Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

P. Harremoes and I. Vajda:

Evaluation of Tight Bounds for Divergences

No. 2273 January 2010

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+42) (2) 688 4903
E-mail: utia@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the Institute.

Evaluation of Tight Bounds for Divergences

Peter Harremoës and Igor Vajda

Abstract

The paper develops a general method for evaluation of the joint range of f-divergences for two different functions f. Via topological arguments it demonstrates that the joint range for general distributions equals the joint range achieved by the much simpler and easily tractable distributions on binary or at most quarternary observation spaces. The joint range provides in a straightforward manner achievable upper and lower bounds for one f-divergence in terms of other f-divergence. As well known, such bounds play important role in information theory, identification of parameters and detection of signals.

1. Divergences and divergence statistics

Many of the divergence measures used in the information theory and statistics are of the f-divergence type introduced by Csiszár in 1963 and independently by Ali and Silvey in 1966. Divergences of this type have been systematically studied in detail by Liese and Vajda (1987). Let $f : R_+ \to R$ denote a convex function satisfying $f(1) = 0$. The $f(0)$ is defined as the limit $\lim_{t\to 0} f(t)$. We define $f^*(t) = tf(t^{-1})$. Then f^* is a convex function and $f^*(0)$ is defined as $\lim_{t\to 0} tf(t^{-1}) = \lim_{t\to\infty} \frac{f(t)}{t}$.

Assume that P and Q are absolutely continuous with respect to a measure μ, and that $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$. For arbitrary distributions P and Q the f-divergence $D_f(P, Q) \geq 0$ is defined by the formula

$$D_f(P,Q) = \int_{\{q>0\}} f\left(\frac{p}{q}\right)\, dQ + f^*(0)\, P(q=0)\,. \tag{1.1}$$

For details about this definition and properties of the f-divergences, see Liese and Vajda (1987), Read and Cressie (1087), Liese and Vajda (2006). This definition implies

$$D_f(P,Q) = D_{f^*}(Q,P)\,.$$

Example 1. *The function $f(t) = |t-1|$ defines the L^1-distance*

$$\|P-Q\| = \sum_{j=1}^{k} q_j \left|\frac{p_j}{q_j} - 1\right| = \sum_{j=1}^{k} |p_j - q_j| \quad \textit{(cf. (1.1))} \tag{1.2}$$

which plays an important role in information theory and mathematical statistics (cf. Barron et al. (1992) or Fedotov and Topsoe (2003).

In (1.1) is often taken the convex function f which is one of the power functions ϕ_α of order $\alpha \in \mathbb{R}$ given in the domain $t > 0$ by the formula

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)} \quad \text{when} \quad \alpha(\alpha-1) \neq 0 \tag{1.3}$$

and by the corresponding limits

$$\phi_0(t) = -\ln t + t - 1 \quad \text{and} \quad \phi_1(t) = t \ln t - t + 1. \tag{1.4}$$

The ϕ-divergences

$$D_\alpha(P,Q) \stackrel{def}{=} D_{\phi_\alpha}(P,Q), \quad \alpha \in \mathbb{R} \tag{1.5}$$

based on (1.3) and (1.4) are usually referred to as power divergences of orders α. For details about the properties of power divergences, see Liese and Vajda (2006) or Read and Cressie (1087). Next we mention the best known members of the family of statistics (1.5), with a reference to the skew symmetry $D_\alpha(P,Q) = D_{1-\alpha}(Q,P)$ of the power divergences (1.5).

Example 2. *The χ^2-divergence or quadratic divergence*

$$D_2(P,Q) = D_{-1}(Q,P) = \frac{1}{2}\sum_{j=1}^{k} \frac{(p_j - q_j)^2}{q_j} \tag{1.6}$$

leads to the well known Pearson and Neyman statistics. The information divergence

$$D_1(P,Q) = D_0(Q,P) = \sum_{j=1}^{k} p_j \ln \frac{p_j}{q_j} \tag{1.7}$$

leads to the log-likelihood ratio and reversed log-likelihood ratio statistics. The symmetric Hellinger divergence

$$D_{1/2}(P,Q) = D_{1/2}(Q,P) = H(P,Q)$$

leads to the Freeman–Tukey statistic.

Metric divergences $D_\phi(P,Q)$ must be symmetric in P, Q. The symmetry condition is

$$t\phi(1/t) = \phi(t), \quad t > 0 \quad \text{(cf. Vajda (1972) or Liese and Vajda (2006))}. \tag{1.8}$$

The metric divergences $D_{f_\alpha}(P,Q)$ from Examples 1 - 4 can be obtained by the symmetrization of some ϕ-divergences $D_\phi(P,Q)$ based on the formulas

$$D_\phi(P,Q) = D_{f^{(1)}}(P,(P+Q)/2),$$
$$D_\phi(P,Q) = D_{f^{(2)}}(Q,(P+Q)/2)$$

for the convex functions

$$f^{(1)}(u) = (2-u)\phi\left(\frac{u}{2-u}\right),$$
$$f^{(2)}(u) = u\phi\left(\frac{2-u}{u}\right), \quad 0 < u < 2$$

(cf. (9) in Vajda (1972)). This leads for every convex $\phi(t)$, $t > 0$ to the inverse formulas

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

where

$$\phi^{(1)}(t) = \frac{1+t}{2}\phi\left(\frac{2t}{1+t}\right), \tag{1.9}$$
$$\phi^{(2)}(t) = \frac{1+t}{2}\phi\left(\frac{2}{1+t}\right), \quad t > 0$$

are the convex functions functions studied previously in Vajda (1972) and Vajda (1989). As a result we get the symmetrized version of arbitrary ϕ-divergence

$$D_\phi(P, (P+Q)/2) + D_\phi(Q, (P+Q)/2) = D_{\phi^{(1+2)}}(P, Q) \tag{1.10}$$

for the convex function

$$\phi^{(1+2)}(t) = \phi^{(1)}(t) + \phi^{(2)}(t), \quad t > 0.$$

Since it holds

$$t\phi^{(1)}(1/t) = \phi^{(2)}(t) \quad \text{and} \quad t\phi^{(2)}(1/t) = \phi^{(1)}(t),$$

the symmetry condition (1.8) holds for $\phi^{(1+2)}(t)$ as it is expected.

Example 3. *By definition, for the total variation $f_0 = f_0^{(1)} = f_0^{(2)}$ so that the symmetrized total variation is the total variation itself. For the symmetric Hellinger divergence the corresponding power function $\phi_{1/2}$ leads to new symmetrized function $\phi_{1/2}^{(1+2)}$ with the corresponding $\phi_{1/2}^{(1+2)}$-divergence different from the Hellinger divergence. Therefore the symmetrized Hellinger divergence is not the Hellinger divergence itself. For the quadratic power function ϕ_2 of (1.3) it holds $\phi_2^{(1+2)}(t) = f_{-1}(t)$ where $f_{-1}(t)$. Therefore the LeCam divergence is nothing but the symmetrized Pearson divergence.*

If the original ϕ-divergence is symmetric then its symmetrized version may be identical (e.g. the total variation) or not identical (e.g. the Hellinger divergence). Similarly, the symmetrization may preserve an already symmetrized divergence (see again the total variation) or it may change it (see e.g. the symmetrization of the symmetrized Pearson divergence).

2. Joint range of f-divergences

In this section we are interested in the range of the map $(P,Q) \to (D_f(P,Q), D_g(P,Q))$ where P and Q are probability distributions on the same set.

Definition 1. *A point $(x,y) \in \mathbb{R}^2$ is a* (f,g)-divergence pair *if there exist a Borel space $(\mathcal{X},\mathcal{F})$ with probability measures P and Q such $(x,y) = (D_f(P,Q), D_g(P,Q))$. A (f,g)-divergence pair (x,y) is* achievable in $\mathbb{R}^d$ *if there exist probability vectors $P,Q \in \mathbb{R}^d$ such that*

$$(x,y) = (D_f(P,Q), D_g(P,Q)).$$

Lemma 1. *Assume that*

$$P_0(A) = Q_0(A) = 1$$

and

$$P_1(B) = Q_1(B) = 1$$

and that $A \cap B = \varnothing$. If $P_\alpha = (1-\alpha)P_0 + \alpha P_1$ and $Q_\alpha = (1-\alpha)Q_0 + \alpha Q_1$ then

$$D_f(P_\alpha, Q_\alpha) = (1-\alpha) D_f(P_0,Q_0) + \alpha D_f(P_1,Q_1).$$

Theorem 2. *The set of (f,g)-divergence pairs is convex.*

Proof. Assume that (P,Q) and $\left(\tilde{P},\tilde{Q}\right)$ are two pairs of probability distributions on a space $(\mathcal{X},\mathcal{F})$. Introduce a two-element set $B = \{0,1\}$ and the product space $\mathcal{X}\times B$ as a measurable space. Let ϕ denote projection on B. Now we define a pair $\left(\tilde{P},\tilde{Q}\right)$of joint distribution on $\mathcal{X}\times B$. The marginal distribution of both $\tilde{P}$ is $\tilde{Q}$ on B is $(1-\alpha,\alpha)$. The conditional distributions are given by $P(\cdot \mid \phi = i) = P_i$ and $Q(\cdot \mid \phi = i) = Q_i$ where $i = 0,1$. Then

$$\begin{pmatrix} D_f(P_\alpha,Q_\alpha) \\ D_g(P_\alpha,Q_\alpha) \end{pmatrix} = \begin{pmatrix} (1-\alpha) D_f(P_0,Q_0) + \alpha D_f(P_1,Q_1) \\ (1-\alpha) D_g(P_0,Q_0) + \alpha D_g(P_1,Q_1) \end{pmatrix}$$
$$= (1-\alpha)\begin{pmatrix} D_f(P_0,Q_0) \\ D_g(P_0,Q_0) \end{pmatrix} + \alpha \begin{pmatrix} D_f(P_1,Q_1) \\ D_g(P_1,Q_1) \end{pmatrix}$$
$$= (1-\alpha)\begin{pmatrix} D_f(P,Q) \\ D_g(P,Q) \end{pmatrix} + \alpha \begin{pmatrix} D_f\left(\tilde{P},\tilde{Q}\right) \\ D_g\left(\tilde{P},\tilde{Q}\right) \end{pmatrix}.$$

■

Example 4. *Briët and Harremoës (2009) have studies the relation between total variation with Jensen Shannon divergence. They found that the set of pairs achievable in* $\mathbb{R}^2$ *is not convex but the set of pairs achievable in* $\mathbb{R}^3$ *is convex and equals the set of all* (f, g)*-divergence pairs achievable in the union of all observation spaces.*

Theorem 3. *Any* (f, g)*-divergence pair is a convex combination of two* (f, g)*-divergence pairs, both of them achievable in* $\mathbb{R}^2$*. Consequently, any* (f, g)*-divergence pair is achievable in* $\mathbb{R}^4$*.*

Proof. Let P and Q denote probability measures on the same measurable space. Define the set $A = \{q > 0\}$ and the function $X = p/q$ on A. Then Q satisfies

$$Q(A) = 1 \quad \text{and} \quad X \, dQ \leq 1.$$

Now we fix X and A. The formulas for the divergences become

$$\begin{aligned} D_f(P, Q) &= \int_A f(X) \, dQ + f^*(0) \, P\left(\complement A\right) \\ &= \int_A f(X) \, dQ + f^*(0) \left(1 - \int_A X \, dQ\right) \\ &= \int_A \left(f(X) + f^*(0)(1 - X)\right) dQ \\ &= E\left[f(X) + f^*(0)(1 - X)\right] \end{aligned}$$

and similarly

$$D_g(P, Q) = E\left[g(X) + g^*(0)(1 - X)\right].$$

Hence, the divergences only depend on the distribution of X. Therefore we may without loss of generality assume that Q is a probability measure on $\mathbb{R}_{0,+}$.

Define C as the set of probability measures on $\mathbb{R}_{0,+}$ satisfying $E[X] \leq 1$. Let C^+ be the set of additive measures μ on $\mathbb{R}_{0,+}$ satisfying $\mu(A) \leq 1$ and $\int_A X \, d\mu \leq 1$. Then C^+ is convex and compact under setwise convergence. According to the Choquet–Bishop–de Leeuw theorem any other point in C^+ is the barycenter of a probability measure over such extreme points. In particular an element $Q \in C$ is the barycenter of a probability measure P_{bary} over extreme points of C^+ and these extreme points must in addition be probability measures with P_{bary}-probability 1. Hence $Q \in C$ is a barycenter of a probability measure over extreme points in C.

Let Q be an element in C. Let $A_i, i = 1, 2, 3$ be a disjoint cover of $\mathbb{R}_{0,+}$ and assume that $Q(A_i) > 0$. Then

$$Q = \sum_{i=1}^{3} Q(A_i) \, Q(\cdot \mid A_i).$$

For a probability vector $\lambda = (\lambda_1, \lambda_2, \lambda_2)$ let Q_λ denote the distribution

$$Q_\lambda = \sum_{i=1}^{3} \lambda_i Q\left(\cdot \mid A_i\right).$$

Then Q_λ is element in C if and only if

$$\sum_{i=1}^{3} \lambda_i \int_A X \; dQ\left(\cdot \mid A_i\right) \leq 1. \tag{2.1}$$

An extreme probability vector λ that satisfies (2.1) has one or two of its components equal to 0. Hence, if Q is extreme in C and $A_i, i = 1, 2, 3$ is a disjoint cover of A, then at least one of the three sets satisfies $Q\left(A_i\right) = 0$. Therefore an extreme point $Q \in C$ is of one of the following two types:

1. Q is concentrated in one point
2. Q has support on two points. In this case the inequality $\int_A X \; dQ \leq 1$ holds with equality and $P\left(A\right) = 1$ so that P is absolutely continuous with respect to Q and therefore supported by the same two-element set.

The formulas for divergence are linear in Q. Hence any (f, g)-divergence pair is a the barycenter of a probability measure P_{bary} over pairs generated by extreme distributions $Q \in C$. The extreme distributions of type 2 generate pairs achievable in $\mathbb{R}^2$.

For extreme points Q concentrated in a single point we can reverse the argument at make a barycentric decomposition with respect to P. If an extreme P has a two-point support then Q is absolutely continuous with respect to P and generates a (f, g)-divergence pair achievable in $\mathbb{R}^2$. If P is concentrated in a point then this point may either be identical with the support of Q and the two probability measures are identical, or the support points are different and P and Q are singular but still (P, Q) is supported on two points. Therefore any (f, g)-divergence pair has a barycentric decomposition into pairs achievable in $\mathbb{R}^2$.

Let $\mathbf{y} = (y, z)$ be a (f, g)-divergence pair. Then (y, z) is a the barycenter of (f, g)-divergence pairs achievable in $\mathbb{R}^2$. According to the Carathéodory's theorem barycentric decomposition may be obtained as a convex combination of at most three points $\mathbf{y}_i, \; i = 1, 2, 3$. Assume that all three points have positive weight. Let ℓ_i be the line through $\mathbf{y}$ and $\mathbf{y}_i$. The point $\mathbf{y}$ divides the line ℓ_i in two half-lines ℓ_i^+ and ℓ_i^- where ℓ_i^- is contains $\mathbf{y}_i$. The lines $\ell_i^+, i = 1, 2, 3$ divide $\mathbb{R}^2$ into three sectors, each of them containing one of the points $\mathbf{y}_i, i = 1, 2, 3$. The set of (f, g)-divergence pairs achievable in $\mathbb{R}^3$ is curve-connected so there exist a continuous curve of (f, g)-divergence pairs achievable in $\mathbb{R}^2$ from $\mathbf{y}_1$ to $\mathbf{y}_2$ that must intersect $\ell_1^+ \cup \ell_3^+$ in a point $\mathbf{z}$. If $\mathbf{z}$ lies on ℓ_i^+ then $\mathbf{y}$ is a convex combination of the two points $\mathbf{y}_i$ and $\mathbf{z}$. Hence, any (f, g)-divergence pair is a convex combination of two points that are (f, g)-divergence pairs achievable in $\mathbb{R}^2$. From the construction in the proof of Theorem 2 we see that any (f, g)-divergence pair is achievable in $\mathbb{R}^4$. ■

Remark 1. *We do not have any example of functions (f,g) such that the set of pairs achievable in $\mathbb{R}^3$ is not convex.*

Remark 2. *An f-divergence on a arbitrary σ-algebra can be approximated by the f-divergence on its finite sub-algebras. Any finite σ-algebra is a Borel σ-algebra for discrete space so for probability measures P,Q on a σ-algebra the point $(D_f(P,Q), D_g(P,Q))$ is in the closure of the pairs achievable in $\mathbb{R}^4$. For many function pairs $((f,g))$ the set of pairs achievable in $\mathbb{R}^2$ is closed and then the set of all (f,g)-divergence pairs is closed and contains $(D_f(P,Q), D_g(P,Q))$ even if P,Q are measures on a non-atomic σ-algebra.*

The (f,g)-divergence pair achievable in $\mathbb{R}^2$ can be parametrized as $P=(1-p,p)$ and $Q=(1-q,q)$. If we define $\overline{(1-p,p)}=(p,1-p)$ then $D_f(P,Q)=D_f\left(\overline{P},\overline{Q}\right)$. Hence we may assume without loss of generality assume that $p\leq q$ and just have to determine the image of the simplex $\Delta=\{(p,q)\mid 0\leq p\leq q\leq 1\}$. This result makes it very easy to make a numerical plot of the (f,g)-divergence pair achievable in $\mathbb{R}^2$ and the joint range is just the convex hull.

3. Image of the triangle

In order to determine the image of the triangle Δ we have to check what happens at inner points and what happens at or near the boundary. Most inner points are mapped into inner points of the range. On subsets of Δ where the derivative matrix is non-singular the mapping $(P,Q)\to(D_f,D_g)$ is open according to the open mapping theorem from calculus. Hence, all inner points that are not mapped into interior points of the range must satisfy

$$\begin{vmatrix} \frac{\partial D_f}{\partial p} & \frac{\partial D_g}{\partial p} \\ \frac{\partial D_f}{\partial q} & \frac{\partial D_g}{\partial q} \end{vmatrix} = 0.$$

Depending on functions f and g this equation may be easy or difficult to solve, but in most cases the solutions will lie on a 1-dimensional manifold that will cut the triangle Δ into pieces, such that each piece is mapped isomorphically into subsets of the range of $(P,Q)\to(D_f,D_g)$. Each pair of functions (f,g) will require its own analysis.

The diagonal $p=q$ in Δ is easy to analyze. It is mapped into $(D_f,D_g)=(0,0)$.

Lemma 1. ***(i)*** *If $f(0)=\infty$, and $\lim_{t\to 0}\inf\frac{g(t)}{f(t)}=\beta_0$, then the supremum of*

$$\beta\cdot D_f(P,Q)-D_g(P,Q)$$

over all distributions P,Q is ∞ if $\beta>\beta_0$.

(ii) *If $f^*(0)=\infty$, and $\lim_{t\to\infty}\inf\frac{g(t)}{f(t)}=\beta_0$, then the supremum of*

$$\beta\cdot D_f(P,Q)-D_g(P,Q)$$

over all distributions P, Q *is* ∞ *if* $\beta > \beta_0$.

(iii) *If* $g(0) = \infty$, *and* $\lim_{t\to 0} \sup \frac{g(t)}{f(t)} = \gamma_0$, *then the supremum of*

$$D_g(P, Q) - \gamma D_f(P, Q)$$

over all distributions P, Q *is* ∞ *if* $\gamma < \gamma_0$.

(iv) *If* $g^*(0) = \infty$, *and* $\lim_{t\to\infty} \sup \frac{g(t)}{f(t)} = \gamma_0$, *then the supremum of*

$$D_g(Q, P) - \gamma D_f(Q, P)$$

over all distributions P, Q *is* ∞ *if* $\gamma < \gamma_0$.

Proof. **(i)** Assume that

$$f(0) = \infty \quad \text{and} \quad \liminf_{t\to 0} \frac{g(t)}{f(t)} = \beta_0.$$

The first condition implies

$$D_f((1, 0), (1/2, 1/2)) = \infty$$

and the second condition implies that $g(0) = \infty$ and

$$D_g((1, 0), (1/2, 1/2)) = \infty.$$

We have

$$\frac{D_g((p, 1-p), (1/2, 1/2))}{D_f((p, 1-p), (1/2, 1/2))} = \frac{g(2p)/2 + g(2(1-p))/2}{f(2p)/2 + f(2(1-p))/2} = \frac{g(2p) + g(2(1-p))}{f(2p) + f(2(1-p))}.$$

Let $(t_n)_n$ be a sequence such that $\frac{g(t_n)}{f(t_n)} \to \beta$ for $n \to \infty$. Then

$$\frac{D_g\left(\left(\frac{t_n}{2}, 1 - \frac{t_n}{2}\right), (1/2, 1/2)\right)}{D_f\left(\left(\frac{t_n}{2}, 1 - \frac{t_n}{2}\right), (1/2, 1/2)\right)} \to \beta$$

and the first result follows. The remaining cases **(ii)** - **(iv)** follow by interchanging f and g, and/or replacing f by f^* and g by g^*, using the fact that $\lim_{t\to 0} \inf \frac{g^*(t)}{f^*(t)} = \lim_{t\to 0} \inf \frac{tg\left(t^{-1}\right)}{tf(t^{-1})} = \lim_{t\to\infty} \inf \frac{g(t)}{f(t)}$. ■

Proposition 2. *Assume that f and g are C^2 and that $f''(1) > 0$ and $g''(1) > 0$. Assume that $\lim_{t\to 0} \inf \frac{g(t)}{f(t)} > 0$, and that $\lim_{t\to\infty} \inf \frac{g(t)}{f(t)} > 0$. Then there exists $\beta > 0$ such that*

$$D_g(P,Q) \geq \beta \cdot D_f(P,Q) \tag{3.1}$$

for all distributions P, Q.

Proof. The inequality $\lim_{t\to 0} \inf \frac{g(t)}{f(t)} > 0$ implies that there exist β_0,$t_0 > 0$ such that $g(t) \geq \beta_0 f(t)$ for $t < t_0$. The Inequality $\lim_{t\to\infty} \inf \frac{g(t)}{f(t)} > 0$ implies that there exists $\beta_\infty > 0$ and $t_\infty > 0$ such that $g(t) \geq \beta_\infty f(t)$ for $t > t_\infty$. According to Taylor's formula we have

$$f(t) = \frac{f''(\theta)}{2}(t-1)^2 \quad \text{and} \quad g(t) = \frac{g''(\eta)}{2}(t-1)^2$$

for some θ and η between 1 and t. Hence

$$\frac{g(t)}{f(t)} = \frac{f''(\theta)}{g''(\eta)} \to \frac{f''(1)}{g''(1)} \text{ for } t \to 1.$$

Therefore there there exists $\beta_1 > 0$ and an interval $]t_-, t_+[$ around 1 such that $\frac{g(t)}{f(t)} \geq \beta_1$ for $t \in]t_-, t_+[$. The function $t \to \frac{g(t)}{f(t)}$ is continuous on the compact set $[t_0, t_-] \cup [t_+, t_\infty]$ so it has a minimum $\tilde{\beta} > 0$ on this set. Inequality 3.1 holds for $\beta = \min\left\{\beta_0, \beta_1, \beta_\infty, \tilde{\beta}\right\}$. ■

4. Bounds for power divergences

As an example we shall determine the exact range of a pair of power divergences. We have

$$f(t) = \phi_2(t),$$
$$g(t) = \phi_3(t).$$

In this case we have

$$D_f((p,1-p),(q,1-q)) = \frac{1}{2}\left(\frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q}\right),$$
$$D_g((p,1-p),(q,1-q)) = \frac{1}{6}\left(\left(\frac{p}{q}\right)^3 q + \left(\frac{1-p}{1-q}\right)^3 (1-q) - 1\right).$$

First we determine the image of the triangle. The derivatives are

$$\begin{aligned}
\frac{\partial D_f}{\partial p} &= \frac{1}{2}\frac{\partial}{\partial p}\left(\frac{(p-q)^2}{q}+\frac{(p-q)^2}{1-q}\right)\\
&= \frac{2}{2}\cdot\frac{(p-q)}{(1-q)\,q}\,,\\
\frac{\partial D_f}{\partial q} &= \frac{1}{2}\frac{\partial}{\partial q}\left(\frac{(p-q)^2}{q}+\frac{(p-q)^2}{1-q}-1\right)\\
&= \frac{1}{2}\cdot\frac{(2pq-q-p)\,(p-q)}{(1-q)^2\,q^2}\,,\\
\frac{\partial D_g}{\partial p} &= \frac{1}{6}\frac{\partial}{\partial p}\left(\left(\frac{p}{q}\right)^3 q+\left(\frac{1-p}{1-q}\right)^3(1-q)\right)\\
&= -\frac{3}{6}\cdot\frac{(2pq-q-p)\,(p-q)}{(1-q)^2\,q^2}\,,\\
\frac{\partial D_g}{\partial q} &= \frac{1}{6}\frac{\partial}{\partial q}\left(\left(\frac{p}{q}\right)^3 q+\left(\frac{1-p}{1-q}\right)^3(1-q)-1\right)\\
&= \frac{2}{6}\cdot\frac{\left(\begin{array}{c}pq+p^2+q^2-\\ 3pq^2-3p^2q+3p^2q^2\end{array}\right)(p-q)}{(q-1)^3\,q^3}\,.
\end{aligned}$$

The determinant of derivatives is

$$\begin{vmatrix} \frac{\partial D_f}{\partial p} & \frac{\partial D_g}{\partial p} \\ \frac{\partial D_f}{\partial q} & \frac{\partial D_g}{\partial q} \end{vmatrix} =$$

$$\frac{(p-q)^2}{12q^4\,(1-q)^4}\begin{vmatrix} 2 & 3p+3q-6pq \\ 2pq-q-p & \left(\begin{array}{c}6pq^2-2p^2-2q^2\\ -2pq+6p^2q-6p^2q^2\end{array}\right)\end{vmatrix}$$

$$= -\frac{1}{12}\left(\frac{p-q}{q\,(1-q)}\right)^4.$$

We see that the determinant of derivatives is different from zero for $p \neq q$ so the interior of Δ is mapped one-to-one to the image. Hence we just have to determine the image of points on or near the boundary of Δ.

For $P=(1,0)$ and $Q=(1-q,q)$ we get

$$\begin{aligned}
D_f\,(P,Q) &= \frac{1}{2}\left(q+\frac{q^2}{1-q}\right) = \frac{1}{2}\left(\frac{1}{1-q}-1\right),\\
D_g\,(P,Q) &= \frac{1}{6}\left(\frac{1}{(1-q)^2}-1\right) = \frac{1}{6}\frac{(2-q)\,q}{(1-q)^2}.
\end{aligned}$$

The first equation leads to

$$q = \left(1 - \frac{1}{2D_f + 1}\right)$$

and hence

$$D_g = \frac{2}{3} D_f \left(D_f + 1\right).$$

We have

$$\frac{f\left(t\right)}{g\left(t\right)} = \frac{\frac{t^2-2(t-1)-1}{2}}{\frac{t^3-3(t-1)-1}{6}} \to \infty \text{ for } t \to \infty.$$

All points $\left(0, s\right), s \in \mathbb{R}_{0,+}$ are in the closure of the range of $\left(P, Q\right) \to \left(D_f, D_g\right)$. By combing these two results we the range consists of the point $\left(0, 0\right)$, all points on the curve $\left(x, \frac{2}{3} x \left(x + 1\right)\right), x \in \mathbb{R}_+$, and all point above this curve. Therefore

$$\text{Inf}_{D_2(P,Q)=x} \; D_3(P, Q) = \frac{2}{3} x \left(x + 1\right) \quad \text{for } x \in [0, \infty) \tag{4.1}$$

and

$$\text{Sup}_{D_2(P,Q)=x} \; D_3(P, Q) = \infty \quad \text{for } x \in [0, \infty). \tag{4.2}$$

In other words, for arbitrary $D_2(P, Q)$ and $D_3(P, Q)$ it holds

$$\frac{2D_2(P, Q)\left[D_2(P, Q) + 1\right]}{3} \leq D_3(P, Q) \leq \infty \tag{4.3}$$

and both these bounds are tight. Similar results holds for any pair of power divergences (D_α, D_β), but for other pairs than (D_2, D_3) the computations become much more involved.

Let us conclude the paper by the remark that the Rényi divergences are monotone functions of the power divergences so our results easily translate into the results on the Rényi divergences.

5. Acknowledgement

The authors thank Job Briët and Tim van Erven for comments to a draft of this paper.This reaearch was supported by the European Network of Excellence and by the GAČR grant 102/07/1131. The paper is a preliminary version of the planned submission for the International Symposium on Information Theory organized by the IEEE in Austin, Texas, in June 13-18, 2010.

6. References

Barron, A., Györfi, L. and van der Meulen, E.C. (1992): Distribution estimates consistent in total variation and in two types of information divergence. *IEEE Trans. on Information Therory*, vol. 38, pp. 1437-1454.

Briët, J. and Harremoës, P. (2009): Properties of classical and Shannon Jensen divergence. *Physical Review A,* vol. 79.

Fedotov, A., Harremoës, P. and Topsoe, F. (2003): Refinements of Pinsker's inequality. *IEEE Trans. on Information Therory*, vol. 49, pp. 1491-1498.

Liese, F. and Vajda, I. (1987): *Convex Statistical Distances*. Teubner Verlag, Leipzig.

Liese, F. and Vajda, I. (2006): On divergence and information in statistics and information theory. *IEEE Trans. on Information Therory*, vol. 52, pp. 4394-4412.

Read, T.R.C. and Cressie N. (1988): *Goodness of Fit Statistics for Discrete Multivariate Data*. Springer Verlag, Berlin.

Vajda, I. (1972): On the f-divergence and singularity of probability measures. *Periodica Math. Hungarica*, vol. 2, pp. 223-234.

Vajda, I. (1989): *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, Boston.