

Applied Mathematical Sciences

Shun-ichi Amari

Information Geometry and Its Applications



Springer

Applied Mathematical Sciences

Volume 194

Editors

S.S. Antman, Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA
ssa@math.umd.edu

Leslie Greengard, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
Greengard@cims.nyu.edu

P.J. Holmes, Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA
pholmes@math.princeton.edu

Advisors

J. Bell, Lawrence Berkeley National Lab, Center for Computational Sciences and Engineering, Berkeley, CA, USA

P. Constantin, Department of Mathematics, Princeton University, Princeton, NJ, USA
J. Keller, Department of Mathematics, Stanford University, Stanford, CA, USA

R. Kohn, Courant Institute of Mathematical Sciences, New York University, New York, USA

R. Pego, Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

L. Ryzhik, Department of Mathematics, Stanford University, Stanford, CA, USA

A. Singer, Department of Mathematics, Princeton University, Princeton, NJ, USA

A. Stevens, Department of Applied Mathematics, University of Münster, Münster, Germany

A. Stuart, Mathematics Institute, University of Warwick, Coventry, United Kingdom

S. Wright, Computer Sciences Department, University of Wisconsin, Madison, WI, USA

Founding Editors

Fritz John, Joseph P. LaSalle and Lawrence Sirovich

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
Applied Mathematical Sciences
ISBN 978-4-431-55977-1
DOI 10.1007/978-4-431-55978-8

ISSN 2196-968X (electronic)
ISBN 978-4-431-55978-8 (eBook)

Library of Congress Control Number: 2015958849

© Springer Japan 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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

1	Manifold, Divergence and Dually Flat Structure	3
1.1	Manifolds	3
1.1.1	Manifold and Coordinate Systems	3
1.1.2	Examples of Manifolds	5
1.2	Divergence Between Two Points	9
1.2.1	Divergence	9
1.2.2	Examples of Divergence	11
1.3	Convex Function and Bregman Divergence	12
1.3.1	Convex Function	12
1.3.2	Bregman Divergence	13
1.4	Legendre Transformation.	16
1.5	Dually Flat Riemannian Structure Derived from Convex Function	19
1.5.1	Affine and Dual Affine Coordinate Systems	19
1.5.2	Tangent Space, Basis Vectors and Riemannian Metric	20
1.5.3	Parallel Transport of Vector	23
1.6	Generalized Pythagorean Theorem and Projection Theorem	24
1.6.1	Generalized Pythagorean Theorem	24
1.6.2	Projection Theorem	26
1.6.3	Divergence Between Submanifolds: Alternating Minimization Algorithm	27
2	Exponential Families and Mixture Families of Probability Distributions	31
2.1	Exponential Family of Probability Distributions	31

2.2	Examples of Exponential Family: Gaussian and Discrete Distributions	34
2.2.1	Gaussian Distribution	34
2.2.2	Discrete Distribution	35
2.3	Mixture Family of Probability Distributions	36
2.4	Flat Structure: e -flat and m -flat	37
2.5	On Infinite-Dimensional Manifold of Probability Distributions	39
2.6	Kernel Exponential Family	42
2.7	Bregman Divergence and Exponential Family	43
2.8	Applications of Pythagorean Theorem	44
2.8.1	Maximum Entropy Principle	44
2.8.2	Mutual Information	46
2.8.3	Repeated Observations and Maximum Likelihood Estimator	47
3	Invariant Geometry of Manifold of Probability Distributions	51
3.1	Invariance Criterion	51
3.2	Information Monotonicity Under Coarse Graining	53
3.2.1	Coarse Graining and Sufficient Statistics in S_n	53
3.2.2	Invariant Divergence	54
3.3	Examples of f -Divergence in S_n	57
3.3.1	KL-Divergence	57
3.3.2	χ^2 -Divergence	57
3.3.3	α -Divergence	57
3.4	General Properties of f -Divergence and KL-Divergence	59
3.4.1	Properties of f -Divergence	59
3.4.2	Properties of KL-Divergence	60
3.5	Fisher Information: The Unique Invariant Metric	62
3.6	f -Divergence in Manifold of Positive Measures	65
4	α-Geometry, Tsallis q-Entropy and Positive-Definite Matrices	71
4.1	Invariant and Flat Divergence	71
4.1.1	KL-Divergence Is Unique	71
4.1.2	α -Divergence Is Unique in R_+^n	72
4.2	α -Geometry in S_n and R_+^n	75
4.2.1	α -Geodesic and α -Pythagorean Theorem in R_+^n	75
4.2.2	α -Geodesic in S_n	76
4.2.3	α -Pythagorean Theorem and α -Projection Theorem in S_n	76
4.2.4	Apportionment Due to α -Divergence	77
4.2.5	α -Mean	77
4.2.6	α -Families of Probability Distributions	80
4.2.7	Optimality of α -Integration	82
4.2.8	Application to α -Integration of Experts	83

4.3	Geometry of Tsallis q -Entropy	84
4.3.1	q -Logarithm and q -Exponential Function	85
4.3.2	q -Exponential Family (α -Family) of Probability Distributions	86
4.3.3	q -Escort Geometry	87
4.3.4	Deformed Exponential Family: χ -Escort Geometry	89
4.3.5	Conformal Character of q -Escort Geometry	91
4.4	(u, v)-Divergence: Dually Flat Divergence in Manifold of Positive Measures	92
4.4.1	Decomposable (u, v)-Divergence	92
4.4.2	General (u, v) Flat Structure in R_+^n	95
4.5	Invariant Flat Divergence in Manifold of Positive-Definite Matrices	96
4.5.1	Bregman Divergence and Invariance Under $Gl(n)$	96
4.5.2	Invariant Flat Decomposable Divergences Under $O(n)$	98
4.5.3	Non-flat Invariant Divergences	101
4.6	Miscellaneous Divergences	102
4.6.1	γ -Divergence	102
4.6.2	Other Types of (α, β) -Divergences	102
4.6.3	Burbea–Rao Divergence and Jensen–Shannon Divergence	103
4.6.4	(ρ, τ) -Structure and (F, G, H) -Structure	104

Part II Introduction to Dual Differential Geometry

5	Elements of Differential Geometry	109
5.1	Manifold and Tangent Space	109
5.2	Riemannian Metric	111
5.3	Affine Connection	112
5.4	Tensors	114
5.5	Covariant Derivative	116
5.6	Geodesic	117
5.7	Parallel Transport of Vector	118
5.8	Riemann–Christoffel Curvature	119
5.8.1	Round-the-World Transport of Vector	120
5.8.2	Covariant Derivative and RC Curvature	122
5.8.3	Flat Manifold	123
5.9	Levi–Civita (Riemannian) Connection	124
5.10	Submanifold and Embedding Curvature	126
5.10.1	Submanifold	126
5.10.2	Embedding Curvature	127

6	Dual Affine Connections and Dually Flat Manifold	131
6.1	Dual Connections	131
6.2	Metric and Cubic Tensor Derived from Divergence	134
6.3	Invariant Metric and Cubic Tensor	136
6.4	α -Geometry	136
6.5	Dually Flat Manifold	137
6.6	Canonical Divergence in Dually Flat Manifold	138
6.7	Canonical Divergence in General Manifold of Dual Connections	141
6.8	Dual Foliations of Flat Manifold and Mixed Coordinates	143
6.8.1	k -cut of Dual Coordinate Systems: Mixed Coordinates and Foliation	144
6.8.2	Decomposition of Canonical Divergence	145
6.8.3	A Simple Illustrative Example: Neural Firing	146
6.8.4	Higher-Order Interactions of Neuronal Spikes	148
6.9	System Complexity and Integrated Information	150
6.10	Input–Output Analysis in Economics	157
Part III Information Geometry of Statistical Inference		
7	Asymptotic Theory of Statistical Inference	165
7.1	Estimation	165
7.2	Estimation in Exponential Family	166
7.3	Estimation in Curved Exponential Family	168
7.4	First-Order Asymptotic Theory of Estimation	171
7.5	Higher-Order Asymptotic Theory of Estimation	173
7.6	Asymptotic Theory of Hypothesis Testing	175
8	Estimation in the Presence of Hidden Variables	179
8.1	EM Algorithm	179
8.1.1	Statistical Model with Hidden Variables	179
8.1.2	Minimizing Divergence Between Model Manifold and Data Manifold	182
8.1.3	EM Algorithm	184
8.1.4	Example: Gaussian Mixture	184
8.2	Loss of Information by Data Reduction	185
8.3	Estimation Based on Misspecified Statistical Model	186
9	Neyman–Scott Problem: Estimating Function and Semiparametric Statistical Model	191
9.1	Statistical Model Including Nuisance Parameters	191
9.2	Neyman–Scott Problem and Semiparametrics	194
9.3	Estimating Function	197
9.4	Information Geometry of Estimating Function	199

9.5	Solutions to Neyman–Scott Problems	206
9.5.1	Estimating Function in the Exponential Case	206
9.5.2	Coefficient of Linear Dependence	208
9.5.3	Scale Problem	209
9.5.4	Temporal Firing Pattern of Single Neuron	211
10	Linear Systems and Time Series	215
10.1	Stationary Time Series and Linear System	215
10.2	Typical Finite-Dimensional Manifolds of Time Series	217
10.3	Dual Geometry of System Manifold	219
10.4	Geometry of AR, MA and ARMA Models	223

Part IV Applications of Information Geometry

11	Machine Learning	231
11.1	Clustering Patterns	231
11.1.1	Pattern Space and Divergence	231
11.1.2	Center of Cluster	232
11.1.3	<i>k</i> -Means: Clustering Algorithm	233
11.1.4	Voronoi Diagram	234
11.1.5	Stochastic Version of Classification and Clustering . . .	236
11.1.6	Robust Cluster Center	238
11.1.7	Asymptotic Evaluation of Error Probability in Pattern Recognition: Chernoff Information	240
11.2	Geometry of Support Vector Machine	242
11.2.1	Linear Classifier	242
11.2.2	Embedding into High-Dimensional Space	245
11.2.3	Kernel Method	246
11.2.4	Riemannian Metric Induced by Kernel	247
11.3	Stochastic Reasoning: Belief Propagation and CCCP Algorithms	249
11.3.1	Graphical Model	250
11.3.2	Mean Field Approximation and <i>m</i> -Projection	252
11.3.3	Belief Propagation	255
11.3.4	Solution of BP Algorithm	257
11.3.5	CCCP (Convex–Concave Computational Procedure)	259
11.4	Information Geometry of Boosting	260
11.4.1	Boosting: Integration of Weak Machines	261
11.4.2	Stochastic Interpretation of Machine	262
11.4.3	Construction of New Weak Machines	263
11.4.4	Determination of the Weights of Weak Machines	263

11.5	Bayesian Inference and Deep Learning	265
11.5.1	Bayesian Duality in Exponential Family	266
11.5.2	Restricted Boltzmann Machine.	268
11.5.3	Unsupervised Learning of RBM.	269
11.5.4	Geometry of Contrastive Divergence.	273
11.5.5	Gaussian RBM	275
12	Natural Gradient Learning and Its Dynamics in Singular Regions	279
12.1	Natural Gradient Stochastic Descent Learning	279
12.1.1	On-Line Learning and Batch Learning	279
12.1.2	Natural Gradient: Steepest Descent Direction in Riemannian Manifold	282
12.1.3	Riemannian Metric, Hessian and Absolute Hessian	284
12.1.4	Stochastic Relaxation of Optimization Problem	286
12.1.5	Natural Policy Gradient in Reinforcement Learning	287
12.1.6	Mirror Descent and Natural Gradient	289
12.1.7	Properties of Natural Gradient Learning	290
12.2	Singularity in Learning: Multilayer Perceptron	296
12.2.1	Multilayer Perceptron	296
12.2.2	Singularities in M	298
12.2.3	Dynamics of Learning in M	302
12.2.4	Critical Slowdown of Dynamics.	305
12.2.5	Natural Gradient Learning Is Free of Plateaus	309
12.2.6	Singular Statistical Models	310
12.2.7	Bayesian Inference and Singular Model.	312
13	Signal Processing and Optimization	315
13.1	Principal Component Analysis	315
13.1.1	Eigenvalue Analysis	315
13.1.2	Principal Components, Minor Components and Whitening	316
13.1.3	Dynamics of Learning of Principal and Minor Components	319
13.2	Independent Component Analysis.	322
13.2.3	Estimating Function of ICA: Semiparametric Approach	330
13.3	Non-negative Matrix Factorization	333
13.4	Sparse Signal Processing.	336
13.4.1	Linear Regression and Sparse Solution	337
13.4.2	Minimization of Convex Function Under L_1 Constraint	338
13.4.3	Analysis of Solution Path	341
13.4.4	Minkovskian Gradient Flow.	343
13.4.5	Underdetermined Case	344

Contents	xiii
13.5 Optimization in Convex Programming	345
13.5.1 Convex Programming	345
13.5.2 Dually Flat Structure Derived from Barrier Function	347
13.5.3 Computational Complexity and m -curvature.	348
13.6 Dual Geometry Derived from Game Theory	349
13.6.1 Minimization of Game-Score	349
13.6.2 Hyvärinen Score	353
References	359
Index	371

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).

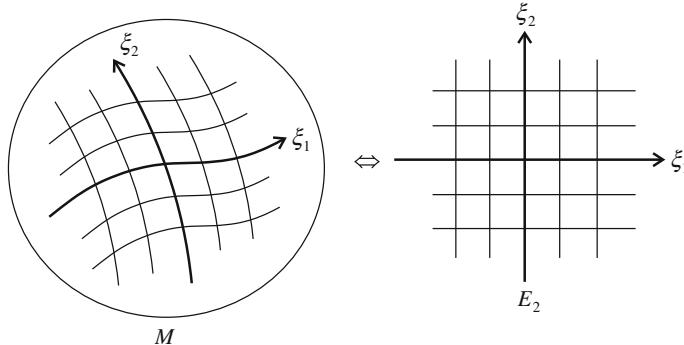


Fig. 1.1 Manifold M and coordinate system ξ . E_2 is a two-dimensional Euclidean space

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

$$\xi = (\xi_1, \dots, \xi_n) \quad (1.1)$$

composed of n components ξ_1, \dots, ξ_n such that each point is uniquely specified by its coordinates ξ 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 $\zeta = (\zeta_1, \dots, \zeta_n)$ be another coordinate system. When a point $P \in M$ is represented in two coordinate systems ξ and ζ , there is a one-to-one correspondence between them and we have relations

$$\xi = f(\zeta_1, \dots, \zeta_n), \quad (1.2)$$

$$\zeta = f^{-1}(\xi_1, \dots, \xi_n), \quad (1.3)$$

where f and 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

¹Mathematically 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 $\xi = (\xi_1, \xi_2)$. A polar coordinate system $\zeta = (r, \theta)$ 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

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

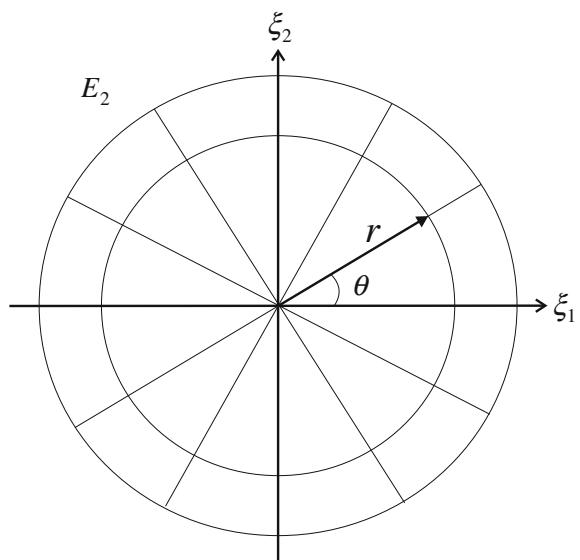
$$\xi_1 = r \cos \theta, \quad \xi_2 = r \sin \theta. \quad (1.5)$$

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 $\xi = (\xi_1, \xi_2)$ 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

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

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

$$\xi = (\mu, \sigma), \quad \sigma > 0 \quad (1.7)$$

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 $\xi = (\mu, \sigma)$.

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

$$m_1 = E[x] = \mu, \quad m_2 = E[x^2] = \mu^2 + \sigma^2, \quad (1.8)$$

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

$$\zeta = (m_1, m_2) \quad (1.9)$$

is a coordinate system (the moment coordinate system).

It will be shown later that the coordinate system defined by θ ,

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

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 $X = \{0, 1, \dots, n\}$. A probability distribution $p(x)$ is specified by $n + 1$ probabilities

$$p_i = \text{Prob}\{x = i\}, \quad i = 0, 1, \dots, n, \quad (1.11)$$

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

$$\mathbf{p} = (p_0, p_1, \dots, p_n). \quad (1.12)$$

Because of the restriction

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

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

$$\xi = (p_1, \dots, p_n) \quad (1.14)$$

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

$$p_0 = 1 - \sum \xi_i. \quad (1.15)$$

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.

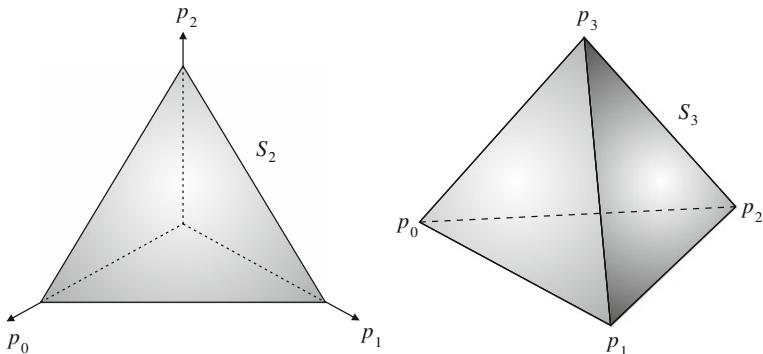


Fig. 1.3 Probability simplex: S_2 and S_3

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

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

Then, a probability distribution of x is denoted by

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

in terms of coordinates ξ .

We shall use another coordinate system θ later, given by

$$\theta_i = \log \frac{p_i}{p_0}, \quad i = 1, \dots, n, \quad (1.18)$$

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 $M = \{p(x, \xi)\}$ specified by a vector parameter ξ . When it satisfies certain regularity conditions, it is called a regular statistical model. Such an M is a manifold, where ξ 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 $N = \{1, 2, \dots, n\}$. We assign a positive measure (or a weight) m_i to element i , $i = 1, \dots, n$. Then

$$\xi = (m_1, \dots, m_n), \quad m_i > 0 \quad (1.19)$$

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

$$m = \sum_{i=1}^n m_i \quad (1.20)$$

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

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

$$\sum m_i = 1, \quad (1.21)$$

it is a probability distribution belonging to S_{n-1} . Hence, S_{n-1} is included in \mathbf{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,

$$s(x, y) > 0. \quad (1.22)$$

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 $\mathbf{R}_+^{n^2}$. 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 $\mathbf{W} = (w_{ji})$ 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 ξ_P and ξ_Q . A divergence $D[P : Q]$ is a function of ξ_P and ξ_Q which satisfies certain

criteria. See Basseville (2013) for a detailed bibliography. We may write it as

$$D[P : Q] = D[\xi_P : \xi_Q]. \quad (1.23)$$

We assume that it is a differentiable function of ξ_P and ξ_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 ξ_P and $\xi_Q = \xi_P + d\xi$, the Taylor expansion of D is written as

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

and matrix $\mathbf{G} = (g_{ij})$ is positive-definite, depending on ξ_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

$$D[P : Q] \neq D[Q : P]. \quad (1.25)$$

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

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

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

$$ds^2 = 2D[\xi : \xi + d\xi] = \sum g_{ij} d\xi_i d\xi_j. \quad (1.27)$$

A manifold M is said to be Riemannian when a positive-definite matrix $\mathbf{G}(\xi)$ is defined on M and the square of the local distance between two nearby points ξ and $\xi + d\xi$ 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,

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

The matrix \mathbf{G} is the identity matrix in this case, so that

$$ds^2 = \sum (d\xi_i)^2. \quad (1.29)$$

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:

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

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 \mathbf{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

$$D_{KL}[\mathbf{m}_1 : \mathbf{m}_2] = \sum m_{1i} \log \frac{m_{1i}}{m_{2i}} - \sum m_{1i} + \sum m_{2i}. \quad (1.31)$$

When the total masses of two measures \mathbf{m}_1 and \mathbf{m}_2 are 1, they are probability distributions and $D_{KL}[\mathbf{m}_1 : \mathbf{m}_2]$ 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 \mathbf{P} and \mathbf{Q} be two positive-definite matrices. The following are typical examples of divergence:

$$D[\mathbf{P} : \mathbf{Q}] = \text{tr}(\mathbf{P} \log \mathbf{P} - \mathbf{P} \log \mathbf{Q} - \mathbf{P} + \mathbf{Q}), \quad (1.32)$$

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

$$D[\mathbf{P} : \mathbf{Q}] = \text{tr}(\mathbf{P}\mathbf{Q}^{-1}) - \log |\mathbf{P}\mathbf{Q}^{-1}| - n, \quad (1.33)$$

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

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

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

1.3 Convex Function and Bregman Divergence

1.3.1 Convex Function

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

$$\lambda\psi(\xi_1) + (1-\lambda)\psi(\xi_2) \geq \psi\{\lambda\xi_1 + (1-\lambda)\xi_2\} \quad (1.35)$$

is satisfied for any ξ_1, ξ_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

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

is positive-definite.

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

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

which is a half of the square of the Euclidean distance from the origin to point ξ . Let p be a probability distribution belonging to S_n . Then, its entropy

$$H(p) = -\sum p_i \log p_i \quad (1.38)$$

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

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

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

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

$$\int p(x, \theta) dx = 1 \quad (1.40)$$

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

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

$M = \{p(x, \theta)\}$ is regarded as a manifold, where θ is a coordinate system. By differentiating (1.41), we can prove that its Hessian is positive-definite (see the next subsection). Hence, $\psi(\theta)$ 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

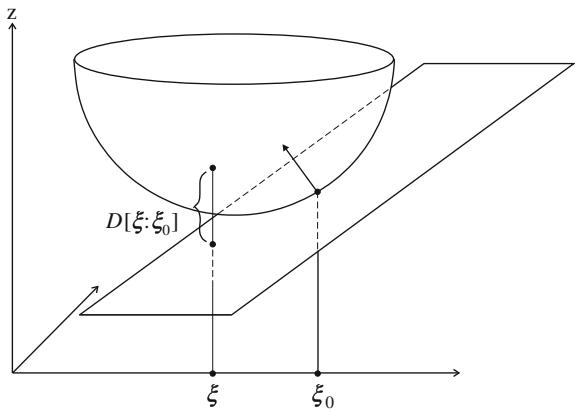
A graph of a convex function is shown in Fig. 1.4. We draw a tangent hyperplane touching it at point ξ_0 (Fig. 1.4). It is given by the equation

$$z = \psi(\xi_0) + \nabla \psi(\xi_0) \cdot (\xi - \xi_0), \quad (1.42)$$

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

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

Fig. 1.4 Convex function $z = \psi(\xi)$, its supporting hyperplane with normal vector $\mathbf{n} = \nabla\psi(\xi_0)$ 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 ξ_0 . Hence, it is a supporting hyperplane of ψ at ξ_0 (Fig. 1.4).

We evaluate how high the function $\psi(\xi)$ is at ξ from the hyperplane (1.42). This depends on the point ξ_0 at which the supporting hyperplane is defined. The difference from (1.42) is written as

$$D_\psi [\xi : \xi_0] = \psi(\xi) - \psi(\xi_0) - \nabla\psi(\xi_0) \cdot (\xi - \xi_0). \quad (1.44)$$

Considering it as a function of two points ξ and ξ_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

$$D[\xi : \xi_0] = \frac{1}{2} |\xi - \xi_0|^2, \quad (1.45)$$

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

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

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

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

Hence, the Bregman divergence is

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

For another convex function

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

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

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

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

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

with respect to θ_i . We then have

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

or

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

$$\nabla \psi(\boldsymbol{\theta}) = \mathbf{E}[\mathbf{x}], \quad (1.54)$$

where \mathbf{E} denotes the expectation with respect to $p(\mathbf{x}, \boldsymbol{\theta})$ 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

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

or

$$\nabla \nabla \psi(\boldsymbol{\theta}) = \mathbf{E}[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T] = \text{Var}[\mathbf{x}], \quad (1.56)$$

where \mathbf{x}^T is the transpose of column vector \mathbf{x} and $\text{Var}[\mathbf{x}]$ is the covariance matrix of \mathbf{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 \mathbf{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

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

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

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

1.4 Legendre Transformation

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

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

is equal to the normal vector \mathbf{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

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

where

$$\boldsymbol{\xi} \cdot \boldsymbol{\xi}^* = \sum_i \xi_i \xi_i^* \quad (1.61)$$

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

$$\boldsymbol{\xi} = f(\boldsymbol{\xi}^*), \quad (1.62)$$

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

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

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

$$\xi^* = \nabla\psi(\xi), \quad \xi = \nabla\psi^*(\xi^*). \quad (1.64)$$

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

$$\psi^*(\xi^*) = \max_{\xi'} \{\xi' \cdot \xi^* - \psi(\xi')\}, \quad (1.65)$$

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 $\psi^*(\xi^*)$ is written as

$$\mathbf{G}^*(\xi^*) = \nabla\nabla\psi^*(\xi^*) = \frac{\partial\xi}{\partial\xi^*}, \quad (1.66)$$

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

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

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

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

$$D_{\psi^*}[\xi^* : \xi^{*\prime}] = D_\psi[\xi' : \xi]. \quad (1.68)$$

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 $\psi(\xi)$ is written as*

$$D_\psi[P : Q] = \psi(\xi_P) + \psi^*(\xi_Q^*) - \xi_P \cdot \xi_Q^*, \quad (1.69)$$

where ξ_P is the coordinates of P in ξ coordinate system and ξ_Q^* is the coordinates of Q in ξ^* coordinate system.

Proof From (1.57), we have

$$\psi^*(\xi_Q^*) = \xi_Q \cdot \xi_Q^* - \psi(\xi_Q). \quad (1.70)$$

Substituting (1.70) in (1.69) and using $\nabla\psi(\xi_Q) = \xi_Q^*$, we have the theorem.