DIFFERENTIAL GEOMETRY DERIVED FROM DIVERGENCE FUNCTIONS: INFORMATION GEOMETRY APPROACH

Shun-ichi Amari

Abstract: *We study differential-geometrical structure of an information manifold equipped with a divergence function. A divergence function generates a Riemannian metric and furthermore it provides a symmetric third-order tensor, when the divergence is asymmetric. This induces a pair of affine connections dually coupled to each other with respect to the Riemannian metric. This is the arising emerged from information geometry. When a manifold is dually flat (it may be curved in the sense of the Levi-Civita connection), we have a canonical divergence and a pair of convex functions from which the original dual geometry is reconstructed. The generalized Pythagorean theorem and projection theorem hold in such a manifold. This structure has lots of applications in information sciences including statistics, machine learning, optimization, computer vision and Tsallis statistical mechanics. The present article reviews the structure of information geometry and its relation to the divergence function. We further consider the conformal structure given rise to by the generalized statistical model in relation to the power law.*

Keywords: *divergence, information geometry, dual affine connection, Bregman divergence, generalized Pythagorean theorem*

ACM Classification Keywords*: G.3 PROBABILITY AND STATISTICS*

MSC*:* 52, 53, 60

Introduction

A divergence function $D[P : Q]$ between two points P and Q in a manifold M plays a fundamental role in many engineering problems, including information theory, statistics, machine learning, computer vision, optimization and brain science. It has dimension of the square of distance but is not in general symmetric with respect to P and Q. The present article surveys the differential-geometric structure generated by a divergence function [Vos, 1991], [Amari and Nagaoka, 2000], [Eguchi, 1983], [Amari and Cichocki, 2010]. When it is symmetric, it gives a Riemannian metric and the Levi-Civita affine connection follows as was studied by [Rao, 1945]. However, when it is asymmetric, we have a third-order symmetric tensor, which gives a pair of affine connections dually coupled to each other with respect to the Riemannian metric. These are the central structure of information geometry [Amari and Nagaoka, 2000] which studies an invariant structure of a manifold of probability distributions [Amari, 1985].

Many asymmetric divergence functions are used in application of information theory. For example, the Kullback-Leibler divergence, which is frequently used in statistics, information theory and others, is a typical example of asymmetric divergence. We study the geometrical structure arising from by an asymmetric divergence. It consists of a Riemannian metric which is symmetric positive-definite second-order tensor and a symmetric third-order tensor which vanishes in the symmetric case. A pair of affine connections is defined by using the two tensors. The Levi-Civita connection is the average of the two connections. They are not metric connections but are dually coupled with respect to the Riemannian metric in the sense that the parallel transports of a vector by the two affine connections keep their inner product invariant with respect to the Riemannian metric. The duality can be expressed in terms of the related covariant derivatives. See [Amari and Nagaoka, 2000] for details.

When a Riemann-Christoffel curvature vanishes with respect to one affine connection, it vanishes automatically with respect to the dual affine connection. We have a dually flat manifold in this case, even though the Riemannian curvature with respect to the Levi-Civita connection does not vanish in general. A dually flat Riemannian manifold has nice properties such as the generalized Pythagorean theorem and projection theorem. Moreover, when a manifold is dually flat, we have two convex functions from which a canonical divergence is uniquely determined. The canonical divergence generates the original dually flat Riemannian structure. Euclidean space is a special example of the dually flat manifold which has a symmetric divergence given by the square of the Euclidean distance.

A dually flat manifold has two affine coordinate systems related to the two flat affine connections. They are connected by the Legendre transformation of the two convex functions. The canonical divergence is given as the Bregman divergence [Bregman, 1967]. The Legendre structure has a geometrical foundation in the framework of the dually flat geometry.

What is the natural divergence function to be introduced in a manifold? We study this question in the case of a manifold of probability distributions. We impose an invariant criterion such that the geometry is invariant under bijective transformations of random variables [Chentsov, 1982], [Picard, 1992] (more generally invariant by using sufficient statistics). Then, the Kullback-Leibler divergence is given as the unique canonical divergence. We further extend our notions to the manifold of positive measures [Amari, 2009]. We have invariant structure [Chentsov, 1982; Amari, 1985] and non-invariant flat structure [Amari and Ohara, 2011] which is related to the Tsallis entropy
[Tsallis, 2009].

We finally study the structure of deformed exponential family [Naudts, 2011] which includes the Tsallis q-structure [Tsallis, 2009]. We can introduce a dually flat geometry in this manifold, which is not invariant in general. We prove that the invariant structure is limited to the α- or q-geometry, which gives non-flat dual geometry. However, we can define a dually flat structure in the q-family which is not invariant, extending the geometry. We prove that the q- or α-structure is unique in the sense that the flat geometry is given by a conformal transformation of the invariant geometry [Amari and Ohara, 2011], [Amari, Ohara and Matsuzoe, 2012].

We do not mention applications of dual geometry, which are now hot topics of research in many fields.
See, e.g., [Banerjee et al., 2005], [Ikeda, Tanaka and Amari, 2004], [Takenouchi et al., 2008],
[Boissonnat, Nielsen and Nock, 2010], [Vemuri et al., 2011], [Liu et al, 2010], [Amari, 2009] and [Cichocki et al., 2009].

Divergence Function and Differential Geometry

Let us consider a manifold M homeomorphic to $\boldsymbol{R}^n$. We use a coordinate system in M and denote the coordinates of a point P by $\boldsymbol{x} = (x_1, \cdots, x_n)$. We consider a function $D[P : Q]$ of two points P and Q in M, which is written as $D[\boldsymbol{x} : \boldsymbol{y}]$ by using the coordinates $\boldsymbol{x}$ and $\boldsymbol{y}$ of P and Q.

Definition: A function $D[P : Q]$ is called a divergence, when the following conditions are satisfied
[Amari and Cichocki, 2010]:

1) $D[\boldsymbol{x} : \boldsymbol{y}] \geq 0$ with equality when and only when $\boldsymbol{x} = \boldsymbol{y}$.

2) $D[\boldsymbol{x} : \boldsymbol{y}]$ is differentiable and the Hessian with respect to $\boldsymbol{x}$ at $\boldsymbol{y} = \boldsymbol{x}$ is positive definite.

It should be noted that $D[\boldsymbol{x} : \boldsymbol{y}]$ is not necessarily symmetric with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$. Hence, it is not a distance. The triangular inequality does not hold either. It has dimension of the square of distance as will be seen below (cf. [Chen, Chen and Rao, 2008]).

Given a divergence function D, for infinitesimally close two points $\boldsymbol{x}$ and $\boldsymbol{y} = \boldsymbol{x} + d\boldsymbol{x}$, we have by Taylor expansion

$$D\left[\boldsymbol{x} : \boldsymbol{x} + d\boldsymbol{x}\right] = \sum g_{ij}(\boldsymbol{x}) dx^i dx^j + O\left(|d\boldsymbol{x}|^3\right), \tag{1}$$

where

$$g_{ij}(\boldsymbol{x}) = \frac{\partial^2}{\partial x_i \partial x_j} D\left[\boldsymbol{x} : \boldsymbol{y}\right] \big|_{\boldsymbol{y}=\boldsymbol{x}} . \tag{2}$$

Since g_{ij} is a positive-definite matrix, it gives a Riemannian metric derived from the divergence. For the sake of notational convenience, we introduce the following abbreviation by using the partial differential operator (natural basis of the tangent space),

$$\partial_i = \frac{\partial}{\partial x_i} \tag{3}$$

and define, for a number of operators $\partial_i, \partial_j, \partial_k$, etc.,

$$D\left[\partial_i \partial_j : \partial_k ; \boldsymbol{x}\right] = \frac{\partial^3}{\partial x_i \partial x_j \partial y_k} D[\boldsymbol{x} : \boldsymbol{y}] \big|_{\boldsymbol{y}=\boldsymbol{x}} . \tag{4}$$

Here, the operators in the left part of D operate on $\boldsymbol{x}$, while those on the right operate on $\boldsymbol{y}$, and finally the result is evaluated at $\boldsymbol{y} = \boldsymbol{x}$. Then, we have, for example,

$$D\left[\partial_i : \cdot ; \boldsymbol{x}\right] = \frac{\partial}{\partial x_i} D[\boldsymbol{x} : \boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{x}} = 0 \tag{5}$$

$$D\left[\cdot : \partial_j ; \boldsymbol{x}\right] = \frac{\partial}{\partial y_j} D[\boldsymbol{x} : \boldsymbol{y}]_{\boldsymbol{y}=\boldsymbol{x}} = 0 \tag{6}$$

$$g_{ij} = D\left[\partial_i \partial_j : \cdot ; \boldsymbol{x}\right] = D\left[\cdot : \partial_i \partial_j ; \boldsymbol{x}\right] = -D\left[\partial_i : \partial_j ; \boldsymbol{x}\right]. \tag{7}$$

The above properties are proved as follow. When $\boldsymbol{y} - \boldsymbol{x}$ is small, we have from (1)

$$D[\boldsymbol{x} : \boldsymbol{y}] = \sum g_{ij}(\boldsymbol{x}) \left(x_1 - y_i\right) \left(x_j - y_j\right) + O\left(|\boldsymbol{x} - \boldsymbol{y}|^3\right). \tag{8}$$

By differentiating this with respect to x_i and/or y_j, and then putting $\boldsymbol{y} = \boldsymbol{x}$, we have (5), (6), (7).

We can easily prove that g_{ij} is a tensor.

Define

$$T_{ijk}(\boldsymbol{x}) = D\left[\partial_k : \partial_i \partial_j ; \boldsymbol{x}\right] - D\left[\partial_i \partial_j : \partial_k ; \boldsymbol{x}\right]. \tag{9}$$

We can prove that T_{ijk} is a tensor symmetric with respect to the three indices[Eguchi, 1983]. When $D[\boldsymbol{x} : \boldsymbol{y}]$ is symmetric, i.e., $D[\boldsymbol{x} : \boldsymbol{y}] = D[\boldsymbol{y} : \boldsymbol{x}]$,

$$T_{ijk} = 0. \tag{10}$$

A manifold M having a divergence function is equipped with two quantities g_{ij} and T_{ijk} derived from it. We write it as $\{M, g_{ij}, T_{ijk}\}$. Obviously the inner product of ∂_i and ∂_j is $\langle \partial_i, \partial_j \rangle = g_{ij}$.

We introduce affine connections in manifold M equipped with metric g_{ij} and cubic form T_{ijk}, that is, $\{M, g_{ij}, T_{ijk}\}$. When g_{ij} is given, the Levi-Civita or Riemannian connection is given by the Christoffel symbol

$$\Gamma^0_{ijk}(\boldsymbol{x}) = [i, j; k] = \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\right). \tag{11}$$

This is a symmetric connection (torsion-free) and the related covariant derivative ∇^0 in the direction ∂_i satisfies

$$\nabla^0_{\partial_i} \langle \partial_j, \partial_k \rangle = \nabla^0_{\partial_i} g_{jk} = 0. \tag{12}$$

When a cubic tensor T_{ijk} is given in addition to g_{jk}, we define the α-connection [Amari and Nagaoka, 2000] by

$$\Gamma^{\alpha}_{ijk}(\boldsymbol{x}) = \Gamma^{0}_{ijk} - \frac{\alpha}{2} T_{ijk}, \tag{13}$$

where α is a real scalar parameter. The α-covariant derivative ∇^{α} is characterized by

$$\nabla^{\alpha}_{\partial i} g_{jk} = \frac{\alpha}{2} T_{ijk}. \tag{14}$$

The Levi-Civita connection is a special case of α-connection because it is given by $\alpha = 0$. When $\alpha = 1$, we call the 1-connection simply the (primal) connection and when $\alpha = -1$, the dual connection, respectively, derived from the divergence. The duality will be explained in the next section.

Dual connections with respect to Riemannian metric

We search for geometrical structure of manifold $\{M, g_{ij}, T_{ijk}\}$. Two affine connections Γ_{ijk} and Γ^{*}_{ijk} (or their covariant derivatives ∇ and ∇^{*}) in M are said to be dual with respect to Riemannian metric g_{ij}, when the following relation holds for three vector fields A, B and C

$$A\langle B, C\rangle = \langle \nabla_A B, C\rangle + \langle B, \nabla^{*}_A C\rangle, \tag{15}$$

where

$$\langle B, C\rangle = \sum g_{ij} B^i C^j \tag{16}$$

is the inner product of B and C, and A is the directional derivative operator, $\sum A^i \partial_i$, where $A = \sum A^i \partial_i$, $B = \sum B^i \partial_i$ and $C = \sum C^i \partial_i$. The following theorem is known [Amari, 1985], [Amari and Nagaoka, 2000].

Theorem 1. The α-connection and $-\alpha$-connection are dual with respect to the Riemannian metric g_{ij}.

Dually flat manifold

Manifold $\{M, g_{ij}, T_{ijk}\}$ derived from a divergence function is equipped with a Riemannian metric and two dual affine connections. Hence, we may represent it by $\{M, g, \nabla, \nabla^{*}\}$ in terms of metric $g = (g_{ij})$ and two dual affine connections (covariant derivatives) ∇ and ∇^{*}. Manifold M is in general curved, having non-zero Riemann-Christoffel curvature.

We prove that, when the Riemann-Christoffel curvature vanishes for the primal connection ∇, the Riemann-Christoffel curvature of the dual connection ∇^{*} vanishes automatically. But the Riemann-Christoffel curvature of the Levi-Civita connection does not vanish in general. A manifold $\{M, g, \nabla, \nabla^{*}\}$ is said to be dually flat, when the curvatures for ∇ and ∇^{*} vanish.

Theorem 2. When the primal curvature vanishes, the dual curvature vanishes at the same time.

Proof. Let $\prod$ and $\prod^{*}$ be parallel transport operators of a vector due to ∇ and ∇^{*}, respectively. The duality implies, in terms of the parallel transports, that

$$\langle A, B\rangle_P = \langle \prod A, \overset{*}{\prod} B\rangle_Q, \tag{17}$$

when two vectors A and B are transported from point P to Q along a curve connecting P and Q by the two parallel transports $\prod$ and $\prod^{*}$. Let us consider a loop c starting from P and ending at Q. Then, when the primal curvature vanishes, we have

$$\prod_{c} A = A \tag{18}$$

for any A. This implies

$$\prod_{c}^{*} B = B \tag{19}$$

for any B. This proves that the curvature vanishes for the dual connection.

When M is dually flat, the following properties hold. See [Amari and Nagaoka, 2000] for mathematical details.

Theorem 3. Let $\{M, g, \nabla, \nabla^*\}$ be a dually flat manifold. Then, the following holds.

1) There are two affine coordinate systems $\boldsymbol{\theta} = \left(\theta^1, \cdots, \theta^n\right)$ and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$ in which the coefficients of the primal and dual connections vanish, respectively,

$$\Gamma_{ijk}(\boldsymbol{\theta}) = 0, \quad \Gamma^{*ijk}(\boldsymbol{\eta}) = 0. \tag{20}$$

We denote the components of $\boldsymbol{\theta}$ by θ^i using the upper index (contravariant) and those of $\boldsymbol{\eta}$ by η_i using the lower index (covariant) because of the duality. The two affine coordinates are unique up to affine transformations.

2) There exist two potential functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ which are convex.

3) The Riemannian metric is given by the Hessian of the potential functions in the respective coordinate systems,

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}), \quad \partial_i = \frac{\partial}{\partial \theta^i}, \tag{21}$$

$$g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta}), \quad \partial^i = \frac{\partial}{\partial \eta_i}. \tag{22}$$

Here, g_{ij} and g^{ij} are inverse matrices,

$$\sum g_{ij} g^{jk} = \delta_i^k, \tag{23}$$

where $\left(\delta_i^k\right)$is the identity matrix and the cubic tensor T is given by

$$T_{ijk}(\boldsymbol{\theta}) = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}), \quad T^{ijk}(\boldsymbol{\eta}) = \partial^i \partial^j \partial^k \varphi(\boldsymbol{\eta}). \tag{24}$$

4) There exists a canonical divergence $D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]$, which is unique up to a scalar constant and is defined by

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi(\boldsymbol{\theta}) + \varphi\left(\boldsymbol{\eta}'\right) - \sum \theta^i \eta_i'. \tag{25}$$

The geometrical structure, the Riemannian metric and the dual affine connections ∇ and ∇^*, derived from this divergence is the same as that of the original $\{M, g, T\}$.

5) A primal geodesic is linear in the $\boldsymbol{\theta}$ coordinate system and the dual geodesic is linear in the $\boldsymbol{\eta}$ coordinate system, so that they are given, respectively, by

$$\boldsymbol{\theta}(t) = t\boldsymbol{a} + \boldsymbol{c}, \quad \boldsymbol{\eta}(t) = t\boldsymbol{b} + \boldsymbol{c}', \tag{26}$$

where $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{c}'$ are constant vectors.

Convex function and Legendre duality

A dually flat manifold M has two convex functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$. Given a convex function $\psi(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ in M, the Legendre transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$ is

$$\eta_i = \partial_i \psi(\boldsymbol{\theta}). \tag{27}$$

There exists a dual potential $\varphi(\boldsymbol{\eta})$ which is convex with respect to $\boldsymbol{\eta}$ and the inverse transformation is given by

$$\theta^i = \partial^i \varphi(\boldsymbol{\eta}). \tag{28}$$

The two potential functions can be chosen to satisfy

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \sum \theta^i \eta_i = 0. \tag{29}$$

This gives the coordinate transformation between the primal affine coordinates $\boldsymbol{\theta}$ and the dual affine coordinates $\boldsymbol{\eta}$ on a dually flat manifold M.

We can define the canonical divergence between two points $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ (or $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$) by using the potential function.

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi(\boldsymbol{\theta}) - \psi\left(\boldsymbol{\theta}'\right) - \sum \partial_i \psi\left(\boldsymbol{\theta}'\right)\left(\theta^i - \theta'^i\right). \tag{30}$$

This is known as the Bregman divergence [Bregman, 1967] and is written in the dual form as

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi(\boldsymbol{\theta}) + \varphi\left(\boldsymbol{\eta}'\right) - \sum \theta^i \eta'_i. \tag{31}$$

When a convex function $\psi(\boldsymbol{\theta})$ is given on M, a dually flat structure is constructed on M and conversely a dually flat M possesses an affine coordinate system $\boldsymbol{\theta}$, a convex potential $\psi(\boldsymbol{\theta})$ and the canonical divergence $D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]$.

Generalized Pythagorean theorem and projection theorem in flat M

The generalized Pythagorean theorem holds in a dually flat manifold M. Let P, Q, and R be three points in M.

Theorem 4. When the primal geodesic connecting Q and R is orthogonal in the sense of Riemannian metric g to the dual geodesic connecting P and Q, then

$$D[P : Q] + D[Q : R] = D[P : R] \tag{32}$$

and when the dual geodesic connecting Q and R is orthogonal to the geodesic connecting P and Q, then

$$D[Q : P] + D[R : Q] = D[R : P]. \tag{33}$$

The generalized Pythagorean theorem is a natural extension of that in a Euclidean space. Indeed, a Euclidean space is a special case of the dually flat manifold, where $T_{ijk} = 0$. In such a case, a primal geodesic and a dual geodesic are the same because the two coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are the same, and g_{ij} reduces to δ_{ij} (identity matrix) in this coordinate system. The potential functions are written as

$$\psi(\boldsymbol{\theta}) = \varphi(\boldsymbol{\eta}) = \frac{1}{2}\sum \left(\theta^i\right)^2 = \frac{1}{2}\sum \left(\eta_i\right)^2 \tag{34}$$

and the divergence is

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \frac{1}{2}\sum \left(\theta^i - \theta'^i\right)^2. \tag{35}$$

Therefore, theorem 4 is a natural extension of the Pythagorean theorem.

The projection theorem is a direct consequence of the Pythagorean theorem, which has many applications in real problems. See, e.g., [Boissonnat, Nielsen and Nock, 2010], [Takenouchi et al., 2008], [Ikeda, Tanaka and Amari, 2004], [Amari et al., 2003], etc.

Theorem 5. Let S be a submanifold in a dually flat manifold M. Given a point P, the point $\hat{P} \in S$ that is closest to P in the sense of minimizing divergences $D[P:Q]$, $Q \in S$ is called the geodesic projection of P to S. The dual geodesic projection $\hat{P}^*$ of P to S is the point that minimize $D[Q:P]$, $Q \in S$. The geodesic projection $\hat{P}$ (dual geodesic projection $\hat{P}^*$) is given by the point that satisfies the following: The geodesic (dual geodesic) connecting P and $\hat{P}$ (P and $\hat{P}^*$) is orthogonal to S.

Invariant divergence in the manifold of probability distributions

Let us consider a manifold of probability distributions. We consider the discrete case where random variable x takes values on finite set $X = \{1, 2, \cdots, n\}$. Then the set of all the probability distributions

$$S_n = \{p(x)\} \tag{36}$$

forms an $(n-1)$-dimensional manifold, called the probability $(n-1)$-simplex. We may write

$$p_i = \text{Prob}\,\{x = i\}\,. \tag{37}$$

Then the probability simplex is specified by $\boldsymbol{p} = (p_i)$, where

$$\sum p_i = 1, \quad p_i > 0. \tag{38}$$

We introduce a divergence $D[\boldsymbol{p}:\boldsymbol{q}]$ between two distributions $\boldsymbol{p}$ and $\boldsymbol{q} \in S_n$ that satisfies the following invariance criterion [Chentsov, 1982], [Csiszar, 1991]. For this purpose, we partition X into disjoint subsets $X_1, \cdots, X_m$,

$$X = \{X_1, \cdots, X_m\}, \quad \cup X_i = X, \quad X_i \cap X_j = \phi. \tag{39}$$

Then, the partition naturally induces a reduced probability distribution $\bar{\boldsymbol{p}}$ on $\bar{X} = \{X_1, \cdots, X_m\}$ such that

$$\bar{p}_i = \sum_{j \in X_i} p_j. \tag{40}$$

This is a coarse graining of observation of x such that we do not know x but know the subclass X_i to which x belongs, so that there is loss of information. A loss is expressed in terms of the divergence as follows.

Invariance Criterion: A divergence is said to be monotone when

$$D[\boldsymbol{p}:\boldsymbol{q}] \geq D\left[\bar{\boldsymbol{p}}:\bar{\boldsymbol{q}}\right] \tag{41}$$

holds for any partition of X. Moreover, it is said to be invariant, when

$$D[\boldsymbol{p}:\boldsymbol{q}] = D\left[\bar{\boldsymbol{p}}:\bar{\boldsymbol{q}}\right] \tag{42}$$

holds, if and only if the conditional probabilities of x conditioned on any X_i are equal for both $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$p\left(x|X_i\right) = q\left(x|X_i\right), \quad i = 1, \cdots, m \tag{43}$$

or

$$p_i = c_j q_i, \quad i \in X_j \tag{44}$$

for constant c_j.

A divergence is said to be decomposable, when it is written as

$$D[\boldsymbol{p}:\boldsymbol{q}] = \sum d\left(p_i, q_i\right) \tag{45}$$

for some function d. The following theorem is known [Amari, 2009], [Amari and Nagaoka, 2000].

Theorem 6. The invariant divergence that gives dually flat structure is unique in S_n and is given by the Kullback-Leibler (KL) divergence.

$$KL\left[\boldsymbol{p} : \boldsymbol{q}\right] = \sum p_i \log \frac{p_i}{q_i}. \tag{46}$$

Theorem 7. The invariant Riemannian metric of S_n is given by the Fisher information matrix,

$$g_{ij}(\boldsymbol{p}) = \sum_x p(x) \frac{\partial \log p(x)}{\partial p_i} \frac{\partial \log p(x)}{\partial p_j}, \quad i, j = 1, \cdots, n-1. \tag{47}$$

The invariant third-order tensor is given by

$$T_{ijk}(\boldsymbol{p}) = \sum_x p(x) \frac{\partial \log p(x)}{\partial p_i} \frac{\partial \log p(x)}{\partial p_j} \frac{\partial \log p(x)}{\partial p_k}. \tag{48}$$

Theorem 8. S_n is a dually flat manifold and the $\boldsymbol{\theta}$ coordinates are

$$\theta^i = \log \frac{p_i}{p_n}, \quad i = 1, \cdots, n-1 \tag{49}$$

the dual $\boldsymbol{\eta}$ coordinates are

$$\eta_i = p_i, \quad i = 1, \cdots, n-1, \tag{50}$$

the potential function is

$$\psi(\boldsymbol{\theta}) = \log\left\{1 + \sum \exp\left(\theta^i\right)\right\}, \tag{51}$$

which is the cumulant generating function, the dual potential function is

$$\varphi(\boldsymbol{\eta}) = \sum \eta_i \log \eta_i + \left(1 - \sum \eta_i\right) \log\left(1 - \sum \eta_i\right), \tag{52}$$

which is the negative of Shannon entropy, and the canonical divergence is the KL divergence.

Information geometry is applied to statistical inference, in particular to higher-order asymptotic theory of statistical inference [Amari, 1985]. For example, consider a statistical model $M = \{p(x, \boldsymbol{u})\} \subset S_n$ specified by parameter $\boldsymbol{u} = (u_1, \cdots, u_m)$, where m is smaller than n. When x is observed N times, the observation defines the empirical distribution $\hat{\boldsymbol{p}}$,

$$\hat{p}_i = \frac{1}{N} \sharp \{x = i\}. \tag{53}$$

The maximum likelihood estimator $\hat{\boldsymbol{u}}$ is given by the geodesic projection of $\hat{\boldsymbol{p}}$ to the statistical model $M = \{p(x, \boldsymbol{u})\}$ in S_n. The error of estimation is evaluated by the Fisher information (Cramér-Rao theorem and by the embedding curvature of M in S_n). Information geometry can also be applied to semiparametric statistical inference, where a theory of estimating functions is established [Amari and Kawanabe, 1997]. We use a fiber-bundle-like structure in this case, when x is a continuous variable and the manifold is infinite-dimensional. However, we should be careful for mathematical difficulties in generalizing the above theory of the discrete case to an infinite-dimensional case. There are lots of applications of information geometry to optimization, machine learning, computer vision and neural networks.

Divergence introduced in the space of positive measures

Let us consider the set $\boldsymbol{R}_+^n$ of positive measures on X, where $m(x)$ gives a measure of $x \in X$. By introducing the delta function $\delta_i(x)$,

$$\delta_i(x) = \begin{cases} 1, & x = i \\ 0, & \text{otherwise,} \end{cases} \tag{54}$$

we can write

$$m(x, \boldsymbol{z}) = \sum z_i \delta_i(x), \tag{55}$$

where $z_i > 0$. $\boldsymbol{R}_n^+$ is a manifold where $\boldsymbol{z}$ is a coordinate system. The probability simplex S_n is its submanifold satisfying

$$\sum z_i = 1. \tag{56}$$

Let $u(z)$ and $v(z)$ be two differentiable and monotonically increasing functions satisfying

$$u(0) = v(0) = 0. \tag{57}$$

Define two coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ by

$$\theta^i = u\left(z_i\right), \tag{58}$$

$$\eta_i = v\left(z_i\right), \tag{59}$$

which are non-linear rescalings of z_i. We introduce the dually flat structure to $\boldsymbol{R}_+^n$ such that $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are dually flat affine coordinate systems. The structure is called the (u, v)-structure, since it is defined by using two functions u and v. Note that the dual invariant affine coordinates $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ of (49) and (50) in S_n are given by

$$u(z) = \log z, \tag{60}$$

$$v(z) = z, \tag{61}$$

within the constraint of (56). The following theorem is given by [Amari, Ohara and Matsuzoe, 2012].

Theorem 9. The potential functions of the (u, v)-structure are given by

$$\psi(\boldsymbol{\theta}) = \sum \int d\theta_i \int \frac{v'\left\{u^{-1}\left(\theta_i\right)\right\}}{u'\left\{u^{-1}\left(\theta_i\right)\right\}} d\theta_i, \tag{62}$$

$$\varphi(\boldsymbol{\eta}) = \sum \int d\eta_i \int \frac{u'\left\{v^{-1}\left(\eta_i\right)\right\}}{v'\left\{v^{-1}\left(\eta_i\right)\right\}} d\eta_i. \tag{63}$$

Proof. Since the two coordinates are connected by Legendre transformations

$$\eta_i = \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta^i}, \quad \theta^i = \frac{\partial \varphi(\boldsymbol{\eta})}{\partial \eta_i}, \tag{64}$$

by integrating them, we have (62) and (63).

Theorem 10. The (u, v)-structure of $\boldsymbol{R}_n^+$ gives a Riemannian metric

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}) \delta_{ij}, \tag{65}$$

and a cubic tensor

$$T_{ijk}(\boldsymbol{\theta}) = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}) \delta_{ijk} \tag{66}$$

which includes only diagonal components since δ_{ij} and δ_{ijk} are 1 when $i = j$ and $i = j = k$, respectively, and 0 otherwise. The metric is Euclidean so that the curvature due to the Levi-Civita connection vanishes.

There are lots of applications of the (u, v)-divergences, in particular in the following form. The (u, v)-structure gives a dually flat affine structure to the Euclidean manifold $\boldsymbol{R}_n^+$ by rescaling the axes. We give examples of (u, v)-structures. See [Cichocki, Cruces and Amari, 2011].

1. Logarithmic structure

When

$$u(z) = \log z, \quad v(z) = z, \tag{67}$$

we have

$$\theta^i = \log z_i, \quad \eta_i = z_i \tag{68}$$

so that

$$m(x, \boldsymbol{z}) = \exp\left\{\sum \theta^i \delta_i(x)\right\}. \tag{69}$$

The potential functions are

$$\psi(\boldsymbol{\theta}) = \sum e^{\theta_i}, \quad \varphi(\boldsymbol{\eta}) = \sum \eta_i \log \eta_i - \sum \eta_i \tag{70}$$

and the canonical divergence is

$$D\left[\boldsymbol{z} : \boldsymbol{z}'\right] = \sum \left(z_i - z_i' + z_i \log \frac{z_i}{z_i'}\right), \tag{71}$$

which is the generalized KL divergence. The probability simplex S_n is the linear subspace given by

$$\sum \eta_i = 1 \tag{72}$$

and therefore is also dually flat. This gives the invariant structure satisfying the invariance criterion.

2. (α, β)**-structure** [Cichocki, Cruces and Amari, 2011]

Define

$$u(z) = z^{\alpha}, \quad v(z) = z^{\beta} \tag{73}$$

for two real parameters α and β. Then

$$\theta^i = (z_i)^{\alpha}, \quad \eta_i = (z_i)^{\beta} \tag{74}$$

and the potential functions are

$$\psi(\boldsymbol{\theta}) = \frac{\alpha}{\alpha+\beta} \sum \left(\theta^i\right)^{\frac{\alpha+\beta}{\alpha}}, \quad \varphi(\boldsymbol{\eta}) = \frac{\beta}{\alpha+\beta} \sum \left(\eta_i\right)^{\frac{\alpha+\beta}{\beta}}. \tag{75}$$

The divergence, named the (α, β)-divergence, is given by

$$D_{\alpha,\beta}\left[\boldsymbol{z} : \boldsymbol{z}'\right] = \sum \left\{\frac{\alpha}{\alpha+\beta} (z_i)^{\alpha+\beta} + \frac{\beta}{\alpha+\beta} \left(z_i'\right)^{\alpha+\beta} - (z_i)^{\alpha} \left(z_i'\right)^{\beta}\right\}. \tag{76}$$

The probability simplex S_n is a subspace, but the derived structure is neither invariant nor dually flat in general.

It should be remarked that, taking limit $\alpha \to 0$ carefully and putting $\beta = 1$, we have

$$\theta^i = \log z^i, \quad \eta_i = z_i. \tag{77}$$

Hence, $(0, 1)$-structure is the logarithmic structure.

3. α**-structure (**q**-structure)** [Amari and Nagaoka, 2000; Amari, 2007]

When $\alpha + \beta = 1$, we replace α by $1 - q = (1 - \tilde{\alpha})/2$ and β by $q = (1 + \tilde{\alpha})/2$, having

$$u(z) = z^{\frac{1-\tilde{\alpha}}{2}} = z^{1-q}, \quad v(z) = z^{\frac{1+\tilde{\alpha}}{2}} = z^q. \tag{78}$$

The structure is called the $\tilde{\alpha}$-structure or q-structure, where $\tilde{\alpha} = 1 - 2q$. We simply replace $\tilde{\alpha}$ by α, omitting $\tilde{}$. The α-structure is used in information geometry, but the same structure is used in non-extensive statistical mechanics by [Tsallis, 2009] and others under the name of q-entropy, etc.

Theorem 11. The $\alpha(q)$-affine coordinates are given by

$$\theta^i = (z_i)^{\frac{1-\alpha}{2}} = (z_i)^{1-q}\ ,\ ;\ \eta_i = (z_i)^{\frac{1+\alpha}{2}} = (z_i)^q \tag{79}$$

with potential functions

$$\psi(\boldsymbol{\theta}) = \frac{1-\alpha}{2}\sum (\theta_i)^{\frac{2}{1-\alpha}} = \frac{1-\alpha}{2}\sum z_i\ ; \tag{80}$$

$$\varphi(\boldsymbol{\eta}) = \frac{1+\alpha}{2}\sum (\eta_i)^{\frac{2}{1+\alpha}} = \frac{1+\alpha}{2}\sum z_i. \tag{81}$$

The α-divergence is defined by

$$D_\alpha\left[\boldsymbol{z}:\boldsymbol{z}'\right] = \sum\left\{\frac{1-\alpha}{2}z_i + \frac{1+\alpha}{2}z_i - (z_i)^{\frac{1-\alpha}{2}}(z_i)^{\frac{1+\alpha}{2}}\right\}. \tag{82}$$

The α-structure introduced in the probability simplex S_n is not dually flat. However, the α-divergence is given in S_n by

$$D_\alpha\left[\boldsymbol{p}:\boldsymbol{p}'\right] = \sum\left\{1 - (p_i)^{\frac{1-\alpha}{2}}\left(p_i'\right)^{\frac{1+\alpha}{2}}\right\}. \tag{83}$$

This is an invariant divergence. We extend the invariance principle to be applicable to $\boldsymbol{R}_+^n$. Then we have the following theorem [Amari, 2009].

Theorem 12. The α-divergence is the unique class of invariant divergence that gives dually flat structure to $\boldsymbol{R}_+^n$.

q exponential family and deformed exponential family

An exponential family S of probability distributions is a statistical model defined by

$$S = \left\{p(\boldsymbol{x},\boldsymbol{\theta}) = \exp\left[\sum \theta^i x_i - \psi(\boldsymbol{\theta})\right]\right\}, \tag{84}$$

where $\boldsymbol{x}$ is a vector random variable and $\boldsymbol{\theta} = \left(\theta^i\right)$ is called the natural parameters to specify a distribution. The invariant geometry introduced in S is dually flat, where $\boldsymbol{\theta}$ is the affine coordinate system, $\psi(\boldsymbol{\theta})$ is the potential function and the dual affine coordinates are given by

$$\eta_i = \partial_i\psi(\boldsymbol{\theta}) = \int x_i p(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}. \tag{85}$$

This is called the expectation parameter. The dual potential is the negative entropy,

$$\varphi(\boldsymbol{\eta}) = \int p(\boldsymbol{x},\boldsymbol{\theta})\log p(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}. \tag{86}$$

Instead of the logarithm, we introduce the q-logarithm (see [Tsallis, 2009], [Naudts, 2011]) defined by

$$\log_q z = \frac{1}{1-q}\left(z^{1-q} - 1\right). \tag{87}$$

Then, a family of probability distributions of the form

$$\log_q p(\boldsymbol{x},\boldsymbol{\theta}) = \sum \theta^i x_i - \psi(\boldsymbol{\theta}) \tag{88}$$

is called a q-exponential family.

More generally, we define χ-logarithm [Naudts, 2011] by

$$\log_\chi(z) = \int_1^z \frac{1}{\chi(t)} dt, \tag{89}$$

where $\chi(t)$ is a monotonically increasing positive function. When

$$\chi(t) = t, \tag{90}$$

this gives the logarithm and when

$$\chi(t) = t^q, \tag{91}$$

this gives the q-logarithm.

The inverse function of the χ-logarithm is the χ-exponential given by

$$\exp_\chi(z) = 1 + \int_0^z \lambda(t) dt \tag{92}$$

where $\lambda(t)$ is related to χ by

$$\lambda\left(\log_\chi z\right) = \chi(z). \tag{93}$$

A family of probability distributions is called a χ-exponential family or χ-family in short when they are written as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \chi\left\{\sum \theta^i x_i - \psi_\chi(\boldsymbol{\theta})\right\}. \tag{94}$$

with respect to dominating measure $\mu(\boldsymbol{x})$. The potential function $\psi_\chi(\boldsymbol{\theta})$ is determined from the normalization condition

$$\int p(\boldsymbol{x}, \boldsymbol{\theta}) d\mu(\boldsymbol{x}) = 1. \tag{95}$$

We may call $\psi_\chi(\boldsymbol{\theta})$ the χ-free energy. We can prove that it is a convex function [Amari, Ohara and Matsuzoe, 2012].

The probability simplex S_n is a χ-exponential family for any χ, since any distribution $p(x)$ on X is written as

$$\log_\chi p(x) = \sum_{i=1}^{n-1} \theta^i \delta_i(x) + \log_\chi\left(1 - \sum p_i\right), \tag{96}$$

where

$$\begin{aligned} \theta^i &= \log_\chi p_i - \log_\chi p_n & (97)\\ x_i &= \delta_i(x), \quad i = 1, \cdots, n-1. & (98) \end{aligned}$$

We can introduce the geometrical structure (the metric and dual affine connections) from the invariance principle for any χ-family. However, except for exponential families and mixture families, the induced structure is not dually flat (except for the 1-dimensional case where the curvature always vanishes).

Instead, we can introduce a dually flat χ-structure to a χ-family, which in general is different from the invariant structure. In particular, the probability simplex S_n has χ-dually flat structure for any χ, which is different from the invariant structure in general.

By using the χ-free energy or χ-potential function $\psi_\chi(\boldsymbol{\theta})$, we define the χ-metric and χ-cubic tensor by

$$\begin{aligned} g_{ij}^\chi &= \partial_i \partial_j \psi_\chi(\boldsymbol{\theta}), & (99)\\ T_{ijk}^\chi &= \partial_i \partial_j \partial_k \psi_\chi(\boldsymbol{\theta}). & (100) \end{aligned}$$

They give a dually flat geometrical structure. The dual affine coordinates are given by

$$\eta_i = \partial_i \psi_\chi(\boldsymbol{\theta}). \tag{101}$$

The dual potential function in S_n is given by

$$\varphi(\boldsymbol{\eta}) = \frac{1}{h_\chi} \sum u' \left\{ v\left(p_i\right) \right\} v\left(p_i\right). \tag{102}$$

In the case of the q-family in S_n, we have

$$\varphi_q(\boldsymbol{\eta}) = \frac{1}{1-q} \left(\frac{1}{h_q} - 1 \right). \tag{103}$$

Therefore, it is natural to define the q-entropy by

$$H_q(\boldsymbol{p}) = \frac{1}{h_q(\boldsymbol{p})} \tag{104}$$

up to a scale and constant. This is different from the Tsallis q-entropy defined by

$$H_{\text{Tsallis}} = -h_q(\boldsymbol{p}). \tag{105}$$

From the geometrical point of view, our definition of the q-entropy is natural.

We finally study how the χ-metric is related to the invariant Fisher metric in S_n [Amari and Ohara, 2011]. The following is an interesting observation, connecting q-geometry and conformal geometry [Kurose, 1994]. When a metric g_{ij} is changed to

$$\tilde{g}_{ij}(\boldsymbol{p}) = \sigma(\boldsymbol{p}) g_{ij}(\boldsymbol{p}), \tag{106}$$

where $\sigma(\boldsymbol{p})$ is a positive scalar function in a Riemannian manifold, the transformation is said to be conformal. In the case of $\{M, g, T\}$, a conformal transformation changes T_{ijk} as

$$\tilde{T}_{ijk}(\boldsymbol{p}) = \sigma(\boldsymbol{p}) T_{ijk} + \left(\partial_i \sigma\right) g_{jk} + \left(\partial_j \sigma\right) g_{ik} + \left(\partial_k \sigma\right) g_{ij}. \tag{107}$$

Theorem 13. The q-geometry is the unique class of probability distributions that is conformally connected to the invariant geometry with Fisher information metric. The conformal factor is given by

$$\sigma(\boldsymbol{p}) = \frac{1}{h_q(\boldsymbol{p})}. \tag{108}$$

Conclusion

We have reviewed the current status of information geometry which emerged from the study of invariant geometrical structure of the manifold of probability distributions. The structure is related to the divergence function and hence is regarded as the geometry of divergence. It consists of a Riemannian metric together with a pair of dual affine connections. A dually flat Riemannian manifold is of particular interest in applications. We have given its mathematical structure and showed that it gives the canonical divergence in the form of the Bregman divergence and vice versa.

We also show the invariance principle to be applied to the manifold of probability distributions. The Kullback-Leibler divergence is its canonical divergence. We also show various types of divergence functions which give the dually flat structure. The Tsallis entropy and the deformed exponential family arising from it are studied in detail. We have proved that the q-structure is the unique class that is derived from conformal transformation of the invariant geometry having the Fisher information metric.

It is natural to use the geometry derived from the invariance principle, when we study statistical inference. However, the invariance geometry is given by α-geometry including a free parameter α. We usually treat the case of $\alpha = \pm 1$. The $\alpha = 0$ case reduces to the Riemannian geometry. There are interesting applications using other α [Matsuyama, 2002]. The invariance principle is applicable only to a manifold of probability distributions. Hence, we can construct a dually flat geometry in many problems such as vision and machine learning, when a convex function is used. This widens the applicability of information geometry.

Finally, we point out its extension. The extension to the infinite-dimensional function space is studied by [Cena and Pistone, 2007]. Its extension to Finsler geometry and Wasserstein geometry is also expected in future.

Bibliography

[Amari, 1985] S. Amari. Differential-Geometrical Methods in Statistics. Springer Lecture Notes in Statistics, 28, 1985

[Amari, 2007] S. Amari. Integration of Stochastic Models by Minimizing α-Divergence. Neural Computation, 19, 10, 2780–2796, 2007.

[Amari, 2009] S. Amari. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Transactions on Information Theory, 55, 11, 4925–4931, 2009.

[Amari and Cichocki, 2010] S. Amari and A. Cichocki. Information geometry of divergence functions. Bulletin of the polish academy of sciences technical sciences, 58, 1, 183–195, 2010.

[Amari and Kawanabe, 1997] S. Amari and M. Kawanabe. Information geometry of estimating functions in semi-parametric statistical models. Bernoulli, 3, 29–54, 1997.

[Amari and Nagaoka, 2000] S. Amari and H. Nagaoka. Methods of Information Geometry. Translations of Mathematical Monographs, 191, Oxford University Press, 2000.

[Amari et al., 2003] S. Amari, H. Nakahara, S. Wu and Y. Sakai. Synchronous Firing and Higher-Order Interactions in Neuron Pool. Neural Computation, 15, 127–142, 2003.

[Amari and Ohara, 2011] S. Amari and A. Ohara. Geometry of q-exponential family of probability distributions. Entropy, 13, 1170–1185; doi:10.3390/e13061170, 2011.

[Amari, Ohara and Matsuzoe, 2012] S. Amari, A. Ohara and H. Matsuzoe. Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries. Physica A, accepted for publication.

[Banerjee et al., 2005] A. Banerjee, S. Merugu, I. S. Dhillon and J. Gosh. Clustering with Bregman divergences. J. Machine Learning Research, 6, 1705–1749, 2005.

[Boissonnat, Nielsen and Nock, 2010] J.-D. Boissonnat, F. Nielsen and R. Nock. Bregman Voronoi diagrams. Discrete and Computational Geometry, 44, 281–307, 2010.

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

[Cena and Pistone, 2007] A. Cena and G. Pistone. Exponential statistical manifold. Annals of Institute of Statist. Math., 59, 27–56, 2007.

[Chen, Chen and Rao, 2008] P. Chen, Y. Chen and M. Rao. Metrics defined by Bregman divergences: Part 2. Commun. Math. Sci., 6, 927–948, 2008.

[Chentsov, 1982] N. N. Chentsov. Statistical Decision Rules and Optimal Inference. AMS Translation Series, 1982.

[Cichocki, Cruces and Amari, 2011] A. Cichocki, S. Cruces and S. Amari. Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. Entropy, 13, 134–170, 2011.

[Cichocki et al., 2009] A. Cichocki, R. Zdunek, A. H. Phan and S. Amari. Nonnegative Matrix and Tensor Factorizations. John Wiley and Sons, U.K., 2009.

[Csiszar, 1991] I. Csiszar. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Annals of Statistics, 19, 2032–2066, 1991.

[Eguchi, 1983] S. Eguchi. Second-order efficiency of minimum contrast estimators in a curved exponential family. Annals of Statistics, 11, 793–803, 1983.

[Ikeda, Tanaka and Amari, 2004] S. Ikeda, T. Tanaka and S. Amari. Information Geometry of Turbo and Low-Density Parity-Check Codes. IEEE Transactions on Information Theory, 50, 6, 1097–1114, 2004.

[Kurose, 1994] T. Kurose. On the divergence of 1-conformally flat statistical manifold. Tohoku Math. J., 46, 427–433, 1994.

[Liu et al, 2010] M. Liu, B. C. Vemuri, S. Amari and F. Nielsen. Total Bregman divergence and its applications to shape retrieval. IEEE CVPR, 3463–3468, 2010.

[Matsuyama, 2002] Y. Matsuyama. The α-EM algorithm: Surrogate likelihood maximization using αlogarithmic information measure. IEEE Trans. Information Theory, 49, 672–706, 2002.

[Naudts, 2011] J. Naudts. Generalized Thermostatistics, Springer, 2011.

[Picard, 1992] D. Picard. Statistical morphisms and related invariance properties. Annals of the Institute of Statistical Mathematics, 44, 45–61, 1992.

[Rao, 1945] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. Bulletin of Calcutta Mathematical Society, 37, 81–91, 1945.

[Takenouchi et al., 2008] T. Takenouchi, S. Eguchi, N. Murata and T. Kanamori. Robust boosting algorithm against mislabeling in multi-class problems. Neural Computation, 20, 6, 1596–1630, 2008

[Tsallis, 2009] C. Tsallis. Introduction to Non-Extensive Statistical Mechanics: Approaching a Complex World. Springer, 2009.

[Vemuri et al., 2011] B. C. Vemuri, M. Lie, S. Amari and F. Nielsen. Total Bregman divergence and its applications to DTI analysis. IEEE Trans. on Medical Imaging, 30, 475–483, 2011.

[Vos, 1991] P. W. Vos. Geometry of f-divergence. Annals of the Institute of Statistical Mathematics, 43, 515–537, 1991.

Authors' Information

Shun-ichi Amari *- Senior Advisor, RIKEN Brain Science Institute, Hirosawa 2-1, Wako-shi, Saitama 351-0198, Japan; e-mail: amari@brain.riken.jp*

Major Fields of Scientific Research: Information geometry, mathematical neuroscience, machine learning, statistics