entropy

MDPI

Article

An Optimization Approach of Deriving Bounds between Entropy and Error from Joint Distribution: Case Study for Binary Classifications

Bao-Gang Hu [1,*] and Hong-Jie Xing [2]

[1] NLPR/LIAMA, Institute of Automation, Chinese Academy of Science, Beijing 100190, China
[2] College of Mathematics and Information Science, Hebei University, Baoding 071002, China; hjxing@hbu.edu.cn
* Correspondence: hubg@nlpr.ia.ac.cn; Tel.: +86-10-8254-4698

Academic Editors: Badong Chen and Jose C. Principe
Received: 3 December 2015; Accepted: 4 February 2016; Published: 19 February 2016

Abstract: In this work, we propose a new approach of deriving the bounds between entropy and error from a joint distribution through an optimization means. The specific case study is given on binary classifications. Two basic types of classification errors are investigated, namely, the Bayesian and non-Bayesian errors. The consideration of non-Bayesian errors is due to the facts that most classifiers result in non-Bayesian solutions. For both types of errors, we derive the closed-form relations between each bound and error components. When Fano's lower bound in a diagram of "Error Probability *vs*. Conditional Entropy" is realized based on the approach, its interpretations are enlarged by including non-Bayesian errors and the two situations along with independent properties of the variables. A new upper bound for the Bayesian error is derived with respect to the minimum prior probability, which is generally tighter than Kovalevskij's upper bound.

Keywords: entropy; error probability; Bayesian errors; error types; upper bound; lower bound

1. Introduction

In information theory, the relations between entropy and error probability are one of the important fundamentals. Among the related studies, one milestone is Fano's inequality (also known as Fano's lower bound for the error probability of decoders), which was originally proposed in 1952 by Fano but formally published in 1961 [1]. It is well known that Fano's inequality plays a critical role in deriving other theorems and criteria in information theory [2–4]. However, within the research community, it has not been widely accepted exactly who was first to develop the upper bound for the error probability [5]. According to [6,7], Kovalevskij [8] was recognized as the first to derive the upper bound of the error probability in relation to entropy in 1965. Later, several researchers, such as Chu and Chueh in 1966 [9], Tebbe and Dwyer in 1968 [10], Hellman and Raviv in 1970 [11], independently developed upper bounds.

The lower and upper bounds of error probability have been a long-standing topic in studies on information theory [6,7,12–21]. However, we consider two issues that have received less attention in these studies:

I. What are the closed-form relations between each bound and error components in a diagram of entropy and error probability?
II. What are the lower and upper bounds in terms of the non-Bayesian errors if a non-Bayesian rule is applied in the information processing?

The first issue implies a need for a complete set of interpretations to the bounds in relation to joint distributions, so that both error probability and its error components are known for a deeper understanding. We will discuss the reasons of the need in the later sections of this paper. Up to now, most existing studies derived the bounds through an inequality means without using joint distribution information. Therefore, their bounds are not described by a generic relation to joint distributions so that their error component information cannot be gained. Several significant studies have achieved Fano's bound from the joint distributions but through different means [16,20,21]. They all did not show the explicit relations to error components. Regarding the second issue, to the best of our knowledge, it seems that no study is shown in open literature on the bounds in terms of the non-Bayesian errors. We will define the Bayesian and non-Bayesian errors in Section 3. The non-Bayesian errors are also of importance because most classifications are realized within this category.

The issues above form the motivation behind this work. We take binary classifications as a problem background since it is more common and understandable from our daily-life experiences. Moreover, we intend to simplify settings within a binary state and Shannon entropy definitions for a case study from an expectation that the central principle of the approach is well highlighted by simple examples. The novel contribution of the present work is given from the following three aspects:

I. A new approach is proposed for deriving bounds directly through the optimization process based on a joint distribution, which is significantly different from all other existing approaches. One advantage of using the approach is the closed-form expressions to the bounds and their error components.
II. A new upper bound in a diagram of "Error Probability *vs*. Conditional Entropy" for the Bayesian errors is derived with a closed-form expression in the binary state, which has not been reported before. The new bound is generally tighter than Kovalevskij's upper bound. Fano's lower bound receives novel interpretations.
III. The comparison study on the bounds in terms of the Bayesian and non-Bayesian errors are made in the binary state. The bounds of non-Bayesian errors are explored for a first time in information theory and imply a significant role in the study of machine learning and classification applications.

In the first aspect, we also conduct the actual derivation using a symbolic software tool, which presents a standard and comprehensive solution in the approach. The rest of this paper is organized as follows. In Section 2, we present related works on the bounds. For a problem background of binary classifications, several related definitions are given in Section 3. The bounds are given and discussed for the Bayesian and non-Bayesian errors in Sections 4 and 5, respectively. Interpretations to some key points are presented in Section 6. We summarize the work in Section 7 and present some discussions in Section 8. The source code from using symbolic software for the derivation is included in Figures A1 and A2.

2. Related Works

Two important bounds are introduced first, which form the baselines for the comparisons with the new bounds. They were both derived from inequality conditions [1,8]. Suppose the random variables X and Y representing input and output messages (out of m possible messages), and the conditional entropy $H(X|Y)$ representing the average amount of information lost on X when given Y [22]. Fano's lower bound for the error probability [1,22] is given in a form of:

$$H(X|Y) \leq H(P_e) + P_e \log_2(m-1), \tag{1}$$

where P_e is the *error probability* (sometimes, also called *error rate* or *error* for short), and $H(P_e)$ is the binary entropy function defined by [23]:

$$H(P_e) = -P_e \log_2 P_e - (1 - P_e) \log_2 (1 - P_e). \tag{2}$$

The base of the logarithm is two so that the units are *bits*.

The upper bound for the error probability is given by Kovalevskij [8] in a piecewise linear form [10]:

$$H(X|Y) \geq \log_2 k + k(k+1)(\log_2 \frac{k+1}{k})(P_e - \frac{k-1}{k}) \text{ and } k < m,\ m \geq 2, \tag{3}$$

where k is a positive integer number, but defined to be smaller than m. For a binary classification ($m = 2$), Fano–Kovalevskij bounds become:

$$H^{-1}(P_e) = G(H(X|Y)) \leq P_e \leq \frac{H(X|Y)}{2}, \tag{4}$$

where $H^{-1}(P_e)$ denotes an inverse function of $H(P_e)$ and has no closed-form expression. Hence, we set it as a function form, $G(H(X|Y))$, in terms of the variable $H(X|Y)$. Feder and Merhav [24] depicted bounds of Equation (4) and presented interpretations on the two specific points from the background of data compression problems.

Studies from the different perspectives have been reported on the bounds between error probability and entropy. The initial difference is made from the entropy definitions, such as Shannon entropy in [12,14,25,26], and Rényi entropy in [6,7,15]. The second difference is the selection of bound relations, such as "P_e *vs.* $H(X|Y)$" in [12,24], "$H(X|Y)$ *vs.* P_e" in [6,7,14,15,20], "P_e *vs.* $MI(X;Y)$" in [27,28], and "$NMI(X;Y)$ *vs.* A" in [25], where A is the accuracy rate, $MI(X;Y)$ and $NMI(X;Y)$ are the mutual information and normalized mutual information between variables X and Y, respectively. Another important study is made on the tightness of bounds. Several investigations [17,19,20,29] have been reported on the improvement of bound tightness. Recently, a study in [26] suggested that an upper bound from the Bayesian errors should be added, which is generally neglected in the bound analysis.

3. Binary Classifications and Related Definitions

Classifications can be viewed as one component in pattern recognition systems [30]. Figure 1 shows a schematic diagram of the pattern recognition systems. The first unit in the systems is termed *representation* in the present problem background but called *encoder* in communication background. This unit processes the tasks of *feature selection*, or *feature extraction*. The second unit is called *classification* or *classifier* in applications. Three sets of variables are involved in the systems, namely, *target variable* T, *feature variables* X, and *prediction variable* Y. While T and Y are univariate discrete random variables for representing labels of the samples, X can be high-dimensional random variables either in forms of discrete, continuous, or their combinations.

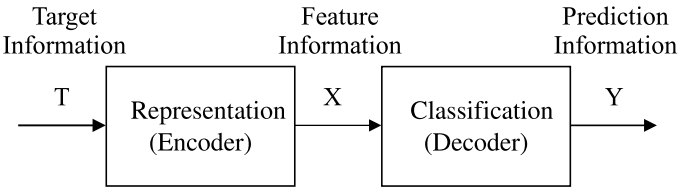

Figure 1. Schematic diagram of the pattern recognition systems (adapted from Figure 1.7 in [30]).

In this work, binary classifications are considered as a case study because they are more fundamental in applications. Sometimes, multi-class classifications are processed by binary classifiers [31]. In this section, we will present several necessary definitions for the present case study.

Let $\mathbf{x}$ be a random sample satisfying $\mathbf{x} \in \mathcal{X} \subset R^d$, which is in a d-dimensional feature space and will be classified. The true (or target) state t of $\mathbf{x}$ is within the finite set of two classes, $t \in \mathcal{T} = \{t_1, t_2\}$, and the prediction (or output) state $y = f(\mathbf{x})$ is within the two classes, $y \in \mathcal{Y} = \{y_1, y_2\}$, where f is a function for classifications. Let $p(t_i)$ be the *prior probability* of class t_i and $p(\mathbf{x}|t_i)$ be the *conditional probability density function* (or *conditional probability*) of $\mathbf{x}$ given that it belongs to class t_i.

Definition 1. *(Bayesian error in binary classification)* In a binary classification, the *Bayesian error*, denoted by P_e, is defined by [30]:

$$P_e = \int_{R_2} p(\mathbf{x}|t_1)p(t_1)d\mathbf{x} + \int_{R_1} p(\mathbf{x}|t_2)p(t_2)d\mathbf{x}, \tag{5}$$

where R_i is the *decision region* for class t_i. The two regions are determined by the Bayesian rule:

$$\begin{cases} \text{Decide } R_1 \text{ if } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \geq 1 \\ \text{Decide } R_2 \text{ if } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} < 1 \end{cases}. \tag{6}$$

In statistical classifications, the Bayesian error is the *theoretically lowest* probability of error [30].

Definition 2. *(Non-Bayesian error)* The *non-Bayesian error*, denoted by P_E, is defined to be any error which is larger than the Bayesian error, that is:

$$P_E > P_e, \tag{7}$$

for the given information of $p(t_i)$ and $p(\mathbf{x}|t_i)$.

Remark 1. Based on the definitions above, for the given joint distribution, the Bayesian error is unique, but the non-Bayesian errors are multiple. Figure 2 shows the Bayesian *decision boundary*, x_b, on a univariate feature variable x for equal priors. The Bayesian error is $P_e = e_1 + e_2$. Any other decision boundary different from x_b will generate the non-Bayesian error for $P_E > P_e$.

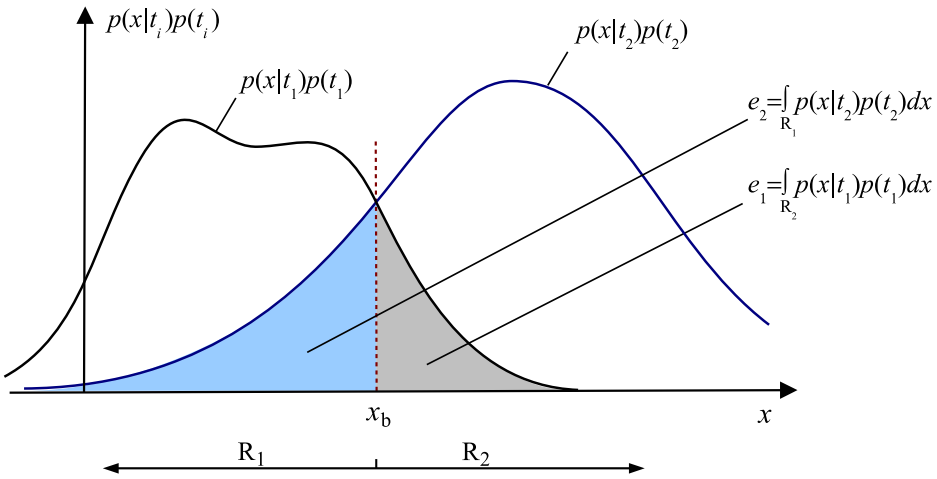

Figure 2. Bayesian decision boundary x_b for equal priors $p(t_i)$ in a binary classification (adapted from Figure 2.17 in [30]).

In a binary classification, the *joint distribution*, $p(t,y) = p(t = t_i, y = y_j) = p_{ij}$, is given in a general form of:

$$\begin{aligned} p_{11} &= p_1 - e_1, \; p_{12} = e_1, \\ p_{21} &= e_2, \quad p_{22} = p_2 - e_2, \end{aligned} \tag{8}$$

where $p_1 = p(t_1)$ and $p_2 = p(t_2)$ are the prior probabilities of Class 1 and Class 2, respectively; their associated errors (also called *error components*) are denoted by e_1 and e_2. Figure 3 shows a graphic diagram of the probability transformation between target variable T and prediction variable Y via their joint distribution $p(t,y)$ in a binary classification. The constraints in Equation (8) are given by [30]:

$$\begin{aligned} 0 < p_1 < 1, \; 0 < p_2 < 1, \; p_1 + p_2 = 1 \\ 0 \le e_1 \le p_1, \; 0 \le e_2 \le p_2. \end{aligned} \tag{9}$$

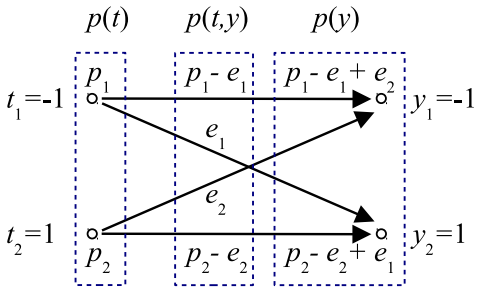

Figure 3. Graphic diagram of the probability transformation between variables T and Y in a binary classification (or channel). Instead of using *conditional probability* $p(y|t)$, *joint probability distributions* $p(t,y)$ are applied to describe the channel.

In this work, we use e to denote error probability, or error variable, for representing either the Bayesian error or non-Bayesian error. They are calculated from the same formula:

$$e = e_1 + e_2 = \begin{cases} P_e & if\ e\ is\ the\ minimum, \\ P_E & otherwise. \end{cases} . \tag{10}$$

Definition 3. *(Minimum and maximum error bounds in binary classifications)* Classifications suggest the minimum error bound as:

$$(P_E)_{min} = (P_e)_{min} = 0, \tag{11}$$

where the subscript *min* denotes the minimum value. The maximum error bound for the Bayesian error in binary classifications is [26]:

$$(P_e)_{max} = p_{min} = \min\{p_1, p_2\}, \tag{12}$$

where the symbol *min* denotes a *minimum* operation. For the non-Bayesian error, its maximum error bound becomes

$$(P_E)_{max} = 1. \tag{13}$$

The Equations from Equations (11) to (13) describe the *initial* ranges of Bayesian and non-Bayesian errors respectively. When they share the same minimum, their maximums are always different.

Remark 2. For a given set of joint distributions in the bound studies, one may fail to tell if it is the solution from using the Bayesian rule or not. Only when $e > p_{min}$, we can say the set is corresponding to the non-Bayesian solution.

In a binary classification, the *conditional entropy*, $H(T|Y)$, is calculated from the joint distribution in Equation (8):

$$\begin{aligned} H(T|Y) &= H(T) - MI(T;Y) \\ &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - e_1 \log_2 \frac{e_1}{(p_2 + e_1 - e_2)p_1} - e_2 \log_2 \frac{e_2}{(p_1 - e_1 + e_2)p_2} \\ &\quad -(p_1 - e_1) \log_2 \frac{(p_1 - e_1)}{(p_1 - e_1 + e_2)p_1} - (p_2 - e_2) \log_2 \frac{(p_2 - e_2)}{(p_2 + e_1 - e_2)p_2}, \end{aligned} \tag{14}$$

where $H(T)$ is a *binary entropy* of the random variable T, and $MI(T;Y)$ is *mutual information* between variables T and Y.

Remark 3. When a joint distribution $p(t,y)$ is given, its associated conditional entropy $H(T|Y)$ is uniquely determined. However, for the given $H(T|Y)$, it is generally unable to reach a unique solution to $p(t,y)$ but receives multiple solutions shown later in this work.

Definition 4. *(Admissible point, admissible set, and their properties in diagram of entropy and error probability)* In a given diagram of entropy and error probability, if a point in the diagram is possibly to be realized from a non-empty set of joint distributions for the given classification information, it is defined to be an *admissible point*. Otherwise, it is a *non-admissible point*. All admissible points will form an *admissible set* (or *admissible region(s)*), which is enclosed by the bounds (also called *boundary*). If every point located on the boundary is admissible (or non-admissible), we call this admissible set *closed* (or *open*). If only a partial portion of boundary points is admissible, the set is said to be *partially closed*. For an admissible point with the given conditions, if it is realized only by a unique joint distribution, it is called a *one-to-one mapping* point. If more than one joint distribution is associated to the same admissible point, it is called a *one-to-many mapping* point.

We consider that classifications present an exemplary justification of raising the first issue in Section 1 about the bound studies. The main reason behind the issue is that a single index of error probability may not be sufficient for dealing with classification problems. For example, when processing class-imbalance problems [32,33], we need to distinguish *error types*. In other words, for the same error probability e (or even the same admissible point), we are required to know the error components of e_1 and e_2 as well. Suppose one encounters a medical diagnosis problem, where p_1 (say, $p_1 = 0.98$) generally represents the *majority class* for *healthy* persons (labeled with *negative* or -1 in Figure 3), and p_2 ($= 0.02$) the *minority class* for *abnormal* persons (labeled with *positive* or 1). A class-imbalance problem is then formed. While e_1 (also called *type I error*) is tolerable (say, $e_1 = 0.01$), e_2 (or *type II error*) seems intolerable (say, $e_2 = 0.01$) because abnormal persons are considered to be "*healthy*". In class-imbalance problems, the performance measure from error probability may become useless. For example, a classification having $e = e_2 = p_2 = 0.02$ does not support a high, yet reasonable, performance. Hence, from either theoretical or application viewpoint, it is necessary for establishing relations between bounds and joint distributions, which can provide error type information within error probability for better interpretations to the bounds.

4. Lower and Upper Bounds for Bayesian Errors

In this work, we select the bound relations between entropy and error probability. Furthermore, the bounds and their associated error components are also given by the following two theorems in a context of binary classifications.

Theorem 1. *(Lower bound and associated error components) The lower bound in a diagram of "P_e vs. $H(T|Y)$" and the associated error components with constraints Equations (9) and (12) are given by:*

$$P_e \geq \max\{0, G_1(H(T|Y))\}, \tag{15a}$$

$$\begin{aligned} \text{for } G_1^{-1}(P_e) \quad &= -P_e \log_2 P_e - (1-P_e)\log_2(1-P_e), \\ P_e \quad &= e_1 + e_2 \leq p_{min}, \end{aligned} \tag{15b}$$

$$(e_1, e_2) = \begin{cases} (0.5, 0) \text{ or } (0, 0.5), & \text{if } P_e = 0.5, \\ (\frac{P_e(1-p_1-P_e)}{1-2P_e}, \frac{P_e(p_1-P_e)}{1-2P_e}), & \text{otherwise}, \end{cases} \tag{15c}$$

where $H(T|Y)$ is the conditional entropy of T when given Y, and G_1 is called the lower bound function (or lower bound). However, one can only achieve the closed-form solution on its inverse function, $G_1^{-1}(\cdot)$, not on $G_1(\cdot)$ itself.

Proof. Based on Equation (14), the lower bound function is derived from the following definition:

$$\begin{aligned} G_1^{-1}(e) &= \arg\max_e H(T|Y) \\ &\text{subject to Equations (9) and (12)}, \end{aligned} \tag{16}$$

where we take e for the input variable in the derivations Equation (16) describes the function of the maximum $H(T|Y)$ with respect to e, and the function needs to satisfy the general constraints of joint distributions in Equation (9). $H(T|Y)$ seems to be governed by the four variables from p_i and e_i in Equation (14). However, only two independent parameter variables determine the solutions of Equations (14) and (16). The variable reduction from four to two is due to the two specific constrains imposed between parameters, that is, $p_1 + p_2 = 1$ and $e_1 + e_2 = e$. When we set p_1 and e_1 as two independent variables, (16) is then equivalent to solving the following problem:

$$\begin{aligned} G_1^{-1}(p_1, e_1) &= \arg\max_{e=P_e} H(T|Y) \\ &\text{subject to Equations (9) and (12)}. \end{aligned} \tag{17}$$

$G_1^{-1}(p_1, e_1)$ is a continuous and differentiable function with respect to the two variables. A differential approach is applied analytically for searching the *critical points* of the optimizations in Equation (17). We achieve the two differential equations below and set them to be zeros:

$$\begin{cases} \frac{\partial H(T|Y)}{\partial e_1} = \log_2 \frac{(p_1-e_1)(P_e-e_1)(1+2e_1-p_1-P_e)^2}{e_1(1+e_1-p_1-P_e)(p_1+P_e-2e_1)^2} = 0, \\ \frac{\partial H(T|Y)}{\partial p_1} = \log_2 \frac{(p_1-2e_1+P_e)(1+e_1-p_1-P_e)}{(p_1-e_1)(1+2e_1-p_1-P_e)} = 0. \end{cases} \tag{18}$$

By solving them simultaneously, we obtain the three pairs of the critical points through analytical derivations:

$$\begin{cases} e_1 = \frac{P_e(1-p_1-P_e)}{1-2P_e}, \\ p_1 = \frac{P_e+2e_1P_e-e_1-P_e^2}{P_e}, \end{cases} \tag{19a}$$

$$\begin{cases} e_1 = \frac{p_1(p_1+P_e-1)}{2p_1-1}, \\ p_1 = \frac{1-P_e}{2} + e_1 + \frac{1}{2}\sqrt{1+P_e^2+4e_1^2-4e_1P_e-2P_e}, \end{cases} \tag{19b}$$

$$\begin{cases} e_1 = \frac{p_1(p_1+P_e-1)}{2p_1-1}, \\ p_1 = \frac{1-P_e}{2} + e_1 - \frac{1}{2}\sqrt{1+P_e^2+4e_1^2-4e_1P_e-2P_e}. \end{cases} \tag{19c}$$

The highest order of each variable, e_1 and p_1, in Equation (18) is four. However, we can see the quadratic component within the first function in Equation (18), $(\frac{1+2e_1-p_1-P_e}{p_1+P_e-2e_1})^2$, will degenerate the total solution order from four to three. Therefore, the three pairs of critical points exhibit a complete set of *possible solutions* to the problem in Equation (17). The *final solution* should be the pair(s) that satisfies both the maximum $H(T|Y)$ with respect to e_1 for the given $e = P_e$ and the Equations constraints (9) and (12). Due to high complexity of the nonlinearity of the second-order partial differential equations on $H(T|Y)$, it seems intractable to examine the three pairs analytically for the final solution.

To overcome the difficulty above, we apply a symbolic software tool, Maple^{TM}9.5 (a registered trademark of Waterloo Maple, Inc.), for a *semi-analytical* solution to the problem (see Maple code in Figure A1). For simplicity and without loss of generality in classifications, we consider p_1 and P_e are known constants in the function. The concavity property of $H(T|Y)$ with respect to e_1 in the ranges defined in Equation (19a) is confirmed numerically by varying data on p_1 and P_e. Hence, a maximum solution on $H(T|Y)$ is always received from the possible solutions of the critical points. Among them, only Equation (19a) satisfies the constraints to be the final solution. When e_1 is set, the expression of e_2 is known as shown in Equation (15c). The singular case is given specifically and the solution of $(e_1, e_2) = (0, 0.5)$ is obtained when p_2 is used in the error expressions. □

Remark 4. Theorem 1 achieves the same lower bound found by Fano [1] (Figure 4), which is general for finite alphabets (or multiclass classifications). One specific relation to Fano's bound is given by the *marginal probability* (see (2-144) in [2]):

$$p(y) = (1 - P_e, \tfrac{P_e}{m-1}, \ldots, \tfrac{P_e}{m-1}), \tag{20}$$

which is termed *sharp* for attaining equality in Equation (1) [2]. We call Fano's bound an *exact* lower bound because every point on it is sharp. The sharp conditions in terms of error components in Equation (15c) are a special case of the study in [20], and can be derived directly from their Theorem 1.

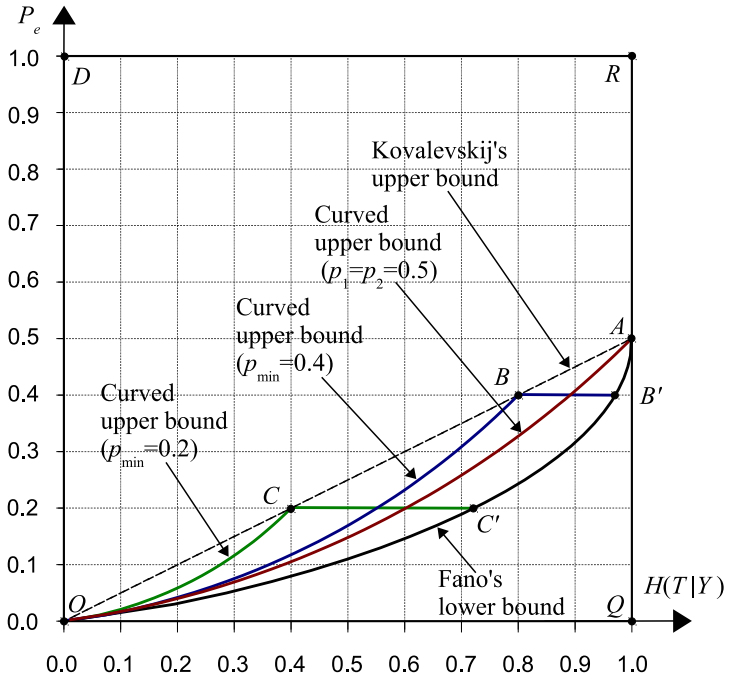

Figure 4. Plot of bounds in a "P_e *vs.* $H(T|Y)$" diagram.

Theorem 2. *(Upper bound and associated error components) The upper bound and the associated error components with constraints Equations (9) and (12) are given by:*

$$P_e \leq \min\{p_{min}, G_2(H(T|Y))\}, \tag{21a}$$

$$for\ G_2^{-1}(P_e) \quad = -p_{min}\log_2\tfrac{p_{min}}{P_e+p_{min}} - P_e\log_2\tfrac{P_e}{P_e+p_{min}}, \tag{21b}$$

$$\begin{aligned} and\ P_e &= e_1 + e_2 \leq p_{min}, \\ e_i &= p_j,\ e_j = 0,\ p_i \geq p_j,\ i \neq j,\ i,j = 1,2, \end{aligned} \tag{21c}$$

where G_2 is called the upper bound function (or upper bound). The closed-form solution can be achieved only on its inverse function, $G_2^{-1}(\cdot)$.

Proof. The upper bound function is obtained from solving the following equation:

$$\begin{aligned} G_2^{-1}(p_1, e_1) &= \arg\min_{e=P_e} H(T|Y), \\ &subject\ to\ Equations\ (9)\ \text{and}\ (12). \end{aligned} \tag{22}$$

Because the concavity property holds for $H(T|Y)$ with respect to e_1 as discussed in the proof of Theorem 1, the possible solutions of e_1 should be located at the two ending points, that is, either at $e_1 = 0$ or at $e_1 = P_e$. We can take the point which produces the smaller $H(T|Y)$ and satisfies the constraints as the final solution. The solution from Maple code in Figure A2 confirms the closed-form expressions in (21). □

Remark 5. Theorem 2 describes a novel set of upper bounds which is in general *tighter* than Kovalevskij's bound [8] for binary classifications (Figure 4). For example, when $p_{min} = 0.2$ is given, the upper bounds defined in Equation (21) shows a curve "$O - C$" plus a line "$C - C'$". Kovalevskij's upper bound, given by a line "$O - C - A$", is sharp only at Point O and Point C. The solution in Equation (21c) confirms an advantage of using the proposed optimization approach in derivations so that a closed-form expression of the exact bound is possibly achieved.

In comparison, Kovalevskij's upper bound described in Equation (3) is general for multiclass classifications. This bound misses a general relation to error components like Equation (21c), although the relation is restricted to a binary state. For distinguishing from the Kovalevskij's upper bound, we also call G_2 a *curved upper bound*. The new *linear upper bound*, $(P_e)_{max} = p_{min}$, shows the maximum error for the Bayesian decisions in binary classifications [26], which is also equivalent to the solution of a blind guess when using the maximum-likelihood decision [30]. If $p_1 = p_2$, the upper bound becomes a single curved one.

Remark 6. The lower and upper bounds defined by Equations (15) and (21) form a closed admissible region in the diagram of "P_e *vs.* $H(X|Y)$". The shape of the admissible region changes depending on a single parameter of p_{min}.

5. Lower and Upper Bounds for Non-Bayesian Errors

In classification problems, the Bayesian errors can be realized only if one has the exact information about all probability distributions of classes. The assumption above is generally impossible in real applications. In addition, various classifiers are designed by employing the non-Bayesian rules or resulted in non-Bayesian errors, from the conventional decision trees, artificial neural networks, and supporting vector machines [30], to the emerging deep learning [34]. Therefore, the analysis of the non-Bayesian errors presents significant interests in classification studies, although the conventional information theory does not distinguish the error types.

Definition 5. *(Label-switching in binary classifications)* In binary classifications, a label-switching operation is an exchange between two labels. Suppose the original joint distribution is denoted by:

$$p_A(t,y): \quad \begin{array}{l} p_{11}=a,\ p_{12}=b, \\ p_{21}=c,\ p_{22}=d. \end{array} \tag{23a}$$

A label-switching operation will change the prediction labels in Figure 3 to be $y_1 = 1$ and $y_2 = -1$, and generate the following joint distribution:

$$p_B(t,y): \quad \begin{array}{l} p_{11}=b,\ p_{12}=a, \\ p_{21}=d,\ p_{22}=c. \end{array} \tag{23b}$$

Proposition 1. *(Invariant property from label-switching) The related entropy measures, including $H(T)$, $H(Y)$, $MI(T;Y)$, and $H(T|Y)$, will be invariant to labels, or unchanged from a label-switching operation in binary classifications. However, the error e will be changed to be $1-e$.*

Proof. Substituting the two sets of joint distributions in Equation (23) into each entropy measure formula respectively, one can obtain the same results. The error change is obvious. □

Theorem 3. *(Lower bound and upper bound for non-Bayesian error without information of p_1 and p_2) In a context of binary classifications, when information about p_1 and p_2 is unknown (say, before classifications), the lower bound and upper bound for the non-Bayesian error with constraints Equations (9) and (13) are given by:*

$$G_1(H(T|Y)) \le P_E \le 1 - G_1(H(T|Y)), \tag{24a}$$

$$\begin{array}{rl} for\ G_1^{-1}(P_E) & = -P_E \log_2 P_E - (1-P_E)\log_2(1-P_E), \\ P_E & = e_1 + e_2 \le 1, \end{array} \tag{24b}$$

$$(e_1,e_2) = \begin{cases} (0.5,0)\ \ or\ \ (0,0.5), & if\ p_1 = p_2 = P_E = 0.5, \\ (\frac{P_E(1-p_1-P_E)}{1-2P_E}, \frac{P_E(p_1-P_E)}{1-2P_E}), & if\ (1-p_1-P_E)(p_1-P_E) \ge 0 \\ (\frac{p_1(p_1+P_E-1)}{2p_1-1}, \frac{(1-p_1)(p_1-P_E)}{2p_1-1}), & otherwise, \end{cases} \tag{24c}$$

where we call the upper bound in Equation (24a), $1 - G_1(H(T|Y))$, the general upper bound (or mirrored lower bound), which is a mirror of Fano's lower bound with the mirror axis along $P_E = 0.5$. Both bounds share the same expression for calculating the associated error components in Equation (24c). When $P_E \le 0.5$, their components, e_1 and e_2, correspond to the lower bound, otherwise, to the upper bound.

Proof. When the error probability is relaxed by Equation (13), all possible solutions in Equation (19) are applicable but within the special ranges respectively. Suppose an admissible point is located at the lower bound which shows $P_E \le 0.5$. By a label-switching operation, one can obtain the mirrored admissible point at $1 - P_E \ge 0.5$, which is located at the mirrored lower bound. Proposition 1 suggests both points share the same value of $H(T|Y)$. Because P_E is the smallest one for the given conditional entropy $H(T|Y)$, its mirrored point is the biggest one for creating the general upper bound. □

Remark 7. Han and Verdù [16] achieved Fano's bound from the joint distributions by including the independent condition $p_{ij} = p(t_i)p(y_j)$ [2]. The condition will only lead to the last set of error equations in Equation (24c), not to the complete sets. In addition, the set is only applicable to the non-Bayesian errors, not to the Bayesian ones except for a special case in Equation (20). Equation (24c) confirms again the advantage of using the optimization in derivations which achieves the complete sets of solutions to describe Fano's bound for non-Bayesian errors.

Remark 8. The bounds from Equation (24) are derived only when p_1 and P_e are given. They exist even one does not have such information. In this situation, Fano's lower bound, its mirror bound, and the axis of P_E form an admissible region, denoted by a boundary "$O - F' - A - F - D - O$" in Figure 5. The axis of P_E encloses the region, but only Points O and D are admissible. Hence, the admissible region is partially closed.

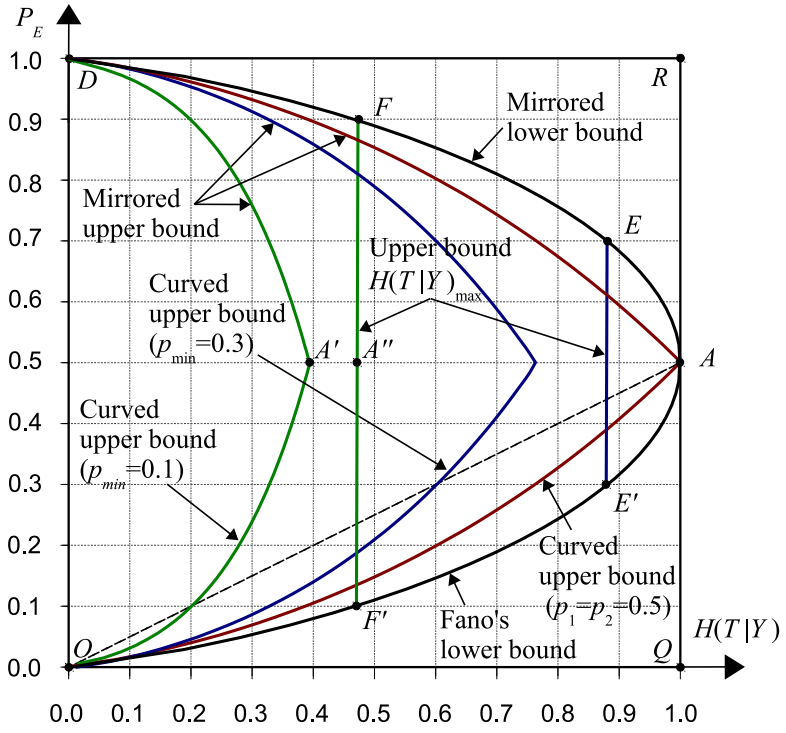

Figure 5. Plot of bounds in a "P_E vs. $H(T|Y)$" diagram.

Theorem 4. *(Admissible region(s) for non-Bayesian error with known information of p_1 and p_2) In binary classifications, when information about p_1 and p_2 is known, a closed admissible region for the non-Bayesian error with constraints Equations (9) and (13) is generally formed (Figure 5) by Fano's lower bound, the general upper bound, the curved upper bound $G_2^{-1}(\cdot)$, the mirrored upper bound of $G_2^{-1}(\cdot)$, and the upper bound $H(T|Y)_{max}$. For the $H(T|Y)_{max}$ bound, its associated error components are given by:*

$$(e_1, e_2) = \begin{array}{c} for\ H(T|Y) = H(T|Y)_{max} = H(p_{min}), \\ \begin{cases} (0.25, 0.25), & if\ p_1 = p_2 = P_E = 0.5, \\ (\frac{p_1(1-p_1-P_E)}{1-2p_1}, \frac{P_E(1-p_1)-p_1(1-p_1)}{1-2p_1}), & otherwise. \end{cases} \end{array} \tag{25}$$

Proof. Following the proof in Theorem 3, one can get the mirrored upper bound of $G_2^{-1}(\cdot)$. The upper bound $H(T|Y)_{max}$ is calculated from the condition of $H(T|Y) \leq H(T)$ [2]. For the given p_1 and p_2, $H(T|Y)_{max}$ is a constant. Because $H(T|Y)_{max}$ also implies a minimization of $MI(T;Y)$ in Equation (14), its associated error components can be obtained from the following equivalent relation (see (11) in [35]):

$$MI(T;Y) = 0 \ \leftrightarrow \ \frac{p_{11}}{p_{21}} = \frac{p_{12}}{p_{22}}. \tag{26}$$

□

Remark 9. Equations (25) and (26) equivalently represent a zero value for the mutual information, which suggests *no correlation* [30] or *statistically independent* [2] between two variables T and Y.

Remark 10. When information of p_1 and p_2 is known, the admissible region(s) is much compact than that when without such information. The shape of the admissible region(s) is fully dependent on a single parameter p_{min}. For example, if $p_{min} = 0.1$, the area is enclosed by the four-curve-one-line boundary "$O - F' - F - D - A' - O$" in Figure 5. However, if $p_1 = p_2 = 0.5$, two admissible areas are formed. They are "$O - F' - A - O$" and "$D - F - A - D$", respectively.

6. Classification Interpretations to Some Key Points

For a better understanding of the theoretical insights between the bounds and errors, some key points shown in Figures 4 and 5 are discussed. Those key points may hold special features in classifications. Novel interpretations are expected from the following discussions.

Point O: This point represents a zero value of $H(T|Y)$. It also suggests an *exact classification* without any error ($P_e = P_E = 0$) by a specific setting of the joint distribution:

$$\begin{aligned} p_{11} = p_1,\ p_{12} = 0, \\ p_{21} = 0,\ \ p_{22} = p_2. \end{aligned} \tag{27}$$

This point is always admissible and independent of error types.

Point A: This point shows the maximum ranges of $H(T|Y) = 1$ for *class-balanced* classifications ($p_1 = p_2$). Three specific classification settings can be obtained for representing this point. The two settings from Equation (24c) are actually *no classification*:

$$\begin{aligned} p_{11} = 1/2,\ p_{12} = 0, \\ p_{21} = 1/2,\ p_{22} = 0, \end{aligned} \quad \text{or} \quad \begin{aligned} p_{11} = 0,\ p_{12} = 1/2, \\ p_{21} = 0,\ p_{22} = 1/2. \end{aligned} \tag{28}$$

They also indicate *zero information* [36] from the classification decisions. The other setting is a *random guessing* from Equation (25):

$$\begin{aligned} p_{11} = 1/4,\ p_{12} = 1/4, \\ p_{21} = 1/4,\ p_{22} = 1/4. \end{aligned} \tag{29}$$

For the Bayesian errors, this point is always included by both Fano's bound and Kovalevskij's bound. However, according to the upper bounds defined in Equation (21a), this point is non-admissible whenever the relation $p_1 \neq p_2$ holds. For the non-Bayesian errors, the point is either admissible or non-admissible depending on the given information about p_1 and p_2. This example suggests that the admissible property of a point should generally rely on the given information in classifications.

Point D: This point occurs for the non-Bayesian classifications in a form of:

$$\begin{aligned} p_{11} = 0,\ \ p_{12} = p_1, \\ p_{21} = p_2,\ p_{22} = 0. \end{aligned} \tag{30}$$

In this case, one can exchange the labels for a perfect classification.

Point B: This point is located at the corner formed by the curved and linear upper bounds, with $H(T|Y) = 0.8$ and $e = 0.4$. In apart from Point O, this is another point obtained from Equation (21) that locates at Kovalevskij's upper bound. The point can be realized from either

Bayesian or non-Bayesian classifications. Suppose $p_1 > p_2 = 0.4$ for the Bayesian classifications. One will achieve Point B by Equation (21):

$$\begin{array}{l} p_{11} = 0.2,\ p_{12} = 0.4, \\ p_{21} = 0.0,\ p_{22} = 0.4, \end{array} \tag{31}$$

for a one-to-one mapping. In other words, this point is uniquely determined by Equation (31) and only corresponding to $p_{min} = 0.4$ within the Bayesian classifications. If non-Bayesian classifications are considered, this point becomes a one-to-many mapping and shows $p_{min} \neq 0.4$. For example, one can get another setting of joint distribution from solving $H(p_{min}) = 0.8$ for $p_{min} = 0.2430$ first. Then, by substituting the relations of $p_2 = p_{min}$ and $P_E = 0.4$ into Equation (25), one can get the error components, that is, $e_1 = 0.2312$ and $e_2 = 0.1688$, for the new setting of joint distribution on Point B.

Point B becomes non-admissible when $p_{min} = 0.5$ (Figure 4), which means no joint distribution exists to satisfy Equation (9). In this situation, we can understand why the new upper bound is generally tighter than Kovalevskij's upper bound.

Point B': The point is with $H(T|Y) = 0.9710$ and $e = 0.4$ in the diagram (Figure 4). It is exactly located at the lower bound and is able to produce a one-to-many mapping for either the Bayesian errors or non-Bayesian errors. One specific setting in terms of the Bayesian errors is:

$$\begin{array}{l} p_{11} = 0.6,\ p_{12} = 0.0, \\ p_{21} = 0.4,\ p_{22} = 0.0, \end{array} \tag{32}$$

which suggests zero information from classifications. More settings can be obtained from Equation (15). For example, if given $p_1 = 0.55$, $p_2 = 0.45$ and $P_e = 0.4$, one can have:

$$\begin{array}{l} p_{11} = 0.45,\ \ p_{12} = 0.10, \\ p_{21} = 0.30,\ \ p_{22} = 0.15. \end{array} \tag{33}$$

Another setting is for the balanced error components:

$$\begin{array}{l} p_{11} = 0.3,\ p_{12} = 0.2, \\ p_{21} = 0.2,\ p_{22} = 0.3. \end{array} \tag{34}$$

The non-Bayesian errors will enlarge the set of one-to-many mapping for an admissible point due to the relaxed condition of Equation (13). Equation (24c) will be applicable for deriving a specific setting when p_1 and e are given. For example, two settings can be obtained:

$$\begin{array}{ll} if & p_1 = 0.250,\ \ P_E = 0.400, \\ then & p_{11} = 0.075,\ p_{12} = 0.175, \\ & p_{21} = 0.225,\ p_{22} = 0.525, \end{array} \tag{35}$$

$$\begin{array}{ll} if & p_1 = 0.300,\ \ P_E = 0.400, \\ then & p_{11} = 0.075,\ p_{12} = 0.225, \\ & p_{21} = 0.175,\ p_{22} = 0.525, \end{array} \tag{36}$$

for representing the same Point B'.

Remark 11. One can observe that Equations (35) and (36) will lead to a zero mutual information, but Equations (33) and (34) are not. The observations reveal new interpretations about Fano's bound in association with two situations in the independent properties of the variables, which have not been reported before.

Points E and E′: All points located at the general upper bound, like Point E, will correspond to the settings from the non-Bayesian errors. If a point located at the lower bound, say E', it can represent settings from either the Bayesian or non-Bayesian errors depending on the given information in classifications. Points E and E' form the mirrored points. Their settings can be connected by a relation in Equation (23) but are not necessary. For example, one specific setting for Point E' with $p_1 = 0.3$ and $p_2 = 0.7$ is:

$$\begin{aligned} p_{11} = 0.0,\ p_{12} = 0.3, \\ p_{21} = 0.0,\ p_{22} = 0.7, \end{aligned} \tag{37}$$

the other for Point E with $p_1 = 0.8$ and $p_2 = 0.2$ is:

$$\begin{aligned} p_{11} = \tfrac{20}{30},\ p_{12} = \tfrac{4}{30}, \\ p_{21} = \tfrac{5}{30},\ p_{22} = \tfrac{1}{30}. \end{aligned} \tag{38}$$

They are mirrored to each other but have no label-switching relation.

Points A′ and A″: When $P_E = 0.5$ and $p_{min} = 0.1$, Points A' and A'' form a pair as the ending points for the given conditions. Supposing $p_1 = 0.9$ and $p_2 = 0.1$, one can get the specific setting for Point A' from Equation (21c):

$$\begin{aligned} p_{11} = 0.4,\ p_{12} = 0.5, \\ p_{21} = 0.0,\ p_{22} = 0.4, \end{aligned} \tag{39}$$

and one for Point A'' from Equation (25):

$$\begin{aligned} p_{11} = 0.45,\ p_{12} = 0.45, \\ p_{21} = 0.05,\ p_{22} = 0.05. \end{aligned} \tag{40}$$

Points Q and R: The two points are specific due to their positions in the diagrams. For either type of errors, both points are non-admissible in the diagrams, because no joint distribution exists in binary classifications which can represent the points.

7. Summary

This work investigates into lower and upper bounds between entropy and error probability. An optimization approach is proposed to the derivations of the bound functions from a joint distribution. As a preliminary work, we consider binary classifications for a case study. Through the approach, Fano's lower bound receives novel interpretations. A new upper bound is derived and shows tighter in general than Kovalevskij's upper bound. The closed-form relations between bounds and error components are presented. The analytical results lead to a better understanding about the sharp conditions of bounds in terms of error components. Because classifications involve either Bayesian errors or non-Bayesian ones, we demonstrate the bounds comparatively for both types of errors.

We recognize that analytical tractability is an issue for the proposed approach. Fortunately, a symbolic software tool is helpful for solving complex problems successfully with different semi-analytical means (such as in [37,38]). The semi-analytical solution used in this work refers to the analytical derivation of all possible solutions, but the numerical verification of the final solution(s).

8. Discussions

To emphasize the importance of the study, we present discussions below from the perspective of machine learning in the context of big-data classifications. We consider that binary classifications will be one of key techniques to implement a *divide-and-conquer* strategy [39] for efficiently processing large quantities of data. Class-imbalance problems with extremely-skewed ratios are mostly formed from a *one-against-other* division scheme [31] in binary classifications. Researchers and users, of course, concern error components in types for performance evaluations [32]. The knowledge of bounds in relation to error components is desirable for theoretical and application purposes.

From a viewpoint of machine learning, the bounds derived in this work provide a basic solution to link learning targets between error and entropy in the related studies. *Error-based learning* is more conventional because of its compatibility with our intuitions in daily life, such as "*trial and error*". Significant studies have been reported under this category. In comparison, *information-based learning* [40] is relatively new and uncommon in some applications, such as classifications. Entropy is not a well-accepted concept related to our intuition in decision making. This is one of the reasons why the learning target is chosen mainly based on error, rather than on entropy. However, error is an empirical concept, whereas entropy is theoretical and general [41]. In [35], we demonstrated that entropy can deal with both notions of *error* and *reject* in abstaining classifications. Information-based learning [40] presents a promising and wider perspective for exploring and interpreting learning mechanisms.

When considering all sides of the issues stemming from machine learning studies, we believe that "*what to learn*" is a primary problem. However, it seems that more investigations focused on the issue of "*how to learn*", which should be put as the second-level problem. Moreover, in comparison with the long-standing yet hot theme of *feature selection*, little study has been done from the perspective of *learning target selection*. We propose that this theme should be emphasized in the study of machine learning. Hence, the relations studied in this work are fundamental and crucial to the extent that researchers, using either error-based or entropy-based approaches, are able to reach a better understanding about its counterpart.

Acknowledgments: This work is supported in part by National Natural Science Foundation of China No. 61273196 and 61573348 for Bao-Gang Hu, and National Natural Science Foundation of China No. 60903089 for Hong-Jie Xing. The first version of this work, entitled "Analytical bounds between entropy and error probability in binary classifications", was appeared as arXiv:1205.6602v1[cs.IT] in 30 May 2012. Thanks to T. Uyematsu and the anonymous reviewers for the valuable comments and suggestions.

Author Contributions: Bao-Gang Hu proposed the core concepts, derived the theorems, implemented the Maple codes, and wrote the paper. Hong-Jie Xing provided comments and made the proofreading of the paper. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix

```
> restart;                              # Clean the memory
> p2:=1-p1;e2:=Pe-e1;                   # Describe the bound with respect to p1 and e1
> HT:=-p1*log[2](p1)-p2*log[2](p2); # Shannon entropy
> p11:=(p1-e1);p12:=e1;p22:=p2-e2;p21:=e2;            # Terms of joint probability
> q1:=p11+p21;q2:=p12+p22;                            # Intermediate variables
> MI:=p11*log[2](p11/q1/p1)+p12*log[2](p12/q2/p1);
> MI=MI+p22*log[2](p22/q2/(1-p1))+p21*log[2](p21/q1/(1-p1));# Mutual information
> HTY:=(HT-MI);                                         # Conditional entropy
> HTY_dif_p1:=simplify(combine(diff(HTY,p1),ln, symbolic)); # Differential w.r.t. p1
                     /(p1 - 2 e1 + Pe) (-1 + p1 + Pe - e1)\
                  ln|-------------------------------------|
                     \  (p1 - e1) (-2 e1 - 1 + p1 + Pe)   /
     HTY_dif_p1 := ---------------------------------------
                                  ln(2)
> HTY_dif_e1:=simplify(combine(diff(HTY,e1),ln, symbolic)); # Differential w.r.t. e1
                  /                                  2          \
                  |  (p1 - e1) (-2 e1 - 1 + p1 + Pe)  (Pe - e1)|
                ln|- ----------------------------------------- |
                  |                    2                      |
                  \   (p1 - 2 e1 + Pe)  e1 (-1 + p1 + Pe - e1) /
 HTY_dif_e1 := -------------------------------------------------
                                  ln(2)
> solve({HTY_dif_p1=0,HTY_dif_e1=0}, {e1, p1});             # not a complete set of
                                                            # possible solutions
                  /                  2                      \
                  |                Pe  + e1 - Pe - 2 e1 Pe|
                 < e1 = e1, p1 = - ---------------------- >
                  |                          Pe            |
                  \                                        /
> E1:=solve(HTY_dif_e1, e1);     # a complete set of possible solutions when p1 is known
                     Pe (-1 + p1 + Pe)  p1 (-1 + p1 + Pe)
                  E1 := -----------------, -----------------
                              2 Pe - 1           2 p1 - 1
> P1_a:=solve(E1[1]=e1, {p1});P1_bc:=solve(E1[2]=e1, {p1}); # a complete set of possible
                                                            # solutions when e1 is known
                            /         2                    \
                            |       Pe  + e1 - Pe - 2 e1 Pe|
                  P1_a := < p1 = - ---------------------- >
                            |                Pe            |
                            \                              /
                             /                                                  (1/2)\
                             |          1   1      1 /    2                   2\     |
                  P1_bc := < p1 = e1 + - - - Pe + - \4 e1  - 4 e1 Pe + 1 - 2 Pe + Pe /      >,
                             |          2   2      2                                        |
                             \                                                             /

                             /                                                  (1/2)\
                             |          1   1      1 /    2                   2\     |
                            < p1 = e1 + - - - Pe - - \4 e1  - 4 e1 Pe + 1 - 2 Pe + Pe /      >
                             |          2   2      2                                        |
                             \                                                             /
> simplify(combine(simplify(eval(HTY, e1=E1[1])),ln,symbolic)); # failed to show it explicitly
> simplify(eval(HTY, e1=E1[2]));    # Display of the lower bound function in terms of p1
                    p1 ln(p1) + ln(1 - p1) - ln(1 - p1) p1
                  - --------------------------------------
                                      ln(2)
> # verification of concavity of HTY by a numerical way (changing Pe and p1 arbitrarily
> Pe:=0.5;p1:=0.6;plot(HTY_graph,e1=0..Pe);                    # with the constraints)
```

Figure A1. Maple code for deriving the lower bound.

```
> restart;                                      # Clean the memory
> HT:=-p1*log[2](p1)-p2*log[2](p2);             # Shannon entropy
> p11:=(p1-e1);p12:=e1;p22:=p2-e2;p21:=e2;  # Terms of joint distribution
> # To examine the HTY on two ending points for e2, i.e., e2 = 0 and e2=e
> # For derivation of the upper bound function when e2=0
> e1:=e;e2:=0;p1:=1-p2;
> q1:=p11+p21;q2:=p12+p22;                      # Intermediate variables
> MI:=p11*log[2](p11/q1/p1)+p12*log[2](p12/q2/p1); # Mutual information
> MI:=MI+p22*log[2](p22/q2/(1-p1));             # Neglect one term when 0*log(0)=0
> HTY_1:=combine(simplify(combine(simplify(HT-MI),ln,symbolic)));
> # Display of the upper bound function when e2=e
                            /e + p2\       /e + p2\
                       p2 ln|------| + e ln|------|
                            \  p2  /       \  e   /
              HTY_1 := ----------------------------
                                  ln(2)
> # For derivation of the upper bound function when e2=e
> e1:=0;e2:=e;
> q1:=p11+p21;q2:=p12+p22;                      # Intermediate variables
> MI:=p11*log[2](p11/q1/p1);                    # Neglect one term when 0*log(0)=0
> MI:=MI+p22*log[2](p22/q2/(1-p1))+p21*log[2](p21/q1/(1-p1));
> HTY:=eval(HT-MI,p2=1-P1);     # Using P1 for p1
> HTY_2:=combine(simplify(combine(simplify(HTY),ln,symbolic)));
> # Display of the upper bound function in terms of e and p2
                             /  P1  \       /  e   \
                       -P1 ln|------| - e ln|------|
                             \P1 + e/       \P1 + e/
              HTY_2 := ------------------------------
                                  ln(2)
> # To calculate the difference between HTY_1 and HTY_2
> delta_HTY:=combine(simplify(HTY_1-HTY_2),ln,symbolic);
                      /e + p2\       /e + p2\        /  P1  \
                 p2 ln|------| + e ln|------| + P1 ln|------|
                      \  p2  /       \P1 + e/        \P1 + e/
    delta_HTY := ------------------------------------------
                                  ln(2)
> # numerical verification of the solution to HTY below:
> # changing p2 arbitrarily with the constraint
> # when p2<0.5, delta_HTY<0, HTY_1 is the final solution,
> # when p2>0.5, delta_HTY>0, HTY_2 is the final solution,
> # when p2=0.5, delta_HTY=0, both are the solutions.
> p2:=0.4;P1:=1-p2;p_min:=min(P1,p2);plot(delta_HTY,e=0..p_min);
```

Figure A2. Maple code for deriving the upper bound.

References

1. Fano, R.M. Transmission of Information: A Statistical Theory of Communication. *Am. J. Phys.* **1961**, doi:10.1119/1.1937609.
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley: New York, NY, USA, 2006.
3. Verdú, S. Fifty years of Shannon theory. *IEEE Trans. Inf. Theory* **1998**, *44*, 2057–2078.
4. Yeung, R.W. *A First Course in Information Theory*; Kluwer Academic: London, UK, 2002.
5. Golic, J.D. Comment on "Relations between entropy and error probability". *IEEE Trans. Inf. Theory* **1999**, doi:10.1109/18.746849.
6. Vajda, I.; Zvárová, J. On generalized entropies, Bayesian decisions and statistical diversity. *Kybernetika* **2007**, *43*, 675–696.
7. Morales, D.; Vajda, I. Generalized information criteria for optimal Bayes decisions. *Kybernetika* **2012**, *48*, 714–749.
8. Kovalevskij, V.A. The Problem of Character Recognition from the Point of View of Mathematical Statistics. In *Character Readers and Pattern Recognition*; Spartan: New York, NY, USA, 1968; pp. 3–30.
9. Chu, J.T.; Chueh, J.C. Inequalities between information measures and error probability. *J. Frankl. Inst.* **1966**, *282*, 121–125.

10. Tebe, D.L.; Dwyer, S.J. Uncertainty and probability of error. *IEEE Trans. Inf. Theory* **1968**, *16*, 516–518.
11. Hellman, M.E.; Raviv, J. Probability of error, equivocation, and the Chernoff bound. *IEEE Trans. Inf. Theory* **1970**, *16*, 368–372.
12. Chen, C.H. Theoretical comparison of a class of feature selection criteria in pattern recognition. *IEEE Trans. Comput.* **1971**, *20*, 1054–1056.
13. Ben-Bassat, M.; Raviv, J. Rényi's entropy and the probability of error. *IEEE Trans. Inf. Theory* **1978**, *24*, 324–330.
14. Golić, J.D. On the relationship between the information measures and the Bayes probability of error. *IEEE Trans. Inf. Theory* **1987**, *35*, 681–690.
15. Feder, M.; Merhav, N. Relations between entropy and error probability. *IEEE Trans. Inf. Theory* **1994**, *40*, 259–266.
16. Han, T.S.; Verdú, S. Generalizing the Fano inequality. *IEEE Trans. Inf. Theory* **1994**, *40*, 1247–1251.
17. Poor, H.V.; Verdú, S. A Lower bound on the probability of error in multihypothesis testing. *IEEE Trans. Inf. Theory* **1995**, *41*, 1992–1994.
18. Harremoës, P.; Topsøe, F. Inequalities between entropy and index of coincidence derived from information diagrams. *IEEE Trans. Inf. Theory* **2001**, *47*, 2944–2960.
19. Erdogmus, D.; Principe, J.C. Lower and upper bounds for misclassification probability based on Renyi's information. *J. VLSI Signal Process.* **2004**, *37*, 305–317.
20. Ho, S.-W.; Verdú, S. On the interplay between conditional entropy and error probability. *IEEE Trans. Inf. Theory* **2010**, *56*, 5930–5942.
21. Liang, X.-B. A note on Fano's inequality. In Proceedings of the 2011 45th Annual Conference on Information Sciences and Systems, Baltimore, MD, USA, 23–25 March 2011.
22. Fano, R.M. Fano inequality. *Scholarpedia* **2008**, doi:10.4249/scholarpedia.6648.
23. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
24. Feder, M.; Merhav, N. Universal prediction of individual sequences. *IEEE Trans. Inf. Theory* **1992**, *38*, 1258–1270.
25. Wang, Y.; Hu, B.-G. Derivations of normalized mutual information in binary classifications. In Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, 14–16 August 2009; pp. 155–163.
26. Hu, B.-G. What are the differences between Bayesian classifiers and mutual-information classifiers? *IEEE Trans. Neural Net. Learn. Syst.* **2014**, *25*, 249–264.
27. Eriksson, T.; Kim, S.; Kang, H.-G.; Lee, C. An information-theoretic perspective on feature selection in speaker recognition. *IEEE Signal Process. Lett.* **2005**, *12*, 500–503.
28. Fisher, J.W.; Siracusa, M.; Tieu, K. Estimation of signal information content for classification. In Proceedings of the IEEE DSP Workshop, Marco Island, FL, USA, 4–7 January 2009; pp. 353–358.
29. Taneja, I.J. Generalized error bounds in pattern recognition. *Pattern Recogni. Lett.* **1985**, *3*, 361–368.
30. Duda, R.O.; Hart, P.E.; Stork, D. *Pattern Classification*, 2nd ed.; John Wiley: New York, NY, USA, 2001.
31. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*; Cambridge University Press: London, UK, 2000.
32. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
33. Sun, Y.M.; Wong, A.K.C.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719.
34. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.
35. Hu, B.-G.; He, R.; Yuan, X.-T. Information-theoretic measures for objective evaluation of classifications. *Acta Autom. Sin.* **2012**, *38*, 1160–1173.
36. Mackay, D.J.C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
37. Subramanian, V.R.; White, R.E. Symbolic solutions for boundary value problems using Maple. *Comput. Chem. Eng.* **2000**, *24*, 2405–2416.
38. Temimi, H.; Ansari, A.R. A semi-analytical iterative technique for solving nonlinear problems. *Comput. Math. Appl.* **2011**, *61*, 203–210.
39. Jordan, M.I. On statistics, computation and scalability. *Bernoulli* **2013**, *19*, 1378–1390.

40. Principe, J.C. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer: New York, NY, USA, 2010.
41. Hu, B.-G. Information theory and its relation to machine learning. In *Proceedings of the 2015 Chinese Intelligent Automation Conference*; Springer-Verlag: Berlin/Heidelberg, Germany, 2015; pp. 1–11.