Contemporary Mathematics
Volume 203, 1997

Information Geometry

S.-I. AMARI

ABSTRACT. Information geometry has emerged from the study of invariant properties of the manifold of probability distributions. It proposes a Riemannian manifold with two torsion-free dually coupled affine connections. This is called the dual structure. Affine differential geometry and Kählerian manifolds give special cases of the dual structure and statistical models give more general cases. A dually flat manifold has special properties such as dual orthogonal foliations, generalized Pythagoras theorem, the invariant divergence, and the Legendre structure. The present paper reviews the mathematical structure of information geometry and some unsolved mathematical problems are shown. Information geometry is regarded as mathematical sciences having vast developing areas of applications as well as givining new ideas to pure mathematics. The paper touches upon such applications, too.

CONTENTS

1. Introduction
2. Riemannian Manifold with Dual Affine Connections
3. Dually flat manifolds
4. Some Unsolved Mathematical Questions
5. Manifold of Probability Distributions
6. Applications of Information Geometry
7. Conclusions
References

1. Introduction

Information geometry originated from the study of the manifold of probability distributions. Consider the set of all the Gaussian distributions as a simple example. Any Gaussian distribution is specified by the mean value μ and variance σ^2, so that it is regarded as a two-dimensional manifold where (μ, σ^2) is a coordinate system. What is the *natural* geometric structure to be introduced in such a manifold? Information geometry gives a definite answer to this problem, and an invariant geometrical structure can be introduced into the manifold consisting

1991 *Mathematics Subject Classification.* Primary 53A15, 62B10; Secondary 95A17.

of a smooth family of general probability distributions. Information geometry is useful not only for applications to statistics, information theory, systems theory where probability distributions play a major role, but also gives a new concept to differential geometry.

Let us consider a manifold M with a Riemannian metric g, $\{M, g\}$. Let $\nabla^{(0)}$ be the covariant derivative due to the Levi-Civita connection. For vector fields X, Y, Z, it satisfies the metric condition

$$X\langle Y, Z\rangle = \langle \nabla^{(0)}_X Y, Z\rangle + \langle Y, \nabla^{(0)}_X Z\rangle, \tag{1.1}$$

where $\langle \cdot, \cdot \rangle$ is the inner product due to g. This is the Riemannian structure. It was proved that the manifold of probability distributions has a unique invariant Riemannian metric. Moreover, it has a pair of torsion-free affine connections or covariant derivatives ∇ and ∇^* which satisfy the duality condition,

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle. \tag{1.2}$$

When the Riemann-Christoffel curvatures R and R^* due to ∇ and ∇^* vanish (but $R^{(0)}$ does not necessarily vanish), the manifold is said to be dually flat. The dually flat manifold has rich structures related to the Legendre transformation and a generalized Pythagoras theorem holds with respect to the invariant divergence. Information geometry gives a powerful method to information sciences and physics based on the dual structure of affine connections.

It is rather surprising that the dual structure has not studied sufficiently from the point of view of pure mathematics. It originated from statistics and information sciences (see Rao [1945], Chentsov [1972], Amari [1985], Murray and Rice [1993], Amari and Nagaoka [1994]). However, a special case of such a structure originated from the theory of hypersurfaces in affine differential geometry (Blaschke [1923], Nomizu and Sasaki [1994]). Another special case was proposed by Shima [1980; 1986; 1995] as the Hessian manifolds which are real manifolds keeping some properties of Kählerian structures.

The present paper reviews the geometry of the dual structure and presents mathematical problems to be solved further. The paper also touches upon wide ranges of applications of information geometry.

2. Riemannian Manifold with Dual Affine Connections

Let $\{S, g, \nabla\}$ be a structure on an n-dimensional manifold S, with Riemannian metric g and affine connection ∇. Let T_p be the tangent space at point $p \in S$. For tangent vectors $X_p, Y_p \in T_p$, the Riemannian metric defines their inner product $g : T_p \times T_p \to \mathrm{R}$,

$$g(X_p, Y_p) = \langle X_p, Y_p\rangle. \tag{2.1}$$

The affine connection is represented by the covariant derivative $\nabla : \mathcal{T} \times \mathcal{T} \to \mathcal{T}$, where $\mathcal{T}$ is the set of all tangent vector fields. The covariant derivative of vector field Y in the direction X is denoted by $\nabla_X Y$. An affine connection is said to be metric when it satisfies

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle. \tag{2.2}$$

The Levi-Civita connection is the only torsion-free metric connection. In this case, $\{S, g, \nabla\}$ is called the Riemannian structure.

Let ξ^i, $i = 1, \cdots, n$, be a local coordinate system of an n-dimensional manifold S. It defines the natural basis vector fields denoted by

$$\mathrm{e}_i = \frac{\partial}{\partial \xi^i} = \partial_i.$$

The metric tensor is given in this coordinate system ∇ by

$$g_{ij} = \langle \mathrm{e}_i, \mathrm{e}_j \rangle \tag{2.3}$$

and the parameters of the affine connection by

$$\Gamma_{ijk} = \langle \nabla_{\mathrm{e}_i} \mathrm{e}_j, \mathrm{e}_k \rangle \tag{2.4}$$

in this coordinate system.

Let us consider a manifold S which has a Riemannian metric g and two affine connections ∇ and ∇^*. This gives a structure $\{S, g, \nabla, \nabla^*\}$. The two affine connections are said to be dual or conjugate with respect to the metric, when

$$X\langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla^*_X Z \rangle \tag{2.5}$$

holds for any vector fields X, Y and Z. Refer to Amari [1985], Murray and Rice [1993], Amari and Nagaoka [1994] for the details of the following discussions. By substituting the natural vector fields e_i, e_j, e_k for X, Y and Z, we have

$$\partial_i g_{jk}(\xi) = \Gamma_{ijk}(\xi) + \Gamma^*_{ikj}(\xi) \tag{2.6}$$

for the duality of ∇ and ∇^* in the component form. When $\nabla = \nabla^*$, they are self-dual and (2.5) reduces to the metric condition (2.2). Therefore, the Riemannian geometry $\{S, g, \nabla\}$ with the Levi-Civita connection is the self-dual case of the general dual Riemannian structure.

In general,

$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk} \tag{2.7}$$

is a tensor. It is easy to prove

$$T_{ijk} = \nabla_{\mathrm{e}_i} \langle \mathrm{e}_j, \mathrm{e}_k \rangle = \nabla_i g_{jk}. \tag{2.8}$$

In particular, when both ∇ and ∇^* are torsion-free, T_{ijk} is a symmetric tensor. Hence, the torsion-free dual structure is characterized by $\{S, g, T\}$, where g and T are symmetric tensors. On the other hand, given symmetric g and T, we can define, for a scalar parameter α, a family of torsion-free connections by

$$\Gamma^{(\alpha)}_{ijk} = \{ij; k\} - \frac{\alpha}{2} T_{ijk}, \tag{2.9}$$

where

$$\Gamma^{(0)}_{ijk} = \{ij; k\} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}) \tag{2.10}$$

is the Levi-Civita connection. In this case, for any α, $\{S, g, \nabla^\alpha, \nabla^{-\alpha}\}$ defines the dual structure called the α-structure. The Riemannian structure is the self-dual case with $\alpha = 0$.

The duality can be expressed in the following integral form. Let c be a curve connecting two points p and q. Let Π_c and Π^*_c be the parallel transport operators of vectors from T_p to T_q along c by ∇ and ∇^*, respectively. Then,

$$\langle X, Y \rangle_p = \langle \Pi_c X,\ \Pi^*_c Y \rangle_q \tag{2.11}$$

is locally equivalent to (2.5), where $\langle\ ,\ \rangle_p$ is the inner product at p.

When ∇ and ∇^* are dual, their Riemann-Christoffel curvatures R and R^* are related to each other by the duality. When R vanishes, R^* automatically vanishes and vice versa.

THEOREM 2.1. *For the α-structure, its $\pm\alpha$-curvatures satisfy*

$$R^{(\alpha)}_{ijkl} = -R^{(-\alpha)}_{ijlk}, \tag{2.12}$$

where

$$R^{(\alpha)}_{ijkl} = R(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k, \mathbf{e}_l)$$

is the Riemann-Christoffel curvature tensor.

The dual structure can be extended to any linear fibre bundles with inner product. Moreover, it can be incorporated with the conformal structure. The structure $\{S, \tilde{g}, \tilde{T}\}$ is said to be conformally equivalent to $\{S, g, T\}$ (Okamoto, Amari and Takeuchi [1991]), when there exists a positive scalar field σ such that

$$\tilde{g}_{ij} = \sigma g_{ij} \tag{2.13}$$

$$\tilde{T}_{ijk} = \sigma(T_{ijk} + g_{ij}s_k + g_{ik}s_j + g_{jk}s_i), \tag{2.14}$$

where

$$s_i = \frac{\partial}{\partial \xi^i} \log \sigma. \tag{2.15}$$

The dual projective structure is studied by Burdet and Perrin [1996].

3. Dually flat manifolds

When $\{S, g, \nabla, \nabla^*\}$ is torsion- and curvature-free, the manifold S is said to be dually flat. The Riemannian (Levi-Civita) curvature does not vanish in general. When it also vanishes, S is proved to be self-dual and flat, so that it reduces to the Euclidean space locally. Therefore, a dually flat manifold is a generalization of the Euclidean space. The dually flat structure inherits some of the Euclidean properties. See Nagaoka and Amari [1982], Amari [1995], Amari and Nagaoka [1994].

Since ∇ and ∇^* are flat, we have two local affine coordinate systems $\theta = (\theta^i)$ and $\eta = (\eta_i)$ such that the coordinate curves are ∇- and ∇^*-geodesics and the natural basis fields $\{\mathrm{e}_i\}$ and $\{\mathrm{e}^{*i}\}$ are ∇- and ∇^*-parallel, that is,

$$\nabla_{\mathrm{e}_i}\mathrm{e}_j = 0, \qquad \nabla_{\mathrm{e}^*_i}\mathrm{e}^*_j = 0, \tag{3.1}$$

respectively, where

$$\mathrm{e}_i = \frac{\partial}{\partial \theta^i}, \qquad \mathrm{e}^{*i} = \frac{\partial}{\partial \eta_i}. \tag{3.2}$$

They are determined to within affine transformations. We can choose a pair of ∇- and ∇^*-affine coordinates such that they are biorthogonal everywhere,

$$\langle \mathrm{e}_i, \mathrm{e}^{*j} \rangle = \delta^j_i. \tag{3.3}$$

The following theorem holds in this coordinate system.

THEOREM 3.1. *In a dually flat manifold, there exist two convex potential functions $\psi(\theta)$ and $\varphi(\eta)$ locally, such that the two coordinates are connected by the Legendre transformation,*

$$\psi(\theta) + \varphi(\eta) - \sum \theta^i \eta_i = 0 \tag{3.4}$$

$$\theta^i = \frac{\partial}{\partial \eta_i}\varphi(\eta), \quad \eta_i = \frac{\partial}{\partial \theta^i}\psi(\theta). \tag{3.5}$$

The metric tensors $g_{ij} = \langle \mathrm{e}_i, \mathrm{e}_j \rangle$ and $g^{ij} = \langle \mathrm{e}^{*i}, \mathrm{e}^{*j} \rangle$ are mutually inverse matrices and are given by

$$g_{ij} = \frac{\partial^2}{\partial \theta^i \partial \theta^j}\psi(\theta), \quad g^{ij} = \frac{\partial^2}{\partial \eta_i \partial \eta_j}\varphi(\eta). \tag{3.6}$$

On the contrary, when a differentiable convex function $\psi(\theta)$ exists in a manifold M with coodinates θ, we can introduce the following dually flat structure

$$\{M, g, \nabla, \nabla^*\}$$

given by

$$g_{ij}(\theta) = \frac{\partial^2}{\partial \theta^i \partial \theta^j}\psi(\theta),$$
$$\Gamma_{ijk}(\theta) = 0$$

and

$$\overset{*}{\Gamma}{}^{ijk}(\eta) = 0$$

in the ∇^*-affine coordinates

$$\eta_i = \frac{\partial}{\partial \theta^i}\psi(\theta).$$

Therefore, the dually flat structure is regarded as the natural geometry of the Legendre transformation.

A dually flat manifold has moreover the following structure which is not known by the theory of the Legendre transformation. Let us define a local function $D : S \times S \to \mathrm{R}$ in a dually flat S. This is a function $D(p : q)$ of two points p, q of S, and is defined by

$$D(p : q) = \psi(\theta_p) + \varphi(\eta_q) - \sum \theta_p^i \eta_{qi}, \tag{3.7}$$

where $\theta_p = (\theta_p^i)$ and $\eta_q = (\eta_{qi})$ are the ∇- and ∇^*-affine coordinates of p and q, respectively.

It is easy to prove that

$$D(p : q) \geq 0,$$

where the equality holds if $p = q$, and, when $q = p + dp$ is infinitesimally close to p, it is expanded as

$$D(p : p + dp) = \frac{1}{2}\sum g_{ij} d\theta^i d\theta^j.$$

Hence, D is a generalization of a half of the square of the Riemannian distance. The symmetry condition

$$D(p : q) = D(q : p)$$

does not in general hold nor the triangular inequality. The function D is called the divergence function. When the dually flat manifold is self-dual, it is Euclidean and the divergence reduces to a half of the squared Euclidean distance between p and q. It is symmetric in this case. The divergence in general satisfies the following generalized Pythagoras theorem.

THEOREM 3.2. *Let p, q and r be three points in a dually flat S such that the ∇^*-geodesic connecting p and q is orthogonal at q to the ∇-geodesic connecting q and r. Then,*

$$D(p : q) + D(q : r) = D(p : r). \tag{3.8}$$

An important consequence of this theorem is the following projection theorem.

THEOREM 3.3. *Let V be a closed subset of a dually flat S that has a smooth boundary hypersurface ∂V. Let p be a point outside V. Then, the point $\hat{p} \in V$ that minimizes $D(p : q)$, $q \in V$, is the ∇^*-geodesic projection of p to V, that is, when $D(p : \hat{p})$ is the minimum of $D(p : q)$, $q \in V$, $p \in \partial V$ and the ∇^*-geodesic connecting p and $\hat{p}$ is orthogonal to ∂V at $\hat{p}$.*

There is another property which is important for applications. For dual local coordinate systems $\theta = (\theta^i)$ and $\eta = (\eta_i)$, we partition the coordinates into two parts,

$$\theta = (\theta_1, \theta_2), \qquad \eta = (\eta_1, \eta_2)$$

where the θ_1 and η_1 parts have k-dimensions and the remaining θ_2 and η_2 have n - k dimensions. Let us consider a family of k-dimensional hyper-surfaces parametrized by n - k real parameters $c = (c_{k+1}, \cdots, c_n)$,

$$A^k(c) = \{\theta \mid \theta_2 = c\}$$

Then, A^k is a ∇-flat hypersurface, and $A^k(c)$'s form a (local) foliation of S,

$$A^k(c) \bigcap A^k(c') = \phi$$

when $c \neq c'$ and

$$\bigcup_c A^k(c) \supset S.$$

Dually to the above, we have another foliation,

$$B^{n-k}(d) = \{\eta \mid \eta_1 = d\}$$

where $d = (d_1, \cdots, d_k)$ is the parameter to specify $(n-k)$-dimensional submanifolds. Any $B^{n-k}(d)$ is ∇^*-flat.

THEOREM 3.4. *For a dual pair of foliations $\mathcal{A} = \{A^k(c)\}$ and $\mathcal{B} = \{B^{n-k}(d)\}$, a point p is determined uniquely as the intersection of a $A^k(c)$ and a $B^{n-k}(d)$. Hence, (c, d) can be used as a local coordinate system, called the mixed coordinate system. Let $p = (c_p, d_p)$ and $q = (c_q, d_q)$ in the mixed coordinates. Then,*

$$D(p : q) = D(p : \hat{p}) + D(\hat{p} : q), \tag{3.9}$$

$$D(q : p) = D(q : \hat{q}) + D(\hat{q} : p), \tag{3.10}$$

where $\hat{p}$ is the ∇^*-projection of p to $A^k(c_p)$ and $\hat{q}$ is the ∇-projection of q to $B^{n-k}(d_q)$.

We have discussed the properties of dually flat manifolds. When S is a manifold of dually constant curvatures, it also enjoys some nice properties expressed by transformations and potential functions. They are studied by Kurose [1993].

4. Some Unsolved Mathematical Questions

The dual structure originated from information sciences, especially statistics. Some special cases of the dual structure is given rise to by the affine differential geometry (Blashkke [1923], Nomizu and Sasaki [1994], Nomizu and Pinkall [1987], Simon et al. [1991]) and also by the Hessian manifold of Shima [1995]. However, there remain many problems to be studied from the mathematical point of view. We show some of them.

1. Embedding problem : It is known that an n-dimensional Riemannian manifold is embedded isometrically in an $n(n+1)/2$-dimensional flat (i.e. Euclidean) manifold. Is it possible to embed an n-dimensional manifold M of the dual structure in an m-dimensional dually flat manifold S such that the dual structure of M is induced from that of S? How large m is necessary? It is conjectured from statistics that $m = \infty$ is in general necessary. This is because a general statistical model $M = \{p(x, \xi)\}$ is not a curved exponential family. When and only when M has a finite-dimensional sufficient statistics for repeated observations, it is a curved exponential family which is embedded in a dually flat exponential family. Then, what is the characterization of a manifold which is embedded in a finite-dimensional dually flat manifold compatibly?
2. Affine embedding problem : Let M be an n-dimensional submanifold in an $(n+1)$-dimensional affine space. Then, it is known from affine differential geometry that a dual structure is induced in M. However, it produces only a special type of the dual structure. More precisely, when and only when a dual structure is conformally flat, it can be realized as a hypersurface of an $(n+1)$-dimensional affine space (Nomizu and Sasaki [1994], Kurose [1990; 1991]). Then, what type of the dual structure is induced in an n-dimensional submanifold of an m-dimensional affine space? See Lauritzen [1987]. In other words, what is the characterization of the dual structure which is realized as a submanifold of a finite-dimensional affine space?
3. Dual flatness : Given a Riemannian manifold $\{M, g\}$ it is sometimes possible to introduce a dual flat structure when one can find T such that $\{M, g, T\}$ is dually flat. Is this always possible? If possible, is T unique? If it is not always, what is the characterizaton of a manifold having such a T?
4. Global structure : We have so far studied only local properties of the dual structure. How are local structures related to global properties? Some preliminary studies are found in Shima [1995]. This is an interesting mathematical problem.
5. Quantum probability distributions : The manifold of quantum probability has attracted much attention in relation with problems of quantum observation and information transmission by quantum states. The non-commutativity plays a fundamental role, and the related manifold has torsion structure. Construction of non-commutative information geometry is an important problem. See Petz [1994a, b], Nagaoka [1989; 1994], Hasegawa [1993] and others.

6. Infinite-dimensional statistical manifolds : We need mathematical foundations on the infinite-dimensional manifold of probability distributions (Friedrich [1991], Lafferty [1988], Kambayashi [1994]). Here, the topology of the manifold is different from that derived from the Fisher information metric. It is interesting to know the statistical role of the symplectic structure induced in it (Friedrich [1991], Lafferty [1988]; see also Barndorff-Nielsen and Jupp [1989]). Infinite dimensional structures are related to nonparametric and semiparametric statistical inference (Amari and Kawanabe [1996]).
7. Dual flatness and completely integrable dynamical systems : There are discussions concerning intrinsic relations between the dually flat structure and complete-integrability of dynamical systems. It is pointed out that the affine projection method of Linear Programming can be formulated as a completely integrable dynamical system (Bayer and Lagarias [1989]). On the other hand, the trajectory of the solution by the inner method is proved to be a ∇^*-geodesic of the related dually flat manifold. Such relations to information geometry are found in the Brockett flow and in the Lax form. See Nakamura [1993], [1994] and Fujiwara and Amari [1995].
8. Finslerian structure : When the probability distributions $\{p(x, \xi)\}$ are statistically singular such as they do not have finite Fisher information, we cannot introduce a Riemannian structure. A Finslerian structure was suggested in such a case (Amari). This is related to asymptotics of stable distributions instead of the central limit theorem and asymptotic Gaussian distributions.

5. Manifold of Probability Distributions

We touch upon the invariant structure of the manifold of probability distributions from which the dual geometry originated. See Amari [1985], Amari et al. [1987], Barndorff-Nielsen [1988], Barndorff-Nielsen, Cox and Reid [1986], Murray and Rice [1993], Kass [1989], etc. Let S be a set of the probability distributions

$$S = \{p(x, \xi)\}$$

of a random variable x parametrized by an n-dimensional real vector parameter $\xi = (\xi^1, \cdots, \xi^n) \in \mathrm{R}^n$. Here, $p(x, \xi)$ is a probability density with respect to a common measure $d\mu$ of the sample space $\mathcal{X}$. Under a certain regularity conditions, S is regarded as an n-dimensional manifold with a local coordinate system ξ. We give some examples.

1. Discrete distributions S_n.

Let $\mathcal{X}$ be the finite sample space $\mathcal{X} = \{0, 1, 2, \cdots, n\}$ where x takes on $n+1$ values $0, 1, \cdots, n$. The probability of x is written as

$$p(x) = \sum_{i=0}^{n} \delta_i(x) p_i, \tag{5.1}$$

where $\delta_i(x) = 1$ when $x = 1$ and 0 otherwise and

$$p_i = \mathrm{Prob}\{x = i\}. \tag{5.2}$$

Hence, the set of all such distributions is parametrized by $\eta = (\eta_i)$,

$$\eta_i = p_i, \quad i = 1, \cdots, n$$

because of $\sum_{i=0}^{n} p_i = 1$. Hence, this is an n-dimensional manifold denoted by S_n.

2. Exponential family.

When the probability density function is written as

$$r(x,\theta) = \exp\left\{\sum_{i=1}^{n} \theta^i k_i(x) - m(x) - \psi(\theta)\right\}, \tag{5.3}$$

the family of such distributions are said to be of the exponential type. Here, $\theta = (\theta^1, \cdots, \theta^n)$ is called the natural parameter to specify distributions. By introducing a vector $x = (x_i)$, $x_i = k_i(x)$, and by taking a common measure

$$d\mu(x) = \exp\{m(x)\}dx,$$

it is rewritten in the simple form

$$p(x,\theta) = \exp\{\sum \theta^i x_i - \psi(\theta)\}, \tag{5.4}$$

where $p(x,\theta)$ is the probability density with respect to the measure $d\mu(x)$. A family $S = \{p(x,\theta)\}$ is said to be an exponential family when it can be represented in the form (5.4). The manifold S_n of discrete distributions is an exponential family because it is rewritten as

$$p(x,\theta) = \exp\{\sum \theta^i x_i - \psi(\theta)\}$$

where

$$x_i = \delta_i(x),$$
$$\theta^i = \log \frac{p_i}{p_0}.$$

A smooth submanifold M of an exponential family is said to be a curved exponential family. Let $u = (u^1, \cdots, u^m)$, $m < n$, be local coordinates of M, and is represented by the equation

$$M:\ \theta = \theta(u).$$

Then, the probability distributions of M are parametrized by u in the form

$$p(x,u) = \exp\{\sum \theta^i(u)x_i - \psi[\theta(u)]\}. \tag{5.5}$$

This is not an exponential family unless $\theta(u)$ is linear.

3. Infinite-dimensional space.

Let us consider the set of all the probability distributions which have smooth non-zero densities with respect to the Lebesgue measure,

$$p(x) > 0.$$

Such a set $S = \{p(x)\}$ does not form a manifold unless we introduce some conditions in the behaviors of $p(x)$ at $x \to \pm\infty$. The set $\mathcal{P}$ of all the smooth distributions over the one-dimensional sphere (circle) S^1 was studied by Friedrich [1991] and Lafferty [1988], see also Kambayashi [1994]. Friedrich and Lafferty showed that $\mathcal{P}$ is homeomorphic to the manifold of all the smooth deformations of S^1, while Kambayashi used the manifold of paths. Geometry of the manifolds of Gaussian and Poissonian processes were studied by Takahashi and Sekine.

What is the natural geometrical structure to be introduced in the manifold of probability distributions? It was Rao [1945] who proposed the Riemannian structure by using the Fisher information matrix. The matrix $g = (g_{ij})$ defined by

$$g_{ij}(\xi) = E\left[\frac{\partial \log p(x,\xi)}{\partial \xi^i} \frac{\partial \log p(x,\xi)}{\partial \xi^j}\right]$$

is an important quantity known as the Fisher information, where E denotes the expectation with respect to the distribution $p(x, \xi)$. Chenstov [1972] proved that the Fisher information is the unique metric under the Markovian morphisms among $\{S_m\}$, $m = 1, 2, \cdots$ in the discrete distributions (see also Campbell [1985]). He also introduced a family of affine connections called the α-connections (Amari [1982], [1985]) defined by

$$\Gamma^{(\alpha)}_{ijk}(\xi) = \{ij;k\} - \frac{\alpha}{2}T_{ijk}, \tag{5.6}$$

$$T_{ijk}(\xi) = E\left[\frac{\partial \log p(x,\xi)}{\partial \xi^i} \frac{\partial \log p(x,\xi)}{\partial \xi^j} \frac{\partial \log p(x,\xi)}{\partial \xi^k}\right]. \tag{5.7}$$

The duality of the affine connections was formulated by Nagaoka and Amari [1982] (see also Amari [1985], Murray and Rice [1993]).

Under a certain regularity conditions, the α-structure given by $\{S, g, \alpha T\}$ or $\{S, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)}\}$ is the unique geometrical structure under the transformation of random variable x to a sufficient statistics $y = f(x)$. A typical case of such f is a bijective transformation. (See Amari and Nagaoka [1994], see also Li [1993] for regularity conditions). A variable $y = f(x)$ is called a sufficient statistics, when the probability distribution $p(x, \xi)$ is of the form

$$p(x, \xi) = h(y; \xi)r(y)$$

where r does not depend on ξ. In this case, y carries all the necessary information concerning the underlying probability distributions.

6. Applications of Information Geometry

6.1. Statistical inference. Because information geometry originated from the manifold of probabilities, it gives naturally a mathematical foundation of statistical inference. The basic structures will be explained here intuitively. To this end, let us consider a curved exponential family $M = \{p(x, u)\}$, which is embedded in an exponential family $S = \{q(x, \theta)\}$,

$$q(x, \theta) = \exp\{\sum \theta^i x_i - \psi(\theta)\}, \tag{6.1}$$

$$p(x, u) = q\{x, \theta(u)\} = \exp\{\sum \theta^i(u)x_i - \psi(\theta)\}. \tag{6.2}$$

The manifold M is a curved manifold in S, where $u = (u^a)$ is a coordinate system in M.

Let $x_1, x_2, \cdots, x_N$ be N independent observations from an unknown probability distribution $p(x, u)$. From the observed sample, we need to estimate the true value u by $\hat{u} = \hat{u}(x_1, \cdots, x_N)$. In the framework of the exponential family S, we take the η-coordinate system in it given by

$$\eta_i = \frac{\partial}{\partial \theta^i}\psi(\theta) = E_\theta[x_i]. \tag{6.3}$$

Then, the natural estimator $\hat{\eta}$ in S is given by replacing the expectation by the sample average,

$$\hat{\eta}_i = \frac{1}{N}\sum x_i. \tag{6.4}$$

This is the maximum likelihood estimator in S and the $\hat{\eta}$ summarizes all the Fisher information included in N observations. Let $\hat{p}$ be the point in S whose η-coordinates are given by $\hat{\eta}$. We call $\hat{p}$ the observed point. However, $\hat{p}$ does not in general belong to M so that it does not give an estimator $\hat{u}$. Geometrically speaking, an estimator e is a mapping from S to M,

$$e : S \to M \; ; \; \hat{p} \mapsto \hat{u},$$

where $\hat{u}$ is the parameter representing $p(x, \hat{u})$ in M. Let

$$e^{-1}(u) = \{p|\, e(p) = u, \; p \in S\}. \tag{6.5}$$

We call $e^{-1}(u)$ the ancillary manifold of u associated with the estimator. When and only when the observed $\hat{p}$ belongs to $e^{-1}(u)$, the estimated value is u.

In order that an estimator is consistent, that is, $\hat{u}$ converges to the true u in probability as N tends to infinity, it is necessary and sufficient that

$$e^{-1}(u) \ni p(x, u). \tag{6.6}$$

Now, let us consider the expected estimation error given by the covariance matrix

$$C^{ab} = \frac{1}{N}E[(\hat{u}^a - u^a)(\hat{u}^b - u^b)]. \tag{6.7}$$

When N is large, the law of large numbers guarantees that $\hat{p}$ is in an $(1/\sqrt{N})$-neighborhood of the true distribution. Hence, in order to evaluate the performance of a consistent estimator, we can linearize the estimation problem in the tangent space $T_S(u)$ and $T_M(u)$ of S and M at $p(x, u)$. The linear analysis gives the following well known universal result of the asymptotic theory: When and only when $e^{-1}(u)$ is orthogonal (with respect to the Fisher information metric) to M, a consistent estimator has the minimum expected error given by

$$C^{ab} = g^{ab} + O\left(\frac{1}{N}\right) \tag{6.8}$$

where $g^{-1} = (g^{ab})$ is the inverse of the Fisher metric of M.

When N is not so large, we need to evaluate the term $O\left(\frac{1}{N}\right)$. In this case, the linear approximation in the tangent space T_S is not sufficient, and we need to evaluate the curvatures which show how T_S changes. For a bias-corrected efficient estimator, we have the following universal result:

$$C^{ab} = g^{ab} + \frac{1}{2N}\left(2\left[H_M^{(e)}\right]^{2ab} + \left[H_A^{(m)}\right]^{2ab} + \left[\Gamma_M^{(m)}\right]^{2ab}\right) + O\left(\frac{1}{N^2}\right), \tag{6.9}$$

where $\left[H_M^{(e)}\right]^2$, $\left[H_A^{(m)}\right]^2$ and $\left[\Gamma_M^{(m)}\right]^2$ are the squares of the e-curvature of M, m-curvature of $e^{-1}(u)$ and the m-connection of M, respectively. Hence, an estimator is second-order efficient when the associated $e^{-1}(u)$ has vanishing m-curvature. The maximum-likelihood estimator attains this.

We can similarily construct the higher-order asymptotic theory of testing hypothesis and interval estimation by using information geometry. The theory is extended to general non-curved exponential cases where the fibre bundle structure is used (Amari [1987a], Barndorff-Nielsen and Jupp [1989]), to sequential estimation where conformal information geometry is used (Okamoto, Amari and Takeuchi [1991]), and to semiparametric cases where the Hilbert bundle structure is introduced (Amari [1982], Amari and Kawanabe [1996]).

6.2. Linear systems and time series. Let us consider a time series $\{x_t\}$, $t = 0, \pm 1, \pm 2, \cdots$, generated by white Gaussian noises $\{z_t\}$ as

$$\sum_{i=0}^{p} a_i x_{t-i} = \sum_{i=1}^{q} b_i z_{t-i}. \tag{6.10}$$

The probability distribution of $\{x_t\}$ is specified by the parameter $\mathrm{u} = (a_0, \cdots, a_p; b_1, \cdots, b_q)$ and is written symbolically as $p(\{x_t\}, \mathrm{u})$. The transformation from $\{z_t\}$ to $\{x_t\}$ is called a linear system, so that we can study the set of the linear systems of the form (6.10) by information geometry. See Amari [1987b].

A control system modifies its performance by adding a feedback loop from the output. The set of linear feedback systems also form a differentiable manifold, where the information-geometric dual structure is introduced. It plays an important role in control systems theory. See Ohara and Amari [1994]; Ohara, Suda and Amari [1995].

6.3. Neural networks and nonlinear systems. Let us consider a nonlinear input-output system

$$\mathrm{x} = \mathrm{f}(\mathrm{z}; \mathrm{u}) \tag{6.11}$$

parametrized by parameter u, where z is an input signal and x is the corresponding output signal. The set of such nonlinear systems forms a differentiable manifold where u is a (local) coordinate system. When z is a random input, (6.11) defines a probability distribution $p(\mathrm{x}; \mathrm{u})$ specified by u. Hence, the dual structure is introduced in the manifold of nonlinear systems.

Sometimes, the input-output relation is stochastic, for example,

$$\mathrm{x} = \mathrm{f}(\mathrm{z}; \mathrm{u}) + \mathrm{n} \tag{6.12}$$

where n is a random noise. In this case, the behaviour of a system is specified by a conditional probability distribution $p(\mathrm{x}|\mathrm{z}; \mathrm{u})$ of x conditioned on z. This also introduces the dual structure of information geometry. It can be extended to the nonlinear dynamical systems

$$\mathrm{x}_{t+1} = \mathrm{f}(\mathrm{x}_t; \mathrm{u}). \tag{6.13}$$

Neural networks are nonlinear systems which include both signal transformations (feedforward networks) and dynamics (recurrently connected networks). They are regarded as neural manifolds, where the synaptic connection weights play the role of a coordinate system. The neural information processing is studied by information geometry. See Amari [1991], [1995]; Amari, Kurata and Nagaoka [1992].

6.4. Information theory and statistics. Information theory studies problems of information transmission where signals are generated from an information source which is subject to a probability distribution. Hence, the dual structure is naturally introduced when we consider a family of information sources. Information geometry plays a fundamental role in the multiterminal statistical inference problem where statistical inference is discussed under multiterminal information transmission. The dual foliation of the underlying manifold plays a fundamental role, because it decomposes the marginal structures from correlational structures. See Amari [1989], Amari and Han [1989].

6.5. Linear programming, convex analysis and completely integrable dynamical systems. A dually flat manifold is naturally constructed from a convex function and the related Legendre transformation. For example, we take the linear programming problem. Let M be a convex region in R^n,

$$M = \{\theta | \theta \in \mathrm{R}^n,\ \sum_{i=1}^{n} A_i^\mu \theta^i - b^\mu \geq 0,\ \mu = 1, \cdots, m\}. \tag{6.14}$$

The linear programming problem is to find a point $\boldsymbol{\theta}$ that minimizes a linear function

$$f(\theta) = \sum c_i \theta^i. \tag{6.15}$$

We can introduce a convex function ψ in the inner region of M by

$$\psi(\theta) = -\sum_{\mu=1}^{m} \log\left(\sum A_i^\mu \theta^i - b^\mu\right). \tag{6.16}$$

This automatically defines the dually flat structure in M. It was shown by Fujiwara and Amari [1995] that the Karmarkar affine projection method is the inner method of obtaining the optimal θ along the ∇^*-geodesic. It was also shown by Bayer and Lagarias [1989] that the method can be reformulated as a completely integrable dynamical system. This is also related to the Brockett flow and solitons. The relation between the dually flat structure and the completely integrable dynamical systems is an important unsolved problem which attracts much attention. See Nakamura [1993], [1994], Fujiwara and Amari [1995].

6.6. Quantum information geometry. Information geometry originated from probability distributions. Quantum mechanics defines another type of probability, the complex probability, where non-commutative operators play a fundamental role. It is an important problem to construct non-commutative information geometry to understand the quantum observation problem. Information transmission via quantum states, and quantum-statistical estimation of states are also an important problem. There are lots of efforts along this line. See Petz [1994a, b], Nagaoka [1989, 1994], Fujiwara and Nagaoka, Hasegawa [1993].

7. Conclusions

Information geometry is a field of mathematical sciences which includes wide areas of applications. It also suggests mathematics a new but classic concept, in its nature, of the dual affine connections, which should be studied further in the modern context. The present paper reviews the fundamental concepts of information geometry and surveys its applications briefly. The field is developing rapidly, and there are lots of problems to be solved. It is hoped that the present review ignites cooperation between mathematical scientists and pure mathematicians.

References

[1] AMARI, S., *Differential geometry of curved exponential families — curvature and information loss*, Ann. Statist. **10**, pp. 357–385 (1982)

[2] AMARI, S., *Differential-Geometrical Methods in Statistics*, Springer lecture Notes in Statistics **28**, (1985).

[3] AMARI, S., *Differential geometry of a parametric family of invertible linear systems — Riemannian metric, dual affine connections and divergence*, Mathematical systems Theory **20**, pp. 53–82 (1987b).

[4] AMARI, S., *Fisher information under restriction of Shannon information*, AISM **41**, pp. 623–648 (1989).

[5] AMARI, S., *Dualistic Geometry of the Manifold of Higher-Order Neurons*, Neural Networks **4**, pp. 217–227 (1989)

[6] AMARI, S., BARNDORFF-NIELSEN, O. E., KASS, R. E., LAURITZEN, S. L. and RAO, C. R,, *Differential Geometry in Statistical Inferences*, IMS Lecture Notes Monograph Series **10**, Hayward California IMS (1987).

[7] AMARI, S. and HAN, T. S., *Statistical inference under multi-terminal rate restrictions — a differential geometrical approach*, IEEE Trans. on Information Theory **IT-35**, pp. 217–227 (1989).

[8] AMARI, S. and KAWANABE, M., *Information geometry of semiparametric statistical models and estimating functions*, to appear (1996).

[9] AMARI, S., KURATA, K. and NAGAOKA, H., *Information Geometry of Boltzmann Machines*, IEEE Trans. on Neural Networks **3** No.2, pp. 260–271 (1992).

[10] AMARI, S. and NAGAOKA, H., *Method of Information Geometry*, Iwanami Lecture Series in Applied Mathematics (in Japanese, to be translated into English by AMS and Oxford University Press) (1994).

[11] BARNDORFF-NIELSEN, O., *Parametric Statistical Models and Likelihood*, Springer Lecture Notes in Statistics **50**, (1988).

[12] BARNDORFF-NIELSEN, O., COX, R. D. and REID, N., *The role of differential geometry in statistical theory*, Int. Staatist. Rev. **54**, pp. 89–96 (1986).

[13] BARNDORFF-NIELSEN, O. and JUPP, P. E., *Approximating exponential models*, AISM **41**, pp. 247–267 (1989).

[14] BAYER, D. A. and LAGARIAS, J. C., *The nonlinear geometry of linear programming*, Trans. AMS. **314**, pp. 499–527 (1989)

[15] BLASCHKE, W., *Vorlesungen über Differentialgeometrie II*, Affine Differentialgeometrie, Springer Berlin (1923).

[16] BURDET, G. and PERRIN, M., *Conformal implementation of the projective equivalence over statistical manifolds*, Indagationes Mathematical, to appear (1996).

[17] CHENTSOV, N. N., *Statistical Decision Rules and Optimal Inference* (in Russian), Nauka Moscow; translated in English, (1982), AMS Rhode Island (1972).

[18] FRIEDRICH, V. T., *Die Fisher-information und symplektische Strukturen*, Math. Nachr. **153**, pp. 273–296 (1991).

[19] FUJIWARA, A. and AMARI, S., *Gradient systems in view of information geometry*, Physica D **80**, pp. 317–327 (1995).

[20] HASEGAWA, H., *α-divergence of the non-commutative information geometry*, Rep. Math. Phys. **33**, pp. 87–93 (1993).

[21] KAMBAYASHI, T., *Statistical manifold of infinite dimension*, Transactions of the Japan Society for Industrial and Applied Mathematics **4** No3, pp. 211–228 (in Japanese) (1994).

[22] KASS, R. E., *The geometry of asymptotic inference* (with discussions), Statistical Science **4**, pp. 188–234 (1989).

[23] KUROSE, T., *On the divergence of 1-conformally flat statistical manifold*, Tohoku Math.Ê J. **46**, pp 427–433 (1994).

[24] KUROSE, T., *On the realization of statistical manifolds in affine space*, (1994).

[25] KUROSE, T., *Dual connections and affine geometry*, Math. Z. **203**, pp. 115–121 (1990).

[26] LAFFERTY, J. D., *The density manifold and configuration space quantization*, Trans. AMS **305**, pp. 699–741 (1988)

[27] LAURITZEN, S. L., *Conjugate connection in statistical theory*, in Dodson, C. T. J. ed., Geometrization of Statistical Theory, ULDM Publications, pp. 33–51 (1987).

[28] LI, H., *Determination of geometrical structures on statistical experiments, probability and statistics*, Proceedings of the Wuhan Meeting (eds. A. Badrkian, P-A Meyer, J-A Yan), World Scientific, pp. 133–1148 (1993).

[29] MURRAY, M. K. and RICE, J. W., *Differential Geometry and Statistics*, Chapman & Hall (1993).

[30] NAGAOKA, H., *On the relation between Kullback divergence and Fisher information — from classical systems to quantum systems*, an outline of IEICE Tech. Rep. **89**, pp. 9–14 (1989).

[31] NAGAOKA, H., *Differential geometrical aspects of quantum state estimation and relative entropy*, METR94-14 (1994).

[32] NAGAOKA, H. and AMARI, S., *Differential geometry of smooth families of probability distributions*, METR **82-7**, University of Tokyo (1982).

[33] NAKAMURA, Y., *Completely integrable gradient systems on the Manifolds of Gaussian and multinomial distributions*, Japan Journal of Industrial and Applied Mathematics **10** No.2, pp. 179–189 (1993).

[34] NAKAMURA, Y., *A Tau-function for the finite Toda molecule and information spaces*, Contemporary Mathematics **179** (1994).

[35] NOMIZU, K. and PINKALL, O., *On the geometry of affine immersions*, Math. Z. **195**, pp. 165–178 (1987).

[36] NOMIZU, K. and SASAKI, T., *Affine Differential Geometry*, Cambridge Trans. in Mathematics (1994).

[37] OHARA, A. and AMARI, S., *Differential geometric structures of stable state feedback systems with dual connections*, Kybernetica **30**, pp. 369–386 (1994).

[38] OHARA, A., SUDA, N. and AMARI, S., *Dualistic differential geometry of positive definite matrices and its applications to related problems*, Linear Algebra Appl. (to appear).

[39] OKAMOTO, I, AMARI, S. and TAKEUCHI, *Asymptotic theory of sequential estimation procedures for curved exponential families*, Ann. Statist. **19**, pp. 961–981 (1991).

[40] PETZ, D., *Geometry of canonical correlation on the state space of a quantum system*, J. Math. Phys. **35**, pp. 780–795 (1994a).

[41] PETZ, D., *Monotone metric on matrix spaces*, Linear Algebra Appl. (1994b).

[42] RAO, C. R., *Information and accuracy attainable in the estimation of statistical parameters*, Bull. Calcutta. Math. Soc. **37**, pp. 81–91 (1945).

[43] SHIMA, H., *Homogeneous Hessian Manifolds*, Ann. Inst. Fourier Grenoble **30**, pp. 91–128 (1980).

[44] SHIMA, H., *Vanishing theorems for compact Hessian manifolds*, Ann. Inst. fourier Grenoble **36** 3, pp. 183–205 (1986).

[45] SHIMA, H., *Hessian manifolds of constant Hessian sectional curvature*, J. Math. Soc. Japan **47** 4, pp. 735–753 (1995).

[46] SIMON, U., SCHWENK-SCHELLSCHMIDT, A. and BIESEL, H., *Introduction to the Affine Differential Geometry of Hypersurfaces*, Lecture Notes, Science U. of Tokyo (1991).

DEPARTMENT OF MATHEMATICAL ENGINEERING AND INFORMATION PHYSICS, UNIVERSITY OF TOKYO, TOKYO 113, JAPAN; RIKEN FRONTIER RESEARCH PROGRAM, BRAIN INFORMATION PROCESSING

Current address: RIKEN FRP, Wako-shi, Hirosawa, Saitama 351-01, Japan

E-mail address: `amari@sat.t.u-tokyo.ac.jp`