BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 58, No. 1, 2010

Information geometry of divergence functions

S. AMARI* and A. CICHOCKI

RIKEN Brain Science Institute, Wako-shi, Hirosawa 2-1, Saitama 351-0198, Japan

Abstract. Measures of divergence between two points play a key role in many engineering problems. One such measure is a distance function, but there are many important measures which do not satisfy the properties of the distance. The Bregman divergence, Kullback-Leibler divergence and f-divergence are such measures. In the present article, we study the differential-geometrical structure of a manifold induced by a divergence function. It consists of a Riemannian metric, and a pair of dually coupled affine connections, which are studied in information geometry. The class of Bregman divergences are characterized by a dually flat structure, which is originated from the Legendre duality. A dually flat space admits a generalized Pythagorean theorem. The class of f-divergences, defined on a manifold of probability distributions, is characterized by information monotonicity, and the Kullback-Leibler divergence belongs to the intersection of both classes. The f-divergence always gives the α-geometry, which consists of the Fisher information metric and a dual pair of $\pm\alpha$-connections. The α-divergence is a special class of f-divergences. This is unique, sitting at the intersection of the f-divergence and Bregman divergence classes in a manifold of positive measures. The geometry derived from the Tsallis q-entropy and related divergences are also addressed.

Key words: information geometry, divergence functions.

1. Introduction

Given two points P and Q in a space S, we may define a divergence $D[P : Q]$ which measures their discrepancy. The standard distance is indeed such a measure. However, there are many other measures frequently used in many areas of applications (see, e.g., [1, 2]). In particular, for two probability distributions $p(x)$ and $q(x)$, one can define various measures $D[p(x) : q(x)]$ such as the Kullback-Leibler divergence and the Hellinger distance. A divergence is not necessarily symmetric, that is, the relation $D[P : Q] = D[Q : P]$ does not generally hold, nor does it satisfy the triangular inequality. It usually has the dimension of squared distance, and a Pythagorean-like relation holds in some cases.

The present paper aims at elucidating the differential-geometrical structure of a manifold equipped with a divergence function. We study the geometry induced by a divergence function, and demonstrate that it endows a Riemannian metric and a pair of dually coupled affine connections [1]. One of the original contribution of the present paper is to give a unified geometrical framework to various divergence functions, which have a lot of applications [2, 3]. We use modern differential geometry, but do not follow the rigorous mathematical formalism, and instead try to give intuitive explanations to be understandable to those who are not familiar with modern differential geometry.

We begin with two typical classes of divergences: One is the class of Bregman divergences [4], introduced through a convex function (see, e.g., [5]). The other is the class of invariant divergences, called f-divergences, where f is a convex function [6–9]. Both are frequently used in engineering applications. Csiszár [10] gives an axiomatic characterization of these divergences. The geometrical structures of these classes are explained intuitively. Most properties are already known, but here we give their new geometrical explanations.

Bregman divergences are derived from convex functions. The Bregman divergence induces a dual structure through the Legendré transformation. It gives a geometrical structure consisting of a Riemannian metric and dually flat affine connections, called the dually flat Riemannian structure [1]. A dually flat Riemannian manifold is a generalization of the Euclidean space, in which the generalized Pythagorean theorem and projection theorem hold. These two theorems provide powerful tools for solving problems in optimization, statistical inference and signal processing. We show that the Bregman type divergence is automatically induced from the dual flatness of a Riemannian manifold.

Then we study the class of invariant divergences [1, 11]. The invariance requirement comes from information monotonicity, which states that a divergence measure does not increase by coarse graining of information [12]. This leads to the class of f-divergences. The α-divergences are typical examples belonging to this class, which also includes the Kullback-Leibler divergence as a special case. This class of divergences induces an invariant Riemannian metric given by the Fisher information matrix and a pair of invariant dual affine connections, the $\pm\alpha$-connections, which are not necessarily flat. See [13] for more delicate problems occurring in the function space.

When a family of unnormalized probability distributions, that is, a family of positive measures or arrays, is considered, we show that the α-divergence is the only class that is both invariant and dually flat at the same time [14].

We further study the geometry derived from a general divergence in detail. This part is the original contribution of the present paper. A divergence endows a geometrical struc-

*e-mail: amari@brain.riken.jp

ture to the underlying space. We discuss the inverse problem of constructing a divergence function from the geometrical structure of a manifold. The study also comprises various examples of divergence functions, and provides an insight into the geometry derived from the Tsallis q-entropy [15, 17–19]. This extends our views to study how the geometry changes by modifying a divergence function.

2. Convex function, Bregman divergence and dual geometry

2.1. Bregman divergence and Riemannian metric. Let $k(\boldsymbol{z})$ be a strictly convex differentiable function defined in a space S with a local coordinate system $\boldsymbol{z}$. Then, for two points $\boldsymbol{z}$ and $\boldsymbol{y}$ in S, we can define the following function

$$D[\boldsymbol{z} : \boldsymbol{y}] = k(\boldsymbol{z}) - k(\boldsymbol{y}) - \text{Grad } k(\boldsymbol{y}) \cdot (\boldsymbol{z} - \boldsymbol{y}), \quad (1)$$

where, Grad k is the gradient vector

$$\text{Grad } k(\boldsymbol{z}) = (\partial k(\boldsymbol{z})/\partial z_i), \quad (2)$$

and the operator '$\cdot$' denotes the inner product,

$$\text{Grad } k(\boldsymbol{y}) \cdot (\boldsymbol{z} - \boldsymbol{y}) = \sum_i \frac{\partial k}{\partial y_i}(z_i - y_i). \quad (3)$$

The function $D[\boldsymbol{z} : \boldsymbol{y}]$ satisfies the following condition for divergences:

1) $D[\boldsymbol{z} : \boldsymbol{y}] \geq 0$,
2) $D[\boldsymbol{z} : \boldsymbol{y}] = 0$, when and only when $\boldsymbol{z} = \boldsymbol{y}$,
3) For small $d\boldsymbol{z}$, Taylor expansion

$$D[\boldsymbol{z} + d\boldsymbol{z} : \boldsymbol{z}] \approx \frac{1}{2} \sum g_{ij} dz_i dz_j \quad (4)$$

gives a positive-definite quadratic form.

We call $D[\boldsymbol{z} : \boldsymbol{y}]$ the Bregman divergence between two points $\boldsymbol{z}$ and $\boldsymbol{y}$. In general, the divergence is not symmetric with respect to $\boldsymbol{z}$ and $\boldsymbol{y}$ so that

$$D[\boldsymbol{y} : \boldsymbol{z}] \neq D[\boldsymbol{z} : \boldsymbol{y}]. \quad (5)$$

When $\boldsymbol{z}$ is infinitesimally close to $\boldsymbol{y}, \boldsymbol{y} = \boldsymbol{z} + d\boldsymbol{z}$, we have

$$D[\boldsymbol{z} : \boldsymbol{z} + d\boldsymbol{z}] = \frac{1}{2} \sum \frac{\partial^2 k(\boldsymbol{z})}{\partial z_i \partial z_j} dz_i dz_j = D[\boldsymbol{z} + d\boldsymbol{z} : \boldsymbol{z}] \quad (6)$$

by Taylor expansion. This is regarded as a half of the squared distance between $\boldsymbol{z}$ and $\boldsymbol{z} + d\boldsymbol{z}$, defined in the following.

We can use the Hessian of k,

$$g_{ij}(\boldsymbol{z}) = \frac{\partial^2}{\partial z_i \partial z_j} k(\boldsymbol{z}), \quad (7)$$

to define the squared local distance as

$$ds^2 = \sum g_{ij}(\boldsymbol{z}) dz_i dz_j. \quad (8)$$

A space S is called a Riemannian manifold, when a positive-definite matrix $g(\boldsymbol{z}) = (g_{ij}(\boldsymbol{z}))$ is defined at each point $\boldsymbol{z} \in S$ such that (8) is the squared local distance.

When S is a Riemannian manifold, we can define a Riemannian geodesic $\boldsymbol{z}(t)$ parameterized by t. A length of a curve $\boldsymbol{z}(t)$ connecting two points $\boldsymbol{z}_1 = \boldsymbol{z}(t_1)$ and $\boldsymbol{z}_2 = \boldsymbol{z}(t_2)$ is defined by the integral

$$s = \int_{t_1}^{t_2} \sum \sqrt{g_{ij}\left(\boldsymbol{z}(t)\right) \dot{z}_i(t) \dot{z}_j} dt, \quad (9)$$

where $\dot{z}_i(t) = (d/dt) z_i(t)$. This is the Riemannian distance along the curve between the two points. A Riemannian geodesic $\boldsymbol{z}(t)$ is the curve that minimizes the above distance.

We next introduce a dual structure in S, defined by the Legendre transformation (see, e.g., [1, 20]. The gradient vector

$$\boldsymbol{z}^* = \text{Grad } k(\boldsymbol{z}) \quad (10)$$

of a convex function $k(\boldsymbol{z})$ is in one-to-one correspondence with $\boldsymbol{z}$. This is the Legendre transformation, and $\boldsymbol{z}^*$ can be regarded as another coordinate system of S different from $\boldsymbol{z}$. We can calculate the dual function of k, defined by

$$k^*(\boldsymbol{z}^*) = \max_{\boldsymbol{z}} \left\{ \boldsymbol{z} \cdot \boldsymbol{z}^* - k(\boldsymbol{z}) \right\}, \quad (11)$$

which is a convex function of $\boldsymbol{z}^*$. Hence we can describe the geometry of S by using the dual convex function k^* and the dual coordinates $\boldsymbol{z}^*$. Obviously, $\boldsymbol{z}$ and $\boldsymbol{z}^*$ are dual, since we have

$$\boldsymbol{z} = \text{Grad } k^*(\boldsymbol{z}^*). \quad (12)$$

The Riemannian metric is given in the dual coordinates by

$$g^*_{ij}(\boldsymbol{z}^*) = \frac{\partial^2}{\partial z^*_i \partial z^*_j} k^*(\boldsymbol{z}^*). \quad (13)$$

Theorem 1. The Riemannian metrics g_{ij} and g^*_{ij} in their matrix form are mutually inverse. They are the same tensor represented in different coordinate systems $\boldsymbol{z}$ and $\boldsymbol{z}^*$, giving the same local distance.

Proof. From $\boldsymbol{z}^* = \text{Grad } k(\boldsymbol{z})$, we have

$$dz^*_i = \sum \frac{\partial^2 k(\boldsymbol{z})}{\partial z_i \partial z_j} dz_j = \sum g_{ij} dz_j. \quad (14)$$

In a similar way, we have

$$dz_i = \sum g^*_{ij} dz^*_j. \quad (15)$$

By using the vector-matrix notation, they are rewritten as

$$d\boldsymbol{z}^* = g d\boldsymbol{z}, \qquad d\boldsymbol{z} = g^* d\boldsymbol{z}^*, \quad (16)$$

where $g = (g_{ij})$ and $g^* = (g^*_{ij})$. This shows

$$g^* = g^{-1}. \quad (17)$$

The local distances of $d\boldsymbol{z}$ and $d\boldsymbol{z}^*$ are given by the quadratic form

$$ds^2 = d\boldsymbol{z}^T g d\boldsymbol{z}, \quad (18)$$

$$ds^{*2} = d\boldsymbol{z}^{*T} g^* d\boldsymbol{z}^*, \quad (19)$$

where $d\boldsymbol{z}^T$ is the transpose of $d\boldsymbol{z}$. Hence, from (16) and (17), we have

$$ds^2 = ds^{*2}. \quad (20)$$

In other words, g and g^* are the same Riemannian metric represented in two different coordinate systems.

The dual function $k^*(\boldsymbol{z}^*)$ also induces a divergence,

$$D^*[\boldsymbol{y}^* : \boldsymbol{z}^*] = k^*(\boldsymbol{y}^*) - k^*(\boldsymbol{z}^*) - \text{grad}k^*(\boldsymbol{z}^*) \cdot (\boldsymbol{z}^* - \boldsymbol{y}^*). \tag{21}$$

Theorem 2.

The two divergences D and D^* are mutually reciprocal, in the sense of

$$D^*[\boldsymbol{y}^* : \boldsymbol{z}^*] = D[\boldsymbol{z} : \boldsymbol{y}]. \tag{22}$$

The divergence between two points $\boldsymbol{z}$ and $\boldsymbol{y}$ is written in the dual form

$$D[\boldsymbol{z} : \boldsymbol{y}] = k(\boldsymbol{z}) + k^*(\boldsymbol{y}^*) - \boldsymbol{z} \cdot \boldsymbol{y}^*. \tag{23}$$

Proof. The right-hand side of (11) is maximized when $\boldsymbol{z}$ and $\boldsymbol{z}^*$ correspond to each other, that is,

$$\frac{\partial}{\partial \boldsymbol{z}} \{\boldsymbol{z} \cdot \boldsymbol{z}^* - k(\boldsymbol{z})\} = \boldsymbol{z}^* - \text{Grad } k(\boldsymbol{z}) = 0. \tag{24}$$

Hence, we have the identity from (11),

$$k(\boldsymbol{z}) + k^*(\boldsymbol{z}^*) - \boldsymbol{z} \cdot \boldsymbol{z}^* = 0. \tag{25}$$

By using this relation for $\boldsymbol{y}$, we have

$$\begin{aligned} D[\boldsymbol{z} : \boldsymbol{y}] &= k(\boldsymbol{z}) - k(\boldsymbol{y}) - \boldsymbol{y}^* \cdot (\boldsymbol{z} - \boldsymbol{y}) && (26)\\ &= k(\boldsymbol{z}) + k^*(\boldsymbol{y}^*) - \boldsymbol{z} \cdot \boldsymbol{y}^*. && (27) \end{aligned}$$

Analogously, we have for $k(\boldsymbol{z})$,

$$D[\boldsymbol{z} : \boldsymbol{y}] = D^*[\boldsymbol{y}^* : \boldsymbol{z}^*]. \tag{28}$$

The theorem shows that it suffices to consider only one divergence function.

2.2. Dual affine structure and Pythagorean theorem. We now introduce an affine structure in S [20], which is different from the Riemannian structure defined by g. We simply assume that the coordinate system $\boldsymbol{z}$ is affine. Hence, a curve represented in the form

$$\boldsymbol{z}(t) = t\boldsymbol{a} + \boldsymbol{b} \tag{29}$$

is a geodesic in this sense, where t is the parameter along the curve and $\boldsymbol{a}$ and $\boldsymbol{b}$ are constant vectors. This geodesic is not a Riemannian geodesic that minimizes the length of a curve.

Further we define a dual affine structure. A dual geodesic $\boldsymbol{z}^*(t)$ is defined by a linear curve

$$\boldsymbol{z}^*(t) = t\boldsymbol{a}^* + \boldsymbol{b}^*, \tag{30}$$

for constant vectors $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$, regarding $\boldsymbol{z}^*$ as another affine coordinate system of S. This defines a dual affine structure of S, which is different from the primal affine structure. However, we will later show their differential-geometrical formalism and prove that the two affine structures are dually coupled with respect to the Riemannian metric.

We have shown that space S equipped with a Bregman divergence is Riemannian, but has two dually flat affine structures. This gives rise to the following generalized Pythagorean theorem [1, 20]. As a preliminary, we mention the orthogonality of two curves in S. This is defined by using the Riemannian metric. Let $\boldsymbol{z}_1(t)$ and $\boldsymbol{z}_2(t)$ be two curves intersecting at $t = 0$, $\boldsymbol{z}_1(0) = \boldsymbol{z}_2(0)$, where t is a parameter along the curves. Then, the two curves intersect orthogonally at $t = 0$, when their tangent vectors

$$\boldsymbol{t}_1 = \frac{d}{dt}\boldsymbol{z}_1(0), \qquad \boldsymbol{t}_2 = \frac{d}{dt}\boldsymbol{z}_2(0) \tag{31}$$

satisfy the Riemannian orthogonality condition,

$$\langle \boldsymbol{t}_1, \boldsymbol{t}_2 \rangle = \sum g_{ij} t_{1i} t_{2j} = 0, \tag{32}$$

where $\langle \boldsymbol{t}_1, \boldsymbol{t}_2 \rangle$ is the inner product with respect to the Riemannian metric g. When the dual coordinate system is used for one curve, say, $\boldsymbol{z}_2^*(t)$, the orthogonality condition is simplified to

$$\sum \frac{d}{dt} z_1^i(0) \frac{d}{dt} z_2^{*i}(0) = 0, \tag{33}$$

because

$$\langle \dot{\boldsymbol{z}}_1, \dot{\boldsymbol{z}}_2 \rangle = \dot{\boldsymbol{z}}_1^T g \dot{\boldsymbol{z}}_2 = \dot{\boldsymbol{z}}_1^T \dot{\boldsymbol{z}}_2^*. \tag{34}$$

Pythagorean Theorem.

Let P, Q, R be three points in S whose coordinates (and dual coordinates) are represented by $\boldsymbol{z}_P$, $\boldsymbol{z}_Q$, $\boldsymbol{z}_R$ ($\boldsymbol{z}_P^*, \boldsymbol{z}_Q^*, \boldsymbol{z}_R^*$), respectively. When the dual geodesic connecting P and Q is orthogonal at Q to the geodesic connecting Q and R, then

$$D[P : R] = D[P : Q] + D[Q : R]. \tag{35}$$

Dually, when the geodesic connecting P and Q is orthogonal at Q to the dual geodesic connecting Q and R, we have

$$D[R : P] = D[Q : P] + D[R : Q]. \tag{36}$$

Proof. By using (23), we have

$$\begin{aligned} D[R : Q] + D[Q : P] &= k(\boldsymbol{z}_R) + k^*(\boldsymbol{z}_Q^*) + k(\boldsymbol{z}_Q) + k^*(\boldsymbol{z}_P^*) \\ &\quad - \boldsymbol{z}_R \cdot \boldsymbol{z}_Q^* - \boldsymbol{z}_Q \cdot \boldsymbol{z}_P^* && (37)\\ &= k(\boldsymbol{z}_R) + k^*(\boldsymbol{z}_P^*) + \boldsymbol{z}_Q \cdot \boldsymbol{z}_Q^* \\ &\quad - \boldsymbol{z}_R \cdot \boldsymbol{z}_Q^* - \boldsymbol{z}_Q \cdot \boldsymbol{z}_P^* && (38)\\ &= D[\boldsymbol{z}_R : \boldsymbol{z}_P^*] + (\boldsymbol{z}_Q - \boldsymbol{z}_R) \\ &\quad \cdot (\boldsymbol{z}_Q^* - \boldsymbol{z}_P^*). && (39) \end{aligned}$$

The tangent vector of the geodesic connecting Q and R is $\boldsymbol{z}_Q - \boldsymbol{z}_R$, and the tangent vector of the dual geodesic connecting Q and P is $\boldsymbol{z}_Q^* - \boldsymbol{z}_P^*$ in the dual coordinate system. Hence, the second term of the right-hand side of the above equation vanishes, because the primal and dual geodesics connecting Q and R, and Q and P are orthogonal.

The following projection theorem is a consequence of the generalized Pythagorean theorem. Let M be a smooth submanifold of S. Given a point P outside M, we connect it to a point Q in M by geodesic (dual geodesic). When the geodesic (dual geodesic) connecting P and Q is orthogonal to M (that is, orthogonal to any tangent vectors of M), Q is said to be the geodesic projection (dual geodesic projection) of P to M.

Projection Theorem.

Given P and M, the point $Q(Q^*)$ that minimizes divergence $D(P : R)$, $R \in Q$ ($D(R : P), R \in M$) is the projection (dual projection) of P to Q.

This theorem is useful, when we search for the point belonging to M that minimizes the divergence $D(P:Q)$ or $D(Q:P)$ for preassigned P. In many engineering problems, P is given from observed data, and M is a model to describe the underlying structure.

2.3. Examples of dually flat geometry. The following examples illustrate our approach:

1) Euclidean geometry:

When $k(\boldsymbol{z})$ has a quadratic form,

$$k(\boldsymbol{z}) = \frac{1}{2}\sum z_i^2, \tag{40}$$

the induced Riemannian metric g is the identity matrix. Hence the space is Euclidean. The primal and dual coordinates are the same, $\boldsymbol{z}^* = \boldsymbol{z}$, so that the space is self-dual. The divergence is then a half of the square of Euclidean distance,

$$D[\boldsymbol{z}:\boldsymbol{y}] = \frac{1}{2}\sum |z_i - y_i|^2 . \tag{41}$$

The Pythagorean theorem and the projection theorem are exactly the same as the well-known counterparts in a Euclidean space.

2) Entropy geometry on discrete probability distributions:

Consider the set S_n of all discrete probability distributions over $n+1$ elements $X = \{x_0, x_1, \cdots, x_n\}$. A probability distribution is given by

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x), \tag{42}$$

where

$$p_i = \text{Prob}\{x = x_i\}, \qquad \delta_i(x) = \begin{cases} 1, & x = x_i, \\ 0, & \text{otherwise.} \end{cases} \tag{43}$$

Obviously,

$$\sum_{i=0}^{n} p_i = 1. \tag{44}$$

We can use a coordinate system $\boldsymbol{z} = (p_1, \cdots, p_n)$ for the set S_n of all such distributions, where $z_0 = p_0$ is regarded as a function of the other coordinates,

$$p_0 = 1 - \sum_{i=1}^{n} z_i. \tag{45}$$

The Shannon entropy,

$$H(\boldsymbol{z}) = -\sum z_i \log z_i - \Big(1 - \sum z_i\Big)\log\Big(1 - \sum z_i\Big), \tag{46}$$

is concave, so that $k(\boldsymbol{z}) = -H(\boldsymbol{z})$ is a convex function of $\boldsymbol{z}$.

The Riemannian metric induced from $k(\boldsymbol{z})$ is calculated as

$$g_{ij}(\boldsymbol{z}) = \frac{1}{p_i}\delta_{ij} + \frac{1}{p_0}, \tag{47}$$

which is the Fisher information matrix. The divergence function is given by

$$D[\boldsymbol{z}:\boldsymbol{y}] = \sum_{i=0}^{n} z_i \log \frac{z_i}{y_i}. \tag{48}$$

which is known as the Kullback-Leibler divergence. It is written in general as

$$KL[p(x):q(x)] = \sum_x p(x)\log\frac{p(x)}{q(x)}. \tag{49}$$

The dual coordinates are given by

$$z_i^* = \log\frac{p_i}{p_0}. \tag{50}$$

and are are known as the natural parameters of an exponential family, where the probability distribution is rewritten in the form

$$p(x, \boldsymbol{z}^*) = \exp\Big\{\sum z_i^* \delta_i(x) - k^*(\boldsymbol{z}^*)\Big\}. \tag{51}$$

An exponential family of probability distributions is usually represented as

$$p(x, \boldsymbol{\theta}) = \exp\Big\{\sum \theta_i \delta_i(x) - \psi(\boldsymbol{\theta})\Big\}. \tag{52}$$

In our case, $\theta_i = z_i^*$, and $\psi(\boldsymbol{\theta}) = k^*(\boldsymbol{z}^*)$ is the cumulant generating function. This is the dual potential,

$$\psi(\boldsymbol{\theta}) = k^*(\boldsymbol{z}^*) = -\log p_0. \tag{53}$$

The induced geometrical structure is studied in detail in information geometry [1], where the Pythagorean theorem and Projection theorem hold.

It should be mentioned that the usual definition begins with the exponential form (52), where ψ is the convex function to define the geometry and $\boldsymbol{\theta} = \boldsymbol{z}^*$ is the primal affine coordinates. Therefore, $\boldsymbol{z} = \boldsymbol{p}$ is the dual affine coordinates. Hence, the definition of the primal and dual coordinates are reversed.

3) Positive-definite matrices:

Let $\mathcal{P}$ be the set of $n \times n$ positive-definite matrices. It is an $n(n+1)/2$-dimensional manifold, since $P \in \mathcal{P}$ is a symmetric matrix. When $|P|$ is the determinant of P,

$$k(P) = -\log|P| \tag{54}$$

is a convex function of P. Its gradient is

$$\text{Grad}\,\psi(P) = -P^{-1}. \tag{55}$$

Hence, the induced divergence is

$$D[P:Q] = -\log\left|PQ^{-1}\right| + \text{tr}(PQ^{-1}) - n, \tag{56}$$

where the operator tr is the trace of a matrix. The dual affine coordinates are

$$P^* = -P^{-1}, \tag{57}$$

and the dual potential is

$$k^*(P^*) = -\log|P^*|. \tag{58}$$

Consider a family of probability distributions of $\boldsymbol{x}$,

$$P(\boldsymbol{x}, P) = \exp\left\{-\frac{1}{2}\boldsymbol{x}^T P^{-1}\boldsymbol{x} - \psi(P)\right\}. \tag{59}$$

This is a multi-variate Gaussian distribution with mean 0 and covariance matrix P. The geometrical structure introduced here is the same as that derived from the exponential family of distributions [1].

Quantum information geometry [1, 21–23] uses the convex function

$$k(P) = \text{tr}(P \log P - P). \tag{60}$$

Its gradient is

$$P^* = \text{Grad}\, k(P) = \log P \tag{61}$$

and hence the dual affine coordinates are $\log P$. The divergence is

$$D(P:Q) = \text{tr}\left\{P(\log P - \log Q) + P + Q\right\}, \tag{62}$$

which is the von Neumann divergence.

We further define the following function by using a convex function f,

$$k_f(P) = \text{tr} f(P) = \sum f(\lambda_i), \tag{63}$$

where λ_i are the eigenvalues of P. Then, k_f is a convex function of P, from which we derive a dual geometrical structure depending on f [24].

For $f(\lambda) = (1/2)\lambda^2$, we have

$$k_f(P) = \frac{1}{2}\,\text{tr}(P^T P), \tag{64}$$

and the derived divergence is

$$D_f[P:Q] = \frac{1}{2}\|P-Q\|^2 = \frac{1}{2}\sum |p_{ij} - q_{ij}|^2. \tag{65}$$

The dual is $P^* = P$, so that the geometry is self-dual and Euclidean.

The statistical case of (54) is derived from

$$f(\lambda) = -\log(\lambda), \tag{66}$$

while

$$f(\lambda) = \lambda \log \lambda - \lambda \tag{67}$$

gives the quantum-information structure (60).

More generally, by putting [22]

$$f_\alpha(\lambda) = \frac{-4}{1-\alpha^2}(\lambda^{\frac{1+\alpha}{2}} - \lambda), \qquad (-1 < \alpha < 1) \tag{68}$$

we have the α-divergence,

$$D_\alpha[P:Q] = $$

$$= \frac{4}{1-\alpha^2}\,\text{tr}\left[\frac{1-\alpha}{2}P + \frac{1+\alpha}{2}Q - P^{\frac{1+\alpha}{2}}Q^{\frac{1-\alpha}{2}}\right]. \tag{69}$$

This is a generalization of the α-divergence defined in the space of positive measures defined later.

4) Linear programming:

Let us consider the following LP (linear programming) problem.

Problem. Minimize the cost function

$$c(\boldsymbol{x}) = \sum c_i x_i \tag{70}$$

under the constraints on $\boldsymbol{x} \in \boldsymbol{R}^n$ given by

$$\sum_{j=1}^{n} A_{ij} x_j \geq b_i, \quad i = 1, \cdots, m. \tag{71}$$

Let $S = \{\boldsymbol{x}\}$ be the open region satisfying the constraints

$$\sum A_{ij} x_j > b_i. \tag{72}$$

Then,

$$k(\boldsymbol{x}) = -\sum_{i=1}^{m} \log \left\{ \sum_{j=1}^{n} A_{ij} x_j - b_i \right\} \tag{73}$$

is a convex function. Hence, we can introduce a dually flat Riemannian structure to S.

The inner method of LP makes use of this structure [25]. The primal and dual curvatures play an important role in evaluating the efficiency of algorithms [26]. This can be generalized to the cone programming and semi-definite programming problems [27].

5) Other divergences:

There are many other divergences in S_n derived from a convex function $f(z)$ in the form of $k(\boldsymbol{z}) = \sum f(z_i)$. We show some of them. An extensive list of divergences is given in [2].

i) For $f(z) = -\log z$, we have the Itakua-Saito divergence

$$D_{IS}[\boldsymbol{p}:\boldsymbol{q}] = \sum (\log \frac{q_i}{p_i} + \frac{p_i}{q_i} - 1), \tag{74}$$

which is useful for spectral analysis of speech signals.

ii) For

$$f(z) = z \log z + (1-z)\log(1-z), \tag{75}$$

we have the Fermi-Dirac divergence

$$D_{FD}[\boldsymbol{p}:\boldsymbol{q}] = \sum \left\{ p_i \log \frac{p_i}{q_i} + (1-p_i)\log \frac{1-p_i}{1-q_i} \right\} \tag{76}$$

iii) The β-divergence is a family divergences derived from the following β-functions,

$$f_\beta(z) = \begin{cases} \dfrac{1}{\beta(\beta+1)}\left\{z^{\beta+1} - (\beta+1)z + \beta\right\} & \\ z\log z - z, & \beta = 0 \\ z - \log z, & \beta = -1. \end{cases} \tag{77}$$

The divergences are

$$D_\beta[\boldsymbol{p}:\boldsymbol{q}] = \begin{cases} \dfrac{1}{\beta(\beta+1)} \sum \Big\{ p_i^{\beta+1} - q_i^{\beta+1} & \\ \qquad -(\beta+1) q_i^\beta (p_i - q_i) \Big\}, & \beta > -1 \\ \sum p_i \log \dfrac{p_i}{q_i}, & \beta = 0 \\ \sum \left(\log \dfrac{q_i}{p_i} + \dfrac{p_i}{q_i} - 1 \right), & \beta = -1. \end{cases} \tag{78}$$

Hence, the family includes the KL-divergence and Itakura-Saito divergence. It is used in machine learning [28] and robust estimation [29–31].

3. Invariant divergence in manifolds of probability and positive measures

3.1. Information monotonicity. Let us consider again the space S_n of all probability distributions over $n+1$ atoms $X = \{x_0, x_1, \cdots, x_n\}$. The probability distributions are given by $\boldsymbol{p} = (p_0, p_1, \cdots, p_n)$, $p_i = \text{Prob}\{x = x_i\}$, $i = 0, 1, \cdots, n$, $\sum p_i = 1$. We try to define a new divergence measure $D[\boldsymbol{p} : \boldsymbol{q}]$ between two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. To this end, we introduce the concept of information monotonicity [12, 14], of which original idea was proposed by Chentsov in 1972 [11].

Let us divide X into m groups, $G_1, G_2, \cdots, G_m (m < n+1)$, say

$$G_1 = \{x_1, x_2, x_5\}, \qquad G_2 = \{x_3, x_8, \cdots\}, \cdots \tag{79}$$

This is a partition of X,

$$X = \cup G_i, \tag{80}$$

$$G_i \cap G_j = \phi. \tag{81}$$

Assume that we do not know the outcome x_i directly, but can observe which group G_j it belongs to. This is called coarse-graining of X.

The coarse-graining generates a new probability distributions $\overline{\boldsymbol{p}} = (\overline{p}_1, \cdots, \overline{p}_m)$ over $G_1, \cdots, G_m$,

$$\overline{p}_j = \text{Prob}\{G_j\} = \sum_{x_i \in G_j} \text{Prob}\{x_i\}. \tag{82}$$

Let $\overline{D}[\overline{\boldsymbol{p}} : \overline{\boldsymbol{q}}]$ be an induced divergence between $\overline{\boldsymbol{p}}$ and $\overline{\boldsymbol{q}}$. Since coarse-graining summarizes some of elements into one group, detailed information of the outcome in each group is lost. Therefore, it is natural to require

$$\overline{D}[\overline{\boldsymbol{p}} : \overline{\boldsymbol{q}}] \leq D[\boldsymbol{p} : \boldsymbol{q}]. \tag{83}$$

When does the equality hold? For two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, assume that the outcome x_i is known to belong to G_j. Then, we require more information to distinguish the two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ by knowing further detail inside group G_j. Since x_i belongs to group G_j, we consider the conditional probability distributions

$$p(x_i | x_i \in G_j), \qquad q(x_i | x_i \in G_j) \tag{84}$$

inside group G_j. If they are equal, we cannot obtain further information to distinguish $\boldsymbol{p}$ from $\boldsymbol{q}$ by observing elements inside G_j. Hence,

$$\overline{D}[\overline{\boldsymbol{p}} : \overline{\boldsymbol{q}}] = D[\boldsymbol{p} : \boldsymbol{q}] \tag{85}$$

holds, when and only when

$$p(x_i | G_j) = q(x_i | G_j) \tag{86}$$

for all G_j and all $x_i \in G_j$, or

$$\frac{p_i}{q_i} = \lambda_j \tag{87}$$

for all $x_i \in G_j$ for some constant λ_j.

A divergence satisfying the above requirements is called an invariant divergence, and such a property is termed as information monotonicity.

3.2. f-divergence and information monotonicity. The f-divergence was introduced by Csiszár [7] and also by Ali and Silvey [6]. It is defined by

$$D_f[\boldsymbol{p} : \boldsymbol{q}] = \sum p_i f\left(\frac{q_i}{p_i}\right), \tag{88}$$

where f is a convex function satisfying

$$f(1) = 0. \tag{89}$$

For $f = cf$ for a constant c, we have

$$D_{cf}[\boldsymbol{p} : \boldsymbol{q}] = cD[\boldsymbol{p} : \boldsymbol{q}]. \tag{90}$$

Hence, f and cf give the same divergence except for the scale factor. In order to standardize the scale of divergence, we may assume that

$$f''(1) = 1, \tag{91}$$

provided f is differentiable. Further, for $f_c(u) = f(u) - c(u-1)$ where c is constant, we have

$$D_{f_c}[\boldsymbol{p} : \boldsymbol{q}] = D_f[\boldsymbol{p} : \boldsymbol{q}]. \tag{92}$$

Hence, we may use such an f that satisfies

$$f'(1) = 0 \tag{93}$$

without loss of generality. A convex function satisfying the above three conditions (89), (91), (93) is called a standard f function.

The f-divergence (88) is written as a sum of functions of two variables p_i and q_i. Such a divergence is said to be decomposable.

Csiszár found that an f-divergence satisfies information monotonicity. Moreover, the class of f-divergences is unique in the sense that any decomposable divergence satisfying information monotonicity is an f-divergence.

Theorem 3.

The f-divergence satisfies the information monotonicity. Conversely, any decomposable information monotonic divergence is written in the form of f-divergence.

The proof is found, e.g., in [14].

The Riemannian metric and affine connections derived from the f-divergence has a common invariant structure. They are given by the Fisher information Riemannian metric and $\pm\alpha$-connections, which are shown in a later section.

3.3. Examples of f-divergence in S_n. An extensive list of f-divergences is given in [2]. Some of them are listed below.

1) Total variation: $f(u) = |u - 1|$.
   The total variation distance is defined by
   $$D[\boldsymbol{p} : \boldsymbol{q}] = \sum |p_i - q_i|. \tag{94}$$
   Note that it is not differentiable, and no Riemannian metric is derived. However, this gives a Minkovskian metric.
2) Squared Hellinger distance: $f(u) = (\sqrt{u} - 1)^2$ for which
   $$D[\boldsymbol{p} : \boldsymbol{q}] = \sum (\sqrt{p_i} - \sqrt{q_i})^2. \tag{95}$$

3) Pearson and Neyman Chi-square divergence: $f(u) = (1/2)(u-1)^2$ and $f(u) = (1/2)(u-1)^2/u$, for which

$$D[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{2}\sum \frac{(p_i - q_i)^2}{p_i}, \tag{96}$$

$$D[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{2}\sum \frac{(p_i - q_i)^2}{q_i}. \tag{97}$$

4) The KL-divergence: $f(u) = u - 1 - \log u$, and

$$D[\boldsymbol{p}:\boldsymbol{q}] = \sum p_i \log \frac{q_i}{p_i} \tag{98}$$

5) The α-divergence:

$$f(u) = \frac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right) - \frac{2}{1-\alpha}(u-1), \tag{99}$$

$$D_\alpha[\boldsymbol{p}:\boldsymbol{q}] = \frac{4}{1-\alpha^2}\sum\left(1 - p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}}\right). \tag{100}$$

The α-divergence was introduced by Havdra and Charvát in 1967 [32], and has been studied extensively by Amari and Nagaoka [1]. Its applications were described earlier by Chernoff in 1952 [33], and later in [34, 35] etc. to mention a few. It is the squared Hellinger distance for $\alpha = 0$, and the KL-divergence and its reciprocal are obtained in the limit of $\alpha \to \pm 1$.

3.4. f-divergence in the space of positive measures. We have studied both invariant and flat geometrical structures in the manifold S_n of probability distributions. Here, we study a similar structure in the space of positive measures M_n over $X = \{x_1, \cdots, x_n\}$, whose points are denoted by $\boldsymbol{z} = (z_1, \cdots, z_n)$, $z_i > 0$. Here z_i is the mass (measure) of x_i, where the total mass $\sum z_i$ is arbitrary. In many applications, $\boldsymbol{z}$ is a non-negative array, and we can extend it to a non-negative double array $\boldsymbol{z} = (z_{ij})$ etc., that is, matrices and tensors. We first derive an f-divergence in M_n: for two positive measures $\boldsymbol{z}$ and $\boldsymbol{y}$, an f-divergence is given by

$$D_f[\boldsymbol{z}:\boldsymbol{y}] = \sum z_i f\left(\frac{y_i}{z_i}\right), \tag{101}$$

where f is a standard convex function. It should be noted that an f-divergence is no more invariant under the transformation from $f(u)$ to

$$f_c(u) = f(u) - c(u-1). \tag{102}$$

Hence, it is absolutely necessary to use a standard f in the case of M_n, because the condition of divergence is violated otherwise.

Among all f-divergences, the α divergence given in the form of

$$D_\alpha[\boldsymbol{z}:\boldsymbol{y}] = \sum z_i f_\alpha\left(\frac{y_i}{z_i}\right), \tag{103}$$

where

$$f_\alpha(u) = \begin{cases} \dfrac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right) + \dfrac{2}{1-\alpha}(u-1), & \alpha \neq \pm 1, \\ u\log u - (u-1), & \alpha = 1, \\ -\log u + (u-1), & \alpha = -1, \end{cases} \tag{104}$$

plays a special role in M_n. This $f_\alpha(u)$ is a standard convex function.

The α-divergence is given in the following form,

$$D_\alpha[\boldsymbol{z}:\boldsymbol{y}] = \begin{cases} \dfrac{4}{1-\alpha^2}\sum\left(\dfrac{1-\alpha}{2}z_i + \dfrac{1+\alpha}{2}y_i - z_i^{\frac{1-\alpha}{2}} y_i^{\frac{1+\alpha}{2}}\right), & \alpha \neq \pm 1, \\ \sum\left(z_i - y_i + y_i \log \dfrac{y_i}{z_i}\right), & \alpha = 1, \\ \sum\left(y_i - z_i + z_i \log \dfrac{z_i}{y_i}\right), & \alpha = -1. \end{cases} \tag{105}$$

3.5. Characterization of α-divergence. We show that the α-divergence is not only invariant, but also has a dually flat structure for any α in M_n. This is different from the case of S_n, because it is not dually flat in S_n, as will be shown later. If M_n is dually flat, it has primal and dual affine coordinate systems $\boldsymbol{w}, \boldsymbol{w}^*$ such that primal and dual geodesics are represented in the linear forms (29), (30) in these coordinate systems. To this end, we introduce

$$r_\alpha(u) = \begin{cases} \dfrac{2}{1-\alpha}\left(u^{\frac{1-\alpha}{2}} - 1\right), & \alpha \neq 1 \\ \log u, & \alpha = 1, \end{cases} \tag{106}$$

which is called the α-representation of u. The new coordinates $\boldsymbol{w}$ and $\boldsymbol{w}^*$ of M_n are defined by

$$w_i = r_\alpha(z_i), \qquad w_i^* = r_{-\alpha}(z_i). \tag{107}$$

We further define convex functions $k(\boldsymbol{w})$ and $k^*(\boldsymbol{w}^*)$ by

$$k(\boldsymbol{w}) = k_\alpha(\boldsymbol{w}) = \sum \frac{2}{1+\alpha}\left\{1 + \frac{1-\alpha}{2}w_i\right\}^{\frac{2}{1-\alpha}}, \tag{108}$$

$$k^*(\boldsymbol{w}^*) = k_{-\alpha}(\boldsymbol{w}^*). \tag{109}$$

Theorem 4. The α-divergence is a Bregman divergence, where $\boldsymbol{w}$ and $\boldsymbol{w}^*$ are affine coordinate systems, having dual convex functions $k_\alpha(\boldsymbol{w})$ and $k_\alpha^*(\boldsymbol{w}^*)$, respectively.

Proof. The divergence D_α between two points $\boldsymbol{w}$ and $\boldsymbol{s}$ (dual coordinates are $\boldsymbol{w}^*$ and $\boldsymbol{s}^*$, respectively) derived from $k_\alpha(\boldsymbol{w})$ is written as

$$D_\alpha[\boldsymbol{w}:\boldsymbol{s}] = k_\alpha(\boldsymbol{w}) + k_{-\alpha}(\boldsymbol{s}^*) - \boldsymbol{w}\cdot\boldsymbol{s}^*. \tag{110}$$

For $\boldsymbol{w} = r_\alpha(\boldsymbol{z})$, $\boldsymbol{s} = r_\alpha(\boldsymbol{y})$, we have

$$k_\alpha(\boldsymbol{w}) = \frac{2}{1+\alpha}\sum z_i, \tag{111}$$

$$k_{-\alpha}(\boldsymbol{s}^*) = \frac{2}{1-\alpha}\sum y_i \tag{112}$$

and

$$\boldsymbol{w}\cdot\boldsymbol{s}^* = \frac{1}{1-\alpha^2} z_i^{\frac{1-\alpha}{2}} y_i^{\frac{1+\alpha}{2}}. \tag{113}$$

By substituting , (111)–(112) in (109), we prove that D_α is the α-divergence given in (105). This demonstrates that M_n has the dually flat α-structure for any α. We can furthermore prove that the α-divergence is unique in the sense that it belongs to both classes of f-divergences and Bregman divergences [14].

4. Geometry derived from general divergence function

4.1. Tangent space, Riemannian metric and affine connection. We have so far studied the geometry derived from Bregman and f-divergences without mentioning underlying mathematical background. Here, we study the geometry derived from a general divergence, together with basic mathematical concepts from differential geometry. Let us consider a manifold S having a local coordinate system $\boldsymbol{z} = (z_1, ..., z_n)$. Let us consider a differentiable function $D[\boldsymbol{y} : \boldsymbol{z}]$, satisfying the condition of divergence. Then, the Taylor expansion,

$$D(\boldsymbol{z} + d\boldsymbol{z}, \boldsymbol{z}) = \frac{1}{2} \sum g_{ij}(\boldsymbol{z}) dz_i dz_j \tag{114}$$

is a positive definite quadratic form. Hence, a Riemannian metric is induced in S by the second-order derivatives of D at $\boldsymbol{y} = \boldsymbol{z}$, that is,

$$g_{ij}(\boldsymbol{z}) = \frac{\partial^2}{\partial z_i \partial z_j} D(\boldsymbol{z} : \boldsymbol{y}) \big|_{\boldsymbol{y}=\boldsymbol{z}} . \tag{115}$$

It is easy to prove that this is a tensor.

The Bregman divergence induces a dually flat geometrical structure, but this does not necessarily hold in the case of general divergences. However, we also have a dually coupled (non-flat) affine connections together with a Riemannian metric. In order to elucidate the induced geometry, we briefly provide here intuitive explanations of the notion of a tangent space and an affine connection.

Let us consider the tangent space $T_{\boldsymbol{z}}$ of S at point $\boldsymbol{z}$. It is a linear space spanned by basis vectors $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n$, where $\boldsymbol{e}_i$ represents the tangent vector along the coordinate axis z_i. One may regard it as a local linearization of S in a neighborhood of $\boldsymbol{z}$. Any tangent vector $\boldsymbol{X}$ of $T_{\boldsymbol{z}}$ is represented as

$$\boldsymbol{X} = \sum X_i \boldsymbol{e}_i. \tag{116}$$

In particular, the tangent vector of a curve $\boldsymbol{z}(t)$ is represented by

$$\dot{\boldsymbol{z}}(t) = \sum \frac{d}{dt} z_i(t) \boldsymbol{e}_i. \tag{117}$$

The Riemannian metric tensor is expressed by the inner product of two basis vectors,

$$g_{ij}(\boldsymbol{z}) = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle, \tag{118}$$

so that the inner product of two tangent vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \sum g_{ij} X_i Y_j. \tag{119}$$

We next consider an affine connection which is necessary for defining a geodesic. When we consider tangent spaces at all points $\boldsymbol{z} \in S$, they are a collection of local linear approximations of S at different points. Such a collection is a fibre bundle called the tangent bundle. A geodesic is an extension of the "straight line", and is defined as a curve of which the tangent directions are the same along the curve. To define this "sameness", we need to connect tangent spaces defined at different points and define which directions are the same in different tangent spaces. The basis vectors $\boldsymbol{e}_i$ have always the same directions in a flat space, provided affine coordinates are taken. In this special case, $\boldsymbol{e}_i(\boldsymbol{z})$ and $\boldsymbol{e}_i(\boldsymbol{z}+d\boldsymbol{z})$ have the same direction. This is the case with a Bregman divergence. When a space is curved, the directions of $\boldsymbol{e}_i(\boldsymbol{z})$ and $\boldsymbol{e}_i(\boldsymbol{z} + d\boldsymbol{z})$ are different, and we cannot have an affine coordinate system in general.

Let us compare a basis vector $\boldsymbol{e}_i(\boldsymbol{z}) \in T_{\boldsymbol{z}}$ with $\boldsymbol{e}_i(\boldsymbol{z} + d\boldsymbol{z}) \in T_{\boldsymbol{z}+d\boldsymbol{z}}$. Their intrinsic change $\delta \boldsymbol{e}_i$ is defined by $\delta \boldsymbol{e}_i = \widetilde{\boldsymbol{e}}_i(\boldsymbol{z} + d\boldsymbol{z}) - \boldsymbol{e}_i(\boldsymbol{z})$, where $\widetilde{\boldsymbol{e}}_i(\boldsymbol{z} + d\boldsymbol{z})$ is the vector in $T_{\boldsymbol{z}}$ corresponding to $\boldsymbol{e}_i(\boldsymbol{z} + d\boldsymbol{z}) \in T_{\boldsymbol{z}+d\boldsymbol{z}}$ in $T_{\boldsymbol{z}}$. We need to define such correspondence by connecting $T_{\boldsymbol{z}}$ and $T_{\boldsymbol{z}+d\boldsymbol{z}}$. By this correspondence, $\widetilde{\boldsymbol{e}}_i(\boldsymbol{z}+d\boldsymbol{z})$ is a vector in $T_{\boldsymbol{z}}$ so that it is spanned by $\{\boldsymbol{e}_i\}$, and it reduces to 0 as $d\boldsymbol{z} \to 0$. Hence, it may be written in the linear form

$$\delta \boldsymbol{e}_j = \sum \Gamma^k_{ij}(\boldsymbol{z}) dz^i \boldsymbol{e}_k. \tag{120}$$

Therefore, if we define the coefficients Γ^k_{ij}, we have a correspondence between two nearby tangent spaces. This is an affine connection, and Γ^k_{ij} are called the coefficients of the affine connection. Note that $\boldsymbol{e}_i(\boldsymbol{z} + d\boldsymbol{z})$ and $\boldsymbol{e}_i(\boldsymbol{z})$ belong to different tangent spaces, so that we cannot subtract one from the other directly. We have defined the intrinsic difference $\delta \boldsymbol{e}_i$ by the above equation.

Formally, an affine connection is defined by using the covariant derivative operator $\nabla_{\boldsymbol{X}} \boldsymbol{Y}$, which operates on vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$, and shows how the field $\boldsymbol{Y}$ changes as points move in the direction of $\boldsymbol{X}$. A covariant derivative $\nabla_{\boldsymbol{X}} \boldsymbol{Y}$ is defined by using an affine connection as

$$\nabla_{\boldsymbol{X}} \boldsymbol{Y} = \sum \left(\frac{\partial Y^k}{\partial z^i} + \sum \Gamma^k_{ij} Y^j \right) X_i \boldsymbol{e}_k, \tag{121}$$

for two vector fields

$$\boldsymbol{X} = \sum X_i \boldsymbol{e}_i, \tag{122}$$

$$\boldsymbol{Y} = \sum Y_j \boldsymbol{e}_j. \tag{123}$$

The basis vectors are regarded as vector fields, and the covariant derivative of vector field $\boldsymbol{e}_j(\boldsymbol{z})$ in the direction of $\boldsymbol{e}_i(\boldsymbol{z})$ is

$$\nabla_{\boldsymbol{e}_i} \boldsymbol{e}_j = \sum \Gamma^k_{ij} \boldsymbol{e}_k. \tag{124}$$

This shows how $\boldsymbol{e}_j(\boldsymbol{z})$ change intrinsically as points move in the direction of $\boldsymbol{e}_i$. The covariant version of Γ^k_{ij} is

$$\Gamma_{ijk} = \sum_m \Gamma^m_{ij} g_{mk}. \tag{125}$$

4.2. Affine connection derived from divergence. An affine connection is derived from a divergence function $D[\boldsymbol{y} : \boldsymbol{z}]$. It was shown by Eguchi [37] that the coefficients of the derived affine connection are given by

$$\Gamma_{ijk}(\boldsymbol{z}) = -\frac{\partial^3}{\partial z_i \partial z_j \partial y_k} D[\boldsymbol{z} : \boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{z}} . \tag{126}$$

A curve $\boldsymbol{z}(t)$ is a geodesic, when

$$\nabla_{\dot{\boldsymbol{z}}} \dot{\boldsymbol{z}} = 0, \tag{127}$$

where

$$\dot{\boldsymbol{z}} = \frac{d}{dt}\boldsymbol{z}(t).$$

This can be rewritten in the component form,

$$\frac{d^2 z_k(t)}{dt^2} + \sum \Gamma_{ijk}\dot{z}_j\dot{z}_i = 0. \tag{128}$$

Since our affine connection is different from that derived from the Riemannian metric, it is not a curve of minimal distance connecting two points. But it keeps straightness in the sense of this affine connection, because the tangent direction never changes along the curve.

We can define another affine connection, the dual connection, from the dual divergence

$$D^*[\boldsymbol{z} : \boldsymbol{y}] = D[\boldsymbol{y} : \boldsymbol{z}]. \tag{129}$$

The dual affine connection Γ^*_{ijk} is given, in terms of the coefficients, as

$$\Gamma^*_{ijk} = -\frac{\partial^3}{\partial y_i \partial y_j \partial z_k} D[\boldsymbol{z} : \boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{z}}. \tag{130}$$

Two affine connections are said to be mutually dual when

$$D_{\boldsymbol{X}}\langle \boldsymbol{Y}, \boldsymbol{Z}\rangle = \langle \nabla_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{Z}\rangle + \langle \boldsymbol{Y}, \nabla^*_{\boldsymbol{X}}\boldsymbol{Z}\rangle \tag{131}$$

holds for three vector fields $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$. In terms of the components, this relation can be written as

$$\Gamma_{kij} + \Gamma^*_{kji} = \frac{\partial}{\partial z_k} g_{ij}. \tag{132}$$

The Riemannian connection (Levi-Civita connection) Γ^0_{ijk} is given by

$$\Gamma^0_{ijk} = \frac{1}{2}\left(\frac{\partial}{\partial z_i}g_{jk} + \frac{\partial}{\partial z_j}g_{ki} - \frac{\partial}{\partial z_k}g_{ij}\right). \tag{133}$$

It is the average of the two connections,

$$\Gamma^0_{ijk} = \frac{1}{2}\left(\Gamma_{ijk} + \Gamma^*_{ijk}\right). \tag{134}$$

The Riemannian geodesic which minimizes the Riemannian length of a curve is given by

$$\ddot{z}_k + \sum \Gamma^0_{ijk}\dot{z}_i\dot{z}_j = 0. \tag{135}$$

When two affine connections are dually coupled, we have a tensor

$$T_{ijk} = \Gamma_{ijk} - \Gamma^*_{ijk}, \tag{136}$$

which is symmetric with respect to the three indices i, j and k. Therefore, by using this tensor, the two dually coupled affine connections can be written as

$$\begin{aligned} \Gamma_{ijk} &= \Gamma^0_{ijk} - \frac{1}{2}T_{ijk}, \\ \Gamma^*_{ijk} &= \Gamma^0_{ijk} + \frac{1}{2}T_{ijk}. \end{aligned} \tag{137}$$

4.3. Geometry of Bregman divergence. We have already studied the geometry derived from a Bregman divergence,

$$D[\boldsymbol{y} : \boldsymbol{z}] = k(\boldsymbol{y}) - k(\boldsymbol{z}) - \{\text{Grad } k(\boldsymbol{z})\}\cdot(\boldsymbol{y} - \boldsymbol{z}). \tag{138}$$

From (115), the Riemannian metric is given by the Hessian

$$g_{ij}(\boldsymbol{z}) = \frac{\partial^2}{\partial z_i \partial z_j}k(\boldsymbol{z}). \tag{139}$$

Furthermore, we see from (126) that

$$\Gamma_{ijk}(\boldsymbol{z}) = 0 \tag{140}$$

holds. Hence, the space is flat with respect to this connection, and the related coordinates $\boldsymbol{z}$ are affine. However, note that $\Gamma^*_{ijk}(\boldsymbol{z}) \neq 0$. If we use the dual affine coordinate system $\boldsymbol{z}^*$, then the dual affine connection $\Gamma^*_{ijk}(\boldsymbol{z}^*)$ vanishes in the dual coordinate system.

4.4. Geometry of f-divergence. We now study the geometrical structure of S_n induced by an f-divergence which has information monotonicity. We begin with the induced Riemannian metric.

Theorem 5. Any f-divergence induces a unique information monotonic Riemannian metric, which is given by the Fisher information matrix.

Proof. The Riemannian metric induced by D_f is calculated from (88) as

$$D_f[\boldsymbol{p} : \boldsymbol{p} + d\boldsymbol{p}] = \frac{1}{2}\sum_{i=0}^{n}\frac{(dp_i)^2}{p_i}. \tag{141}$$

Let $\boldsymbol{z} = (p_1, \cdots, p_n)$ be the coordinate system, where $p_i = z_i$ $(i = 1, \cdots, n)$ and $p_0 = 1 - \sum z_i$. By eliminating dp_0 by using $dp_0 = -\sum dz_i$, we have

$$g_{ij} = \frac{1}{p_i}\delta_{ij} + \frac{1}{p_0}. \tag{142}$$

This coincides with the Fisher information matrix defined by

$$g_{ij}(\boldsymbol{z}) = \int p(\boldsymbol{x}, \boldsymbol{z})\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{z})}{\partial z_i}\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{z})}{\partial z_j}d\boldsymbol{x}. \tag{143}$$

Hence for any f, the Riemannian metric is the same and given by the Fisher information matrix. This is the only invariant metric of S_n.

Since an f-divergence $D_f[\boldsymbol{p} : \boldsymbol{q}]$ cannot be written in the form of Bregman divergence, except for the case of the KL-divergence, there are no affine and dual affine coordinate systems such that $\Gamma_{ijk}(z) = 0$ or $\Gamma^*_{ijk}(\boldsymbol{z}^*) = 0$ holds. This implies that D_f does not induce a flat structure in S_n. However, it induces a pair of dual affine connections, which define primal and dual geodesics.

A symmetric tensor is defined by

$$T_{ijk}(\boldsymbol{z}) = \int p(\boldsymbol{x}, \boldsymbol{z})\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{z})}{\partial z_i}\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{z})}{\partial z_j}\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{z})}{\partial z_k}d\boldsymbol{x}. \tag{144}$$

The integration reduces to the summation in the discrete case. We can then calculate the induced primal and dual affine connections.

Theorem 6. An f-divergence $D_f[\boldsymbol{z} : \boldsymbol{y}]$ induces the primal and dual affine connections defined by

$$\Gamma_{ijk} = \Gamma^{\alpha}_{ijk} = \Gamma^{0}_{ijk} - \frac{\alpha}{2} T_{ijk}, \tag{145}$$

$$\Gamma^{*}_{ijk} = \Gamma^{-\alpha}_{ijk} = \Gamma^{0}_{ijk} + \frac{\alpha}{2} T_{ijk}, \tag{146}$$

where $\alpha = 3 + 2f'''(1)$.

Proof. By simple calculations, we have

$$\begin{aligned} \Gamma_{ijk} &= -\frac{\partial^3}{\partial z_i \partial z_j \partial y_k} D_f[\boldsymbol{z} : \boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{z}} \\ &= \Gamma^{0}_{ijk} - \left(f'''(1) + \frac{3}{2} \right) T_{ijk}, \end{aligned} \tag{147}$$

and its dual,

$$\Gamma^{*}_{ijk} = \Gamma^{0}_{ijk} + \left\{ f'''(1) + \frac{3}{2} \right\} T_{ijk}. \tag{148}$$

They are two dually coupled affine connections, given by

$$\Gamma^{\pm\alpha}_{ijk} = \Gamma^{0}_{ijk} \pm \frac{\alpha}{2} T_{ijk}, \tag{149}$$

where $\alpha = 3 + 2f'''(1)$. The geometrical structure given by the triplet $\{S, g_{ij}, T_{ijk}\}$ is called the α geometry.

The α geometry is a natural consequence of information monotonicity. This is explained from the invariancy principle that the geometrical structure is invariant by transformation from $\boldsymbol{x}$ to $\boldsymbol{t}(\boldsymbol{x})$, when $\boldsymbol{t}$ is a sufficient statistics. For more details, see [1, 11, 36].

Among these divergences, the KL-divergence and its dual are unique in the sense that the space is dually flat in S_n. In other words, there exist affine and dual affine coordinate systems $\boldsymbol{z}$ and $\boldsymbol{z}^*$ such that the two affine connections Γ_{ijk} and Γ^{*}_{ijk} vanish in the respective coordinate systems. Hence, it is written in the form of a Bregman divergence, as is given in Subsec. 2.3. A dually flat manifold has the Pythagorean and Projection theorems, which are useful for a wide range of applications.

4.5. Invariant Geometry of M_n. An f-divergence induces a dually flat structure in M_n, as we saw in the previous section. For positive measures $\boldsymbol{z} = (z_1, \cdots, z_n)$, any f-divergence gives

$$D_f(\boldsymbol{z} : \boldsymbol{z} + d\boldsymbol{z}) = \frac{1}{2} \sum \frac{1}{z_i} (dz_i)^2. \tag{150}$$

Hence, the Riemannian metric is

$$g_{ij}(\boldsymbol{z}) = \frac{1}{z_i} \delta_{ij}. \tag{151}$$

It is Euclidean, because it can be changed into

$$\widetilde{g}_{ij}\,(\widetilde{\boldsymbol{z}}) = \delta_{ij}, \tag{152}$$

by the coordinate transformation

$$\widetilde{z}_i = 2\sqrt{z_i}. \tag{153}$$

We can also calculate the coefficients of the primal and dual affine connections.

Theorem 7. An f-divergence $D_f[\boldsymbol{z} : \boldsymbol{y}]$ induces the primal and dual affine connections in M_n,

$$\Gamma_{ijk}(\boldsymbol{z}) = \frac{1}{2z_i^2}(-1-\alpha)\delta_{ijk}, \tag{154}$$

$$\Gamma^{*}_{ijk}(\boldsymbol{z}) = \frac{1}{2z_i^2}(-1+\alpha)\delta_{ijk}, \tag{155}$$

where $\alpha = 3 + 2f'''(1)$, and $\delta_{ijk} = 1$ when $i = j = k$ and 0 otherwise.

This shows that any f-divergence induces a family of affine connections indexed by $\alpha = 3+2f'''(1)$. We can further prove the following flatness theorem.

Theorem 8. The primal and dual affine connections of M_n induced by an f-divergence are flat.

Proof. The theorem can be proved by calculating the Riemann-Christoffel curvature tensors, which we omit.

A flat manifold has an affine coordinate system. We have already shown that the α-representation given in (107) is an affine coordinate system, and $-\alpha$-representation (107) is a dual affine coordinate system.

4.6. Canonical divergence. Given a divergence $D[\boldsymbol{z} : \boldsymbol{y}]$, we have a Riemannian metric and a dual pair of affine connections. However, given a Riemannian manifold with a dual pair of affine connections, we have infinitely many divergences that generate the same geometrical structure. This is because the differential geometrical structure depends only on local properties of a divergence function. For example, given a divergence $D[\boldsymbol{z} : \boldsymbol{y}]$, its modification

$$\widetilde{D}[\boldsymbol{z} : \boldsymbol{y}] = D[\boldsymbol{z} : \boldsymbol{y}] + c \sum |z_i - y_i|^4, \quad c > 0 \tag{156}$$

gives the same geometry.

When a manifold S is dually flat, there exist two dual affine coordinate systems $\boldsymbol{z}$ and $\boldsymbol{z}^*$, accompanied by two convex functions $k(\boldsymbol{z})$ and $k^*(\boldsymbol{z}^*)$. These coordinate systems are unique up to affine transformations, and convex functions are unique up to linear terms. However, the divergence is uniquely determined,

$$D[\boldsymbol{y} : \boldsymbol{z}] = k(\boldsymbol{y}) + k^*(\boldsymbol{z}^*) - \boldsymbol{y} \cdot \boldsymbol{z}^*. \tag{157}$$

We call it the canonical divergence of a dually flat manifold. The Bregman divergence is the canonical divergence of a dually flat Riemannian manifold. So we have:

Theorem 9. The α-divergence is the canonical divergence in the dually flat space M_n of positive measures. The KL-divergence is the canonical divergence in the dually flat space S_n of discrete probability distributions.

4.7. Symmetric divergence. A divergence is not symmetric in general. However, we can construct a symmetric divergence $D_s[\boldsymbol{z}:\boldsymbol{y}]$ from an asymmetric $D[\boldsymbol{z}:\boldsymbol{y}]$ by

$$D_s[\boldsymbol{z}:\boldsymbol{y}] = \frac{1}{2}\left\{D[\boldsymbol{z}:\boldsymbol{y}] + D[\boldsymbol{y}:\boldsymbol{z}]\right\}. \tag{158}$$

Its geometry is elucidated by the following theorem.

Theorem 10. The symmetrized divergence gives the same Riemannian metric as the asymmetric one. The derived affine connections are self-dual, and is the Riemannian connection,

$$\Gamma^S_{ijk} = \Gamma^{*S}_{ijk} = \Gamma^0_{ijk}. \tag{159}$$

Proof. From

$$\frac{\partial^2}{\partial z_i \partial z_j} D[\boldsymbol{z}:\boldsymbol{y}]_{\boldsymbol{y}=\boldsymbol{z}} = \frac{\partial^2}{\partial y_i \partial y_j} D[\boldsymbol{z}:\boldsymbol{y}]_{\boldsymbol{y}=\boldsymbol{z}}, \tag{160}$$

we have

$$g^S_{ij} = g_{ij}. \tag{161}$$

It is easy to see

$$-\frac{\partial^3}{\partial z_i \partial z_j \partial y_k} D_s[\boldsymbol{z}:\boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{z}} = \frac{1}{2}\left\{\Gamma_{ijk} + \Gamma^*_{ijk}\right\}. \tag{162}$$

We may say that the squared Riemannian distance is the canonical divergence of a space with a symmetric divergence.

4.8. Tsallis entropy, q-exponential family, and conformal geometry. Let us define

$$h_q(\boldsymbol{p}) = \sum p_i^q, \qquad 0 < q < 1. \tag{163}$$

Then, the Tsallis q-entropy [15] is defined by

$$H_q = \frac{1}{1-q}\left(h_q - 1\right). \tag{164}$$

It is connected with the Rényi entropy [16] defined as

$$H^R_q = \frac{1}{1-q}\log h_q, \tag{165}$$

by a monotonic transformation. Since both $H_q(\boldsymbol{p})$ and $H^R_q(\boldsymbol{p})$ are concave functions, we can introduce dual geometric structures by using $k(\boldsymbol{p}) = -H_q$ and $\widetilde{k}(\boldsymbol{p}) = -H^R_q$. We next study how the geometry changes by a nonlinear transformation of a convex function.

Given a convex function $k(\boldsymbol{z})$, we can obtain a new function $\widetilde{k}(\boldsymbol{z})$ by a nonlinear transformation,

$$\widetilde{k}(\boldsymbol{z}) = f\left\{k(\boldsymbol{z})\right\}. \tag{166}$$

When $f(u)$ is a monotonically increasing convex function, $\widetilde{k}(\boldsymbol{z})$ is a convex function. The geometry derived by $\widetilde{k}(\boldsymbol{z})$ is compared with the original geometry derived by $k(\boldsymbol{z})$.

The Riemannian metric $\widetilde{g}_{ij}$ is given by

$$\widetilde{g}_{ij}(\boldsymbol{z}) = f'\left\{k(\boldsymbol{z})\right\} g_{ij}(\boldsymbol{z}) + f''\left\{k(\boldsymbol{z})\right\} \frac{\partial k(\boldsymbol{z})}{\partial z_i}\frac{\partial k(\boldsymbol{z})}{\partial z_j}. \tag{167}$$

As for the affine structure, $\boldsymbol{z}$ is a common affine coordinate system, so that the geodesics are the same in both cases. The dual coordinates $\widetilde{\boldsymbol{z}}^*$ are related to $\boldsymbol{z}^*$ by

$$\widetilde{\boldsymbol{z}}^* = \mathrm{Grad}\,\widetilde{k}(\boldsymbol{z}) = f'\left\{k(\boldsymbol{z})\right\}\boldsymbol{z}^*. \tag{168}$$

For a dual geodesic in the original geometry given by

$$\boldsymbol{z}^*(t) = t\boldsymbol{z}^* + (1-t)\boldsymbol{b}^*, \tag{169}$$

the curve in the new dual coordinate system is given by

$$\widetilde{\boldsymbol{z}}^*(t) = tf'(t)\boldsymbol{a}^* + (1-t)f'(t)\boldsymbol{b}^*, \tag{170}$$

where

$$f'(t) = f'\left[k\left\{\boldsymbol{z}\left(\boldsymbol{z}^*(t)\right)\right\}\right]. \tag{171}$$

This is not a dual geodesic of the derived geometry.

However, when $\boldsymbol{b}^* = 0$, a geodesic passes through the origin $\boldsymbol{z}^* = \widetilde{\boldsymbol{z}}^*$, which is the point minimizing the convex function $k(\boldsymbol{z})$ and $\widetilde{k}(\boldsymbol{z})$. Such a dual geodesic is again a dual geodesic with respect to the transformed geometry, since

$$\boldsymbol{z}^*(t) = t\boldsymbol{a}^* \tag{172}$$

is transformed to

$$\widetilde{\boldsymbol{z}}^*(t) = tf'(t)\boldsymbol{a}^*. \tag{173}$$

When $k(\boldsymbol{z})$ is the negative entropy, this represents the fact that the maximum-entropy theorem holds commonly both for $k(\boldsymbol{z})$ and $\widetilde{k}(\boldsymbol{z})$.

We further study the geometry of S_n derived from the q-exponential family [17, 18, 27]. Let us define the q-exponential function by

$$\exp_q(z) = \begin{cases} \left\{1 + (1-q)z\right\}^{\frac{1}{1-q}}, & q \neq 1, \\ \exp z, & q = 1. \end{cases} \tag{174}$$

Then, a family of probability distributions parameterized by $\boldsymbol{z}$,

$$p(\boldsymbol{x}, \boldsymbol{z}) = \exp_q\left\{\boldsymbol{z}\cdot\boldsymbol{x} - \psi(\boldsymbol{z})\right\} \tag{175}$$

is called a q-exponential family. This is an ordinary exponential family when $q = 1$. Here, $\psi(\boldsymbol{z})$ is a normalizing factor to ensure

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{z}) = 1. \tag{176}$$

We can prove that $\psi(\boldsymbol{z})$ is a convex function.

In the case of S_n, we have

$$\log_q p(x, \boldsymbol{z}) = \frac{1}{1-q}\left\{\sum_{i=1}^n z_i\delta_i(x) - \psi(\boldsymbol{z})\right\}, \tag{177}$$

where we used the definition of

$$z_i = \log_q p_i - \log_q p_0, \tag{178}$$

$$\psi(\boldsymbol{z}) = -\log_q p_0(\boldsymbol{z}), \tag{179}$$

and

$$\log_q(z) = \begin{cases} \dfrac{1}{1-q}\left(z^{1-q} - 1\right), & q \neq 1, \\ \log z, & q = 1. \end{cases} \tag{180}$$

By studying the geometrical structure derived from $\psi(\boldsymbol{z})$, we obtain the following new theorems. We omit their proofs.

Theorem 11. The q-Riemannian metric g^q_{ij} derived from $\psi(\boldsymbol{z})$ is a conformal transform of the Fisher information metric,

$$g^q_{ij}(\boldsymbol{p}) = \frac{q}{h_q(\boldsymbol{p})} g_{ij}(\boldsymbol{p}). \tag{181}$$

Theorem 12. The q-divergence is a conformal transform of the α-divergence, with $q = (1+\alpha)/2$,

$$D_q[\boldsymbol{p} : \boldsymbol{q}] = \frac{1}{(1-q)h_q(\boldsymbol{p})}\left(1 - \sum p_i^q q_i^{1-q}\right). \qquad (182)$$

These results pave a way to develop conformal information geometry further.

5. Conclusions

We have studied various divergence measures and the differential-geometrical structures derived therefrom. This connects many important engineering problems in computational vision, optimization, signal processing, neural networks, etc. with a modern information geometry. A divergence function gives a Riemannian metric to the underlying space, and furthermore a pair of dually coupled affine connections. It has been shown that the Bregman divergence always gives a dually flat Riemannian structure, and conversely, a dually flat Riemannian manifold always gives a canonical divergence of the Bregman type. We have presented a number of important divergences of the Bregman type.

Information monotonicity is a natural requirement for a divergence defined in a manifold of probability distributions. This leads to the class of f-divergences. We have proved that an f-divergence gives the unique Riemannian metric, which is the Fisher information matrix. It also gives a class of α-connections, where α- and $-\alpha$-connections are dually coupled. By extending it to the manifold of positive measures, we have shown that the α-divergences are canonical divergences.

We have addressed the geometrical structure inspired from the Tsallis q-entropy and the corresponding divergences. This will open a new field of conformal information geometry.

REFERENCES

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Oxford University Press, New York, 2000.

[2] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations*, John Wiley, New York, 2009.

[3] F. Nielsen, "Emerging trends in visual computing", *Lecture Notes in Computer Science* 6, CD-ROM (2009).

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

[5] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences", *J. Machine Learning Research* 6, 1705–1749 (2005).

[6] M.S. Ali and S.D. Silvey, "A general class of coefficients of divergence of one distribution from another", *J. Royal Statistical Society*, B(28), 131–142 (1966).

[7] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations", *Studia Sci. Math.* 2, 299–318 (1967).

[8] I. Csiszár, "Information measures: a critical survey", *Transaction of the 7th Prague Conf.* 1, 83–86 (1974).

[9] I. Taneja and P. Kumar, "Relative information of type s, Csiszár's f-divergence, and information inequalities", *Information Sciences* 166, 105–125 (2004).

[10] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear problems", *Annuals of Statistics* 19, 2032–2066 (1991).

[11] N.N. Chentsov, *Statistical Decision Rules and Optimal Inference*, American Mathematical Society, New York, 1972.

[12] I. Csiszár, "Axiomatic characterizations of information measures", *Entropy* 10, 261–273 (2008).

[13] G. Pistone and C. Sempi, "An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given on", *Annals of Statistics* 23, 1543–1561 (1995).

[14] S. Amari, "Alpha divergence is unique, belonging to both classes of f-divergence and Bregman divergence", *IEEE Trans. Information Theory* B 55, 4925–4931 (2009).

[15] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics", *J. Stat. Phys.* 52, 479–487 (1988).

[16] A. Rényi, "On measures of entropy and information", *Proc. 4th Berk. Symp. Math. Statist. and Probl.* 1, 547–561 (1961).

[17] J. Naudts, "Estimators, escort probabilities, and phi-exponential families in statistical physics", *J. Ineq. Pure App. Math.* 5, 102 (2004).

[18] J. Naudts, "Generalized exponential families and associated entropy functions", *Entropy* 10, 131–149 (2008).

[19] H. Suyari, "Mathematical structures derived from q-multinomial coefficient in Tsallis statistics", *Physica* A 368, 63–82 (2006).

[20] S. Amari, "Information geometry and its applications: convex function and dually flat manifold", *Emerging Trends in Visual Computing* A2, 5416 (2009).

[21] M.R. Grasselli, "Duality, monotonicity and Wigner-Yanase-Dyson metrics", *Infinite Dimensional Analysis, Quantum Probability and Related Topics* 7, 215–232 (2004).

[22] H. Hasegawa, "α-divergence of the non-commutative information geometry", *Reports on Mathematical Physics* 33, 87–93 (1993).

[23] D. Petz, "Monotone metrics on matrix spaces", *Linear Algebra and its Applications* 244, 81–96 (1996).

[24] I.S. Dhillon and J.A. Tropp, "Matrix nearness problem with Bregman divergences", *SIAM J. on Matrix Analysis and Applications* 29, 1120–1146 (2007).

[25] Yu. Nesterov and M.J. Todd, "On the Riemannian geometry defined by self-concordant barriers and interior-point methods", *Foundations of Computational Mathematics* 2, 333–361 (2002).

[26] A. Ohara and T. Tsuchiya, *An Information Geometric Approach to Polynomial-time Interior-time Algorithms*, (to be published).

[27] A. Ohara, "Information geometric analysis of an interior point method for semidefinite programming", *Geometry in Present Day Science* 1, 49–74 (1999).

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

[29] S. Eguchi and J. Copas, "A class of logistic type discriminant function", *Biometrika* 89, 1–22 (2002).

[30] M. Minami and S. Eguchi," Robust blind source separation by beta-divergence", *Neural Commutation* 14, 1859–1886 (2004).

[31] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination", *J. Multivariate Analysis* 99, 2053–2081 (2008).

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

[33] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations", *Annals of Mathematical Statistics* 23, 493–507 (1952).

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

[35] Y. Matsuyama, "The α-EM algorithm: Surrogate likelihood maximization using α-logarithmic information measures", *IEEE Trans. on Information Theory* 49, 672–706 (2002).

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

[37] S. Eguchi, "Second order efficiency of minimum contrast estimations in a curved exponential family", *Annals of Statistics* 11, 793–803 (1983).