Journal of Inequalities in Pure and Applied Mathematics

http://jipam.vu.edu.au/

Volume 7, Issue 2, Article 59, 2006

ENTROPY LOWER BOUNDS RELATED TO A PROBLEM OF UNIVERSAL CODING AND PREDICTION

FLEMMING TOPSØE

UNIVERSITY OF COPENHAGEN
DEPARTMENT OF MATHEMATICS
DENMARK
topsoe@math.ku.dk
URL: http://www.math.ku.dk/~topsoe

Received 18 August, 2004; accepted 20 March, 2006
Communicated by F. Hansen

ABSTRACT. Second order lower bounds for the entropy function expressed in terms of the index of coincidence are derived. Equivalently, these bounds involve entropy and Rényi entropy of order 2. The constants found either explicitly or implicitly are best possible in a natural sense. The inequalities developed originated with certain problems in universal prediction and coding which are briefly discussed.

Key words and phrases: Entropy, Index of coincidence, Rényi entropy, Measure of roughness, Universal coding, Universal prediction.

2000 *Mathematics Subject Classification.* 94A15, 94A17.

1. BACKGROUND, INTRODUCTION

We study probability distributions over the natural numbers. The set of all such distributions is denoted $M_+^1(\mathbb{N})$ and the set of $P \in M_+^1(\mathbb{N})$ which are supported by $\{1, 2, \dots, n\}$ is denoted $M_+^1(n)$.

We use U_k to denote a generic uniform distribution over a k-element set, and if also U_{k+1}, $U_{k+2}, \dots$ are considered, it is assumed that the supports are increasing. By H and by IC we denote, respectively *entropy* and *index of coincidence*, i.e.

$$H(P) = -\sum_{k=1}^{\infty} p_k \ln p_k \,,$$

$$IC(P) = \sum_{k=1}^{\infty} p_k^2 \,.$$

ISSN (electronic): 1443-5756

153-04

Results involving index of coincidence may be reformulated in terms of *Rényi entropy of order* 2 (H_2) as

$$H_2(P) = -\ln IC(P).$$

In Harremoës and Topsøe [5] the exact range of the map $P \curvearrowright (IC(P), H(P))$ with P varying over either $M^1_+(n)$ or $M^1_+(\mathbb{N})$ was determined. Earlier related work includes Kovalevskij [7], Tebbe and Dwyer [9], Ben-Bassat [1], Vajda and Vašek [13], Golic [4] and Feder and Merhav [2]. The ranges in question, termed *IC/H-diagrams*, were denoted Δ, respectively Δ_n:

$$\Delta = \{(IC(P), H(P)) \mid P \in M^1_+(\mathbb{N})\}\,,$$
$$\Delta_n = \{(IC(P), H(P)) \mid P \in M^1_+(n)\}\,.$$

By Jensen's inequality we find that $H(P) \geq -\ln IC(P)$, thus the logarithmic curve $t \curvearrowright (t, -\ln t)$; $0 < t \leq 1$ is a lower bounding curve for the IC/H-diagrams. The points $Q_k = \left(\frac{1}{k}, \ln k\right)$; $k \geq 1$ all lie on this curve. They correspond to the uniform distributions: $(IC(U_k), H(U_k)) = (\frac{1}{k}, \ln k)$. No other points in the diagram Δ lie on the logarithmic curve, in fact, Q_k; $k \geq 1$ are extremal points of Δ in the sense that the convex hull they determine contains Δ. No smaller set has this property.

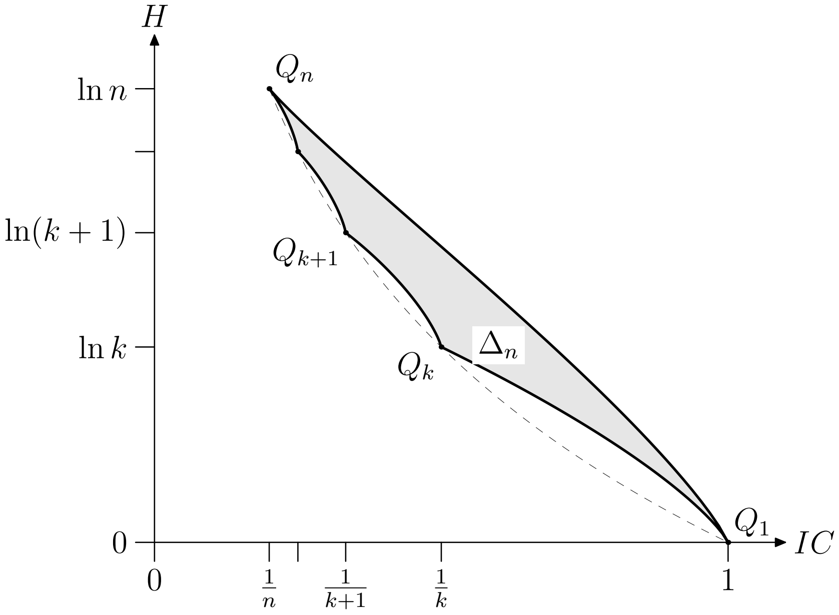

Figure 1.1: The restricted IC/H-diagram Δ_n, ($n = 5$).

Figure 1.1, adapted from [5], illustrates the situation for the restricted diagrams Δ_n. The key result of [5] states that Δ_n is the bounded region determinated by a certain Jordan curve determined by n smooth arcs, viz. the "upper arc" connecting Q_1 and Q_n and then $n-1$ "lower arcs" connecting Q_n with Q_{n-1}, Q_{n-1} with Q_{n-2} etc. until Q_2 which is connected with Q_1.

In [5], see also [11], the main result was used to develop concrete upper bounds for the entropy function. Our concern here will be lower bounds. The study depends crucially on the nature of the lower arcs. In [5] these arcs were identified. Indeed, the arc connecting Q_{k+1} with Q_k is the curve which may be parametrized as follows:

$$s \curvearrowright \vec{\varphi}((1-s)U_{k+1} + s\,U_k)$$

with s running through the unit interval and with $\vec{\varphi}$ denoting the *IC/H-map* given by $\vec{\varphi}(P) = (IC(P), H(P))$; $P \in M_+^1(\mathbb{N})$.

The distributions in $M_+^1(\mathbb{N})$ fall in *IC-complexity classes*. The kth class consists of all $P \in M_+^1(\mathbb{N})$ for which $IC(U_{k+1}) < IC(P) \le IC(U_k)$ or, equivalently, for which $\frac{1}{k+1} < IC(P) \le \frac{1}{k}$. In order to determine good lower bounds for the entropy of a distribution P, one first determines the IC-complexity class k of P. One then determines that value of $s \in]0,1]$ for which $IC(P_s) = IC(P)$ with $P_s = (1-s)U_{k+1} + s\,U_k$. Then $H(P) \ge H(P_s)$ is the theoretically best lower bound of $H(P)$ in terms of $IC(P)$.

In order to write the sought lower bounds for $H(P)$ in a convenient form, we introduce the kth *relative measure of roughness* by

$$\overline{MR}_k(P) = \frac{IC(P) - IC(U_{k+1})}{IC(U_k) - IC(U_{k+1})} = k(k+1)\left(IC(P) - \frac{1}{k+1}\right). \tag{1.1}$$

This definition applies to any $P \in M_+^1(\mathbb{N})$ but really, only distributions of IC-complexity class k will be of relevance to us. Clearly, $\overline{MR}_k(U_{k+1}) = 0$, $\overline{MR}_k(U_k) = 1$ and for any distribution of IC-complexity class k, $0 \le \overline{MR}_k(P) \le 1$. For a distribution on the lower arc connecting Q_{k+1} with Q_k one finds that

$$\overline{MR}_k((1-s)U_{k+1} + s\,U_k) = s^2\,. \tag{1.2}$$

In view of the above, it follows that for any distribution P of IC-complexity class k, the theoretically best lower bound for $H(P)$ in terms of $IC(P)$ is given by the inequality

$$H(P) \ge H\big((1-x)U_{k+1} + x\,U_k\big), \tag{1.3}$$

where x is determined so that P and $(1-x)U_{k+1} + x\,U_k$ have the same index of coincidence, i.e.

$$x^2 = \overline{MR}_k(P)\,. \tag{1.4}$$

By writing out the right-hand-side of (1.3) we then obtain the best lower bound of the type discussed. Doing so one obtains a quantity of mixed type, involving logarithmic and rational functions. It is desirable to search for structurally simpler bounds, getting rid of logarithmic terms. The simplest and possibly most useful bound of this type is the linear bound

$$H(P) \ge H(U_k)\overline{MR}_k(P) + H(U_{k+1})(1 - \overline{MR}_k(P)), \tag{1.5}$$

which expresses the fact mentioned above regarding the extremal position of the points Q_k in relation to the set Δ. Note that (1.5) is the best linear lower bound as equality holds for $P = U_{k+1}$ as well as for $P = U_k$. Another comment is that though (1.5) was developed with a view to distributions of IC-complexity class k, the inequality holds for all $P \in M_+^1(\mathbb{N})$ (but is weaker than the trivial bound $H \ge -\ln IC$ for distributions of other IC-complexity classes).

Writing (1.5) directly in terms of $IC(P)$ we obtain the inequalities

$$H(P) \ge \alpha_k - \beta_k\, IC(P); \quad k \ge 1 \tag{1.6}$$

with α_k and β_k given via the constants

$$u_k = \ln\left(1 + \frac{1}{k}\right)^k = k \ln\left(1 + \frac{1}{k}\right) \tag{1.7}$$

by

$$\begin{aligned} \alpha_k &= \ln(k+1) + u_k\,, \\ \beta_k &= (k+1)u_k\,. \end{aligned}$$

Note that the $u_k \uparrow 1$.[1]

In the present paper we shall develop sharper inequalities than those above by adding a second order term. More precisely, for $k \geq 1$, we denote by γ_k the largest constant such that the inequality

$$H \geq \ln k \, \overline{MR}_k + \ln(k+1)\,(1-\overline{MR}_k) + \frac{\gamma_k}{2k}\overline{MR}_k(1-\overline{MR}_k) \tag{1.8}$$

holds for all $P \in M^1_+(\mathbb{N})$. Here, $H = H(P)$ and $\overline{MR}_k = \overline{MR}_k(P)$. Expressed directly in terms of $IC = IC(P)$, (1.8) states that

$$H \geq \alpha_k - \beta_k\, IC + \frac{\gamma_k}{2}k(k+1)^2\left(IC - \frac{1}{k+1}\right)\left(\frac{1}{k} - IC\right) \tag{1.9}$$

for $P \in M^1_+(\mathbb{N})$.

The basic results of our paper may be summarized as follows: *The constants* $(\gamma_k)_{k\geq 1}$ *increase with* $\gamma_1 = \ln 4 - 1 \approx 0.3863$ *and with limit value* $\gamma \approx 0.9640$.

More substance will be given to this result by developing rather narrow bounds for the γ_k's in terms of γ and by other means.

The refined second order inequalities are here presented in their own right. However, we shall indicate in the next section how the author was led to consider inequalities of this type. This is related to problems of universal coding and prediction. The reader who is not interested in these problems can pass directly to Section 3.

2. A Problem of Universal Coding and Prediction

Let $\mathbb{A} = \{a_1, \dots, a_n\}$ be a finite *alphabet*. The models we shall consider are defined in terms of a subset $\mathcal{P}$ of $M^1_+(\mathbb{A})$ and a decomposition $\theta = \{A_1, \dots, A_k\}$ of $\mathbb{A}$ representing *partial information*.

A *predictor* (*θ-predictor*) is a map $P^* : \mathbb{A} \to [0,1]$ such that, for each $i \leq k$, the restriction $P^*_{|A_i}$ is a distribution in $M^1_+(A_i)$. The predictor P^* is *induced by* $P_0 \in M^1_+(\mathbb{A})$, and we write $P_0 \rightsquigarrow P^*$, if, for all $x \in \mathbb{A}$, $P^*_{|A_i} = (P_0)_{|A_i}$, the conditional probability of P_0 given A_i.

When we think of a predictor P^* in relation to the model $\mathcal{P}$, we say that P^* is a *universal predictor* (since the model may contain many distributions) and we measure its performance by the *guaranteed expected redundancy given* θ:

$$R(P^*) = \sup_{P\in\mathcal{P}} D^\theta(P\|P^*)\,. \tag{2.1}$$

Here, *expected redundancy (or divergence) given* θ is defined by

$$D^\theta(P\|P^*) = \sum_{i\leq k} P(A_i) D(P_{|A_i}\|P^*_{|A_i}) \tag{2.2}$$

with $D(\cdot\|\cdot)$ denoting standard Kullback-Leibler divergence. By R_{min} we denote the quantity

$$R_{min} = \inf_{P^*} R(P^*) \tag{2.3}$$

and we say that P^* is the *optimal universal predictor for $\mathcal{P}$ given θ* (or just the *optimal predictor*) if $R(P^*) = R_{min}$ and P^* is the only predictor with this property.

[1] Concrete algebraic bounds for the u_k, which, via (1.6), may be used to obtain concrete lower bounds for $H(P)$, are given by $\frac{2k}{2k+1} \leq u_k \leq \frac{2k+1}{2k+2}$. This follows directly from (1.6) of [12] (as $u_k = \lambda(\frac{1}{k})$ in the notation of that manuscript).

In parallel to predictors we consider quantities related to coding. A *θ-coding strategy* is a map $\kappa^* : \mathbb{A} \to [0, \infty]$ such that, for each $i \leq k$, *Kraft's equality*

$$\sum_{x \in A_i} \exp(-\kappa^*(x)) = 1 \tag{2.4}$$

holds. Note that there is a natural one-to-one correspondence, notationally written $P^* \leftrightarrow \kappa^*$, between predictors and coding strategies which is given by the relations

$$\kappa^* = -\ln P^* \quad \text{and} \quad P^* = \exp(-\kappa^*)\,. \tag{2.5}$$

When $P^* \leftrightarrow \kappa^*$, we may apply the *linking identity*

$$D^\theta(P\|P^*) = \langle \kappa^*, P\rangle - H^\theta(P) \tag{2.6}$$

which is often useful for practical calculations. Here, $H^\theta(P) = \sum_i P(A_i)H(P_{|A_i})$ is standard conditional entropy and $\langle \cdot, P\rangle$ denotes expectation w.r.t. P.

From Harremoës and Topsøe [6] we borrow the following result:

Theorem 2.1 (Kuhn-Tucker criterion). *Assume that $A_1, \ldots, A_m$ are distributions in $\mathcal{P}$, that $P_0 = \sum_{\nu \leq m} \alpha_\nu A_\nu$ is a convex combination of the A_ν's with positive weights which induces the predictor P^*, that, for some finite constant R, $D^\theta(A_\nu\|P^*) = R$ for all $\nu \leq m$ and, finally, that $R(P^*) \leq R$.*

Then P^* is the optimal predictor and $R_{min} = R$. Furthermore, the convex set $\overline{\mathcal{P}}$ given by

$$\overline{\mathcal{P}} = \{P \in M_+^1(\mathbb{A}) | D^\theta(P\|P^*) \leq R\} \tag{2.7}$$

can be characterized as the largest model with P^* as optimal predictor and $R_{min} = R$.

This result is applicable in a great variety of cases. For indications of the proof, see [6] and Section 4.3 of [10][1]. The distributions A_ν of the result are referred to as *anchors* and the model $\overline{\mathcal{P}}$ as the *maximal model*.

The concrete instances of Theorem 2.1 which we shall now discuss have a certain philosophical flavour which is related to the following general and loosely formulated question: If we think of "Nature" or "God" as deciding which distribution $P \in \mathcal{P}$ to choose as the "true" distribution, and if we assume that the model we consider is really basic and does not lend itself to further fragmentation, one may ask if any other choice than a uniform distribution is really feasible. In other words, one may maintain the view that "God only knows the uniform distribution".

Whether or not the above view can be formulated more precisely and meaningfully, say within physics, is not that clear. Anyhow, motivated by this kind of thinking, we shall look at some models involving only uniform distributions. For models based on large alphabets, the technicalities become quite involved and highly combinatorial. Here we present models with $\mathbb{A}_0 = \{0, 1\}$ consisting of the two binary digits as the *source alphabet*. The three uniform distributions pertaining to $\mathbb{A}_0$ are denoted U_0 and U_1 for the two deterministic distributions and U_{01} for the uniform distribution over $\mathbb{A}_0$. For an integer $t \geq 2$ consider the model $\mathcal{P} = \{U_0^t, U_1^t, U_{01}^t\}$ with exponentiation indicating product measures. We are interested in *universal coding* or, equivalently, *universal prediction* of Bernoulli trials $x_1^t = x_1 x_2 \cdots x_t \in \mathbb{A}_0^t$ from this model, assuming that partial information corresponding to observation of $x_1^s = x_1 \cdots x_s$ for a fixed s is available. This model is of interest for any integers s and t with $0 \leq s < t$. However, in order to further simplify, we assume that $s = t - 1$. The model we arrive at is then closely related to the classical "problem of succession" going back to Laplace, cf. Feller [3]. For a modern treatment, see Krichevsky [8].

[1] The former source is just a short proceedings contribution. For various reasons, documentation in the form of comprehensive publications is not yet available. However, the second source which reveals the character of the simple proof, may be helpful.

Common sense has it that the optimal coding strategy and the optimal predictor, respectively κ^* and P^*, are given by expressions of the form

$$\kappa^*(x_1^t) = \begin{cases} \kappa_1 & \text{if} \quad x_1^t = 0\cdots 00 \quad \text{or} \quad 1\cdots 11 \\ \kappa_2 & \text{if} \quad x_1^t = 0\cdots 01 \quad \text{or} \quad 1\cdots 10 \\ \ln 2 & \text{otherwise} \end{cases} \tag{2.8}$$

and

$$P^*(x_1^t) = \begin{cases} p_1 & \text{if} \quad x_1^t = 0\cdots 00 \quad \text{or} \quad 1\cdots 11 \\ p_2 & \text{if} \quad x_1^t = 0\cdots 01 \quad \text{or} \quad 1\cdots 10 \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{2.9}$$

with $p_1 = \exp(-\kappa_1)$ and $p_2 = \exp(-\kappa_2)$. Note that p_i is the weight P^* assigns to the occurrence of i binary digits in x_1^t in case only one binary digit occurs in x_1^s. Clearly, if both binary digits occur in x_1^s, it is sensible to predict the following binary digit to be a 0 or a 1 with equal weights as also shown in (2.9).

With $t|s$ as superscript to indicate partial information we find from (2.6) that

$$D^{t|s}(U_0^t \| P^*) = D^{t|s}(U_1^t \| P^*) = \kappa_1 \,,$$
$$D^{t|s}(U_{01}^t \| P^*) = 2^{-s}(\kappa_1 + \kappa_2 - \ln 4)\,.$$

With an eye to Theorem 2.1 we equate these numbers and find that

$$(2^s - 1)\kappa_1 = \kappa_2 - \ln 4\,. \tag{2.10}$$

Expressed in terms of p_1 and p_2, we have $p_1 = 1 - p_2$ and

$$4p_2 = (1 - p_2)^{2^s - 1}\,. \tag{2.11}$$

Note that (2.11) determines $p_2 \in [0,1]$ uniquely for any s.

It is a simple matter to check that the conditions of Theorem 2.1 are fulfilled (with U_0^t, U_1^t and U_{01}^t as anchors). With reference to the discussion above, we have then obtained the following result:

Theorem 2.2. *The optimal predictor for prediction of x_t with $t = s+1$, given $x_1^s = x_1 \cdots x_s$ for the Bernoulli model $\mathcal{P} = \{U_0^t, U_1^t, U_{01}^t\}$ is given by* (2.9) *with $p_1 = 1 - p_2$ and p_2 determined by* (2.11)*. Furthermore, for this model, $R_{min} = -\ln p_1 = \kappa_1$ and the maximal model, $\overline{\mathcal{P}}$, consists of all $Q \in M_1^+(\mathbb{A}_0^t)$ for which $D^{t|s}(Q\|P^*) \le \kappa_1$.*

It is natural to ask about the type of distributions included in the maximal model $\overline{\mathcal{P}}$ of Theorem 2.2. In particular, we ask, sticking to the framework of a Bernoulli model, which product distributions are included? Applying (2.6), this is in principle easy to answer. We shall only comment on the three cases $s = 1, 2, 3$.

For $s = 1$ or $s = 2$ one finds that the inequality $D^{t|s}(P^t\|P^*) \le R_{min}$ is equivalent to the inequality $H \ge \ln 4(1 - IC)$ which, by (1.6) for $k = 1$, is known to hold for any distribution. Accordingly, in these cases, $\overline{\mathcal{P}}$ contains every product distribution.

For the case $s = 3$ the situation is different. Then, as the reader can easily check, the crucial inequality $D^{t|s}(P^t\|P^*) \le R_{min}$ is equivalent to the following inequality (with $H = H(P)$, $IC = IC(P)$):

$$H \ge (1 - IC)\left(\ln 4 + (\ln 2 + 3\kappa_1)\left(IC - \frac{1}{2}\right)\right)\,. \tag{2.12}$$

This is a second order lower bound of the entropy function of the type discussed in Section 1. In fact this is the way we were led to consider such inequalities. As stated in Section 1

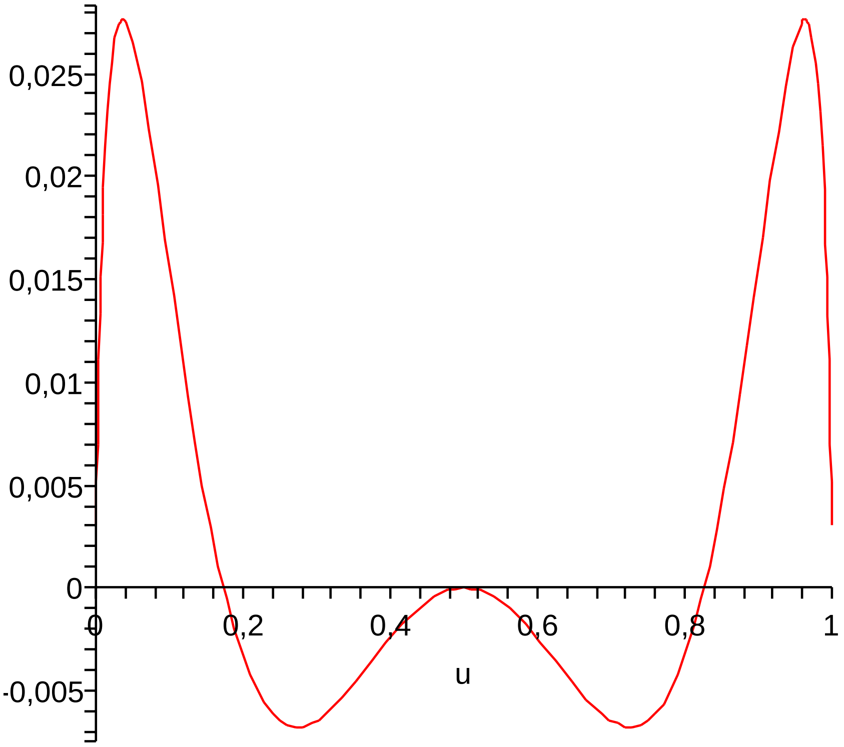

Figure 2.1: A plot of $R_{min} - D^{4|3}(P^t \| P^*)$ as function of p with $P = (p, q)$.

and proved rigorously in Lemma 3.5 of Section 3, the largest constant which can be inserted in place of the constant $\ln 2 + \kappa_1 \approx 1.0438$ in (2.12), if we want the inequality to hold for all $P \in M_+^1(\mathbb{A}_0)$, is $2\gamma_1 = 2(\ln 4 - 1) \approx 0.7726$. Thus (2.12) does *not* hold for all $P \in M_+^1(\mathbb{A}_0)$. In fact, considering the difference between the left hand and the right hand side of (2.12), shown in Figure 2.1, we realize that when $s = 3$, P^4 with $P = (p, q)$ belongs to the maximal model if and only if either $P = U_{01}$ or else one of the probabilities p or q is smaller than or equal to some constant (≈ 0.1734).

3. Basic Results

The key to our results is the inequality (1.3) with x determined by (1.4)[1] . This leads to the following analytical expression for γ_k:

Lemma 3.1. *For $k \geq 1$ define $f_k : [0, 1] \to [0, \infty]$ by*

$$f_k(x) = \frac{2k}{x^2(1-x^2)} \left[-\frac{k+x}{k+1} \ln\left(1 + \frac{x}{k}\right) - \frac{1-x}{k+1} \ln(1-x) + x^2 \ln\left(1 + \frac{1}{k}\right) \right] .$$

Then $\gamma_k = \inf\{f_k(x) \mid 0 < x < 1\}$.

[1] For the benefit of the reader we point out that this inequality can be derived rather directly from the *lemma of replacement* developed in [5]. The relevant part of that lemma is the following result: If $f : [0, 1] \to \mathbb{R}$ is concave/convex (i.e. concave on $[0, \xi]$, convex on $[\xi, 1]$ for some $\xi \in [0, 1]$), then, for any $P \in M_+^1(\mathbb{N})$, there exists $k \geq 1$ and a convex combination P_0 of U_{k+1} and U_k such that $F(P_0) \leq F(P)$ with F defined by $F(Q) = \sum_1^\infty f(q_n)$; $Q \in M_+^1(\mathbb{N})$.

Proof. By the defining relation (1.8) and by (1.3) with x given by (1.4), recalling also the relation (1.2), we realize that γ_k is the infimum over $x \in]0,1[$ of

$$\frac{2k}{x^2(1-x^2)}\left[H\left((1-x)U_{k+1}+xU_k\right)-\ln k\cdot x^2-\ln(k+1)\cdot(1-x^2)\right] .$$

Writing out the entropy of $(1-x)U_{k+1}+xU_k$ we find that the function defined by this expression is, indeed, the function f_k. □

It is understood that $f_k(x)$ is defined by continuity for $x=0$ and $x=1$. An application of l'Hôspitals rule shows that

$$f_k(0)=2u_k-1\,,\quad f_k(1)=\infty\,. \tag{3.1}$$

Then we investigate the limiting behaviour of $(f_k)_{k\geq 1}$ for $k\to\infty$.

Lemma 3.2. *The pointwise limit $f=\lim_{k\to\infty} f_k$ exists in $[0,1]$ and is given by*

$$f(x)=\frac{2\left(-x-\ln(1-x)\right)}{x^2(1+x)}\,;\quad 0<x<1 \tag{3.2}$$

with $f(0)=1$ and $f(1)=\infty$. Alternatively,

$$f(x)=\frac{2}{1+x}\sum_{n=0}^{\infty}\frac{x^n}{n+2}\,;\quad 0\leq x\leq 1.[1] \tag{3.3}$$

The simple proof, based directly on Lemma 3.1, is left to the reader. We then investigate some of the properties of f:

Lemma 3.3. *The function f is convex, $f(0)=1$, $f(1)=\infty$ and $f'(0)=-\frac{1}{3}$. The real number $x_0=$ argmin$f$ is uniquely determined by one of the following equivalent conditions:*

(i) $f'(x_0)=0$,

(ii) $-\ln(1-x_0)=\frac{2x_0(1+x_0-x_0^2)}{(3x_0+2)(1-x_0)}$,

(iii) $\sum_{n=1}^{\infty}\left(\frac{n+1}{n+3}+\frac{n-1}{n+2}\right)x_0^n=\frac{1}{6}$

One has $x_0\approx 0.2204$ and $\gamma\approx 0.9640$ with $\gamma=f(x_0)=\min f$.

Proof. By standard differentiation, say based on (3.2), one can evaluate f and f'. One also finds that (i) and (ii) are equivalent. The equivalence with (iii) is based on the expansion

$$f'(x)=\frac{2}{(1+x)^2}\sum_{n=0}^{\infty}\left(\frac{n+1}{n+3}+\frac{n-1}{n+2}\right)x^n$$

which follows readily from (4).

The convexity, even strict, of f follows as f can be written in the form

$$f(x)=\left(\frac{2}{3}+\frac{1}{3}\cdot\frac{1}{1+x}\right)+\sum_{n=2}^{\infty}\frac{2}{n+2}\cdot\frac{x^n}{1+x}\,,$$

easily recognizable as a sum of two convex functions.

The approximate values of x_0 and γ were obtained numerically, based on the expression in (ii). □

The convergence of f_k to f is in fact increasing:

Lemma 3.4. *For every $k\geq 1$, $f_k\leq f_{k+1}$.*

[1] or, as a power series in x, $f(x)=2\sum_0^\infty(-1)^n(1-l_{n+2})x^n$ with $l_n=-\sum_1^n(-1)^k\frac{1}{k}$.

Proof. As a more general result will be proved as part (i) of Theorem 4.1, we only indicate that a direct proof involving three times differentiation of the function

$$\Delta_k(x) = \frac{1}{2}x^2(1-x^2)(f_{k+1}(x) - f_k(x))$$

is rather straightforward. □

Lemma 3.5. $\gamma_1 = \ln 4 - 1 \approx 0.3863$.

Proof. We wish to find the best (largest) constant c such that

$$H(P) \geq \ln 4 \cdot (1 - IC(P)) + 2c\left(IC(P) - \frac{1}{2}\right)(1 - IC(P)) \tag{3.4}$$

holds for all $P \in M_+^1(\mathbb{N})$, cf. (1.9), and know that we only need to worry about distributions $P \in M_+^1(2)$. Let $P = (p,q)$ be such a distribution, i.e. $0 \leq p \leq 1$, $q = 1-p$. Take p as an independent variable and define the auxiliary function $h = h(p)$ by

$$h = H - \ln 4 \cdot (1 - IC) - 2c\left(IC - \frac{1}{2}\right)(1 - IC)\,.$$

Here, $H = -p\ln p - q\ln q$ and $IC = p^2 + q^2$. Then:

$$h' = \ln\frac{q}{p} + 2(p-q)\ln 4 - 2c(p-q)(3 - 4IC)\,,$$
$$h'' = -\frac{1}{pq} + 4\ln 4 - 2c(-10 + 48pq)\,.$$

Thus $h(0) = h(\frac{1}{2}) = h(1) = 0$, $h'(0) = \infty$, $h'(\frac{1}{2}) = 0$ and $h'(1) = -\infty$. Further, $h''(\frac{1}{2}) = -4 + 4\ln 4 - 4c$, hence h assumes negative values if $c > \ln 4 - 1$. Assume now that $c < \ln 4 - 1$. Then $h''(\frac{1}{2}) > 0$. As h has (at most) two inflection points (follows from the formula for h'') we must conclude that $h \geq 0$ (otherwise h would have at least six inflection points!).

Thus $h \geq 0$ if $c < \ln 4 - 1$. Then $h \geq 0$ also holds if $c = \ln 4 - 1$. □

The lemma is an improvement over an inequality established in [11] as we shall comment more on in Section 4.

With relatively little extra effort we can find reasonable bounds for each of the γ_k's in terms of γ. What we need is the following lemma:

Lemma 3.6. *For* $k \geq 1$ *and* $0 \leq x < 1$,

$$f_k(x) = \frac{2k}{(k+1)(1-x^2)}\sum_{n=0}^{\infty}\frac{1}{2n+2} \times \left[\frac{1-x^{2n+1}}{2n+3}\left(1 - \frac{1}{k^{2n+2}}\right) + \frac{1-x^{2n}}{2n+1}\left(1 + \frac{1}{k^{2n+1}}\right)\right] \tag{3.5}$$

and

$$f(x) = \frac{2}{1-x^2}\sum_{n=0}^{\infty}\frac{1}{2n+2}\left(\frac{1-x^{2n+1}}{2n+3} + \frac{1-x^{2n}}{2n+1}\right)\,. \tag{3.6}$$

Proof. Based on the expansions

$$-x - \ln(1-x) = x^2\sum_{n=0}^{\infty}\frac{x^n}{n+2}$$

and

$$(k+x)\ln\left(1+\frac{x}{k}\right) = x + x^2\sum_{n=0}^{\infty}\frac{(-1)^n x^n}{(n+2)(n+1)k^{n+1}}$$

(which is also used for $k = 1$ with x replaced by $-x$), one readily finds that

$$\begin{aligned} &-(k+x)\ln\left(1+\frac{x}{k}\right) - (1-x)\ln(1-x) + (k+1)x^2\ln\left(1+\frac{1}{k}\right) \\ &\qquad = x^2\left[1+\sum_{n=0}^{\infty}\frac{(-1)^n}{(n+2)(n+1)}\cdot\frac{1}{k^{n+1}}\right. \\ &\qquad\qquad \left. - \sum_{n=0}^{\infty}\frac{x^n}{(n+2)(n+1)}\left(\frac{(-1)^n}{k^{n+1}}+1\right)\right]. \end{aligned}$$

Upon writing 1 in the form

$$1 = \sum_{n=0}^{\infty}\frac{1}{2n+2}\left(\frac{1}{2n+1}+\frac{1}{2n+3}\right)$$

and collecting terms two-by-two, and subsequent division by $1-x^2$ and multiplication by $2k$, (3.5) emerges. Clearly, (3.6) follows from (3.5) by taking the limit as k converges to infinity. □

Putting things together, we can now prove the following result:

Theorem 3.7. *We have* $\gamma_1 \le \gamma_2 \le \cdots$, $\gamma_1 = \ln 4 - 1 \approx 0.3863$ *and* $\gamma_k \to \gamma$ *where* $\gamma \approx 0.9640$ *can be defined as*

$$\gamma = \min_{0<x<1}\left\{\frac{2}{x^2(1+x)}\left(\ln\frac{1}{1-x} - x\right)\right\}.$$

Furthermore, for each $k \ge 1$,

$$\left(1-\frac{1}{k}\right)\gamma \le \gamma_k \le \left(1-\frac{1}{k}+\frac{1}{k^2}\right)\gamma\,. \tag{3.7}$$

Proof. The first parts follow directly from Lemmas 3.1 – 3.5. To prove the last statement, note that, for $n \ge 0$,

$$1-\frac{1}{k^{2n+2}} \ge 1-\frac{1}{k^2}\,.$$

It then follows from Lemma 3.6 that $(1+\frac{1}{k})f_k \ge \left(1-\frac{1}{k^2}\right)f$, hence $f_k \ge (1-\frac{1}{k})f$ and $\gamma_k \ge (1-\frac{1}{k})\gamma$ follows.

Similarly, note that $1+k^{-(2n+1)} \le 1+k^{-3}$ for $n \ge 1$ (and that, for $n = 0$, the second term in the summation in (3.5) vanishes). Then use Lemma 3.6 to conclude that $(1+\frac{1}{k})f_k \le (1+\frac{1}{k^3})f$. The inequality $\gamma_k \le (1-\frac{1}{k}+\frac{1}{k^2})\gamma$ follows. □

The discussion contains more results, especially, the bounds in (3.7) are sharpened.

4. Discussion and Further Results

Justification:

The justification for the study undertaken here is two-fold: As a study of certain aspects of the relationship between entropy and index of coincidence – which is part of the wider theme of comparing one Rényi entropy with another, cf. [5] and [14] – and as a preparation for certain results of exact prediction in Bernoulli trials. Both types of justification were carefully dealt with in Sections 1 and 2.

Lower bounds for distributions over two elements:

Regarding Lemma 3.5, the key result proved is really the following inequality for a two-element probability distribution $P = (p, q)$:

$$4pq\left(\ln 2 + \left(\ln 2 - \frac{1}{2}\right)(1 - 4pq)\right) \leq H(p, q)\,. \tag{4.1}$$

Let us compare this with the lower bounds contained in the following inequalities proved in [11]:

$$\ln p \ln q \leq H(p, q) \leq \frac{\ln p \ln q}{\ln 2}\,, \tag{4.2}$$

$$\ln 2 \cdot 4pq \leq H(p, q) \leq \ln 2 (4pq)^{1/\ln 4}\,. \tag{4.3}$$

Clearly, (4.1) is sharper than the lower bound in (4.3). Numerical evidence shows that "normally" (4.1) is also sharper than the lower bound in (4.2) but, for distributions close to a deterministic distribution, (4.2) is in fact the sharper of the two.

More on the convergence of f_k to f:

Although Theorem 3.7 ought to satisfy most readers, we shall continue and derive sharper bounds than those in (3.7). This will be achieved by a closer study of the functions f_k and their convergence to f as $k \to \infty$. By looking at previous results, notably perhaps Lemma 3.1 and the proof of Theorem 3.7, one gets the suspicion that it is the sequence of functions $(1 + \frac{1}{k})f_k$ rather than the sequence of f_k's that are well behaved. This is supported by the results assembled in the theorem below, which, at least for parts (ii) and (iii), are the most cumbersome ones to derive of the present research:

Theorem 4.1.

(i) $(1 + \frac{1}{k})f_k \uparrow f$, *i.e.* $2f_1 \leq \frac{3}{2}f_2 \leq \frac{4}{3}f_3 \leq \cdots \to f\,.$

(ii) *For each* $k \geq 1$, *the function* $f - (1 + \frac{1}{k})f_k$ *is decreasing in* $[0, 1]\,.$

(iii) *For each* $k \geq 1$, *the function* $(1 + \frac{1}{k})f_k/f$ *is increasing in* $[0, 1]$.

The technique of proof will be elementary, mainly via torturous differentiations (which may be replaced by MAPLE look-ups, though) and will rely also on certain inequalities for the logarithmic function in terms of rational functions. A sketch of the proof is relegated to the appendix.

An analogous result appears to hold for convergence from above to f. Indeed, experiments on MAPLE indicate that $(1 + \frac{1}{k} + \frac{1}{k^2})f_k \downarrow f$ and that natural analogs of (ii) and (iii) of Theorem 4.1 hold. However, this will not lead to improved bounds over those derived below in Theorem 4.2.

Refined bounds for γ_k in terms of γ:

Such bounds follow easily from (ii) and (iii) of Theorem 4.1:

Theorem 4.2. *For each* $k \geq 1$, *the following inequalities hold:*

$$(2u_k - 1)\gamma \leq \gamma_k \leq \frac{k}{k+1}\gamma + \frac{2k}{k+1} - \frac{2k+1}{k+1}u_k\,. \tag{4.4}$$

Proof. Define constants a_k and b_k by

$$a_k = \inf_{0 \leq x \leq 1}\left(f(x) - \left(1 + \frac{1}{k}\right)f_k(x)\right)\,,$$

$$b_k = \inf_{0 \leq x \leq 1}\frac{\left(1 + \frac{1}{k}\right)f_k(x)}{f(x)}\,.$$

Then

$$b_k\gamma \leq \left(1+\frac{1}{k}\right)\gamma_k \leq \gamma - a_k\,.$$

Now, by (ii) and (iii) of Theorem 4.1 and by an application of l'Hôpitals rule, we find that

$$a_k = \left(2+\frac{1}{k}\right)u_k - 2\,,$$

$$b_k = \left(1+\frac{1}{k}\right)(2u_k-1)\,.$$

The inequalities of (4.4) follow. □

Note that another set of inequalities can be obtained by working with $\sup$ instead of $\inf$ in the definitions of a_k and b_k. However, inspection shows that the inequalities obtained that way are weaker than those given by (4.4).

The inequalities (4.4) are sharper than (3.7) of Theorem 3.7 but less transparent. Simpler bounds can be obtained by exploiting lower bounds for u_k (obtained from lower bounds for $\ln(1+x)$, cf. [12]). One such lower bound is given in footnote [1] and leads to the inequalities

$$\frac{2k-1}{2k+1}\gamma \leq \gamma_k \leq \frac{k}{k+1}\gamma\,. \tag{4.5}$$

Of course, the upper bound here is also a consequence of the relatively simple property (i) of Theorem 4.1. Applying sharper bounds of the logarithmic function leads to the bounds

$$\frac{2k-1}{2k+1}\gamma \leq \gamma_k \leq \frac{k}{k+1}\left(\gamma - \frac{1}{6k^2+6k+1}\right)\,. \tag{4.6}$$

APPENDIX

We shall here give an outline of the proof of Theorem 4.1. We need some auxiliary bounds for the logarithmic function which are available from [12]. In particular, for the function λ defined by

$$\lambda(x) = \frac{\ln(1+x)}{x}\,,$$

one has

$$(2-x)\lambda(y) - \frac{1-x}{1+y} \leq \lambda(xy) \leq x\lambda(y) + (1-x)\,, \tag{4.7}$$

valid for $0 \leq x \leq 1$ and $0 \leq y < \infty$, cf. (16) of [12].

Proof of (i) of Theorem 4.1. Fix $0 \leq x \leq 1$ and introduce the parameter $y = \frac{1}{k}$. Put

$$\psi(y) = \left(1+\frac{1}{k}\right)\frac{x^2(1-x^2)}{2}f_k(x) + (1-x)\ln(1-x)$$

(with $k = \frac{1}{y}$). Then, simple differentiation and an application of the right hand inequality of (4.7) shows that ψ is a decreasing function of y in $]0,1]$. This implies the desired result. □

Proof of (ii) of Theorem 4.1. Fix $k \geq 1$ and put $\varphi = f - \left(1+\frac{1}{k}\right)f_k$. Then φ' can be written in the form

$$\varphi'(x) = \frac{2kx}{x^4(1-x^2)^2}\psi(x)\,.$$

We have to prove that $\psi \leq 0$ in $[0,1]$. After differentiations, one finds that $\psi(0) = \psi(1) = \psi'(0) = \psi'(1) = \psi''(0) = 0$.

Furthermore, we claim that $\psi''(1) < 0$. This amounts to the inequality

$$\ln(1+y) > \frac{y(8+7y)}{(1+y)(8+3y)} \quad \text{with} \quad y = \frac{1}{k}\,. \tag{4.8}$$

This is valid for $y > 0$, as may be proved directly or deduced from a known stronger inequality (related to the function ϕ_2 listed in Table 1 of [12]).

Further differentiation shows that $\psi'''(0) = -y^3 < 0$. With two more differentiations we find that

$$\psi^{(5)}(x) = -\frac{18y^3}{(1+xy)^2} - \frac{20y^3}{(1+xy)^3} - \frac{6y^3(1-y^2)}{(1+xy)^4} + \frac{24y^3(1-y^2)}{(1+xy)^5}\,.$$

Now, if ψ assumes positive values in $[0,1]$, $\psi''(x) = 0$ would have at least 4 solutions in $]0,1[$, hence $\psi^{(5)}$ would have at least one solution in $]0,1[$. In order to arrive at a contradiction, we put $X = 1+xy$ and note that $\psi^{(5)}(x) = 0$ is equivalent to the equality

$$-9X^3 - 10X^2 - 3(1-y^2)X + 12(1-y^2) = 0\,.$$

However, it is easy to show that the left hand side here is upper bounded by a negative number. Hence we have arrived at the desired contradiction, and conclude that $\psi \le 0$ in $[0,1]$. □

Proof of (iii) of Theorem 4.1. Again, fix k and put

$$\psi(x) = 1 - \frac{(1+\frac{1}{k})f_k(x)}{f(x)}\,.$$

Then, once more with $y = \frac{1}{k}$,

$$\psi(x) = \frac{(1+xy)\ln(1+xy) - x^2(1+y)\ln(1+y) - xy(1-x)}{y(1-x)(-x-\ln(1-x))}\,.$$

We will show that $\psi' \le 0$. Write ψ' in the form

$$\psi' = \frac{y}{\text{denominator}^2}\xi\,,$$

where "denominator" refers to the denominator in the expression for ψ. Then $\xi(0) = \xi(1) = 0$. Regarding the continuity of ξ at 1 with $\xi(1) = 0$, the key fact needed is the limit relation

$$\lim_{x-1^-} \ln(1-x)\cdot \ln\frac{1+xy}{1+y} = 0\,.$$

Differentiation shows that $\xi'(0) = -2y < 0$ and that $\xi'(1) = \infty$. Further differentiation and exploitation of the left hand inequality of (4.7) gives:

$$\begin{aligned}\xi''(x) &\ge y\left(-10x - 2xy - \frac{1}{1+xy} + 6 + \frac{1}{1-x}\right)\\ &\ge y\left(-12x - \frac{1}{1+x} + \frac{1}{1-x} + 6\right)\,,\end{aligned}$$

and this quantity is ≥ 0 in $[0,1[$. We conclude that $\xi \le 0$ in $[0,1[$. The desired result follows.

All parts of Theorem 4.1 are hereby proved. □

REFERENCES

[1] M. BEN-BASSAT, f-entropies, probability of error, and feature selection, *Information and Control*, **39** (1978), 227–242.

[2] M. FEDER AND N. MERHAV, Relations between entropy and error probability, *IEEE Trans. Inform. Theory*, **40** (1994), 259–266.

[3] W. FELLER, *An Introduction to Probability Theory and its Applications*, Volume I. Wiley, New York, 1950.

[4] J. Dj. GOLIĆ, On the relationship between the information measures and the Bayes probability of error. *IEEE Trans. Inform. Theory*, **IT-33**(5) (1987), 681–690.

[5] P. HARREMOËS AND F. TOPSØE, Inequalities between entropy and index of coincidence derived from information diagrams. *IEEE Trans. Inform. Theory*, **47**(7) (2001), 2944–2960.

[6] P. HARREMOËS AND F. TOPSØE, Unified approach to optimization techniques in Shannon theory, in *Proceedings, 2002 IEEE International Symposium on Information Theory*, page 238. IEEE, 2002.

[7] V.A. KOVALEVSKIJ, *The Problem of Character Recognition from the Point of View of Mathematical Statistics*, pages 3–30. Spartan, New York, 1967.

[8] R. KRICHEVSKII, Laplace's law of succession and universal encoding, *IEEE Trans. Inform. Theory*, **44** (1998), 296–303.

[9] D.L. TEBBE AND S.J. DWYER, Uncertainty and the probability of error, *IEEE Trans. Inform. Theory*, **14** (1968), 516–518.

[10] F. TOPSØE, Information theory at the service of science, in *Bolyai Society Mathematical Studies*, Gyula O.H. Katona (Ed.), Springer Publishers, Berlin, Heidelberg, New York, 2006. (to appear).

[11] F. TOPSØE, Bounds for entropy and divergence of distributions over a two-element set, *J. Ineq. Pure Appl. Math.*, **2**(2) (2001), Art. 25. [ONLINE: `http://jipam.vu.edu.au/article.php?sid=141`].

[12] F. TOPSØE, Some bounds for the logarithmic function, in *Inequality Theory and Applications, Vol. 4*, Yeol Je Cho and Jung Kyo Kim and Sever S. Dragomir (Eds.), Nova Science Publishers, New York, 2006 (to appear). *RGMIA Res. Rep. Coll.*, **7**(2) (2004), [ONLINE: `http://rgmia.vu.edu.au/v7n2.html`].

[13] I. VAJDA AND K. VAŠEK, Majorization, concave entropies, and comparison of experiments. *Problems Control Inform. Theory*, **14** (1985), 105–115.

[14] K. ZYCZKOWSKI, Rényi extrapolation of Shannon entropy, *Open Systems and Information Dynamics*, **10** (2003), 297–310.