CSISZÁR'S f-DIVERGENCES - BASIC PROPERTIES

Ferdinand Österreicher
Institute of Mathematics, University of Salzburg, Austria

Abstract

In this talk basic general properties of f-divergences, including their axiomatic, and some important classes of f-divergences are presented.

Without essential loss of insight we restrict ourselves to discrete probability distributions and note that the extension to the general case relies strongly on the *Lebesgue-Radon-Nikodym* Theorem.

This talk was presented while participating in a workshop of the Research Group in Mathematical Inequalities and Applications at the Victoria University, Melbourne, Australia, in October 2002.

1 BASIC NOTIONS

Let $\Omega = \{x_1, x_2, ...\}$ be a set with at least two elements, $\mathfrak{P}(\Omega)$ the set of all subsets of Ω and $\mathcal{P}$ the set of all probability distributions $P = (p(x) : x \in \Omega)$ on Ω .

A pair $(P, Q) \in \mathcal{P}^2$ of probability distributions is called a *(simple versus simple) testing problem.*

Two probability distributions P and Q are called *orthogonal* ($P \perp Q$) if there exists an element $A \in \mathcal{P}(\Omega)$ such that $P(A) = Q(A^c) = 0$ where $A^c = \Omega \backslash A$.

A testing problem $(P, Q) \in \mathcal{P}^2$ is called *least infomative* if $P = Q$ and is called *most informative* if $P \perp Q$.

Furthermore, let $\mathcal{F}$ be the set of convex functions $f : [0, \infty) \mapsto (-\infty, \infty]$ continuous at 0 (i.e. $f(0) = \lim_{u \downarrow 0} f(u)$) , $\mathcal{F}_0 = \{f \in \mathcal{F} : f(1) = 0\}$ and let $D_- f$ and $D_+ f$ denote the *left-hand side* and *right-hand side derivative* of f , respectively. Further let $f^* \in \mathcal{F}$, defined by

$$f^*(u) = u f\left(\frac{1}{u}\right) \ , \ u \in (0, \infty) \ ,$$

the $*$-*conjugate* (convex) function of f , let a function $f \in \mathcal{F}$ satisfying $f^* \equiv f$ be called $*$-*self conjugate* and let $\tilde{f} = f + f^*$. Then

$$\begin{array}{rcccl} 0 \cdot f^*\left(\frac{x}{0}\right) & = & x \cdot f\left(\frac{0}{x}\right) & = & x \cdot f(0) \quad \text{for} \quad x \in (0,\infty) \\ 0 \cdot f\left(\frac{y}{0}\right) & = & y \cdot f^*\left(\frac{0}{y}\right) & = & y \cdot f^*(0) \quad \text{for} \quad y \in (0,\infty) \\ 0 \cdot f\left(\frac{0}{0}\right) & = & 0 \cdot f^*\left(\frac{0}{0}\right) & = & 0 \ . \end{array}$$

Definition (*Csiszár* (1963), *Ali & Silvey* (1966)): Let $P, Q \in \mathcal{P}$. Then

$$I_f(Q,P) = \sum_{x \in \Omega} p(x) f\left(\frac{q(x)}{p(x)}\right)$$

is called the f-*divergence* of the probability distributions Q and P .

Remark 1: Because of $p(x) f\left(\frac{q(x)}{p(x)}\right) = q(x) f^*\left(\frac{q(x)}{p(x)}\right) \quad \forall\ x \in \Omega$ it holds

$$I_f(Q,P) = I_{f^*}(P,Q) \quad \forall\ (P,Q) \in \mathcal{P}^2 \ .$$

EXAMPLES: Total Variation Distance ($f(u) = |u-1| = f^*(u)$)

$$I_f(Q,P) = \sum_{x \in \Omega} p(x) \left| \frac{q(x)}{p(x)} - 1 \right| = \sum_{x \in \Omega} |q(x) - p(x)|$$

χ^2-Divergence ($f(u) = (u-1)^2$, $f^*(u) = \frac{(u-1)^2}{u}$)

$$I_f(Q,P) = \sum_{x \in \Omega} p(x) \left(\frac{q(x)}{p(x)} - 1 \right)^2 = \sum_{x \in \Omega} \frac{(q(x) - p(x))^2}{p(x)} = I_{f^*}(P,Q)$$

Kullback-Leibler Divergence ($f(u) = u \ln(u)$, $f^*(u) = -\ln(u)$)

$$I_f(Q,P) = \sum_{x \in \Omega} p(x) \frac{q(x)}{p(x)} \ln\left(\frac{q(x)}{p(x)}\right) = \sum_{x \in \Omega} q(x) \ln\left(\frac{q(x)}{p(x)}\right) = I_{f^*}(P,Q)$$

2 BASIC PROPERTIES (Part 1)

Uniqueness Theorem (*Liese & Vajda* (1987))**:** Let $f,\ f_1 \in \mathcal{F}$. Then

(1) $I_{f_1}(Q,P) = I_f(Q,P) \quad \forall\ (P,Q) \in \mathcal{P}^2$ iff (2) $\exists c \in \mathbb{R} : f_1(u) - f(u) = c\,(u-1)$.

Proof: (2) $\Longrightarrow$ (1): The f-divergence of the function $f_1 - f$ vanishes because of

$$I_{f_1 - f}(Q,P) = c \sum_{x \in \Omega} (q(x) - p(x)) = c \left(\sum_{x \in \Omega} q(x) - \sum_{x \in \Omega} p(x) \right) = c\,(1-1) = 0 \ .$$

(1) $\Longrightarrow$ (2): For this direction we restrict ourselves to the case $\tilde{f}(0) < \infty$. Then the Range of Values Theorem stated below implies $\tilde{f}_1(0) = \tilde{f}(0)$ and therefore $c = f_1^*(0) - f^*(0) = f(0) - f_1(0)$.

For $u \leq 1$ let $P = (1, 0)$ and $Q = (u, 1-u)$. Then (1) implies in view of $I_f(Q,P) = f(u) + 0f(\frac{1-u}{0}) = f(u) + (1-u) f^*(0)$

$$f_1(u) - f(u) = (1-u)(f^*(0) - f_1^*(0)) = c(u-1) .$$

For $u > 1$ let $P = (\frac{1}{u}, 1 - \frac{1}{u})$ and $Q = (1, 0)$. Then (1) implies in view of $I_f(Q,P) = \frac{1}{u} f(u) + \left(1 - \frac{1}{u}\right) f(0)$

$$f_1(u) - f(u) = (u-1)(f(0) - f_1(0)) = c(u-1) .$$

Remark 2: a) Owing to $(u-1)^2 = u^2 - 1 - 2(u-1)$

$$\chi^2(Q,P) = \sum_{x \in \Omega} p(x) \left(\frac{q(x)}{p(x)} - 1 \right)^2 = \sum_{x \in \Omega} p(x) \left(\left(\frac{q(x)}{p(x)} \right)^2 - 1 \right)$$

b) Let $f \in \mathcal{F}$ and $c \in [D_- f(1), D_+ f(1)]$ then $f_1(u) = f(u) - c(u-1)$ satisfies $f_1(u) \geq f(1) \ \ \forall \ u \in [0, \infty)$ while not changing the f-divergence. Hence we can assume $f(u) \geq f(1) \ \ \forall \ u \in [0, \infty)$ without loss of generality. For theoretical purposes and purposes of unification of specific f-divergences it is often convenient to switch to such functions f . (See e.g. the making of Class II).

Symmetry Theorem (*Liese & Vajda* (1987))**:** Let $f \in \mathcal{F}$ and f^* be its $*$-conjugate. Then

$$I_{f^*}(Q,P) = I_f(Q,P) \ \ \forall \ (P,Q) \in \mathcal{P}^2 \quad \text{iff} \quad \exists c \in \mathbb{R} : f^*(u) - f(u) = c(u-1) .$$

In words: An f-divergence is symmetric iff - apart from an additional linear term $c(u-1)$ - f is $*$-self conjugate.

Remark 3: a) Obviously the functions $\tilde{f} = f + f^*$ and $\tilde{f}/2$ are $*$-self conjugate. Owing to

$$\frac{f(u) + f^*(u)}{u+1} = \frac{1}{u+1} f(u) + \frac{u}{u+1} f(\frac{1}{u}) \geq f(1)$$

it holds $\tilde{f}(u)/2 - f(1) \geq \frac{f(1)}{2}(u-1)$ and hence $f(u) - f(1) \geq \frac{f(1)}{2}(u-1)$ provided f is $*$-self conjugate.

b) The maximum of f and f^*, namely $\hat{f}(u) = \max(f(u), f^*(u))$ is also $*$-self conjugate. This provides another possibility to obtain a $*$-self conjugate function from a given function $f \in \mathcal{F}$.

Remark 4: Note that

$$I_f(Q,P) = f(0)\cdot P(\{x : q(x)=0\}) + f^*(0)\cdot Q(\{x : p(x)=0\}) + \\ + \sum_{x:\ q(x)\cdot p(x)>0} p(x) f\left(\frac{q(x)}{p(x)}\right)$$

and that $P(\{x : q(x)=0\})$ is the amount of singularity of the distribution P with respect to Q and $Q(\{x : p(x)=0\})$ is the amount of singularity of the distribution Q with respect to P. Therefore $f(0)=\infty$ and $f^*(0)=\infty$ imply $I_f(Q,P)=\infty$ unless $\{x\in\Omega :\ q(x)\cdot p(x)>0\} = \Omega$, i.e. all probabilities are positive.

Range of Values Theorem (*Vajda* (1972))**:** It holds

$$f(1) \le I_f(Q,P) \le f(0)+f^*(0) \qquad \forall\ Q,P\in\mathcal{P}\ .$$

In the first inequality, equality holds if / iff $Q=P$. The latter provided f is strictly convex at 1.
In the second, equality holds if / iff $Q\perp P$. The latter provided $\tilde{f}(0) = f(0)+f^*(0)<\infty$.

Remark 5: In order to exclude the trivial case $I_f(Q,P)\equiv f(1)$ we will assume from now on that $f\in\mathcal{F}$ is not trivial, i.e. it satisfies $\tilde{f}(0)-f(1)>0$.

Measures of Similarity [1]

In this case we concentrate on the first inequality. The difference $I_f(Q,P)-f(1)$ is a quantity which compares the given testing problem $(P,Q)\in\mathcal{P}^2$ with the least informative testing problem. These quantities are therefore appropriate for applications where the two probability distributions are or get very close.

In order that $I_f(Q,P)$ fulfils the basic property (M1) of a measure of similarity, namely

$$I_f(Q,P)\ge 0 \qquad \text{with equality iff} \qquad Q=P\ , \tag{M1}$$

f needs to have the properties (i,0) $f(1)=0$ and (i,1) f is strictly convex at 1.

Given a function $f\in\mathcal{F}$, property (i,0) can easily be achieved by setting $f(u):=f(u)-f(1)$. Hence we will assume $f\in\mathcal{F}_0$ without loss of generality.

Measures of (approximate) Orthogonality

In this case we concentrate on the second inequality. The difference $\tilde{f}(0)-I_f(Q,P)$ is a quantity which compares the given testing problem $(P,Q)\in$

[1] The notions 'Measures of Similarity' and 'Measures of Orthogonality', which are not common in literature, are intended to distinguish between the two major types of applications of f-divergences.

$\mathcal{P}^2$ with the most informative testing problem. These quantities are therefore appropriate for applications where the two probability distributions are or get nearly orthogonal.

To ensure that this difference exists we have to assume $\tilde{f}(0) < \infty$ and hence $f(0) < \infty$ and $f^*(0) < \infty$.

We attribute to such a (convex) function $f \in \mathcal{F}$ the concave function $g : [0, \infty) \mapsto [0, \infty)$ given by

$$g(u) = f(0) + u \cdot f^*(0) - f(u) \ ,$$

which - obviously - satisfies $g(0) = g^*(0) = 0$, $g(1) = \tilde{f}(0) - f(1)$ and is monotone increasing, and define

$$I_g(Q, P) = \sum_{x \in \Omega} p(x) g\left(\frac{q(x)}{p(x)}\right) \ .$$

Then owing to $I_g(Q, P) = \tilde{f}(0) - I_f(Q, P)$ our 'Measure of Orthogonality' can be expressed in terms of $I_g(Q, P)$ more appropriately.

For all $f \in \mathcal{F}$ satisfying $0 < \tilde{f}(0) - f(1) < \infty$ the quantity $I_g(Q, P)$ is defined and fulfils the basic property (O) of a measure of orthogonality, namely

$$I_g(Q, P) \geq 0 \qquad \text{with equality iff} \qquad Q \perp P \ . \tag{O}$$

Remark 6: It is important to note that both types of measures have their specific applications, whereby many applications of the 'Measures of Similarity' rely heavily on the convexity of the function f, whereas those of the 'Measures of Orthogonality' rely heavily on the concavity of the function g .

3 CLASSES OF f-DIVERGENCES

In this section we present some of the more intensively studied classes of f-divergences in terms of their convex functions f . The historic references are intended to give some hints as to their making.

Some of these and further f-divergences have also been investigated by members of the RGMIA. The paper by *Barnett, Cerone, Dragomir & Sofo* (2002) may serve as but one example.

(I) The class of χ^α-divergences

Total Variation Distance

$$f(u) = |u - 1|$$

K. Pearson (1900)

$$\chi^2(u) = (u - 1)^2$$

Kagan (1963), *Vajda* (1973), *Boekee* (1977)

$$\chi^{\alpha}(u) = |u-1|^{\alpha} \ , \ \alpha \geq 1$$

(II) Dichotomy Class

Kullback & Leibler (1951)

$$f(u) = u\ln(u)$$

Likelihood Disparity

$$f^{*}(u) = -\ln(u)$$

K. Pearson (1900)

$$\chi^{2}(u) = (u-1)^{2}$$

Neyman (1949)

$$\left(\chi^{2}\right)^{*}(u) = \frac{(u-1)^{2}}{u}$$

Liese & Vajda (1987)

$$f^{\alpha}(u) = \begin{cases} u-1-\ln u & \text{for} \quad \alpha = 0 \\ \frac{\alpha u+1-\alpha-u^{\alpha}}{\alpha(1-\alpha)} & \text{for} \quad \alpha \in \mathbb{R}\setminus\{0,1\} \\ 1-u+u\ln u & \text{for} \quad \alpha = 1 \end{cases}$$

Read & Cressie (1988): $f_{\lambda}(u) = \frac{u^{\lambda+1}-1}{\lambda(\lambda+1)}$ with $\lambda = \alpha - 1 \in \mathbb{R}\setminus\{-1,0\}$

Remark 7: According to *Feldman* (1972, for $\alpha \in (0,1)$) and *Leidinger* (1996, for the general case) this class of f-divergences is characterized by the dichotomy with respect to testing problems.

(II') Symmetrized Dichotomy Class

Jeffreys (1946)

$$\tilde{f}(u) = (u-1)\ln(u)$$

Csiszár & Fischer (1962)

$$f^{(s)}(u) = 1+u-\left(u^{s}+u^{1-s}\right) \ , \ 0 < s < 1$$

$$\tilde{f}^{(s)}(u) = \begin{cases} (u-1)\ln(u) & \text{for} \quad s = 1 \\ \frac{1+u-\left(u^{s}+u^{1-s}\right)}{s(1-s)} & \text{for} \quad s \in (0,1)\cup(1,\infty) \end{cases}$$

(III) Matusita's Divergences

Matusita (1954)

$$f^{\frac{1}{2}}(u) = \left(\sqrt{u}-1\right)^{2}$$

Matusita (1964), *Boekee* (1977)

$$f^{\alpha}(u) = |u^{\alpha} - 1|^{\frac{1}{\alpha}} \ , \ 0 < \alpha \leq 1$$

Renyi's Divergences [2]

(*Hellinger* (1909): $g^{\frac{1}{2}}(u) = \sqrt{u}$)

Bhattacharyya (1946)

$$-\ln(\sum_{x\in\Omega} \sqrt{p(x)q(x)})$$

Chernoff (1952)

$$-\min_{0\leq\alpha\leq 1} \ln(\sum_{x\in\Omega} p(x) \left(\frac{q(x)}{p(x)}\right)^{\alpha})$$

Renyi (1961)

$$R_{\alpha}(Q,P) = \begin{cases} \sum_{x\in\Omega} q(x) \ln(\frac{q(x)}{p(x)}) & \text{for} \quad \alpha = 1 \\ \frac{1}{\alpha-1} \ln(\sum_{x\in\Omega} p(x) \left(\frac{q(x)}{p(x)}\right)^{\alpha}) & \text{for} \quad \alpha \in (0,\infty)\backslash\{1\} \end{cases}$$

(IV) Elementary Divergences

Österreicher & *Feldman* (1981)

$$f_t(u) = \max(u - t, 0) \ , \ t \geq 0$$

(V) Puri-Vincze Divergences

Le Cam (1986), *Topsøe* (1999)

$$\Phi_2(u) = \frac{1}{2} \frac{(1-u)^2}{u+1}$$

Puri & *Vincze* (1990), *Kafka, Österreicher* & *Vincze* (1989)

$$\Phi_k(u) = \frac{1}{2} \frac{|1-u|^k}{(u+1)^{k-1}} \ , \ k \geq 1$$

(VI) Divergences of Arimoto-type

Perimeter Divergence: *Österreicher* (1982), *Reschenhofer* & *Bomze* (1991)

$$f(u) = \sqrt{1+u^2} - \frac{1+u}{\sqrt{2}}$$

[2] Note that this class doesn't belong to the family of f-divergences and the functions $g^{\alpha}(u) = u^{\alpha}$, $\alpha \in (0,1)$ are concave.

Perimeter-type Divergences: *Österreicher* (1996)

$$f_p(u) = \begin{cases} (1+u^p)^{\frac{1}{p}} - 2^{\frac{1}{p}-1}(1+u) & \text{for} \quad p \in (1,\infty) \\ \frac{|1-u|}{2} & \text{for} \quad p = \infty \end{cases}$$

Österreicher & *Vajda* (1997)

$$f_\beta(u) = \begin{cases} \frac{1}{1-1/\beta}\left[\left(1+u^\beta\right)^{1/\beta} - 2^{1/\beta-1}(1+u)\right] & \text{if} \quad \beta \in (0,\infty)\backslash\{1\} \\ (1+u)\ln(2) + u\ln(u) - (1+u)\ln(1+u) & \text{if} \quad \beta = 1 \\ |1-u|/2 & \text{if} \quad \beta = \infty\,. \end{cases}$$

Remark 8: *Lin* (1991) proposed his f-divergence in terms of the convex function

$$f(u) = \ln(2) + u\ln\left(\frac{u}{1+u}\right)\ .$$

Owing to

$$\tilde{f}(u) = f(u) + f^*(u) = (1+u)\ln(2) + u\ln(u) - (1+u)\ln(1+u)$$

Lin's (in this way) symmetrized f-divergence equals our special case $\beta = 1$.

4 BASIC PROPERTIES (Part 2: Axiomatic)

Characterization Theorem (*Csiszár*, 1974)**:** Given a mapping $I : \mathcal{P}^2 \mapsto (-\infty, \infty]$ then the following two statements are equivalent

$(*)$ I is an f-divergence
i.e. there exists an $f \in \mathcal{F}$ such that $I(Q,P) = I_f(Q,P)\ \ \forall\ (P,Q) \in \mathcal{P}^2$

$(**)$ I satisfies the following three properties.

(a) $I(Q,P)$ is invariant under permutation of Ω ,

(b) Let $\mathcal{A} = (A_i,\ i \geq 1)$ be a partition of Ω and let

$$P_{\mathcal{A}} = (P(A_i),\ i \geq 1) \quad \text{and} \quad Q_{\mathcal{A}} = (Q(A_i),\ i \geq 1)$$

be the restrictions of the probability distributions P and Q to $\mathcal{A}$. Then

$$I(Q,P) \geq I(Q_{\mathcal{A}}, P_{\mathcal{A}})$$

with equality holding if $Q(A_i) \times p(x) = P(A_i) \times q(x)\ \ \forall\ x \in A_i\ ,\ i \geq 1$ and

(c) Let P_1, P_2 and Q_1, Q_2 probability distributions on Ω . Then

$$I(\alpha P_1 + (1-\alpha)P_2, \alpha Q_1 + (1-\alpha)Q_2) \leq \alpha I(P_1,Q_1) + (1-\alpha)I(P_2,Q_2)\ .$$

Remark 9: a) Since the proof of the direction $(*) \Rightarrow (**)$ will be an immediate consequence of the Representation Theorem (*Österreicher & Feldman*, 1982) we skip it here and present, instead, a proof of the direction $(**) \Rightarrow (*)$ under the assumption that all probabilities are positive.

b) The properties (b) and/or (c) are crucial for many applications of f - divergences. We will concentrate on the applications of f-divergences in a later talk.

Proof of the direction $(**) \Rightarrow (*)$: Consequences of (a):

Let $P = (p_1, ..., p_m),\ Q = (q_1, ..., q_m)$ such that $\{x \in \Omega : p(x)\, q(x) > 0\} = \Omega$. Then (a) implies that there exists a function $v : (0, \infty)^2 \mapsto \mathbb{R}$ such that

$$I(Q, P) = \sum_{i=1}^{m} v(p_i, q_i) \ .$$

We have to show

$$q f^* \left(\frac{p}{q}\right) = q v(1, \frac{p}{q}) = v(p, q) = p v(\frac{q}{p}, 1) = p f\left(\frac{q}{p}\right) \quad \forall\ 0 < p, q < 1 \ .$$

Consequences of (b):

Let $m \geq 2,\ 1 \leq r \leq m$ and $0 < t < \frac{m}{r}$ and let $\Omega = \{x_1, ..., x_r, x_{r+1}\}$. Furthermore let

$P_r = (\frac{1}{m}, ..., \frac{1}{m}, 1 - \frac{r}{m}),\ Q_r = (\frac{t}{m}, ..., \frac{t}{m}, 1 - t\frac{r}{m}),\ A = \{x_1, ..., x_r\}$ and

$\mathcal{A} = \{A, \{x_{r+1}\}, \emptyset, \Omega\},\ \ \tilde{P}_r = (\frac{r}{m}, 1 - \frac{r}{m}),\ \tilde{Q}_r = (t\frac{r}{m}, 1 - t\frac{r}{m})$. Then, owing to

$$Q(A) \cdot p(x) = t\frac{r}{m^2} = Q(A) \cdot p(x) \quad \forall\ x \in A \ ,$$

(b) implies

$$\begin{aligned} 0 &= I(\tilde{Q}_r, \tilde{P}_r) - I(P_r, Q_r) = \\ &= \left[v\left(\frac{r}{m}, t\frac{r}{m}\right) + v\left(1 - \frac{r}{m}, 1 - t\frac{r}{m}\right)\right] - \left[\sum_{i=1}^{r} v\left(\frac{1}{m}, t\frac{1}{m}\right) + v\left(1 - \frac{r}{m}, 1 - t\frac{r}{m}\right)\right] \\ &= v\left(\frac{r}{m}, t\frac{r}{m}\right) - r v\left(\frac{1}{m}, t\frac{1}{m}\right) \end{aligned}$$

and hence

$$v\left(\frac{r}{m}, t\frac{r}{m}\right) = r v\left(\frac{1}{m}, t\frac{1}{m}\right) \ .$$

For $r = m$ this yields $v\left(\frac{1}{m}, t\frac{1}{m}\right) = \frac{1}{m} v(1, t)$ and consequently

$$v\left(\frac{r}{m}, t\frac{r}{m}\right) = \frac{r}{m} r(1, t) \ .$$

Therefore it holds

$$v\,(p,q) = pr(1,\frac{q}{p})$$

for $p = \frac{r}{m}$ and $q = t\frac{r}{m}$ and all $0 < t < \frac{m}{r}$, $1 \le r \le m,\ m \ge 2$.

Consequences of (c):

Let $0 < x, y < 1,\ 0 < p < x,\ 0 < q < y$ and $p, q, x, y \in \mathbb{Q}$. Furthermore let $\Omega = \{x_1, x_2, x_3\},\ P_1 = (p, x-p, 1-x),\ P_2 = (x-p, p, 1-x)$,
$Q_1 = (q, y-q, 1-y),\ Q_2 = (y-q, q, 1-y)$ and finally $\alpha = \frac{1}{2}$. Then

$$\alpha P_1 + (1-\alpha)P_2 = \frac{P_1+P_2}{2} = (\frac{x}{2}, \frac{x}{2}, 1-x) \quad \text{and} \quad \alpha Q_1 + (1-\alpha)Q_2 = \frac{Q_1+Q_2}{2} = (\frac{y}{2}, \frac{y}{2}, 1-y)$$

and hence (c) implies

$$\begin{aligned}
\Delta &= \alpha I(P_1,Q_1) + (1-\alpha)I(P_2,Q_2) - I(\alpha P_1 + (1-\alpha)P_2, \alpha Q_1 + (1-\alpha)Q_2) \\
&= \frac{I(P_1,Q_1)+I(P_2,Q_2)}{2} - I(\frac{P_1+P_2}{2}, \frac{Q_1+Q_2}{2}) \\
&= \frac{1}{2}\left[\begin{array}{l} pv\left(1,\frac{q}{p}\right) + (x-p)v\left(1,\frac{y-q}{x-p}\right) + (1-x)v\left(1,\frac{1-y}{1-x}\right) + \\ +(x-p)\ v\left(1,\frac{y-q}{x-p}\right) + pv\left(1,\frac{q}{p}\right) + (1-x)v\left(1,\frac{1-y}{1-x}\right) \end{array}\right] - \\
&\quad - \left[xv\left(1,\frac{y}{x}\right) + (1-x)\,v\left(1,\frac{1-y}{1-x}\right)\right] \\
&= pv\left(1,\frac{q}{p}\right) + (x-p)v\left(1,\frac{y-q}{x-p}\right) - xv\left(1,\frac{y}{x}\right) \ge 0
\end{aligned}$$

and, after dividing by x ,

$$\frac{p}{x}v\left(1,\frac{q}{p}\right) + \frac{x-p}{x}v\left(1,\frac{y-q}{x-p}\right) \ge v\left(1,\frac{y}{x}\right)\ .$$

In view of $\frac{p}{x}\frac{q}{p} + \frac{x-p}{x}\frac{y-q}{x-p} = \frac{y}{x}$ the convexity of the function

$$f : (0,\infty) \cap \mathbb{Q} \mapsto \mathbb{R} \ \text{ defined by } \ f(u) = v(1,u)$$

is verified. Let f also denote the continuous extension of this convex function from $(0,\infty)\cap\mathbb{Q}$ to $(0,\infty)$. Then f is convex and fulfils, due to its continuity,

$$v\,(p,q) = pr(1,\frac{q}{p}) = pf(\frac{q}{p})$$

for all $0 < p, q$. By setting $f^*(u) = v(u,1)$ we similarly obtain

$$v\,(p,q) = qr(\frac{p}{q},1) = qf^*(\frac{p}{q})$$

for all $0 < p, q$.

Remark 10: By this approach the introduction of the $*$-conjugate f^* of a convex function f , defined by

$$f^*(u) \equiv uf(\frac{1}{u})$$

is straightforward.

References

Ali, S. M. and Silvey, S. D. (1966): A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B*, **28**, 131–142.

Barnett, N.S., Cerone, P., Dragomir, S.S. and A. Sofo (2002): Approximating Csiszár f-divergence by the use of Taylor's formula with integral remainder. *Math. Inequ. & Appl.*, **5**/3, 417-434.

Bhattacharyya, A. (1946): On some analogues to the amount of information and their uses in statistical estimation. *Sankhya* **8**, 1-14.

Boekee, D. E.: A generalization of the Fisher information measure. Delft University Press, Delft 1977.

Chernoff, H. (1952): A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.*, **30**, 493-507.

Csiszár, I. and Fischer, J. (1962): Informationsentfernungen im Raum der Wahrscheinlichkeitsverteilungen. *Magyar Tud. Akad. Mat. Kutató Int. Kösl.*, **7**, 159–180.

Csiszár, I. (1963): Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.*, **8**, 85–107.

Csiszár, I. (1974): Information measures: A critical survey. In: *Trans. 7^{slth} Prague Conf. on Information Theory*, Academia Prague, Vol. A, 73-86.

Feldman, D. (1972): Some properties of Bayesian orderings of experiments. *Ann. Math. Statist.*, **43**, 1428-1440.

Feldman, D. and Österreicher, F. (1981): Divergenzen von Wahrscheinlichkeitsverteilungen – integralgeometrisch betrachtet. *Acta Math. Acad. Sci. Hungar.*, **37**/4, 329–337.

Hellinger, E. (1909): Neue Begründung der Theorie der quadratischen Formen von unendlichen vielen Veränderlichen. *J. Reine Ang. Math.*, **136**, 210-271.

Jeffreys, H. (1946): An invariant form for the prior probability in estimating problems. *Proc. Roy. Soc.*, Ser. A, **186**, 453-461.

Kullback, S. and R. Leibler (1951): On information and sufficiency. *Ann. Math. Stat.*, **22**, 79-86.

Kullback, S.: Information Theory and Statistics. Dover Publications, New York 1968

Kafka, P., Österreicher, F. and Vincze I. (1991): On powers of f-divergences defining a distance. *Studia Sci. Math. Hungar.*, **26**, 415–422.

Kagan, A.M. (1963): On the theory of Fisher's amount of information (in Russian). *Dokl. Akad. Nauk SSSR*, **151**, 277-278.

Le Cam, L.: Asymptotic Methods in Statistical Decision Theory. Springer-Verlag, New York-Berlin 1986

Leidinger, J. (1996): Zur Charakterisierung der Dichotomie beim Vergleich des Informationsgehalts statistischer Experimente, Dissertation, Salzburg

Lin, J. (1991): Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Th.*, **37**, 145-151.

Liese, F. and Vajda, I.: Convex Statistical Distances. Teubner–Texte zur Mathematik, Band **95**, Leipzig 1987

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

Matusita, K. (1964): Distances and decision rules. *Ann. Inst. Statist. Math.*, **16**, 305–320.

Neyman, J. (1949): Contribution to the theory of χ^2-test. *Proc. 1^{slst} Berkeley Symp. on Math. Statist.*, 239-273, Univ. Calif. Press, Berkeley.

Österreicher, F. (1982): The construction of least favourable distributions is traceable to a minimal perimeter problem. *Studia Sci. Math. Hungar.*, **17**, 341–351.

Österreicher, F.: Informationstheorie, Skriptum zur Vorlesung, Salzburg 1991

Österreicher, F. (1996): On a class of perimeter-type distances of probability distributions. *Kybernetika*, **32**/4, 389–393.

Österreicher, F. and Vajda, I. (1997): A new class of metric divergences on probability spaces and its statistical applicability. Submitted to the Annals of the Institute of Statistical Mathematics, Japan.

Pearson, K. (1900): On the criterion that a given system of deviations from the probable in the case of correlated system of variables in such that it can be reasonable supposed to have arisen from random sampling. *Phil. Mag.*, **50**, 157-172.

Ruri, M.L. and Vincze, I. (1990): Measures of information and contiguity. *Statist. Probab. Letters*, **9**, 223-228.

Read, T. C. R. and Cressie, N. A.: Goodness–of–Fit Statistics for Discrete Multivariate Data. Springer, New York 1988

Renyi, A. (1961): On measures of entropy and information. *Proc. 4^{slth} Berkeley Symp. on Math. Statist.*, Vol. 1, 547-561, Univ. Calif. Press, Berkeley.

Reschenhofer, E. and Bomze, I. M. (1991): Length tests for goodness of fit. *Biometrika* **78**, 207–216.

Topsøe, F. (1999): Some inequalities for information divergence and related measures of discrimination. *Res. Rep. Call., RGMIA,* **2**/1, 85-98.

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

Vajda, I. (1973): χ^{α}-divergences and generalized Fisher's information. In: *Trans. 6^{slth} Prague Conf. on Information Theory*, Vol. B, 873-886, Academia Prague.

Vajda, I.: Theory of Statistical Inference and Information, Kluwer Academic Publishers, Dordrecht-Boston 1989