Applied Mathematical Sciences

Shun-ichi Amari

Information Geometry and Its Applications

Springer

Applied Mathematical Sciences

Volume 194

More information about this series at http://www.springer.com/series/34

Shun-ichi Amari

Information Geometry and Its Applications

Springer

Shun-ichi Amari
Brain Science Institute
RIKEN
Wako, Saitama
Japan

ISSN 0066-5452 ISSN 2196-968X (electronic)
Applied Mathematical Sciences
ISBN 978-4-431-55977-1 ISBN 978-4-431-55978-8 (eBook)
DOI 10.1007/978-4-431-55978-8

Library of Congress Control Number: 2015958849

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer Japan KK

Preface

Information geometry is a method of exploring the world of information by means of modern geometry. Theories of information have so far been studied mostly by using algebraic, logical, analytical, and probabilistic methods. Since geometry studies mutual relations between elements such as distance and curvature, it should provide the information sciences with powerful tools.

Information geometry has emerged from studies of invariant geometrical structure involved in statistical inference. It defines a Riemannian metric together with dually coupled affine connections in a manifold of probability distributions. These structures play important roles not only in statistical inference but also in wider areas of information sciences, such as machine learning, signal processing, optimization, and even neuroscience, not to mention mathematics and physics.

It is intended that the present monograph will give an introduction to information geometry and an overview of wide areas of application. For this purpose, Part I begins with a divergence function in a manifold. We then show that this provides the manifold with a dually flat structure equipped with a Riemannian metric. A highlight is a generalized Pythagorean theorem in a dually flat information manifold. The results are understandable without knowledge of differential geometry.

Part II gives an introduction to modern differential geometry without tears. We try to present concepts in a way which is intuitively understandable, not sticking to rigorous mathematics. Throughout the monograph, we do not pursue a rigorous mathematical basis but rather develop a framework which gives practically useful and understandable descriptions.

Part III is devoted to statistical inference, where various topics will be found, including the Neyman–Scott problem, semiparametric models, and the EM algorithm. Part IV overviews various applications of information geometry in the fields of machine learning, signal processing, and others.

Allow me to review my own personal history in information geometry. It was in 1958, when I was a graduate student on a master's course, that I followed a seminar on statistics. The text was "Information Theory and Statistics" by S. Kullback, and a professor suggested to me that the Fisher information might be regarded as a Riemannian metric. I calculated the Riemannian metric and curvature of the manifold of Gaussian distributions and found that it is a manifold of constant curvature, which is no different from the famous Poincaré half-plane in non-Euclidean geometry. I was enchanted by its beauty. I believed that a beautiful structure must have important practical significance, but I was not able to pursue its consequences further.

Fifteen years later, I was stimulated by a paper by Prof. B. Efron and accompanying discussions by Prof. A.P. Dawid, and restarted my investigation into information geometry. Later, I found that Prof. N.N. Chentsov had developed a theory along similar lines. I was lucky that Sir D. Cox noticed my approach and organized an international workshop on information geometry in 1984, in which many active statisticians participated. This was a good start for information geometry.

Now information geometry has been developed worldwide and many symposia and workshops have been organized around the world. Its areas of application have been enlarged from statistical inference to wider fields of information sciences.

To my regret, I have not been able to introduce many excellent works by other researchers around the world. For example, I have not been able to touch upon quantum information geometry. Also I have not been able to refer to many important works, because of my limited capability.

Last but not least, I would like to thank Dr. M. Kumon and Prof. H. Nagaoka, who collaborated in the early period of the infancy of information geometry. I also thank the many researchers who have supported me in the process of construction of information geometry, Profs. D. Cox, C.R. Rao, O. Barndorff-Nielsen, S. Lauritzen, B. Efron, A.P. Dawid, K. Takeuchi, and the late N.N. Chentsov, among many many others. Finally, I would like to thank Ms. Emi Namioka who arranged my handwritten manuscripts in the beautiful TEX form. Without her devotion, the monograph would not have appeared.

April 2015 Shun-ichi Amari

Contents

Part I Geometry of Divergence Functions: Dually Flat Riemannian Structure

Part II Introduction to Dual Differential Geometry

Part III Information Geometry of Statistical Inference

Part IV Applications of Information Geometry

Part I
Geometry of Divergence Functions: Dually Flat Riemannian Structure

Chapter 1
Manifold, Divergence and Dually Flat Structure

The present chapter begins with a manifold and a coordinate system within it. Then, a divergence between two points is defined. We use an intuitive style of explanation for manifolds, followed by typical examples. A divergence represents a degree of separation of two points, but it is not a distance since it is not symmetric with respect to the two points. Here is the origin of dually coupled asymmetry, leading us to a dual world. When a divergence is derived from a convex function in the form of the Bregman divergence, two affine structures are induced in the manifold. They are dually coupled via the Legendre transformation. Thus, a convex function provides a manifold with a dually flat affine structure in addition to a Riemannian metric derived from it. The dually flat structure plays a pivotal role in information geometry, as is shown in the generalized Pythagorean theorem. The dually flat structure is a special case of Riemannian geometry equipped with non-flat dual affine connections, which will be studied in Part II.

1.1 Manifolds

1.1.1 Manifold and Coordinate Systems

An n-dimensional manifold M is a set of points such that each point has n-dimensional extensions in its neighborhood. That is, such a neighborhood is topologically equivalent to an n-dimensional Euclidean space. Intuitively speaking, a manifold is a deformed Euclidean space, like a curved surface in the two-dimensional case. But it may have a different global topology. A sphere is an example which is locally equivalent to a two-dimensional Euclidean space, but is curved and has a different global topology because it is compact (bounded and closed).

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_1

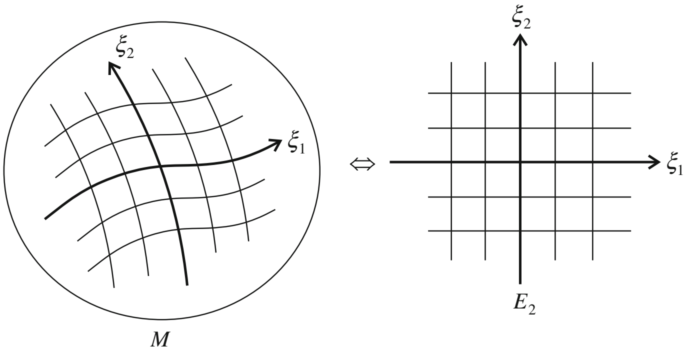


Fig. 1.1 Manifold M and coordinate system ξ. E_2 is a two-dimensional Euclidean space

Since a manifold M is locally equivalent to an n-dimensional Euclidean space E_n, we can introduce a local coordinate system

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n) \tag{1.1}$$

composed of n components $\xi_1, \ldots, \xi_n$ such that each point is uniquely specified by its coordinates $\boldsymbol{\xi}$ in a neighborhood. See Fig. 1.1 for the two-dimensional case. Since a manifold may have a topology different from a Euclidean space, in general we need more than one coordinate neighborhood and coordinate system to cover all the points of a manifold.

The coordinate system is not unique even in a coordinate neighborhood, and there are many coordinate systems. Let $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)$ be another coordinate system. When a point $P \in M$ is represented in two coordinate systems $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$, there is a one-to-one correspondence between them and we have relations

$$\boldsymbol{\xi} = \boldsymbol{f}\left(\zeta_1, \ldots, \zeta_n\right), \tag{1.2}$$

$$\boldsymbol{\zeta} = \boldsymbol{f}^{-1}\left(\xi_1, \ldots, \xi_n\right), \tag{1.3}$$

where $\boldsymbol{f}$ and $\boldsymbol{f}^{-1}$ are mutually inverse vector-valued functions. They are a coordinate transformation and its inverse transformation. We usually assume that (1.2) and (1.3) are differentiable functions of n coordinate variables.[1]

[1] Mathematically trained readers may know the rigorous definition of a manifold: A manifold M is a Hausdorff space which is covered by a number of open sets called coordinate neighborhoods, such that there exists an isomorphism between a coordinate neighborhood and a Euclidean space. The isomorphism defines a local coordinate system in the neighborhood. M is called a differentiable manifold when the coordinate transformations are differentiable. See textbooks on modern differential geometry. Our definition is intuitive, not mathematically rigorous, but is sufficient for understanding information geometry and its applications.

1.1.2 Examples of Manifolds

A. Euclidean Space

Consider a two-dimensional Euclidean space, which is a flat plane. It is convenient to use an orthonormal Cartesian coordinate system $\boldsymbol{\xi} = (\xi_1, \xi_2)$. A polar coordinate system $\boldsymbol{\zeta} = (r, \theta)$ is sometimes used, where r is the radius and θ is the angle of a point from one axis (see Fig. 1.2). The coordinate transformation between them is given by

$$r = \sqrt{\xi_1^2 + \xi_2^2}, \quad \theta = \tan^{-1}\left(\frac{\xi_2}{\xi_1}\right), \tag{1.4}$$

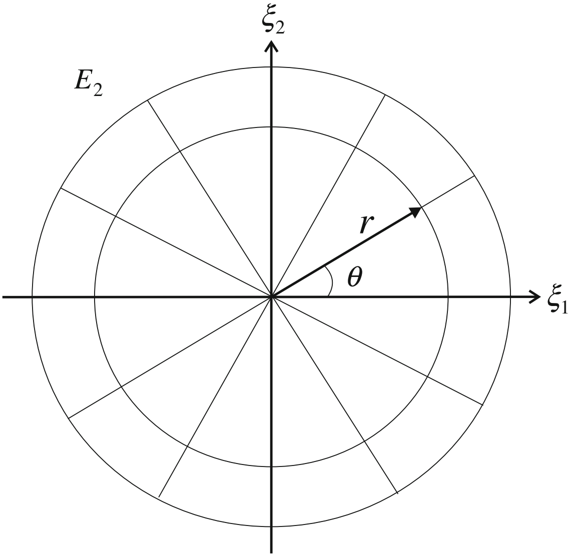
$$\xi_1 = r\cos\theta, \quad \xi_2 = r\sin\theta. \tag{1.5}$$

The transformation is analytic except for the origin.

B. Sphere

A sphere is the surface of a three-dimensional ball. The surface of the earth is regarded as a sphere, where each point has a two-dimensional neighborhood, so that we can draw a local geographic map on a flat sheet. The pair of latitude and longitude gives a local coordinate system. However, a sphere is topologically different from a Euclidean space and it cannot be covered by one coordinate system. At least two

Fig. 1.2 Cartesian coordinate system $\boldsymbol{\xi} = (\xi_1, \xi_2)$ and polar coordinate system (r, θ) in E_2



coordinate systems are required to cover it. If we delete one point, say the north pole of the earth, it is topologically equivalent to a Euclidean space. Hence, at least two overlapping coordinate neighborhoods, one including the north pole and the other including the south pole, for example, are necessary and they are sufficient to cover the entire sphere.

C. Manifold of Probability Distributions

C1. Gaussian Distributions

The probability density function of Gaussian random variable x is given by

$$p\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{1.6}$$

where μ is the mean and σ^2 is the variance. Hence, the set of all the Gaussian distributions is a two-dimensional manifold, where a point denotes a probability density function and

$$\boldsymbol{\xi}=(\mu,\sigma), \quad \sigma>0 \tag{1.7}$$

is a coordinate system. This is topologically equivalent to the upper half of a two-dimensional Euclidean space. The manifold of Gaussian distributions is covered by one coordinate system $\boldsymbol{\xi}=(\mu,\sigma)$.

There are other coordinate systems. For example, let m_1 and m_2 be the first and second moments of x, given by

$$m_1=\mathrm{E}[x]=\mu, \quad m_2=\mathrm{E}\left[x^2\right]=\mu^2+\sigma^2, \tag{1.8}$$

where E denotes the expectation of a random variable. Then,

$$\boldsymbol{\zeta}=(m_1,m_2) \tag{1.9}$$

is a coordinate system (the moment coordinate system).

It will be shown later that the coordinate system defined by $\boldsymbol{\theta}$,

$$\theta_1=\frac{\mu}{\sigma^2}, \quad \theta_2=-\frac{1}{2\sigma^2}, \tag{1.10}$$

is referred to as the natural parameters, and is convenient for studying properties of Gaussian distributions.

C2. Discrete Distributions

Let x be a discrete random variable taking values on $X = \{0, 1, \ldots, n\}$. A probability distribution $p(x)$ is specified by $n + 1$ probabilities

$$p_i = \mathrm{Prob}\{x = i\}, \quad i = 0, 1, \ldots, n, \tag{1.11}$$

so that $p(x)$ is represented by a probability vector

$$\boldsymbol{p} = (p_0, p_1, \ldots, p_n) \, . \tag{1.12}$$

Because of the restriction

$$\sum_{i=0}^{n} p_i = 1, \quad p_i > 0, \tag{1.13}$$

the set of all probability distributions $\boldsymbol{p}$ forms an n-dimensional manifold. Its coordinate system is given, for example, by

$$\boldsymbol{\xi} = (p_1, \ldots, p_n) \tag{1.14}$$

and p_0 is not free but is a function of the coordinates,

$$p_0 = 1 - \sum \xi_i \, . \tag{1.15}$$

The manifold is an n-dimensional simplex, called the probability simplex, and is denoted by S_n. When $n = 2$, S_2 is the interior of a triangle and when $n = 3$, it is the interior of a 3-simplex, as is shown in Fig. 1.3.

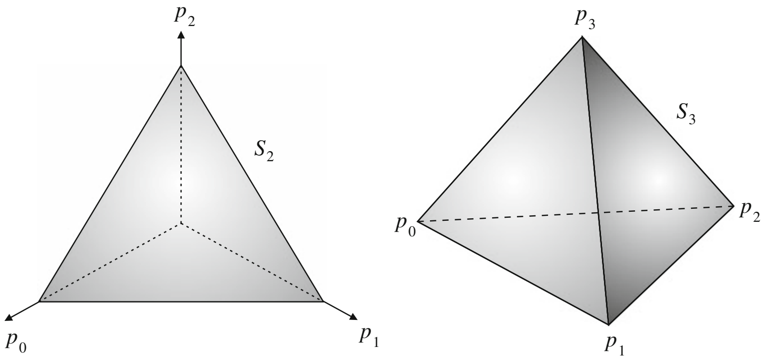


Fig. 1.3 Probability simplex: S_2 and S_3

Let us introduce $n + 1$ random variables $\delta_i(x)$, $i = 0, 1, \ldots, n$, such that

$$\delta_i(x) = \begin{cases} 1, & x = i, \\ 0, & x \neq i. \end{cases} \tag{1.16}$$

Then, a probability distribution of x is denoted by

$$p(x, \boldsymbol{\xi}) = \sum_{i=1}^{n} \xi_i \delta_i(x) + p_0(\boldsymbol{\xi})\delta_0(x) \tag{1.17}$$

in terms of coordinates $\boldsymbol{\xi}$.

We shall use another coordinate system $\boldsymbol{\theta}$ later, given by

$$\theta_i = \log \frac{p_i}{p_0}, \quad i = 1, \ldots, n, \tag{1.18}$$

which is also very useful.

C3. Regular Statistical Model

Let x be a random variable which may take discrete, scalar or vector continuous values. A statistical model is a family of probability distributions $M = \{p(x, \boldsymbol{\xi})\}$ specified by a vector parameter $\boldsymbol{\xi}$. When it satisfies certain regularity conditions, it is called a regular statistical model. Such an M is a manifold, where $\boldsymbol{\xi}$ plays the role of a coordinate system. The family of Gaussian distributions and the family of discrete probability distributions are examples of the regular statistical model. Information geometry has emerged from a study of invariant geometrical structures of regular statistical models.

D. Manifold of Positive Measures

Let x be a variable taking values in set $N = \{1, 2, \ldots, n\}$. We assign a positive measure (or a weight) m_i to element i, $i = 1, \ldots, n$. Then

$$\boldsymbol{\xi} = (m_1, \ldots, m_n), \quad m_i > 0 \tag{1.19}$$

defines a distribution of measures over N. The set of all such measures sits in the first quadrant $\boldsymbol{R}^n_+$ of an n-dimensional Euclidean space. The sum

$$m = \sum_{i=1}^{n} m_i \tag{1.20}$$

is called the total mass of $\boldsymbol{m} = (m_1, \ldots, m_n)$.

When $\boldsymbol{m}$ satisfies the constraint that the total mass is equal to 1,

$$\sum m_i = 1, \tag{1.21}$$

it is a probability distribution belonging to S_{n-1}. Hence, S_{n-1} is included in $\boldsymbol{R}^n_+$ as its submanifold.

A positive measure (unnormalized probability distribution) appears in many engineering problems. For example, image $s(x, y)$ drawn on the x–y plane is a positive measure when the brightness is positive,

$$s(x, y) > 0. \tag{1.22}$$

When we discretize the x–y plane into n^2 pixels (i, j), the discretized pictures $\{s(i, j)\}$ form a positive measure belonging to $\boldsymbol{R}^{n^2}_+$. Similarly, when we consider a discretized power spectrum of a sound, it is a positive measure. The histogram of observed data defines a positive measure, too.

E. Positive-Definite Matrices

Let $\mathbf{A}$ be an $n \times n$ matrix. All such matrices form an n^2-dimensional manifold. When $\mathbf{A}$ is symmetric and positive-definite, they form a $\frac{n(n+1)}{2}$-dimensional manifold. This is a submanifold embedded in the manifold of all the matrices. We may use the upper right elements of $\mathbf{A}$ as a coordinate system. Positive-definite matrices appear in statistics, physics, operations research, control theory, etc.

F. Neural Manifold

A neural network is composed of a large number of neurons connected with each other, where the dynamics of information processing takes place. A network is specified by connection weights w_{ji} connecting neuron i with neuron j. The set of all such networks forms a manifold, where matrix $\mathbf{W} = \left(w_{ji}\right)$ is a coordinate system. We will later analyze behaviors of such networks from the information geometry point of view.

1.2 Divergence Between Two Points

1.2.1 Divergence

Let us consider two points P and Q in a manifold M, of which coordinates are $\boldsymbol{\xi}_P$ and $\boldsymbol{\xi}_Q$. A divergence $D[P : Q]$ is a function of $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_Q$ which satisfies certain criteria. See Basseville (2013) for a detailed bibliography. We may write it as

$$D[P : Q] = D\left[\boldsymbol{\xi}_P : \boldsymbol{\xi}_Q\right]. \tag{1.23}$$

We assume that it is a differentiable function of $\boldsymbol{\xi}_P$ and $\boldsymbol{\xi}_Q$.

Definition 1.1 $D[P : Q]$ is called a divergence when it satisfies the following criteria:

(1) $D[P : Q] \geq 0$.
(2) $D[P : Q] = 0$, when and only when $P = Q$.
(3) When P and Q are sufficiently close, by denoting their coordinates by $\boldsymbol{\xi}_P$ and $\boldsymbol{\xi}_Q = \boldsymbol{\xi}_P + d\boldsymbol{\xi}$, the Taylor expansion of D is written as

$$D[\boldsymbol{\xi}_P : \boldsymbol{\xi}_P + d\boldsymbol{\xi}] = \frac{1}{2}\sum g_{ij}(\boldsymbol{\xi}_P)d\xi_i d\xi_j + O(|d\boldsymbol{\xi}|^3), \tag{1.24}$$

and matrix $\mathbf{G} = \left(g_{ij}\right)$ is positive-definite, depending on $\boldsymbol{\xi}_P$.

A divergence represents a degree of separation of two points P and Q, but it or its square root is not a distance. It does not necessarily satisfy the symmetry condition, so that in general

$$D[P : Q] \neq D[Q : P]. \tag{1.25}$$

We may call $D[P : Q]$ divergence from P to Q. Moreover, the triangular inequality does not hold. It has the dimension of the square of distance, as is suggested by (1.24). It is possible to symmetrize a divergence by

$$D_S[P : Q] = \frac{1}{2}\left(D[P : Q] + D[Q : P]\right). \tag{1.26}$$

However, the asymmetry of divergence plays an important role in information geometry, as will be seen later.

When P and Q are sufficiently close, we define the square of an infinitesimal distance ds between them by using (1.24) as

$$ds^2 = 2D\left[\boldsymbol{\xi} : \boldsymbol{\xi} + d\boldsymbol{\xi}\right] = \sum g_{ij} d\xi_i d\xi_j. \tag{1.27}$$

A manifold M is said to be Riemannian when a positive-definite matrix $\mathbf{G}(\boldsymbol{\xi})$ is defined on M and the square of the local distance between two nearby points $\boldsymbol{\xi}$ and $\boldsymbol{\xi} + d\boldsymbol{\xi}$ is given by (1.27). A divergence D provides M with a Riemannian structure.

1.2.2 *Examples of Divergence*

A. Euclidean Divergence

When we use an orthonormal Cartesian coordinate system in a Euclidean space, we define a divergence by a half of the square of the Euclidean distance,

$$D[P : Q] = \frac{1}{2} \sum \left(\xi_{Pi} - \xi_{Qi} \right)^2 . \tag{1.28}$$

The matrix **G** is the identity matrix in this case, so that

$$ds^2 = \sum (d\xi_i)^2 . \tag{1.29}$$

B. Kullback–Leibler Divergence

Let $p(x)$ and $q(x)$ be two probability distributions of random variable x in a manifold of probability distributions. The following is called the Kullback–Leibler (KL) divergence:

$$D_{KL}[p(x) : q(x)] = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{1.30}$$

When x is discrete, integration is replaced by summation. We can easily check that it satisfies the criteria of divergence. It is asymmetric in general and is useful in statistics, information theory, physics, etc. Many other divergences will be introduced later in a manifold of probability distributions.

C. KL-Divergence for Positive Measures

A manifold of positive measures $\boldsymbol{R}_+^n$ is a subset of a Euclidean space. Hence, we can introduce the Euclidean divergence (1.28) in it. However, we can extend the KL-divergence to give

$$D_{KL}\left[\boldsymbol{m}_1 : \boldsymbol{m}_2\right] = \sum m_{1i} \log \frac{m_{1i}}{m_{2i}} - \sum m_{1i} + \sum m_{2i}. \tag{1.31}$$

When the total masses of two measures $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are 1, they are probability distributions and $D_{KL}\left[\boldsymbol{m}_1 : \boldsymbol{m}_2\right]$ reduces to the KL-divergence D_{KL} in (1.30).

D. Divergences for Positive-Definite Matrices

There is a family of useful divergences introduced in the manifold of positive-definite matrices. Let **P** and **Q** be two positive-definite matrices. The following are typical examples of divergence:

$$D[\mathbf{P} : \mathbf{Q}] = \mathrm{tr}\left(\mathbf{P}\log\mathbf{P} - \mathbf{P}\log\mathbf{Q} - \mathbf{P} + \mathbf{Q}\right), \tag{1.32}$$

which is related to the Von Neumann entropy of quantum mechanics,

$$D[\mathbf{P} : \mathbf{Q}] = \mathrm{tr}\left(\mathbf{P}\mathbf{Q}^{-1}\right) - \log\left|\mathbf{P}\mathbf{Q}^{-1}\right| - n, \tag{1.33}$$

which is due to the KL-divergence of multivariate Gaussian distribution, and

$$D[\mathbf{P} : \mathbf{Q}] = \frac{4}{1-\alpha^2}\mathrm{tr}\left(-\mathbf{P}^{\frac{1-\alpha}{2}}\mathbf{Q}^{\frac{1+\alpha}{2}} + \frac{1-\alpha}{2}\mathbf{P} + \frac{1+\alpha}{2}\mathbf{Q}\right), \tag{1.34}$$

which is called the α-divergence, where α is a real parameter. Here, tr **P** denotes the trace of matrix **P** and $|\mathbf{P}|$ is the determinant of **P**.

1.3 Convex Function and Bregman Divergence

1.3.1 Convex Function

A nonlinear function $\psi(\boldsymbol{\xi})$ of coordinates $\boldsymbol{\xi}$ is said to be convex when the inequality

$$\lambda\psi\left(\boldsymbol{\xi}_1\right) + (1-\lambda)\psi\left(\boldsymbol{\xi}_2\right) \geq \psi\left\{\lambda\boldsymbol{\xi}_1 + (1-\lambda)\boldsymbol{\xi}_2\right\} \tag{1.35}$$

is satisfied for any $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$ and scalar $0 \leq \lambda \leq 1$. We consider a differentiable convex function. Then, a function is convex if and only if its Hessian

$$\mathbf{H}(\boldsymbol{\xi}) = \left(\frac{\partial^2}{\partial\xi_i\partial\xi_j}\psi(\boldsymbol{\xi})\right) \tag{1.36}$$

is positive-definite.

There are many convex functions appearing in physics, optimization and engineering problems. One simple example is

$$\psi(\boldsymbol{\xi}) = \frac{1}{2}\sum\xi_i^2 \tag{1.37}$$

which is a half of the square of the Euclidean distance from the origin to point $\boldsymbol{\xi}$. Let $\boldsymbol{p}$ be a probability distribution belonging to S_n. Then, its entropy

$$H(\boldsymbol{p}) = -\sum p_i \log p_i \tag{1.38}$$

is a concave function, so that its negative, $\varphi(\boldsymbol{p}) = -H(\boldsymbol{p})$, is a convex function.

We give one more example from a probability model. An exponential family of probability distributions is written as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\sum \theta_i x_i + k(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{1.39}$$

where $p(\boldsymbol{x}, \boldsymbol{\theta})$ is the probability density function of vector random variable $\boldsymbol{x}$ specified by vector parameter $\boldsymbol{\theta}$ and $k(\boldsymbol{x})$ is a function of $\boldsymbol{x}$. The term $\exp\{-\psi(\boldsymbol{\theta})\}$ is the normalization factor with which

$$\int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = 1 \tag{1.40}$$

is satisfied. Therefore, $\psi(\boldsymbol{\theta})$ is given by

$$\psi(\boldsymbol{\theta}) = \log \int \exp\left\{\sum \theta_i x_i + k(\boldsymbol{x})\right\} d\boldsymbol{x}. \tag{1.41}$$

$M = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$ is regarded as a manifold, where $\boldsymbol{\theta}$ is a coordinate system. By differentiating (1.41), we can prove that its Hessian is positive-definite (see the next subsection). Hence, $\psi(\boldsymbol{\theta})$ is a convex function. It is known as the cumulant generating function in statistics and free energy in statistical physics. The exponential family plays a fundamental role in information geometry.

1.3.2 Bregman Divergence

A graph of a convex function is shown in Fig. 1.4. We draw a tangent hyperplane touching it at point $\boldsymbol{\xi}_0$ (Fig. 1.4). It is given by the equation

$$z = \psi\left(\boldsymbol{\xi}_0\right) + \nabla\psi\left(\boldsymbol{\xi}_0\right) \cdot \left(\boldsymbol{\xi} - \boldsymbol{\xi}_0\right), \tag{1.42}$$

where z is the vertical axis of the graph. Here, ∇ is the gradient operator such that $\nabla\psi$ is the gradient vector defined by

$$\nabla\psi = \left(\frac{\partial}{\partial \xi_i}\psi(\boldsymbol{\xi})\right), \quad i = 1, \ldots, n \tag{1.43}$$

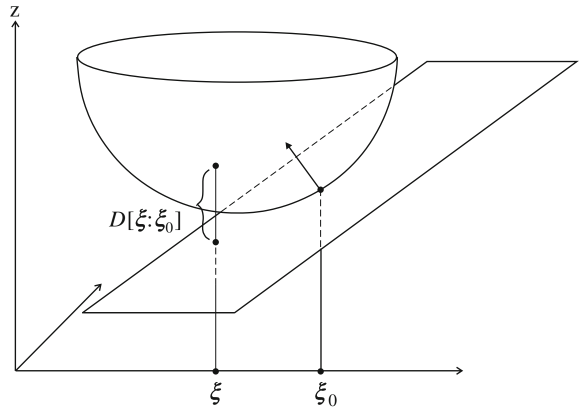


Fig. 1.4 Convex function $z = \psi(\xi)$, its supporting hyperplane with normal vector $\boldsymbol{n} = \nabla\psi(\xi_0)$ and divergence $D[\xi : \xi_0]$

in the component form. Since ψ is convex, the graph of ψ is always above the hyperplane, touching it at $\boldsymbol{\xi}_0$. Hence, it is a supporting hyperplane of ψ at $\boldsymbol{\xi}_0$ (Fig. 1.4).

We evaluate how high the function $\psi(\boldsymbol{\xi})$ is at $\boldsymbol{\xi}$ from the hyperplane (1.42). This depends on the point $\boldsymbol{\xi}_0$ at which the supporting hyperplane is defined. The difference from (1.42) is written as

$$D_\psi\left[\boldsymbol{\xi} : \boldsymbol{\xi}_0\right] = \psi(\boldsymbol{\xi}) - \psi\left(\boldsymbol{\xi}_0\right) - \nabla\psi\left(\boldsymbol{\xi}_0\right) \cdot \left(\boldsymbol{\xi} - \boldsymbol{\xi}_0\right). \tag{1.44}$$

Considering it as a function of two points $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_0$, we can easily prove that it satisfies the criteria of divergence. This is called the Bregman divergence [Bregman (1967)] derived from a convex function ψ.

We show examples of Bregman divergence.

Example 1.1 (*Euclidean divergence*) For ψ defined by (1.37) in a Euclidean space, we easily see that the divergence is

$$D\left[\boldsymbol{\xi} : \boldsymbol{\xi}_0\right] = \frac{1}{2}\left|\boldsymbol{\xi} - \boldsymbol{\xi}_0\right|^2, \tag{1.45}$$

that is, the same as a half of the square of the Euclidean distance. It is symmetric.

Example 1.2 (*Logarithmic divergence*) We consider a convex function

$$\psi(\boldsymbol{\xi}) = -\sum_{i=1}^{n} \log \xi_i \tag{1.46}$$

in the manifold $\boldsymbol{R}_+^n$ of positive measures. Its gradient is

$$\nabla\psi(\boldsymbol{\xi}) = \left(-\frac{1}{\xi_i}\right). \tag{1.47}$$

Hence, the Bregman divergence is

$$D_{\psi}\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] = \sum_{i=1}^{n} \left(\log \frac{\xi_i'}{\xi_i} + \frac{\xi_i}{\xi_i'} - 1 \right). \tag{1.48}$$

For another convex function

$$\varphi(\boldsymbol{\xi}) = \sum \xi_i \log \xi_i, \tag{1.49}$$

the Bregman divergence is the same as the KL-divergence (1.31), given by

$$D_{\varphi}\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] = \sum \left(\xi_i \log \frac{\xi_i}{\xi_i'} - \xi_i + \xi_i' \right). \tag{1.50}$$

When $\sum \xi_i = \sum \xi_i' = 1$, this is the KL-divergence from probability vector $\boldsymbol{\xi}$ to another $\boldsymbol{\xi}'$.

Example 1.3 (*Free energy of exponential family*) We calculate the divergence given by the normalization factor $\psi(\boldsymbol{\theta})$ (1.41) of an exponential family. To this end, we differentiate the identity

$$1 = \int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = \int \exp\left\{ \sum \theta_i x_i + k(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) \right\} d\boldsymbol{x} \tag{1.51}$$

with respect to θ_i. We then have

$$\int \left\{ x_i - \frac{\partial}{\partial \theta_i} \psi(\boldsymbol{\theta}) \right\} p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = 0 \tag{1.52}$$

or

$$\frac{\partial}{\partial \theta_i} \psi(\boldsymbol{\theta}) = \int x_i p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = \mathbf{E}\left[x_i\right] = \bar{x}_i, \tag{1.53}$$

$$\nabla \psi(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{x}\right], \tag{1.54}$$

where E denotes the expectation with respect to $p(\boldsymbol{x}, \boldsymbol{\theta})$ and $\bar{x}_i$ is the expectation of x_i. We then differentiate (2.12) again with respect to θ_j and, after some calculations, obtain

$$-\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} + \mathrm{E}\left[\left(x_i - \bar{x}_i\right)\left(x_j - \bar{x}_j\right)\right] = 0 \tag{1.55}$$

or

$$\nabla\nabla \psi(\boldsymbol{\theta}) = \mathrm{E}\left[\left(\boldsymbol{x} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x} - \bar{\boldsymbol{x}}\right)^T\right] = \mathrm{Var}[\boldsymbol{x}], \tag{1.56}$$

where $\boldsymbol{x}^T$ is the transpose of column vector $\boldsymbol{x}$ and Var$[\boldsymbol{x}]$ is the covariance matrix of $\boldsymbol{x}$, which is positive-definite. This shows that $\psi(\boldsymbol{\theta})$ is a convex function. It is useful to see that the expectation and covariance of $\boldsymbol{x}$ are derived from $\psi(\boldsymbol{\theta})$ by differentiation.

The Bregman divergence from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ derived from ψ of an exponential family is calculated from

$$D_\psi \left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi\left(\boldsymbol{\theta}\right) - \psi(\boldsymbol{\theta}') - \nabla\psi(\boldsymbol{\theta}') \cdot \left(\boldsymbol{\theta} - \boldsymbol{\theta}'\right), \tag{1.57}$$

proving that it is equal to the KL-divergence from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$ after careful calculations,

$$D_{KL}\left[p\left(\boldsymbol{x}, \boldsymbol{\theta}'\right) : p(\boldsymbol{x}, \boldsymbol{\theta})\right] = \int p\left(\boldsymbol{x}, \boldsymbol{\theta}'\right) \log \frac{p\left(\boldsymbol{x}, \boldsymbol{\theta}'\right)}{p(\boldsymbol{x}, \boldsymbol{\theta})} d\boldsymbol{x}. \tag{1.58}$$

1.4 Legendre Transformation

The gradient of $\psi(\boldsymbol{\xi})$

$$\boldsymbol{\xi}^* = \nabla\psi(\boldsymbol{\xi}) \tag{1.59}$$

is equal to the normal vector $\boldsymbol{n}$ of the supporting tangent hyperplane at $\boldsymbol{\xi}$, as is easily seen from Fig. 1.4. Different points have different normal vectors. Hence, it is possible to specify a point of M by its normal vector. In other words, the transformation between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ is one-to-one and differentiable. This shows that $\boldsymbol{\xi}^*$ is used as another coordinate system of M, which is connected with $\boldsymbol{\xi}$ by (1.59).

The transformation (1.59) is known as the Legendre transformation. The Legendre transformation has a dualistic structure concerning the two coupled coordinate systems $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$. To show this, we define a new function of $\boldsymbol{\xi}^*$ by

$$\psi^*\left(\boldsymbol{\xi}^*\right) = \boldsymbol{\xi} \cdot \boldsymbol{\xi}^* - \psi(\boldsymbol{\xi}), \tag{1.60}$$

where

$$\boldsymbol{\xi} \cdot \boldsymbol{\xi}^* = \sum_i \xi_i \xi_i^* \tag{1.61}$$

and $\boldsymbol{\xi}$ is not free but is a function of $\boldsymbol{\xi}^*$,

$$\boldsymbol{\xi} = \boldsymbol{f}\left(\boldsymbol{\xi}^*\right), \tag{1.62}$$

which is the inverse function of $\boldsymbol{\xi}^* = \nabla\psi(\boldsymbol{\xi})$. By differentiating (1.60) with respect to $\boldsymbol{\xi}^*$, we have

$$\nabla\psi^*\left(\boldsymbol{\xi}^*\right) = \boldsymbol{\xi} + \frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\xi}^*}\boldsymbol{\xi}^* - \nabla\psi(\boldsymbol{\xi})\frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\xi}^*}. \tag{1.63}$$

Since the last two terms of (1.63) cancel out because of (1.59), we have a dualistic structure

$$\boldsymbol{\xi}^* = \nabla\psi(\boldsymbol{\xi}), \quad \boldsymbol{\xi} = \nabla\psi^*\left(\boldsymbol{\xi}^*\right). \tag{1.64}$$

ψ^* is called the Legendre dual of ψ. The dual function ψ^* satisfies

$$\psi^*\left(\boldsymbol{\xi}^*\right) = \max_{\boldsymbol{\xi}'}\left\{\boldsymbol{\xi}'\cdot\boldsymbol{\xi}^* - \psi(\boldsymbol{\xi}')\right\}, \tag{1.65}$$

which is usually used as the definition of ψ^*. Our definition (1.60) is direct. We need to show ψ^* is a convex function. The Hessian of $\psi^*\left(\boldsymbol{\xi}^*\right)$ is written as

$$\mathbf{G}^*\left(\boldsymbol{\xi}^*\right) = \nabla\nabla\psi^*\left(\boldsymbol{\xi}^*\right) = \frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\xi}^*}, \tag{1.66}$$

which is the Jacobian matrix of the inverse transformation from $\boldsymbol{\xi}^*$ to $\boldsymbol{\xi}$. This is the inverse of the Hessian $\mathbf{G} = \nabla\nabla\psi(\boldsymbol{\xi})$, since it is the Jacobian matrix of the transformation from $\boldsymbol{\xi}$ to $\boldsymbol{\xi}^*$. Hence, it is a positive-definite matrix. This shows that $\psi^*\left(\boldsymbol{\xi}^*\right)$ is a convex function of $\boldsymbol{\xi}^*$.

A new Bregman divergence is derived from the dual convex function $\psi^*\left(\boldsymbol{\xi}^*\right)$,

$$D_{\psi^*}\left[\boldsymbol{\xi}^* : \boldsymbol{\xi}^{*\prime}\right] = \psi^*\left(\boldsymbol{\xi}^*\right) - \psi^*\left(\boldsymbol{\xi}^{*\prime}\right) - \nabla\psi^*\left(\boldsymbol{\xi}^{*\prime}\right)\cdot\left(\boldsymbol{\xi}^* - \boldsymbol{\xi}^{*\prime}\right), \tag{1.67}$$

which we call the dual divergence. However, by calculating carefully, one can easily derive

$$D_{\psi^*}\left[\boldsymbol{\xi}^* : \boldsymbol{\xi}^{*\prime}\right] = D_{\psi}\left[\boldsymbol{\xi}' : \boldsymbol{\xi}\right]. \tag{1.68}$$

Hence, the dual divergence is equal to the primal one if the order of two points is exchanged. Therefore, the divergences derived from the two convex functions are substantially the same, except for the order.

It is convenient to use a self-dual expression of divergence by using the two coordinate systems.

Theorem 1.1 *The divergence from P to Q derived from a convex $\psi(\boldsymbol{\xi})$ is written as*

$$D_{\psi}[P : Q] = \psi\left(\boldsymbol{\xi}_P\right) + \psi^*\left(\boldsymbol{\xi}^*_Q\right) - \boldsymbol{\xi}_P\cdot\boldsymbol{\xi}^*_Q, \tag{1.69}$$

where $\boldsymbol{\xi}_P$ is the coordinates of P in $\boldsymbol{\xi}$ coordinate system and $\boldsymbol{\xi}^*_Q$ is the coordinates of Q in $\boldsymbol{\xi}^*$ coordinate system.

Proof From (1.57), we have

$$\psi^*\left(\boldsymbol{\xi}^*_Q\right) = \boldsymbol{\xi}_Q\cdot\boldsymbol{\xi}^*_Q - \psi(\boldsymbol{\xi}_Q). \tag{1.70}$$

Substituting (1.70) in (1.69) and using $\nabla\psi\left(\boldsymbol{\xi}_Q\right) = \boldsymbol{\xi}^*_Q$, we have the theorem.