

Probability Theory

On the minimum f -divergence for given total variation

Gustavo L. Gilardoni¹

Departamento de Estatística, Universidade de Brasília, Brasília, DF 70910-900, Brazil

Received 17 April 2006; accepted 2 October 2006

Available online 1 December 2006

Presented by Marc Yor

Abstract

We want to find a lower bound for an f -divergence D_f in terms of variational distance V which is best possible for any given V . In other words, we want to find $L_{D_f}(v) = \inf\{D_f(P, Q) : V(P, Q) = v\}$. In this note we solve this problem for any convex f . Although the form of $L_{D_f}(V)$ depends on inverting some expressions which may be difficult in general, simplifications can occur when f has some kind of symmetry. For instance, if D_f is symmetric in the sense that $D_f(P, Q) = D_f(Q, P)$, we show that $L_{D_f}(v) = \frac{2-v}{2} f\left(\frac{2+v}{2-v}\right) - f'(1)v$. For the Kullback–Leibler divergence K we obtain an expression of L_K in terms of the two real branches of Lambert's W function. **To cite this article:** *G.L. Gilardoni, C. R. Acad. Sci. Paris, Ser. I 343 (2006).*

© 2006 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

Résumé

Sur la f -divergence minimale pour variation totale donnée. Pour chaque distance variationnelle V donnée on veut trouver la meilleure borne inférieure possible pour une f -divergence D_f . En d'autres termes, on veut trouver $L_{D_f}(v) = \inf\{D_f(P, Q) : V(P, Q) = v\}$. Dans cette note on résout ce problème pour toute fonction f convexe. Bien que la forme de $L_{D_f}(V)$ dépende de l'inversion de quelques expressions, ce qui peut être difficile en général, des simplifications peuvent se produire quand f a une certaine symétrie. Par exemple, si D_f est symétrique dans le sens : $D_f(P, Q) = D_f(Q, P)$, on prouve que $L_{D_f}(v) = \frac{2-v}{2} f\left(\frac{2+v}{2-v}\right) - f'(1)v$. Pour la divergence de Kullback–Leibler K nous obtenons une expression de L_K à l'aide des deux branches réelles de la fonction W de Lambert. **Pour citer cet article :** *G.L. Gilardoni, C. R. Acad. Sci. Paris, Ser. I 343 (2006).*

© 2006 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

1. Main results

We consider the problem of finding $L_{D_f}(v) = \inf\{D_f(P, Q) : P, Q \text{ such that } V(P, Q) = v\}$, where $f :]0, \infty[\rightarrow \mathbf{R}$ is convex, $f(1) = 0$, P and Q are probability measures with densities p and q with respect to a common dominating measure μ , $V(P, Q) = \int |q - p| d\mu = 2 \sup_A \{|Q(A) - P(A)|\}$ is the total variation (or L^1) distance and $D_f(P, Q) = \int f(q/p) p d\mu$ is the f -divergence between P and Q .

f -divergences were introduced in [3,1] and include many well known measures of discrepancy between probability measures. Besides V and the cross entropy or Kullback–Leibler divergence $K(P, Q) = \int p \log(p/q) d\mu$, all

E-mail address: gilardon@unb.br (G.L. Gilardoni).

¹ Partially supported by a CNPq and two PROCAD/CAPES grants.

of the following are f -divergences: $\chi^2(P, Q) = \int (q - p)^2 / p \, d\mu$, the square of the Hellinger distance $h^2(P, Q) = \int (\sqrt{q} - \sqrt{p})^2 \, d\mu$, Triangular divergence $\Delta(P, Q) = \int (q - p)^2 / (p + q) \, d\mu$, Jensen–Shannon divergence $S(P, Q) = 2^{-1} [K(P, M) + K(Q, M)]$ where $M = (Q + P)/2$ and Jeffrey’s divergence $J(P, Q) = K(P, Q) + K(Q, P)$. For future reference we note here that if $\tilde{f}(u) = f(u) - f'(1)(u - 1)$, then the D_f and the $D_{\tilde{f}}$ divergences are identical. For instance, $K = D_{-\log u} = D_{u^{-1} - \log u}$.

The study of inequalities between information measures in general and between divergences and variational distance in particular has been of interest in several areas including physics, probability, statistics and, of course, information theory. Recently, these kind of results and its relation with Gagliardo–Nirenberg and generalized Sobolev inequalities have been used in order to obtain the decay rate of solutions of nonlinear diffusion equations—see [4] and references therein. In the cross entropy case, $L_K(v)$ was introduced in [6] and is usually called *Vajda’s tight lower bound*. The problem of finding L_K was recently solved by Fedotov, Harremoës and Topsøe [5], who found a parametric expression of the curve $(v, L_K(v))$ in terms of trigonometric hyperbolic functions of $t = t(v) = L'_K(v)$. Our approach here seems to be more intuitive while, at the same time, solving the problem for any f -divergence.

Provided that derivatives and inverses below are conveniently defined, no assumptions other than convexity of f and the usual conventions $f(0) = \lim_{u \downarrow 0} f(u)$, $0 \cdot f(0/0) = 0$ and $0 \cdot f(a/0) = \lim_{\epsilon \downarrow 0} \epsilon f(a/\epsilon) = a \lim_{u \rightarrow +\infty} f(u)/u$ are needed for our results to hold. However, for reason of space, in order to simplify the proofs we will assume throughout this note that f is twice differentiable with $f''(u) > 0$ for $u \neq 1$.

Let $B = \{\omega: q(\omega) \geq p(\omega)\}$. Since $V(P, Q) = 2[Q(B) - P(B)]$, it follows that $0 \leq V(P, Q) \leq 2$ with equality holding respectively if and only if $P = Q$ or $P \perp Q$. This implies that $L_{D_f}(0) = 0$ and

$$L_{D_f}(2) = \int_{\{p>0, q=0\}} f(q/p) p \, d\mu + \int_{\{p=0, q>0\}} f(q/p) (p/q) q \, d\mu = f(0) + \lim_{u \rightarrow +\infty} f(u)/u.$$

Hence, from now on we will consider that $0 < v < 2$. Our main result states that

$$L_{D_f}(v) = v \int_{1/v}^{\infty} k(w) \, dw = \frac{v}{2} \left\{ \frac{f[g_R^{-1}(k(1/v))]}{g_R^{-1}(k(1/v)) - 1} + \frac{f[g_L^{-1}(k(1/v))]}{1 - g_L^{-1}(k(1/v))} \right\} \quad (1)$$

where

$$k^{-1}(t) = \frac{1}{2} \left(\frac{1}{1 - g_L^{-1}(t)} + \frac{1}{g_R^{-1}(t) - 1} \right),$$

$k(u) = (k^{-1})^{-1}(u)$ is the inverse of k^{-1} and g_R^{-1} and g_L^{-1} are the two inverses of $g(u) = (u - 1)f'(u) - f(u)$, respectively to the right and to the left of $u = 1$ (i.e. $g_R^{-1}[g(u)] \equiv u$ for $u \geq 1$ and $g_L^{-1}[g(u)] \equiv u$ for $u \leq 1$), which are well defined because $g'(u) = (u - 1)f''(u)$ is negative for $u < 1$ and positive for $u > 1$. The curve $(v, L_{D_f}(v))$ can be parametrized in terms of $t = t(v) = k(1/v)$. Indeed, define $d(a, v) = af(1 + \frac{v}{2a}) + (1 - a)f(1 - \frac{v}{2(1-a)})$ for $0 < a < 1 - v/2$. We will show that $L_{D_f}(v) = d(a(v), v)$, where $a(v)$ is obtained from the system of equations

$$1 + \frac{v}{2a(v)} = g_R^{-1}(t); \quad 1 - \frac{v}{2(1-a(v))} = g_L^{-1}(t) \quad (2)$$

or equivalently from

$$\frac{1}{v} = \frac{1}{2} \left(\frac{1}{1 - g_L^{-1}(t)} + \frac{1}{g_R^{-1}(t) - 1} \right) = k^{-1}(t); \quad a(v) = \frac{1 - g_L^{-1}(t)}{g_R^{-1}(t) - g_L^{-1}(t)}. \quad (3)$$

A nice geometrical interpretation (cf. Fig. 1) is obtained after noting that $-g(u)$ is the ordinate of the intersection of the tangent to f at u and the straight line $u = 1$. Hence, (2) shows that the two tangents to the curve $(u, f(u))$ which intersect at the point $(1, -t) = (1, -t(v))$ should touch the curve at points with abscises $[1 - \frac{v}{2(1-a(v))}]$ and $[1 + \frac{v}{2a(v)}]$.

Although in general it may be difficult to obtain explicit expressions for the inverses g_R^{-1} , g_L^{-1} and $k = (k^{-1})^{-1}$ involved in (1), considerable simplification can occur when f has some kind of symmetry. For instance, if $f(1 - u) = f(1 + u) - 2f'(1)u$ (i.e. $\tilde{f}(1 + u) = \tilde{f}(1 - u)$) for $0 < u \leq 1$, it can be shown that values of $f(u)$ for $u > 2$ are

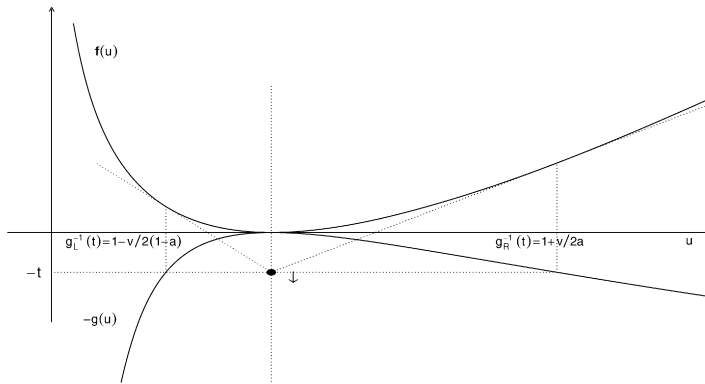


Fig. 1. Geometrical interpretation of the system of Eqs. (2).

irrelevant to determine L_{D_f} . Hence, for the relevant range, we have that $g(1 - u) = g(1 + u)$, $g_R^{-1}(t) - 1 = 1 - g_L^{-1}(t)$, $k^{-1}(t) = 1/[g_R^{-1}(t) - 1] = 1/[1 - g_L^{-1}(t)]$ and $k(u) = g(1 + 1/u)$. Therefore, (1) becomes in this case $L_{D_f}(v) = f(1 + v) - f'(1)v$. For instance, $L_{\chi^2}(v) = v^2$. Maybe more important, suppose that D_f itself is symmetric, in the sense that $D_f(P, Q) = D_f(Q, P)$ for every P and Q . In this case one should have that $f(u) = uf(1/u) + 2f'(1)(u - 1)$ (i.e. $\tilde{f}(u) = u\tilde{f}(1/u)$) and hence

$$g(1/u) = g(u), \quad g_L^{-1}(t) = 1/g_R^{-1}(t), \quad k^{-1}(t) = \frac{1}{2}[g_R^{-1}(t) + 1]/[g_R^{-1}(t) - 1] \quad \text{and}$$

$$k(u) = g[(2u + 1)/(2u - 1)].$$

Substituting into (1) we obtain that $L_{D_f}(v) = \frac{2-v}{2}f(\frac{2+v}{2-v}) - f'(1)v$. For instance, $L_J(v) = v \log \frac{2+v}{2-v}$, $L_{h^2}(v) = 2 - \sqrt{4 - v^2}$, $L_\Delta(v) = \frac{1}{2}v^2$ and $L_S(v) = (2 - v) \log \frac{2-v}{2} + (2 + v) \log \frac{2+v}{2}$.

Consider now *Vajda's tight lower bound* $L_K(v)$. In this case $g(u) = \log u - (u - 1)/u$. Solving $g(u) = t$ is equivalent to $(-1/u)e^{-1/u} = -e^{-(t+1)}$. Hence $g_R^{-1}(t) = -1/W_0(-e^{-(t+1)})$ and $g_L^{-1}(t) = -1/W_{-1}(-e^{-(t+1)})$, where W_0 and W_{-1} are the main and secondary real branches of Lambert's W function (i.e. $W_0(x)e^{W_0(x)} = x$ for $W_0(x) \geq -1$ and $W_{-1}(x)e^{W_{-1}(x)} = x$ for $W_{-1}(x) \leq -1$, cf. [2]). Letting $k(1/v) = t$ in (1) and noting that $f(u) = (u - 1)f'(u) - g(u)$, we obtain after some algebra that $L_K(v)$ can be parametrized as

$$v(t) = 2 \frac{[1 + W_0(-e^{-(1+t)})][1 + W_{-1}(-e^{-(1+t)})]}{W_{-1}(-e^{-(1+t)}) - W_0(-e^{-(1+t)})},$$

$$L_K(v(t)) = \frac{v(t)}{2} [W_0(-e^{-(1+t)}) - W_{-1}(-e^{-(1+t)})] - t.$$

Alternatively, we can write that $L_K(v) = \frac{v}{2}[w_0 - w_{-1}] - k(1/v)$, where

$$w_i = W_i(-e^{-[1+k(1/v)]}),$$

$$k^{-1}(t) = \frac{1}{2} [W_{-1}(-e^{-(1+t)}) - W_0(-e^{-(1+t)})] / [1 + W_0(-e^{-(1+t)})][1 + W_{-1}(-e^{-(1+t)})]$$

and, as before, $k(u) = (k^{-1})^{-1}(u)$.

2. Proofs

We begin by stating precisely the well known fact that in order to find $L_{D_f}(v)$ one needs to consider only binary spaces (cf. [6] or [5] for the relative entropy case).

Proposition 2.1. *Let $d(a, v) = af(1 + \frac{v}{2a}) + (1 - a)f(1 - \frac{v}{2(1-a)})$ and $d(0, v) = f(1 - \frac{v}{2}) + \frac{v}{2} \lim_{u \rightarrow +\infty} \frac{f(u)}{u}$. Then, for every $0 < v < 2$, $L_{D_f}(v) = \inf_{a: 0 \leq a < 1-v/2} d(a, v)$.*

Proof. First, consider a sample space $\Omega = \{0, 1\}$ and probability measures P and Q with $p(0) = (1 - a)$, $p(1) = a$, $q(0) = (1 - a - v/2)$ and $q(1) = (a + v/2)$, so that $V(P, Q) = v$ and $D_f(P, Q) = d(a, v)$. This shows that $L_{D_f}(v) \leq \inf_{a: 0 \leq a < 1 - v/2} d(a, v)$. To show the reversed inequality let $B = \{\omega \in \Omega: q(\omega) \geq p(\omega)\}$, $v = V(P, Q) = 2[Q(B) - P(B)]$ and use Jensen's inequality to obtain that $D_f(P, Q) \geq d(P(B), v)$. \square

Since

$$\frac{\partial^2 d}{\partial a^2}(a, v) = \frac{v^2}{4a^3} f''\left(1 + \frac{v}{2a}\right) + \frac{v^2}{4(1-a)^3} f''\left(1 - \frac{v}{2(1-a)}\right) \geq 0,$$

the map $a \mapsto d(a, v)$ is convex. Hence, if we can solve for $a = a(v)$ the equation

$$\frac{\partial d}{\partial a}(a, v) = f\left(1 + \frac{v}{2a}\right) - \frac{v}{2a} f'\left(1 + \frac{v}{2a}\right) - f\left(1 - \frac{v}{2(1-a)}\right) - \frac{v}{2(1-a)} f'\left(1 - \frac{v}{2(1-a)}\right) = 0, \quad (4)$$

we can then write that $L_{D_f}(v) = d(a(v), v)$. Since (4) implies that $g\left(1 + \frac{v}{2a(v)}\right) = g\left(1 - \frac{v}{2(1-a(v))}\right)$, defining $t = t(v) = g\left(1 + \frac{v}{2a(v)}\right) = g\left(1 - \frac{v}{2(1-a(v))}\right)$ we obtain (2).

Proposition 2.2. $L_{D_f}(v)$ satisfies the differential equation $vL'_{D_f}(v) - L_{D_f}(v) = t(v) = k(1/v)$ and hence is given by Eq. (1).

Proof. First, use that $\frac{\partial d}{\partial a}(a(v), v) = 0$, $f'(u) = \frac{f(u)+g(u)}{u-1}$ and the definition of $t(v)$ to obtain that

$$\begin{aligned} L'_{D_f}(v) &= \frac{\partial d}{\partial a}(a(v), v)a'(v) + \frac{\partial d}{\partial v}(a(v), v) = \frac{1}{2}f'\left(1 + \frac{v}{2a(v)}\right) - \frac{1}{2}f'\left(1 - \frac{v}{2(1-a(v))}\right) \\ &= \frac{a(v)}{v} \left[f\left(1 + \frac{v}{2a(v)}\right) + t(v) \right] + \frac{1-a(v)}{v} \left[f\left(1 - \frac{v}{2(1-a(v))}\right) + t(v) \right] = \frac{L_{D_f}(v)}{v} + \frac{t(v)}{v}. \end{aligned}$$

Since the homogeneous equation has general solution $L_{D_f}(v) = cv$, taking $c = c(v)$ and doing variation of the constant we obtain that $v[c'(v)v + c(v)] - c(v)v = c'(v)v^2 = t(v)$, hence that $c(v) = \int_0^v u^{-2}t(u) du$ and $L_{D_f}(v) = vc(v) = v \int_0^v u^{-2}t(u) du = v \int_0^v u^{-2}k(1/u) du$. Substituting $u = 1/w$ we obtain the first equality in (1). Finally, the last equality in (1) is obtained from Proposition 2.1 after noting that the rightmost term in (1) equals $d(a(v), v)$, where v and $a(v)$ are given in (3). \square

Acknowledgements

I thank P.N. Rathie and C. Caiado for calling my attention about Lambert's W and L.A. Maia for encouragement and helpful discussions.

References

- [1] S.M. Ali, S.D. Silvey, A general class of coefficients of divergence of one distribution from another, *J. Roy. Statist. Soc. Ser. B* 28 (1966) 131–142.
- [2] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, D.E. Knuth, On the Lambert W function, *Adv. Comput. Math.* 5 (4) (1996) 329–359.
- [3] I. Csizsár, Information-type measures of difference of probability distributions and indirect observations, *Studia Sci. Math. Hungar.* 2 (1967) 299–318.
- [4] M. Del Pino, J. Dolbeault, Nonlinear diffusions and optimal constants in Sobolev type inequalities: asymptotic behaviour of equations involving the p -Laplacian, *C. R. Acad. Sci. Paris, Ser. I* 334 (5) (2002) 365–370.
- [5] A. Fedotov, P. Harremoës, F. Topsøe, Refinements of Pinsker's inequality, *IEEE Trans. Inform. Theory* 49 (2003) 1491–1498.
- [6] I. Vajda, Note on discrimination information and variation, *IEEE Trans. Inform. Theory* 16 (1970) 771–773.