Inequalities between Entropy and Index of Coincidence derived from Information Diagrams

Peter Harremoës and Flemming Topsøe *
Rønne Allé, Søborg, Denmark
and
Department of Mathematics
University of Copenhagen, Denmark

Abstract

To any discrete probability distribution P we can associate its entropy $H(P) = -\sum p_i \ln p_i$ and its index of coincidence $IC(P) = \sum p_i^2$. The main result of the paper is the determination of the precise range of the map $P \curvearrowright (IC(P), H(P))$. The range looks much like that of the map $P \curvearrowright (P_{\max}, H(P))$ where $P_{\max}$ is the maximal point probability, cf. research from 1965 (Kovalevskij [18]) to 1994 (Feder and Merhav [7]). The earlier results, which actually focus on the probability of error $1 - P_{\max}$ rather than $P_{\max}$, can be conceived as limiting cases of results obtained by methods here presented. Ranges of maps as those indicated are called *Information Diagrams.*

The main result gives rise to precise lower as well as upper bounds for the entropy function. Some of these bounds are essential for the exact solution of certain problems of universal coding and prediction for Bernoulli sources. Other applications concern Shannon theory (relations betweeen various measures of divergence), statistical decision theory and rate distortion theory.

Two methods are developed. One is topological, another involves convex analysis and is based on a "lemma of replacement" which is of independent interest in relation to problems of optimization of mixed type (concave/convex optimization).

*research of both authors has been supported by the COWI Foundation as well as by the Danish Natural Science Research Council.

Keywords. Entropy, index of coincidence, measure of roughness, information divergence, chi-squared divergence, Rényi entropy, entropy inequalities, concave/convex optimization, lemma of replacement, information diagrams.

1 Introduction, definitions

In information theory proper and in applications to probability theory, statistics and other fields, one often needs inequalities which relate entropy or relative entropy of distributions (i.e. divergence) to entropies of other distributions or to other quantities of interest. One of the most well known and useful inequalities of this type is *Pinskers inequality* ($D \geq \frac{1}{2}V^2$), cf. Pinsker [21], Csiszár and Körner [5], Problem I.3.17 or Fedotov, Harremoës and Topsøe [8]. This inequality allows one to conclude convergence of distributions from smallness of divergence. Some of our inequalities allow for the same conclusion but, in contrast to several other inequalities, they are of significance also when the deviation or "divergence" considered is moderate or large. However, the main results presented in this paper are more restrictive in another way as they only involve deviation from a uniform distribution. The very nature of our results, including also that they are best possible in a certain sense, open up for various applications which are indicated in more detail in the final section which offers a discussion of results obtained. Here we only mention that the key motivation of the authors is the link to certain problems of exact prediction, a theme that will be the subject of a forthcoming publication.

Returning to a comparison with Pinskers inequality, we note that this relates divergence between two distributions to the square of the l_1-distance between the distributions. Basically, what we shall investigate, is the relation to the l_2-distance.

By $M_+^1(\mathbb{N})$ we denote the set of discrete probability distributions P over $\mathbb{N}$, identified by the vector $(p_1, p_2, \dots)$ of point probabilites. We use the notation U_k for the generic uniform distribution over a k-set: $U_k = (\frac{1}{k}, \dots, \frac{1}{k}, 0, \dots)$ or isomorphic versions of this distribution. If, say U_k and U_{k+1} (or U_k and U_j with $j > k$) are considered at the same time, it is assumed without further comment that the support of U_k is contained in that of U_{k+1} (or U_j): $\text{supp}(U_k) \subseteq \text{supp}(U_{k+1})$ (or $\text{supp}(U_k) \subseteq \text{supp}(U_j)$). This convention corresponds to the ordering $p_1 \geq p_2 \geq \cdots$ of the point probabilities of the distributions concerned. Often we restrict attention to distributions with fixed finite support. The set of all distributions $P = (p_1, \dots, p_n)$ on the n-set $\{1, 2, \dots, n\}$ is denoted by $M_+^1(n)$.

As usual we denote by $H(P)$ the *entropy* of P:

$$H(P) = -\sum_i p_i \ln p_i.$$

We shall work with natural logarithms and use standard conventions, e.g. regarding $0 \cdot \ln 0$ (set to 0). *Relative entropy* or *divergence*, defined by

$$D(P\|Q) = \sum_i p_i \ln \frac{p_i}{q_i},$$

also plays a role for our investigations, though mainly as motivation for the study we shall embark on. In fact, we shall only have cases in mind for which Q is a uniform distribution, say $Q = U_n$, and for which $P \in M_+^1(n)$. Then, as $D(P\|U_n) = \ln n - H(P)$, it is possible to express bounds for $D(P\|U_n)$ as bounds for the entropy $H(P)$, and this is what we shall do.

With $P \in M_+^1(\mathbb{N})$ we also associate its *index of coincidence,* denoted $IC(P)$, and defined by

$$IC(P) = \sum_i p_i^2. \tag{1}$$

This quantity, which is the probability of getting "two of a kind" in two independent trials governed by the distribution P, is of significance in cryptoanalysis, cf. Friedman [9], Stinson [25] and Menezes et al. [20]. Simple transformations of the index of coincidence occur elsewhere in the literature as we shall comment on later.

Note the trivial inequality

$$IC(P) \leq \max_i p_i, \tag{2}$$

in particular $IC(P) \leq 1$, with equality for deterministic distributions. The difference $1 - IC(P)$ is sometimes called the *concentration measure.*

The index $IC(P)$ may also be thought of as the square of the l^2-norm of P, i.e., when using $\|\cdot\|$ to denote l^2-norm, as $IC(P) = \|P\|^2$. The deviation of a distribution $P \in M_+^1(n)$ from U_n (the most "flat" distribution in $M_+^1(n)$) can be measured by the norm-squared deviation:

$$\|P - U_n\|^2 = \sum_{i=1}^n (p_i - \frac{1}{n})^2 = IC(P) - \frac{1}{n}.$$

This quantity is the *measure of roughness,* and we denote it by $\mathrm{MR}_n(P)$. In the cited references this measure is only used for $P \in M_+^1(n)$. However, we shall find it convenient to use this terminology and notation for any distribution, i.e. we define

$$\mathrm{MR}_n(P) = IC(P) - \frac{1}{n}, \quad P \in M_+^1(\mathbb{N}). \tag{3}$$

If we do restrict attention to distributions in $M_+^1(n)$, it follows from the above that $IC(P) \geq \frac{1}{n}$ for $P \in M_+^1(n)$. Therefore, referring also to (2), we see that for $P \in M_+^1(n)$, $IC(P)$ varies between $\frac{1}{n}$ and 1. When one works in $M_+^1(n)$ it is often convenient to use the measure of roughness rather than the index of coincidence. Note that then $\mathrm{MR}_n(P)$ is closely related to the often used *chi-squared divergence* between P and U_n, indeed, then

$$\mathrm{MR}_n(P) = \frac{1}{n}\chi^2(P, U_n) \tag{4}$$

holds where, as usual,

$$\chi^2(P, Q) = \sum_i \frac{|p_i - q_i|^2}{q_i}.$$

For our purposes we find it useful also to introduce some "relative" quantities associated with a distribution. Often, this leads to simpler formulas. The new quantities may be thought of either as relative indices of coincidence or as relative measures of roughness. We prefer the latter terminology. Thus, let $1 \leq k < j$ be natural numbers. The (j,k)-*relative measure of roughness* of $P \in M_+^1(\mathbb{N})$ is defined by

$$\overline{\mathrm{MR}}_{j,k}(P) = \frac{\mathrm{MR}_j(P)}{\mathrm{MR}_j(U_k)} = \frac{IC(P) - \frac{1}{j}}{\frac{1}{k} - \frac{1}{j}}. \tag{5}$$

Cases with $0 \leq \overline{\mathrm{MR}}_{j,k} \leq 1$ are the most interesting ones but clearly, $\overline{\mathrm{MR}}_{j,k}$ may be negative (take $P = U_{j+1}$, for instance) or larger than 1 (take $P = U_1$ when $k > 1$). Essentially, the validity of the inequalities $0 \leq \overline{\mathrm{MR}}_{j,k} \leq 1$ depends on a kind of "complexity" of P with respect to the index of coincidence. To be precise, we say that P is of *IC-complexity class* k if

$$IC(U_{k+1}) = \frac{1}{k+1} < IC(P) \leq \frac{1}{k} = IC(U_k). \tag{6}$$

The relevance of this kind of division has been noted before, first it seems in 1965 by Kovalevskij [18]. We shall return to this in the discussion. Through this definition, $M_+^1(\mathbb{N})$ is decomposed into complexity classes, one for each $k = 1, 2, \ldots$.

We note the following curious relation which will be useful later on and right now can be taken to motivate the introduction of the relative measures of roughness.

Lemma 1.1. *Let $j > k$. Then the (j,k)-relative measure of roughness of any mixture of a uniform distribution U_k over a k-set and a uniform distribution U_j over a larger j-set is given by the formula*

$$\overline{\mathrm{MR}}_{j,k}\left((1-x)U_j + xU_k\right) = x^2, \quad 0 \le x \le 1.$$

The simple purely computational proof is left to the reader.

Except for the technical Section 5, we only need the $\overline{\mathrm{MR}}$-quantities when $k = j-1$ and when $k = 1$. For these two cases we introduce special notation, using sub- and superscripts in a way which will later be suggestive (see, respectively lower, and upper curves in Figure 1 below). As to the first case of special importance, $j = k+1$, we shall write $\overline{\mathrm{MR}}_k$ in place of $\overline{\mathrm{MR}}_{k+1,k}$, i.e.

$$\overline{\mathrm{MR}}_k(P) = \frac{\mathrm{MR}_{k+1}(P)}{\mathrm{MR}_{k+1}(U_k)} = \frac{IC(P) - \frac{1}{k+1}}{\frac{1}{k} - \frac{1}{k+1}}. \tag{7}$$

For the second case of special interest, $k = 1$, the value of j will be the maximal value under consideration in a given context, denoted n below. We shall then write $\overline{\mathrm{MR}}^n(P)$ in place of $\overline{\mathrm{MR}}_{n,1}(P)$, i.e.

$$\overline{\mathrm{MR}}^n(P) = \frac{\mathrm{MR}_n(P)}{\mathrm{MR}_n(U_1)} = \frac{IC(P) - \frac{1}{n}}{1 - \frac{1}{n}}. \tag{8}$$

If $P \in M_+^1(n)$, then

$$\overline{\mathrm{MR}}^n(P) = \frac{1}{n-1}\chi^2(P, U_n).$$

Usually, the value of n is understood from the context.

We shall see later that the quantity $1 - \overline{\mathrm{MR}}^n(P)$ plays a special role for some of the inequalities we will study. We note that this quantity behaves as a kind of entropy and belongs, apart from a constant, to a class of entropy-like functions first considered by Havrda and Charvát, cf. [17]. The important case considered here was called, again apart from a constant, "quadratic entropy" by Vajda [32] and reintroduced by Daróczy [6]. More details about previous research in this area (axiomatics and basic properties) can be found in Vajda and Vašek [33] and in references quoted there.

The main purpose of the paper is to study the relationship between $D(P\|U_n)$ and $\mathrm{MR}_n(P)$ (or, equivalently, $\chi^2(P,U_n)$). Qualitatively, $D(P\|U_n)$ and $\mathrm{MR}_n(P)$ both measure deviation from the uniform distribution U_n. It is therefore natural to compare these quantities. If P is close to U_n, an approximate comparison is easy as we have

$$D(P\|U_n) = \sum_{i=1}^{n} p_i \ln\left(1 + \frac{p_i - \frac{1}{n}}{\frac{1}{n}}\right),$$

and using the approximation $\ln(1+x) \approx x - x^2/2$ we then find that

$$D(P\|U_n) \approx \sum_{i=1}^{n}\left(\left(p_i - \frac{1}{n}\right) + \frac{1}{n}\right)\cdot a\left(n\left(p_i - \frac{1}{n}\right) - \frac{n^2}{2}\left(p_i - \frac{1}{n}\right)^2\right),$$

hence, for $P \approx U_n$,

$$D(P\|U_n) \approx \frac{n}{2}\,\mathrm{MR}_n(P) = \frac{1}{2}\chi^2(P,U_n) \tag{9}$$

(compare also with Corollary 2.9 of the next section).

In other situations, approximation arguments as the above cannot be invoked. Instead, we shall investigate the relationship between $D(P\|U_n)$ and $\mathrm{MR}_n(P)$ by determining the exact range of the map $P \curvearrowright (\mathrm{MR}_n(P), D(P\|U_n))$ of $M_+^1(n)$ into $\mathbb{R}^2$. As $\mathrm{MR}_n(P) = IC(P) - \frac{1}{n}$ and $D(P\|U_n) = \ln n - H(P)$, it appears simpler to work with the map $P \curvearrowright (IC(P), H(P))$. Note that this map, referred to as the *IC/H-map*, is well defined on all of $M_+^1(\mathbb{N})$ if we allow for infinite values in the range (points of the form (x,∞) with $0 < x < 1$).

The main problem then is to determine the range of the IC/H-map and, in fact more important, to determine the range of the restriction of this map to $M_+^1(n)$. The sets in question we denote by Δ and Δ_n, respectively:

$$\begin{aligned} \Delta &= \left\{(IC(P), H(P)) \mid P \in M_+^1(\mathbb{N})\right\} && (10)\\ \Delta_n &= \left\{(IC(P), H(P)) \mid P \in M_+^1(n)\right\}. && (11) \end{aligned}$$

We refer to Δ as well as to Δ_n as *IC/H-diagrams*. If necessary in a given context, Δ will be called the *full IC/H-diagram* and Δ_n the *restricted IC/H-diagram*. These diagrams are special instances of so-called *Information Diagrams* which will occur in various variants in the sequel and be further discussed in the final section.

2 Statements of the main results

Regarding the nature of the IC/H-diagrams Δ and Δ_n, $n \geq 2$, we first note that the points Q_k, $k \geq 1$, which correspond to uniform distributions, i.e. the points given by

$$Q_k = (IC(U_k),\, H(U_k)) = \left(\frac{1}{k}, \ln k\right), \quad k \geq 1$$

all belong to the diagram Δ. These points lie on the smooth curve $y = -\ln x$, $0 < x \leq 1$. In fact, this curve is a lower bounding curve for Δ as is easily seen by an application of Jensen's inequality:

$$H(P) = -\sum_i p_i \ln p_i \geq -\ln \sum_i p_i^2 = -\ln IC(P). \tag{12}$$

This inequality holds for all $P \in M_+^1(\mathbb{N})$, and the derivation shows that the points Q_k, $k \geq 1$, are the only points in Δ which actually lie *on* the curve $y = -\ln x$. All other points in Δ lie above this curve.

It is interesting—and perhaps a bit surprising—that the theoretically best lower bounding curve for the IC/H-diagrams is not smooth but has certain singularities at the points Q_k. This is illustrated in Figure 1 which displays the diagram Δ_n (for $n = 5$).

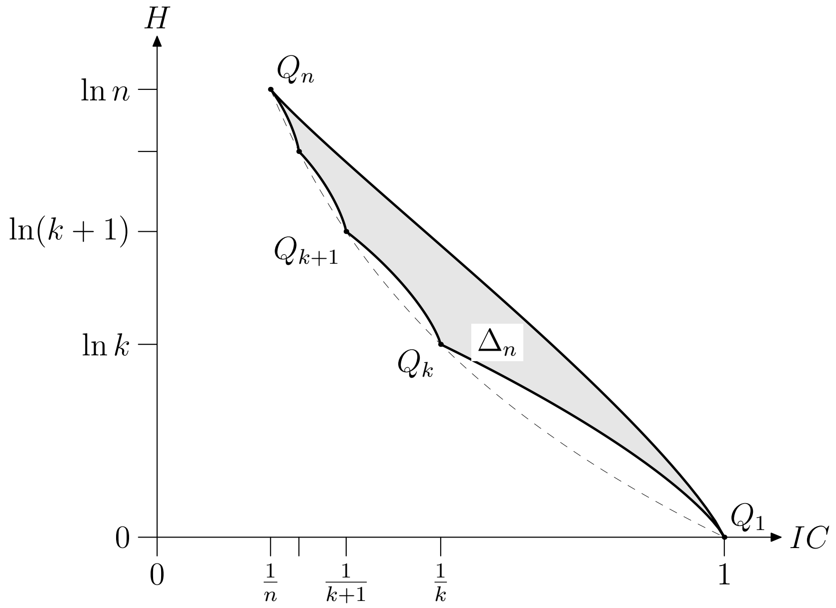

Fig.1. The restricted IC/H-diagram $\Delta_n (n = 5,\, k = 2)$.

The shaded region is Δ_n. This region lies between the *lower cascade* which is composed of certain curves connecting the points Q_{k+1} and Q_k, $k = n-1, n-2, \ldots, 1$, and the *upper arc* which connects Q_n and Q_1. We have also indicated the lower bounding curve $x \curvearrowright -\ln x$ (the dashed curve).

In order to understand the content of Figure 1 fully, it only remains to clarify the meaning of the curves occurring there. It is convenient first to introduce a notation for the IC/H-map:

$$\vec{\varphi}(P) = (IC(P), H(P)), \quad P \in M_+^1(\mathbb{N}). \tag{13}$$

For $k \geq 1$, the *arc joining Q_{k+1} and Q_k* is the oriented curve from Q_{k+1} to Q_k given by the parametrization

$$s \curvearrowright \vec{\varphi}\left((1-s)U_{k+1} + sU_k\right), \quad 0 \leq s \leq 1. \tag{14}$$

This curve, taken with orientation as given by (14), we denote by $\frown Q_{k+1}Q_k$. The *lower cascade* referred to above is the piecewise smooth curve, denoted $\frown\frown Q_nQ_1$, and given by

$$\frown\frown Q_nQ_1 = \frown Q_nQ_{n-1} + \frown Q_{n-1}Q_{n-2} + \cdots + \frown Q_2Q_1. \tag{15}$$

The *upper arc* is determined by the parametrization

$$s \curvearrowright \vec{\varphi}\left((1-s)U_n + sU_1\right), \quad 0 \leq s \leq 1.$$

This curve we denote $\frown Q_nQ_1$ and the same curve, but taken in opposite direction, we denote by $\frown Q_1Q_n$.

The main result can then be formulated as follows:

Theorem 2.1. *Let $n \geq 3$. The curve $\mathscr{J}_n = \frown\frown Q_nQ_1 + \frown Q_1Q_n$ is a positively oriented Jordan curve in the plane, and the bounded region which this curve determines (including $\mathscr{J}_n$ itself) coincides with the restricted IC/H-diagram Δ_n.*

Remark. For $n = 2$, we find that Δ_2 consists of all points on the arc $\frown Q_2Q_1$. This may be considered to be a degenerate case of the theorem which then holds for $n \geq 2$.

From Theorem 2.1 it is easy to deduce the following corollaries:

Corollary 2.2. *For $n \geq 3$, the restricted IC/H-diagram Δ_n consists of the points (x, y) with $1/n \leq x \leq 1$ for which there exists y_* and y^* such that $y_* \leq y \leq y^*$ and such that (x, y_*) lies on the lower cascade and (x, y^*) on the upper arc.*

Corollary 2.3. *The full IC/H-diagram Δ consists of the point $(1,0)$ and of all points (x,y) with $0<x<1$ for which there exists $k\geq 1$ and $y_*\leq y$ such that $(x,y_*)\in {}^\frown Q_{k+1}Q_k$ (here, we also allow the value $y=\infty$)*

Yet another corollary to Theorem 2.1 is obtained by a closer inspection of the nature of the curves ${}^\frown Q_{k+1}Q_k$. Indeed, the entropy function considered as a function of the index of coincidence on the curves in question turns out to be concave. This then shows that the following result holds:

Theorem 2.4. (i). *For $n\geq 3$, $Q_1,Q_2,\dots,Q_n$ are extremal points of the restricted diagram Δ_n and these points are the only extremal points on the lower cascade ${}^{\frown\frown}Q_nQ_1$.*

(ii). *The set of extremal points for the full IC/H-diagram Δ coincides with the set of points $Q_1,Q_2,\dots$.*

(iii). *The diagram Δ lies above all secants joining neighbouring points Q_{k+1} and Q_k, $k\geq 1$.*

We remark that for $n=2$, all points in Δ_2 $(= {}^\frown Q_2Q_1)$ are extremal for Δ_2. We warn the reader that for $n\geq 3$ only some points on the upper arc ${}^\frown Q_nQ_1$ are extremal points of Δ_n (since this curve has an inflection point).

Concerning all three parts of Theorem 2.4 it is important to note that the lower bounding curve $y=-\ln x$, $0<x\leq 1$, is convex.

A simple transformation applied to the restricted IC/H-diagram gives the shape of the $\overline{\mathrm{MR}}^n$*/D-diagram* which by definition is the set of points $(\overline{\mathrm{MR}}^n(P), D(P\|U_n))$ for $P\in M_+^1(n)$. Note that the relative measure of roughness $\overline{\mathrm{MR}}^n(P)$ varies between 0 and 1 and appears to be a natural choice of parameter for the kind of diagram considered. Of course, if we had divided the divergence with its maximal value $(\ln n)$, also the other parameter would vary between 0 and 1. For $n=5$ the $\overline{\mathrm{MR}}^n/D$-diagram is shown in Figure 2.

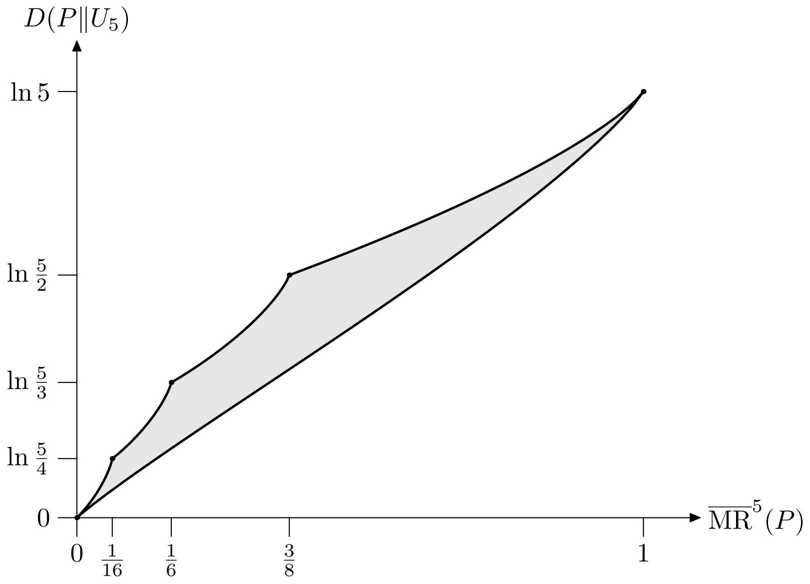

Fig.2. The $\overline{\mathrm{MR}}/D$-diagram for $n = 5$.

Another equivalent form of Theorem 2.1 is obtained by replacing $IC(P)$ by the Rényi entropy $H_2(P)$ of order 2, cf. Rényi [22], Cover and Thomas [3] or Csiszár and J. Körner [5], for example. As

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

it is a simple matter to transform the IC/H-diagram Δ_n into the equivalent H_2/H-diagram. The result of this transformation, again for $n = 5$, is shown in Figure 3. We note that the logarithmic lower bounding curve for the IC/H-diagram is transformed into the diagonal (the identity map) for the H_2/H-diagram. It should also be remarked that, apparently, the bounding curves have better convexity properties in the transformed diagram.

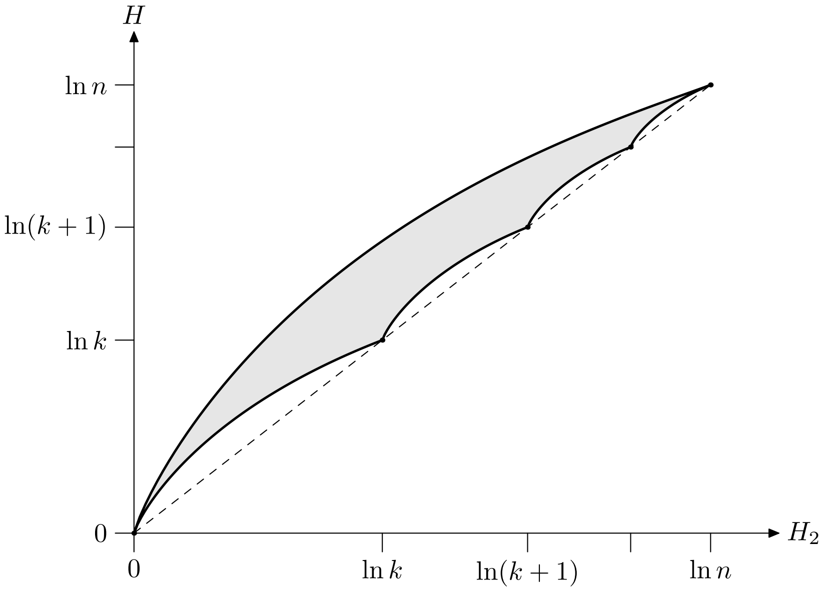

Fig.3. The H_2/H-diagram ($n = 5$, $k = 2$).

The results so far were expressed in a geometric way. The analytic equivalents are also important. First, let us reformulate Corollaries 2.2 and 2.3 using the relative measure of roughness:

Theorem 2.5. (i) *Let $P \in M_+^1(n)$ for some $n \geq 2$ and put $r = \overline{\mathrm{MR}}^n(P)$. Then*

$$H(P) \leq H\left((1 - r^{1/2})U_n + r^{1/2}\, U_1\right). \tag{16}$$

(ii) *Let $P \in M_+^1(\mathbb{N})$ and assume that P is of IC-complexity class k. Put $r_k = \overline{\mathrm{MR}}_k(P)$. Then*

$$H(P) \geq H\left((1 - r_k^{1/2})U_{k+1} + r_k^{1/2}\, U_k\right). \tag{17}$$

Proof. (i): By Lemma 1.1, the distributions P and $(1-r^{1/2}\,)U_n+r^{1/2}\,U_1$ which occur in (16) have the same $(n, 1)$-relative measure of roughness, hence also the same index of coincidence. The result then follows from Corollary 2.2. The proof of (ii) is similar, referring this time to Corollary 2.3. □

The lower bound (17) (and certain extensions of these bounds) have been obtained recently also by György and Linder [13] who applied the results to the study of problems of quantization and rate distortion theory (for general background, see Cover and Thomas [3] and the recent survey paper by Gray and Neuhoff [11]). That this type of research also leads to certain *diagrams* as the IC/H-diagrams can be seen as follows (brief indication): Consider a random variable X which is uniformly distributed on $[0,1]$. If Q is a nearest neighbour quantizer with finite range and if we use Euclidean distance as distortion measure, then – after a simple computation – it is seen that the distortion of Q equals $\frac{1}{4}IC(Q')$ where Q' denotes the distribution of $Q(X)$. As the *rate* of the quantizer is the entropy $H(Q')$, we see that a study of the connection between rate and distortion essentially amounts to a study of the IC/H-diagram.

Exploiting part (iii) of Theorem 2.4 analytically we are led to a set of inequalities which to a great extent provoked the research reported on in this paper. To formulate the inequalities in a convenient way, we first introduce the constants

$$e_k = \left(1+\frac{1}{k}\right)^k, \quad k \geq 1. \tag{18}$$

Note that as k increases, e_k increases and has e as its limit value.

We can then state another key result, announced in Topsøe [29]:

Theorem 2.6. *For any $P \in M_+^1(\mathbb{N})$ and any $k \geq 1$, the inequality*

$$H(P) \geq \alpha_k - \beta_k IC(P) \tag{19}$$

holds, with the constants α_k and β_k defined by

$$\alpha_k = \ln(k+1) + \ln e_k\,, \qquad \beta_k = (k+1)\ln e_k, \quad k \geq 1. \tag{20}$$

Proof. The proof follows upon noting that as the straight line with equation $y = \alpha_k - \beta_k x$ contains both points Q_{k+1} and Q_k, it must be the equation for the secant joining these points. The result then follows from Theorem 2.4, (iii). □

The inequality (19) is "rigid" in the sense that the constants cannot be improved. This follows as equality holds if P is either a U_{k+1} - or a U_k - type distribution. We also express this by saying that the inequality is *anchored* in the U_{k+1}- and in the U_k- type distributions.

It is pretty clear that inequalities as the above are of relevance for error probability analysis, cf. e.g. Gallager [10]. There we also find an inequality which is closely related to (19) in the case $k = 1$, viz. the inequality $H \geq 1 - IC$ (in our notation), cf. Exercise 4.7 of [10]. Note that (19) contains the following strengthened version of this inequality: $H \geq \ln 4(1 - IC)$. Here, $\ln 4$ is the best constant as equality holds for the distributions U_1 and U_2. It should be noted that in the quoted exercise this optimal version of the inequality is also noted, but, strictly speaking, only for $P \in M_+^1(2)$.

It is noteworthy, in fact surprising, that though previous research regarding information diagrams have worked with the probability of error in place of index of coincidence, exactly the same constants α_k and β_k have appeared in related inequalities (due to the form of inequalities considered it is in fact $\alpha_k + \beta_k$ and β_k that appears in previous research). To be precise, we refer to equations (12) in Kovalevskij [18], (6) in Tebbe and Dwyer [26], (29) in Ben-Bassat [1] and, finally, to equation (14) in Feder and Merhav [7]. An explanation for this phenomenon is given in the discussion.

As indicated in the announcement [29], the inequality (19) can be proved in a straightforward way by induction (over n with $P \in M_+^1(n)$) in case $k = 1$. Simple direct proofs of (19) for other values of k are not known to the authors.

The inequality (19) is really only of significance for distributions P of IC-index class k since, for other distributions, this inequality is weaker than the more elementary inequality (12).

Again, a reformulation in terms of the relative measure of roughness is illuminating:

Corollary 2.7. *Let $P \in M_+^1(\mathbb{N})$, let $k \geq 1$ and put $r = \overline{\mathrm{MR}}_k(P)$. Then*

$$H(P) \geq (1 - r)H(U_{k+1}) + rH(U_k).$$

In connection with this corollary, it lies nearby to notice the following strengthening of the usual concavity property of the entropy function:

$$H\left((1 - x)U_{k+1} + xU_k\right) \geq (1 - x^2)H\left(U_{k+1}\right) + x^2 H\left(U_k\right) \tag{21}$$

which, of course, is lower bounded by $(1-x)H\left(U_{k+1}\right)+xH(U_k)$. The inequality (21) follows from Corollary 2.7, using also Lemma 1.1. The exponent in (21) is best possible in the sense that if we replace the two occurrences of x^2 by x^β for some exponent β, the inequality you get fails unless $\beta \leq 2$ (consider small values of x).

The full proof of the geometric results, especially Theorem 2.1, uses facts from topology and will be postponed until Section 5. The reader may turn directly to that section. However, in Sections 3 and 4 we show that it is possible to prove important parts of the main result, viz. Theorem 2.6, and from that Theorem 2.4, by a direct analytical method which we believe many will find more elementary. The technique consists of two parts, a general part and a specific part. The general part—which we believe is of independent interest—is a method of reduction, called the "lemma of replacement", which simplifies the study of certain optimization problems which involve a basic function of "mixed type"—first concave, then convex. In Section 4 the lemma of replacement is applied to a specific optimization problem, thereby proving Theorem 2.6. We point out that this technique gives a further reduction than would be obtainable had we introduced Lagrange multipliers in the usual fashion.

In Section 6 we consider Δ_n and use the nature of the upper arc $\frown Q_n Q_1$ to derive upper bounds for the entropy function. A main result from that section may be stated as follows, expressed in terms of the relevant relative measure of roughness:

Theorem 2.8. *There exists an increasing sequence $(\tau_n)_{n\geq 2}$ of constants with $\tau_2 = (\ln 4)^{-1} \approx 0.7213$ and $\lim_{n\to\infty} \tau_n = 1$ such that the inequality*

$$H(P) \leq \ln n \cdot \left(1 - \overline{\mathrm{MR}}^n(P)\right)^{\tau_n} \tag{22}$$

holds for $n \geq 2$ and all $P \in M_+^1(n)$.

Section 6 contains more information about the best (i.e. largest) constants for this inequality. For now we only point out the following corollary:

Corollary 2.9. *For any $n \geq 2$ and any $P \in M_+^1(n)$ the following inequalities hold:*

$$H(P) \leq \ln n \left(1 - \tau_n \overline{\mathrm{MR}}^n(P)\right), \tag{23}$$

$$D(P\|U_n) \geq \frac{\tau_n \ln n}{n-1} \chi^2(P, U_n). \tag{24}$$

Proof. The function $x \curvearrowright \ln n \cdot (1-x)^{\tau_n}$ which appears in (22) is concave, hence lies below any of its tangents. Considering the tangent through $(0, \ln n)$, the inequality (23) results. Rewriting (23), we obtain (24). □

Figure 4 illustrates the various upper bounds obtained for the entropy function. The example $P \in M_+^1(5)$ has been chosen (as for Figure 1). The closest bound is the one obtained from (22) for which the exponent is $\tau_5 \approx 0.8473$. A somewhat looser bound is obtained by considering the exponent $\tau_2 = (\ln 4)^{-1} \approx 0.7213$ which can always be used according to Theorem

2.8 (whereas the maximal exponents are difficult to calculate and may not be available). Finally, Figure 4 also shows a straight-line upper bound for the entropy function, viz. the one obtained from (23).

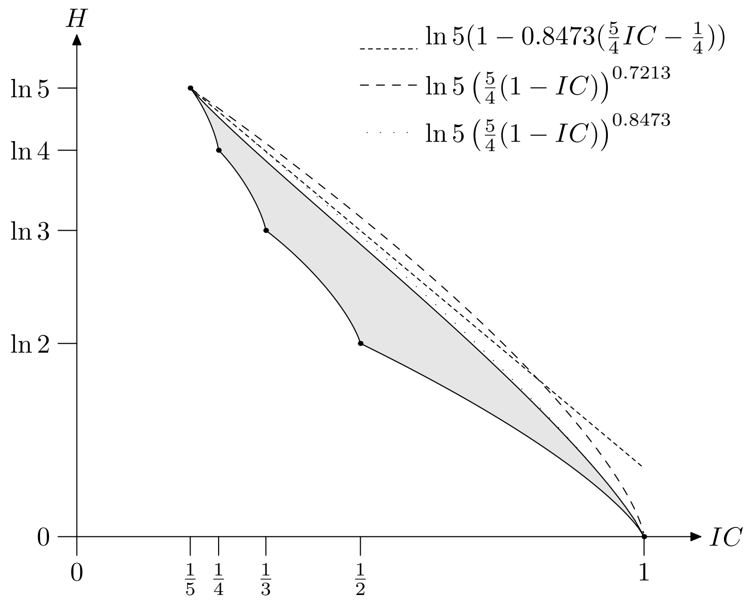

Fig.4. Upper bounds for entropy in terms of index of coincidence for $n = 5$.

Simple but not very sharp further bounds are the following:

$$1 \leq H(P) + IC(P) \leq \ln n + \frac{1}{n}. \tag{25}$$

The left-hand inequality follows from (12) and the right-hand inequality – previously announced as equation (30) of Topsøe [30]– is obtained by noting that as $\tau_n \geq \tau_2 = (\ln 4)^{-1}$, $\tau_n \ln n \geq 1 - \frac{1}{n}$ and then the inequality follows from (8) and (23).

Note that the lower bound in (25) is the sum of the minimum of H and the maximum of IC, whereas the upper bound is the sum of the maximum of H and the minimum of IC. Graphically, the inequalities (25) give a parallel band in the $IC/H-$plane within which the information diagram Δ_n must lie. Of course, much more narrow confinements on the position of Δ_n, and still with easily computed limits follow from other inequalities, say (12) for lower bounds and the more sophisticated upper bounds in (22) and (23) when combined with specific lower bounds of τ_n, in particular, we point to the bound $\tau_n \geq \frac{\ln n}{\ln n+1}$.

The final section offers a discussion which also indicates some ideas for further development.

3 The lemma of replacement

Throughout this section $f : [0,1] \to \mathbb{R}$ will denote a function with $f(0) = 0$ for which there exists $0 \le \xi \le 1$ such that the restriction of f to $[0,\xi]$ is concave and the restriction of f to $[\xi,1]$ is convex. Though the value ξ may not be uniquely determined and though f may not be twice differentiable, we call ξ the *inflection point* of f. In "standard" applications, ξ will in fact be uniquely determined from f and given as the solution to the equation $f''(\xi) = 0$.

We refer to the situation described by saying that f is a *concave/convex function with inflection point* ξ.

Well known arguments involving convex (or concave) functions show that f is bounded on $[0,1]$ and continuous except possibly at the points $0, \xi$ and 1.

With the concave/convex function f we associate the map $F : M_+^1(\mathbb{N}) \to \mathbb{R} \cup \{\infty\}$ defined by

$$F(P) = \sum_i f(p_i). \tag{26}$$

where, as usual, the p_i's denote the point probabilities of P.

Note that F is bounded below. To see this, let $f \ge c$ on $[0,1]$ and put $\nu = \lfloor (1/\xi) \rfloor$. Noting also that $f(t) \ge (f(\xi)/\xi)t$ for all $t \in [0,\xi]$, it is easy to show that $F(P) \ge -|f(\xi)/\xi| + \nu c$ for all $P \in M_+^1(\mathbb{N})$.

Consider now the problem to determine either the maximum (supremum) or the minimum (infimum) of $F(P)$, possibly with P restricted to $M_+^1(n)$. What we shall show is that any "initial" distribution P may be replaced by a distribution of special type for which the value of F is closer to the extreme value.

The distributions of "special type" which we shall encounter are in fact those we met in Section 2, viz. mixtures of two uniform distributions, either a mixture of U_n and U_1 for some $n \ge 3$ (when we seek upper bounds) or a mixture of U_{k+1} and U_k for some $k \ge 1$ (when we seek lower bounds).

We shall use "co" to denote "convex hull", e.g. $co\{U_n, U_1\}$ denotes the set of mixtures of U_n and U_1.

Theorem 3.1 (Lemma of replacement). *Let f be a concave/convex function and let $P \in M_+^1(\mathbb{N})$. Then the following holds:*

(i). *There exists a $k \ge 1$ and a distribution $P_0 \in co\{U_{k+1}, U_k\}$ such that $F(P_0) \le F(P)$.*

(ii). *If, for a fixed* $n \geq 3$, $P \in M_+^1(n)$, *there exists a distribution* $P_1 \in co\{U_n, U_1\}$ *such that* $F(P_1) \geq F(P)$.

Proof. Let f have ξ as inflection point. We first pay attention to the restriction of F to $M_+^1(n)$ for a fixed $n \geq 2$. Let $P \in M_+^1(n)$ and assume that the point probabilities in P are ordered: $p_1 \geq p_2 \geq \cdots \geq p_n$. Let $\nu \in \{1, \dots, n+1\}$ be determined so that $p_i \geq \xi$ for $i < \nu$ and $p_i < \xi$ for $i \geq \nu$. The cases $\nu = 1$ and $\nu = n+1$ are extreme cases for which either no point probability or all point probabilities are $\geq \xi$. We may assume that $\nu \leq n$.

Denote by K the set of $(q_\nu, \dots, q_n)$ for which $0 \leq q_i \leq \xi$ for $\nu \leq i \leq n$ and

$$\sum_{i=\nu}^{n} q_i = \sum_{i=\nu}^{n} p_i$$

hold. Then K is a compact and convex subset of $\mathbb{R}^{n-\nu+1}$. Let $G : K \to \mathbb{R}$ denote the function defined by

$$G(q_\nu, \dots, q_n) = F(p_1, \dots, p_{\nu-1}, q_\nu, \dots, q_n), \quad (q_\nu, \dots, q_n) \in K.$$

Then G is a concave and lower semi-continuous function on K, hence G assumes its minimal value at an extremal point of K. The extremal elements are those elements $(q_\nu, \dots, q_n)$ for which the strict inequalities $0 < q_j < \xi$ at most hold for one value of j. We can thus determine an element of K of the form $(\xi, \xi, \dots, \xi, r, 0, \dots, 0)$ with $0 \leq r < \xi$ such that

$$F(P) \geq F(p_1, \dots, p_{\nu-1}, \xi, \xi, \dots, \xi, r, 0, \dots, 0).$$

Possibly, certain extreme cases occur: no ξ's, no r or no 0's. In fact we may assume that no ξ's occur since otherwise they can be moved to the group of large ($\geq \xi$) point probabilities. We may thus assume that $F(P) \geq F(p_1, \dots, p_{\nu-1}, r, 0, \dots, 0)$. As f is convex on $[\xi, 1]$, it is easy to see that

$$F(p_1, \dots, p_{\nu-1}, r, 0, \dots, 0) \geq F(p, \dots, p, r, 0, \dots, 0)$$

with $p = (\sum_1^{\nu-1} p_i)/(\nu - 1)$. The probability vector $(p, \dots, p, r, 0, \dots, 0)$ is a mixture of U_{k+1} and U_k for a certain k (note that $p > r \geq 0$).

To prove the general validity of (i), consider an arbitrary distribution $P = (p_1, p_2, \dots)$ in $M_+^1(\mathbb{N})$. We may assume that $p_1 \geq p_2 \geq \dots$, that all the p_i are positive and also, that $F(P) < \infty$. It follows that f is continuous at 0. Put $P_n = (p_1, \dots, p_n, q_n)$ with $q_n = \sum_{n+1}^{\infty} p_i$. If we apply the same procedure to P_n as the one we applied to the distribution given in the first part of the proof, we find that provided $q_n < \xi$, the distribution we end up with, now conceived as a distribution in $M_+^1(\mathbb{N})$, is independent of n. Thus, there exists k and $P_0 \in co\{U_{k+1}, U_k\}$ such that $F(P_n) \geq F(P_0)$ for n sufficiently large. Now, for n sufficiently large,

$$F(P) = F(P_n) + \sum_{i=n+1}^{\infty} f(p_i) - f(q_n) \geq F(P_0) + \frac{f(\xi)}{\xi} q_n - f(q_n)$$

and it follows that $F(P) \geq F(P_0)$. This proves (i).

The situation dealt with in (ii) can be treated in a similar manner. Details are left to the reader. □

4 Entropy inequalities via the lemma of replacement

We shall here apply the technique developed in the previous section in order to establish the inequalities (19) of Theorem 2.6.

Let us fix $k \geq 1$ and define α_k and β_k as in (20). For $P \in M_+^1(\mathbb{N})$ we must then prove that $H(P) \geq \alpha_k - \beta_k IC(P)$. In order to do so, consider the function $f_k : [0,1] \to \mathbb{R}$ defined by

$$f_k(x) = -x \ln x - \alpha_k x + \beta_k x^2.$$

As

$$f_k''(x) = -\frac{1}{x} + 2\beta_k,$$

the function f_k is concave/convex with inflection point $\xi_k = (2\beta_k)^{-1}$, i.e.

$$\xi_k = \frac{1}{2k(k+1)\ln(1+\frac{1}{k})}.$$

Clearly, $\frac{1}{2(k+1)} < \xi_k < \frac{1}{2k}$. The function associated with f_k we denote by F_k, cf. (26). We find that

$$F_k(P) = H(P) - \alpha_k + \beta_k \cdot IC(P), \quad P \in M_+^1(\mathbb{N}).$$

Thus, what we have to prove is that $F_k \geq 0$. From the lemma of replacement we see that it suffices to prove that $F_k(P) \geq 0$ if $P \in co\{U_{j+1}, U_j\}$ for some

$j \geq 1$. This inequality is trivial if $j \neq k$ since then the stronger inequality $H(P) \geq -\ln IC(P)$, i.e. $H(P) \geq H_2(P)$, holds (look at it geometrically, cf. Figure 1, observing that the line containing Q_{k+1} and Q_k has the equation $y = \alpha_k - \beta_k x$). We have thus reduced the problem to proving that $F_k(P) \geq 0$ for all $P \in co\{U_{k+1}, U_k\}$.

Define the function $\varphi_k : [0,1] \to \mathbb{R}$ by

$$\varphi_k(x) = F_k\left((1-x)U_{k+1} + xU_k\right).$$

The desired inequality $\varphi_k \geq 0$ can then be proved by elementary considerations as follows: Simple calculations show that

$$\varphi_k'(x) = \frac{1}{k+1}\ln\frac{1-x}{k+x} + 2x\ln\left(1+\frac{1}{k}\right) + \frac{1}{k+1}\ln k\,,$$

$$\varphi_k''(x) = -\frac{1}{(1-x)(k+x)} + 2\ln\left(1+\frac{1}{k}\right).$$

Therefore, apart from the already known relations $\varphi_k(0) = \varphi_k(1) = 0$, we see that $\varphi_k'(0) = 0$, that $\varphi_k'(1) = -\infty$ and that

$$\varphi_k''(0) = -\frac{1}{k} + 2\ln\left(1+\frac{1}{k}\right) > \frac{k-1}{k(k+1)}) \geq 0.$$

It is then easy to show that $\varphi_k \geq 0$. Indeed, if φ_k assumed a negative value, φ_k would have at least three inflection points which is impossible (in fact, φ_k only has one inflection point).

Backtracing our steps, we see that Theorem 2.6 is proved. Theorem 2.4 is obtained as a simple consequence. The further results can, in part, be obtained by more elaborate considerations based on the lemma of replacement. However, one will invariably enter into more and more topologically oriented considerations and the full power of Theorem 2.1 can hardly be obtained by further elaborations based only on the lemma of replacement. In particular, it is difficult to exclude the possibility of holes in the diagram Δ_n situated between the lower cascade and the upper arc. The further study which is needed calls for genuine topological considerations.

5 The IC/H-diagram

In this section we give the full proof of Theorem 2.1. Difficulties inherent in the problem are that there are interior points in the simplex $M_+^1(n)$ which are mapped by the IC/H-map to boundary points of the range and also, there are boundary points of the simplex which are mapped to interior points of the range. By decomposing the simplex properly, the first difficulty is avoided and then methods from topology can be invoked. Therefore, our proof will combine general topological facts with specific computations.

We first remind the reader of some basic topological notions, cf. Bredon [2] or Greenberg and Harber [12] or, for recent new proofs, Thomassen [27].

For a subset A of a topological space, $\mathrm{int}(A) = A^\circ$ denotes the interior of A and $\partial(A)$ the boundary of A. We shall also work with connectedness components, just called components, but only need this concept for open subsets of the plane $\mathbb{R}^2$. Thus, if $G \subseteq \mathbb{R}^2$ is open, the *components* of G are the maximal connected subsets of G. Recall that a *Jordan curve* is a set $\mathscr{L} \subseteq \mathbb{R}^2$ which is homeomorphic to the circle. By the *Jordan curve theorem*, a Jordan curve $\mathscr{L}$ divides the plane in exactly two components, say G_0 and G_1, i.e. these sets are the components of the complement $\mathbb{R}^2 \setminus \mathscr{L}$. Both components have $\mathscr{L}$ as their boundary: $\partial(G_0) = \partial(G_1) = \mathscr{L}$, and one of the components, let it be G_0, is bounded. We refer to the compact set $G_0 \cup \mathscr{L}$ as the *compact region determined by* $\mathscr{L}$. Apart from the Jordan curve theorem we need another non-trivial fact from topology, viz. that no proper subset of a Jordan curve $\mathscr{L}$ divides the plane in two components, i.e. if $\mathscr{L}' \subseteq \mathscr{L}$ and $\mathscr{L}' \neq \mathscr{L}$, then $\mathbb{R}^2 \setminus \mathscr{L}'$ is connected.

We shall also find it convenient to extend the more specific notions introduced in Section 1. For natural numbers $1 \leq i, j \leq n$, $\gamma_{j,i}$ denotes the oriented curve in $\mathbb{R}^2$ which connects Q_j and Q_i, and which is given by the parametrization $s \curvearrowright \vec{\varphi}(P_s)$, $0 \leq s \leq 1$ where

$$P_s = (1-s)U_j + sU_i.$$

Formally this definition allows for the singular case $i = j$, which will, however not be of relevance below. If necessary to stress the dependency on i and j, we write $P_s^{j,i}$ in place of P_s. If $i_1, i_2, \ldots, i_k$ are integers in $[1, n]$, we denote by $\gamma_{i_1,\ldots,i_k}$ the curve $\gamma_{i_1,i_2} + \cdots + \gamma_{i_{k-1},i_k}$.

The lower cascade $\frown\frown Q_n Q_1$ is then the curve $\gamma_{n,n-1,\ldots,2,1}$ and the curve considered in Theorem 2.1 is $\mathscr{J}_n = \gamma_{n,n-1,\ldots,2,1,n}$. Thus, what we have to prove is that $\mathscr{J}_n$ is a Jordan curve and that the compact region it determines coincides with the IC/H-diagram Δ_n.

In fact, our method of proof will lead in a natural way to a more general result. In order to formulate that, we introduce certain new concepts. First, an *index vector* $\mathbf{i}$ is a vector of the form $\mathbf{i} = (i_1, \ldots, i_\nu)$ for some $1 \leq \nu \leq n$ with $1 \leq i_1 < i_2 < \cdots < i_\nu \leq n$. The number ν is the *length* of $\mathbf{i}$ and we write

$|\mathbf{i}| = \nu$. With any index vector $\mathbf{i}$ of length at least 3 we associate the curve $\mathscr{J}_{\mathbf{i}} = \gamma_{i_\nu,\dots,i_1,i_\nu}$ and the domain $\Delta_{\mathbf{i}}$ defined as the union of $\mathscr{J}_{\mathbf{i}}$ and all bounded components of $\mathbb{R}^2 \setminus \mathscr{J}_{\mathbf{i}}$. As we shall see below, $\mathscr{J}_{\mathbf{i}}$ is a Jordan curve so that $\Delta_{\mathbf{i}}$ is in fact the compact region determined by $\mathscr{J}_{\mathbf{i}}$. For $\mathbf{i} = (1, 2, \dots, n)$ we find $\mathscr{J}_{\mathbf{i}} = \mathscr{J}_n$ and $\Delta_{\mathbf{i}} = \Delta_n$.

With an index vector $\mathbf{i} = (i_1, \dots, i_\nu)$ we also associate the subset $M_{\mathbf{i}} \subseteq M_+^1(n)$ of all distributions P which can be written as a convex combination of the form

$$P = \alpha_1 U_{i_1} + \alpha_2 U_{i_2} + \cdots + \alpha_\nu U_{i_\nu}. \tag{27}$$

As usual, this notation implies that the U_{i_j}'s are uniform distributions over i_j-element sets which are nested: $\mathrm{supp}(U_{i_1}) \subseteq \cdots \subseteq \mathrm{supp}(U_{i_\nu})$. By $M_{\mathbf{i}}'$ we denote the set of P's of this form but corresponding to the "standard choice" of the uniform distributions U_{i_j}, i.e. the choice for which $\mathrm{supp}(U_{i_j}) = \{1, 2, \dots, i_j\}$, $j = 1, \dots, \nu$. We refer to $M_{\mathbf{i}}'$ as the *standard cell* of $M_{\mathbf{i}}$. Note that by symmetry considerations, $M_{\mathbf{i}}$ can be written as the union of sets, each of which is an isomorphic copy (via permutation) of the standard cell.

The set $M_{\mathbf{i}}$ can also be characterized as the set of $P \in M_+^1(n)$ for which the point probabilities can be grouped into $\nu + 1$ classes with $i_1, i_2 - i_1, \dots, i_\nu - i_{\nu-1}$, respectively $n - i_\nu$ point probabilities in each class and such that, firstly within a given class, the point probabilities are the same, secondly, the point probabilities in one class are grater than or equal to the point probabilities in the next class and, thirdly, all point probabilities in the last class are 0. Note that all classes except possibly the last one (when $i_\nu = n$) contain at least one point probability. Note also that $M_{1,2,\dots,n} = M_+^1(n)$ and that $\mathbf{i} = (1, 2, \dots, n)$ is the only index vector for which $M_{\mathbf{i}}$ is convex.

Consider a fixed $P \in M_+^1(n)$. Then P can be written in the form (27). Moreover, if we insist that the weights all be positive, this representation of P is unique. We refer to this representation as the *representation of P as a proper mixture of nested uniform distributions* and we write $\mathbf{i}(P)$ for the associated index vector $\mathbf{i}$. Via $\mathbf{i}(P)$, $M_+^1(n)$ is split into distributions of various *types*, more specifically called *types defined by nesting*. Distributions with $|\mathbf{i}(P)| \geq 3$, i.e. distributions which, in more loose terms, are mixtures of at least three nested uniform distributions, will play a special role in the sequel as distributions that behave "regularly", cf. Lemma 5.2 below. Distributions with $|\mathbf{i}(P)| = 2$ are mixtures of just 2 nested uniform distributions and are those we used to define the curves $\gamma_{j,i}$ which are the building blocks for the planar curves that appear in our study. And distributions with $|\mathbf{i}(P)| = 1$ are the uniform distributions themselves, the key building blocks among individual distributions.

After the above introductory definitions, we embark on a sequence of lemmas which will, eventually, lead to the desired result. The first one is purely topological:

Lemma 5.1. *Let K be a compact subset of the plane $\mathbb{R}^2$ with $K^\circ \neq \emptyset$ and let $\mathscr{L}$ be a Jordan curve in $\mathbb{R}^2$. If $\partial K \subseteq \mathscr{L}$, then K is the compact region determined by $\mathscr{L}$.*

Remark. Note that the conditions on K and $\mathscr{L}$ may be expressed equivalently by demanding that $K \setminus \mathscr{L}$ be a non-empty subset of K°.

The proof below is stated for the convenience of the reader. It uses basic principles, and as such is less general but simpler than corresponding results in modern textbooks (as the already cited ones, [2] and [12]).

Proof. As both $\mathbb{R}^2 \setminus K$ and K° are non-empty open subsets of $\mathbb{R}^2 \setminus \partial K$, the boundary ∂K divides the plane in at least two components. Then, since no proper subset of $\mathscr{L}$ has this property and since $\partial K \subseteq \mathscr{L}$, $\partial K = \mathscr{L}$ must hold. Let G_0 be the bounded, and G_1 the unbounded component of $\mathbb{R}^2 \setminus \mathscr{L}$. We claim that $K^\circ = G_0$.

Note that $\mathbb{R}^2 \setminus \mathscr{L} = (\mathbb{R}^2 \setminus K) \cup K^\circ$ is a representation of $\mathbb{R}^2 \setminus \mathscr{L}$ as a disjoint union of two non-empty open sets of which the second set is bounded. Clearly then, $K^\circ = G_0$ and $K = K^\circ \cup \partial K$ must be the compact region determined by $\mathscr{L}$. □

Then we show that certain mixtures of uniform distributions behave "regularly" under $\vec{\varphi}$:

Lemma 5.2. *Distributions $P \in M_+^1(n)$ which are mixtures of at least three nested uniform distributions are mapped by $\vec{\varphi}$ into interior points of the range of $\vec{\varphi}$.*

Proof. Let $x_1, x_2, \cdots, x_n$ be the point probabilities of P. The functional matrix of the map $\vec{\varphi}$ under the condition $\sum_1^n x_i = 1$ is

$$\mathbf{F} = \left\{ \begin{array}{cccc} 2x_1 & 2x_2 & \cdots & 2x_n \\ -1-\ln x_1 & -1-\ln x_2 & \cdots & -1-\ln x_n \\ 1 & 1 & \cdots & 1 \end{array} \right\}.$$

We have to show that the rank of this matrix is 3.[1] By assumption, there are at least three distinct non-zero x's. Assume, as we may, that $0 < x_1 < x_2 < x_3$

[1]Less streamlined and more elementary, what is really involved in this proof is the restriction of $\vec{\varphi}$ to a 2-dimensional domain which—with the assumption $0 < x_1 < x_2 <$

and consider the submatrix

$$\mathbf{F}_0 = \left\{ \begin{array}{ccc} 2x_1 & 2x_2 & 2x_3 \\ -1-\ln x_1 & -1-\ln x_2 & -1-\ln x \\ 1 & 1 & 1 \end{array} \right\}.$$

Clearly, $\mathbf{F}_0$ has the same rank as the matrix

$$\left\{ \begin{array}{ccc} x_1 & x_2 & x_3 \\ \ln x_1 & \ln x_2 & \ln x_3 \\ 1 & 1 & 1 \end{array} \right\}$$

and this must be 3 because x_2 is a convex combination of x_1 and x_3 and 1 is of course the same convex combination of 1 and 1, whereas $\ln x_2$ is strictly larger than the convex combination in question of $\ln x_1$ and $\ln x_3$ as the logarithmic function is strictly concave. □

Then we turn to a closer study of mixtures of just two nested uniform distributions and their images under $\vec{\varphi}$, i.e. we turn to a study of the curves $\gamma_{j,i}$. Let i and j be fixed with $1 \leq i < j \leq n$. Without overburdening the notation, we let IC and H denote the coordinate functions, i.e., with $P_s = P_s^{j,i}$,

$$\vec{\varphi}(P_s) = (IC(s), H(s)), \quad 0 \leq s \leq 1.$$

Further, IC', IC'', IC''', H', H'' and H''' denote the derivatives (first, second and third) of the coordinate–functions w.r.t. s. For the formulas below we put

$$p' = \frac{1-s}{j} + \frac{s}{i}\,, \quad p'' = \frac{1-s}{j}, \tag{28}$$

hence also suppressing the dependency on s. A simple computation leads to

x_3—is the domain obtained by fixing the 4'th, 5'th,. . . ,n'th point probabilities to those of P and by allowing the 1'st, 2'nd and 3'rd point probabilities to vary in a neighbourhood of those of P. Eliminating one of the parameters, say the one corresponding to x_3, you obtain a map from a domain in $\mathbb{R}^2$ into $\mathbb{R}^2$. By direct computation of determinants, the 2×2 functional matrix of this map is non-singular at the point corresponding to P if and only if $\mathbf{F}_0$ below is non-singular.

the following formulas:

$$
\begin{aligned}
IC' &= 2\sum_{a=1}^{n} P_s(a)\cdot(U_i(a)-U_j(a)) = 2\big(1-\frac{i}{j}\big)(p'-p''), && (29)\\
IC'' &= 2\sum_{a=1}^{n}(U_i(a)-U_j(a))^2, && (30)\\
H' &= -\sum_{a=1}^{n}\ln P_s(a)\cdot(U_i(a)-U_j(a)) = \big(1-\frac{i}{j}\big)\ln\frac{p''}{p'}, && (31)\\
H'' &= -\sum_{a=1}^{n}\frac{1}{P_s(a)}\cdot(U_i(a)-U_j(a))^2 && (32)\\
H''' &= \sum_{a=1}^{n}\frac{1}{P_s(a)^2}\cdot(U_i(a)-U_j(a))^3\\
&= \frac{j-1}{j^3}\bigg(\big(\frac{j-i}{i}\big)^2(p')^{-2}-(p'')^{-2}\bigg). && (33)
\end{aligned}
$$

Of course, summation in these formulas could be constrained to $a\in\operatorname{supp}(U_j)$. We need the following consequences of the formulas:

Lemma 5.3. *With assumptions and notation as above, we have*

(i). $IC'\geq 0$ *with equality if and only if* $s=0$*,*

(ii). $IC''>0$*,*

(iii). $IC'''=0$*,*

(iv). $-\infty\leq H'\leq 0$ *with* $H'=-\infty$ *if and only if* $s=1$*, and with* $H'=0$ *if and only if* $s=0$*,*

(v). $-\infty\leq H''<0$ *with* $H''=-\infty$ *if and only if* $s=1$*,*

(vi). H''' *assumes negative values (for* $s=1$ *where* $H'''=-\infty$ *and in the vicinity of* $s=1$*). If* $j\leq 2i$*, in particular if* $j=i+1$*,* $H'''\leq 0$*. If* $j>2i$*,* H''' *assumes positive values (for* $s=0$ *and in the vicinity of* $s=0$*).*

Proof. (i)–(v) follow from (29)–(32) when noting that $p''=0$ for $s=1$ and that $p'\geq p''$ with equality for $s=0$. Regarding (vi) we note that the inequality $p'\geq p''$ shows that

$$H'''\leq (p'')^{-2}(j-1)(ij)^{-2}(j-2i)$$

with equality if $s=0$ or if $s=1$ (when you find $H'''=-\infty$). ☐

We depict the curves $\gamma_{j,i}$ as curves in the (IC, H)-plane with index of coincidence as abscissa and entropy as ordinate. Though $s \in [0,1]$ is chosen as parameter via the map $s \curvearrowright \vec{\varphi}(P_s)$, the curves can just as well be parametrized by the index of coincidence. This follows from Lemma 1.1 which shows that $\overline{\mathrm{MR}}_{j,i}(P_s) = s^2$ in connection with the definition of $\overline{\mathrm{MR}}_{j,i}$, cf. (5).

Recall that a plane curve is said to be *convex* if it lies at the same side of the tangent at each curve point. Furthermore, the sign of the curvature determines at which side of the tangent the curve lies.

Lemma 5.4. *Let $1 \le i < j \le n$ and consider the curve $\gamma_{j,i}$.*

(i). H is a decreasing function of IC on the curve.

(ii). The curve has a tangent at each point. Considered as a curve in the (IC, H)-plane, the tangent is vertical at the point Q_i and has the slope $-j/2$ at the point Q_j.

(iii). The curve is a convex curve if and only if $j \le 2i$, and when this condition is fulfilled, H considered as a function of IC is a concave function (for $1/j \le IC \le 1/i$).

Proof. Property (i) follows from (i) and (iv) of Lemma 5.3. Property (ii) follows from the same facts and, for $s = 0$ (when $P_s = Q_j$), from (30) and (32). In order to establish the claim of (iii), recall that the curvature of $\gamma_{j,i}$ at $\vec{\varphi}(P_s)$ is

$$((IC')^2 + (H')^2)^{-3/2} \cdot \kappa,$$

where

$$\kappa = \det\left(\frac{d\vec{\varphi}(P_s)}{ds}, \quad \frac{d^2\vec{\varphi}(P_s)}{ds^2}\right) = \begin{vmatrix} IC' & IC'' \\ H' & H'' \end{vmatrix}.$$

By Lemma 5.3 $\kappa(0) = 0$ and

$$\frac{d\kappa}{ds} = \begin{vmatrix} IC'' & IC'' \\ H'' & H'' \end{vmatrix} + \begin{vmatrix} IC' & IC''' \\ H' & H''' \end{vmatrix} = IC' \cdot H'''.$$

By Lemma 5.3, (vi), we see that if $j \le 2i$, $\kappa \le 0$ whereas, if $j > 2i$, κ assumes both positive and negative values. This proves (iii). $\square$

Lemma 5.5. *If $\mathbf{i}$ is an index vector of length 3, then $\vec{\varphi}$ restricted to the standard cell $M'_{\mathbf{i}}$ is an embedding of $M'_{\mathbf{i}}$ in $\mathbb{R}^2$.*

Proof. Assume that $\mathbf{i} = (i, j, k)$ and put $M = M'_{\mathbf{i}}$. Then M is the convex hull of the uniform distributions U_i, U_j and U_k supported by, respectively, $\{1, \dots, i\}$, $\{1, \dots, j\}$ and $\{1, \dots, k\}$. We consider M as a subset of its affine hull (homeomorphic to $\mathbb{R}^2$). In this proof, $\vec{\varphi}$ denotes the restriction of the IC/H-map to M, and we put $K = \vec{\varphi}(M)$.

By Lemma 5.2 or, rather, by a natural variant of this Lemma,

$$\vec{\varphi}(M^\circ) \subseteq K^\circ. \tag{34}$$

This follows as the functional matrix $\vec{\varphi}$ is non-sigular at points in M°. This fact also implies that locally, at each point of M°, $\vec{\varphi}$ is a homeomorphism. In other words, $\vec{\varphi}$ is a covering on M°. By (34),

$$\partial K \subseteq \vec{\varphi}(\partial M).$$

Now, $\vec{\varphi}(\partial M)$ consists of all points on one of the arcs $\gamma_{k,j}$, $\gamma_{j,i}$ and $\gamma_{k,i}$, i.e. $\vec{\varphi}(\partial M) = \mathscr{J}_{\mathbf{i}}$. We shall prove that $\mathscr{J}_{\mathbf{i}}$ is a Jordan curve. We see that $\mathscr{J}_{\mathbf{i}}$ consists of two curves $\gamma_{k,j,i}$ and $\gamma_{k,i}$. According to Lemma 5.4, (i), both curves can be considered as graphs of decreasing functions of IC for $1/k \le IC \le 1/i$, and both graphs connect Q_k with Q_i. We claim that we can connect Q_k with Q_i by a curve entirely in K° (except for the two endpoints). This follows from (34) as we can connect U_k with U_i by a path in M° (again except for the endpoints). Thus, to any value of IC with $1/k < IC < 1/i$ we can find $Q = (IC, y) \in K^\circ$. Then there must exist y_1, y_2 with $y_1 < y < y_2$ such that (IC, y_1) and (IC, y_2) both lie in ∂K. Thus (IC, y_1) must lie on one of the graphs considered and (IC, y_2) on the other. Considering the nature of the two graphs and the fact that $Q_j \in \gamma_{k,j,i}$, it now follows that $\gamma_{k,j,i}$ lies strictly below $\gamma_{k,i}$ between Q_k and Q_i. Thus $\mathscr{J}_{k,j,i}$ is a Jordan curve.

From Lemma 5.1 it now follows that $K = \Delta_{\mathbf{i}}$. This is the key fact we need in the sequel. To prove that $\vec{\varphi}$ is in fact an embedding it remains to be proved that $\vec{\varphi}$ is injective.

As $\vec{\varphi}$ is a covering on the simply connected region M° and as K° is also simply connected, the restriction of $\vec{\varphi}$ to M° is a homeomorphism.

□

With the proof of Lemma 5.5 available it is now easy to determine the relative geometric positions of the curves $\gamma_{j,i}$:

Lemma 5.6. *Let $\gamma_1 = \gamma_{j_1,i_1}$ and $\gamma_2 = \gamma_{j_2,i_2}$ be two different curves for which the indices fulfill the conditions $1 \le i_1 < j_1 \le n$ and $1 \le i_2 < j_2 \le n$ and $i_2 \le i_1$. Then:*

If γ_1 and γ_2 have one endpoint in common, this is their only intersection.

If γ_1 and γ_2 have no endpoint in common, they intersect if and only if $i_1 < j_2 < j_1$.

Proof. The cases when γ_1 and γ_2 have a common endpoint are dealt with by reasoning as in the proof of Lemma 5.5, which implies that the curves then only intersect at the common endpoint.

Assume that $i_2 < i_1 < j_2 < j_1$. Then γ_2 divides $\{(x, y) \mid y \geq -\ln x\}$ into two components and Q_{i_1} and Q_{j_1} belong to different components. Therefore, γ_1 intersects γ_2.

Then assume that $i_2 < i_1 < j_1 < j_2$. Then we can either argue along similar lines as above or we can refer to the proof of Lemma 5.5 which shows that γ_2 lies above γ_{j_1,i_2} which lies above γ_1.

Clearly, if $i_2 < j_2 < i_1 < j_1$, γ_1 and γ_2 do not intersect. □

Finally, we can prove the main result of the section:

Theorem 5.7. *Let $\mathbf{i}$ be an index vector. Then $\mathscr{J}_{\mathbf{i}}$ is a positively oriented Jordan curve and $\vec{\varphi}(M_{\mathbf{i}}) = \Delta_{\mathbf{i}}$, the compact region determined by $\mathscr{J}_{\mathbf{i}}$.*

Proof. By Lemma 5.6, $\mathscr{J}_{\mathbf{i}}$ is a Jordan curve. Put $K = \vec{\varphi}(M_{\mathbf{i}})$ and let M denote the standard cell $M'_{\mathbf{i}}$. By symmetry, $K = \vec{\varphi}(M)$. Let $\mathbf{i} = (i_1, \dots, i_\nu)$. Then, by Lemma 5.6,

$$\Delta_{\mathbf{i}} = \Delta_{i_1,i_2,i_3} \cup \Delta_{i_1,i_3,i_4} \cup \cdots \cup \Delta_{i_1,i_{\nu-1},i_\nu}$$

and combining this with Lemma 5.5 it follows that $K \supseteq \Delta_{\mathbf{i}}$. By Lemma 5.2, points in M which are mixtures of three or more of the dirstributions $U_{i_1}, U_{i_2}, \dots, U_{i_\nu}$ (with U_{i_j} the uniform distribution over $\{1, \dots, i_j\}$, $j = 1, \dots, \nu$), are mapped by $\vec{\varphi}$ into K°. Therefore,

$$\partial K \subseteq \bigcup_{j<k} \gamma_{i_{k,j}}.$$

However, as $K \supseteq \Delta_{\mathbf{i}}$, Lemma 5.6 shows that most of the sets here are contained in K° (except for their endpoints). More precisely, it follows by these considerations that $\partial K \subseteq \mathscr{J}_{\mathbf{i}}$. Then, by Lemma 5.1, $K = \Delta_{\mathbf{i}}$. □

6 Upper bounds of power-type for the entropy function

In this section we study distributions $P \in M_+^1(n)$, and derive upper bounds for the entropy $H(P)$ given in terms of the index of coincidence $IC(P)$. The theoretically best upper bound was already given in Corollary 2.2 which shows that

$$H(P) \le H\left((1-x)U_n + xU_1\right) \tag{35}$$

if x is determined so that P and $(1-x)U_n + xU_1$ have the same index of coincidence.

This bound is perhaps somewhat awkward to handle. In any case, we seek analytically simpler bounds. It appears most natural to express such bounds in terms of the relative measure of roughness, $\overline{\mathrm{MR}}^n(P)$, rather than in terms of the index of coincidence itself.

By τ_n we denote the largest constant such that the inequality of "power-type",

$$H(P) \le \ln n \cdot \left(1 - \overline{\mathrm{MR}}^n(P)\right)^{\tau_n}, \tag{36}$$

holds for all $P \in M_+^1(n)$. We refer to the τ_n's as the *maximal exponents*.

It appears impossible, except for $n = 2$, to express τ_n in closed form. It is natural to expect that $\tau_n \le 1$ for all n. This fact follows from Lemmas 6.1 and 6.2 below, but more precise bounds will follow. We consider it a key non-trivial fact that the maximal exponents converge to 1 as $n \to \infty$, cf. Theorem 6.4.

In principle, the study of the maximal exponents is elementary and depends on a closer inspection of the functions h_n, $n \ge 2$, defined by

$$h_n(x) = H((1-x)U_n + xU_1), \quad 0 \le x \le 1\,. \tag{37}$$

We find it convenient to introduce the abbreviation

$$n' = n - 1. \tag{38}$$

For the function h_n, and its first two derivatives, which we shall need later on, we find the expressions

$$h_n(x) = \frac{n'(1-x)}{n} \ln \frac{n}{1-x} + \frac{1+n'x}{n} \ln \frac{n}{1+n'x} \tag{39}$$

$$= \ln n - \frac{n'}{n}(1-x)\ln(1-x) - \frac{1}{n}(1+n'x)\ln(1+n'x)\,, \tag{40}$$

$$h_n'(x) = -\frac{n'}{n} \ln \frac{1+n'x}{1-x}\,, \tag{41}$$

$$h_n''(x) = -\frac{n'}{(1-x)(1+n'x)}\,. \tag{42}$$

Introduce also the function s by

$$s(x) = 1 - x^2, \quad 0 \le x \le 1. \tag{43}$$

By Lemma 1.1 and by (35) we see that the validity of the power inequality (36) for all $P \in M_+^1(n)$ is equivalent with

$$\frac{h_n(x)}{\ln n} \le s(x)^{\tau_n}$$

for $0 < x < 1$. Hence, defining the function g_n by

$$g_n(x) = \frac{\ln(\frac{h_n(x)}{\ln n})}{\ln s(x)}, \quad 0 < x < 1, \tag{44}$$

we find the following expression for the τ_n's:

Lemma 6.1. *For each $n \ge 2$, the maximal exponent τ_n is given in terms of g_n by*

$$\tau_n = \inf_{0<x<1} g_n(x).$$

At the more technical level we start by collecting basic facts about the individual functions g_n:

Lemma 6.2. (i) *The limit values of g_n at the endpoints 0 and 1 are:*

$$g_n(0^+) = \frac{n'}{2\ln n}\,, \quad g_n(1^-) = 1. \tag{45}$$

(ii) *For $n \ge 2$, $g_n(x) < 1$ for all x sufficiently close to 1 and for $n \ge 3$, $g_n(x) < g_n(0^+)$ for all x sufficiently close to 0.*

(iii) *For $n \geq 3$, there exists x_0 such that $0 < x_0 < 1$ and $g_n(x_0) = \tau_n$.*

Proof. (i): The routine proof, using proper approximations and/or l'Hospitals rule, is left to the reader.

(ii): First note that by (39),

$$h_n(x) \geq \frac{n'(1-x)}{n} \ln \frac{n}{1-x}. \tag{46}$$

Then, for x sufficiently close to 1, it follows that $h_n(x) > s(x) \ln n$, hence $g_n(x) < 1$ for such values of x.

Regarding the behaviour near the other endpoint, one may again appeal to a (tedious!) application of l'Hospitals rule. One finds that

$$\lim_{x \to 0^+} \frac{g_n(x) - g_n(0^+)}{x} = -\frac{(n-1)(n-2)}{6 \ln n}.$$

Thus, for $n \geq 3$, $g_n(x) < g_n(0+)$ for x sufficiently close to 0.

(iii): This follows from (ii). □

According to (i) above, we can now define the g_n's by continuity on all of $[0, 1]$.

Then we study the behaviour of the g_n's as n varies:

Lemma 6.3. (i) *For $0 < x < 1$, $g_2(x) < g_3(x) < g_4(x) < \cdots$.*

(ii) *For $0 < x \leq 1$ $\lim_{n\to\infty} g_n(x) = g_\infty$ exists and is given by*

$$g_\infty(x) = \frac{\ln(1-x)}{\ln(1-x^2)}, \quad 0 < x \leq 1 \tag{47}$$

(with $g_\infty(1) = 1$).

Proof. (i): Fix $n \geq 2$ and consider the auxiliary function

$$f(x) = \frac{h_n(x)}{\ln n} - \frac{h_{n+1}(x)}{\ln(n+1)}, \quad 0 \leq x \leq 1.$$

Then $f(0) = f(1) = 0$ and, by (41), $f'(0) = 0$ and $f'(1) = -\infty$ (note that $n/((n+1)\ln(n+1)) - n'/(n \ln n) < 0$). By (42), we see that

$$f''(x) = \frac{1}{1-x}\left(-\frac{n'}{\ln n(1+n'x)} + \frac{n}{\ln(n+1)(1+nx)} \right).$$

This formula implies, firstly, that $f''(0) = n/\ln(n+1) - n'/\ln n > 0$ and, secondly, that f only has one inflection point in $]0,1[$. Collectiong these facts we see that $f(x) > 0$ for $x \in]0,1[$ (since otherwise, f would have at least 3 inflection points in $]0,1[$). As $f(x) > 0$ in $]0,1[$, $g_n(x) < g_{n+1}(x)$ in $]0,1[$ follows.

(ii): By (i) we may define g_∞ on $]0,1]$ as the pointwise limit of the g_n's. Clearly, $g_\infty(1) = 1$. By (40) it follows that for $0 < x < 1$,

$$\lim_{n\to\infty} \frac{h_n(x)}{\ln n} = 1 - x,$$

hence g_∞ is given by (47). □

Theorem 6.4.

$$\frac{1}{\ln 4} = \tau_2 < \tau_3 < \cdots$$

and $\lim_{n\to\infty} \tau_n = 1$.

Proof. In view of Lemma 6.3, (i), the inequalities $\tau_2 < \tau_3 < \cdots$ follow readily. The determination of τ_2 can be found in Topsøe [31].

Clearly, g_∞ is decreasing in $]0,1]$ and thus assumes its minimal value 1 for $x = 1$. Choose n_0 such that $g_{n_0}(\frac{1}{2}) > 1$. Then, for $n \geq n_0$, g_n assumes its minimal value (τ_n) in $[\frac{1}{2}, 1]$. As g_n converges uniformly to g_∞ in $[\frac{1}{2}, 1]$, the minima of the g_n's converge to the minimum of g_∞. This is the assertion of the Theorem. □

At this point we find it convenient to collect some information about the g_n's in the form of a table (Table 1). Except for the two last columns, the information given is obtained by numerical- and graphical experimentation. For each of the investigated values of n, we have found experimentally that there is a unique minimum point of g_n, and the argmin-value as well as the function value is quoted. The bounds given in the last two columns are taken from Lemmas 6.5 and 6.8 below (a more precise lower bound is indicated in the discussion).

n	argmin	τ_n	upper bound	lower bound
2	0.0000	0.7213	–	0.4094
3	0.8625	0.7991	–	0.5231
4	0.9442	0.8292	0.8327	0.5809
5	0.9704	0.8473	0.8559	0.6168
6	0.9821	0.8597	0.8705	0.6418
7	0.9882	0.8690	0.8808	0.6605
8	0.9917	0.8762	0.8885	0.6753
9	0.9940	0.8820	0.8944	0.6872
10	0.9955	0.8868	0.8993	0.6972
11	0.9965	0.8909	0.9033	0.7057
12	0.9972	0.8944	0.9067	0.7131
13	0.9978	0.8975	0.9096	0.7195
14	0.9982	0.9002	0.9121	0.7252
15	0.9985	0.9026	0.9143	0.7303
16	0.9987	0.9048	0.9163	0.7349
17	0.9989	0.9068	0.9181	0.7391
18	0.9991	0.9086	0.9197	0.7430
19	0.9992	0.9102	0.9212	0.7465
20	0.9993	0.9117	0.9226	0.7497
50	$1 - 4.8 \cdot 10^{-5}$	0.9329	0.9407	0.7964
100	$1 - 4.7 \cdot 10^{-6}$	0.9436	0.9496	0.8216
1000	$1 - 3.1 \cdot 10^{-9}$	0.9636	0.9664	0.8732
10000	$1 - 1.5 \cdot 10^{-12}$	0.9733	0.9748	0.9021

Table 1

Lemma 6.5. *Let c be the constant determined uniquely by the two requirements*

$$0 < c < 1, \quad c = 1 + \ln(2c) \tag{48}$$

($c \approx 0.231961$). Then, for all $n \geq 4$,

$$\tau_n \leq 1 - \frac{c}{\ln n} \tag{49}$$

Proof. The validity of (49) for $n = 4, 5, \dots, 18$ is established on a case-by-case basis. This can be done quite precisely via Table 1, even without having to rely on the exactness of this table. Indeed, for each of the values of n in question, one chooses a value of x close to the stated value of argmin and checks, numerically, that $g_n(x) \le 1 - \frac{c}{\ln n}$.

We shall now establish (49) for $n \ge 19$. Put

$$a = \arg\max_{x>0} \frac{\ln\left(\frac{1+x}{2}\right)}{x} \tag{50}$$

($a \approx 3.31107$). As the function appearing here is strictly concave, a is uniquely determined by the vanishing of the derivative, i.e. by the equation

$$\frac{1}{1+a} = 1 + \ln\left(\frac{2}{1+a}\right).$$

It follows that $1/(1+a) = c$ and then, that

$$\frac{\ln\left(\frac{1+a}{2}\right)}{a} = c. \tag{51}$$

By Lemma 6.1,

$$\tau_n \le g_n(1 - n^{-a}) \tag{52}$$

and it only remains to estimate this function value.

By (46) and by rewriting g_n in the form

$$g_n(x) = 1 - \frac{\ln\left(\frac{h_n(x)}{s(x)\ln n}\right)}{\ln\left(\frac{1}{s(x)}\right)},$$

we find that

$$g_n(1 - n^{-a}) \le 1 - \frac{\ln\left(\frac{n'(a+1)}{n(2-n^{-a})}\right)}{\ln\left(\frac{n^a}{2-n^{-a}}\right)}.$$

Thus

$$\begin{aligned} 1 - \tau_n \quad &\ge \quad \frac{\ln\left(\frac{n'(a+1)}{2n}\right)}{\ln\left(\frac{n^a}{2-n^{-a}}\right)} = \frac{\ln\left(\frac{n'}{n}\right) + ac}{a\ln n - \ln(2-n^{-a})} \\ &= \quad \frac{c}{\ln n} + \frac{c\ln(2-n^{-a}) - \ln n \cdot \ln\left(1 + \frac{1}{n'}\right)}{(a\ln n - \ln(2-n^{-a}))\ln n}. \end{aligned}$$

Recalling the (approximate) values of c and a, it is easy to check numerically that for $n \geq 19$,

$$\ln n \cdot \ln(1 + 1/n) \leq c \ln(2 - n^{-a})$$

holds. We then see that for these values of n, $1 - \tau_n \geq c/\ln n$, i.e. (49) holds. $\square$

In order to derive specific lower bounds for the τ_n's, a closer study of the g_n's is necessary. This presents certain problems as these functions are unstable close to the endpoint 1. This is illustrated in Figure 5. Table 1 is also helpful in revealing the nature of the g_n's.

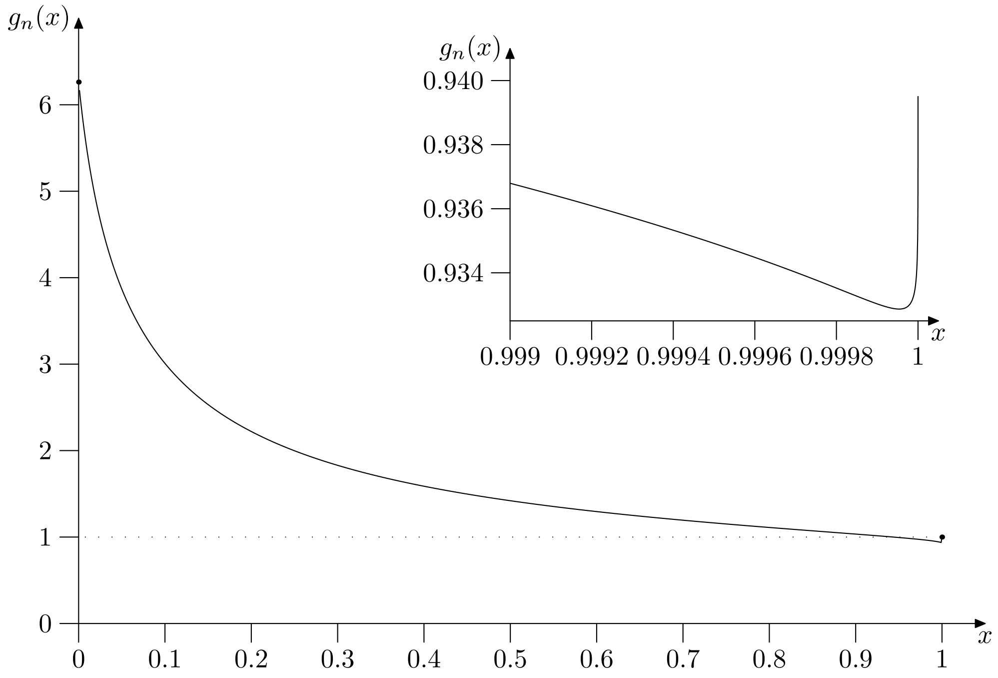

Fig.5. Illustration of the special behaviour of g_n ($n = 50$).

For the further study, we consider the derived function g_n'. One finds that

$$g_n'(x) = \frac{s'(x)}{s(x)\ln s(x)}(\zeta_n(x) - g_n(x)), \tag{53}$$

where ζ_n denotes the auxiliary function

$$\zeta_n(x) = \frac{(\ln h_n)'(x)}{(\ln s)'(x)} = \frac{h_n'(x)s(x)}{h_n(x)s'(x)}, \quad 0 < x < 1. \tag{54}$$

Now, let us aim at deriving what we consider to be the most interesting results, viz. lower bounds for the maximal exponents. It appears difficult to do so by direct inspection of the functions g_n. Instead, the auxiliary functions ζ_n come into play.

Lemma 6.6. *For all $n \geq 3$,*

$$\tau_n \geq \inf_{0<x<1} \zeta_n(x).$$

Proof. By Lemma 6.1 and Lemma 6.2 (iii), there exists $0 < x_0 < 1$ with $g_n(x_0) = \tau_n$ and $g_n'(x_0) = 0$. From (53) we conclude that $\zeta_n(x_0) = \tau_n$, hence $\tau_n \geq \inf \zeta_n$. □

A direct calculation, cf. (40) and (41), shows that

$$\zeta_n(x) = \frac{\ln(1+n'x) + \ln\left(\frac{1}{1-x}\right)}{\ln n + \ln\left(\frac{1}{1-x}\right) + \frac{1+n'x}{n'(1-x)}\ln\left(\frac{n}{1+n'x}\right)} \cdot \frac{1+x}{2x}. \tag{55}$$

Using the inequality $\ln(a) \leq a - 1$ in the denominator, we find that

$$\zeta_n(x) \geq \frac{\ln(1+n'x) + \ln\left(\frac{1}{1-x}\right)}{\ln n + \ln\left(\frac{1}{1-x}\right) + 1} \cdot \frac{1+x}{2x}. \tag{56}$$

At this point we need a simple lemma:

Lemma 6.7. *For $n \geq 5$ and all $0 \leq x \leq 1$,*

$$\frac{\ln(1+n'x)}{\ln n} \geq \frac{2x}{1+x}. \tag{57}$$

Proof. Put

$$\varphi(x) = (1+x)\ln(1+n'x) - 2x\ln n.$$

We have to prove that $\varphi \geq 0$. One finds that $\varphi(0) = \varphi(1) = 0$ and that $\varphi'(0) = n' - 2\ln n > 0$ (as $n \geq 4$) as well as $\varphi'(1) = 2n'/n - \ln n < 0$ (as $n \geq 5$) hold. Investigating φ'', it becomes evident that φ only has one inflection point in $]0,1[$. We conclude that $\varphi \geq 0$. □

It is now easy to establish a lower bound for τ_n which gives more substance to the convergence established in Theorem 6.4.

Lemma 6.8. *For $n \geq 2$,*

$$\tau_n \geq 1 - \frac{1}{\ln n + 1}. \tag{58}$$

Proof. From (56) and (57) we get, for any $0 < x < 1$,

$$\begin{aligned}
\zeta_n(x) &\geq \frac{\ln n + \frac{1+x}{2x}\ln\left(\frac{1}{1-x}\right)}{\ln\left(\frac{n}{1-x}\right)+1} \geq \frac{\ln\left(\frac{n}{1-x}\right)}{\ln\left(\frac{n}{1-x}\right)+1} \\
&= 1 - \frac{1}{\ln n + 1 + \ln\left(\frac{1}{1-x}\right)} \geq 1 - \frac{1}{\ln n + 1}.
\end{aligned}$$

As this bound is independent of x, the result now follows from Lemma 6.6. □

Collecting facts, we can now summarize the main results of this section:

Theorem 6.9. *The maximal exponents $(\tau_n)_{n\geq 2}$ in the power-type inequality (36) are increasing with limit 1 and bounded below by $\tau_2 = (\ln 4)^{-1}$. Furthermore, for $n \geq 2$, the inequalities*

$$\frac{c}{\ln n} \leq 1 - \tau_n \leq \frac{1}{\ln n + 1} \tag{59}$$

hold where $c \approx 0.2320$ is defined by (48).

7 Discussion

Relation to planned further work

The present paper belongs to a series of three papers (the other two are [14] and [15]). The overall goal is to consolidate and further develop a game theoretical viewpoint underlying certain basic parts of information theory for which optimization plays an important role. A starting point is Topsøe [28]. Several other authors have also stressed the importance of the game theoretical view, cf. for example Haussler [16] and Xie and Barron [34]. In [15] we will collect the main theoretical results, [14] will contain specific results of universal coding and prediction and in the present paper we develop special techniques which are needed in [14] but which appear to be more general, so that they can be presented – as is done here – without special reference to the problems studied in [14] and in [15].

Extensions to the study of other diagrams, work of other authors

Our main result, Theorem 2.1 can be extended in various directions. First we note that the actual result proved in Section 5, Theorem 5.7, is more general than Theorem 2.1. We may for instance use it to study a restriction of the IC/H-diagram to distributions with a lower bound of the form $\frac{1}{m}$ for integer m on the individual point probabilities. However, other extensions are more interesting. First we point to extensions to a study of certain IC/H_f-diagrams where

$$H_f(P) = \sum_i f(p_i).$$

In order to carry out an analysis like the one in Section 5, the essential condition is that f''' be positive.

Another direction of generalization is to consider Rényi entropies in place of the ordinary entropy. Recall the definition of the *Rényi entropy* H_α *of order* $\alpha > 0$:

$$H_\alpha(P) = \frac{1}{1-\alpha} \ln \left(\sum_i p_i^\alpha \right).$$

The limit as $\alpha \to 1$ is the usual entropy, and the limit as $\alpha \to \infty$ is $-\ln P_{\max}$, where $P_{\max}$ denotes the maximal point probability. Recall that the diagram in Figure 3 is a comparison of two Rényi entropies (of orders 2 and 1, respectively). It is natural to ask what happens if Rényi entropies of other orders are compared. If $0 < \nu < \mu < \infty$ the arguments in Section 5 still apply and you obtain a diagram very similar to the one in Figure 3 with smooth curves connecting the points of the form $(\ln k, \ln k), k = 0, 1, \cdots, n$. If we let μ go to infinity while ν is kept fixed, we get in the limit a diagram comparing Rényi entropies of order ∞ and ν. Now, $1 - P_{\max}$ can be identified with the *probability of error*, and this explains why diagrams similar to Figures 1, 2 and 3 appear when entropy and probability of error are compared.

The first to study in detail the P_e/H-diagram was Kovalevskij [18] in 1965. Further research, partly extending Kovalevskij's results, partly rediscovering them, includes papers by Tebbe and Dwyer [26], Ben-Bassat [1] and Feder and Merhav [7].

The study by Sayir [24] shows how diagrams with a shape as those considered here and in the previous literature come up in an experimental study.

A recent independent study by György and Linder [13] deals with problems of quantization and rate distortion and in this connection they also discovered the lower cascade related to the IC_α/H-diagrams. Finally, the paper [23] by Santis, Gaggia and Vaccaro may be the last published research in this direction.

Let us also briefly discuss an extension of the index of coincidence to *indices of order* α defined by

$$IC_\alpha(P) = \sum_i p_i^\alpha.$$

The quantities $1-IC_\alpha$ behave like an entropy in some respects and have been introduced, apart from a proportionality factor, in Havrda and Charvát [17]. However, right now, the connection with Rényi entropies:

$$H_\alpha(P) = \frac{1}{1-\alpha} \ln IC_\alpha(P)$$

is more important.

When a diagram comparing Rényi entropies is transformed into a diagram like figure 1 or 2 based on IC_α by the non-linear transformation above, the convexity of the bounding curves is generally not conserved. The convexity of bounding curves as well as the determination of the extremal points of the range for generalized diagrams is an important subject which we hope to return to.

Universality of constants in Theorem 2.6

In connection with Theorem 2.6, we noted in Section 2, that the constants in (19) are the same as those that have appeared in previous research when studying P_e/H-diagrams. To understand this, consider for any $\alpha > 1$ the (e^{-H_α}/H)-diagram and note that the dividing points will be $(\frac{1}{k}, \ln k)$ for all α. Also note that that for $\alpha = 2$, we simply obtain the IC/H-diagram. As $e^{-H_\alpha} \to P_{\max} = 1 - P_e$ when $\alpha \to \infty$, these facts explain the phenomenon regarding coincidence of constants.

The significance of IC

Many of the theorems presented in this paper are also valid for $\alpha \neq 2$, but by focusing on the index of coincidence of order 2, the technical problems are kept at a minimum while the main ideas are carried through. It appears that there are mainly three reasons why the case $\alpha = 2$ is simple. Firstly, Lemma 1.1 is a computationally convenient structural property which does not carry over to the indices IC_α of arbitrary orders. Secondly, the vanishing of the third derivative in Lemma 5.3, (iii) is of course a special property for the case $\alpha = 2$ and, lastly, we point to the connection between Rényi entropies and the indices of order α which is particularly simple for $\alpha = 2$.

On the topological method

The combination of well-known qualitative methods from topology and simple specific considerations related to the specific quantities under consideration (here index of coincidence and entropy) appears to have a potential which could take one quite far beyond the present research as indicated also above. The method will enable one to establish a basic result on the interrelationship between two quantities of interest and from there on one may develop more specific inequalities as the need may be. The paper illustrates this point: Theorem 2.1 as the basic result and Theorems 2.6 and 2.8 as specific inequalities that follow (note that when Theorem 2.6 is derived directly from Theorem 2.1, the only extra observations you need is Lemma 5.4, (iii) and the fact that $y = -\ln x$ is a lower bounding curve for the IC/H-diagram).

Information Diagrams

As is seen from previous research pointed to and from research about to be published, "diagrams" as those discussed here play a significant role for several areas. Right now the application areas we can point to are the following ones: Shannon theory, prediction and universal coding, rate distortion analysis and statistical decision theory with a common denominator for the two last mentioned areas being error probability analysis. Justified by this rather wide range of applicability, we suggest that one uses the term *Information Diagram* for these objects. A precise definition is not sensible. What we have in mind is situations where quantities of significance in information theory – two or more such quantities – are studied with the aim of obtaining information about the global relationship between the quantities – information which goes beyond partial information as obtained, e.g. by asymptotic or local approximations.

Complexity classes

For an information diagram as here considered, the non-smooth bounding curve does vary smoothly when restricted to certain intervals determined by IC, P_e or what the case may be. These "bounding intervals" (terminology of Kovalevskij [18]) also determine a decomposition of the set $M_+^1(n)$ (or $M_+^1(\mathbb{N})$) and thus gives rise to various "complexity classes", as those we introduced in connection with our study of the IC/H-diagram. A further study of these for various information diagrams should be interesting.

Wider perspectives

Wider perspectives open up when we speculate over higher-dimensional versions of the topological method. Then, what will be involved, is the study of interrelationships between more than two quantities. The extension of the topological method to higher dimensions could also be essential as then the special method developed in Section 3 may not be available in a suitably generalized form.

Taking these wider perspectives into account, it may be reasonable to use the term "bigram" for the main case when only two quantities are compared, e.g. we may speak about the IC/H-bigram, the H_2/H-bigram etc.

Possible extensions of Theorem 2.6

The original aim of the paper was to establish the inequalities of Theorem 2.6 which are essential for the solution of certain problems of exact universal coding and prediction in Bernoulli sources, cf. Harremoës and Topsøe [14]. The tools developed are, however, rather general and may lead to a number of other inequalities. Especially, the tools appear to be suitable for the investigation of inequalities between divergence measures. For this, the lemma of replacement, Theorem 3.1, is quite sufficient. However, in order that the arguments run smoothly, it is most natural to extend the reasoning behind the lemma of replacement so that it applies to a generalization of f-divergences, cf. Csiszàr [4], to cases with a function f of mixed type (concave/convex or more general). We hope to return to this in a subsequent publication (announced proofs of certain results from Topsøe [30] which were planned for this paper have thus been postponed).

The Lemma of Replacement

When we inspect the proof of Theorem 3.1, we realize that the "replacement distributions" P_0 and P_1 can in fact be given directly and quite simply in terms of P and the inflection point ξ. We need some preparations: To any $P \in M_+^1(\mathbb{N})$ and any $0 < \xi < 1$ we associate two numbers defined by:

$$\#(P:\xi) = \#\{i \mid p_i \geq \xi\} \tag{60}$$

and

$$\sigma(P:\xi) = \sum\{p_i \mid p_i \geq \xi\} \tag{61}$$

(with # denoting "number of elements such that"). These numbers may be called "ξ–significance numbers" of P.

Theorem 7.1 (Lemma of replacement, specific form). *Let f be a concave/convex function with inflection point ξ and F the associated map defined on $M_+^1(\mathbb{N})$. Let $P \in M_+^1(\mathbb{N})$ and put $\nu = \#(P:\xi)$, $\sigma = \sigma(P:\xi)$ and $\sigma' = 1-\sigma$.*

(i). *Determine the integer* l *such that* $\sigma' \in [l\xi, (l+1)\xi[$ *and put*

$$k = \nu + l, \quad r = \sigma' - l\xi, \tag{62}$$
$$P_0 = (k+1)rU_{k+1} + (1-(k+1)r)U_k. \tag{63}$$

Then $P_0 \in co\{U_{k+1}, U_k\}$ *and* $F(P) \geq F(P_0)$.

(ii). *Let* $n \geq 2$ *and assume that* $P \in M_+^1(n)$. *Define* $s \geq \xi$ *and* P_1 *by*

$$s = \sigma - (\nu - 1)\xi, \tag{64}$$
$$P_1 = \tfrac{1-s}{1-\frac{1}{n}}U_n + \tfrac{s-\frac{1}{n}}{1-\frac{1}{n}}U_1. \tag{65}$$

Then $P_1 \in co\{U_n, U_1\}$ *and* $F(P) \leq F(P_1)$.

The power-type inequalities

As follows from Section 6, the power-type inequalities developed there require a subtle analysis of in principle elementary functions. This was illustrated in Figure 5 and in Table 1 and may be further illuminated by considering Figure 6 which considers the difference f_n between the right-hand-side and the left-hand-side involved in the power inequality, i.e.

$$f_n(x) = \ln n(1-x^2)^{\tau_n} - h_n(x), \quad x \in [0,1],$$

cf. (36) and (37).

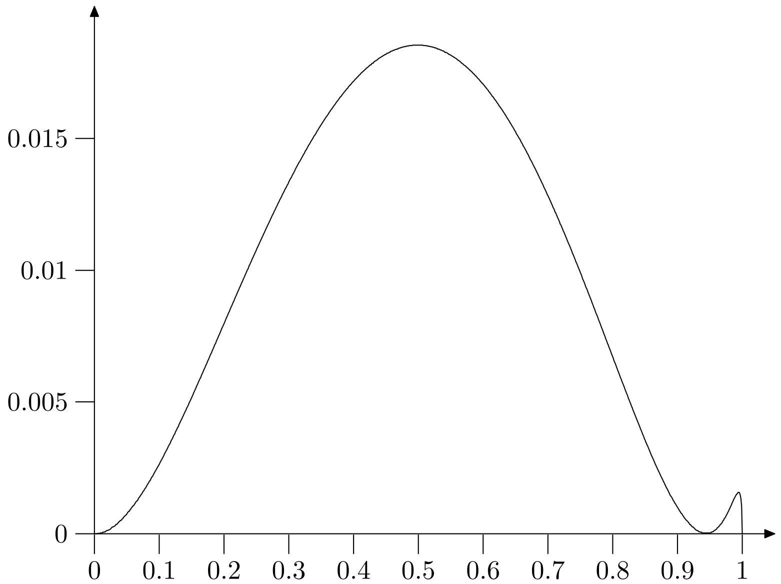

Fig.6. Illustration of the special behaviour of $f_n (n = 4)$.

Without going into the details we mention that by an analysis of the functions f_n, appealing to standard theory of Csiszár divergences, it is easy to see that the maximal exponents are all $\geq \frac{2}{3}$ (this value of the exponent corresponds to a much simpler looking function f_n). It is noted that even the inequality $\tau_n \geq \frac{1}{\ln 4}$ is stronger than the result you get when using the exponent $\frac{2}{3}$.

As far as the authors know, no previous instances of the power-type inequality have occured before for general n. However, Lin [19] proved a partial result which amounts to the inequality $\tau_2 \geq \frac{1}{2}$.

If one plots the upper bound in the power-inequality for $n = 2$ against the entropy function, one will not be able to tell the difference between the bound and the entropy function. Therefore, it is sensible to plot instead a quotient. First let us denote the generic distribution in $M_+^1(2)$ by $P = (p, q)$. Then $H(p,q) = -p\ln p - q\ln q$ and $1 - \overline{\mathrm{MR}}^2(p,q) = 4pq$, so that the upper bound obtained from Theorem 6.9 is

$$H(p,q) \leq \ln 2(4pq)^{\frac{1}{\ln 4}}. \tag{66}$$

In Figure 7 we have plotted the quotient between the upper bound in (66) and $H(p,q)$ for $p = (1+x)/2, q = (1-x)/2, x \in [0,1]$. The dashed curve is the similar plot but corresponding to Lin's upper bound $H(p,q) \leq \ln 2\sqrt{4pq}$.

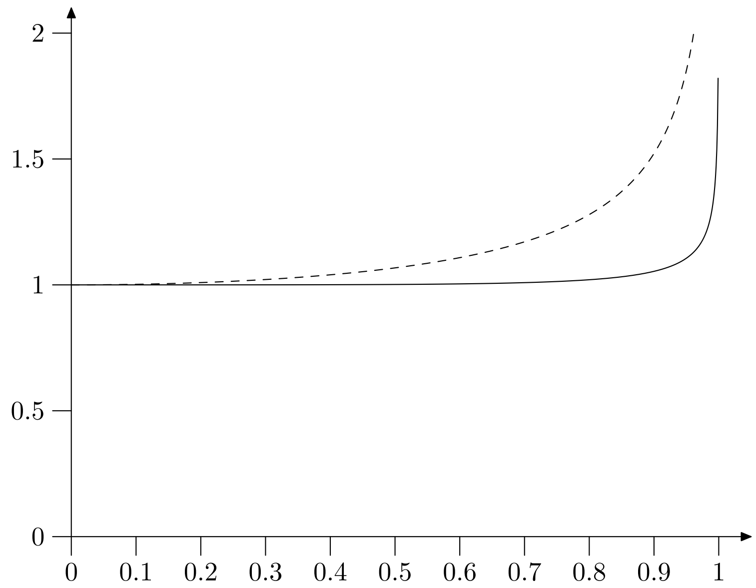

Fig.7. Ratio between power-bound and entropy for $n = 2$.

Finally, we conjecture that the upper bound of Lemma 6.5 is in fact precise in the sense that the asymptotic result

$$\lim_{n\to\infty}(1-\tau_n)\ln n = c, \tag{67}$$

holds with c the constant from Lemma 6.5. In order to support this conjecture, we note that there are various possibilities for obtaining sharper lower bounds than the bound in Lemma 6.8. For instance, one may apply Lemma 6.7 directly to (55). One then finds that

$$\zeta_n \geq \xi_n,$$

where ξ_n is defined by

$$\xi_n = \frac{\ln n + \ln \frac{1}{1-x}}{\ln n + \ln \frac{1}{1-x} + \frac{1+n'x}{n'(1-x)} \ln \frac{n}{1+n'x}}, \tag{68}$$

and here the last term in the denominator may be bounded more sharply than in the main text. Without going into details we mention that it is easy to obtain in this way a better theoretical lower bound. For instance, the lower bounds in Table 1 corresponding to the values $n = 10, 20, 100$ and 1000 may be replaced by the sharper bounds $0.8022, 0.8329, 0.8739$ and 0.9051, respectively.

Acknowledgements

In 1997-98 it was discovered that the inequalities (19) are essential in order to solve certain problems of exact prediction and universal coding. However, only the simplest of the inequalities (corresponding to the case $k = 1$) was established (by induction). The remaining inequalities were discussed with Boris Ryabko and strong evidence for their validity established, though full proofs still remained. The fruitfull discussions with Ryabko, who also followed the further work, are gratefully acknowledged.

The authors have also had fruitfull discussions with Rasmus Borup Hansen (student of F. Topsøe) who demonstrated the then surprising shape of the $IC/H-$ diagrams via numerical experiments. He also assisted with the technical production of the figures. For the results on upper bounds of the entropy function, we acknowledge useful input from Jan Caesar (student of F. Topsøe).

Finally, we acknowledge last minute input obtained from Igor Vajda and from András György when a preliminary version of the manuscript was discussed at the workshop *Information Theory in Mathematics*, Balatonlelle, Hungary, July 2000. In particular, Igor Vajda provided several important references and comments regarding the earlier literature and András György commented on his then forthcoming joint paper with Linder, [13].

References

[1] M. Ben–Bassat, "f-Entropies, Probability of Error, and Feature Selection," *Information and Control*, vol. 39, pp. 227–242, 1978.

[2] G.E. Bredon, *Topology and Geometry*, New York. Springer, 1993.

[3] T.M. Cover and J.A. Thomas, *Information Theory*, New York. Wiley, 1991.

[4] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[6] Z. Daróczy,"Generalized information functions," *Inform. Contr.*, vol. 16, pp. 36–51, 1970.

[7] M. Feder and N. Merhav,"Relations Between Entropy and Error Probability," IEEE Trans. Inform. Theory, vol. 40, pp. 259–266, 1994.

[8] A.A. Fedotov, P. Harremoës and F. Topsøe, "Refinements of Pinsker's Inequality," submitted.

[9] W.F. Friedman, *The Index of Coincidence and its Applications in Cryptoanalysis.* Laguna Hills: Aegean Park Press, 1987.

[10] R.G. Gallager, *Information Theory and reliable Communication.* New York: Wiley, 1968.

[11] R.M. Gray and D.L. Neuhoff,"Quantization," IEEE Trans. Inform. Theory, vol. 44, pp. 2325–2383, 1998.

[12] M.J. Greenberg and J.R. Harber, *Algebraic Topology,* Redwood City, CA: Addison-Wesley, 1981.

[13] A. György and T. Linder, "Optimal Entropy-Constrained Scalar Quantization of a Uniform Source," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2704–2711, 2000.

[14] P. Harremoës and F. Topsøe, "Instances of Exact Prediction and Universal Coding," in preparation.

[15] P. Harremoës and F. Topsøe, "Universal Coding via Games," in preparation.

[16] D. Haussler, "A general minimax result for relative entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, 1997.

[17] J. Havrda and F. Charvát, "The concept of structural α-entropy," *Kybernetika*, vol. 3, pp. 30–35, 1967.

[18] V.A. Kovalevskij, "The problem of character recognition from the point of view of mathematical statistics," in *Character Readers and Pattern Recognition* (eds. V.A. Kovalevskij), pp. 3–30. New York: Spartan, 1967. Russian edition 1965.

[19] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inform. Theory*, vol. 37, pp. 145–151, 1991.

[20] A.J. Menezes, P. van Oorschot and S. Vanstone, *Handbook of applied cryptography.* Boca Raton, Florida: CRC Press, 1997.

[21] M.S. Pinsker, *Information and Information Stability of Random Variables and Processes.* San-Francisco, CA: Holden-Day, 1964. Russion original 1960.

[22] A. Rényi, " On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, Berkeley, Univ. of Calif. Press, vol 1, pp. 547–561, 1961.

[23] A.De Santis, A.G. Gaggia and U. Vaccaro, "Bounds on Entropy in a Guessing Game," IEEE Trans. Inform. Theory, vol. 47, pp. 468–473, 2001.

[24] J. Sayir, *On Coding by Probability Transformation.* Konstantz: Hartung-Gorre, 1999.

[25] D.R. Stinson, *Cryptography: Theory and Practice.* Boca Raton, Florida: CRC Press, 1995.

[26] D.L. Tebbe and S.J. Dwyer, "Uncertainty and the Probability of Error," IEEE Trans. Inform. Theory, vol. 14, pp. 516–518, 1968.

[27] C. Thomassen, "The Jordan-Schönflies Theorem and the Classification of Surfaces," *Am. Math. Mon.*, vol. 99, pp. 116–130, 1992.

[28] F. Topsøe, "Information theoretical Optimization Techniques," *Kybernetika*, vol. 15, pp. 8–27, 1979.

[29] F. Topsøe, "Instances of exact prediction and a new type of inequalities obtained by anchoring," in *Proceedings of the 1999 IEEE Information Theory and Communications Workshop*, Kruger National Park, South Africa, editors: Francis Swarts, Jacobus Swarts, p.99, 1999.

[30] F. Topsøe, "Some Inequalities for Information Divergence and Related Measures of Discrimination," *IEEE trans. Inform. Theory*, vol. 46, pp. 1602–1609, 2000.

[31] F. Topsøe, "Bounds for entropy and divergence for distributions over a two-element set," *J. Ineq. Pure Appl. Math.*, 2001. Accepted. [ONLINE] http://jipam.vu.edu.au/accepted_papers/044_00.html.

[32] I. Vajda, "Bounds on the minimal error probability and checking a finite or countable number of hypothesis," *Information Transmission Problems*, vol. 4, pp. 9–17, 1968.

[33] I. Vajda and K. Vašek, "Majorization, concave entropies and comparison of experiments," *Problems of Control and Information Theory*, vol. 14, pp. 105–115, 1985.

[34] Q. Xie and A.R. Barron, "Asymptotic Minimax Regret for Data Compression, Gambling, and Prediction," *IEEE Trans. Inform. Theory*, vol. 46, pp. 431–445, 2000.