Bounds for entropy and divergence for distributions over a two-element set

Flemming Topsøe
Department of Mathematics
University of Copenhagen, Denmark

18.12.00

Abstract

Three results dealing with probability distributions (p, q) over a two-element set are presented. The two first give bounds for the entropy function $H(p, q)$ and are referred to as the *logarithmic* and the *power-type* bounds, respectively. The last result is a refinement of well known *Pinsker-type inequalities* for information divergence. The refinement readily extends to general distributions, but the key case to consider involves distributions on a two-element set.

The discussion points to some elementary, yet non-trivial problems concerning seemingly simple concrete functions.

Keywords. Entropy, divergence, Pinsker's inequality.

1 Introduction and statements of results

Denote by $M_+^1(\mathbb{N})$ the set of discrete probability distributions over $\mathbb{N}$, typically identified by the set of point probabilities $P = (p_1, p_2, \dots), Q = (q_1, q_2, \dots)$ or what the case may be. *Entropy*, (Kullback-Leibler–) *divergence* and (total) *variation* are defined as usual:

$$H(P) = -\sum_{i=1}^{\infty} p_i \ln p_i, \tag{1}$$

$$D(P\|Q) = \sum_{i=1}^{\infty} p_i \ln \frac{p_i}{q_i}, \tag{2}$$

$$V(P, Q) = \sum_{i=1}^{\infty} |p_i - q_i|. \tag{3}$$

Here, “ln” denotes natural logarithm. Thus we measure entropy and divergence in “nits” (natural units) rather than in “bits”. Admitedly, some of our results, especially the power–type bounds, would look more appealing had we chosen to work with logarithms to the base 2, i.e. with bits.

By $M_+^1(n)$ we denote the set of $P \in M_+^1(\mathbb{N})$ with $p_i = 0$ for $i > n$.

We shall pay special attention to $M_+^1(2)$. Our two first results give bounds for $H(P)$ with $P = (p, q) = (p, q, 0, 0, \dots) \in M_+^1(2)$:

Theorem 1 (Logarithmic bounds). *For any* $P = (p, q) \in M_+^1(2)$,

$$\ln p \cdot \ln q \leq H(p, q) \leq \frac{\ln p \cdot \ln q}{\ln 2}. \tag{4}$$

Theorem 2 (Power–type bounds). *For any* $P = (p, q) \in M_+^1(2)$,

$$\ln 2 \cdot (4pq) \leq H(p, q) \leq \ln 2 \cdot (4pq)^{1/\ln 4}. \tag{5}$$

The proofs are given in sections 2 and 3 and the final section contains a discussions of these inequalities. Here we only remark that the results are best possible in a natural sense, e.g. in Theorem 2 the exponent $1/\ln 4$ is the largest one possible.

The last inequality we shall prove concerns the relation between $D = D(P\|Q)$ and $V = V(P, Q)$. We are interested in lower bounds of D in terms of V. The start of research in this direction is Pinskers inequality

$$D \geq \frac{1}{2} V^2, \tag{6}$$

cf. Pinsker [10] and a later improvement by Csiszár [1], where the best constant for this inequality is found (1/2 as stated in (6)). The best two term inequality of this type is

$$D \geq \frac{1}{2} V^2 + \frac{1}{36} V^4 \tag{7}$$

as proved by Krafft [6].

A further term $1/288\, V^6$ was added by Krafft and Schmitz [7] and Toussaint [12]. For further details see Vajda [13] and also Topsøe [11] where an improvement of the results in [7] and [12] was announced. For present purposes, the *best constants* $c_\nu^{\max}$, $\nu = 0, 1, 2, \dots$, are defined recursively by taking $c_\nu^{\max}$ to be the largest constant c for which the inequality

$$D \geq \sum_{i<\nu} c_i^{\max} V^i + c V^\nu \tag{8}$$

holds generally (for any P and Q in $M_+^1(\mathbb{N})$). Clearly $c_\nu^{\max}$, $\nu = 0, 1, 2, \ldots$, are well defined non-negative real constants.

By the datareduction inequality, cf. Kullback and Leibler [8] and also Csiszár [1], it follows that the determination of lower bounds of the type considered only depends on the interrelationship between D and V for distributions P, Q in $M_+^1(2)$. In particular, in the relation (8) defining the best constants, we may restrict attention to distributions P and Q in $M_+^1(2)$. Thus, researching lower bounds as here, belongs to the theme of the present paper as it essentially amounts to a study of distributions in $M_+^1(2)$. Our contribution is easily summarized:

Theorem 3.

$$c_6^{\max} = \frac{1}{270}, \tag{9}$$

$$c_8^{\max} = \frac{221}{340200}. \tag{10}$$

Corollary 4 (Refinements of Pinsker's inequality). *For any set of probability distributions P and Q, the inequality*

$$D \geq \frac{1}{2}V^2 + \frac{1}{36}V^4 + \frac{1}{270}V^6 + \frac{221}{340200}V^8 \tag{11}$$

holds with $D = D(P\|Q)$ and $V = V(P, Q)$.

Note also that the term $1/270\,V^6$ is better than the term $1/288\,V^6$ which is the term given in the papers by Krafft and Schmitz and by Toussaint. Indeed, the term is the best one in the sense described. And so is the last term in (11). The proofs of these facts depend on an expansion of D in terms of V which is of independent interest. The expansion in question is due to Kambo and Kotz, [5], and is presented in Section 4. The proof of (9) is given in all details in Section 5, whereas the proof of (10), which is similar, is here left to the reader (it may be included in a later publication).

We stress once more that though the proofs deal with distributions on a two-element set, Corollary 4 applies to general distributions.

2 The logarithmic bounds

In this section we prove Theorem 1. The original proof found by the author and supplied for the first version of the manuscript was not elegant but cumbersome (with seven differentiations!). The idea of the simple proof we shall now present is due to O.N. Arjomand, M. Bahramgiri and B.D. Rouhani, Teheran, (private communication). These authors remark that the function f given by

$$f(p) = \frac{H(p,q)}{\ln\, p \cdot \ln\, q}\,;\ 0 \leq p \leq 1 \tag{12}$$

(with $q = 1 - p$ and $f(0)$ and $f(1)$ defined by continuity for $p = 0$ and $p = 1$) can be written in the form

$$f(p) = \varphi(p) + \varphi(q)$$

where φ denotes the function given by

$$\varphi(x) = \frac{x-1}{\ln x}\,;\ x \geq 0 \tag{13}$$

(with $\varphi(0) = 1$), and they observe that φ is concave (details below). It follows that f is concave too, and as f is also symmetric around $p = \frac{1}{2}$, f must be increasing in $[0, \frac{1}{2}]$, decreasing in $[\frac{1}{2}, 1]$. Thus $f(0) \leq f \leq f(\frac{1}{2})$ which is the inequalities claimed in Theorem 1.

The essential concavity of φ is proved by differentiation. Indeed,

$$\varphi''(x) = \frac{-1}{x^2(\ln\, x)^3}\psi(x)$$

with

$$\psi(x) = (x+1)\ln\, x + 2(1-x)\,.$$

As

$$\psi'(x) = \ln\, x - \left(1 - \frac{1}{x}\right) \geq 0\,,$$

and as $\psi(1) = 0$, inspection of the sign of φ'' shows that $\varphi''(x) \leq 0$ for all $x > 0$, and concavity of φ follows.

3 The power–type bounds

In this section we prove Theorem 2.

The lower bound of $H(p,q)$ is a special case of Theorem 2.6 of Harremoës and Topsøe, [3].

A direct proof of this bound is quite easy. We may also apply the technique of the previous section. Indeed, let f^* and φ^* be the functions "dual" to f and φ:

$$f^*(p) = \frac{H(p,q)}{pq}\,;\ 0 \le p \le 1\,, \tag{14}$$

$$\varphi^*(x) = \frac{1}{\varphi(x)} = \frac{\ln\,x}{x-1}\,;\ x \ge 0 \tag{15}$$

($f^*(0) = f^*(1) = \varphi^*(0) = \infty$). Then φ^* is convex and $f^*(p) = \varphi^*(p) + \varphi^*(q)$, so f^* is convex too. Noting also the symmetry of f^*, we see that f^* is decreasing in $\left[0, \frac{1}{2}\right]$, increasing in $[\frac{1}{2}, 1]$. Thus $f^*(\frac{1}{2}) \le f^* \le f^*(0)$ which shows that $4\ln\,2 \le f^* \le \infty$, thereby establishing the lower bound in Theorem 2.

For the proof of the upper bound, we parametrize $P = (p,q)$ by $p = \frac{1+x}{2}$, $q = \frac{1-x}{2}$ and consider only values of x in $[0,1]$. From the cited reference it follows that for no larger exponent α than $\alpha = (\ln 4)^{-1}$ can the inequality

$$H(p,q) \le \ln 2 \cdot (4pq)^{\alpha} \tag{16}$$

hold generally (see also the discussion). For the remainder of this section we put

$$\alpha = \frac{1}{\ln 4}. \tag{17}$$

With this choice of α we have to prove that (16) holds generally. Let ψ denote the auxiliary function

$$\psi = \ln 2 \cdot (4pq)^{\alpha} - H(p,q), \tag{18}$$

conceived as a function of $x \in [0,1]$, i.e.

$$\psi(x) = \ln 2 \cdot (1-x^2)^{\alpha} - \ln 2 + \frac{1+x}{2}\ln(1+x) + \frac{1-x}{2}\ln(1-x). \tag{19}$$

We have to prove that $\psi \ge 0$. Clearly $\psi(0) = \psi(1) = 0$. In contrast to the method used in the previous section we now prefer to base the analysis mainly on the technique of power series expansion. From (19) we find that, at least for $0 \leq x < 1$,

$$\psi(x) = \sum_{\nu=2}^{\infty} \frac{1}{2\nu}\left(\frac{1}{2\nu-1} - \big(1-\alpha\big)\big(1-\frac{\alpha}{2}\big)\cdots\big(1-\frac{\alpha}{\nu-1}\big)\right)x^{2\nu}. \tag{20}$$

Actually (20) also holds for $x = 1$ but we do not need this fact. The computation behind this formula is straight forward when noting that the coefficient $\ln 2 \cdot \binom{\alpha}{\nu}(-1)^{\nu}$ which occurs in the expansion of the first term in (19) can be written as $-\frac{1}{2\nu}(1-\alpha)(1-\frac{\alpha}{2})\cdots(1-\frac{\alpha}{\nu-1})$.

We cannot conclude directly from (20) that $\psi \geq 0$, as (20) contains negative terms, but (20) does show that $\psi'(0) = 0$ and that $\psi(x) > 0$ for $0 < x < \varepsilon$ with $\varepsilon > 0$ sufficiently small. For $0 < x < 1$, we find from (20) that

$$\psi''(x)\frac{1-x^2}{x^2} = 3\alpha - 2 - \sum_{\nu=1}^{\infty}\big(2-2\alpha-\frac{\alpha}{\nu+1}\big)\big(1-\alpha\big)\cdots\big(1-\frac{\alpha}{\nu}\big)x^{2\nu},$$

thus, still for $0 < x < 1$, the equivalence

$$\psi''(x) = 0 \Leftrightarrow \sum_{\nu=1}^{\infty}\big(2-2\alpha-\frac{\alpha}{\nu+1}\big)\big(1-\alpha\big)\cdots\big(1-\frac{\alpha}{\nu}\big)x^{2\nu} = 3\alpha - 2$$

holds. As all terms in the infinite series occuring here are positive, it is clear that ψ only has one inflection point in $]0,1[$. Combining with the facts stated regarding the behaviour of ψ at (or near) the end points, we conclude that $\psi > 0$ in $]0,1[$, thus $\psi \geq 0$.

4 The Kambo–Kotz expansion

The proof of Theorem 3 will be based on the *Kambo–Kotz expansion*, cf. Kambo and Kotz [5] [1], which we shall now discuss. Two distributions P and Q in $M_{+}^{1}(2)$ are involved. For these we choose the basic parametrization

$$P = \left(\frac{1-\alpha}{2}, \frac{1+\alpha}{2}\right), \quad Q = \left(\frac{1+\beta}{2}, \frac{1-\beta}{2}\right), \tag{21}$$

and we consider values of the parameters as follows: $-1 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$. We shall also work with another parametrization (ρ, V) where

$$\rho = \frac{\alpha}{\beta}, \quad V = |\alpha + \beta|. \tag{22}$$

[1] The result is contained in the proof of Lemma 3 of that paper; there is a minor numerical error in the statement of this lemma, cf. Krafft, [6]

Here, V is the total variation $V(P,Q)$, the essential parameter in Pinsker-type inequalities. We may avoid the inconvenient case $\beta = 0$ simply by noting that this case corresponds to $Q = U_2$ (the uniform distribution $(\frac{1}{2}, \frac{1}{2})$) which will never cause difficulties in view of the simple expansion

$$D(P\|U_2) = \sum_{\nu=1}^{\infty} \frac{V^{2\nu}}{2\nu(2\nu-1)} \tag{23}$$

with $V = V(P,Q)$ (actually derived in section 3 in view of the identity $D(P\|U_2) = \ln 2 - H(P)$).

Fig. 1. Parameter domain for the Kambo-Kotz expansion with indication of the *critical domain* (for explanation see further on in the text).

Denote by Ω the subset of the (ρ, V)-plane scetched in Figure 1. To be precise,

$$\Omega = \{(-1,0)\} \cup \Omega_1 \cup \Omega_2 \cup \Omega_3 \tag{24}$$

with

$$\Omega_1 = \{(\rho,V) \mid \rho < -1,\, 0 < V \leq 1 + 1/\rho\}, \tag{25}$$
$$\Omega_2 = \{(\rho,V) \mid -1 < \rho \leq 1,\, 0 < V \leq 1 + \rho\}, \tag{26}$$
$$\Omega_3 = \{(\rho,V) \mid 1 < \rho,\, 0 < V \leq 1 + 1/\rho\}. \tag{27}$$

From [5] we have (adapting notation etc. to our setting):

Theorem 5 (Kambo-Kotz expansion). *Consider P and Q of the form* (21)*, assume that $\beta > 0$ and define ρ and V by* (22)*. Then $(\rho, V) \in \Omega$ and*

$$D(P\|Q) = \sum_{\nu=1}^{\infty} \frac{f_\nu(\rho)}{2\nu(2\nu-1)} V^{2\nu}, \tag{28}$$

where f_ν; $\nu \geq 1$, are rational funtions defined by

$$f_\nu(\rho) = \frac{\rho^{2\nu} + 2\nu\rho + 2\nu - 1}{(\rho+1)^{2\nu}}; \quad \rho \neq -1. \tag{29}$$

We note that the value of f_ν for $\rho = -1$ is immaterial in (28) as $V = 0$ when $\rho = -1$ hence, with usual conventions, (28) gives the correct value $D = 0$ in this case too. However, we do find it natural to define $f_1(-1) = 1$ and $f_\nu(-1) = \infty$ for $\nu \geq 2$.

The functions f_ν are essential for the further analysis. We shall refer to them as the *Kambo–Kotz functions*. We need the following result:

Lemma 6 (Basic properties of the Kambo–Kotz functions). *All functions f_ν; $\nu \geq 1$, are everywhere positive, f_1 is the constant function 1 and all other functions f_ν assume their minimal value at a uniquely determined point ρ_ν which is the only stationary point of f_ν. We have $\rho_2 = 2$, $1 < \rho_\nu < 2$ for $\nu \geq 3$ and $\rho_\nu \to 1$ as $\nu \to \infty$.*

For $\nu \geq 2$, f_ν is strictly increasing in the two intervals $]-\infty, -1[$ and $[2, \infty[$ and f_ν is strictly decreasing in $]-1, 1]$. Furthermore, f_ν is strictly convex in $[1, 2]$ and, finally, $f_\nu(\rho) \to 1$ for $\rho \to \pm\infty$.

Proof. Clearly, $f_1 \equiv 1$. For the rest of the proof assume that $\nu \geq 2$. For $\rho \geq 0$, $f_\nu(\rho) > 0$ by (29) and for $\rho < 0$, we can use the formula

$$f_\nu(\rho) = (\rho+1)^{-(2\nu-2)} \sum_{k=2}^{2\nu} (-1)^k (k-1)\rho^{2\nu-k} \tag{30}$$

and realize that $f_\nu(\rho) > 0$ in this case, too.

We need the following formulas:

$$f'_\nu(\rho) = 2\nu(\rho+1)^{-(2\nu+1)}(\rho^{2\nu-1} - (2\nu-1)\rho - (2\nu-2)) \tag{31}$$

and

$$f''_\nu(\rho) = 2\nu(\rho+1)^{-(2\nu+2)} \cdot g_\nu(\rho), \tag{32}$$

with the auxiliary function g_ν given by

$$g_\nu(\rho) = -2\rho^{2\nu-1} + (2\nu-1)\rho^{2\nu-2} + 2\nu(2\nu-1)\rho + 4\nu^2 - 4\nu - 1. \tag{33}$$

By (31), $f'_\nu > 0$ in $]-\infty, -1]$ and $f'_\nu < 0$ in $]-1, 1]$. The sign of f'_ν in $[1, 2]$ is the same as that of $\rho^{2\nu-1} - (2\nu-1)\rho - (2\nu-2)$ and by differentiation and evaluation at $\rho = 2$, we see that $f'_\nu(\rho) = 0$ at a unique point $\rho = \rho_\nu$ in $]1, 2]$. Furthermore, $\rho_2 = 2$, $1 < \rho_\nu < 2$ for $\nu \geq 3$ and $\rho_\nu \to 1$ for $\nu \to \infty$. Investigating further the sign of f'_ν, we find that f_ν is strictly increasing in $[2, \infty[$. As $f_\nu(\rho) \to 1$ for $\rho \to \pm\infty$ by (29), we now conclude that f_ν has the stated monotonicity behaviour. To prove the convexity assertion, note that g_ν defined by (33) determines the sign of f''_ν. For $\nu = 2$, $g_2(\rho) = 2(2-\rho)\rho^2 + \rho(12-\rho) + 7$ which is positive in $[1, 2]$. A similar conclusion can be drawn in case $\nu = 3$ since $g_3(\rho) = 2\rho^4(2-\rho) + \rho^4 + 30\rho + 23$. For the general case $\nu \geq 4$, we note that $g_\nu(1) = 4(\nu-1)(2\nu+1) > 0$ and we can then close the proof by showing that g_ν is increasing in $[1, 2]$. Indeed, $g'_\nu = (2\nu-1)h_\nu$ with $h_\nu(\rho) = -2\rho^{2\nu-2} + (2\nu-2)\rho^{2\nu-3} + 2\nu$, hence $h_\nu(1) = 4(\nu-1) > 0$ and $h'_\nu(\rho) = (2\nu-2)(2\nu-3-2\rho)\rho^{2\nu-4}$ which is positive in $[1, 2]$. □

In the sequel, we shall write $D(\rho, V)$ in place of $D(P\|Q)$ with P and Q parametrized as explained by (21) and (22).

Fig. 2. A typical Kambo-Kotz function shown in normal/logarithmic scale.

Figure 2 illustrates the behaviour of the Kambo–Kotz functions. In order to illustrate as clearly as possible the nature of these functions, the graph shown is actually that of the logarithm of one of the Kambo-Kotz functions.

Note that if we extend the domain Ω by the points $(\pm\infty, V)$ with $0 < V \leq 1$, then (28) reduces to (23). Therefore, we may consider the case $\beta = 0$ as a singular or limiting case for which (28) also holds.

Motivated by the lemma, we define the *critical domain* as the set

$$\begin{aligned} \Omega^* &= \{(\rho, V) \in \Omega \mid 1 \leq \rho \leq 2\} \\ &= \{(\rho, V) \in \Omega \mid 1 \leq \rho \leq 2,\ 0 < V < 1 + 1/\rho\}. \end{aligned} \tag{34}$$

We then realize that in the search for lower bounds of D in terms of V we may restrict the attention to the critical domain. In particular:

Corollary 7. *For each* $\nu_0 \geq 1$

$$c_{\nu_0}^{\max} = \inf\Big\{V^{-\nu_0}\big(D(\rho,V) - \sum_{\nu<\nu_0} c_\nu^{\max} V^\nu\big) \mid (\rho,V) \in \Omega^*\Big\}. \tag{35}$$

5 A refinement of Pinskers inequality

In this section we prove Theorem 3.

We use notation and results from the previous section. We shall determine the best constants $c_\nu^{\max}$, $\nu = 0, 1, \ldots, 8$ in the inequality $D \geq \sum_{\nu=0}^\infty c_\nu V^\nu$, cf. the explanation in the introductory section. In fact, we shall mainly focus on the determination of $c_6^{\max}$. The reason for this is that the value of $c_\nu^{\max}$ for $\nu \leq 4$ is known and that it is pretty clear (see analysis below) that $c_5^{\max} = c_7^{\max} = 0$. Further, the determination of $c_8^{\max}$, though more complicated, is rather similar to that of $c_6^{\max}$.

Before we continue, let us briefly indicate that from the Kambo–Kotz expansion and the identities $f_1 \equiv 1$ and

$$f_2(\rho) = \frac{1}{3}\left(1 + \frac{2(2-\rho)^2}{(1+\rho)^2}\right) \tag{36}$$

one deduces the results regarding $c_\nu^{\max}$ for $\nu \leq 4$ (in fact for $\nu \leq 5$).

Now then, let us determine $c_6^{\max}$. From the identity

$$\begin{aligned} & D(\rho,V) - \frac{1}{2}V^2 - \frac{1}{36}V^4 \\ = \; & \frac{1}{18}\left(\frac{2-\rho}{1+\rho}\right)^2 V^4 + \frac{1}{30}\frac{\rho^6+6\rho+5}{(1+\rho)^6}V^6 + \sum_{\nu=4}^\infty \frac{f_\nu(\rho)}{2\nu(2\nu-1)}V^{2\nu}, \end{aligned} \tag{37}$$

we see that $c_6^{\max} \leq 1/270$ (take $\rho = 2$ and consider small V's). In order to show that $c_6^{\max} \geq 1/270$, we recall (Lemma 6) that each term in the sum $\sum_4^\infty$ in (37) is non-negative, hence it suffices to show, that

$$\frac{1}{18}\left(\frac{2-\rho}{1+\rho}\right)^2 V^{-2} + \frac{f_3(\rho)}{30} + \frac{f_4(\rho)}{56}V^2 \geq \frac{1}{270}. \tag{38}$$

Here we could restrict (ρ, V) to the critical domain Ω^*, but we may also argue more directly as follows: If $\rho \geq 2$, the middle term alone in (38) dominates

1/270. Then, since for fixed non-negative s and t, the minimal value of $sV^{-2}+tV^2$ is $2\sqrt{st}$, it suffices to show that

$$\frac{f_3(\rho)}{30}+2\sqrt{\frac{(2-\rho)^2(\rho^8+8\rho+7)}{18\cdot 56\cdot(1+\rho)^{10}}}\geq\frac{1}{270}$$

for $\rho<2$, i.e. we must check that

$$8\rho^3-6\rho^2+9\rho-22\leq\frac{45}{\sqrt{7}}\sqrt{\rho^6-2\rho^5+3\rho^4-4\rho^3+5\rho^2-6\rho+7}$$

holds (here, factors of $1+\rho$ and $2-\rho$ have been taken out). In fact, even the square of the left-hand term is dominated by the square of the right-hand term for all $\rho\in\mathbb{R}$. This claim amounts to the inequality

$$45^2(\rho^6-2\rho^5+3\rho^4-4\rho^3+5\rho^2-6\rho+7)\geq 7(8\rho^3-6\rho^2+9\rho-22)^2. \quad (39)$$

An elementary way to verify (39) runs as follows: Write the equation in the form

$$\sum_{\nu=0}^{6}(-1)^\nu a_\nu\rho^\nu\geq 0, \quad (40)$$

and note that, for all $\rho\in\mathbb{R}$

$$\sum_{\nu=0}^{6}(-1)^\nu a_\nu\rho^\nu\geq x\rho^4+\sum_{\nu=0}^{3}(-1)^\nu a_\nu\rho^\nu\geq y\rho^2+\sum_{\nu=0}^{1}(-1)^\nu a_\nu\rho^\nu\geq z,$$

with

$$x=a_4-\frac{a_5^2}{4a_6},\quad y=a_2-\frac{a_3^2}{4x},\quad z=a_6-\frac{a_1^2}{4y}$$

(since a_6, x and y are all positive). Since $z>0$ (in fact, $z\approx 6949.51$), (40) and therefore also (39) follow. Thus $c_6^{\max}=1/270$.

6 Discussion

Theorem 1:

Emphasis here is on the quite precise upper bound of $H(p,q)$. An explanation of the origin of the upper bound may not be all that helpful to the reader.

Basically, the author stumbled over the inequality (in the search for a natural proof of Theorem 2, cf. below), and has no special use in mind for it. The reader may take it as a curiosity, an ad-hoc inequality. It is not known if the inequality has natural generalisations to distributions in $M_+^1(3)$, $M_+^1(4)$, $\ldots$.

Theorem 2:

This result, again with emphasis on the upper bound, is believed to be of greater significance. It is discussed, together with generalizations to $M_+^1(n)$, in Harremoës and Topsøe [3]. Applications to statistics (decision theory, Chernoff bound) appear promising. The term $4pq$ in the inequality should best be thought of as 1 minus the *relative measure of roughness* introduced in [3]. The term may, qualitatively, be taken to measure the closeness to the "flat" uniform distribution $(1/2, 1/2)$. It varies from 0 (for a deterministic distribution) to 1 (for the uniform distribution).

As stated in the introduction, the exponent $1/\ln 4 \approx 0.7213$ is best possible. A previous result by Lin [9] establishes the inequality with exponent 1/2, i.e. $H(p,q) \leq \ln 2\sqrt{4pq}$.

Theorem 2 was stated in [3] but not proved there.

Comparing the logarithmic and the power-type bounds:

The two lower bounds are shown graphically in Figure 3. The power bound is normally much sharper and it is the best bound, except for distributions close to a deterministic distribution ($\max(p,q)$ >0.9100).

Both upper bounds are quite accurate for all distributions in $M_+^1(2)$ but, again, the power bound is slightly better, except when (p,q) is very close to a deterministic distribution ($\max(p,q)$ >0.9884). Because of the accuracy of the two upper bounds, a simple graphical presentation together with the entropy function will not enable us to distinguish between the three functions. Instead, we have shown in Figure 4 the difference between the two upper bounds (logarithmic bound minus power-type bound).

Fig. 3: Lower bounds

Fig. 4: Difference of upper bounds

Fig. 5: Ratios regarding lower bounds

Fig. 6: Ratios regarding upper bounds

Thus, for both upper and lower bounds, the power–type bound is usually the best one. However, an attractive feature of the logarithmic bounds is that the quotient between the entropy function and the $\ln p \ln q$ function is bounded. On Figures 5 and 6 we have shown the ratios: entropy to lower bounds, and: upper bounds to entropy. Note (hardly visible on the graphs in Figure 6), that for the upper bounds, the ratios shown approaches infinity for the power bound but has a finite limit ($1/\ln 2 \approx 1.44$) for the logarithmic bound when (p, q) approaches a deterministic distribution.

Other proofs of Theorem 1:

As indicated already, the first proof found by the author was not very satisfactory, and the author asked for more natural proofs, which should also display the monotonicity property of the function f given by (12). Several responses were received. The one by Arjomand, Bahramgiri and Rouhani was reflected in section 2. Another suggestion came from Iosif Pinelis, Houghton, Michigan (private communication), who showed that the following general l'Hospital – type of result may be taken as the basis for a proof:

Lemma. *Let f and g be differentiable functions on an interval $]a,b[$ such that $f(a+) = g(a+) = 0$ or $f(b-) = g(b-) = 0$, g' is nonzero and does not change sign, and f'/g' is increasing (decreasing) on (a,b). Then f/g is increasing (respectively, decreasing) on $]a,b[$.*

Other proofs have been obtained as response to the authors suggestion to work with power series expansions. As the feed-back obtained may be of interest in other connections (dealing with other inequalities or other type of problems), we shall indicate the considerations involved, though for the specific problem, the methods discussed above are more elementary and also more expedient.

Let us parametrize $(p,q) = (p, 1-p)$ by $x \in [-1,1]$ via the formula

$$p = \frac{1+x}{2},$$

and let us first consider the analytic function

$$\varphi(x) = \frac{1}{\ln\frac{1+x}{2}}; \quad |x| < 1.$$

Let

$$\varphi(x) = \sum_{\nu=0}^{\infty} \gamma_\nu x^\nu; \quad |x| < 1, \tag{41}$$

be the Taylor expansion of φ and introduce the abbreviation $\lambda = \ln 2$. One finds that $\gamma_0 = -1/\lambda$ and that

$$f\Big(\frac{1+x}{2}\Big) = \frac{1}{\lambda} - \sum_{\nu=1}^{\infty}(\gamma_{2\nu} - \gamma_{2\nu-1})x^{2\nu}; \quad |x| < 1. \tag{42}$$

Numerical evidence indicates that $\gamma_2 \geq \gamma_4 \geq \gamma_6 \geq \cdots$, that $\gamma_1 \leq \gamma_3 \leq \gamma_5 \leq \cdots$ and that both sequences converge to -2. However, it appears that the natural question to ask concerns the Taylor coefficients of the analytic function

$$\psi(x) = \frac{2}{1+x} + \frac{1}{\ln(\frac{1-x}{2})}\,; |x| < 1\,. \tag{43}$$

Let us denote these coefficients by β_ν; $\nu \le 0$, i.e.

$$\psi(x) = \sum_{k=0}^{\infty} \beta_k x^k \,; |x| < 1\,. \tag{44}$$

The following conjecture is easily seen to imply the desired monotonicity property of f as well as the special behaviour of the γ's:

Conjecture 1. *The sequence $(\beta_\nu)_{\nu \ge 0}$ is decreasing with limit* 0.

In fact, this conjecture was settled in the positive, independently, by Christian Berg, Copenhagen, and by Miklós Laczkovich, Budapest (private communications). Laczkovich used the residue calculus in a straightforward manner and Berg appealed to the theory of so-called Pick-functions – a theory which is of great significance for the study of many inequalities, including matrix type inequalities. In both cases the result is an integral representation for the coefficients β_ν, which immediately implies the conjecture.

It may be worth while to note that the β_ν's can be expressed as combinations involving certain symmetric functions, thus the settlement of the conjecture gives information about these functions. What we have in mind is the following: Guided by the advice contained in Henrici [4] we obtain expressions for the coefficients β_ν which depend on numbers $h_{\nu,j}$ defined for $\nu \ge 0$ and each $j = 0, 1, \dots, \nu$, by $h_{\nu,0} = 1$ and

$$h_{\nu,j} = \sum_{1 \le i_1 < \cdots < i_j \le \nu} (i_1 i_2 \cdots i_j)^{-1}.$$

Then, for $k \ge 1$,

$$\beta_k = 2(-1)^k - \frac{1}{k\lambda} \sum_{\nu=1}^{k} \frac{(-1)^\nu \nu!}{\lambda^\nu} h_{k-1,\nu-1}\,. \tag{45}$$

A natural proof of Theorem 2:

Denote by g the function

$$g(p) = \frac{\ln\left(\frac{H(p,q)}{\ln 2}\right)}{\ln(4pq)}; \quad 0 \le p \le 1, \tag{46}$$

with $q = 1-p$. This function is defined by continuity in the critical points, i.e. $g(0) = g(1) = 1$ and $g(1/2) = 1/\ln 4$. Clearly, g is symmetric around $p = 1/2$ and the power-type bounds of Theorem 2 are equivalent to the inequalities

$$g(1/2) \le g(p) \le g(1). \tag{47}$$

Our proof (in Section 3) of these inequalities was somewhat ad hoc. Numerical or graphical evidence points to a possible natural proof which will even establish monotonicity of g in each of the intervals $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. The natural conjecture to propose which implies these empirical facts is the following:

Conjecture 2. *The function g is convex.*

Last minute input obtained from Iosif Pinelis established the desired monotonicity properties of g. Pinelis' proof of this fact is elementary, relying once more on the above l'Hospital type of lemma.

Pinsker type inequalities:

While completing the manuscript, new results were obtained in collaboration with Alexei Fedotov and Peter Harremoës. These results will be published in a separate paper. Among other things, a determination in closed form (via a parametrization) of *Vajda's tight lower bound*, cf. [13], has been obtained. This research also points to some obstacles when studying further terms in refinements of Pinsker's inequality. It may be that an extension beyond the result in Corollary 4 will need new ideas.

Acknowledgements

The author thanks Alexei Fedotov and Peter Harremoës for useful discussions, and further, he thanks O. Naghshineh Arjomand, M. Bahramgiri, Behzad Djafari Rouhani, Christian Berg, Miklós Laczkovich and Iosif Pinelis for contributions which settled open questions contained in the first version of the paper, and for accepting the inclusion of hints or full proofs of these results in the final version.

References

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

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

[3] P. Harremoës. and F. Topsøe, *Inequalities between Entropy and Index of Coincidence derived from Information Digrams*, submitted for publication.

[4] P. Henrici, *Applied and Computational Complex Analysis, vol. 1*, New York: Wiley, 1988.

[5] N. S. Kambo, and S. Kotz, *On exponential bounds for binomial probabilities*, Ann. Inst. Stat. Math., vol. 18, pp. 277–287, 1966.

[6] O. Krafft, *A note on exponential bounds for binomial probabilities*, Ann. Inst. Stat. Math., vol. 21, pp. 219–220, 1969.

[7] O. Krafft, and N. Schmitz, *A note on Hoefding's inequality,* J. Amer. Statist. Assoc., vol. 64 , pp. 907–912, 1969.

[8] S. Kullback, and R. Leibler, *On information and sufficiency*, Ann. Math. Statist., vol. 22, pp. 79–86, 1951

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

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

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

[12] G.T. Toussaint, *Sharper lower bounds for discrimination information in terms of variation,*IEEE Trans. Inform. Theory, vol 21, pp. 99–100, 1975.

[13] I. Vajda, *Note on discrimination information and variation,* IEEE Trans. Inform. Theory, vol. 16, pp. 771–773, 1970.