Tamsui Oxford Journal of Mathematical Sciences **24(2)** (2008) 141-151
Aletheia University

Some Properties for the Exponential of the Kullback-Leibler Divergence*

Sergiu C. Dragomir†

Austral Internet Publishing, P. O. Box 156,
Endeavour Hills, 3802 Victoria, Australia

Received September 21, 2006, Accepted November 30, 2006.

Abstract

In this paper we have established for the Kullback-Leibler divergence $D(\cdot||\cdot)$ that the functional $exp[-D(p||.)]$ is *supperadditive*, preserves the bounds under some *likelihood ratio conditions* and is *concave* on the *convex cone* of all probability distributions of given length $n \geq 2$. Some lower bounds for $exp[D(p||H(q,r))]$, where $H(q,r)$ is the *harmonic mean* of the probability distributions q and r are also given.

1. Introduction

In Probability and Information Theory, the *Kullback–Leibler divergence* (or *information divergence*, or *information gain*, or *relative entropy*) is a natural distance measure from a "true" probability distribution p to an arbitrary probability distribution q. Typically p represents data, observations, or a precise calculated probability distribution. The measure q typically represents a theory, a model, a description or an approximation of p.

It can be interpreted as the expected extra message-length per datum that must be communicated if a code that is optimal for a given (wrong) distribution q is used, compared to using a code based on the true distribution p.

*2000 *Mathematics Subject Classification.* Primary 26D15.

†E-mail: sergiu.dragomir@ajmaa.org

The Kullback–Leibler divergence can also be interpreted as the expected discrimination information for H_1 over H_0: the mean information per sample for discriminating in favour of a hypothesis H_1 against a hypothesis H_0, when hypothesis H_1 is true.

In Bayesian Statistics the Kullback–Leibler divergence can be used as a measure of the information gain in moving from a prior distribution to a posterior distribution.

Originally introduced by Solomon Kullback and Richard Leibler in 1951, [7] as the directed divergence between two distributions, it is not the same as a divergence in calculus: the term "divergence" in the terminology should not be misinterpreted. One might be tempted to call it a "distance metric" on the space of probability distributions, but this would not be correct as the Kullback-Leibler divergence is not symmetric. Mistaking p for q is not the same as mistaking q for p. Moreover, $D(p||q)$ does not satisfy the triangle inequality (see http://en.wikipedia.org/wiki/Kullback-Leibler_divergence). To be more specific, let $p = (p_1, ..., p_n)$, $q = (q_1, ..., q_n)$ be two discrete probability distributions. Define the *Kullback-Leibler divergence* (see [7] or [3]) by

$$D(p||q) := \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right), \tag{1}$$

the χ^2-*distance* (see for example [3]) by

$$D_{\chi^2}(p||q) := \sum_{i=1}^{n} \frac{p_i^2 - q_i^2}{q_i} \tag{2}$$

and the *variational distance* (see for example [3]) by

$$V(p||q) := \sum_{i=1}^{n} |p_i - q_i|. \tag{3}$$

The following result is of fundamental importance in Information Theory [3, p. 26].

Under the above assumptions for p and q, we have (the Information Inequality, Gibbs' inequality):

$$D(p||q) \geq 0, \tag{4}$$

with equality iff $p_i = q_i$ for all $i \in \{1, ..., n\}$.

As a matter of fact, the inequality (4) can be improved as follows (see [3, p. 300]).

Let p, q be as above. Then

$$D(p||q) \geq \frac{1}{2}V^2(p||q)\,(\geq 0), \tag{5}$$

with equality iff $p_i = q_i$ for all $i \in \{1, ..., n\}$.

The following counterpart of (5) is also known

$$D_{\chi^2}(p||q) \geq D(p||q), \tag{6}$$

with equality iff $p_i = q_i$ for all $i \in \{1, ..., n\}$.

For various other bounds involving the Kullback–Leibler divergence see for instance [5], [6], [7], [8], [9] and the book [3].

The aim of the present note is to explore some properties for the exponential of the Kullback-Leibler divergence when the second probability q is replaced by either the convex combination of two probabilities, the sum of those probabilities or even the harmonic mean of them. As a consequence, we have established that the functional $exp[-D(p||.)]$ is *supperadditive*, preserve the bounds under some *likelihood ratio conditions* and it is *concave* on the *convex cone* of all probability distributions of given length $n \geq 2$. Some lower bounds for $exp[D(p||H(q,r))]$, where $H(q,r)$ denotes the *harmonic mean* of the probability distributions q and r are also given. Some numerical experiments for densities of length 2 which depict the behavior of the terms in the obtained inequalities are also provided.

2. Some Preliminary Results

For the n-tuples of nonnegative real numbers $a = (a_1, ..., a_n)$ and the probability distribution $p = (p_1, ..., p_n)$ we can consider the *weighted geometric mean* denoted by $G_n(p, a)$ and defined by the equation

$$G_n(p, a) := \prod_{i=1}^{n} a_i^{p_i}. \tag{7}$$

The weighted geometric mean has an important property as a function in the second variable a, namely $G_n(p, \cdot)$ is *superadditive*, which means that

$$G_n(p, a + b) \geq G_n(p, a) + G_n(p, b), \tag{8}$$

for any choice of the nonnegative n-tuples $a = (a_1, ..., a_n)$ and $b = (b_1, ..., b_n)$ and each probability distribution $p = (p_1, ..., p_n)$.

This is a well known fact and a proof based on the quasi-linearization method can be seen in [1, p. 214].

For the sake of completeness we point out here a simple proof that can be derived from the Jensen's inequality [4].

First, recall that if $f : \mathbb{R} \to \mathbb{R}$ is a convex function and $x = (x_1, ..., x_n)$ is an n-tuple of real numbers while $p = (p_1, ..., p_n)$ is a probability distribution, then [1, p. 30]

$$\sum_{i=1}^{n} p_i f(x_i) \geq f\left(\sum_{i=1}^{n} p_i x_i\right). \tag{9}$$

Further, if we consider the function $f(x) = \ln(1 + e^x)$, then [4]

$$f'(x) = \frac{e^x}{1 + e^x} \text{ and } f''(x) = \frac{e^x}{(1 + e^x)^2}, x \in \mathbb{R}$$

which, due to the fact that $f''(x) > 0$ for any $x \in \mathbb{R}$, shows that f is strictly convex where is defined.

Now if we apply Jensen's inequality to the function $f(x) = \ln(1 + e^x)$ and $x_i = \ln\left(\frac{a_i}{b_i}\right)$ (provided $b_i > 0$), $i = \{1, ...n\}$ then we can write the inequality:

$$\prod_{i=1}^{n} \left(1 + \frac{a_i}{bi}\right)^{p_i} \geq 1 + \prod_{i=1}^{n} \left(\frac{a_i}{b_i}\right)^{p_i},$$

which gives the desired superadditivity property (8).

A simple consequence that is worth mentioning is the following *monotonicity property* of the weighted geometric mean $G_n(p, \cdot)$:

$$G_n(p, b) \geq G_n(p, a) \tag{10}$$

provided $b \geq a$, which means that $b_i \geq a_i$ for each $i \in \{1, ..., n\}$. This fact follows from the superadditivity property on noticing that

$$\begin{aligned} G_n(p, b) &= G_n(p, b - a + a) \\ &\geq G_n(p, a) + G_n(p, b - a) \geq G_n(p, b - a) \geq 0 \end{aligned}$$

since $b - a$ is nonnegative and $G_n(p, b - a) \geq 0$.

3. The Superadditive Property of $\exp\left[-D\left(p\|\cdot\right)\right]$

We consider X, Y, Z three discrete random variables having the probability distributions $p = (p_1, \ldots, p_n)$, $q = (q_1, \ldots, q_n)$ and $r = (r_1, \ldots, r_n)$. If we write the weighted geometric mean of $\frac{q}{p} = \left(\frac{q_1}{p_1}, \ldots, \frac{q_n}{p_n}\right)$ with the weights $p = (p_1, \ldots, p_n), (p_i \neq 0, i \in \{1, \ldots, n\})$ we have

$$\begin{aligned} G_n\left(p, \frac{q}{p}\right) &= \prod_{i=1}^{n}\left(\frac{q_i}{p_i}\right)^{p_i} = \exp\left\{\ln\left[\prod_{i=1}^{n}\left(\frac{q_i}{p_i}\right)^{p_i}\right]\right\} \\ &= \exp\left[\sum_{i=1}^{n} p_i \ln\left(\frac{q_i}{p_i}\right)\right] = \exp\left[-D\left(p\|q\right)\right] \end{aligned}$$

and in a similar fashion

$$G_n\left(p, \frac{r}{p}\right) = \exp\left[-D\left(p\|r\right)\right].$$

Also, the weighted geometric mean with the weights $p = (p_1, ..., p_n)$ and the nonnegative sequence $\frac{q}{p} + \frac{r}{p} = \left(\frac{q_1+r_1}{p_1}, ..., \frac{q_n+r_n}{p_n}\right)$ gives

$$\begin{aligned} G_n\left(p, \frac{q+r}{p}\right) &= \prod_{i=1}^{n}\left(\frac{q_i+r_i}{p_i}\right)^{p_i} = \exp\left[\sum_{i=1}^{n} p_i \ln\left(\frac{q_i+r_i}{p_i}\right)\right] \\ &= \exp\left[-D\left(p\|q+r\right)\right]. \end{aligned}$$

Now, by the superadditivity of the weighted geometric mean we can conclude that

$$G_n = \left(p, \frac{q+r}{p}\right) \geq G_n\left(p, \frac{q}{p}\right) + G_n\left(p, \frac{r}{p}\right) \tag{11}$$

which shows that *the function* $\exp\left[-D\left(p\|\cdot\right)\right]$ *is superadditive* as claimed in the title of the section.

It is also a natural problem to ask how far the exponential quantities $\exp\left[-D\left(p\|q\right)\right]$ and $\exp\left[-D\left(p\|r\right)\right]$ are from each other when some bounds to the likelihood ratio $\frac{q_i}{r_i}, i \in \{1, ..., n\}$ are *a priory* known.

To be more specific, we assume that there exists the positive quantities m, M where $M > m$ and so that

$$0 < m \leq \frac{q_i}{r_i} \leq M \text{for all } i \in \{1, ..., n\}. \tag{12}$$

This condition obviously implies that $m < \frac{r_i}{p_i} \leq \frac{q_i}{p_i} \leq M\frac{r_i}{p_i}$ for each $i \in \{1, ..., n\}$. By utilising the motonicity properties of the geometric mean, we conclude that

$$G_n\left(p, \frac{mr}{p}\right) \leq G_n\left(p, \frac{q}{p}\right). \tag{13}$$

Since

$$\begin{aligned} G_n\left(p, \frac{mr}{p}\right) &= \prod_{i=1}^{n}\left(m\frac{r_i}{p_i}\right)^{p_i} = m\prod_{i=1}^{n}\left(\frac{r_i}{p_i}\right)^{p_i} \\ &= m\exp\left[-D\left(p||r\right)\right], \end{aligned}$$

hence by (13) we deduce the following bounds:

$$m\exp\left[-D\left(p||r\right)\right] \leq \exp\left[-D\left(p||q\right)\right] \leq M\exp\left[-D\left(p||r\right)\right] \tag{14}$$

provided the probability densities q and r satisfy the condition (12).

To investigate further the properties of the function $\exp\left[-D\left(p||\cdot\right)\right]$ we notice that if $\alpha, \beta \in [0,1]$ and $\alpha + \beta = 1$ and q, r are probability distributions, then the convex combination $\alpha q + \beta r$ is also a probability distribution and it is natural then to ask how the value $\exp\left[-D\left(p||\left(\alpha r + \beta q\right)\right)\right]$ relates to the original values $\exp\left[-D\left(p||r\right)\right]$ and $\exp\left[-D\left(p||q\right)\right]$.

Utilising the superadditivity properties of the geometrical mean we have:

$$\begin{aligned} \exp\left[-D\left(p||\left(\alpha r + \beta q\right)\right)\right] &= G_n\left(p, \frac{\alpha r + \beta q}{p}\right) \\ &\geq G_n\left(p, \alpha\frac{r}{p}\right) + G_n\left(p, \beta\frac{q}{p}\right) \\ &= \alpha G_n\left(p, \frac{r}{p}\right) + \beta G_n\left(p, \frac{q}{p}\right) \\ &= \alpha\exp\left[-D\left(p||r\right)\right] + \beta\exp\left[-D\left(p||q\right)\right] \end{aligned}$$

showing that *the function* $\exp\left[-D\left(p||\cdot\right)\right]$ *is concave* on the convex cone of all probability distributions of given length n $(n \geq 2)$.

4. Other Properties

It is obvious that different choices for the nonnegative n-tuple in the superadditivity property of the weighted geometric mean $G_n(p, a)$ would provide other inequalities for the Kullback-Leibler divergence measure $D(p||q)$. In this section we establish such a result where in the second variable of $D(\cdot||\cdot)$ the harmonic mean $\frac{2qr}{q+r}$ of the two distributions q and r is considered.

For this purpose, we observe that for $a_i = \frac{p_i}{2q_i}, b_i = \frac{p_i}{2r_i}$ we have:

$$\begin{aligned} G_n\left(p, \frac{p}{2q}\right) &= \prod_{i=1}^{n}\left(\frac{p_i}{2q_i}\right)^{p_i} = \exp\left\{\ln\left[\prod_{i=1}^{n}\left(\frac{p_i}{2q_i}\right)^{p_i}\right]\right\} \\ &= \exp\left[\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{2q_i}\right)\right] \\ &= \exp\left[\ln\frac{1}{2} + D(p||q)\right] = \frac{1}{2}\exp\left[D(p||q)\right] \end{aligned} \tag{15}$$

and, similarly,

$$G_n\left(p, \frac{p}{2r}\right) = \frac{1}{2}\exp\left[D(p||r)\right]. \tag{16}$$

Also

$$\begin{aligned} G_n\left(p, \frac{p}{2q} + \frac{p}{2r}\right) &= \prod_{i=1}^{n}\left(\frac{p_i}{2q_i} + \frac{p_i}{2r_i}\right)^{p_i} \\ &= \exp\left\{\ln\left[\prod_{i=1}^{n}\left(\frac{p_i}{2q_i} + \frac{p_i}{2r_i}\right)^{p_i}\right]\right\} \\ &= \exp\left\{\sum_{i=1}^{n} p_i \ln\left[p_i\left(\frac{q_i + r_i}{2q_i r_i}\right)\right]\right\} \\ &= \exp\left[\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{\frac{2q_i r_i}{q_i + r_i}}\right)\right] \\ &= \exp D\left(p||\frac{2qr}{q+r}\right). \end{aligned} \tag{17}$$

Therefore, by $(15)-(17)$ and the superadditive properties of the geometric mean we have the following inequality:

$$\exp D\left(p||\frac{2qr}{q+r}\right) \geq \frac{1}{2}\left[\exp D(p||q) + \exp D(p||r)\right]. \tag{18}$$

If by $H(q,r)$ we denote the harmonic mean of the distribution q,r, i.e., $H(q,r)=\frac{2qr}{q+r}$ and by $A(x,y)$ we denote the arithmetic mean of the nonnegative quantities x and y, then (18) can be stated as:

$$\exp D(p||H(q,r)) \geq A[\exp D(p||q), \exp D(p||r)]. \tag{19}$$

Finally, since always the arithmetic mean of two positive quantities is greater than the harmonic mean of the same two quantities, we deduce from (19) the following result as well

$$\exp D(p||H(q,r)) \geq H(\exp D(p||q), \exp D(p||r)). \tag{20}$$

5. Some numerical experiments

Consider the probability distributions $p=(x,1-x)$, $q=(y,1-y)$ and $r=(z,1-z)$ where $x,y,z\in(0,1)$. Then

$$exp[-D(p||q+r)] = \left(\frac{y+z}{x}\right)^{x}\cdot\left(\frac{2-y-z}{1-x}\right)^{1-x},$$

$$exp[-D(p||q)] = \left(\frac{y}{x}\right)^{x}\cdot\left(\frac{1-y}{1-x}\right)^{1-x}$$

and

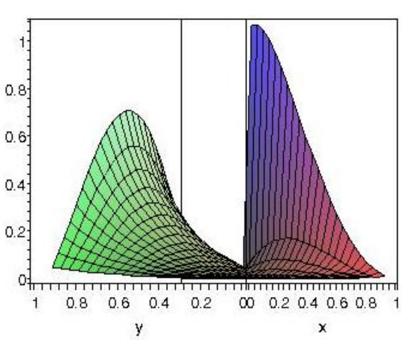

Figure 1: The plot of $\delta(\cdot,\cdot,z)$ for $z=1/4$.

$$exp\left[-D\left(p||r\right)\right]=\left(\frac{z}{x}\right)^{x}\cdot\left(\frac{1-z}{1-x}\right)^{1-x}$$

where $x,y,z\in(0,1)$.

Utilising the fact that the mapping $exp\left[-D\left(p||\cdot\right)\right]$ is superadditive, we have that

$$\Delta\left(x,y,z\right):=exp\left[-D\left(p||q+r\right)\right]-exp\left[-D\left(p||q\right)\right]-exp\left[-D\left(p||r\right)\right]\geq 0$$

for any $x,y,z\in(0,1)$. The plot depicted in Figure 1 show the behavior of $\Delta\left(\cdot,\cdot,z\right)$ for the value of $z=1/4$, in the box $(0,1)\times(0,1)$.

We also have that

$$expD\left(p||\frac{2qr}{q+r}\right)=\left[\frac{x\left(y+z\right)}{2\left(yz\right)}\right]^{x}\cdot\left[\frac{\left(1-x\right)\left(2-y-z\right)}{2\left(1-y\right)\left(1-z\right)}\right]^{1-x},$$

$$\frac{1}{2}expD\left(p||q\right)=\left(\frac{x}{2y}\right)^{x}\cdot\left[\frac{1-x}{2\left(1-y\right)}\right]^{1-x},$$

and

$$\frac{1}{2}expD\left(p||r\right)=\left(\frac{x}{2z}\right)^{x}\cdot\left[\frac{1-x}{2\left(1-z\right)}\right]^{1-x}.$$

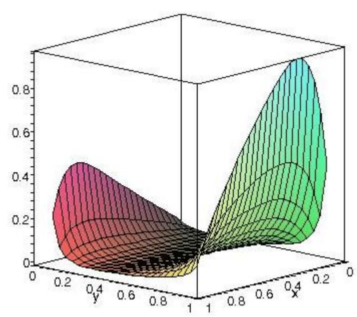

Figure 2: The plot of $\gamma\left(\cdot,\cdot,z\right)$ for $z=1/3$.

By (18) we know that the function

$$\gamma\left(x,y,z\right):=expD\left(p||\frac{2qr}{q+r}\right)-\frac{1}{2}expD\left(p||q\right)-\frac{1}{2}expD\left(p||r\right)$$

is nonnegative for any $x, y, z \in (0,1)$.

The plot depicted in Figure 2 shows the behavior of $\gamma(\cdot,\cdot,z)$ for the value $z = 1/3$ in the box $(0,1) \times (0,1)$.

References

[1] P. S. Bullen, *Handbook of Means and Their Inequalities*, Kluwer Academic Publishers, Dordrecht/Boston/London, 2003.

[2] A. Charnes, W. W. Cooper and J. Tyssedal, Khinčin-Kullback-Leibler estimation with inequality constraints. *Math. Operationsforsch. Statist. Ser. Optim.* **14** (1983), no. 3, 377–380.

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

[4] S. S. Dragomir, D. Comanescu and C. E. M. Pearce, On some mappings associated with geometric and arithmetic means, *Bull.Austral. Math. Soc.* **55**(1997), 299-309.

[5] A. A. Fedotov, P. Harremoës and F. Topsøe, Refinements of Pinsker's inequality. *IEEE Trans. Inform. Theory* **49** (2003), no. 6, 1491–1498.

[6] J. N. Kapur, Some inequalities involving $\ln x$ and their applications to information theory. *Bull. Math. Assoc. India* **21** (1989), no. 1-4, 55–68.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, **22**(1):79–86, March 1951.

[8] I. J. Taneja, Generalized relative information and information inequalities. *J. Inequal. Pure Appl. Math.* **5** (2004), No. 1, Article 21, 19 pp. (electronic).

[9] I. J. Taneja and P. Kumar, Relative information of type s, Csiszár's f-divergence, and information inequalities. *Inform. Sci.* **166** (2004), No. 1-4.

[10] W. H. Wong and X. Shen, Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** (1995), no. 2, 339–362.