta(x) > \theta^*$ for $x \in B$, it holds that $\int_B \eta(x)\mathrm{d}\mu \geqslant \theta^* \mu(B)$; whereas the equality $\int_A \eta(x)\mathrm{d}\mu = \theta^* \cdot [\pi + \mu(A)]$ remains true. So, again, we obtain $\Gamma(\theta) \leqslant \theta^* = \Gamma(\theta^*)$. ■

In summary, to determine the maximum F-score for a given classification problem, one needs only to find the unique solution θ^* to the equation (48). The maximum F-score is then $\overline{\text{FSC}} = 2 \cdot \Gamma(\theta^*) = 2\theta^*$; and the corresponding optimal classifier is $\hat{y}(x) = [\![\eta(x) > \theta^*]\!]$. An interesting implication of the equality $\overline{\text{FSC}} = 2\theta^*$ is that $\theta^* \leqslant \frac{1}{2}$ (as $\overline{\text{FSC}} \leqslant 1$). That is, for F-score the optimal threshold is always less than or equal to 0.5.

7. Bounds on the Maximum F-score in Terms of Conditional Entropy

In this section, we derive bounds on maximum F-score, $\overline{\text{FSC}}$, in terms of the conditional entropy $H(\mathrm{y}|\mathrm{x})$ as defined by Equation (21). As before, we shall first examine the range of $H(\mathrm{y}|\mathrm{x})$ for any given value of $\overline{\text{FSC}}$.

In the preceding section we have proved that $\theta^* := \frac{1}{2}\overline{\mathrm{FSC}} \in [0, 0.5]$ (as $\overline{\mathrm{FSC}} \leqslant 1$) is the unique solution to the equation (48), which, by Equation (4), can be rewritten as

$$\textstyle\int_X \{\theta^* \cdot \eta(x) - [\eta(x) - \theta^*]^+\}\, \mathrm{d}\mu = 0; \quad \text{i.e.,} \quad \mathbb{E}_{\mathsf{x}\sim\mu}[u(\eta(\mathsf{x}))] = 0,$$

where $u(\eta) := \theta^*\eta - (\eta - \theta^*)^+$ is a function on $[0,1]$. For any fixed value of θ^*, we know from the above equation and Equations (4) and (21) that

$$[\pi, 0, H(\mathsf{y}|\mathsf{x})] = \mathbb{E}_{\mathsf{x}\sim\mu}[\eta(\mathsf{x}), u(\eta(\mathsf{x})), h(\eta(\mathsf{x}))]. \tag{49}$$

By Theorem 1, in the η-u-h space $[\pi, 0, H(\mathsf{y}|\mathsf{x})]$ is a point in the convex hull of the curve $\ell = \{[\eta, u(\eta), h(\eta)] \mid \eta \in [0,1]\}$.—*We have completed (a variant of) the first step in Scheme 3.*

In Figure 7-a the graph of the curve ℓ is plotted for $\theta^* = 0.3$ and the Shannon conditional entropy. By the definition of $u(\eta)$, we have $u(\eta) = \theta^*\eta$ for $\eta \leqslant \theta^*$ and $u(\eta) = \theta^* - \tilde{\theta}^*\eta$ for $\eta \geqslant \theta^*$. Thus, as in the case of balanced error rate, here the curve ℓ consists also of two parts each of which is in a plane. Consequently, its convex hull is bounded by three flat facets and one curved surface OAB (O is the origin) which is the minimum concave surface with line segment OB and curve ℓ as its boundary.—*These are the second and third steps in Scheme 3.*

As its second coordinate is a constant 0, the point $[\pi, 0, H(\mathsf{y}|\mathsf{x})]$ lies in the intersection of the plane $u = 0$ and the convex hull of ℓ. Therefore, as shown by Figure 7-a, $H(\mathsf{y}|\mathsf{x})$ is lower bounded by the line OD; and upper bounded by the curve OE that is the intersection of the plane $u = 0$ and the curved surface OAB we just mentioned. Notice here that, π is not fixed, but may take values between the points O and C (C is the intersection point of line $u = h = 0$ and plane $u = \theta^* - \tilde{\theta}^*\eta$). That is, $\pi \in [0, \theta^*/\tilde{\theta}^*]$. Thus, the lower bound of $H(\mathsf{y}|\mathsf{x})$ is given by the minimum h-coordinate of points on the line segment OD. Obviously, this equals to 0, the h-coordinate of the origin point O, which happens when π tends to zero. That means, for any given value of θ^*, the *tight* lower bound of $H(\mathsf{y}|\mathsf{x})$ is always 0.

Similarly, the maximum h-coordinate of points on the curve OE is the upper bound of $H(\mathsf{y}|\mathsf{x})$. For *symmetric* functions $h(\cdot)$, we shall soon prove that the endpoint E has the maximum h-coordinate (over all points on the curve OE). Furthermore, from Figure 7-a we know the η-coordinate of E is $\mathrm{E}_\eta = \theta^*/\tilde{\theta}^*$, so its h-coordinate $\mathrm{E}_h = h(\theta^*/\tilde{\theta}^*) = h\left(\frac{\overline{\mathrm{FSC}}}{2-\overline{\mathrm{FSC}}}\right)$.

From the above discussion we obtain the tight lower and upper bounds on $H(\mathsf{y}|\mathsf{x})$, as follows (*the fourth step in Scheme 3*).

Theorem 15 (tight bounds on $H(\mathsf{y}|\mathsf{x})$ in terms of $\overline{\mathrm{FSC}}$) *Let $H(\mathsf{y}|\mathsf{x})$ be the conditional entropy defined by a symmetric concave function $h : [0,1] \to \mathbb{R}$. Then for any two-class problem (μ, η), it holds that*

$$0 \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant h\left(\frac{\overline{\mathrm{FSC}}}{2-\overline{\mathrm{FSC}}}\right), \tag{50}$$

and the two inequalities are sharp.

A remark on Theorem 15: As we have used a variant of Scheme 3 to *derive* the inequality (50), the two general theorems 1 and 5 cannot be directly applied here and we need to *prove* it separately. In fact, in the above we have established the one-to-one correspondence between $\overline{\mathrm{FSC}}$ and the equation $\mathbb{E}_{\mathsf{x}\sim\mu}[u(\eta(\mathsf{x}))] = 0$ with $u(\eta) = \theta^*\eta - (\eta - \theta^*)^+$ and $\theta^* = \frac{1}{2}\overline{\mathrm{FSC}}$. That is, a task with the optimum F-score $\overline{\mathrm{FSC}}$ must satisfy the condition $\mathbb{E}_{\mathsf{x}\sim\mu}[u(\eta(\mathsf{x}))] = 0$; and conversely, a task satisfying this condition must have the optimum F-score $\overline{\mathrm{FSC}}$. Therefore, although we cannot find a

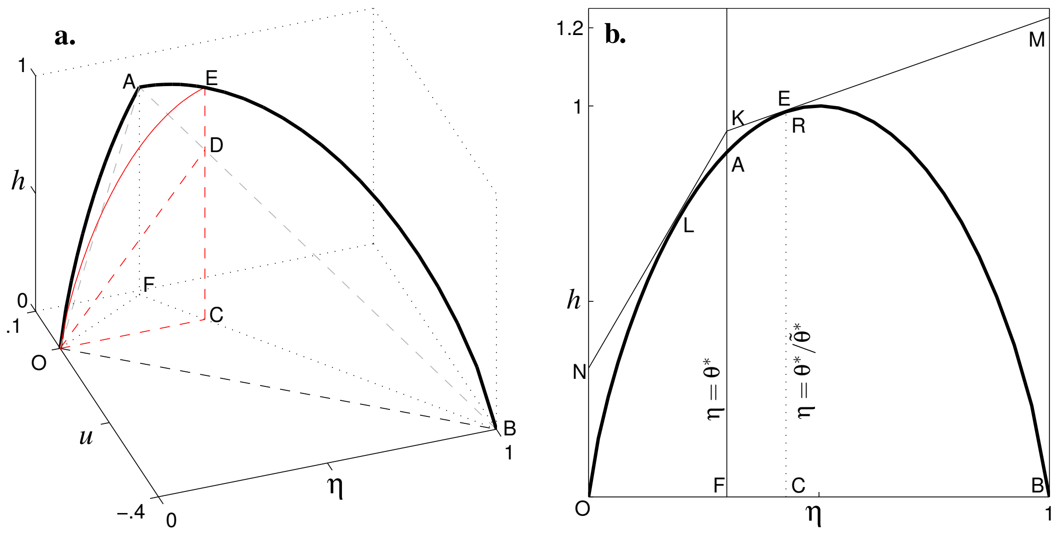


Figure 7: **a.** The curve $\ell = \{[\eta, u(\eta), h(\eta)] \mid \eta \in [0,1]\}$ and its convex hull in the η-u-h space; ODE is the intersection of the plane $u = 0$ and the convex hull, in which we are interested.
b. An geometric interpretation of Lemma 18: draw from a point K on the line $\eta = \theta \leqslant \frac{1}{2}$ two tangent lines of $h(\eta)$, the height of the left tangent point L is less than that of the right one R, provided that $h(\eta)$ is symmetric. To prove the second inequality in Equation (50), here the right tangent point R is set to be on the line $\eta = \tilde{\theta}^{-1}\theta$; so it corresponds to the point E in Figure a.

random variable with expectation $\overline{\text{FSC}}$, we still can apply Theorem 1 to the auxiliary random variable $u(\eta(\mathrm{x}))$ in which $\theta^* = \frac{1}{2}\overline{\text{FSC}}$ serves as a parameter. The resulting tight bounds on $H(\mathrm{y}|\mathrm{x})$ are in fact $0 \leqslant H(\mathrm{y}|\mathrm{x}) \leqslant h(\theta^*/\tilde{\theta}^*)$, which can be rewritten as Equation (50).—*So here we see an implicit use of Scheme 3.*

Analogous to the analysis in page 1050, from Theorem 15 we can easily derive the lower and upper bounds on the maximum F-score by means of conditional entropy. This is best illustrated by Figure 8, where the Shannon conditional entropy is used for $H(\mathrm{y}|\mathrm{x})$. In general, as the function $h(\cdot)$ is symmetric and concave, for any given value of $H(\mathrm{y}|\mathrm{x})$ in the range of $h(\cdot)$, there exists a unique $\beta \in [0, 0.5]$ such that $h(\beta) = h(1-\beta) = H(\mathrm{y}|\mathrm{x})$. So from Figure 8 we know the value of $\overline{\text{FSC}}$ must satisfy $\beta \leqslant \frac{\overline{\text{FSC}}}{2-\overline{\text{FSC}}} \leqslant 1-\beta$, that is, $\frac{2\beta}{1+\beta} \leqslant \overline{\text{FSC}} \leqslant \frac{2-2\beta}{2-\beta}$. An interesting observation of this inequality is that $\overline{\text{FSC}}$ can only assume the value $\frac{2}{3}$ when $\beta = \frac{1}{2}$, that is, when $H(\mathrm{y}|\mathrm{x}) = 1$. This can be explained as follows.

For the Shannon entropy it holds that $1 \geqslant H(\mathrm{y}) \geqslant H(\mathrm{y}|\mathrm{x})$; so $H(\mathrm{y}|\mathrm{x}) = 1$ would imply that $H(\mathrm{y}) = 1$ and $I(\mathrm{x};\mathrm{y}) = H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}) = 0$, where $I(\mathrm{x};\mathrm{y})$ is the mutual information between x and y. From $I(\mathrm{x};\mathrm{y}) = 0$ we know x and y are independent; and from $H(\mathrm{y}) = 1$, $\Pr\{\mathrm{y} = 0\} = \Pr\{\mathrm{y} = 1\} = \frac{1}{2}$. In other words, essentially there is only one classification task whose conditional entropy is 1; and it is actually the most "uncertain" one. It is also the most difficult problem in that the feature vector is completely uninformative and the class label totally random. For such a task, the error probability on any object is always 0.5, regardless of which class is predicted. Hence, $\text{TP} = \text{FP}$ and $\text{TN} = \text{FN}$. It then follows from the second expression of FSC in Equation (17) that $\text{FSC} = \frac{2\times\text{TP}}{3\times\text{TP}+\text{FN}}$. Now, as $\text{FN} \geqslant 0$, the F-score has the maximum value $\frac{2}{3}$; and this happens when $\text{FN} = 0$, which requires that all objects be regarded as positive.

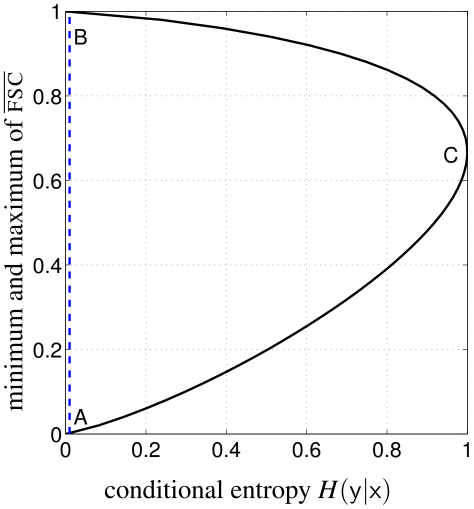


Figure 8: The relationship between $\overline{\text{FSC}}$ and $H(\mathsf{y}|\mathsf{x})$. For given $\overline{\text{FSC}}$, the dashed line AB is the lower bound on $H(\mathsf{y}|\mathsf{x})$; and the solid line ACB the upper bound. Therefore, for given value of $H(\mathsf{y}|\mathsf{x})$, the curve BC is the upper bound of $\overline{\text{FSC}}$; and the curve AC is the lower bound.

For the remainder of the section we will complete the derivation/proof of Theorem 15 by showing that E is the "highest" point on the curve OE, under the assumption that the concave function $h(\cdot)$ is *symmetric*, that is, $h(\eta) = h(\tilde{\eta})$ for any $\eta \in [0,1]$. To simplify the proof, we further assume that $h(\cdot)$ is a differentiable function.[12] The following two lemmas are easy to see. They are there only because they will be referenced several times and so help to shorten the argument that follows.

Lemma 16 *Let $h : [0,1] \to \mathbb{R}$ be a symmetric, differentiable and concave function. Then for $\eta \in [0,\frac{1}{2}]$, $h'(\eta) \geqslant 0$, that is, $h(\eta)$ is monotonically non-decreasing; and for $\eta \in [\frac{1}{2},1]$, $h(\eta)$ is monotonically non-increasing. Moreover, $h'(1-\eta) = -h'(\eta)$ for $\eta \in [0,1]$.*

Lemma 17 *Let $h : [0,1] \to \mathbb{R}$ be a differentiable concave function and $a \in [0,1]$. Then $f(t) := h(t) + h'(t)\cdot(a-t)$ is non-increasing on $[0,a]$; and non-decreasing on $[a,1]$.*

Proof For $s,t \in [0,a]$ satisfying $s < t$, the mean value theorem implies that, for some $u \in (s,t)$, $h(s) - h(t) = h'(u)\cdot(s-t)$. Since h is concave, its derivative h' is monotonically non-increasing. In

[12] This assumption is in fact unnecessary: if $h(\cdot)$ is non-differentiable at some point η_0, we can use any number between its *right derivative* $h'(\eta_0+)$ and *left derivative* $h'(\eta_0-)$ to replace $h'(\eta_0)$.

particular, we have $h'(s) \geqslant h'(u) \geqslant h'(t)$. It thus follows that

$$\begin{aligned} f(s) &= h(s)+h'(s)\cdot(a-s) \\ &= h(t)+h'(u)\cdot(s-t)+h'(s)\cdot(a-s) \\ &\geqslant h(t)+h'(u)\cdot(s-t)+h'(u)\cdot(a-s) \\ &\geqslant h(t)+h'(t)\cdot(a-t) \;=\; f(t)\,. \end{aligned}$$

So $f(t)$ is a non-increasing function on $[0,a]$. Following a similar line, one can prove that $f(t)$ is non-decreasing on the interval $[a,1]$. ■

Lemma 18 *Let $h(\eta)$ be as in Lemma 16. For $\theta \in [0,\frac{1}{2}]$, let $a \leqslant \theta$ and $b \geqslant \theta$ be such that $h(a)+h'(a)(\theta-a) = h(b)-h'(b)(b-\theta)$. Then $h(a) \leqslant h(b)$.*

Proof If $b \leqslant \frac{1}{2}$, by Lemma 16 we at once get $h(a) \leqslant h(b)$. Assume now $b > \frac{1}{2}$, by Lemma 16 we know $h'(a) \geqslant 0$ and $h'(b) \leqslant 0$. Since $\theta \leqslant \frac{1}{2}$, the assumed equality implies

$$\begin{aligned} h(b)-h'(b)(b-\tfrac{1}{2}) &\leqslant h(b)-h'(b)(b-\theta) && \text{as } \theta \leqslant \tfrac{1}{2} \text{ and } h'(b) \leqslant 0 \\ &= h(a)+h'(a)(\theta-a) && \text{the assumed equality} \\ &\leqslant h(a)+h'(a)(\tfrac{1}{2}-a) && \text{as } \theta \leqslant \tfrac{1}{2} \text{ and } h'(a) \geqslant 0 \\ &= h(\tilde{a})-h'(\tilde{a})(\tilde{a}-\tfrac{1}{2})\,. && \text{by Lemma 16} \end{aligned}$$

By $a \leqslant \theta \leqslant \frac{1}{2}$, we have $\tilde{a} \geqslant \frac{1}{2}$. By Lemma 17, $h(b)-h'(b)(b-\frac{1}{2})$ is non-decreasing with respect to $b \in [\frac{1}{2},1]$. So the above inequality implies $b \leqslant \tilde{a}$. As $b > \frac{1}{2}$, by Lemma 16 we know $h(b) \geqslant h(\tilde{a}) = h(a)$. ■

In geometry (see Figure 7-b), $h(a)+h'(a)(\theta-a)$ represents the "height" of the intersection point of the vertical line $\eta=\theta$ and the tangent line of $h(\eta)$ at $\eta = a$. With this in mind, we see that the assumed equality in Lemma 18 means the two tangent lines are drawn from one point on the line $\eta=\theta$. Thus, for a symmetric function $h(\eta)$ and $\theta=\frac{1}{2}$, this would give us the tangent points a and $b=\tilde{a}$ (by symmetry); and so $h(a)=h(b)$. When the line $\eta=\theta$ moves left, that is, $\theta<\frac{1}{2}$, the tangent points a and b also move left. This would result in $h(a) \leqslant h(b)$.

In Figure 7-a, we draw in the plane AFB (i.e., $u=\theta^*-\tilde{\theta}^*\eta$) the tangent line of the curve ℓ at point E, intersecting with line AF at K. From this point K we draw the tangent line of ℓ in the plane OAF ($u=\theta^*\eta$). These are represented in Figure 7-b as their projection on the plane $u=0$. Assume M and N are the intersection points of the two tangent lines with the vertical lines at B and O, respectively. Then, by Lemma 16, the slope of KN is larger than zero; so the h-coordinate of N, N_h, is less than that of the left tangent point L, which, by Lemma 18, is further less than that of the right tangent point E. We thus get $\mathrm{N}_h \leqslant \mathrm{E}_h$.

We now "transfer" the graph in Figure 7-b back to Figure 7-a. Intuitively, one can imagine that Figure 7-b is folded along the line FA; and then put on the broken-line OFB in Figure 7-a (after the obvious lengthening operation). As h is concave, in the η-u-h space the broken line NKM is obviously "above" curve ℓ, that is, the curve OAEB. So the plane KMN is above the convex hull of ℓ. It follows that line NE, the intersection line of the two planes KMN and $u=0$, is above the curve OE (the one in Figure 7-a), the intersection of plane $u=0$ and the convex hull of ℓ. Therefore, the maximum h-coordinate of points on line NE is larger than that of the curve OE. But we have already shown that $\mathrm{N}_h \leqslant \mathrm{E}_h$, so $\mathrm{E}_h = h(\theta^*/\tilde{\theta}^*)$ is larger than the maximum h-coordinate of points on the curve OE. The second inequality in Equation (50) now gets proved.

8. Infomax Is Not Proper for Optimizing Cost-Sensitive Risk or F-Score

In the introduction section, we pointed out that the Infomax principle is consistent with the learning target of minimizing the error rate. The reason is that both Fano's bound and Hellman's bound are monotonically increasing with the conditional entropy; so minimizing the conditional entropy normally results in lower error rate. The same phenomenon is also observed between conditional entropy and balanced error rate (see Figure 6-b). In this sense, Infomax is suitable also for minimizing the balanced error rate.

As for F-score, however, the lower bound on the maximum F-score, $\overline{\text{FSC}}$, is an *increasing* function of conditional entropy, as is depected in Figure 8. This implies a counterintuitive situation. Usually with classification problems, as we decrease the conditional entropy, we can expect the worst case (measured by the maximum F-score) classification scenario to improve. With F-score, however, as we decrease entropy, the worst case F-score gets even worse, decreasing to zero when $H(\mathsf{y}|\mathsf{x})$ tends to zero. As we have briefly mentioned at the end of Section 4, the same non-intuitve scenario is observed for the upper bound on the minimum cost-sensitive risk—see, Figure 4-b. Moreover, from Figure 8 we see that $\overline{\text{FSC}}$ may take any value between 0 and 1 when $H(\mathsf{y}|\mathsf{x})$ tends to zero. This also seems non-intuitive as $H(\mathsf{y}|\mathsf{x}) = 0$ means y is a deterministic function of x, for which the best classifier should have F-score 1.

In this section, we discuss the possible reasons of these inconsistencies through some simple examples. The first example is constructed to illustrate that for any given value of $\overline{\text{FSC}}$, there are classification problems whose maximum F-score is $\overline{\text{FSC}}$; but the conditional entropy $H(\mathsf{y}|\mathsf{x})$ can be arbitrary small.—If the maximum F-score of a problem is small, we would think of it as a difficult task, since no classifier would perform well (as measured by F-score) on it. Intuitively, this means the relationship between x and y is quite uncertain, hence the conditional entropy $H(\mathsf{y}|\mathsf{x})$ should be large. However, our first example shows that is is not necessarily the case.

Example 1 *Here the feature space consists only of two distinct vectors, say, $\mathcal{X} = \{x^{(1)}, x^{(2)}\}$. The joint distribution of* x *and* y *is given by*

$$[\Pr\{x^{(1)},0\},\ \Pr\{x^{(2)},0\},\ \Pr\{x^{(1)},1\},\ \Pr\{x^{(2)},1\}] = [a,\ b,\ 0,\ \pi]\,,$$

where a, b, π are positive numbers with sum 1.

For this task there are four different classifiers which can be encoded naturally as 00, 01, 10 and 11, according to the predicted label on $x^{(1)}$ and $x^{(2)}$. The F-score of these classifiers are $\text{FSC}(00) = \text{FSC}(10) = 0$, $\text{FSC}(01) = \frac{2\pi}{2\pi+b}$ and $\text{FSC}(11) = \frac{2\pi}{2\pi+a+b}$. The computation procedure is detailed in Table 3, where the top-left corner is the joint distribution of x and y. It thus follows that $\overline{\text{FSC}} = \text{FSC}(01) = \frac{2\pi}{2\pi+b}$. The (Shannon) conditional entropy of this task is calculated as

$$H(\mathsf{y}|\mathsf{x}) = H(\mathsf{xy}) - H(\mathsf{x}) = (b+\pi)\log(b+\pi) - b\log b - \pi\log\pi.$$

Now let $\lambda = \frac{b}{\pi}$, then the above $\overline{\text{FSC}}$ and $H(\mathsf{y}|\mathsf{x})$ can be written respectively as

$$\overline{\text{FSC}} = \tfrac{2}{2+\lambda}\,, \quad H(\mathsf{y}|\mathsf{x}) = \pi \cdot [(\lambda+1)\log(\lambda+1) - \lambda\log\lambda] =: \pi \cdot f(\lambda)\,.$$

In the above computation, we have factorized $H(\mathsf{y}|\mathsf{x})$ as the product of two terms. The first term π describes the imbalance between the two classes; the second term $f(\lambda)$ is an increasing function

$\mathsf{y} =$	0	1	$\hat{y}(\mathsf{x})$	$\hat{y}(\mathsf{x})$	$\hat{y}(\mathsf{x})$	$\hat{y}(\mathsf{x})$
$\mathsf{x} = x^{(1)}$	a	0	0	0	1	1
$\mathsf{x} = x^{(2)}$	b	π	0	1	0	1
		TP	0	π	0	π
		TN	$a+b$	a	b	0
		FN	π	0	π	0
		FP	0	b	a	$a+b$
		FSC	0	$\frac{2\pi}{2\pi+b}$	0	$\frac{2\pi}{2\pi+a+b}$

Table 3: Computing the maximum F-score for Example 1

of the ratio $\lambda = \frac{b}{\pi} = \frac{\Pr(0|x_2)}{\Pr(1|x_2)}$, which reflects the uncertainty of the task. In particular, for any fixed $\overline{\text{FSC}}$ or λ, $H(\mathsf{y}|\mathsf{x})$ can be arbitrarily small if we let $\pi \to 0$. This certainly does not hint the involved problem is deterministic.

Next, we will see a more instructive example. It can be seen as the "dual" of Example 1, in that now the value of $\overline{\text{FSC}}$ is shown to be variable (by tuning a free parameter) with the conditional entropy being fixed.

Example 2 *In this example, the feature space consists of three vectors, $\mathcal{X} = \{x^{(1)}, x^{(2)}, x^{(3)}\}$. The marginal distribution of $\mathsf{x} \in \mathcal{X}$ (i.e., the probability measure μ, see Table 1) is given by*

$$[\Pr\{x^{(1)}\},\, \Pr\{x^{(2)}\},\, \Pr\{x^{(3)}\}] = [a,\, b,\, c]\,,$$

where a,b,c are positive numbers with sum 1. The conditional probability of class 1 is denoted as $\eta_i = \Pr\{\mathsf{y} = 1 \mid \mathsf{x} = x^{(i)}\}$ for $i = 1,2,3$; and set to be $[\eta_1, \eta_2, \eta_3] = [0.5, 0, 1]$.

According to Definition 2, the conditional entropy of the above task can be written as

$$H(\mathsf{y}|\mathsf{x}) = a \cdot h(\eta_1) + b \cdot h(\eta_2) + c \cdot h(\eta_3) = a\,,$$

since for the binary entropy function it holds that $h(0.5) = 1$ and $h(0) = h(1) = 0$. There are eight different classifiers though, we actually need only to compute the F-score for two of them to get the maximum F-score. This is because the object $x^{(2)}$ should be classified as negative and $x^{(3)}$ as positive for sure, by any F-score-maximizing classifier—see, the results of Section 6. Only the classification of $x^{(1)}$ is unclear; so we calculate the F-score of the two classifiers 001 and 101. Letting $\eta_1 = 0.5$ in Table 4 we get $\text{FSC}(001) = \frac{2c}{2c+0.5a}$ and $\text{FSC}(101) = \frac{2c+a}{2c+1.5a}$. It is clear that $\text{FSC}(001) \leqslant \text{FSC}(101)$, therefore

$$\overline{\text{FSC}} = \text{FSC}(101) = \frac{2c+a}{2c+1.5a}\,.$$

Although the example is very simple, it does reveal quite a few insights into the notions of conditional entropy and F-score. First of all, for any given value of $H(\mathsf{y}|\mathsf{x}) = a$, one can freely adjust the value of the maximum F-score by tuning the parameter c. In more detail, $\overline{\text{FSC}}$ is an increasing function of c. As $0 \leqslant c \leqslant 1-a$, it is easy to see that $\overline{\text{FSC}}$ ranges from $\frac{2}{3}$ (at $c = 0$) to $\frac{2-a}{2-0.5a}$ (at $c = 1-a$) for the particular problem considered here.

Secondly, the quantity b does not present in the expressions of $H(\mathsf{y}|\mathsf{x})$ and $\overline{\text{FSC}}$. In general, based on the conditional probability $\eta(x) = \Pr\{\mathsf{y} = 1 \mid \mathsf{x} = x\}$, we can classify the objects (feature

$\mathsf{y} =$	0	1	$\hat{y}(\mathsf{x})$	$\hat{y}(\mathsf{x})$
$\mathsf{x} = x^{(1)}$	$\tilde{\eta}_1 a$	$\eta_1 a$	0	1
$\mathsf{x} = x^{(2)}$	b	0	0	0
$\mathsf{x} = x^{(3)}$	0	c	1	1
		TP	c	$\eta_1 a + c$
		TN	$\tilde{\eta}_1 a + b$	b
		FN	$\eta_1 a$	0
		FP	0	$\tilde{\eta}_1 a$
		FSC	$\frac{2c}{2c+\eta_1 a}$	$\frac{2c+2\eta_1 a}{2c+(1+\eta_1)a}$

Table 4: Computing the maximum F-score for Example 2

vectors) in a given task into three catrgories, namely, those belong surely to the positive ($\eta(x) = 1$) or the negative ($\eta(x) = 0$) class and those might be in either class ($0 < \eta(x) < 1$). The proportion of the three types are denoted here as c, b, and a, respectively. Then from the example we see that the conditional entropy $H(\mathsf{y}|\mathsf{x})$ is independent of the "certain" objects (due to the fact that $h(0) = h(1) = 0$). In other words, it measures purely the amount of "uncertainty" for a classification task, which includes two factors, a and $h(\eta_1)$. The former factor represents the "population" of uncertain objects; and the latter represents the (average) degree of uncertainty of these objects.

On the other hand, the maximum F-score depends on the uncertain objects and the positive objects; but not on the negative objects. This is because the definition of F-score, Equation (17), does not take the true negative term, TN, into account. Consequently, classifiers aiming to maximize the F-score would intend to classify objects as positive, as this will increase the true positive and so increase the F-score—it will decrease the true negative at the same time, which however is not captured by F-score. This phenomenon is also reflected in the expression of the optimal classifier, $\hat{y}(x) = [\![\eta(x) > \theta^*]\!]$ (cf. the last paragraph of Section 6). Here the threshold θ^* is determined by the condition $\text{FSC}(\theta^*) = 2\theta^*$, which, as has been explained at the end of Section 6, is below 0.5. So an object would be regarded as positive even if the conditional probability of the positive class is less than half.

Finally, in this example the minimum value of $\overline{\text{FSC}}$ is $\frac{2}{3}$ (the horizontal line through the point C in Figure 8), far from the lower bound. This is due to that we have set $\eta_1 = 0.5$; by using a lower value for η_1, we can in principle hit the lower bound curve AC in Figure 8. For instance, in the next example, we will see a setup with $\overline{\text{FSC}} = 0.625 < \frac{2}{3}$.

8.1 On Information-Theoretic Feature Filtering Methods

Feature selection is a key step when dealing with high-dimensional data; it aims to find useful features and discard others, hence reduces the dimensionality. There are three major categories of feature selection techniques (Guyon and Elisseeff, 2003). *Embedded* methods (Lal et al., 2006) exploit the structure of specific classes of classifiers to guide the feature selection process. *Wrapper* methods (Kohavi and John, 1997) search the space of feature subsets, using the training/validation performance of a particular classifier to measure the utility of a candidate subset. These two are classifier-dependent, with the disadvantage of a considerable computational load, and may produce subsets that are overly specific to the classifiers used. In contrast, *filter* methods (Duch, 2006) separate the classification and feature selection components, and select features using a heuristic scoring criterion that measures how potentially useful a feature or feature subset may be when used in a classifier.

Information-theoretic feature filters use an information measure (usually the mutual information between the selected features and class label) as the scoring criterion. The idea behind is that features showing maximum mutual information with class label are usually most useful for predicting the class label. This is well justified when the (balanced) error rate is concerned, as we have argued earlier. In this section, however, we will illustrate, using a simple example, that feature selection methods based on mutual information may fail to choose the optimal features when the classification performance is measured by F-score or cost-sensitive risk. Here we assume the perfect classifier, that is, the classifier with maximum F-score or minimum cost-sensitive risk, can be derived once the feature subset is determined.

Example 3 *In this example, we assume the objects are described by two features,* x_1 *and* x_2*, both of which take three distinct values. That is,* $\mathsf{x}_i \in \{x_i^{(1)}, x_i^{(2)}, x_i^{(3)}\}$ *for* $i = 1,2$*. The joint distribution of* x_1*,* x_2*, and* y *are set to be*

$\mathsf{y} = 0$	$\mathsf{x}_2 = x_2^{(1)}$	$x_2^{(2)}$	$x_2^{(3)}$
$\mathsf{x}_1 = x_1^{(1)}$	0.33	0	0
$x_1^{(2)}$	0.174	0.1	0
$x_1^{(3)}$	0	0	0

$\mathsf{y} = 1$	$x_2^{(1)}$	$x_2^{(2)}$	$x_2^{(3)}$
$x_1^{(1)}$	0.15	0	0.18
$x_1^{(2)}$	0	0	0
$x_1^{(3)}$	0.066	0	0

The target here is to select one feature to predict the class label.

As we can see here, by selecting either feature we are actually comparing two different problems that are described respectively by the distribution of the pairs $(\mathsf{x}_1, \mathsf{y})$ and $(\mathsf{x}_2, \mathsf{y})$. So we compute the two distributions from the given joint distribution of $(\mathsf{x}_1, \mathsf{x}_2, \mathsf{y})$, which gives us

$$\Pr\{\mathsf{x}_1, \mathsf{y}\} = \begin{bmatrix} 0.33 & 0.33 \\ 0.274 & 0 \\ 0 & 0.066 \end{bmatrix}; \quad \Pr\{\mathsf{x}_2, \mathsf{y}\} = \begin{bmatrix} 0.504 & 0.216 \\ 0.1 & 0 \\ 0 & 0.18 \end{bmatrix}. \tag{51}$$

Both $\Pr\{\mathsf{x}_1, \mathsf{y}\}$ and $\Pr\{\mathsf{x}_2, \mathsf{y}\}$ are structurally similar to the one in Example 2, with the parameters $[a, b, c; \eta_1] = [0.66, 0.274, 0.066; 0.5]$ and $[0.72, 0.1, 0.18; 0.3]$ respectively. So we can reuse the computation there to obtain the Shannon conditional entropy

$$\begin{aligned} H(\mathsf{y}|\mathsf{x}_1) &= a \cdot h(\eta_1) = 0.66 \cdot h(0.5) = 0.66\,, \\ H(\mathsf{y}|\mathsf{x}_2) &= a \cdot h(\eta_1) = 0.72 \cdot h(0.3) = 0.6345\,. \end{aligned}$$

Thus, according to the Infomax principle, the second feature x_2 should be selected as the class label predictor.

However, the maximum F-score of the two problems tells us a different story. For $(\mathsf{x}_1, \mathsf{y})$, we already have (see Example 2)

$$\overline{\mathrm{FSC}} = \mathrm{FSC}(101) = \frac{2c + a}{2c + 1.5a} = \frac{2 \times 0.066 + 0.66}{2 \times 0.066 + 1.5 \times 0.66} = 0.7059\,. \tag{52}$$

For $(\mathsf{x}_2, \mathsf{y})$, we need to compare the F-score of the classifiers 001 and 101. It follows from Table 4 that

$$\begin{aligned} \text{FSC}(001) &= \frac{2c}{2c+\eta_1 a} = \frac{2\times 0.18}{2\times 0.18+0.3\times 0.72} = 0.625\,, \\ \text{FSC}(101) &= \frac{2c+2\eta_1 a}{2c+(1+\eta_1)a} = \frac{2\times 0.18+0.6\times 0.72}{2\times 0.18+1.3\times 0.72} = 0.6111\,. \end{aligned}$$

Thus, $\overline{\text{FSC}} = \text{FSC}(001) = 0.625$, which is less than that of $(\mathsf{x}_1, \mathsf{y})$ at 0.7059 in Equation (52). This reveals that while the feature x_2 is selected by Infomax, it is in fact possible to design a better classifier (as measured by F-score) using the first feature x_1. The constructed problem shows that to minimize error rate and balanced error rate we should pick feature x_2; whereas to minimize cost-sensitive risk we should pick a different feature, x_1.

We now examine the minimum cost-sensitive risk of the two problems $(\mathsf{x}_1, \mathsf{y})$ and $(\mathsf{x}_2, \mathsf{y})$. Assume the cost of a false negative is $c_1 = 2.5$ and that of a false positive is $c_0 = 0.625$. If the feature x_1 is used, then, by Equation (51), the optimal classifier is 101 (which produces a false positive on $x^{(1)}$ with probability 0.33); and the corresponding minimum cost-sensitive risk is $\underline{\text{CSR}} = \text{CSR}(101) = 0.625\times 0.33 = 0.2063$. When the feature x_2 is selected, we compute $\text{CSR}(001) = 2.5\times 0.216 = 0.54$ and $\text{CSR}(101) = 0.625\times 0.504 = 0.315$. Thus, $\underline{\text{CSR}} = \text{CSR}(101) = 0.315$. It thus follows that choosing the feature x_1 would (potentially) obtain a lower cost-sensitive risk 0.2063, contradicting the selection suggested by Infomax.

On the other hand, the minimum (balanced) error rate of the problem $(\mathsf{x}_1, \mathsf{y})$ is

$$\underline{\text{ERR}} = 0.33\,, \quad \underline{\text{BER}} = \tfrac{0.33}{2\times(0.33+0.274)} = 0.2732\,.$$

For both criteria, the optimal classifier is 101. For the problem $(\mathsf{x}_2, \mathsf{y})$, we have

$$\underline{\text{ERR}} = 0.216\,, \quad \underline{\text{BER}} = \tfrac{0.216}{2\times(0.216+0.18)} = 0.2727\,,$$

with both minima obtained at the classifier 001. Therefore, selecting x_2 will do better than x_1 as to minimize the (balanced) error rate, in agreement with Infomax.

8.2 Towards Proper Information Measures for Cost-Sensitive Risk

In the preceding section, we constructed an example demonstrating that Shannon's mutual information is generally not a proper criterion for feature selection when the cost-sensitive risk or F-score is concerned. A natural question one would immediately raise is *what is the proper information measure for the two criteria then?* So far, this problem is not completely solved; and we only have partial solution.

In Section 4 we derived the tight lower and upper bounds on the conditional entropy $H(\mathsf{y}|\mathsf{x})$ in terms of the minimum cost-sensitive risk $\underline{\text{CSR}}$ (see Figure 4-b); and noticed that the upper bound is not an increasing function of $\underline{\text{CSR}}$. On the other hand, as we have emphasized several times, it is the monotonicity of Fano's and Hellman's bounds that justifies the Infomax principle. This motivates us to construct concave functions $h(\eta)$ such that the conditional entropy $H(\mathsf{y}|\mathsf{x})$ as defined in Definition 2 has lower and upper bounds that are monotonically increasing with respect to $\underline{\text{CSR}}$.

As we can see in Figure 4, the curve OCDF plays an important role in determining the lower and upper bounds on $H(\mathsf{y}|\mathsf{x})$. It is obtained from the graph of the function $h(\eta)$ by a simple piecewise

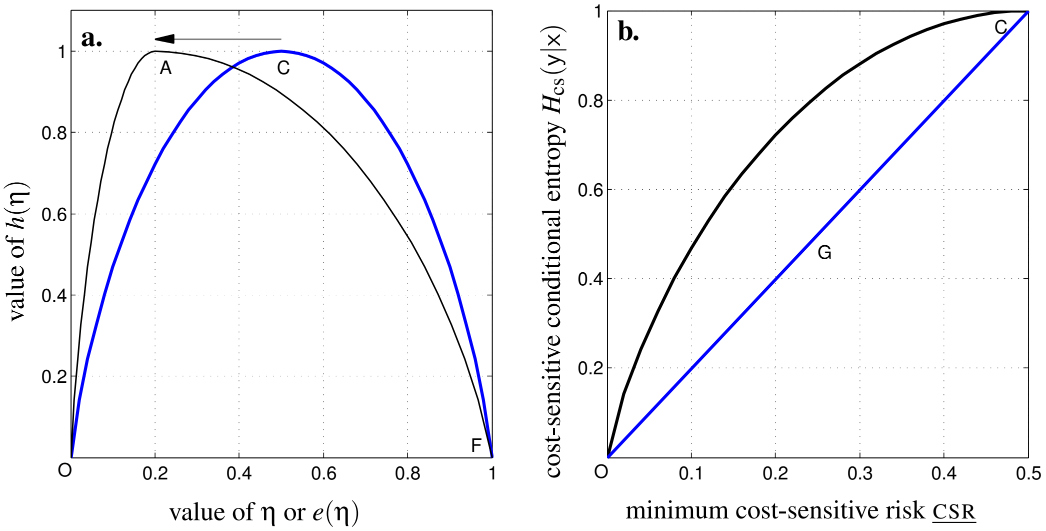


Figure 9: **a.** To get a symmetric concave curve OCF after the transform as indicated by the arrow AC in Figure 4-a, we apply an inverse transform to the target curve OCF. That is, we move the peak point C back to the point A with the η-coordinate $\frac{1}{2c_1}$, yielding the curve OAF that represents the function $h(\eta)$ in the definition of conditional entropy—see Equation (21). **b.** The lower (the line OGC) and upper (the curve OC) bounds on $H_{\text{cs}}(\mathsf{y}|\mathsf{x})$ are obtained from the curve OCF in Figure 9-a using the same transform as in Figure 4. Both bounds are now monotonically increasing functions of $\underline{\text{CSR}}$. This can be contrasted with the standard conditional entropy bounds in Figure 4-b.

linear transformation on the input η. Its left part OC determines the lower bound; and its right part CDF (after flipping along the central vertical line) corresponds to the upper bound. Thus, if we can construct a concave function $h(\eta)$ that makes the curve OCDF symmetric (for example, coincide with the curve OABF), then the lower and upper bounds on $H(\mathsf{y}|\mathsf{x})$ would be very similar to Fano's and Hellman's, respectively. As shown in Figure 9, this can be done by applying a piecewise linear transform to the input variable of a symmetric concave function, so that the peak point C on its graph is moved left to the point A with the first coordinate $\frac{1}{2c_1}$.

Denote by $g(\eta)$, $\eta \in [0,1]$, the function corresponding to the curve OCF in Figure 9-a. Then the curve OAF is described by the function[13]

$$h(\eta) = \begin{cases} g(c_1\eta) & \text{if } \eta \in [0, \frac{1}{2c_1}), \\ g(1 - c_0\tilde{\eta}) & \text{if } \eta \in [\frac{1}{2c_1}, 1], \end{cases} \tag{53}$$

[13] This can be easily verified by checking the value of $h(\eta)$ at $\eta = 0, \frac{1}{2c_1}, 1$, which should be $g(0)$, $g(\frac{1}{2})$ and $g(1)$, respectively.

where the costs c_0 and c_1 satisfy the conditions $c_1 \geqslant c_0$ and $c_0^{-1} + c_1^{-1} = 2$. In particular, when $g(\eta) = h_{\text{bin}}(\eta) = -\eta \log \eta - (1-\eta)\log(1-\eta)$ (Shannon), we have

$$h(\eta) = \begin{cases} -(c_1\eta)\log(c_1\eta) - (1-c_1\eta)\log(1-c_1\eta) = h_{\text{bin}}(c_1\eta) & \text{if } \eta \in [0, \frac{1}{2c_1}), \\ -(c_0\tilde{\eta})\log(c_0\tilde{\eta}) - (1-c_0\tilde{\eta})\log(1-c_0\tilde{\eta}) = h_{\text{bin}}(c_0\tilde{\eta}) & \text{if } \eta \in [\frac{1}{2c_1}, 1]. \end{cases} \tag{54}$$

Substituting the above expression into Equation (21), we get a new definition of conditional entropy which we call the *cost-sensitive conditional entropy* and denote as $H_{\text{cs}}(\mathsf{y}|\mathsf{x})$. That is, $H_{\text{cs}}(\mathsf{y}|\mathsf{x}) := \mathbb{E}_{\mathsf{x}\sim\mu}[h(\eta(\mathsf{x}))]$, with $h(\eta)$ defined as above.

We now compute the value of $H_{\text{cs}}(\mathsf{y}|\mathsf{x})$ for the two classification tasks $(\mathsf{x}_1, \mathsf{y})$ and $(\mathsf{x}_2, \mathsf{y})$ as defined in Example 3, under the settings of $c_0 = 0.625$, $c_1 = 0.25$. By Equation (51), we know that the marginal distribution of x_1 is $\Pr\{\mathsf{x}_1 = x_1^{(1,2,3)}\} = [0.66, 0.274, 0.066]$; the conditional probability of the positive class is $\eta_{1,2,3} = \Pr\{\mathsf{y} = 1 \mid \mathsf{x}_1 = x_1^{(1,2,3)}\} = [0.5, 0, 1]$. By Equation (54), we have $h(\eta_1) = h_{\text{bin}}(c_0\tilde{\eta}_1) = h_{\text{bin}}(0.625 \cdot 0.5) = 0.6211$; $h(\eta_2) = h_{\text{bin}}(c_1\eta_2) = h_{\text{bin}}(0) = 0$; and $h(\eta_3) = h_{\text{bin}}(c_0\tilde{\eta}_3) = h_{\text{bin}}(0) = 0$. Thus,

$$H_{\text{cs}}(\mathsf{y}|\mathsf{x}_1) = \Pr\{\mathsf{x}_1 = x_1^{(1)}\} \cdot h(\eta_1) = 0.66 \times 0.6211 = 0.4099\,.$$

Similarly, for the feature x_2, we have $\eta_{1,2,3} = \Pr\{\mathsf{y} = 1 \mid \mathsf{x}_2 = x_2^{(1,2,3)}\} = [0.3, 0, 1]$ and $\Pr\{\mathsf{x}_2 = x_2^{(1,2,3)}\} = [0.72, 0.1, 0.18]$. Thus, $h(\eta_1) = h_{\text{bin}}(c_0\tilde{\eta}_1) = h_{\text{bin}}(0.625 \cdot 0.7) = 0.6853$ and $h(\eta_2) = h(\eta_3) = 0$. We thus get

$$H_{\text{cs}}(\mathsf{y}|\mathsf{x}_2) = \Pr\{\mathsf{x}_2 = x_2^{(1)}\} \cdot h(\eta_1) = 0.72 \times 0.6853 = 0.4934\,.$$

Since $H_{\text{cs}}(\mathsf{y}|\mathsf{x}_1) < H_{\text{cs}}(\mathsf{y}|\mathsf{x}_2)$, the feature x_1 would be selected according to the cost-sensitive conditional entropy. This coincides with the decision we previously obtained by directly comparing the value of CSR.

In conclusion, Shannon's mutual information or conditional entropy is not a proper surrogate learning objective in dealing with a cost-sensitive situation or when the subsequent classification process is assessed by the metric of F-score. Conversely, we have proven the positive result that Shannon's information is appropriate for balanced error rate. For cost-sensitive risk minimization problems, we suggest to use a cost-sensitive variant of normal symmetric conditional entropies as defined by Equation (53). As far as the authors know, this definition of conditional entropy has not been studied in the context of feature selection. The work by Elkan (2001) might be the closest to ours, where he investigated the possibility of adapting a given learning algorithm to the cost-sensitive situation by simply adjusting the prior probability of each class (whereas here we intend to change the posterior probabilities). For F-score maximization, we have not found a proper information measure so far.

9. Conclusion and Future Work

Inspired by the widespread use of Fano's inequality in machine learning—in particular, in feature selection, the paper has extended Fano's and Hellman's bounds (on error probability) to the bounds on other commonly used criteria including balanced error rate, F-score and cost-sensitive risk. To this end, we developed a general geometric method which enables us to derive the tight bounds on the above mentioned criteria using a general definition of conditional entropy (see Definition

2), in a uniform way. These bounds are presented in three main theorems of the paper: Theorems 7, 11 and 15. Our work extends previous knowledge on the relationship between classification performance criteria and conditional entropy (Ben-Bassat, 1978; Golic, 1987; Feder and Merhav, 1994; Erdogmus and Principe, 2004).

The advantage of the proposed geometric approach is clear: it provides a visible and intuitive insight into the relationship between the concerned criteria and information measures. Moreover, defining the conditional entropy through a general concave function $h(\eta)$ in fact gives us much more than what we have stated so far. For example, let $h(\eta) = \min\{\eta, \tilde{\eta}\}$ in Theorem 15, we immediately get the bounds on the Bayes error rate in terms of maximum F-score: $0 \leqslant \underline{\mathrm{ERR}} \leqslant \min\{\frac{\overline{\mathrm{FSC}}}{2-\overline{\mathrm{FSC}}}, 1 - \frac{\overline{\mathrm{FSC}}}{2-\overline{\mathrm{FSC}}}\}$.

When deriving the bounds on the maximum F-score and the minimum cost-sensitive risk, some new findings were noticed, which, interestingly, might be of more interest than the bounds themselves. Firstly, as a by-product of the bounds on the maximum F-score, $\overline{\mathrm{FSC}}$, in Section 6 we proved that the optimal classifiers for maximizing the F-score have the form $\hat{y}(x) = [\![\eta(x) > \theta]\!]$. This property is called *the probability thresholding principle for binary classifications* by Lewis (1995); and has been proved by Lewis (1995) and Jansche (2007) independently for finite input spaces $\mathcal{X}$. Here we presented a proof for the general case where $\mathcal{X}$ is an arbitrary set, which, to the best of our knowledge, is novel.

The most important new finding in the paper is that the Infomax principle based on standard information measures could be misleading when F-score or cost-sensitive risk is used as the performance measure. We illustrated this by analytical argument and a simple example in the field of feature selection. For cost-sensitive risk, we proposed an alternative information measure, whose usefulness is justified by the same example (and by the monotonicity of the resulting bounds, see Figure 9-b). To summarize,

> *Shannon's conditional entropy is* ***not*** *a proper criterion for feature selection when the subsequent classification process is measured by F-score or cost-sensitive risk. Instead, we suggest to use a cost-sensitive variant as defined by Equation* (53).

A corresponding measure for F-score is left as an open problem for further research. This is a challenging question due to the fact that F-score is defined on the *whole* object space, whereas information measures are usually defined through the conditional probabilities on *single* objects, $\Pr\{\mathrm{y} = 1 \mid \mathrm{x} = x\}$. To find a proper information measure for the F-score maximization problem is a research topic in our group to be pursued in the future. As the presented bounds hold only for binary problems, extending them to the multi-class problem is also a topic of interest in the group.

We finish the paper with an important remark. The paper is theory-oriented; it is concerned with *problems*, not with *classifiers* or *algorithms*. More precisely, while the performance of a classifier could be measured by error rate, balanced error rate, F-score or cost-sensitive risk, their optimum value over *all* classifiers can be seen as different difficulty measures of the concerned problem. On the other hand, the conditional entropy $H(\mathrm{y}|\mathrm{x})$ measures the amount of uncertainty about the class label remaining after we have observed the object. Thus, it can also be seen as a difficulty measure of classification tasks. From this perspective, in this paper we are examining the relationship between two different types of difficulty measures of classification problems. Our main finding is that Shannon's conditional entropy as a difficulty measure is inconsistent with the maximum F-score and the minimum cost-sensitive risk. This fact has serious implications in the field of feature selection, as we have discussed in Section 8.

Acknowledgments

This research was conducted with support from the UK Engineering and Physical Sciences Research Council. Ming-Jie Zhao and Narayanan Edakunni were supported by grants EP/F023855/1, EP/G000662/1, and EP/G013500/1, and Adam Pocock was supported on an EPSRC studentship from EP/G000662/1.

Appendix A. Bounds on the Bayes Error Rate: the Multi-Class Case

We have already mentioned in the introduction section the main work in the literature that are related to ours. These are all about bounding the Bayes error rate by means of different conditional entropies. This section briefly introduces a unifying derivation of these bounds, based on the work of Tebbe and Dwyer III (1968), Ben-Bassat (1978) and Golic (1987). As Fano's bound and others' actually hold for the *multi-class* problem, we need to extend the notations introduced in Section 2 to catch up with the multi-class case.[14] These new notations are used only in this section and not listed in Table 1.

Assume there are m classes which are labeled by the integers 1 to m. Then a classifier can be written as a function $\hat{y}(x)$ on $\mathcal{X}$ that takes values in the set $\{1,\ldots,m\}$. Similar to the binary case, we decompose the joint distribution of (x,y) as the product of the marginal distribution of x and the conditional distribution of y given x. As such, the definition of $\mu(A)$ is unchanged, see Equation (2). But the quantity $\eta(x)$ is now replaced by an m-dimensional vector $\boldsymbol{\eta}(x) = [\eta_1(x),\ldots,\eta_m(x)]$, with

$$\eta_y(x) := \Pr\{\mathsf{y} = y \mid \mathsf{x} = x\}, \qquad \forall x \in \mathcal{X},\ \forall y \in \{1,\ldots,m\}.$$

By the above definition we see that the elements of $\boldsymbol{\eta}(x)$ are non-negative and sum to 1. Such vectors are called *probability vectors* in statistics. We shall denote by $\mathcal{P}_m$ the set of probability vectors of dimension m, which is also known as the *probability simplex* in $\mathbb{R}^m$:

$$\mathcal{P}_m := \{\boldsymbol{\eta} \in \mathbb{R}^m \mid \eta_y \geqslant 0 \text{ for all } y = 1,\ldots,m; \text{ and } \textstyle\sum_{y=1}^m \eta_y = 1\}.$$

In terms of μ and $\boldsymbol{\eta}$, the joint distribution of (x,y) can be written as

$$\Pr\{\mathsf{x} \in A, \mathsf{y} = y\} = \textstyle\int_A \eta_y(x)\mathrm{d}\mu, \quad \forall A \subseteq \mathcal{X} \text{ measurable}, \forall y \in \{1,\ldots,m\}.$$

Letting $A = \mathcal{X}$ in the above formula, we get the (prior) probability of each class,

$$\pi_y := \Pr\{\mathsf{y} = y\} = \textstyle\int_{\mathcal{X}} \eta_y(x)\mathrm{d}\mu, \qquad \forall y \in \{1,\ldots,m\}.$$

Note that the vector of class probabilities, $\boldsymbol{\pi} := [\pi_1,\ldots,\pi_m]$, is also a probability vector.

As before, we call the pair $(\mu,\boldsymbol{\eta})$ a (classification) task, whose conditional entropy is defined as follows.

Definition 19 *Let $h : \mathcal{P}_m \to \mathbb{R}$ be a symmetric concave function—the word "symmetric" refers to that, for any $\boldsymbol{\eta} = [\eta_1,\ldots,\eta_m] \in \mathcal{P}_m$ and any permutation $(i_1,\ldots,i_m)$ of $\{1,\ldots,m\}$, it holds that $h(\eta_1,\ldots,\eta_m) = h(\eta_{i_1},\ldots,\eta_{i_m})$. The conditional entropy of a task $(\mu,\boldsymbol{\eta})$ is*

$$H(\mathsf{y}|\mathsf{x}) := \textstyle\int_{\mathcal{X}} h(\boldsymbol{\eta}(x))\mathrm{d}\mu = \mathbb{E}_{\mathsf{x}\sim\mu}[h(\boldsymbol{\eta}(\mathsf{x}))].$$

[14] Only in this section we discuss the multi-class problem; the rest of the paper is devoted to the binary case.

In particular, letting $h(\boldsymbol{\eta}) = -\sum_{i=1}^{m} \eta_i \log \eta_i$, we get the Shannon conditional entropy.

For any classification task $(\mu, \boldsymbol{\eta})$, the (expected) error rate of a given classifier $\hat{y} : \mathcal{X} \to \{1, \ldots, m\}$ can be computed as

$$\begin{aligned} \text{ERR} &= \Pr\{\mathsf{y} \neq \hat{y}(\mathsf{x})\} \\ &= 1 - \Pr\{\mathsf{y} = \hat{y}(\mathsf{x})\} \\ &= 1 - \textstyle\sum_{y=1}^{m} \Pr\{\mathsf{y} = y, \hat{y}(\mathsf{x}) = y\} \\ &= 1 - \textstyle\sum_{y=1}^{m} \int_{\mathcal{X}_y} \eta_y(x) \mathrm{d}\mu , \end{aligned}$$

where $\mathcal{X}_y$ are subsets of the feature space $\mathcal{X}$ determined by the classifier $\hat{y}(x)$ via $\mathcal{X}_y := \{x \in \mathcal{X} \mid \hat{y}(x) = y\}$, for $y = 1, \ldots, m$. Therefore, the error rate is minimized when $\eta_y(x)$ is the maximum element in the whole vector $\boldsymbol{\eta}(x)$ on the set $\mathcal{X}_y$. That is,

$$\underline{\text{ERR}} = 1 - \textstyle\sum_{y=1}^{m} \int_{\mathcal{X}_y} \max\{\boldsymbol{\eta}(x)\} \mathrm{d}\mu = 1 - \int_{\mathcal{X}} \max\{\boldsymbol{\eta}(x)\} \mathrm{d}\mu = \mathbb{E}_{\mathsf{x}\sim\mu}[1 - \max\{\boldsymbol{\eta}(\mathsf{x})\}] ,$$

where, for any vector $\boldsymbol{\eta}$, $\max\{\boldsymbol{\eta}\}$ denotes its maximum entry.

Following the line presented in Section 3, here we need to examine the range of the point $[\underline{\text{ERR}}, H(\mathsf{y}|\mathsf{x})] = \mathbb{E}_{\mathsf{x}\sim\mu}[e(\boldsymbol{\eta}(\mathsf{x})), h(\boldsymbol{\eta}(\mathsf{x}))]$ in the error rate versus conditional entropy plane. Here the function $e(\boldsymbol{\eta})$ is defined as $e(\boldsymbol{\eta}) = 1 - \max\{\boldsymbol{\eta}\}$; for the binary case, this becomes $e(\eta) = \min\{\eta, \tilde{\eta}\}$. So the problem is now reduced to finding the convex hull of the set $\{[e(\boldsymbol{\eta}), h(\boldsymbol{\eta})] \mid \boldsymbol{\eta} \in \mathcal{P}_m\}$, which further amounts to computing the extreme values of $h(\boldsymbol{\eta})$ given that $e(\boldsymbol{\eta})$ is fixed.

Lemma 20 *Let $h : \mathcal{P}_m \to \mathbb{R}$ be a symmetric concave function. Let $\boldsymbol{\eta} \in \mathcal{P}_m$ be such that $e(\boldsymbol{\eta}) = r$. Then $h(\boldsymbol{\eta})$ is maximized when one element of $\boldsymbol{\eta}$ equals to $1 - r$ and the others are all $\frac{r}{m-1}$; and it is minimized when all entries of $\boldsymbol{\eta}$ are either $1 - r$ or 0, except one whose value is determined by the condition that $\boldsymbol{\eta}$ has element sum 1.*

In particular, for the function $h(\boldsymbol{\eta}) = -\sum_{i=1}^{m} \eta_i \log \eta_i$, we have

$$\begin{aligned} h_{\max}(r) &:= \max_{\boldsymbol{\eta}\in\mathcal{P}_m, e(\boldsymbol{\eta})=r} h(\boldsymbol{\eta}) = -(1-r)\cdot\log(1-r) - r\cdot\log\left(\tfrac{r}{m-1}\right) , \\ h_{\min}(r) &:= \min_{\boldsymbol{\eta}\in\mathcal{P}_m, e(\boldsymbol{\eta})=r} h(\boldsymbol{\eta}) = -k\cdot(1-r)\cdot\log(1-r) - \beta\cdot\log\beta , \end{aligned}$$

where k is the maximum integer such that $k \cdot (1 - r) \leqslant 1$ and $\beta = 1 - k \cdot (1 - r)$.

The graphs of $h_{\max}(r)$ and $h_{\min}(r)$ are plotted in Figure 10 for the case of $m = 5$ classes. Notice that for $m = 5$, $\max\{\boldsymbol{\eta}\} \geqslant 0.2$, so the range of $r = e(\boldsymbol{\eta}) = 1 - \max\{\boldsymbol{\eta}\}$ is $[0, 0.8]$. From the figure we see that while $h_{\max}(r)$ is a smooth function, the curve of $h_{\min}(r)$ consists of $m - 1 = 4$ segments, connected by the endpoints A, B, and C. The r-coordinate of these endpoints are determined by the condition $k \cdot (1 - r) = 1$, for $k = 1, \ldots, m$ corresponding to the points O, A, ..., D, respectively.

By definition, the region between the curves of $h_{\max}(r)$ and $h_{\min}(r)$ is exactly the set $\{[e(\boldsymbol{\eta}), h(\boldsymbol{\eta})] \mid \boldsymbol{\eta} \in \mathcal{P}_m\}$. Moreover, it can be proven that $h_{\max}(r)$ is concave and $h_{\min}(r)$ concave within each segment; their graph also shows this. Therefore, the convex hull of $\{[e(\boldsymbol{\eta}), h(\boldsymbol{\eta})] \mid \boldsymbol{\eta} \in \mathcal{P}_m\}$ is bounded by the curve OD and the broken line OABCD, which represent the *tight* lower (Fano) and upper (Tebbe) bounds on the Bayes error rate, $\underline{\text{ERR}}$, in terms of the conditional entropy, $H(\mathsf{y}|\mathsf{x})$. Furthermore, the broken line OABCD forms a convex function, so it is lower bounded by its most left segment OA. This actually gives us the Hellman inequality, $\underline{\text{ERR}} \leqslant \frac{1}{2} H(\mathsf{y}|\mathsf{x})$.

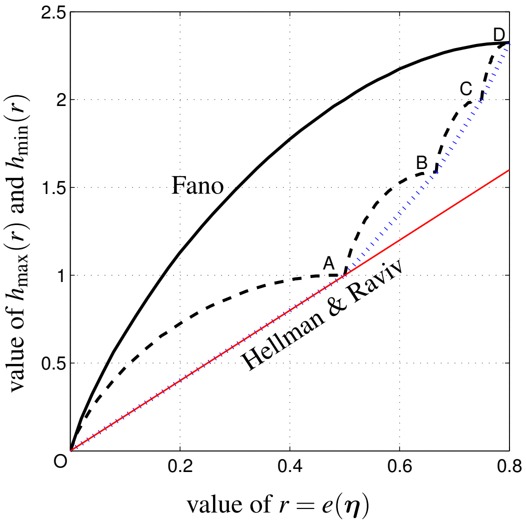


Figure 10: Graphs of the functions $h_{\max}(r)$ (solid line) and $h_{\min}(r)$ (dashed line) for the $m = 5$ class problem. From the two curves we obtain the tight lower bound (Fano), the tight upper bound (Tebbe, the dotted broken line OABCD) on the Bayes error rate, and the Hellman bound.

The above derivation is simple and elegant. However, it can not be directly applied to the case of balanced error rate (for multi-class problems). The main difficulty is that now we have an extra condition on the posterior probabilities $\boldsymbol{\eta}(x)$, namely, its integral over the space $\mathcal{X}$ should be equal to the class probability $\boldsymbol{\pi}$. So the problem of bounding balanced error rate actually amounts to

$$\begin{array}{ll} \text{min or max} & H(\mathrm{y}|\mathrm{x}) = \int_{\mathcal{X}} h(\boldsymbol{\eta}(x))\mathrm{d}\mu \\ \text{subject to} & \int_{\mathcal{X}} \boldsymbol{\eta}(x)\mathrm{d}\mu = \boldsymbol{\pi}\,, \\ & \frac{1}{m}\int_{\mathcal{X}} \min_{y=1,\ldots,m}\{\pi_y^{-1}\eta_y(x)\}\mathrm{d}\mu = \underline{\mathrm{BER}}\,, \end{array}$$

which is a very difficult optimization problem, even for the case of $m = 2$.

In this paper, we restrict ourselves to the binary case (so the vector $\boldsymbol{\eta}(x)$ can be represented by a scalar $\eta(x)$, as is shown in the paper) and consider the reachable region of the 3-dimensional point $[\eta, r(\eta), h(\eta)]$—where $r(\eta) = \min\{\pi^{-1}\eta, \tilde{\pi}^{-1}\tilde{\eta}\}$, rather than the 2-dimensional point $[r(\eta), h(\eta)]$. As such, we avoid solving the above optimization problem, which is extremely hard as we can tell.

Finally, we should also mention that extending our method to the multi-class case is also difficult, if not impossible. Because that will involve the reachable region of the high $(m+2)$ dimensional point $[\boldsymbol{\eta}, r(\boldsymbol{\eta}), h(\boldsymbol{\eta})]$ with $r(\boldsymbol{\eta}) := \min_{y=1,\ldots,m}\{\pi_y^{-1}\eta_y\}$.

Appendix B. Mathematical Foundation

In this section we discuss three fundamental points which support the content of the paper and provide a mathematically more rigorous foundation.[15]

[15] We thank two anonymous reviewers for pointing out these imperfection to us.

1. The concept of a *true label* can be very messy in the commonly used setup for pattern classification which involves only the jointly-distributed random variables (x,y).

2. The convex hull of the curves ℓ in the paper are obtained heuristically, a more rigorous treatement is needed.

3. The geometric arguments in this paper are based on the proposition: *the expectation of a random vector* $\mathsf{u} \in \mathbb{R}^n$ *lies in the convex hull of its range*. While this is correct in intuition, it needs a mathematical proof.

As the above three problems make sense not only for the specific topic studied here, but also from a broader viewpoint of pattern classification and probability theory, we place the discussion of them in a separate section.

B.1 A Mathematical Definition of True Class Labels

To describe the classification problem in a mathematical framework, many textbooks on machine learning start off with a joint distribution of the feature vector x and the class label y. For example, in Devroye et al. (1996, Chapter 1) the authors wrote:

> ...More formally, an *observation* is a d-dimensional vector x. The unknown nature of the observation is called a *class*. It is denoted by y ...

and in Chapter 2 they continued with:

> ...The random pair (x,y) may be described in a variety of ways: for example, it is defined by the pair (μ,η), where μ is the probability measure for x and η is the regression of y on x. ...

While this treatment has the advantages of simplicity and ease to understand, it fails to capture some natural notions rised in real applications such as the *true label* of an object. It is also not a uniform framework in the sense that, whenever a new feature is added, we have to extend the vector x by one component and redefine the joint distribution.

We now introduce an alternative framework that allows for a clear definition of true class labels. The key idea is to distinguish between an object and the features describing it. For this, we denote by Ω the set of all objects ω[16] in the considered problem. In medical diagnosis, for example, this could be (the set of) all people in a country or an area. We can then define a σ-algebra $\mathscr{F}$ of subsets of Ω and a probability measure P on $\mathscr{F}$, yielding a probability space $(\Omega,\mathscr{F},P)$. Note that, here $(\Omega,\mathscr{F},P)$ serves only as a uniform base for discussion; and the concrete definition of $\mathscr{F}$ and P are not important.

According to the problem at hand, the set Ω is often naturally divided into several (measurable) subsets that are *pairwise disjoint*, say $\Omega = \Omega_0 \cup \Omega_1$. For example, Ω_1 may represent those people who are affected by a certain disease; and Ω_0 is the set of the others. This can be conveniently described by a function $\mathsf{y} : \Omega \to \{0,1\}$ that sends each $\omega \in \Omega_i$ to the value i $(i = 0,1)$. Since the subsets Ω_i are pairwise disjoint and their union equals to Ω, the function y is well defined. *We can now define the value of* y *at a particular* ω *as the* true class label *of the object* ω.

[16] More rigorously speaking, here ω is in fact the "name" of the object it represented.

One main goal in pattern classification is to predict the true class label $\mathsf{y}(\omega)$, for which we need to make some measurements on the object ω—obviously, the object *itself* cannot be used as a predictor here. The measuring procedure is also described as a *measurable* function $\mathsf{x} : \Omega \to \mathcal{X}$. For example, $\mathsf{x}(\omega)$ might be the vector of heart rate, body temperature and blood pressure of the person ω. The joint distribution of x and y is induced from the probability measure P, as follows: for any $A \subseteq \mathcal{X}$ measurable and $y \in \{0,1\}$,

$$\Pr\{\mathsf{x} \in A, \mathsf{y} = y\} := P(\{\omega \in \Omega \mid \mathsf{x}(\omega) \in A, \mathsf{y}(\omega) = y\}) .$$

Thus, every notion in the traditional framework can also be well defined in the new framework, but not vice versa.

The target of pattern classification is to design a classifier $\hat{y} : \mathcal{X} \to \{0,1\}$ so that some criterion is minimized or maximized. Since $\hat{y}$ is a function on $\mathcal{X}$ rather than Ω, for each $x \in \mathcal{X}$ it does not discriminate between the objects in the set $\Omega_x = \{\omega \in \Omega \mid \mathsf{x}(\omega) = x\}$, which may belong to different classes—that is, $\mathsf{y}(\omega)$ may not assume a constant value on Ω_x. So the best thing we can do is to choose the *best class label* (according to the concerned criterion) for each feature vector $x \in \mathcal{X}$; and assign it to *all* objects in the set Ω_x, regardless of their *true class label*.

To recapitulate, the concept of a true class label can only be defined for objects, not for feature vectors; so it is not well defined under the traditional probabilistic framework, which identify an object with its feature vector. At the feature vector level, the notion of best class labels can be defined; and its definition depends on the performance criterion used. We however had better keep using the term "true label" anyway, for otherwise some commonly used notions such as true positive and misclassification rate would cause even more confusion. Also, dropping this term will make the discussion in Elkan (2001) about "reasonableness" conditions of cost matrices (see also page 1040 of this paper) meaningless. For this reason, we have abused the notion of true labels in the paper even though the traditional framework is adopted. It actually should be understood as *the true class label of the particular object* ω *we are talking about whose feature vector is x*.

B.2 On the Convex Hull of a Given Set in $\mathbb{R}^m$

In this section, we propose a recursive procedure to construct the convex hull of a general subset in the Euclidean space $\mathbb{R}^m$, for which we introduce some basic terminologies first. A set $D \subseteq \mathbb{R}^m$ is said to be *convex* if $\alpha \boldsymbol{u} + \tilde{\alpha} \boldsymbol{v} \in D$ for any $\boldsymbol{u}, \boldsymbol{v} \in D$ and any $\alpha \in [0,1]$—recall that $\tilde{\alpha} := 1 - \alpha$. Let D be a convex set in $\mathbb{R}^m$. A function $f : D \to \mathbb{R}$ is *convex* if for any $\boldsymbol{u}, \boldsymbol{v} \in D$ and $\alpha \in [0,1]$, it holds that $f(\alpha \boldsymbol{u} + \tilde{\alpha} \boldsymbol{v}) \leqslant \alpha \cdot f(\boldsymbol{u}) + \tilde{\alpha} \cdot f(\boldsymbol{v})$. If, instead, the reversed inequality holds, then f is a *concave* function. Notice that convex (concave) functions are defined on convex sets. For any $D \subseteq \mathbb{R}^m$, its *convex hull*, denoted as $\operatorname{co} D$ in the paper, is defined as the set of all (finite) convex combinations of points in D,

$$\operatorname{co} D := \{\textstyle\sum_{i=1}^n \alpha_i \boldsymbol{u}_i \mid n \in \mathbb{N}, \boldsymbol{u}_i \in D, \alpha_i \geqslant 0, \sum_{i=1}^n \alpha_i = 1\} .$$

Another equivalent definition of $\operatorname{co} D$ is that it is the smallest convex set that contains D as a subset. Both definitions will be usee (in proving certain propositions). Furthermore, for any $\boldsymbol{u} \in \mathbb{R}^m$ we shall call the sum $\sum_{i=1}^n \alpha_i \boldsymbol{u}_i$ or the set $\{(\alpha_i, \boldsymbol{u}_i)\}_{i=1}^n$ a *convex decomposition of* $\boldsymbol{u}$ *in* D if it holds that $\alpha_i \geqslant 0$, $\sum_{i=1}^n \alpha_i = 1$, $\boldsymbol{u}_i \in D$ and $\boldsymbol{u} = \sum_{i=1}^n \alpha_i \boldsymbol{u}_i$. Notice that a vector $\boldsymbol{u} \in \operatorname{co} D$ if and only if it has at least one convex decomposition in D.

Lemma 21 *The convex hull of any subset D of the real line* $\mathbb{R}$ *is an interval with the endpoints* $a = \inf D$ *and* $b = \sup D$. *Moreover,* $a \in \operatorname{co} D$ *iff* $a \in D$ *and* $b \in \operatorname{co} D$ *iff* $b \in D$.

Proof We only consider the case where $a \in D$ and $b \notin D$; the other three possibilities can be discussed analogously. It is clear that $D \subseteq [a,b)$ and that $[a,b)$ is a convex set. But $\mathrm{co}D$ is the smallest convex set that contains D, thus $\mathrm{co}D \subseteq [a,b)$. It now remains to show that $[a,b) \subseteq \mathrm{co}D$ (hence $\mathrm{co}D = [a,b)$ and we are done). For any $c \in [a,b)$, as $b = \sup D$, there exists a $t \in D$ such that $c < t < b$. Put $\alpha = \frac{t-c}{t-a}$, then $c = \alpha \cdot a + \tilde{\alpha} \cdot t \in \mathrm{co}D$. ∎

Although simple, the above lemma characterizes completely the convex hull of subsets in the 1-dimensional space. For the high dimensional case, we need further to introduce some new symbols. For any $m \in \mathbb{N}$ and any $E \subseteq \mathbb{R}^{m+1}$, denote by $E^{\downarrow}$ the projection of E onto $\mathbb{R}^m$ (the subspace of $\mathbb{R}^{m+1}$ spanned by the first m unit vectors):

$$E^{\downarrow} := \{\boldsymbol{u} = [u_1, \ldots, u_m] \in \mathbb{R}^m \mid [\boldsymbol{u}, s] = [u_1, \ldots, u_m, s] \in E \text{ for some } s \in \mathbb{R}\}.$$

For each $\boldsymbol{u} \in \mathbb{R}^m$, we define $E^{\uparrow}_{\boldsymbol{u}} := \{s \in \mathbb{R} \mid [\boldsymbol{u}, s] \in E\}$. Intuitively, $E^{\uparrow}_{\boldsymbol{u}} \subseteq \mathbb{R}$ can be seen as the intersection of the set E and the real line "vertically" placed at the point $\boldsymbol{u}$. Observe that $E^{\uparrow}_{\boldsymbol{u}} \neq \varnothing$ iff $\boldsymbol{u} \in E^{\downarrow}$ and that $[\boldsymbol{u}, s] \in E$ iff $s \in E^{\uparrow}_{\boldsymbol{u}}$. Furthermore, with the notions of $E^{\downarrow}$ and $E^{\uparrow}_{\boldsymbol{u}}$, any $E \subseteq \mathbb{R}^{m+1}$ can be expressed as $E = \{[\boldsymbol{u}, s] \mid \boldsymbol{u} \in E^{\downarrow}, s \in E^{\uparrow}_{\boldsymbol{u}}\}$. In particular, replacing the set E by its convex hull in this identity, we obtain

$$\mathrm{co}E = \{[\boldsymbol{u}, s] \mid \boldsymbol{u} \in (\mathrm{co}E)^{\downarrow}, s \in (\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}\}. \tag{55}$$

Lemma 22 *For any $E \subseteq \mathbb{R}^{m+1}$, it holds that $(\mathrm{co}E)^{\downarrow} = \mathrm{co}E^{\downarrow}$.*

Proof The set $\mathrm{co}E$ is convex, so is its projection $(\mathrm{co}E)^{\downarrow}$—see, for example, Rockafellar (1970, p. 19, Corollary 3.4.1). Moreover, from $E \subseteq \mathrm{co}E$ we know $E^{\downarrow} \subseteq (\mathrm{co}E)^{\downarrow}$. It then follows from the minimality of $\mathrm{co}E^{\downarrow}$ that $\mathrm{co}E^{\downarrow} \subseteq (\mathrm{co}E)^{\downarrow}$. We now show that $(\mathrm{co}E)^{\downarrow} \subseteq \mathrm{co}E^{\downarrow}$. Let $\boldsymbol{u} \in (\mathrm{co}E)^{\downarrow}$, then $[\boldsymbol{u}, s] \in \mathrm{co}E$ for some $s \in \mathbb{R}$. Hence the vector $[\boldsymbol{u}, s]$ has a convex decomposition in E, say $[\boldsymbol{u}, s] = \sum_{i=1}^{n} \alpha_i \cdot [\boldsymbol{u}_i, s_i]$. It follows from $[\boldsymbol{u}_i, s_i] \in E$ that $\boldsymbol{u}_i \in E^{\downarrow}$ and so $\boldsymbol{u} = \sum_{i=1}^{n} \alpha_i \boldsymbol{u}_i \in \mathrm{co}E^{\downarrow}$. This proves $(\mathrm{co}E)^{\downarrow} \subseteq \mathrm{co}E^{\downarrow}$. ∎

Lemma 22 links the convex hull of a set in $\mathbb{R}^{m+1}$ to that in $\mathbb{R}^m$ and hence simplifies the first ingredient of Equation (55), $(\mathrm{co}E)^{\downarrow}$. We now analyze its second ingredient, $(\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$ with $\boldsymbol{u} \in (\mathrm{co}E)^{\downarrow} = \mathrm{co}E^{\downarrow}$. First of all, since $\mathrm{co}E$ is a convex set, so is $(\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$. To see this, let $s_1, s_2 \in (\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$, then $[\boldsymbol{u}, s_1], [\boldsymbol{u}, s_2] \in \mathrm{co}E$. But $\mathrm{co}E$ is a convex set, so for any $\alpha \in [0,1]$ it holds that $\alpha[\boldsymbol{u}, s_1] + \tilde{\alpha}[\boldsymbol{u}, s_2] = [\boldsymbol{u}, \alpha s_1 + \tilde{\alpha} s_2] \in \mathrm{co}E$, that is, $\alpha s_1 + \tilde{\alpha} s_2 \in (\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$. Secondly, $(\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$ is a subset of $\mathbb{R}$, it hence must be an interval—one should have no difficulty to see that every convex set in $\mathbb{R}$ is an interval. The problem is thus reduced to determining the two endpoints of the interval, that is, the infimum and supremum of $(\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$, for which the following two symbols $\overline{g}(\cdot|\cdot)$ and $\underline{g}(\cdot|\cdot)$ are useful.

By the definition of $E^{\downarrow}$ and $E^{\uparrow}_{\boldsymbol{u}}$, it is obvious that $E^{\uparrow}_{\boldsymbol{u}}$ is nonempty for each $\boldsymbol{u} \in E^{\downarrow}$. Now assume that $E \subseteq \mathbb{R}^{m+1}$ is bounded, that is, $E \subseteq [-b, b]^{m+1}$ for some $b \in \mathbb{R}$, then the set $E^{\uparrow}_{\boldsymbol{u}}$ is also bounded—in fact, $E^{\uparrow}_{\boldsymbol{u}} \subseteq [-b, b]$ for any $\boldsymbol{u} \in E^{\downarrow}$. For such sets E, the functions

$$\overline{g}(\cdot|E) : E^{\downarrow} \to \mathbb{R}, \boldsymbol{u} \mapsto \sup E^{\uparrow}_{\boldsymbol{u}}, \qquad \underline{g}(\cdot|E) : E^{\downarrow} \to \mathbb{R}, \boldsymbol{u} \mapsto \inf E^{\uparrow}_{\boldsymbol{u}} \tag{56}$$

are well defined (and bounded). Note that for general sets $E \subseteq \mathbb{R}^{m+1}$, the above functions could be $\pm\infty$ at some points $\boldsymbol{u}$. So the boundness of the set E is necessary for $\overline{g}$ and $\underline{g}$ to be real-valued. The notation $\overline{g}(\cdot|\cdot)$ and $\underline{g}(\cdot|\cdot)$ allows us to rewrite the supremum of the set $(\mathrm{co}\,E)^{\uparrow}_{\boldsymbol{u}}$ as $\overline{g}(\boldsymbol{u}|\,\mathrm{co}\,E)$ and its infimum as $\underline{g}(\boldsymbol{u}|\,\mathrm{co}\,E)$. Here the functions $\overline{g}(\cdot|\,\mathrm{co}\,E)$ and $\underline{g}(\cdot|\,\mathrm{co}\,E)$ are also defined by Equation (56), but with the set E replaced by $\mathrm{co}\,E$. That is,

$$\begin{aligned} &\overline{g}(\cdot|\,\mathrm{co}\,E) : (\mathrm{co}\,E)^{\downarrow} \to \mathbb{R}, \boldsymbol{u} \mapsto \sup(\mathrm{co}\,E)^{\uparrow}_{\boldsymbol{u}}, \\ &\underline{g}(\cdot|\,\mathrm{co}\,E) : (\mathrm{co}\,E)^{\downarrow} \to \mathbb{R}, \boldsymbol{u} \mapsto \inf(\mathrm{co}\,E)^{\uparrow}_{\boldsymbol{u}}. \end{aligned} \tag{57}$$

In the following we aim to relate the above two functions to $\overline{g}(\cdot|E)$ and $\underline{g}(\cdot|E)$.

Lemma 23 *Let $E \subseteq \mathbb{R}^{m+1}$ be a bounded convex set. Then $\underline{g}(\cdot|E)$ is a convex function and $\overline{g}(\cdot|E)$ a concave function on $E^{\downarrow}$.*

Proof Since E is convex, so is its projection $E^{\downarrow}$. Thus, $\alpha\boldsymbol{u} + \tilde{\alpha}\boldsymbol{v} \in E^{\downarrow}$ for any $\boldsymbol{u}, \boldsymbol{v} \in E^{\downarrow}$ and $\alpha \in [0,1]$. By the definition of $\underline{g}(\cdot|E)$, Equation (56), to prove its convexity we need to show

$$\inf E^{\uparrow}_{\alpha\boldsymbol{u}+\tilde{\alpha}\boldsymbol{v}} \leqslant \alpha \cdot \inf E^{\uparrow}_{\boldsymbol{u}} + \tilde{\alpha} \cdot \inf E^{\uparrow}_{\boldsymbol{v}}. \tag{58}$$

For any $\varepsilon > 0$, by the definition of $E^{\uparrow}_{\boldsymbol{u}}$ we know there is an $s \in \mathbb{R}$ such that $[\boldsymbol{u}, s] \in E$ and $s < \inf E^{\uparrow}_{\boldsymbol{u}} + \varepsilon$. Similarly, there exists a $t \in \mathbb{R}$ satisfying $[\boldsymbol{v}, t] \in E$ and $t < \inf E^{\uparrow}_{\boldsymbol{v}} + \varepsilon$. Then $[\alpha\boldsymbol{u} + \tilde{\alpha}\boldsymbol{v}, \alpha s + \tilde{\alpha} t] \in E$ since E is convex. This means that $\alpha s + \tilde{\alpha} t \in E^{\uparrow}_{\alpha\boldsymbol{u}+\tilde{\alpha}\boldsymbol{v}}$ and so

$$\inf E^{\uparrow}_{\alpha\boldsymbol{u}+\tilde{\alpha}\boldsymbol{v}} \leqslant \alpha s + \tilde{\alpha} t < \alpha \cdot \inf E^{\uparrow}_{\boldsymbol{u}} + \tilde{\alpha} \cdot \inf E^{\uparrow}_{\boldsymbol{v}} + \varepsilon.$$

Since $\varepsilon > 0$ can be arbitrarily small, we get the desired inequality (58).

The concavity of the function $\overline{g}(\cdot|E)$ can be proven in the similar way. ■

By this lemma, we at once see that the function $\underline{g}(\cdot|\,\mathrm{co}\,E)$ is convex and $\overline{g}(\cdot|\,\mathrm{co}\,E)$ concave. Moreover, as $E \subseteq \mathrm{co}\,E$ and hence $E^{\uparrow}_{\boldsymbol{u}} \subseteq (\mathrm{co}\,E)^{\uparrow}_{\boldsymbol{u}}$ for any $\boldsymbol{u} \in \mathbb{R}^m$, we know from Equations (56)–(57) that

$$\underline{g}(\boldsymbol{u}|\,\mathrm{co}\,E) \leqslant \underline{g}(\boldsymbol{u}|E) \leqslant \overline{g}(\boldsymbol{u}|E) \leqslant \overline{g}(\boldsymbol{u}|\,\mathrm{co}\,E), \qquad \forall \boldsymbol{u} \in E^{\downarrow}.$$

Here the domain of $\overline{g}(\cdot|E)$ and $\underline{g}(\cdot|E)$, $E^{\downarrow}$, does not need to be convex; and the domain of $\overline{g}(\cdot|\,\mathrm{co}\,E)$ and $\underline{g}(\cdot|\,\mathrm{co}\,E)$, $(\mathrm{co}\,E)^{\downarrow} = \mathrm{co}\,E^{\downarrow}$, is the convex hull of the domain of $\overline{g}(\cdot|E)$ and $\underline{g}(\cdot|E)$. These observations motivate us to introduce the concepts of the convex/concave hull of functions defined on a subset of $\mathbb{R}^m$ which is not necessarily convex.

Let $D \subseteq \mathbb{R}^m$ and $f : D \to \mathbb{R}$. The *concave hull* of f is the smallest concave function $f^{\frown} : \mathrm{co}\,D \to \mathbb{R}$ such that $f^{\frown}(\boldsymbol{u}) \geqslant f(\boldsymbol{u})$ for all $\boldsymbol{u} \in D$; and the *convex hull* of f is the greatest convex function $f_{\smile} : \mathrm{co}\,D \to \mathbb{R}$ with $f_{\smile}(\boldsymbol{u}) \leqslant f(\boldsymbol{u})$ for $\boldsymbol{u} \in D$. In particular, if the domain D is itself a convex set, then $\mathrm{co}\,D = D$ and our definition of $f_{\smile}$ and $f^{\frown}$ degenerates into the standard definition. Here both $f_{\smile}$ and $f^{\frown}$ are required to be real-valued. As such, some functions might have no convex or concave hull. For instance, the function $f(t) = t^2$, $t \in \mathbb{R}$ does not have concave hull—it would be $f^{\frown}(t) = \infty$ if the extended real line is considered instead of $\mathbb{R}$.

Lemma 24 *For any bounded subset $E \subseteq \mathbb{R}^{m+1}$, the function $\underline{g}(\cdot|\operatorname{co}E)$ is the convex hull of $\underline{g}(\cdot|E)$; and $\overline{g}(\cdot|\operatorname{co}E)$ is the concave hull of $\overline{g}(\cdot|E)$.*

Proof We have shown that $\underline{g}(\cdot|\operatorname{co}E) : \operatorname{co}E^{\downarrow} \to \mathbb{R}$ is a convex function which for any $\boldsymbol{u} \in E^{\downarrow}$ satisfies $\underline{g}(\boldsymbol{u}|\operatorname{co}E) \leqslant \underline{g}(\boldsymbol{u}|E)$. It thus remains to show that $\underline{g}(\cdot|\operatorname{co}E) \geqslant f(\cdot)$ for any convex function $f : \operatorname{co}E^{\downarrow} \to \mathbb{R}$ satisfying the same condition.

Let $\boldsymbol{u} \in \operatorname{co}E^{\downarrow}$, by definition, $\underline{g}(\boldsymbol{u}|\operatorname{co}E) = \inf(\operatorname{co}E)^{\uparrow}_{\boldsymbol{u}}$. Thus, for any $\varepsilon > 0$, there is an $s \in (\operatorname{co}E)^{\uparrow}_{\boldsymbol{u}}$ such that $s < \underline{g}(\boldsymbol{u}|\operatorname{co}E) + \varepsilon$. By $s \in (\operatorname{co}E)^{\uparrow}_{\boldsymbol{u}}$ we know $[\boldsymbol{u}, s] \in \operatorname{co}E$, so it has a convex decomposition in E, say $[\boldsymbol{u}, s] = \sum_{i=1}^{n} \alpha_i \cdot [\boldsymbol{u}_i, s_i]$. It follows from $[\boldsymbol{u}_i, s_i] \in E$ that $s_i \in E^{\uparrow}_{\boldsymbol{u}_i}$ and hence $\underline{g}(\boldsymbol{u}_i|E) = \inf E^{\uparrow}_{\boldsymbol{u}_i} \leqslant s_i$. Since $f : \operatorname{co}E^{\downarrow} \to \mathbb{R}$ is a convex function and since $f(\cdot) \leqslant \underline{g}(\cdot|E)$ on $E^{\downarrow}$, by Jensen's inequality we have

$$\begin{aligned} f(\boldsymbol{u}) = f(\textstyle\sum_{i=1}^{n} \alpha_i \boldsymbol{u}_i) \leqslant \sum_{i=1}^{n} \alpha_i \cdot f(\boldsymbol{u}_i) \leqslant \sum_{i=1}^{n} \alpha_i \cdot \underline{g}(\boldsymbol{u}_i|E) \quad \ldots \\ \ldots \quad \leqslant \textstyle\sum_{i=1}^{n} \alpha_i s_i = s < \underline{g}(\boldsymbol{u}|\operatorname{co}E) + \varepsilon. \end{aligned}$$

As $\varepsilon > 0$ can be arbitrarily small, the above inequality implies $f(\boldsymbol{u}) \leqslant \underline{g}(\boldsymbol{u}|\operatorname{co}E)$. We thus have proved that $\underline{g}(\cdot|\operatorname{co}E)$ is the convex hull of $\underline{g}(\cdot|E)$. By the similar argument, we can prove $\overline{g}(\cdot|\operatorname{co}E)$ is the concave hull of $\overline{g}(\cdot|E)$. ■

Lemma 25 *Let $D \subset \mathbb{R}^m$ be an arbitrary set. Then any lower (upper) bounded function $f : D \to \mathbb{R}$ allows for a convex (concave) hull $f_{\smile}(f^{\frown}) : \operatorname{co}D \to \mathbb{R}$.*

Proof On the set $\operatorname{co}D$ define two functions $f^*(\boldsymbol{u})$ and $f_*(\boldsymbol{u})$ by

$$f^*(\boldsymbol{u}) := \sup\{\textstyle\sum_{i=1}^{n} \alpha_i \cdot f(\boldsymbol{u}_i) \mid \{(\alpha_i, \boldsymbol{u}_i)\}_{i=1}^{n} \text{ a conv. decomp. of } \boldsymbol{u} \text{ in } D\}, \tag{59}$$

$$f_*(\boldsymbol{u}) := \inf\{\textstyle\sum_{i=1}^{n} \alpha_i \cdot f(\boldsymbol{u}_i) \mid \{(\alpha_i, \boldsymbol{u}_i)\}_{i=1}^{n} \text{ a conv. decomp. of } \boldsymbol{u} \text{ in } D\}. \tag{60}$$

As any $\boldsymbol{u} \in \operatorname{co}D$ allows for at least one convex decomposition in D, the above set $\{\sum \ldots\}$ is nonempty and hence its supremum and infimum are well defined. We claim that $f_{\smile} = f_*$ when f is lower bounded and that $f^{\frown} = f^*$ when f is upper bounded.

By the definition of $f_{\smile}$, to see that $f_{\smile} = f_*$ it suffices to show

(a) *$f_*(\cdot)$ is a convex function on* $\operatorname{co}D$: Let $\boldsymbol{u}, \boldsymbol{v} \in \operatorname{co}D$ and $t \in [0,1]$, we need to prove $f_*(t\boldsymbol{u} + \tilde{t}\boldsymbol{v}) \leqslant t \cdot f_*(\boldsymbol{u}) + \tilde{t} \cdot f_*(\boldsymbol{v})$. For any $\varepsilon > 0$, by the definition of $f_*(\boldsymbol{u})$, there is a convex decomposition of $\boldsymbol{u}$ in D, $\{(\alpha_i, \boldsymbol{u}_i)\}_{i=1}^{n}$, such that $f_*(\boldsymbol{u}) > \sum_{i=1}^{n} \alpha_i \cdot f(\boldsymbol{u}_i) - \varepsilon$. Analogously, $f_*(\boldsymbol{v}) > \sum_{i=1}^{k} \beta_i \cdot f(\boldsymbol{v}_i) - \varepsilon$ for some convex decomposition of $\boldsymbol{v}$, $\{(\beta_i, \boldsymbol{v}_i)\}_{i=1}^{k}$. We thus get

$$t \cdot f_*(\boldsymbol{u}) + \tilde{t} \cdot f_*(\boldsymbol{v}) > \textstyle\sum_{i=1}^{n} t\alpha_i \cdot f(\boldsymbol{u}_i) + \sum_{i=1}^{k} \tilde{t}\beta_i \cdot f(\boldsymbol{v}_i) - \varepsilon.$$

But the set $\{(t\alpha_i, \boldsymbol{u}_i)\}_{i=1}^{n} \cup \{(\tilde{t}\beta_i, \boldsymbol{v}_i)\}_{i=1}^{k}$ forms a convex decomposition of $t\boldsymbol{u} + \tilde{t}\boldsymbol{v}$, so

$$f_*(t\boldsymbol{u} + \tilde{t}\boldsymbol{v}) \leqslant \textstyle\sum_{i=1}^{n} t\alpha_i \cdot f(\boldsymbol{u}_i) + \sum_{i=1}^{k} \tilde{t}\beta_i \cdot f(\boldsymbol{v}_i).$$

It hence follows that $f_*(t\boldsymbol{u} + \tilde{t}\boldsymbol{v}) < t \cdot f_*(\boldsymbol{u}) + \tilde{t} \cdot f_*(\boldsymbol{v}) + \varepsilon$. Since $\varepsilon > 0$ can be arbitrarily small, we conclude that $f_*(t\boldsymbol{u} + \tilde{t}\boldsymbol{v}) \leqslant t \cdot f_*(\boldsymbol{u}) + \tilde{t} \cdot f_*(\boldsymbol{v})$.

(b) *$f_*(\boldsymbol{u}) \leqslant f(\boldsymbol{u})$ for all $\boldsymbol{u} \in D$*: This is obvious as $\{(1,\boldsymbol{u})\}$ is a convex decomposition of $\boldsymbol{u}$ in D.

(c) *$g(\boldsymbol{u}) \leqslant f_*(\boldsymbol{u})$ for any $g : \mathrm{co}D \to \mathbb{R}$ satisfying the above conditions (a) and (b), and any $\boldsymbol{u} \in \mathrm{co}D$*: For any $\varepsilon > 0$, by the definition of $f_*(\boldsymbol{u})$, there is a convex decomposition of $\boldsymbol{u}$ in D, $\{(\alpha_i,\boldsymbol{u}_i)\}_{i=1}^n$, such that $f_*(\boldsymbol{u}) > \sum_{i=1}^n \alpha_i \cdot f(\boldsymbol{u}_i) - \varepsilon$. As $g(\cdot)$ is a convex function, and as $g \leqslant f$ on D, by Jensen's inequality we have

$$\textstyle\sum_{i=1}^n \alpha_i \cdot f(\boldsymbol{u}_i) \geqslant \sum_{i=1}^n \alpha_i \cdot g(\boldsymbol{u}_i) \geqslant g(\sum_{i=1}^n \alpha_i \boldsymbol{u}_i) = g(\boldsymbol{u}),$$

where the last equality follows from that $\{(\alpha_i,\boldsymbol{u}_i)\}_{i=1}^n$ is a convex decomposition of $\boldsymbol{u}$. We thus know $f_*(\boldsymbol{u}) + \varepsilon > g(\boldsymbol{u})$ and so $f_*(\boldsymbol{u}) \geqslant g(\boldsymbol{u})$, since $\varepsilon > 0$ can be arbitrarily small.

By the similar argument, one shows that $f^{\frown} = f^*$ for upper bounded functions f. ■

The above two lemmas enable us to describe the functions $\overline{g}(\cdot|\mathrm{co}E)$ and $\underline{g}(\cdot|\mathrm{co}E)$ in terms of $\overline{g}(\cdot|E)$ and $\underline{g}(\cdot|E)$, respectively. In fact, by putting $f(\cdot) = \overline{g}(\cdot|E)$ in Equation (59) and $f(\cdot) = \underline{g}(\cdot|E)$ in Equation (60), we get

$$\overline{g}(\boldsymbol{u}|\mathrm{co}E) = \sup\{\textstyle\sum_{i=1}^n \alpha_i \overline{g}(\boldsymbol{u}_i|E) \mid \{(\alpha_i,\boldsymbol{u}_i)\}_{i=1}^n \text{ a conv. decomp. of } \boldsymbol{u} \text{ in } E^{\downarrow}\}, \tag{61}$$

$$\underline{g}(\boldsymbol{u}|\mathrm{co}E) = \inf\{\textstyle\sum_{i=1}^n \alpha_i \underline{g}(\boldsymbol{u}_i|E) \mid \{(\alpha_i,\boldsymbol{u}_i)\}_{i=1}^n \text{ a conv. decomp. of } \boldsymbol{u} \text{ in } E^{\downarrow}\}. \tag{62}$$

Now let us return to the expression (55), $\mathrm{co}E = \{[\boldsymbol{u},s] \mid \boldsymbol{u} \in (\mathrm{co}E)^{\downarrow}, s \in (\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}\}$. As has been pointed out earlier, for any $\boldsymbol{u} \in (\mathrm{co}E)^{\downarrow} = \mathrm{co}E^{\downarrow}$, the set $(\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$ is an interval in $\mathbb{R}$ with the two endpoints $\overline{g}(\boldsymbol{u}|\mathrm{co}E)$, $\underline{g}(\boldsymbol{u}|\mathrm{co}E)$ determined respectively by Equation (61) and Equation (62). This interval might be open, closed, or half-open-half-closed, depending on whether or not the respective endpoint is in the interval. For simplicity we restrict ourselves to bounded and closed sets E. Then their convex hull $\mathrm{co}E$ are also bounded and closed—see, for example, Aliprantis and Border (2006, p. 185, Corollary 5.33), which in turn implies the set $(\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}}$ can only be a closed interval, $(\mathrm{co}E)^{\uparrow}_{\boldsymbol{u}} = [\underline{g}(\boldsymbol{u}|\mathrm{co}E), \overline{g}(\boldsymbol{u}|\mathrm{co}E)]$. Equation (55) can thus be rewritten as

$$\mathrm{co}E = \{[\boldsymbol{u},s] \mid \boldsymbol{u} \in \mathrm{co}E^{\downarrow},\ \underline{g}(\boldsymbol{u}|\mathrm{co}E) \leqslant s \leqslant \overline{g}(\boldsymbol{u}|\mathrm{co}E)\}. \tag{63}$$

The projection $E^{\downarrow}$ of a bounded closed set E is also bounded and closed, so the above expression of $\mathrm{co}E$ gives naturally rise to a recursive algorithm to construct the convex hull of any bounded and closed set E, as follows. To get $\mathrm{co}E$ we need only to find $\mathrm{co}E^{\downarrow}$ and the functions $\underline{g}(\cdot|\mathrm{co}E)$ and $\overline{g}(\cdot|\mathrm{co}E)$ as given by Equations (56), (61) and (62); to get $\mathrm{co}E^{\downarrow}$ we need to find $\mathrm{co}E^{\downarrow\downarrow}$ and the functions $\underline{g}(\cdot|\mathrm{co}E^{\downarrow})$ and $\overline{g}(\cdot|\mathrm{co}E^{\downarrow})$; and so forth. As $E^{\downarrow} \subseteq \mathbb{R}^m$ for any $E \subseteq \mathbb{R}^{m+1}$, this procedure terminates with the 1-dimensional case after m steps, which has been fully discussed in Lemma 21.

B.3 The Convex Hull of Three Curves ℓ in the Paper

We now apply the recursive procedure presented in the preceding section to three curves occurred in the paper, to get their convex hull. These curves have been parameterized by the posterior probability $\eta \in [0,1]$, as listed below—to distinguish, a subscript is used to indicate the quantity with which the curve is associated:

$$\ell_{\underline{\mathrm{CSR}}} = \{[e(\eta), h(\eta)] \mid \eta \in [0,1]\}\,, \qquad e(\eta) = \min\{c_1\eta, c_0\tilde{\eta}\}\,; \tag{64}$$

$$\ell_{\underline{\mathrm{BER}}} = \{[\eta, r(\eta), h(\eta)] \mid \eta \in [0,1]\}\,, \qquad r(\eta) = \min\{\pi^{-1}\eta, \tilde{\pi}^{-1}\tilde{\eta}\}\,; \tag{65}$$

$$\ell_{\overline{\mathrm{FSC}}} = \{[\eta, u(\eta), h(\eta)] \mid \eta \in [0,1]\}\,, \qquad u(\eta) = \theta^*\eta - (\eta - \theta^*)^+ \text{ and } \theta^* = \tfrac{1}{2}\overline{\mathrm{FSC}}\,. \tag{66}$$

In the above, the function $h : [0,1] \to \mathbb{R}$ is concave and satisfies $h(0) = h(1) = 0$.

B.3.1 THE CONVEX HULL OF $\ell_{\underline{\mathrm{CSR}}}$

The curve $\ell_{\underline{\mathrm{CSR}}}$ lies in the e-h plane; and, by Equation (63), its convex hull can be expressed as

$$\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}} = \{[e_0, h_0] \mid e_0 \in \mathrm{co}\,\ell^{\downarrow}_{\underline{\mathrm{CSR}}},\ \underline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}}) \leqslant h_0 \leqslant \overline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}})\}\,. \tag{67}$$

As $c_0^{-1} + c_1^{-1} = 2$, the range of $e(\eta)$ is $[0, 0.5]$—see the analysis in page 1048. We thus have $\ell^{\downarrow}_{\underline{\mathrm{CSR}}} = \{e(\eta) \mid \eta \in [0,1]\} = [0, 0.5]$ and hence $\mathrm{co}\,\ell^{\downarrow}_{\underline{\mathrm{CSR}}} = [0, 0.5]$. For each $e_0 \in [0, 0.5]$, the set $(\ell_{\underline{\mathrm{CSR}}})^{\uparrow}_{e_0}$ is computed as follows: $(\ell_{\underline{\mathrm{CSR}}})^{\uparrow}_{e_0} = \{h_0 \in \mathbb{R} \mid [e_0, h_0] \in \ell_{\underline{\mathrm{CSR}}}\} = \{h(\eta) \mid e(\eta) = e_0\}$. But $e(\eta) = e_0$ implies $\eta = c_1^{-1}e_0$ or $\eta = 1 - c_0^{-1}e_0$, so $(\ell_{\underline{\mathrm{CSR}}})^{\uparrow}_{e_0} = \{h(c_1^{-1}e_0), h(1 - c_0^{-1}e_0)\}$. It then follows from Equation (56) that

$$\overline{g}(e_0|\ell_{\underline{\mathrm{CSR}}}) = \max\{h(c_1^{-1}e_0), h(1 - c_0^{-1}e_0)\}\,, \qquad \forall\, e_0 \in [0, 0.5]\,;$$
$$\underline{g}(e_0|\ell_{\underline{\mathrm{CSR}}}) = \min\{h(c_1^{-1}e_0), h(1 - c_0^{-1}e_0)\}\,, \qquad \forall\, e_0 \in [0, 0.5]\,.$$

By Lemma 24, we know $\overline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}})$ is the concave hull of $\overline{g}(e_0|\ell_{\underline{\mathrm{CSR}}})$ and $\underline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}})$ the convex hull of $\underline{g}(e_0|\ell_{\underline{\mathrm{CSR}}})$, that is,

$$\overline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}}) = [\max\{h(c_1^{-1}e_0), h(1 - c_0^{-1}e_0)\}]^{\frown}\,,$$
$$\underline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}}) = [\min\{h(c_1^{-1}e_0), h(1 - c_0^{-1}e_0)\}]_{\smile} = 2 \cdot h(\tfrac{1}{2c_1}) \cdot e_0\,,$$

where the last equality holds because both $h(c_1^{-1}e_0)$ and $h(1 - c_0^{-1}e_0)$ are concave functions of e_0 and they have the same endpoints: $[0,0]$ and $[\frac{1}{2}, h(\frac{1}{2c_1})]$. Substituting the identity $\mathrm{co}\,\ell^{\downarrow}_{\underline{\mathrm{CSR}}} = [0, 0.5]$ and the above expressions of $\overline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}})$ and $\underline{g}(e_0|\,\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}})$ into Equation (67), we obtain

$$\mathrm{co}\,\ell_{\underline{\mathrm{CSR}}} = \{[e_0, h_0] \mid e_0 \in [0, 0.5],\ 2e_0 \cdot h(\tfrac{1}{2c_1}) \leqslant h_0 \leqslant [\ldots]^{\frown}\}\,, \tag{68}$$

where the expression in the brackets $[\ldots]$ is $\max\{h(c_1^{-1}e_0), h(1 - c_0^{-1}e_0)\}$.

B.3.2 THE CONVEX HULL OF $\ell_{\underline{\mathrm{BER}}}$ AND $\ell_{\overline{\mathrm{FSC}}}$

The curves $\ell_{\underline{\mathrm{BER}}}$ and $\ell_{\overline{\mathrm{FSC}}}$ are of the same nature: both $r(\eta)$ and $u(\eta)$ are piecewise affine functions whose graph consists of two line segments. They can hence be treated together. By the definition of $r(\eta)$ and $u(\eta)$, we have

$$r(\eta) = \begin{cases} \pi^{-1}\eta & \text{if } \eta \leqslant \pi \\ \tilde{\pi}^{-1}\tilde{\eta} & \text{otherwise} \end{cases}\,, \qquad u(\eta) = \begin{cases} \theta^*\eta & \text{if } \eta \leqslant \theta^* \\ \theta^* - \tilde{\theta}^*\eta & \text{otherwise} \end{cases}\,. \tag{69}$$

The graph of the two functions for $\pi = \theta^* = 0.3$ are shown in Figure 11; they also represent the curves $\ell^{\downarrow}_{\underline{\mathrm{BER}}} = \{[\eta, r(\eta)] \mid \eta \in [0,1]\}$ and $\ell^{\downarrow}_{\overline{\mathrm{FSC}}} = \{[\eta, u(\eta)] \mid \eta \in [0,1]\}$, respectively. As both $r(\eta)$

and $u(\eta)$ are concave functions, by Equation (63) it is easy to see that (the detailed derivation is just a routine work and omitted here)

$$\operatorname{co}\ell^{\downarrow}_{\underline{\mathrm{BER}}} = \{[\eta, r_0] \mid \eta \in [0,1],\, 0 \leqslant r_0 \leqslant r(\eta)\}\,,$$
$$\operatorname{co}\ell^{\downarrow}_{\overline{\mathrm{FSC}}} = \{[\eta, u_0] \mid \eta \in [0,1],\, (\theta^* - \tilde{\theta}^*)\eta \leqslant u_0 \leqslant u(\eta)\}\,. \tag{70}$$

This is also clear from Figure 11: they are just the triangles OAB.

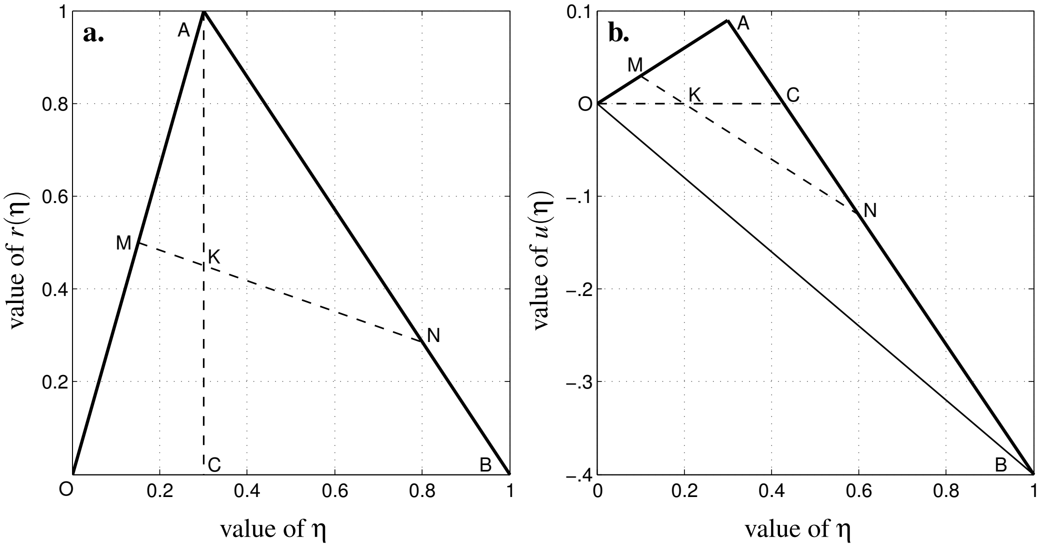


Figure 11: **a.** The graph of the function $r(\eta)$ (the solid line OAB), which can also be expressed as $\ell^{\downarrow}_{\underline{\mathrm{BER}}} = \{[\eta, r(\eta)] \mid \eta \in [0,1]\}$. From the graph one easily sees that the convex hull of $\ell^{\downarrow}_{\underline{\mathrm{BER}}}$ is the area bounded by the triangle OAB.
b. The same graph and curve ($\ell^{\downarrow}_{\underline{\mathrm{CSR}}}$) for the function $u(\eta)$.

As before, to derive the convex hull of the curve $\ell_{\overline{\mathrm{FSC}}}$, we use Equation (63) and obtain

$$\operatorname{co}\ell_{\overline{\mathrm{FSC}}} = \{[\eta, u_0, h_0] \mid [\eta, u_0] \in \operatorname{co}\ell^{\downarrow}_{\overline{\mathrm{FSC}}},\, \underline{g}(\eta, u_0 | \operatorname{co}\ell_{\overline{\mathrm{FSC}}}) \leqslant h_0 \leqslant \overline{g}(\eta, u_0 | \operatorname{co}\ell_{\overline{\mathrm{FSC}}})\}. \tag{71}$$

The set $\operatorname{co}\ell^{\downarrow}_{\overline{\mathrm{FSC}}}$ is already known, so it remains to find the expressions of $\overline{g}(\eta, u_0 | \operatorname{co}\ell_{\overline{\mathrm{FSC}}})$ and $\underline{g}(\eta, u_0 | \operatorname{co}\ell_{\overline{\mathrm{FSC}}})$, for which we need first to determine the values of $\overline{g}(\eta, u_0 | \ell_{\overline{\mathrm{FSC}}})$ and $\underline{g}(\eta, u_0 | \ell_{\overline{\mathrm{FSC}}})$ for $[\eta, u_0] \in \ell^{\downarrow}_{\overline{\mathrm{FSC}}}$—see Equation (61) and Equation (62). By the definition of $\ell_{\overline{\mathrm{FSC}}}$, we know $[\eta, u_0] \in \ell^{\downarrow}_{\overline{\mathrm{FSC}}}$ if and only if $u_0 = u(\eta)$; and $(\ell_{\overline{\mathrm{FSC}}})^{\uparrow}_{[\eta, u_0]} = \{h(\eta)\}$ for any $[\eta, u_0] \in \ell^{\downarrow}_{\overline{\mathrm{FSC}}}$. It thus follows from Equation (56) that $\overline{g}(\eta, u_0 | \ell_{\overline{\mathrm{FSC}}}) = \underline{g}(\eta, u_0 | \ell_{\overline{\mathrm{FSC}}}) = h(\eta)$ for any point $[\eta, u_0] \in \ell^{\downarrow}_{\overline{\mathrm{FSC}}}$, that is, for any $\eta \in [0,1]$ and $u_0 = u(\eta)$.

Based upon the above discussion and Equations (61) and (62), we have

$$\overline{g}(\eta, u_0 | \operatorname{co}\ell_{\overline{\mathrm{FSC}}}) = \sup\{\textstyle\sum_{i=1}^{n} \alpha_i \cdot h(\eta_i) \mid \text{condition on } (\alpha_i, \eta_i)\}\,, \tag{72}$$
$$\underline{g}(\eta, u_0 | \operatorname{co}\ell_{\overline{\mathrm{FSC}}}) = \inf\{\textstyle\sum_{i=1}^{n} \alpha_i \cdot h(\eta_i) \mid \text{condition on } (\alpha_i, \eta_i)\}\,, \tag{73}$$

for any $[\eta, u_0] \in \operatorname{co} \ell^{\downarrow}_{\overline{\mathrm{FSC}}}$, where the unspecified condition is that $\{(\alpha_i, \eta_i, u(\eta_i))\}_{i=1}^{n}$ forms a convex decomposition of the point $[\eta, u_0]$ in $\ell^{\downarrow}_{\overline{\mathrm{FSC}}}$. In other words, here the parameters $\alpha_i, \eta_i \in [0,1]$ should satisfy $\sum_{i=1}^{n} \alpha_i = 1$, $\sum_{i=1}^{n} \alpha_i \eta_i = \eta$ and $\sum_{i=1}^{n} \alpha_i \cdot u(\eta_i) = u_0$. Next we shall prove that $n \leqslant 2$ when the supremum in Equation (72) is obtained; and that the infimum in Equation (73) is attained at $n \leqslant 3$ with $\eta_i \in \{0, \theta^*, 1\}$.

We discuss Equation (73) first. For each η_i in a convex decomposition $\{(\alpha_i, \eta_i, u(\eta_i))\}_{i=1}^{n}$ of $[\eta, u_0]$, $\eta_i \in [0, \theta^*]$ or $\eta_i \in [\theta^*, 1]$. For the former case, we "split up" the item $(\alpha_i, \eta_i, u(\eta_i))$ into two items at $\eta = 0$ and θ^*, with the α-parameter computed from the condition that the weighted sum of the new items equals to the original one. That is, we construct $(\alpha_i^1, \eta_i^1, u(\eta_i^1))$ and $(\alpha_i^2, \eta_i^2, u(\eta_i^2))$ with $\eta_i^1 = 0$ and $\eta_i^2 = \theta^*$, such that

$$\alpha_i^1 + \alpha_i^2 = \alpha_i, \quad \alpha_i^1 \eta_i^1 + \alpha_i^2 \eta_i^2 = \alpha_i \eta_i, \quad \alpha_i^1 \cdot u(\eta_i^1) + \alpha_i^2 \cdot u(\eta_i^2) = \alpha_i \cdot u(\eta_i). \tag{74}$$

The third equality of Equation (74) is actually an implication of the first two, because $u(\eta)$ is an affine function on the interval $[0, \theta^*]$—see Equation (69). By the first two equations in Equation (74), we know $\alpha_i^1 = \alpha_i \cdot \frac{\theta^* - \eta_i}{\theta^*}$ and $\alpha_i^2 = \alpha_i \cdot \frac{\eta_i}{\theta^*}$. For the case of $\eta_i \in [\theta^*, 1]$, the split is computed also from Equation (74), but with $\eta_i^1 = \theta^*$ and $\eta_i^2 = 1$. This gives us $\alpha_i^1 = \alpha_i \cdot \frac{1-\eta_i}{1-\theta^*}$ and $\alpha_i^2 = \alpha_i \cdot \frac{\eta_i - \theta^*}{1-\theta^*}$.

In geometry (see Figure 11), the above splitting operation replaces any point M (resp. N) on the line segment OA (resp. AB) by a (unique) convex combination of the two endpoints O and A (resp. A and B). We thus get a new set $\{(\alpha_i^1, \eta_i^1, u(\eta_i^1)), (\alpha_i^2, \eta_i^2, u(\eta_i^2))\}_{i=1}^{n}$, which, by Equation (74), is obviously a convex decomposition of the point $[\eta, u_0]$ in $\ell^{\downarrow}_{\overline{\mathrm{FSC}}}$. Now, as $h(\eta)$ is a concave function, we know from Lemma 9 that $\alpha_i^1 \cdot h(\eta_i^1) + \alpha_i^2 \cdot h(\eta_i^2) \leqslant \alpha_i \cdot h(\eta_i)$. This implies that the sum $\sum_i \alpha_i \cdot h(\eta_i)$ of the new convex decomposition is no more than that of the original one. Moreover, by its construction, the η-parameter of this new convex decomposition assumes one of the three values: 0, θ^* and 1. We can thus "merge" all items with same η-value into one item (in an obvious way), yielding a convex decomposition of $[\eta, u_0]$ with no more than three items—we wrote "no more than" because any item with $\alpha_i = 0$ can be removed without changing the whole convex decomposition and the value of $\sum_i \alpha_i \cdot h(\eta_i)$.

Thus far, we have shown that for any convex decomposition of $[\eta, u_0]$ another convex decomposition can be constructed which has at most three items whose η-parameter are in the set $\{0, \theta^*, 1\}$, and for which the sum $\sum_i \alpha_i \cdot h(\eta_i)$ is less than or equal to that of the original convex decomposition. Therefore, Equation (73) can be simplified to

$$\underline{g}(\eta, u_0 | \operatorname{co} \ell_{\overline{\mathrm{FSC}}}) = \inf\{\alpha_1 \cdot h(0) + \alpha_2 \cdot h(\theta^*) + \alpha_3 \cdot h(1) \mid \text{condition on } \alpha_{1,2,3}\}.$$

In the above expression, $\alpha_i \geqslant 0$ are the coefficients occurred when $[\eta, u_0] \in \operatorname{co} \ell^{\downarrow}_{\overline{\mathrm{FSC}}}$ is written as the (unique) convex combination of the three points $[0, u(0)]$, $[\theta^*, u(\theta^*)]$ and $[1, u(1)]$. In Figure 11, this corresponds with that a point K in the triangle OAB is written as a convex combination of the three extreme points O, A and B. As is well know in geometry, such a convex combination is unique.

By the above discussion, the expression of $\underline{g}(\eta, u_0 | \operatorname{co} \ell_{\overline{\mathrm{FSC}}})$ can be simplified further to

$$\underline{g}(\eta, u_0 | \operatorname{co} \ell_{\overline{\mathrm{FSC}}}) = \alpha_1 \cdot h(0) + \alpha_2 \cdot h(\theta^*) + \alpha_3 \cdot h(1) = \alpha_2 \cdot h(\theta^*),$$

where we have employed the fact $h(0) = h(1) = 0$, and the coefficients α_i are (uniquely) determined by the linear equations with η and u_0 as known constants:

$$\alpha_1 + \alpha_2 + \alpha_3 = 0, \qquad \alpha_1 \cdot [0, u(0)] + \alpha_2 \cdot [\theta^*, u(\theta^*)] + \alpha_3 \cdot [1, u(1)] = [\eta, u_0].$$

As $u(0)=0$, $u(\theta^*)=(\theta^*)^2$ and $u(1)=\theta^*-\tilde{\theta}^*$, solving the above equations results in $\alpha_2=(\theta^*\tilde{\theta}^*)^{-1}\cdot[u_0+\eta(\tilde{\theta}^*-\theta^*)]$. Therefore,

$$\underline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})=(\theta^*\tilde{\theta}^*)^{-1}\cdot h(\theta^*)\cdot[u_0+\eta(\tilde{\theta}^*-\theta^*)]\,,\qquad\forall\,[\eta,u_0]\in\operatorname{co}\ell^{\downarrow}_{\overline{\mathrm{FSC}}}\,. \tag{75}$$

The expression of $\underline{g}(\eta,r_0|\operatorname{co}\ell_{\underline{\mathrm{BER}}})$ can be derived in a similar way, yielding

$$\underline{g}(\eta,r_0|\operatorname{co}\ell_{\underline{\mathrm{BER}}})=r_0\cdot h(\pi)\,,\qquad\forall\,[\eta,r_0]\in\operatorname{co}\ell^{\downarrow}_{\underline{\mathrm{BER}}}\,. \tag{76}$$

Note that both $\underline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})$ and $\underline{g}(\eta,r_0|\operatorname{co}\ell_{\underline{\mathrm{BER}}})$ are affine functions. This fact and Equation (71) reveal that in Figure 6-a (resp. Figure 7-a) the convex hull of the curve $\ell_{\underline{\mathrm{BER}}}$ (resp. $\ell_{\overline{\mathrm{FSC}}}$) is bounded from below by the triangle OAB in the η-r-h (resp. η-u-h) space.

We now study the expression of $\overline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})$, Equation (72). For simplicity, assume that a convex decomposition $\{(\alpha_i,\eta_i,u(\eta_i))\}_{i=1}^n$ of $[\eta,u_0]$ have been ordered such that $\eta_i<\theta^*$ for $i<k$ and $\eta_i\geqslant\theta^*$ for $i\geqslant k$. We can then "merge" the items $\{(\alpha_i,\eta_i,u(\eta_i))\}_{i=1}^{k-1}$ into one, namely, their weighted sum $(\alpha',\eta',u(\eta'))$ with $\alpha'=\sum_{i=1}^{k-1}\alpha_i$ and $\eta'=\frac{1}{\alpha'}\cdot\sum_{i=1}^{k-1}\alpha_i\eta_i$. Similarly, $\{(\alpha_i,\eta_i,u(\eta_i))\}_{i=k}^n$ can be "merged" into $(\alpha'',\eta'',u(\eta''))$ with $\alpha''=\sum_{i=k}^n\alpha_i$ and $\eta''=\frac{1}{\alpha''}\cdot\sum_{i=k}^n\alpha_i\eta_i$. As $u(\eta)$ is an affine function on the intervals $[0,\theta^*]$ and $[\theta^*,1]$—see Equation (69), one easily verifies that $\{(\alpha',\eta',u(\eta')),(\alpha'',\eta'',u(\eta''))\}$ is a convex decomposition of $[\eta,u_0]$ in $\ell^{\downarrow}_{\overline{\mathrm{FSC}}}$. Furthermore, by the concavity of $h(\eta)$ we know $\alpha'\cdot h(\eta')\geqslant\sum_{i=1}^{k-1}\alpha_i\cdot h(\eta_i)$ and $\alpha''\cdot h(\eta'')\geqslant\sum_{i=k}^n\alpha_i\cdot h(\eta_i)$. Hence $\alpha'\cdot h(\eta')+\alpha''\cdot h(\eta'')\geqslant\sum_{i=1}^n\alpha_i\cdot h(\eta_i)$. This enables us to consider only convex decompositions with at most two items when dealing with the function $\overline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})$. That is, Equation (72) can now be simplified to

$$\overline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})=\sup\{\tilde{t}\cdot h(\eta')+t\cdot h(\eta'')\mid\text{condition on }t,\eta'\text{ and }\eta''\}\,, \tag{77}$$

where t, η' and η'' should be such that $\eta'<\theta^*\leqslant\eta''$ and $\{(\tilde{t},\eta',u(\eta')),(t,\eta'',u(\eta''))\}$ forms a convex decomposition of $[\eta,u_0]$.

In Figure 11, Equation (77) means that for any point $\mathrm{K}=[\eta,u_0]$ in the triangle OAB, we need to find a point $\mathrm{M}=[\eta',u(\eta')]$ on the line segment OA and a point $\mathrm{N}=[\eta'',u(\eta'')]$ on the line segment AB, such that K is on the line segment MN, that is, $\mathrm{K}=\tilde{t}\cdot\mathrm{M}+t\cdot\mathrm{N}$. The value of $\overline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})$ is then the supremum of $t\cdot h(\eta')+\tilde{t}\cdot h(\eta'')$ over all such pairs (M,N). For the curve $\ell_{\underline{\mathrm{BER}}}$, we actually have already carried out this computation in Section 5—see Equations (36)–(38) and (41), whose correctness gets verified by the discussion here. Moreover, the analysis in this section shows that calculating Equations (39) and (40) is in fact unnecessary, which was previously proven in Lemma 10. The similar method can be used to simplify the expression of $\overline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})$ to the maximum of a function of t, like Equation (41).

For the purpose of deriving Theorem 11 and Theorem 15, we will focus only on the case of $u_0=0$ for $\overline{g}(\eta,u_0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})$ and the case of $\eta=\pi$ for $\overline{g}(\eta,r_0|\operatorname{co}\ell_{\underline{\mathrm{BER}}})$. The corresponding expressions for these two cases are listed below (the detailed computation is omitted):

$$\overline{g}(\eta,0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}})=\sup\{\tilde{t}\cdot h(\tilde{t}^{-1}(\eta\tilde{\theta}^*-t\theta^*))+t\cdot h(\theta^*+t^{-1}\eta\theta^*)\mid t\in[\eta\tfrac{\theta^*}{\tilde{\theta}^*},\eta\tfrac{\tilde{\theta}^*}{\theta^*}]\}\,, \tag{78}$$

$$\overline{g}(\pi,r_0|\operatorname{co}\ell_{\underline{\mathrm{BER}}})=\sup\{\tilde{t}\cdot h(\pi-\tilde{t}^{-1}\pi\tilde{\pi}\tilde{r}_0)+t\cdot h(\pi+t^{-1}\pi\tilde{\pi}\tilde{r}_0)\mid t\in[\pi\tilde{r}_0,r_0+\pi\tilde{r}_0]\}\,. \tag{79}$$

Note that the above Equation (79) is same as Equation (41) (if we replace ρ by r_0).

B.4 The Expectation of a Random Vector and the Convex Hull of Its Range

This section is devoted to proving that *any random vector in $\mathbb{R}^m$ has the expectation lying in the convex hull of its range*. We actually will prove a stronger theorem, which are to be stated in a formal way after we have introduced the necessary definitions and notations.

Modern probability theory defines a random variable as a measurable function on some probability space $(\Omega, \mathscr{F}, P)$. In particular, a random vector is a measurable function from Ω into the Euclidean space $\mathbb{R}^m$ equipped with the σ-algebra of Borel sets. For any $A \in \mathscr{F}$ with $P(A) > 0$ and any random vector $\mathsf{u} : \Omega \to \mathbb{R}^m$, we write

$$\mathsf{u}(A) := \{\mathsf{u}(\omega) \mid \omega \in A\},$$
$$\mathbb{E}_A[\mathsf{u}] := P(A)^{-1} \cdot \textstyle\int_A \mathsf{u}(\omega)\mathrm{d}P.$$

Intuitively, $\mathsf{u}(A) \subseteq \mathbb{R}^m$ is the image of the set $A \subseteq \Omega$ under the mapping u; and $\mathbb{E}_A[\mathsf{u}]$ is the average value (weighted by probability) of u on the set A. Note that, when $A = \Omega$ the above two quantities are the range and the expectation of u, respectively.

We are now ready to formally state the main result of this section.

Theorem 26 *Let $\mathsf{u} : \Omega \to \mathbb{R}^m$ be a random vector and $A \in \mathscr{F}$ satisfy $P(A) > 0$. Then $\mathbb{E}_A[\mathsf{u}] \in \operatorname{co}\mathsf{u}(A)$.*

The following two lemmas discuss the 1-dimensional case (i.e., $m = 1$) and are useful for proving the theorem.

Lemma 27 *Let $A \in \mathscr{F}$ be such that $P(A) > 0$ and the random variable $\mathsf{u} : \Omega \to \mathbb{R}$ satisfy $\mathsf{u}(\omega) > 0$ for any $\omega \in A$. Then $\int_A \mathsf{u}(\omega)\mathrm{d}P > 0$.*

Proof For each $n \in \mathbb{N}$, define $A_n := \{\omega \in A \mid \mathsf{u}(\omega) \geqslant \frac{1}{n}\}$, then $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$. Furthermore, as $\mathsf{u}(\omega) > 0$ for $\omega \in A$, we have $A = \bigcup_{n=1}^{\infty} A_n$. The continuity of probability measures then implies $\lim_{n\to\infty} P(A_n) = P(A) > 0$. Thus, there exists an $N \in \mathbb{N}$ such that $P(A_N) > 0$. We thus get $\int_A \mathsf{u}(\omega)\mathrm{d}P = \int_{A_N} \mathsf{u}(\omega)\mathrm{d}P + \int_{A\setminus A_N} \mathsf{u}(\omega)\mathrm{d}P \geqslant \frac{1}{N} \cdot P(A_N) > 0$. ■

Lemma 28 *Let $\mathsf{u} : \Omega \to \mathbb{R}$ be a real-valued random variable and let $A \in \mathscr{F}$ be such that $P(A) > 0$. Then $\mathbb{E}_A[\mathsf{u}] \in \operatorname{co}\mathsf{u}(A)$.*

Proof Write $a = \inf \mathsf{u}(A)$, $b = \sup \mathsf{u}(A)$ and assume that $a \in \mathsf{u}(A)$ and $b \notin \mathsf{u}(A)$—there are three other possibilities to which a similar discussion to the one presented here applies. Then by Lemma 21 we have $\operatorname{co}\mathsf{u}(A) = [a, b)$; so it suffices to show $a \leqslant \mathbb{E}_A[\mathsf{u}] < b$.

As $a = \inf \mathsf{u}(A)$, we have $\mathsf{u}(\omega) \geqslant a$ for all $\omega \in A$. Thus, $P(A) \cdot \mathbb{E}_A[\mathsf{u}] = \int_A \mathsf{u}(\omega)\mathrm{d}P \geqslant a \cdot P(A)$ and hence $\mathbb{E}_A[\mathsf{u}] \geqslant a$. To show that $\mathbb{E}_A[\mathsf{u}] < b$, we define $\mathsf{v} = b - \mathsf{u}$. Then $\mathsf{v} > 0$ is a random variable; and it follows from Lemma 27 that $\mathbb{E}_A[\mathsf{v}] = P(A)^{-1} \cdot \int_A \mathsf{v}(\omega)\mathrm{d}P > 0$. But $\mathbb{E}_A[\mathsf{v}] = b - \mathbb{E}_A[\mathsf{u}]$, we thus get $\mathbb{E}_A[\mathsf{u}] < b$. ■

We now prove Theorem 26, by inducting on the dimensionality m.

Proof The case of $m = 1$ has been established in Lemma 28. Assume that the theorem is true in $\mathbb{R}^{m-1}$; and we want to show that it holds also for $\mathbb{R}^m$. If this is not the case, then there exist a random vector $\mathsf{u} : \Omega \to \mathbb{R}^m$ and a set $A \in \mathscr{A}$ such that $P(A) > 0$ and $\mathbb{E}_A[\mathsf{u}] \notin \operatorname{co}\mathsf{u}(A)$. Without loss of generality, we can, and do, further assume that $\mathbb{E}_A[\mathsf{u}] = 0$ (otherwise we turn to considering the random vector $\mathsf{u}'(\omega) := \mathsf{u}(\omega) - \mathbb{E}_A[\mathsf{u}]$).

As $\mathbb{E}_A[\mathsf{u}] = 0$ is a point not in the *convex* set $\operatorname{co}\mathsf{u}(A)$, there is a hyperplane separating the two—see for example, Boyd and Vandenberghe (2004, Chapter 2.5). That is, there exist $\boldsymbol{w} \in \mathbb{R}^m$ and $c \in \mathbb{R}$ such that $\boldsymbol{w}\cdot\boldsymbol{u}+c \geqslant 0$ for all $\boldsymbol{u} \in \operatorname{co}\mathsf{u}(A)$ and that $\boldsymbol{w}\cdot\mathbf{0}+c = c \leqslant 0$, where $\boldsymbol{w}\cdot\boldsymbol{u}$ denotes the standard inner product of $\boldsymbol{w}$ and $\boldsymbol{u}$. Thus, $\boldsymbol{w}\cdot\boldsymbol{u} \geqslant -c \geqslant 0$ for any $\boldsymbol{u} \in \operatorname{co}\mathsf{u}(A)$. To simplify the discussion, we assume $\boldsymbol{w}$ is the first standard unit vector, $\boldsymbol{w} = [1,0,\ldots,0]$—this can always be obtained by applying a proper rotation operator on the random vector u, so it causes no loss of generality. Under this assumption, the inequality $\boldsymbol{w}\cdot\boldsymbol{u} \geqslant 0$ now reads $u_1 \geqslant 0$, for any $\boldsymbol{u} = [u_1,\ldots,u_m] \in \operatorname{co}\mathsf{u}(A)$.

A side remark: intuitively, the above argument says that, since $\operatorname{co}\mathsf{u}(A)$ *is convex and* $\mathbb{E}_A[\mathsf{u}] \notin \operatorname{co}\mathsf{u}(A)$*, we can first move the origin to the point* $\mathbb{E}_A[\mathsf{u}]$*; then rotate the axes so that* $\operatorname{co}\mathsf{u}(A)$ *lies in the half space* $H_{\geqslant 0} := \{\boldsymbol{u} = [u_1,\ldots,u_m] \in \mathbb{R}^m \mid u_1 \geqslant 0\}$ *after the rotation.*

We return and continue the proof. Define

$$\begin{aligned} H_0 &:= \{\boldsymbol{u} \in \mathbb{R}^m \mid u_1 = 0\}, & A_0 &:= \{\omega \in A \mid \mathsf{u}(\omega) \in H_0\}, \\ H_{>0} &:= \{\boldsymbol{u} \in \mathbb{R}^m \mid u_1 > 0\}, & A_1 &:= \{\omega \in A \mid \mathsf{u}(\omega) \in H_{>0}\}. \end{aligned}$$

Then it is clear that $A_0 \cap A_1 = \varnothing$. Furthermore, from $\mathsf{u}(A) \subseteq \operatorname{co}\mathsf{u}(A) \subseteq H_{\geqslant 0}$ we know $A_0 \cup A_1 = A$. It hence follows from $\mathbb{E}_A[\mathsf{u}] = 0$ that

$$0 = P(A)\cdot\mathbb{E}_A[\mathsf{u}] = \textstyle\int_A \mathsf{u}(\omega)\mathrm{d}P = \int_{A_0} \mathsf{u}(\omega)\mathrm{d}P + \int_{A_1} \mathsf{u}(\omega)\mathrm{d}P. \tag{80}$$

Extracting the first component of this equality results in $\int_{A_1} \mathsf{u}_1(\omega)\mathrm{d}P = 0$. This is because $\mathsf{u}_1(\omega) = 0$ on A_0 and hence $\int_{A_0} \mathsf{u}_1(\omega)\mathrm{d}P = 0$. But $\mathsf{u}_1(\omega) > 0$ for $\omega \in A_1$, so by Lemma 27 we know $P(A_1) = 0$, which in turn implies $\int_{A_1} \mathsf{u}(\omega)\mathrm{d}P = \mathbf{0}$ and $P(A_0) = P(A) > 0$ (as $A = A_0 \cup A_1$). Equation (80) can then be rewritten as $0 = P(A_0)^{-1}\cdot\mathbb{E}_A[\mathsf{u}] = \int_{A_0} \mathsf{u}(\omega)\mathrm{d}P$, that is, $\mathbb{E}_A[\mathsf{u}] = 0 = P(A_0)^{-1}\cdot\int_{A_0} \mathsf{u}(\omega)\mathrm{d}P = \mathbb{E}_{A_0}[\mathsf{u}]$.

On the other hand, $A_0 \subseteq A$ implies $\operatorname{co}\mathsf{u}(A_0) \subseteq \operatorname{co}\mathsf{u}(A)$. So it follows from the assumptions $\mathbb{E}_A[\mathsf{u}] \notin \operatorname{co}\mathsf{u}(A)$ and $\mathbb{E}_A[\mathsf{u}] = \mathbf{0}$ that $\mathbf{0} \notin \operatorname{co}\mathsf{u}(A_0)$. Write $\mathsf{u} = [\mathsf{u}_1,\ldots,\mathsf{u}_m] = [\mathsf{u}_1,\mathsf{v}]$, that is, $\mathsf{v} = [\mathsf{u}_2,\ldots,\mathsf{u}_m]$. As $\mathsf{u}_1(\omega) = 0$ for all $\omega \in A_0$, we have the following "decomposition":

$$\begin{aligned} \mathsf{u}(A_0) &= \{[\mathsf{u}_1(\omega),\mathsf{v}(\omega)] \mid \omega \in A_0\} \\ &= \{(0,\mathsf{v}(\omega)) \mid \omega \in A_0\} \\ &= \{0\} \times \{\mathsf{v}(\omega) \mid \omega \in A_0\} \\ &= \{0\} \times \mathsf{v}(A_0), \end{aligned}$$

and hence $\operatorname{co}\mathsf{u}(A_0) = \{0\} \times \operatorname{co}\mathsf{v}(A_0)$. This fact together with $\mathbf{0} \notin \operatorname{co}\mathsf{u}(A_0)$ implies that $\mathbf{0} \notin \operatorname{co}\mathsf{v}(A_0)$. For the $(m-1)$-dimensional random vector v, the induction hypothesis gives $\mathbb{E}_{A_0}[\mathsf{v}] \in \operatorname{co}\mathsf{v}(A_0)$. It thus follows that $\mathbb{E}_{A_0}[\mathsf{v}] \neq \mathbf{0}$, which further implies $\mathbb{E}_{A_0}[\mathsf{u}] \neq \mathbf{0}$.

We have proved both $\mathbb{E}_{A_0}[\mathsf{u}] = \mathbf{0}$ and $\mathbb{E}_{A_0}[\mathsf{u}] \neq \mathbf{0}$; this contradiction reveals that the assumption $\mathbb{E}_A[\mathsf{u}] \notin \operatorname{co}\mathsf{u}(A)$ must not be true. We thus accomplished the proof. ■

Appendix C. Proofs to the Main Theorems

For those readers who are not satisfied with the presented derivations and who are really enthusiastic about rigorous mathematical proofs, we translate in this section the geometric proofs to the main theorems into the analytical one. We have already done the main job in the preceding section; all we need to do here is to assemble the discussion presented in that section into a proper proof.

C.1 Proof to Theorem 7

By Equations (28), (64) and Theorem 1, we have $[\underline{\mathrm{CSR}}, H(\mathsf{y}|\mathsf{x})] \in \mathrm{co}\,\ell_{\underline{\mathrm{CSR}}}$. It then follows from Equation (68) that

$$2 \cdot h(\tfrac{1}{2c_1}) \cdot \underline{\mathrm{CSR}} \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant [\max\{h(c_1^{-1} \cdot \underline{\mathrm{CSR}}),\, h(1-c_0^{-1} \cdot \underline{\mathrm{CSR}})\}]^{\frown}.$$

This proves Equation (30). Furthermore, the proofs to Theorem 5 and Corollary 6 can be moved straightforwardly here to show the tightness of the two bounds in Equation (30).

C.2 Proof to Theorem 11

By Equations (34), (65) and Theorem 1, we know $[\pi, 2\underline{\mathrm{BER}}, H(\mathsf{y}|\mathsf{x})]$ is in the set $\mathrm{co}\,\ell_{\underline{\mathrm{BER}}}$, which, by Equation (63), can be written as

$$\mathrm{co}\,\ell_{\underline{\mathrm{BER}}} = \{[\eta, r_0, h_0] \mid [\eta, r_0] \in \mathrm{co}\,\ell^{\downarrow}_{\underline{\mathrm{BER}}},\, \underline{g}(\eta, r_0 | \mathrm{co}\,\ell_{\underline{\mathrm{BER}}}) \leqslant h_0 \leqslant \overline{g}(\eta, r_0 | \mathrm{co}\,\ell_{\underline{\mathrm{BER}}})\}.$$

Now fix $\eta = \pi$ in the above expression, and we obtain from Equations (76) and (79) the desired inequality, Equation (44). The tightness of the obtained bounds can be proven similarly to that in Theorem 5 and Corollary 6.

C.3 Theorem 11 is Stronger Than Corollary 8

In this section, we intend to show the upper bound of $H(\mathsf{y}|\mathsf{x})$ as given by Equation (33) is never tighter than that in Equation (44). Mathematically, this amount to proving that

$$\max\{f_3(t) \mid t \in [\pi\tilde{\rho}, \pi\tilde{\rho}+\rho]\} \leqslant \begin{cases} h(\tilde{\pi}\rho) & \text{if } \pi \leqslant 0.5 \\ h(\pi\rho) & \text{if } \pi > 0.5 \end{cases},$$

where $f_3(t) = \tilde{t} \cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) + t \cdot h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho})$, $\rho = 2\underline{\mathrm{BER}} \in [0,1]$, and $h : [0,1] \to \mathbb{R}$ is a symmetric concave function satisfying $h(0) = h(1) = 1$.

To simplify the proof and notation, we shall consider only the case of $\pi \leqslant 0.5$ under an additional condition that the function $h(\cdot)$ is differentiable[17]. For any function as such and any numbers $\eta, \eta_0 \in [0,1]$, by the concavity of $h(\eta)$ we know $h(\eta) \leqslant h(\eta_0) + h'(\eta_0) \cdot (\eta - \eta_0)$. If $\tilde{\pi}\rho \geqslant \frac{1}{2}$, put $\eta_0 = 1 - \tilde{\pi}\rho = \pi + \tilde{\pi}\tilde{\rho} \leqslant \frac{1}{2}$. Since $h(\eta)$ is symmetric and concave, we know $h(\eta_0) = h(\tilde{\pi}\rho)$ and $h'(\eta_0) \geqslant 0$. It then follows that

$$h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) \leqslant h(\eta_0) + h'(\eta_0) \cdot (\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho} - \eta_0) = h(\tilde{\pi}\rho) - h'(\eta_0) \cdot \tilde{\pi}\tilde{\rho} \cdot (1 + \tilde{t}^{-1}\pi),$$
$$h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho}) \leqslant h(\eta_0) + h'(\eta_0) \cdot (\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho} - \eta_0) = h(\tilde{\pi}\rho) - h'(\eta_0) \cdot \tilde{\pi}\tilde{\rho} \cdot (1 - t^{-1}\pi),$$

[17] The case where $\pi > 0.5$ can be discussed similarly. As before, the differentiability assumption is unnecessary: if $h(\cdot)$ is non-differentiable at some point η_0, we can use any number between its right derivative $h'(\eta_0+)$ and left derivative $h'(\eta_0-)$ to replace $h'(\eta_0)$.

and hence $f_3(t) = \tilde{t}\cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) + t\cdot h(\pi + t^{-1}\pi\tilde{\pi}\tilde{\rho}) \leqslant h(\tilde{\pi}\rho) - h'(\eta_0)\cdot\tilde{\pi}\tilde{\rho} \leqslant h(\tilde{\pi}\rho)$ for any $t \in [0,1]$.

Now suppose that $\tilde{\pi}\rho < \frac{1}{2}$. As $h(\cdot)$ is symmetric, by Jensen's inequality we know

$$f_3(t) = \tilde{t}\cdot h(\pi - \tilde{t}^{-1}\pi\tilde{\pi}\tilde{\rho}) + t\cdot h(\tilde{\pi} - t^{-1}\pi\tilde{\pi}\tilde{\rho}) \leqslant h(\tilde{t}\pi + t\tilde{\pi} - 2\pi\tilde{\pi}\tilde{\rho}).$$

For $t \in [\pi\tilde{\rho}, \pi\tilde{\rho}+\rho]$, by direct computation we have $\tilde{t}\pi + t\tilde{\pi} - 2\pi\tilde{\pi}\tilde{\rho} \in [\pi\rho, \tilde{\pi}\rho]$. By $\pi \leqslant 0.5$ we know $\pi\rho \leqslant \tilde{\pi}\rho < \frac{1}{2}$ and so $h(\pi\rho) \leqslant h(\tilde{\pi}\rho)$. Thus $f_3(t) \leqslant h(\tilde{\pi}\rho)$ for any $t \in [\pi\tilde{\rho}, \pi\tilde{\rho}+\rho]$. So far we have proved that $\max\{f_3(t) \mid t \in [\pi\tilde{\rho}, \rho+\pi\tilde{\rho}]\} \leqslant h(\tilde{\pi}\rho)$.

C.4 Proof to Theorem 15

For any binary classification task (μ,η), let $\theta^* = \frac{1}{2}\overline{\mathrm{FSC}}(\mu,\eta)$ and let the function $u(\eta)$ be as in Equation (69). Then $\mathbb{E}_{\mathsf{x}\sim\mu}[u(\eta(\mathsf{x}))] = 0$ and hence Equation (49) holds. It follows from Equation (66) and Theorem 1 that $[\pi, 0, H(\mathsf{y}|\mathsf{x})]$ is in the set $\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}}$. So by Equation (71) we know $\underline{g}(\pi, 0|\,\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}}) \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant \overline{g}(\pi, 0|\,\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}})$, where the range of π is determined by the condition $[\pi, 0] \in \mathrm{co}\,\ell^{\downarrow}_{\overline{\mathrm{FSC}}}$, which by Equation (70) implies $\pi \in [0, \theta^*/\tilde{\theta}^*]$. It thus follows that

$$\inf_{\pi\in[0,\theta^*/\tilde{\theta}^*]} \underline{g}(\pi, 0|\,\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}}) \leqslant H(\mathsf{y}|\mathsf{x}) \leqslant \sup_{\pi\in[0,\theta^*/\tilde{\theta}^*]} \overline{g}(\pi, 0|\,\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}}). \tag{81}$$

By Equation (75), we have $\underline{g}(\pi, 0|\,\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}}) = (\theta^*\tilde{\theta}^*)^{-1}\cdot h(\theta^*)\cdot\pi(\tilde{\theta}^* - \theta^*)$, so the infimum in Equation (81) is 0, which is obtained at $\pi = 0$. Next we will prove briefly that the right hand side of Equation (81) equals to $h(\theta^*/\tilde{\theta}^*) = h\left(\frac{\overline{\mathrm{FSC}}}{2-\overline{\mathrm{FSC}}}\right)$, with the help of Figure 7-b.

For any concave function $h : [0,1] \to \mathbb{R}$ and $\eta_0 \in (0,1)$, it is well known that the left derivative $h'(\eta_0-)$ and the right derivative $h'(\eta_0+)$ exist and satisfy $h'(\eta_0+) \leqslant h'(\eta_0-)$. Moreover, for $\eta_1, \eta_2 \in (0,1)$ with $\eta_1 > \eta_2$, we have $h'(\eta_1-) \leqslant h'(\eta_2+)$. Let $s(\eta_0)$ be a number between $h'(\eta_0+)$ and $h'(\eta_0-)$ and define $f(\eta) := s(\eta_0)\cdot(\eta - \eta_0) + h(\eta_0)$, $\eta \in [0,1]$. As is well known, the affine function $f(\eta)$ satisfies $f(\eta_0) = h(\eta_0)$ and $f(\eta) \geqslant h(\eta)$ for any $\eta \in [0,1]$. Such an affine function is called a *supporting line* of $h(\eta)$ and η_0.

Let $f_1(\eta) = s(\eta_1)\cdot(\eta - \eta_1) + h(\eta_1)$ be a supporting line of $h(\eta)$ at $\eta_1 = \theta^*/\tilde{\theta}^*$. This line intersects with the line $\eta = \theta^*$ at point $\mathrm{K} = [\theta^*, f_1(\theta^*)]$. Through the point K there is a supporting line of $h(\eta)$ at $\eta_2 \leqslant \theta^*$, which we denote as $f_2(\eta) = s(\eta_2)\cdot(\eta - \eta_2) + h(\eta_2)$. As $h(\eta)$ is symmetric and $\eta_2 \leqslant \theta^* \leqslant \frac{1}{2}$, we have $s(\eta_2) \geqslant 0$. Moreover, since $f_1(\theta^*) = f_2(\theta^*)$ and $\eta_2 \leqslant \theta^* \leqslant \eta_1$, by Lemma 18 we know $h(\eta_2) \leqslant h(\eta_1)$.

In Equation (78) let $\eta = \pi$ and relax the resulting expression to

$$\begin{aligned}\overline{g}(\pi, 0|\,\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}}) &\leqslant \sup\{\tilde{t}\cdot f_2(\tilde{t}^{-1}(\pi\tilde{\theta}^* - t\theta^*)) + t\cdot f_1(\theta^* + t^{-1}\pi\theta^*) \mid t \in [\pi\tfrac{\theta^*}{\tilde{\theta}^*}, \pi\tfrac{\tilde{\theta}^*}{\theta^*}]\} \\ &= f_1(\theta^*) - s(\eta_2)\cdot\theta^* + \pi\cdot[s(\eta_1)\theta^* + s(\eta_2)\tilde{\theta}^*] =: f_0(\pi).\end{aligned}$$

Since $\pi \in [0, \theta^*/\tilde{\theta}^*]$ and $f_0(\pi)$ is an affine function, the above inequality further implies

$$\overline{g}(\pi, 0|\,\mathrm{co}\,\ell_{\overline{\mathrm{FSC}}}) \leqslant \max\{f_0(0), f_0(\theta^*/\tilde{\theta}^*)\}.$$

As $s(\eta_2) \geqslant 0$, we know $f_0(0) = f_2(\theta^*) - s(\eta_2)\cdot\theta^* = f_2(0) \leqslant f_2(\eta_2) = h(\eta_2)$. Furthermore,

$$f_0(\theta^*/\tilde{\theta}^*) = f_1(\theta^*) + s(\eta_1)\cdot(\theta^*)^2/\tilde{\theta}^* = f_1\left(\theta^* + (\theta^*)^2/\tilde{\theta}^*\right) = f_1(\eta_1) = h(\eta_1).$$

It thus follows that $f_0(0) \leqslant f_0(\theta^*/\tilde{\theta}^*)$ and so $\overline{g}(\pi,0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}}) \leqslant h(\eta_1) = h(\theta^*/\tilde{\theta}^*)$ for any $\pi \in [0,\theta^*/\tilde{\theta}^*]$. Thus, $\sup_{\pi\in[0,\theta^*/\tilde{\theta}^*]} \overline{g}(\pi,0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}}) \leqslant h(\theta^*/\tilde{\theta}^*)$.

On the other hand, let $\eta = \pi$ and $t = \pi \cdot \tilde{\theta}^*/\theta^*$ in Equation (78), we obtain

$$\overline{g}(\pi,0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}}) \geqslant \tilde{t} \cdot h(0) + t \cdot h(\theta^*/\tilde{\theta}^*) = \pi \cdot \tilde{\theta}^*/\theta^* \cdot h(\theta^*/\tilde{\theta}^*).$$

Thus, $\sup_{\pi\in[0,\theta^*/\tilde{\theta}^*]} \overline{g}(\pi,0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}}) \geqslant \overline{g}(\theta^*/\tilde{\theta}^*,0|\operatorname{co}\ell_{\overline{\mathrm{FSC}}}) \geqslant h(\theta^*/\tilde{\theta}^*)$.

So far, we have finished the proof to Equation (50). The tightness of the two inequalities in Equation (50) can be proven by considering the convex decomposition of the extreme points $\mathrm{O} = [0,0,0]$ (or a point arbitrary close to O) and $\mathrm{E} = [\theta^*/\tilde{\theta}^*, 0, h(\theta^*/\tilde{\theta}^*)]$ in Figure 7-a. The detail is similar to that in Theorem 5 and Corollary 6 and omitted here.

References

C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, 2006.

A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

M. Ben-Bassat. f-entropies, probability of error, and feature selection. *Information and Control*, 39 (3):227–242, 1978.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Camgridge University Press, 2004.

R.P. Brent. *Algorithms for Minimization with Derivatives*. Prentice Hall, 1973.

G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 2012.

R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–78. ACM, 2004.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer, 1996.

W. Duch. *Feature Extraction: Foundation and Applications*, chapter 3, pages 89–117. Springer, 2006. ISBN 3-540-35487-5.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

D. Erdogmus and J.C. Principe. Lower and upper bounds for misclassification probability based on Rényi's information. *Journal of VLSI Signal Processing*, 37:305–317, 2004.

R.M. Fano. *Transmission of Information: a Statistical Theory of Communications*. MIT Press, 1961.

M. Feder and N. Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40:259–266, 1994.

A. Garg and D. Roth. Understanding probabilistic classifiers. In *12th European Conference on Machine Learning (ECML)*, pages 179–191, 2001.

J.D. Golic. On the relationship between the information measures and the bayes probability of error. *IEEE Transactions on Information Theory*, IT-33(5):681–693, 1987.

S. Guiasu. Weighted entropy. *Reports on Mathematical Physics*, 2(3):165–179, 1971.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

M.E. Hellman and J. Raviv. Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Infomation Theory*, IT-16(4):368–372, 1970.

K.E. Hild II, D. Erdogmus, K. Torkkola, and J.C. Principe. Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1385–1392, 2006.

M. Jansche. A maximum expected utility framework for binary sequence labeling. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 736–743, 2007.

T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, February 1967.

R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, special issue on relevance:273–324, 1997.

V.A. Kovalevsky. The problem of character recognition from the point of view of mathematical statistics. In V. A. Kovalevski, editor, *Character Readers and Pattern Recognition*, pages 3–30, New York, 1968.

T.N. Lal, O. Chapelle, J. Weston, and A. Elisseeff. *Feature Extraction: Foundation and Applications*, chapter 5, pages 137–165. Springer, 2006.

D. Lewis. Evaluating and optimizing autonomous text classification systems. In *SIGIR*, pages 246–254, 1995.

R. Linsker. Towards an organizing principle for a layered perceptual network. In *Advances in Neural Information Processing Systems (NIPS)*, volume 0, pages 485–494, 1988.

R. Linsker. An application of the priciple of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems (NIPS)*, volume 1, pages 186–194, 1989.

C.D. Manning, P. Raghavan, and H. Schuetze. *An Introduction to Information Retrieval*. Cambridge University Press, 2008.

G.H. Nguyen, A. Bouzerdoum, and S.L. Phung. Learning pattern classification tasks with imbalanced data sets. In P. Yin, editor, *Pattern recognition*, chapter 10, pages 193–208. Vukovar, Croatia: In-Teh., 2009.

J.C. Principe and D. Xu. Information-theoretic learning using Renyi's quadratic entropy. In *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation*, pages 407–412, 1999.

A. Rényi. On measures of entropy and information. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27: 379–423, 623–656, 1948.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

I.J. Taneja. *Generalized Information Measures and Their Applications*. on-line book, 2001. URL `www.mtm.ufsc.br/ taneja/book/book.html`.

D.L. Tebbe and S.J. Dwyer III. Uncertainty and the probability of error. 14(3):516–518, May 1968.

A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.