APPLICATIONES MATHEMATICAE
24,4 (1997), pp. 415–423

A. L. RUKHIN (Baltimore)

INFORMATION-TYPE DIVERGENCE WHEN THE LIKELIHOOD RATIOS ARE BOUNDED

Abstract. The so-called ϕ-divergence is an important characteristic describing "dissimilarity" of two probability distributions. Many traditional measures of separation used in mathematical statistics and information theory, some of which are mentioned in the note, correspond to particular choices of this divergence. An upper bound on a ϕ-divergence between two probability distributions is derived when the likelihood ratio is bounded. The usefulness of this sharp bound is illustrated by several examples of familiar ϕ-divergences. An extension of this inequality to ϕ-divergences between a finite number of probability distributions with pairwise bounded likelihood ratios is also given.

1. Information-type divergences. Let ϕ be a convex function defined on the positive half-line, and let F and G be two different probability distributions such that F is absolutely continuous with respect to G. The *ϕ-divergence* between F and G is defined as

$$\phi(F|G) = \int \phi\left(\frac{dF}{dG}\right) dG = E_G \phi\left(\frac{dF}{dG}\right)$$

(see for example, Vajda, 1989). Clearly

$$\phi(1) = \phi(F|F) \le \phi(F|G).$$

This inequality and the fact that many familiar separation characteristics used in mathematical statistics and information theory correspond to particular choices of ϕ justify the interest in ϕ-divergences.

Out of these choices perhaps the most important is

$$\phi_I(u) = -\log u + u - 1,$$

1991 *Mathematics Subject Classification*: 60E15, 94A17.
Key words and phrases: convexity, information measures, likelihood ratio, multiple decisions.

in which case

$$\phi_I(F|G) = E^G \log\left(\frac{dF}{dG}\right) = K(G,F)$$

is the classical information number. Another information number $K(F,G)$ corresponds to the function $\phi(u) = u\log u - u + 1$, and the sum of these information numbers (the so-called J-divergence, see Cover and Thomas, 1991) is determined by $\phi_J(u) = (u-1)\log u$.

The probability of correct discrimination between F and G in the Bayesian setting is another example of ϕ-divergence. Indeed, let λ be the prior probability of distribution F, so that $1-\lambda$ is the prior probability of G. Then the probability of the correct decision is

$$\lambda \int\limits_{\lambda dF \geq (1-\lambda)dG} dF + (1-\lambda) \int\limits_{\lambda dF < (1-\lambda)dG} dG$$
$$= \int \max[\lambda dF, (1-\lambda)dG] = \phi_C(F|G),$$

which is another version of ϕ-divergence with $\phi_C(u) = \max[\lambda u, 1-\lambda]$.

A further classical example of ϕ-divergence is provided by χ^2-separation with $\phi(u) = (u-1)^2$, or by more general functions of the form

$$\phi_r(u) = \begin{cases} |1-u^r|^{1/r}, & 0 < r < 1, \\ |1-u|^r, & r \geq 1. \end{cases}$$

For a fixed number $w, 0 < w < 1$, the ϕ-divergence with $\phi(u) = -u/(wu + 1 - w)$ or, somewhat more conveniently, with

$$\phi_M(u) = u\left[1 - w - \frac{1}{wu + 1 - w}\right], \qquad u > 0,$$

appears in the statistical estimation problems of the mixture parameter and of the change-point parameter (Rukhin, 1996).

In this note the interest is in obtaining an upper bound on a ϕ-divergence when the likelihood ratio, dF/dG, is bounded. Intuitively it is clear that the closer the probability distributions F and G are to each other, the smaller any ϕ-divergence must be. This intuition is confirmed by the inequality (2) in the next section.

One of the motivations for the study of the bounded likelihood ratios family is statistical inference with finite memory (see Cover, Freedman and Hellman, 1976) or recurrent multiple decision-making (Rukhin, 1994). In the latter problem a recursive procedure can be consistent only if the distribution of the likelihood ratio is supported by the whole positive half-line. It is demonstrated by Rukhin (1993) that in the bounded likelihood ratio situation the probability of the correct decision is bounded from above by an explicitly given constant, which is strictly smaller than one. Theorem 2.1 generalizes this result.

Another reason for interest in distributions with bounded likelihood ratio is importance sampling in Monte-Carlo methods (see Fishman (1996), Sec. 4.1). This technique, designed to reduce the variance of an estimate of an integral, replaces sampling from the distribution F by sampling from a suitably chosen G under condition (1). Similar situation appears in the rejection method of generating of non-uniform random variables (cf. Devroy (1986), II.3). The inequality (2) gives a bound on possible gain (or loss) obtained from such a replacement.

2. A bound for ϕ-divergence. Suppose that with G-probability one

$$b_{\min} \le \frac{dF}{dG} \le b_{\max}. \tag{1}$$

Then $b_{\min} < 1 < b_{\max}$.

Notice that all functions ϕ considered above have minimum at $u = 1$ and that they are bowl-shaped, i.e. are non-increasing in the interval $(0, 1)$ and are non-decreasing for $u > 1$. Only this condition is needed in the following theorem.

THEOREM 2.1. *Assume that the function ϕ is bowl-shaped with the minimum at $u = 1$. Under the condition* (1),

$$\phi(F|G) \le \frac{b_{\max} - 1}{b_{\max} - b_{\min}}\phi(b_{\min}) + \frac{1 - b_{\min}}{b_{\max} - b_{\min}}\phi(b_{\max}). \tag{2}$$

P r o o f. Let

$$A_1 = \left\{u : \frac{dF}{dG}(u) = b_{\max}\right\} \quad \text{and} \quad A_2 = \left\{u : \frac{dF}{dG}(u) = b_{\min}\right\}.$$

If the set $(A_1 \cup A_2)^{\mathrm{c}}$ is not empty, the value of $\int \phi(dF/dG)\, dG$, for fixed distribution G, can get only larger by the inclusion of the points of this set either in A_1 or in A_2. Thus for any F, under condition (1),

$$\begin{aligned}\phi(F|G) &\le \int\limits_{A_1} \phi\left(\frac{dF}{dG}\right) dG + \int\limits_{A_2} \phi\left(\frac{dF}{dG}\right) dG \\ &= \phi(b_{\max})G(A_1) + \phi(b_{\min})G(A_2).\end{aligned}$$

Since

$$F(A_1) = b_{\max}G(A_1) \quad \text{and} \quad F(A_2) = b_{\min}G(A_2),$$

one obtains

$$G(A_1) = \frac{1 - b_{\min}}{b_{\max} - b_{\min}} \quad \text{and} \quad G(A_2) = \frac{b_{\max} - 1}{b_{\max} - b_{\min}},$$

which proves (2). ∎

Let us illustrate this theorem by the particular versions of ϕ from Section 1.

1. For $\phi_I(u) = -\log u + u - 1$, Theorem 2.1 shows that

$$K(G,F) \le -\frac{(1-b_{\min})\log b_{\max} + (b_{\max}-1)\log b_{\min}}{b_{\max}-b_{\min}}.$$

Similarly,

$$K(G,F) + K(F,G) \le \frac{(b_{\max}-1)(1-b_{\min})}{b_{\max}-b_{\min}} \log \frac{b_{\max}}{b_{\min}}.$$

2. The function $\phi_C(u) = \max[\lambda u, 1-\lambda]$ has a (non-unique) minimum at $u = 1$ if $\lambda \le 1/2$. The inequality (2) shows that in this case

$$\phi_C(F|G) \le \frac{(1-b_{\min})\max[\lambda b_{\max}, 1-\lambda] + (b_{\max}-1)(1-\lambda)}{b_{\max}-b_{\min}},$$

which is equivalent to the inequality (3.3) in Rukhin (1993).

3. For $\phi_2(u) = (u-1)^2$, one concludes from Theorem 2.1 that

$$E_G\left(\frac{dF}{dG}\right)^2 \le 1 + (b_{\max}-1)(1-b_{\min}). \tag{3}$$

For two discrete distributions with probabilities $p_1, \dots, p_n$ and $q_1, \dots, q_n$ such that $b_{\min} \le p_i/q_i \le b_{\max}$, this inequality means that

$$\sum \frac{p_i^2}{q_i} \le 2 + (b_{\max}-1)(1-b_{\min}).$$

For arbitrary non-negative numbers $\alpha_1, \dots, \alpha_n$ and $\beta_1, \dots, \beta_n$ put $q_i = \beta_i^2 / \sum \beta_k^2$, and $p_i = \alpha_i\beta_i / \sum_k \alpha_k\beta_k$. Then

$$\frac{\sum \alpha_i^2 \sum \beta_i^2}{(\sum \alpha_i\beta_i)^2} \le 2 + (b_{\max}-1)(1-b_{\min}),$$

where

$$\beta_{\max} = \max_i \frac{\alpha_i}{\beta_i} \cdot \frac{\sum \beta_i^2}{\sum \alpha_i\beta_i}, \qquad \beta_{\min} = \min_i \frac{\alpha_i}{\beta_i} \cdot \frac{\sum \beta_i^2}{\sum \alpha_i\beta_i}.$$

By maximizing the right-hand side of (3) when $b_{\max}/b_{\min} = B$, one obtains

$$E_G\left(\frac{dF}{dG}\right)^2 \le \frac{(B+1)^2}{4B}. \tag{4}$$

For discrete distributions, as above, this inequality reduces to a well known inequality

$$\frac{\sum \alpha_i^2 \sum \beta_i^2}{\left(\sum \alpha_i\beta_i\right)^2} \le \frac{(B+1)^2}{4B}$$

with $B = \max_i(\alpha_i/\beta_i) / \min_i(\alpha_i/\beta_i)$ (see Pólya and Szegő, 1972).

The latter inequality has been used by Tukey (1948) and Bloch and Moses (1988) in the problem of statistical estimation of the common mean by weighted means statistics with measurements of different precision. Both of these papers comment on the numerical accuracy of the bound (4) (which is weaker than (2)).

4. For ϕ_M, Theorem 2.1 implies

$$\int \frac{dF\,dG}{wdF + (1-w)dG} \geq \frac{1 - w + wb_{\min}b_{\max}}{(wb_{\min} + 1 - w)(wb_{\max} + 1 - w)}.$$

The example of two Bernoulli distributions with probabilities of success $(1-b_{\min})/(b_{\max}-b_{\min})$ and $b_{\max}(1-b_{\min})/(b_{\max}-b_{\min})$, respectively, shows that the inequality (2) is sharp. Its sharpness can also be seen by the limiting cases when $w = 0$ or $w = 1$.

As another example, let F be the exponential distribution with mean ω and G be the exponential distribution with mean 1. Then

$$\frac{dF}{dG}(x) = \omega \exp\{(1-\omega)x\}, \qquad x > 0,$$

so that for $\omega > 1$, $b_{\max} = \omega$ and $b_{\min} = 0$. Therefore for any bowl-shaped function ϕ with minimum at $u = 1$, for $\omega > 1$ we have

$$\phi(F|G) = \int_0^1 \phi(\omega u^{\omega-1})\,du \leq \frac{(\omega-1)\phi(0) + \phi(\omega)}{\omega}.$$

When $\omega \downarrow 1$, this inequality reduces to equality.

3. Information divergence for several probability distributions. In this section we derive an inequality similar to the one in Theorem 2.1 for the information divergence between several probability distributions. This divergence is defined in the following way (see Györfi and Nemetz, 1975).

Let $\phi(u_1, \ldots, u_m)$ be a non-negative convex function defined over the positive quadrant of m-dimensional Euclidean space. Assume that ϕ is a homogeneous function, i.e. for all positive u,

$$\phi(uu_1, \ldots, uu_m) = u\phi(u_1, \ldots, u_m).$$

Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space, and let different probability distributions $P_1, \ldots, P_m$ defined on $\mathcal{A}$ be absolutely continuous with respect to μ. The ϕ-divergence between $P_1, \ldots, P_m$ is defined as

$$\phi(P_1, \ldots, P_m) = \int_{\mathcal{X}} \phi\left(\frac{dP_1}{d\mu}, \ldots, \frac{dP_m}{d\mu}\right) d\mu.$$

The homogeneity property of ϕ guarantees independence of $\phi(P_1, \ldots, P_m)$ from the dominating measure μ. When $m = 2$, this information divergence reduces to the one in Section 1 with the function $\phi(u)$ there equal to $\phi(u, 1)$.

The examples of ϕ-divergence include the error probability in a multiple decision problem for $\phi_C(u_1,\dots,u_m)=\max_i[\lambda_i u_i]$ with probabilities $\lambda_1,\dots,\lambda_m$; the analogues of Kullback–Leibler divergences,

$$\phi_I(u_1,\dots,u_m)=\sum_{i,k}\left[u_i-u_k-u_k\log\frac{u_i}{u_k}\right],$$

$$\phi_J(u_1,\dots,u_m)=\sum_{i,k}(u_i-u_k)\log\frac{u_i}{u_k};$$

and Hellinger-type transforms with $\phi(u_1,\dots,u_m)=u_1^{\alpha_1}u_2^{\alpha_2}\dots u_m^{\alpha_m}$ for $\alpha_1+\dots+\alpha_m=1$.

Assume now that the ratios of the densities $p_i=dP_i/d\mu$, $i=1,\dots,m$, are bounded, i.e.

$$b_{ki}\le\frac{p_k(x)}{p_i(x)}\le\frac{1}{b_{ik}}\qquad\mu\text{-a.s.} \tag{5}$$

Moreover, assume that b_{ik} are the largest (positive) quantities satisfying (5). Then $b_{ki}b_{il}<b_{kl}$ for $i\ne k,l$. In particular, $b_{ki}b_{ik}<1$ for $i\ne k$. The set $\mathcal{P}$ of all probability distributions satisfying this condition is convex and closed under weak convergence. Since the functional $\phi(P_1,\dots,P_m)$ is convex, its maximum is attained on the set $\mathbf{ext}(\mathcal{P})$ of the extreme points of $\mathcal{P}$.

The next result gives a necessary condition for $(P_1^0,\dots,P_m^0)$ to belong to $\mathbf{ext}(\mathcal{P})$.

Proposition 3.1. *If $(P_1^0,\dots,P_m^0)$ is an extreme point of $\mathcal{P}$, then for any k,*

$$\mu\left\{\max_{i:i\ne k}b_{ki}p_i^0(x)<p_k^0(x)<\min_{i:i\ne k}\frac{p_i^0(x)}{b_{ik}}\right\}=0. \tag{6}$$

Proof. We show first of all that the conditions

$$\frac{p_i(x)}{b_{ik}}=\min_{l:l\ne k}\frac{p_l(x)}{b_{lk}}$$

and

$$b_{ki}p_i(x)=\max_{l:l\ne k}b_{kl}p_l(x)$$

are equivalent. Indeed, according to the first condition, for any $l\ne k$,

$$b_{kl}p_l(x)\ge\frac{b_{kl}b_{lk}p_i(x)}{b_{ik}},$$

so that

$$\max_{l:l\ne k}b_{kl}p_l(x)\ge\max_{l:l\ne k}b_{kl}b_{lk}\frac{p_i(x)}{b_{ik}}\ge\max_{l:l\ne k}b_{kl}b_{lk}\max_{l:l\ne k}\frac{b_{kl}p_l(x)}{b_{ki}b_{ik}}.$$

It follows that

$$\max_{l:l\neq k} b_{kl}b_{lk} = b_{ki}b_{ik} \tag{7}$$

and that

$$\max_{l:l\neq k} b_{kl}p_l(x) = b_{ki}p_i(x).$$

Suppose now that for some $i \neq k$, (6) does not hold for $(P_1, \dots, P_m) \in \mathcal{P}$, i.e. on a set of μ-positive measure,

$$\max_{l:l\neq k} b_{kl}p_l(x) = b_{ki}p_i(x) < p_k(x) < \frac{p_i(x)}{b_{ik}} = \min_{l:l\neq k} \frac{p_l(x)}{b_{lk}}. \tag{8}$$

Then for sufficiently small positive w, the μ-measure of the set

$$b_{ki} + w\left(\frac{1}{b_{ik}} - b_{ki}\right) \leq \frac{p_k(x)}{p_i(x)} \leq \frac{1}{b_{ik}} - w\left(\frac{1}{b_{ik}} - b_{ki}\right)$$

is positive. For any number a such that $b_{ki} < a < 1/b_{ik}$, this set is contained in the region

$$C = \left\{b_{ki} + w(a - b_{ki}) \leq \frac{p_k(x)}{p_i(x)} \leq \frac{1}{b_{ik}} - w\left(\frac{1}{b_{ik}} - a\right)\right\},$$

With $a = P_k(C)/P_i(C)$, the set C must have μ-positive measure.

For $x \in C$ put

$$r(x) = \frac{p_k(x) - wap_i(x)}{1 - w}, \qquad q(x) = ap_i(x),$$

and for $x \notin C$,

$$r(x) = q(x) = p_k(x).$$

Then for all x,

$$p_k(x) = wq(x) + (1 - w)r(x),$$

and q and r are probability densities. We now show that $(P_1, \dots, Q, \dots, P_m) \in \mathcal{P}$ and $(P_1, \dots, R, \dots, P_m) \in \mathcal{P}$. Indeed, for $x \in C$,

$$b_{ki} \leq \frac{q(x)}{p_i(x)}a \leq \frac{1}{b_{ik}},$$

and these inequalities trivially hold for $x \notin C$. Also,

$$b_{ki} \leq \frac{r(x)}{p_i(x)} = \frac{\frac{p_k(x)}{p_i(x)} - wa}{1 - w} \leq \frac{1}{b_{ik}}$$

for $x \in C$, by the definition of C. Because of (8), for any $l \neq k$,

$$b_{kl} \leq \frac{r(x) \wedge q(x)}{p_l(x)} \leq \frac{r(x) \vee q(x)}{p_l(x)} \leq \frac{1}{b_{lk}}.$$

Therefore $(P_1, \dots, P_m) \notin \mathbf{ext}(\mathcal{P})$, which concludes the proof. ∎

According to this proposition, if $(P_1^0, \dots, P_m^0) \in \mathbf{ext}(\mathcal{P})$, then for any k there exists i, $i \neq k$, which can be found from (7), such that the sets

$$A_k^+ = \left\{ p_k^0(x) = \frac{p_i^0(x)}{b_{ik}} = \min_{l:l\neq k} \frac{p_l^0(x)}{b_{lk}} \right\}$$

and

$$A_k^- = \{p_k^0(x) = p_i^0(x) b_{ki} = \max_{l:l\neq k} p_l^0(x) b_{kl}\}$$

form a partition of $\mathcal{X}$. Clearly

$$P_k^0(A_k^+) = \frac{P_i^0(A_k^+)}{b_{ik}} \le \min_{l:l\neq k} \frac{P_l^0(A_k^+)}{b_{lk}},$$
$$P_k^0(A_k^-) = P_i^0(A_k^-) b_{ki} \ge \max_{l:l\neq k} P_l^0(A_k^-) b_{kl}.$$

As in Section 2,

$$P_k^0(A_k^+) = \frac{1 - b_{ik}}{1 - b_{ik} b_{ki}}, \qquad P_k^0(A_k^-) = \frac{b_{ki}(1 - b_{ik})}{1 - b_{ik} b_{ki}}.$$

If $\phi(u_1, \dots, u_m)$ attains its minimum at $(1, \dots, 1)$ then

$$\begin{aligned}
(9)\quad \phi(P_1, \dots, P_m) &\le \max_{(P_1^0,\dots,P_m^0)\in \mathbf{ext}(\mathcal{P})} \phi(P_1^0, \dots, P_m^0) \\
&= \max_{(P_1^0,\dots,P_m^0)\in \mathbf{ext}(\mathcal{P})} \int_{\mathcal{X}} \phi\left(p_1^0, \dots, p_m^0\right) \, d\mu \\
&\le \max_k \max_{(P_1^0,\dots,P_m^0)\in \mathbf{ext}(\mathcal{P})} \Big[\int_{A_k^+} \phi(p_1^0, \dots, p_m^0) \, d\mu \\
&\quad + \int_{A_k^-} \phi(p_1^0, \dots, p_m^0) \, d\mu \Big] \\
&\le \max_k \max_{(P_1^0,\dots,P_m^0)\in \mathbf{ext}(\mathcal{P})} \Big[\phi(b_{1k}, \dots, 1, \dots, b_{mk}) P_k^0(A_k^+) \\
&\quad + \phi\left(\frac{1}{b_{k1}}, \dots, 1, \dots, \frac{1}{b_{km}}\right) P_k^0(A_k^-) \Big] \\
&= \max_k \Big[\phi(b_{1k}, \dots, 1, \dots, b_{mk}) \frac{1 - b_{ik}}{1 - b_{ik} b_{ki}} \\
&\quad + \phi\left(\frac{1}{b_{k1}}, \dots, 1, \dots, \frac{1}{b_{km}}\right) \frac{b_{ki}(1 - b_{ik})}{1 - b_{ik} b_{ki}} \Big] \\
&\le \max_{k\neq l} \Big[\phi\Big(b_{1k}, \dots, 1, \dots, b_{mk}\Big) \frac{1 - b_{lk}}{1 - b_{lk} b_{kl}} \\
&\quad + \phi\left(\frac{1}{b_{k1}}, \dots, 1, \dots, \frac{1}{b_{km}}\right) \frac{b_{kl}(1 - b_{kl})}{1 - b_{lk} b_{kl}} \Big].
\end{aligned}$$

We formulate the obtained result.

THEOREM 3.2. *Under the boundedness condition* (5), *the inequality* (9) *holds for any information divergence* $\phi(P_1, \dots, P_m)$ *such that the convex function* $\phi(u_1, \dots, u_m)$ *attains its minimum at* $(1, \dots, 1)$.

It is easy to see that for convex functions ϕ the inequality (9) implies that of Theorem 2.1.

References

[1] D. A. Bloch and L. E. Moses, *Nonoptimally weighted least squares*, Amer. Statist. 42 (1988), 50–53.

[2] T. M. Cover, M. A. Freedman and M. E. Hellman, *Optimal finite memory learning algorithms for the finite sample problem*, Information Control 30 (1976), 49–85.

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

[4] L. Devroy, *Non-Uniform Random Variate Generation*, Springer, New York, 1986.

[5] G. S. Fishman, *Monte Carlo*: *Concepts*, *Algorithms and Applications*, Springer, New York, 1996.

[6] L. Györfi and T. Nemetz, *f-dissimilarity*: *A generalization of the affinity of several distributions*, Ann. Inst. Statist. Math. 30 (1978), 105–113.

[7] G. Pólya and G. Szegő, *Problems and Theorems in Analysis. Volume 1*: *Series*, *Integral Calculus*, *Theory of Functions*, Springer, New York, 1972.

[8] A. L. Rukhin, *Lower bound on the error probability for families with bounded likelihood ratios*, Proc. Amer. Math. Soc. 119 (1993), 1307–1314.

[9] —, *Recursive testing of multiple hypotheses*: *Consistency and efficiency of the Bayes rule*, Ann. Statist. 22 (1994), 616–633.

[10] —, *Change-point estimation*: *linear statistics and asymptotic Bayes risk*, Math. Methods Statist. 5 (1996), 412–431.

[11] J. W. Tukey, *Approximate weights*, Ann. Math. Statist. 19 (1948), 91–92.

[12] I. Vajda, *Theory of Statistical Inference and Information*, Kluwer, Dordrecht, 1989.

Andrew L. Rukhin
Department of Mathematics and Statistics
University of Maryland at Baltimore County
1000 Hilltop Circle
Baltimore, Maryland 21250
U.S.A.
E-mail: rukhin@math.umbc.edu

Received on 16.9.1996;
revised version on 10.12.1996