α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes

Shun-Ichi Amari, *Life Fellow, IEEE*

***Abstract*—A divergence measure between two probability distributions or positive arrays (positive measures) is a useful tool for solving optimization problems in optimization, signal processing, machine learning, and statistical inference. The Csiszár f-divergence is a unique class of divergences having information monotonicity, from which the dual α geometrical structure with the Fisher metric is derived. The Bregman divergence is another class of divergences that gives a dually flat geometrical structure different from the α-structure in general. Csiszár gave an axiomatic characterization of divergences related to inference problems. The Kullback–Leibler divergence is proved to belong to both classes, and this is the only such one in the space of probability distributions. This paper proves that the α-divergences constitute a unique class belonging to both classes when the space of positive measures or positive arrays is considered. They are the canonical divergences derived from the dually flat geometrical structure of the space of positive measures.**

***Index Terms*—Bregman divergence, canonical divergence, dually flat structure, f-divergence, Fisher information, information geometry, information monotonicity.**

I. Introduction

VARIOUS divergence measures are used to show how two points are separated in a space S of signals.

A divergence is used to solve an optimization problem. Given a point p find q in a model set that minimizes the divergence $D[p:q]$ from p to q. There are many problems of this type in signal processing and vision analysis; see [26] and [27], for example. A divergence is also used for a clustering problem, where a set of points is divided into a number of clusters such that these points close to each other in the sense of the divergence belong to the same cluster. The center of a cluster is also defined by using the divergence [27], [7].

When the space is a set of probability distributions, the most well-known divergences belong to the class of f-divergences and the class of Bregman divergences. These classes have their own characteristics. Csiszár [1] gave an axiomatic approach to elucidate the structure of divergences. The concepts of regularity, locality, symmetricity, and transitivity are introduced, which played fundamental roles to characterize the classes f-divergences and Bregman divergences. It was proved that the Kullback–Leibler divergence is the only divergence belonging to the intersection of the two classes in the framework of inference under linear constraints.

In many applications [8]–[10], the normalizing constraint of a probability distribution that the total mass is 1 is relaxed, and one treats signals represented by positive measures or positive arrays. For example, a visual signal is a two-dimensional array with nonnegative elements. The set of nonnegative symmetric matrices is also useful in many engineering problems. This paper studies the relation between the two classes of divergence in the space of positive measures.

Let $X = \{x_1, \dots, x_n\}$ be a set of n elements on which a positive measure $\boldsymbol{m} = (m_1, \dots, m_n)$ is defined

$$m_i = \text{Measure}\,[x_i] > 0, \qquad i = 1, \dots, n. \tag{1}$$

We denote the set of all positive measures by $\mathcal{M}$. When

$$\sum m_i = 1 \tag{2}$$

it is a probability measure, denoted by $\boldsymbol{p} = (p_1, \dots, p_n)$. The set of all probability distributions is a subset of $\mathcal{M}$ and is denoted by $\mathcal{P}, \mathcal{P} \subset \mathcal{M}$.

This paper uses a geometrical approach, where dual flatness in information geometry [2] plays a fundamental role. The main result of this paper is to prove that the class of α-divergences [2] is the intersection of the classes of f-divergences and Bregman divergences in the manifold of positive measures. This extends the results of Csiszár [1] to problems of nonlinear constraints with a flat structure.

The f-divergence was introduced by Csiszár [3], [4] and Ali and Silvey [11]. See also [12] for its detailed properties. It is characterized by information monotonicity [5]. Csiszár gave a more fundamental characterization by using statisticity [1]. From the point of view of geometry, it induces a unique geometrical structure consisting of the Fisher information Riemannian metric together with dually coupled $\pm\alpha$ affine connections [2]. The induced geometry is not flat in general, except for the Kullback–Leibler (KL) divergence $(\alpha = \pm 1)$. The α-divergence [2], [13] is a special class of f-divergences. It was used by Chernoff [14] to evaluate classification errors, and also Renyi [15] to generalize the concept of entropy. Tsallis [16] proposed a similar concept from the physics point of view of nonextensive entropy. Its information geometrical structure was given by Amari and Nagaoka [2]. We show that the α-divergences form a class of canonical divergences in the dually flat manifold of positive measures.

The Bregman divergence was introduced by Bregman [6] to solve convex optimization problems. It is derived from a convex function and is characterized by transitivity in the Csiszár

Manuscript received April 22, 2009; revised July 06, 2009. Current version published October 21, 2009.

The author is with the Riken Brain Science Institute, Saitama 351-0198, Japan (e-mail: amari@brain.riken.jp).

Communicated by H. Bölcskei, Associate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2009.2030485

approach [1]. The Bregman divergence [6], [7] induces a geometrical structure in S, which is a dually flat Riemannian space in information geometry. However, it is not information monotone, and the derived Riemannian metric is in general different from the Fisher information matrix, except for the KL-divergence. The dually flat Riemannian structure is induced through the Legendre transformation. Conversely, a dually flat Riemannian manifold always possesses a convex function, from which a canonical divergence is introduced as the related Bregman divergence.

In this paper, we consider only a decomposable divergence, such that the divergence measure $D[\boldsymbol{m} : \boldsymbol{n}]$ between $\boldsymbol{m}, \boldsymbol{n} \in \mathcal{M}$ is written as a sum of components

$$D[\boldsymbol{m} : \boldsymbol{n}] = \sum_{i=1}^{n} D(m_i, n_i). \tag{3}$$

This is characterized by semisymmetricity [1]. The f-divergence belongs to this class, although a Bregman divergence can be more general.

II. f-Divergence

A. f-Divergence in $\mathcal{P}$

Let $f(u)$ be a convex function satisfying $f(1) = 0$. Then, the f-divergence from $\boldsymbol{p}$ to $\boldsymbol{q}$ is defined in $\mathcal{P}$ as

$$D_f[\boldsymbol{p} : \boldsymbol{q}] = \sum_{i=1}^{n} p_i f\left(\frac{q_i}{p_i}\right). \tag{4}$$

It has the following properties:

$$D_f[\boldsymbol{p} : \boldsymbol{q}] \geq 0 \tag{5}$$

$$D_f[\boldsymbol{p} : \boldsymbol{q}] = 0, \qquad \text{iff } \boldsymbol{p} = \boldsymbol{q}. \tag{6}$$

A typical example is

$$f(u) = -\log u \tag{7}$$

which gives the KL-divergence

$$D_{\mathrm{KL}}(\boldsymbol{p} : \boldsymbol{q}) = \sum p_i \log \frac{p_i}{q_i}. \tag{8}$$

Its dual, that is, $D_{\mathrm{KL}}(\boldsymbol{q} : \boldsymbol{p})$, is given by

$$f(u) = u \log u. \tag{9}$$

The f-divergence further satisfies the following relations:

1) for $f_c(u) = f(u) - c(u-1)$

$$D_f[\boldsymbol{p} : \boldsymbol{q}] = D_{fc}[\boldsymbol{p} : \boldsymbol{q}]; \tag{10}$$

2) for $c > 0$

$$D_{cf}[\boldsymbol{p} : \boldsymbol{q}] = c D_f[\boldsymbol{p} : \boldsymbol{q}]. \tag{11}$$

Because of 1), we may choose f such that it satisfies $f'(1) = 0$, without loss of generality, by putting $c = f'(1)$, when f is differentiable. Because of 2), we may normalize f such that $f''(1) = 1$. We call such f a standard convex function.

B. Information Monotonicity

A partition of X into m groups $(m < n)$ is a coarse-grained version of X. We then have m subsets $G_1, \ldots, G_m$ of X such that

$$G_i \cap G_j = \phi, \quad \cup G_i = X. \tag{12}$$

The partition naturally induces a probability distribution $\bar{\boldsymbol{p}}$ over $G_1, \ldots, G_m$

$$\bar{p}_i = \mathrm{Prob}\left\{G_i\right\} = \sum_{x_k \in G_i} p_k. \tag{13}$$

Since coarse-graining loses information by summarizing elements within each subset G_i, it is natural to stipulate a monotonic relation

$$D[\boldsymbol{p} : \boldsymbol{q}] \geq D[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}]. \tag{14}$$

Consider the case where p_j and q_j are proportional inside each class G_i

$$\frac{q_j}{p_j} = \lambda_i \quad x_j \in G_i, \qquad i = 1, \ldots, m. \tag{15}$$

In other words, the conditional distributions of $\boldsymbol{p}$ and $\boldsymbol{q}$ are equal, conditioned on $x \in G_i$. Then, it is natural to assume

$$D[\boldsymbol{p} : \boldsymbol{q}] = D[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}] \tag{16}$$

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

The above properties are called information monotonicity. The f-divergence is the only class of decomposable information monotonic divergences [4]. We give a short proof of this fact in the Appendix. Information monotonicity is essentially the same as the assumption used in [17] for characterizing invariant geometry to be introduced in a manifold of probability distributions. From this, the Fisher information metric and α-affine connections were derived (see [2] for more details). Information monotonicity is also a basis of quantum information geometry [18].

The f-divergence is also characterized by the statisticity of Csiszár [1]. It is known that any f-divergence induces a geometrical structure with the Fisher information metric and $\pm\alpha$ affine connections with $\alpha = 3 + f'''(1)$ [2].

C. f-Divergence and α-Divergence in $\mathcal{M}$

The f-divergence can be formally extended to $\mathcal{M}$

$$D_f[\boldsymbol{m} : \boldsymbol{n}] = \sum m_i f\left(\frac{n_i}{m_i}\right). \tag{17}$$

However, the relation

$$D_f[\boldsymbol{m} : \boldsymbol{n}] = D_{f_c}[\boldsymbol{m} : \boldsymbol{n}] \tag{18}$$

no longer holds for $f_c(u) = f(u) - c(u-1)$ [1].

Lemma: When f is a standard convex function, $D_f[\boldsymbol{m} : \boldsymbol{n}]$ is a divergence in $\mathcal{M}$ satisfying (5) and (6).

The proof is easy. This lemma shows that we need to use only a standard convex function in the case of $\mathcal{M}$.

The α-divergence is a special case of the f-divergence. Let us define the following α-function parameterized by α:

$$f_\alpha(u) = \begin{cases} \dfrac{4}{1-\alpha^2}\left(1-u^{\frac{1+\alpha}{2}}\right) - \dfrac{2}{1-\alpha}(u-1) & \alpha \neq \pm 1 \\ u\log u - (u-1), & \alpha = 1 \\ -\log u + (u-1), & \alpha = -1. \end{cases} \tag{19}$$

These functions are standard. The related f-divergence is thus called the α-divergence [2]

$$D_\alpha[\boldsymbol{m}:\boldsymbol{n}] = \begin{cases} \dfrac{4}{1-\alpha^2}\sum\left\{\dfrac{1-\alpha}{2}m_i + \dfrac{1+\alpha}{2}n_i - m_i^{\frac{1+\alpha}{2}} n_i^{\frac{1-\alpha}{2}}\right\}, & \alpha \neq \pm 1 \\ \sum\left(m_i - n_i + n_i \log\dfrac{n_i}{m_i}\right), & \alpha = 1 \\ \sum\left(n_i - m_i + m_i \log\dfrac{m_i}{n_i}\right), & \alpha = -1. \end{cases} \tag{20}$$

When $\sum p_i = \sum q_i = 1$ is satisfied, we have the α-divergence $D_\alpha[\boldsymbol{p}:\boldsymbol{q}]$ in $\mathcal{P}$. The -1-divergence is the KL-divergence, the 1-divergence is its dual, and the 0-divergence is the squared Hellinger distance.

Because of the information monotonicity, all the f-divergences give the same Fisher information matrix g_{ij} as a Riemannian metric in $\mathcal{P}$. The induced affine connection is given by the $\pm\alpha$ connections, where $\alpha = 3 + f'''(1)$. Although the α affine connection is not flat in $\mathcal{P}$, in general, we will prove in a later section that it is dually flat in $\mathcal{M}$ for any α.

III. Bregman Divergence

A. *Convex Function and Bregman Divergence*

Let ψ be a convex function defined on $\boldsymbol{r} \in V \subset \boldsymbol{R}^n$, where V is an open convex set covered by a single coordinate system $\boldsymbol{r}$. The Bregman divergence between $\boldsymbol{r}, \boldsymbol{s} \in V$ is defined by

$$D_\psi[\boldsymbol{r}:\boldsymbol{s}] = \psi(\boldsymbol{r}) - \psi(\boldsymbol{s}) - \nabla\psi(\boldsymbol{s})\cdot(\boldsymbol{r}-\boldsymbol{s}) \tag{21}$$

where $\nabla\psi$ is the gradient of ψ.

Let us define

$$\boldsymbol{r}^* = \nabla\psi(\boldsymbol{r}). \tag{22}$$

This is the Legendre transformation, and the correspondence between $\boldsymbol{r}$ and $\boldsymbol{r}^*$ is one-to-one. Hence, $\boldsymbol{r}^*$ is regarded as another (nonlinear) coordinate system of V. We can obtain a dual potential function by

$$\psi^*(\boldsymbol{r}^*) = \max_{\boldsymbol{r}}\{\boldsymbol{r}\cdot\boldsymbol{r}^* - \psi(\boldsymbol{r})\} \tag{23}$$

which is convex in $\boldsymbol{r}^*$, and a pair of $\boldsymbol{r}$ and $\boldsymbol{r}^*$ satisfies the following relation:

$$\psi(\boldsymbol{r}) + \psi^*(\boldsymbol{r}^*) - \boldsymbol{r}\cdot\boldsymbol{r}^* = 0. \tag{24}$$

The inverse transformation is given by

$$\boldsymbol{r} = \nabla\psi^*(\boldsymbol{r}^*). \tag{25}$$

The Bregman divergence can be rewritten in the dual form as

$$D[\boldsymbol{r}:\boldsymbol{s}] = \psi(\boldsymbol{r}) + \psi^*(\boldsymbol{s}^*) - \boldsymbol{r}\cdot\boldsymbol{s}^*. \tag{26}$$

We have thus two coordinate systems $\boldsymbol{r}$ and $\boldsymbol{r}^*$ in V. They define two flat structures in V such that a curve in V with parameter t

$$\boldsymbol{r}(t) = t\boldsymbol{a} + \boldsymbol{b} \tag{27}$$

is a ψ-geodesic curve, and

$$\boldsymbol{r}^*(t) = t\boldsymbol{c}^* + \boldsymbol{d}^* \tag{28}$$

is a ψ^*-geodesic curve, where $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}^*$, and $\boldsymbol{d}^*$ are constants.

The Riemannian metric is defined by the metric tensors $G = (g_{ij})$ and $G^* = (g^*_{ij})$

$$g_{ij} = \frac{\partial^2}{\partial r_i \partial r_j}\psi(\boldsymbol{r}) \tag{29}$$

$$g^*_{ij} = \frac{\partial^2}{\partial r^*_i \partial r^*_j}\psi^*(\boldsymbol{r}^*) \tag{30}$$

in the two coordinate systems, and they are mutual inverses $G^* = G^{-1}$. Because the squared local distance of two nearby points $\boldsymbol{r}$ and $\boldsymbol{r} + d\boldsymbol{r}$ is given by

$$D[\boldsymbol{r}:\boldsymbol{r}+d\boldsymbol{r}] = \frac{1}{2}\sum g_{ij}(\boldsymbol{r})dr_i dr_j \tag{31}$$

$$= \frac{1}{2}\sum g^*_{ij}(\boldsymbol{r}^*)dr^*_i dr^*_j \tag{32}$$

they give the same Riemannian distance. The above two types of geodesics (27) and (28) are different from the Riemannian geodesic derived from the Riemannian metric, and the minimality of the curve length does not hold for them. Two geodesic curves (27) and (28) are orthogonal when they intersect and their tangent vectors are orthogonal in the sense of the Riemannian metric. However, the orthogonality condition can be represented in a simple form in terms of the two dual coordinates as

$$\langle\boldsymbol{a}, \boldsymbol{b}^*\rangle = \sum a_i b^*_i = 0. \tag{33}$$

A generalized Pythagorean theorem and projection theorem hold for a dually flat space [2], [7], [19].

Generalized Pythagorean Theorem: Let $\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{q}$ be three points in V such that the ψ^*-geodesic connecting $\boldsymbol{r}$ and $\boldsymbol{s}$ is orthogonal to the ψ-geodesic connecting $\boldsymbol{s}$ and $\boldsymbol{q}$. Then

$$D_\psi[\boldsymbol{r}:\boldsymbol{s}] + D_\psi[\boldsymbol{s}:\boldsymbol{q}] = D_\psi[\boldsymbol{r}:\boldsymbol{q}]. \tag{34}$$

Let S be a submanifold of V. Given $\boldsymbol{p} \in V$, a point $\boldsymbol{q} \in S$ is said to be the ψ-projection (ψ^*-projection) of $\boldsymbol{p}$ to S when the ψ-geodesic (ψ^*-geodesic) connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ is orthogonal to S with respect to the Riemannian metric g_{ij}.

Projection Theorem: Given $\boldsymbol{p} \in V$, the minimizer of $D_\psi(\boldsymbol{p}, \boldsymbol{r}), \boldsymbol{r} \in S$ is the ψ^*-projection of $\boldsymbol{p}$ to S, and the minimizer of $D_\psi(\boldsymbol{r}, \boldsymbol{p}), \boldsymbol{r} \in S$, is the ψ-projection of $\boldsymbol{p}$ to S.

B. Decomposable Divergence

We will consider only the case of $\mathcal{M}$, although our discussion holds true in the general case. Let $k(u)$ be a monotone function, and let us introduce a nonlinear coordinate system $\boldsymbol{r}$ in $\mathcal{M}$ defined by

$$r_i = k(m_i). \tag{35}$$

Zhang [20] used a special type of such representation separately from a convex function to define a Bregman divergence. Let $U(z)$ be a convex function. Then, we have a convex function of $\boldsymbol{r}$

$$\psi(\boldsymbol{r}) = \sum U(r_i). \tag{36}$$

The dual of U is

$$U^*(r^*) = \max_r \{rr^* - U(r)\} \tag{37}$$

and the dual coordinates are

$$r_i^* = U'(r_i). \tag{38}$$

The dual convex function is

$$\psi^*(\boldsymbol{r}^*) = \sum U^*(r_i^*). \tag{39}$$

The ψ-divergence between $\boldsymbol{r}$ and $\boldsymbol{s}$ is decomposable and given by the sum of components [8]

$$D_\psi[\boldsymbol{r} : \boldsymbol{s}] = \sum \{U(r_i) + U^*(s_i^*) - r_i s_i^*\}. \tag{40}$$

Given two functions $k(m)$ and $U(r)$, we can introduce a dually flat Riemannian structure in $\mathcal{M}$, where the flat coordinates are not $\boldsymbol{m}$ but rather $\boldsymbol{r} = k(\boldsymbol{m})$. There are infinitely many such structures depending on the choice of k and U. We show two typical examples, the α-divergence and β-divergence, both of which use a power function of the type u^q including the log and exponential function as the limiting cases.

The Bregman divergence is characterized by the property of transitivity [1]. Geometrically, it is characterized by the property of dual flatness [4], which is applicable to the case of nonlinear constraints through a nonlinear representation function.

C. β-Divergence

Use $\boldsymbol{m}$ itself as a coordinate system of $\mathcal{M}$; that is, the representation function is $k(u) = u$. The β-divergence [21] is induced by the potential function

$$U_\beta(z) = \begin{cases} \dfrac{1}{\beta+1}(1+\beta z)^{\frac{\beta+1}{\beta}}, & \beta > 0 \\ \exp z, & \beta = 0. \end{cases} \tag{41}$$

The β-divergence is thus written as

$$\begin{aligned} &D_\beta[\boldsymbol{m} : \boldsymbol{n}] \\ &\quad = \begin{cases} \dfrac{1}{\beta+1}\sum\left(n_i^{\beta+1} - m_i^{\beta+1}\right) - \frac{1}{\beta}\sum m_i^\beta (n_i - m_i), & \beta > 0 \\ \sum\left(m_i \log \dfrac{m_i}{n_i} + n_i - m_i\right), & \beta = 0. \end{cases} \end{aligned} \tag{42}$$

It is the KL-divergence when $\beta = 0$, but it is different from an α-divergence when $\beta > 0$.

This is the unique class of divergence satisfying the regular, local, transitive, and scale-invariant properties of Csiszár [1].

Minami and Eguchi [22] demonstrated that statistical inference based on the β-divergence ($\beta > 0$) is robust. Such idea has been applied to machine learning in [8]. The β-divergence induces a dually flat structure in $\mathcal{M}$. Since its flat coordinates are $\boldsymbol{m}$, the restriction

$$\sum m_i = 1 \tag{43}$$

is a linear constraint. Hence, the manifold $\mathcal{P}$ is also dually flat, where $\boldsymbol{p}$ is its flat coordinates. The dual flat coordinates are

$$m_i^* = \begin{cases} (1+\beta m_i)^{\frac{1}{\beta}}, & \beta \neq 0 \\ \exp m_i, & \beta = 0 \end{cases} \tag{44}$$

depending on β.

D. α-Divergence as Bregman Divergence

Let us define the α-representation by

$$k_\alpha(u) = \begin{cases} \dfrac{2}{1-\alpha}\left(u^{\frac{1-\alpha}{2}} - 1\right), & \alpha \neq 1 \\ \log u, & \alpha = 1. \end{cases} \tag{45}$$

Then, the α-coordinates $\boldsymbol{r}_\alpha$ of $\mathcal{M}$ are given by

$$r_i^{(\alpha)} = k_\alpha(m_i). \tag{46}$$

Next we use a convex function

$$U_\alpha(r) = \frac{2}{1+\alpha} k_\alpha^{-1}(r). \tag{47}$$

In terms of m, this is a linear function

$$U_\alpha\{r(m)\} = \frac{2}{1+\alpha} m \tag{48}$$

which is not a (strictly) convex function of m. The α-potential function defined by

$$\begin{aligned} \psi_\alpha(\boldsymbol{r}) &= \frac{2}{1+\alpha} \sum k_\alpha^{-1}(r_i) \\ &= \frac{2}{1+\alpha} \sum \left(1 + \frac{1-\alpha}{2} r_i\right)^{\frac{2}{1-\alpha}} \end{aligned} \tag{49}$$

is a convex function of $\boldsymbol{r}$.

The dual potential is simply given by

$$\psi^*_\alpha(\boldsymbol{r}^*) = \psi_{-\alpha}(\boldsymbol{r}^*) \tag{50}$$

and the dual affine coordinates are

$$\boldsymbol{r}^{*(\alpha)} = \boldsymbol{r}^{(-\alpha)} = k_{-\alpha}(\boldsymbol{m}). \tag{51}$$

We can then prove the following theorem.

Theorem: The α-divergence is the Bregman divergence derived from the α-representation k_α and the linear U_α function of m.

Proof: We have the Bregman divergence between $\boldsymbol{r} = k(\boldsymbol{m})$ and $\boldsymbol{s} = k(\boldsymbol{n})$ based on ψ_α as

$$D_\alpha[\boldsymbol{r} : \boldsymbol{s}] = \psi_\alpha(\boldsymbol{r}) + \psi_{-\alpha}(\boldsymbol{s}^*) - \boldsymbol{r} \cdot \boldsymbol{s}^* \tag{52}$$

where $\boldsymbol{s}^*$ is the dual of $\boldsymbol{s}$. By substituting (49), (50), and (51) in (52), we see that $D_\alpha[\boldsymbol{r}(\boldsymbol{m}) : \boldsymbol{s}(\boldsymbol{n})]$ is equal to the α-divergence defined in (20).

This proves that the α-divergence for any α belongs to the intersection of the classes of f-divergences and Bregman divergences in $\mathcal{M}$. Hence, it possesses information monotonicity and the induced geometry is dually flat at the same time. However, the constraint $\sum m_i = 1$ is not linear in the flat coordinates $\boldsymbol{r}^\alpha$ or $\boldsymbol{r}^{-\alpha}$ except for the case $\alpha = \pm 1$. Hence, $\mathcal{P}$ is not dually flat except for $\alpha = \pm 1$, and it is a curved submanifold in $\mathcal{M}$. Hence, the α-divergences $(\alpha \neq \pm 1)$ do not belong to the class of Bregman divergences in $\mathcal{P}$. This proves that the KL-divergence belongs to both classes of divergences in $\mathcal{P}$ and is unique.

The α-divergences are used in many applications, e.g., [23] and [24].

IV. Uniqueness of the α-Divergence

Theorem: The α-divergence is the unique class of divergences sitting at the intersection of the f-divergence and Bregman divergence classes.

Proof: The f-divergence is of the form (17) and the decomposable Bregman divergence is of the form

$$\sum \tilde{U}(m_i) + \sum \tilde{U}^*(n_i) - \sum r_i(m_i) r_i^*(n_i) \tag{53}$$

where

$$\tilde{U}(m) = U(k(m)) \tag{54}$$

and so on. When they are equal, we have f, r, and r^* that satisfy

$$mf\left(\frac{n}{m}\right) = r(m) r^*(n) \tag{55}$$

except for additive terms depending only on m or n. Differentiating the above by n, we get

$$f'\left(\frac{n}{m}\right) = r(m) r^{*'}(n). \tag{56}$$

By putting $x = n, y = 1/m$, we get

$$f'(xy) = r\left(\frac{1}{y}\right) r^{*'}(x). \tag{57}$$

Hence, by putting $h(u) = \log f'(u)$, we have

$$h(xy) = s(x) + t(y) \tag{58}$$

where $s(x) = \log r^{*'}(x), t(y) = r(1/y)$. By differentiating the above equation with respect to x and putting $x = 1$, we get

$$h'(y) = \frac{c}{y}. \tag{59}$$

This proves that f is of the form

$$f(u) = \begin{cases} cu^{\frac{1+\alpha}{2}}, & \alpha \neq \pm 1 \\ cu \log u, & \alpha = 1 \\ c \log u, & \alpha = -1. \end{cases} \tag{60}$$

By changing the above into the standard form, we arrive at the α-divergence.

Corollary: The KL-divergence and its dual are unique divergences belonging to the f-divergence and Bregman divergence classes.

V. Divergence and Geometry

A divergence $D[\boldsymbol{r} : \boldsymbol{s}]$ uniquely determines the geometry of S, consisting of a Riemannian metric and a dual pair of affine connections [2], [25]. However, a dual geometry does not uniquely determine a divergence.

In the case of the f-divergence, the Riemannian metric is always the Fisher information matrix, and affine connections are determined from $\alpha = 3 + f'''(1)$. Since the two standard convex functions $f(u)$ and

$$\bar{f}(u) = f(u) + c(u-1)^4 \tag{61}$$

satisfy $f'''(1) = \bar{f}'''(1)$, they give the same dual pair of affine connections.

Given a Bregman divergence $D[\boldsymbol{r} : \boldsymbol{s}]$, the new divergence

$$\bar{D}[\boldsymbol{r} : \boldsymbol{s}] = D[\boldsymbol{r} : \boldsymbol{s}] + c \sum (r_i - s_i)^4 \tag{62}$$

gives the same dually flat geometry, although it is not a Bregman divergence.

When a manifold S is dually flat, we can always find affine coordinate systems $\boldsymbol{r}$ and its dual $\boldsymbol{r}^*$, as well as convex functions $k(\boldsymbol{r})$ and $k(\boldsymbol{r}^*)$. The affine coordinates are determined up to affine transformations and convex functions up to linear terms. However, the derived Bregman divergence is unique. We call this the canonical divergence of a dually flat manifold. The KL-divergence is the canonical divergence of the $\alpha = \pm 1$ dually flat structure of $\mathcal{P}$ and the α-divergence is the canonical divergence of $\mathcal{M}$.

When S is not flat, we do not have a canonical divergence. When S is embedded in a higher dimensional dually flat manifold $\tilde{S}$ such that the dual structure of S is inherited from that of $\tilde{S}$, we can construct a canonical divergence $\tilde{D}[\tilde{\boldsymbol{r}} : \tilde{\boldsymbol{s}}]$ in S. Let

$$\tilde{\boldsymbol{r}} = h(\boldsymbol{r}) \tag{63}$$

be the embedding function. We then can write the canonical divergence induced to S, given by

$$D[\boldsymbol{r} : \boldsymbol{s}] = \tilde{D}[h(\boldsymbol{r}) : h(\boldsymbol{s})]. \tag{64}$$

Theorem: The α-divergence is the canonical divergence of $\mathcal{P}$ embedded in $\mathcal{M}$.

VI. Concluding Remarks

The class of f-divergences has information monotonicity, and induces a geometrical structure with the Fisher metric and $\pm\alpha$-connections. However, it is not necessarily dually flat. The class of Bregman divergences always gives a dually flat geometrical structure but does not generally have information monotonicity.

We proved that the intersection of the f-divergence and Bregman divergence classes in $\mathcal{M}$ consists of the α-divergences. The KL-divergence and its dual are the unique members belonging to the both classes in $\mathcal{P}$.

We have so far considered only decomposable divergences. For a monotonically increasing function k with $k(0)=0$, we can obtain a new divergence function

$$\tilde{D}[\boldsymbol{p}:\boldsymbol{q}] = k(D_f[\boldsymbol{p}:\boldsymbol{q}]) \tag{65}$$

from an f-divergence D_f. This function is again a divergence with information monotonicity, but is not decomposable.

The following is an interesting conjecture.

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

Appendix
Proof of Information Monotonicity of f-Divergence

We first show that, when f is convex, the f-divergence has information monotonicity. We summarize x_1 and x_2 into a new group G, keeping all the other x_i as singletons. Then, it suffices to prove that

$$p_1 f\left(\frac{q_1}{p_1}\right) + p_2 f\left(\frac{q_2}{p_2}\right) \geq (p_1+p_2) f\left(\frac{q_1+q_2}{p_1+p_2}\right). \tag{66}$$

All the other cases are proved in a similar way. We put

$$\lambda_i = \frac{q_i}{p_i}, \qquad i=1,2. \tag{67}$$

Then, the right-hand side of (66) satisfies

$$(p_1+p_2) f\left(\frac{p_1}{p_1+p_2}\lambda_1 + \frac{p_2}{p_1+p_2}\lambda_2\right) \leq p_1 f(\lambda_1) + p_2 f(\lambda_2) \tag{68}$$

from the Jensen inequality. This proves information monotonicity.

Conversely, we assume that information monotonicity holds for

$$D[\boldsymbol{p}:\boldsymbol{q}] = \sum_i D[p_i:q_i]. \tag{69}$$

For a group consisting of two elements x_1 and x_2, this implies

$$D(p_1:q_1) + D(p_2:q_2) = D(p_1+p_2:q_1+q_2) \tag{70}$$

when

$$q_1 = \lambda p_1 \quad q_2 = \lambda p_2. \tag{71}$$

By putting

$$k(p,\lambda) = D(p,\lambda p) \tag{72}$$

we have

$$k(p_1,\lambda) + k(p_2,\lambda) = k(p_1+p_2,\lambda) \tag{73}$$

for any $\lambda > 0$. This holds for

$$k(p,\lambda) = f(\lambda)p. \tag{74}$$

Hence, D is written in the form of

$$D(p,q) = pf\left(\frac{q}{p}\right) \tag{75}$$

proving the theorem.

References

[1] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, pp. 2032–2066, 1991.

[2] S. Amari and H. Nagaoka, *Methods of Information Geometry*. New York: Oxford Unive. Press, 2000.

[3] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Magyar Tud. Akad. Mat. Kutat Int. Kzl*, vol. 8, pp. 85–108, 1963.

[4] I. Csiszár, "Information measures: A critical survey," in *Proc. 7th Conf. Inf. Theory*, Prague, Czech Republic, 1974, pp. 83–86.

[5] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, pp. 261–273, 2008.

[6] L. Bregman, "The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming," *Comput. Math. Phys. USSR*, vol. 7, pp. 200–217, 1967.

[7] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[8] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-boost and Bregman divergence," *Neural Comput.*, vol. 26, pp. 1651–1686, 2004.

[9] A. Cichocki, R. Adunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations*. New York: Wiley, 2009.

[10] F. Nielsen, Ed., *Emerging Trends in Visual Computing*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 2009, vol. 6.

[11] M. S. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. R. Statist. Soc. B*, no. 28, pp. 131–142, 1966.

[12] I. Taneja and P. Kumar, "Relative information of type s, Csiszár's f-divergence, and information inequalities," *Inf. Sci.*, vol. 166, pp. 105–125, 2004.

[13] J. Havrda and F. Charvát, "Quantification method of classification process. Concept of structural α-entropy," *Kybernetika*, vol. 3, pp. 30–35, 1967.

[14] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493–507, 1952.

[15] A. Rényi, "On measures of entropy and information," in *Proc. 4th Symp. Math. Statist. Probl.*, Berkeley, CA, 1961, vol. 1, pp. 547–561.

[16] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J. Statist. Phys.*, vol. 52, pp. 479–487, 1988.

[17] N. N. Chentsov, *Statistical Decision Rules and Optimal Inference*. Providence, RI: American Mathematical Society, 1982.

[18] D. Petz, "Monotone metrics on matrix spaces," *Linear Algebra Appl.*, vol. 244, pp. 235–249, 1996.

[19] S. Amari, “Information geometry and its applications: Convex function and dually flat manifold,” in *Emerging Trends in Visual Computing*, ser. Lecture Notes in Computer Science, F. Nielsen, Ed. Berlin, Germany: Springer-Verlag, 2009, vol. 5416, pp. 75–102.

[20] J. Zhang, “Divergence function, duality, and convex analysis,” *Neural Comput.*, vol. 16, pp. 159–195, 2004.

[21] S. Eguchi and J. Copas, “A class of logistic type discriminant function,” *Biometrika*, vol. 89, pp. 1–22, 2002.

[22] M. Minami and S. Eguchi, “Robust blind source separation by beta-divergence,” *Neural Comput.*, vol. 14, pp. 1859–1886, 2004.

[23] S. Amari, “Integration of stochastic models by minimizing α-divergence,” *Neural Comput.*, vol. 19, pp. 2780–2796, 2007.

[24] Y. Matsuyama, “The α-EM algorithm: Surrogate likelihood maximization using α-logarithmic information measures,” *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 672–706, Mar. 2002.

[25] S. Amari and A. Cichocki, “Information Geometry of Divergence Functions,” *Bull. Polish Acad. Sci.*, to be published.

[26] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari*, Non-Negative Matrix and Tensor Factorizations: Applications to Explanatory Multi-Way Data Analysis and Blind Source Separation*. New York: Wiley, 2009.

[27] F. Nielsen and R. Noch, “Sided and symmetrized Bregman divergence,” *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2882–2904, Jun. 2009.

Shun-Ichi Amari (M’71–SM’92–F’94–LF’06) was born in Tokyo, Japan, on January 3, 1936. He graduated from the Graduate School, University of Tokyo, Tokyo, Japan, in 1963, majoring in mathematical engineering, and received the Dr. Eng. degree.

He worked as an Associate Professor at Kyushu University and the University of Tokyo, and then a Full Professor at the University of Tokyo, where he is now Professor-Emeritus. He served as Director of RIKEN Brain Science Institute for five years, and is now its senior advisor. He has been engaged in research in wide areas of mathematical engineering, in particular, mathematical foundations of neural networks, including statistical neurodynamics, dynamical theory of neural fields, associative memory, self-organization, and general learning theory. Another main subject of his research is information geometry initiated by himself, which provides a new powerful method to information sciences and neural networks.

Dr. Amari served as President of the Institute of Electronics, Information and Communication Engineers, Japan, and President of the International Neural Networks Society. He received Emanuel A. Piore Award and Neural Networks Pioneer Award from IEEE, the Japan Academy Award, Caianiello Award, C&C award, among many others.