f-DIVERGENCES - REPRESENTATION THEOREM AND METRIZABILITY

Ferdinand Österreicher
Institute of Mathematics, University of Salzburg, Austria

Abstract

In this talk we are first going to state the so-called 'Representation Theorem' which provides the representation of general f-divergences in terms of the class of elementary divergences. Then we consider the risk set of a (simple versus simple) testing problem and its characterization by the above mentioned class. These ingredients enable us to prove a sharpening of the 'Range of Values Theorem' presented in Talk 1.

In the sequel, necessary and sufficient conditions are given so that f-divergences allow for a topology, respectively a metric. In addition, sufficient conditions are given which permits a suitable power of an f-divergence to be a distance.

Finally, we investigate the classes of f-divergences discussed in Talk 1 along these lines.

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 November 2002.

Let $\Omega = \{x_1, x_2, ...\}$ be a set, $\mathfrak{P}(\Omega)$ the set of all subsets of Ω , $\mathcal{P}$ the set of all probability distributions $P = (p(x) : x \in \Omega)$ on Ω and $\mathcal{P}^2$ the set of all (simple) testing problems.

Furthermore, let $\mathcal{F}_0$ be the set of convex functions $f : [0, \infty) \mapsto (-\infty, \infty]$ continuous at 0 and satisfying $f(1) = 0$, $f^* \in \mathcal{F}_0$ the $*$-conjugate (convex) function of f and $\tilde{f} = f + f^*$.

1 REPRESENTATION THEOREM

1.1 REPRESENTATION BY ELEMENTARY DIVERGENCES

At the beginning we investigate the f-divergences of the Class (IV) of Elementary Divergences given by

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

and the divergences of the associated class of concave functions

$$\begin{aligned} g_t(u) &= f_t(0) + u f_t^*(0) - f_t(u) = u - \max(u-t, 0) \\ &= \min(u, t) \ , \ t \geq 0 \ . \end{aligned}$$

Let $A_t = \{q > tp\}$ and $A_t^+ = \{q \geq tp\}$. Then the corresponding divergences are, as can easily be verified, the 'Measures of Similarity'

$$\begin{aligned} (Q - tP)^+ (\Omega) &= \sum_{x \in \Omega} \max \left(q(x) - tp(x), 0 \right) \\ &= Q\left(A_t\right) - tP\left(A_t\right) \end{aligned}$$

and the 'Measures of Orthogonality'

$$b(Q, tP) = \sum_{x \in \Omega} \min \left(q(x), tp(x) \right) = Q(A_t^c) + tP(A_t)$$

respectively. Obviously it holds

$$Q(A^c) + tP(A) \geq b(Q, tP) \quad \forall \ A \in \mathfrak{P}(\Omega)$$

with equality iff $A_t \subseteq A \subseteq A_t^+$.

Remark 1: a) Let A be a *test*, i.e. a subset $A \subset \Omega$ so that we decide in favour of the hypothesis Q if $x \in A$ is observed and in favour of P if $x \in A^c$ is observed. Then $P(A)$ and $Q(A^c)$ is the probability of type I error and type II error respectively. Furthermore, let us associate with $t \geq 0$ a probability distribution $\left(\frac{t}{1+t}, \frac{1}{1+t} \right)$ on the set $\{P, Q\} \subset \mathcal{P}$ of the given testing problem (P, Q) . Then $\frac{t}{1+t} P(A) + \frac{1}{1+t} Q(A^c)$ is the *Bayes risk* of the test A with respect to the prior distribution $\left(\frac{t}{1+t}, \frac{1}{1+t} \right)$. Consequently $\frac{1}{1+t} b(Q, tP)$ is the corresponding *minimal Bayes risk*.

b) As can easily be seen from

$$|u - 1| = \max(u - 1, 0) + \max(1 - u, 0)$$

the special case of the term $(Q - tP)^+ (\Omega)$ for $t = 1$ is half of the Total Variation Distance of P and Q , i.e.

$$(Q - P)^+ (\Omega) = V(Q, P)/2 \ .$$

c) The general term $(Q - tP)^+ (\Omega)$ can be described more appropriately in the language of measure theory as the total mass of the positive part of the signed measure $Q - tP$. Note in this context that there is an interesting correspondence between the *Neyman-Pearson* lemma in statistics and the *Radon-Nikodym* theorem in measure theory. For details we refer to *Österreicher & Thaler* (1978).

Representation Theorem (*Feldman* & *Österreicher* (1981)) [1]: Let $f \in \mathcal{F}_0$ and let D_+f denote the right-hand side derivative of the convex function f. Then

$$\begin{aligned} I_f(Q,P) &= \int_0^\infty [\min(1,t) - b(Q,tP)]\, dD_+f(t) \\ &= \int_0^\infty \left[(Q-tP)^+(\Omega) - (P-tP)^+(\Omega)\right] dD_+f(t)\ . \end{aligned}$$

Remark 2: a) The main ingredients of the proof are the following obvious representation of convex functions $f \in \mathcal{F}$ satisfying $f(0), D_+f(0) \in \mathbb{R}$

$$f(u) = f(0) + uD_+f(0) + \int_0^\infty \max(u-t,0)\ dD_+f(t)$$

and *Fubini's* theorem.

b) Provided $\tilde{f}(0) < \infty$ the Representation Theorem for the 'Measure of Orthogonality' has the form

$$I_g(Q,P)\ (= \tilde{f}(0) - I_f(Q,P)\ ) = \int_0^\infty b(Q,tP)\ dD_+f(t)\ .$$

1.2 RISK SETS

Definition: Let $(P,Q) \in \mathcal{P}^2$ be a testing problem. Then the set

$$R(P,Q) = co\left\{(P(A), Q(A^c)) : A \in \mathfrak{P}(\Omega),\ P(A) + Q(A^c) \le 1\right\}$$

is called the *risk set of the testing problem* (P,Q), whereby 'co' stands for 'the convex hull of'.

The bulkiness of a risk set provides a measure for the deviation of P and Q. In fact, the family of risk sets define a uniform structure on the set $\mathcal{P}$. Cf. *Linhart* & *Österreicher* (1985).

Properties of Risk Sets

(R1) $R(P,Q)$ is a convex subset of the triangle $\Delta = \left\{(\alpha,\beta) \in [0,1]^2 : \alpha + \beta \le 1\right\}$ containing the diagonal $D = \left\{(\alpha,\beta) \in [0,1]^2 : \alpha + \beta = 1\right\}$. More specifically it holds

$$D \subseteq R(P,Q) \subseteq \Delta$$

with equality iff $P = Q$ and $P \perp Q$ respectively.

(R2) Let $t \ge 0$ and $b(Q,tP)$ be the $(1+t)$-multiple of the minimal *Bayes* risk with respect to the prior distribution $\left(\frac{t}{1+t}, \frac{1}{1+t}\right)$. Then the risk set $R(P,Q)$ of a testing problem is determined by its family of supporting lines from below, namely

$$\beta = b(Q,tP) - t \cdot \alpha\ ,\quad t \ge 0\ .$$

[1] Cf. also *Österreicher* & *Vajda* (1993)

(R2) and the Representation Theorem imply that an f-divergence $I_f(Q,P)$ depends on the testing problem (P,Q) only via its risk set $R(P,Q)$ [2]. In fact, the following holds.

Remark 3: Let $R(P,Q)$ and $R(\tilde{P},\tilde{Q})$ be two testing problems. Then obviously

$$R(P,Q) \supseteq R(\tilde{P},\tilde{Q}) \quad \Leftrightarrow \quad b(Q,tP) \leq b(\tilde{Q},t\tilde{P}) \ \ \forall\ t \geq 0$$

and hence by virtue of the Representation Theorem

$$R(P,Q) \supseteq R(\tilde{P},\tilde{Q}) \quad \Rightarrow \quad I_f\,(Q,P) \geq I_f(\tilde{Q},\tilde{P})$$

If, in addition, f is strictly convex and $I_f(\tilde{Q},\tilde{P}) < \infty$ then equality holds for the f-divergence iff it holds for the risk sets.

1.3 REFINEMENT OF THE RANGE OF VALUES THEOREM

Let $0 < x < 1$ and let $\alpha \in [0, 1-x]$, $P_\alpha = (\alpha, 1-\alpha)$ and consequently $P_{\alpha+x} = (\alpha + x, 1-(\alpha+x))$. Then the testing problem $(P_{\alpha+x}, P_\alpha)$ has the risk set

$$R(P_\alpha, P_{\alpha+x}) = co\,\{(0,1),(\alpha,1-(x+\alpha)),(1,0)\}\ ,$$

the f-divergence

$$I_f\,(P_{\alpha+x}, P_\alpha) = \alpha f\left(1+\frac{x}{\alpha}\right) + (1-\alpha)\,f\left(1-\frac{x}{1-\alpha}\right)$$

and consequently the Total Variation Distance $V\,(P_{\alpha+x}, P_\alpha) = 2x$.

On the other hand let $P_{(x)} = (0, 1-x, x)$ and $Q_{(x)} = (x, 1-x, 0)$. Then the testing problem $(P_{(x)}, Q_{(x)})$ has the risk set

$$R(P_{(x)}, Q_{(x)}) = co\,\{(0,1)\,,(0,1-x),(1-x,0),(1,0)\}\ ,$$

the f-divergence

$$I_f(P_{(x)}, Q_{(x)}) = \tilde{f}(0)\cdot x$$

and consequently the Total Variation Distance $V(P_{(x)}, Q_{(x)}) = 2\cdot x$ which equals $V(P_{\alpha+x}, P_\alpha)$. Therefore every testing problem

$$(P,Q) \in \mathcal{P} \quad \text{such that} \quad R(P_\alpha, P_{\alpha+x}) \subseteq R(P,Q) \subseteq R(P_{(x)}, Q_{(x)})$$

satisfies

$$V(Q,P) = 2x \qquad \text{and} \qquad I_f(P_\alpha, P_{\alpha+x}) \leq I_f(Q,P) \leq \tilde{f}(0)\cdot x \ \ .$$

[2] This fact and Remark 3 match the properties (a) and (b) of the Characterization Theorem presented in Talk 1.

Now let

$$c_f(x) = \min \{I_f(P_{\alpha+x}, P_\alpha) : 0 \le \alpha \le 1-x\} \ .$$

Then every testing problem (P,Q) with Variation Distance $V(Q,P) = 2x$ satisfies

$$c_f(x) \le I_f(Q,P) \le \tilde{f}(0) \cdot x \ .$$

Since obviously $c_f(0) = f(1) = 0$ we have achieved the following

Refinement of the Range of Values Theorem (*Feldman & Österreicher* (1989)**:** Let the function $c_f : [0,1] \mapsto \mathbb{R}$ be defined as above. Then

$$c_f(V(Q,P)/2) \le I_f(Q,P) \le \tilde{f}(0) \cdot V(Q,P)/2 \ .$$

The function c_f is characteristic for the f-divergence and therefore important for more elaborate investigations.

Properties of c_f : Let $f \in \mathcal{F}_0$. Then the function $c_f : [0,1] \mapsto \mathbb{R}$ has the following properties:

(a)

$$0 \le \tilde{f}(1-x) \le c_f(x) \quad \forall \ x \in [0,1]$$

with strict second inequality for $x \in (0,1]$ if f is strictly convex at 1 . Furthermore $c_f(0) = 0$ and $c_f(1) = \tilde{f}(0)$,

(b) c_f is convex and continuous on $[0,1]$,

(c) c_f is increasing (strictly increasing iff f is strictly convex at 1) and

(d) $c_{f^*} \equiv c_f \le \frac{1}{2} c_{\tilde{f}}$. If, in addition, $f^* \equiv f$ then

$$c_f(x) = \frac{1}{2} c_{\tilde{f}}(x) = (1+x)\, f\left(\frac{1-x}{1+x}\right) \ .$$

2 METRIC f-DIVERGENCES

2.1 GENERATION OF A TOPOLOGY

Let both Ω and $f \in \mathcal{F}_0$ be non-trivial, $\mathcal{P}$ be the set of all probability distributions on Ω , $P \in \mathcal{P}$ and $\varepsilon > 0$,

$$U_f(P,\varepsilon) = \{Q \in \mathcal{P} : I_f(Q,P) < \varepsilon\}$$

be an I_f-ball around P with radius ε ,

$$\mathcal{V}_f(P) = \{U_f(P,\varepsilon) : \varepsilon > 0\}$$

the set of all such balls and

$$\mathcal{U}_f(P) = \{U \subset \mathcal{P} : \exists V \in \mathcal{V}_f(P) \text{ such that } V \subset U\}$$

the associated system of neighbourhoods of P .

Definition: f is said to *generate a topology* $\mathcal{T}_f$ *on* $\mathcal{P}$ if $\{\mathcal{U}_f(P) : P \in \mathcal{P}\}$ generates a topology on $\mathcal{P}$.

In the following we state two criteria under which f generates a topology $\mathcal{T}_f$ on $\mathcal{P}$.

Theorem (*Csiszár* (1967))**:** Let Ω be infinite and $f \in \mathcal{F}_0$. Then f generates a topology on $\mathcal{P}$ iff (i) f is strictly convex at 1 and (iii) $\tilde{f}(0) < \infty$.

Theorem (*Kafka* (1995))**:** Let Ω be finite and $f \in \mathcal{F}_0$. Then f generates a topology on $\mathcal{P}$ iff (i) f is strictly convex at 1 .

Corollary of the Refinement of the Range of Values Theorem: If the properties (i) and (iii) are satisfied then the topology on $\mathcal{P}$ generated by f coincides with the topology induced by the Total Variation Distance.

2.2 BASIC PROPERTIES AND RESULTS

In the sequel we assume Ω to be infinite and $f \in \mathcal{F}_0$ to satisfy (without loss of generality) $f(u) \geq 0 \quad \forall\ u \in [0, \infty)$. We consider the 'Measures of Similarity' and concentrate especially on those (further) properties of the convex function f which make metric divergences possible.

As we know already $I_f(Q, P)$ fulfils the basic property (M1) of a metric divergence, namely

$$I_f(Q, P) \geq 0 \quad \forall\ P, Q \in \mathcal{P} \qquad \text{with equality iff} \qquad Q = P\ , \tag{M1}$$

provided (i) f is strictly convex at 1 .

In addition $I_f(Q, P)$ is symmetric, i.e. satisfies

$$I_f(Q, P) = I_f(P, Q) \quad \forall\ P, Q \in \mathcal{P} \tag{M2}$$

iff (ii) f is $*$-self conjugate, i.e. satisfies $f \equiv f^*$.

As we know from *Csiszár's* Theorem f allows for a topology (namely the topology induced by the Total Variation Distance) iff, in addition to (i), (iii) $\tilde{f}(0) < \infty$ holds. Therefore the properties (i), (ii) and (iii) are crucial for $f \in \mathcal{F}_0$ to allow for the definition of a metric divergence.

Now we state two theorems given in *Kafka*, *Österreicher* & *Vincze* (1991). Theorem 1 offers a class (iii,α), $\alpha \in (0, 1]$ of conditions which are sufficient for guaranteeing the power $[I_f(Q, P)]^\alpha$ to be a distance on $\mathcal{P}$. Theorem 2 determines, in dependence of the behaviour of f in the neighbourhood of 1 and of $g(u) = f(0)(1 + u) - f(u)$ in the neighbourhood of 0 , the maximal α providing a distance.

Theorem 1: Let $\alpha \in (0,1]$ and let $f \in \mathcal{F}_0$ fulfil, in addition to (ii), the condition

(iii,α) the function $h(u) = \frac{(1-u^\alpha)^{\frac{1}{\alpha}}}{f(u)}$, $u \in [0,1)$, is non-increasing.

Then

$$\rho_\alpha(Q,P) = [I_f(Q,P)]^\alpha$$

satisfies the triangle inequality

$$\rho_\alpha(Q,P) \leq \rho_\alpha(Q,R) + \rho_\alpha(R,P) \quad \forall\ P,Q,R \in \mathcal{P}\ . \qquad \text{(M3,}\alpha\text{)}$$

Remark 4: The conditions (ii) and (iii,α) imply both (i) and (iii).

Theorem 2: Let (i),(ii) and (iii) hold true and let $\alpha_0 \in (0,1]$ be the maximal α for which (iii,α) is satisfied. Then the following statement concerning α_0 holds. If for some k_0, k_1, c_0, $c_1 \in (0,\infty)$

$$\begin{aligned} f(0)\cdot(1+u) - f(u) &\sim c_0 \cdot u^{k_0} \\ f(u) &\sim c_1 \cdot |u-1|^{k_1} \end{aligned}$$

then $k_0 \leq 1$, $k_1 \geq 1$ and $\alpha_0 \leq \min(k_0, 1/k_1) \leq 1$.

2.3 EXAMPLES OF METRIC f-DIVERGENCES

In the sequel we investigate the classes of f-divergences discussed in Talk 1. First we check for which parameters the conditions (i), (ii), (iii) are satisfied. The results are summarized in the following table.

Class	(i) given for	(ii) given for	(iii) given for
(I)	$\forall\ \alpha$	$\alpha = 1$	$\alpha = 1$
(II)	$\forall\ \alpha$	$\alpha = 1/2$	$0 < \alpha < 1$
(II')	$\forall\ s$	$\forall\ s$	$0 < s < 1$
(III)	$\forall\ \alpha$	$\forall\ \alpha$	$\forall\ \alpha$
(IV)	$t = 1$	($t = 1$)	$\forall\ t$
(V)	$\forall\ k$	$\forall\ k$	$\forall\ k$
(VI)	$\forall\ \beta$	$\forall\ \beta$	$\forall\ \beta$

Now we proceed with those parameters of the different classes which are not ruled out and present the maximal powers of the f-divergences defining a distance.

(I) The class of χ^α-Divergences

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

From this class only the parameter $\alpha = 1$ provides a distance, namely the Total Variation Distance $V(Q,P)$.

(II) Dichotomy Class

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

From this class only the parameter $\alpha = \frac{1}{2}$ $\left(f^{\frac{1}{2}}(u) = (\sqrt{u} - 1)^2 \right)$ provides a distance, namely the *Hellinger Distance*

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

(II') Symmetrized Dichotomy Class

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

As shown by *Csiszár* & *Fischer* (1962) the parameters $s \in (0,1)$ provide the distances $\left[I_{\tilde{f}^{(s)}}(Q,P) \right]^{\min(s,1-s)}$.

(III) Matusita's Divergences

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

are the prototypes of metric divergences, providing the distances $[I_{f^{\alpha}}(Q,P)]^{\alpha}$.

(IV) Elementary Divergences

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

This class provides only for $t = 1$ ($f_1(u) = \max(u-1, 0)$) a metric, namely $V(Q,P)/2$.

(V) Puri-Vincze Divergences

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

As shown by *Kafka*, *Österreicher* & *Vincze* (1991) this class provides the distances $[I_{\Phi_k}(Q,P)]^{\frac{1}{k}}$.

(VI) Divergences of Arimoto-type

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

As shown by *Österreicher* & *Vajda* (1997) this class provides the distances $\left[I_{f_{\beta}}(Q,P) \right]^{\min(\beta, \frac{1}{2})}$ for $\beta \in (0,\infty)$ and $V(Q,P)/2$ for $\beta = \infty$.

References

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

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

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

Feldman, D. and Österreicher, F. (1989): A note on f-divergences. *Studia Sci. Math. Hungar.*, **24**, 191–200.

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

Kafka, P.: Erzeugen von Topologien und Projektionen in Wahrscheinlichkeitsräumen mittels f-Divergenzen, Diplomarbeit, Salzburg 1995

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

Linhart, J. und Österreicher, F. (1985): Uniformity and distance - a vivid example from statistics. *Int. J. Edu. Sci. Technol.,* **16**/5, 645-649.

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

Österreicher, F. and Thaler, M. (1978): The fundamental Neyman-Pearson lemma and the Radon-Nikodym theorem from a common statistical point of view. *Int. J. Edu. Sci. Technol.,* **9**, 163-176.

Österreicher, F. (1983): Least favourable distributions. *Entry from Kotz-Johnson: Encyclopedia of Statistical Sciences*, Vol. 4, 588-592, John Wiley & Sons, New York

Österreicher, F. (1985) Die Risikomenge eines statistischen Testproblems - ein anschauliches mathematisches Hilfsmittel. *Österreichische Zeitschrift f. Statistik u. Informatik*, 15. Jg. Heft 2-3, 216-228.

Österreicher, F. (1990): The risk set of a testing problem – A vivid statistical tool. In: *Trans. of the 11^{th} Prague Conference on Information Theory*, Academia Prague, Vol. A, 175–188.

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

Österreicher, F. and Vajda, I. (1993): Statistical information and discrimination. *IEEE Trans. Inform. Theory*, **39**/3, 1036–1039.