Some Upper Bounds for Relative Entropy and Applications

SOME UPPER BOUNDS FOR RELATIVE ENTROPY AND APPLICATIONS

S.S. DRAGOMIR, M. L. SCHOLZ, AND J. SUNDE

ABSTRACT. In this paper we derive some upper bounds for the relative entropy $D(p \parallel q)$ of two probability distribution and apply them to mutual information and entropy mapping. To achieve this we use an inequality for the logarithm function, (2.3) below, and some classical inequalities such as the Kantorovič Inequality and Diaz-Metcalf Inequality.

1. INTRODUCTION

To design a communication system with a specific message handling capability we need a measure of information content to be transmitted. The *entropy* of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable.

The *relative entropy* is a measure of the distance between two distributions. In statistics, it arises as the expectation of the logarithm of the likelihood ratio. The relative entropy $D(p \parallel q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. For example, if we knew the true distribution of the random variable, then we could construct a code with average description length $H(p)$. If, instead, we used the code for a distribution q, we would need $H(p) + D(p\|q)$ bits on the average to describe the random variable [6, p. 18].

Definition. *(Relative Entropy) The relative entropy or, Kullback-Leibler distance, between two probability mass functions $p(x)$ and $q(x)$ is defined by*

$$\begin{aligned} D(p\|q) \quad : \quad &= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) \\ &= E_p \log\left(\frac{p}{q}\right), \end{aligned}$$

where log *will always denote the natural logarithm.*

In the above definition, we use the convention (based on continuity arguments) that $0\log\left(\frac{0}{q}\right) = 0$ and $p\log\left(\frac{p}{0}\right) = \infty$.

It is well-known that relative entropy is always non-negative and zero if and only if $p = q$. However, this is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality.

The following theorem is of fundamental importance [6, p. 26].

Date: November, 1998.

1991 *Mathematics Subject Classification.* 94Xxx, 26D15.

Key words and phrases. Relative Entropy, Mutual Information, Log-mapping, Kantorovič Inequality, Diaz-Metcalf Inequality.

Thanks for DSTO TSS funding support.

Theorem A. *(Information Inequality) Let $p(x)$, $q(x), x \in \mathcal{X}$, be two probability mass functions. Then*

$$D(p\|q) \geq 0 \tag{1.1}$$

with equality if and only if

$$p(x) = q(x) \text{ for all } x \in \mathcal{X}.$$

Proof. Let $\mathcal{A} := \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned} -D(p\|q) &= -\sum_{x\in\mathcal{A}} p(x)\log\left(\frac{p(x)}{q(x)}\right) = \sum_{x\in\mathcal{A}} p(x)\log\left(\frac{q(x)}{p(x)}\right) \\ &\leq \log\left(\sum_{x\in\mathcal{A}} p(x)\frac{q(x)}{p(x)}\right) \\ &= \log\left(\sum_{x\in\mathcal{A}} q(x)\right) \leq \log\left(\sum_{x\in\mathcal{X}} q(x)\right) \\ &= \log 1 = 0 \end{aligned}$$

where the first inequality follows from Jensen's inequality.

Since log is strictly concave, we have equality above if and only if $q(x)/p(x) = 1$ everywhere. Hence we have $D(p \parallel q) = 0$ if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$. ∎

Actually, the inequality (1.1) can be improved as follows (see , [6, p. 300]):

Theorem B. *Let p, q be as above. Then*

$$D(p \parallel q) \geq \frac{1}{2}\|p-q\|_1^2 \tag{1.2}$$

where $\|p-q\|_1 = \sum_{x\in\mathcal{X}} |p(x) - q(x)|$ is the usual 1-norm of $p - q$.

We remark that the argument of (1.2) is not based on the convexity of $-\log$.

To estimate the relative entropy $D(p \parallel q)$ it would be interesting to establish some upper bounds.

Before we do this, let us recall some other important concepts in Information Theory.

We consider *mutual information*, which is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other [6, p. 18].

Definition. *(Mutual Information) Consider two random variables X and Y with a joint probability mass function $t(x, y)$ and marginal probability mass function $p(x)$ and $q(y)$. The mutual information is the relative entropy between the joint distribution and the product distribution, i.e.*

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} t(x,y)\log\left(\frac{t(x,y)}{p(x)q(y)}\right) \\ &= D\left(t(x,y) \parallel p(x)q(y)\right) \\ &= E_t \log\left(\frac{t}{pq}\right). \end{aligned}$$

The following equalities are well-known [6, p. 20]

$$I(X;Y) = H(X) - H(X|Y),$$

$$I(X;Y) = H(Y) - H(Y|X),$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$

$$I(X;Y) = I(Y;X)$$

and

$$I(X;X) = H(X)$$

where

$$H(X|Y) := \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} t(x,y)\log\frac{1}{t(x|y)}, t(x|y) = t(x,y)/q(y)$$

is the *conditional entropy* of X provided Y and

$$H(X,Y) := \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} t(x,y)\log\frac{1}{t(x,y)}$$

is the *joint entropy* of X and Y (see for example [6, p. 15-16]).

The following corollaries of Theorem A are important [6, p. 27].

Corollary C. *(Non-negativity of mutual information): For any two random variables, X, Y*

$$I(X;Y) \geq 0 \tag{1.3}$$

with equality if and only if X and Y are independent.

Now, let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform probability mass function on $\mathcal{X}$ and let $p(x)$ be the probability mass function for X.

It is well-known that [6, p. 27]

$$\begin{aligned} D(p\|u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} \\ &= \log|\mathcal{X}| - H(X). \end{aligned}$$

Corollary D. *Let X be a random variable and $|\mathcal{X}|$ denotes the number of elements in the range of $\mathcal{X}$. Then*

$$H(X) \leq \log|\mathcal{X}|$$

with equality if and only if X has a uniform distribution over $\mathcal{X}$.

An improvement of (1.3) can be obtained via the inequality (1.2) as follows:

Corollary 1. *Under the above assumptions, we have*

$$I(X;Y) \geq \frac{1}{2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |t(x,y) - p(x)q(y)| \geq 0.$$

2. An Upper Bound for the relative entropy

We start this section with the new upper bound for $D(p \parallel q)$.

Theorem 1. *Let $p(x)$, $q(x) > 0$, $x \in \mathcal{X}$ be two probability mass functions. Then*

$$D(p\|q) \leq \sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} - 1 = \frac{1}{2}\sum_{x,y\in\mathcal{X}} p(x)p(y)\left(\frac{p(x)}{q(x)} - \frac{p(y)}{q(y)}\right)\left(\frac{q(y)}{p(y)} - \frac{q(x)}{p(x)}\right) \tag{2.1}$$

with equality if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof. We know that for every differentiable real valued strictly convex function f defined on an interval I of the real line, we have the inequality

$$f(b) - f(a) \geq f^{'}(a)(b-a) \tag{2.2}$$

for all a, $b \in I$. The equality holds if and only if $a = b$.

Now, apply (2.2) to $f(x) = -\log x$ and $I = (0, \infty)$ to get

$$\log a - \log b \geq \frac{1}{a}(a-b) \tag{2.3}$$

for all a, $b > 0$.

Choose $a = q(x)$, $b = p(x)$, $x \in \mathcal{X}$. Then, by (2.3), we get

$$\log q(x) - \log p(x) \geq \frac{1}{q(x)}\left(q(x) - p(x)\right), \quad x \in \mathcal{X}.$$

Multiplying by $p(x) \geq 0$, we get

$$p(x)\log p(x) - p(x)\log q(x) \leq \frac{p(x)}{q(x)}\left(p(x) - q(x)\right)$$

i.e.

$$p(x)\log\frac{p(x)}{q(x)} \leq \frac{p(x)}{q(x)}\left(p(x) - q(x)\right) = \frac{p^2(x)}{q(x)} - p(x)$$

for all $x \in \mathcal{X}$.

Summing over $x \in \mathcal{X}$, we get

$$\begin{aligned} D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\ &\leq \sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} - \sum_{x\in\mathcal{X}} p(x) \\ &= \sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} - 1. \end{aligned}$$

The case of equality follows by the strict convexity of $-\log$ and we omit the details.

To prove the last equality, let observe that

$$
\begin{aligned}
& \frac{1}{2}\sum_{x,y\in\mathcal{X}} p(x)p(y)\left(\frac{p(x)}{q(x)}-\frac{p(y)}{q(y)}\right)\left(\frac{q(y)}{p(y)}-\frac{q(x)}{p(x)}\right) \\
= & \frac{1}{2}\sum_{x,y\in\mathcal{X}} p(x)p(y)\left(\frac{p(x)q(y)}{q(x)p(y)}+\frac{p(y)q(x)}{q(y)p(x)}-\frac{p(x)q(y)}{q(y)p(y)}-\frac{p(x)q(y)}{q(x)p(x)}\right) \\
= & \frac{1}{2}\left[\sum_{x,y\in\mathcal{X}} p^2(x)\frac{q(y)}{q(x)}+\sum_{x,y\in\mathcal{X}}\frac{p^2(y)q(x)}{q(y)}\right. \\
& \left.-\sum_{x,y\in\mathcal{X}} p(x)p(y)-\sum_{x,y\in\mathcal{X}} p(x)p(y)\right] \\
= & \sum_{x\in\mathcal{X}}\frac{p^2(x)}{q(x)}-1
\end{aligned}
$$

and the last part of (2.1) is also proved. ∎

Remark 1. In paper [8] (related to The Noiseless Coding Theorem) a similar result was obtained. For further developments and counterparts see [3] and [4] .

We provide two corollaries.

Corollary 2. *Let X and Y be two random variables. Then*

$$I(X;Y) \leq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{t^2(x,y)}{p(x)q(y)} - 1.$$

The equality holds if and only if X and Y are independent.

Proof. We know that

$$
\begin{aligned}
I(X;Y) &= D\left(t(x,y)\parallel p(x)q(y)\right) \\
&\leq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{t^2(x,y)}{p(x)q(y)} - 1, \text{ (by Theorem 1),}
\end{aligned}
$$

with equality if and only if $t(x,y) = p(x)q(y)$ for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$. ∎

Corollary 3. *Let X be a random variable whose range has $|\mathcal{X}|$ elements. Then*

$$
\begin{aligned}
0 &\leq \log|\mathcal{X}| - H(X) \leq |\mathcal{X}|\sum_{x\in\mathcal{X}} p^2(x) - 1 \\
&= \frac{1}{2}\sum_{x,y\in\mathcal{X}} \left(p(x)-p(y)\right)^2 .
\end{aligned} \tag{2.4}
$$

The equality holds if and only if X has a uniform distribution over $\mathcal{X}$.

Proof. We know that

$$
\begin{aligned}
\log|\mathcal{X}| - H(X) &= D\left(p\parallel u\right) \leq \sum_{x\in\mathcal{X}} \frac{p^2(x)}{\frac{1}{|\mathcal{X}|}} - 1 \\
&= |\mathcal{X}|\sum_{x\in\mathcal{X}} p^2(x) - 1 = \frac{1}{2}\sum_{x,y\in\mathcal{X}} \left(p(x)-p(y)\right)^2 .
\end{aligned}
$$

and the corollary is proved. ∎

Remark 2. The inequality (2.4) has been proved by S.S. Dragomir and C.J. Goh in the paper [2] as a particular case of a more general inequality which is a counterpart of Jensen's Inequality.

It would be useful for practical applications to find out sufficient condition for the probabilities $p(x)$ and $q(x)$, $x \in \mathcal{X}$ such that $D(p \parallel q) \leq \varepsilon$ where $\varepsilon > 0$ is sufficiently small.

For this purpose, let us define $r(x) = p(x)/q(x)$, $x \in \mathcal{X}$ where we have assumed that $p(x) > 0, q(x) > 0$ for all $x \in \mathcal{X}$. Put

$$R := \max_{x \in \mathcal{X}} r(x), \quad r := \min_{x \in \mathcal{X}} r(x).$$

Consider also the quotient

$$S := \frac{R}{r} \geq 1.$$

The following theorem holds:

Theorem 2. *Let $\varepsilon > 0$ and $p(x)$, $q(x)$, $x \in \mathcal{X}$ be two probability mass functions so that*

$$S \leq 1 + \varepsilon + \sqrt{\varepsilon(\varepsilon + 2)}. \tag{2.5}$$

Then

$$D\left(p \parallel q\right) \leq \varepsilon. \tag{2.6}$$

Proof. Define

$$\begin{aligned} K \quad : \quad &= \frac{1}{2} \sum_{x,y \in \mathcal{X}} p(x)p(y) \left(\frac{p(x)}{q(x)} - \frac{p(y)}{q(y)} \right) \left(\frac{q(y)}{p(y)} - \frac{q(x)}{p(x)} \right) \\ &= \frac{1}{2} \sum_{x,y \in \mathcal{X}} p(x)p(y) \left(r(x) - r(y) \right) \left(\frac{1}{r(y)} - \frac{1}{r(x)} \right) \\ &= \frac{1}{2} \sum_{x,y \in \mathcal{X}} p(x)p(y) \frac{\left(r(x) - r(y) \right)^2}{r(x)r(y)}. \end{aligned}$$

Now, observe that (for all $x, y \in \mathcal{X}$)

$$\frac{\left(r(x) - r(y) \right)^2}{2r(x)r(y)} \leq \varepsilon$$

is equivalent to

$$r^2(x) - 2(1+\varepsilon)r(x)r(y) + r^2(y) \leq 0$$

or, moreover, to

$$\left[\frac{r(x)}{r(y)} \right]^2 - 2(1+\varepsilon)\frac{r(x)}{r(y)} + 1 \leq 0 \text{ for all } x, y \in \mathcal{X},$$

which in turn is equivalent to:

$$1 + \varepsilon - \sqrt{\varepsilon(\varepsilon+2)} \leq \frac{r(x)}{r(y)} \leq 1 + \varepsilon + \sqrt{\varepsilon(\varepsilon+2)}. \tag{2.7}$$

Now, if (2.5) holds, then

$$\frac{r(x)}{r(y)} \leq S \leq 1 + \varepsilon + \sqrt{\varepsilon(\varepsilon+2)},$$

i.e. the second part of (2.7).

Also, we have

$$\frac{r(x)}{r(y)} \geq \frac{1}{S} \geq \frac{1}{1+\varepsilon+\sqrt{\varepsilon(\varepsilon+2)}} = 1+\varepsilon-\sqrt{\varepsilon(\varepsilon+2)},$$

i.e. the first part of (2.7).

Consequently, the condition (2.2) implies that

$$\frac{(r(x)-r(y))^2}{2r(x)r(y)} \leq \varepsilon$$

and then

$$K \leq \varepsilon \sum_{x,y\in\mathcal{X}} p(x)p(y) = \varepsilon.$$

Using the inequality (2.1) we deduce the desired estimation (2.6). ∎

We have the following corollaries.

Corollary 4. *Let X, Y be two random variables so that $p(x), q(y), t(x,y) > 0$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Put*

$$M = \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{t(x,y)}{p(x)q(y)}, \qquad m = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{t(x,y)}{p(x)q(y)},$$

and define

$$\mu = \frac{M}{m} \geq 1.$$

If

$$\mu \leq 1+\varepsilon+\sqrt{\varepsilon(\varepsilon+2)}, \quad \varepsilon > 0$$

then we have

$$I(X;Y) \leq \varepsilon.$$

Corollary 5. *Let X be a random variable whose range has $|\mathcal{X}|$ elements and $p(x) > 0$, $x \in \mathcal{X}$. Put*

$$P = \max_{x\in\mathcal{X}} p(x), \qquad p = \min_{x\in\mathcal{X}} p(x),$$

and define

$$\Pi := \frac{P}{p} \geq 1.$$

If

$$\Pi \leq 1+\varepsilon+\sqrt{\varepsilon(\varepsilon+2)}, \qquad \varepsilon > 0$$

then

$$0 \leq \log|\mathcal{X}| - H(X) \leq \varepsilon.$$

3. An Upper Bound Using Kantorovič Inequality

In 1948, L.B. Kantorovič proved the following inequality for sequences of real numbers

$$\sum_{k=1}^{n} r_k u_k^2 \sum_{k=1}^{n} \frac{1}{r_k} u_k^2 \leq \frac{1}{4}\left(\sqrt{\frac{M}{m}}+\sqrt{\frac{m}{M}}\right)^2 \left(\sum_{k=1}^{n} u_k^2\right)^2 \tag{3.1}$$

where

$$0 < m \leq r_k \leq M \quad \text{for } k = 1, ..., n.$$

For other results of this type see for example the classical book in Theory of Inequalities by D.S. Mitrinović [9].

Using this result, we can provide the following upper bound for $D(p \parallel q)$.

Theorem 3. *Let $p(x), q(x) > 0, x \in \mathcal{X}$ be two probability mass functions satisfying the condition*

$$0 < r \leq \frac{p(x)}{q(x)} \leq R \quad \textit{for } x \in \mathcal{X}. \tag{3.2}$$

then we have the bound

$$D(p\|q) \leq \frac{(R-r)^2}{4rR}. \tag{3.3}$$

The equality holds if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof. Define $r(x) = p(x)/q(x), x \in \mathcal{X}$. Then by (3.2) we have that $r \leq r(x) \leq R, x \in \mathcal{X}$. Put also $u(x) = \sqrt{p(x)}$, $x \in \mathcal{X}$ and use (3.1) to get

$$\sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} \sum_{x\in\mathcal{X}} \frac{q(x)}{p(x)} p(x) \leq \frac{1}{4}\left(\sqrt{\frac{R}{r}}+\sqrt{\frac{r}{R}}\right)^2 \left(\sum_{x\in\mathcal{X}} p(x)\right)^2$$

which is clearly equivalent to

$$\sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} \leq \frac{1}{4}\left(\sqrt{\frac{R}{r}}+\sqrt{\frac{r}{R}}\right)^2. \tag{3.4}$$

Now, by (2.1) and (3.4) we can state

$$\begin{aligned} D(p\|q) &\leq \sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} - 1 \leq \frac{1}{4}\left(\sqrt{\frac{R}{r}}+\sqrt{\frac{r}{R}}\right)^2 - 1 \\ &= \frac{1}{4}\left(\sqrt{\frac{R}{r}}-\sqrt{\frac{r}{R}}\right)^2 \end{aligned}$$

and the inequality (3.3) is proved.

The case of equality holds in (3.3) from the fact that in the Kantorovič Inequality we have equality if and only if $r_k = 1$ for all k.

We omit the details. ∎

Remark 3. A similar result was obtained by M. Matić in his Ph.D. Thesis [7]. Note that Matić's proof used a Grüss type inequality for sequences of real numbers.

Corollary 6. *Let X, Y be two random variables so that $p(x), q(y), t(x,y) > 0$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, and put*

$$M = \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{t(x,y)}{p(x)q(y)}, \qquad m = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{t(x,y)}{p(x)q(y)}.$$

Then we have

$$0 \le I(X;Y) \le \frac{(M-m)^2}{4Mm}.$$

The equality holds if and only if X and Y are independent.

We can also state the following corollary.

Corollary 7. *Let X be a random variable whose range has $|\mathcal{X}|$ elements and $p(x) > 0$, $x \in \mathcal{X}$. Put $P = \max_{x\in\mathcal{X}} p(x)$ and $p = \min_{x\in\mathcal{X}} p(x)$. Then*

$$0 \le \log|\mathcal{X}| - H(X) \le \frac{(P-p)^2}{4Pp}.$$

The equality holds if and only if p is the uniform distribution.

The above Theorem 3 allows us to point out a sufficient condition for the probabilities p and q such that $D(p \parallel q) \le \varepsilon$, where ε is a given small number.

Theorem 4. *Let $p(x), q(x) > 0, x \in \mathcal{X}$ be two probability mass functions and define*

$$S := \frac{R}{r} \ge 1.$$

If $\varepsilon > 0$ and

$$S \le 1 + 2\varepsilon + 2\sqrt{\varepsilon(\varepsilon+1)}, \tag{3.5}$$

then

$$D(p \parallel q) \le \varepsilon. \tag{3.6}$$

Proof. Observe that for a given $\varepsilon > 0$, the inequality

$$\frac{(R-r)^2}{4rR} \le \varepsilon$$

is equivalent to

$$R^2 - 2(1+2\varepsilon)rR + r^2 \le 0$$

i.e.,

$$\left(\frac{R}{r}\right)^2 - 2(1+2\varepsilon)\frac{R}{r} + 1 \le 0$$

or

$$S^2 - 2(1+2\varepsilon)S + 1 \le 0$$

which is clearly equivalent to

$$S \le \left[1 + 2\varepsilon - 2\sqrt{\varepsilon(\varepsilon+1)}, 1 + 2\varepsilon + 2\sqrt{\varepsilon(\varepsilon+1)}\right]. \tag{3.7}$$

Furthermore, as $S \ge 1$ then (3.7) follows by (3.5) and then (3.5) implies that

$$\frac{(R-r)^2}{4rR} \le \varepsilon.$$

Using (3.7), we get the desired inequality (3.6). ∎

Remark 4. Considering the fact that the bound $1+2\varepsilon+2\sqrt{\varepsilon(\varepsilon+1)}$ provided by (3.5) is greater than the bound $1+\varepsilon+\sqrt{\varepsilon(\varepsilon+2)}$ provided by (2.2) for any $\varepsilon>0$, then Theorem 4 is a better result than Theorem 1. This fact illustrates the importance of the Kantorovič Inequality in Information Theory.

4. An Upper Bound Using Diaz-Metcalf Inequality

The following result is well known in the literature as Diaz-Metcalf Inequality for real numbers (see e.g. [9, p. 61]):

Theorem 5. *Let $p_k>0(k=1,...,n)$ with $\sum_{k=1}^{n}p_k=1$. If $a_k(\neq 0)$ and $b_k(k=1,...,n)$ are real numbers and if*

$$m\leq\frac{b_k}{a_k}\leq M\quad \text{for } k=1,...,n; \tag{4.1}$$

then

$$\sum_{k=1}^{n}p_kb_k^2+mM\sum_{k=1}^{n}p_ka_k^2\leq(M+m)\sum_{k=1}^{n}p_ka_kb_k. \tag{4.2}$$

Equality holds in (4.2) if and only if for each $k, 1\leq k\leq n$ either $b_k=ma_k$ or $b_k=Ma_k$.

Using this inequality, we can point out another bound for $D(p\parallel q)$ as follows.

Theorem 6. *Let $p(x),q(x)>0,x\in\mathcal{X}$ be two probability mass functions satisfying the condition:*

$$0<r\leq\frac{p(x)}{q(x)}\leq R\quad \text{for all } x\in\mathcal{X}.$$

Then we have the bound

$$D(p\parallel q)\leq(1-r)(R-1)\leq\frac{1}{4}(R-r)^2. \tag{4.3}$$

Proof. Define

$$b(x)=\sqrt{\frac{p(x)}{q(x)}},a(x)=\sqrt{\frac{q(x)}{p(x)}},x\in\mathcal{X}.$$

Then

$$\frac{b(x)}{a(x)}=\frac{p(x)}{q(x)}\in[r,R]\quad\text{for all } x\in\mathcal{X}$$

and applying the inequality (4.2), we get

$$\begin{aligned}&\sum_{x\in\mathcal{X}}p(x)\left(\sqrt{\frac{p(x)}{q(x)}}\right)^2+Rr\sum_{x\in\mathcal{X}}p(x)\left(\sqrt{\frac{q(x)}{p(x)}}\right)^2\\ \leq\ &(r+R)\sum_{x\in\mathcal{X}}\sqrt{\frac{p(x)}{q(x)}}\cdot\sqrt{\frac{q(x)}{p(x)}}p(x)\end{aligned}$$

i.e.,

$$\sum_{x\in\mathcal{X}}\frac{p^2(x)}{q(x)}+Rr\sum_{x\in\mathcal{X}}q(x)\leq(r+R)\sum_{x\in\mathcal{X}}p(x)$$

which is equivalent to

$$\sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} \leq r + R - rR.$$

Using Theorem 1, we get

$$\begin{aligned} D(p\|q) &\leq \sum_{x\in\mathcal{X}} \frac{p^2(x)}{q(x)} - 1 \leq r + R - rR - 1 \\ &= (1-r)(R-1) \end{aligned}$$

and the first part of inequality (4.3) is obtained. The last inequality is obvious. ∎

The following two corollaries are natural application.

Corollary 8. *Let X, Y be two random variables and have M, m defined as in Corollary 6. Then*

$$0 \leq I(X;Y) \leq (1-m)(M-1) \leq \frac{1}{4}(M-m)^2 .$$

Finally, we also have:

Corollary 9. *Let X be as in Corollary 7. Then*

$$(4.4) \qquad 0 \leq \log|\mathcal{X}| - H(X) \leq |\mathcal{X}|^2 \left(\frac{1}{|\mathcal{X}|} - p\right)\left(P - \frac{1}{|\mathcal{X}|}\right) \leq \frac{|\mathcal{X}|^2}{4}(P-p)^2 .$$

Proof. As $p \leq p(x) \leq P$ for all $x \in \mathcal{X}$, we get that $|\mathcal{X}|p \leq p(x)/(1/|\mathcal{X}|) \leq P|\mathcal{X}|$. Now, if we apply Theorem 6 for $R = P|\mathcal{X}|$, $r = p|\mathcal{X}|$ we get the desired inequality (4.4). ∎

Acknowledgement. The authors would like to thank the referee for his valuable comments in improving the paper.

References

[1] S. S. Dragomir and C. J. Goh, Some bounds on entropy measures in information theory, *Appl. Math. Lett.,* **10**(3) (1997) 23–28.

[2] S. S. Dragomir and C. J. Goh, A counterpart of Jensen's discrete inequality for differentiable convex mappings and applications in information theory, *Math. Comput. Modelling,* **24**(2) (1996) 1–11.

[3] M. Matić, C. E. M. Pearce and J. Pečarić, Improvements of some bounds on entropy measures in information theory, *Math. Inequal. and Appl.*, **1**(1998) 295–304.

[4] M. Matić, C. E. M. Pearce and J. Pečarić, Refinements of some bounds in information theory, *J. Austr. Math. Soc. Ser. B,* to appear.

[5] R. J. McEliece, *The Theory of Information and Coding,* Addison-Wesley, Reading, Mass. 1977.

[6] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[7] M. Matić, Jensen's Inequality and Applications in Information Theory (in *Croatian*), *Ph.D. Thesis*, University of Zagreb, Croatia.

[8] S.S. Dragomir and N.M. Dragomir, An inequality for logarithms and its application in coding theory, *Indian Journal of Mathematics, (in press).*

[9] D.S. Mitrinović, *Analytic Inequalities*, Springer Verlag, 1970.

Department of Computer and Mathematical Sciences, Victoria University of Technology, PO Box 14428, MCMC, Melbourne, Victoria 8001
E-mail address: sever@matilda.vut.edu.au

Communication Division, DSTO, Po Box 1500, Salisbury, SA 5108
E-mail address: Marcel.Scholz@dsto.defence.gov.au

Communication Division, DSTO, Po Box 1500, Salisbury, SA 5108
E-mail address: Jadranka.Sunde@dsto.defence.gov.au