

Information Geometry and Its Applications: Convex Function and Dually Flat Manifold

Shun-ichi Amari

RIKEN Brain Science Institute
Amari Research Unit for Mathematical Neuroscience
`amari@brain.riken.jp`

Abstract. Information geometry emerged from studies on invariant properties of a manifold of probability distributions. It includes convex analysis and its duality as a special but important part. Here, we begin with a convex function, and construct a dually flat manifold. The manifold possesses a Riemannian metric, two types of geodesics, and a divergence function. The generalized Pythagorean theorem and dual projections theorem are derived therefrom. We construct alpha-geometry, extending this convex analysis. In this review, geometry of a manifold of probability distributions is then given, and a plenty of applications are touched upon. Appendix presents an easily understandable introduction to differential geometry and its duality.

Keywords: Information geometry, convex function, Riemannian geometry, dual affine connections, dually flat manifold, Legendre transformation, generalized Pythagorean theorem.

1 Introduction

Information geometry emerged from a study on the invariant geometrical structure of a family of probability distributions. We consider a family $S = \{p(x, \boldsymbol{\theta})\}$ of probability distributions, where x is a random variable and $\boldsymbol{\theta}$ is an n -dimensional vector parameter. This forms a geometrical manifold where $\boldsymbol{\theta}$ plays the role of a coordinate system.

We searched for the invariant structure to be introduced in S , and found a Riemannian structure together with a dual pair of affine connections (see Chentsov, 12; Amari and Nagaoka, 8). Such a structure has scarcely been studied in traditional differential geometry, and is still not familiar.

Typical families of probability distributions, e.g., exponential families and mixture families, are dually flat together with non-trivial Riemannian metrics. Some non-flat families are curved submanifolds of flat manifolds. For example, the family of Gaussian distributions

$$S = \left\{ p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} \right\}, \quad (1)$$

where μ is the mean and σ^2 is the variance, is a flat 2-dimensional manifold. However, when $\sigma^2 = \mu^2$ holds, the family of distributions

$$M = \left\{ p(x, \mu) = -\frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu)^2}{2\mu^2} \right\} \right\} \quad (2)$$

is a curved 1-dimensional submanifold (curve) embedded in S . Therefore, it is important to study the properties of a dually flat Riemannian space.

A dually flat Riemannian manifold possesses dual convex potential functions, and all the geometrical structure can be derived from them. In particular, a Riemannian metric, canonical divergence, generalized Pythagorean relation and projection theorem are their outcomes.

Conversely, given a convex function, we can construct a dually-flat Riemannian structure, which is an extension and foundation of the early approach by (Bregman, 10) and a geometrical foundation of the Legendre duality. The present paper focuses on a convex function, and reconstructs dually-flat Riemannian structure therefrom. See (Zhang, 28) for details.

Applications of information geometry are expanding, and we touch upon some of them. See Appendix for an understandable introduction to differential geometry.

2 Convex Function and Legendre Structure

A dually flat Riemannian manifold possesses a pair of convex functions. The set of all the discrete probability distributions gives a dually flat manifold. The geometrical structures are derived from the convex functions. Therefore, many useful results are derived from the analysis of a convex function. On the other hand, given a convex function together with its dual convex function, a dually flat Riemannian manifold is automatically derived. In the present section, we begin with a convex function and derive its fundamental properties from the dual geometry point of view. Convex analysis is important in many fields of science such as physics, optimization, information theory, signal processing and vision.

2.1 Metric Derived from Convex Function

Let us consider a smooth convex function $\psi(\boldsymbol{\theta})$ defined in an open set S of $\mathbf{R}^n$, where $\boldsymbol{\theta}$ plays the role of a coordinate system. Its second derivative, that is, the Hessian of ψ ,

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}) \quad (3)$$

is a positive definite matrix depending on position $\boldsymbol{\theta}$, where $\partial_i = \partial/\partial\theta^i$, and $\boldsymbol{\theta} = (\theta^1, \dots, \theta^n)$.

Consider two infinitesimally nearby points $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$, and define the square of their distance by

$$ds^2 = \langle d\boldsymbol{\theta}, d\boldsymbol{\theta} \rangle = \sum g_{ij}(\boldsymbol{\theta}) d\theta^i d\theta^j, \quad (4)$$

where $\langle d\boldsymbol{\theta}, d\boldsymbol{\theta} \rangle$ is the inner product defined in the above. This is the second-order term of the Taylor expansion of $\psi(\boldsymbol{\theta} + d\boldsymbol{\theta})$,

$$\psi(\boldsymbol{\theta} + d\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}) + \sum \partial_i \psi d\theta^i + \frac{1}{2} \sum g_{ij} d\theta^i d\theta^j, \quad (5)$$

so that it is defined by the curvature of function ψ . A manifold in which an infinitesimal distance is defined by (4) is called a Riemannian manifold, where the matrix $g = (g_{ij})$ is called a Riemannian metric.

When

$$\psi(\boldsymbol{\theta}) = \frac{1}{2} \sum (\theta^i)^2, \quad (6)$$

we have

$$g_{ij} = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad (7)$$

and the space is Euclidean, because the squared distance is given by

$$ds^2 = \sum (d\theta^i)^2. \quad (8)$$

Hence our framework includes the Euclidean geometry as its special case.

We have fixed a coordinate system $\boldsymbol{\theta}$, and the metric is derived by using the convex function $\psi(\boldsymbol{\theta})$. In order to have a general geometrical structure, the geometry should be invariant under coordinate transformations. However, the convexity is not invariant under coordinate transformations. We define a geometrical structure by introducing a metric and affine connection in a manifold with respect to a specific coordinate system, that is $\boldsymbol{\theta}$ in our case, and then extend it invariantly to any coordinate systems.

The Riemannian metric is defined by (3). We define a geodesic in S such that a straight line connecting two points, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$,

$$\boldsymbol{\theta}(t) = (1 - t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2 \quad (9)$$

is a geodesic of S . The coordinates lines of θ^i of $\boldsymbol{\theta}$ are all geodesics. Mathematically speaking, a geodesic is defined by using an affine connection. In the present case, the covariant derivative reduces to the ordinary derivative in this coordinate system, implying the space is flat (that is, Riemann-Christoffel curvature vanishes and the coefficients of the connection are 0 in this coordinate system). See appendix.

The affine connection implicitly introduced here is not a metric connection, that is, it is not a Riemannian connection derived from the metric g_{ij} . Therefore, a geodesic is no more a minimal distance line connecting two points. In other words, the straightness and minimality of distance holding in a Euclidean space splits in a space of general affine connection. This looks ugly at first glance, but its deep structure becomes clear when we consider duality.

2.2 Legendre Transformation and Dual Coordinates

The gradient of $\psi(\boldsymbol{\theta})$ is given by partial derivatives

$$\boldsymbol{\eta} = \text{Grad}\psi(\boldsymbol{\theta}), \quad \eta_i = \frac{\partial}{\partial \theta^i} \psi(\boldsymbol{\theta}). \quad (10)$$

It is remarkable that, given $\boldsymbol{\eta}$, its original $\boldsymbol{\theta}$ is uniquely determined, that is, the correspondence between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are one-to one. Hence we can use $\boldsymbol{\eta}$ as another coordinate system of S .

We use here lower indices like η_i to represent quantities related to the $\boldsymbol{\eta}$ coordinate system. This makes it clear that θ^i and η_i are mutually reciprocal.

The transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$ is called the Legendre transformation. We can find a convex function of $\boldsymbol{\eta}$, defined by

$$\varphi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}} \{ \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}) \}. \quad (11)$$

The two potential functions satisfy the relation

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \boldsymbol{\theta} \cdot \boldsymbol{\eta} = 0 \quad (12)$$

when $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are respective coordinates of the same point, and the inverse transformation from $\boldsymbol{\eta}$ to $\boldsymbol{\theta}$ is given by the gradient

$$\boldsymbol{\theta} = \text{Grad}\varphi(\boldsymbol{\eta}), \quad \theta^i = \frac{\partial}{\partial \eta_i} \varphi(\boldsymbol{\eta}). \quad (13)$$

Hence, they are dually coupled.

Since we have another coordinates $\boldsymbol{\eta}$ and another convex function $\varphi(\boldsymbol{\eta})$, we can define another geometric structure based on them in a similar manner. In this dual setting, η_i are dual geodesic coordinates, and any straight line connecting two points in the $\boldsymbol{\eta}$ coordinates is a dual geodesic. The Riemannian metric is given by

$$g^{ij} = \frac{\partial^2}{\partial \eta_i \partial \eta_j} \varphi(\boldsymbol{\eta}). \quad (14)$$

Since the Jacobian matrices of coordinate transformations are written as

$$g_{ij} = \frac{\partial \eta_i}{\partial \theta^j}, \quad g^{ij} = \frac{\partial \theta^i}{\partial \eta_j}, \quad (15)$$

we can prove the following theorem.

Theorem 1. The metric g^{ij} is the inverse matrix of the metric g_{ij} . Moreover, $\sum g_{ij} d\theta^i d\theta^j = \sum g^{ij} d\eta_i d\eta_j$ implying the two metrics defined in terms of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are the same.

Proof. It is easy to see from (15) that $G = (g_{ij})$ and $G^{-1} = (g^{ij})$ are mutually inverse. We also see

$$d\boldsymbol{\eta} = G d\boldsymbol{\theta}, \quad d\boldsymbol{\theta} = G^{-1} d\boldsymbol{\eta}. \quad (16)$$

Hence

$$ds^2 = d\boldsymbol{\theta}^T G d\boldsymbol{\theta} = d\boldsymbol{\eta}^T G^{-1} d\boldsymbol{\eta}. \quad (17)$$

When we consider two points P and $P + dP$ whose coordinates are $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$, and $\boldsymbol{\eta}$ and $\boldsymbol{\eta} + d\boldsymbol{\eta}$ in the respective coordinates, the above theorem shows that the squared distance between P and $P + dP$ is the same whichever coordinates we use. That is, the two Riemannian metrics are the same, and they are represented in different coordinate systems. This is because it is a tensor. However, the two types of geodesics are different.

2.3 Divergence

We introduce a divergence function between two points P and Q , of which coordinates are written as $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_Q$, and also $\boldsymbol{\eta}_P$ and $\boldsymbol{\eta}_Q$ in the dual coordinates. The divergence is defined by

$$D(P : Q) = \psi(\boldsymbol{\theta}_P) + \varphi(\boldsymbol{\eta}_Q) - \boldsymbol{\theta}_P \cdot \boldsymbol{\eta}_Q, \quad (18)$$

where

$$\boldsymbol{\theta} \cdot \boldsymbol{\eta} = \sum \theta^i \eta_i. \quad (19)$$

It is easy to see

$$D(P : P) = 0, \quad (20)$$

and is positive otherwise. The divergence is not symmetric in general,

$$D(P : Q) \neq D(Q : P). \quad (21)$$

Changing the role of P and Q (or $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$), we can define the dual divergence

$$D^*(P : Q) = \varphi(\boldsymbol{\eta}_P) + \psi(\boldsymbol{\theta}_Q) - \boldsymbol{\eta}_P \cdot \boldsymbol{\theta}_Q. \quad (22)$$

However, it is easy to see

$$D(P : Q) = D^*(Q : P). \quad (23)$$

Hence we need only one divergence. The other is obtained by changing the order of two points.

When Q is close to P , by putting $Q = P + dP$, we have

$$D(P : P + dP) = \frac{1}{2} \sum g_{ij} d\theta^i d\theta^j = \frac{1}{2} \sum g^{ij} d\eta_i d\eta_j \quad (24)$$

which is the squared Riemannian distance. Hence, the divergence is similar to the square of distance. Therefore, the triangle inequality does not hold. Instead, the Pythagoras-like theorem holds.

2.4 Generalized Pythagorean Theorem

We first define the orthogonality of two curves intersecting at a point. Let us consider two curves $\boldsymbol{\theta}(t)$ and $\boldsymbol{\theta}'(t)$ parameterized by scalar t and assume they

intersect at $t = 0$, that is $\boldsymbol{\theta}(0) = \boldsymbol{\theta}'(0)$. The same curves can be represented by the dual coordinates as $\boldsymbol{\eta}(t)$ and $\boldsymbol{\eta}'(t)$. The tangent vector of a curve $\boldsymbol{\theta}(t)$ is a vector represented by $\dot{\boldsymbol{\theta}}(t) = (\dot{\theta}^1(t), \dots, \dot{\theta}^n(t))$ where $\dot{} = d/dt$ represents the derivative with respect to t .

The two curves are said to be orthogonal at the intersection point, when their inner product vanishes,

$$\langle \dot{\boldsymbol{\theta}}, \dot{\boldsymbol{\theta}}' \rangle = \sum g_{ij} \dot{\theta}^i \dot{\theta}'^j = 0. \quad (25)$$

This can be represented as

$$\sum g^{ij} \dot{\eta}_i \dot{\eta}'_j = 0 \quad (26)$$

or

$$\dot{\boldsymbol{\theta}} \cdot \dot{\boldsymbol{\eta}}' = \sum \dot{\theta}^i \dot{\eta}'_i = 0 \quad (27)$$

by using the two coordinate systems.

Let us take three points P, Q and R in S .

Theorem 2. When the dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and R ,

$$D(P : R) = D(P : Q) + D(Q : R). \quad (28)$$

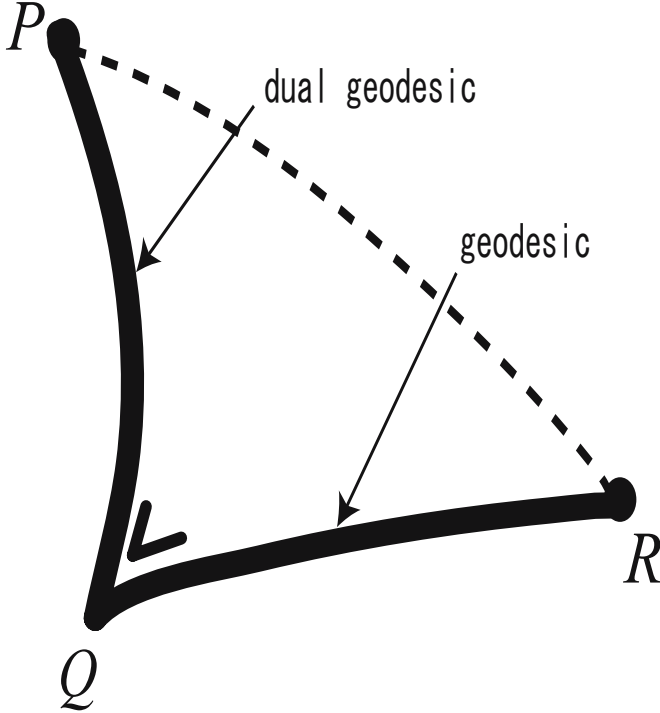


Fig. 1. Pythagorean theorem

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

$$D^*(P : R) = D^*(P : Q) + D^*(Q : R). \quad (29)$$

See Fig.1.

When the potential function $\psi(\boldsymbol{\theta})$ is quadratic,

$$\psi(\boldsymbol{\theta}) = \frac{1}{2} \sum (\theta^i)^2, \quad (30)$$

the metric reduces to the identity matrix $g_{ij} = \delta_{ij}$. The dual coordinates are the same as the primal ones, $\boldsymbol{\theta} = \boldsymbol{\eta}$, and two geodesics are the same. In this case, the above theorem is exactly the same as the Pythagorean theorem. Hence, our dually-flat manifold includes the Euclidean space as a special example.

2.5 Projection Theorem

Let us consider a closed region V included in S . Let P be a point in S not included in V . We search for the point in V which is closest to P in the sense of divergence. That is, we search for the point $P^\perp \in V$ that minimizes $D(P : Q)$, $Q \in V$. It is obvious that $P^\perp$ is on the boundary of V . This point $P^\perp$ is called the reverse I-projection or dual geodesic projection of P to V (Amari and Nagaoka, 2000; see also Csiszár, 1975).

Dually to the above, we may think about the point $P^{*\perp}$ in V that minimizes $D^*(P : Q) = D(Q : P)$, $Q \in V$. This is called the geodesic projection or the I-projection of P to V (Csiszár, 1975; Amari and Nagaoka, 2000).

We say V is convex, when the geodesic connecting two points R and R' are in V , that is,

$$(1 - t)\boldsymbol{\theta}_R + t\boldsymbol{\theta}_{R'} \in V. \quad (31)$$

We say that V is dual-convex, when the dual geodesic connecting R and R' is included in V , that is,

$$(1 - t)\boldsymbol{\eta}_R + t\boldsymbol{\eta}_{R'} \in V. \quad (32)$$

We have the following projection theorem.

Theorem 3. When $P^\perp$ is the dual projection of P to V , the dual geodesic connecting P and $P^\perp$ is orthogonal to the boundary surface of V . The dual projection is unique when V is convex. When $P^{*\perp}$ is the projection of P to V , the geodesic connecting P to $P^{*\perp}$ is orthogonal to the boundary surface of V . The projection is unique when V is dual convex. See Fig.2.

Corollary. The dual projection of P to a flat subspace is unique. The projection of P to a dual flat subspace is unique.

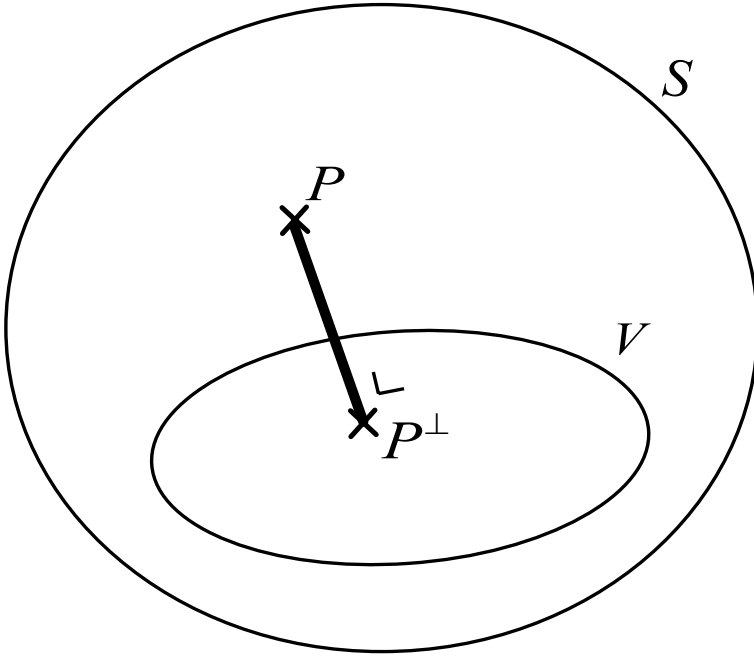


Fig. 2. Projection of P to V

Let V_1 and V_2 be two regions in S . Let us define the divergence (dual divergence) between V_1 and V_2 by

$$D(V_1 : V_2) = \min_{P \in V_1, Q \in V_2} D(P : Q) \quad (33)$$

$$D^*(V_2 : V_1) = \min_{P \in V_1, Q \in V_2} D^*(Q : P) \quad (34)$$

We search for the points $P^* \in V_1$ and $Q^* \in V_2$ such that

$$D(V_1 : V_2) = D^*(V_2 : V_1) = D(P^* : Q^*), \quad (35)$$

Then, the two extremal points P^* and Q^* satisfy the following: The dual projection of P^* to V_2 is Q^* , and the projection of Q^* to V_1 is P^* .

There is an iterative algorithm to find P^* and Q^* :

1. Start from $R_1 = R$
2. For $i = 1, 2, \dots$, dual-project R_i to V_2 , and let it be Q_i
3. Then, project Q_i to V_1 , and let it be R_{i+1}
4. End when $R_i = R_{i+1}$ (or equivalently $Q_i = Q_{i+1}$) is attained, or their divergence is sufficiently small.

This procedure is called the alternative minimization algorithm (Bregman, 1967). Csiszár and Tusnady (15) proposed this as an iterative procedure for the EM algorithm. The algorithm was studied in information geometry (Amari, 3), see also (Amari, Kurata and Nagaoka, 7) and (Byrne, 11) for application to Boltzmann machine. There are many recent applications in relation to the EM algorithm and Bregman divergence.

2.6 Tangent Space and Foliation

The tangent space T_P of S at point P is a local linearization of S at P (see Appendix). Since S is flat, all the tangent spaces look equivalent. However, S is a Riemannian space, so its metric structure g_{ij} depends on P .

T_P is a vector space. Its basis is spanned by n vectors, $e_1, \dots, e_n$ which represent the tangent vector of the coordinate curve $\theta^i, i = 1, \dots, n$. Such a basis is called the natural basis induced by the coordinate system θ (Fig. 3). The inner product of two basis vectors gives the Riemannian metric,

$$\langle e_i, e_j \rangle = g_{ij} \quad (36)$$

which in general depends on P .

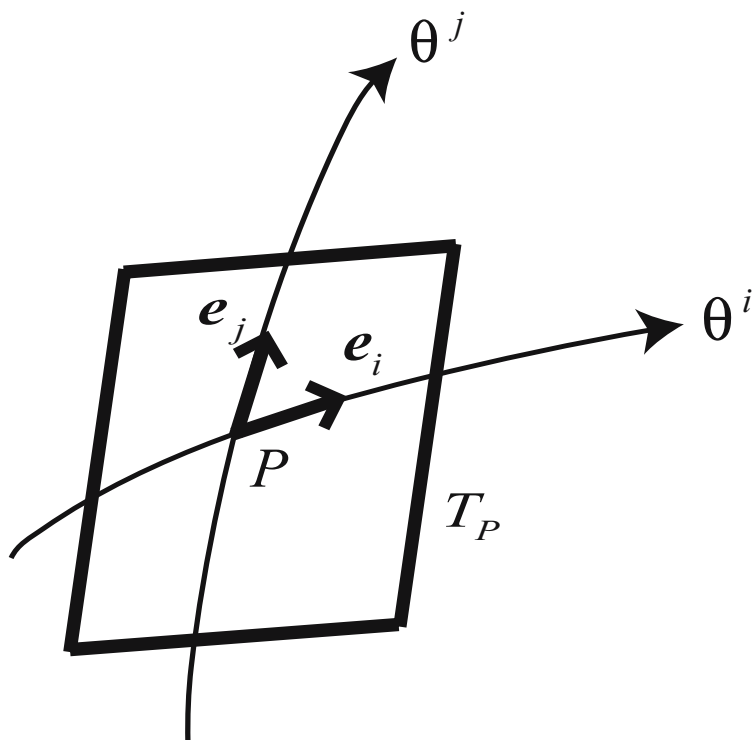


Fig. 3. Tangent space T_P and natural basis

Dually to the above, we have the natural basis with respect to the dual coordinate system, spanned by $\mathbf{e}_1^*, \dots, \mathbf{e}_n^*$ which are the tangent vectors of the coordinate curves η_i . The inner product of two basis vectors gives

$$\langle \mathbf{e}_i^*, \mathbf{e}_j^* \rangle = g^{ij} \quad (37)$$

which is the Riemannian metric in the dual coordinate system.

The two bases of T_P are reciprocal in the following sense.

Theorem 4. The two sets of basis vectors, are mutually orthogonal,

$$\langle \mathbf{e}_i, \mathbf{e}_j^* \rangle = \delta_i^j \quad (38)$$

Proof. Since we have

$$\mathbf{e}_j^* = \sum g^{ji} \mathbf{e}_i \quad (39)$$

it is easy to derive (38).

Let us fix p dual coordinates $\boldsymbol{\eta}_p = (\eta_1, \dots, \eta_p)$ in $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)$, $p < n$, and consider the $(n - p)$ -dimensional dually flat subspaces

$$M(\boldsymbol{\eta}_p) = \{P : \text{its first } p \text{ } \boldsymbol{\eta}\text{-coordinates are } (\eta_1, \dots, \eta_p)\}. \quad (40)$$

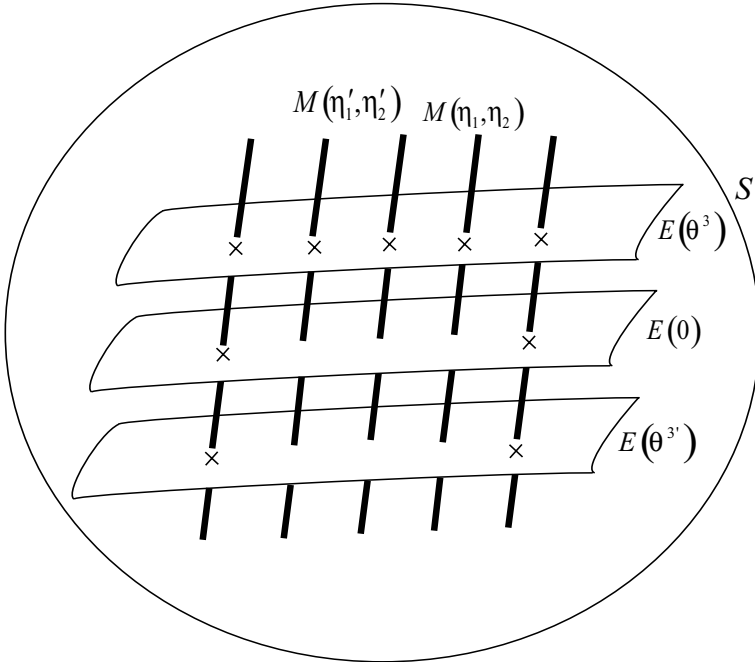


Fig. 4. Orthogonal foliation $(\eta_1, \eta_2; \theta^3)$

A family of subspaces $M(\boldsymbol{\eta}_p)$ for all $\boldsymbol{\eta}_p$, which do not overlap, and fills the entire space S . Such a family is called a foliation of S . Similarly, by fixing $n-p$ coordinates $\boldsymbol{\theta}_{n-p} = (\theta^{n-p+1}, \dots, \theta^n)$, and consider p dimensional flat subspace

$$E(\boldsymbol{\theta}_{n-p}) = \{P : \text{its last } n-p \text{ coordinates are } \boldsymbol{\theta}_{n-p}\}. \quad (41)$$

Their family gives another foliation of S .

The two foliations are mutually orthogonal, in the sense that the tangent vectors of the respective subspaces at the intersection are orthogonal. The intersection of $M(\boldsymbol{\eta}_p)$ and $E(\boldsymbol{\theta}_{n-p})$ is the point whose first p $\boldsymbol{\eta}$ -coordinates are $\boldsymbol{\eta}_p$ and the last $n-p$ coordinates are $\boldsymbol{\theta}_{n-p}$ (Fig. 4). We may use a new coordinate system $\boldsymbol{\xi} = (\boldsymbol{\eta}_p, \boldsymbol{\theta}_{n-p})$, which is very useful for applications.

The dual orthogonal geodesic foliation has lots of applications (Amari, 5).

2.7 α -Divergence and α -Geodesic

We have studied the dually flat structure of a manifold originated from a convex function, consisting of divergence function, Riemannian metric and dually flat affine connections. It is possible to extend this framework to the α -connections, where α is a real parameter (Zhang, 28). When $\alpha = 1$, it gives the original $\boldsymbol{\theta}$ structure and when $\alpha = -1$, it gives the dual $\boldsymbol{\eta}$ structure. We then have α -divergence and α -geodesic by extending our definition. Although the α -affine connection is not flat, but the α -connection and $-\alpha$ -connection are dually coupled.

When $\alpha = 1$, we have a flat affine connection, and when $\alpha = -1$, we have a dual flat affine connection. Therefore, the α -divergence connects the primal and dual divergences continuously. Similarly, the α -geodesic connects the primal and dual geodesics continuously.

The α -divergence (Zhang, 28) between P and P' is defined by

$$D_\alpha(P : P') = \frac{4}{1-\alpha^2} \left[1 - \exp \left\{ \psi \left(\frac{1+\alpha}{2} \boldsymbol{\theta} + \frac{1-\alpha}{2} \boldsymbol{\theta}' \right) - \frac{1+\alpha}{2} \psi(\boldsymbol{\theta}) - \frac{1-\alpha}{2} \psi(\boldsymbol{\theta}') \right\} \right]. \quad (42)$$

When $\alpha = \pm 1$, we take the limit $\alpha \rightarrow \pm 1$, and then have

$$D_{-1}(P : P') = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}') - \sum \theta_i \eta'_i, \quad (43)$$

$$D_1(P : P') = \psi(\boldsymbol{\theta}') + \varphi(\boldsymbol{\eta}) - \sum \theta'_i \eta_i. \quad (44)$$

The Riemannian metric derived from the α -divergence does not depend on α , giving the same metric.

The α -connection is derived from the α -divergence by calculating

$$\frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} D_\alpha(\boldsymbol{\theta} : \boldsymbol{\theta}')|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}. \quad (45)$$

This gives the α -connection in the primal coordinate system as

$$\Gamma_{ijk}^{(\alpha)} = \frac{1-\alpha}{2} T_{ijk}. \quad (46)$$

Here, T_{ijk} is a tensor defined by

$$T_{ijk} = \frac{\partial^3}{\partial \theta^i \partial \theta^j \partial \theta^k} \psi(\boldsymbol{\theta}) \quad (47)$$

and $\Gamma^{(\alpha)}$ is the coefficients of the α -connection.

The α -geodesic $\boldsymbol{\theta}(t)$ is given by the geodesic equation

$$\ddot{\theta}^i + \sum \Gamma_{jk}^{(\alpha)} \dot{\theta}^j \dot{\theta}^k = 0. \quad (48)$$

We do not discuss more details of the α -structure here, but it has wide applications. Sometimes, more robust results are obtained by using the α -projection instead of the primal or dual geodesic projections. See, for example, Matsushima (19).

3 Information Geometry of Probability Distributions

Let us consider a family of probability distributions $S = \{p(x, \boldsymbol{\xi})\}$, where x is a random variable and $\boldsymbol{\xi}$ is an n -dimensional vector parameter to specify a distribution. This is considered as an n -dimensional manifold, where $\boldsymbol{\xi}$ is coordinates, and a point $\boldsymbol{\xi}$ is regarded as a probability distribution $p(x, \boldsymbol{\xi})$.

In order to introduce an invariant geometrical structure to M , we require the following criterion:

Invariance Criterion: The geometry should be invariant under coordinate transformations of $\boldsymbol{\xi}$ and also under one-to-one transformations of random variable x .

When we transform x to y by

$$y = f(x), \quad x = f^{-1}(y), \quad (49)$$

the probability density function is changed from $p(x, \boldsymbol{\xi})$ to

$$\bar{p}(y, \boldsymbol{\xi}) = \frac{1}{\left| \frac{\partial y}{\partial x} \right|} p(f^{-1}(y), \boldsymbol{\xi}). \quad (50)$$

The above criterion requires that the geometrical structure is the same for $M = \{p(x, \boldsymbol{\xi})\}$ and $\bar{M} = \{\bar{p}(y, \boldsymbol{\xi})\}$.

Based on the criterion, we have the only invariant Riemannian metric, which is given by the Fisher information matrix (Chentsov, 12). We also have a one-parameter family of invariant affine connections.

Theorem 5. The invariant Riemannian metric is given by

$$g_{ij} = E \left[\frac{\partial}{\partial \xi_i} \log p(x, \boldsymbol{\xi}) \frac{\partial}{\partial \xi_j} \log p(x, \boldsymbol{\xi}) \right] \quad (51)$$

and the invariant affine connection parameterized by α is

$$\Gamma_{ijk}^{(\alpha)} = [ij; k] - \alpha T_{ijk} \quad (52)$$

where $[ij; k]$ is the Christoffel symbol calculated from g_{ij} and

$$T_{ijk} = E \left[\frac{\partial}{\partial \xi_i} \log p(x, \boldsymbol{\xi}) \frac{\partial}{\partial \xi_j} \log p(x, \boldsymbol{\xi}) \frac{\partial}{\partial \xi_k} \log p(x, \boldsymbol{\xi}) \right] \quad (53)$$

is an invariant tensor symmetric with respect to three indices.

The α - and $-\alpha$ -affine connections are dual, in the sense that the two covariant derivatives ∇^α and $\nabla^{-\alpha}$ satisfy

$$X \langle Y, Z \rangle = \langle \nabla_X^\alpha Y, Z \rangle + \langle \nabla_X^{-\alpha} Z, Y \rangle \quad (54)$$

for three vector fields X, Y and Z . We do not go into details here. See Appendix and (Amari and Nagaoka, 8).

It is also proved that, when the manifold is flat with respect to one affine connection, it is also flat with respect to the dual affine connection. Hence, we have a dually flat manifold. When a manifold is flat, we have a geodesic coordinate system $\boldsymbol{\theta}$ in which

$$\Gamma_{ijk}^{(1)}(\boldsymbol{\theta}) = 0. \quad (55)$$

Such a manifold has a convex potential function $\psi(\boldsymbol{\theta})$, by which two tensors are given by

$$g_{ij} = \frac{\partial^2}{\partial \theta^i \partial \theta^j} \psi(\boldsymbol{\theta}), \quad T_{ijk} = \frac{\partial^3}{\partial \theta^i \partial \theta^j \partial \theta^k} \psi(\boldsymbol{\theta}). \quad (56)$$

We also have a dual coordinate system, a dual potential function, and an invariant canonical divergence function $D(P : Q)$. An important family of probability distributions is the exponential family represented in the following form by taking an adequate dominating measure,

$$p(x, \boldsymbol{\theta}) = \exp \left\{ \sum \theta^i x_i - \psi(\boldsymbol{\theta}) \right\}. \quad (57)$$

In this case, the parameter $\boldsymbol{\theta}$ is called the canonical or natural parameter, and the function $\psi(\boldsymbol{\theta})$ is the cumulant generating function (in physics it is called the free energy), which is convex.

The manifold of an exponential family is dually flat, given by the convex function $\psi(\boldsymbol{\theta})$. The dual coordinates are the expectation parameters, $\boldsymbol{\eta}$

$$\eta_i = E[x_i]. \quad (58)$$

We have the dual structure explained in Section 2.

Another important dually flat manifold is a mixture family, which is written as

$$p(x, \xi) = \sum \xi_i p_i(x), \quad \sum \xi_i = 1, \quad \xi_i \geq 0, \quad (59)$$

where $p_i(x)$ are fixed probability distributions.

There are many probability distributions which are submanifolds of an exponential family. They are called curved exponential families. Other families exist which are not embedded in an exponential family. They also have a dual structure.

When random variable x is discrete, taking a finite number of values, $x = 0, 1, 2, \dots, n$, the set of all such distributions form an exponential family. This is because we have

$$p(x, \theta) = \exp \left\{ \sum \theta^i \delta_i(x) - \psi(\theta) \right\}, \quad (60)$$

where $\theta^i = \log p_i$ is the log probability of $x = i$ and $\delta_i(x)$ is the delta function

$$\delta_i(x) = \begin{cases} 1, & x = i, \\ 0, & \text{otherwise.} \end{cases} \quad (61)$$

This implies any parameterized family of probability distributions over discrete random variables is a curved exponential family, and is hence a submanifold in a dually flat manifold. This is the reason why a dually flat manifold is important.

Procedures of statistical inference, such as estimation and testing statistical hypothesis is well founded in the framework of information geometry.

Semiparametric statistical inference has been established in the fibre bundle theory of statistical inference (Amari and Kawanabe, 6).

4 Applications of Information Geometry

Plenty of applications of information geometry have been widely prevailing in many fields and are growing. Here, we mention only a simple list of applications.

1) Applications to statistical inference

Higher-order evaluation of statistical inference, including estimation and hypothesis testing, is the oldest applications (Amari, 1). Here, both of the primal e-curvature and dual m-curvature play a fundamental role. Other applications are given in the area of EM algorithm, semiparametric estimation and robust estimation.

2) Control theory and time series analysis

A linear system and a nonlinear system can be identified by the probability distribution of the system output when white noise is applied to its input. This gives one-to-one correspondence of a family of probability distributions and a

family of systems. Hence, the geometrical structure of the manifold of systems can be studied by information geometry (Amari, 2).

3) Applications to neural networks and systems

One application is found in multilayer perceptrons, where a set of perceptrons forms a Riemannian space. However, due to the symmetric structure of neurons in the hidden layer, the manifold is singular, and the Riemannian metric degenerates in symmetry regions (Amari, Park and Ozeki, 9). This causes serious difficulty in learning behaviors. We can overcome this difficulty by taking the Riemannian structure into account and modifying a learning algorithm to fit the Riemannian structure (the natural gradient method, Amari, 1998).

Another application is analysis of spike trains in a pool of neurons, where their firing is correlated. We can decompose correlated structure orthogonally in a sum of firing rates, pairwise correlations, third-order and higher-order correlations (Nakahara and Amari, 21). Dual geodesic foliations of information geometry plays a fundamental role in this decomposition.

4) Machine learning

Machine learning deals with stochastic situations and extracts necessary information from examples. In the case of a graphical model, the belief propagation algorithm uses e- and m-projections effectively (Ikeda, Tanaka and Amari, 16).

Another application is in the analysis of the boosting method of combining weak learners (Lebanon and Lafferty, 18; Murata et al., 20; Lebanon, 17). We can use the geometric idea of conformal transformation for improving the kernel function in the area of support vector machines (Wu and Amari, 27).

5) Convex programming

In the case of convex programming such as linear programming and semi-definite programming, the inner method can be used effectively. A convex potential function is given from the convex region as a barrier function from which the dual structure is introduced. The m-curvature plays a fundamental role in evaluating the complexity of the algorithm (Ohara, 23; Ohara and Tsuchiya, 24).

6) Signal processing and ICA (independent component analysis)

Independent component analysis is a technique to extract hidden signals from their mixtures. Here, the geometry of the manifold of mixing and unmixing matrices plays a fundamental role, together with the manifold of joint probability distributions which includes the submanifold of independent distributions in its inside (Cichocki and Amari, 13).

7) Other applications

We can find many applications in mathematics, physics and information theory.

4.1 Conclusions

Information geometry, I-projection, and alternative procedure of dual projections are important tools of frequent use. The present review paper constructs a dually flat manifold starting from a convex function. We have elucidated its du-

alistic differential-geometrical structure without going into details of differential geometry. This is possible because the manifold is dually flat. However, deeper mathematical framework would be required when we go further, and the present paper plays a role of introduction to information geometry.

It is also useful to point out the relation of information geometry to affine differential geometry (Nomizu and Sasaki, 22). A potential function $\psi(\boldsymbol{\theta})$ is regarded as an n -dimensional submanifold defined by $\boldsymbol{\theta}^0 = \psi(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^n)$ in the extended space $(\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^n)$. Then, affine differential geometry studies the geometry of this surface, where the space is regarded as an affine space. We may note that the Rényi entropy (Rényi, 25) and Tsallis entropy (Tsallis, 26) are closely related to the α -geometry.

Applications of information geometry are expanding. One can see some of them in the present monograph, e.g., applications to signal processing, vision analysis, shape analysis and others.

References

- Amari, S.: Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics, vol. 28. Springer, Heidelberg (1985)
- Amari, S.: Differential geometry of a parametric family of invertible linear systems-Riemannian metric, dual affine connections and divergence. *Mathematical Systems Theory* 20, 53–82 (1987)
- Amari, S.: Information geometry of the EM and em algorithms for neural networks. *Neural Networks* 8-9, 1379–1408 (1995)
- Amari, S.: Natural gradient works efficiently in learning. *Neural Computation* 10, 251–276 (1998)
- Amari, S.: Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory* 47, 1701–1711 (2001)
- Amari, S., Kawanabe, M.: Information geometry of estimating functions in semi parametric statistical models. *Bernoulli*. 3(1), 29–54 (1997)
- Amari, S., Kurata, K., Nagaoka, H.: Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks* 3, 260–271 (1992)
- Amari, S., Nagaoka, H.: *Methods of Information Geometry*. Translations of Mathematical Monographs, vol. 191. AMS & Oxford University Press (2000)
- Amari, S., Park, H., Ozeki, T.: Singularities affect dynamics of learning in neuromanifolds. *Neural Computation* 18, 1007–1065 (2006)
- Bregman, L.M.: 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)
- Byrne, W.: Alternating minimization and Boltzmann machine learning. *IEEE Transactions on Neural Networks* 3, 612–620 (1992)
- Chentsov (Čencov), N.N.: *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, Rhode Islandm U.S.A. (1982); originally published in Russian, Nauka, Moscow (1972)
- Cichocki, A., Amari, S.: *Adaptive Blind Signal and Image Processing*. John Wiley, Chichester (2002)
- Csiszár, I.: I -divergence geometry of probability distributions and minimization problems. *The Annals of Probability* 3, 146–158 (1975)

- Csiszár, I., Tusnády, G.: Information geometry and alternating minimization procedures. In: Dedewicz, E.F., et al. (eds.) *Statistics and Decisions*, vol. (1), pp. 205–237. R. Oldenbourg Verlag, Munich (1984)
- Ikeda, S., Tanaka, T., Amari, S.: Stochastic reasoning, free energy, and information geometry. *Neural Computation* 16, 1779–1810 (2004)
- Lebanon, G.: *Riemannian Geometry and Statistical Machine Learning*. CMU-LTI-05-189, Carnegie-Mellon University (2005)
- Lebanon, G., Lafferty, J.: Boosting and maximum likelihood for exponential models. *Advances in Neural Information Processing Systems*, vol. 14, pp. 447–451. MIT Press, Cambridge (2002)
- Matsushima, Y.: The alpha EM algorithms: Surrogate likelihood maximization using alpha-logarithmic information measures. *IEEE Transactions on Information Theory* 49, 692–706 (2003)
- Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S.: Information geometry of U -boost and Bregman divergence. *Neural Computation* 16, 1437–1481 (2004)
- Nakahara, H., Amari, S.: Information-geometric measure for neural spikes. *Neural Computation* 14, 2269–2316 (2002)
- Nomizu, K., Sasaki, T.: *Affine Differential Geometry*. Cambridge University Press, Cambridge (1994)
- Ohara, A.: Geodesics for dual connections and means on symmetric cone. *Integral Equation and Operator Theory* 50, 537–548 (2004)
- Ohara, A., Tsuchiya, T.: An information geometric approach to polynomial-time interior-point algorithms (submitted, 2008)
- Rényi, A.: On measures of entropy and information. In: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 547–561. University of California Press (1961)
- Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* 52, 479 (1988)
- Wu, S., Amari, S.: Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers. *Neural Processing Letters* 15, 59–67 (2002)
- Zhang, J.: Divergence function, duality, and convex analysis. *Neural Computation* 16, 159–195 (2004)

Appendix

Introduction to Differential Geometry — Intuitive Explanation

We give a brief and intuitively-understandable introduction to differential geometry with dual affine connections. We do not enter in mathematically rigorous definitions and explanations but use intuitive ideas to understand the concepts.

For more rigorous treatments, refer to Boothby (2003), Kobayashi and Nomizu (1969), and Amari and Nagaoka (2000).

A Manifold and Coordinates

An n -dimensional manifold S is a space in which a neighborhood of each point has an n -dimensional Euclidean-like structure. This means that S is a curved

version of a Euclidean space. In order to specify a point in S , we use a coordinate system

$$\boldsymbol{\xi} = (\xi^1, \dots, \xi^n). \quad (62)$$

There are many coordinate systems, for example,

$$\boldsymbol{\zeta} = (\zeta^1, \dots, \zeta^n) \quad (63)$$

in which $\zeta^i, i = 1, \dots, n$, are smooth and invertible functions of $\boldsymbol{\xi}$. Any such coordinate system is eligible.

A manifold S may have global topological structure different from a Euclidean space. For example, a sphere or a torus is a two-dimensional manifold having a different topological structure from a Euclidean space, although a small neighborhood of each point looks like a two-dimensional Euclidean space. In such a case a single coordinate system might not cover the entire manifold. We use in such a case a number of coordinate systems to cover it. Global structure of a manifold is an important topic of geometry and topology. However, we study here only local structure of a manifold covered by a single coordinate system. We denote by $\{S, \boldsymbol{\xi}\}$ the manifold S with coordinate system $\boldsymbol{\xi}$.

B Tangent Space

Let us consider an infinitesimally small neighborhood of point P in manifold S . Even when S is curved, we may regard a small neighborhood of P as a flat space. This is linear approximation of S in a small area around P . We show one example. The earth is a ball, and its surface is a sphere which is curved. However, we may regard its local area, say Tokyo area, as a flat surface approximately. This idea leads us to the concept of the tangent space T_P at P .

T_P is a linear space, which is locally an approximation of S around P . When one think of a curved surface in a 3-dimensional Euclidean space, we can easily think of a tangent space T_P at any point P . When S is curved, T_P depends on P and T_P and $T_{P'}$ are different in general, that is, having different two-dimensional orientations, when $P \neq P'$.

T_P is a vector space whose origin is at P . It is spanned by n vectors

$$\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\} \quad (64)$$

which form a basis. We may use any basis provided $\mathbf{e}_i$ are linearly independent. When we use a coordinate system $\boldsymbol{\xi}$, a convenient basis is to use n tangent vectors of the coordinate lines ξ^i . This is because n coordinate lines have different directions, and their tangent lines are linearly independent. Such a basis is called the natural basis associated with $\{S, \boldsymbol{\xi}\}$.

Mathematicians like to denote a natural basis vector $\mathbf{e}_i$ by the differential operators,

$$\partial_i = \frac{\partial}{\partial \xi^i}. \quad (65)$$

Given a function $f(\boldsymbol{\xi})$ on the manifold, $\partial_i f(\boldsymbol{\xi})$ denotes a rate of change of function f in the direction of the coordinate line ξ^i . Hence, they identify the direction $\mathbf{e}_i$ with its directional differentiation.

Any tangent vector $X \in T_P$ is written as

$$X = \sum X^i \mathbf{e}_i \quad (66)$$

by using a basis $\{\mathbf{e}_i\}$. The coefficients X^i are the components of X in this basis $\{\mathbf{e}_i\}$. By using different basis $\{\mathbf{e}'_i\}$, we have

$$X = \sum X'^i \mathbf{e}'_i \quad (67)$$

and components change, although the vector X is the same. So components depend on the basis.

From mathematician's point of view, a vector

$$X = \sum X^i \mathbf{e}_i \quad (68)$$

is a differential operator, where

$$Xf = \sum X^i \partial_i f(\boldsymbol{\xi}) \quad (69)$$

gives a directional derivative of function $f(\boldsymbol{\xi})$.

Let us come back to the idea of local approximation. Let P be a point whose coordinates are $\boldsymbol{\xi}$, and let P' is a nearby point whose coordinates are $\boldsymbol{\xi} + d\boldsymbol{\xi}$. Then, the line element PP' is regarded as a tangent vector

$$PP' = dP = \sum d\xi^i \mathbf{e}_i \quad (70)$$

where we use the natural basis. Then, we have

$$dPf = \sum d\xi^i \partial_i f = df \quad (71)$$

giving the total differential of function in the direction of dP .

C Riemannian Manifold

When the tangent space is Euclidean, we have the inner product of two tangent vectors. In particular, the square of its length is defined by the inner product of itself. When each tangent space is equipped with the inner product, the space is said to be Riemannian. A Euclidean space is a special case of the Riemannian manifold.

Let $G = (g_{ij})$ be a matrix defined by the inner product of basis vectors,

$$g_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle. \quad (72)$$

It is a positive-definite symmetric matrix called the Riemannian metric tensor.

The inner product of two vectors

$$X = \sum X^i e_i, \quad Y = \sum Y^j e_j \quad (73)$$

is given in their component form by

$$\langle X, Y \rangle = \sum X^i Y^j g_{ij}. \quad (74)$$

In particular, the square of the absolute length of a vector X is

$$\|X\|^2 = \langle X, X \rangle = \sum X^i X^j g_{ij}. \quad (75)$$

Two vectors X and Y are said to be orthogonal, when their inner product vanishes,

$$\langle X, Y \rangle = 0. \quad (76)$$

Let us consider two curves $\xi(t)$ and $\xi'(t)$ parameterized by t , such that they intersect at $t = 0$. Their tangent vectors are denoted by

$$\dot{\xi}(t) = \sum \frac{d}{dt} \xi^i(t) e_i, \quad \dot{\xi}' = \sum \frac{d}{dt} \xi'^i(t) e_i. \quad (77)$$

The two curves are said to be orthogonal when the inner product of the two tangent vectors vanishes,

$$\langle \dot{\xi}(0), \dot{\xi}'(0) \rangle = 0. \quad (78)$$

Let PP' be again a small line-element of S . The squared length of this line-element is given by

$$ds^2 = \langle d\xi, d\xi \rangle = \sum d\xi^i d\xi^j g_{ij}. \quad (79)$$

We now consider two points P and Q which are connected by a curve $\xi(t)$, where P has coordinates $\xi(1)$ and Q has coordinates $\xi(0)$. Their length along the curve is given by the integral

$$s = \int_0^1 \sqrt{\sum \dot{\xi}^i(t) \dot{\xi}^j(t) g_{ij}} dt. \quad (80)$$

This obviously depends on the curve connecting them.

We sometimes consider a surface or m -dimensional submanifold M in $S(m < n)$. It is specified in a parametric form

$$\xi = \xi(u_1, \dots, u_m). \quad (81)$$

Then, we consider the tangent space T_P^M of M at $P \in M$. This is a subspace of the tangent space of S at P ,

$$T_P^M \subset T_P. \quad (82)$$

The tangent space T_P^M is spanned by m vectors

$$e_1^M = \frac{\partial \xi}{\partial u_1}, \dots, e_m^M = \frac{\partial \xi}{\partial u_m}. \quad (83)$$

Let us consider a curve $\xi(t)$ which transverses M at point P . Then, it is said to be orthogonal to M when the tangent vector $\xi(t)$ is orthogonal to T_P^M at P ,

$$\langle \dot{\xi}, e_i^M \rangle = 0, \quad i = 1, \dots, m. \quad (84)$$

D Affine Connection

We have so far studied the metric structure of the tangent space of S at each point separately. This looks like cutting S into small local pieces, and study each piece separately. Since S is curved in general, tangent spaces of T_P and $T_{P'}$ are different. However, the manifold is smoothly connected, and when the point deviates from P slightly to P' , the tangent space T_P changes slightly to $T_{P'}$. By studying how they change, we may recover the curved structure of S by connecting these neighboring tangent spaces.

Let us consider two tangent spaces T_P and $T_{P'}$, where $P' = P + dP$ is infinitesimally close to P . They are very similar, and become identical as P' approaches P . But $T_{P'}$ is different from T_P . So we map basis vector e_i of T_P to $T_{P'}$. The counterpart of e_i in $T_{P'}$ is given by this mapping. Since the manifold and its coordinates are smooth, we may consider e_i is very close to the basis vector e'_i of $T_{P'}$ at P' . So we assume that e_i is mapped to

$$\tilde{e}_i = e'_i + \delta e_i \in T_{P'}, \quad (85)$$

where the deviation term δe_i is small (Fig. A1). When the manifold is flat like a Euclidean space, by using a linear coordinate system, e_i and e'_i are exactly the same. However, when the coordinate system is curved (e.g., a polar coordinate system), e_i and e'_i are not identical even in a Euclidean space. Hence, e_i is mapped to $T_{P'}$ in the above form even in the case of a Euclidean space.

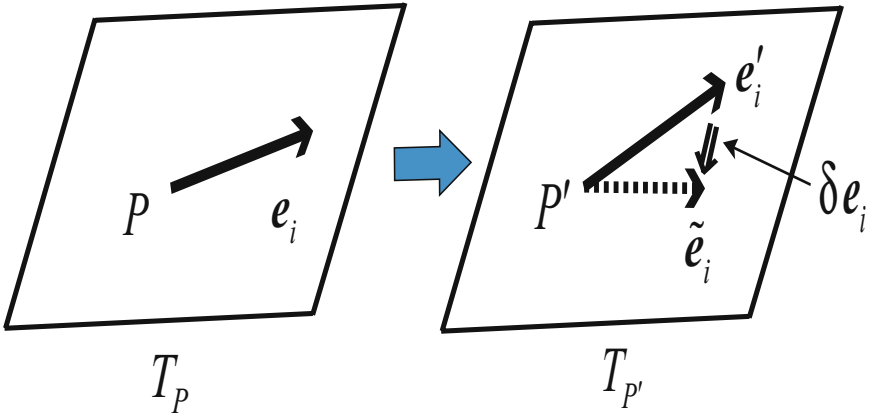


Fig. A1. Affine mapping of $e_i \in T_P$ to $T_{P'}$

Since $\delta \mathbf{e}_i$ is a vector in $T_{P'}$, we represent it in the component form,

$$\delta \mathbf{e}_i = \sum_k \delta_i^k \mathbf{e}'_k. \quad (86)$$

The components δ_i^k are small, and depend on the difference $d\xi$ of P and P' , that is in which direction P' deviates from P . The correction term becomes 0 as $d\xi$ becomes 0. Hence, when $d\xi$ is infinitesimally small, the linear form in $d\xi$ suffices. So we express $\delta \mathbf{e}_i$ in the form

$$\delta \mathbf{e}_i = - \sum_{j,k} \Gamma_{ji}^k d\xi^j \mathbf{e}'_k \quad (87)$$

This mapping establishes the relation between two local linear approximations of S at P and P' , because the relation between $\mathbf{e}_i \in T_P$ and $\mathbf{e}'_i \in T_{P'}$ are established. The relation is represented by a quantity Γ_{ij}^k having three indices i, j, k . This correspondence is called an affine connection, and the quantities Γ_{ji}^k are called the coefficients of affine connection. They are given by using the inner product as

$$\langle \delta \mathbf{e}_i, \mathbf{e}'_k \rangle = - \sum_{j,m} \Gamma_{ji}^m g_{mk} d\xi^j. \quad (88)$$

We put

$$\Gamma_{jik} = \sum_m \Gamma_{ji}^m g_{mk}. \quad (89)$$

We consider the case where

$$\Gamma_{jik} = \Gamma_{ijk} \quad (90)$$

holds. When this is not the case, S is said to have torsion. We do not treat the case with torsion in this appendix.

Once the mapping is established for the basis vectors, it is generalized to the mapping of any vector

$$\mathbf{X} = \sum X^i \mathbf{e}_i \in T_P. \quad (91)$$

It is mapped to $T_{P'}$ by

$$\tilde{\mathbf{X}} = \sum X^i \tilde{\mathbf{e}}_i = \sum X^i \left(\mathbf{e}'_i - \sum \Gamma_{ji}^k \mathbf{e}'_k d\xi^j \right) \in T_{P'}. \quad (92)$$

A curved manifold is regarded as a collection $\{T_P, P \in S\}$ of local linear approximations. They are connected by an affine connection, and the entire curved structure is recovered by it.

E Geodesic and Parallel Transport

Let us consider a curve

$$\xi = \xi(t), \quad (93)$$

where $P = \xi(t)$ and $P' = \xi(t + dt)$. Its tangent vector at P and P' are

$$\dot{\xi}(t) = \sum \dot{\xi}^i(t) e_i, \quad \dot{\xi}(t + dt) = \sum \dot{\xi}^i(t + dt) e'_i, \quad (94)$$

respectively. When the tangent vector $\dot{\xi}(t)$ at P is mapped to $T_{P'}$, it is in general different from the tangent vector $\dot{\xi}(t + dt)$ at P' . When they are the same, the curve does not change its tangent direction as t increases.

The tangent vector $\dot{\xi}(t)$ at P is mapped to

$$\tilde{\xi}(t) = \sum \dot{\xi}^i(t) e'_i - \sum \Gamma_{ji}^k d\xi^j e'_k \dot{\xi}^i(t) \quad (95)$$

of $T_{P'}$. However, the tangent vector $\dot{\xi}(t + dt)$ at $T_{P'}$ is

$$\dot{\xi}(t + dt) = \sum \left(\dot{\xi}^i(t) + \ddot{\xi}^i(t) dt \right) e'_i. \quad (96)$$

Hence, they are equal when

$$\ddot{\xi}^i + \sum \Gamma_{jk}^i \dot{\xi}^j \dot{\xi}^k = 0. \quad (97)$$

That is, when a curve $\xi(t)$ satisfies the above equation, its tangent vectors are always the same by this affine correspondence.

Such a curve is called a geodesic. That is, a geodesic is a curve not changing its direction, so that it is an extension of a straight line.

Here, we assume that the tangent vector does not change its length, too. When we reparameterize t , this curve does not change its direction but its length may change. Here we consider only the case that both the direction and length do not change.

A vector

$$X = \sum X^i e_i \quad (98)$$

at P is mapped to

$$\tilde{X} = \sum_k X^k \left(e'_k - \sum_j \Gamma_{ji}^k d\xi^j e'_k \right) \quad (99)$$

at P' . Let us consider a curve $\xi(t)$ connecting two points P and Q , where P is denoted by $\xi(0)$ and Q by $\xi(1)$. Given vector $X(0)$ at P , we map it to a neighboring point P' along the curve. We may continue this process such that $X(t)$ is mapped to $X(t + dt)$

$$X(t + dt) = \tilde{X}(t) = \sum X^k(t) e'_k - \sum_{i,k} \Gamma_{ji}^k \dot{\xi}^j X^i(t) e'_k \quad (100)$$

along the curve. Since, we have

$$X^i(t + dt) = \sum X^i + \frac{\partial}{\partial \xi^j} X^i \dot{\xi}^j dt, \quad (101)$$

such a tangent vector $X(t)$ defined on the curve satisfies

$$\dot{X}^i(t) + \sum \Gamma_{jk}^i \dot{\xi}^j X^k = 0. \quad (102)$$

We say that $X(t)$ is a parallel transport of $X(0)$ along the curve (Fig. A2). A geodesic is the curve where $X(t) = \dot{\xi}(t)$ is the tangent vector of the curve $\xi(t)$.

Let $X(t)$ satisfy (102). This defines a mapping of $X(0)$ at P to $X(1)$ at Q along the curve $\xi(t)$. We call $X(1)$ the parallel transport of $X(0)$ along the curve, and denote it as

$$X(1) = \prod_{\xi(t)} X(0). \quad (103)$$

This obviously depends on which curve, connecting P and Q , is used to transport it. When the transport does not depend on the curve, the manifold is said to be flat. When S is flat, we have a coordinate system such that the coordinate curves are geodesics. Therefore, S is an affine space, except for the metric structure.

A manifold is flat when its Riemann-Christoffel curvature vanishes, where the curvature is calculated from the components of an affine connection.

F Covariant Derivative

We can compare two vectors defined at different positions by using an affine connection, although the tangent spaces they belong to are different. Here we introduce the notion of the covariant derivative, which makes the differentiation or derivative of a vector function possible. This also makes it possible to have derivative of a tensor, but we do not mention it.

Let X be a vector field, that is a vector-valued function defined on S . It is represented as

$$X = \sum X^i(\xi) e_i(\xi), \quad (104)$$

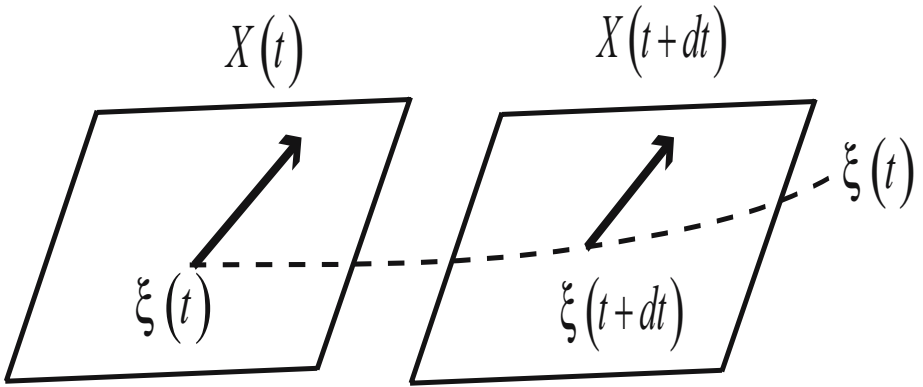


Fig. A2. Parallel transport $X(t)$

where $e_i(\xi)$ are the natural basis vectors at point ξ . We want to know how $X(\xi)$ changes as the position P moves to P' , where their coordinates are ξ and $\xi + d\xi$, respectively. This is given by the derivative of X in the direction of $d\xi$. However, it is not mere derivatives of the components X , because the basis vectors changes from P to P' . Hence in order to have an intrinsic change of vector X , we need to subtract from $X(\xi + d\xi)$ the map of $X(\xi)$. So the intrinsic change of vector X is given by

$$\delta X^i = X^i(\xi + d\xi) - (X^i(\xi) - X^i \delta e_i) \quad (105)$$

Letting $d\xi$ small, the derivative of X in the direction of the coordinate curve ξ^i is given by

$$\nabla_i X(\xi) = \sum_k \left(\frac{\partial}{\partial \xi^i} X^k(\xi) + \sum_j \Gamma_{ij}^k X^j \right) e_k \quad (106)$$

This is the covariant derivative of X in the direction of ξ^i . In general, the covariant derivative of X in the direction of vector Y is denoted by $\nabla_Y X$, and is

$$\nabla_Y X = \frac{\partial}{\partial \xi^i} X^k Y^i + \Gamma_{ij}^k Y^i X^j. \quad (107)$$

By using the covariant derivative, the equation of a geodesic is written as

$$\nabla_{\dot{\xi}} \dot{\xi} = 0. \quad (108)$$

Similarly, when vector $X(t)$ is parallelly transported along a curve $\xi(t)$, we have

$$\nabla_{\dot{\xi}} X = 0. \quad (109)$$

G Riemannian Connection

We have defined an affine connection independently of the metric structure. An affine connection is related to the metric by the metric condition which we state now. We will see a more general idea of dual connections in the next section, which is also related to the metric.

Metric condition: The length of a vector X does not change by the parallel transport.

Under the metric condition, the affine connection is uniquely determined from the metric, provided S does not have torsion. This unique connection is called the Levi-Civita connection or Riemannian connection. It is a good exercise to derive the following theorem.

Theorem A1. The Riemannian connection is given by

$$\Gamma_{ijk} = [ij; k] = \frac{1}{2} \left(\frac{\partial}{\partial \xi^i} g_{jk} + \frac{\partial}{\partial \xi^j} g_{ik} - \frac{\partial}{\partial \xi^k} g_{ij} \right) \quad (110)$$

where $[ij; k]$ is called the Christoffel symbol.

A straight line connects two points P and Q by the shortest distance in a Euclidean space. The notions of straightness and minimality of distance coincide when we use the Riemannian connection.

Theorem A2. The minimal distance line connecting two points is given by the geodesic connecting them.

Proof is also an exercise, and is obtained by calculating the variational equation of (80).

This nice property connecting the straightness and minimal distance is valid only under the metric condition. When we use another affine connection, it does not hold. However, as we will see, there is nice dualistic extension of this.

H Dual Connections

Let us consider a manifold having a Riemannian metric g_{ij} and two affine connections $\Gamma_{ijk}, \Gamma_{ijk}^*$. We denote the two connections by their covariant derivatives ∇, ∇^* or equivalently by the two parallel transport operators Π, Π^* . Each of the affine connections does not satisfy the metric condition in general. However, a pair of affine connections are said to be dual with respect to the metric, when they satisfy the following dual metric condition.

Dual metric condition: The inner product of two vectors X and Y does not change when X is parallelly transported from P to Q by one affine connection and Y by the other affine connection,

$$\langle X, Y \rangle_P = \left\langle \Pi_{\xi} X, \Pi_{\xi_Y}^* \right\rangle_Q \quad (111)$$

When $X = Y$, we have

$$\langle X, X \rangle_P = \left\langle \Pi_{\xi} X, \Pi_{\xi}^* X \right\rangle_Q \quad (112)$$

Hence, this is a generalization of the metric condition. When the affine connection happens to be self-dual, that is $\nabla = \nabla^*$, the dual metric condition reduces to the metric condition.

A space with dual affine connections is specified by its structure

$$\{S, \nabla, \nabla^*, g\}. \quad (113)$$

We have the differential version of the dual metric condition: Two affine connections ∇ and ∇^* are dual with respect to the metric, when the following equation holds for three vector fields:

$$X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X^* Z \rangle, \quad (114)$$

where $\langle Y, Z \rangle$ is the inner product of Y and Z ,

$$\langle Y, Z \rangle = \sum g_{ij} Y^i Z^j. \quad (115)$$

In terms of the coefficients of affine connections, the duality holds when

$$\frac{\partial}{\partial \xi^i} g_{jk} = \Gamma_{ijk} + \Gamma_{ikj}^*. \quad (116)$$

I Dually Flat Manifold

When S is flat with respect to one affine connection, we can prove that it is automatically flat with respect to the dual affine connection. We call such a space a dually flat manifold.

A dually flat manifold plays a role of the Euclidean space in Riemannian geometry: Any n -dimensional Riemannian space is embedded in $n(n+1)/2$ -dimensional Euclidean space as its curved submanifold. Similarly, a manifold with dual affine connections can be embed in a dually flat manifold of high dimensions, possibly of infinite dimensions.

The following is a fundamental theorem on the dually flat manifold.

Theorem A3. A dually flat manifold S has two special coordinate systems denoted by $\boldsymbol{\theta} = (\theta^1, \dots, \theta^n)$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)$ such that $\boldsymbol{\theta}$ is an affine coordinate system of ∇ -connection and $\boldsymbol{\eta}$ is an affine coordinate system of ∇^* -connection. There exist two potential functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ which are strictly convex, and are connected by the Legendre transformation such that

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \sum \theta^i \eta_i = 0, \quad (117)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are the respective coordinates of the same point. S has a canonical divergence between two points P and Q defined by

$$D[P : Q] = \psi(\boldsymbol{\theta}_P) + \varphi(\boldsymbol{\eta}_Q) - \sum \theta_P^i \eta_{Qi} \quad (118)$$

where $\boldsymbol{\theta}_P$ and $\boldsymbol{\eta}_Q$ are respective coordinates of points P and Q .

Theorem A4. The canonical divergence function of a dually flat manifold satisfies the Pythagorean relation,

$$D = [P : R] = D[P : Q] + D[Q : R] \quad (119)$$

when ∇^* -geodesic connection P and Q is orthogonal at Q to ∇ -geodesic connecting Q and R .

J α -Connections

Given a dual manifold $\{S, \nabla, \nabla^*, g_{ij}\}$, we can construct the α -connection by

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2} \nabla + \frac{1-\alpha}{2} \nabla^*. \quad (120)$$

A family of the α -connections connect ∇ -and ∇^* -connections by one-parameter family. Even when ∇ -and ∇^* -connections are flat, the intermediate $\nabla^{(\alpha)}$ connections are not necessarily flat.

It is remarkable that $\alpha = 0$ connection is the Riemannian connection. Hence starting with a Riemannian space with 0-connection, straightness is split into two pairs of dual straight concepts by the α -connections, which satisfy the dual metric condition. This is possible when and only when we have symmetric tensor T_{ijk} . Then, we can define the induced connection ∇ , which satisfies

$$T_{ijk} = \nabla_i g_{jk}, \quad (121)$$

and is symmetric with respect to three indices i, j, k .

Given the metric g_{ij} and affine connection ∇ , we define its derivative

$$T_{ijk} = \nabla_i g_{jk}. \quad (122)$$

When this vanishes, ∇ is the Riemannian connection satisfying the metric condition. When T_{ijk} is symmetric with respect to the three indices, we have a torsion-less dual connection

$$\Gamma_{ijk}^* = \Gamma_{ijk} - T_{ijk}, \quad (123)$$

and manifold $\{S, \nabla, \nabla^*, g_{ij}\}$ has dual structure. We construct the α -connection by

$$\Gamma_{ijk}^{(\alpha)} = \Gamma_{ijk}^{(0)} + \frac{\alpha}{2} T_{ijk}, \quad (124)$$

where $\Gamma_{ijk}^{(0)}$ with $\alpha = 0$ is the Riemannian connection.

The α -connection sometimes plays an interesting role in applications.