KYBERNETIKA — VOLUME 45 (2009), NUMBER 6, PAGES 885 – 900

# ON METRIC DIVERGENCES OF PROBABILITY MEASURES

Igor Vajda

Standard properties of $\phi$-divergences of probability measures are widely applied in various areas of information processing. Among the desirable supplementary properties facilitating employment of mathematical methods is the metricity of $\phi$-divergences, or the metricity of their powers. This paper extends the previously known family of $\phi$-divergences with these properties. The extension consists of a continuum of $\phi$-divergences which are squared metric distances and which are mostly new but include also some classical cases like e. g. the Le Cam squared distance. The paper establishes also basic properties of the $\phi$-divergences from the extended class including the range of values and the upper and lower bounds attained under fixed total variation.

*Keywords:* Total variation, Hellinger divergence, Le Cam divergence, Information divergence, Jensen–Shannon divergence, Metric divergences

*AMS Subject Classification:* 94A17, 62B10, 62H30, 68T10

## 1. INTRODUCTION

Let us consider divergences $D(P,Q)$ of probability measures $P,Q$ from the general space $\mathbb{P}$ of all probability measures on a measurable space $(\mathcal{X},\mathcal{A})$. For the obvious mathematical reasons, the most interesting are the metric divergences satisfying the conditions of *reflexivity*

$$D(P,Q) \geq 0 \quad \text{for all } P,Q \in \mathbb{P} \tag{1}$$

with the equality if and only if $P = Q$, *symmetry*

$$D(P,Q) = D(Q,P) \quad \text{for all } P,Q \in \mathbb{P} \tag{2}$$

and *triangle inequality*

$$D(P,Q) \leq D(P,R) + D(R,Q) \quad \text{for all } P,Q,R \in \mathbb{P}. \tag{3}$$

This paper is restricted to the narrower class of metric divergences of the form

$$D(P,Q) = D_\phi(P,Q)^\pi \tag{4}$$

where $\pi$ is a positive power of the $\phi$-divergence $D_\phi(P,Q)$ defined for arbitrary $P,Q \in \mathbb{P}$ by the formula

$$D_\phi(P,Q) = \int g\,\phi\left(\frac{f}{g}\right)\mathrm{d}\mu, \qquad f = \frac{\mathrm{d}P}{\mathrm{d}\mu},\ \ g = \frac{\mathrm{d}Q}{\mathrm{d}\mu}. \tag{5}$$

Here and in the sequel, $\mu$ denotes a $\sigma$-finite measure on $(\mathcal{X},\mathcal{A})$ dominating $P,Q$ and $\phi(t)$ a nonnegative convex function on the domain $0 < t < \infty$ strictly convex and vanishing at $t = 1$. Behind the integral in (5) are considered the extensions and conventions

$$\phi(0) = \lim_{t\downarrow 0}\phi(t) \quad \text{and} \quad 0\,\phi\left(\frac{f}{0}\right) = f\,\phi^*(0) \quad \text{for all } f \geq 0 \tag{6}$$

where

$$\phi^*(0) = \lim_{t\downarrow 0}\phi^*(t) \ \text{ and } \phi^*(t) = t\,\phi\left(\frac{1}{t}\right),\ t > 0. \tag{7}$$

By (5), for each constant $c \in \mathbb{R}$,

$$D_{\phi(t)}(P,Q) = D_{\phi(t)+c.(t-1)}(P,Q) \quad \text{for all } P,Q \in \mathbb{P} \tag{8}$$

but if the function $\phi(t)$ under consideration is differentiable at $t = 1$ then $\phi'(1) = 0$ so that $\phi(t) + c\cdot(t-1)$ is nonnegative for $0 < t < \infty$ and vanishing at $t = 1$ only if $c = 0$.

Note that in the definition (4) we consider powers of $\phi$-divergences rather than the $\phi$-divergences themselves because, by the Corollary to Theorem A.5 in the Appendix, the triangle inequality (3) with $D(P,Q)$ replaced by $D_\phi(P,Q)$ holds only if $D_\phi(P,Q) = \delta\cdot V(P,Q)$ where $\delta$ is a positive constant and

$$V(P,Q) = \int |g - f|\,\mathrm{d}\mu \tag{9}$$

is the $\phi$-divergence for $\phi(t) = |t-1|$ called *total variation* of $P,Q$.

The metric properties (1) – (3) together with the following basic $\phi$-divergence properties **(a)** – **(e)** valid for all $P,Q \in \mathbb{P}$ guarantee wide applicability of the divergences studied in this paper. The properties **(a)** – **(e)** are stated here for references later. For their detailed proofs, and also for details about applications of (6) – (8) in the definition (5), see Csiszár [1] and Liese and Vajda [8, 9].

**(a)** The *range of values* is

$$0 \leq D_\phi(P,Q) \leq \phi(0) + \phi^*(0) \tag{10}$$

for $\phi(0)$, $\phi^*(0)$ given by (6), (7). Here $D_\phi(P,Q) = 0$ if and only if $P = Q$ so that the reflexivity (1) holds for every power of $\phi$-divergence considered in (4). On the other hand, the upper bound $D_\phi(P,Q) = \phi(0) + \phi^*(0)$ is achieved if $P \perp Q$ (orthogonality, i. e. disjoint supports of $P$ and $Q$). The "if" condition can be replaced by "if and only if" when

$$\phi(0) + \phi^*(0) < \infty. \tag{11}$$

**(b)** The function $\phi^*(t)$ adjoint to $\phi(t)$ in the sense of (7) is nonnegative, convex on the domain $t > 0$ and strictly convex and vanishing at $t = 1$. Thus it defines the $\phi^*$-divergence

$$D_{\phi^*}(P, Q) = D_\phi(Q, P) \quad \text{for all } P, Q \in \mathbb{P}. \tag{12}$$

The *symmetry*

$$D_\phi(P, Q) = D_\phi(Q, P) \quad \text{for all } P, Q \in \mathbb{P} \tag{13}$$

takes place if and only if there exists a constant $c \in \mathbb{R}$ such that

$$\phi^*(t) = \phi(t) + c \cdot (t - 1). \tag{14}$$

For symmetric $\phi$-divergences $\phi^*(0) = \phi(0) - c$ so that (11) reduces to the simpler boundedness condition

$$\phi(0) < \infty. \tag{15}$$

**(c)** The *monotonicity* deals with relations between the $\phi$-divergences $D_\phi(P, Q)$ of probability measures $P, Q$ and the $\phi$-divergences of restrictions $P_\mathcal{B}, Q_\mathcal{B}$ on sub-$\sigma$-algebras $\mathcal{B} \subset \mathcal{A}$. It states the ordering

$$D_\phi(P_\mathcal{B}, Q_\mathcal{B}) \le D_\phi(P, Q) \tag{16}$$

with the equality if $\mathcal{B}$ is sufficient for the pair $\{P, Q\}$. The "if" condition can be replaced by "if and only if" when $\phi(t)$ is strictly convex on the whole domain $t > 0$ and $D_\phi(P, Q)$ is finite.

**(d)** If the $\sigma$-algebra $\mathcal{A}$ is generated by an at most countable $\mathcal{A}$-measurable partition $\mathcal{S}$ of $\mathcal{X}$ (called spectrum of $\mathcal{A}$) then the *spectral representation* provides the simpler $\phi$-divergence formula

$$D_\phi(P, Q) = \sum_{A \in \mathcal{S}} Q(A) \phi\left(\frac{P(A)}{Q(A)}\right). \tag{17}$$

**(e)** Finally, the *finite approximation*

$$D_\phi(P, Q) = \sup_{\mathcal{S}} \sum_{A \in \mathcal{S}} Q(A) \phi\left(\frac{P(A)}{Q(A)}\right) \tag{18}$$

is a general alternative to the definition (5). Here the supremum is assumed to run over all finite $\mathcal{A}$-measurable partitions $\mathcal{S}$ of $\mathcal{X}$. Let us note that the conventions (6), (7) are supposed to be applied behind the sums in (18) and (17).

**Remark 1.** As said above, if the nonnegative convex $\phi(t)$ under consideration is differentiable at $t = 1$ then $\phi'(1) = 0$ which is equivalent to $(\phi^*)'(1) = 0$. Thus by (14)

$$\phi^*(t) = \phi(t) \quad \text{for all } t > 0 \tag{19}$$

is necessary and sufficient condition for the symmetry (13). Therefore in this case the symmetric $\phi$-divergences satisfy the identity $\phi^*(0) = \phi(0)$. Some situations where it is not so are illustrated in Example 1 below.

**Remark 2.** The fact that $\phi^*$ considered in **(b)** satisfies the conditions imposed in (5) provided $\phi$ does so follows from Theorem A.1 in the Appendix. By Theorem A.2 in the Appendix, (15) is the Csiszár necessary condition for the metricity of a $\phi$-divergence power $D(P,Q) = D_\phi(P,Q)^\pi, \pi > 0$. Hence the metrizable divergences $D_\phi(P,Q)$ are uniformly finitely bounded on $\mathbb{P}$ and achieve the upper bound

$$\phi(0) + \phi^*(0) = 2\phi(0) - c \qquad \text{(cf. (14))} \tag{20}$$

if and only if $P \perp Q$.

**Example 1.** Relation (8) and the symmetry conditions in **(b)** can be illustrated by the class of functions

$$\phi_{\alpha,\beta}(t) = \alpha(1-t)\mathbb{I}(t<1) + \beta(t-1)\mathbb{I}(t>1) \quad \text{for } \alpha,\beta \geq 0,\ \alpha+\beta>0 \tag{21}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Obviously, all these function belong to the class of convex functions considered in this paper and satisfy the relation $\phi^*_{\alpha,\beta}(t) = \phi_{\beta,\alpha}(t)$ and also the symmetry condition

$$\phi^*_{\alpha,\beta}(t) = \phi_{\alpha,\beta}(t) + (\beta-\alpha)(t-1) \qquad \text{(cf. (14))}$$

considered in **(b)**. Hence if $\alpha \neq \beta$ then

$$\phi_{\alpha,\beta}(0) = \alpha \neq \beta = \phi^*_{\alpha,\beta}(0).$$

It is easy to see that the functions (21) define the metric divergences

$$D_{\phi_{\alpha,\beta}}(P,Q) = \frac{\alpha+\beta}{2} V(P,Q) \tag{22}$$

where $V(P,Q)$ is the total variation (9).

The present paper follows [6] where the authors emphasized statistical applicability of some $\phi$-divergence classes with functions $\phi = \phi_\alpha$ continuously depending on a real parameter $\alpha$. According to what was said above, it is important to select among them them the classes satisfying for positive powers $\pi = \pi_\alpha$ the metric divergence properties (1)–(3). One such class with parameters $\alpha \in \mathbb{R}$ is investigated in the present paper. It is an extension of the class introduced previously for $\alpha \geq 0$ in Österreicher and Vajda [12].

## 2. METRIC DIVERGENCES

Let us start with some historical examples of $\phi$-divergences which are or are not metric divergences in the sense defined above, and with a review of previous results in this area.

As it was said above, the triangle inequality

$$V(P,Q) \leq V(P,R) + V(R,Q) \quad \text{for all } P,Q,R \in \mathbb{P} \tag{23}$$

holds for the total variation (9) since it is the $L_1$-distance of probability densities which fulfills also the remaining metric properties (1) and (2). Among the well known $\phi$-divergences (4) satisfying in the power $\pi = 1/2$ all metric properties (1)–(3) is the *squared Hellinger distance*

$$H^2(P,Q) = 2\int \left(\sqrt{g} - \sqrt{f}\right)^2 \mathrm{d}\mu \quad \text{for } \phi(t) = 2\left(\sqrt{t}-1\right)^2 \tag{24}$$

since it is the squared $L_2$-distance in the space of the square roots $\sqrt{f}$ and $\sqrt{g}$ of probability densities. The reflexivity and symmetry (1), (2) of the *squared Le Cam distance* (sometimes called *the Vincze–Le Cam distance*)

$$LC^2(P,Q) = \frac{1}{2}\int \frac{(f-g)^2}{f+g}\,\mathrm{d}\mu \quad \text{for } \phi(t) = \frac{(t-1)^2}{2(t+1)} \tag{25}$$

introduced independently by Vincze [14] and Le Cam [7] are easily seen from **(a)** and **(b)** above, but the triangle inequality (3) for $LC(P,Q) = (LC^2(P,Q))^{1/2}$ is a nontrivial problem to which we return in the next section. The best known example of $\phi$-divergence is the *information divergence*

$$I(P,Q) = \int f \ln\left(\frac{f}{g}\right)\mathrm{d}\mu \quad \text{for } \phi(t) = t\ln t - t + 1. \tag{26}$$

Here (4) with no $\pi$ is metric. Obviously, all powers $D(P,Q) = I(P,Q)^\pi$, $\pi > 0$ are reflexive but none of them is symmetric in the sense of (2). The powers $J(P,Q)^\pi$ of the reflexive symmetrized version

$$J(P,Q) = I(P,Q) + I(Q,P)$$

called *Jeffrey's divergence* do not satisfy the triangle inequality. Indeed, $J(P,Q)$ is the $\tilde{\phi}$-divergence for the sum $\tilde{\phi} = \phi + \phi^*$ of the logarithmic function (26) with the adjoint function $\phi^*(t) = -\ln t + t - 1$ where $\tilde{\phi}(0) = \infty$ violates the necessary metricity condition (15).

Metric properties of $\phi$-divergences were for the first time studied by Csiszár [2] who introduced the metricity condition (15). However, the classes of metric divergences in the sense stated above started to be systematically studied by Kafka et al. [4]. These authors proved the sufficient metricity condition of Theorem A.3 in the Appendix cited in the sequel simply as the *Kafka condition.* In the selected examples above we find this condition with $\pi = 1$ satisfied by the total variation function $\phi(t) = |t-1|$, and with $\pi = 1/2$ by the Hellinger function $\phi(t) = 2\left(\sqrt{t}-1\right)^2$ and the Le Cam function $\phi(t) = (t-1)^2/2(t+1)$. In [4] this condition was used to establish the metricity of the classes of $\phi$-divergences of the form

$$D(P,Q) = D_{\phi_\alpha}(P,Q)^{1/\alpha} \qquad \text{(cf. (4))} \tag{27}$$

generated by the class

$$\phi_\alpha(t) = |1 - t^{1/\alpha}|^\alpha \quad \text{for } 0 < \alpha \le 1 \tag{28}$$

of Matusita [10] functions as well as by the the extensions

$$\phi_\alpha(t) = \frac{|1-t|^\alpha}{(t+1)^{\alpha-1}} \quad \text{for } \alpha \geq 1$$

of the Le Cam function $\phi(t) = \phi_2(t)$.

Later Österreicher [11] used the Kafka condition to prove the metricity of the class of $\phi_\beta$-divergences

$$D(P,Q) = D_{\phi_\beta}(P,Q)^{1/2} \quad \text{(cf. (4))} \tag{29}$$

defined by the class of convex functions

$$\phi_\beta(t) = (1+t^\beta)^{1/\beta} - 2^{(1/\beta)-1}(1+t), \quad \beta > 1 \tag{30}$$

including the limit $\phi_\infty(t) = |t-1|/2$.

Österreicher and Vajda [12] normalized and extended the functions (30) into the class

$$f_\beta(t) = \frac{\phi_\beta(t)}{1-1/\beta} = \frac{(1+t^\beta)^{1/\beta} - 2^{(1/\beta)-1}(1+t)}{1-1/\beta}, \quad \beta > 0,\ \beta \neq 1 \tag{31}$$

including the limits $f_\infty(t) = \phi_\infty(t)$ and

$$f_1(t) = \lim_{\beta\to 1} f_\beta(t) = t\ln t + (t+1)\ln\frac{2}{t+1}$$

and proved that the class of the corresponding $f_\beta$-divergence powers

$$D(P,Q) = D_{f_\beta}(P,Q)^{\min\{1/2,\beta\}}, \quad 0 < \beta \leq \infty$$

satisfy the metric properties (1) – (3).

The present paper further extends the previous extension, namely to the domain $\beta < 0$. The basic step is the reparametrization of the class (31) by $\alpha = 1/\beta \geq 0$. A slight modification consisting in the multiplication by

$$\text{sign } \alpha = \frac{\alpha}{|\alpha|} \quad \text{for } \alpha \neq 0$$

allowed to extend the convexity of the functions $\phi_\alpha(t) = f_{1/\alpha}(t)$ to the domain $\alpha < 0$. As a result, we introduce here the class of $\phi_\alpha$-divergences

$$\mathcal{D}_\alpha(P,Q) = D_{\phi_\alpha}(P,Q), \quad \alpha \in \mathbb{R} \tag{32}$$

for the convex functions $\phi_\alpha(t)$ given in the domain $t > 0$ by the formula

$$\phi_\alpha(t) = \frac{\text{sign } \alpha}{1-\alpha}\left[(t^{1/\alpha}+1)^\alpha - 2^{\alpha-1}(t+1)\right] \tag{33}$$

if $\alpha(\alpha-1) \neq 0$ and by the corresponding limits

$$\phi_1(t) = \lim_{\alpha\to 1}\phi_\alpha(t) = t\ln t + (t+1)\ln\frac{2}{t+1} \tag{34}$$

and

$$\phi_0(t) = \lim_{\alpha\to 0}\phi_\alpha(t) = |t-1|/2 \tag{35}$$

otherwise. Our main result is the next theorem proved in Section 3 below.

**Theorem 1.**

**(i)** The functions $\phi_\alpha(t)$ given by (33) – (35) and used in the definition of $\mathcal{D}_\alpha(P,Q)$ in (32) satisfy the assumptions concerning $\phi$ in the definition of $\phi$-divergence (5), with the strict convexity instead of the ordinary convexity on the domain $t>0$ unless $\alpha=0$. Moreover, they are are self-adjoint in the sense $\phi_\alpha(t) = t\phi_\alpha(1/t) \equiv \phi^*_\alpha(t)$ on this domain so that the $\phi_\alpha$-divergences $\mathcal{D}_\alpha(P,Q)$ are symmetric in the sense of (2).

**(ii)** The powers

$$D(P,Q) = \mathcal{D}_\alpha(P,Q)^{\pi(\alpha)} \quad \text{for} \quad \pi(\alpha) = \begin{cases} \frac{1}{2} & \text{when } -\infty < \alpha \le 2 \\ \\ \frac{1}{\alpha} & \text{when } \alpha > 2 \end{cases} \tag{36}$$

of the extended $\phi_\alpha$-divergences satisfy the metric properties (1) – (3).

**Example 2.** It is easy to verify for all $P, Q$ the formulas

$$\mathcal{D}_0(P,Q) = V(P,Q)/2 \quad \text{(total variation (9))}, \tag{37}$$

$$\mathcal{D}_2(P,Q) = H^2(P,Q)/4 \quad \text{(Hellinger (24))}, \tag{38}$$

$$\mathcal{D}_{-1}(P,Q) = LC^2(P,Q)/4 \quad \text{(Le Cam (25))}, \tag{39}$$

$$\mathcal{D}_1(P,Q) = I\left(P,(P+Q)/2\right) + I(Q,(P+Q)/2) \quad \text{(normalized and symmetrized } I\text{-divergence (26))}. \tag{40}$$

Thus the roots $\sqrt{\mathcal{D}_k(P,Q)}$ for $k=-1,0,1,2$ are metrics on the space of probability measures $\mathbb{P}$.

In connection with the normalized $I$-divergence $I\left(P,(P+Q)/2\right)$ and $I(Q,(P+Q)/2)$ appearing in (40) and sometimes called *the Jensen–Shannon divergence* (see e. g. Fuglede and Topsøe [3]), one can mention that the general normalized versions

$$D_\phi\left(P,(P+Q)/2\right) = D_{\phi^{(1)}}(P,Q), \quad \phi^{(1)}(t) = \frac{1+t}{2}\,\phi\left(\frac{2t}{1+t}\right) \tag{41}$$

and

$$D_\phi\left(Q,(P+Q)/2\right) = D_{\phi^{(2)}}(P,Q), \quad \phi^{(2)}(t) = \frac{1+t}{2}\,\phi\left(\frac{2}{1+t}\right) \tag{42}$$

of arbitrary $\phi$-divergences were introduced and studied previously in [13]. Their symmetrized variants

$$\begin{aligned} D_{\phi^{(1)}+\phi^{(2)}}(P,Q) &= \int \frac{f+g}{2}\left[\phi\left(\frac{2f}{f+g}\right) + \phi\left(\frac{2g}{f+g}\right)\right] \\ &= D_\phi\left(P,(P+Q)/2\right) + D_\phi\left(Q,(P+Q)/2\right) \end{aligned} \tag{43}$$

contain as a special case the symmetrized and normalized $I$-divergence $\mathcal{D}_1(P,Q)$ given by (40). However, powers of the general normalized and symmetrized divergences $D_{\phi^{(1)}+\phi^{(2)}}(P,Q)$ usually do not satisfy the triangle inequality – see e. g. the next example. In this sense the Jensen–Shannon divergence $\mathcal{D}_1(P,Q)$ represents an exception.

**Example 3.** Consider the nonnegative convex function $\phi(t) = -\ln t + t - 1$ leading to the *reversed information divergence* $D_\phi(P,Q) = I(Q,P)$. Here

$$\phi^{(1)}(t) + \phi^{(2)}(t) = (1+t)\ln\left(\frac{1+t}{2\sqrt{t}}\right) \qquad \text{(cf. (41), (42))}$$

defines the normalized and symmetrized reversed $I$-divergence

$$D_{\phi^{(1)}+\phi^{(2)}}(P,Q) = I\left((P+Q)/2, P\right) + I\left((P+Q)/2, Q\right).$$

Since $\phi^{(1)}(0)+\phi^{(2)}(0) = \infty$ violates the necessary metricity condition (15), no power of this divergence fulfills the triangle inequality on $\mathbb{P}$.

## 3. SUPPLEMENT AND PROOF OF THEOREM 1

The following supplement of Theorem 1 presents bounds obtained for the class of the divergences $\mathcal{D}_\alpha(P,Q)$, $\alpha \in \mathbb{R}$ defined by (33) – (35). Among them are the tight lower and upper bounds

$$L_\alpha(V) \le \mathcal{D}_\alpha(P,Q) \le U_\alpha(V) \tag{44}$$

attained for fixed values $\alpha \in \mathbb{R}$ by the $\phi_\alpha$-divergences $\mathcal{D}_\alpha(P,Q)$ of distributions $P, Q \in \mathbb{P}$ with a given total variation value $V(P,Q) = V, \quad 0 \le V \le 2$.

**Theorem 2.**

**(i)** The divergences $\mathcal{D}_\alpha(P,Q)$ take on values between 0 and and the strictly positive values

$$\phi_\alpha(0) + \phi_\alpha^*(0) = 2\phi_\alpha(0) = \begin{cases} \dfrac{2^\alpha}{|\alpha|+1} & \text{when } \alpha < 0 \\ \dfrac{2^\alpha - 2}{\alpha - 1} & \text{when } \alpha \ge 0,\ \alpha \ne 1 \\ 2\ln 2 & \text{when } \alpha = 1. \end{cases} \tag{45}$$

The bounds $\mathcal{D}_\alpha(P,Q) = 0$ or $\mathcal{D}_\alpha(P,Q) = 2\phi_\alpha(0)$ are attained if and only if $P = Q$ or $P \perp Q$ (disjoint supports of $P$ and $Q$), respectively.

**(ii)** The attainable lower bounds in (44) are for every argument $0 \le V \le 2$ continuous in the variable $\alpha \in \mathbb{R}$ and given by the formulas

$$L_\alpha(V) = \frac{|\alpha|}{\alpha(\alpha-1)}\left(2^\alpha - \left[(1+V/2)^{1/\alpha} + (1-V/2)^{1/\alpha}\right]^\alpha\right) \tag{46}$$

if $\alpha(\alpha-1)\neq 0$ and by their limits

$$L_0(V)=V/2,\quad L_1(V)=(1+V/2)\ln(1+V/2)+(1-V/2)\ln(1-V/2) \quad (47)$$

for $\alpha\to 0$ and $\alpha\to 1$ otherwise.

**(iii)** If the observation space $(\mathcal{X},\mathcal{A})$ is dichotomous in the sense that $\mathcal{A}$ contains only two nonvoid sets $A$, $\mathcal{X}-A$ then the attainable upper bound defined in (44) is the convex function

$$U_\alpha(V)=\frac{V}{2}c_\alpha+\frac{2-V}{2}\phi\left(\frac{2}{2-V}\right) \quad (48)$$

of variable $0\le V\le 2$, where the constants $c_\alpha>0$ continuously depend on the parameter $\alpha\in\mathbb{R}$ and are given by the formula

$$c_\alpha=\phi_\alpha(0)=\begin{cases}2^{\alpha-1}/(|\alpha|+1) & \text{when } \alpha<0\\ (2^{\alpha-1}-1)/(\alpha-1) & \text{when } \alpha\ge 0,\ \alpha\neq 1\\ \ln 2 & \text{when } \alpha=1.\end{cases} \quad (49)$$

**(iv)** If the $\sigma$-algebra $\mathcal{A}$ contains more than two nonvoid sets then the attainable upper bounds defined in (44) are linear functions of the form $U_\alpha(V)=c_\alpha V$ where the constants $c_\alpha>0$ are the same as in (49).

Proof of Theorem 1. For $\alpha\ge 0$ the desired results follow from what was proved in Österreicher and Vajda [12] so that it suffices to prove the extensions for $\alpha<0$. Next follows a proof covering all cases $\alpha\in\mathbb{R}$. It uses the auxiliary functions

$$\psi_\alpha^{(k)}(t)=\left(1+t^{1/\alpha}\right)^{\alpha-k},\quad t>0$$

for $k=0,1,2,\ldots$ and $\alpha\neq 0$.

**(i)** For $\alpha=0$ the statement is obvious so let $\alpha\neq 0$. The nonnegativity and strict convexity of $\phi_\alpha(t)$ follow from the first and second derivative formulas

$$\phi_\alpha'(1)=0\quad\text{and}\quad \phi_\alpha''(t)=\frac{1}{|\alpha|}\,\psi_\alpha^{(2)}(t)t^{(1-2\alpha)/\alpha}>0.$$

Further, it is easy to verify that the functions $\psi_\alpha^{(0)}(t)$ are self-adjoint in the sense $t\psi_\alpha^{(0)}(1/t)=\psi_\alpha^{(0)}(t)$. Then the self-adjointness of $\phi_\alpha(t)$ follows easily and implies the symmetry of $\mathcal{D}_\alpha(P,Q)$ directly from the definition of $\phi$-divergence in (5).

**(ii)** The reflexivity and symmetry of $\mathcal{D}_\alpha(P,Q)^{\pi(\alpha)}$ follow from **(i)**. It remains to prove the triangle inequality

$$\mathcal{D}_\alpha(P,Q)^{\pi(\alpha)}\le\mathcal{D}_\alpha(P,R)^{\pi(\alpha)}+\mathcal{D}_\alpha(R,Q)^{\pi(\alpha)} \quad (50)$$

for arbitrary probability measures $P, Q, R \in \mathbb{P}$, and arbitrary $\alpha \in \mathbb{R}$. In fact, this suffices to prove for all $\alpha$ from a dense subset $\mathbb{R}_* \subset \mathbb{R}$ Namely, for restrictions $P_{\mathcal{B}}, Q_{\mathcal{B}}, R_{\mathcal{B}}$ of $P, Q, R \in \mathbb{P}$ on the sub-algebras $\mathcal{B} \subset \mathcal{A}$ generated by arbitrary finite $\mathcal{A}$-measurable spectra $\mathcal{S}$ the inequality

$$\mathcal{D}_\alpha(P_{\mathcal{B}}, Q_{\mathcal{B}})^{\pi(\alpha)} \leq \mathcal{D}_\alpha(P_{\mathcal{B}}, R_{\mathcal{B}})^{\pi(\alpha)} + \mathcal{D}_\alpha(R_{\mathcal{B}}, Q_{\mathcal{B}})^{\pi(\alpha)} \tag{51}$$

can be extended from $\mathbb{R}_*$ to $\mathbb{R}$ by using the spectral representations of the divergences in (51) stated in **(c)** above, and applying the continuity of the expressions

$$Q(A)\phi_\alpha\left(\frac{P(A)}{Q(A)}\right), \quad R(A)\phi_\alpha\left(\frac{P(A)}{R(A)}\right), \quad Q(A)\phi_\alpha\left(\frac{R(A)}{Q(A)}\right), \quad A \in \mathcal{S}$$

in the variable $\alpha \in \mathbb{R}$. The inequality (51) established in this manner for all $\alpha \in \mathbb{R}$ can be further extended into the general form (50) by applying for every fixed $\alpha \in \mathbb{R}$ the finite approximation of the divergences appearing in (50) in the sense of **(d)** above. Here we shall prove (50) for all $\alpha \in \mathbb{R}_* = \mathbb{R} - \{0, 1\}$, i. e. for $\alpha(1-\alpha) \neq 0$. Fix such $\alpha$ and consider the ratio (54) of the Kafka criterion for $\pi = 1/2$, i. e. let

$$f_\alpha(t) = \frac{(\sqrt{t}-1)^2}{\phi_\alpha(t)}, \quad \phi_\alpha(t) = \frac{\text{sign }\alpha}{1-\alpha}\left[\psi_\alpha^{(0)}(t) - 2^{\alpha-1}(1+t)\right], \quad t > 0.$$

If $-\infty < \alpha \leq 2$ then it suffices to prove that the derivative $f'_\alpha(t)$ is nonpositive in the domain $0 < t < 1$. But

$$f'_\alpha(t) = \left(\frac{1}{\sqrt{t}} - 1\right)\left[\frac{g_\alpha(t)}{\phi_\alpha^2(t)}\right]$$

for

$$g_\alpha(t) = \frac{\text{sign }\alpha}{1-\alpha}\left[2^{\alpha-1}(1+\sqrt{t}) - \psi_\alpha^{(1)}(t)\left(1 + t^{(2-\alpha)/2\alpha}\right)\right]$$

where $g_\alpha(1) = 0$. Hence it suffices to prove that the function

$$h_\alpha(t) = \sqrt{t}g'_\alpha(t) = \frac{\text{sign }\alpha}{1-\alpha}\left\{2^{\alpha-2} - \psi_\alpha^{(2)}(t)t^{1/\alpha}\left[\frac{1-\alpha}{\alpha}(1-\sqrt{t}) + \frac{1+\sqrt{t}}{2}\right]\right\}$$

is nonnegative in the domain $0 < t < 1$. However, $h_\alpha(1) = 0$ so that the nonnegativity of $h_\alpha(t)$ follows from the obvious nonpositivity of the derivative

$$h'_\alpha(t) = -\frac{2-\alpha}{2\alpha^2}\psi_\alpha^{(3)}(t)t^{\frac{1}{\alpha}-2}.$$

If $\alpha > 2$ then it suffices to apply the Kafka criterion for $\pi = 1/\alpha$. This step is skipped here as it can be realized by obvious modifications of the steps in Österreicher and Vajda [12]. □

Proof of Theorem 2. **(i)** It follows from (33) – (35) that the limits $\phi_\alpha(0)$ satisfy (45). The range of values and the conditions for equalities thus follow from the self-adjointness in **(i)** of Theorem 1 and from the general range of values property **(a)** above.

**(ii)** For $\alpha = 0$ the coinciding bounds $L_0(V) = U_0(V) = V/2$ are trivial consequences of the equality $\mathcal{D}_0(P,Q) = V(P,Q)/2$ established in (37). If $\alpha \neq 0$ then the function $\phi_\alpha$ satisfies for all $0 < V < 2$ the relation

$$(2+V)\,\phi_\alpha\left(\frac{2-V}{2+V}\right) = (2-V)\,\phi_\alpha\left(\frac{2+V}{2-V}\right) = L_\alpha(V)$$

for $L_\alpha(V)$ given in (46). Hence Proposition 8.28 in Liese and Vajda [9] implies that the functions $L_\alpha(V)$ given by (46) are the desired lower bounds. The continuity of $L_\alpha(V)$ in $\alpha \in \mathbb{R}$ can be verified with the help of the l'Hospital rule.

**(iii)** By Theorem A.6 in the Appendix, the attainable upper bound is

$$U_{\phi_\alpha}(V) = \max\left\{\frac{V}{2}\phi_\alpha(0) + \frac{2-V}{2}\phi_\alpha\left(\frac{2}{2-V}\right), \frac{V}{2}\phi^*_\alpha(0) + \frac{2-V}{2}\phi^*_\alpha\left(\frac{2}{2-V}\right)\right\}$$

where by **(b)**,

$$\phi^*_\alpha(t) = \phi_\alpha(t) + c_\alpha.(t-1) \quad \text{for all } t > 0.$$

This implies that both maximized expressions coincide. By (45), $\phi_\alpha(0) = c_\alpha$ for $c_\alpha$ given by (49). It is easy to verify the continuity of the constant $c_\alpha$ of (49) in the parameter $\alpha \in \mathbb{R}$.

**(iv)** By Proposition A.6 in the Appendix and (45),

$$U_{\phi_\alpha}(V) = V \cdot c_\alpha \quad \text{where} \quad c_\alpha = \frac{\phi_\alpha(0) + \phi^*_\alpha(0)}{2} = \phi_\alpha(0). \tag{52}$$

The rest is clear from the previous step. □

**Remark 2.** For the particular parameter $\alpha = 0$ and the corresponding divergence power $(2\mathcal{D}_0(P,Q))^{1/2} = \sqrt{V(P,Q)}$ we get the triangle inequality

$$\sqrt{V(P,Q)} \leq \sqrt{V(P,R)} + \sqrt{V(R,Q)}$$

which is weaker than the classical inequality (23). This indicates that also for $\alpha \neq 0$ are not excluded stronger triangular inequalities than those obtained from Theorem 1.

**Example 4.** Take the divergence $D_{-1}(P,Q) = LC^2(P,Q)/4$, i. e. the modified squared Le Cam from Example 2. Then assertion **(i)** of Theorem 2 leads to the lower bound

$$\begin{aligned} L_{-1}(V) &= \frac{1}{2}\left(\frac{1}{2} - \left[\frac{1}{1+V/2} + \frac{1}{1-V/2}\right]^{-1}\right) \\ &= \frac{1}{2}\left[\frac{1}{2} - \frac{1-(V/2)^2}{2}\right] = \left(\frac{V}{4}\right)^2. \end{aligned}$$

Assertion **(ii)** of the same theorem implies that $c_{-1} = 1/8$ which leads to the upper bound $U_{-1}(V) = V/8$. Thus we obtained the relation

$$V(P,Q)/2 \le LC(P,Q) \le \sqrt{V(P,Q)/2} \tag{53}$$

for the Le Cam distance where both the inequalities are tight. This result seems to be new.

## APPENDIX

Here are stated the assertions needed in Sections 1 – 3.

**Theorem A.1.** If $\phi : (0,\infty) \longmapsto \mathbb{R}$ is convex then the function

$$\psi(u,v) = v\phi\left(\frac{u}{v}\right)$$

of two variables is convex on the domain $(u,v) \in (0,\infty)^2$.

P r o o f. Consider $\lambda \in (0,1)$ and two points $x_i = (u_i, v_i)$ from the domain of $\phi$. For

$$w = \frac{\lambda v_1}{\lambda v_1 + (1-\lambda)v_2} \quad \text{and} \quad t_i = \frac{u_i}{v_i}$$

we get

$$\phi(wt_1 + (1-w)t_2) \le w\phi(t_1) + (1-w)\phi(t_2)$$

so that

$$(\lambda v_1 + (1-\lambda)v_2)\phi\left(\frac{\lambda u_1 + (1-\lambda)u_2}{\lambda v_1 + (1-\lambda)v_2}\right) \le \lambda v_1\phi\left(\frac{u_1}{v_1}\right) + (1-\lambda)v_2\phi\left(\frac{u_2}{v_2}\right)$$

or, equivalently,

$$\psi(\lambda x_1 + (1-\lambda)x_2) \le \lambda\psi(x_1) + (1-\lambda)\psi(x_2)$$

which completes the proof. □

**Theorem A.2.** If a positive power $D(P,Q) = D_\phi(P,Q)^\pi$ of a symmetric $\phi$-divergence satisfies the triangle inequality (3) then $\phi(0) < \infty$.

P r o o f. Let the triangle inequality hold an let by contrary $\phi(0) = \infty$. Then the desired contradiction is based on the possibility to choose $P, Q, R \in \mathbb{P}$ such that $D_\phi(P,Q) = \infty$ and $D_\phi(P,R) + D_\phi(Q,R) < \infty$. For details we refer to Csiszár [2] who established this metricity criterion. □

**Theorem A.3** If for a convex function $\phi(t)$ considered in (4) defines symmetric $\phi$-divergence and for some $\pi > 0$ the ratio

$$\frac{(1-t^\pi)^{1/\pi}}{\phi(t)} \tag{54}$$

is nonincreasing in the variable $t \in (0,1)$ then the power (10) satisfies the triangle inequality (3).

P r o o f. For the proof we refer to Kafka et al. [4] where this metricity criterion was established. □

**Theorem A.4.** If a convex function $\phi : (a,b) \longmapsto \mathbb{R}$ is strictly convex at no $t \in (a,b)$ then $\phi$ is linear on $(a,b)$.

P r o o f. By definition, $\phi$ is not strictly convex at $t \in (a,b)$ if and only if there is an open neighborhood $U_t \subset (a,b)$ of $t$ such that $\phi$ is linear on it. By choosing a countable subcovering of $(a,b)$ from the covering $\{U_t : t \in (a,b)\}$ and using the mathematical induction, the linearity of can be extended from any neighborhood $U_t$ to the whole interval $(a,b)$. □

**Theorem A.5.** If the $\phi$-divergence $D(P,Q) = D_\phi(P,Q)$ under consideration satisfies the triangle inequality (3) then $\phi(t)$ is linear on the subdomains $(0,1)$ and $(1,\infty)$.

P r o o f. Similar result was obtained recently by Khosravifard et al. [5]. Next follows a more transparent and simpler proof. Consider the Bernoulli probability measures

$$P = (p,q) \quad \text{and} \quad Q = (q,p) \quad \text{where} \quad q = 1-p \quad \text{and let} \quad t = \frac{q}{p} \in (0,1) \cup (1,\infty).$$

Under the assumptions of theorem,

$$D_\phi(P,Q) \le D_\phi\left(P, \frac{P+Q}{2}\right) + D_\phi\left(\frac{P+Q}{2}, Q\right)$$

where using the convexity of $\phi^*(t)$ and the assumption $\phi(1) \equiv \phi^*(1) = 0$ we get from (12) and from the spectral representation formula (17)

$$D_\phi\left(P, \frac{P+Q}{2}\right) = D_{\phi^*}\left(\frac{P+Q}{2}, P\right) = p\phi^*\left(\frac{t+1}{2}\right) + q\phi^*\left(\frac{t^{-1}+1}{2}\right)$$
$$\leq \frac{p}{2}\left[\phi^*(t) + t\phi^*\left(t^{-1}\right)\right] = \frac{p}{2}\left[t\phi\left(t^{-1}\right) + \phi(t)\right] = \frac{1}{2}D_\phi(P,Q).$$

By means of the auxiliary functions $A_\phi(t) = t\left[\phi\left(t^{-1}\right) + \phi(1)\right]/2$ and $B_\phi(t) = \left[\phi(t) + \phi(1)\right]/2$ and the assumption $\phi(1) = 0$ the get from here

$$\begin{aligned}\frac{1}{2}D_\phi(P,Q) &= p\left[A_\phi(t) + B_\phi(t)\right] \\ &\leq D_\phi\left(\frac{P+Q}{2}, Q\right) = p\left[t\phi\left(\frac{t^{-1}+1}{2}\right) + \phi\left(\frac{t+1}{2}\right)\right]\end{aligned}$$

where the convexity of $\phi(t)$ implies

$$A_\phi(t) \geq t\phi\left(\frac{t^{-1}+1}{2}\right) \quad \text{and} \quad B_\phi(t) \geq \phi\left(\frac{t+1}{2}\right).$$

Therefore the equalities

$$A_\phi(t) \equiv t\frac{\phi\left(t^{-1}\right) + \phi(1)}{2} = t\phi\left(\frac{t^{-1}+1}{2}\right) \quad \text{and} \quad B_\phi(t) \equiv \frac{\phi(t) + \phi(1)}{2} = \phi\left(\frac{t+1}{2}\right)$$

hold for all $t \in (0,1) \cup (1,\infty)$. This implies that $\phi(t)$ is strictly convex at no point of the form

$$\frac{t^{-1}+1}{2} \quad \text{or} \quad \frac{t+1}{2} \quad \text{for} \quad t \in (0,1) \cup (1,\infty).$$

Since such points cover the whole domain $(0,1) \cup (1,\infty)$, the rest is clear from Theorem A.4. □

**Corollary.** If $D_\phi(P,Q)$ is the $\phi$-divergence considered in (5) then the metricity conditions (1) – (3) with $D(P,Q)$ replaced by $D_\phi(P,Q)$ hold if and only if $D_\phi(P,Q) = \delta.V(P,Q)$ where $\delta$ is a positive constant and $V(P,Q)$ is the total variation (9).

P r o o f. This assertion was previously obtained in [5]. The verification of metricity for $D_\phi(P,Q) = \delta.V(P,Q)$ is the same as at the beginning of Section 2. If conversely $D_\phi(P,Q)$ is a metric then Theorem A.5 together with the condition $\phi(1) = 0$ implies that $\phi(t)$ coincides with one of the functions $\phi_{\alpha,\beta}(t)$ of Example 1. From the symmetry discussed in Example 1 we get that $D_\phi(P,Q)$ must be of the form (22).□

The last theorem of this section deals with the tight upper $\phi$-divergence bound

$$U_\phi(V) = \sup_{(P,Q)\in\mathbb{Q}_V} D_\phi(P,Q), \quad 0 \leq V \leq 2 \tag{55}$$

where $\mathbb{Q}_V = \{(P,Q) \in \mathbb{P}\otimes\mathbb{P} : V(P,Q) = V\}$. Special cases $U_\alpha(V) := U_{\phi_\alpha}(V)$ for the family of functions $\phi_\alpha$ defined in (33) were introduced in (44). This theorem deals with observation spaces $(\mathcal{X}, \mathcal{A})$ nontrivial in the sense $\mathcal{A} \neq \{\emptyset, \mathcal{X}\}$ and sharpens Proposition 8.27 of Liese and Vajda [8].

**Theorem A.6.** If the observation space $(\mathcal{X}, \mathcal{A})$ is dichotomous in the sense that $\mathcal{A}$ contains only two nonvoid sets $A$, $\mathcal{X} - A$ then the attainable upper bound (55) is

$$U_\phi(V) = \max\left\{\frac{V}{2}\phi(0) + \frac{2-V}{2}\phi\left(\frac{2}{2-V}\right), \frac{V}{2}\phi^*(0) + \frac{2-V}{2}\phi^*\left(\frac{2}{2-V}\right)\right\}. \quad (56)$$

In all remaining observation spaces this bound is linear of the form

$$U_\phi(V) = \frac{\phi(0) + \phi^*(0)}{2}V. \quad (57)$$

P r o o f . In the dichotomous case with

$$Q(A) = t \quad \text{and} \quad P(A) = t + V/2 \quad \text{for} \quad 0 \le V \le 2 \quad \text{and} \quad 0 \le t \le 1 - V/2$$

we get from the spectral representation formula (17)

$$D_\phi(P, Q) = t\phi\left(\frac{t + V/2}{t}\right) + (1-t)\phi\left(\frac{1 - t - V/2}{1-t}\right).$$

This function is convex in $t$ and it is easy to verify that the arguments of the maxima in (56) are the extremal values of this function at $t = 0$ and $t = 1 - V/2$. If $\mathcal{A}$ contains three nonvoid sets $A, B, C = \mathcal{X} - A \cup B$ and

$$P(A) = Q(A) = 1 - V/2, \quad P(B) = Q(C) = 0, \quad P(c) = Q(B) = V/2$$

then

$$V(P, Q) = V \quad \text{and} \quad D_\phi(P, Q) = V(\phi(0) + \phi^*(0))/2$$

so that the values (57) are attained. The fact that (57) is the bound follows from the inequalities

$$\phi(t) \le \phi(0)\,(1-t) + t\phi(1) = \phi(0)\,(1-t)$$

and

$$\phi^*(t) \le \phi^*(0)\,(1-t) + t\phi^*(1) = \phi^*(0)\,(1-t)$$

valid for all $0 \le t \le 1$. Indeed, using these inequalities we get from from the definitions (5) and (9) for the set $A = \{f \ge g\}$

$$\begin{aligned} D_\phi(P, Q) &= \textstyle\int_A f\phi^*\,(g/f)\,\mathrm{d}\mu + \int_{\mathcal{X}-A} g\phi\,(f/g)\,\mathrm{d}\mu \\ &\le \phi^*(0)\textstyle\int_A (f-g)\mathrm{d}\mu + \phi(0)\int_{\mathcal{X}-A}(g-f)\mathrm{d}\mu \\ &= \phi^*(0)V(P, Q)/2 + \phi(0)V(P, Q)/2 \end{aligned}$$

which implies the desired result. □

ACKNOWLEDGEMENT

This research was supported by the Czech Science Foundation under Grant 102/07/1131 and by the Ministry of Education, Youth and Sport of the Czech Republic under Project 1M0572. The author thanks referees for careful reading of the manuscript and for many useful suggestions.

(Received June 23, 2008.)

REFERENCES

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

[2] I. Csiszár: On topological properties of $f$-divergences. Studia Sci. Math. Hungar. *2* (1967), 329–339.

[3] B. Fuglede and T. Topsøe: Jensen–Shannon divergence and Hilbert space embedding. In: Proc. IEEE Internat. Symposium on Inform. Theory, IEEE Publications, New York 2004, p. 31.

[4] P. Kafka, F. Österreicher, and I. Vincze: On powers of $f$-divergences defining a distance. Stud. Sci. Math. Hungar. *26* (1991), 329–339.

[5] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver: Confliction of the convexity and metric properties in $f$-divergences. IEICE Trans. on Fundamentals *E90-A* (2007), 1848–1853.

[6] V. Kůs, D. Morales, and I. Vajda: Extensions of the parametric families of divergences used in statistical inference. Kybernetika *44* (2008), 95–112.

[7] L. Le Cam: Asymptotic Methods in Statistical Decision Theory. Springer, New York 1986.

[8] F. Liese and I. Vajda: Convex Statistical Distances. Teubner, Leipzig 1987.

[9] F. Liese and I. Vajda: On divergence and informations in statistics and information theory. IEEE Trans. Inform. Theory *52* (2006), 4394–4412.

[10] K. Matusita: Decision rules based on the distance for problems of fit, two samples and estimation. Ann. Math. Statist. *26* (1955), 631–640.

[11] F. Öesterreicher: On a class of perimeter-type distances of probability distributions. Kybernetika *32* (1996), 389–393.

[12] F. Österreicher and I. Vajda: A new class of metric divergences on probability spaces and its statistical applications. Ann. Inst. Statist. Math. *55* (2003), 639–653.

[13] I. Vajda: On the $f$-divergence and singularity of probability measures. Period. Math. Hungar. *2* (1972), 223–234.

[14] I. Vincze: On the concept and measure of information contained in an observation. In: Contributions to Probability (J. Gani and V. F. Rohatgi, eds.), Academic Press, New York 1981, pp. 207–214.

*Igor Vajda, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.*
*e-mail: vajda@utia.cas.cz*