NOTE **Communicated by Shun-ichi Amari**

A Note on Divergences

Xiao Liang
xlianguw@uw.edu
Institute of Technology, University of Washington Tacoma,
Tacoma, WA 98402, U.S.A.

In many areas of neural computation, like learning, optimization, estimation, and inference, suitable divergences play a key role. In this note, we study the conjecture presented by Amari (2009) and find a counterexample to show that the conjecture does not hold generally. Moreover, we investigate two classes of Φ-divergence (Zhang, 2004), weighted *f*-divergence and weighted *α*-divergence, and prove that if a divergence is a weighted *f*-divergence, as well as a Bregman divergence, then it is a weighted *α*-divergence. This result reduces in form to the main theorem established by Amari (2009) when $\mu_i = 1$ $(i = 1, \ldots, n)$.

1 Introduction

In many areas of neural computation like learning, optimization, estimation, and inference, various divergences, which measure discrepancy between two points or between two probability distributions or positive measures, play a key role. Therefore, divergences are fundamental objects in information theory, statistics, mathematical programming, computational vision, and neural networks. The well-known important divergences are Kullback-Leibler divergence, *f*-divergence (Csiszár, 1963), Bregman divergence (Bregman, 1967), *α*-divergence (Amari, 1985), Jenson difference (Rao, 1987), Φ-divergence (Zhang, 2004), and *U*-divergence (Eguchi, 2008). So far, the theory of divergences and its applications have been well developed (see the references just mentioned, and Ackley, Hinton, & Sejnowski, 1985; Amari, 2007, 2009, 2016; Amari & Nagaoka, 2000; Cichocki, Adunek, Phan, & Amari, 2009; Csiszár, 1974; Eguchi & Copas, 2002; Jiao, Courtade, No, Venkat, & Weissman, 2015; Murata, Takenouchi, Kanamori, & Eguchi, 2004; Nielsen, 2009; Nielsen & Noch, 2009; Taneja & Kumar, 2004).

Inspired by these significant works, especially the work of Amari (2009) and Zhang (2004), in this note, we study the relationships between some divergences in the space of positive measures. As shown in previous literature (Cichocki et al., 2009; Murata et al., 2004; Nielsen, 2009), in order to deal with more complex problems from the real world, the ordinary constraint of a probability distribution that the total mass is 1 needs to be relaxed in many cases. A typical example is the visual signal, which is

Neural Computation **28, 2045–2062** (2016)
doi:10.1162/NECO_a_00878

normally a two-dimensional array with nonnegative elements. Therefore, we still choose the space of positive measures as our basic setting of the space in the work. We first study the following conjecture, put forward by Amari (2009):

Conjecture: When a divergence $D[\mathbf{p}, \mathbf{q}]$ satisfies information monotonicity, it is a function of an f-divergence.

We find a counterexample to show that this conjecture does not hold generally.

Second, we investigate two classes of Φ-divergence (Zhang, 2004): weighted f-divergence and weighted α-divergence, which are generalizations of the well-known f-divergence and α-divergence, respectively. We prove that if a divergence is a weighted f-divergence as well as a Bregman divergence, then it is a weighted α-divergence. This result reduces in form to the main theorem established by Amari (2009) when $\mu_i = 1$ $(i = 1, \ldots, n)$.

2 Several Classical Divergences and Their Basic Properties

In this section, based on the work cited in section 1, we recall several classical divergences and present their basic properties and the corresponding proofs.

2.1 Divergence. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n elements on which a positive measure $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ is defined as

$$\text{Measure}[x_i] = m_i, \quad i = 1, 2, \ldots, n.$$

When

$$\sum_{i=1}^{n} m_i = 1,$$

it is a probability measure.

We use $\mathcal{M}$ and $\mathcal{P}$ to stand for the set of all positive measures on X and the set of all probability measures on X, respectively. It is clear that

$$\mathcal{P} \subset \mathcal{M}.$$

A function $D[\mathbf{m} : \mathbf{n}]$, $\mathbf{m}, \mathbf{n} \in \mathcal{M}$, is called a divergence when it satisfies the following conditions:

1. $D[\mathbf{n} : \mathbf{m}] \geq 0$.
2. $$D[\mathbf{n} : \mathbf{m}] = 0 \quad \text{iff} \quad \mathbf{n} = \mathbf{m}.$$

3. For small $d\mathbf{n}$,

$$D[\mathbf{n} + d\mathbf{n} : \mathbf{n}] \approx \frac{1}{2} \sum_{i,j=1}^{n} a_{ij} dn_i dn_j$$

gives a positive-definite quadratic form.

We refer readers to the important work of Zhang (2004) for another more general definition of divergences.

A basic example of divergence is the square of the Euclidean distance:

$$D[\mathbf{m} : \mathbf{n}] = \frac{1}{2} \sum_{i=1}^{n} |m_i - n_i|^2.$$

Another basic example of divergence is the (squared) Hellinger divergence,

$$D[\mathbf{p} : \mathbf{q}] = \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2,$$

for any $\mathbf{p}, \mathbf{q} \in \mathcal{P}$.

2.2 *f*-Divergence. Let f be a convex and twice continuously differentiable function on $R^+ = [0, \infty)$ satisfying $f(1) = 0$ and $f'(1) = 0$. For every $\mathbf{m}, \mathbf{n} \in \mathcal{M}$, define

$$D_f[\mathbf{m} : \mathbf{n}] = \sum_{i=1}^{n} m_i f\left(\frac{n_i}{m_i}\right).$$

Then $D_f[\mathbf{m} : \mathbf{n}]$ is called an f-divergence in $\mathcal{M}$.

The f-divergence was introduced by Csiszár (1963). (See also Taneja & Kumar, 2004, for its detailed properties.)

2.2.1 Case: In $\mathcal{P}$. In this case, we see that

1. The convexity of f implies that

$$D_f[\mathbf{p} : \mathbf{q}] = \sum_{i=1}^{n} p_i f\left(\frac{q_i}{p_i}\right) \geq f\left(\sum_{i=1}^{n} p_i \frac{q_i}{p_i}\right) = f(1) = 0,$$

for any $\mathbf{p}, \mathbf{q} \in \mathcal{P}$.

2. If f is strictly convex, then

$$D_f[\mathbf{p} : \mathbf{q}] = 0 \quad \text{iff} \quad \mathbf{p} = \mathbf{q}.$$

3. If we take $f_c(u) = f(u) - c(u-1)$, we get

$$D_f[\mathbf{p} : \mathbf{q}] = D_{f_c}[\mathbf{p} : \mathbf{q}], \quad \text{for any } \mathbf{p}, \mathbf{q} \in \mathcal{P}.$$

4.

$$D_{cf}[\mathbf{p} : \mathbf{q}] = cD_f[\mathbf{p} : \mathbf{q}], \quad \text{for all } c > 0, \ \mathbf{p}, \ \mathbf{q} \in \mathcal{P}.$$

Properties 3 and 4 show that without loss of generality, we can require that f satisfies

$$f''(1) = 1.$$

As in Amari (2009), we call f a standard convex function on R^+ if f is a strictly convex and twice continuously differentiable function on R^+ satisfying

$$f(1) = 0, \quad f'(1) = 0, \quad f''(1) = 1.$$

2.2.2 Case: In the general $\mathcal{M}$. Let f be a standard convex function on R^+. Then we have

$$\begin{aligned} f(x) &\geq 0 \qquad \text{for } x \geq 0, \\ f(x) &= 0 \qquad \text{iff } x = 0. \end{aligned}$$

Hence,

- For every $\mathbf{m}, \mathbf{n} \in \mathcal{M}$, we obtain

$$\begin{aligned} D_f[\mathbf{m} : \mathbf{n}] &= \sum_{i=1}^{n} m_i f\left(\frac{n_i}{m_i}\right) \\ &= \left(\sum_{i=1}^{n} m_i\right) \sum_{i=1}^{n} \frac{m_i}{\sum_{i=1}^{n} m_i} f\left(\frac{n_i}{m_i}\right) \\ &\geq \left(\sum_{i=1}^{n} m_i\right) f\left(\sum_{i=1}^{n} \frac{m_i}{\sum_{i=1}^{n} m_i} \frac{n_i}{m_i}\right) \\ &= \sum_{i=1}^{n} m_i f\left(\frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} m_i}\right) \\ &\geq 0. \end{aligned}$$

- The strict convexity of f implies that

$$D_f[\mathbf{m} : \mathbf{n}] = 0 \quad \text{iff} \quad \mathbf{m} = \mathbf{n}.$$

2.3 α-Divergence in $\mathcal{M}$. The Amari α-divergence (Amari, 1985) is another important parametric family of divergence functionals.

The definition of α-divergence is as follows. For any $\alpha \in R = (-\infty, +\infty)$, set

$$f_\alpha(u) = \begin{cases} \dfrac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right) + \dfrac{2}{1-\alpha}(u-1), & \alpha \neq \pm 1, \\ u \log u - (u-1), & \alpha = 1, \\ -\log u + (u-1), & \alpha = -1. \end{cases}$$

Then the related f_α-divergence is called the α-divergence, and

$$D_\alpha[\mathbf{m} : \mathbf{n}] = \begin{cases} \dfrac{4}{1-\alpha^2} \displaystyle\sum_{i=1}^{n} \left(\frac{1-\alpha}{2} m_i + \frac{1+\alpha}{2} n_i - m_i^{\frac{1-\alpha}{2}} n_i^{\frac{1+\alpha}{2}} \right), & \alpha \neq \pm 1, \\ \displaystyle\sum_{i=1}^{n} \left(m_i - n_i + n_i \log \frac{n_i}{m_i} \right), & \alpha = 1, \\ \displaystyle\sum_{i=1}^{n} \left(n_i - m_i + m_i \log \frac{m_i}{n_i} \right), & \alpha = -1. \end{cases}$$

2.4 Bregman Divergence. Let $\Omega \subset R^n$ be a convex set, and let $\Phi : \Omega \to R$ be a continuously differentiable real-valued and strictly convex function. Recall that the Bregman divergence associated with Φ for points $\mathbf{p}, \mathbf{q} \in \Omega$ is defined by

$$D_\Phi[\mathbf{p} : \mathbf{q}] = \Phi(\mathbf{p}) - \Phi(\mathbf{q}) - \nabla\Phi(\mathbf{q}) \bullet (\mathbf{p} - \mathbf{q}),$$

where $\nabla\Phi$ is the gradient of Φ.

2.5 Bregman Divergence in $\mathcal{M}$. Let $k_i \in C^1(R^+)$ $(i = 1, 2, \ldots, n)$ be strictly monotone functions and

$$\Omega_0 = k_1(R^+) \times k_2(R^+) \times \cdots \times k_n(R^+).$$

Let $\Phi : \Omega_0 \to R$ be a twice continuously differentiable strictly convex function, that is, $\Phi \in C^2(\Omega_0)$, and Φ is strictly convex.

For every $\mathbf{r}, \mathbf{s} \subset \mathcal{M}$, define

$$D_\Phi[\mathbf{r} : \mathbf{s}] = \Phi(k(\mathbf{r})) - \Phi(k(\mathbf{s})) - \sum_{i=1}^{n} \Phi_i'(k(\mathbf{s}))(k_i(r_i) - k_i(s_i)),$$

where

$$\Phi(k(\mathbf{r})) = \Phi(k_1(r_1), \ldots, k_n(r_n)),$$
$$\Phi(k(\mathbf{s})) = \Phi(k_1(s_1), \ldots, k_n(s_n)), \quad \Phi_i'(k(\mathbf{s})) = \frac{\partial \Phi}{\partial x_i}(k(\mathbf{s})).$$

Then D_Φ is called a Bregman divergence in $\mathcal{M}$, which is associated with Φ.

3 On the Conjecture Presented by Amari (2009)

In this section, we give a counterexample to show that the conjecture presented by Amari (2009) is not generally true.

First, we recall the concept of information monotonicity.

Let $G_1, G_2, \ldots, G_m$ $(m < n)$ be subsets of X such that

$$G_i \cap G_j = \emptyset \ \ (i \neq j), \quad \cup_{i=1}^{m} G_i = X.$$

Then $\{G_1, G_2, \ldots, G_m\}$ is called a partition of X, which is a coarsely grained version of X.

The partition naturally induces some kind of distribution $\overline{\mathbf{p}}$ over $G_1, G_2, \ldots, G_m$:

$$\overline{\mathbf{p}} = \{\overline{p}_1, \ldots, \overline{p}_m\} \text{ with } \overline{p}_i = \text{measure}\{G_i\} = \sum_{x_k \in G_i} p_k. \tag{3.1}$$

As a coarsely grained version of X, $\{G_1, G_2, \ldots, G_m\}$ loses information by summarizing elements within each subset G_i. So it is natural to stipulate a monotonic relation,

$$D[\mathbf{p} : \mathbf{q}] \geq D[\overline{\mathbf{p}} : \overline{\mathbf{q}}],$$

where $\mathbf{p}, \mathbf{q} \in \mathcal{M}$, $\overline{\mathbf{p}}, \overline{\mathbf{q}}$ are related distributions of $\{G_1, G_2, \ldots, G_m\}$, defined as in equation 3.1.

Consider the case

$$\frac{q_k}{p_k} = \lambda_i \quad x_k \in G_i, \quad i = 1, 2, \ldots, m. \tag{3.2}$$

Here p_k and q_k are proportional inside each class G_i, that is, the conditional distributions of $\mathbf{p}$ and $\mathbf{q}$ are equal, conditioned on $x \in G_i$. Then it is natural to assume that

$$D[\mathbf{p} : \mathbf{q}] = D[\overline{\mathbf{p}} : \overline{\mathbf{q}}]$$

because details of G_i do not give any information distinguishing $\mathbf{p}$ from $\mathbf{q}$. The equality holds only in this case.

The above properties are called information monotonicity.

As Csiszár (1974) found, every f-divergence has information monotonicity. Actually, since f is convex function, we have

$$\begin{aligned}
D_f[\mathbf{p} : \mathbf{q}] &= \sum_{i=1}^{n} p_i f\left(\frac{q_i}{p_i}\right) \\
&= \sum_{i=1}^{m} \sum_{x_k \in G_i} p_k f\left(\frac{q_k}{p_k}\right) \\
&= \sum_{i=1}^{m} \sum_{x_k \in G_i} \left[\left(\sum_{x_k \in G_i} p_k \right) \frac{p_k}{\sum_{x_k \in G_i} p_k} f\left(\frac{q_k}{p_k}\right) \right] \\
&\geq \sum_{i=1}^{m} \left(\sum_{x_k \in G_i} p_k \right) f\left(\frac{\sum_{x_k \in G_i} q_k}{\sum_{x_k \in G_i} p_k} \right) \\
&= D_f[\overline{\mathbf{p}} : \overline{\mathbf{q}}],
\end{aligned}$$

and for the case of equation 3.2, we get

$$\begin{aligned}
D_f[\mathbf{p} : \mathbf{q}] &= \sum_{i=1}^{n} p_i f\left(\frac{q_i}{p_i}\right) \\
&= \sum_{i=1}^{m} \sum_{x_k \in G_i} p_k f\left(\lambda_i\right) \\
&= \sum_{i=1}^{m} \left[\left(\sum_{x_k \in G_i} p_k \right) f\left(\lambda_i\right) \right] \\
&= D_f[\overline{\mathbf{p}} : \overline{\mathbf{q}}].
\end{aligned}$$

Therefore, *f*-divergence has information monotonicity. Moreover, from Csiszár (1974) and Amari (2009), we know that the *f*-divergence is the only class of decomposable information monotonic divergences.

We are now in a position to give a counterexample to show that the conjecture above is not generally true.

Set

$$f_1(u) = \frac{1}{2}(u-1)^2, \tag{3.3}$$

$$f_2(u) = \frac{1}{2}(u-1)^2 + (u-1)^4. \tag{3.4}$$

Then, f_1 and f_2 are two standard convex functions on R^+. Let D_1 and D_2 be *f*-divergences derived from f_1 and f_2 , respectively, and

$$D[\mathbf{p} : \mathbf{q}] := D_2[\mathbf{p} : \mathbf{q}] + \{D_1[\mathbf{p} : \mathbf{q}]\}^2, \tag{3.5}$$

where $\mathbf{p},\ \mathbf{q} \in \mathcal{M}$. Then it is clear that

$$D[\mathbf{p} : \mathbf{q}] \geq 0$$

and

$$D[\mathbf{p} : \mathbf{q}] = 0 \quad \text{iff} \ \ \mathbf{p} = \mathbf{q}.$$

Moreover, for small $d\mathbf{p}$,

$$D[\mathbf{p} + d\mathbf{p} : \mathbf{p}] \approx \frac{1}{2} \sum_{i,j=1}^{n} a_{ij} dp_i dp_j$$

gives a positive-definite quadratic form. Therefore, $D[\mathbf{p} : \mathbf{q}]$ is a divergence.

Since D_1 and D_2 are *f*-divergences, they have an information monotonicity property. Thus, it is easy to see that $D[\mathbf{p} : \mathbf{q}]$ in equation 3.5 has an information monotonicity property.

Next, we show by contradiction (i.e., the reductio ad absurdum argument) that $D[\mathbf{p} : \mathbf{q}]$ is not a function of an *f*-divergence. Suppose this is not true. Then there exists an *f*-divergence D_f and a function κ such that

$$\kappa(D_f[\mathbf{p} : \mathbf{q}]) = D[\mathbf{p} : \mathbf{q}] \quad \text{for all } \mathbf{p},\ \mathbf{q} \in \mathcal{M}; \tag{3.6}$$

that is, for all $\mathbf{p},\ \mathbf{q} \in \mathcal{M}$ with $\mathbf{p} = \{p_1, \ldots, p_n\}$ and $\mathbf{q} = \{q_1, \ldots, q_n\}$, we have

$$\kappa\left(\sum_{i=1}^{n} p_i f\left(\frac{q_i}{p_i}\right)\right) = \sum_{i=1}^{n} p_i f_2\left(\frac{q_i}{p_i}\right) + \left(\sum_{i=1}^{n} p_i f_1\left(\frac{q_i}{p_i}\right)\right)^2. \tag{3.7}$$

Clearly, equation 3.7 holds for the special case

$$p_i = p_1,\ q_i = q_1, \quad i = 1, 2, \ldots, n;$$

that is,

$$\kappa\left(\sum_{i=1}^{n} p_1 f\left(\frac{q_1}{p_1}\right)\right) = \sum_{i=1}^{n} p_1 f_2\left(\frac{q_1}{p_1}\right) + \left(\sum_{i=1}^{n} p_1 f_1\left(\frac{q_1}{p_1}\right)\right)^2.$$

Hence,

$$\kappa\left(np_1 f\left(\frac{q_1}{p_1}\right)\right) = np_1 f_2\left(\frac{q_1}{p_1}\right) + \left(np_1 f_1\left(\frac{q_1}{p_1}\right)\right)^2. \tag{3.8}$$

Next, we prove that equation 3.8 leads to a contradiction, which means that equation 3.6 leads to a contradiction. This implies that equation 3.6 (and even equation 3.8) does not hold; that is, $D[\mathbf{p} : \mathbf{q}]$ is not a function of an f-divergence.

Set

$$x := np_1, \quad y := \frac{q_1}{p_1} - 1.$$

Then by equations 3.8, 3.3, and 3.4, we have

$$\begin{aligned} &\kappa(xf(y+1)) \\ &= xf_2(y+1) + x^2 f_1^2(y+1) \\ &= \frac{1}{2}xy^2(2y^2+1) + \frac{1}{4}x^2y^4, \quad \text{for } x > 0, y > -1. \end{aligned} \tag{3.9}$$

Clearly, there exists a $y_0 > -1$ such that $f(y_0 + 1) \neq 0$. Otherwise, $f(y + 1) = 0$ for all $y > -1$, which means that the left side of the equality 3.9 is a constant $\kappa(0)$. Taking $y = 0$ on the right side of equation 3.9, we get $\kappa(0) = 0$. Then, taking $x = y = 1$ on the right side of equation 3.9, we see that $0 = \frac{7}{4}$, but this is false.

Moreover, we know that

$$y_0 \neq 0$$

by noting that $f(1) = 0$. For this fixed y_0, we set

$$t := xf(y_0 + 1), \quad A := \frac{y_0^2(2y_0^2 + 1)}{2f(y_0 + 1)}, \quad B := \frac{y_0^4}{4f^2(y_0 + 1)}. \tag{3.10}$$

Then we obtain

$$\kappa(t) = At + Bt^2.$$

In view of equation 3.9, we get, for $x > 0, y > -1$,

$$Axf(y + 1) + Bx^2 f^2(y + 1) = \frac{1}{2}xy^2(2y^2 + 1) + \frac{1}{4}x^2y^4. \tag{3.11}$$

Since f is a standard convex function, that is, f is a twice continuously differentiable convex function on R^+ such that

$$f(1) = 0, \quad f'(1) = 0, \quad f''(1) = 1,$$

it follows from L'Hospital's rule that

$$\lim_{y\to 0} \frac{f(y + 1)}{y^2} = \lim_{y\to 0} \frac{f'(y + 1)}{2y} = \frac{1}{2}.$$

Therefore,

$$\lim_{y\to 0} \frac{Axf(y + 1) + Bx^2 f^2(y + 1)}{y^2} = \frac{1}{2}Ax.$$

Thus, by equation 3.11, we have

$$\frac{1}{2}Ax = \lim_{y\to 0} \left\{ \frac{1}{2}x(2y^2 + 1) + \frac{1}{4}x^2y^2 \right\} = \frac{1}{2}x.$$

Hence,

$$A = 1. \tag{3.12}$$

Moreover, differentiating equation 3.11 twice with respect to x (or comparing the coefficients before x^2 in equation 3.11) gives that

$$Bf^2(y + 1) = \frac{y^4}{4}. \tag{3.13}$$

Clearly,

$$\lim_{y\to 0} B\frac{f^2(y+1)}{y^4} = \frac{1}{4}B.$$

So by equation 3.13, we have

$$B = 1. \tag{3.14}$$

By equations 3.12 and 3.14, we get

$$A^2 = B.$$

This, together with equation 3.10, implies that

$$\frac{y_0^4(2y_0^2+1)^2}{4f^2(y_0+1)} = \frac{y_0^4}{4f^2(y_0+1)}.$$

So,

$$y_0^4 + y_0^2 = 0,$$

that is,

$$y_0 = 0$$

This contradicts $y_0 \neq 0$.

Consequently, $D[\mathbf{p} : \mathbf{q}]$ is not a function of an *f*-divergence, that is, the conjecture above is not true for $D[\mathbf{p} : \mathbf{q}]$.

Remark 1. For the special case of $n = 2$, Amari's conjecture has already been disproved by Jiao et al. (2015). (See also Amari, 2016.) Jiao et al. (2015) proved that the $n = 2$ case (binary X) is different from the general case and that an information monotone divergence of the Bregman type is not necessarily a KL-divergence, when $n = 2$, giving a general form of information-monotone divergence. It includes divergence that cannot be written as a function of an *f*-divergence.

We just disproved Amari's conjecture in the general case $n \geq 2$.

4 Weighted *f*-Divergence and Weighted α-Divergence

In this section, we are concerned with two classes of Φ-divergences (Zhang, 2004), weighted *f*-divergence and weighted α-divergence, which are generalizations of the well-known *f*-divergence and α-divergence, respectively.

Definition 1. *Let f be a standard convex function on R^+, and $\mu_i \in (0, \infty)$ ($i = 1, \ldots, n$). For every $\boldsymbol{m}, \boldsymbol{n} \in \mathcal{M}$, define*

$$D_f^w[\boldsymbol{m} : \boldsymbol{n}] = \sum_{i=1}^{n} \mu_i m_i f\left(\frac{n_i}{m_i}\right).$$

Then, $D_f^w[\boldsymbol{m} : \boldsymbol{n}]$ is called a weighted f-divergence in $\mathcal{M}$.

The related weighted f_α-divergence is thus called the weighted α-divergence.

Obviously the *f*-divergence (resp. α-divergence) is a special case of weighted *f*-divergence (resp. weighted α-divergence) when

$$\mu_i = 1, \quad i = 1, \ldots, n.$$

Moreover, since f is a standard convex function on R^+, we have

$$f(x) \geq 0 \quad \text{for } x \geq 0,$$
$$f(x) = 0 \quad \text{iff } x = 1.$$

Hence,

- For every $\mathbf{m}, \mathbf{n} \in \mathcal{M}$,

$$D_f^w[\mathbf{m} : \mathbf{n}] \geq 0.$$

- The strict convexity of f implies that

$$D_f^w[\mathbf{m} : \mathbf{n}] = 0 \quad \text{iff } \mathbf{m} = \mathbf{n}.$$

Lemma 1. *Let D_Φ be a Bregman divergence in $\mathcal{M}$. If D_Φ is a decomposable divergence, then there exist $\psi_i \in C^2(k_i(R^+))$ ($i = 1, 2, \ldots, n$) such that ψ_i ($i = 1, 2, \ldots, n$) are strictly convex functions, and for every $i = 1, 2, \ldots, n$,*

$$\frac{\partial\left(\Phi(\boldsymbol{x}) - \sum_{i=1}^{n} \psi_i(x_i)\right)}{\partial x_i} = 0, \quad \textit{for every } \boldsymbol{x} \in \mathcal{M}. \tag{4.1}$$

Moreover, in this case, we have, for every $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{M}$,

$$\frac{\partial(D_{\Phi}[\boldsymbol{r} : \boldsymbol{s}])}{\partial r_i} = \psi_i'(k_i(r_i))k_i'(r_i) - \psi_i'(k_i(s_i))k_i'(r_i), \quad \textit{for any } i = 1, 2, \ldots, n. \tag{4.2}$$

Proof. Since the divergence D_{Φ} is a decomposable divergence, we know that there are divergences D_i on $R^+ \times R^+$ $(i = 1, 2, \ldots, n)$ such that for any $\mathbf{r}, \mathbf{s} \in \mathcal{M}$,

$$D_{\Phi}[\mathbf{r} : \mathbf{s}] = \sum_{i=1}^{n} D_i[r_i, s_i]. \tag{4.3}$$

By the definition of a Bregman divergence in $\mathcal{M}$, we have

$$\sum_{i=1}^{n} D_i[r_i : s_i] = \Phi(k(\mathbf{r})) - \Phi(k(\mathbf{s})) - \nabla\Phi(k(\mathbf{s})) \bullet (k(\mathbf{r}) - k(\mathbf{s})),$$
$$\text{for any } \mathbf{r}, \mathbf{s} \in \mathcal{M}.$$

Hence,

$$\frac{\partial \sum D_i[r_i : s_i]}{\partial r_i} = \frac{\partial(\Phi(k(\mathbf{r})) - \Phi(k(\mathbf{s})) - \nabla\Phi(k(\mathbf{s})) \bullet (k(\mathbf{r}) - k(\mathbf{s})))}{\partial r_i}.$$

So,

$$\frac{\partial D_i[r_i : s_i]}{\partial r_i} = \Phi_i'(k(\mathbf{r}))k_i'(r_i) - \Phi_i'(k(\mathbf{s}))k_i'(r_i). \tag{4.4}$$

Therefore,

$$\Phi_i'(k(\mathbf{r})) = \frac{\partial D_i[r_i : s_i]}{\partial x_i}\frac{1}{k_i'(r_i)} + \Phi_i'(k(\mathbf{s})).$$

Thus,

$$\Phi_i'(k(\mathbf{r})) = \frac{\partial D_i[r_i, 1]}{\partial r_i}\frac{1}{k_i'(r_i)} + \Phi_i'(k(\mathbf{1})), \tag{4.5}$$

where $\mathbf{1} = (1, 1, \ldots, 1)$, and $k(\mathbf{1}) = (k_1(1), k_2(1), \ldots, k_n(1))$.

For each $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \Omega_0$ and $i = 1, 2, \ldots, n$, we set

$$g_i(x_i) := \frac{\partial D_i[k_i^{-1}(x_i), 1]}{\partial r_i} \frac{1}{k_i'(k_i^{-1}(x_i))} + \Phi_i'(k(\mathbf{1})),$$

$$\psi_i(x_i) := \int_{p_i}^{x_i} g_i(t)dt \quad \text{for a fixed } p_i \in k_i(R^+).$$

Then by equation 4.5, we see that for every $i = 1, 2, \ldots, n$,

$$\psi_i'(x_i) = g_i(x_i) = \Phi_i'(k(k^{-1}(\mathbf{x}))) = \Phi_i'(\mathbf{x}), \tag{4.6}$$

$\psi_i \in C^2(k_i(R^+))$ and ψ_i is a strictly convex function. The equality 4.6 implies that equation 4.1 is true.

Morover, combining equations 4.3, 4.4, and 4.6, we deduce that equation 4.2 holds.

Now, we are in a position to prove that if a divergence is a weighted *f*-divergence as well as a Bregman divergence, then it is a weighted α-divergence. This means that the weighted α-divergence is the unique class of divergences sitting at the intersection of the weighted *f*-divergence and Bregman divergence classes. This result generalizes the main theorem of Amari (2009). Moreover, the approach in our proof is somewhat different from that in Amari (2009).

Theorem 1. *If a divergence is a weighted f-divergence as well as a Bregman divergence, then it is a weighted α-divergence.*

Proof. Let the divergence $D[\mathbf{m} : \mathbf{n}]$ be both the *f*-divergence and Bregman divergence on $\mathcal{M}$. Then the definition of a weighted *f*-divergence in $\mathcal{M}$ says that

$$D[\mathbf{m} : \mathbf{n}] = D_f^w[\mathbf{m} : \mathbf{n}] = \sum_{i=1}^{n} \mu_i m_i f\left(\frac{n_i}{m_i}\right)$$

for every $\mathbf{m}$, $\mathbf{n} \in \mathcal{M}$. Clearly, *f*-divergence is decomposable. Therefore, by lemma 1, we know that there exist $\psi_i \in C^2(k_i(R^+))$ $(i = 1, 2, \ldots, n)$ such that ψ_i $(i = 1, 2, \ldots, n)$ are strictly convex functions, and for every $\mathbf{m}, \mathbf{n} \in \mathcal{M}$,

$$\frac{\partial(D[\mathbf{m} : \mathbf{n}])}{\partial m_i} = \psi_i'(k_i(m_i))k_i'(m_i)) - \psi_i'(k_i(n_i))k_i'(m_i),$$

$$\text{for any } i = 1, 2, \ldots, n.$$

Therefore,

$$\begin{aligned}\frac{\partial^2\left[\mu_1 m_1 f\left(\frac{n_1}{m_1}\right)\right]}{\partial n_1 \partial m_1} &= \frac{\partial^2(D[\mathbf{m}:\mathbf{n}])}{\partial n_1 \partial m_1} \\ &= \frac{\partial[\psi_1'(k_1(m_1))k_1'(m_1)) - \psi_1'(k_1(n_1))k_1'(m_1)]}{\partial n_1} \\ &= -\psi_1''(k_1(n_1))k_1'(n_1)k_1'(m_1).\end{aligned}$$

On the other hand

$$\begin{aligned}\frac{\partial^2\left[\mu_1 m_1 f\left(\frac{n_1}{m_1}\right)\right]}{\partial n_1 \partial m_1} &= \frac{\partial\left[\mu_1 f\left(\frac{n_1}{m_1}\right) - \mu_1 \frac{n_1}{m_1} f'\left(\frac{n_1}{m_1}\right)\right]}{\partial n_1} \\ &= \mu_1 f'\left(\frac{n_1}{m_1}\right)\frac{1}{m_1} - \mu_1 \frac{1}{m_1} f'\left(\frac{n_1}{m_1}\right) - \mu_1 \frac{n_1}{m_1^2} f''\left(\frac{n_1}{m_1}\right) \\ &= -\mu_1 \frac{n_1}{m_1^2} f''\left(\frac{n_1}{m_1}\right).\end{aligned}$$

So, we obtain

$$-\mu_1 \frac{n_1}{m_1^2} f''\left(\frac{n_1}{m_1}\right) = -\psi_1''(k_1(n_1))k_1'(n_1)k_1'(m_1).$$

Hence,

$$f''\left(\frac{n_1}{m_1}\right) = \frac{\psi_1''(k_1(n_1))k_1'(n_1)}{\mu_1 n_1} k_1'(m_1)m_1^2.$$

Let

$$\begin{aligned} x &= n_1, \quad h(x) = \frac{\psi_1''(k_1(x))k_1'(x)}{x}, \\ y &= \frac{1}{m_1}, \quad t(y) = \frac{k_1'(1/y)}{\mu_1 y^2}. \end{aligned}$$

Then we have

$$f''(xy) = h(x)t(y), \quad \text{for all } x, y \in (0, \infty).$$

Since f is a standard convex function, that is, f is a twice continuously differentiable convex function on R^+ such that

$$f(1) = 0, \quad f'(1) = 0, \quad f''(1) = 1,$$

we see that for all $x, y \in (0, \infty)$,

$$f''(xy) = h(x)t(y)f''(1) = h(x)t(1)h(1)t(y) = f''(x)f''(y).$$

It follows from a basic theorem about the functional equations that

$$f''(x) = x^\beta \quad \text{for all } x \in R^+,$$

where β is a constant.

Therefore, we obtain

$$f(x) = \begin{cases} \dfrac{x^{\beta+2}}{(\beta+1)(\beta+2)} - \dfrac{x}{\beta+1} + \dfrac{1}{\beta+2} & \beta \neq -1, -2, \\ x\ln x - x + 1 & \beta = -1, \\ -\ln x + x - 1 & \beta = -2. \end{cases}$$

Let $\alpha = 2\beta + 3$. Then

$$f(x) = \begin{cases} \dfrac{4}{1-\alpha^2}(1 - x^{\frac{1+\alpha}{2}}) + \dfrac{2}{1-\alpha}(x-1) & \alpha \neq \pm 1, \\ x\ln x - x + 1 & \alpha = 1, \\ -\ln x + x - 1 & \alpha = -1. \end{cases}$$

This means that the related weighted f-divergence is a weighted α-divergence.

5 Conclusion

In this note, we have studied the conjecture presented by Amari (2009) and found a counterexample to show that the conjecture does not hold generally. Moreover, we have investigated two classes of Φ-divergence (Zhang, 2004), weighted f-divergence and weighted, α-divergence and proved that if a divergence is a weighted f-divergence as well as a Bregman divergence, then it is a weighted α-divergence. This result reduces in form to the main theorem established by Amari (2009) when $\mu_i = 1$ $(i = 1, \ldots, n)$.

Acknowledgments

I thank the referees very much for helpful comments and suggestions, which led to my revision of section 3. Moreover, I am very grateful to the referees for bringing the references Amari (2016) and Jiao et al. (2015) to my attention.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

Amari, S. (1985). *Differential geometric methods in statistics.* New York: Springer.

Amari, S. (2007). Integration of stochastic models by minimizing α-divergence. *Neural Computation 19*, 2780–2796.

Amari, S. (2009). α-divergence is unique, belonging to both *f*-divergence and Bregman divergence. *IEEE Transaction on Information Theory*, *55*, 4925–4931.

Amari, S. (2016). *Information geometry and its applications.* Berlin: Springer.

Amari S., & Nagaoka, H. (2000). *Methods of information geometry*. New York: Oxford University Press.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, *7*, 200–217.

Cichocki, A., Adunek, R., Phan, A. H., & Amari S. (2009). *Non-negative matrix and tensor factorizations: Applications to explanatory multi-way data analysis and blind source separation*. New York: Wiley.

Csiszár, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tudomnyos Akadmia Matematikai Kutat Intzetnek Kzlemnyei*, *8*, 85–108.

Csiszár, I. (1974). Information measures: A critical survey. In *Transactions of the 7th Conference on Information Theory Statistical Decision Functions, Random Processes* (pp. 73–86). Prague: Academia.

Eguchi, S. (2008). Information divergence geometry and the application to statistical machine learning. In F. Emmert-Streib & M. Dehmer (Eds.), *Information theory and statistical learning* (pp. 309–332). Berlin: Springer.

Eguchi, S., & Copas, J. (2002). A class of logistic type discriminant function. *Biometrika*, *89*, 1–22.

Jiao, J. T., Courtade, T. A., No, A., Venkat, K. & Weissman, T. (2015), Information measure: The curious case of the binary alphabet. *IEEE Transactions on Information Theory*, *60*, 7616–7626.

Murata, N., Takenouchi, T., Kanamori, T., & Eguchi, S. (2004). Information geometry of *U*-boost and Bregman divergence. *Neural Computation*, *26*, 1651–1686.

Nielsen, F. (Ed.). (2009). *Emerging trends in visual computing*. Berlin: Springer-Verlag.

Nielsen, F., & Noch, R. (2009). Sided and symmetrized Bregman divergence. *IEEE Transactions on Information Theory*, *55*, 2882–2904.

Rao, C. R. (1987). Differential metrics in probability spaces. In S. Amari, O. Barndorff-Nielsen, R. Kass, S. Lauritzen, & C.R. Rao (Eds.), *Differential geometry*

in statistical interference (pp. 217–240). Hayward, CA: Institute of Mathematical Interference.

Taneja, I., & Kumar, P. (2004). Relative information of type *s*, Csiszár's *f*-divergence, and information inequalities. *Information Science*, *166*, 105–125.

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, *16*, 159–195.

Received April 15, 2016; accepted May 28, 2016.