THE APPROXIMATION OF CSISZÁR f–DIVERGENCE FOR ABSOLUTELY CONTINUOUS MAPPINGS

S. S. DRAGOMIR, V. GLUŠČEVIĆ, AND C. E. M. PEARCE

Abstract. We consider the approximation of the Csiszár f–divergence of two probability distributions over a finite alphabet when the first (or second) derivative of f is absolutely continuous and the second (respectively third) essentially bounded. The approximants are suggested by results from numerical integration theory, which also provides the means of deriving error bounds. Tools are also developed for obtaining tight bounds.

1. Introduction

A common situation in information theory concerns an alphabet $\{a_i\}_{i=1}^n$ on which two probability distributions $p = (p_1, \ldots, p_n)$, $q = (q_1, \ldots, q_n)$ are defined. For example, p_i, q_i may represent respectively the frequency with which the symbol a_i occurs before and after transmission through a noisy channel.

A variety of divergence or discrimination measures are in common use for quantifying the difference between such a pair p, q. The authors give a number of these in [4]. Two of these, that we shall invoke shortly in this introduction, are the chi–squared and (absolute) chi–cubed measures, which are given by

$$D_{|\chi|^m}(p, q) := \sum_{i=1}^{n} \frac{|p_i - q_i|^m}{q_i^{m-1}}$$

for $m = 2$ and $m = 3$ respectively. The diversity of the measures appearing in the literature is remarkable. Csiszár [1]–[3] has systematised the field by showing that a majority of these can be represented in the form

$$I_f(p, q) = \sum_{i=1}^{n} q_i f(p_i/q_i)$$

for a suitable choice of function f. Typically but not invariably $f(1) = 0$ and f is nonnegative and strictly convex. With these assumptions, I_f automatically carries the physically natural properties that $I_f(p, p) = 0$ and $I_f(p, q) > 0$ for p and q distinct.

A basic question is that of approximation: how closeness of f and g is reflected in closeness of I_f and I_g. As an illustrative example, the following was considered in [5].

Suppose that there exist distinct real numbers r, R with

$$0 < r \le p_i/q_i \le R < \infty \text{ for all } i \in \{1, ..., n\}. \tag{1.1}$$

1991 *Mathematics Subject Classification.* 94A17, 26D15.

Key words and phrases. Csiszár f–divergence, Iyengar inequality, trapezoid inequality, divergence measures, absolute continuity.

We suppose also that $f : [r, R] \to \mathbf{R}$ is such that $f^{'}$ is absolutely continuous on $[r, R]$ and $f^{''} \in L_\infty[r, R]$, that is,

$$\left\| f^{''} \right\|_\infty := ess \sup_{t \in [r,R]} \left| f^{''}(t) \right| < \infty.$$

Define $g : [r, R] \to \mathbf{R}$ by

$$g(u) := f(1) + (u-1) f^{'}(1).$$

Then

$$|I_f(p,q) - I_g(p,q)| \le \frac{(R-r)^2}{8} \left\| f^{''} \right\|_\infty .$$

Similarly if $f^{''}$ is absolutely continuous on $[r, R]$ and $f^{'''} \in L_\infty[r, R]$, and we define $g : [r, R] \to \mathbf{R}$ by

$$g(u) := f(1) + (u-1) f^{'}(1) + \frac{1}{2}(u-1)^2 f^{''}(1),$$

then

$$|I_f(p,q) - I_g(p,q)| \le \frac{(R-r)^3}{48} \left\| f^{'''} \right\|_\infty .$$

These examples are prompted naturally by consideration of Taylor series. However, even within the same framework of assumptions on derivatives, substantially stronger results exist. The following were established in [5] with the choice $g = f^*$, where

$$f^*(u) := f(1) + (u-1) f^{'}\left(\frac{1+u}{2}\right).$$

Theorem A. *Suppose that (1.1) holds and assume that $f : [r, R] \to \mathbf{R}$ is such that $f^{'}$ is absolutely continuous on $[r, R]$ and $f^{''} \in L_\infty[r, R]$. Then*

$$\begin{aligned} |I_f(p,q) - I_{f^*}(p,q)| &\le \frac{1}{4} \left\| f^{''} \right\|_\infty D_{\chi^2}(p,q) \\ &\le \frac{1}{4} \left\| f^{''} \right\|_\infty (R-1)(1-r) \\ &\le \frac{1}{16} \left\| f^{''} \right\|_\infty (R-r)^2. \end{aligned}$$

Theorem B. *Suppose that (1.1) holds and assume that $f : [r, R] \to \mathbf{R}$ is such that $f^{''}$ is absolutely continuous on $[r, R]$ and $f^{'''} \in L_\infty[r, R]$. Then*

$$\begin{aligned} |I_f(p,q) - I_{f^*}(p,q)| &\le \frac{1}{24} \left\| f^{'''} \right\|_\infty D_{|\chi|^3}(p,q) \\ &\le \frac{1}{24} \left\| f^{'''} \right\|_\infty \frac{(R-1)(1-r)}{R-r} \left[(1-r)^2 + (R-1)^2 \right] \\ &\le \frac{1}{192} \left\| f^{'''} \right\|_\infty (R-r)^3. \end{aligned}$$

The constants are best–possible.

In this article, we show that these ideas can be pushed further, again without extending the raft of assumptions. We assume (1.1) throughout without further comment and keep the assumption that r, R are distinct, which is easily seen to entail that $r < 1 < R$. Suppose $f^{\dagger} : [r, R] \to \mathbf{R}$ is given by

$$f^{\dagger}(u) := f(1) + \frac{u-1}{2} f^{'}(u).$$

In Section 2 we derive Theorems 1 and 2, which are versions of Theorems A and B with $f^{\dagger}$ in place of f^{*}. Theorem 1 has a tighter error bound than Theorem A and Theorem B a less tight bound than Theorem B. The first bound in Theorem 1 is quite complicated. In Section 3 we introduce one– and two–point distributions and show how these may be used to clarify the bounds in Theorem 1 and indeed also to produce best–possible functional improvements of the bounds, in terms of r and R only, in both theorems. These take into account the detailed structure of the function f. In Section 4, we codify these ideas into a pair of connected geometric constructions. We conclude in Section 5 by illustrating the ideas involved by reference to some choices of f in common use.

2. Basic theorems

The analogue of Theorem A we are about to derive is based on the Iyengar inequality [7], which states the following.

Theorem C. *Suppose $g : [a, b] \to \mathbf{R}$ is absolutely continuous on $[a, b]$ and $g^{'} : [a, b] \to \mathbf{R}$ is essentially bounded, that is, $g^{'} \in L_{\infty}[a, b]$. Then*

$$\left| \int_a^b g(t)dt - \frac{1}{2}(b-a)[g(a) + g(b)] \right| \leq \frac{1}{4} \left\| g^{'} \right\|_{\infty} (b-a)^2 - \frac{1}{4 \left\| g^{'} \right\|_{\infty}} [g(b) - g(a)]^2.$$

We shall also make use of Proposition 1 of [5], which provides the following.

Proposition A. *Suppose that (1.1) is satisfied with $r < R$ and that $m \geq 1$. Then*

$$D_{|\chi|^m}(p, q) \leq \frac{(R-1)(1-r)}{R-r} \left[(1-r)^{m-1} + (R-1)^{m-1} \right] \leq \left(\frac{R-r}{2} \right)^m.$$

The first inequality is an equality if and only if p, q form a boundary pair with respect to r and R, that is, for each i either $p_i/q_i = r$ or $p_i/q_i = R$. The second inequality is an equality if and only if $R + r = 2$, that is, r and R are equidistant from unity.

We now proceed to our first basic result. We shall make use of $f_0 : [r, R] \to \mathbf{R}$ given by

$$f_0(u) := \left[f^{'}(u) - f^{'}(1) \right]^2.$$

Theorem 1. *Suppose $f : [r, R] \to \mathbf{R}$ with f' absolutely continuous on $[r, R]$ and $f'' \in L_\infty[r, R]$. Then*

$$\left|I_f(p,q) - I_{f^\dagger}(p,q)\right| \leq \frac{1}{4}\left\|f''\right\|_\infty D_{\chi^2}(p,q) - \frac{1}{4\left\|f''\right\|_\infty} I_{f_0}(p,q) \tag{2.1}$$
$$\leq \frac{1}{4}\left\|f''\right\|_\infty D_{\chi^2}(p,q)$$
$$\leq \frac{1}{4}\left\|f''\right\|_\infty (R-1)(1-r)$$
$$\leq \frac{1}{16}\left\|f''\right\|_\infty (R-r)^2.$$

Proof. We choose $a = 1$, $b = x$ and $g = f'$ in Theorem C to obtain

$$\left|\int_1^x f'(t)dt - \frac{1}{2}(x-1)\left[f'(1) + f'(x)\right]\right| \leq \frac{1}{4}\left\|f''\right\|_\infty (x-1)^2 - \frac{1}{4\left\|f''\right\|_\infty} f_0(x),$$

or equivalently

$$\left|f(x) - f(1) - \frac{1}{2}f'(1)(x-1) - \frac{1}{2}f'(x)(x-1)\right| \leq \frac{1}{4}\left\|f''\right\|_\infty (x-1)^2 - \frac{1}{4\left\|f''\right\|_\infty} f_0(x)$$

for all $x \in [r, R]$.

If we choose $x = p_i/q_i$, multiply by q_i and sum over i from 1 to n, we derive the first inequality in (2.1) *via* the extended triangle inequality. The second inequality is immediate, since f_0 and so also I_{f_0} is nonnegative. The remaining inequalities are given by Proposition A with $m = 2$. □

Corollary 1. *Suppose the assumptions of Theorem 1 hold. If $\varepsilon > 0$ and*

$$0 \leq R - r \leq 4 \cdot \sqrt{\varepsilon / \left\|f''\right\|_\infty},$$

then

$$\left|I_f(p,q) - I_{f^\dagger}(p,q)\right| \leq \varepsilon.$$

The following corollary emphasizes the approximation aspect of Theorem 1 for distributions p and q which are close.

Corollary 2. *Let $f : [0, 2] \to \mathbf{R}$ be such that $f' : [0, 2] \to \mathbf{R}$ is absolutely continuous and $f'' \in L_\infty[0, 2]$. If $\eta \in (0, 1)$ and $p(\eta)$, $q(\eta)$ satisfy*

$$\left|\frac{p_i(\eta)}{q_i(\eta)} - 1\right| \leq \eta \quad \text{for all } i \in \{1, ..., n\}, \tag{2.2}$$

then

$$I_f\left(p(\eta), q(\eta)\right) = I_{f^\dagger}\left(p(\eta), q(\eta)\right) + R_f(p, q, \eta) \tag{2.3}$$

and the remainder $R_f(p,q,\eta)$ satisfies

$$R_f(p,q,\eta) \le \frac{1}{4}\left\|f''\right\|_\infty \eta^2.$$

Proof. Choose $r = 1-\eta$, $R = 1+\eta$ in Theorem 1. □

Our other basic theorem makes use of the trapezoid inequality

$$\left|\int_a^b g(t)\,dt - \frac{1}{2}(b-a)\left[g(a)+g(b)\right]\right| \le \frac{1}{12}(b-a)^3\left\|g''\right\|_\infty \tag{2.4}$$

from numerical integration, which holds provided $g'' \in L_\infty[a,b]$.

Theorem 2. *If $f:[r,R]\to\mathbf{R}$ with f'' absolutely continuous on $[r,R]$ and $f''' \in L_\infty[r,R]$, then*

$$\begin{aligned}(2.5)\qquad \left|I_f(p,q) - I_{f^\dagger}(p,q)\right| &\le \frac{1}{12}\left\|f'''\right\|_\infty D_{|\chi|^3}(p,q)\\ &\le \frac{1}{24}\left\|f'''\right\|_\infty \frac{(R-1)(1-r)}{R-r}\left[(1-r)^2+(R-1)^2\right]\\ &\le \frac{1}{96}\left\|f'''\right\|_\infty (R-r)^3.\end{aligned}$$

The constants on the right are best–possible.

Proof. The first inequality is derived from (2.4) along the same lines as the first inequality in Theorem 1. The remaing inequalities follow from Proposition A with $m = 3$.

For $f(u) = |u-1|^3$ we have $I_f(p,q) = D_{|\chi|^3}(p,q)$. Also $f^\dagger(u) = (3/2)f(u)$, so as I is linear in f we have

$$|I_f(p,q) - I_{f^\dagger}(p,q)| = \frac{1}{2}D_{|\chi|^3}(p,q).$$

Since $\left\|f'''\right\| = 6$, the first inequality in (2.5) is thus an equality for this choice of f and the corresponding constant is best–possible. That the following constants are best–possible is inherited from Proposition A. □

Corollary 3. *Let f be as in Theorem 2. If $\varepsilon > 0$ and*

$$0 \le R - r \le 2\cdot\sqrt[3]{12\varepsilon/\left\|f'''\right\|_\infty},$$

then

$$\left|I_f(p,q) - I_{f^\dagger}(p,q)\right| \le \varepsilon.$$

Also, the following approximation result holds.

Corollary 4. *Let $f : [0, 2] \to \mathbf{R}$ with f'' absolutely continuous on $[0, 2]$ and $f''' \in L_\infty[0, 2]$. If $\eta \in (0, 1)$ and $p(\eta)$, $q(\eta)$ are such that (2.2) holds, then we have the representation (2.3) and the remainder $R_f(p, q, \eta)$ satisfies*

$$|R_f(p, q, \eta)| \leq \frac{1}{12} \left\| f''' \right\|_\infty \eta^3.$$

Remark 1. *The last bound in Theorem 2 is tighter than that of Theorem 1 if*

$$\frac{6 \left\| f'' \right\|_\infty}{\| f''' \|_\infty} > R - r,$$

while the reverse is true if

$$\frac{6 \left\| f'' \right\|_\infty}{\| f''' \|_\infty} < R - r.$$

As we shall see in Section 5, both possibilities can arise in practice. In the examples we consider, Theorem 2 gives the better bound when r/R is large and Theorem 1 when it is small.

3. One– and two–point distributions

Suppose we wish to obtain the error bound involved in estimating $I_f(p, q)$ by $I_g(p, q)$ for some function g. Since

$$I_f(p, q) - I_g(p, q) = I_{f-g}(p, q),$$

we wish to find $\sup |I_h(p, q)|$, where $h = f - g$ and the supremum is taken over all n–point probability distribution pairs (p, q) satisfying (1.1). In this section we approach this question directly and establish some basic results. We shall find it convenient for clarity and succinctness to adopt the notation $I_{h,n}(p, q)$ and to write $u_i := p_i/q_i$ for $i \in \{1, 2, \ldots, n\}$ in this and the following section.

Proposition 1. *Let $n \geq 1$ and assume p, q are n–point distributions all of whose components are nonzero.*
(a) There exist k–point distributions p^U, q^U, where k takes one of the values 1 or 2 and p^U, q^U depend on p and q, such that

$$p_i^U / q_i^U \in \{u_1, \ldots, u_n\} \text{ for } i = 1, \ldots, k$$

and

$$I_{h,k}(p^U, q^U) \geq I_{h,n}(p, q).$$

(b) There exist k–point distributions p^L, q^L, where k takes one of the values 1 or 2 and p^L, q^L depend on p and q, such that

$$p_i^L / q_i^L \in \{u_1, \ldots, u_n\} \text{ for } i = 1, \ldots, k$$

and

$$I_{h,k}(p^L, q^L) \leq I_{h,n}(p, q).$$

Proof. Consider the first half of the enunciation. If there are j distinct values $u_{(1)}, u_{(2)}, \ldots, u_{(j)}$ $(1 \leq j \leq n)$, then for each such we may sum the associated values q_i to obtain $q_1^{(j)}, q_2^{(j)}, \ldots, q_j^{(j)}$. Likewise we derive $p_\ell^{(j)}$ associated with $u_{(\ell)}$ $(1 \leq \ell \leq j)$. We have at once that $p_\ell^{(j)}/q_\ell^{(j)} = u_{(\ell)}$, and that $p^{(j)} = (p_1^{(j)}, \ldots, p_j^{(j)})$ and

$q^{(j)} = (q_1^{(j)}, \ldots, q_j^{(j)})$ are j–point probability distributions for which $I_{h,j}(p^{(j)}, q^{(j)})$ has the same value as $I_{h,n}(p, q)$.

To derive the desired result (a) by mathematical induction, we need to show that if m is such that $3 \le m \le j$ then there exist $(m-t)$–point probability distributions $p^{(m-t)}$, $q^{(m-t)}$ (with t equal to 1 or 2) depending on $p^{(m)}$ and $q^{(m)}$ with

$$p_\ell^{(m-t)}/q_\ell^{(m-t)} = u_{(\ell)} \quad (1 \le \ell \le m-t)$$

and

$$I_{h,m-t}(p^{(m-t)}, q^{(m-t)}) \ge I_{h,j}(p^{(m)}, q^{(m)}).$$

We may without loss of generality assume that there are at least three distinct values u_i, for otherwise there is nothing to prove.

To achieve the induction, we shall show that such a reduction from m–point support to $(m-t)$–point support can be brought about by replacing the last three components of $p^{(m)}$ by two suitably chosen components and a zero or one suitably chosen component and two zeros, with a corresponding replacement in $q^{(m)}$, the zeros being in the same position or positions.

For notational convenience, put $v_1 := u_{(m-2)}$, $v_2 := u_{(m-1)}$, $v_3 := u_{(m)}$. With relabelling if necessary, we may assume $v_1 < v_2 < v_3$. Likewise we put $p_{m-2}^{(m)} = \rho_1$, $p_{m-1}^{(m)} = \rho_2$, $p_m^{(m)} = \rho_3$ and $q_{m-2}^{(m)} = \sigma_1$, $q_{m-1}^{(m)} = \sigma_2$, $q_m^{(m)} = \sigma_3$, so that $\rho_i/\sigma_i = v_i$ for $i = 1, 2, 3$. Define

$$\lambda := \frac{v_3 - v_2}{v_3 - v_1},$$

so that $0 < \lambda, 1-\lambda < 1$ and $v_2 = \lambda v_1 + (1-\lambda)v_3$. We address in turn three possible cases.

(i) If $h(v_2) \le \lambda h(v_1) + (1-\lambda)h(v_3)$, then

$$\sum_{i=1}^{3} \sigma_i h(v_i) \le \sum_{i=1,3} \sigma_i' h(v_i),$$

where $\sigma_1' = \sigma_1 + \lambda\sigma_2$ and $\sigma_3' = \sigma_3 + (1-\lambda)\sigma_2$. Note that $\sum_{i=1}^{3} \sigma_i = \sum_{i=1,3} \sigma_i'$. If we define $\rho_i' = \sigma_i' v_i$ for $i = 1, 3$, then $\rho_i'/\sigma_i' = v_i$ $(i = 1, 3)$ and

$$\sum_{i=1,3} \rho_i' = \sum_{i=1,3} \sigma_i v_i + \sigma_2[\lambda_1 v_1 + (1-\lambda)v_3] = \sum_{i=1}^{3} \sigma_i v_i = \sum_{i=1}^{3} \rho_i.$$

This shows that the reduction can be effected in this case with $t = 1$.

(ii) Next we suppose

$$h(v_2) > \lambda h(v_1) + (1-\lambda)h(v_3) \tag{3.1}$$

with

$$\sigma_1/\lambda \le \sigma_3/(1-\lambda). \tag{3.2}$$

Put

$$\sigma_2' := \sigma_2 + \sigma_1/\lambda, \quad \sigma_3' := \sigma_3 - \frac{1-\lambda}{\lambda}\sigma_1.$$

Then $\sigma_2' > 0$ and by (3.2) $\sigma_3' \geq 0$. Further $\sum_{i=2,3} \sigma_i' = \sum_{i=1}^{3} \sigma_i$. Also by (3.1),

$$\sum_{i=2,3} \sigma_i' h(v_i) > \frac{\sigma_1}{\lambda}[\lambda h(v_1) + (1-\lambda)h(v_3)] + \sigma_2 h(v_2) + \left[\sigma_3 - \frac{1-\lambda}{\lambda}\sigma_1\right] h(v_3) = \sum_{i=1}^{3} \sigma_i h(v_i).$$

If we define $\rho_i' = \sigma_i' v_i$ for $i = 2, 3$, then $\rho_i'/\sigma_i' = v_i$ $(i = 2, 3)$ and much as above

$$\sum_{i=2,3} \rho_i' = \sum_{i=1}^{3} \rho_i.$$

If (3.2) holds with equality, then $\sigma_3' = 0$ and the reduction holds with $t = 2$. Otherwise it holds with $t = 1$.

(iii) Finally we have the possibility that (3.1) holds with

$$\sigma_1/\lambda > \sigma_3/(1-\lambda). \tag{3.3}$$

We may argue as in (ii), this time starting with

$$\sigma_1' := \sigma_1 - \frac{\lambda}{1-\lambda}\sigma_3, \quad \sigma_2' := \sigma_2 + \frac{1}{1-\lambda}\sigma_3.$$

Then σ_1', σ_2' are positive and

$$\sum_{i=1,2} \sigma_i' = \sum_{i=1^3} \sigma_i.$$

By (3.1),

$$\sum_{i=1,2} \sigma_i' h(v_i) = \sum_{i=1,2} \sigma_i h(v_i) + \frac{\sigma_3}{1-\lambda}[h(v_2) - \lambda h(v_1)] > \sum_{i=1}^{3} \sigma_i h)v_i).$$

If $\rho_i' = \sigma_i' v_i$ for $i = 1, 2$, then $\rho_i'/\sigma_i' = v_i$ $(i = 1, 2)$ and

$$\sum_{i=1,2} \rho_i' = \sum_{i=1,2} \sigma_i' v_i = \sum_{i=1,2} \sigma_i v_i + \frac{\sigma_3}{1-\lambda}[v_2 - \lambda v_1] = \sum_{i=1}^{3} \sigma_i v_i = \sum_{i=1}^{3} \rho_i.$$

Thus we have a reduction with $t = 1$.

This completes the proof of part (a). Part (b) follows by applying part (a) to the function $-h$. □

Suppose h is bounded on $[r, R]$. By letting p_i, q_i tend to zero for different choices of i successively while keeping $p_\ell/q_\ell \in [r, R]$ for all $\ell \in \{1, 2, \ldots, n\}$, we can obtain one– and two–point distributions satisfying (1.1) as limiting cases of n–point distributions. With this convention, the following result is natural.

Proposition 2. *Suppose h is continuous and bounded on $[r, R]$. Then $I_{h,n}$ achieves its supremum and infimum over n–point distributions p, q satisfying (1.1). These are realised by one– or two–point distributions.*

Proof. The first part is immediate. The second follows *via* Proposition 1, by relating extremum–achieving distributions to I_h evaluated at one– or two–point distributions at which I_h is dominating (in the supremum case) or dominated (in the infimum case). □

Corollary 5. *The supremum and infimum of I_h subject to (1.1) take one of the forms*

$$qh(u) + (1-q)h\left(\frac{1-qu}{1-q}\right), \quad h(1), \tag{3.4}$$

where $u \in [r, R]$.

Proof. The first form in (3.4) is immediate, since if $p/q = u$, then

$$\frac{1-p}{1-q} = \frac{1-qu}{1-q}.$$

The second is trivial, since $p/q = 1$ when p and q are one–point distributions. □

The ideas of this section provide the means for obtaining tight bounds on $|I_h|$ in terms of r, R. This we pursue in the following section. In practice the calculations can be quite intricate even when h has a relatively simple functional form, although they are suited to efficient numerical implementation, such as by bifurcation search. For this reason, it is very convenient to have even the largest and second largest bounds on the right–hand sides of (2.1) and (2.5).

Finite–point distributions have a special role in extremal theory. For a general discussion, the reader is referred to [6].

4. Evaluating extrema

We now draw together the ideas of the preceding section to codify the treatment of some broad classes of function h. For notational convenience we introduce

$$F(x,y) := \frac{y-1}{y-x}h(x) + \frac{1-x}{y-x}h(y),$$

which gives the value of I_h for two–point distributions p, q with support at $u = x, y$. We assume throughout that (1.1) applies and that $x, y \in [r, R]$.

Theorem 3. *Suppose r_T, R_T satisfy $r \le r_T < 1 < R_T \le R$ and the line joining $(r_T, h(r_T))$ to $(R_T, h(R_T))$ lies strictly above the graph of $h(u)$ for $u \in [r, R] \setminus \{r_T, R_T\}$. Then*

$$\sup\ I_h(p,q) = F(r_T, R_T).$$

Proof. Put $v_1 = r_T$, $v_3 = R_T$ and suppose if possible $u = v_2 \in (r_T, R_T)$ is in the support of one– or two–point distributions p_0, q_0 for which I_f realises its supremum. By assumption

$$h(v_2) < \lambda h(v_1) + (1-\lambda)h(v_3)$$

in the notation of Proposition 1(a) case (i). The argument of case (i) shows that $I_f(p_0, q_0)$ experiences a strict increase if p_0, q_0 are modified by a suitable redistribution of probability mass from v_2 to v_1 and v_3, a contradiction to the extremality of $I_f(p_0, q_0)$. Thus the support of p_0 and q_0 must have empty intersection with (r_T, R_T). There is nothing more to prove if $r_T = r$ and $R_T = R$. In any case, we see that since $1 \in (r_T, R_T)$, p_0 and q_0 must have two–point support.

If $R_T < R$, suppose if possible p_0 and q_0 have a point of support $r_0 \in (R_T, R]$. Then $(R_T, h(R_T))$ lies above the chord joining $(r_T, h(r_T))$ to $(r_0, h(r_0))$, so we may derive a contradiction by the construction of case (ii) or case (iii) of Proposition 1 (a). Since p_0 and q_0 must have a point of support greater than unity, that point must therefore be $u = R_T$. A similar argument show that the support point less than unity must be at $u = r_T$. □

By taking $-h$ in place of h in the preceding theorem, we derive the following corresponding theorem for infima.

Theorem 4. *Suppose r_S, R_S satisfy $r \le r_S < 1 < R_S \le R$ and the line joining $(r_S, h(r_S))$ to $(R_S, h(R_S))$ lies strictly below the graph of $h(u)$ for $u \in [r, R] \setminus \{r_S, R_S\}$. Then*

$$\inf\ I_h(p, q) = F(r_S, R_S).$$

Whenever the above theorems are applicable, we may derive $\sup\ |I_f(p, q)|$ from

$$\sup\ |I_f(p, q)| = \max\left[\sup\ I_f(p, q), -\inf\ I_f(p, q)\right].$$

Some modification to Theorem 3 is necessary if $r_T < 1 < R_T$ is violated. This has not been found to occur in the examples we have looked at but could be dealt with on an *ad hoc* basis. For this reason we have not seen fit to strive for further generality in Theorem 3 at the cost of complicating it. The same comment applies to Theorem 4.

5. Examples

For the mapping $f : (0, \infty) \to R$ given by $f(u) = u \ln u$, $I_f(p, q)$ becomes the Kulback–Leibler distance

$$D(p, q) := \sum_{i=1}^{n} p_i \ln(u_i).$$

We have

$$\begin{aligned} I_f(p, q) &= \frac{1}{2} \sum_{i=1}^{n} \left[\ln\left(\frac{p_i}{q_i}\right) + 1 \right] (p_i - q_i) \\ &= \frac{1}{2} \sum_{i=1}^{n} (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right) \\ &= \frac{1}{2} [D(p, q) + D(q, p)], \end{aligned}$$

$$\sum_{i=1}^{n} q_i f_0(u_i) = \sum_{i=1}^{n} q_i \left[\ln(u_i)\right]^2$$

and

$$\left\| f^{''} \right\|_{\infty} = \sup_{u \in [r, R]} \left| f^{''}(u) \right| = 1/r.$$

Consequently Theorem 1 provides

$$
\begin{aligned}
(5.1) \qquad 0 &\le |D(p,q) - D(q,p)| \\
&\le \frac{1}{2r} D_{\chi^2}(p,q) - \frac{r}{2} \sum_{i=1}^{n} q_i \left[\ln \left(\frac{p_i}{q_i} \right) \right]^2 \\
&\le \frac{1}{2r} D_{\chi^2}(p,q) \\
&\le \frac{1}{2r} (R-1)(1-r) \\
&\le \frac{1}{8r} (R-r)^2
\end{aligned}
$$

which measures the asymmetry of the Kullback–Leibler distance.

A simple calculation gives

$$\left\| f^{'''} \right\|_\infty = 1/r^2,$$

so that Theorem 2 gives the bounds

$$0 \le |D(p,q) - D(q,p)| \le \frac{1}{6r^2} D_{|\chi|^3}(p,q) \le \frac{1}{48r^2}(R-r)^3.$$

The last bound here can be seen to be strictly better than that in (5.1) if $r > R/7$.

Tight bounds for $\frac{1}{2}|D(p,q) - D(q,p)|$ involving only r, R can be derived using the ideas of the previous section. We have

$$h(u) = \frac{u+1}{2} \left[\ln u - (u-1) \right],$$

so that

$$h^{'}(u) = \frac{1}{2} \left[\ln u + \frac{1}{u} \right] \text{ and } h^{''}(u) = \frac{u-1}{2u^2}.$$

Thus $h^{'}(u) > 0$ for $u \ge 1$. By the elementary inequality

$$\frac{1}{u} + \ln u > 1 \text{ for } 0 < u < 1,$$

we have that $h^{'}(u) > 0$ holds for $0 < u < 1$ as well and so h is strictly increasing for all $u < 0$. Also h is strictly concave for $0 < u < 1$ and strictly convex for $u > 1$.

This falls within the scope of Theorems 3 and 4. If r_1 is the demonstrably unique value of u less than unity at which the tangent to the graph of h passes through $(R, h(R))$, then we may choose $r_T = \max(r, r_1)$. If R_1 is the demonstrably unique value of u greater than unity at which the tangent to the graph of h passes through $(r, h(r))$, then we may choose $R_T = \min(R, R_1)$.

Now consider the mapping $f : (0, \infty) \to R$ given by $f(u) = \ln u$. We have

$$I_f(p,q) = \sum_{i=1}^{n} q_i \ln(u_i) = -D(q,p),$$

$$\sum_{i=1}^{n} f^{'}(u_i)(p_i - q_i) = \sum_{i=1}^{n} \frac{q_i}{p_i}(p_i - q_i) = 1 - \sum_{i=1}^{n} \frac{q_i^2}{p_i} = -D_{\chi^2}(q,p),$$

$$\sum_{i=1}^{n} q_i f_0(u_i) = \sum_{i=1}^{n} q_i \left(\frac{q_i}{p_i} - 1 \right)^2 = \sum_{i=1}^{n} \frac{q_i(q_i - p_i)^2}{p_i^2}$$

and $\left\|f^{''}\right\|_\infty = 1/r^2$.

Consequently, by (2.1), we have

$$0 \le \left|D(q,p) + \frac{1}{2}D_{\chi^2}(q,p)\right| \le \frac{1}{4r^2}D_{\chi^2}(p,q) - \frac{r^2}{4}\sum_{i=1}^n \frac{q_i(q_i-p_i)}{p_i^2}^2 \le \frac{1}{4r^2}D_{\chi^2}(p,q) \le \frac{1}{4r^2}(R-1)(1-r) \le \frac{1}{16r^2}(R-r)^2. \tag{5.2}$$

A simple calculation shows that $\left\|f^{'''}\right\|_\infty = 2/r^3$, so that by Theorem 2

$$0 \le \left[D(q,p) + \frac{1}{2}D_{\chi^2}(q,p)\right] \le \frac{1}{6r^3}D_{|\chi|^3}(p,q) \le \frac{1}{48}\left(\frac{R}{r}-1\right)^3.$$

The last bound here can be seen to be better than that in (5.2) if $r > R/4$.

Again we may obtain a tight bound for $|D(q,p) - \frac{1}{2}D_{\chi^2}(q,p)|$ in terms of r, R alone by use of Theorems 3 and 4. We have

$$h(u) = \ln u - \frac{u-1}{2u},$$

so that

$$h^{'}(u) = \frac{2u-1}{2u^2} \text{ and } h^{''}(u) = \frac{1-u}{u^3}.$$

Thus h is decreasing for $0 < u < 1/2$ and increasing for $u > 1/2$. Further it is strictly convex for $0 < u < 1$ and strictly concave for $u > 1$, and so is quasiconvex. We may define r_T and R_T exactly as in the previous example.

Finally suppose $f : (0,\infty) \to R$ is given by $f(u) = \frac{1}{2}\left(\sqrt{u}-1\right)^2$. Then $I_f(p,q)$ becomes the Hellinger discrimination

$$h^2(p,q) := \frac{1}{2}\sum_{i=1}^n q_i\left(\sqrt{u_i}-1\right)^2.$$

We have

$$\sum_{i=1}^n f^{'}\left(\frac{p_i}{q_i}\right)(p_i - q_i) = \frac{1}{2}\sum_{i=1}^n (q_i - p_i)\frac{\sqrt{q_i}}{\sqrt{p_i}},$$

$$\sum_{i=1}^n q_i f_0(u_i) = \sum_{i=1}^n \frac{q_i}{p_i}(\sqrt{p_i}-\sqrt{q_i})^2$$

and

$$\left\|f^{''}\right\|_\infty = \sup_{u\in[r,R]}\left|f^{''}(u)\right| = \frac{1}{4r^{3/2}}.$$

Consequently Theorem 1 provides

$$\begin{aligned}\left|h^2(p,q)-\frac{1}{4}\sum_{i=1}^{n}(q_i-p_i)\sqrt{\frac{q_i}{p_i}}\right| &\le \frac{1}{16r^{3/2}}D_{\chi^2}(p,q)-r^{3/2}\sum_{i=1}^{n}\frac{q_i}{p_i}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2\\ &\le \frac{1}{16r^{3/2}}D_{\chi^2}(p,q)\\ &\le \frac{1}{16r^{3/2}}(R-1)(1-r)\\ &\le \frac{1}{64r^{3/2}}(R-r)^2.\end{aligned}$$

Also, as $f^{'''}(u) = -\frac{3}{8}u^{-5/2}$, we have $\left\|f^{'''}\right\|_\infty = \sup_{u\in[r,R]}\left|f^{'''}(u)\right| = \frac{3}{8}r^{-5/2}$, and Theorem 2 gives

$$\left|h^2(p,q)-\frac{1}{4}\sum_{i=1}^{n}(q_i-p_i)\sqrt{\frac{q_i}{p_i}}\right| \le \frac{1}{32r^{5/2}}D_{|\chi|^3}(p,q) \le \frac{1}{256r^{5/2}}(R-r)^3. \tag{5.3}$$

The largest bound here is better than the largest provided by Theorem 1 provided $r > R/5$.

The use of Theorems 3 and 4 to obtain an absolute upper bound for the left–hand side of the first inequality in (5.3) is more complicated than in the previous examples. We have

$$h(u) = \frac{\left(\sqrt{u}-1\right)^3}{4\sqrt{u}}.$$

Hence

$$h^{'}(u) = \frac{(\sqrt{u}-1)^2}{8u^{3/2}}\left[2\sqrt{u}+1\right] \text{ and } h^{''}(u) = \frac{\sqrt{u}-1}{16u^3}\left[2+3u^{1/2}+3u-6u^{3/2}\right].$$

Thus h is strictly increasing for $u > 0$. It is strictly concave for $0 < u < 1$, strictly convex for $1 < u < u_0$ and strictly concave for $u > u_0$, where u_0 is the unique zero exceeding unity of the cubic polynomial $2 + 3x + 3x^2 - 6x^3$.

There exist a unique pair of points $(r_1, h(r_1))$, $(R_1, h(R_1))$ with $r_1 < 1 < R_1$ at which the graph of h has a common tangent which lies above the graph for all $u > 0$ except at the two osculating points. If $r \le r_1 < R_1 \le R$, we may take $r_T = r_1$ and $R_T = R_1$. Suppose $r_1 < r$. Then there exists a unique $u = R_2 > 1$ such that the tangent to the graph at $(R_2, h(R_2))$ passes through $(r, h(r))$ and lies above the graph for $r < u < R_2$. We may choose $r_T = r$, $R_T = \min(R_2, R)$.

If the join of $(r, h(r))$ to $(R, h(R))$ lies below the graph for $r < u < R$, or is tangential to the graph at an intermediate point, we may take $r_S = r$, $R_S = R$. Otherwise, $(r, h(r))$ lies on the tangent to the graph at a unique point $(r_2, h(r_2))$ with $1 < r_2 < u_0$ and this tangent meets the graph again at $(r_3, h(r_3))$ with $r_3 > R$. Similarly $(R, h(R))$ lies on the tangent to the graph at a unique point $(R_2, h(R_2))$ with $1 < R_2 < u_0$ and this tangent meets the graph again at $(R_3, h(R_3))$ with $R_3 < r$. At least one of r_2, R_2 is not unity. If $r_2 \ne 1$, we may take $r_S = r$, $R_T = r_2$. If $R_2 \ne 1$, we may take $r_S = R_2$, $R_S = R$.

References

[1] I. Csiszár, *A note on Jensen's inequality*, Studia Sci. Math. Hung. **1** (1966), 185–188.

[2] I. Csiszár, *Information–type measures of differences of probability distributions and indirect observation*, Studia Sci. Math. Hung. **2** (1967), 299–318.

[3] I. Csiszár, *On topological properties of f–divergence*, Studia Sci. Math. Hung. **2** (1967), 329–339.

[4] S. S. Dragomir, V. Gluščević and C. E. M. Pearce, *Csiszár f–divergence, Ostrowski's inequality and mutual information*, to appear in J. Nonlin. Anal. Ser. A.

[5] S. S. Dragomir, V. Gluščević and C. E. M. Pearce, *Approximations for Csiszár f–divergence via midpoint inequalities*, submitted.

[6] M. G. Kreĭn and A. A. Nudel'man, *The Markov moment problem and extremal problems*, Transl. Math. Monographs **50**, Amer. Math. Soc., 1977.

[7] D. S. Mitrinović, J. E. Pečarić and A. M. Fink, *Inequalities for functions and their integrals and derivatives*, Kluwer Academic, Dordrecht, 1994.

School of Communications and Informatics, Victoria University, PO Box 14428, MCMC, Melbourne, Victoria 8001, Australia

E-mail address: `sever@matilda.vu.edu.au`

Royal Australian Air Force, ARDU, PO Box 1500, Salisbury SA 5108, Australia

E-mail address: `vgluscev@maths.adelaide.edu.au`

Applied Mathematics Department, Adelaide University, Adelaide, SA 5005, Australia

E-mail address: `cpearce@maths.adelaide.edu.au`