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

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 *k*-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 Asmptotic 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 *m*-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

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).

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

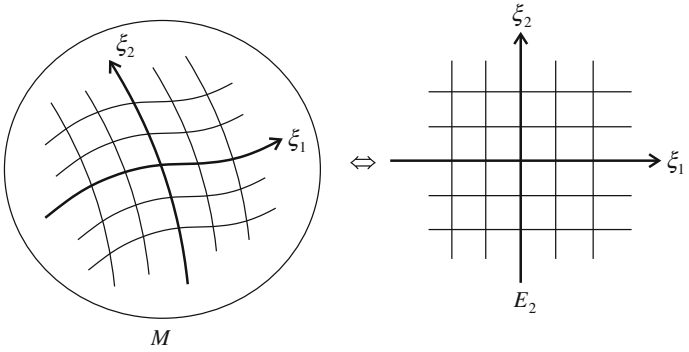


Fig. 1.1 Manifold M and coordinate system ξ. E_2 is a two-dimensional Euclidean space

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

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

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

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

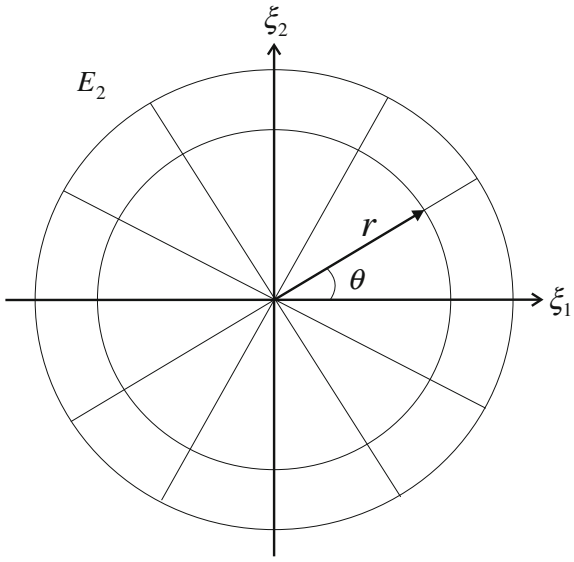


Fig. 1.2 Cartesian coordinate system $\boldsymbol{\xi} = (\xi_1, \xi_2)$ and polar coordinate system (r, θ) in E_2

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

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

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.

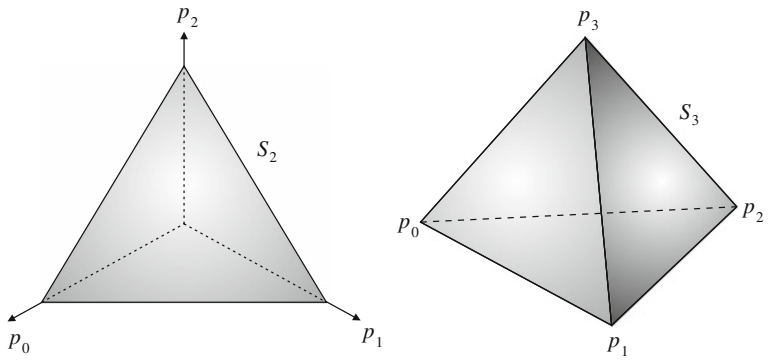


Fig. 1.3 Probability simplex: S_2 and S_3

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

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

When the total masses of two measures $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are 1, they are probability distributions and $D_{KL}[\boldsymbol{m}_1 : \boldsymbol{m}_2]$ 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:

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

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 $\mathbf{P}$ 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 $\psi(\boldsymbol{\xi})$ 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

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

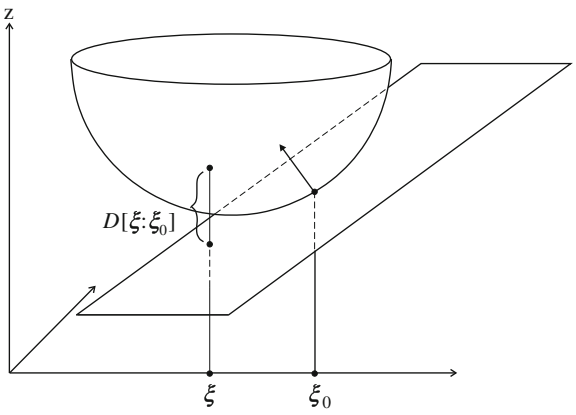


Fig. 1.4 Convex function $z = \psi(\xi)$, its supporting hyperplane with normal vector $\boldsymbol{n} = \nabla\psi(\xi_0)$ and divergence $D[\xi : \xi_0]$

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 $\mathrm{Var}[\boldsymbol{x}]$ 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.

We give examples of dual convex functions. For convex function (1.37) in Example 1.1, we easily have

$$\psi^* \left(\boldsymbol{\xi}^*\right) = \frac{1}{2} \left|\boldsymbol{\xi}^*\right|^2 \tag{1.71}$$

and

$$\boldsymbol{\xi}^* = \boldsymbol{\xi}. \tag{1.72}$$

Hence, the dual convex function is the same as the primal one, implying that the structure is self-dual. □

In the case of Example 1.2, the duals of ψ and φ in (1.46) and (1.49) are

$$\psi^* \left(\boldsymbol{\xi}^*\right) = -\sum \left\{1 + \log\left(-\xi_i^*\right)\right\}, \tag{1.73}$$

$$\varphi^* \left(\boldsymbol{\xi}^*\right) = \sum \exp\left\{\xi_i^* - 1\right\}, \tag{1.74}$$

by which

$$\nabla\psi^* \left(\boldsymbol{\xi}^*\right) = \boldsymbol{\xi}, \quad \nabla\varphi^* \left(\xi^*\right) = \boldsymbol{\xi} \tag{1.75}$$

hold, respectively.

In the case of the free energy $\psi(\boldsymbol{\theta})$ in Example 1.3, its Legendre transformation is

$$\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{x}], \tag{1.76}$$

where $\mathrm{E}_{\boldsymbol{\theta}}$ is the expectation with respect to $p(\boldsymbol{x}, \boldsymbol{\theta})$. Because of this, $\boldsymbol{\theta}^*$ is called the expectation parameter in statistics. The dual convex function $\psi^* \left(\boldsymbol{\theta}^*\right)$ derived from (1.65) is calculated from

$$\psi^* \left(\boldsymbol{\theta}^*\right) = \boldsymbol{\theta}^* \cdot \boldsymbol{\theta} - \psi(\boldsymbol{\theta}), \tag{1.77}$$

where $\boldsymbol{\theta}$ is a function of $\boldsymbol{\theta}^*$ given by $\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta})$. This proves that ψ^* is the negative entropy,

$$\psi^* \left(\boldsymbol{\theta}^*\right) = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}. \tag{1.78}$$

The dual divergence derived from $\psi^* \left(\boldsymbol{\theta}^*\right)$ is the KL-divergence

$$D_{\psi^*} \left[\boldsymbol{\theta}^* : \boldsymbol{\theta}^{*\prime}\right] = D_{KL} \left[p(\boldsymbol{x}, \boldsymbol{\theta}) : p\left(\boldsymbol{x}, \boldsymbol{\theta}'\right)\right], \tag{1.79}$$

where $\boldsymbol{\theta} = \nabla\psi^*(\boldsymbol{\theta}^*)$ and $\boldsymbol{\theta}' = \nabla\psi^* \left(\boldsymbol{\theta}^{*\prime}\right)$.

1.5 Dually Flat Riemannian Structure Derived from Convex Function

1.5.1 Affine and Dual Affine Coordinate Systems

When a function $\psi(\boldsymbol{\theta})$ is convex in a coordinate system $\boldsymbol{\theta}$, the same function expressed in another coordinate system $\boldsymbol{\xi}$,

$$\tilde{\psi}(\boldsymbol{\xi}) = \psi\left\{\boldsymbol{\theta}(\boldsymbol{\xi})\right\}, \tag{1.80}$$

is not necessarily convex as a function of $\boldsymbol{\xi}$. Hence, the convexity of a function depends on the coordinate system of M. But a convex function remains convex under affine transformations

$$\boldsymbol{\theta}' = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{b}, \tag{1.81}$$

where $\mathbf{A}$ is a non-singular constant matrix and $\boldsymbol{b}$ is a constant vector.

We fix a coordinate system $\boldsymbol{\theta}$ in which $\psi(\boldsymbol{\theta})$ is convex and introduce geometric structures to M based on it. We consider $\boldsymbol{\theta}$ as an affine coordinate system, which provides M with an affine flat structure: M is a flat manifold and each coordinate axis of $\boldsymbol{\theta}$ is a straight line. Any curve $\boldsymbol{\theta}(t)$ of M written in the linear form of parameter t,

$$\boldsymbol{\theta}(t) = \boldsymbol{a}t + \boldsymbol{b}, \tag{1.82}$$

is a straight line, where and $\boldsymbol{a}$ and $\boldsymbol{b}$ are constant vectors. We call it a geodesic of an affine manifold. Here, the term "geodesic" is used to represent a straight line and does not mean the shortest path connecting two points. A geodesic is invariant under affine transformations (1.81), but this is not true under nonlinear coordinate transformations.

Dually, we can define another coordinate system $\boldsymbol{\theta}^*$ by the Legendre transformation,

$$\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta}), \tag{1.83}$$

and consider it as another type of affine coordinates. This defines another affine structure. Each coordinate axis of $\boldsymbol{\theta}^*$ is a dual straight line or dual geodesic. A dual straight line is written as

$$\boldsymbol{\theta}^*(t) = \boldsymbol{a}t + \boldsymbol{b}. \tag{1.84}$$

This is the dual affine structure derived from the convex function $\psi^*\left(\boldsymbol{\theta}^*\right)$. Since the coordinate transformation between the two affine coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ is not linear in general, a geodesic is not a dual geodesic and vice versa. This implies that we have introduced two different criteria of straightness or flatness in M, namely primal and dual flatness. M is dually flat and the two flat coordinates are connected by the Legendre transformation.

1.5.2 Tangent Space, Basis Vectors and Riemannian Metric

When $d\boldsymbol{\theta}$ is an (infinitesimally) small line element, the square of its length ds is given by

$$ds^2 = 2D_\psi [\boldsymbol{\theta} : \boldsymbol{\theta} + d\boldsymbol{\theta}] = \sum g_{ij} d\theta^i d\theta^j. \tag{1.85}$$

Here, we use the upper indices i, j to represent components of $\boldsymbol{\theta}$. It is easy to see that the Riemannian metric g_{ij} is given by the Hessian of ψ

$$g_{ij}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta^i \partial\theta^j}\psi(\boldsymbol{\theta}). \tag{1.86}$$

Let $\{\boldsymbol{e}_i, i = 1, \ldots, n\}$ be the set of tangent vectors along the coordinate curves of $\boldsymbol{\theta}$ (Fig. 1.5). The vector space spanned by $\{\boldsymbol{e}_i\}$ is the tangent space of M at each point. Since $\boldsymbol{\theta}$ is an affine coordinate system, $\{\boldsymbol{e}_i\}$ looks the same at any point. A tangent vector $\boldsymbol{A}$ is represented as

$$\boldsymbol{A} = \sum A^i \boldsymbol{e}_i, \tag{1.87}$$

where A^i are the components of $\boldsymbol{A}$ with respect to the basis vectors $\{\boldsymbol{e}_i\}, i = 1, \ldots, n$. The small line element $d\boldsymbol{\theta}$ is a tangent vector expressed as

$$d\boldsymbol{\theta} = \sum d\theta^i \boldsymbol{e}_i. \tag{1.88}$$

Dually, we introduce a set of basis vectors $\{\boldsymbol{e}^{*i}\}$ which are tangent vectors of the dual affine coordinate curves of $\boldsymbol{\theta}^*$ (Fig. 1.6). The small line element $d\boldsymbol{\theta}^*$ is expressed as

$$d\boldsymbol{\theta}^* = \sum d\theta^*_i \boldsymbol{e}^{*i} \tag{1.89}$$

in this basis. A vector $\boldsymbol{A}$ is represented in this basis as

$$\boldsymbol{A} = \sum A_i \boldsymbol{e}^{*i}. \tag{1.90}$$

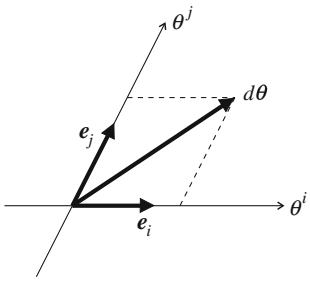


Fig. 1.5 Basis vectors $\boldsymbol{e}_i$ and small line element $d\boldsymbol{\theta}$

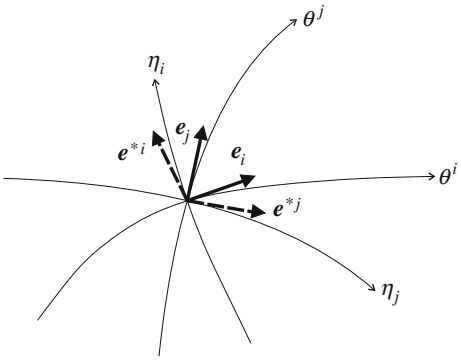


Fig. 1.6 Two dual bases $\{e_i\}$ and $\{e^{*i}\}$

In order to distinguish affine and dual affine bases, we use the lower index as in $\boldsymbol{e}_i$ for the affine basis and the upper index as in $\boldsymbol{e}^{*i}$ for the dual affine basis. Then, by using the lower and upper indices as in A^i and A_i in the two bases, the components of a vector are naturally expressed without changing the letter A but by changing the position of the index to upper or lower. Since they are the same vector expressed in different bases,

$$\boldsymbol{A} = \sum A^i \boldsymbol{e}_i = \sum A_i \boldsymbol{e}^{*i}, \tag{1.91}$$

and $A_i \neq A^i$ in general.

It is cumbersome to use the summation symbol in Eqs. (1.87)–(1.91) and others. Even if the summation symbol is discarded, the reader may consider from the context that it has been omitted by mistake. In most cases, index i appearing twice in one term, once as an upper index and the other time as a lower index, is summed over from 1 to n. A. Einstein introduced the following summation convention:

Einstein Summation Convention: When the same index appears twice in one term, once as an upper index and the other time as a lower index, summation is automatically taken over this index even without the summation symbol.

We use this convention throughout the monograph, unless specified otherwise. Then, (1.91) is rewritten as

$$\boldsymbol{A} = A^i \boldsymbol{e}_i = A_i \boldsymbol{e}^{*i}. \tag{1.92}$$

Since the square of the length ds of a small line element $d\boldsymbol{\theta}$ is given by the inner product of $d\boldsymbol{\theta}$, we have

$$ds^2 = \langle d\boldsymbol{\theta}, d\boldsymbol{\theta} \rangle = g_{ij} d\theta^i d\theta^j, \tag{1.93}$$

which is rewritten as

$$ds^2 = \langle d\theta^i \boldsymbol{e}_i, d\theta^j \boldsymbol{e}_j \rangle = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle d\theta^i d\theta^j. \tag{1.94}$$

Therefore, we have

$$g_{ij}(\boldsymbol{\theta}) = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle. \tag{1.95}$$

This is the inner product of basis vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, which depends on position $\boldsymbol{\theta}$.

A manifold equipped with $\mathbf{G} = \left(g_{ij}\right)$, by which the length of a small line element $d\boldsymbol{\theta}$ is given by (1.93), is a Riemannian manifold. In the case of a Euclidean space with an orthonormal coordinate system, g_{ij} is given by

$$g_{ij} = \delta_{ij}, \tag{1.96}$$

where δ_{ij} is the Kronecker delta, which is equal to 1 for $i = j$ and 0 otherwise. This is derived from convex function (1.37). A Euclidean space is a special case of the Riemannian manifold in which there is a coordinate system such that g_{ij} does not depend on position, in particular, written as (1.96). A manifold induced from a convex function is not Euclidean in general.

The Riemannian metric can also be represented in the dual affine coordinate system $\boldsymbol{\theta}^*$. From the representation of a small line element $d\boldsymbol{\theta}^*$ as

$$d\boldsymbol{\theta}^* = d\theta_i^* \boldsymbol{e}^{*i}, \tag{1.97}$$

we have

$$ds^2 = \langle d\boldsymbol{\theta}^*, d\boldsymbol{\theta}^* \rangle = g^{*ij} d\theta_i^* d\theta_j^*, \tag{1.98}$$

where g^{*ij} is given by

$$g^{*ij} = \langle \boldsymbol{e}^{*i}, \boldsymbol{e}^{*j} \rangle. \tag{1.99}$$

From (1.66), we see that the components of the small line elements $d\boldsymbol{\theta}$ and $d\boldsymbol{\theta}^*$ are related as

$$d\boldsymbol{\theta}^* = \mathbf{G} d\boldsymbol{\theta}, \quad d\boldsymbol{\theta} = \mathbf{G}^{-1} d\boldsymbol{\theta}^*, \tag{1.100}$$

$$d\theta_i^* = g_{ij} d\theta^j, \quad d\theta^j = g^{*ji} d\theta_i^*, \tag{1.101}$$

where $\mathbf{G} = \mathbf{G}^{*-1}$. So the two Riemannian metric tensors are mutually inverse.

This also implies that the two bases are related as

$$\boldsymbol{e}^{*i} = g^{ij} \boldsymbol{e}_j, \quad \boldsymbol{e}_i = g_{ij} \boldsymbol{e}^{*j}. \tag{1.102}$$

Hence, the inner product of two basis vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}_j^*$ satisfies

$$\langle \boldsymbol{e}_i, \boldsymbol{e}^{*j} \rangle = \delta_i^j \tag{1.103}$$

because $\mathbf{G} = \mathbf{G}^{*-1}$. So the two bases $\{\boldsymbol{e}_i\}$ and $\left\{\boldsymbol{e}^{*i}\right\}$ are mutually dual or reciprocal (Fig. 1.6). Neither of the bases is orthonormal by itself in general, but the two together are complementarily orthogonal. Such a set of bases is useful, because the

components of a vector $\boldsymbol{A}$ are given by the inner product,

$$A^i = \langle \boldsymbol{A}, \boldsymbol{e}^{*i} \rangle, \quad A_i = \langle \boldsymbol{A}, \boldsymbol{e}_i \rangle. \tag{1.104}$$

The two components are connected by

$$A_i = g_{ij} A^j, \quad A^j = g^{*ij} A_i. \tag{1.105}$$

1.5.3 Parallel Transport of Vector

A tangent vector $\boldsymbol{A} = A^i \boldsymbol{e}_i$ defined at a point $\boldsymbol{\theta}$ is transported to another point $\boldsymbol{\theta}'$ without changing the components A^i, because $\boldsymbol{e}_i$ are the same everywhere in a dually flat manifold. This is a special case of parallel transport of a vector in a general non-flat manifold. As will be seen in Part II, the parallel transport of a vector needs to use an affine connection in the general case. But in our case of a dually flat manifold derived from a convex function $\psi(\boldsymbol{\theta})$, the parallel transport is very simple.

The dual parallel transport of $\boldsymbol{A}$ is different from the parallel transport of $\boldsymbol{A}$. When $\boldsymbol{A}$ is represented in the dual basis as

$$\boldsymbol{A} = A_i \boldsymbol{e}^{*i}, \tag{1.106}$$

the dual transport does not change the components A_i. However, it changes the components A^i, because the relation between A_i and A^i depends on position $\boldsymbol{\theta}$ or $\boldsymbol{\theta}^*$, as is seen from (1.105), where g_{ij} and g^{*ij} depend on $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$.

Since M is Riemannian and is not Euclidean in general, even though the parallel transport is defined easily, the length of a vector changes by the parallel transport and the dual parallel transport. The square of the magnitude of $\boldsymbol{A}$ is written as

$$|\boldsymbol{A}|^2 = \langle \boldsymbol{A}, \boldsymbol{A} \rangle = g_{ij}(\boldsymbol{\theta}) A^i A^j = A^i A_i. \tag{1.107}$$

Therefore, it depends on the position $\boldsymbol{\theta}$, even though the components of A^i do not change by parallel transport. The inner product of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ is represented by various forms,

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = g_{ij} A^i B^j = g^{*ij} A_i B_j = A_i B^i. \tag{1.108}$$

Two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal when $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = 0$. However, when both $\boldsymbol{A}$ and $\boldsymbol{B}$ are parallelly transported from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, the orthogonality does not hold in general at $\boldsymbol{\theta}'$ even when it holds at $\boldsymbol{\theta}$. However, when $\boldsymbol{A}$ is transported in parallel and $\boldsymbol{B}$ is transported in dual parallel, the orthogonality is kept invariant, because $A^i B_i$ is invariant. This is an important property of two dually coupled parallel transports.

1.6 Generalized Pythagorean Theorem and Projection Theorem

1.6.1 Generalized Pythagorean Theorem

Two curves $\boldsymbol{\theta}_1(t)$ and $\boldsymbol{\theta}_2(t)$ intersect orthogonally when their tangent vectors

$$\dot{\boldsymbol{\theta}}_1(t) = \frac{d}{dt}\boldsymbol{\theta}_1(t), \tag{1.109}$$

$$\dot{\boldsymbol{\theta}}_2(t) = \frac{d}{dt}\boldsymbol{\theta}_2(t) \tag{1.110}$$

are orthogonal, that is,

$$\langle \dot{\boldsymbol{\theta}}_1(t), \dot{\boldsymbol{\theta}}_2(t) \rangle = g_{ij}\dot{\theta}_1^i(t)\dot{\theta}_2^j(t) = 0 \tag{1.111}$$

at the intersection point $t = 0$, $\boldsymbol{\theta}_1(0) = \boldsymbol{\theta}_2(0)$ and $\dot{}$ denotes d/dt.

Even though a manifold is flat from the point of view of affine structures, it is different from a Euclidean space. A dually flat manifold is a generalization of the Euclidean space. A generalized Pythagorean theorem holds in a dually flat manifold M.

Let us consider three points P, Q, R in a dually flat manifold M, which form a triangle. We call it an orthogonal triangle when the dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and R (Fig. 1.7).

Theorem 1.2 (Generalized Pythagorean Theorem) *When triangle PQR is orthogonal such that the dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and R, the following generalized Pythagorean relation holds:*

$$D_\psi[P : R] = D_\psi[P : Q] + D_\psi[Q : R]. \tag{1.112}$$

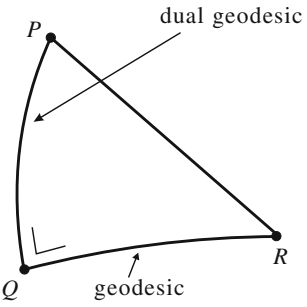


Fig. 1.7 Generalized orthogonal triangle ΔPQR and Pythagorean theorem

Proof By using the relation

$$D_\psi[P:Q] = \psi(\boldsymbol{\theta}_P) + \psi^*\left(\boldsymbol{\theta}^*_Q\right) - \boldsymbol{\theta}_P \cdot \boldsymbol{\theta}^*_Q, \tag{1.113}$$

we have

$$D_\psi[P:Q] + D_\psi[Q:R] - D_\psi[P:R] = \left(\boldsymbol{\theta}^*_P - \boldsymbol{\theta}^*_Q\right) \cdot \left(\boldsymbol{\theta}_Q - \boldsymbol{\theta}_R\right) \tag{1.114}$$

after some calculations. The dual geodesic connecting P and Q is written as

$$\boldsymbol{\theta}^*_{PQ}(t) = (1-t)\boldsymbol{\theta}^*_P + t\boldsymbol{\theta}^*_Q, \tag{1.115}$$

in the parametric form. Its tangent vector is given by

$$\dot{\boldsymbol{\theta}}^*_{PQ}(t) = \boldsymbol{\theta}^*_Q - \boldsymbol{\theta}^*_P. \tag{1.116}$$

Dually, the geodesic connecting Q and R is

$$\boldsymbol{\theta}_{QR}(t) = (1-t)\boldsymbol{\theta}_Q + t\boldsymbol{\theta}_R \tag{1.117}$$

and its tangent vector is

$$\dot{\boldsymbol{\theta}}_{QR}(t) = \boldsymbol{\theta}_R - \boldsymbol{\theta}_Q. \tag{1.118}$$

Since the two tangent vectors are orthogonal, we have

$$\left(\boldsymbol{\theta}^*_P - \boldsymbol{\theta}^*_Q\right) \cdot \left(\boldsymbol{\theta}_Q - \boldsymbol{\theta}_R\right) = 0. \tag{1.119}$$

The Pythagorean relation is proved from (1.114). □

Since the divergence is asymmetric, we have the dual statement.

Theorem 1.3 (Dual Pythagorean Theorem) *When triangle PQR is orthogonal such that the geodesic connecting P and Q is orthogonal to the dual geodesic connecting Q and R, the dual of the generalized Pythagorean relation holds,*

$$D_{\psi^*}[P:R] = D_{\psi^*}[P:Q] + D_{\psi^*}[Q:R]. \tag{1.120}$$

In the special case of convex function (1.37), the divergence is exactly a half of the square of the Euclidean distance. Moreover, the affine coordinate system is exactly the same as the dual affine coordinate system, because the affine structure is self-dual. Hence, a geodesic is a dual geodesic at the same time. In this case, the generalized Pythagorean relation reduces to the Pythagorean relation in a Euclidean space. The theorems are indeed a generalization of the Pythagorean theorem of a Euclidean space to a dually flat manifold.

1.6.2 Projection Theorem

Consider a point P and a smooth submanifold S in a dually flat manifold M. Then, the divergence from a point P to submanifold S is defined by

$$D_\psi[P : S] = \min_{R \in S} D_\psi[P : R]. \tag{1.121}$$

We study the problem of finding the point in S that is closest to P in the sense of divergence. This gives an approximation of P by using a point inside S. The Pythagorean theorem is useful for solving various approximation problems.

We define the geodesic projection and the dual geodesic projection of P to $S \subset M$. A curve $\boldsymbol{\theta}(t)$ is said to be orthogonal to S when its tangent vector $\dot{\boldsymbol{\theta}}(t)$ is orthogonal to any tangent vectors of S at the intersection (Fig. 1.8).

Definition 1.2 $\hat{P}_S$ is the geodesic projection of P to S when the geodesic connecting P and $\hat{P}_S \in S$ is orthogonal to S. Dually, $\hat{P}_S^*$ is the dual geodesic projection of P to S, when the dual geodesic connecting P and $\hat{P}_S^* \in S$ is orthogonal to S. See Fig. 1.8.

We then have the projection theorem:

Theorem 1.4 (Projection Theorem) *Given $P \in M$ and a smooth submanifold $S \subset M$, the point $\hat{P}_S^*$ that minimizes the divergence $D_\psi[P : R]$, $R \in S$, is the dual geodesic projection of P to S. The point $\hat{P}_S$ that minimizes the dual divergence $D_{\psi^*}[P : R]$, $R \in S$, is the geodesic projection of P to S.*

Proof Let $\hat{P}_S^*$ be the dual geodesic projection of P to S. Consider a point $Q \in S$ which is (infinitesimally) close to $\hat{P}_S^*$. Then, three points P, $\hat{P}_S^*$ and Q form an orthogonal triangle, because the small line element connecting $\hat{P}_S^*$ and Q is orthogonal to the the dual geodesic connecting P and $\hat{P}_S^*$. Hence, the Pythagorean theorem shows

$$D_\psi[P : Q] = D_\psi\left[P : \hat{P}_S^*\right] + D_\psi\left[\hat{P}_S^* : Q\right] \tag{1.122}$$

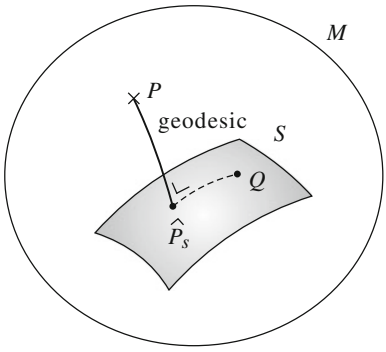


Fig. 1.8 Geodesic projection of P to S

for any neighboring Q. This shows that $\hat{P}_S^*$ is a critical point of $D_\psi[P : Q]$, $Q \in S$, proving the theorem. The dual part is proved similarly. □

It should be noted that the projection theorem gives a necessary condition for the point $\hat{P}_S^*$ to minimize the divergence, but is not sufficient. The projection or dual projection can give the maximum or saddle point of the divergence. The following theorem gives a sufficient condition for the minimality of the projection and its uniqueness.

Theorem 1.5 *When S is a flat submanifold of a dually flat manifold M, the dual projection of P to S is unique and minimizes the divergence. Dually, when S is a dual flat submanifold of a dually flat manifold M, the projection of P to S is unique and minimizes the dual divergence.*

Proof The Pythagorean relations (1.112), (1.120) hold for any $Q \in S$. Hence the projection (dual projection) is unique and minimizes the dual divergence (divergence). □

1.6.3 Divergence Between Submanifolds: Alternating Minimization Algorithm

When there are two submanifolds K and S in a dually flat M, we define a divergence between K and S by

$$D[K : S] = \min_{P \in K, Q \in S} D[P : Q] = D\left[\bar{P} : \bar{Q}\right]. \tag{1.123}$$

The two points $\bar{P} \in K$ and $\bar{Q} \in S$ are the closest pair between K and S. In order to obtain the closest pair, the following iterative algorithm, the alternating minimization algorithm, is proposed. See Fig. 1.9.

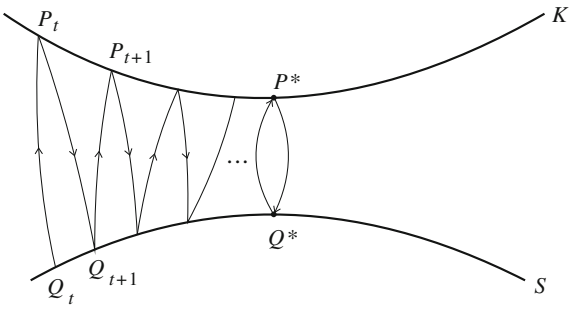


Fig. 1.9 Iterated dual geodesic projections (*em* algorithm)

Begin with an arbitrary $Q_t \in S, t = 0, 1, \ldots$, and search for $P \in K$ that minimizes $D[P : Q_t]$. This is given by the geodesic projection of Q_t to K. Let it be $P_t \in K$. Then search for the point in S that minimizes $D[P_t : Q]$. Let it be Q_{t+1}. This is given by the dual geodesic projection of P_t to S. Since we have

$$D\left[P_{t-1} : Q_t\right] \geq D[P_t : Q_t] \geq D\left[P_t : Q_{t+1}\right], \tag{1.124}$$

the procedure converges. It is unique when S is flat and K is dual flat. Otherwise, the converging point is not necessarily unique.

In later sections, the geodesic projection is called the e-projection, signifying the exponential projection, and the dual geodesic projection is called the m-projection, signifying the mixture projection. By this reason, this alternating primal and dual geodesic projection algorithm is called the *em* algorithm.

Remarks

A dually flat Riemannian structure is derived from the Bregman divergence by using a convex function. It has a dualistic structure. However, not all divergences are Bregman divergences, that is, not necessarily derived from convex functions. An interesting question is what type of geometry is induced from such a general divergence. This question will be studied in Part II. Briefly speaking, it gives a Riemannian manifold with a dual pair of affine connections which are not flat. There are no affine coordinate systems in such cases.

A dually flat manifold is a generalization of a Euclidean space, inheriting useful properties from it. A general non-flat manifold is regarded as a curved submanifold of a dually flat manifold, as a Riemannian manifold is a curved submanifold of a Euclidean space with higher dimensions. Therefore, it is important to study the properties of a dually flat manifold.

The Pythagorean theorem and related projection theorem are highlights of a dually flat manifold, proposed in Nagaoka and Amari (1982). However, this work was not published in a journal, because, unfortunately, it was rejected by major journals. These theorems play important roles in most applications of information geometry. The Pythagorean theorem has been known for many years in the case of the KL-divergence. It is information geometry that has generalized the Pythagorean relation applicable to any Bregman divergence. Conversely, when a manifold is dually flat from the geometrical point of view, we can prove that there is a convex function from which the dually flat structure is derived. This will be explained later.

We add a comment on the notation. There are many coordinate systems in a coordinate neighborhood of a manifold, because when $\boldsymbol{\xi}$ is a coordinate system, its transform $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)$,

$$\boldsymbol{\zeta} = \boldsymbol{f}(\boldsymbol{\xi}) \; ; \; \zeta_\kappa = f_\kappa(\xi_1, \ldots, \xi_n), \quad \kappa = 1, \ldots, n \tag{1.125}$$

is another coordinate system, provided $\boldsymbol{f}$ is differentiable and invertible. The Jacobian matrix $\mathbf{J} = (J_{\kappa i})$ of the coordinate transformation

$$J_{\kappa i} = \frac{\partial f_\kappa}{\partial \xi_i}, \quad i = 1, \ldots, n \tag{1.126}$$

is non-degenerate, that is, matrix $\mathbf{J}$ is invertible.

Here we use indices $i, j, \ldots$ to represent components in the coordinate system $\boldsymbol{\xi} = (\xi_i), i = 1, \ldots, n$ and Greek indices $\kappa, \lambda, \nu, \ldots$ for the coordinate system $\boldsymbol{\zeta} = (\zeta_\kappa), \kappa = 1, \ldots, n$. This is a convenient way of distinguishing coordinate systems. For example, a small line element connecting P and $P + dP$ is $d\boldsymbol{\xi} = (d\xi_i)$ in coordinate system $\boldsymbol{\xi}$ and $d\boldsymbol{\zeta} = (d\zeta_\kappa)$ in coordinate system $\boldsymbol{\zeta}$, and they are linearly connected by

$$d\zeta_\kappa = \sum_i J_{\kappa i} d\xi_i. \tag{1.127}$$

When ds is a local distance written as

$$ds^2 = \sum g_{ij} d\xi_i d\xi_j \tag{1.128}$$

in the coordinate system $\boldsymbol{\xi}$, it can be written as

$$ds^2 = \sum g_{\kappa\lambda} d\zeta_\kappa d\zeta_\lambda \tag{1.129}$$

in coordinate system $\boldsymbol{\zeta}$. Here, $\left(g_{ij}\right)$ and $(g_{\kappa\lambda})$ are different matrices connected by

$$g_{ij} = \sum_{\kappa,\lambda} J_{\kappa i} J_{\lambda j} g_{\kappa\lambda}. \tag{1.130}$$

Such a quantity is called a tensor. We use the same letter g for the Riemannian metric tensor, but indices i, j or κ, λ distinguish the coordinate system in which it is represented. In general, we may use the same letter for a quantity even if it is represented in different coordinate systems, distinguishing them by the letter types of indices. This is convenient for the index notation, introduced by Schouten (1954). We mainly follow this idea.

We may choose any coordinate system. The geometry should be the same whichever coordinate system we use. Mathematicians often do not like to use a coordinate system, because geometry should not depend on it. They say that the index notation is an ugly classic method of differential geometry, where tensors are represented by quantities having indices. So they use the coordinate-free method of abstract description. This is sometimes elegant. However, it is wiser to choose an adequate coordinate system, because the geometry is the same in whichever coordinate system it is analyzed. For Euclidean geometry, an orthonormal coordinate system is usually preferable. However, when we analyze a boundary value problem of the

heat equation in a Euclidean space, if the boundary is a circle, the polar coordinate system makes the boundary condition very simple. So in such a case, we use this.

Any coordinate system is permissible, but it is advisable to use a convenient one, instead of rejecting the usage of a coordinate system. This is the way in which engineers and physicists work.

Chapter 2
Exponential Families and Mixture Families of Probability Distributions

The present chapter studies the geometry of the exponential family of probability distributions. It is not only a typical statistical model, including many well-known families of probability distributions such as discrete probability distributions S_n, Gaussian distributions, multinomial distributions, gamma distributions, etc., but is associated with a convex function known as the cumulant generating function or free energy. The induced Bregman divergence is the KL-divergence. It defines a dually flat Riemannian structure. The derived Riemannian metric is the Fisher information matrix and the two affine coordinate systems are the natural (canonical) parameters and expectation parameters, well-known in statistics. An exponential family is a universal model of dually flat manifolds, because any Bregman divergence has a corresponding exponential family of probability distributions (Banerjee et al. 2005).

We also study the mixture family of probability distributions, which is the dual of the exponential family. Applications of the generalized Pythagorean theorem demonstrate how useful this is.

2.1 Exponential Family of Probability Distributions

The standard form of an exponential family is given by the probability density function

$$p(x, \boldsymbol{\theta}) = \exp\left\{\theta^i h_i(x) + k(x) - \psi(\boldsymbol{\theta})\right\}, \tag{2.1}$$

where x is a random variable, $\boldsymbol{\theta} = \left(\theta^1, \ldots, \theta^n\right)$ is an n-dimensional vector parameter to specify a distribution, $h_i(x)$ are n functions of x which are linearly independent, $k(x)$ is a function of x, ψ corresponds to the normalization factor and the Einstein summation convention is working. We introduce a new vector random variable $\boldsymbol{x} = (x_1, \ldots, x_n)$ by

$$x_i = h_i(x). \tag{2.2}$$

We further introduce a measure in the sample space $X = \{\boldsymbol{x}\}$ by

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_2

$$d\mu(\boldsymbol{x}) = \exp\{k(x)\}\, dx. \tag{2.3}$$

Then, (2.1) is rewritten as

$$p(x, \boldsymbol{\theta})dx = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}\, d\mu(\boldsymbol{x}). \tag{2.4}$$

Hence, we may put

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}, \tag{2.5}$$

which is a probability density function of $\boldsymbol{x}$ with respect to measure $d\mu(\boldsymbol{x})$.

The family of distributions

$$M = \{p(\boldsymbol{x}, \boldsymbol{\theta})\} \tag{2.6}$$

forms an n-dimensional manifold, where $\boldsymbol{\theta}$ is a coordinate system. From the normalization condition

$$\int p(\boldsymbol{x}, \boldsymbol{\theta})d\mu(\boldsymbol{x}) = 1, \tag{2.7}$$

ψ is written as

$$\psi(\boldsymbol{\theta}) = \log \int \exp(\boldsymbol{\theta} \cdot \boldsymbol{x})d\mu(\boldsymbol{x}). \tag{2.8}$$

We proved in Chap. 1 that $\psi(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$, known as the cumulant generating function in statistics and free energy in physics. A dually flat Riemannian structure is introduced in M by using $\psi(\boldsymbol{\theta})$. The affine coordinate system is $\boldsymbol{\theta}$, which is called the natural or canonical parameter of an exponential family. The dual affine parameter is given by the Legendre transformation,

$$\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta}), \tag{2.9}$$

which is the expectation of $\boldsymbol{x}$ denoted by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$,

$$\boldsymbol{\eta} = \mathbf{E}[\boldsymbol{x}] = \int \boldsymbol{x}\, p(\boldsymbol{x}, \boldsymbol{\theta})d\mu(\boldsymbol{x}). \tag{2.10}$$

This $\boldsymbol{\eta}$ is called the expectation parameter in statistics. Since the dual affine parameter $\boldsymbol{\theta}^*$ is nothing other than $\boldsymbol{\eta}$, we hereafter use $\boldsymbol{\eta}$, instead of $\boldsymbol{\theta}^*$, to represent the dual affine parameter in an exponential family. This is a conventional notation used in Amari and Nagaoka (2000), avoiding the cumbersome $*$ notation. So we have

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}). \tag{2.11}$$

Hence, $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are two affine coordinate systems connected by the Legendre transformation.

We use $\varphi(\boldsymbol{\eta})$ to denote the dual convex function $\psi^*\left(\boldsymbol{\theta}^*\right)$, the Legendre dual of ψ, which is defined by

$$\varphi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}} \{\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})\}. \tag{2.12}$$

In order to obtain $\varphi(\boldsymbol{\eta})$, we calculate the negative entropy of $p(\boldsymbol{x}, \boldsymbol{\theta})$, obtaining

$$\mathrm{E}\left[\log p(\boldsymbol{x}, \boldsymbol{\theta})\right] = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\mu(\boldsymbol{x}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}). \tag{2.13}$$

Given $\boldsymbol{\eta}$, the $\boldsymbol{\theta}$ that maximizes the right-hand side of (2.12) is given by the solution of $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$. Hence, the dual convex function ψ^* of ψ, which we hereafter denote as φ, is given by the negative entropy,

$$\varphi(\boldsymbol{\eta}) = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}, \tag{2.14}$$

where $\boldsymbol{\theta}$ is regarded as a function of $\boldsymbol{\eta}$ through $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$. The inverse transformation is given by

$$\boldsymbol{\theta} = \nabla\varphi(\boldsymbol{\eta}). \tag{2.15}$$

The divergence from $p(\boldsymbol{x}, \boldsymbol{\theta}')$ to $p(\boldsymbol{x}, \boldsymbol{\theta})$ is written as

$$\begin{aligned} D_\psi\left[\boldsymbol{\theta}' : \boldsymbol{\theta}\right] &= \psi\left(\boldsymbol{\theta}'\right) - \psi(\boldsymbol{\theta}) - \boldsymbol{\eta} \cdot \left(\boldsymbol{\theta}' - \boldsymbol{\theta}\right) \\ &= \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p\left(\boldsymbol{x}, \boldsymbol{\theta}'\right)} d\mu(\boldsymbol{x}) = D_{KL}\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]. \end{aligned} \tag{2.16}$$

The Riemannian metric is given by

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}), \tag{2.17}$$

$$g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta}), \tag{2.18}$$

for which we hereafter use the abbreviation

$$\partial_i = \frac{\partial}{\partial\theta^i}, \quad \partial^i = \frac{\partial}{\partial\eta_i}. \tag{2.19}$$

Here, the position of the index i is important. If it is lower, as in ∂_i, the differentiation is with respect to θ^i, whereas, if it is upper as in ∂^i, the differentiation is with respect to η_i.

The Fisher information matrix plays a fundamental role in statistics. We prove the following theorem which connects geometry and statistics.

Theorem 2.1 *The Riemannian metric in an exponential family is the Fisher information matrix defined by*

$$g_{ij} = \mathrm{E}\left[\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_j \log p(\boldsymbol{x}, \boldsymbol{\theta})\right]. \tag{2.20}$$

Proof From

$$\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) = x_i - \partial_i \psi(\boldsymbol{\theta}) = x_i - \eta_i, \tag{2.21}$$

we have

$$\mathbf{E}\left[\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_j \log p(\boldsymbol{x}, \boldsymbol{\theta})\right] = \mathbf{E}\left[(x_i - \eta_i)\left(x_j - \eta_j\right)\right], \tag{2.22}$$

which is equal to $\nabla\nabla\psi(\boldsymbol{\theta})$. This is the Riemannian metric derived from $\psi(\boldsymbol{\theta})$, as is shown in (1.56). □

2.2 Examples of Exponential Family: Gaussian and Discrete Distributions

There are many statistical models belonging to the exponential family. Here, we show only two well-known, important distributions.

2.2.1 Gaussian Distribution

The Gaussian distribution with mean μ and variance σ^2 has the probability density function

$$p(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{2.23}$$

We introduce a new vector random variable $\boldsymbol{x} = (x_1, x_2)$,

$$x_1 = h_1(x) = x, \tag{2.24}$$

$$x_2 = h_2(x) = x^2. \tag{2.25}$$

Note that x and x^2 are dependent, but are linearly independent. We further introduce new parameters

$$\theta^1 = \frac{\mu}{\sigma^2}, \tag{2.26}$$

$$\theta^2 = -\frac{1}{2\sigma^2}. \tag{2.27}$$

Then, (2.23) is written in the standard form,

$$p(x, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\right\}. \tag{2.28}$$

The convex function $\psi(\boldsymbol{\theta})$ is given by

$$\begin{aligned}\psi(\boldsymbol{\theta}) &= \frac{\mu^2}{2\sigma^2} + \log\left(\sqrt{2\pi}\sigma\right) \\ &= -\frac{\left(\theta^1\right)^2}{4\theta^2} - \frac{1}{2}\log\left(-\theta^2\right) + \frac{1}{2}\log\pi. \end{aligned} \tag{2.29}$$

Since x_1 and x_2 are not independent but satisfy the relation

$$x_2 = (x_1)^2, \tag{2.30}$$

we use the dominating measure of

$$d\mu(\boldsymbol{x}) = \delta\left(x_2 - x_1^2\right) dx, \tag{2.31}$$

where δ is the delta function.

The dual affine coordinates $\boldsymbol{\eta}$ are given from (2.10) as

$$\eta_1 = \mu, \quad \eta_2 = \mu^2 + \sigma^2. \tag{2.32}$$

2.2.2 Discrete Distribution

Distributions of discrete random variable x over $X = \{0, 1, \ldots, n\}$ form a probability simplex S_n. A distribution $\boldsymbol{p} = (p_0, p_1, \ldots, p_n)$ is represented by

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x). \tag{2.33}$$

We show that S_n is an exponential family. We have

$$\begin{aligned}\log p(x) &= \sum_{i=0}^{n} (\log p_i)\, \delta_i(x) = \sum_{i=1}^{n} (\log p_i)\, \delta_i(x) + (\log p_0)\, \delta_0(x) \\ &= \sum_{i=1}^{n} \left(\log \frac{p_i}{p_0}\right) \delta_i(x) + \log p_0, \end{aligned} \tag{2.34}$$

because of

$$\delta_0(x) = 1 - \sum_{i=1}^{n} \delta_i(x). \tag{2.35}$$

We introduce new random variables x_i,

$$x_i = h_i(x) = \delta_i(x), \quad i = 1, \ldots, n \tag{2.36}$$

and new parameters

$$\theta^i = \log \frac{p_i}{p_0}. \tag{2.37}$$

Then, a discrete distribution $\boldsymbol{p}$ is written from (2.34) as

$$p(x, \boldsymbol{\theta}) = \exp\left\{ \sum_{i=1}^{n} \theta^i x_i - \psi(\boldsymbol{\theta}) \right\}, \tag{2.38}$$

where the cumulant generating function is

$$\psi(\boldsymbol{\theta}) = -\log p_0 = \log\left\{ 1 + \sum_{i=1}^{n} \exp\left(\theta^i\right) \right\}. \tag{2.39}$$

The dual affine coordinates $\boldsymbol{\eta}$ are

$$\eta_i = E\left[h_i(x)\right] = p_i, \quad i = 1, \ldots, n. \tag{2.40}$$

The dual convex function is the negative entropy,

$$\varphi(\boldsymbol{\eta}) = \sum \eta_i \log \eta_i + \left(1 - \sum \eta_i\right) \log\left(1 - \sum \eta_i\right). \tag{2.41}$$

By differentiating it, we have $\boldsymbol{\theta} = \nabla\varphi(\boldsymbol{\eta})$.

$$\theta^i = \log \frac{\eta_i}{1 - \sum \eta_i}. \tag{2.42}$$

2.3 Mixture Family of Probability Distributions

A mixture family is in general different from an exponential family, but family S_n of discrete distributions is an exponential family and a mixture family at the same time. We show that the two families play a dual role.

Given $n + 1$ probability distributions $q_0(x), q_1(x), \ldots, q_n(x)$ which are linearly independent, we compose a family of probability distributions given by

$$p(x, \boldsymbol{\eta}) = \sum_{i=0}^{n} \eta_i q_i(x), \tag{2.43}$$

where

$$\sum_{i=0}^{n} \eta_i = 1, \quad \eta_i > 0. \tag{2.44}$$

This is a statistical model called a mixture family, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ is a coordinate system and $\eta_0 = 1 - \sum \eta_i$. (We sometimes consider the closure of the above family, where $\eta_i \geq 0$.)

As is easily seen from (2.33), a discrete distribution $p(x) \in S_n$ is a mixture family, where

$$q_i(x) = \delta_i(x), \quad \eta_i = p_i, \quad i = 0, 1, \ldots, n. \tag{2.45}$$

Hence, $\boldsymbol{\eta}$ is a dual affine coordinate system of the exponential family S_n. We consider a general mixture family (2.43) which is not an exponential family. Even in this case, the negative entropy

$$\varphi(\boldsymbol{\eta}) = \int p(x, \boldsymbol{\eta}) \log p(x, \boldsymbol{\eta}) dx \tag{2.46}$$

is a convex function of $\boldsymbol{\eta}$. Therefore, we regard it as a dual convex function and introduce the dually flat structure to $M = \{p(x, \boldsymbol{\eta})\}$, having $\boldsymbol{\eta}$ as the dual affine coordinate system. Then, the primary affine coordinates are given by the gradient,

$$\boldsymbol{\theta} = \nabla \varphi(\boldsymbol{\eta}). \tag{2.47}$$

It defines the primal affine structure dually coupled with $\boldsymbol{\eta}$, although $\boldsymbol{\theta}$ is not the natural parameter of an exponential family, except for the case of S_n where $\boldsymbol{\theta}$ is the natural parameter.

The divergence given by $\varphi(\boldsymbol{\eta})$ is the KL-divergence

$$D_{\varphi}\left[\boldsymbol{\eta} : \boldsymbol{\eta}'\right] = \int p(x, \boldsymbol{\eta}) \log \frac{p(x, \boldsymbol{\eta})}{p\,(x, \boldsymbol{\eta}')} dx. \tag{2.48}$$

2.4 Flat Structure: *e*-flat and *m*-flat

The manifold M of exponential family is dually flat. The primal affine coordinates which define straightness or flatness are the natural parameter $\boldsymbol{\theta}$ in an exponential family. Let us consider the straight line, that is a geodesic, connecting two distributions $p(\boldsymbol{x}, \boldsymbol{\theta}_1)$ and $p(\boldsymbol{x}, \boldsymbol{\theta}_2)$. This is written in the $\boldsymbol{\theta}$ coordinate system as

$$\boldsymbol{\theta}(t) = (1 - t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2, \tag{2.49}$$

where t is the parameter. The probability distributions on the geodesic are

$$p(\boldsymbol{x}, t) = p\{\boldsymbol{x}, \boldsymbol{\theta}(t)\} = \exp\{t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \cdot \boldsymbol{x} + \boldsymbol{\theta}_1 \boldsymbol{x} - \psi(t)\}. \tag{2.50}$$

Hence, a geodesic itself is a one-dimensional exponential family, where t is the natural parameter.

By taking the logarithm, we have

$$\log p(\boldsymbol{x}, t) = (1-t) \log p(\boldsymbol{x}, \boldsymbol{\theta}_1) + t \log p(\boldsymbol{x}, \boldsymbol{\theta}_2) - \psi(t). \tag{2.51}$$

Therefore, a geodesic consists of a linear interpolation of the two distributions in the logarithmic scale. Since (2.51) is an exponential family, we call it an e-geodesic, e standing for "exponential". More generally, a submanifold which is defined by linear constraints in $\boldsymbol{\theta}$ is said to be e-flat. The affine parameter $\boldsymbol{\theta}$ is called the e-affine parameter.

The dual affine coordinates are $\boldsymbol{\eta}$, and define the dual flat structure. The dual geodesic connecting two distributions specified by $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ is given by

$$\boldsymbol{\eta}(t) = (1-t)\boldsymbol{\eta}_1 + t\boldsymbol{\eta}_2 \tag{2.52}$$

in terms of the dual coordinate system. Along the dual geodesic, the expectation of $\boldsymbol{x}$ is linearly interpolated,

$$\mathbf{E}_{\eta(t)}[\boldsymbol{x}] = (1-t)\mathbf{E}_{\eta_1}[\boldsymbol{x}] + t\mathbf{E}_{\eta_2}[\boldsymbol{x}]. \tag{2.53}$$

In the case of discrete probability distributions S_n, the dual geodesic connecting $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ is

$$\boldsymbol{p}(t) = (1-t)\boldsymbol{p}_1 + t\boldsymbol{p}_2, \tag{2.54}$$

which is a mixture of two distributions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. Hence, a dual geodesic is a mixture of two probability distributions. We call a dual geodesic an m-geodesic and, by this reasoning, $\boldsymbol{\eta}$ is called the m-affine parameter, where m stands for "mixture". A submanifold which is defined by linear constraints in $\boldsymbol{\eta}$ is said to be m-flat. The linear mixture

$$(1-t)p\left(\boldsymbol{x}, \boldsymbol{\eta}_1\right) + tp\left(\boldsymbol{x}, \boldsymbol{\eta}_2\right) \tag{2.55}$$

is not included in M in general, but $p\left(\boldsymbol{x}, (1-t)\boldsymbol{\eta}_1 + t\boldsymbol{\eta}_2\right)$ is in M, where we used the abuse of notation $p(\boldsymbol{x}, \boldsymbol{\eta})$ to specify the distribution of M of which dual coordinates are $\boldsymbol{\eta}$.

Remark An m-geodesic (2.52) is not a linear mixture of two distributions specified by $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ in the case of a general exponential family. However, we use the term m-geodesic even in this case.

2.5 On Infinite-Dimensional Manifold of Probability Distributions

We have shown that S_n of discrete probability distributions is an exponential family and a mixture family at the same time. It is a super-manifold, in which any statistical model of a discrete random variable is embedded as a submanifold. When x is a continuous random variable, we are apt to consider the geometry of the manifold F of all probability density functions $p(x)$ in a similar way. It is a super-manifold including all statistical models of a continuous random variable. It is considered to be an exponential family and a mixture family at the same time. However, the problem is not mathematically easy, since it is a function space of infinite dimensions. We show a naive idea of studying the geometry of F. This is not mathematically justified, although it works well in most cases, except for "pathological" situations.

Let $p(x)$ be a probability density function of real random variable $x \in \boldsymbol{R}$, which is mutually absolutely continuous with respect to the Lebesgue measure.[1] We put

$$F = \left\{ p(x) \,\middle|\, p(x) > 0, \ \int p(x)dx = 1 \right\}. \tag{2.56}$$

Then, F is a function space consisting of L_1 functions. For two distributions $p_1(x)$ and $p_2(x)$, the exponential family connecting them is written as

$$p_{\exp}(x, t) = \exp\{(1-t)\log p_1(x) + t \log p_2(x) - \psi(t)\}, \tag{2.57}$$

provided it exists in F. Also the mixture family connecting them

$$p_{\text{mix}}(x, t) = (1-t)p_1(x) + tp_2(x) \tag{2.58}$$

is assumed to belong to F. Then, F is regarded as an exponential and a mixture family at the same time as S_n is. Mathematically, there is a delicate problem concerning the topology of F. The L_1-topology and L_2-topology of the function space F are different. Also the topology induced by $p(x)$ is different from that induced by $\log p(x)$.

Disregarding such mathematical problems, we discretize the real line $\boldsymbol{R}$ into $n+1$ intervals, $I_0, I_1, \ldots, I_n$. Then, the discretized version of $p(x)$ is given by the discrete probability distribution $\boldsymbol{p} = (p_0, p_1, \ldots, p_n)$,

$$p_i = \int_{I_i} p(x)dx, \quad i = 0, 1, \ldots, n. \tag{2.59}$$

[1] It would be better to use density function $p(x)$ with respect to the Gaussian measure

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

rather than the Lebsesque measure dx.

This gives a mapping from F to S_n, which approximates $p(x)$ by $\boldsymbol{p} \in S_n$. When the discretization is done in such a way that p_i in each interval converges to 0 as n tends to infinity, the approximation looks fine. Then, the geometry of F would be defined by the limit of S_n consisting of discretized $\boldsymbol{p}$. However, we have difficulty in this approach. The limit $n \to \infty$ of the geometry of S_n might not be unique, depending on the method of discretization. Moreover, an admissible discretization would be different for different $p(\boldsymbol{x})$.

Forgetting about the difficulty, by using the delta function $\delta(x)$, let us introduce a family of random variables $\delta(s - x)$ indexed by a real parameter s, which plays the role of index i in $\delta_i(x)$ of S_n. Then, we have

$$p(x) = \int p(s)\delta(x - s)ds, \tag{2.60}$$

which shows that F is a mixture family generated by the delta distributions $\delta(s - x)$, $s \in \boldsymbol{R}$. Here, $p(s)$ are mixing coefficients. Similarly, we have

$$p(x) = \exp\left\{\int \theta(s)\delta(s - x)dx - \psi\right\}, \tag{2.61}$$

where

$$\theta(s) = \log p(s) + \psi \tag{2.62}$$

and ψ is a functional of $\theta(s)$ formally given by

$$\psi[\theta(s)] = \log\left\{\int \exp\{\theta(s)\}\, ds\right\}. \tag{2.63}$$

Hence, F is an exponential family where $\theta(s) = \log p(s) + \psi$ is the $\boldsymbol{\theta}$ affine coordinates and $\eta(s) = p(s)$ is the dual affine coordinates $\boldsymbol{\eta}$. The dual convex function is

$$\varphi[\eta(s)] = \int \eta(s) \log \eta(s)ds. \tag{2.64}$$

Indeed the dual coordinates are given by

$$\eta(s) = \mathbf{E}_p[\delta(s - x)] = p(s) \tag{2.65}$$

and we have

$$\eta(s) = \nabla\psi[\theta(s)], \tag{2.66}$$

where ∇ is the Fréchet-derivative with respect to function $\theta(s)$. The e-geodesic connecting $p(x)$ and $q(x)$ is (2.57) and the m-geodesic (2.58). The tangent vector of an e-geodesic is

$$\frac{d}{dt}\log p(x,t) = \dot{l}(x,t) = \log q(x) - \log p(x) \tag{2.67}$$

in the e-coordinates, and that of an m-geodesic is

$$\dot{p}(x,t) = q(x) - p(x) \tag{2.68}$$

in the m-coordinates.

The KL-divergence is

$$D_{KL}\left[p(x) : q(x)\right] = \int p(x)\log\left\{\frac{p(x)}{q(x)}\right\}dx, \tag{2.69}$$

which is the Bregman divergence derived from $\psi[\theta]$ and it gives F a dually flat structure. The Pythagorean theorem is written, for three distributions $p(x)$, $q(x)$ and $r(x)$, as

$$D_{KL}\left[p(x) : r(x)\right] = D_{KL}\left[p(x) : q(x)\right] + D_{KL}\left[q(x) : r(x)\right], \tag{2.70}$$

when the mixture geodesic connecting p and q is orthogonal to the exponential-geodesic connecting q and r, that is, when

$$\int \{p(x) - q(x)\}\{\log r(x) - \log q(x)\}\,dx = 0. \tag{2.71}$$

It is easy to prove this directly. The projection theorem follows similarly.

The KL-divergence between two nearby distributions $p(x)$ and $p(x) + \delta p(x)$ is expanded as

$$\begin{aligned} D_{KL}\left[p(x) : p(x) + \delta p(x)\right] &= \int p(x)\log\left\{1 - \frac{\delta p(x)}{p(x)}\right\}dx \\ &= \frac{1}{2}\int \frac{\{\delta p(x)\}^2}{p(x)}dx. \end{aligned} \tag{2.72}$$

Hence, the squared distance of an infinitesimal deviation $\delta p(x)$ is

$$ds^2 = \int \frac{\{\delta p(x)\}^2}{p(x)}dx, \tag{2.73}$$

which defines the Riemannian metric given by the Fisher information.

Indeed, the Riemannian metric in $\boldsymbol{\theta}$-coordinates are given by

$$g(s,t) = \nabla\nabla\psi = p(s)\delta(s-t) \tag{2.74}$$

and its inverse is

$$g^{-1}(s,t) = \frac{1}{p(s)}\delta(s-t) \tag{2.75}$$

in $\boldsymbol{\eta}$-coordinates.

It appears that most of the results we have studied in S_n hold well even in the function space F with naive treatment. They are practically useful even though no mathematical justification is given. Unfortunately, we are not free from mathematical difficulties. We show some examples.

The pathological nature in the continuous case has long been known. The following fact was pointed out by Csiszár (1967). We define a quasi-ε-neighborhood of $p(x)$ based on the KL-divergence,

$$N_\varepsilon = \{q(x) \,|\, D_{KL}\,[p(x) : q(x)] < \varepsilon\}\,. \tag{2.76}$$

However, the set of the quasi-ε-neighborhoods does not satisfy the axiom of the topological subbase. Hence, we cannot use the KL-divergence to define the topology. More simply, it is demonstrated that the entropy functional

$$\varphi[p(x)] = \int p(x)\log p(x)dx \tag{2.77}$$

is not continuous in F, whereas it is continuous and differentiable in S_n (Ho and Yeung 2009).

G. Pistone and his co-workers studied the geometrical properties of F based on the theory of Orlicz space, where F is not a Hilbert space but a Banach space. See Pistone and Sempi (1995), Gibilisco and Pistone (1998), Pistone and Rogathin (1999), Cena and Pistone (2007). This was further developed by Grasselli (2010). See recent works by Pistone (2013) and Newton (2012), where trials for mathematical justification using innocent ideas have been developed.

2.6 Kernel Exponential Family

Fukumizu (2009) proposed a kernel exponential family, which is a model of probability distributions of function degrees of freedom. Let $k(x, y)$ be a kernel function satisfying positivity,

$$\int k(x,y)f(x)f(y)dxdy > 0 \tag{2.78}$$

for any $f(x)$ not equal to 0. A typical example is the Gaussian kernel

$$k_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-y)^2\right\}, \tag{2.79}$$

where σ is a free parameter.

A kernel exponential family is defined by

$$p(x,\theta) = \exp\left\{\int \theta(y)k(x,y)dx - \psi[\theta]\right\} \tag{2.80}$$

with respect to suitable measure $d\mu(x)$, e.g.,

$$d\mu(x) = \exp\left\{-\frac{x^2}{2\tau^2}\right\} dx. \tag{2.81}$$

The natural or canonical parameter is a function $\theta(y)$ indexed by y instead of θ^i and the dual parameter is

$$\eta(y) = \mathrm{E}[k(x,y)], \tag{2.82}$$

where expectation is taken by using $p(x,\theta)$. $\psi[\theta]$ is a convex functional of $\theta(y)$. This exponential family does not cover all $p(x)$ of probability density functions. So there are many such models, depending on $k(x,y)$ and $d\mu(x)$. The naive treatment in Sect. 2.5 may be regarded as the special case where the kernel $k(x,y)$ is put equal to the delta function $\delta(x-y)$.

2.7 Bregman Divergence and Exponential Family

An exponential family induces a Bregman divergence $D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]$ given in (2.16). Conversely, when a Bregman divergence $D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]$ is given, is it possible to find a corresponding exponential family $p(\boldsymbol{x},\boldsymbol{\theta})$? The problem is solved positively by Banerjee et al. (2005). Consider a random variable $\boldsymbol{x}$. It specifies a point $\boldsymbol{\eta}' = \boldsymbol{x}$ in the $\boldsymbol{\eta}$-coordinates of a dually flat manifold given by ψ. Let $\boldsymbol{\theta}'$ be its $\boldsymbol{\theta}$-coordinates. The ψ-divergence from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, the latter of which is the $\boldsymbol{\theta}$-coordinates of $\boldsymbol{\eta}' = \boldsymbol{x}$, is written as

$$D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'(\boldsymbol{x})\right] = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{x}) - \boldsymbol{\theta}\cdot\boldsymbol{x}. \tag{2.83}$$

Using this, we define a probability density function written in terms of the divergence as

$$p(\boldsymbol{x},\boldsymbol{\theta}) = \exp\left\{-D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] + \varphi(\boldsymbol{x})\right\} = \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \psi(\boldsymbol{\theta})\right\}, \tag{2.84}$$

where $\boldsymbol{\theta}'$ is determined from $\boldsymbol{x}$ as the $\boldsymbol{\theta}$-coordinates of $\boldsymbol{\eta}' = \boldsymbol{x}$. Thus, we have an exponential family derived from D_ψ.

The problem is restated as follows: Given a convex function $\psi(\boldsymbol{\theta})$, find a measure $d\mu(\boldsymbol{x})$ such that (2.8), or equivalently

$$\exp\left\{\psi(\boldsymbol{\theta})\right\} = \int \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x}\right\} d\mu(\boldsymbol{x}), \tag{2.85}$$

is satisfied. This is the inverse of the Laplace transform. A mathematical theory concerning the one-to-one correspondence between (regular) exponential families and (regular) Bregman divergences is established in Banerjee et al. (2005).

Theorem 2.2 *There is a bijection between regular exponential families and regular Bregman divergences.*

The theorem shows that a Bregman divergence has a probabilistic expression given by an exponential family of probability distributions. A Bregman divergence is always written in the form of the KL-divergence of the corresponding exponential family.

Remark A mixture family $M = \{p(x, \boldsymbol{\eta})\}$ has a dually flat structure, where the negative entropy $\varphi(\boldsymbol{\eta})$ is a convex function. We can define an exponential family of which the convex function is $\varphi(\boldsymbol{\theta})$. However, this is different from the original M. Hence, Theorem 2.2 does not imply that a mixture family is an exponential family, even though it is dually flat.

2.8 Applications of Pythagorean Theorem

A few applications of the generalized Pythagorean Theorem are shown here to illustrate its usefulness.

2.8.1 Maximum Entropy Principle

Let us consider discrete probability distributions $S_n = \{p(x)\}$, although the following arguments hold even when x is a continuous vector random variable. Let $c_1(x), \ldots, c_k(x)$ be k random variables, that is, k functions of x. Their expectations are

$$\mathrm{E}\,[c_i(x)] = \sum p(x)c_i(x), \quad i = 1, 2, \ldots, k. \tag{2.86}$$

We consider a probability distribution $p(x)$ for which the expectations of $c_i(x)$ take prescribed values $\boldsymbol{a} = (a_1, \ldots, a_k)$,

$$\mathrm{E}\,[c_i(x)] = a_i, \quad i = 1, 2, \ldots, k. \tag{2.87}$$

There are many such distributions and they form an $(n-k)$-dimensional submanifold $M_{n-k}(\boldsymbol{a}) \subset S_n$ specified by $\boldsymbol{a}$, because k restrictions given by (2.87) are imposed. This M_{n-k} is m-flat, because any mixtures of distributions in M_{n-k} belong to the same M_{n-k}.

When one needs to choose a distribution from $M_{n-k}(\boldsymbol{a})$, if there are no other considerations, one would choose the distribution that maximizes the entropy. This is called the maximum entropy principle.

Let P_0 be the uniform distribution that maximizes the entropy in S_n. The dual divergence between $P \in S_n$ and P_0 is written as

$$D_\psi [P_0 : P] = \psi(\boldsymbol{\theta}_0) + \varphi(\boldsymbol{\eta}) - \boldsymbol{\theta}_0 \cdot \boldsymbol{\eta}, \tag{2.88}$$

where the e-coordinates of P_0 are given by $\boldsymbol{\theta}_0$, $\boldsymbol{\eta}$ is the m-coordinates of P and $\varphi(\boldsymbol{\eta})$ is the negative entropy. This is the KL-divergence $D_{KL}[P : P_0]$ from P to P_0. Since P_0 is the uniform distribution, $\boldsymbol{\theta}_0 = 0$. Hence, maximizing the entropy $\varphi(\boldsymbol{\eta})$ is equivalent to minimizing the divergence. Let $\hat{P} \in M_{n-k}$ be the point that maximizes the entropy. Then, triangle $P\hat{P}P_0$ is orthogonal and the Pythagorean relation

$$D_{KL}[P : P_0] = D_{KL}\left[P : \hat{P}\right] + D_{KL}\left[\hat{P} : P_0\right] \tag{2.89}$$

holds (Fig. 2.1). This implies that the entropy maximizer $\hat{P}$ is given by the e-projection of P_0 to $M_{n-k}(\boldsymbol{a})$.

Each $M_{n-k}(\boldsymbol{a})$ includes the entropy maximizer $\hat{P}(\boldsymbol{a})$. By changing $\boldsymbol{a}$, all of these $\hat{P}(\boldsymbol{a})$ form a k-dimensional submanifold E_k which is an exponential family, where the natural coordinates are specified by $\boldsymbol{\theta} = \boldsymbol{a}$ (Fig. 2.1),

$$\hat{p}(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{c}(x) - \psi(\boldsymbol{\theta})\}. \tag{2.90}$$

It is easy to obtain this result by the variational method that maximizes the entropy $\varphi(\boldsymbol{\eta})$ under constraints (2.87).

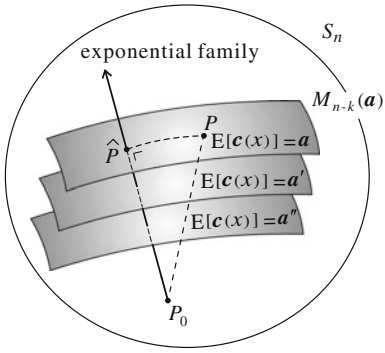


Fig. 2.1 The family maximizing entropy under linear constraints is an exponential family

2.8.2 Mutual Information

Let us consider two random variables x and y and the manifold M consisting of all $p(x, y)$. When x and y are independent, the probability can be written in the product form as

$$p(x, y) = p_X(x)p_Y(y), \tag{2.91}$$

where $p_X(x)$ and $p_Y(y)$ are respective marginal distributions.

Let the family of all the independent distributions be M_I. Since the exponential family connecting two independent distributions is again independent, the e-geodesic connecting them consists of independent distributions. Therefore, M_I is an e-flat submanifold.

Given a non-independent distribution $p(x, y)$, we search for the independent distribution which is closest to $p(x, y)$ in the sense of KL-divergence. This is given by the m-projection of $p(x, y)$ to M_I (Fig. 2.2). The projection is unique and given by the product of the marginal distributions

$$\hat{p}(x, y) = p_X(x)p_Y(y). \tag{2.92}$$

The divergence between $p(x, y)$ and its projection is

$$D_{KL}\left[p(x, y) : \hat{p}(x, y)\right] = \int p(x, y) \log \frac{p(x, y)}{\hat{p}(x, y)} dxdy, \tag{2.93}$$

which is the mutual information of two random variables x and y. Hence, the mutual information is a measure of discrepancy of $p(x, y)$ from independence.

The reverse problem is also interesting. Given an independent distribution (2.92), find the distribution $p(x, y)$ that maximizes $D_{KL}\left[p : \hat{p}\right]$ in the class of distributions having the same marginal distributions as $\hat{p}$. These distributions are the inverse image of the m-projection. This problem is studied by Ay and Knauf (2006) and Rauh (2011). See Ay (2002), Ay et al. (2011) for applications of information geometry to complex systems.

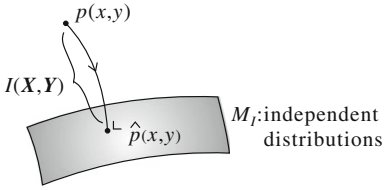


Fig. 2.2 Projection of $p(x, y)$ to the family M_I of independent distributions is the m-projection. The mutual information $I(\boldsymbol{X}, \boldsymbol{Y})$ is the KL-divergence $D_{KL}[p(x, y) : p_X(x)p_Y(y)]$

2.8.3 Repeated Observations and Maximum Likelihood Estimator

Statisticians use a number of independently observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from the same probability distribution $p(\boldsymbol{x}, \boldsymbol{\theta})$ in an exponential family M for estimating $\boldsymbol{\theta}$. The joint probability density of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ is given by

$$p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}\right) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right) \tag{2.94}$$

having the same parameter $\boldsymbol{\theta}$. We see how the geometry of M changes by multiple observations.

Let the arithmetic average of $\boldsymbol{x}_i$ be

$$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i. \tag{2.95}$$

Then, (2.94) is rewritten as

$$p_N\left(\bar{\boldsymbol{x}}, \boldsymbol{\theta}\right) = p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}\right) = \exp\left\{N\boldsymbol{\theta} \cdot \bar{\boldsymbol{x}} - N\psi(\boldsymbol{\theta})\right\}. \tag{2.96}$$

Therefore, the probability density of $\bar{\boldsymbol{x}}$ has the same form as $p(\boldsymbol{x}, \boldsymbol{\theta})$, except that $\boldsymbol{x}$ is replaced by $\bar{\boldsymbol{x}}$ and the term $\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})$ becomes N times larger.

This implies that the convex function becomes N times larger and hence the KL-divergence and Riemannian metric (Fisher information matrix) also become N times larger. The dual affine structure of M does not change. Hence, we may use the original M and the same coordinates $\boldsymbol{\theta}$ even when multiple observations take place for statistical inference. The binomial distributions and multinomial distributions are exponential families derived from S_2 and S_n by multiple observations.

Let M be an exponential family and consider a statistical model $S = \{p(\boldsymbol{x}, \boldsymbol{u})\}$ included in it as a submanifold, where S is specified by parameter $\boldsymbol{u} = (u_1, \ldots, u_k)$, $k < n$. Since it is included in M, the e-coordinates of $p(\boldsymbol{x}, \boldsymbol{u})$ in M are determined by $\boldsymbol{u}$ in the form of $\boldsymbol{\theta}(\boldsymbol{u})$. Given N independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, we estimate the parameter $\boldsymbol{u}$ based on them.

The observed data specifies a distribution in the entire M, such that its m-coordinates are

$$\bar{\boldsymbol{\eta}} = \frac{1}{N} \sum \boldsymbol{x}_i = \bar{\boldsymbol{x}}. \tag{2.97}$$

This is called an observed point. The KL-divergence from the observed $\bar{\boldsymbol{\eta}}$ to a distribution $p(\boldsymbol{x}, \boldsymbol{u})$ in S is written as $D_{KL}\left[\bar{\boldsymbol{\theta}} : \boldsymbol{\theta}(\boldsymbol{u})\right]$, where $\bar{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}$-coordinates of the observed point $\bar{\boldsymbol{\eta}}$. We consider a simple case of S_n, where the observed point is given by the histogram

$$\bar{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta (x - x_i) . \tag{2.98}$$

Then, except for a constant term, minimizing $D_{KL} [\bar{p}(x) : p(x, \boldsymbol{u})]$ is equivalent to maximizing the log-likelihood

$$L = \sum_{i=1}^{N} \log p (x_i, \boldsymbol{u}) . \tag{2.99}$$

Hence, the maximum likelihood estimator that minimizes the divergence is given by the m-projection of $\bar{p}(x)$ to S. See Fig. 2.3. In other words, the maximum likelihood estimator is characterized by the m-projection.

Remarks

An exponential family is an ideal model to study the dually flat structure and also statistical inference. The Legendre duality between the natural and expectation parameter was pointed out by Barndorff-Nielsen (1978). It is good news that the family S_n of discrete distributions is an exponential family, because any statistical model having a discrete random variable is regarded as a submanifold of an exponential family. Therefore, it is wise to study the properties of the exponential family first and then see how they are transferred to curved subfamilies.

Unfortunately, this is not the case with continuous random variable x. There are many statistical models which are not subfamilies of exponential families, even though many are curved-exponential families, that is, submanifolds of exponential families. Again, the study of the exponential family is useful. In the case of a truly non-exponential model, we use its local approximation by using a larger exponential family. This gives an exponential fibre-bundle-like structure to statistical models. This is useful for studying the asymptotic theory of statistical inference. See Amari (1985).

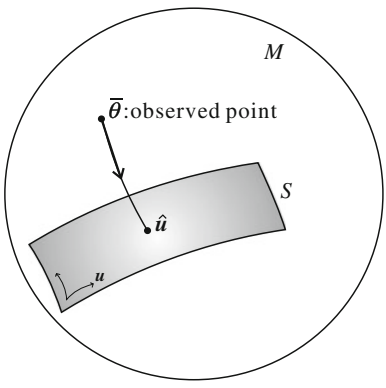


Fig. 2.3 The maximum likelihood estimator is the m-projection of observed point to S

It should be remarked that a generalized linear model provides a dually flat structure, although it is not an exponential family. See Vos (1991). A mixture model also has remarkable characteristics from the point of view of geometry. See Marriott (2002), Critchley et al. (1993).

Chapter 3
Invariant Geometry of Manifold of Probability Distributions

We have introduced a dually flat Riemannian structure in the manifolds of the exponential family and the mixture family based on the convexity of the cumulant generating function (free energy) and the negative entropy, respectively. The KL-divergence is derived from these convex functions. However, we need justification for this selection of convex function and divergence. Moreover, such a convex function does not exist for a general statistical model. Therefore, a reasonable criterion is needed for introducing a geometrical structure to a manifold of probability distributions. It is invariance that justifies the above selection.

Invariance requires that a geometrical structure should be invariant when random variable $\boldsymbol{x}$ is represented in a different form $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$, provided $\boldsymbol{y}(\boldsymbol{x})$ is invertible. This is an idea introduced by Chentsov (1972). We begin with a simpler idea of information monotonicity by coarse graining, due to Csiszár (1974), a simplified version of Chentsov's invariance. There exists a unique class of decomposable invariant divergences, known as f-divergences.

3.1 Invariance Criterion

We treat a statistical model

$$M = \{p(x, \boldsymbol{\xi})\}, \tag{3.1}$$

parameterized by $\boldsymbol{\xi}$, which forms a manifold with coordinate system $\boldsymbol{\xi}$. Here, x may take discrete, continuous and vector values. What is a natural divergence $D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]$ between two probability distributions $p(x, \boldsymbol{\xi})$ and $p(x, \boldsymbol{\xi}')$? In answering this question, we consider the invariance criterion, which states that the geometry is the same when random variable x is transformed into y without losing information. We consider a mapping of x to y

$$y = k(x), \tag{3.2}$$

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_3

which is in general many-to-one, so we cannot recover x from y. Then, information is lost by this mapping. Let the probability distribution of y be $\bar{p}(y, \boldsymbol{\xi})$,

$$\bar{p}(y, \boldsymbol{\xi}) = \sum_{x: k(x)=y} p(x, \boldsymbol{\xi}), \tag{3.3}$$

in the discrete case, which is induced from $p(x, \boldsymbol{\xi})$ by the mapping $y = k(x)$. In the continuous case, the probability density $\bar{p}(y, \boldsymbol{\xi})$ is given by integration. The divergence $D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]$ between $p(x, \boldsymbol{\xi})$ and $p\left(x, \boldsymbol{\xi}'\right)$ changes to $\bar{D}\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]$ between $\bar{p}(y, \boldsymbol{\xi})$ and $\bar{p}(y, \boldsymbol{\xi}')$. Since divergence $D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]$ represents the dissimilarity of two distributions $p(x, \boldsymbol{\xi})$ and $p\left(x, \boldsymbol{\xi}'\right)$, it is postulated that it decreases in general by this mapping,

$$\bar{D}\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] \leq D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]. \tag{3.4}$$

We call this relation information monotonicity.

Obviously, when k is one-to-one, that is invertible, there is no loss of information and the equality is required to hold in (3.4). However, there is a case when information is not lost even when k is not invertible. This is the case when x includes a redundant part, the distribution of which does not depend on $\boldsymbol{\xi}$. We may abandon this part without losing information concerning $\boldsymbol{\xi}$. The remaining part retains full information. Statisticians call such a part a sufficient statistic. Its definition is given below.

A function

$$s = k(x) \tag{3.5}$$

is called a sufficient statistic when the probability density function $p(x, \boldsymbol{\xi})$ is decomposed as

$$p(x, \boldsymbol{\xi}) = \bar{p}(s, \boldsymbol{\xi}) r(x). \tag{3.6}$$

This implies that the probability $p(x, \boldsymbol{\xi})$ is written as a function of s, except for a multiplicative term $r(x)$ which does not depend on $\boldsymbol{\xi}$. The equality is required to hold in (3.4) when and only when y is a sufficient statistic.

We formally state the invariance criterion, for which the basic idea was originally due to Chentsov (1972) and which was formulated in this way by Amari and Nagaoka (2000).

Invariance Criterion: A geometrical structure of M is invariant when it satisfies the monotonicity (3.4), where the equality holds if and only if $y = k(x)$ is a sufficient statistic.

We study the class of invariant divergences and invariant Riemannian metrics. The invariant metric is unique, given by the Fisher information matrix except for a scale constant (Chentsov 1972).

3.2 Information Monotonicity Under Coarse Graining

3.2.1 Coarse Graining and Sufficient Statistics in S_n

We consider a family S_n of discrete probability distributions, where random variable x takes on values $X = \{0, 1, \ldots, n\}$. Let us denote a probability distribution by an $(n+1)$-dimensional probability vector $\boldsymbol{p}$. We divide X into $m+1$ subsets $X_0, X_1, \ldots, X_m$ such that

$$\cup X_a = X, \quad X_a \cap X_b = \emptyset, \quad a \neq b, \tag{3.7}$$

where ϕ is the empty set. This is a partition of X (Fig. 3.1).

Assume that we cannot observe x directly, but know the subset to which x belongs. This is the case when X is coarse-grained. We then introduce a coarse-grained random variable y, taking on values $\{0, 1, \ldots, m\}$, where $y = a$ implies that x belongs to X_a. Let its distribution be denoted by $(m+1)$-dimensional probability vector $\bar{\boldsymbol{p}} = (\bar{p}_0, \ldots, \bar{p}_m)$. Coarse graining leads to a new distribution $\bar{\boldsymbol{p}}$ in S_m given by

$$\bar{p}_a = \sum_{i \in X_a} p_i. \tag{3.8}$$

Let $D\left[\boldsymbol{p} : \boldsymbol{q}\right]$ be a divergence between two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. It is said to be additive or decomposable when it is written in an additive form of componentwise divergences,

$$D[\boldsymbol{p} : \boldsymbol{q}] = \sum_{i=0}^{n} d\left(p_i, q_i\right) \tag{3.9}$$

for some function $d(p, q)$. The divergence $D[\boldsymbol{p} : \boldsymbol{q}]$ changes to $\bar{D}\left[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}\right]$ by coarse graining,

$$\bar{D}\left[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}\right] = \sum_{a=0}^{m} d\left(\bar{p}_a, \bar{q}_a\right). \tag{3.10}$$

The information monotonicity criterion requires

$$D[\boldsymbol{p} : \boldsymbol{q}] \geq \bar{D}\left[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}\right]. \tag{3.11}$$

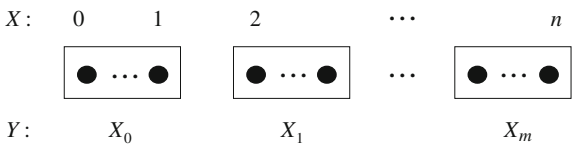


Fig. 3.1 Partition of X into $m+1$ subsets

When does the equality hold in (3.11)? This occurs in the case when there is no loss of information by coarse graining. Since y is a function of x, we have the following decomposition:

$$p(x, \boldsymbol{\xi}) = p(x, y, \boldsymbol{\xi}) = p(y, \boldsymbol{\xi}) p(x|y, \boldsymbol{\xi}), \tag{3.12}$$

where $\boldsymbol{\xi}$ is a coordinate system of S_n. We see that y is a sufficient statistic when $p(x|y, \boldsymbol{\xi})$ does not depend on $\boldsymbol{\xi}$. In this case, the conditional distributions of $p(x|y, \boldsymbol{\xi})$ and $q(x|y, \boldsymbol{\xi}')$ are equal for two distributions $p(x) = p(x, \boldsymbol{\xi})$ and $q(x) = p(x, \boldsymbol{\xi}')$, that is,

$$p\left(x = j \,|y = a, \boldsymbol{\xi}\right) = p\left(x = j \,\middle|y = a, \boldsymbol{\xi}'\right). \tag{3.13}$$

3.2.2 *Invariant Divergence*

When a divergence is written in the form

$$D_f[\boldsymbol{p} : \boldsymbol{q} = \sum p_i f\left(\frac{q_i}{p_i}\right), \tag{3.14}$$

where f is a differentiable convex function satisfying

$$f(1) = 0, \tag{3.15}$$

it is called an f-divergence. The f-divergence was introduced by Morimoto (1963), Ali and Silvey (1966) and Csiszár (1967). It is easy to prove that this satisfies the criteria of divergence, by expanding $D_f[\boldsymbol{p} : \boldsymbol{p} + d\boldsymbol{p}]$ in the Taylor series, although it is not a Bregman divergence in general.

Theorem 3.1 *An f-divergence is invariant and decomposable. Conversely an invariant and decomposable divergence is an f-divergence, except for the case of $n = 1$.*

Proof We first prove that an f-divergence satisfies the criterion of information monotonicity. Consider a simple partition where $X_0 = \{1, 2\}$ and all the other X_a are singleton sets. That is, $x = 1, 2$ are put in a subset X_0 but all the other x remain as they are. We prove only this case, but other cases can be proved similarly. We need to prove

$$p_1 f\left(\frac{q_1}{p_1}\right) + p_2 f\left(\frac{q_2}{p_2}\right) \geq (p_1 + p_2) f\left(\frac{q_1 + q_2}{p_1 + p_2}\right). \tag{3.16}$$

By introducing

$$u_1 = \frac{q_1}{p_1}, \quad u_2 = \frac{q_2}{p_2}, \tag{3.17}$$

the right-hand side of (3.16) is written as

$$(p_1 + p_2)\, f\left(\frac{p_1}{p_1 + p_2} u_1 + \frac{p_2}{p_1 + p_2} u_2\right). \tag{3.18}$$

Since f is convex,

$$(p_1 + p_2)\, f\left(\frac{p_1}{p_1 + p_2} u_1 + \frac{p_2}{p_1 + p_2} u_2\right) \leq p_1 f\,(u_1) + p_2 f\,(u_2)\,, \tag{3.19}$$

which proves the information monotonicity.

Conversely, assume that the information monotonicity holds for a decomposable divergence (3.9). Then, the equality holds when (3.13) is satisfied, that is, $u_1 = u_2$ in the present case. The equality is written as

$$d\,(p_1, q_1) + d\,(p_2, q_2) = d\,(p_1 + p_2, q_1 + q_2)\,. \tag{3.20}$$

By putting

$$k(p, u) = d(p, up), \tag{3.21}$$

we have

$$k\,(p_1, u) + k\,(p_2, u) = k\,(p_1 + p_2, u) \tag{3.22}$$

for $u > 0$, and hence $k(p, u)$ is linear in p. So we have

$$k(p, u) = f(u)p, \tag{3.23}$$

implying

$$d(p, q) = pf\left(\frac{q}{p}\right). \tag{3.24}$$

This proves the theorem. □

Remark 1 The above proof is not valid when $n = 1$, because coarse graining causes $m = 0$. The following is shown by Jiao et al. (2015): There exists a class of invariant divergences which are not necessarily f-divergences when $n = 1$. So the case with $n = 1$ is special and Jiao et al. (2015) derived a general class of invariant divergences when $n = 1$.

Remark 2 When we treat non-decomposable divergences, there are invariant divergences which are not f-divergences. A function of f-divergence is invariant but is not decomposable in general. A simple example is

$$D[\boldsymbol{p} : \boldsymbol{q}] = D_f[\boldsymbol{p} : \boldsymbol{q}] + \left\{D_f[\boldsymbol{p} : \boldsymbol{q}]\right\}^2. \tag{3.25}$$

Further, an adequate nonlinear function of two f-divergences D_{f_1} and D_{f_2} is invariant but is not an f-divergence.

We will show in Part II that any invariant divergence gives the same geometry called the α-structure.

When a linear term is added to a convex function f,

$$\bar{f}(u) = f(u) + c(u-1), \tag{3.26}$$

where c is a constant, $\bar{f}$ is also convex. It is easy to see

$$D_{\bar{f}}[\boldsymbol{p} : \boldsymbol{q}] = D_f\left[\boldsymbol{p} : \boldsymbol{q}\right], \tag{3.27}$$

so (3.26) does not change the divergence. Hence, without loss of generality, we can always use a convex function satisfying

$$f(1) = 0, \quad f'(1) = 0. \tag{3.28}$$

Moreover, since

$$D_{cf}[\boldsymbol{p} : \boldsymbol{q}] = cD_f[\boldsymbol{p} : \boldsymbol{q}] \tag{3.29}$$

holds for another constant $c > 0$, the constant c determines the scale of divergence. To fix the scale, we use f that satisfies

$$f''(1) = 1. \tag{3.30}$$

Definition 3.1 A convex function f satisfying (3.28) and (3.30) is said to be standard. An f-divergence derived from a standard f is a standard f-divergence.

When $D_f\left[\boldsymbol{p} : \boldsymbol{q}\right]$ is a standard f-divergence, its dual $D_f^*[\boldsymbol{p} : \boldsymbol{q}] = D_f[\boldsymbol{q} : \boldsymbol{p}]$ is also a standard f-divergence. To show this, define

$$f^*(u) = uf\left(\frac{1}{u}\right). \tag{3.31}$$

Then, f^* is a standard convex function when f is, and we have

$$D_{f^*}[\boldsymbol{p} : \boldsymbol{q}] = D_f[\boldsymbol{q} : \boldsymbol{p}]. \tag{3.32}$$

3.3 Examples of f-Divergence in S_n

3.3.1 KL-Divergence

For

$$f(u) = -\log u, \tag{3.33}$$

the derived divergence is the KL-divergence

$$D_f[\boldsymbol{p} : \boldsymbol{q}] = \sum p_i \log \frac{p_i}{q_i}. \tag{3.34}$$

The dual of f is

$$f^*(u) = u \log u. \tag{3.35}$$

The derived divergence is the dual of the KL-divergence

$$D_{f^*}[\boldsymbol{p} : \boldsymbol{q}] = D_{KL}[\boldsymbol{q} : \boldsymbol{p}], \tag{3.36}$$

which coincides with the divergence derived from the cumulant generating function ψ.

3.3.2 χ^2-Divergence

For

$$f(u) = \frac{1}{2}(u-1)^2, \quad D_f[\boldsymbol{p} : \boldsymbol{q}] = \frac{1}{2} \sum \frac{(p_i - q_i)^2}{p_i}. \tag{3.37}$$

This is known as the Pearson χ^2-divergence.

3.3.3 α-Divergence

Let α be a real parameter. We define the α-function by

$$f_\alpha(u) = \frac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right), \quad \alpha \neq \pm 1. \tag{3.38}$$

The derived divergence is the α-divergence (Amari 1985; Amari and Nagaoka 2000) given by

$$D_\alpha[\boldsymbol{p}:\boldsymbol{q}] = \frac{4}{1-\alpha^2}\left(1 - \sum p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}}\right), \quad \alpha \neq \pm 1. \tag{3.39}$$

The dual of the α-function is the $-\alpha$-function. Hence, the dual of the α-divergence is the $-\alpha$-divergence,

$$D_\alpha\left[\boldsymbol{p}:\boldsymbol{q}\right] = D_{-\alpha}\left[\boldsymbol{q}:\boldsymbol{p}\right]. \tag{3.40}$$

When $\alpha = 0$, we have

$$f(u) = 4(1-\sqrt{u}), \quad D_f[\boldsymbol{p}:\boldsymbol{q}] = 2\sum\left(\sqrt{p_i} - \sqrt{q_i}\right)^2, \tag{3.41}$$

which is known as the square of the Hellinger distance.

We extend the α-function (3.38) to the case of $\alpha = \pm 1$, by taking the limit $\alpha \to \pm 1$. Then,

$$f_\alpha(u) = \begin{cases} u\log u, & \alpha = 1, \\ -\log u, & \alpha = -1. \end{cases} \tag{3.42}$$

The derived divergences are

$$D_\alpha[\boldsymbol{p}:\boldsymbol{q}] = \begin{cases} \sum q_i \log \frac{q_i}{p_i}, & \alpha = 1, \\ \sum p_i \log \frac{p_i}{q_i}, & \alpha = -1. \end{cases} \tag{3.43}$$

Hence, the KL-divergence is -1-divergence and its dual is 1-divergence.

For

$$f(u) = |1-u| \tag{3.44}$$

which is not differentiable, and hence D_f is not a divergence by our definition, D_f is a symmetric function of $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$D_f[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{2}\sum |p_i - q_i|, \tag{3.45}$$

known as the variational distance.

The square of the Euclidean distance,

$$D[\boldsymbol{p}:\boldsymbol{q}] = \sum (p_i - q_i)^2, \tag{3.46}$$

is a divergence. But it is not an f-divergence and is not invariant.

3.4 General Properties of f-Divergence and KL-Divergence

3.4.1 Properties of f-Divergence

The following properties hold in S_n.

(1) An f-divergence $D_f[\boldsymbol{p} : \boldsymbol{q}]$ is convex with respect to both $\boldsymbol{p}$ and $\boldsymbol{q}$.
(2) It is bounded from above as

$$0 \leq D_f[\boldsymbol{p} : \boldsymbol{q}] \leq \lim_{u \to 0} \left\{ f(u) + uf\left(\frac{1}{u}\right) \right\}, \tag{3.47}$$

$$0 \leq D_f[\boldsymbol{p} : \boldsymbol{q}] \leq \sum (p_i - q_i)\, f'\left(\frac{p_i}{q_i}\right). \tag{3.48}$$

(3) For $\alpha \geq 1$,

$$D_\alpha[\boldsymbol{p} : \boldsymbol{q}] = \infty, \tag{3.49}$$

when $p(x) = 0$ and $q(x) \neq 0$ hold for some x.
(4) For $\alpha \leq -1$,

$$D_\alpha[\boldsymbol{p} : \boldsymbol{q}] = \infty, \tag{3.50}$$

when $p(x) \neq 0$ and $q(x) = 0$ hold for some x.

Properties (3) and (4) hold for the KL-divergence and its dual, because they are ± 1-divergences. They lead to the following results of approximation of a probability distribution by using the α-divergence. Given $\boldsymbol{p} \in S_n$, we search for the distribution $\hat{\boldsymbol{p}}_S$ that minimizes the divergence from $\boldsymbol{p}$ to a smooth submanifold $S \subset S_n$,

$$\hat{\boldsymbol{p}}_S = \arg\min_{\boldsymbol{q} \in S} D_\alpha[\boldsymbol{p} : \boldsymbol{q}]. \tag{3.51}$$

Then, the following holds:

(5) Zero-forcing: When $\alpha \geq 1$, the best approximation $\hat{\boldsymbol{p}}_S$ in the closure of S satisfies

$$\hat{p}_S(x) = 0 \tag{3.52}$$

for x at which $p(x) = 0$.
(6) Zero-avoidance: When $\alpha \leq -1$, the best approximation $\hat{\boldsymbol{p}}_S$ in the closure of S satisfies

$$\hat{p}_S(x) \neq 0 \tag{3.53}$$

for x at which $p(x) \neq 0$.

3.4.2 Properties of KL-Divergence

A. Large deviation

Let $\boldsymbol{p}$ be a distribution in S_n from which N independent data $x(1), \ldots, x(N)$ are generated. The empirical distribution of the observed data is given by $\hat{\boldsymbol{p}}$,

$$\hat{p}_i = \frac{1}{N} \sum_{t=1}^{N} \delta_i \{x(t)\} = \frac{N_i}{N}, \tag{3.54}$$

where N_i is the number that $x = i$ is observed among N data. This is the maximum likelihood estimator. How far is $\hat{\boldsymbol{p}}$ from the true $\boldsymbol{p}$? The probability distribution of $\hat{\boldsymbol{p}}$ is evaluated by the KL-divergence asymptotically when N is large.

Sanov Lemma. The probability of $\hat{\boldsymbol{p}}$ is asymptotically given by

$$\text{Prob}\left\{\hat{\boldsymbol{p}}; \boldsymbol{p}\right\} = \exp\left\{-N\ D_{KL}\left[\hat{\boldsymbol{p}} : \boldsymbol{p}\right]\right\}, \tag{3.55}$$

that is, the probability decays exponentially as N increases where the exponent of decay is $D_{KL}\left[\hat{\boldsymbol{p}} : \boldsymbol{p}\right]$.

The proof is given by evaluating the distribution of $\hat{\boldsymbol{p}}$, a multinomial distribution, when N is large, which we omit. When $\hat{\boldsymbol{p}}$ is close to $\boldsymbol{p}$, by putting

$$\varepsilon = \frac{1}{\sqrt{N}} \left(\hat{\boldsymbol{p}} - \boldsymbol{p}\right) \tag{3.56}$$

and expanding $N\ D_{KL}\left[\hat{\boldsymbol{p}} : \boldsymbol{p}\right]$, we have the central limit theorem.

Central Limit Theorem The distribution of $\hat{\boldsymbol{p}}$ is asymptotically Gaussian with mean $\boldsymbol{p}$ and covariance

$$E\left[\left(\hat{p}_i - p_i\right)\left(\hat{p}_j - p_j\right)\right] = \frac{1}{N} g_{ij}. \tag{3.57}$$

Let A be a region in S_n. Then, we have the theorem of large deviation, which is useful in information theory and statistics (Fig. 3.2).

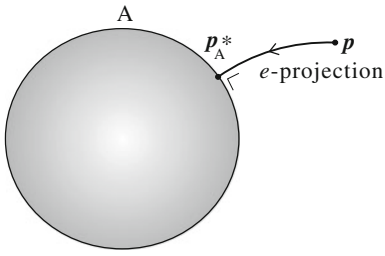


Fig. 3.2 e-projection of $\boldsymbol{p}$ to A

Large Deviation Theorem The probability that $\hat{\boldsymbol{p}}$ is included in A is given asymptotically by

$$\text{Prob}\left\{\hat{\boldsymbol{p}} \in A\right\} = \exp\left\{-N\ D_{KL}\left[\boldsymbol{p}_A^* : \boldsymbol{p}\right]\right\}, \tag{3.58}$$

where

$$\boldsymbol{p}_A^* = \arg\min_{\boldsymbol{q}\in A} D_{KL}\left[\boldsymbol{q} : \boldsymbol{p}\right]. \tag{3.59}$$

When A is a closed set having a boundary, $\boldsymbol{p}_A^*$ is given by e-projecting $\boldsymbol{p}$ to the boundary of A.

B. Symmetrized KL-divergence and Fisher information

The Riemannian distance between two points $\boldsymbol{p}$ and $\boldsymbol{q}$ is given by the minimum of the distance along all curves $\boldsymbol{\xi}(t)$ connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ such that $\boldsymbol{\xi}(0) = \boldsymbol{p}$, $\boldsymbol{\xi}(1) = \boldsymbol{q}$, that is,

$$s = \min \int_0^1 \sqrt{g_{ij}(t)\dot{\xi}^i\dot{\xi}^j}dt. \tag{3.60}$$

Since the KL-divergence is

$$D_{KL}\left[\boldsymbol{\xi}(t) : \boldsymbol{\xi}(t+dt)\right] = \frac{1}{2}g_{ij}\dot{\xi}^i\dot{\xi}^j dt^2, \tag{3.61}$$

there would be some relation between the KL-divergence and the integration of the Fisher information along a curve. Let us consider the e-geodesic and the m-geodesic connecting two points $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$\gamma_e : \boldsymbol{\xi}_e(t) = \exp\left\{(1-t)\log \boldsymbol{p} + t\log \boldsymbol{q} - \psi(t)\right\}, \tag{3.62}$$

$$\gamma_m : \boldsymbol{\xi}_m(t) = (1-t)\boldsymbol{p} + t\boldsymbol{q}. \tag{3.63}$$

They are exponential and mixture families, respectively. Let $g_e(t)$ and $g_m(t)$ be the Fisher information along the curves,

$$g_e(t) = g_{ij}\dot{\xi}_e^i(t)\dot{\xi}_e^j(t), \tag{3.64}$$

$$g_m(t) = g_{ij}\dot{\xi}_m^i(t)\dot{\xi}_m^j(t). \tag{3.65}$$

Then, we have the following theorem.

Theorem 3.2 *The symmetrized KL-divergence is given by the integration of the Fisher information along the e-geodesic and the m-geodesic,*

$$\frac{1}{2}\{D_{KL}[\boldsymbol{p}:\boldsymbol{q}]+D_{KL}[\boldsymbol{q}:\boldsymbol{p}]\}$$
$$=\int_0^1 g_e(t)dt=\int_0^1 g_m(t)dt. \tag{3.66}$$

The proof is technical and is omitted.

3.5 Fisher Information: The Unique Invariant Metric

Since an f-divergence is invariant, the Riemannian metric derived from it is invariant. We can easily calculate the metric g_{ij} from an f-divergence by the Taylor expansion,

$$D_f\left[p(x,\boldsymbol{\xi}):p(x:\boldsymbol{\xi}+d\boldsymbol{\xi})\right]=\int p(x,\boldsymbol{\xi})f\left\{\frac{p(x,\boldsymbol{\xi}+d\boldsymbol{\xi})}{p(x,\boldsymbol{\xi})}\right\}dx=\frac{1}{2}g_{ij}(\boldsymbol{\xi})d\xi^i d\xi^j. \tag{3.67}$$

A simple calculation gives the following lemma.

Lemma *Any standard f-divergence gives the same Riemannian metric which is the Fisher information matrix*

$$g_{ij}=E\left[\partial_i \log p(\boldsymbol{x},\boldsymbol{\xi})\partial_j \log p(\boldsymbol{x},\boldsymbol{\xi})\right], \tag{3.68}$$

where

$$\partial_i=\frac{\partial}{\partial\xi_i}. \tag{3.69}$$

Chentsov (1972) proved a stronger theorem that the Fisher information matrix is the unique invariant metric of S_n. He used the framework of category theory. We show a simpler proof due to Campbell (1986).

Consider a series of S_n, $n=1,2,3,\ldots$, and reformulate the invariance criterion. We consider coarse graining of S_n by a partition of $X=\{0,1,2,\ldots,n\}$ to $Y=\{A_0,A_1,\ldots,A_m\}$, where $n\geq m$. Random variable x taking on values $0,1,\ldots,n$ is reduced to random variable y taking on values $0,1,\ldots,m$, such that $y=i$ when x is included in A_i. Obviously, probability distribution $\boldsymbol{p}\in S_n$ is mapped to $\boldsymbol{q}\in S_m$ by this coarse graining. It defines a mapping f from S_n to S_m

$$f:\boldsymbol{p}\mapsto\boldsymbol{q}\ ;\ q_i=\sum_{j\in A_i}p_j. \tag{3.70}$$

Conversely, we consider a mapping h from S_m to S_n, which is determined by an arbitrary conditional probability distribution,

$$r_{ij}=\text{Prob}\{x=i\mid y=j\}. \tag{3.71}$$

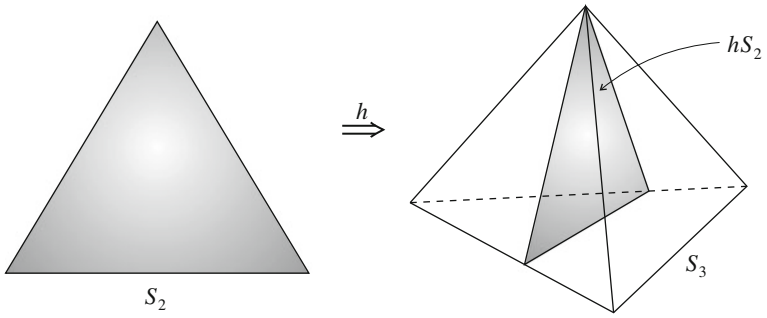


Fig. 3.3 Embedding of S_2 in S_3 (m = 2, n = 3)

Given $y = j$, it generates $x = i$ stochastically based on r_{ij}. We define a mapping by

$$h : \boldsymbol{q} \mapsto \boldsymbol{p} \; ; \; p_i = r_{ij} q_j . \tag{3.72}$$

Given y of which the probability is $\boldsymbol{q}$, the probability distribution $\boldsymbol{p} = h\boldsymbol{q}$ of x is given by (3.72). The mapping h which depends on r_{ij} embeds S_m in S_n and it satisfies

$$f \circ h = \mathrm{Id}, \tag{3.73}$$

where Id is the identity mapping (see Fig. 3.3).

Consider a problem of estimation of $\boldsymbol{q} \in S_m$ by observing random variable y. When S_m is embedded in a larger manifold S_n by (3.72), the random variable is x. However, x includes a redundant part for estimating $\boldsymbol{q}$. y is a sufficient statistic for estimating $\boldsymbol{q}$.

The invariance criterion claims that the geometry of S_m is the same as the geometry of embedded hS_m in the larger manifold S_n. In particular, the inner product of two basis vectors in S_m should be the same as that in the embedded image. Now we state the theorem of Chentsov.

Theorem 3.3 *The invariant metric is unique, given by the Fisher information to within a constant factor.*

Proof We use $\boldsymbol{R}^n_+$ to prove the theorem, considering S_{n-1} as its subspace constrained by $\sum p_i = 1$. When $m = n$, the mapping f is only a permutation of indices. We consider the center of S_{n-1},

$$\bar{\boldsymbol{p}} = \left(\frac{1}{n}, \ldots, \frac{1}{n} \right) \in \boldsymbol{R}^n_+ . \tag{3.74}$$

It is invariant under the permutation group of index i. So the inner product of two basis vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ in $\boldsymbol{R}_n^+$ is invariant under the permutation of indices. Hence, we put

$$g_{ij}^n(\bar{\boldsymbol{p}}) = B(n), \text{ for any } i, j\ ;\ i \neq j, \tag{3.75}$$

$$g_{ii}^n(\bar{\boldsymbol{p}}) = A(n) + B(n), \text{ for any } i, \tag{3.76}$$

or

$$g_{ij}^n(\bar{\boldsymbol{p}}) = A(n)\delta_{ij} + B(n). \tag{3.77}$$

When $\boldsymbol{p}$ is in S_{n-1}, its small deviation $\delta\boldsymbol{p}$ inside S_{n-1} satisfies

$$\sum_{i=1}^{n} \delta p_i = 0. \tag{3.78}$$

Since $\delta\boldsymbol{p}$ is a tangent vector of S_{n-1},

$$\sum Z^i = 0 \tag{3.79}$$

holds for any tangent vector $Z = Z^i\boldsymbol{e}_i$ of S_{n-1}.

Therefore, we may put $B(n) = 0$ when calculating the inner product of two tangent vectors of S_{n-1}. $B(n)$ is responsible only for the normal direction to S_{n-1}. So, we put

$$g_{ij}^n(\bar{\boldsymbol{p}}) = A(n)\delta_{ij}. \tag{3.80}$$

Let us consider a point

$$\boldsymbol{q} = \left(\frac{k_1}{n}, \frac{k_2}{n}, \ldots, \frac{k_m}{n}\right) \in S_{m-1}, \tag{3.81}$$

where k_i are integers, satisfying $\sum k_i = n$. We then consider the following embedding of S_{m-1} in S_{n-1} given by the conditional distributions

$$r_{ij} = \begin{cases} \frac{1}{k_j}, & i \in A_j, \\ 0, & \text{otherwise}, \end{cases} \tag{3.82}$$

where $\{A_j\}$ is a partition of $\{0, 1, \ldots, n\}$ such that A_j includes k_j elements. Then, $\boldsymbol{q}$ is mapped to the center of S_{n-1},

$$h\boldsymbol{q} = \bar{\boldsymbol{p}} = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right). \tag{3.83}$$

The basis vector $\boldsymbol{e}_1^m \in \boldsymbol{R}_+^m$ is mapped to

$$\tilde{\boldsymbol{e}}_1^n = \frac{1}{k_1}\left(\boldsymbol{e}_1^n + \cdots \boldsymbol{e}_{k_1}^n\right) \tag{3.84}$$

in $\boldsymbol{R}_+^n$ by this embedding. Similar equations hold for other $\boldsymbol{e}_i^n$, $i = 2, 3, \ldots, m$. The inner product is equal to

$$g_{11}^m(\boldsymbol{q}) = \langle \boldsymbol{e}_1^m, \boldsymbol{e}_1^m \rangle = \langle \tilde{\boldsymbol{e}}_1^n, \tilde{\boldsymbol{e}}_1^n \rangle = \left\langle \frac{1}{k_1}\sum_{i=1}^{k_1} \boldsymbol{e}_i^n, \frac{1}{k_1}\sum_{i=1}^{k_1} \boldsymbol{e}_i^n \right\rangle = \frac{1}{k_1} g_{11}^n\left(\bar{\boldsymbol{p}}\right) = \frac{A(n)}{k_1}. \tag{3.85}$$

Hence, we have

$$g_{11}^m(\boldsymbol{q}) = \frac{nc}{k_1} = \frac{c}{q_1}. \tag{3.86}$$

Since the constant c is used only to determine the scale of the Fisher information, we may put $c = 1$. Similarly,

$$g_{ii}^m(\boldsymbol{q}) = \frac{n}{k_i} = \frac{1}{q_i}. \tag{3.87}$$

This holds only at the points where q_i are rational numbers, but because of the continuity, it holds for any $\boldsymbol{q}$. This proves the theorem. □

Remark We can prove the uniqueness of the cubic tensor T_{ijk} defined by

$$T_{ijk} = \mathrm{E}\left[\partial_i \log p(x, \boldsymbol{\xi}) \partial_j \log p(x, \boldsymbol{\xi}) \partial_k \log p(x, \boldsymbol{\xi})\right] \tag{3.88}$$

under the invariance criterion in a similar way. This will be used to study the uniqueness of the α-connection in Part II.

3.6 f-Divergence in Manifold of Positive Measures

We extend the notion of invariance from S_n to $\boldsymbol{R}_+^n$ by using the information monotonicity under coarse graining. We can prove that the only invariant decomposable divergence is an f-divergence, since the proof of Theorem 4.1 is also valid for $\boldsymbol{R}_+^n$. An f-divergence is

$$D_f[\boldsymbol{m} : \boldsymbol{n}] = \sum m_i f\left(\frac{n_i}{m_i}\right), \tag{3.89}$$

$\boldsymbol{m}, \boldsymbol{n} \in \boldsymbol{R}_+^n$, for a manifold of positive measures $\boldsymbol{R}_+^n$, where f is a standard convex function satisfying (3.28) and (3.30). We need to use a standard convex function to define a divergence in $\boldsymbol{R}_+^n$, because (3.89) does not satisfy the criteria of divergence for a general convex f. The criteria are satisfied when a standard convex function f is used.

We can calculate the invariant Riemannian metric induced in $\boldsymbol{R}_+^n$ by an f-divergence.

Theorem 3.4 *The Riemannian metric in $\boldsymbol{R}_+^n$ induced by an invariant divergence is the Euclidean metric*

$$g_{ij}(\boldsymbol{m}) = \frac{1}{m_i}\delta_{ij}. \tag{3.90}$$

Proof It is easy to derive (3.90) by the Taylor expansion of (3.89)

$$\begin{aligned} D_f[\boldsymbol{m} : \boldsymbol{m} + d\boldsymbol{m}] &= \sum m_i f\left(1 + \frac{dm_i}{m_i}\right) \\ &= \sum \frac{f''(1)}{2m_i} dm_i^2, \end{aligned} \tag{3.91}$$

where $f''(1) = 1$. By using a new coordinate system given by

$$\xi^i = 2\sqrt{m_i}, \tag{3.92}$$

the square of an infinitesimal distance is given as

$$ds^2 = \sum \left(d\xi^i\right)^2, \tag{3.93}$$

showing that the manifold is Euclidean and the coordinate system is orthonormal. □

It should be noted that manifold S_n is a submanifold of $\boldsymbol{R}_{n+1}^+$. The constraint $\sum p_i = 1$ becomes

$$\sum \left(\xi^i\right)^2 = 4 \tag{3.94}$$

in the new coordinate system. Hence, S_n is a sphere in a Euclidean space, so it is curved.

As an important special case of f-divergence, we introduce the α-divergence, which is previously defined in S_n, to $\boldsymbol{R}_+^n$. It is defined by using the standard α-function,

$$f_\alpha(u) = \begin{cases} \frac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right) - \frac{2}{1-\alpha}(u-1), & \alpha \neq \pm 1, \\ u\log u - (u-1), & \alpha = 1, \\ -\log u + (u-1), & \alpha = -1. \end{cases} \tag{3.95}$$

Definition The α-divergence is defined in $\boldsymbol{R}_+^n$ by

$$D_\alpha[\boldsymbol{m}:\boldsymbol{n}] = \begin{cases} \frac{4}{1-\alpha^2}\sum\left\{\frac{1-\alpha}{2}m_i + \frac{1+\alpha}{2}n_i - m_i^{\frac{1-\alpha}{2}} n_i^{\frac{1+\alpha}{2}}\right\}, & \alpha \neq \pm 1, \\ \sum\left\{m_i - n_i + n_i \log \frac{n_i}{m_i}\right\}, & \alpha = 1, \\ \sum\left\{n_i - m_i + m_i \log \frac{m_i}{n_i}\right\}, & \alpha = -1. \end{cases} \tag{3.96}$$

When both $\boldsymbol{m}$ and $\boldsymbol{n}$ satisfy the normalization condition,

$$\sum m_i = \sum n_i = 1, \tag{3.97}$$

they are probability distributions and the α-divergence is equal to that in a manifold of probability distributions.

Remarks

There is a long history of studies on geometry of manifolds of probability distributions. C.R. Rao is believed to have been the first who introduced a Riemannian metric by using the Fisher information matrix (Rao 1945). This was work he did at the age of twenty-four, and the famous Crámer–Rao theorem was presented in the same seminal paper. It is a monumental work from which Information Geometry has emerged. Jeffreys (1946) used the square root of the determinant of the Fisher metric, which is the Riemannian volume element, as an invariant prior distribution over the manifold in Bayesian statistics. However, there was no such concept in the first edition of his famous book, "Probability Theory", published in 1939 (Jeffreys 1939). It appeared in the second edition (Jeffreys 1948; see also Jeffreys 1946).

It was a big surprise that a hidden prehistory was uncovered by Stigler (2007) (Frank Nielsen kindly let me know of this paper). In 1929, Harold Hotelling spent nearly half a year at Rothamsted working with R.A. Fisher on establishing a foundation for mathematical statistics. He submitted a paper entitled "Spaces of statistical parameters" to the American Mathematical Society Meeting in 1929 (which, in his absence, was read by O. Ore). The paper has never been published, so his idea has become entombed and remains unknown. He stated in the paper that the Riemannian metric is given by the Fisher information matrix in a statistical manifold. Moreover, he remarked that the manifold of a location-scale statistical model has a constant negative curvature. Incidentally, I discovered this fact in 1958 when I was a master's student, and this was the origin of my study of information geometry.

After Rao, there appeared a number of works using the Riemannian structure, e.g., James (1973). It was Chentsov (1972) who introduced the invariance criterion for defining the geometry of a statistical manifold. He proved that the Fisher information matrix is the only invariant metric in S_n. Moreover, he obtained the class of invariant affine connections (α-connections studied in Part II). Unfortunately, his work was published only in Russian, so his contributions did not become popular in the western world until an English translation appeared in 1982. Later, Efron (1975) investigated old unpublished calculations by R.A. Fisher and elucidated the results by defining the

statistical curvature of a statistical model. He showed that the higher-order efficiency of statistical estimation is given by the statistical curvature which is the e-curvature defined in Part II. This work was commented on by A.P. Dawid in discussions of Efron's paper, where the e- and m-connections were suggested.

Following Efron's and Dawid's works, Amari (1982) further developed the differential geometry of statistical models and elucidated its dualistic nature. It was applied to statistical inference to establish a higher-order statistical theory (Amari 1982, 1985; Kumon and Amari 1983). The formal theory of a dually flat manifold was first proposed by Nagaoka and Amari (1982), which included the Pythagorean theorem and projection theorem. However, it was not published as a journal paper, because it was rejected by major journals. The editor of the *Annals of Probability* asked me to withdraw the paper, because he had approached seven reviewers but none reviewed it seriously. So he concluded that most probabilists would not have any interest in the direction of this research. A reviewer for the *Zeitschrift für Wahrsceinlichkeitstheorie und Verwandte Gebiete* (*Theory of Probability and its Applications*) sent me a letter stating that the paper was useless, because no essential relation exists between statistics and differential geometry. He also pointed out that the differential geometry of this paper is different from that in textbooks so it would be dubious. (We proposed a new framework of duality in differential geometry.) So it was rejected. Several years passed and thirdly a reviewer in *IEEE Transactions on Information Theory* wrote that the theory was now well known around the world and the paper submitted included few new ideas. This was because a workshop on this subject was organized in London in 1984 by Sir D. Cox, and my "Springer Lecture Notes" (Amari 1985) were published. Since then, information geometry has become widely known and a number of competent researchers have joined from the fields of statistics, vision, optimization, machine learning, etc. Many international conferences have been organized on this subject.

However, a mathematically rigorous foundation involves difficulty in the case of the function space of probability density functions. This is because the topology of the space of $p(x)$ is different from that of the space of $\log p(x)$. There is a series of studies given by Pistone and his coworkers (Pistone and Sempi 1995; Pistone and Rogatin 1999; Cena and Pistone 2007; Pistone 2013). See also Grasselli (2010). Newton (2012) gave a theory based on a Hilbert space, in the framework that $p(x)$ has finite entropy. Here, $p(x)$, a probability density function with respect to measure $\mu(x)$, is mapped onto a Hilbert space by using the following representation of $p(x)$:

$$\Phi[p] = p(x) + \log p(x), \tag{3.98}$$

where

$$E_\mu\left[\{p(x)\}^2\right] < \infty, \quad E_\mu\left[\log^2 p(x)\right] < \infty \tag{3.99}$$

are presumed. J. Jost and his coworkers are developing a rigorous theory in Leipzig, preparing a monograph. See Ay et al. (2013).

Information geometry uses the e- and m-geodesic connecting two distributions $p(x)$ and $q(x)$, KL-divergence $D_{KL}[p(x) : q(x)]$, the Pythagorean and projection theorems and the orthogonality of two curves. Therefore, we want to have a framework in which the above structures are guaranteed. We need to search for mild regularity conditions which give such a framework (cf. Pistone 2013; Newton 2012). Fukumizu (2009) proposed a novel idea of the kernel exponential family for treating statistical manifolds with function degrees of freedom.

Chapter 4
α-Geometry, Tsallis q-Entropy and Positive-Definite Matrices

An f-divergence is not necessarily of the Bregman type. Hence, the invariant geometry induced from an f-divergence does not necessarily give a dually flat structure. It is proved that the KL-divergence, which is an f-divergence, is the unique class of decomposable divergences that are invariant and flat in $S_n (n > 1)$. However, when we study a manifold $\boldsymbol{R}_+^n$ of positive measures, there are other invariant, flat and decomposable divergences. They are α-divergences, including the KL-divergence as a special case. The present chapter studies the invariant α-structure originating from the α-divergence. It includes the α-geodesic, α-mean, α-projection, α-optimization and α-family of probability distributions.

We also remark that the geometry originating from Tsallis q-entropy (Tsallis 1988, 2009; Naudts 2011) is nothing other than the α-geometry, where $\alpha = 2q - 1$. We show another type of flat structure, called conformal flattening, induced from the Tsallis q-entropy. It is related to the escort probability distribution. Extending it, we identify a universal class of dually flat divergences in $\boldsymbol{R}_+^n$. We further study a general invariant flat structure of the manifold of positive-definite matrices, which is important in its own right.

4.1 Invariant and Flat Divergence

4.1.1 KL-Divergence Is Unique

A divergence is flat when it induces a flat structure in the underlying manifold. A Bregman divergence is flat. We begin with the following well-known result in S_n. See Csiszár (1991) for the characterization of the KL-divergence.

Theorem 4.1 *The KL-divergence and its dual are the only decomposable, flat and invariant divergences, except for the special case of* $n = 1$.

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_4

A proof of the present theorem is given as a corollary of Theorem 4.2 in the next subsection. It will be shown in Part II without assuming the decomposability that the KL-divergence is the unique canonical divergence in a dually flat manifold of probability distributions.

4.1.2 α-Divergence Is Unique in $\boldsymbol{R}^n_+$

We begin with a theorem due to Amari (2009).

Theorem 4.2 *The α-divergences form the unique class of decomposable, flat and invariant divergences of $\boldsymbol{R}^n_+$.*

Proof We first prove that an α-divergence is a Bregman divergence in the manifold $\boldsymbol{R}^n_+$. This does not imply that its affine coordinate system is the measure vector $\boldsymbol{m} = (m_i) \in \boldsymbol{R}^n_+$ itself. We define a new coordinate system $\boldsymbol{\theta} = \left(\theta^i\right)$ by

$$\theta^i = h_\alpha\left(m_i\right) = m_i^{\frac{1-\alpha}{2}}, \quad \alpha \neq 1 \tag{4.1}$$

and call θ^i the α-representation of a positive measure m_i. Then,

$$m_i = h_\alpha^{-1}(\theta^i) = \left(\theta^i\right)^{\frac{2}{1-\alpha}} \tag{4.2}$$

is a convex function of θ^i when $|\alpha| < 1$. Therefore,

$$\psi_\alpha(\boldsymbol{\theta}) = \frac{1-\alpha}{2} \sum \left(\theta^i\right)^{\frac{2}{1-\alpha}} = \frac{1-\alpha}{2} \sum m_i \tag{4.3}$$

is a convex function of $\boldsymbol{\theta}$ for $\alpha > -1$ and the accompanying affine coordinate system is $\boldsymbol{\theta}$. The dual affine coordinate system $\boldsymbol{\eta}$ is given by $\boldsymbol{\eta} = \nabla\psi_\alpha(\boldsymbol{\theta})$ as

$$\eta_i = \left(\theta^i\right)^{\frac{1+\alpha}{1-\alpha}} = h_{-\alpha}\left(m_i\right). \tag{4.4}$$

Hence, it is the $-\alpha$-representation of m_i. The dual convex function is

$$\varphi_\alpha\left(\boldsymbol{\eta}\right) = \psi_{-\alpha}\left(\boldsymbol{\eta}\right). \tag{4.5}$$

Calculations show that the Bregman divergence

$$D_\alpha\left[\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2\right] = \psi_\alpha\left(\boldsymbol{\theta}_1\right) + \psi_{-\alpha}\left(\boldsymbol{\eta}_2\right) - \boldsymbol{\theta}_1 \cdot \boldsymbol{\eta}_2 \tag{4.6}$$

is the α-divergence defined in (3.96).

Conversely, assume that an f-divergence

$$D_f[\boldsymbol{m} : \boldsymbol{n}] = \sum m_i f\left(\frac{n_i}{m_i}\right) \tag{4.7}$$

is a Bregman divergence and further that its affine coordinate system $\boldsymbol{\theta} = \left(\theta^i\right)$ is connected with m_i componentwise as

$$\theta^i = k\left(m_i\right). \tag{4.8}$$

The dual affine coordinates are

$$\eta_i = k^*\left(m_i\right) \tag{4.9}$$

for some function k^*. Since the cross term of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ in the divergence is included only in the last term of (4.6), the relation

$$m_i f\left(\frac{n_i}{m_i}\right) = k(m_i)k^*(n_i) \tag{4.10}$$

must hold for each i. By differentiating it with respect to n_i and omitting suffix i for brevity, we have

$$f'\left(\frac{n}{m}\right) = k(m)k^{*\prime}(n). \tag{4.11}$$

By putting $x = n$ and $y = 1/m$, we have

$$f'(xy) = k\left(\frac{1}{y}\right)k^*(x). \tag{4.12}$$

Further, by putting

$$h(u) = \log f'(u), \tag{4.13}$$

the logarithm of (4.12) is written in the form

$$h(xy) = s(x) + t(y). \tag{4.14}$$

for some functions s and t. By differentiating both sides with respect to x, we have

$$h(u) = -c \log u, \tag{4.15}$$

where c is a constant. From this, we see that f is of the form

$$f(u) = -u^{\frac{1+\alpha}{2}} \tag{4.16}$$

except for a scale factor and a constant. This is a convex function for $|\alpha| < 1$ but is not a standard f-function. By transforming it to the standard form, we have

$$f(u) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2} + \frac{1+\alpha}{2}u - u^{\frac{1+\alpha}{2}}\right), \tag{4.17}$$

and the theorem is proved. □

We did not mention the case of $\alpha = 1$. If we modify the definition (4.1) of the α-representation as

$$h_\alpha(m) = \frac{2}{1-\alpha}\left(m^{\frac{1-\alpha}{2}} - 1\right), \tag{4.18}$$

$\log m$ is given by the limit $\alpha \to 1$. By using this, the proof holds even in the limiting case of $\alpha = \pm 1$.

S_n is a submanifold of $\boldsymbol{R}_+^{n+1}$, where the constraint

$$\sum_{i=0}^{n} m_i = 1 \tag{4.19}$$

is imposed. The constraint is rewritten in the $\boldsymbol{\theta}$-coordinate system as

$$\sum_{i=0}^{m} h_\alpha^{-1}\left(\theta^i\right) = 1. \tag{4.20}$$

This is a nonlinear constraint for $\alpha \neq -1$. So S_n is not dually flat but curved for general α, except for the linear constraint case of $\alpha = -1$. When $\alpha = 1$, it is linear in the dual coordinate system. Hence, the α-divergence gives a flat structure to S_n only when $\alpha = \pm 1$, that is the KL-divergence and its dual. Therefore, the KL-divergence is the only invariant, flat and decomposable divergence in S_n, proving Theorem 4.1.

Remark Jiao et al. (2015) proved that the KL-divergence is the only invariant divergence of the Bregman type in S_n without assuming the decomposability. It is also proved in the geometrical framework that the canonical divergence of S_n is the KL-divergence in Part II. The case of $n = 1$ is fully studied in Jiao et al. (2015), characterizing the class of invariant Bregman-type divergences in S_n. The following is proved:

(1) An invariant decomposable divergence is an f-divergence when $n > 1$, but there is a new class of divergences which are not necessarily f-divergences when $n = 1$.
(2) An invariant Bregman divergence is the KL-divergence for any n.

From the point of view of geometry, a one-dimensional manifold S_1 is a curve so its curvature always vanishes. The case with $n = 1$ is special in this sense.

4.2 α-Geometry in S_n and $\boldsymbol{R_+^n}$

4.2.1 α-Geodesic and α-Pythagorean Theorem in $\boldsymbol{R_+^n}$

The affine and dual affine coordinates of $\boldsymbol{R}_+^n$ due to the α-divergence are given by (4.1) and (4.4), respectively. An α-geodesic passing through $\boldsymbol{\theta}_0$ is linear in the α-representation $\boldsymbol{\theta}$ of (4.1), written as

$$\boldsymbol{\theta}(t) = t\boldsymbol{a} + \boldsymbol{\theta}_0, \tag{4.21}$$

where t is the parameter of the geodesic and $\boldsymbol{a}$ is a constant vector, representing the tangent direction of the geodesic. In particular, the α-geodesic connecting two measures $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ is

$$m_i(t)^{\frac{1-\alpha}{2}} = \left\{(1-t)m_{1i}^{\frac{1-\alpha}{2}} + tm_{2i}^{\frac{1-\alpha}{2}}\right\}. \tag{4.22}$$

Dually, a $-\alpha$-geodesic is linear in the $-\alpha$-representation $\boldsymbol{\eta}$ of (4.4),

$$\boldsymbol{\eta}(t) = t\boldsymbol{a} + \boldsymbol{\eta}_0. \tag{4.23}$$

The $-\alpha$-geodesic connecting $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ is

$$m_i(t)^{\frac{1+\alpha}{2}} = \left\{(1-t)m_{1i}^{\frac{1+\alpha}{2}} + tm_{2i}^{\frac{1+\alpha}{2}}\right\}. \tag{4.24}$$

We have the α-version of the Pythagorean theorem and projection theorem.

Theorem 4.3 *Given three positive measures $\boldsymbol{m}, \boldsymbol{n}, \boldsymbol{k}$, when the α-geodesic connecting $\boldsymbol{m}$ and $\boldsymbol{n}$ is orthogonal to the $-\alpha$-geodesic connecting $\boldsymbol{n}$ and $\boldsymbol{k}$,*

$$D_\alpha[\boldsymbol{m} : \boldsymbol{k}] = D_\alpha[\boldsymbol{m} : \boldsymbol{n}] + D_\alpha[\boldsymbol{n} : \boldsymbol{k}]. \tag{4.25}$$

Theorem 4.4 *Given $\boldsymbol{m}$ and a submanifold S in $\boldsymbol{R}_+^n$, the point $\hat{\boldsymbol{k}}$ in S that minimizes the α-divergence*

$$\hat{\boldsymbol{k}} = \arg\min_{\boldsymbol{k}} D_\alpha[\boldsymbol{m} : \boldsymbol{k}], \quad \boldsymbol{k} \in S, \tag{4.26}$$

is the α-projection of $\boldsymbol{m}$ to S. When S is an $-\alpha$-flat submanifold, the projection is unique.

Remark When $\alpha = -1$, $D_\alpha[\boldsymbol{m} : \boldsymbol{n}]$ is the KL-divergence and the theorems are the Pythagorean and projection theorems given in Chap. 1.

4.2.2 α-Geodesic in S_n

Although the α-divergence is a Bregman divergence in $\boldsymbol{R}_+^{n+1}$, it is not a flat divergence in S_n for $\alpha \neq \pm 1$. The α-geodesic connecting two probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ in $\boldsymbol{R}_+^{n+1}$, given by (4.22) with $\boldsymbol{m}_1 = \boldsymbol{p}$ and $\boldsymbol{m}_2 = \boldsymbol{q}$, is not included in S_n. However, we can normalize (4.22) to obtain the probability vector $\boldsymbol{p}(t)$,

$$p_i^{\frac{1-\alpha}{2}}(t) = c(t)\left\{(1-t)p_i^{\frac{1-\alpha}{2}} + tq_i^{\frac{1-\alpha}{2}}\right\}, \tag{4.27}$$

where $c(t)$ is determined from

$$\sum_{i=0}^{n} p_i(t) = 1. \tag{4.28}$$

This is included in S_n. We call it the α-geodesic of S_n. We can define the α-projection in S_n by using the α-geodesic.

4.2.3 α-Pythagorean Theorem and α-Projection Theorem in S_n

Since $\boldsymbol{R}_+^{n+1}$ is α-flat, its submanifold S_n enjoys an extended version of the Pythagorean theorem. The following theorem is due to Kurose (1994) and it holds for a general dual manifold having a constant curvature.

Theorem 4.5 *Let $\boldsymbol{p}, \boldsymbol{q}$ and $\boldsymbol{r}$ be three points in S_n. When the α-geodesic connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ is orthogonal to the $-\alpha$-geodesic connecting $\boldsymbol{q}$ and $\boldsymbol{r}$,*

$$D_\alpha[\boldsymbol{p}:\boldsymbol{r}] = D_\alpha[\boldsymbol{p}:\boldsymbol{q}] + D_\alpha[\boldsymbol{q}:\boldsymbol{r}] - \frac{1-\alpha^2}{4} D_\alpha[\boldsymbol{p}:\boldsymbol{q}] D_\alpha[\boldsymbol{q}:\boldsymbol{r}]. \tag{4.29}$$

We omit the proof. This is a generalization of a theorem in the spherical geometry, which has a constant curvature.

The projection theorem follows from it.

Theorem 4.6 *Let M be a submanifold of S_n. Given $\boldsymbol{p}$, the point in M that minimizes the α-divergence from $\boldsymbol{p}$ to M is given by the α-geodesic projection of $\boldsymbol{p}$ to M.*

We can easily see from (4.29) that the α-projection gives the critical point of the α-divergence. See Matsuyama (2003) for the minimization of α-divergence and α-projection in ICA (independent component analysis).

4.2.4 Apportionment Due to α-Divergence

We show an interesting application of α-divergence in social science. There are many methods of deciding the numbers of seats proportionately to the populations in states, since the number of seats in a state must be an integer, whereas the ratios of populations are rational numbers. Let $\boldsymbol{p} = (p_i)$ be the population quotient vector

$$p_i = \frac{N_i}{N}, \tag{4.30}$$

where N_i is the populations of state i and $N = \sum N_i$. Let $\boldsymbol{q} = (q_i)$ be the apportionment quotient vector and n be the total number of seats such that nq_i is the number of seats assigned to state i.

We cannot simply put $\boldsymbol{q} = \boldsymbol{p}$, because nq_i should be an integer. Hence, we search for a $\boldsymbol{q}$ that is a rational vector of the form $q_i = n_i/n$ closest to $\boldsymbol{p}$. We can use the α-divergence $D_\alpha[\boldsymbol{p} : \boldsymbol{q}]$ to show the closeness of $\boldsymbol{p}$ and $\boldsymbol{q}$ and search for a rational vector $\boldsymbol{q}$ that minimizes $D_\alpha[\boldsymbol{p} : \boldsymbol{q}]$. There have been proposed many algorithms to decide $\boldsymbol{q}$. Ichimori (2011) and Wada (2012) showed that most existing methods are interpreted as minimization of some α-divergence and their differences are only in the values of α.

4.2.5 α-Mean

By using the α-representation, we define the α-mean. Let us consider two positive numbers x and y. We rescale them by

$$\tilde{x} = h(x), \quad \tilde{y} = h(y), \tag{4.31}$$

where $h(x)$ is a monotonically increasing differentiable function satisfying $h(0) = 0$. We may call $h(x)$ the h-representation of x. The α-representation is the case of $h(x) = h_\alpha(x)$.

The quantity called the h-mean of x and y,

$$m_h(x, y) = h^{-1}\left\{\frac{h(x) + h(y)}{2}\right\}, \tag{4.32}$$

is obtained by using the h-representations of x and y, taking their arithmetic mean, and then rescaling it back by using h^{-1}. The α-mean of x and y is

$$m_\alpha(x, y) = \left\{\frac{1}{2}\left(x^{\frac{1-\alpha}{2}} + y^{\frac{1-\alpha}{2}}\right)\right\}^{\frac{2}{1-\alpha}}. \tag{4.33}$$

We further require that the h-mean is scale-free, implying that, for $c > 0$, the h-mean of cx and cy is c times their h-mean,

$$m_h(cx, cy) = cm_h(x, y). \tag{4.34}$$

The following theorem characterizes the α-mean.

Theorem 4.7 (Hardy et al. 1952) *The α-mean using*

$$h(u) = h_\alpha(u) = \begin{cases} u^{\frac{1-\alpha}{2}}, & \alpha \neq 1, \\ \log u, & \alpha = 1, \end{cases} \tag{4.35}$$

is the only scale-free means among h-means.

Proof We show the proof given by Amari (2007). Let h be a monotonically increasing differentiable function such that the h-mean is scale-free,

$$h(cm) = \frac{1}{2}\{h(cx) + h(cy)\}. \tag{4.36}$$

By differentiating Eq. (4.36) with respect to x, we derive

$$ch'(cm)m' = \frac{1}{2}ch'(cx), \tag{4.37}$$

where

$$m' = \frac{\partial}{\partial x}m(x, y). \tag{4.38}$$

By putting $c = 1$, we have

$$h'(m)m' = \frac{1}{2}h'(x). \tag{4.39}$$

Hence, we derive from (4.37) and (4.39),

$$\frac{h'(cx)}{h'(x)} = \frac{h'(cm)}{h'(m)}. \tag{4.40}$$

Since m takes an arbitrary value as y varies, we have

$$\frac{h'(cx)}{h'(x)} = g(c) \tag{4.41}$$

for a function $g(c)$ of c. By putting

$$k(x) = \log h'(x), \tag{4.42}$$

we have

$$k(cx) - k(x) = \log g(c). \tag{4.43}$$

Hence, we have

$$ck'(cx) = k'(x). \tag{4.44}$$

By putting $x = 1$,

$$k'(c) = \frac{b}{c} \tag{4.45}$$

for constant $b = k'(1)$. We finally derive

$$h(x) = \begin{cases} x^{\frac{1-\alpha}{2}}, & \alpha \neq 1, \\ \log x, & \alpha = 1, \end{cases} \tag{4.46}$$

neglecting a constant of proportionality. In the case of $\alpha = 1$, we have $\log x$. □

One sees that the family of α-means includes various known means:

$$\begin{aligned}
\alpha = 1 \text{ (geometric mean)}: &\quad m_1(a,b) = \sqrt{ab} \\
\alpha = -1 \text{ (arithmetic mean)}: &\quad m_{-1}(a,b) = \frac{1}{2}(a+b) \\
\alpha = 0: &\quad m_0(a,b) = \frac{1}{4}\left(\sqrt{a}+\sqrt{b}\right)^2 = \frac{1}{2}\left(\frac{1}{2}(a+b)+\sqrt{ab}\right) \\
\alpha = 3 \text{ (harmonic mean)}: &\quad m_3(a,b) = \frac{2}{\frac{1}{a}+\frac{1}{b}} \\
\alpha = \infty: &\quad m_\infty(a,b) = \min\{a,b\} \\
\alpha = -\infty: &\quad m_{-\infty}(a,b) = \max\{a,b\}.
\end{aligned}$$

The last two cases show that fuzzy logic is naturally included in the α-mean.

The α-mean is inversely monotone with respect to α,

$$m_\alpha(a,b) \geq m_{\alpha'}(a,b), \quad \alpha \leq \alpha'. \tag{4.47}$$

This is a generalization of the well-known inequalities

$$\frac{a+b}{2} \geq \sqrt{ab} \geq \frac{2}{a^{-1}+b^{-1}}. \tag{4.48}$$

As α increases, the α-mean relies more on the smaller element of $\{a, b\}$, while, as α decreases, the larger one is more emphasized. We may say that the α-mean with smaller α is pessimistic, and with larger α is more optimistic. See Fig. 4.1.

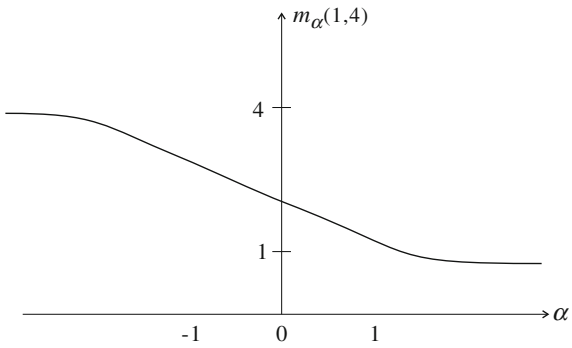


Fig. 4.1 α-mean of 1 and 4 for various α

We can further define the weighted α-mean of $a_1, \ldots, a_k$ with weights $w_1, \ldots, w_k$ by

$$m_\alpha(a_1, \ldots, a_k; \boldsymbol{w}) = h_\alpha^{-1}\left\{\sum w_i h_\alpha(a_i)\right\}, \tag{4.49}$$

where $\boldsymbol{w} = (w_1, \ldots, w_k)$ and $w_1 + \cdots + w_k = 1$. This leads us to the α-family of probability distributions in the next subsection.

4.2.6 α-Families of Probability Distributions

Given k probability distributions $p_i(x)$, $i = 1, \ldots, k$, we can define their α-mixture by using the α-mean.

The α-representation of probability density function $p(x)$ is given (Amari and Nagaoka 2000) by

$$h_\alpha[p(x)] = \begin{cases} p(x)^{(1-\alpha)/2}, & \alpha \neq 1, \\ \log p(x), & \alpha = 1. \end{cases} \tag{4.50}$$

Their α-mixture is defined by

$$\tilde{p}_\alpha(x) = c h_\alpha^{-1}\left\{\frac{1}{k}\sum_{i=1}^{k} h_\alpha\{p_i(x)\}\right\}, \tag{4.51}$$

where normalization constant c is necessary to make it a probability distribution. It is given by

$$c = \frac{1}{\int h_\alpha^{-1}\left\{\frac{1}{k}\sum h_\alpha[p_i(x)]\right\} dx}. \tag{4.52}$$

The $\alpha = -1$ mixture is the ordinary mixture and the $\alpha = 1$ mixture is the exponential mixture. The $\alpha = -\infty$ mixture,

$$\tilde{p}_{-\infty}(x) = c \max_i \{p_i(x)\}, \tag{4.53}$$

is the optimistic integration of component distributions in the sense that, for each x, it takes the largest values of the component probabilities. On the contrary, the $\alpha = \infty$ mixture is pessimistic, taking the minimum of the component probabilities,

$$\tilde{p}_{\infty}(x) = c \min \{p_i(x)\}. \tag{4.54}$$

The exponential mixture is more pessimistic than the ordinary mixture in the sense that the resulting probability density is close to 0 at x where some of the components are close to 0.

Let us next consider weighted mixtures. The weighted α-mixture with weights $w_1, \ldots, w_k$ satisfying $\sum w_i = 1$ is given by

$$\tilde{p}_{\alpha}(x; \boldsymbol{w}) = c h_{\alpha}^{-1} \left\{ \sum w_i h_{\alpha} \{p_i(x)\} \right\}. \tag{4.55}$$

This is called the α-integration of $p_1(x), \ldots, p_k(x)$ with weights $w_1, \ldots, w_k$. It connects k component distributions $p_1(x), \ldots, p_k(x)$ continuously by using the parameter $\boldsymbol{w} = (w_1, \ldots, w_k)$. It is called the α-family of probability distributions where $\boldsymbol{w}$ plays the role of its coordinate system. When $\alpha = -1$, this is an ordinary mixture family,

$$\tilde{p}_{-1}(x; \boldsymbol{w}) = \sum w_i p_i(x), \tag{4.56}$$

where $\sum w_i = 1$ is imposed. When $\alpha = 1$, this is an exponential family,

$$\tilde{p}_1(x, \boldsymbol{w}) = \exp \left\{ \sum w_i \log p_i(x) - \psi(\boldsymbol{w}) \right\}, \tag{4.57}$$

where the normalization constant is given by

$$c = \exp \{-\psi\}. \tag{4.58}$$

The probability simplex S_n (and the function space F of probability distributions) are special, satisfying the following theorem.

Theorem 4.8 *The probability simplex S_n is an α-family for any α.*

Proof S_n is a mixture of $\delta_i(x)$,

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x). \tag{4.59}$$

The α-mixture family of $\delta_i(x)$ is

$$\tilde{p}_\alpha(x, \boldsymbol{w}) = c \left[\sum w_i \delta_i(x) \right]^{\frac{2}{1-\alpha}}, \tag{4.60}$$

where

$$w_i = p_i^{\frac{1-\alpha}{2}}, \quad i = 0, 1, \ldots, n. \tag{4.61}$$

They cover the entire S_n so that S_n is an α-family. □

We can also show that an α-geodesic connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ in S_n is a one-dimensional α-family.

4.2.7 Optimality of α-Integration

When a cluster of k distributions $p_1(x), \ldots, p_k(x)$ is given, we search for $q(x)$ that is close to all of $p_1(x), \ldots, p_k(x)$. It is regarded as the center of the cluster. Let $w_1, \ldots, w_k$ be weights assigned to $p_i(x), i = 1, \ldots, k$, and we use the weighted average of divergences from $p_i(x)$'s to $q(x)$,

$$R_D[q(x)] = \sum w_i D[p_i(x) : q(x)] \tag{4.62}$$

as a risk function. We search for the distribution $q(x)$ that minimizes $R_D[q(x)]$. The minimizer of R_D is called the D-optimal integration of $p_1(x), \ldots, p_k(x)$ with weights $w_1, \ldots, w_k$. The following theorem characterizes the α-integration (Amari 2007).

Theorem 4.9 (Optimality of α-integration) *The α-integration of probability distributions $p_1(x), \ldots, p_k(x)$ with weights $w_1, \ldots, w_k$ is optimal under the α-risk,*

$$R_\alpha[q(x)] = \sum w_i D_\alpha[p_i(x) : q(x)], \tag{4.63}$$

where D_α is the α-divergence.

Proof Let us first prove the case of $\alpha \neq \pm 1$. By taking the variation of $R_\alpha[q(x)]$ under the normalizing constraint

$$\int q(x)dx = 1, \tag{4.64}$$

we derive

$$\begin{aligned} &\delta R_\alpha[q(x)] - \lambda \int \delta q(x)dx \\ &= \frac{2}{1-\alpha} \sum w_i \int p_i(x)^{\frac{1-\alpha}{2}} q(x)^{-\frac{1+\alpha}{2}} \delta q(x)dx - \lambda \int \delta q(x)dx \\ &= 0, \end{aligned} \tag{4.65}$$

where λ is the Lagrange multiplier. This gives

$$q(x)^{-\frac{1-\alpha}{2}} \sum w_i p_i(x)^{\frac{1-\alpha}{2}} = \text{const} \tag{4.66}$$

and hence, the optimal $q(x)$ is

$$q(x) = c h_\alpha^{-1} \left[\sum w_i h_\alpha \{p_i(x)\} \right]. \tag{4.67}$$

When $\alpha = \pm 1$, we obtain

$$\delta R_1[q(x)] = \sum w_i \int \log \frac{q(x)}{p_i(x)} \delta q(x) dx, \tag{4.68}$$

$$\delta R_{-1}[q(x)] = -\sum w_i \int \left\{ \frac{p_i(x)}{q(x)} + const \right\} \delta q(x) dx, \tag{4.69}$$

respectively. Hence, the optimal q is proved to be the α-integration for any α. □

The case with unnormalized probabilities, i.e., positive measures, is similar. The optimal integration $\tilde{m}_\alpha(x)$ of $m_1(x), \ldots, m_k(x)$ under the α-divergence criterion is

$$\tilde{m}_\alpha(x) = h_\alpha^{-1} \left[\sum w_i h_\alpha \{m_i(x)\} \right], \tag{4.70}$$

where the normalization constant is not necessary.

There are interesting papers concerning applications of the α-integration of stochastic evidences, see e.g., Wu (2009), Choi et al. (2013) and Soriano and Vergara (2015).

4.2.8 Application to α-Integration of Experts

Let us consider a system composed of k experts $M_1, \ldots, M_k$, each of which processes input signal $\boldsymbol{x}$ and emits its own answer. The answer of M_i is a response y corresponding to $\boldsymbol{x}$. More generally, consider the case that the output of M_i is a probability distribution of y, $p_i(y|\boldsymbol{x})$, or a positive measure, $m_i(y|\boldsymbol{x})$. The entire system integrates these answers and provides an integrated answer concerning the distribution of y given $\boldsymbol{x}$ (Fig. 4.2).

Let us assume that $w_i(\boldsymbol{x})$ is given as the weight or reliability of M_i for input $\boldsymbol{x}$. The α-risk of an integrated answer $q(y|\boldsymbol{x})$ is given by

$$R_\alpha [q(y|\boldsymbol{x})] = \sum w_i(\boldsymbol{x}) D_\alpha [p_i(y|\boldsymbol{x}) : q(y|\boldsymbol{x})]. \tag{4.71}$$

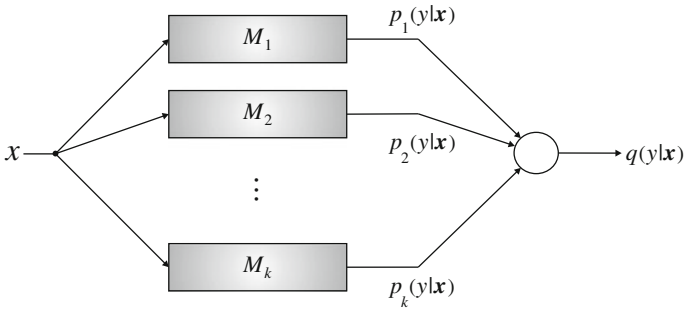


Fig. 4.2 Integration of answers of expert machines

Theorem 4.10 *The α-expert machine*

$$q(y|\boldsymbol{x}) = h_\alpha^{-1}\left[\sum w_i(\boldsymbol{x}) h_\alpha\{p_i(y|\boldsymbol{x})\}\right] \tag{4.72}$$

is optimal under the α-risk $R_\alpha[q(y|\boldsymbol{x})]$.

Similar assertions hold for the case of positive measures.

The $\alpha = 1$ machine is the mixture of experts (Jacobs et al. 1991) and the $\alpha = -1$ machine is the product of experts (Hinton 2002).

It is important to determine the weights or reliability functions $w_i(\boldsymbol{x})$. When a teacher output $q^*(y|\boldsymbol{x})$ is available, one may use the soft-max function

$$w_i(\boldsymbol{x}) = c\exp\left\{-\beta D_\alpha\left[p_i(y|\boldsymbol{x}) : q^*(y|\boldsymbol{x})\right]\right\} \tag{4.73}$$

as the weight of M_i, where c is the normalization constant and β is the "inverse temperature", indicating the effectiveness of the weights.

4.3 Geometry of Tsallis q-Entropy

The Boltzmann–Gibbs distribution in statistical physics is an exponential family, such that an invariant flat structure is given to the underlying manifold. Its convex function is free energy and its dual convex function is the negative of the Shannon entropy. C. Tsallis proposed a generalized entropy called the q-entropy for studying various phenomena not included in the conventional Boltzmann–Gibbs framework (Tsallis 1988, 2009). The induced probability distributions are not exponential families which are subject to exponential decay of tail probabilities. This has opened the door to a new world of physics and beyond. The q-logarithm and q-exponential are introduced to this end. However, the q-logarithm is essentially the same as the α-representation, where q and α are connected by $\alpha = 2q - 1$. Therefore, the α-geometry covers the

geometry of q-entropy physics (Ohara 2007). We treat the discrete case of S_n mostly, but the results hold in the continuous case, too.

We further extend the q-framework by using the q-escort distribution. This gives a new dually flat structure to S_n, although it is not invariant (Amari and Ohara 2011). It is conformally related to the invariant geometry (Amari et al. 2012). This framework is extended further to deformed exponential families proposed by Naudts (2011).

4.3.1 q-Logarithm and q-Exponential Function

Tsallis introduced a generalized logarithm, called the q-logarithm, by

$$\log_q(u) = \frac{1}{1-q}\left(u^{1-q} - 1\right), \tag{4.74}$$

which gives $\log u$ in the limit $q \to 1$. The inverse of the q logarithm is the q-exponential,

$$\exp_q(u) = \{1 + (1-q)u\}^{\frac{1}{1-q}}, \tag{4.75}$$

which gives the ordinary exponential function in the limit $q \to 1$. These functions are the same as the α-representation $h_\alpha(u)$ and its inverse, where $\alpha = 2q - 1$, except for a scaling factor and a constant. However, we keep the original q-notation rather than the α-notation in this section, respecting the original q-terminology by C. Tsallis.

The Tsallis q-entropy is defined by

$$H_q(\boldsymbol{p}) = \sum p_i \log_q \frac{1}{p_i}, \tag{4.76}$$

by replacing log by $\log_q$, which is concave for $0 < q \leq 1$ and is the Shannon entropy when $q = 1$. This is closely related to the Rényi entropy (Rényi 1961). Similarly, the q-divergence is defined by

$$D_q[\boldsymbol{p} : \boldsymbol{r}] = \mathrm{E}\left[\log_q \frac{r(x)}{p(x)}\right] = \frac{1}{1-q}\left(1 - \sum p_i^q r_i^{1-q}\right), \tag{4.77}$$

where E is the expectation with respect to $\boldsymbol{p}$. This is the same as the α-divergence (3.39) with $\alpha = 2q - 1$.

The geometry derived from the q-divergence satisfies the invariance criterion, since it belongs to the class of f-divergence. So the Riemannian metric is given by the Fisher information matrix. Further, it is not dually flat except for the limiting cases of $q = 0$ ($\alpha = -1$) and 1 ($\alpha = 1$). However, if we extend it to the manifold of positive measures, it is both invariant and dually flat.

4.3.2 *q*-Exponential Family (α-Family) of Probability Distributions

We define the q-exponential family by

$$\log_q p(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_q(\boldsymbol{\theta}), \tag{4.78}$$

or equivalently by

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp_q \left\{ \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_q(\boldsymbol{\theta}) \right\}, \tag{4.79}$$

where $\log_q$ and $\exp_q$ are used instead of log and exp in the ordinary exponential family. This is an α-family (4.60) of S_n, in which h_α is used instead of $\log_q$, $\boldsymbol{\theta} = (w_i)$ and $\boldsymbol{x} = \{\delta_i(x)\}$. Here, $\psi_q(\boldsymbol{\theta})$ is determined from the normalization constraint

$$\int \exp_q \left\{ \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_q(\boldsymbol{\theta}) \right\} d\boldsymbol{x} = 1. \tag{4.80}$$

Another example is the q-Gaussian distribution, given by

$$\log_q(x, \boldsymbol{\theta}) = -\frac{(x-\mu)^2}{2\sigma^2} = \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta}), \tag{4.81}$$

$$\boldsymbol{\theta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right), \quad \boldsymbol{x} = \left(x, x^2 \right), \tag{4.82}$$

where random variable x takes continuous values. Different from a Gaussian distribution, the values of x are limited within a finite range. Another important q-family is S_n. We rewrite Theorem 4.8 in the following form.

Theorem 4.11 *The family S_n of all the discrete distributions is a q-family for any q, that is an α-family for any α.*

Proof By introducing random variables $\delta_i(x)$ and putting $\boldsymbol{x} = (\delta_1(x), \ldots, \delta_n(x))$, a probability $\boldsymbol{p} \in S_n$ is written, using parameter $\boldsymbol{\theta}$, in the form

$$\log_q p(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1-q} \left\{ \sum_{i=1}^{n} \left(p_i^{1-q} - p_0^{1-q} \right) \delta_i(x) + p_0^{1-q} - 1 \right\}, \tag{4.83}$$

where the coordinate system $\boldsymbol{\theta}$ is

$$\theta^i = \frac{1}{1-q} \left(p_i^{1-q} - p_0^{1-q} \right), \quad x_i = \delta_i(x). \tag{4.84}$$

Hence, it is a q-family (α-family). The function corresponding to the free energy is

$$\psi_q(\boldsymbol{\theta}) = -\log_q p_0, \tag{4.85}$$

where p_0 is a function of $\boldsymbol{\theta}$. We call it the q-free energy. □

4.3.3 *q-Escort Geometry*

The q-geometry (α-geometry) is induced in a q-exponential family from the q-divergence. It consists of the Fisher information metric (3.68) and cubic tensor defined in (3.88). It is invariant but not flat in general. This is because the q-divergence (α-divergence) is not a Bregman divergence in general. However, it is possible to modify it conformally to obtain a new dually flat structure. To begin with, we show that the q-free energy $\psi_q(\boldsymbol{\theta})$ defined by (4.80) is a convex function of $\boldsymbol{\theta}$.

Lemma 4.1 *The q-free energy is convex.*

Proof By differentiating (4.79) with respect to $\boldsymbol{\theta}$, we have

$$\partial_i p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{\theta})^q \left(x_i - \partial_i \psi_q\right). \tag{4.86}$$

Its second derivatives are

$$\partial_i \partial_j p(\boldsymbol{x}, \boldsymbol{\theta}) = q p(\boldsymbol{x}, \boldsymbol{\theta})^{2q-1} \left(x_i - \partial_i \psi_q\right)\left(x_j - \partial_j \psi_q\right) - p(\boldsymbol{x}, \boldsymbol{\theta})^q \partial_i \partial_j \psi_q. \tag{4.87}$$

We introduce a functional

$$h_q\left[p(\boldsymbol{x})\right] = \int p(\boldsymbol{x})^q d\boldsymbol{x}, \tag{4.88}$$

which is the Tsallis q-entropy except for a scale and constant. Then, from (4.86) and (4.87) and by using the identities

$$\partial_i \int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = \partial_i \partial_j \int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = 0, \tag{4.89}$$

we have

$$\partial_i \psi_q(\boldsymbol{\theta}) = \frac{1}{h_q(\boldsymbol{\theta})} \int x_i p(\boldsymbol{x}, \boldsymbol{\theta})^q d\boldsymbol{x}, \tag{4.90}$$

$$\partial_i \partial_j \psi_q(\boldsymbol{\theta}) = \frac{q}{h_q(\boldsymbol{\theta})} \int \left(x_i - \partial_i \psi_q\right)\left(x_j - \partial_j \psi_q\right) p(\boldsymbol{x}, \boldsymbol{\theta})^{2q-1} d\boldsymbol{x}, \tag{4.91}$$

the latter of which shows that the Hessian of ψ_q is positive-definite. This is called the q-metric

$$g_{ij}^q = \partial_i \partial_j \psi_q(\boldsymbol{\theta}), \tag{4.92}$$

which is different from the invariant Fisher metric.

A new dually flat structure is introduced in S_n by the q-free-energy, which is different from the free energy. The affine coordinates are θ^i given by (4.84). The dual affine coordinate system $\boldsymbol{\eta}$ is given by

$$\eta_i = \partial_i \psi_q(\boldsymbol{\theta}) = \frac{p_i^q}{h_q(\boldsymbol{p})}. \tag{4.93}$$

The dual convex function is the inverse of the q-entropy

$$\varphi_q(\boldsymbol{\eta}) = \frac{1}{1-q}\left\{\frac{1}{h_q(\boldsymbol{p})} - 1\right\}, \tag{4.94}$$

except for a scale and constant.

The Bregman divergence derived by ψ_q is

$$\tilde{D}_q\left[p(\boldsymbol{x}) : r(\boldsymbol{x})\right] = \frac{1}{(1-q)h_q[r(\boldsymbol{x})]}\left(1 - \int p(\boldsymbol{x})^{1-q} r(\boldsymbol{x})^q d\boldsymbol{x}\right), \tag{4.95}$$

which is different from the q-divergence D_q. $\tilde{D}_q$ gives another dually flat Riemannian structure to S_n.

By putting

$$\tilde{p}_i = \eta_i, \quad i = 1, \ldots, n, \tag{4.96}$$

$$\tilde{p}_0 = \frac{p_0^q}{h_q(\boldsymbol{p})}, \tag{4.97}$$

$$\sum_{i=0}^{n} \tilde{p}_i = 1 \tag{4.98}$$

holds. So $\boldsymbol{\eta}$ gives another probability distribution $\tilde{\boldsymbol{p}}$ of S_n. We call it the escort probability distribution of $\boldsymbol{p}$. The escort distribution is obtained by changing p_i to p_i^q / h_q which shifts $\boldsymbol{p}$ toward the center (the uniform distribution $\boldsymbol{p}_0$) as q decreases from $q = 1$.

We can define the q-escort geodesic and dual q-escort geodesic in S_n. By using these geodesics, the q-Pythagorean theorem holds with respect to the q-escort divergence. One of the important consequences is the q-max entropy theorem. To this end, we define the q-escort expectation by

$$\tilde{\mathrm{E}}_q\left[a(x)\right] = \int \frac{a(x)p(x)^q}{h_q} dx. \tag{4.99}$$

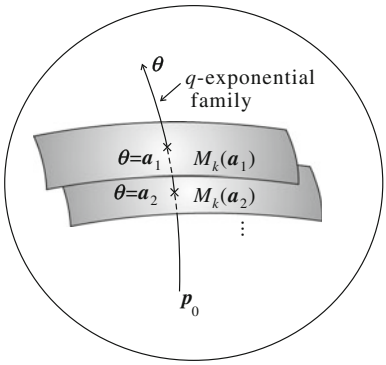


Fig. 4.3 q-max entropy theorem

Theorem 4.12 (q-Max-Entropy Theorem) *Let $M_k(\boldsymbol{a})$ be a submanifold of S_n consisting of probability distributions of which the q-escort expectations of random variables $c_1(\boldsymbol{x}), \ldots, c_k(\boldsymbol{x})$ take fixed values,*

$$\tilde{\mathrm{E}}_q\left[c_i(\boldsymbol{x})\right] = a_i, \quad i = 1, \ldots, k. \tag{4.100}$$

where $\boldsymbol{a} = (a_1, \ldots, a_k)$. The probability distribution $\hat{\boldsymbol{p}}(\boldsymbol{a})$ in $M_k(\boldsymbol{a})$ that maximizes the q-entropy is given by the q-geodesic projection of the uniform distribution $\boldsymbol{p}_0$ to $M_k(\boldsymbol{a})$. The family of such distributions for various $\boldsymbol{a} = \boldsymbol{\theta}$ is a q-exponential family of distributions,

$$\log_q p(\boldsymbol{x}, \boldsymbol{\theta}) = \theta^i c_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}). \tag{4.101}$$

Proof This is clear from the fact that M_k is flat in the dual sense and (4.101) is a flat submanifold in the primal sense. See Fig. 4.3. □

4.3.4 Deformed Exponential Family: χ-Escort Geometry

We used the q-logarithm to define the q-structure in S_n. However, we may use a more general representation to study various dually flat structures of S_n. See, for example, a deformed exponential family called the κ-exponential family (Kaniadakis and Scarfone 2002). Following Naudts (2011), we introduce the χ-logarithm defined by

$$\log_\chi(s) = \int_1^s \frac{1}{\chi(t)} dt, \tag{4.102}$$

where χ is a positive non-decreasing function. We simply put

$$u(s) = \log_\chi(s). \tag{4.103}$$

When χ is a power function

$$\chi(s) = s^q, \quad q > 0, \tag{4.104}$$

it gives the q-logarithm. We use the inverse of u as the v-representation,

$$v(s) = \exp_\chi(s) = u^{-1}(s). \tag{4.105}$$

The χ-deformed exponential family is defined by using (4.105) as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp_\chi \left\{ \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_\chi(\boldsymbol{\theta}) \right\}, \tag{4.106}$$

where ψ_χ is the free-energy corresponding to the normalization factor.

Theorem 4.13 *S_n is a χ-exponential family for any χ function.*

Proof We can prove the theorem in the same way as Theorem 4.11, by replacing $\log_q$ by $\log_\chi$. The affine coordinates are

$$\theta^i = u\left(p_i\right) - \psi_\chi, \quad i = 1, \ldots, n \tag{4.107}$$

and the χ-free-energy is

$$\psi_\chi(\boldsymbol{\theta}) = -u\left(p_0\right). \tag{4.108}$$

The χ-free-energy is a convex function of $\boldsymbol{\theta}$, so we can introduce a new dually flat affine structure together with a Riemannian metric. The Riemannian metric is written anew as

$$\partial_i \partial_j \psi_\chi(\boldsymbol{\theta}) = \frac{\int u''(\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi)\left(x_i - \partial_i \psi\right)\left(x_j - \partial_j \psi\right) d\boldsymbol{x}}{h_\chi(\boldsymbol{\theta})}, \tag{4.109}$$

where $h_\chi(\boldsymbol{\theta})$ is the χ-escort entropy defined by

$$h_\chi(\boldsymbol{\theta}) = \int \chi\left\{p(\boldsymbol{x}, \boldsymbol{\theta})\right\} d\boldsymbol{x} = \sum u'\left(\theta^i - \psi_\chi\right) + u'\left(-\psi_\chi\right). \tag{4.110}$$

The dual affine coordinates are given by

$$\eta_i = \frac{\int u'(\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi) x_i d\boldsymbol{x}}{h_\chi(\boldsymbol{\theta})} = \frac{1}{h_\chi(\boldsymbol{p})} \frac{1}{v'\left(p_i\right)}, \tag{4.111}$$

where

$$u'\left\{v(p)\right\} = \frac{1}{v'(p)} \tag{4.112}$$

is used. The dual $\boldsymbol{\eta}$ in (4.111) defines a probability distribution $\tilde{\boldsymbol{p}}$ called the χ-escort distribution. The dual convex function is

$$\varphi_\chi(\boldsymbol{\eta}) = \frac{1}{h_\chi} \sum_{i=0}^{n} \frac{v\left(p_i\right)}{v'\left(p_i\right)}. \tag{4.113}$$

The χ-divergence is

$$\begin{aligned} D_\chi\left[\boldsymbol{p} : \boldsymbol{q}\right] &= \psi_\chi\left(\boldsymbol{\theta}_p\right) + \varphi_\chi\left(\boldsymbol{\eta}_q\right) - \boldsymbol{\theta}_p \cdot \boldsymbol{\eta}_q \\ &= \frac{1}{h_\chi(\boldsymbol{p})} \sum_{i=0}^{n} \frac{u\left(p_i\right) - u\left(q_i\right)}{v'\left(p_i\right)}. \end{aligned} \tag{4.114}$$

The generalized Pythagorean theorem holds as well. □

Remark The $\exp_\chi(u)$ is a convex function. Vigelis and Cavalcante (2013) introduced a φ-family of probability distributions by using a convex function $\varphi(u)$. A new representation $f(x)$ of a probability density function $p(x)$ is given by

$$f(x) = \varphi\left\{p(x)\right\}. \tag{4.115}$$

This is closely related to the χ-representation. A φ-family of probability distributions and φ-divergence are defined in this framework, giving a dually flat structure. It is possible to extend to the non-parametric case.

4.3.5 Conformal Character of q-Escort Geometry

The q-divergence is an invariant divergence, leading to the Fisher information metric. The q-escort divergence (4.95) is not invariant and the derived metric is not the Fisher information metric. However, we see that the q-metric is connected to the Fisher metric $g_{ij}(\boldsymbol{\theta})$ by

$$\tilde{g}_{ij}(\boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}) g_{ij}(\boldsymbol{\theta}), \quad \sigma(\boldsymbol{\theta}) > 0, \tag{4.116}$$

where

$$\sigma(\boldsymbol{p}) = \frac{1}{h_q(\boldsymbol{p})}. \tag{4.117}$$

This implies that the metric is changed pointwise isotropically, implying that the magnitude of a vector is enlarged or shrunken by a factor $\sigma(\boldsymbol{p})$ but the angle of two vectors never changes, keeping the orthogonality invariant. Such a transformation of metric is called a conformal transformation. Hence, the q-escort structure is given by a conformal transformation from the invariant geometry. However, this property does not hold in the general χ-structure. We show the following theorem without proof. See Amari et al. (2012) and Ohara et al. (2012).

Theorem 4.14 *The q-escort geometry is unique among the χ-escort geometries in the sense that its Riemannian metric is derived by a conformal transformation of the invariant Fisher metric.*

Remark Conformal transformations are used in asymptotic theory of statistical inference (Okamoto et al. 1991; Kumon et al. 2011). They are also used in improving a kernel function in support vector machines, which will be shown later in Chap. 11.

4.4 (u, v)-Divergence: Dually Flat Divergence in Manifold of Positive Measures

We have used p and $\log p$ representations of probability, which play the role of two dual coordinate systems in the invariant geometry. We have further used the α- or q-representations, which lead us to the α-geometry. The generalized deformed exponential family uses the χ-representation. A representation of probability defines the geometry. The importance of representation was emphasized by Zhang (2004). Eguchi et al. (2014) uses a U-representation to define the U structure which is dually flat.

The present section considers $\boldsymbol{R}_+^n$ and introduces a dually flat structure by using a pair of representations. We extend the idea given by Zhang (2011, 2013) and establish a general dually flat structure in $\boldsymbol{R}_+^n$. The present section mostly follows Amari (2014) to define general decomposable and non-decomposable Bregman divergences in a manifold of positive measures. In the next section, they are extended to invariant Bregman divergences of a manifold of positive-definite matrices.

4.4.1 Decomposable (u, v)-Divergence

Let us use two monotonically increasing and differentiable functions $u(m)$ and $v(m)$ and define

$$\theta = u(m), \quad \eta = v(m). \tag{4.118}$$

They are called the u- and v-representations of positive measure m, respectively.

Given $\boldsymbol{m} \in \boldsymbol{R}_+^n$, we call $\boldsymbol{\theta} = \left(\theta^i\right)$ and $\boldsymbol{\eta} = (\eta_i)$ defined by

$$\theta^i = u\left(m_i\right), \quad \eta_i = v\left(m_i\right), \tag{4.119}$$

the u- and v-representations of $\boldsymbol{m}$, respectively. The $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are coordinate systems in $\boldsymbol{R}_n^+$. We search for a dually flat structure such that the u- and v-representations of $\boldsymbol{m}$ become two affine coordinates. To this end, we define a pair of convex functions $\psi_{u,v}(\boldsymbol{\theta})$ and $\varphi_{u,v}(\boldsymbol{\eta})$ from which a Bregman divergence $D_{u,v}\left[\boldsymbol{m} : \boldsymbol{m}'\right]$ is derived.

We define two scalar functions of θ and η by

$$\tilde{\psi}_{u,v}(\theta) = \int_0^{u^{-1}(\theta)} v(m)u'(m)dm, \tag{4.120}$$

$$\tilde{\varphi}_{u,v}(\eta) = \int_0^{v^{-1}(\eta)} u(m)v'(m)dm. \tag{4.121}$$

By differentiation, we have

$$\tilde{\psi}'_{u,v}(\theta) = v(m), \tag{4.122}$$

$$\tilde{\psi}''_{u,v}(\theta) = \frac{v'(m)}{u'(m)}. \tag{4.123}$$

Since $u'(m) > 0$, $v'(m) > 0$, $\tilde{\psi}''_{u,v} > 0$. Hence, $\tilde{\psi}_{u,v}(\theta)$ is a convex function. So is $\tilde{\varphi}_{u,v}(\eta)$. Moreover, they are the Legendre duals, because

$$\tilde{\psi}_{u,v}(\theta) + \tilde{\varphi}_{u,v}(\eta) - \theta\eta = \int_0^m v(m)u'(m)dm + \int_0^m u(m)v'(m)dm - u(m)v(m) = 0. \tag{4.124}$$

We now define decomposable convex functions of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ by

$$\psi_{u,v}(\boldsymbol{\theta}) = \sum \tilde{\psi}_{u,v}\left(\theta^i\right), \tag{4.125}$$

$$\varphi_{u,v}(\boldsymbol{\eta}) = \sum \tilde{\varphi}_{u,v}\left(\eta_i\right). \tag{4.126}$$

Definition 4.1 The (u, v)-divergence between two points $\boldsymbol{m}, \boldsymbol{m}' \in \boldsymbol{R}_n^+$ is defined by

$$\begin{aligned} D_{u,v}\left[\boldsymbol{m} : \boldsymbol{m}'\right] &= \psi_{u,v}(\boldsymbol{\theta}) + \varphi_{u,v}\left(\boldsymbol{\eta}'\right) - \boldsymbol{\theta} \cdot \boldsymbol{\eta}' \\ &= \sum \left[\int_0^{m_i} v(m)u'(m)dm \right. \\ &\quad \left. + \int_0^{m'_i} u(m)v'(m)dm - u\left(m_i\right) v\left(m'_i\right) \right], \end{aligned} \tag{4.127}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}'$ are u- and v-representations of $\boldsymbol{m}$ and $\boldsymbol{m}'$, respectively.

The (u, v)-divergence gives a dually flat structure, where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are affine and dual affine coordinate systems. The transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is simple in the (u, v)-structure, because it can be done componentwise,

$$\theta^i = u\left\{v^{-1}\left(\eta_i\right)\right\}, \tag{4.128}$$

$$\eta_i = v\left\{u^{-1}\left(\theta^i\right)\right\}. \tag{4.129}$$

This is a merit of the (u, v)-divergence. The Riemannian metric is given by

$$g_{ij}(\boldsymbol{m}) = \frac{v'(m_i)}{u'(m_i)}\delta_{ij}. \tag{4.130}$$

It is easy to see that this is a Euclidean metric. We have a new coordinate system $\boldsymbol{r}(\boldsymbol{m})$

$$r(m_i) = \int^{m_i} \sqrt{\frac{v'(m)}{u'(m)}}dm, \tag{4.131}$$

in which the Riemannian metric is $g_{ij} = \delta_{ij}$. The following theorem follows immediately.

Theorem 4.15 *A decomposable and dually flat divergence in $\boldsymbol{R}_+^n$ is a (u, v)-divergence when it is invariant under the permutation of indices.*

Many divergences are written in the form of (u, v)-divergence.

1. (α, β)-divergence

From the following power functions,

$$u(m) = \frac{1}{\alpha}m^{\alpha},\ v(m) = \frac{1}{\beta}m^{\beta}, \tag{4.132}$$

$$D_{\alpha,\beta}[\boldsymbol{p} : \boldsymbol{q}] = \frac{1}{\alpha\beta(\alpha+\beta)}\sum\left\{\alpha p_i^{\alpha+\beta} + \beta q_i^{\alpha+\beta} - (\alpha+\beta)p_i^{\alpha}q_i^{\beta}\right\} \tag{4.133}$$

is derived. This was introduced by Cichocki and Amari (2010) and Cichocki et al. (2011). The affine and dual affine coordinates are

$$\theta^i = \frac{1}{\alpha}(m_i)^{\alpha},\quad \eta_i = \frac{1}{\beta}(m_i)^{\beta} \tag{4.134}$$

and the convex functions are

$$\psi(\boldsymbol{\theta}) = c_{\alpha,\beta}\sum\theta_i^{\frac{\alpha+\beta}{\alpha}},\quad \varphi(\boldsymbol{\eta}) = c_{\beta,\alpha}\sum\eta_i^{\frac{\alpha+\beta}{\beta}}, \tag{4.135}$$

where

$$c_{\alpha,\beta} = \frac{1}{\beta(\alpha+\beta)}\alpha^{\frac{\alpha+\beta}{\alpha}}. \tag{4.136}$$

2. α-divergence

By putting

$$u(m) = \frac{2}{1-\alpha}m^{\frac{1-\alpha}{2}},\quad v(m) = \frac{2}{1+\alpha}m^{\frac{1+\alpha}{2}}, \tag{4.137}$$

we have

$$D_\alpha\left[\boldsymbol{m} : \boldsymbol{m}'\right] = \frac{4}{1-\alpha^2}\sum\left\{\frac{1-\alpha}{2}m_i + \frac{1+\alpha}{2}m_i^{\frac{1-\alpha}{2}} - m_i^{\alpha}\left(m_i'\right)^{\frac{1+\alpha}{2}}\right\}. \quad (4.138)$$

This is a special case of the (α, β)-divergence.

3. β-divergence

From

$$u(m) = m, \quad v(m) = \frac{1}{\beta}m^{1+\beta}, \quad (4.139)$$

we have

$$D_\beta\left[\boldsymbol{m} : \boldsymbol{m}'\right] = \frac{1}{\beta(\beta+1)}\sum_i\left[m_i^{\beta+1} + (\beta+1)m_i' - \left(m_i'\right)^{\beta+1} - (\beta+1)m_i\left(m_i'\right)^{\beta}\right]. \quad (4.140)$$

This is the β-divergence (Minami and Eguchi 2004). It gives a dually flat structure even in S_n. This is because $u(m)$ is linear in m.

4. U-divergence

From

$$u(m) = m, \quad v(m) = U'(m), \quad (4.141)$$

where $U(m)$ is a convex function, we have the U-divergence (Eguchi et al. 2014).

4.4.2 General (u, v) Flat Structure in $\boldsymbol{R}_+^n$

We consider a general dually flat structure of $\boldsymbol{R}_+^n$ which is not necessarily decomposable. Let us introduce a new coordinate system

$$\boldsymbol{\theta} = \boldsymbol{u}(\boldsymbol{m}) \quad (4.142)$$

in $\boldsymbol{R}_+^n$, where $\boldsymbol{u}$ is an arbitrary differentiable bijective vector function. We can define a dually flat structure in $\boldsymbol{R}_+^n$ by using an arbitrary convex function $\psi(\boldsymbol{\theta})$. $\boldsymbol{\theta}$ is the associated affine coordinate system and the dual affine coordinates are

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}). \quad (4.143)$$

We put

$$\boldsymbol{v}(\boldsymbol{m}) = \nabla\psi(\boldsymbol{\theta}). \quad (4.144)$$

This structure is used in Nock et al. (2015).

An arbitrary pair $(\boldsymbol{u}, \boldsymbol{v})$ of coordinate systems do not necessarily give a dually flat structure. They give dually flat structure when and only when there exists a convex function $\psi(\boldsymbol{\theta})$ such that

$$\boldsymbol{\eta} = \boldsymbol{v}\left\{\boldsymbol{u}^{-1}(\boldsymbol{\theta})\right\} \tag{4.145}$$

is its gradient. In the case of a decomposable pair (u, v), the condition is always satisfied and the pair always defines a dually flat structure.

The Riemannian metric induced from a $(\boldsymbol{u}, \boldsymbol{v})$-structure is $G(\boldsymbol{\theta}) = \nabla\nabla\psi(\boldsymbol{\theta})$, which is not Euclidean in general.

4.5 Invariant Flat Divergence in Manifold of Positive-Definite Matrices

The present section studies information geometry of the manifold of positive-definite matrices, following Amari (2014). See also Ohara and Eguchi (2013). An extensive review is found in Cichocki et al. (2015). A positive definite matrix $\mathbf{A}$ is decomposed as

$$\mathbf{A} = \mathbf{O}^T \boldsymbol{\Lambda} \mathbf{O}, \tag{4.146}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix consisting of positive entries (eigenvalues of $\mathbf{A}$) and $\mathbf{O}$ is an orthogonal matrix. A positive-definite diagonal matrix is compared with a positive measure distribution. When its trace is 1, it is compared with a probability distribution. So a positive-definite matrix is an extension of a positive measure. Therefore, one can introduce a dually flat structure to the manifold of positive-definite matrices with the help of the (u, v)-structure. The manifold of positive-definite Hermitian matrices, in particular those with a trace equal to 1, are important in quantum information theory, but we do not study them, treating only the real case.

4.5.1 Bregman Divergence and Invariance Under Gl(n)

Let $\mathbf{P}$ be a positive-definite symmetric matrix and $\psi(\mathbf{P})$ be a convex function. A Bregman divergence is defined between two positive-definite matrices $\mathbf{P}$ and $\mathbf{Q}$ by

$$D[\mathbf{P} : \mathbf{Q}] = \psi(\mathbf{P}) - \psi(\mathbf{Q}) - \nabla\psi(\mathbf{Q}) \cdot (\mathbf{P} - \mathbf{Q}), \tag{4.147}$$

where ∇ is the gradient operator with respect to matrix $\mathbf{P} = \left(P_{ij}\right)$ and hence $\nabla\psi$ is a matrix, and the inner product of two matrices is defined by

$$\mathbf{Q} \cdot \mathbf{P} = \mathrm{tr}\left\{\mathbf{QP}\right\}. \tag{4.148}$$

It induces a dually flat structure in the manifold of positive-definite matrices, where the affine coordinate system is $\mathbf{P}$ itself and the dual affine coordinate system is

$$\mathbf{P}^* = \nabla\psi(\mathbf{P}). \tag{4.149}$$

There is a one-to-one correspondence between positive-definite matrices and zero-mean multivariate Gaussian distributions. Indeed, a zero-mean multivariate Gaussian distribution is given by using a positive-definite matrix $\mathbf{P}$ as

$$p(\boldsymbol{x}, \mathbf{P}) = \exp\left\{\frac{1}{2}\boldsymbol{x}^T\mathbf{P}^{-1}\boldsymbol{x} - \log\sqrt{2\pi\det|\mathbf{P}|}\right\}, \tag{4.150}$$

which is an exponential family. Its e-affine coordinates are $\mathbf{P}^{-1}$. The flat geometry is, therefore, given by the KL-divergence,

$$D[\mathbf{P} : \mathbf{Q}] = \mathrm{tr}\left(\mathbf{P}\mathbf{Q}^{-1}\right) - \log\left(\det\left|\mathbf{P}\mathbf{Q}^{-1}\right|\right) - n, \tag{4.151}$$

which is obtained from the potential function

$$\psi(\mathbf{P}^{-1}) = -\log(\det|\mathbf{P}^{-1}|). \tag{4.152}$$

Let us consider a linear transformation of $\mathbf{P}$ by $\mathbf{L} \in Gl(n)$, which is the set of all non-degenerate $n \times n$ matrices, given by

$$\tilde{\mathbf{P}} = \mathbf{L}^T\mathbf{P}\mathbf{L}. \tag{4.153}$$

This corresponds to the transformation of random variable $\boldsymbol{x}$ to

$$\tilde{\boldsymbol{x}} = \mathbf{L}\boldsymbol{x}. \tag{4.154}$$

A divergence is said to be invariant under $Gl(n)$ when it satisfies

$$D\left[\mathbf{P} : \mathbf{Q}\right] = D\left[\mathbf{L}^T\mathbf{P}\mathbf{L} : \mathbf{L}^T\mathbf{Q}\mathbf{L}\right]. \tag{4.155}$$

Since the KL-divergence is invariant under any transformation of $\boldsymbol{x}$, it is invariant under $Gl(n)$.

Theorem 4.16 *The KL-divergence is a flat divergence which is invariant under $Gl(n)$ in the manifold of positive-definite matrices.*

4.5.2 Invariant Flat Decomposable Divergences Under $O(n)$

The eigenvalues of a positive-definite matrix do not change under an orthogonal transformation $\mathbf{O} \in O(n)$, the group of orthogonal matrices. It is natural to consider a dually flat structure which is invariant under $O(n)$.

4.5.2.1 The Case When P is *e*-Affine

We have a convex function $\psi(\mathbf{P})$ of $\mathbf{P}$ in this case. It is invariant under $O(n)$ when

$$\psi(\mathbf{P}) = \psi\left(\mathbf{O}^T\mathbf{P}\mathbf{O}\right). \tag{4.156}$$

An invariant function is a symmetric function of n eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{P}$ (Dhillon and Tropp 2007). An invariant convex function of $\mathbf{P}$ is written using a convex function f of one variable satisfying $f(0) = 0$ as

$$\psi_f(\mathbf{P}) = \sum f(\lambda_i) = \mathrm{tr} f(\mathbf{P}), \tag{4.157}$$

when it is decomposable in the additive form of λ_i. We study this case. We can prove the following lemma.

Lemma

$$\mathbf{P}^* = \nabla\psi_f(\mathbf{P}) = f'(\mathbf{P}). \tag{4.158}$$

Outline of the proof. We assume that f is an analytic function. Then, $f(\mathbf{P})$ is expanded in a power series of $\mathbf{P}$. Therefore, we prove the lemma in the case of $f(\mathbf{P}) = \mathbf{P}^n$, which is easy. Hence, we have the lemma. □

Let $g(u)$ be a function such that $g'(u)$ is the inverse function of $f'(u)$, satisfying $g(0) = 0$. Then, the inverse transformation from $\mathbf{P}'$ to $\mathbf{P}$ is given by

$$\mathbf{P} = g'\left(\mathbf{P}'\right). \tag{4.159}$$

Hence, the dual potential function is

$$\varphi_f\left(\mathbf{P}^*\right) = \mathrm{tr}\left\{g\left(\mathbf{P}^*\right)\right\}. \tag{4.160}$$

Theorem 4.17 *An e-flat decomposable $O(n)$-invariant divergence is given by*

$$D_f[\mathbf{P} : \mathbf{Q}] = \psi_f(\mathbf{P}) + \varphi_f\left\{f'(\mathbf{Q})\right\} - \mathrm{tr}\left\{\mathbf{P}\mathrm{f}'(\mathbf{Q})\right\}, \tag{4.161}$$

where φ_f is the Legendre dual of ψ_f.

We give well-known examples of invariant symmetric convex functions and dually flat divergences.

(1) For $f(\lambda) = (1/2)\lambda^2$, we have

$$\psi(\mathbf{P}) = \frac{1}{2}\sum \lambda_i^2, \tag{4.162}$$

$$D[\mathbf{P} : \mathbf{Q}] = \frac{1}{2}\|\mathbf{P} - \mathbf{Q}\|^2, \tag{4.163}$$

where $\|\mathbf{P}\|^2$ is the Frobenius norm

$$\|\mathbf{P}\|^2 = \sum P_{ij}^2. \tag{4.164}$$

This gives a Euclidean structure.
(2) For $f(\lambda) = -\log\lambda$, we have (4.152) and (4.151), which are invariant under $Gl(n)$.
(3) For $f(\lambda) = \lambda\log\lambda - \lambda$,

$$\psi(\mathbf{P}) = \mathrm{tr}\,(\mathbf{P}\log\mathbf{P} - \mathbf{P}), \tag{4.165}$$

$$D[\mathbf{P} : \mathbf{Q}] = \mathrm{tr}\,(\mathbf{P}\log\mathbf{P} - \mathbf{P}\log\mathbf{Q} - \mathbf{P} + \mathbf{Q}). \tag{4.166}$$

This divergence is used in quantum information theory. The affine coordinate system is $\mathbf{P}$, the dual affine coordinate system is $\log\mathbf{P}$ and $\psi(\mathbf{P})$ is related to the von Neumann entropy.

4.5.2.2 General Dually Flat Decomposable Case: (u, v)-Divergence

We use the (u, v)-structure to introduce a general dually flat invariant decomposable divergence. Let

$$\boldsymbol{\Theta} = u(\mathbf{P}), \quad \mathbf{H} = v(\mathbf{P}) \tag{4.167}$$

be u- and v-representations of matrices. We use two functions $\tilde{\psi}_{u,v}(\theta)$ and $\tilde{\varphi}_{u,v}(\eta)$ defined by (4.120) and (4.121) for defining a pair of dually coupled invariant convex functions,

$$\psi(\boldsymbol{\Theta}) = \mathrm{tr}\,\tilde{\psi}_{u,v}\{\boldsymbol{\Theta}\}, \tag{4.168}$$

$$\varphi(\boldsymbol{H}) = \mathrm{tr}\,\tilde{\varphi}_{u,v}\{\boldsymbol{H}\}. \tag{4.169}$$

They are not convex with respect to $\mathbf{P}$, but convex with respect to $\boldsymbol{\Theta}$ and $\mathbf{H}$, respectively. The derived Bregman divergence is

$$D_{u,v}[\mathbf{P} : \mathbf{Q}] = \psi\{\Theta(\mathbf{P})\} + \varphi\{\mathbf{H}(\mathbf{Q})\} - \boldsymbol{\Theta}(\mathbf{P}) \cdot \mathbf{H}(\mathbf{Q}). \tag{4.170}$$

It induces a dually flat structure to the manifold of positive-definite matrices.

Theorem 4.18 *A dually flat, invariant and decomposable divergence is a (u, v)-divergence in the manifold of positive-definite matrices.*

The Euclidean, Gaussian and von Neumann divergences given in (4.163), (4.151) and (4.166) are special examples of (u, v)-divergences. They are given by

$$(1)\quad u(m) = v(m) = m, \tag{4.171}$$

$$(2)\quad u(m) = m, \quad v(m) = -\frac{1}{m}, \tag{4.172}$$

$$(3)\quad u(m) = m, \quad v(m) = \log m. \tag{4.173}$$

When u and v are power functions, we have the (α, β)-structure in the manifold of positive-definite matrices.

(4) $(\alpha$-$\beta)$-divergence

By using the (α, β)-structure given by (4.132), we have

$$\psi(\boldsymbol{\Theta}) = \frac{\alpha}{\alpha+\beta}\mathrm{tr}\ \boldsymbol{\Theta}^{\frac{\alpha+\beta}{\alpha}} = \frac{\alpha}{\alpha+\beta}\mathrm{tr}\ \mathbf{P}^{\alpha+\beta}, \tag{4.174}$$

$$\varphi(\mathbf{H}) = \frac{\beta}{\alpha+\beta}\mathrm{tr}\ \mathbf{H}^{\frac{\alpha+\beta}{\beta}} = \frac{\beta}{\alpha+\beta}\mathrm{tr}\ \mathbf{P}^{\alpha+\beta} \tag{4.175}$$

and the (α, β)-divergence of matrices,

$$D_{\alpha\beta}[\mathbf{P} : \mathbf{Q}] = \mathrm{tr}\left\{\frac{\alpha}{\alpha+\beta}\mathbf{P}^{\alpha+\beta} + \frac{\beta}{\alpha+\beta}\mathbf{Q}^{\alpha+\beta} - \mathbf{P}^{\alpha}\mathbf{Q}^{\beta}\right\}. \tag{4.176}$$

This is a Bregman divergence, where the affine coordinate system is $\boldsymbol{\Theta} = \mathbf{P}^{\alpha}$ and its dual is $\mathbf{H} = \mathbf{P}^{\beta}$.

(5) The α-divergence is derived as

$$\boldsymbol{\Theta}(\mathbf{P}) = \frac{2}{1-\alpha}\mathbf{P}^{\frac{1-\alpha}{2}}, \tag{4.177}$$

$$\psi(\boldsymbol{\Theta}) = \frac{2}{1+\alpha}\mathbf{P}, \tag{4.178}$$

$$D_{\alpha}[\mathbf{P} : \mathbf{Q}] = \frac{4}{1-\alpha^2}\mathrm{tr}\left(-\mathbf{P}^{\frac{1-\alpha}{2}}\mathbf{Q}^{\frac{1+\alpha}{2}} + \frac{1-\alpha}{2}\mathbf{P} + \frac{1+\alpha}{2}\mathbf{Q}\right) \tag{4.179}$$

The affine coordinate system is $\frac{2}{1-\alpha}\mathbf{P}^{\frac{1-\alpha}{2}}$ and its dual is $\frac{2}{1+\alpha}\mathbf{P}^{\frac{1+\alpha}{2}}$.

(6) The β-divergence is derived from (4.139) as

$$D_{\beta}[\mathbf{P} : \mathbf{Q}] = \frac{1}{\beta(\beta+1)}\mathrm{tr}\left[\mathbf{P}^{\beta+1} + (\beta+1)\mathbf{Q} - \mathbf{Q}^{\beta+1} - (\beta+1)\mathbf{P}\mathbf{Q}^{\beta}\right]. \tag{4.180}$$

4.5.3 Non-flat Invariant Divergences

We have so far studied invariant flat divergences. There are other types of invariant divergences which are not necessarily flat. We remark that the eigenvalues of $\boldsymbol{P}\,\boldsymbol{Q}^{-1}$ are invariant under $Gl(n)$, because, for $\tilde{\mathbf{P}} = \mathbf{L}^T\mathbf{P}\mathbf{L}$ and $\tilde{\mathbf{W}} = \mathbf{L}^T\mathbf{Q}\mathbf{L}$,

$$\tilde{\mathbf{P}}\tilde{\mathbf{Q}}^{-1} = \mathbf{L}^T\left(\mathbf{P}\mathbf{Q}^{-1}\right)\left(\mathbf{L}^T\right)^{-1} \tag{4.181}$$

holds. So a divergence $D[\mathbf{P} : \mathbf{Q}]$ is invariant when it is written as a function of $\boldsymbol{\Lambda} = diag\,(\lambda_1, \ldots, \lambda_n)$, where λ_i are the eigenvalues of $\mathbf{P}\mathbf{Q}^{-1}$.

Cichocki et al. (2015) introduced the following (α-β)-log-det divergence:

$$D_{\alpha,\beta}^{\log\text{-}\det}[\mathbf{P} : \mathbf{Q}] = \frac{1}{\alpha\beta}\log\det\frac{\alpha\left(\mathbf{P}\mathbf{Q}^{-1}\right)^{\beta} + \beta\left(\mathbf{P}\mathbf{Q}^{-1}\right)^{\alpha}}{\alpha+\beta}, \tag{4.182}$$

which can be written in terms of $\boldsymbol{\Lambda}$ as

$$\begin{aligned} D_{\alpha,\beta}^{\log\text{-}\det}[\mathbf{P} : \mathbf{Q}] &= \frac{1}{\alpha\beta}\log\det\frac{\alpha\boldsymbol{\Lambda}^{\beta} + \beta\boldsymbol{\Lambda}^{\alpha}}{\alpha+\beta} \\ &= \frac{1}{\alpha\beta}\sum\log\frac{\alpha\lambda_i^{\beta} + \beta\lambda_i^{-\alpha}}{\alpha+\beta}. \end{aligned} \tag{4.183}$$

It is extended to the case of $\alpha = 0$ and/or $\beta = 0$ by taking the limit $\alpha, \beta \to 0$. For example,

$$D_{\alpha,0}^{\log\text{-}\det}[\mathbf{P} : \mathbf{Q}] = \frac{1}{\alpha^2}\left[\sum\left\{(\lambda_i)^{-\alpha} + \alpha\log\lambda_i\right\} - n\right], \tag{4.184}$$

$$D_{0,0}^{\log\text{-}\det}[\mathbf{P} : \mathbf{Q}] = \frac{1}{2}\sum(\log\lambda_i)^2\,. \tag{4.185}$$

When $\alpha = \beta$, $D_{\alpha,\beta}^{\log\text{-}\det}[\mathbf{P} : \mathbf{Q}]$ is symmetric with respect to $\mathbf{P}$ and $\mathbf{Q}$ and hence the geometry is self-dual and Riemannian.

It is interesting to see that $D_{\alpha,\beta}^{\log\text{-}\det}[\mathbf{P} : \mathbf{Q}]$ generates the same Riemannian metric not depending on α and β, although the dual affine connections do depend on α and β.

Theorem 4.19 *The Riemannian metric induced from the* (α, β)-log-det *divergence is*

$$\langle d\mathbf{P}, d\mathbf{P}\rangle_{\mathbf{P}} = \frac{1}{2}\mathrm{tr}\left(\mathrm{d}\mathbf{P}\mathbf{P}^{-1}\mathrm{d}\mathbf{P}\mathbf{P}^{-1}\right). \tag{4.186}$$

We omit the proof.

4.6 Miscellaneous Divergences

Many divergences have been defined in the literature. We show some of them. They are not invariant and not flat in general, but have their own characteristics. An extensive survey on divergence is found in Basseville (2013). See also Cichocki et al. (2009, 2011), for example. Only a Bregman divergence generates a dually flat structure. However, any divergence generates a dual pair of affine connections together with a Riemannian metric, as will be shown in Part II.

4.6.1 γ-Divergence

The γ-divergence was proposed by Fujisawa and Eguchi (2008). See also Cichocki and Amari (2010). Let γ be a real parameter. The γ-divergence between two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined by

$$D_\gamma[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{\gamma(\gamma-1)} \log \frac{\sum p_i^\gamma \left(\sum q_i^\gamma\right)^{\gamma-1}}{\left(\sum p_i q_i^{\gamma-1}\right)^\gamma}. \tag{4.187}$$

It is projectively invariant in the sense that, for any positive constants c_1 and c_2,

$$D_\gamma\left[c_1\boldsymbol{p}:c_2\boldsymbol{q}\right] = D_\gamma[\boldsymbol{p}:\boldsymbol{q}] \tag{4.188}$$

holds.

The γ-divergence has a super-robust property when we use it in statistical estimation. It is extremely robust even when outliers are mixed in observed data. It is possible to define the γ-divergence between positive-definite matrices $\mathbf{P}$ and $\mathbf{Q}$ as

$$D_\gamma[\mathbf{P}:\mathbf{Q}] = \frac{1}{\gamma(\gamma-1)} \log \frac{\mathrm{tr}\mathbf{P}^\gamma \{(\mathrm{tr}\mathbf{Q})^\gamma\}^{\gamma-1}}{\left\{\mathrm{tr}\mathbf{P}\mathbf{Q}^{\gamma-1}\right\}^\gamma}. \tag{4.189}$$

4.6.2 Other Types of (α, β)-Divergences

Zhang (2004) introduced the following (α, β)-divergence,

$$D_{Zhang}^{\alpha,\beta}[\boldsymbol{p}:\boldsymbol{q}] = \frac{4}{1-\alpha^2}\frac{2}{1+\beta}\sum\left\{\frac{1-\alpha}{2}p_i + \frac{1+\alpha}{2}q_i - \left(\frac{1-\alpha}{2}p^{\frac{1-\beta}{2}} + \frac{1+\alpha}{2}q_i^{\frac{1-\beta}{2}}\right)^{\frac{2}{1-\beta}}\right\}, \tag{4.190}$$

which is different from that in the previous subsection. The geometry induced from (4.190) is exactly the same as the α-geometry.

Zhang (2011) presented another α-divergence when a convex function $\psi(\boldsymbol{p})$ exists. It is given by

$$D_{\varphi}^{\alpha}[p(x):q(x)] = \frac{4}{1-\alpha^2}\int\left[\frac{1-\alpha}{2}\varphi(p)+\frac{1-\alpha}{2}\varphi(q) - \varphi\left\{\frac{1-\alpha}{2}p+\frac{1+\alpha}{2}q\right\}\right]dx. \tag{4.191}$$

Furuichi (2010) also introduced another (α-β)-divergence,

$$D_{Furuichi}^{\alpha,\beta}[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{\alpha-\beta}\sum\left(p_i^{\alpha}q_i^{1-\alpha}-p_i^{\beta}q_i^{1-\beta}\right). \tag{4.192}$$

4.6.3 Burbea–Rao Divergence and Jensen–Shannon Divergence

For a convex function $F(\boldsymbol{p})$, one can construct a symmetric divergence by

$$D_F[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{2}\{F(\boldsymbol{p})+F(\boldsymbol{q})\} - F\left(\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right). \tag{4.193}$$

This is called the Burbea–Rao divergence (Burbea and Rao 1982). When we use the negative of entropy as a convex function, we have

$$D_{JS}[\boldsymbol{p}:\boldsymbol{q}] = H\left(\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right) - \frac{1}{2}\{H(\boldsymbol{p})+H(\boldsymbol{q})\}. \tag{4.194}$$

This is called the Jensen–Shannon divergence. It can be rewritten using the KL-divergence as

$$D_{JS}[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{2}\left\{D_{KL}\left[\boldsymbol{p}:\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right]+D_{KL}\left[\boldsymbol{q}:\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right]\right\}. \tag{4.195}$$

These are not flat in general.

We have the α-version of the Burbea–Rao divergence

$$D_F^{\alpha}[\boldsymbol{p}:\boldsymbol{q}] = \alpha F(\boldsymbol{p})+(1-\alpha)F(\boldsymbol{q}) - F\{\alpha\boldsymbol{p}+(1-\alpha)\boldsymbol{q}\}. \tag{4.196}$$

This is asymmetric divergence.

4.6.4 (ρ, τ)-Structure and (F, G, H)-Structure

Zhang (2004) considered two representations of probabilities p_i in S_n by generalizing $\pm\alpha$-representations. Let ρ be a positive increasing function, and call

$$\rho_i = \rho(p_i) \tag{4.197}$$

the ρ-representation of probability p_i. In the continuous case, $\rho(x) = \rho\{p(x)\}$ is the ρ-representation. For a differentiable convex function $f(\rho)$, we define a positive increasing function

$$\tau(p) = f'\{\rho(p)\}, \tag{4.198}$$

which is another representation, τ-representation, of probability,

$$\tau_i = \tau(p_i). \tag{4.199}$$

This was proposed earlier and is the same as the (u, v)-structure of Sect. 4.4.1 defined in $\boldsymbol{R}^n_+$.

Harsha and Moosath (2014) introduced a non-invariant dual structure called the (F, G, H)-structure to a manifold of probability distributions. However, it is proved to be equivalent to the (ρ, τ)-structure, Zhang (2015). Let $G(u)$ be a smooth positive function. The G-metric is defined by

$$g^G_{ij}(\boldsymbol{\xi}) = \int \partial_i l(x, \boldsymbol{\xi}) \partial_j l(x, \boldsymbol{\xi}) p(x, \boldsymbol{\xi}) G\{p(x, \boldsymbol{\xi})\} dx, \tag{4.200}$$

which reduces to the invariant Fisher metric when $G(u) = 1$. Let F and H be two differentiable monotonically increasing positive functions. We call $F\{p(x, \boldsymbol{\xi})\}$ and $H\{p(x, \boldsymbol{\xi})\}$ the F- and H-representations of probability, respectively.

We define the (F, G)-connection by

$$\nabla^{F,G}_{\partial_i} \partial_j = \langle \partial_i \partial_j F, \partial_k F \rangle_G \, g^{km} \partial_m, \tag{4.201}$$

where $\langle \cdot, \cdot \rangle_G$ denotes the inner product by using the G-metric. It is represented in the component form as

$$\Gamma^{F,G}_{ijk} = \int \left[\partial_j \partial_j l + \left\{ 1 + \frac{F''(p)}{F'(p)} \right\} \partial_i l \partial_j l \right] \partial_k l G(p) p dx. \tag{4.202}$$

Similarly, we define the (H, G)-connection.

Theorem 4.20 *The (F, G)-connection and (H, G)-connection are dual with respect to the G-metric when the following relation holds:*

$$F'(u) H'(u) = \frac{G(u)}{u}. \tag{4.203}$$

The proof is omitted.

The α-(ρ, τ) divergence is defined by

$$D^{\alpha}_{\rho,\tau}[\boldsymbol{p} : \boldsymbol{q}] = \frac{1}{1-\alpha^2} \sum \left[\frac{1-\alpha}{2} f\{\rho(p_i)\} + \frac{1+\alpha}{2} f\{\rho(q_i)\} - f\left\{ \frac{1-\alpha}{2}\rho(p_i) + \frac{1+\alpha}{2}\rho(q_i) \right\} \right]. \tag{4.204}$$

This is neither a Bregman divergence nor an invariant divergence in general, but covers a wide range of divergences in S_n.

Remarks

We have seen that a dually flat structure is derived from a Bregman divergence. There are many divergences of the Bregman type which lead to different dually flat Riemannian structures. The invariance is a criterion which specifies a reasonable divergence in a manifold of probability distributions. We have searched for the divergence that is invariant and, at the same time, dually flat in the manifold S_n of probability distributions. The KL-divergence is the unique divergence of the Bregman type that is invariant.

If we consider the extended manifold of $\boldsymbol{R}^n_+$, the α-divergences are derived as a unique class of invariant divergences of the Bregman type. This introduces the α-geometry to the manifold of probability distributions. It is invariant geometry but is not necessarily dually flat except for the case of $\alpha = \pm 1$, which gives the KL-divergence. The α-geometry is interesting. We have shown the α-Pythagorean theorem and α-projection theorem in an α-family despite the fact that the manifold is not dually flat. More generally, given a general divergence and a point P in a submanifold $S \subset M$, the set of point Q that minimizes $D[Q : M]$ at $P \in M$ does not form a geodesic submanifold orthogonal to M at P. That is, the minimizer P is not the geodesic projection of Q to M. However, in the case of an α-family, this is given by the α-geodesic projection for the α-divergence. The α-projection is useful in applications. See, e.g., Matsuyama (2003).

It is a happy coincidence that the Tsallis q-geometry of the q-entropy is exactly the same as the geometry where $\alpha = 2q - 1$. Furthermore, the q-geometry introduced the escort probability distributions, which lead us to the conformal flattening of the non-flat q-geometry. This gives a new q-divergence of the Bregman type, from which flat (but non-invariant) geometry is derived. This idea has been generalized to a general deformed exponential family.

Apart from the framework of invariance, we introduced a general class of decomposable and non-decomposable divergences of the Bregman type in $\boldsymbol{R}^n_+$. They are the (u, v)- and $(\boldsymbol{u}, \boldsymbol{v})$-divergence. This is extended to give an invariant dually flat geometry to the manifold of positive-definite matrices. Quantum information geometry deals with a manifold of positive-definite Hermite matrices (which is a complex version of positive-definite real matrices). Therefore, the invariant (u, v)-structure would be useful in studying quantum information geometry, although we cannot explore it in the present monograph.

Divergences are used in various applications. The choice of a divergence function depends on the purpose of the application. An invariant divergence gives a Fisher efficient estimator but is not robust. There are robust divergences like the γ-divergence. A decomposable divergence is used in many applications, because they are simple and the coordinate transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is tractable.

Part II
Introduction to Dual Differential Geometry

Chapter 5
Elements of Differential Geometry

Here is an introduction to Riemannian geometry. The reader does not need to understand the detailed derivations of equations. More important are ideas and concepts of differential geometry. They can be understood "intuitively" without tears.

5.1 Manifold and Tangent Space

Let us consider an n-dimensional manifold M having a (local) coordinate system $\boldsymbol{\xi} = \left(\xi^1, \ldots, \xi^n\right)$. It is in general curved. The tangent space $T_{\boldsymbol{\xi}}$ at point $\boldsymbol{\xi}$ is a vector space spanned by n tangent vectors along the coordinate curves of ξ^i. We denote them as $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$, which is a basis of the tangent space (Fig. 5.1). Tangent space $T_{\boldsymbol{\xi}}$ is regarded as a linearization of M in a neighborhood of $\boldsymbol{\xi}$, since a small line element $d\boldsymbol{\xi}$ of M connecting two nearby points $P = \boldsymbol{\xi}$ and $P' = (\boldsymbol{\xi} + d\boldsymbol{\xi})$ is approximated by an (infinitesimally small) tangent vector

$$\overrightarrow{PP'} = d\boldsymbol{\xi} = d\xi^i \boldsymbol{e}_i. \tag{5.1}$$

See Fig. 5.2.

Mathematicians are not satisfied with this intuitive definition. They ask what the tangent vector along the coordinate curve ξ^i is. They define a tangent vector in terms of a differential operator on a function $f(\boldsymbol{\xi})$ in that direction. That is, they identify tangent vector $\boldsymbol{e}_i$ with the well-established partial derivative operator

$$\boldsymbol{e}_i \approx \partial_i = \frac{\partial}{\partial \xi^i}. \tag{5.2}$$

It operates on a differentiable function $f(\boldsymbol{\xi})$ and gives its derivative in the direction of coordinate curve ξ^i, that is, the partial derivative. Hence, one may write

$$\boldsymbol{e}_i f = \partial_i f(\boldsymbol{\xi}). \tag{5.3}$$

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_5

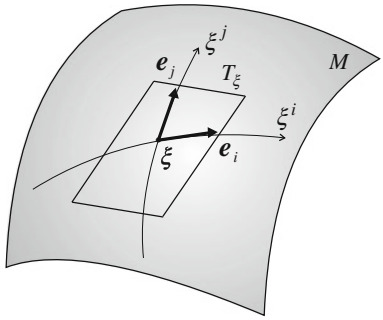


Fig. 5.1 Tangent space T_ξ and basis vectors $\boldsymbol{e}_i$

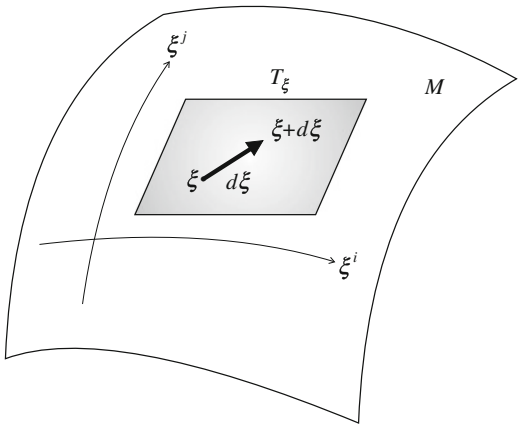


Fig. 5.2 Infinitesimal vector $d\boldsymbol{\xi}$ in T_ξ

A vector

$$A = A^i \boldsymbol{e}_i = A^i \partial_i \tag{5.4}$$

is the directional derivative operator which operates on f as

$$Af = A^i \partial_i f(\boldsymbol{\xi}). \tag{5.5}$$

When the coordinate system is changed from $\boldsymbol{\xi} = \left(\xi^i\right)$ to $\boldsymbol{\zeta} = (\zeta^\kappa)$, the partial derivatives change as follows:

$$\partial_i = J_i^\kappa \partial_\kappa, \quad \partial_\kappa = J_\kappa^i \partial_i, \tag{5.6}$$

where

$$J_i^\kappa = \frac{\partial \zeta^\kappa}{\partial \xi^i}, \quad J_\kappa^i = \frac{\partial \xi^i}{\partial \zeta^\kappa}. \tag{5.7}$$

Therefore, we have the law of transformation for the tangent vectors,

$$\boldsymbol{e}_\kappa = J^i_\kappa \boldsymbol{e}_i, \quad \boldsymbol{e}_i = J^\kappa_i \boldsymbol{e}_\kappa \ ; \tag{5.8}$$

$$\partial_\kappa = J^i_\kappa \partial_i, \quad \partial_i = J^\kappa_i \partial_\kappa. \tag{5.9}$$

For a manifold of probability distributions, we have another expression of a tangent vector. We identify $\boldsymbol{e}_i$ with the score function

$$\boldsymbol{e}_i \approx \partial_i \log p(x, \boldsymbol{\xi}), \tag{5.10}$$

which is a random variable because it is a function of x. Then, the tangent space T_ξ is a linear space spanned by n random variables $\partial_i \log p(x, \boldsymbol{\xi})$, $i = 1, \ldots, n$.

A tangent vector is a geometrical quantity, but it has various representations such as a differentiation operator and a random variable.

5.2 Riemannian Metric

When an inner product is defined in the tangent space T_ξ, we have a matrix $\mathbf{G} = \left(g_{ij}\right)$ consisting of the inner products of basis vectors

$$g_{ij}(\boldsymbol{\xi}) = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle. \tag{5.11}$$

It is a positive-definite matrix depending on $\boldsymbol{\xi}$. It is called the metric tensor and its components change to

$$g_{\kappa\lambda} = J^i_\kappa J^j_\lambda g_{ij} \tag{5.12}$$

by a coordinate transformation. (See Sect. 5.4 for the definition of a tensor.) A manifold is Riemannian when a metric tensor is defined.

For the manifold of probability distributions, we define an inner product by using the stochastic expression

$$\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = \mathrm{E}\left[\partial_i \log p(x, \boldsymbol{\xi}) \partial_j \log p(x, \boldsymbol{\xi})\right]. \tag{5.13}$$

This is the Fisher information matrix which is invariant.

The inner product of two vectors $A = A^i \boldsymbol{e}_i$ and $B = B^j \boldsymbol{e}_j$ is given by

$$\langle A, B \rangle = \langle A^i \boldsymbol{e}_i, B^j \boldsymbol{e}_j \rangle = A^{ij} g_{ij}. \tag{5.14}$$

A Riemannian manifold is Euclidean when there exists a coordinate system in which the metric tensor becomes

$$g_{ij}(\boldsymbol{\xi}) = \delta_{ij}. \tag{5.15}$$

A Riemannian manifold is curved from the metric point of view when it does not have a coordinate system satisfying (5.15). We will see later that a manifold is (locally) flat when and only when the Riemann–Christoffel curvature tensor vanishes. We need an affine connection to define the curvature tensor.

5.3 Affine Connection

Tangent space T_ξ is a local approximation of M at $\boldsymbol{\xi}$. However, a collection of T_ξ's at all $\boldsymbol{\xi}$ does not recover the entire figure of M without specifying how T_ξ and $T_{\xi'}$ ($\boldsymbol{\xi} \neq \boldsymbol{\xi}'$) are related. It is the role of an affine connection to establish a one-to-one mapping between T_ξ and $T_{\xi'}$, in particular when $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ are infinitesimally close. The entire figure of M will be recovered from the aggregate of T_ξ's by using an affine connection.

Let us consider two nearby tangent spaces T_ξ and $T_{\xi+d\xi}$. Let

$$X = X^i \boldsymbol{e}_i(\boldsymbol{\xi}) \in T_\xi, \tag{5.16}$$

$$\tilde{X} = \tilde{X}^i \boldsymbol{e}_i(\boldsymbol{\xi} + d\boldsymbol{\xi}) \in T_{\xi+d\xi} \tag{5.17}$$

be two tangent vectors belonging to T_ξ and $T_{\xi+d\xi}$, respectively. How different are they? We cannot compare them directly, because they belong to different tangent spaces. The basis vectors $\boldsymbol{e}_i = \boldsymbol{e}_i(\boldsymbol{\xi}) \in T_\xi$ and $\tilde{\boldsymbol{e}}_i = \boldsymbol{e}_i(\boldsymbol{\xi} + d\boldsymbol{\xi}) \in T_{\xi+d\xi}$ are different, so even when the components of X^i and $\tilde{X}^i$ are the same, we cannot say they are equal.

A manifold is a continuum, so T_ξ and $T_{\xi+d\xi}$ would be very similar, almost the same intuitively speaking, because the two tangent spaces become identical as $d\boldsymbol{\xi}$ tends to 0. We define a one-to-one affine correspondence between two nearby tangent spaces such that it becomes identical as $d\boldsymbol{\xi}$ tends to 0. As an example, consider a curved surface embedded in a three-dimensional Euclidean space. The tangent spaces at $\boldsymbol{\xi}$ and at $\boldsymbol{\xi} + d\boldsymbol{\xi}$ are slightly different in the three-dimensional space. We shift $T_{\xi+d\xi}$ in parallel such that the origins of T_ξ and $T_{\xi+d\xi}$ coincide in the three-dimensional space. However, the directions of $\boldsymbol{e}_i$ and $\tilde{\boldsymbol{e}}_i$ are slightly different when the surface is curved. We project the shifted $\tilde{\boldsymbol{e}}_i$ to T_ξ (Fig. 5.3) and let it be $\boldsymbol{e}'_i \in T_\xi$. The projected $\boldsymbol{e}'_i$ is the counterpart of $\tilde{\boldsymbol{e}}_i \in T_{\xi+d\xi}$ in T_ξ, so a correspondence between T_ξ and $T_{\xi+d\xi}$ is established by this projection. This is an example of affine connection.

We begin with technical expressions of an affine connection. Let us map the basis vector $\tilde{\boldsymbol{e}}_i$ of $T_{\xi+d\xi}$ to T_ξ, by which an affine correspondence is established. It is the projection in the ambient Euclidean space in the previous example, but we consider a more general situation. The map $\boldsymbol{e}'_i \in T_\xi$ of $\tilde{\boldsymbol{e}}_i \in T_{\xi+d\xi}$ is close to $\boldsymbol{e}_i(\boldsymbol{\xi})$, so it is represented as

$$\boldsymbol{e}'_i = \boldsymbol{e}_i + d\boldsymbol{e}_i. \tag{5.18}$$

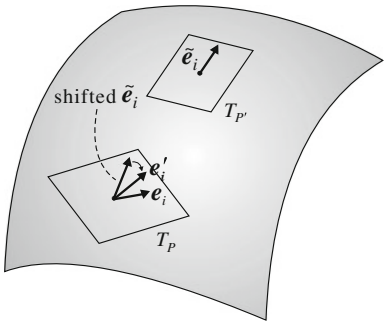


Fig. 5.3 Shift $\tilde{e}_i \in T_{\xi+d\xi}$ to P. The shifted $\tilde{e}_i$ does not belong to T_p. Project it to T_p, obtaining $e'_i \in T_p$ which is slightly different from $e_i \in T_p$

The difference de_i is a vector of T_ξ written in the component form as

$$de_i = \left(de_i^j\right) e_j. \tag{5.19}$$

The components de_i^j become 0 as $d\xi \to 0$. So they are linear in $d\xi$ and we put

$$de_i^j = \Gamma_{ki}^j(\xi) d\xi^k \tag{5.20}$$

as the first-order approximation, where the coefficient Γ_{ki}^j is a quantity having three indices.

A linear correspondence between T_ξ and $T_{\xi+d\xi}$ is established by giving a quantity $\Gamma = \left(\Gamma_{ki}^j\right)$ having three indices. They are called the coefficients of an affine connection which is to be established. The coefficients are given by the inner products of de_i and e_m as

$$\langle de_i, e_m \rangle = \Gamma_{ki}^j g_{jm} d\xi^k = \Gamma_{kim} d\xi^k, \tag{5.21}$$

where

$$\Gamma_{kim} = \Gamma_{ki}^j g_{mj} \tag{5.22}$$

is the covariant expression (lower indices expression) of Γ.

There still remains the problem of determining Γ_{kji}. The traditional way is to use the Riemannian metric g_{ij}. It is the Levi–Civita connection (Riemannian connection) introduced in Sect. 5.9. Another way is to use a divergence $D\left[\xi : \xi'\right]$ defined in M. This leads us to dually coupled affine connections, which we will see in the next chapter.

5.4 Tensors

A tensor is a quantity having a number of components such as $A = \left(A^i\right)$, $G = \left(g_{ij}\right)$ and $T = \left(T_{ijk}\right)$. A vector is a tensor having only one index. More precisely, a tensor is a quantity associated with tangent space T_ξ spanned by n tangent vectors $\{\boldsymbol{e}_i\}$. A vector A is represented in this basis as

$$A = A^i \boldsymbol{e}_i \tag{5.23}$$

in the component form, where the Einstein summation convention is working.

Let $\left\{\boldsymbol{e}^i\right\}$ be the dual basis, which is given by

$$\boldsymbol{e}^i = g^{ij}\boldsymbol{e}_j, \quad \boldsymbol{e}_j = g_{ji}\boldsymbol{e}^i \tag{5.24}$$

by using the metric tensor $\mathbf{G} = \left(g_{ij}\right)$ and its inverse $\mathbf{G}^{-1} = \left(g^{ij}\right)$. Note that the dual basis was denoted previously as $\boldsymbol{e}^{*i}$, but we hereafter omit $*$, because the upper index i of $\boldsymbol{e}^i$ shows that it is a dual basis vector. Vector A is represented in the dual basis as

$$A = A_i \boldsymbol{e}^i \tag{5.25}$$

so that we have

$$A_i = g_{ij}A^j. \tag{5.26}$$

A tensor $K = \left(K_{ij}{}^{klm}\right)$, for example, is a quantity represented, as

$$K = K_{ij}{}^{klm}\boldsymbol{e}^i\boldsymbol{e}^j\boldsymbol{e}_k\boldsymbol{e}_l\boldsymbol{e}_m \tag{5.27}$$

in the linear form of the product $\boldsymbol{e}^i\boldsymbol{e}^j\boldsymbol{e}_k\boldsymbol{e}_l\boldsymbol{e}_m$ of basis vectors. The product is formal and is just a concatenation of basis vectors. When an index is in the upper position, as in A^i, it is said to be contravariant, and when it is in the lower position, as in A_i, it is said to be covariant. A tensor may have contravariant and covariant components at the same time, as in $K_{ij}{}^{klm}$.

When another coordinate system $\boldsymbol{\zeta} = (\zeta^\kappa)$ is adopted, the basis vectors change by the coordinate transformation as in (5.8). Therefore, the component of a vector changes, as in

$$A^\kappa = J^\kappa_i A^i \tag{5.28}$$

for a contravariant (upper index) vector and

$$A_\kappa = J^i_\kappa A_i \tag{5.29}$$

for a covariant vector. Similarly, for a tensor like $K_{ij}{}^{klm}$, the components change as in

$$K_{\kappa\lambda}{}^{\mu\nu\tau} = J^i_\kappa J^j_\lambda J^\mu_k J^\nu_l J^\tau_m K_{ij}{}^{klm}. \tag{5.30}$$

For a scalar function $f(\boldsymbol{\xi})$, its gradient

$$\nabla f(\boldsymbol{\xi}) = (\partial_i f(\boldsymbol{\xi})), \quad \partial_i = \frac{\partial}{\partial \xi^i} \tag{5.31}$$

is a covariant vector, because of

$$\partial_\kappa f = J^i_\kappa \partial_i f, \quad \partial_\kappa = \frac{\partial}{\partial \zeta^\kappa}. \tag{5.32}$$

The Fisher information matrix (5.13) is a tensor. We define a quantity

$$T_{ijk} = \mathrm{E}\left[\partial_i l(x, \boldsymbol{\xi}) \partial_j l(x, \boldsymbol{\xi}) \partial_k l(x, \boldsymbol{\xi})\right]. \tag{5.33}$$

It is a covariant tensor having three indices and is symmetric. We call it a (statistical) cubic tensor for short. Two tensors G and T will play a fundamental role in the manifold of probability distributions.

Not all indexed quantities are tensors. For example, the second derivative of a scalar function f

$$f_{ij} = \partial_i \partial_j f(\boldsymbol{\xi}) \tag{5.34}$$

gives a quantity having two indices, but it is not a tensor. By changing the coordinate system from $\boldsymbol{\xi}$ to $\boldsymbol{\zeta}$, we have

$$f_{\kappa\lambda} = \partial_\kappa \partial_\lambda f = \partial_\kappa \left(J^j_\lambda \partial_j f\right) = J^i_\kappa J^j_\lambda f_{ij} + \left(\partial_\kappa J^j_\lambda\right) \partial_j f. \tag{5.35}$$

This shows that it is not a tensor. (It is a tensor at the critical point where $\partial_j f = 0$ holds.)

It should be noted that Γ is not a tensor. By changing the coordinate system from $\boldsymbol{\xi}$ to $\boldsymbol{\zeta}$, $d\boldsymbol{e}_i$ changes as in

$$d\boldsymbol{e}_\kappa = d\left(J^i_\kappa \boldsymbol{e}_i\right) = \left(\partial_\lambda J^i_\kappa\right) d\zeta^\lambda \boldsymbol{e}_i + J^i_\kappa d\boldsymbol{e}_i \tag{5.36}$$

in the new coordinate system. By using this relation, after some calculations, we have

$$\Gamma_{\kappa\lambda\mu} = J^i_\kappa J^j_\lambda J^k_\mu \Gamma_{ijk} + \left(\partial_\kappa J^j_\lambda\right) J^k_\mu g_{jk}. \tag{5.37}$$

So it is not a tensor. Note that, even if

$$\Gamma_{\kappa\lambda\mu}(\boldsymbol{\zeta}) = 0 \tag{5.38}$$

holds in one coordinate system, it does not mean that

$$\Gamma_{ijk}(\boldsymbol{\xi}) = 0 \tag{5.39}$$

in another coordinate system. Although it is not a tensor, it has its own meaning, representing the nature of the coordinate system. In a Euclidean space,

$$\Gamma_{ijk} = 0 \tag{5.40}$$

in an orthonormal coordinate system $\boldsymbol{\xi}$, but if we use the polar coordinate system $\boldsymbol{\zeta}$

$$\Gamma_{\kappa\lambda\mu} \neq 0, \tag{5.41}$$

because the tangent vector $\boldsymbol{e}_r$ in the radial direction changes depending on the position in the polar coordinate system.

When an equation is written in a tensor form such as

$$K_{ij}{}^{kl}(\boldsymbol{u}, \boldsymbol{v}, \ldots) = 0, \tag{5.42}$$

depending on physical quantities $\boldsymbol{u}, \boldsymbol{v} \ldots$, it has the same form in other coordinate systems

$$K_{\kappa\lambda}{}^{\mu\nu}(\boldsymbol{u}, \boldsymbol{v}, \ldots) = 0. \tag{5.43}$$

In this sense, a tensorial equation is invariant. A. Einstein obtained the equation of gravity in terms of tensors, because he believed that the law of nature should have the same form whichever coordinate system we use, and hence it should be written in a tensorial form.

5.5 Covariant Derivative

A vector field X is a vector-valued function on M, the value of which at $\boldsymbol{\xi}$ is given by a vector

$$X(\boldsymbol{\xi}) = X^i(\boldsymbol{\xi})\boldsymbol{e}_i(\boldsymbol{\xi}) \in T_{\boldsymbol{\xi}}. \tag{5.44}$$

When a vector field is given, it is possible to evaluate the intrinsic change of the vector as position $\boldsymbol{\xi}$ changes, by using the affine connection.

In order to see the intrinsic change between $X(\boldsymbol{\xi})$ and $X(\boldsymbol{\xi} + d\boldsymbol{\xi})$, since they belong to different tangent spaces, we need to map $X(\boldsymbol{\xi} + d\boldsymbol{\xi}) \in T_{\boldsymbol{\xi}+d\boldsymbol{\xi}}$ to $T_{\boldsymbol{\xi}}$ for comparison. Since the basis vector $\tilde{\boldsymbol{e}}_i = \boldsymbol{e}_i(\boldsymbol{\xi} + d\boldsymbol{\xi})$ is mapped to $T_{\boldsymbol{\xi}}$ by

$$\boldsymbol{e}'_i = \boldsymbol{e}_i(\boldsymbol{\xi}) + \Gamma^k_{ji}\boldsymbol{e}_k d\xi^j, \tag{5.45}$$

vector $X(\boldsymbol{\xi} + d\boldsymbol{\xi})$ is mapped to $T_{\boldsymbol{\xi}}$ as

$$\tilde{X} = X^k(\boldsymbol{\xi} + d\boldsymbol{\xi})\tilde{\boldsymbol{e}}_k = \left(X^k\boldsymbol{e}_k + \partial_j X^k \boldsymbol{e}_k d\xi^j + \Gamma^m_{jk} X^k \boldsymbol{e}_m d\xi^j\right), \tag{5.46}$$

where the Taylor expansion of $X^k(\boldsymbol{\xi} + d\boldsymbol{\xi})$ is used. Hence, their difference is evaluated as

$$\tilde{X} - X(\boldsymbol{\xi}) = \left(\partial_i X^k + \Gamma_{ij}^k X^j\right) d\xi^i \boldsymbol{e}_k. \tag{5.47}$$

This shows the intrinsic change of $X(\boldsymbol{\xi})$ as $\boldsymbol{\xi}$ changes by $d\boldsymbol{\xi}$. The rate of intrinsic change along the coordinate curve ξ^i is denoted as

$$\nabla_i X^k = \partial_i X^k + \Gamma_{ij}{}^k X^j. \tag{5.48}$$

This is called the covariant derivative of $X(\boldsymbol{\xi})$ and $\nabla_i X^k$ is a tensor.

Let $Y(\boldsymbol{\xi})$ be another vector field. Then, the directional covariant derivative of X along Y is denoted as

$$\nabla_Y X = Y^i \nabla_i X^k = Y^i \left(\partial_i X^k + \Gamma_{ij}{}^k X^j\right) \boldsymbol{e}_k. \tag{5.49}$$

This is the covariant derivative of X along Y. It is again a vector field.

We can define the covariant derivative of a tensor, e.g.,

$$K = K_k^{ij} \boldsymbol{e}_i \boldsymbol{e}_j \boldsymbol{e}^k \tag{5.50}$$

in a similar way, since it is spanned by multilinear vector products of the basis vectors $\boldsymbol{e}_i, \boldsymbol{e}_j, \boldsymbol{e}^k$.

For a scalar function $f(\boldsymbol{\xi})$, its change is measured by ordinary differentiation. Hence, vector field $Y(\boldsymbol{\xi})$ gives its directional derivative

$$Yf = Y^i \partial_i f. \tag{5.51}$$

Note that, for a vector field X, the partial derivatives of its components $\partial_i X^j(\boldsymbol{\xi})$ are not a tensor. We should use the covariant derivative for evaluating its intrinsic change.

5.6 Geodesic

A curve $\boldsymbol{\xi}(t)$ is called a geodesic when its direction does not change. So it is a generalization of the straight line. Here, a change in direction is measured by the covariant derivative derived from an affine connection. Note that this does not necessarily mean that it is a curve of minimal distance connecting two points, although this is the literal definition of a geodesic. The minimality and straightness can be different in a general manifold. It is possible to define an affine connection by using the metric such that a straight line (geodesic) has the minimality of distance, as is given in Theorem 5.2. But a divergence function gives a more general affine connection.

The tangent vector of curve $\boldsymbol{\xi}(t)$ at t is given by

$$\dot{\boldsymbol{\xi}}(t) = \dot{\xi}^i(t) \boldsymbol{e}_i(t), \tag{5.52}$$

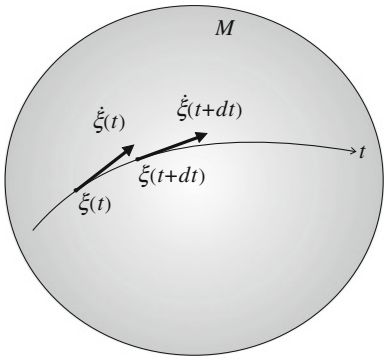


Fig. 5.4 Geodesic: $\dot{\boldsymbol{\xi}}(t+dt)$ corresponds to $\dot{\boldsymbol{\xi}}(t)$

where $\boldsymbol{e}_i(t) = \boldsymbol{e}_i\{\boldsymbol{\xi}(t)\}$ and $\cdot$ denotes the derivative d/dt. When $\boldsymbol{\xi}(t)$ is a geodesic, the tangent vector $\dot{\boldsymbol{\xi}}(t+dt) \in T_{\xi(t+dt)}$ corresponds to $\dot{\boldsymbol{\xi}}(t) \in T_{\xi(t)}$ by the affine connection. See Fig. 5.4. Since the change of the tangent direction of a curve is measured by the covariant derivative of $\dot{\boldsymbol{\xi}}$ along itself, the equation of the geodesic is

$$\nabla_{\dot{\xi}}\dot{\boldsymbol{\xi}} = 0. \tag{5.53}$$

This is given in the component form as

$$\ddot{\xi}^i(t) + \Gamma^i_{jk}\dot{\xi}^j(t)\dot{\xi}^k(t) = 0. \tag{5.54}$$

If we consider the equation

$$\nabla_{\dot{\xi}}\dot{\boldsymbol{\xi}} = c(t)\dot{\boldsymbol{\xi}}, \tag{5.55}$$

$\boldsymbol{\xi}(t)$ does not change the direction $\dot{\boldsymbol{\xi}}(t)$ of the curve, too, but its magnitude may change. However, by choosing the parameter t adequately, it is possible to reduce (5.55) to (5.54). Hence we consider only the case of (5.54).

5.7 Parallel Transport of Vector

We can transport a vector $A \in T_{\xi+d\xi}$ at $\boldsymbol{\xi} + d\boldsymbol{\xi}$ to T_ξ at $\boldsymbol{\xi}$ without changing it "intrinsically". The affine connection determines this parallel transport. For two distant points $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, we continue the process of parallel transport of a vector along a curve $\boldsymbol{\xi}(t)$ connecting $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$.

Define a vector field

$$A(t) = A^i\boldsymbol{e}_i(t) \tag{5.56}$$

along curve $\boldsymbol{\xi}(t)$ connecting P and Q (see Fig. 5.5). When its covariant derivative along the curve vanishes,

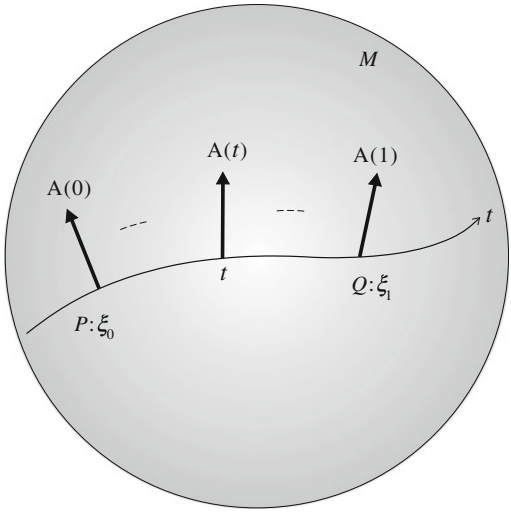


Fig. 5.5 Parallel transport of $A(0)$ at $\boldsymbol{\xi}_0$ to $A(1)$ at $\boldsymbol{\xi}_1$ along curve $\xi(t)$

$$\nabla_{\dot{\xi}} A(t) = 0, \tag{5.57}$$

$A(t)$ is intrinsically the same at any $\boldsymbol{\xi}(t)$. This is written in the component form as

$$\dot{A}^i(t) + \Gamma^i_{jk}(t) A^k(t) \dot{\xi}^j(t) = 0. \tag{5.58}$$

When $A(t)$ satisfies (5.57) or (5.58), we say that $A(0)$ at $T_{\xi(0)}$ is transported parallelly to $A(1)$ at $T_{\xi(1)}$. The parallel transport depends in general on the path along which it is transported. So we denote the parallel transport of a vector A from $\boldsymbol{\xi}_0 = \boldsymbol{\xi}(0)$ to $\boldsymbol{\xi}_1 = \boldsymbol{\xi}(1)$ along curve $c = \boldsymbol{\xi}(t)$ by

$$A(1) = \prod_{c}^{\boldsymbol{\xi}_1} A(0). \tag{5.59}$$

5.8 Riemann–Christoffel Curvature

A manifold is curved in general. When a vector is transported in parallel from one point to another, the resultant vector depends on the path along which it is transported. This never happens in a flat manifold. In order to show how curved a manifold is, we define the Riemann–Christoffel (RC) curvature tensor determined from the affine connection. One may skip this section, since we do not use RC curvature in applications in this monograph. When the RC curvature tensor vanishes, that is, when it is identically equal to 0, the manifold is (locally) flat. When it is flat, there exists an affine coordinate system such that each coordinate curve is a geodesic and its tangent vector coincides at any point by parallel transport.

5.8.1 Round-the-World Transport of Vector

Let us consider two curves $c_1 : \boldsymbol{\xi}_1(t)$ and $c_2 : \boldsymbol{\xi}_2(t)$, $0 \le t \le 1$, both connecting the same two points $\boldsymbol{\xi}_0 = \boldsymbol{\xi}_0(0)$ and $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_1(1)$. When a vector A at $\boldsymbol{\xi}_0$ is transported to $\boldsymbol{\xi}_1$ in parallel along curve c_1, it becomes $\prod_{c_1} A$. If we transport $\prod_{c_1} A$ back to $\boldsymbol{\xi}_0$ along the same curve c_1 in reverse, it is A. Now we transport A in parallel along the two curves c_1 and c_2. The resultant vectors, $\prod_{c_1} A$ and $\prod_{c_2} A$, are different in general (Fig. 5.6). This implies that when we transport a vector from $\boldsymbol{\xi}_0$ to $\boldsymbol{\xi}_1$ along path c_1 and then transport it back to the original point $\boldsymbol{\xi}_0$ along the other path c_2 in reverse, the resultant vector is different from A. So a vector changes when it is transported along a loop (consisting of path c_1 and reverse path of c_2). In other words, a vector is changed by a round-the-world trip.

The change may be used to measure how curved M is. To evaluate the change, we consider an infinitesimally small quadrangle connecting four points P, Q, R and S, where their coordinates are

$$P : \boldsymbol{\xi}, \quad Q : \boldsymbol{\xi} + d_1\boldsymbol{\xi}, \quad R : \boldsymbol{\xi} + d_1\boldsymbol{\xi} + d_2\boldsymbol{\xi}, \quad S : \boldsymbol{\xi} + d_2\boldsymbol{\xi}. \tag{5.60}$$

(See Fig. 5.7.) We transport A in parallel first from P to Q by $d_1\boldsymbol{\xi}$. Then, A becomes $A_1 = A + d_1A$, the components of which are

$$d_1A^i = -\Gamma^i_{jk} A^k d_1\xi^j. \tag{5.61}$$

We further transport A_1 from Q to R along the path $\overrightarrow{QR} = d_2\boldsymbol{\xi}$. Then, the transported vector at R is $A_{12} = A + d_1A + \delta_{12}A$, where the components of additional change $\delta_{12}A$ are

$$\delta_{12}A^i = -\Gamma^i_{jk}(\boldsymbol{\xi} + d_1\boldsymbol{\xi})\left(A^k + d_1A^k\right) d_2\xi^j. \tag{5.62}$$

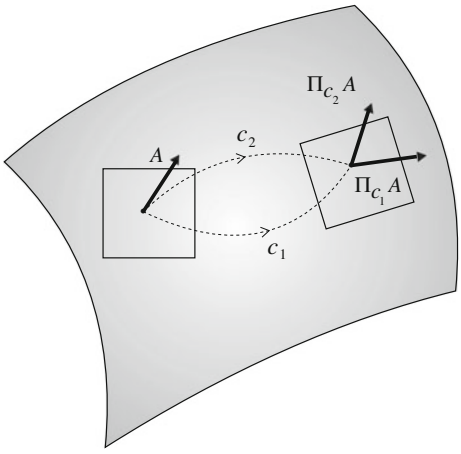


Fig. 5.6 Parallel transport of A via c_1 is different from that via c_2

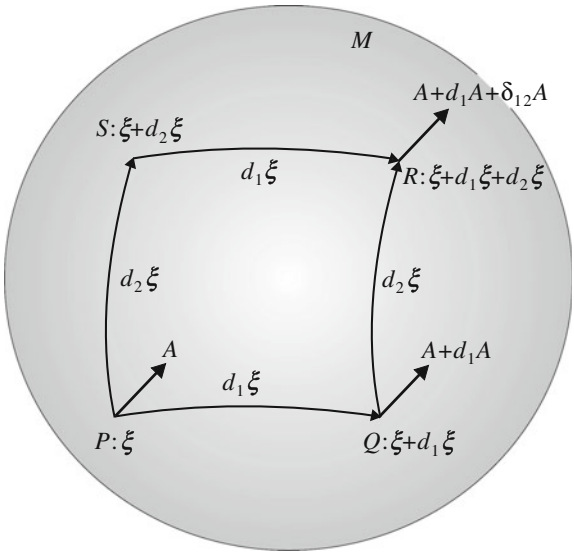


Fig. 5.7 Parallel transports of A along PQR and PSR

Since Γ is evaluated at $\boldsymbol{\xi} + d_1\boldsymbol{\xi}$, the Taylor expansion gives

$$\delta_{12}A^i = -\Gamma^i_{jk} d_2\xi^j A^k - \partial_l \Gamma^i_{jk} d_1\xi^l d_2\xi^j A^k + \Gamma^i_{jk}\Gamma^k_{lm} d_2\xi^j d_1\xi^l A^m. \tag{5.63}$$

Now, we transport A along the different route, first along the path $\overrightarrow{PS} = d_2\boldsymbol{\xi}$ to S and then to R along $\overrightarrow{SR} = d_1\boldsymbol{\xi}$. The resultant change is given by exchanging $d_1\boldsymbol{\xi}$ and $d_2\boldsymbol{\xi}$ in (5.63). The result is

$$\delta_{21}A^i = -\Gamma^i_{jk} d_1\xi^j A^k - \partial_l \Gamma^i_{jk} d_2\xi^l d_1\xi^j A^k + \Gamma^i_{jk}\Gamma^k_{lm} d_2\xi^j d_1\xi^l A^m. \tag{5.64}$$

How different are the resultant vectors? By subtracting A_{21} from A_{12} where (5.63) and (5.64) are used, the result is written as

$$A_{21} - A_{12} = \delta A^i = R_{jkl}{}^i A^l \left(d_1\xi^j d_2\xi^k - d_1\xi^k d_2\xi^j\right), \tag{5.65}$$

where we put

$$R_{ijk}{}^l = \partial_i \Gamma_{jk}{}^l - \partial_j \Gamma_{ik}{}^l + \Gamma_{im}{}^l \Gamma_{jk}{}^m - \Gamma_{jm}{}^l \Gamma_{ik}{}^m. \tag{5.66}$$

We can prove that $R_{ijk}{}^l$ is a tensor. It is called the Riemann–Christoffel (RC) curvature tensor. This shows how a vector changes by the round-the-world trip along an infinitesimal loop. We denote the infinitesimal loop encircling P, Q, R, S and P by a tensor

$$df^{jk} = \left(d_1\xi^j d_2\xi^k - d_1\xi^k d_2\xi^j\right), \tag{5.67}$$

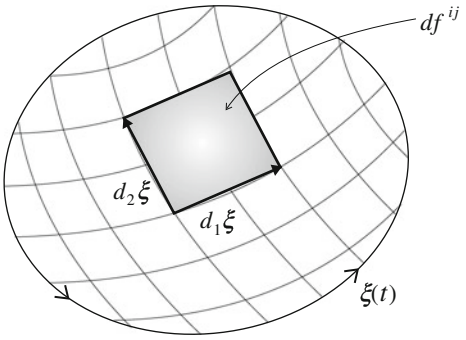


Fig. 5.8 Small surface element, loop and membrane

which is antisymmetric with respect to the two indices i and j,

$$df^{ij} = -df^{ji}. \tag{5.68}$$

This is a small surface element, representing a surface spanned by $d_1\boldsymbol{\xi}$ and $d_2\boldsymbol{\xi}$ (Fig. 5.8). Equation (5.65) is written as

$$\delta A^i = R_{jkl}{}^i A^l df^{jk}. \tag{5.69}$$

When vector A is transported in parallel along a general loop $\boldsymbol{\xi}(t), 0 \leq t \leq 1$, $\boldsymbol{\xi}(0) = \boldsymbol{\xi}(1)$, we span a membrane encircled by the loop (see Fig. 5.8). Then, the changed ΔA due to the round-the-world parallel transportation is given by the surface integral

$$\Delta A^i = \int R_{jkl}{}^i A^l df^{jk}. \tag{5.70}$$

This does not depend on the way of spanning the membrane, as is clear from the Stokes' Theorem.

5.8.2 *Covariant Derivative and RC Curvature*

The partial derivative is always commutative,

$$\partial_i \partial_j = \partial_j \partial_i. \tag{5.71}$$

However, this does not in general hold for the covariant derivative,

$$\nabla_i \nabla_j - \nabla_j \nabla_i \neq 0. \tag{5.72}$$

The covariant derivative of $\boldsymbol{e}_j$ in the direction of basis vector $\boldsymbol{e}_i$ is

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{e}_j = \Gamma_{ij}^{\ k}\boldsymbol{e}_k. \tag{5.73}$$

By using this, we have

$$\left(\nabla_{\boldsymbol{e}_i}\nabla_{\boldsymbol{e}_j} - \nabla_{\boldsymbol{e}_j}\nabla_{\boldsymbol{e}_i}\right)X(\boldsymbol{\xi}) = R_{ijk}^{\ \ l}X^k\boldsymbol{e}_l. \tag{5.74}$$

We omit the proof, but we see that the RC curvature shows the degree of non-commutativity of the covariant derivative.

In general, we can define the RC curvature by

$$R(X, Y)Z = \nabla_X\left(\nabla_Y Z\right) - \nabla_Y\left(\nabla_X Z\right) - \nabla_{[X,Y]}Z, \tag{5.75}$$

where

$$[X, Y] = XY - YX = \left(X^j\partial_j Y^i - Y^j\partial_j X^i\right)\boldsymbol{e}_i. \tag{5.76}$$

This is a sophisticated definition of the RC curvature tensor which one sees in modern textbooks on differential geometry. However, it is difficult to understand the meaning of the RC curvature from it.

5.8.3 *Flat Manifold*

When the RC curvature vanishes, M is said to be flat. The parallel transport of a vector does not depend on the path in this case. Let us consider a set of basis vectors $\{\boldsymbol{e}_i\}$ in the tangent space at a point. We construct n geodesics passing through the point in the directions of $\boldsymbol{e}_i$. We then have n coordinate curves θ^i, of which tangent vectors $\boldsymbol{e}_i$ are the same everywhere. This generates a flat coordinate system $\boldsymbol{\theta} = \left(\theta^i\right)$. Indeed, we transport the tangent vectors $\boldsymbol{e}_i$ to any point and compose the geodesics the directions of which are $\boldsymbol{e}_i$. Vectors $\boldsymbol{e}_i$ are parallel at any point and we have a net of coordinate curves $\boldsymbol{\theta}$.

Since the tangent vectors of a coordinate curve θ^i are all in parallel, we have

$$\nabla_{\boldsymbol{e}_j}\boldsymbol{e}_i = 0. \tag{5.77}$$

Therefore, from (5.73) we have

$$\Gamma_{jik} = 0. \tag{5.78}$$

Hence, when M is flat, we have an affine coordinate system consisting of geodesics in which (5.78) holds at any $\boldsymbol{\xi}$.

5.9 Levi–Civita (Riemannian) Connection

We have so far treated the metric and affine connection separately. However, it is possible to define an affine connection such that it is essentially related to the metric, giving a unified picture. This is Riemannian geometry. It requires that the magnitude of a vector does not change by the parallel transport. This establishes a relation between the metric and the affine connection (see Fig. 5.9).

It is easy to see the equivalence of the following two propositions of parallel transportation: (1) The magnitude of a vector does not change by parallel transportation,

$$\langle A, A\rangle_{\xi_0} = \left\langle \prod A, \prod A \right\rangle_{\xi_1}, \text{ for any } A. \tag{5.79}$$

(2) The inner product of two vectors does not change by parallel transportation,

$$\langle A, B\rangle_{\xi_0} = \left\langle \prod A, \prod B \right\rangle_{\xi_1}, \text{ for any } A, B. \tag{5.80}$$

We consider an infinitesimal parallel transport of two basis vectors along the coordinate curve ξ^i. Then, when the length of a vector does not change, we have

$$g_{ij}(\boldsymbol{\xi} + d\boldsymbol{\xi}) = \langle \boldsymbol{e}_i(\boldsymbol{\xi} + d\boldsymbol{\xi}), \boldsymbol{e}_j(\boldsymbol{\xi} + d\boldsymbol{\xi})\rangle_{\xi + d\xi} = \langle \boldsymbol{e}_i(\boldsymbol{\xi}) + d\boldsymbol{e}_i, \boldsymbol{e}_j(\boldsymbol{\xi}) + d\boldsymbol{e}_j\rangle_{\xi}. \tag{5.81}$$

Because $g_{ij}(\boldsymbol{\xi} + d\boldsymbol{\xi}) = g_{ij}(\boldsymbol{\xi}) + \partial_k g_{ij} d\xi^k$, this condition is written as

$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma_{kji}. \tag{5.82}$$

More generally, this condition is equivalent to

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle, \tag{5.83}$$

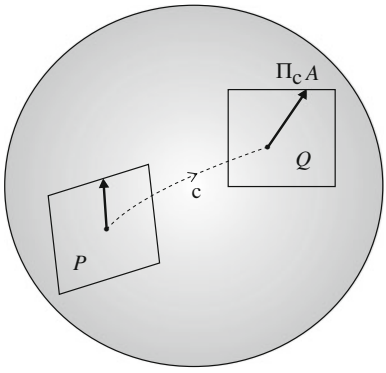


Fig. 5.9 The magnitude of A is equal to the magnitude of $\prod_c A$

for three vector fields X, Y and Z. When an affine connection satisfies this condition, it is said to be metric. The metric affine connection is uniquely determined from metric g_{ij}, provided the symmetric condition

$$\Gamma_{ijk} = \Gamma_{jik} \tag{5.84}$$

holds.

Theorem 5.1 *When the parallel transport does not change the magnitude of a vector, there is a unique symmetric affine connection given by*

$$\Gamma_{ijk}(\boldsymbol{\xi}) = \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\right). \tag{5.85}$$

It is an interesting exercise to prove this from (5.82). It is called the Levi–Civita connection or Riemannian connection. Many conventional textbooks on differential geometry study only the Levi–Civita connection. By using the Levi–Civita connection, we have the following convenient property.

Theorem 5.2 *A curve that connects two points by a minimal distance is a geodesic under the Levi–Civita connection, where the length of a curve $c = \boldsymbol{\xi}(t)$ connecting $\boldsymbol{\xi}(0)$ and $\boldsymbol{\xi}(1)$ is given by*

$$s = \int_0^1 \sqrt{g_{ij}(t)\dot{\boldsymbol{\xi}}^i(t)\dot{\boldsymbol{\xi}}^j(t)}dt. \tag{5.86}$$

It is possible to obtain (5.54) and (5.85) by the variational method, $\delta s = 0$, of minimizing (5.86) with respect to curve $\boldsymbol{\xi}(t)$. This is also a good exercise. See Fig. 5.10.

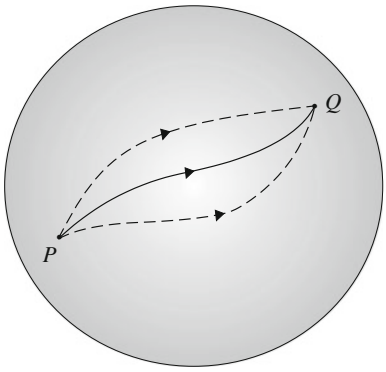


Fig. 5.10 A Riemannian geodesic $\xi(t)$ is a curve which does not change the direction $\dot{\xi}(t)$, and also the distance s is minimized along it

5.10 Submanifold and Embedding Curvature

We consider a submanifold embedded in a larger manifold. It has embedding curvature when it is curved in the ambient manifold. This is different from the RC curvature. It is useful to embed a manifold in a simple (e.g. flat) higher-dimensional ambient manifold and study its properties in the ambient manifold. Geometrical quantities are transferred from a simple ambient manifold to the submanifold by embedding.

5.10.1 Submanifold

Let S be a submanifold embedded in M (Fig. 5.11). Let $\boldsymbol{\xi} = \left(\xi^i\right)$ be a coordinate system of M, $i = 1, \ldots, n$ and $\boldsymbol{u} = (u^a)$ be a coordinate system of S, $a = 1, \ldots, m$, where we assume $n > m$. Since a point $\boldsymbol{u}$ in S is also a point in the ambient M, its coordinates in M are specified by $\boldsymbol{u}$ as

$$\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{u}). \tag{5.87}$$

We consider the case that $\boldsymbol{\xi}(\boldsymbol{u})$ is differentiable with respect to $\boldsymbol{u}$.

The tangent vector $\boldsymbol{e}_a$ along the coordinate curve u^a of S is

$$\boldsymbol{e}_a = \partial_a \tag{5.88}$$

and the tangent space $T_{\boldsymbol{u}}^S$ of S is spanned by them (Fig. 5.12). However, they are regarded as tangent vectors at point $\boldsymbol{\xi}(\boldsymbol{u})$ of M by embedding. They are represented in M as

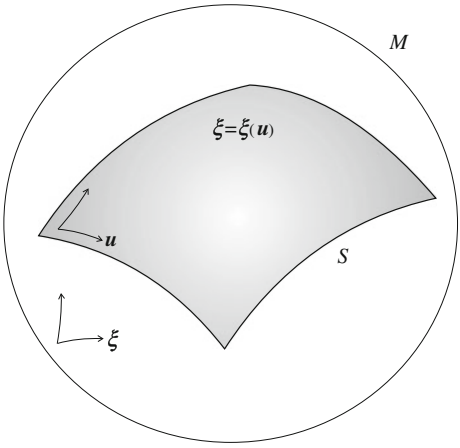


Fig. 5.11 Submanifold S embedded in M

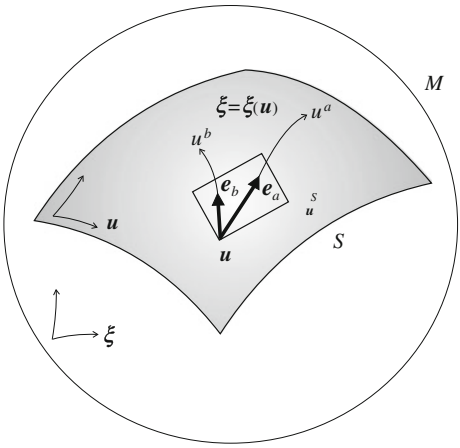


Fig. 5.12 Tangent vectors $\boldsymbol{e}_a$ of submanifold

$$\boldsymbol{e}_a = \frac{\partial \xi^i}{\partial u^a}\boldsymbol{e}_i \qquad \left(\partial_a = \frac{\partial \xi^i}{\partial u^a}\partial_i\right) \tag{5.89}$$

in terms of the basis vectors $\boldsymbol{e}_i \in T_{\boldsymbol{\xi}}$ of M. Defining

$$B_a^i = \frac{\partial \xi^i}{\partial u^a}, \tag{5.90}$$

we have

$$\boldsymbol{e}_a = B_a^i \boldsymbol{e}_i. \tag{5.91}$$

A vector $X = X^a \boldsymbol{e}_a \in T_{\boldsymbol{u}}^S$ is a vector $X = X^i \boldsymbol{e}_i \in T_{\boldsymbol{\xi}}$ and

$$X^i = B_a^i X^a. \tag{5.92}$$

Submanifold S inherits the geometrical structures of M. The magnitude of a tangent vector A in $T_{\boldsymbol{u}}^S$ is given by its magnitude in M. Hence, the metric

$$g_{ab} = \langle \boldsymbol{e}_a, \boldsymbol{e}_b \rangle \tag{5.93}$$

in S is given by

$$g_{ab} = B_a^i B_b^j g_{ij}. \tag{5.94}$$

5.10.2 Embedding Curvature

An affine connection is naturally transferred to S from M. Let $\tilde{\boldsymbol{e}}_a = \boldsymbol{e}_a(\boldsymbol{u} + d\boldsymbol{u}) \in T_{\boldsymbol{u}+d\boldsymbol{u}}^S$ be a basis vector at $\boldsymbol{u} + d\boldsymbol{u}$ of S and we map it in parallel to $T_{\boldsymbol{u}}^S$. We first

transport it in M from $\boldsymbol{\xi}(\boldsymbol{u})$ to $\boldsymbol{\xi}(\boldsymbol{u}+d\boldsymbol{u})$ in parallel. The resultant vector is denoted as $\boldsymbol{e}'_a = \boldsymbol{e}_a + d\boldsymbol{e}_a \in T_{\boldsymbol{\xi}(\boldsymbol{u})}$, where $d\boldsymbol{e}_a$ is given by the covariant derivative of $\boldsymbol{e}_a$ in the direction of $\boldsymbol{e}_b = B^i_b \boldsymbol{e}_i$ in M,

$$d\boldsymbol{e}_a = \nabla_{\boldsymbol{e}_b} \boldsymbol{e}_a du^b. \tag{5.95}$$

We calculate $\nabla_{\boldsymbol{e}_a} \boldsymbol{e}_b$ in M,

$$\begin{aligned}\nabla_{\boldsymbol{e}_a} \boldsymbol{e}_b &= B^i_a \nabla_{\boldsymbol{e}_i} \left(B^j_b \boldsymbol{e}_j \right) = B^i_a \partial_i B^k_b \boldsymbol{e}_k + B^i_a B^j_b \nabla_{\boldsymbol{e}_i} \boldsymbol{e}_j \\ &= \left(B^i_a \partial_i B^k_b + B^i_a B^j_b \Gamma_{ij}{}^k \right) \boldsymbol{e}_k \\ &= \Gamma_{ab}{}^k \boldsymbol{e}_k, \end{aligned} \tag{5.96}$$

where we put

$$\Gamma_{ab}{}^k = B^i_a \partial_i B^k_b + B^i_a B^j_b \Gamma_{ij}{}^k. \tag{5.97}$$

Here, the vector $d\boldsymbol{e}_a$ is not necessarily included in the tangent space of S (Fig. 5.13). So we decompose it in the tangent direction of S and its orthogonal direction,

$$d\boldsymbol{e}_a = d\boldsymbol{e}^{\parallel}_a + d\boldsymbol{e}^{\perp}_a, \tag{5.98}$$

where $d\boldsymbol{e}^{\parallel}_a \in T^S_{\boldsymbol{u}}$ and $d\boldsymbol{e}^{\perp}_a$ is orthogonal to S. We define the parallel transport of $\tilde{\boldsymbol{e}}_a$ within S by the change $d\boldsymbol{e}^{\parallel}_a$, neglecting the change in the orthogonal direction:

$$\tilde{\boldsymbol{e}}_a = \boldsymbol{e}_a + d\boldsymbol{e}^{||}_a. \tag{5.99}$$

Rewriting $d\boldsymbol{e}^{\parallel}_a$ in terms of basis vectors $\{\boldsymbol{e}_b\}$, the induced affine connection of S is given as

$$\Gamma_{abc}(\boldsymbol{u}) = B^i_a B^j_b B^k_c \Gamma_{ijk} + B^j_c \partial_a B^i_b g_{ij}. \tag{5.100}$$

The orthogonal direction of $d\boldsymbol{e}^{\perp}_a$ represents how S is curved in M. To show the orthogonal component, we supplement $T^S_{\boldsymbol{u}}$ with $n-m$ orthogonal vectors $\boldsymbol{e}_\kappa$, $\kappa = n-m+1, \ldots, n$, such that the entire vectors $\{\boldsymbol{e}_a, \boldsymbol{e}_\kappa\}$ span the tangent space of M. Then, the orthogonal part is given by

$$\delta\boldsymbol{e}^{\perp}_a = H_{ab}{}^{\kappa} \boldsymbol{e}_k du^b, \tag{5.101}$$

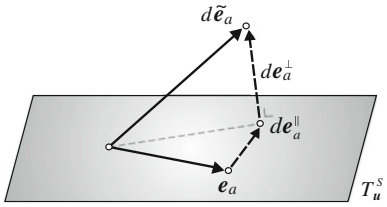


Fig. 5.13 Decomposition of $d\tilde{\boldsymbol{e}}_a \in T_{\xi(u)}$ in the orthogonal part $d\boldsymbol{e}^{\perp}_a$ and parallel part $d\boldsymbol{e}^{||}_a$ with respect to $T^S_{\boldsymbol{u}}$

where we use the covariant derivative in M to define

$$H_{ba\kappa} = \langle \nabla_{\boldsymbol{e}_b} \boldsymbol{e}_a, \boldsymbol{e}_\kappa \rangle. \tag{5.102}$$

This is the embedding curvature of S, sometimes called the Euler–Schouten curvature tensor.

The embedding curvature is different from the RC curvature, which is derived from the affine connection Γ_{abc}. The RC curvature is the intrinsic curvature considering only inside S. As a simple example, let us consider a cylinder S embedded in a three-dimensional Euclidean manifold M. It is curved in M, so it has non-zero embedding curvature. But its RC curvature vanishes and Euclidean geometry holds inside S. If we live in S and do not know the outer world of three dimensions, we enjoy Euclidean geometry in S where the RC curvature is 0. But S has non-zero embedding curvature.

Remarks

Differential geometry studies properties of a manifold by its local structure. A Riemannian manifold is a typical example, where a manifold is equipped with a metric tensor $\mathbf{G} = \left(g_{ij}\right)$ by which the distance of two nearby points is measured. It is locally approximated by a Euclidean space but is curved in general. Modern differential geometry further studies the global topological structure of a manifold. It is interesting to see how the global structure is restricted by the local structure such as the curvature. This is an important perspective. However, we have not mentioned the global properties, because most (though not all) applications use only local structure.

Since differential geometry has been developed as pure mathematics, mathematicians have constructed a rigorous, sophisticated theory, excluding intuitive definitions of geometrical concepts. However, once such a rigorous theory is established, we may use intuitive understanding for applications. Part II is an attempt to introduce the modern concepts of differential geometry without tears to beginners.

After non-Euclidean geometry was developed in the 19th century, people came to know that there existed non-Euclidean spaces. B. Riemann, in his professorship lecture, proposed the concept of Riemannian geometry, which is non-Euclidean and curved. He conjectured that the real world might be Riemannian on a cosmological scale or on an atomic scale. His view was proved true in the 20th century in relativity theory and elementary particle theory.

There are many of applications of differential geometry. Relativity theory is one of them, in which Einstein introduced the concept of a torsion tensor to establish a unified theory (unification of gravity and electromagnetism). Unfortunately, this interesting idea failed. But the torsion tensor survived in mathematics. The torsion tensor is a third-order tensor of which the first two indices are anti-symmetric. This is a new quantity to supplement the Riemannian structure of $\{M, G\}$, although we have not mentioned it here.

A Riemannian manifold with torsion plays a fundamental role in continuum mechanics including dislocations. A dislocation field in a continuum is identified with a torsion field, and a rich theory has been established. See, e.g., Amari [1962, 1968]. Another application is the dynamics of electro-mechanical systems, such as

motors and generators, by Gabriel Kron. Here, non-holonomic constraints play a role, and are converted to torsion. Recent robotics also uses non-holonomic constraints. Differential geometry plays important roles in various areas.

Information geometry also uses differential geometry, where the invariance criterion plays a fundamental role in defining the geometrical structure of a manifold of probability distributions. However, the conventional edifice of differential geometry in textbooks is not enough to explore its structure. We need a new concept of duality of affine connections with respect to the Riemannian metric. In the next chapter, we study a Riemannian manifold equipped with dually coupled affine connections.

Chapter 6
Dual Affine Connections and Dually Flat Manifold

We have considered one affine connection, namely the Levi–Civita connection, in a Riemannian manifold M. However, we can establish a new edifice of differential geometry, by treating a pair of affine connections which are dually coupled with respect to the Riemannian metric. Such a structure has not been described in conventional textbooks, but is the heart of information geometry. Mathematically speaking, in addition to the Riemannian structure $\{M, G\}$, we study the structure $\{M, G, T\}$, which has a third-order symmetric tensor T in addition to G. As an important special case, we study a dually flat Riemannian manifold. It may be regarded as a dualistic extension of the Euclidean space. The generalized Pythagorean theorem and projection theorem hold in such a manifold. They are particularly useful in applications.

6.1 Dual Connections

The Levi–Civita connection is the only metric affine connection (without torsion) that preserves the metric by parallel transport. However, there is another way of preserving the metric by using two affine connections. We consider here two symmetric affine connections Γ and Γ^* and denote the associated parallel transports as $\prod$ and $\prod^*$, respectively. These affine connections are dually coupled when the parallel transports of vectors A and B, one by $\prod$ and the other by $\prod^*$, do not change their inner product,

$$\langle A, B\rangle = \left\langle \prod A, \prod^* B \right\rangle. \tag{6.1}$$

See Fig. 6.1. Such a pair of affine connections are said to be dually coupled with respect to the Riemannian metric, which defines the inner product. A pair of connections collaborate to preserve the inner product by parallel transportation of vectors.

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_6

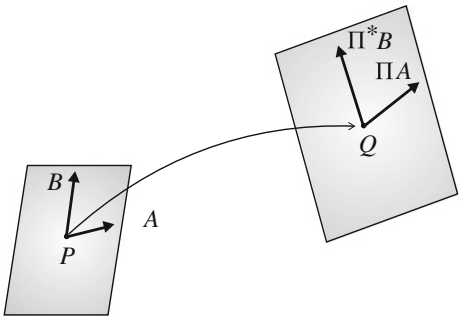


Fig. 6.1 Conservation of inner product by dual parallel transports

When the two connections are identical, (6.1) reduces to the metric condition (5.80), so that this is an extension of the metric connection.

We search for analytical expressions of dual connections. Consider two basis vectors $\tilde{\boldsymbol{e}}_i$ and $\tilde{\boldsymbol{e}}_j$ at point $\boldsymbol{\xi} + d\boldsymbol{\xi}$. We transport them to $\boldsymbol{\xi}$, one by using affine connection Γ and the other by dual connection Γ^*. Then, their parallel transports are, respectively,

$$\boldsymbol{e}_i + d\boldsymbol{e}_i = \boldsymbol{e}_i + \Gamma_{ki}{}^j \boldsymbol{e}_j d\xi^k, \tag{6.2}$$

$$\boldsymbol{e}_j + d^*\boldsymbol{e}_j = \boldsymbol{e}_j + \Gamma^{*\,i}_{kj} \boldsymbol{e}_i d\xi^k, \tag{6.3}$$

where d and d^* denote the changes induced by the parallel transformations due to Γ and Γ^*, respectively. From the conservation of the inner product

$$\langle \tilde{\boldsymbol{e}}_i, \tilde{\boldsymbol{e}}_j \rangle_{\xi + d\xi} = \langle \boldsymbol{e}_i + d\boldsymbol{e}_i,\ \boldsymbol{e}_j + d^*\boldsymbol{e}_j \rangle_{\xi}, \tag{6.4}$$

we have

$$g_{ij}(\boldsymbol{\xi} + d\boldsymbol{\xi}) = g_{ij}(\boldsymbol{\xi}) + \langle d\boldsymbol{e}_i,\ \boldsymbol{e}_j \rangle_{\xi} + \langle \boldsymbol{e}_i, d^*\boldsymbol{e}_j \rangle_{\xi}, \tag{6.5}$$

where higher-order terms are neglected.

By the Taylor expansion, we have the componentwise expression

$$\partial_i g_{jk} = \Gamma_{ijk} + \Gamma^*_{ikj}. \tag{6.6}$$

Compare this with the self-dual case (5.82).

This is rewritten in terms of the covariant derivatives as

$$Z\langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla^*_Z Y \rangle, \tag{6.7}$$

where X, Y and Z are three vector fields. This is confirmed by using three vector fields $Z = \boldsymbol{e}_i$, $X = \boldsymbol{e}_j$ and $Y = \boldsymbol{e}_k$, as

$$\boldsymbol{e}_i \langle \boldsymbol{e}_j, \boldsymbol{e}_k \rangle = \langle \nabla_{\boldsymbol{e}_i} \boldsymbol{e}_j, \boldsymbol{e}_k \rangle + \langle \boldsymbol{e}_j, \nabla^*_{\boldsymbol{e}_i} \boldsymbol{e}_k \rangle, \tag{6.8}$$

which is the same as (6.6).

The average of two dual connections is given by

$$\Gamma^0_{ijk} = \frac{1}{2}\left(\Gamma_{ijk} + \Gamma^*_{ijk}\right). \tag{6.9}$$

The related covariant derivative is

$$\nabla^{(0)} = \frac{1}{2}\left(\nabla + \nabla^*\right). \tag{6.10}$$

From (6.7), we see that ∇^0 satisfies (5.83) and Γ^0_{ijk} is the Levi–Civita connection. Let us define

$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk}. \tag{6.11}$$

Then, the dual connections are written as

$$\Gamma_{ijk} = \Gamma^0_{ijk} - \frac{1}{2}T_{ijk}, \quad \Gamma^*_{ijk} = \Gamma^0_{ijk} + \frac{1}{2}T_{ijk}. \tag{6.12}$$

Theorem 6.1 *When Γ and Γ^* are dual affine connections, T is a symmetric tensor given by*

$$\nabla_i g_{jk} = T_{ijk}, \tag{6.13}$$
$$\nabla^{*i} g^{jk} = -T^{ijk}. \tag{6.14}$$

Proof We calculate the covariant derivative of tensor $G = \left(g_{ij}\right)$. It is given by

$$\nabla_i g_{jk} = \partial_i g_{jk} - \Gamma_{ik}{}^m g_{mj} - \Gamma_{ij}{}^m g_{mk}. \tag{6.15}$$

Since ∇^0 is the metric connection,

$$\nabla^0_i g_{jk} = \partial_i g_{jk} - \Gamma^0_{ikj} - \Gamma^0_{ijk} = 0. \tag{6.16}$$

Hence, we have

$$\nabla_i g_{jk} = \frac{1}{2}\left(T_{ijk} + T_{ikj}\right). \tag{6.17}$$

Since g_{jk} is symmetric with respect to j and k, we have

$$\nabla_i g_{jk} = T_{ijk}. \tag{6.18}$$

Moreover, Γ_{ijk} is a symmetric connection, so T_{ijk} is symmetric with respect to i and j. Hence T_{ijk} is symmetric with respect to three indices, i, j, and k. T_{ijk} is a tensor, because it is the covariant derivative of tensor g_{jk}. (6.14) is also derived similarly. □

Remark S. Lauritzen called T_{ijk} the skewness tensor. However, it is symmetric, and so we hesitate to use the term "skewness", which often implies anti-symmetry. So we use the term cubic tensor. This is called the Amari–Chentsov tensor by some researchers (e.g., Ay et al. 2013), since Chentsov defined it and Amari has developed its theory. The triplet $\{M, G, T\}$ is also called the Amari–Chentsov structure.

Dual affine connections are determined from $\{G, T\}$ by (6.12), where g_{ij} is a positive-definite symmetric matrix and T_{ijk} is a cubic tensor. When $T_{ijk} = 0$, the two affine connections are identical. Hence, the connection is self-dual and M reduces to the ordinary Riemannian manifold, having the Levi–Civita connection.

6.2 Metric and Cubic Tensor Derived from Divergence

When a divergence $D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]$ is defined in M, we show that two tensors g_{ij}^D and T_{ijk}^D are automatically induced from it. We consider a neighborhood of diagonal position $\boldsymbol{\xi} = \boldsymbol{\xi}'$ of D. Since D has two arguments, we introduce the following notation of differentiation at the diagonal:

$$D_i = \frac{\partial}{\partial \xi^i} D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]_{\xi'=\xi}, \tag{6.19}$$

$$D_{;i} = \frac{\partial}{\partial \xi'^i} D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]_{\xi'=\xi}. \tag{6.20}$$

Similarly, for multiple differentiation, we use the notation

$$D_{ij;k} = \frac{\partial^2}{\partial \xi^i \partial \xi^j} \frac{\partial}{\partial \xi'^k} D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]_{\xi'=\xi}, \tag{6.21}$$

etc.

We define the following quantities by using the above notations:

$$g_{ij}^D = -D_{i;j}, \tag{6.22}$$

$$\Gamma_{ijk}^D = -D_{ij;k}, \tag{6.23}$$

$$\Gamma_{ijk}^{D^*} = -D_{k;ij}. \tag{6.24}$$

We can prove that Γ^D and Γ^{D^*} define affine connections, by checking how they are transformed by coordinate transformations. We omit the calculations since they are technical and easy. Their difference

$$T_{ijk}^D = \Gamma_{ijk}^{D^*} - \Gamma_{ijk}^D \tag{6.25}$$

is a third-order symmetric tensor. Hence, we have two characteristic tensors g_{ij}^D and T_{ijk}^D from a divergence D. The following is a key result connecting a divergence and dual geometry, derived by Eguchi (1983).

Theorem 6.2 *The two affine connections Γ^D and Γ^{D^*} are dual with respect to the Riemannian metric g^D.*

Proof By differentiating

$$g_{ij}^D(\boldsymbol{\xi}) = -\frac{\partial^2}{\partial \xi^i \partial \xi'^j} D[\boldsymbol{\xi} : \boldsymbol{\xi}] \tag{6.26}$$

with respect to $\boldsymbol{\xi}$, we have

$$\partial_k g_{ij}^D(\boldsymbol{\xi}) = -D_{ki;j} - D_{i;jk} = \Gamma_{kij}^D + \Gamma_{kji}^{D^*}. \tag{6.27}$$

This satisfies (6.6) so that Γ^D and Γ^{D^*} are dual affine connections. □

When a Legendre pair of convex functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ are given, where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are connected by the Legendre transformation, we have a Bregman divergence

$$D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi(\boldsymbol{\theta}) + \varphi\left(\boldsymbol{\eta}'\right) - \boldsymbol{\theta} \cdot \boldsymbol{\eta}', \tag{6.28}$$

where $\boldsymbol{\eta}'$ is the Legendre dual of $\boldsymbol{\theta}'$. The metric tensor derived from it is

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}) \tag{6.29}$$

in the $\boldsymbol{\theta}$-coordinates and

$$g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta}) \tag{6.30}$$

in the $\boldsymbol{\eta}$-coordinates. Moreover, by differentiating it, we have from (6.23) and (6.24),

$$\Gamma_{ijk}(\boldsymbol{\theta}) = \Gamma^{*ijk}(\boldsymbol{\eta}) = 0 \tag{6.31}$$

in the two coordinate systems. This implies that the geometry derived from a convex function, or the related Bregman divergence, is dually flat and the affine coordinate systems are $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. The cubic tensor is written as

$$T_{ijk} = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}), \quad T^{ijk} = \partial^i \partial^j \partial^k \varphi(\boldsymbol{\eta}) \tag{6.32}$$

in the two coordinate systems. This justifies our former definition of the dually flat structure introduced in Part I without differential geometry.

6.3 Invariant Metric and Cubic Tensor

An f-divergence is an invariant divergence in the manifold of probability distributions. We calculate the two tensors G^f and T^f derived from an f-divergence, which are therefore invariant.

Theorem 6.3 *Invariant tensors derived from a standard f-divergence in the manifold of probability distributions are given as*

$$g_{ij}^f = g_{ij}, \tag{6.33}$$

$$T_{ijk}^f = \alpha T_{ijk}, \tag{6.34}$$

where g_{ij} is the Fisher information matrix and

$$T_{ijk} = E\left[\partial_i l(x,\boldsymbol{\xi})\partial_j l(x,\boldsymbol{\xi})\partial_k l(x,\boldsymbol{\xi})\right], \tag{6.35}$$

$$\alpha = 2f'''(1) + 3. \tag{6.36}$$

Proof By differentiating an arbitrary f-divergence

$$D_f\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] = \int p(x,\boldsymbol{\xi}) f\left\{\frac{p\left(x,\boldsymbol{\xi}'\right)}{p(x,\boldsymbol{\xi})}\right\} dx, \tag{6.37}$$

with respect to $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ and putting $\boldsymbol{\xi}' = \boldsymbol{\xi}$, we have (6.33) and (6.34). □

Remark The uniqueness of the f-divergence under the invariance criterion is derived from the information monotonicity and decomposability. More strongly, the Chentsov theorem proves that g_{ij} and αT_{ijk} are the unique invariant second-order and third-order symmetric tensors in S_n.

6.4 α-Geometry

When T_{ijk} is a symmetric tensor, so is αT_{ijk} for real α. We call the two affine connections derived from $\{G, \alpha T\}$,

$$\Gamma_{ijk}^{\alpha} = \Gamma_{ijk}^{0} - \frac{\alpha}{2} T_{ijk}, \quad \Gamma_{ijk}^{-\alpha} = \Gamma_{ijk}^{0} + \frac{\alpha}{2} T_{ijk}, \tag{6.38}$$

the α-connection and $-\alpha$-connection, respectively.

Theorem 6.4 *Γ^{α} and $\Gamma^{-\alpha}$ are dually coupled and the $\alpha = 0$ connection Γ^0 is the Levi–Civita connection, which is self-dual.*

The proof is easy from (6.12).

When T^f is derived from an f-divergence, it is αT for α satisfying (6.36) and, moreover, $-\alpha T$ is derived from the dual of the f-divergence. The derived dual structure is the only invariant geometry in the case of a manifold of probability distributions. We call it the α-geometry. The α-geometry is derived from the α-divergence defined in (3.39). It is not dually flat in general. When $\alpha = \pm 1$, it reduces to the KL-divergence, giving a dually flat structure.

For any convex function ψ, we can construct a related α-divergence. In this case, the α-geometry is induced from the α-divergence defined by

$$D_{\psi}^{(\alpha)}\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \frac{4}{1-\alpha^2}\left\{\frac{1-\alpha}{2}\psi(\boldsymbol{\theta}) + \frac{1+\alpha}{2}\psi\left(\boldsymbol{\theta}'\right) - \psi\left(\frac{1-\alpha}{2}\boldsymbol{\theta} + \frac{1+\alpha}{2}\boldsymbol{\theta}'\right)\right\}. \tag{6.39}$$

This is a Jensen-type divergence introduced by Zhang (2004).

6.5 Dually Flat Manifold

We have the following theorem concerning dual curvatures.

Theorem 6.5 *When the RC curvature R vanishes with respect to one affine connection, the RC curvature R^* with respect to the dual connection vanishes and vice versa.*

Proof When the RC curvature vanishes, $R = 0$, the round-the-world parallel transportation does not change any A:

$$A = \prod A. \tag{6.40}$$

For vector transportations, we always have

$$\langle A, B\rangle = \left\langle \prod A, \prod{}^{*} B \right\rangle. \tag{6.41}$$

Hence, when (6.40) holds, we have

$$\langle A, B\rangle = \left\langle A, \prod{}^{*} B \right\rangle \tag{6.42}$$

for any A and B. This implies

$$B = \prod{}^{*} B \tag{6.43}$$

showing that the dual RC curvature vanishes, $R^* = 0$. □

A manifold is always dually flat when it is flat with respect to one connection. When M is dually flat, there exists an affine coordinate system $\boldsymbol{\theta}$ for which

$$\Gamma_{ijk}(\boldsymbol{\theta}) = 0. \tag{6.44}$$

Each coordinate curve θ^i is a geodesic. The basis vectors $\{\boldsymbol{e}_i\}$ are transported in parallel to any position, not depending on a path of transportation.

Similarly, there exists a dual affine coordinate system $\boldsymbol{\eta}$ for which

$$\Gamma^{*ijk}(\boldsymbol{\eta}) = 0 \tag{6.45}$$

holds. Each coordinate curve η_i is a dual geodesic. Let its direction be $\boldsymbol{e}^i$. Here, we use a lower index to denote the components of $\boldsymbol{\eta} = (\eta_i)$ and the related basis vectors are denoted by upper-indexed quantities such as $\boldsymbol{e}^i$. This notation fits our index notation of raising and lowering indices by using the metric tensors g_{ij} and its inverse g^{ij}. The Jacobians of the coordinate transformations satisfy

$$g_{ij} = \frac{\partial \eta_i}{\partial \theta^j}, \quad g^{ij} = \frac{\partial \theta^i}{\partial \eta_j}. \tag{6.46}$$

Therefore, two bases $\{\boldsymbol{e}_i\}$ and $\left\{\boldsymbol{e}^i\right\}$ satisfy

$$\boldsymbol{e}_i = g_{ij}\boldsymbol{e}^j, \quad \boldsymbol{e}^j = g^{ji}\boldsymbol{e}_i. \tag{6.47}$$

Theorem 6.6 *In a dually flat manifold, there exists affine coordinate system $\boldsymbol{\theta}$ and dual affine coordinate system $\boldsymbol{\eta}$ such that their tangent vectors are reciprocally orthogonal,*

$$\langle \boldsymbol{e}_i, \boldsymbol{e}^j \rangle = \langle \partial_i, \partial^j \rangle = \delta_i^j. \tag{6.48}$$

Proof From (6.47), we have

$$\langle \boldsymbol{e}_i, \boldsymbol{e}^j \rangle = \langle \boldsymbol{e}_i, g^{jk}\boldsymbol{e}_k \rangle = g_{ik}g^{jk} = \delta_i^j. \tag{6.49}$$

We also have

$$\left\langle \prod \boldsymbol{e}_i, \prod^* \boldsymbol{e}^j \right\rangle = \langle \boldsymbol{e}_i, \boldsymbol{e}^j \rangle = \delta_i^j \tag{6.50}$$

at any point. Note that g_{ij} and g^{ji} depend on the position, but (6.47) holds at any point. □

6.6 Canonical Divergence in Dually Flat Manifold

We have shown that a dual structure is constructed from a divergence function. In particular, a dually flat structure is induced from a Bregman divergence. However, many divergences give the same dual structure. This is because the differential geometry of the metric and connections is defined from the derivatives of divergence $D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]$

at $\boldsymbol{\xi} = \boldsymbol{\xi}'$, given in (6.22)–(6.24). That is, it depends only on the values of $D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]$ for infinitesimally close $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$. There are no unique ways of extending an infinitesimally defined divergence to the entire M. That is, $D[\boldsymbol{\xi} : \boldsymbol{\xi}] + d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right)$ gives the same geometry as $D\left[\boldsymbol{\xi}, \boldsymbol{\xi}'\right]$ when a non-negative function $d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right)$ satisfies

$$d(\boldsymbol{\xi}, \boldsymbol{\xi}) = 0, \tag{6.51}$$

$$\partial_i d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right)_{|\boldsymbol{\xi}} = \partial'_i d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right)_{|\boldsymbol{\xi}=\boldsymbol{\xi}'} = 0 \tag{6.52}$$

$$\partial_i \partial_j d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right)_{|\boldsymbol{\xi}} = \partial'_i \partial'_j d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right)_{|\boldsymbol{\xi}=\boldsymbol{\xi}'} = 0, \tag{6.53}$$

$$\partial_i \partial_j \partial'_k d(\boldsymbol{\xi}, \boldsymbol{\xi}')_{|\boldsymbol{\xi}=\boldsymbol{\xi}'} = \partial'_i \partial'_j \partial_k d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right)_{|\boldsymbol{\xi}=\boldsymbol{\xi}'} = 0, \tag{6.54}$$

where $\partial_i = \partial/\partial\xi^i$ and $\partial'_i = \partial/\partial\xi'^i$. $d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right) = \left\{D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right]\right\}^2$ given in (3.25) is such an example. Interestingly, however, when a manifold is dually flat, we can obtain a unique canonical divergence, despite the fact that there are many locally equivalent divergences. To show this, we begin with the following lemma.

Lemma 6.1 *When M is dually flat, there are a pair of dual affine coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ and of Legendre-dual convex functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ satisfying*

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \theta^i \eta_i = 0, \tag{6.55}$$

such that the metric is given by

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}), \quad g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta}) \tag{6.56}$$

and the cubic tensor by

$$T_{ijk}(\boldsymbol{\theta}) = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}), \tag{6.57}$$

$$T^{ijk}(\boldsymbol{\eta}) = \partial^i \partial^j \partial^k \varphi(\boldsymbol{\eta}). \tag{6.58}$$

Proof By using the affine coordinate system $\boldsymbol{\theta}$ for which $\Gamma_{ijk}(\boldsymbol{\theta}) = 0$, (6.6) reduces to

$$\partial_i g_{jk} = \Gamma^*_{ikj}. \tag{6.59}$$

Because the connections are symmetric, we have

$$\partial_i g_{jk} = \partial_k g_{ji}. \tag{6.60}$$

We fix index j and denote it by $\cdot$. So we have

$$\partial_i g_{k\cdot} = \partial_k g_{i\cdot}. \tag{6.61}$$

Then, there is a function $\psi_\cdot$ satisfying

$$g_{i\cdot} = \partial_i \psi_\cdot, \tag{6.62}$$

or for each j, we have

$$g_{ij} = \partial_i \psi_j. \tag{6.63}$$

Since g_{ij} is symmetric, we have

$$\partial_i \psi_j - \partial_j \psi_i = 0. \tag{6.64}$$

This guarantees the existence of a scalar function ψ such that

$$\psi_j = \partial_j \psi. \tag{6.65}$$

Hence

$$g_{ij} = \partial_i \partial_j \psi, \tag{6.66}$$

where $\psi(\boldsymbol{\theta})$ is convex because g_{ij} is positive-definite. Since $\nabla_i = \partial_i$ for the $\boldsymbol{\theta}$-coordinates, T_{ijk} is given from (6.18) by

$$T_{ijk} = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}). \tag{6.67}$$

By using the dual affine coordinate system, we have a convex function $\varphi(\boldsymbol{\eta})$ that satisfies (6.56) and (6.58). It is easy to see that the two coordinate systems are connected by a Legendre transformation, so that the two functions are the Legendre duals. □

Theorem 6.7 *When M is dually flat, there exists a Lengdre pair of convex functions $\psi(\boldsymbol{\theta})$, $\varphi(\boldsymbol{\eta})$ and a canonical divergence given by the Bregman divergence*

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi(\boldsymbol{\theta}) + \varphi\left(\boldsymbol{\eta}'\right) - \boldsymbol{\theta} \cdot \boldsymbol{\eta}'. \tag{6.68}$$

They are uniquely determined except for affine transformations. Conversely, the canonical divergence gives the original dually flat Riemannian structure.

Theorem 6.8 *The KL-divergence is the canonical divergence of an exponential family of probability distributions which is invariant and dually flat.*

Remark 1 Many studies begin with the KL-divergence given a priori without any justification. However, our theory shows that the KL-divergence is an outcome of dual flatness in the invariant geometry.

Remark 2 The KL-divergence is derived as the unique canonical divergence without assuming decomposability in the above theorem. See also another derivation by Jiao et al. (2015).

For a dually flat M, its affine coordinates $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are not unique. Any affine transformation

$$\tilde{\boldsymbol{\theta}} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{b}, \quad \tilde{\boldsymbol{\eta}} = \mathbf{A}^{-1}\boldsymbol{\eta} + \boldsymbol{c}, \tag{6.69}$$

where $\mathbf{A}$ is an invertible matrix and $\boldsymbol{b}$, $\boldsymbol{c}$ are constants, gives a set of dually coupled coordinate systems. The convex functions $\psi(\boldsymbol{\theta})$ are not unique either, because we may add a linear term, as in

$$\tilde{\psi}(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}) + \boldsymbol{a}\boldsymbol{\theta} + d, \tag{6.70}$$

where $\boldsymbol{a}$ is a vector and d is a scalar. However, the canonical divergence

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi\left(\boldsymbol{\theta}\right) + \varphi\left(\boldsymbol{\eta}'\right) - \boldsymbol{\theta} \cdot \boldsymbol{\eta}' \tag{6.71}$$

is uniquely determined, not depending on a specific choice of affine coordinate systems.

6.7 Canonical Divergence in General Manifold of Dual Connections

It is known that there always exists a divergence in a manifold having dual connections, such that the same dual structure is given by the divergence (Matumoto 1993). There are many such divergences. So it is an interesting problem to define a canonical divergence, if possible, in a manifold M having non-flat dual connections. When M is dually flat, we have a canonical divergence. Kurose (1994) showed that a canonical divergence called the geometrical divergence exists when M is 1-conformally flat. M is embedded in $\boldsymbol{R}^{n+1}$ in this case. Moreover, when it has constant curvature, the generalized Pythagorean theorem (Theorem 4.5) holds. The α-divergence is a canonical divergence of S_n in this sense. Taking these facts into account, we demonstrate an on-going trial by N. Ay and S. Amari to define a canonical divergence in the general case, briefly without proof (Ay and Amari 2015, Henmi and Kobayashi 2000).

Consider $\{M, g, \nabla, \nabla^*\}$ of a Riemannian manifold with dual affine connections. Let $\boldsymbol{\xi}$ be a coordinate system. Given a point $\boldsymbol{\xi}_p$ and a tangent vector X belonging to the tangent space at $\boldsymbol{\xi}_p$, we have a geodesic curve $\boldsymbol{\xi}(t)$,

$$\nabla_{\dot{\boldsymbol{\xi}}}\dot{\boldsymbol{\xi}}(t) = 0, \tag{6.72}$$

passing through $\boldsymbol{\xi}_p$ and its tangent direction is X,

$$\boldsymbol{\xi}(0) = \boldsymbol{\xi}_p, \ \dot{\boldsymbol{\xi}}(0) = X. \tag{6.73}$$

When the geodesic reaches point $\boldsymbol{\xi}_q$ as t increases from 0 to 1,

$$\boldsymbol{\xi}_q = \boldsymbol{\xi}(1), \tag{6.74}$$

$\boldsymbol{\xi}_q$ is called the exponential map of X,

$$\boldsymbol{\xi}_q = \exp_{\boldsymbol{\xi}_p}(X). \tag{6.75}$$

Given $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_q$, we have the inverse of the exponential map,

$$X\left(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q\right) = \exp^{-1}_{\boldsymbol{\xi}_p}\left(\boldsymbol{\xi}_q\right) \tag{6.76}$$

in a neighborhood of $\boldsymbol{\xi}_p$.

We now define a canonical divergence in a general manifold of dual connections. We first define a divergence between $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_q$ by

$$\tilde{D}\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right] = \int_0^1 t g_{ij}\left\{\boldsymbol{\xi}(t)\right\} \dot{\xi}^i(t)\dot{\xi}^j(t)dt, \tag{6.77}$$

where $\boldsymbol{\xi}(t)$ is the primal geodesic connecting $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_q$. It can be rewritten as

$$\begin{aligned}\tilde{D}\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right] &= \int_0^1 t\langle\dot{\boldsymbol{\xi}}(t), \dot{\boldsymbol{\xi}}(t)\rangle dt \\ &= \int_0^1 -\langle\exp^{-1}_{\boldsymbol{\xi}(t)}\left(\boldsymbol{\xi}_p\right), \dot{\boldsymbol{\xi}}(t)\rangle dt.\end{aligned} \tag{6.78}$$

We then define another divergence by using the dual geodesic $\boldsymbol{\xi}^*(t)$ connecting $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_q$:

$$\tilde{D}^*\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right] = \int_0^1 (1-t) g_{ij}\left\{\boldsymbol{\xi}^*(t)\right\} \dot{\xi}^{*i}(t)\dot{\xi}^{*j}(t)dt. \tag{6.79}$$

A canonical divergence is defined by the arithmetic mean of the above two.

Definition A canonical divergence is given by

$$D\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right] = \frac{1}{2}\left(\tilde{D}\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right] + \tilde{D}^*\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right]\right). \tag{6.80}$$

Theorem 6.9 *The geometrical structure derived from the canonical divergence (6.80) coincides with the original geometry. When M is dually flat, it gives the canonical divergence of the Bregman type. When M is Riemannian ($T = 0$), it is a half of the squared Riemannian distance.*

In a dually flat manifold, the projection theorem holds: Given $\boldsymbol{\xi}_p$ and a submanifold $S \subset M$, the point $\hat{\boldsymbol{\xi}}_p$ that minimizes $D\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right]$, $\boldsymbol{\xi}_q \in S$, is the geodesic projection

of $\boldsymbol{\xi}_p$ to S such that the geodesic connecting $\boldsymbol{\xi}_p$ and $\hat{\boldsymbol{\xi}}_p$ is orthogonal to S at $\hat{\boldsymbol{\xi}}_p$. The projection theorem does not hold in general, but we have the following theorem.

Theorem 6.10 *The canonical divergence satisfies the projection theorem when*

$$X^i\left(\boldsymbol{\xi}_q, \boldsymbol{\xi}_p\right) \propto -g^{ij}\left(\boldsymbol{\xi}_q\right) \partial_j' D\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}_q\right], \tag{6.81}$$

where X^i *is the contravariant component of* $X = \exp_{\boldsymbol{\xi}_q}^{-1}\left(\boldsymbol{\xi}_p\right)$ *and*

$$\partial_j' = \frac{\partial}{\partial \xi_q^j}. \tag{6.82}$$

Proof Consider a divergence ball centered at $\boldsymbol{\xi}_p$ and with radius $c \geq 0$,

$$B_c = \left\{\boldsymbol{\xi} \,\middle|\, D\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}\right] = c\right\}. \tag{6.83}$$

Let S be a smooth submanifold of M. Let $\hat{\boldsymbol{\xi}}_p$ be the minimizer of $D\left[\boldsymbol{\xi}_p : \boldsymbol{\xi}\right], \boldsymbol{\xi} \in S$. When c is increasing from 0, the ball B_c touches S at $\hat{\boldsymbol{\xi}}_p$. The tangent hypersurfaces of S and B_c are the same at this point, and its normal vector is given by

$$n^i = g^{ij}\left(\hat{\boldsymbol{\xi}}_p\right) \partial_j' D\left[\boldsymbol{\xi}_p : \hat{\boldsymbol{\xi}}_p\right]. \tag{6.84}$$

The tangent direction of the geodesic connecting $\boldsymbol{\xi}_p$ and $\hat{\boldsymbol{\xi}}_p$ is given by

$$\dot{\boldsymbol{\xi}}\left(\hat{\boldsymbol{\xi}}_p\right) = X\left(\hat{\boldsymbol{\xi}}_p, \boldsymbol{\xi}_p\right). \tag{6.85}$$

So the projection theorem holds when the above two share the same direction. □

It is interesting to study when the geodesic projection theorem (6.81) holds. It holds in the dually flat case. It holds for the α-divergence, and hence it is the canonical divergence of the α-geometry.

6.8 Dual Foliations of Flat Manifold and Mixed Coordinates

A dually flat manifold admits two types of foliations, e-foliation and m-foliation, which are orthogonal to each other. This structure is useful for separating two quantities, one represented in the e-coordinates and the other in the m-coordinates. This fits particularly well for analyzing a hierarchical system (Amari 2001).

6.8.1 k-cut of Dual Coordinate Systems: Mixed Coordinates and Foliation

Let M be a dually flat manifold with dually coupled affine coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. We partition the coordinates into two parts, one consisting of k components and the other consisting of $n-k$ components. We rearrange the suffixes such that the first k components consist of $\theta^1, \ldots \theta^k$ and the last $n-k$ components consist of $\theta^{k+1}, \ldots, \theta^n$. The same rearrangement is done for the $\boldsymbol{\eta}$-coordinates. We call such a partition a k-cut.

Let us compose a new coordinate system $\boldsymbol{\xi}$ of which the first k components are the corresponding $\boldsymbol{\eta}$-coordinates and the last $n-k$ components are $\boldsymbol{\theta}$-coordinates such as

$$\boldsymbol{\xi} = \left(\eta_1, \ldots, \eta_k \,;\, \theta^{k+1}, \ldots, \theta^n\right). \tag{6.86}$$

This is a new coordinate system called a mixed coordinate system, since m-affine coordinates and e-affine coordinates are mixed in it. It is not an affine coordinate system by itself. The basis vectors of the tangent space in the mixed coordinates are composed of two parts, the first part consisting of

$$\boldsymbol{e}^i = \frac{\partial}{\partial \eta_i}, \quad i = 1, \ldots, k \tag{6.87}$$

and the second part consisting of

$$\boldsymbol{e}_j = \frac{\partial}{\partial \theta^j}, \quad j = k+1, \ldots, n. \tag{6.88}$$

They are orthogonal, because of

$$\langle \boldsymbol{e}^i, \boldsymbol{e}_j \rangle = 0, \quad i \neq j. \tag{6.89}$$

Therefore, the Riemannian metric in this coordinate system has a block-diagonal form,

$$\mathbf{G} = \left[\begin{array}{c|c} g^{ij} & 0 \\ \hline 0 & g_{lm} \end{array}\right]. \tag{6.90}$$

Let us consider an $(n-k)$-dimensional submanifold obtained by fixing the first k coordinates ($\boldsymbol{\eta}$-coordinates) to be equal to $\boldsymbol{c} = (c_1, \ldots, c_k)$

$$\eta_i = c_i, \quad i = 1, \ldots, k \tag{6.91}$$

and denote it by $M^*(\boldsymbol{c})$, in which $\theta^{k+1}, \ldots, \theta^n$ run freely. It is an m-flat submanifold, because it is defined by linear constraints on $\boldsymbol{\eta}$-coordinates. For $\boldsymbol{c} \neq \boldsymbol{c}'$, $M^*(\boldsymbol{c})$ and $M^*(\boldsymbol{c}')$ do not intersect,

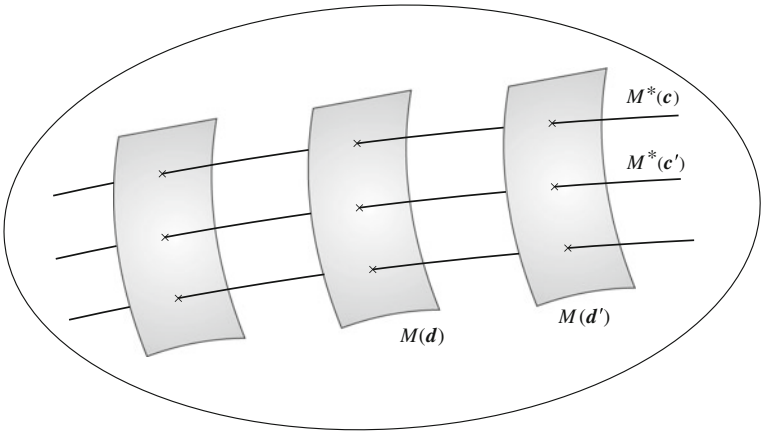


Fig. 6.2 Dual orthogonal foliation of manifold

$$M^*(\boldsymbol{c}) \bigcap M^*\left(\boldsymbol{c}'\right) = \phi. \tag{6.92}$$

Moreover, the entire M is covered by the aggregate of all $M^*(\boldsymbol{c})$'s

$$\bigcup_{\boldsymbol{c}} M^*(\boldsymbol{c}) = M. \tag{6.93}$$

Hence, $M^*(\boldsymbol{c})$'s give a partition of M. Such a partition is called a foliation.

Dually to the above, we fix the second part of the mixed coordinates ($\boldsymbol{\theta}$-coordinates),

$$\theta^j = d_j, \quad j = k+1, \ldots, n, \tag{6.94}$$

where $\boldsymbol{d} = (d_{k+1}, \ldots, d_n)$ and $\eta_1, \ldots, \eta_k$ run freely. We then have a k-dimensional e-flat submanifold denoted by $M(\boldsymbol{d})$. Moreover, $M(\boldsymbol{d})$'s form another foliation of M. We thus have two foliations. Moreover, $M(\boldsymbol{d})$ and $M^*(\boldsymbol{c})$ are orthogonal to each other for any $\boldsymbol{c}$ and $\boldsymbol{d}$. See Fig. 6.2.

Theorem 6.11 *A dually flat M admits a pair of orthogonal k-cut foliations for any k, one of which is m-flat and the other e-flat.*

6.8.2 Decomposition of Canonical Divergence

By using the mixed coordinates, the canonical divergence between two points P and Q can be decomposed into a sum of two divergences, one representing the difference in the first part and the other in the second part. Let

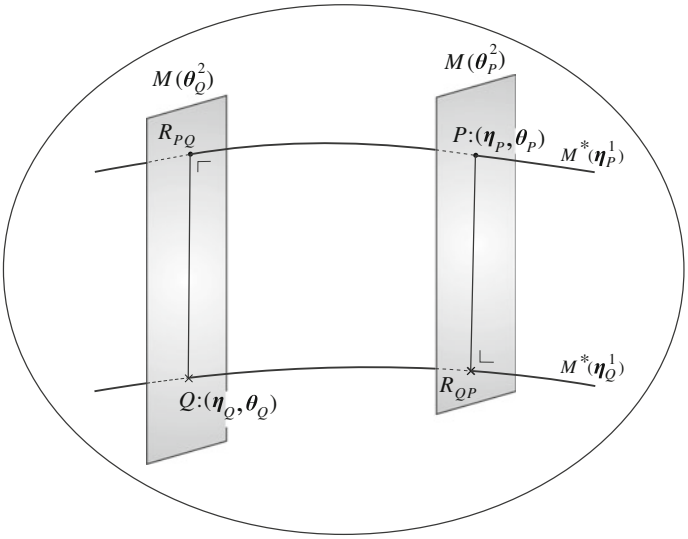


Fig. 6.3 Foliation and decomposition of KL-divergence

$$\xi_P = \left(\eta_P \,;\, \theta_P\right), \quad \xi_Q = \left(\eta_Q \,;\, \theta_Q\right) \tag{6.95}$$

be the mixed coordinates of the two points P and Q. P is located at the intersection of $M^*\left(\eta_P\right)$ and $M\left(\theta_P\right)$ and Q is at the intersection of $M^*\left(\eta_Q\right)$ and $M\left(\theta_Q\right)$. We m-project P to $M\left(\theta_Q\right)$ and let the projected point be R_{PQ}. We also e-project P to $M^*\left(\eta_Q\right)$ and let the projected point be R_{QP}. See Fig. 6.3. Since the m-geodesic connecting P and R_{PQ} is orthogonal to the e-geodesic connecting R_{PQ} and Q, $PR_{PQ}Q$ forms a right triangle so that the Pythagorean theorem is applicable. We can do the same thing for the triangle $PR_{QP}Q$. Then, we have the decomposition theorem.

Theorem 6.12 *The canonical divergence $D[P : Q]$ is decomposed as*

$$D[P : Q] = D\left[P : R_{PQ}\right] + D\left[R_{PQ} : Q\right], \tag{6.96}$$

$D\left[P : R_{PQ}\right]$ representing the difference in the first part and $D\left[R_{PQ} : Q\right]$ representing the difference in the second part.

6.8.3 A Simple Illustrative Example: Neural Firing

We show the usefulness of the orthogonal foliation by a simple example. Let us consider a network consisting of two neurons which emit spikes stochastically. Let x_1 and x_2 be two binary random variables, taking values $x_i = 1, i = 1, 2$, when neuron i is excited (emitting a spike) and 0 otherwise. Joint probability $p\left(x_1, x_2\right)$ specifies the

stochastic behavior of this network. The manifold of all joint probability distributions $M = \{p(x_1, x_2)\}$ forms a three-dimensional exponential family, because

$$p(1,1) + p(1,0) + p(0,1) + p(0,0) = 1. \tag{6.97}$$

This is a set of discrete distributions over four elements, and we can write it in the exponential form,

$$p(x_1, x_2) = \exp\left\{\sum_{i=1}^{2} \theta^i x_i + \theta^{12} x_1 x_2 - \psi(\boldsymbol{\theta})\right\}. \tag{6.98}$$

The affine coordinates are given by

$$\boldsymbol{\theta} = \left(\theta^1, \theta^2, \theta^{12}\right). \tag{6.99}$$

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

$$\eta_i = \mathrm{E}[x_i] = \mathrm{Prob}\{x_i = 1\}, \quad i = 1, 2, \tag{6.100}$$

showing the firing rate (probability of $x_i = 1$) of neuron i and

$$\eta_{12} = \mathrm{E}\left[x_i x_j\right] = \mathrm{Prob}\{x_1 = x_2 = 1\}, \tag{6.101}$$

showing the joint firing rate (the probability of the two neurons firing at the same time).

We construct mixture coordinates such that the first part consists of η_1 and η_2 and the second part consists of θ^{12}. Using the mixed coordinate system

$$\boldsymbol{\xi} = \left(\eta_1, \eta_2\ ;\ \theta^{12}\right), \tag{6.102}$$

we have a dually orthogonal foliation. The one-dimensional submanifold $M^*(\eta_1, \eta_2)$ consists of all the distributions in which the firing rates of the two neurons are fixed to (η_1, η_2). The coordinate θ^{12} in $M^*(\eta_1, \eta_2)$ represents how firing of the two neurons is correlated. When $\theta^{12} = 0$, x_1 and x_2 are independent, as is seen from (6.98). Given θ^{12}, the e-flat submanifold $M\left(\theta^{12}\right)$ represents distributions for which interaction of x_1 and x_2 is fixed to be equal to θ^{12} but the firing rates of the neurons are arbitrary. Thus, the partition is done in such a way that the first part represents firing rates of neurons and the second part represents the interaction of two neurons (Fig. 6.4).

One may measure the degree of interaction by the covariance of x_1 and x_2,

$$v = \mathrm{Cov}[x_1, x_2] = \eta_{12} - \eta_1 \eta_2. \tag{6.103}$$

It is 0 when the two neurons fire independently. If we use v as a coordinate in $M^*(\eta_1, \eta_2)$, we have another coordinate system (η_1, η_2, v) in M. However, the

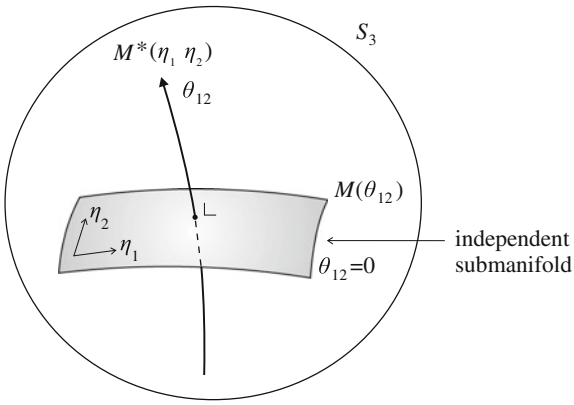


Fig. 6.4 Dual foliation of $S_3 = \{p(x_1, x_2)\}$

v-axis is not orthogonal to the marginal firing rates η_1, η_2, while θ^{12} is. Therefore, the mixed coordinates are successful in decomposing the firing rates and interaction orthogonally but v is not.

Given two distributions $p\,(x_1, x_2)$ and $q\,(x_1, x_2)$, we have the decomposition of their KL-divergence, as

$$KL[p : q] = KL[p : r] + KL[r : q], \tag{6.104}$$

where $r(x_1, x_2)$ is the distribution having the same marginal distributions as p and the same interaction as q. $KL[p : r]$ represents the divergence due to the difference in mutual interaction and $KL[r : q]$ represents that due to marginal firing rates.

6.8.4 Higher-Order Interactions of Neuronal Spikes

We can generalize the idea to a network of n neurons (Amari 2001; Nakahara and Amari 2002; Nakahara et al. 2006; Amari et al. 2003). Let us consider a network consisting of n neurons, which emit spikes stochastically. Let x_i be a binary random variable, representing emission of spikes. The state of the network is represented by $\boldsymbol{x} = (x_1, \ldots, x_n)$. The set of all probability distributions $p(\boldsymbol{x})$ forms S_{N-1}, where $N = 2^n$, since there are N states $\boldsymbol{x}$. This is an exponential family. By expanding $p(\boldsymbol{x})$ as

$$\log p(\boldsymbol{x}) = \sum \theta^i x_i + \sum \theta^{ij} x_i x_j + \cdots + \theta^{1\ldots n} x_1 \ldots x_n - \psi, \tag{6.105}$$

we have

$$p(x, \boldsymbol{\theta}) = \exp\left\{\sum \theta^i x_i + \sum \theta^{ij} x_i x_j + \cdots + \theta^{1\ldots n} x_1 \ldots x_n - \psi\right\}. \tag{6.106}$$

This is called a log linear model. According to the degrees of variables in x_i, we partition the entire $\boldsymbol{\theta}$ in a hierarchical form as

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n), \tag{6.107}$$

$$\boldsymbol{\theta}_1 = \left(\theta^1, \ldots, \theta^n\right), \ \boldsymbol{\theta}_2 = \left(\theta^{12}, \theta^{13}, \ldots, \theta^{n-1n}\right), \ldots \tag{6.108}$$

such that each subvector $\boldsymbol{\theta}_k$ consists of coefficients of monomials $x_{j_1} \ldots x_{j_k}$ of degree k.

The dual affine coordinates are composed of

$$\eta_{i_1 \ldots i_k} = E\left[x_{i_1} \ldots x_{i_k}\right] = \text{Prob}\left\{x_{i_1} = 1, \ldots, x_{i_k} = 1\right\}, \tag{6.109}$$

which are joint firing rates of k neurons, $k = 1, \ldots, n$, and they are hierarchically partitioned as

$$\boldsymbol{\eta} = \left(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_n\right), \tag{6.110}$$

where

$$\boldsymbol{\eta}_k = \left(\eta_{i_1 \ldots i_k}\right), \ k = 1, 2, \ldots, n. \tag{6.111}$$

The kth mixed coordinate system is composed of

$$\boldsymbol{\xi} = \left(\boldsymbol{H}_k; \boldsymbol{\Theta}^k\right) = \left(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k; \boldsymbol{\theta}^{k+1}, \ldots, \boldsymbol{\theta}^n\right). \tag{6.112}$$

Since

$$\boldsymbol{H}_k = \left(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k\right) \tag{6.113}$$

is composed of the joint firing rates up to k neurons, the other coordinates

$$\boldsymbol{\Theta}^k = \left(\boldsymbol{\theta}^{k+1}, \ldots, \boldsymbol{\theta}^n\right) \tag{6.114}$$

represent the directions orthogonal to the joint firing rates up to k neurons. A change in $\boldsymbol{\theta}^{k+1}, \boldsymbol{\theta}^{k+2}, \ldots$ does not affect $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k$ but alters the joint firing rates of more than k neurons. Hence, $\boldsymbol{\Theta}^k$ represents interactions of more than k neurons orthogonal to the firing rates up to k neurons.

Among n terms, $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^n$, we can say that $\boldsymbol{\theta}^k$ represents the degree of mutual interactions among k neurons. $\boldsymbol{\theta}^k$'s ($k \geq 3$) are called the higher-order correlations or interactions of neurons. Although $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^n$ are not mutually orthogonal, $\boldsymbol{\theta}^i$ are orthogonal to $\boldsymbol{\eta}_j$ ($j \neq i$).

We show a simple case of $n = 3$, consisting of three neurons. We have

$$\theta^{123} = \log \frac{p_{111} p_{100} p_{010} p_{001}}{p_{110} p_{101} p_{011} p_{000}}, \tag{6.115}$$

which represents the third-order interactions of the three neurons. It is orthogonal to firing rates of neurons and joint firing rates of any pair of neurons. Similarly,

$$\theta^{12} = \log \frac{p_{110}p_{000}}{p_{100}p_{010}} \tag{6.116}$$

represents pairwise interactions of neurons 1 and 2, which are orthogonal to the firing rates of single neurons.

Remark There are many other hierarchical stochastic systems. One is a Markov chain consisting of various orders. A lower-order system is included in a higher-order system. Hence, we can decompose them in a dually orthogonal way. The auto-regressive (AR) and moving-average (MA) models of time series also form hierarchical stochastic systems, where their degrees compose hierarchy. See Amari (1987, 2001).

6.9 System Complexity and Integrated Information

We consider a stochastic system which receives an input signal $\boldsymbol{x}$, processes it and emits output $\boldsymbol{y}$, and study its complexity by using a mixed coordinate system. We regard it a muliterminal stochastic channel having n input and n output terminals, see Fig. 6.5. Input $\boldsymbol{x} = (x_1, \ldots, x_n)$ and output $\boldsymbol{y} = (y_1, \ldots, y_n)$ are vectors. When a system is very simple, there is no interaction among different terminals. Hence, output y_i depends only on x_i and input x_j $(j \neq i)$ does not affect y_i. A complex system has interaction among different terminals and information is integrated to give an integrated output $\boldsymbol{y}$. The degree of interaction is used to define a measure of complexity of the system (Ay 2002, 2015; Ay et al. 2011). Tononi (2008) initiated a new idea of IIT (integrated information theory) to elucidate consciousness. The degree of information integration distinguishes a conscious state from unconscious states in the brain (Balduzzi and Tononi 2008; Oizumi et al. 2014, etc.).

We propose a measure of complexity, or of information integration, by using a degree of stochastic interaction within a system from the information geometric point of view, based on part of on-going work with M. Oizumi and N. Tsuchiya. This is an extension of the work by Ay (2001, 2015), and is related to the Tononi information integration (Barrett and Seth 2011).

We consider a 2×2 system for simplicity, where input is $\boldsymbol{x} = (x_1, x_2)$ and output is $\boldsymbol{y} = (y_1, y_2)$, having only two terminals (Fig. 6.6), although generalization is easy. We study the binary case where x_i and y_i take on values 0 and 1, and also the Gaussian case where $\boldsymbol{x}$ and $\boldsymbol{y}$ are Gaussian random variables with mean 0. The behavior of a system is described by joint probability distribution $p(\boldsymbol{x}, \boldsymbol{y})$. When the components of $\boldsymbol{x}$ and $\boldsymbol{y}$ are binary, it belongs to an exponential family M_F, called the full model,

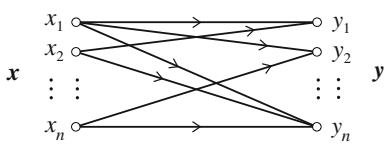


Fig. 6.5 Stochastic information transmission channel

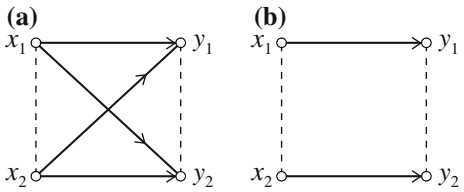


Fig. 6.6 **a** Channel with two terminals and **b** its split version

$$p(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{ \sum \theta_i^X x_i + \sum \theta_i^Y y_i + \theta_{12}^X x_1 x_2 + \theta_{12}^Y y_1 y_2 + \sum \theta_{ij}^{XY} x_i y_j + \text{higher-order terms of } x_i \text{ and } y_j - \psi \right\}, \tag{6.117}$$

described by e-coordinates $\boldsymbol{\theta}$. The higher-order terms are $\theta^{12,1} x_1 x_2 y_1$ and so on. We have the corresponding $\boldsymbol{\eta}$-coordinates. The full model is a graphical model shown in Fig. 6.6a, which is a complete graph, since intrinsic correlations between x_1 and x_2 and also between y_1 and y_2 may exist, as is denoted by the dotted branches in Fig. 6.6a. Refer to information geometry of a graphical model studied in Chap. 11.

The complexity of a system is measured by the degree to which it is different from split systems where no interaction exists between x_i and y_j $(i \neq j)$ (Ay 2002, 2015). So we consider a split system S where no mutual interaction exists, as is shown in Fig. 6.6b. Here, a split model is derived by deleting the branches connecting x_i and y_j $(i \neq j)$. Let the probability distribution of the split model be $q(\boldsymbol{x}, \boldsymbol{y})$, the e-coordinates of which are $\tilde{\boldsymbol{\theta}}$. Since there are no branches connecting (x_1, y_2) and (x_2, y_1), we put $\tilde{\theta}_{12}^{XY} = \tilde{\theta}_{21}^{XY} = 0$. (This is because x_i and y_j $(i \neq j)$ are conditionally independent where the other variables are fixed.) The higher-order terms are also 0. (This is because no cliques exist connecting three or four nodes in the split model.) Hence, a split model has a probability distribution of the form,

$$q(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{ \sum \tilde{\theta}_i^X x_i + \sum \tilde{\theta}_i^Y y_i + \tilde{\theta}_{12}^X x_1 x_2 + \tilde{\theta}_{12}^Y y_1 y_2 + \sum \tilde{\theta}_{ii}^{XY} x_i y_i - \tilde{\psi} \right\}. \tag{6.118}$$

Split models form an exponential family M_S, which has ten degrees of freedom and is a submanifold of M_F.

The split model family M_S defined in the above is slightly different from the one M_S' defined by N. Ay. In a split model belonging to M_S', no direct correlation between y_1 and y_2 exists, so $\tilde{\theta}_{12}^Y = 0$ in addition to $\tilde{\theta}_{12}^{XY} = \tilde{\theta}_{21}^{XY} = 0$. That is, M_S' is derived from M_S by deleting the branch connecting y_1 and y_2. M_S' is an e-flat submanifold of M_S. We do not assume $\tilde{\theta}_{12}^Y = 0$ in M_S, because y_1 and y_2 may be affected by correlated noises directly given from the environment. Since such correlations are given rise to by the environmental situation, even when x_1 and x_2 are independent

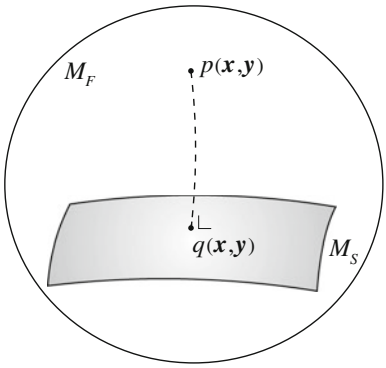


Fig. 6.7 Split model and orthogonal projection

and x_i does not affect y_j $(j \neq i)$, y_1 and y_2 can be correlated in M_S, but not in M'_S. To explain this situation, consider a Gaussian model,

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{\varepsilon} \tag{6.119}$$

where $\mathbf{A}$ is a 2×2 matrix and $\boldsymbol{\varepsilon}$ is a noise term subject to $N(0, \mathbf{V})$, where $\mathbf{V}$ is the covariance matrix of $\boldsymbol{\varepsilon}$. The components ε_1 and ε_2 can be correlated.

The degree of system complexity, or of integrated information, of $p(\boldsymbol{x}, \boldsymbol{y})$ is measured by the KL-divergence from $p(\boldsymbol{x}, \boldsymbol{y})$ to the split distribution $\hat{q}(\boldsymbol{x}, \boldsymbol{y})$ or $\hat{q}'(\boldsymbol{x}, \boldsymbol{y})$ that is closest to $p(\boldsymbol{x}, \boldsymbol{y})$ in M_S or M'_S (Fig. 6.7),

$$\hat{q}(\boldsymbol{x}, \boldsymbol{y}) = D_{KL}\left[p(\boldsymbol{x}, \boldsymbol{y}) : M_S\right] = \underset{q \in M_S}{\arg\min}\, D_{KL}\left[p(\boldsymbol{x}, \boldsymbol{y}) : q(\boldsymbol{x}, \boldsymbol{y})\right], \tag{6.120}$$

$$\hat{q}'(\boldsymbol{x}, \boldsymbol{y}) = D_{KL}\left[p(\boldsymbol{x}, \boldsymbol{y}) : M'_S\right] = \underset{q \in M'_S}{\arg\min}\, D_{KL}\left[p(\boldsymbol{x}, \boldsymbol{y}) : q(\boldsymbol{x}, \boldsymbol{y})\right]. \tag{6.121}$$

They are given by the m-projection of $p(\boldsymbol{x}, \boldsymbol{y})$ to M_S and $M_{S'}$. Since we have two split models M_S and M'_S, we have two definitions of geometric measure of information integration or stochastic interactions.

Definition Geometric measures of information integration, or system complexity, are defined by

$$GI\left[p(\boldsymbol{x}, \boldsymbol{y})\right] = D_{KL}\left[p(\boldsymbol{x}, \boldsymbol{y}) : M_S\right], \tag{6.122}$$

$$GI'[p(\boldsymbol{x}, \boldsymbol{y})] = D_{KL}\left[p(\boldsymbol{x}, \boldsymbol{y}) : M'_S\right]. \tag{6.123}$$

GI' is the same as that of Ay (2002, 2015) and also that of Barrett and Seth (2011). GI is a new measure.

Before comparing these two, we show a criterion which such a measure should satisfy. Oizumi et al. (2015) postulated that a degree ϕ of information integration should satisfy

$$0 \leq \phi \leq I[X : Y], \tag{6.124}$$

where $I[X : Y]$ is the mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$. ϕ should be 0 when $I[X : Y] = 0$, that is, when no information is transmitted from X to Y. They argued that various measures of ϕ so far proposed do not necessarily satisfy the postulate, and defined a new measure ϕ^* based on the concept of mismatched decoding, which satisfies the postulate (Oizumi et al. 2015).

We study properties of GI and GI' and see if they satisfy the postulate. Since M_S is an e-flat submanifold constrained by

$$\theta_{12}^{XY} = \theta_{21}^{XY} = 0, \tag{6.125}$$

where we use $\boldsymbol{\theta}$ instead of $\tilde{\boldsymbol{\theta}}$, we define mixed coordinates

$$\boldsymbol{\xi} = \left(\eta_1^X, \eta_2^X, \eta_{12}^X, \eta_1^Y, \eta_2^Y, \eta_{12}^Y, \eta_{11}^{XY}, \eta_{22}^{XY}\ ;\ \theta_{12}^{XY}, \theta_{21}^{XY}\right). \tag{6.126}$$

Then, the m-projection of $p(\boldsymbol{x}, \boldsymbol{y})$ to M_S,

$$\hat{q}(\boldsymbol{x}, \boldsymbol{y}) = \prod\nolimits_{M_S} p(\boldsymbol{x}, \boldsymbol{y}), \tag{6.127}$$

keeps the $\boldsymbol{\eta}$-part of the mixed coordinates invariant. Therefore, the mixed coordinates $\hat{\boldsymbol{\xi}}$ of $\hat{q}(\boldsymbol{x}, \boldsymbol{y})$ are given by

$$\hat{\eta}_i^X = \eta_i^X, \quad \hat{\eta}_{12}^X = \eta_{12}^X, \quad \hat{\eta}_i^Y = \eta_i^Y, \quad \hat{\eta}_{12}^Y = \eta_{12}^Y, \tag{6.128}$$

$$\hat{\eta}_{ii}^{XY} = \eta_{ii}^{XY}, \quad \hat{\theta}_{12}^{XY} = \hat{\theta}_{21}^{XY} = 0, \tag{6.129}$$

where η_i^X etc. are those of $p(\boldsymbol{x}, \boldsymbol{y})$. These results are directly obtained by minimizing $D_{KL}[p : q],\ q \in M_S$, too. We have a similar result in the case of the m-projection to M_S', where $\hat{\eta}_{12}^Y = \eta_{12}^Y$ is replaced by $\hat{\theta}_{12}^Y = 0$.

We see from (6.128) that the m-projection $\hat{q}(\boldsymbol{x}, \boldsymbol{y})$ is characterized by

$$\hat{q}_X(\boldsymbol{x}) = p_X(\boldsymbol{x}), \quad \hat{q}_Y(\boldsymbol{y}) = p_Y(\boldsymbol{y}), \tag{6.130}$$

where $p_X(\boldsymbol{x})$ etc. are the marginal distributions of $p(\boldsymbol{x}, \boldsymbol{y})$, etc. This means that the marginal distributions of $\hat{q}(\boldsymbol{x}, \boldsymbol{y})$ concerning $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal to those of $p(\boldsymbol{x}, \boldsymbol{y})$, respectively. Moreover, the conditional distributions are equal:

$$\hat{q}\left(y_i|x_i\right) = p\left(y_i|x_i\right), \quad i = 1, 2. \tag{6.131}$$

The m-projection $\hat{q}'(\boldsymbol{x}, \boldsymbol{y})$ to M_S' satisfies

$$\hat{q}_X'(\boldsymbol{x}) = p_Y(\boldsymbol{x}), \tag{6.132}$$

$$\hat{q}'\left(y_i\right) = p\left(y_i\right), \quad i = 1, 2 \tag{6.133}$$

$$\hat{q}'\left(y_i|x_i\right) = p\left(y_i|x_i\right), \quad i = 1, 2. \tag{6.134}$$

Note that $\hat{q}'_Y(\boldsymbol{y}) = p_Y(\boldsymbol{y})$ does not in general hold in M'_S.

Although $\hat{\theta}'^{Y}_{12} = 0$ holds in $\hat{q}'(\boldsymbol{x}, \boldsymbol{y})$, this does not mean that $\hat{y}'_1$ and $\hat{y}'_2$ are uncorrelated. When x_1 and x_2 are correlated, $\hat{y}'_1$ and $\hat{y}'_2$ are correlated even in the split model M'_S.

The measures GI and GI' are represented in terms of entropy and mutual information as follows. Due to the Pythagorean theorem, we have, in the binary case,

$$D_{KL}[p : p_0] = -H[p] + c, \tag{6.135}$$

$$D_{KL}\left[p : \hat{q}\right] = D\left[p : \hat{q}\right] + D\left[\hat{q} : p_0\right], \tag{6.136}$$

where $H[p]$ is the entropy of $p(\boldsymbol{x}, \boldsymbol{y})$ and $p_0(\boldsymbol{x}, \boldsymbol{y})$ is the uniform distribution of which entropy is put equal to c. Therefore, we have

$$GI[p(\boldsymbol{x}, \boldsymbol{y})] = D_{KL}\left[p : \hat{q}\right] = H\left[\hat{q}\right] - H[p]. \tag{6.137}$$

This holds in general, including the Gaussian case, where an independent distribution $p_0(\boldsymbol{x}, \boldsymbol{y})$ is used instead of the uniform p_0. Similarly,

$$GI'[p(\boldsymbol{x}, \boldsymbol{y})] = H\left[\hat{q}'\right] - H[p]. \tag{6.138}$$

Since the entropy is decomposed as

$$H[p] = H[X] + H[Y|X], \tag{6.139}$$

we have the following theorem, which is useful for calculating GI and GI'.

Theorem 6.13 *The two geometrical measures GI and GI' are given, in terms of conditional entropy, as*

$$GI[p(\boldsymbol{x}, \boldsymbol{y})] = H\left[\hat{Y}|X\right] - H[Y|X], \tag{6.140}$$

$$GI'[p(\boldsymbol{x}, \boldsymbol{y})] = H\left[\hat{Y}'|X\right] - H[Y|X], \tag{6.141}$$

where X, and Y denote the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ subject to $p(\boldsymbol{x}, \boldsymbol{y})$, and $\hat{Y}$ and $\hat{Y}'$ denote the random variables $\boldsymbol{y}$ subject to $\hat{q}(\boldsymbol{x}, \boldsymbol{y})$ and $\hat{q}'(\boldsymbol{x}, \boldsymbol{y})$, respectively.

Moreover, we have simpler representations.

Theorem 6.14

$$GI[p] = \sum H[Y_i|X_i] - H[Y|X] - I\left[\hat{Y}_1 : \hat{Y}_2|X\right], \tag{6.142}$$

$$GI'[p] = \sum H[Y_i|X_i] - H[Y|X], \tag{6.143}$$

where $I\left[\hat{Y}_1 : \hat{Y}_2 | X\right]$ *is the conditional mutual information. This elucidates the relation between* GI *and* GI' *as follows:*

$$GI[p] = GI'[p] + D_{KL}\left[\hat{q} : \hat{q}'\right], \tag{6.144}$$

$$D_{KL}\left[\hat{q} : \hat{q}'\right] = H\left[\hat{Y}'|X\right] - H\left[\hat{Y}|X\right], \tag{6.145}$$

$$GI[p] \geq GI'[p]. \tag{6.146}$$

Theorem 6.15 GI *satisfies the postulate (6.124) but* GI' *does not.*

Proof Since both GI and GI' are given by the KL-divergence, they satisfy

$$GI \geq GI' \geq 0. \tag{6.147}$$

Let us next consider the independent distribution

$$p_{\text{ind}}(\boldsymbol{x}, \boldsymbol{y}) = p_X(\boldsymbol{x}) p_Y(\boldsymbol{y}) \tag{6.148}$$

derived from $p(\boldsymbol{x}, \boldsymbol{y})$. The mutual information is

$$I[X : Y] = D_{KL}\left[p(\boldsymbol{x}, \boldsymbol{y}) : p_{\text{ind}}(\boldsymbol{x}, \boldsymbol{y})\right]. \tag{6.149}$$

Since $p_{\text{ind}}(\boldsymbol{x}, \boldsymbol{y})$ satisfies $\theta_{12}^{XY} = \theta_{21}^{XY} = 0$, this is included in M_S. So

$$GI \leq I(X : Y) \tag{6.150}$$

since $\hat{q}$ is the minimizer of divergence in M_S. However, $p_{\text{ind}}(\boldsymbol{x}, \boldsymbol{y})$ does not necessarily satisfy $\theta_{12}^{Y} = 0$ and hence is not included in M_S' in general. Hence,

$$GI' \leq I(X : Y) \tag{6.151}$$

is not guaranteed. Indeed, for $p(\boldsymbol{x}, \boldsymbol{y})$ where X and Y are independent, $I(X : Y) = 0$, but if Y_1 and Y_2 are correlated

$$GI' > 0. \tag{6.152}$$

□

We analyze the Gaussian system given in (6.119) for illustration.

Example 1 (*Gaussian channel*) The joint probability distribution of $(\boldsymbol{x}, \boldsymbol{y})$ in (6.119) is

$$p(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^T \boldsymbol{x} + (\boldsymbol{y} - \mathbf{A}\boldsymbol{x})^T \mathbf{V}^{-1} (\boldsymbol{y} - \mathbf{A}\boldsymbol{x}) - \psi\right)\right\}, \tag{6.153}$$

when $\boldsymbol{x}$ is subject to $N(0, \mathbf{I})$. By putting

$$z = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}, \tag{6.154}$$

it is rewritten as

$$p(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{ -\frac{1}{2} z^T \mathbf{R} z - \psi \right\}, \tag{6.155}$$

where $\mathbf{R}$ is the inverse of the covariance matrix

$$\sum = E\left[z z^T \right], \tag{6.156}$$

and they are given explicitly as functions of the system parameters $\mathbf{A}$ and $\mathbf{V}$.

A full model $p(\boldsymbol{x}, \boldsymbol{y})$ belongs to an exponential family, where the $\boldsymbol{\theta}$-coordinates are $\mathbf{R}$ and $\boldsymbol{\eta}$-coordinates are $\sum$. A split model is given by

$$q(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{ \sum \left(\theta_i^X x_i + \theta_i^Y y_i \right) + \theta_{12}^X x_1 x_2 + \theta_{12}^Y y_1 y_2 + \sum \theta_{ii}^{XY} x_i y_i - \psi \right\}, \tag{6.157}$$

which does not include terms $\theta_{ij}^{XY} x_i y_j$ $(i \neq j)$. By using this expression, we obtain $\hat{q}(\boldsymbol{x}, \boldsymbol{y})$ from $p(\boldsymbol{x}, \boldsymbol{y})$.

However, there is a serious problem concerning the optimal solution. The solution can be written as

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{A}} \boldsymbol{x} + \hat{\varepsilon}, \tag{6.158}$$

but $\hat{\boldsymbol{A}}$ is not diagonal. The solution is split in the sense that $\theta_{ij}^{XY} = 0$ $(i \neq j)$ is satisfied and its graph does not have diagonal branches, but not split in the sense that $\hat{\boldsymbol{A}}$ is not diagonal. Hence, $E[y_i | \boldsymbol{x}]$ depends on both x_1 and x_2. This does not happen in M_S', since $E[y_i | \boldsymbol{x}] = E[y_i | x_i]$ holds.

In order to overcome this flaw, we introduce the third model of split systems,

$$M_S'' = \left\{ q(\boldsymbol{x}, \boldsymbol{y}) \,\middle|\, q\left(x_i, y_j | x_j\right) = q\left(x_i | x_j\right) q\left(y_j | x_j\right),\ i = 1, 2,\ j \neq i \right\} \tag{6.159}$$

This condition can be written as the Markov conditions

$$X_1 \to X_2 \to Y_2, \quad X_2 \to X_1 \to Y_1, \tag{6.160}$$

that is, X_i and Y_j $(i \neq j)$ are conditionally independent when X_j is fixed,

$$I\left(X_1 : Y_2 | X_2\right) = I\left(X_2 : Y_1 | X_1\right) = 0. \tag{6.161}$$

Since M_S'' includes $p_X(\boldsymbol{x}) p_Y(\boldsymbol{y})$, GI'' satisfies the postulate

$$0 \leq GI'' \leq I(X : Y). \tag{6.162}$$

However, $M_S'' \subset M_F$ is neither e-flat nor m-flat. It is curved, so we need to study its properties carefully. This remains as a problem for our future study (Oizumi et al. 2016).

Before finishing this subsection, we show an example in the binary case.

Example 2 (*Binary channel*) We consider two binary transmission channels. One is $C_1(\varepsilon)$, in which y_i chooses x_i with probability $1-\varepsilon$ and chooses x_j $(i \neq i)$ with probability ε. Once x_1 or x_2 is chosen by y_1, the transmission of x_1 (x_2) to y_1 is through a binary symmetric channel with error probability ν. This means that, when $x_1 = 1$, the probability of $y_1 = 1$ is $1-\nu$ and that of $y_1 = 0$ is ν. The other cases are similarly defined. We further consider another channel C_2 which generates $\boldsymbol{z} = (0,0), (1,1)$ with probability $1/2$ each, and its output is $\boldsymbol{y} = \boldsymbol{z}$ irrespective of $\boldsymbol{x}$. So no information transmission takes place in C_2. We study a combined binary channel C that chooses C_1 with probability $1-\delta$ and chooses C_2 with probability δ. The split model M_S is defined by $\varepsilon = 0$, and ν is not necessarily 0. ν plays the role of correlated $\boldsymbol{\varepsilon}$ in the Gaussian case. The split model M_S' is defined by $\varepsilon = 0$ and $\delta = 0$.

Remark 1 We can introduce a hierarchy of split models in a general channel having n input terminals and n output terminals. We partition k inputs $x_1, \ldots, x_n$ into k subsets $X_1, \ldots, X_k$,

$$\cup X_i = \{x_1, \ldots, x_n\}, \quad X_i \cap X_j = \emptyset. \tag{6.163}$$

Similarly, we partition $\boldsymbol{y}$ into $Y_1, \ldots, Y_k$. The split model M_S with respect to this partition is obtained by deleting all the branches connecting terminals in X_i and Y_j $(i \neq j)$. Since a refinement of a partition gives a finer partition, we have a hierarchical structure concerning partitions. Hence, GI forms a hierarchical structure with respect to partitions.

Remark 2 We can extend the above results to the dynamical systems of Markov chains, such that the state $\boldsymbol{x}_{t+1}$ at time $t+1$ is determined stochastically by a conditional probability distribution $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ of a stochastic channel. The initial state distribution $p(\boldsymbol{x}_0)$ is set equal to the stationary distribution of the Markov chain.

6.10 Input–Output Analysis in Economics

We show another example of the dual foliation from the field of economics, due to Morioka and Tsuda (2011). The input–output analysis uses a table $\mathbf{A}$, which is an $n \times n$ matrix, showing the quantities of products and amounts of consumption in n industries and how the products are transferred from one industry to another for consumption. Namely, each row and column of matrix $\mathbf{A} = \left(A_{ij}\right)$ represent an industry and A_{ij} is the amount of product that industry i sells to industry j. A_{ij} are represented in the monetary basis.

Let

$$A_{i\cdot} = \sum_{j=1}^{n} A_{ij} \tag{6.164}$$

be the row sum of the table, which represents the quantity of gross product of industry i. Similarly, the column sum

$$A_{\cdot j} = \sum_{i=1}^{n} A_{ij} \tag{6.165}$$

represents the amount of gross consumption of industry j. They satisfy

$$A_{\cdot\cdot} = \sum_{i} A_{i\cdot} = \sum A_{\cdot j}. \tag{6.166}$$

We have an interest not merely in the gross product and consumption of each industry but more in their interactions, reflecting the structural relationship between industries.

To this end, let us consider the manifold M consisting of all input–output tables

$$M = \{\mathbf{A}\}, \tag{6.167}$$

where A_{ij} form a coordinate system of M. We define another coordinate system by

$$L_{ij} = \log A_{ij}, \quad \mathbf{L} = \left(L_{ij}\right) \tag{6.168}$$

and regard it as an e-flat coordinate system of M. The associated convex function is

$$\psi(\mathbf{L}) = \sum_{ij} \exp\left\{L_{ij}\right\}. \tag{6.169}$$

Then, the dual m-coordinate system is given by $\nabla\psi(\mathbf{L})$ which is merely

$$\mathbf{A} = \left(A_{ij}\right) \tag{6.170}$$

and the dual convex function is

$$\varphi(\mathbf{A}) = \sum_{i,j} \left(A_{ij} \log A_{ij} - A_{ij}\right). \tag{6.171}$$

The canonical divergence between two input–output tables $\mathbf{A}$ and $\mathbf{B}$ is

$$D[\mathbf{A} : \mathbf{B}] = \sum \left\{ B_{ij} \log \frac{B_{ij}}{A_{ij}} - \sum B_{ij} + \sum A_{ij} \right\}. \tag{6.172}$$

In order to separate the distributions of gross products and consumptions from their interrelations, we treat $A_{i\cdot}$ and $A_{\cdot j}$ as a part of new m-affine coordinates, which are linear combinations of m-coordinates A_{ij}. We replace the last row A_{ni} and last column A_{jn} by $A_{i\cdot}$ and $A_{\cdot j}$, respectively. Then we have a modified table in which the last row and column are replaced. We denote the new coordinates by $\tilde{A}_{ij}$. This is given by an affine coordinate transformation from $\mathbf{A}$. The corresponding e-affine coordinates, denoted by $\tilde{L}_{ij}$ are calculated from the invariance relation

$$\sum A_{ij} L_{ij} = \sum \tilde{A}_{ij} \tilde{L}_{ij} \tag{6.173}$$

as

$$\tilde{L}_{ij} = L_{ij} - L_{in} - L_{nj} + L_{nn} = \log \frac{A_{ij} A_{nn}}{A_{in} A_{nj}}, \quad i, j = 1, \ldots, n-1, \tag{6.174}$$

$$\tilde{L}_{in} = L_{in} - L_{nn}, \quad \tilde{L}_{nj} = L_{nj} - L_{nn}, \quad \tilde{L}_{nn} = L_{nn}. \tag{6.175}$$

We partition the coordinates and construct the mixed coordinates. The first part consists of $\left(A_{i\cdot}, A_{\cdot j}, A_{\cdot\cdot}\right), i, j = 1, \ldots, n-1$. The second part consists of $\tilde{L}_{ij}, i, j = 1, \ldots, n-1$. The first m-coordinates represent the gross products and consumptions in industries, while the second part is orthogonal to the first part, representing the interrelations among industries. The divergence between two tables can be decomposed into a sum, the one due to the difference of gross products and consumptions and the second due to the difference in the interrelations.

The $\tilde{L}_{ij}$ are obtained by deleting industry n from the table. Hence, it is not symmetric with respect to all the industries. To overcome this difficulty, let $\tilde{L}_{ij}^{(k)}$ be the e-coordinates where industry k is replaced, instead of industry n, by the total sums. Then, their average defined by

$$\tilde{L}_{ij}^{*} = \frac{1}{n} \sum_{k=1}^{n} \tilde{L}_{ij}^{(k)} \tag{6.176}$$

would be a good measure of interactions among industries.

Instead of replacing one industry k by the gross distributions, we may add $\left(A_{i\cdot}, A_{\cdot j}, A_{\cdot\cdot}\right)$ to the input–output table as its $(n+1)$th row and $(n+1)$th column. Then, the interaction part based on the $(n+1)$th row and column becomes

$$S_{ij} = \log \frac{A_{ij}}{A_{i\cdot} A_{\cdot j}}. \tag{6.177}$$

Morioka and Tsuda (2011) used this for analysis.

Observing the trend of yearly changes in the first part of $\left(A_{i\cdot}, A_{\cdot j}\right)$, one can understand the developments of the gross products in industries. The yearly changes

of the second part $\tilde{L}_{ij}$ represent how the industrial interrelationship changes. This reflects the structural change in the interrelations among industries.

One can try to alter the gross amounts of products of industries from $A_{i\cdot}$ to

$$\bar{A}_{i\cdot} = \mu_i A_{i\cdot} \tag{6.178}$$

by using arbitrary coefficients $\mu_1, \ldots, \mu_n$. By using another set of coefficients $\lambda_1, \ldots, \lambda_n$, the gross consumptions are changed to

$$\bar{A}_{\cdot j} = \lambda_j A_{\cdot j}. \tag{6.179}$$

Such changes can be realized by transforming A_{ij} into

$$\bar{A}_{ij} = \mu_i \lambda_j A_{ij}. \tag{6.180}$$

This is called the RAS transformation, by which the interrelationship $\tilde{L}_{ij}$ does not change but the gross amounts of products and consumptions may change arbitrarily.

Annual statistics of gross amounts $A_{i\cdot}$ and $A_{\cdot j}$ are published every year, but A_{ij} themselves are not, because construction of the entire A_{ij} table is laborious. So, the entire table is published only every five years in Japan, for example. In such a case, we can interpolate the $\tilde{L}_{ij}$ part (or S_{ij} part) in the unknown years by using the e-geodesic in the interaction part from the known S-parts. Morioka and Tsuda (2011) studied the change in the industrial structure of Japan after the War, finding remarkable changes occurring as the Japanese economy developed.

See Marriott and Salmon (2011) for other applications of geometry to economics.

Remarks

The concept of dual affine connections was introduced in a Riemannian manifold by Amari (1982) and Nagaoka and Amari (1982). See also Amari and Nagaoka (2000). The idea emerged from the invariant geometry of a manifold of probability distributions due to Chentsov (1972). However, the late professor K. Nomizu stated that such a concept exists in affine differential geometry (Nomizu and Sasaki 1994).

Affine differential geometry studies properties of n-dimensional hypersurfaces embedded in an $(n+1)$-dimensional affine space. This was originally developed by W. Blaschke and also developed by J. L. Koszul (see Nomizu and Sasaki 1994). The Hessian manifold of Shima (2007) also deals with a dually flat manifold.

The concept of dual (conjugate) affine connections is naturally introduced in affine differential geometry but it has not played a central role. The concept of dual connections in information geometry is more general, since it deals with a manifold which might not be embedded in an $(n+1)$-dimensional affine space. However, there is much overlap between these two fields and they have been developping through mutual interactions. The present monograph does not touch upon affine differential geometry, although there are many common interesting problems. Excellent work

is found in Kurose (1990, 1994, 2002). See also Matsuzoe (1998, 1999), Matsuzoe et al. (2006), Uohashi (2002) and many others.

lnvariant geometry is due to Chentsov (1972), where the uniqueness of two tensors G and T is presented. The invariant geometry (α-geometry) is constructed from these tensors. How is a general dual manifold related to a statistical manifold? Due to a theorem of Banerjee et al. (2005), we know that any dually flat manifold is realized as an exponential family. Lê (2005) proved a stronger theorem that any dual manifold can be realized as a submanifold of an N-dimensional probability simplex S_N for a sufficiently large N. There is another interesting problem: Given a Riemannian manifold $\{M, G\}$, on what condition does it become dually flat by supplementing an adequate T? Such a Riemannian manifold is said to be flattenable. It is interesting to know the characterization of flattenable Riemannian manifolds. When $n = 2$, this is always possible, but when $n > 2$, it is not. This problem was studied by Amari and Armstrong (2014).

The Chentsov invariance theorem was proved in the discrete case of S_n. Amari and Nagaoka (2000) formulated the invariance in a general continuous case in terms of sufficient statistics. However, there is no rigorous proof due to difficulties in dealing with a function space. The Leipzig group, including J. Jost and H. V. Lê, is tackling this problem (Ay et al. 2013).

The global topology of a statistical manifold is another interesting problem of differential geometry. It is interesting to see how a dual pair of local curvatures is related to the global topology of a manifold.

Finally, we give a list of monographs on information geometry. They each have the own characteristics: Amari (1985), Amari and Nagaoka (2000), Arwini and Dodson (2008), Calin and Udriste (2013), Chentsov (1972), Kass and Vos (1997), Murray and Rice (1993).

Part III
Information Geometry of Statistical Inference

Chapter 7
Asymptotic Theory of Statistical Inference

7.1 Estimation

Let $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$ be a statistical model specified by parameter $\boldsymbol{\xi}$, which is to be estimated. When we observe N independent data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from $p(\boldsymbol{x}, \boldsymbol{\xi})$, we want to know the underlying parameter $\boldsymbol{\xi}$. This is a problem of estimation, and an estimator

$$\hat{\boldsymbol{\xi}} = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \tag{7.1}$$

is a function of D. The estimation error is given by

$$\boldsymbol{e} = \hat{\boldsymbol{\xi}} - \boldsymbol{\xi}, \tag{7.2}$$

when $\boldsymbol{\xi}$ is the true value. The bias of the estimator is defined by

$$\boldsymbol{b}(\boldsymbol{\xi}) = \mathrm{E}\left[\hat{\boldsymbol{\xi}}\right] - \boldsymbol{\xi}, \tag{7.3}$$

where the expectation is taken with respect to $p(\boldsymbol{x}, \boldsymbol{\xi})$. An estimator is unbiased when $\boldsymbol{b}(\boldsymbol{\xi}) = 0$.

The asymptotic theory studies the behavior of an estimator when N is large. When the bias satisfies

$$\lim_{N\to\infty} \boldsymbol{b}(\boldsymbol{\xi}) = 0, \tag{7.4}$$

it is asymptotically unbiased.

It is expected that a good estimator converges to the true parameter as N tends to infinity. It is written as

$$\lim_{N\to\infty} \hat{\boldsymbol{\xi}} = \boldsymbol{\xi}. \tag{7.5}$$

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_7

When this holds, an estimator is consistent. The accuracy of an estimator is measured by the error covariance matrix, $\mathbf{V} = \left(V_{ij}\right)$,

$$V_{ij} = \mathrm{E}\left[\left(\hat{\xi}_i - \xi_i\right)\left(\hat{\xi}_j - \xi_j\right)\right]. \tag{7.6}$$

It decreases in general in proportion to $1/N$, so that the estimator $\hat{\boldsymbol{\xi}}$ becomes sufficiently accurate as N increases. The well-known Cramér–Rao Theorem gives a bound of accuracy.

Theorem 7.1 *For an asymptotically unbiased estimator $\hat{\boldsymbol{\xi}}$, the following inequality holds:*

$$\mathbf{V} \geq \frac{1}{N}\mathbf{G}^{-1}, \tag{7.7}$$

$$\mathrm{E}\left[\left(\hat{\xi}_i - \xi_i\right)\left(\hat{\xi}_j - \xi_j\right)\right] \geq \frac{1}{N}g^{ij}, \tag{7.8}$$

where $\mathbf{G} = \left(g_{ij}\right)$ is the Fisher information matrix, $\mathbf{G}^{-1} = \left(g^{ij}\right)$ is its inverse, and the matrix inequality implies that $\mathbf{V} - \mathbf{G}^{-1}/N$ is positive semi-definite.

The maximum likelihood estimator (MLE) is the maximizer of the likelihood,

$$\hat{\boldsymbol{\xi}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\xi}} \prod_{i=1}^{N} p\left(\boldsymbol{x}_i, \boldsymbol{\xi}\right). \tag{7.9}$$

It is known that the MLE is asymptotically unbiased and its error covariance satisfies

$$\mathbf{V}_{\mathrm{MLE}} = \frac{1}{N}\mathbf{G}^{-1} + O\left(\frac{1}{N^2}\right), \tag{7.10}$$

attaining the Cramér–Rao bound (7.7) asymptotically. Such an estimator is said to be Fisher efficient (first-order efficient).

Remark We do not mention Bayes estimators, where a prior distribution of parameters is used. However, when the prior distribution is uniform, the MLE is the maximum a posteriori Bayes estimator. Moreover, it has the same asymptotic properties for any regular Bayes prior. Information geometry of Bayes statistics will be touched upon in a later chapter.

7.2 Estimation in Exponential Family

An exponential family is a model having excellent properties such as dual flatness. We begin with an exponential family

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\} \tag{7.11}$$

to study the statistical theory of estimation, because it is simple and transparent.

Given data D, their joint probability distribution is written as

$$p(D, \boldsymbol{\theta}) = \exp\left[N\left\{(\boldsymbol{\theta} \cdot \bar{\boldsymbol{x}}) - \psi(\boldsymbol{\theta})\right\}\right], \tag{7.12}$$

where $\bar{\boldsymbol{x}}$ is the arithmetic mean of the observed examples,

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i. \tag{7.13}$$

It is a sufficient statistic. The MLE $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is given by differentiating (7.12) and is the solution to

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}) = \bar{\boldsymbol{x}}. \tag{7.14}$$

Using the $\boldsymbol{\eta}$-coordinates, this is written as

$$\hat{\boldsymbol{\eta}}_{\mathrm{MLE}} = \bar{\boldsymbol{x}}. \tag{7.15}$$

Observed data defines a point $\bar{\boldsymbol{\eta}}$ in M of which the coordinates are

$$\bar{\boldsymbol{\eta}} = \bar{\boldsymbol{x}}. \tag{7.16}$$

We call it the observed point determined from data D, which is nothing other than the MLE in the $\boldsymbol{\eta}$-coordinates. The following theorem is easy to prove.

Theorem 7.2 *The MLE is unbiased and efficient:*

$$\mathrm{E}\left[\hat{\boldsymbol{\eta}}_{MLE}\right] = \boldsymbol{\eta}, \tag{7.17}$$

$$\mathbf{V} = \frac{1}{N}\mathbf{G}^{-1}. \tag{7.18}$$

Proof We see from the central limit theorem that $\bar{\boldsymbol{\eta}}$ is asymptotically subject to a Gaussian distribution with mean $\boldsymbol{\eta}$ and covariance matrix $\mathbf{G}^{-1}/N$. Since the MLE attains the Cramér–Rao bound, it is the best estimator in an exponential family. □

Remark The MLE $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ expressed in the $\boldsymbol{\theta}$-coordinates is asymptotically unbiased and asymptotically efficient, but it is not exactly unbiased, nor does it attain the Cramér–Rao bound exactly. This is because the bias and covariance matrix are not tensors so that the results are different in the $\boldsymbol{\theta}$-coordinate system.

7.3 Estimation in Curved Exponential Family

Estimation in an exponential family is too simple. We study estimation in a curved exponential family, which is a submanifold embedded in an exponential family. Many statistical models belong to this class. A curved exponential family of probability distributions with parameter $\boldsymbol{u}$ is written in the following form:

$$p(\boldsymbol{x}, \boldsymbol{u}) = \exp\left[\boldsymbol{\theta}(\boldsymbol{u}) \cdot \boldsymbol{x} - \psi\left\{\boldsymbol{\theta}(\boldsymbol{u})\right\}\right]. \tag{7.19}$$

$S = \{p(\boldsymbol{x}, \boldsymbol{u})\}$ is a submanifold of an exponential family $M = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$, where $\boldsymbol{u}$ is a coordinate system of S.

Observed data D specifies the observed point $\bar{\boldsymbol{\eta}} = \bar{\boldsymbol{x}}$ in the ambient exponential family M, which is not included in S in general. An estimated value of $\boldsymbol{u}$ is derived by mapping the observed point $\bar{\boldsymbol{\eta}}$ to S (Fig. 7.1). That is, an estimator $\hat{\boldsymbol{u}}$ is derived from a mapping from M to S. Let it be

$$f : M \to S \tag{7.20}$$

such that

$$\hat{\boldsymbol{u}} = f(\bar{\boldsymbol{\eta}}). \tag{7.21}$$

The observed point $\bar{\boldsymbol{\eta}}$ converges to the true point as N goes to infinity, as is clear from the law of large numbers. Hence, a consistent estimator satisfies

$$\lim_{N\to\infty} \hat{\boldsymbol{u}} = f\left\{\boldsymbol{\eta}(\boldsymbol{u})\right\}. \tag{7.22}$$

Let us consider the set of points $\boldsymbol{\eta}$ in M which are mapped to $\boldsymbol{u}$ by the estimator $f(\boldsymbol{\eta})$. This is the inverse image of an estimator f, denoted by

$$A(\boldsymbol{u}) = f^{-1}(\boldsymbol{u}) = \left\{\boldsymbol{\eta} \in M \,|\, f(\boldsymbol{\eta}) = \boldsymbol{u}\right\}. \tag{7.23}$$

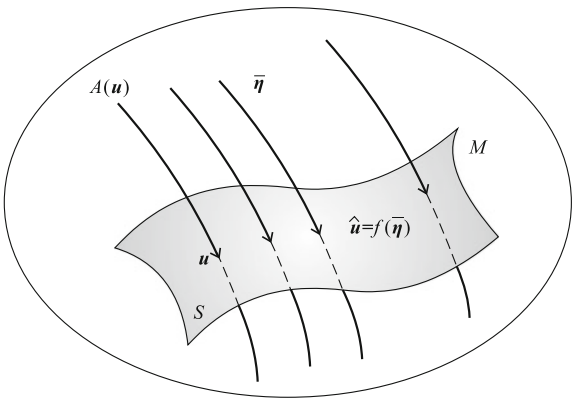


Fig. 7.1 An estimator $f : \boldsymbol{\eta} \to \boldsymbol{\eta} = f(\boldsymbol{\eta})$ defines auxiliary submanifold $A(\boldsymbol{u}) = f^{-1}(\boldsymbol{u})$

It forms an $(n-m)$-dimensional submanifold passing through $\boldsymbol{\eta}(\boldsymbol{u}) \in M$ (Fig. 7.1). We call it an ancillary submanifold associated with estimator f. $A(\boldsymbol{u})$ is defined at each $\boldsymbol{u} \in S$ and they give a foliation of M at least in a neighborhood of S,

$$A(\boldsymbol{u}) \bigcap A\left(\boldsymbol{u}'\right) = \emptyset, \quad \boldsymbol{u} \neq \boldsymbol{u}', \tag{7.24}$$

$$\bigcup_{\boldsymbol{u}} A(\boldsymbol{u}) \supset U, \tag{7.25}$$

where U is a neighborhood of S. When $A(\boldsymbol{u}) \ni \boldsymbol{\eta}(\boldsymbol{u})$, that is, when $A(\boldsymbol{u})$ passes through $\boldsymbol{\eta}(\boldsymbol{u})$, $A(\boldsymbol{u})$ gives a consistent estimator.

An estimator defines an ancillary family $\mathcal{A} = \{A(\boldsymbol{u})\}$ associated with it and conversely an ancillary family $\mathcal{A}$ defines a consistent estimator when f satisfies (7.22). It is possible to study the performance of an estimator in terms of the geometry of an ancillary family. Let us define a coordinate system $\boldsymbol{v}$ inside each $A(\boldsymbol{u})$ such that the origin $\boldsymbol{v} = 0$ is at $\boldsymbol{\eta}(\boldsymbol{u})$ which is the intersection of $A(\boldsymbol{u})$ and S. We denote coordinates of S by

$$\boldsymbol{u} = \left(u^a\right), \quad a = 1, \ldots, m \tag{7.26}$$

and coordinates in $A(\boldsymbol{u})$ by

$$\boldsymbol{v} = \left(v^\kappa\right), \quad \kappa = m+1, \ldots, n. \tag{7.27}$$

Then, combining the two, we have a new coordinate system of M,

$$\boldsymbol{w} = (\boldsymbol{u}, \boldsymbol{v}) = \left(w^\alpha\right), \quad \alpha = 1, 2, \ldots, n, \tag{7.28}$$

defined in a neighborhood $U \subset S$ (Fig. 7.2).

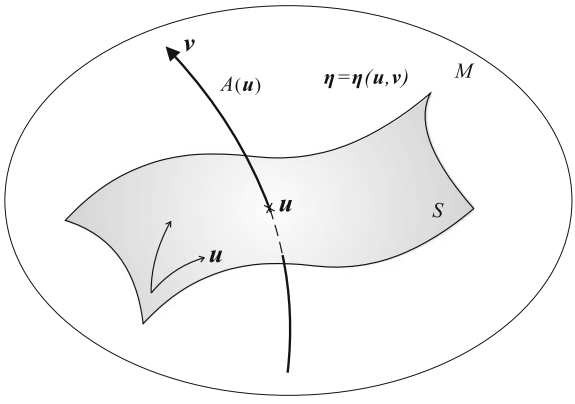


Fig. 7.2 New coordinate system $\boldsymbol{w} = (\boldsymbol{u}, \boldsymbol{v})$ of M

The $\boldsymbol{\theta}$-coordinates and $\boldsymbol{\eta}$-coordinates in M are written in terms of the new coordinates $\boldsymbol{w}$ as

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{w}) = \boldsymbol{\theta}(\boldsymbol{u}, \boldsymbol{v}), \tag{7.29}$$

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{w}) = \boldsymbol{\eta}(\boldsymbol{u}, \boldsymbol{v}). \tag{7.30}$$

Any point in S satisfies $\boldsymbol{v} = 0$, so that S is represented by

$$S = \{\boldsymbol{\eta}(\boldsymbol{u}, \boldsymbol{v}) \,|\boldsymbol{v} = 0\}. \tag{7.31}$$

The Jacobian matrices of the coordinate transformations between $\boldsymbol{w}$ and $\boldsymbol{\theta}$ and $\boldsymbol{w}$ and $\boldsymbol{\eta}$ are expressed as

$$B^i_\alpha = \frac{\partial \theta^i}{\partial w^\alpha}, \tag{7.32}$$

$$B_{\alpha i} = \frac{\partial \eta_i}{\partial w^\alpha}, \tag{7.33}$$

and are decomposed as

$$B^i_a = \frac{\partial \theta^i}{\partial u^a}, \quad B^i_\kappa = \frac{\partial \theta^i}{\partial v^\kappa}; \tag{7.34}$$

$$B_{ai} = \frac{\partial \eta_i}{\partial u^a}, \quad B_{\kappa i} = \frac{\partial \eta_i}{\partial v^\kappa} \tag{7.35}$$

in terms of the $\boldsymbol{u}$ and $\boldsymbol{v}$ coordinates.

The Fisher information is given in the $\boldsymbol{w}$-coordinate system as

$$g_{\alpha\beta} = B^i_\alpha g_{ij} B^j_\beta \tag{7.36}$$

and is decomposed as

$$\mathbf{G} = \begin{bmatrix} g_{ab} & g_{a\lambda} \\ g_{\kappa b} & g_{\kappa\lambda} \end{bmatrix}. \tag{7.37}$$

Given data D, the $\boldsymbol{u}$- and $\boldsymbol{v}$-coordinates $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$ of the observed point $\bar{\boldsymbol{\eta}}$ are determined from

$$\bar{\boldsymbol{\eta}} = \boldsymbol{\eta}(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}). \tag{7.38}$$

The estimator associated with ancillary family $\mathcal{A}$ is given by

$$\hat{\boldsymbol{u}} = \bar{\boldsymbol{u}}. \tag{7.39}$$

7.4 First-Order Asymptotic Theory of Estimation

When the true distribution is $\boldsymbol{u}$ in S, by the law of large numbers, the observed point $\bar{\boldsymbol{\eta}}$ converges to

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{u}, 0), \tag{7.40}$$

as N tends to infinity. We define the error, that is the deviation of the observed point from the true distribution in the $\boldsymbol{\eta}$-coordinates, by

$$\boldsymbol{e} = \bar{\boldsymbol{\eta}} - \boldsymbol{\eta}. \tag{7.41}$$

Since it is small, we normalize it as

$$\tilde{\boldsymbol{e}} = \sqrt{N}\boldsymbol{e}. \tag{7.42}$$

Then, the moments of the error are easily calculated. They are summarized in the following theorem.

Theorem 7.3 *The moments of the error (deviation) $\tilde{\boldsymbol{e}}$ in the $\boldsymbol{\eta}$-coordinates are given by*

$$\mathrm{E}\left[\tilde{e}_i\right] = 0, \tag{7.43}$$
$$\mathrm{E}\left[\tilde{e}_i \tilde{e}_j\right] = g_{ij}, \tag{7.44}$$
$$\mathrm{E}\left[\tilde{e}_i \tilde{e}_j \tilde{e}_k\right] = \frac{1}{\sqrt{N}} T_{ijk}, \tag{7.45}$$

where

$$g_{ij} = \partial_i \partial_j \psi(\boldsymbol{\theta}), \tag{7.46}$$
$$T_{ijk} = \partial_i \partial_j \partial_k \psi(\boldsymbol{\theta}). \tag{7.47}$$

Let us also normalize the error in the $\boldsymbol{w}$-coordinates as

$$\tilde{\boldsymbol{w}} = \sqrt{N}\left(\bar{\boldsymbol{w}} - \boldsymbol{w}\right), \tag{7.48}$$

where $\bar{\boldsymbol{w}}$ is the $\boldsymbol{w}$-coordinates of $\bar{\boldsymbol{\eta}}$. By expanding

$$\bar{\boldsymbol{x}} = \boldsymbol{\eta}\left(\boldsymbol{w} + \frac{\tilde{\boldsymbol{w}}}{\sqrt{N}}\right), \tag{7.49}$$

we have

$$\bar{x}_i = \eta_i + \frac{1}{\sqrt{N}} B_{\alpha i} \tilde{w}^{\alpha} + \frac{1}{2N} B_{\alpha\beta i} \tilde{w}^{\alpha} \tilde{w}^{\beta} + O\left(\frac{1}{N\sqrt{N}}\right), \tag{7.50}$$

where

$$B_{\alpha\beta i} = \frac{\partial^2 \eta_i}{\partial w^\alpha \partial w^\beta}. \tag{7.51}$$

By inverting (7.50), we have

$$\tilde{w}^\alpha = g^{\alpha\beta} B^i_\beta \tilde{e}_i - \frac{1}{2\sqrt{N}} C_{\beta\gamma}{}^\alpha \tilde{w}^\beta \tilde{w}^\gamma, \tag{7.52}$$

where

$$C_{\beta\gamma}{}^\alpha = B^{\alpha i} B_{\beta\gamma i}. \tag{7.53}$$

We have, therefore, an asymptotic evaluation of the error in the $\boldsymbol{w}$-coordinates as

$$\mathrm{E}\left[\tilde{w}^\alpha\right] = -\frac{1}{2\sqrt{N}} C_{\beta\gamma}{}^\alpha g^{\beta\gamma}, \tag{7.54}$$

$$\mathrm{E}\left[\tilde{w}^\alpha \tilde{w}^\beta\right] = g^{\alpha\beta}. \tag{7.55}$$

Since $\tilde{\boldsymbol{e}} = \sqrt{N}\,(\bar{\boldsymbol{x}} - \boldsymbol{\eta})$ are asymptotically Gaussian, the error $\tilde{\boldsymbol{w}} = (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ in (7.48) expressed in the $\boldsymbol{w}$-coordinates is asymptotically

$$p\,(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) = c \exp\left\{-\frac{1}{2} g_{\alpha\beta} \tilde{w}^\alpha \tilde{w}^\beta\right\}. \tag{7.56}$$

By integrating $p\,(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ with respect to $\tilde{\boldsymbol{v}}$, we have the asymptotic distribution of the estimation error

$$p\,(\tilde{\boldsymbol{u}}) = c \exp\left\{-\frac{1}{2} \bar{g}_{ab} \tilde{u}^a \tilde{u}^b\right\}, \tag{7.57}$$

where

$$\bar{g}_{ab} = g_{ab} - g_{a\kappa} g_{b\lambda} g^{\kappa\lambda}. \tag{7.58}$$

When $A(\boldsymbol{u})$ is orthogonal to M,

$$g_{a\kappa} = B^i_a g_{ij} B^j_\kappa = 0, \tag{7.59}$$

so that we have

$$p\,(\tilde{u}) = c \exp\left\{-\frac{1}{2} g_{ab} \tilde{u}^a \tilde{u}^b\right\}. \tag{7.60}$$

In general

$$(\bar{g}_{ab}) \leq (g_{ab}) \tag{7.61}$$

and $(\bar{g}_{ab})$ is maximized in the orthogonal case, where the Cramér–Rao bound is asymptotically attained. An estimator is efficient in this case.

We summarize the results in the following.

Theorem 7.4 *(1) An estimator $\hat{\boldsymbol{u}}$ is consistent when its ancillary family $A(\boldsymbol{u})$ passes through $\boldsymbol{w} = (\boldsymbol{u}, 0) \in S$ in M. (2) A consistent estimator is efficient when $A(\boldsymbol{u})$ is orthogonal to S.*

The maximum likelihood estimator is given by the m-projection of $\bar{\boldsymbol{\eta}}$ to S. Therefore, its $A(\boldsymbol{u})$ is orthogonal to S and it is efficient.

Remark The first-order asymptotic theory is a linear theory in a small neighborhood of the true distribution. Hence, it is enough to consider only the tangent space T_η instead of the entire M for evaluating the performance of an estimator. Therefore, the asymptotic theory is common for all regular statistical models. We may consider the case where the ancillary family $A(\boldsymbol{u})$ depends on N so that it is denoted as $A_N(\boldsymbol{u})$. Then, the theory holds when $A_N(\boldsymbol{u})$ passes through $(\boldsymbol{u}, 0)$ and is orthogonal to S, as N tends to infinity. Such an ancillary family is important for studying the performance of testing hypotheses.

7.5 Higher-Order Asymptotic Theory of Estimation

The covariance matrix of an efficient estimator achieves the CR-bound $\mathbf{G}^{-1}/N$ asymptotically when we ignore the term of order $1/N^2$. The higher-order asymptotic theory evaluates this higher-order term. This makes it possible to compare the performances of various efficient estimators more accurately.

In order to compare the higher-order errors, we introduce asymptotic bias-correction of estimators. The asymptotic bias $\boldsymbol{b}$ of an estimator is given in (7.54), which is of the order $1/N$. If we modify the estimator by

$$\hat{\boldsymbol{u}}^* = \hat{\boldsymbol{u}} - \boldsymbol{b}\left(\hat{\boldsymbol{u}}\right), \tag{7.62}$$

the bias of the new estimator becomes

$$E\left[\hat{\boldsymbol{u}}^*\right] - \boldsymbol{u} = O\left(\frac{1}{N^2}\right). \tag{7.63}$$

We call it a bias-corrected estimator. In order to compare the covariances of various efficient estimators, we use their bias-corrected versions. The idea of bias correction is due to Rao (1962), and is necessary in order to exclude estimators which are good at some specific points but not uniformly good. For example, the trivial estimator

$$\hat{\boldsymbol{u}} = \boldsymbol{u}_0 \tag{7.64}$$

which does not depend on data D, is the best estimator when the true distribution is $\boldsymbol{u}_0$ but very bad for other $\boldsymbol{u}$.

We evaluate the error terms from (7.52) by using the higher-order terms of the Taylor expansion, where we need higher-order moments of the error given in (7.43)–(7.45). We then have the following theorem. The calculations are technical and they are formidably complicated, so we neglect them and give only the results. See Amari (1985).

Theorem 7.5 *The covariance matrix of a bias-corrected efficient estimator is given by*

$$E\left[\tilde{u}^{*a}\tilde{u}^{*b}\right] = g^{ab} + \frac{1}{2N}\left\{\left(\Gamma_S^{m2}\right)^{ab} + 2\left(H_S^{e2}\right)^{ab} + \left(H_A^{m2}\right)^{ab}\right\} + O\left(\frac{1}{N^2}\right), \tag{7.65}$$

where

$$\left(H_S^{e2}\right)^{ab} = H_{ec}^{(e)\kappa} H_{fd}^{(e)\lambda} g^{cd} g_{\kappa\lambda} g^{ae} g^{fb} \tag{7.66}$$

is the square of the e-embedding curvature of S,

$$\left(H_A^{m2}\right)^{ab} = H_{\kappa\lambda}^{(m)a} H_{\mu\nu}^{(m)b} g^{\kappa\mu} g^{\lambda\nu} \tag{7.67}$$

is the square of the m-embedding curvature of the ancillary family $A(\boldsymbol{u})$ and

$$\left(\Gamma_S^{m2}\right)^{ab} = \Gamma_{cd}^{(m)a} \Gamma_{ef}^{(m)b} g^{ce} g^{df} \tag{7.68}$$

is the square of the m-connection of the coordinate system $\boldsymbol{u}$ in S.

Thus, the second-order terms of the covariance of error are decomposed into a sum of three non-negative terms. The *e*-curvature term $\left(H_S^{e2}\right)^{ab}$ depends on the statistical model S, showing the degree of its deviation from an exponential family. This vanishes when S itself is an exponential family. This term was introduced by Efron (1975) and he named it statistical curvature. The term $\left(\Gamma_S^{m2}\right)^{ab}$ depends on the method of parameterization $\boldsymbol{u}$ in S and is common to all estimators. The m-curvature term $\left(H_A^{m2}\right)^{ab}$ depends on the m-embedding curvature of $A(\boldsymbol{u})$. It vanishes when the m-curvature of $A(\boldsymbol{u})$ vanishes. Note that the m-curvature of $A(\boldsymbol{u})$ vanishes for the MLE, since the MLE is given by the m-projection of the observed point to S. This is the only quantity which depends on the estimator.

Theorem 7.6 *A bias-corrected efficient estimator is second-order efficient when the embedding m-curvature of the associated $A(\boldsymbol{u})$ vanishes at S. The bias-corrected MLE is second-order efficient.*

Remark It is intriguing to ask if the higher-order bias-corrected MLE is third-order efficient or not. Unfortunately, it is not. Kano (1997, 1998) disproved the conjecture, showing that the MLE is not third-order efficient. It was Fisher's belief that the MLE would be the best estimator, but the dream of Fisher was shattered in the third-order asymptotic theory.

7.6 Asymptotic Theory of Hypothesis Testing

When the number of observations is large, we have an asymptotic theory of hypothesis testing. A typical situation is to test a null hypothesis

$$H_0 : u = u_0 \tag{7.69}$$

against alternatives

$$H : u > u_0 \tag{7.70}$$

in a one-dimensional curved exponential family $S = \{p(\boldsymbol{x}, \boldsymbol{\theta}(u))\}$. This is a one-sided test but we can treat a two-sided test similarly.

Since S is a curve in M, we design a test by defining a rejection region R in M such that hypothesis H_0 is rejected when the observed point $\bar{\boldsymbol{\eta}}$ is included in R and is not rejected (is accepted) otherwise. The observed point $\bar{\eta}$ converges to u_0 as N increases when hypothesis H_0 is true. Hence, the rejection region should not include u_0, but its boundary $B = \partial R$ lies close to u_0, approaching u_0 as N tends to infinity. See Fig. 7.3. The boundary surface $B\,(u_0)$ of R depends on the null hypothesis u_0. It is an $(n-1)$-dimensional hypersurface crossing S at point u_0' which converges to u_0 as N increases. We denote it by $A_N\,(u_0)$. See Fig. 7.3.

We consider $u = u_0$ as a free scalar parameter, and form an ancillary family of $\mathcal{A} = \{A_N(u)\}$, depending on N. This is a foliation of M consisting of the boundaries of the rejection regions for various $u = u_0$. This is useful for analyzing the performance of a hypothesis testing. The first-order asymptotic theory is easy, since $\bar{\boldsymbol{\eta}}$ converges to $\boldsymbol{\eta}\,(u_0)$ under hypothesis H_0.

Theorem 7.7 *A test is (first-order) efficient when the associated ancillary surface $A_N(u)$ passing through u_N is orthogonal to S and u_N converges to u_0, as N tends to infinity.*

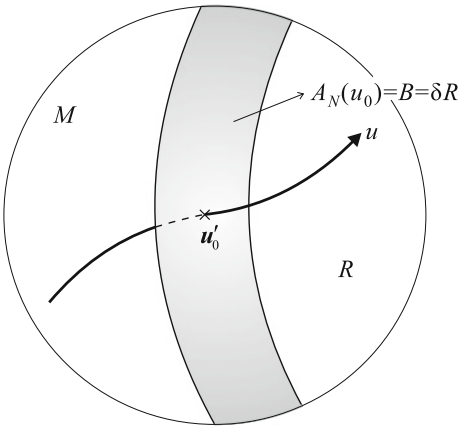


Fig. 7.3 Rejection region R and associated auxiliary submanifold $A_N(\boldsymbol{u}_0)$

There are many first-order efficient tests, the Rao test, Wald test, likelihood-ratio test, locally most powerful test among others. How do these tests differ in their performance? The question is answered by studying the power functions of test T, the probabilities $P_T(u)$ of rejecting H_0 when the true distribution is u, up to the higher order. There are no uniformly most powerful tests in the second order except for the case that S is an exponential family. Therefore, one test is powerful at a specific point, while another is good at a different point. Information geometry characterizes the performances of various tests by the geometry of the associated ancillary surfaces, in particular by the m-embedding curvatures of $A_N(u)$ and the asymptotic angle between $A_N(u)$ and S. The second-order power functions of various tests are analyzed in Kumon and Amari (1983), Amari (1985). See also Amari and Nagaoka (2000).

Remarks

Information geometry was developed for elucidating higher-order characteristics of statistical inference, in particular, estimation and hypothesis testing. The first-order theory was established by the Cramér–Rao theory and the Neyman–Pearson fundamental lemma. Researchers tackled the second-order theories in the late 1970s and many results were obtained independently in Japan, Germany, India, Russia and the U.S.A. See Akahira and Takeuchi (1981). B. Efron was the first to point out the role of statistical curvature in the second-order asymptotic theory (Efron 1975).

Amari (1982) established the second-order theory of estimation by using differential geometry. Kumon and Amari (1983) extended it to the higher-order theory of hypothesis testing. Information geometry was developed further for this purpose, while the duality theory was established by Nagaoka and Amari (1982). See also Amari and Nagaoka (2000).

Sir David Cox, one of most influential statisticians, paid attention to differential geometrical theory when he visited Japan, and he organized a Workshop on Differential Geometry of Statistics in London in 1984. Numerous active statisticians, C.R. Rao, B. Efron, A.P. Dawid, R. Kass, N. Read, O.E. Barndorff-Nielsen, S. Lauritzen, D.V. Hinkley, S. Eguchi and many others, participated in the workshop. It was very fortunate for information geometry that the topic was discussed openly at this workshop in its period of early infancy. But it was unfortunate that N.N. Chentsov could not participate, because it was supported by NATO and the world was divided by the Iron Curtain at that time.

We have shown in this chapter the asymptotic theory of statistics in the framework of a curved exponential family. We have not described details, but shown only intuitive ideas and results. The details are shown in Amari (1985) and also in Amari and Nagaoka (2000) or related journal papers. Since not all regular statistical models are curved exponential families, one might wonder if the theory is valid in a more general regular statistical model. We can prove that most results of higher-order statistical theory hold in a general regular statistical model, by forming a fiber bundle-like structure attached to S, consisting of higher-order derivatives of the score function. This is called a local exponential family. See Amari (1985) for details of higher-order asymptotic theory.

How about non-regular statistical models, where the Fisher information matrix is degenerate or not defined (diverging to infinity)? In the former case, a statistical model includes singularities. There are many such models. Typical examples include the multilayer perceptron. We will study such models in Part IV.

A simple example of the latter type is the location model where x is uniformly distributed in the interval of $[u - 0.5, u + 0.5]$ and u is the unknown parameter. The Fisher information matrix diverges to infinity. In such a statistical model, there is no inner product in the tangent space. The metric is given by a Minkowski metric in the tangent space, which is different from a Riemannian manifold. In this case, M is a Finsler space. An estimator is not asymptotically Gaussian in such a model but is subject to a stable distribution. It is interesting to see the relation between the Finsler metric, stable distribution, and associated fractal structure, comparing them with the Riemannian metric, the Gaussian distribution due to the Central Limit Theorem and the smooth structure of the regular case. However, such a theory has not yet been explored. See a preliminary study by Amari (1984, in Japanese).

Chapter 8
Estimation in the Presence of Hidden Variables

8.1 EM Algorithm

8.1.1 Statistical Model with Hidden Variables

Let us consider a statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$, where vector random variable $\boldsymbol{x}$ is divided into two parts $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{h})$ so that $p(\boldsymbol{x}, \boldsymbol{\xi}) = p(\boldsymbol{y}, \boldsymbol{h}; \boldsymbol{\xi})$. When $\boldsymbol{x}$ is not fully observed but $\boldsymbol{y}$ is observed, $\boldsymbol{h}$ is called a hidden variable. In such a case, we estimate $\boldsymbol{\xi}$ from observed $\boldsymbol{y}$. These situations occur in many applications. One can eliminate the hidden variable $\boldsymbol{h}$ by marginalization such that

$$p_Y(\boldsymbol{y}, \boldsymbol{\xi}) = \int p(\boldsymbol{y}, \boldsymbol{h}; \boldsymbol{\xi}) d\boldsymbol{h}. \tag{8.1}$$

Then, we have a statistical model $M' = \{p_Y(\boldsymbol{y}, \boldsymbol{\xi})\}$ which does not include hidden variables. However, in many cases, the form of $p(\boldsymbol{x}, \boldsymbol{\xi})$ is simple and estimation is tractable in M, but M' is complicated because of integration or summation over $\boldsymbol{h}$. Estimation in such a model is computationally intractable. Typically, M is an exponential family. The EM algorithm is a procedure to estimate $\boldsymbol{\xi}$ by using a large model M from which model M' is derived.

Let us consider a larger model

$$S = \{q(\boldsymbol{y}, \boldsymbol{h})\} \tag{8.2}$$

consisting of all probability density functions of $(\boldsymbol{y}, \boldsymbol{h})$. When both $\boldsymbol{y}$ and $\boldsymbol{h}$ are binary variables, S is a probability simplex so that it is an exponential family. We study the continuous variable case similarly, without considering delicate mathematical problems. Model M is included in S as a submanifold. Observed data give an observed point

$$\bar{q}(\boldsymbol{x}) = \frac{1}{N} \sum \delta(\boldsymbol{x} - \boldsymbol{x}_i) \tag{8.3}$$

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_8

in S when examples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are fully observed. This is the empirical distribution. When S is an exponential family, it is given by the sufficient statistic

$$\bar{\boldsymbol{\eta}} = \bar{\boldsymbol{x}} = \frac{1}{N} \sum \boldsymbol{x}_i \tag{8.4}$$

in the $\boldsymbol{\eta}$-coordinates. The MLE is given by m-projecting $\bar{q}(\boldsymbol{x})$ to M.

We do not have a full observed point $\bar{q}(\boldsymbol{x})$ in the hidden variable case. We observe only $\boldsymbol{y}$ so that we have an empirical distribution $\bar{q}_Y(\boldsymbol{y})$ of $\boldsymbol{y}$ only. In order to have a candidate of a joint distribution $\bar{q}(\boldsymbol{y}, \boldsymbol{h})$, we use an arbitrary conditional distribution $q(\boldsymbol{h}|\boldsymbol{y})$ and put

$$\bar{q}(\boldsymbol{y}, \boldsymbol{h}) = \bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y}). \tag{8.5}$$

Since $q(\boldsymbol{h}|\boldsymbol{y})$ is arbitrary, we take all of them as possible candidates of observed points and consider a submanifold

$$D = \{\bar{q}(\boldsymbol{y}, \boldsymbol{h}) \,|\bar{q}(\boldsymbol{y}, \boldsymbol{h}) = \bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y}),\ q(\boldsymbol{h}|\boldsymbol{y}) \ \text{ is arbitrary}\}\,. \tag{8.6}$$

This is the observed submanifold in S specified by the partially observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$. By using the empirical distribution, it is written as

$$q\,(\boldsymbol{y}, \boldsymbol{h}) = \frac{1}{N} \sum \delta\left(\boldsymbol{y} - \boldsymbol{y}_i\right) q\left(\boldsymbol{h}\,\middle|\,\boldsymbol{y}_i\right) \tag{8.7}$$

The data submanifold D is m-flat, because it is linear with respect to $q(\boldsymbol{h}|\boldsymbol{y})$.

Before analyzing the estimation procedure, we give two simple examples of the hidden variable model.

(1) Gaussian mixture model

Let $N(\mu)$ be a Gaussian distribution of y with mean μ and variance 1. We can treat more general multivariate Gaussian models with unknown covariance matrices in a similar way, but this simple model is enough for the purpose of illustration. The Gaussian mixture model consists of the mixture of k Gaussian distributions having different means $\mu_1, \ldots, \mu_k$,

$$p(y, \boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi}} \sum w_j \exp\left\{-\frac{\left(y - \mu_j\right)^2}{2}\right\}, \tag{8.8}$$

where $\boldsymbol{\xi} = (w_1, \ldots, w_k;\ \mu_1, \ldots, \mu_k)$, $\sum w_i = 1$, are unknown parameters to be estimated. Estimation is easy if, for each $y_1, \ldots, y_N$, we know the Gaussian distribution from which this y_i is generated. So we introduce a hidden variable h, which takes value i when y is generated from the ith distribution $N(\mu_i)$. The h is a random variable, the distribution of which is multinomial, taking value i with probability w_i. Hence, the entire joint distribution is

$$p(y, h, \boldsymbol{\xi}) = \frac{w_h}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y - \mu_h)^2\right\}, \quad h = 1, \ldots, k \tag{8.9}$$

and (8.8) is the marginal distribution of (8.9), obtained by summing h from 1 to k.

(2) Boltzmann machine with hidden units

The Boltzmann machine is a stochastic model having a binary vector random variable $\boldsymbol{x} = (x_1, \ldots, x_n)$. It originates from a model of a spin system in physics and a model of associative memory in machine learning. Consider a Markov chain $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, where state $\boldsymbol{x}_{t+1}$ at time $t + 1$ is stochastically determined from $\boldsymbol{x}_t$. We do not describe here the stochastic dynamics of the state transition, but simply study its stable distribution given by

$$p(\boldsymbol{x}, \boldsymbol{a}, \mathbf{W}) = \exp\left\{\boldsymbol{a} \cdot \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \mathbf{W} \boldsymbol{x} - \psi(\boldsymbol{a}, \mathbf{W})\right\}. \tag{8.10}$$

This is called a Boltzmann machine specified by parameter $\boldsymbol{\xi} = (\mathbf{W}, \boldsymbol{a})$, where an element w_{ij} of matrix $\mathbf{W}$ is regarded as the intensity of connection between units i and j. They are assumed to be symmetric $w_{ij} = w_{ji}$ with $w_{ii} = 0$. The linear term $\boldsymbol{a} \cdot \boldsymbol{x}$ in the exponent is called a bias term, which specifies the tendency of x_i to be 1 rather than 0.

We consider the case where $\boldsymbol{x}$ is divided into two parts, $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{h})$ and $\boldsymbol{y}$ is observable while $\boldsymbol{h}$ is hidden. For the sake of simplicity, we consider the restricted Boltzmann machine (RBM), which consists of two layers, an observable layer and a hidden layer (Fig. 8.1). Connections exist only between units in the observable layer and between units in the hidden layer. No connections exist between units within the observable layer, and no connections exist between units within the hidden layer. Then, the stable distribution is written as

$$p(\boldsymbol{y}, \boldsymbol{h}, \mathbf{W}) = \exp\left\{-\frac{1}{2}\boldsymbol{y}^T \mathbf{W} \boldsymbol{h} - \psi(\mathbf{W})\right\}, \tag{8.11}$$

where, for the sake of simplicity, we ignore the bias term $\boldsymbol{a}$ and let it be 0.

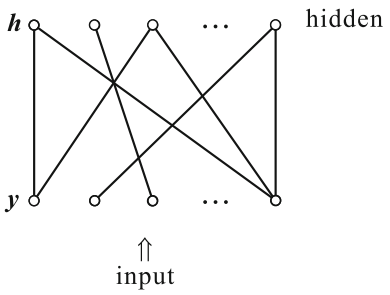


Fig. 8.1 Restricted Boltzmann machine

The marginal distribution of $\boldsymbol{y}$ is

$$p_Y(\boldsymbol{y}, \mathbf{W}) = \sum_{\boldsymbol{h}} \exp\left\{-\frac{1}{2}\boldsymbol{y}^T \mathbf{W}\boldsymbol{h} - \psi(\mathbf{W})\right\}, \tag{8.12}$$

which is a mixture of exponential family distributions. The conditional distribution of $\boldsymbol{h}$ given $\boldsymbol{y}$ is

$$p(\boldsymbol{h}|\boldsymbol{y}; \mathbf{W}) = \frac{p(\boldsymbol{y}, \boldsymbol{h}; \mathbf{W})}{p_Y(\boldsymbol{y}, \mathbf{W})}, \tag{8.13}$$

when the parameters $\mathbf{W}$ are known. This model is used in deep learning and we discuss it in a later chapter from the viewpoint of Bayesian information geometry.

8.1.2 Minimizing Divergence Between Model Manifold and Data Manifold

The MLE is the minimizer of KL-divergence from the observed point $\bar{q}$ to the model manifold in the fully observed case. We have an observed data submanifold D in the hidden case instead of $\bar{q}$. We consider the minimizer of KL-divergence from the data manifold to the model manifold. The problem is then to minimize the divergence between two submanifolds D and M,

$$D_{KL}[D : M] = \min \int \bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y}) \log \frac{\bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y})}{p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi})} d\boldsymbol{y} d\boldsymbol{h}, \tag{8.14}$$

where the minimum between two sets D and M is taken with respect to $\bar{q} \in D$, $p \in M$. The alternating minimization algorithm (*em* algorithm) studied in Chap. 1 is useful for this purpose.

Theorem 8.1 *The MLE is the minimizer of the KL-divergence from D to M.*

Proof The KL-divergence from a distribution $\bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y}) \in D$ to a distribution $p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi}) \in M$ is written as

$$\begin{aligned} D\left[\bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y}) : p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi})\right] = \int \Big[\bar{q}_Y(\boldsymbol{y}) \int q(\boldsymbol{h}|\boldsymbol{y}) \log q(\boldsymbol{h}|\boldsymbol{y}) d\boldsymbol{h} \\ - \bar{q}_Y(\boldsymbol{y}) \int q(\boldsymbol{h}|\boldsymbol{y}) \log p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi}) d\boldsymbol{h}\Big] d\boldsymbol{y} + c, \end{aligned} \tag{8.15}$$

where c is a term not depending on $\boldsymbol{\xi}$ and $q(\boldsymbol{h}|\boldsymbol{y})$. We minimize (8.15) with respect to both $\boldsymbol{\xi}$ and $q(\boldsymbol{h}|\boldsymbol{y})$ alternately by the *em* algorithm, that is, the alternating use of the *e*-projection and *m*-projection. First, assume that $q(\boldsymbol{h}|\boldsymbol{y})$ is given and we minimize (8.15) with respect to $\boldsymbol{\xi}$. We consider one observed $\boldsymbol{y}$ for simplicity, although we

need to consider the expectation with respect to $\bar{q}_Y(\boldsymbol{y})$, which is the summation over all observed $\boldsymbol{y}_i$.

Our task is to maximize the second term of (8.15)

$$L(\boldsymbol{\xi}|q) = \int q(\boldsymbol{h}|\boldsymbol{y}) \log p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi}) d\boldsymbol{h} \tag{8.16}$$

with respect to $\boldsymbol{\xi}$. By differentiating it, the solution is given in the equation

$$\int \frac{q(\boldsymbol{h}|\boldsymbol{y})}{p(\boldsymbol{h}|\boldsymbol{y}, \boldsymbol{\xi})} \frac{\partial}{\partial \boldsymbol{\xi}} p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi}) d\boldsymbol{h} = 0. \tag{8.17}$$

In order to minimize (8.15) with respect to $q(\boldsymbol{h}|\boldsymbol{y})$, we use the following lemma.

Lemma 8.1 *The e-projection from a point of M to D does not alter the conditional distribution* $q(\boldsymbol{h}|\boldsymbol{y})$ *and hence the conditional expectation of* $\boldsymbol{h}$.

Proof Given $\boldsymbol{\xi}$ and observed data $\boldsymbol{y}$, we search for $q(\boldsymbol{h}|\boldsymbol{y})$ that minimizes (8.15). This is to minimize

$$KL\left[\bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y}) : p(\boldsymbol{y}, \boldsymbol{h}; \boldsymbol{\xi})\right] \tag{8.18}$$

under the constraint

$$\int q(\boldsymbol{h}|\boldsymbol{y}) d\boldsymbol{h} = 1. \tag{8.19}$$

The minimizer is given by the e-projection of $p(\boldsymbol{y}, \boldsymbol{h}; \boldsymbol{\xi})$ to D and analytically by solving

$$\int \left[\log \frac{q(\boldsymbol{h}|\boldsymbol{y})}{p(\boldsymbol{h}|\boldsymbol{y}, \boldsymbol{\xi})} - \lambda\right] \delta q(\boldsymbol{h}|\boldsymbol{y}) d\boldsymbol{h} = 0, \tag{8.20}$$

where λ is the Lagrange multiplier corresponding to (8.19). This proves

$$q(\boldsymbol{h}|\boldsymbol{y}) = p\left(\boldsymbol{h}|\boldsymbol{y}; \boldsymbol{\xi}\right), \tag{8.21}$$

which is exactly the same as the conditional probability of $\boldsymbol{h}$ at $\boldsymbol{\xi}$. □

By substituting (8.21) in (8.17), the minimizer of the KL-divergence satisfies

$$\frac{\partial}{\partial \boldsymbol{\xi}} \int p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi}) d\boldsymbol{h} = \frac{\partial}{\partial \boldsymbol{\xi}} p_Y(\boldsymbol{y}, \boldsymbol{\xi}) = 0, \tag{8.22}$$

proving that it is the MLE. □

8.1.3 EM Algorithm

The EM algorithm (expectation maximization algorithm) is an iterative algorithm for obtaining the MLE in a model including hidden variables. It was formulated by Dempster et al. (1977). We show its geometry due to Csiszár and Tusnady (1984), also by Amari et al. (1992), Byrne (1992) and Amari (1995). It is an application of the *em* algorithm from the geometrical point of view. We begin with $\boldsymbol{\xi}_0$ as an initial parameter, and e-project it to D to obtain the conditional distribution $q(\boldsymbol{h}|\boldsymbol{y}) = p\left(\boldsymbol{h}\,\middle|\,\boldsymbol{y}; \boldsymbol{\xi}_0\right)$. This determines a candidate for the observed distribution in D. We calculate the conditional expectation of log likelihood to evaluate the likelihood of a new candidate $\boldsymbol{\xi}$, given by

$$L\left(\boldsymbol{\xi}, \boldsymbol{\xi}_0\right) = \frac{1}{N}\sum_i \int p\left(\boldsymbol{h}|\boldsymbol{y}_i, \boldsymbol{\xi}_0\right) \log p(\boldsymbol{y}_i, \boldsymbol{h}, \boldsymbol{\xi}) d\boldsymbol{h}, \tag{8.23}$$

for observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$. This is called the E-step, because it calculates the conditional expectation. This is the e-projection of $p\left(\boldsymbol{y}, \boldsymbol{R}, \boldsymbol{\xi}_0\right)$ to D.

We then m-project the new candidate in D to M, to obtain a new candidate $\boldsymbol{\xi}_1$ in M. This is obtained by maximizing (8.23). It is called the M-step, because it is the maximization of the log likelihood (8.23). This is the m-projection. We repeat the procedures. See Fig. 8.2 .

It is easy to prove the following theorem.

Theorem 8.2 *The KL-divergence decreases monotonically by repeating the E-step and the M-step. Hence, the algorithm converges to an equilibrium.*

It should be noted that the m-projection is not necessarily unique unless M is e-flat. Hence, there might exist local minima.

8.1.4 Example: Gaussian Mixture

The parameters to be estimated are the weights $w_1, \ldots, w_k$ and the means $\mu_1, \ldots, \mu_k$ of component Gaussian distributions, $\boldsymbol{\xi} = (w_i, \mu_i; i = 1, \ldots, k)$. We begin with

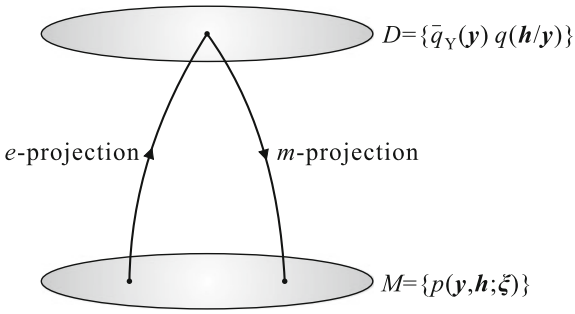


Fig. 8.2 EM algorithm

initial $\boldsymbol{\xi}_0$, and let $\boldsymbol{\xi}^t = \left(w_i^t, \mu_i^t\right)$ be the candidate at t. The E-step is to e-project $p(y, h; \boldsymbol{\xi}^t)$ to D to obtain $q_t(h|y)$. This is the same as that at $\boldsymbol{\xi}^t$,

$$q_t(h|y, \boldsymbol{\xi}^t) = \frac{w_h^t}{\sqrt{2\pi} p\left(y, \boldsymbol{\xi}^t\right)} \exp\left\{-\frac{1}{2}\left(y - \mu_h^t\right)^2\right\}. \tag{8.24}$$

The conditional expectation is

$$L(\boldsymbol{\xi}, \boldsymbol{\xi}^t) = \sum_h p\left(h|y, \boldsymbol{\xi}^t\right)\left\{\log w_h - \frac{1}{2}(y - y_h)^2\right\} \tag{8.25}$$

up to a constant not depending on the parameters.

The M-step is maximization (m-projection) searching for a new $\boldsymbol{\xi}^{t+1}$ that maximizes (8.25). By differentiating (8.25) and making it equal to 0, we easily obtain

$$w_h^{t+1} = \frac{1}{N}\sum p\left(h\left|y_i, \boldsymbol{\xi}^t\right.\right), \quad \mu_h^{t+1} = \frac{\sum_i y_i p\left(h\left|y_i, \boldsymbol{\xi}^t\right.\right)}{\sum_i p\left(h\left|y_i, \boldsymbol{\xi}^t\right.\right)}. \tag{8.26}$$

8.2 Loss of Information by Data Reduction

Given original data $D_X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, assume that we summarize it to a statistic

$$\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \tag{8.27}$$

and use it for estimation. Then, we consider an estimator $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}(\boldsymbol{T})$, which is a function of $\boldsymbol{T}$. When $\boldsymbol{T}$ is a sufficient statistic, there is no loss of information. Otherwise, summarizing the data in $\boldsymbol{T}$ will cause loss of information, which is measured by using the Fisher information. When there is a hidden variable $\boldsymbol{h}$ and we use $\boldsymbol{T} = \left\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\right\}$ for estimation, $\boldsymbol{T}$ is not sufficient in general.

We define the conditional Fisher information of the original data D_X conditioned on $\boldsymbol{T}$. When $\boldsymbol{T} = \boldsymbol{t}$, the probability distribution of D_X is given by the conditional probability $p(D_X|\boldsymbol{t}; \boldsymbol{\xi})$. Hence the Fisher information is given as

$$g_{ij}(\cdot|\boldsymbol{t}; \boldsymbol{\xi}) = \mathrm{E}_X\left[\partial_i \log p(D_X|\boldsymbol{t}; \boldsymbol{\xi})\partial_j \log p(D_X|\boldsymbol{t}; \boldsymbol{\xi})\right], \tag{8.28}$$

where E_X is the conditional expectation of D_X. Taking the average over $\boldsymbol{t}$, we have the conditional Fisher information

$$g_{ij}^{X|T}(\boldsymbol{\xi}) = \mathrm{E}_{\boldsymbol{t}} g_{ij}(\cdot|\boldsymbol{t}; \boldsymbol{\xi}). \tag{8.29}$$

From the equality

$$g_{ij}^X(\boldsymbol{\xi}) = g_{ij}^T(\boldsymbol{\xi}) + g_{ij}^{X|T}(\boldsymbol{\xi}), \tag{8.30}$$

where g_{ij}^X, g_{ij}^T, $g_{ij}^{X|T}$ are the Fisher information based on D_X, $\boldsymbol{T}$ and D_X conditionally on $\boldsymbol{T}$. The loss of Fisher information by summarizing data to statistics $\boldsymbol{T}$ is given by

$$\Delta g_{ij}^T(\boldsymbol{\xi}) = g_{ij}^{X|T}(\boldsymbol{\xi}). \tag{8.31}$$

Oizumi et al. (2011) studied the loss of information in the case of spikes of neurons. Let t firing patterns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ of neurons be observed. These include firing rates of neurons, covariances of spikes of two neurons and higher-order correlations of a number of neurons. Since the brain reduces information in the process of decision making, it loses some information. Consider a curved exponential family

$$p(\boldsymbol{X}, \boldsymbol{\xi}) = \exp\{\boldsymbol{\theta}(\boldsymbol{\xi}) \cdot \boldsymbol{X} - \psi\}, \tag{8.32}$$

where $\boldsymbol{X} = \left(x_i, x_i x_j, \ldots, x_1 \ldots x_n\right)$ and $\boldsymbol{\xi}$ is a parameter to specify the probability distribution based on which $\boldsymbol{x}$ is generated. When a multiple observation is possible, we have the sufficient statistic (observed point)

$$\bar{\boldsymbol{\eta}} = \frac{1}{N} \sum \boldsymbol{X}_i. \tag{8.33}$$

It includes all the information concerning firing rates, pairwise and higher-order interactions. An efficient estimator is obtained by m-projecting it to model M of which the coordinates are $\boldsymbol{\xi}$.

When a part of $\bar{\boldsymbol{\eta}}$ is lost, for example higher-order correlations of spikes are lost, we cannot identify the observed point. We have instead an observed data submanifold D. The optimum estimator is the minimizer of $D_{KL}[D : M]$. The amount of loss of information is calculated when correlational information is lost (Oizumi et al. 2011).

8.3 Estimation Based on Misspecified Statistical Model

When the true statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$ is very complicated, we are apt to use a simplified model $M_q = \{q(\boldsymbol{x}, \boldsymbol{\xi})\}$ to estimate parameter $\boldsymbol{\xi}$. This is a misspecified model. What is the loss of information by using a misspecified model? We begin with a simple example for illustration of the problem. Assume that n neurons are arranged in a one-dimensional neural field. When a stimulus is applied at position u, $0 < u < 1$, the neuron corresponding to that position and neighboring neurons are activated. When the ith neuron corresponds to position

$$u = \frac{i}{n}, \tag{8.34}$$

it is excited strongly, and neighboring neurons are also excited. We assume that, for an arbitrary j, the response of neuron j is $r_j(u)$ when a stimulus is applied at u. The

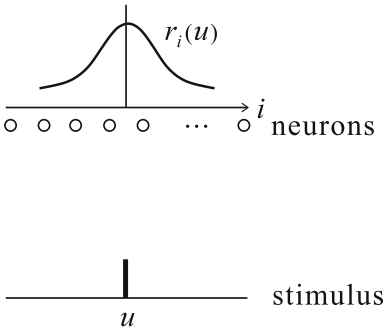


Fig. 8.3 Tuning *curve* of neural field

curve $r_j(u)$ is called the tuning curve of neuron j. See Fig. 8.3. We assume that x_i is the firing rate of neuron i subject to a Gaussian distribution of which the mean is $r_i(u)$ and the covariance matrix is $V = \left(V_{ij}\right)$. Then, the statistical model of excitation is

$$p(\boldsymbol{x}, u) = c \exp\left\{-\frac{1}{2}\left\{\boldsymbol{x} - \boldsymbol{r}(u)\right\}^T V^{-1}\left\{\boldsymbol{x} - \boldsymbol{r}(u)\right\}\right\}. \tag{8.35}$$

Consider a simpler model having the same tuning curves but no correlations,

$$q(\boldsymbol{x}, u) = c \exp\left[-\frac{1}{2}\left\{\boldsymbol{x} - \boldsymbol{r}(u)\right\}^T \left\{\boldsymbol{x} - \boldsymbol{r}(u)\right\}\right]. \tag{8.36}$$

Wu et al. (2002) showed that there is asymptotically no loss of information even if we use the simple misspecified model M_q of (8.36). This is good news for the brain.

We study a general case of a misspecified model to see its loss of information. We consider the case that both $p(\boldsymbol{x}, \boldsymbol{u})$ and $q(\boldsymbol{x}, \boldsymbol{u})$ are curved exponential families lying in a larger exponential family S. The observed point $\bar{\boldsymbol{\eta}}$ is asymptotically subject to the Gaussian distribution with mean $\boldsymbol{\eta}(\boldsymbol{u})$ in the true model M and covariance matrix $\mathbf{G}\left\{\boldsymbol{\eta}(\boldsymbol{u})\right\}/N$ when the true distribution is $p(\boldsymbol{x}, \boldsymbol{u})$. The maximum likelihood estimator using model $M_q = \{q(\boldsymbol{x}, \boldsymbol{u})\}$ is called the q-MLE. The q-MLE is obtained by m-projecting the observed point to M_q by using the q-ancillary family $A_q(\boldsymbol{u})$, which is an m-flat submanifold of S passing through $q(\boldsymbol{x}, \boldsymbol{u})$ and orthogonal to the tangent space of M_q at $\boldsymbol{u}$. Since the observed point converges to $\boldsymbol{\eta}(\boldsymbol{u})$ as N tends to infinity, the q-MLE is consistent when the q-ancillary family passing through $q(\boldsymbol{x}, \boldsymbol{u})$ passes through $p(\boldsymbol{x}, \boldsymbol{u}) \in M$. See Fig. 8.4.

Theorem 8.3 *The q-MLE is consistent when and only when*

$$\mathrm{E}_{p(\boldsymbol{x},\boldsymbol{u})}\left[\partial_a \log q(\boldsymbol{x}, \boldsymbol{u})\right] = 0, \quad \partial_a = \frac{\partial}{\partial u^a}, \tag{8.37}$$

which holds when the q-ancillary family $A_q(\boldsymbol{u})$ passes through $p(\boldsymbol{x}, \boldsymbol{u}) \in M$.

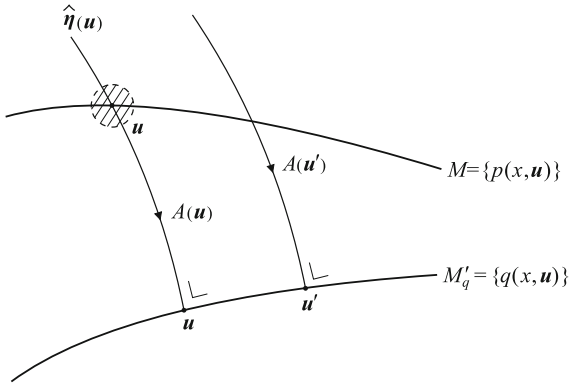


Fig. 8.4 q-auxiliary family and q-MLE

Proof Let

$$r(\boldsymbol{x}, \boldsymbol{u}; t) = (1-t)q(\boldsymbol{x}, \boldsymbol{u}) + tp(\boldsymbol{x}, \boldsymbol{u}) \tag{8.38}$$

be the m-geodesic connecting $q(\boldsymbol{x}, \boldsymbol{u})$ and $p(\boldsymbol{x}, \boldsymbol{u})$. Its tangent vector at M_q is

$$\dot{r} = \frac{d}{dt} \log r(\boldsymbol{x}, \boldsymbol{u}, t)\Big|_{t=0} = \frac{1}{q(\boldsymbol{x}, \boldsymbol{u})} \{q(\boldsymbol{x}, \boldsymbol{u}) - p(\boldsymbol{x}, \boldsymbol{u})\}. \tag{8.39}$$

It is orthogonal to the tangent vectors

$$\dot{l}_q = \frac{\partial}{\partial \boldsymbol{u}} \log q(\boldsymbol{x}, \boldsymbol{u}) \tag{8.40}$$

of M_q, when $\langle \dot{r}, \dot{l}_q \rangle_q = 0$, which is calculated as

$$\begin{aligned} \langle \dot{r}, \dot{l}_q \rangle_q &= \int \{q(\boldsymbol{x}, \boldsymbol{u}) - p(\boldsymbol{x}, \boldsymbol{u})\} \partial_{\boldsymbol{u}} \log q(\boldsymbol{x}, \boldsymbol{u}) d\boldsymbol{x} \\ &= -\int p(\boldsymbol{x}, \boldsymbol{u}) \partial_{\boldsymbol{u}} \log q(\boldsymbol{x}, \boldsymbol{u}) d\boldsymbol{x}. \end{aligned} \tag{8.41}$$

This implies that (8.37) holds and vice versa. □

The q-MLE estimator is Fisher efficient when the m-geodesic connecting $q(\boldsymbol{x}, \boldsymbol{u})$ and $p(\boldsymbol{x}, \boldsymbol{u})$ is orthogonal to both M and M_q, because the ancillary submanifold $A_q(\boldsymbol{u})$ and the true ancillary submanifold $A(\boldsymbol{u})$ of the true MLE coincide. Hence, the observed $\bar{\boldsymbol{\eta}}$ is mapped to the same $\hat{\boldsymbol{u}}$ in both M and M_q by the m-projection. When $A_q(\boldsymbol{u})$ is not orthogonal to M, there is information loss. This is easily evaluated from the angles of the q-ancillary submanifold $A_q(\boldsymbol{u})$ and M.

Theorem 8.4 *The q-MLE estimator is Fisher efficient when the q-ancillary family is orthogonal to M. When it is not orthogonal, the loss of Fisher information is given by*

$$\Delta g_{ab}(\boldsymbol{u}) = g_{a\kappa}(\boldsymbol{u}) g_{b\lambda}(\boldsymbol{u}) g^{\kappa\lambda}(\boldsymbol{u}), \tag{8.42}$$

where v^{κ} is the transversal coordinate system in $A_q(\boldsymbol{u})$.

Proof By using the q-ancillary family, we can map the observed point $\bar{\boldsymbol{\eta}}$ to M_q. This is efficient when and only when $A_q(\boldsymbol{u})$ is orthogonal to the tangent space of M. Otherwise, there is information loss. By using the $(\boldsymbol{u}, \boldsymbol{v})$-coordinates, where $\boldsymbol{u} = (u^a)$ and $\boldsymbol{v} = (v^{\kappa})$ are the coordinates along the ancillary family $A_q(\boldsymbol{u})$, the q-MLE is mapped through it, but this is a non-orthogonal mapping to M. Hence, loss of information occurs, as is given in (7.58) or (8.42).

Remark When the q-ancillary family $A_q(\boldsymbol{u})$ does not pass $p(\boldsymbol{x}, \boldsymbol{u})$, the q-estimator is not consistent. However, when this does not hold, let $\boldsymbol{f}(\boldsymbol{u})$ be the coordinates of M at which $A_q(\boldsymbol{u})$ intersects M. If we reparameterize M_q such that the new parameter of M_q is $\boldsymbol{f}^{-1}(\boldsymbol{u})$, then the consistency always holds.

Remarks

The present short chapter introduces statistical models which are different from a regular model. One is a model with hidden variables, in which some random variables are not observed. The EM algorithm is known in such a model. From the geometrical point of view, it is nothing other than the *em* algorithm, which minimizes the divergence between the model manifold M and data manifold D derived from observed data. This is now a standard method in machine learning. When it was proposed by Csiszár and Tusnady (1984), the paper was rejected by a journal because the reviewer did not admit computationally heavy iterative procedures (I. Csiszár, personal communication). So this remains a conference paper.

Another model is a misspecified model. Its performance is easily understood from geometry, so that it is a good example to show the power of information geometry. The brain might use a misspecified or unfaithful statistical model for decoding information, because the true model is often unknown or too complicated. Therefore, we need to know the performance of the misspecified model. Oizumi et al. (2015) use a misspecified model to evaluate the amount of integrated information to measure the degree of consciousness.

Chapter 9
Neyman-Scott Problem: Estimating Function and Semiparametric Statistical Model

The present chapter studies the famous Neyman–Scott problem, where the number of unknown parameters increases in proportion to the number of observations. The problem gave a shock to the statistics community, because the MLE is not necessarily asymptotically consistent or efficient in this problem. We solve the problem by constructing information geometry of estimating functions. The problem is reformulated in the framework of a semiparametric statistical model, which includes a finite number of parameters of interest and a nuisance parameter of function degrees of freedom. The problem uses a function space but we apply an intuitive description, sacrifying mathematical justification. The results are useful for solving both the semiparametric and Neyman–Scott problems.

9.1 Statistical Model Including Nuisance Parameters

Let us consider a statistical model

$$M = \{p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})\} \tag{9.1}$$

which includes two types of parameters. One is a parameter which we want to estimate, denoted by $\boldsymbol{u}$. This is called the parameter of interest. The other, denoted by $\boldsymbol{v}$, is a parameter the value of which is of no concern to us. It is called a nuisance parameter. We give two examples.

1. Measurement under Gaussian noise:scale problem: Let us measure the weight of a specimen repeatedly by using a scale. The true weight is μ but measurements $x_1, \ldots, x_N$ are independent random Gaussian variables with mean μ and variance σ^2, where σ^2 represents the accuracy of the scale. When we have interest in estimating μ but do not care about σ^2, μ is the parameter of interest and σ^2 is the nuisance parameter. When we are interested in knowing the accuracy σ^2 of the scale but do not care about μ, σ^2 is the parameter of interest and μ the nuisance parameter.

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_9

2. Coefficient of proportionality: We consider a pair (x, y) of Gaussian random variables, where x and y represent the volume and the weight of a specimen, respectively. Here, x is a noisy observation of the volume v of the specimen and y is the noisy observation of its weight uv, where u is the specific gravity of the specimen. We assume that the noises are independent and Gaussian with mean 0 and variance 1. Then, their joint distribution is specified by

$$x \sim N(v, 1), \quad y \sim N(uv, 1). \tag{9.2}$$

When we are interested only in specific gravity u, i.e., the coefficient of proportionality, but do not care about v, u is the parameter of interest and v is the nuisance parameter. The joint probability is written as

$$p(x, y; u, v) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}\left\{(x - v)^2 + (y - uv)^2\right\}\right]. \tag{9.3}$$

The problem is easy, because given observed data $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, we can use the MLE to estimate u and v and simply discard the estimator $\hat{v}$ of the nuisance parameter. Since MLE $\left(\hat{u}, \hat{v}\right)$ is efficient, the estimator $\hat{u}$ is efficient.

Let the Fisher information matrix in the model (9.1) of all the parameters $\boldsymbol{\xi} = (\boldsymbol{u}, \boldsymbol{v})$ be $g_{\alpha\beta}$, where we use suffixes α, β for the entire $\boldsymbol{\xi} = (\xi^{\alpha})$, $a, b, c, \ldots$ for the parameter $\boldsymbol{u} = (u^a)$ of interest and $\kappa, \lambda, \mu, \ldots$ for the nuisance parameter $\boldsymbol{v} = (v^{\kappa})$. The Fisher information matrix is partitioned as

$$g_{a\beta} = \begin{bmatrix} g_{ab} & g_{a\kappa} \\ g_{\lambda b} & g_{\kappa\lambda} \end{bmatrix}, \tag{9.4}$$

where, by putting $l = \log p$,

$$g_{ab} = \mathrm{E}\left[\partial_a l(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \partial_b l\left(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}\right)\right], \tag{9.5}$$

$$g_{a\kappa} = \mathrm{E}\left[\partial_a l(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \partial_{\kappa} l(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})\right], \tag{9.6}$$

$$g_{\kappa\lambda} = \mathrm{E}\left[\partial_{\kappa} l(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \partial_{\lambda} l(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})\right]. \tag{9.7}$$

The asymptotic error covariance of the entire estimator $\hat{\boldsymbol{\xi}} = \left(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}\right)$ is given by using its inverse as

$$\mathrm{E}\left[\left(\hat{\xi}^{\alpha} - \xi^{\alpha}\right)\left(\hat{\xi}^{\beta} - \xi^{\beta}\right)\right] = \frac{1}{N} g^{\alpha\beta}. \tag{9.8}$$

The inverse of the Fisher information matrix is also partitioned as

$$g^{\alpha\beta} = \begin{bmatrix} g^{ab} & g^{a\kappa} \\ g_{\lambda b} & g_{\lambda\kappa} \end{bmatrix}, \tag{9.9}$$

where its (a, b)-part $\left(g^{ab}\right)$ is not the inverse of the (a, b)-part (g_{ab}) of $\left(g_{\alpha\beta}\right)$. It is the (a, b)-part of the inverse $\left(g^{\alpha\beta}\right)$ of $\left(g_{\alpha\beta}\right)$. The two are different and $\left(g^{ab}\right)$ is given by the inverse of

$$\bar{g}_{ab} = g_{ab} - g_{a\kappa} g^{\kappa\lambda} g_{\lambda b}, \tag{9.10}$$

as is clear from the inversion of a partitioned matrix.

We have

$$(\bar{g}_{ab}) \leq (g_{ab}) \tag{9.11}$$

in the sense of a positive-definite matrix, which means that information is lost in the presence of unknown nuisance parameter $\boldsymbol{v}$. This is because, when $\boldsymbol{v}$ is known, the Fisher information is (g_{ab}). Since the covariance of the estimation error, when $\boldsymbol{v}$ is unknown, is asymptotically

$$E\left[\left(\hat{u}^a - u^a\right)\left(\hat{u}^b - u^b\right)\right] = \frac{1}{N}\bar{g}^{ab}, \tag{9.12}$$

$(\bar{g}_{ab})$ is called the efficient Fisher information matrix. The tangent vectors $\boldsymbol{e}_a$ and $\boldsymbol{e}_\kappa$ along the $\boldsymbol{u}$ and $\boldsymbol{v}$ coordinate axes are represented by score functions

$$\boldsymbol{e}_a = \partial_a \log p(\boldsymbol{x}, \boldsymbol{\xi}), \quad \boldsymbol{e}_\kappa = \partial_\kappa \log p(\boldsymbol{x}, \boldsymbol{\xi}). \tag{9.13}$$

Let us project $\boldsymbol{e}_a$ to the space orthogonal to the subspace spanned by $\boldsymbol{e}_\kappa$ (Fig. 9.1). Then, the projected vector is given by

$$\bar{\boldsymbol{e}}_a = \boldsymbol{e}_a - g_{a\lambda} g^{\lambda\kappa} \boldsymbol{e}_\kappa, \tag{9.14}$$

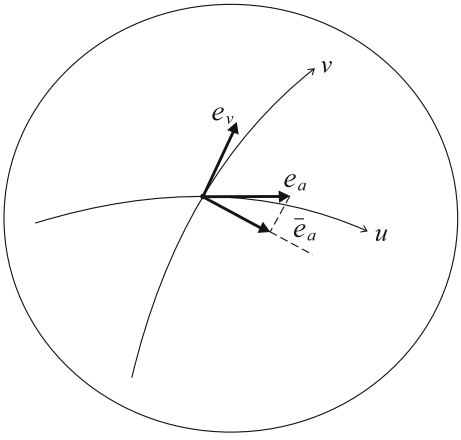


Fig. 9.1 Efficient score $\bar{\boldsymbol{e}}_a$ in the presence of nuisance parameter

or, in terms of the score functions,

$$\bar{\partial}_a l(\boldsymbol{x}, \boldsymbol{\xi}) = \partial_a l(\boldsymbol{x}, \boldsymbol{\xi}) - g_{a\lambda} g^{\lambda\kappa} \partial_\kappa l(\boldsymbol{x}, \boldsymbol{\xi}). \tag{9.15}$$

This is called the efficient score, because the efficient Fisher information matrix is

$$\bar{g}_{ab} = \langle \bar{\boldsymbol{e}}_a, \bar{\boldsymbol{e}}_b \rangle = \mathrm{E}\left[\bar{\partial}_a l \bar{\partial}_b l\right]. \tag{9.16}$$

This shows that only the part orthogonal to the nuisance direction is effective, keeping information for estimating $\boldsymbol{u}$, and the part in the nuisance direction is useless, because $\boldsymbol{v}$ is unknown.

When the subspace spanned by the scores of the parameter of interest is orthogonal to the nuisance parameters, we have $g_{a\kappa} = 0$. In this case,

$$g_{ab} = \bar{g}_{ab} \tag{9.17}$$

holds, so there is asymptotically no loss of information. Therefore, it is desirable to choose the nuisance parameters such that the orthogonality holds. Given a statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})\}$, we consider the problem of reparameterizing $\boldsymbol{v}$ depending on $\boldsymbol{u}$ as

$$\boldsymbol{v}' = \boldsymbol{v}'(\boldsymbol{u}, \boldsymbol{v}) \tag{9.18}$$

such that

$$g_{a\lambda} = \mathrm{E}\left[\partial_a l\left(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}'\right) \partial_\lambda l\left(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}'\right)\right] = 0. \tag{9.19}$$

This is in general impossible (see p. 254 in Amari 1985). However, when $\boldsymbol{u}$ is a scalar parameter, it is always possible.

9.2 Neyman–Scott Problem and Semiparametrics

Neyman and Scott (1948) presented a class of statistical problems and questioned the validity of the MLE, by showing that the asymptotic consistency and efficiency of the MLE are not guaranteed in some of their models. Let $M = \{p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})\}$ be a statistical model and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be N independent observations. The value of $\boldsymbol{u}$ (the parameter of interest) is kept the same (unknown) throughout the observations, but $\boldsymbol{v}$ changes each time. Hence, $\boldsymbol{x}_i$ is subject to $p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}_i)$. This is the Neyman–Scott problem and there are many examples of this type.

The estimation of the radius of the stone circle, Stonehenge in England, is a well-known romantic problem of this type. The stones are supposed to have been arranged in a circle to start with, but their positions have been disturbed in their long history. See Fig. 9.2. The radius u of the stone circle is the parameter of interest, and the declination angle of the ith stones v_i is the nuisance parameter. We show later another problem of estimating the shape parameter from neural spikes under

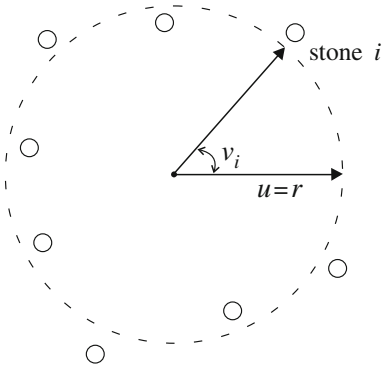


Fig. 9.2 Location of the ith stone

changing firing rates. Independent component analysis, treated in Chap. 13, is also of this type. There are similar problems in computer vision (Kanatani 1998; Okatani and Deguchi 2009).

We use the problem of the coefficient of proportionality as a working example. It consists of N independent observations (x_i, y_i), $i = 1, \ldots, N$, subject to

$$x_i = v_i + \varepsilon_i, \quad y_i = uv_i + \varepsilon_i', \tag{9.20}$$

where ε_i and ε_i' are independent noises subject to Gaussian distributions with mean 0 and common variance σ^2. We assume that σ^2 is known. The joint probability distribution of (x_i, y_i) is

$$p(x_i, y_i, u, v_i) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{(x_i - v_i)^2 + (y_i - uv_i)^2}{2\sigma^2}\right\}. \tag{9.21}$$

Here, u and v_i, are scalar parameters.

Figure 9.3a shows an example of observed data and the problem is to draw a regression line to fit the data. The problem looks very simple, but is not. We show a number of intuitive solutions to this problem.

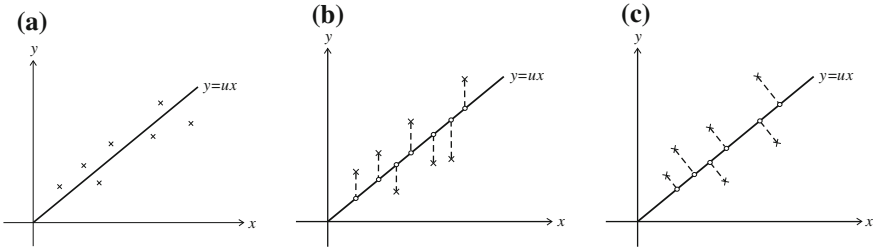


Fig. 9.3 Coefficient of proportionality: **a** Observed data; **b** least squares; **c** total least squares

1. Least squares solution
The least squares solution is the minimizer of

$$L = \frac{1}{2} \sum (y_i - ux_i)^2 , \tag{9.22}$$

which is the sum of the squares of vertical errors to the regression line (Fig. 9.3b). The solution is

$$\hat{u} = \frac{\sum y_i x_i}{\sum x_i^2}. \tag{9.23}$$

However, this is a bad solution. It is not consistent even asymptotically and it does not converge to the correct u even when N increases to become infinitely large.

2. Averaging method
Let $\hat{u}_i = y_i / x_i$ be the ratio obtained from one specimen. Their average

$$\hat{u} = \frac{1}{N} \sum \hat{u}_i \tag{9.24}$$

gives a consistent estimator. This is better than the least squares solution but is not so good in general.

3. Gross average method
Let us sum up all x_i and all y_i separately. Then calculate their ratio,

$$\hat{u} = \frac{\sum y_i}{\sum x_i}. \tag{9.25}$$

This is a good consistent estimator. It is of interest to know how good it is.

4. Total least square solution
Instead of minimizing the vertical errors in the least squares solution, we minimize the square of the lengths of orthogonal projection to the regression line (Fig. 9.3c). This is called the total least squares (TLS) solution. It is given by solving

$$\sum (y_i - ux_i)(uy_i + x_i) = 0. \tag{9.26}$$

5. MLE
We estimate all the parameters $u, v_1, \ldots, v_N$, jointly by maximizing the likelihood, and we disregard all $\hat{v}_i$, keeping $\hat{u}$ only. This is the MLE. We can prove that this is identical with the TLS solution.

We use a semiparametric formulation of the Neyman–Scott problem. Since the sequence $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ is arbitrary and unknown, we assume that it is generated from an unknown probability distribution $k(\boldsymbol{v})$. In order to generate the ith example $\boldsymbol{x}_i$ ((x_i, y_i) in the above example), Nature chooses $\boldsymbol{v}_i$ from distribution $k(\boldsymbol{v})$. Then, $\boldsymbol{x}_i$

is chosen from $p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}_i)$. Thus, each $\boldsymbol{x}_i$ is subject to one and the same probability distribution

$$p(\boldsymbol{x}, \boldsymbol{u}, k) = \int p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) k(\boldsymbol{v}) d\boldsymbol{v}. \tag{9.27}$$

We treat an extended statistical model

$$\tilde{M} = \{p_K(\boldsymbol{x}, \boldsymbol{u}, k)\} \tag{9.28}$$

which includes two parameters: One is $\boldsymbol{u}$, the parameter of interest, and the other is a function $k(\boldsymbol{v})$. Each observation is independently and identically distributed (iid) in this setting, but the underlying model includes the nuisance parameter k of function degrees of freedom. Such a model is called a semiparametric statistical model (Begun et al. 1983). We study the problem under this formulation.

9.3 Estimating Function

An estimating function is a generalization of the score function which is the derivative of the log likelihood and is used to obtain the ML estimator. It is particularly convenient for a model having a nuisance parameter. For a statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})\}$, we consider a differentiable function $f(\boldsymbol{x}, \boldsymbol{u})$ which does not depend on $\boldsymbol{v}$. Here, we treat the case where $\boldsymbol{u}$ and $\boldsymbol{v}$ are scalar parameters for simplicity, but it is easy to generalize it to the case with vector $\boldsymbol{u}$ and vector $\boldsymbol{v}$.

A function $f(\boldsymbol{x}, u)$ is called an estimating function, or more precisely an unbiased estimating function, when

$$\mathrm{E}_{u,v}[f(\boldsymbol{x}, u)] = 0, \tag{9.29}$$

$$\mathrm{E}_{u,v}\left[f\left(\boldsymbol{x}, u'\right)\right] \neq 0, \quad u' \neq u \tag{9.30}$$

hold for any v, where $\mathrm{E}_{u,v}$ is the expectation with respect to $p(\boldsymbol{x}, u, v)$. See Godambe (1991). We further assume

$$\mathrm{E}_{u,v}\left[f'(\boldsymbol{x}, u)\right] \neq 0, \tag{9.31}$$

where f' is the derivative with respect to u. An estimating function of M satisfies

$$\mathrm{E}_{p_K(\boldsymbol{x},u,k)}\left[f\left(\boldsymbol{x}, u'\right)\right] = 0, \text{ when and only when } u' = u, \tag{9.32}$$

for an arbitrary function $k(v)$, when a statistical model M is extended to a semiparametric model $\tilde{M}$ in (9.28). This is because $p_K(\boldsymbol{x}, u, k)$ is a linear mixture of $p(\boldsymbol{x}, u, v)$ with mixing distribution $k(v)$.

The law of large numbers guarantees that the arithmetic mean of $f(\boldsymbol{x}_i, u)$ over the observed data converges to its expectation. Hence, because of (9.29), the solution of

$$\frac{1}{N}\sum f(\boldsymbol{x}_i, u) = 0 \tag{9.33}$$

will give a good estimator; (9.33) is called an estimating equation. In the case of a statistical model without a nuisance parameter, the score function

$$\dot{l}(x, u) = \frac{d}{du}\log p(\boldsymbol{x}, u) \tag{9.34}$$

satisfies (9.29), so it is an estimating function. In this case, (9.33) is the likelihood equation and the derived estimator is the MLE.

We analyze the asymptotic behavior of the estimator derived from an estimating function.

Theorem 9.1 *The estimator $\hat{u}$ derived from an estimating function $f(\boldsymbol{x}, u)$ is asymptotically unbiased and its error covariance is given asymptotically by*

$$E\left[\left(\hat{u} - u_0\right)^2\right] = \frac{1}{N}\frac{\mathrm{E}\left[\{f(\boldsymbol{x}, u_0)\}^2\right]}{\{\mathrm{E}\left[f'(\boldsymbol{x}, u_0)\right]\}^2}, \tag{9.35}$$

when u_0 is the true parameter.

Proof The proof is given by a similar method as the asymptotic analysis of MLE. We expand the left-hand side of (9.33) at u_0,

$$\begin{aligned}&\frac{1}{\sqrt{N}}\sum f\left(\boldsymbol{x}_i, \hat{u}\right)\\ &= \frac{1}{\sqrt{N}}\sum f(\boldsymbol{x}_i, u_0) + \frac{1}{\sqrt{N}}\sum f'(\boldsymbol{x}_i, u_0)\left(\hat{u} - u_0\right).\end{aligned} \tag{9.36}$$

The first term in the right-hand side converges, due to the central limit theorem, to a Gaussian random variable ε with mean 0 and variance

$$\sigma^2 = \mathrm{E}\left[\{f(x, u_0)\}^2\right]. \tag{9.37}$$

The last term of (9.36) converges, due to the law of large numbers, to $\sqrt{N}A$, where

$$A = \mathrm{E}\left[f'(x, u_0)\right] \neq 0. \tag{9.38}$$

Hence, we have

$$\hat{u} - u_0 = \frac{1}{\sqrt{N}}\frac{\varepsilon}{A}, \tag{9.39}$$

from which we have (9.35). □

An estimating function gives an unbiased estimator of which the error covariance converges to 0 in the order of $1/N$. However, there is no guarantee that an estimating function really exists. When does it exist? If there are many estimating functions, how should we choose a good one? These are questions we should address. We use information geometry in answering these questions.

Although we explain the scalar parameter case, our method holds in the vector case. When the parameter $\boldsymbol{u}$ of interest is vector-valued, an estimating function $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ is vector-valued, having the same dimensions as $\boldsymbol{u}$. An $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ is an (unbiased) estimating function when it satisfies

$$\mathrm{E}_{\boldsymbol{u},v}\left[f\left(\boldsymbol{x}, \boldsymbol{u}'\right)\right] \begin{cases} = 0, \ \boldsymbol{u}' = \boldsymbol{u}, \\ \neq 0, \ \boldsymbol{u}' \neq \boldsymbol{u} \end{cases} \tag{9.40}$$

and also the matrix

$$\mathbf{A} = \mathrm{E}_{\boldsymbol{u},v}\left[\frac{\partial}{\partial \boldsymbol{u}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})\right] \tag{9.41}$$

is non-degenerate. The estimating equation is a vector equation

$$\sum \boldsymbol{f}\left(\boldsymbol{x}_i, \boldsymbol{u}\right) = 0. \tag{9.42}$$

The resulting estimator is asymptotically unbiased and Gaussian, having the asymptotic error covariance matrix

$$E\left[\left(\hat{\boldsymbol{u}} - \boldsymbol{u}\right)\left(\hat{\boldsymbol{u}} - \boldsymbol{u}\right)^T\right] = \frac{1}{N}\mathbf{A}^{-1}\mathrm{E}\left[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})\boldsymbol{f}^T(\boldsymbol{x}, \boldsymbol{u})\right]\left(\mathbf{A}^{-1}\right)^T. \tag{9.43}$$

9.4 Information Geometry of Estimating Function

The statistical model $\tilde{M}$ is parameterized by u and $k(v)$, the latter of which has function-degrees of freedom. So we are obliged to use intuitive treatment, not mathematically rigorously justified, but the results are useful. In the function space $F = \{p(\boldsymbol{x})\}$, let us consider a submanifold $M_U(k)$ obtained by fixing the mixing function $k(v)$. It is one-dimensional, that is, it is a curve, having a scalar parameter u. It is denoted by

$$M_U(k) = \{p_K(\boldsymbol{x}, u, k) \mid k \text{ fixed}\}. \tag{9.44}$$

We then consider an infinite-dimensional submanifold

$$M_K(u) = \{p_K(\boldsymbol{x}, u, k) \mid u \text{ fixed}\}, \tag{9.45}$$

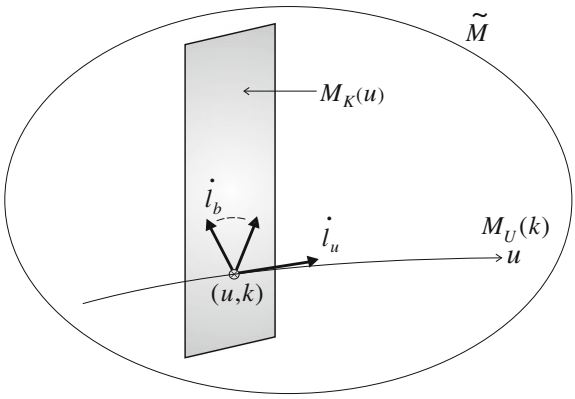


Fig. 9.4 Two submanifolds $M_U(k)$ and $M_K(u)$ and their tangent vectors

where u is fixed but the mixing $k(v)$ is free. One may consider that, for each u, an infinite-dimensional $M_K(u)$ is attached as a fiber. See Fig. 9.4.

The tangent space at a point (u, k) of $\tilde{M}$ is spanned by infinitesimally small deviations $\delta p_K(\boldsymbol{x}, u, k)$ of probability density $p_K(\boldsymbol{x}, u, k)$. By using the logarithmic expression, $l_K(\boldsymbol{x}, u, k) = \log p_K(\boldsymbol{x}, u, k)$, we have

$$\delta l_K(\boldsymbol{x}, u, k) = \frac{\delta p_K(\boldsymbol{x}, u, k)}{p_K(\boldsymbol{x}, u, k)}, \tag{9.46}$$

where

$$\mathrm{E}_{u,k}\left[\delta l_K(\boldsymbol{x}, u, k)\right] = 0, \tag{9.47}$$

$\mathrm{E}_{u,k}$ being the expectation with respect to $p_K(\boldsymbol{x}, u, k)$. This shows that the tangent space $T_{u,k}$ at $(u, k) \in \tilde{M}$ is composed of random variables $r(\boldsymbol{x})$ satisfying

$$\mathrm{E}_{u,k}\left[r(\boldsymbol{x})\right] = 0. \tag{9.48}$$

We assume

$$\mathrm{E}_{u,k}\left[\{r(\boldsymbol{x})\}^2\right] < \infty \tag{9.49}$$

and the inner product of two tangent vectors $r(\boldsymbol{x})$ and $s(\boldsymbol{x})$ are defined by

$$\langle r, s\rangle = \mathrm{E}_{u,k}\left[r(\boldsymbol{x})s(\boldsymbol{x})\right]. \tag{9.50}$$

So the tangent space $T_{u,k}$ is a Hilbert space. An estimating function $f(\boldsymbol{x}, u)$ satisfies (9.48) at any (u, k), so it is a vector belonging to $T_{u,k}$ for any k.

The tangent vector along the u-coordinate axis

$$\frac{d}{du} l_K(\boldsymbol{x}, u, k) = \dot{l}_u(\boldsymbol{x}, u, k) \tag{9.51}$$

satisfies (9.48). The one-dimensional subspace

$$T_U(u,k) = \left\{\dot{l}_u(\boldsymbol{x},u,k)\right\} \tag{9.52}$$

composed of the u-score vector $\dot{l}_u(\boldsymbol{x},u,k)$ is called the tangent subspace of interest at (u,k). In order to define tangent vectors along the nuisance parameter $k(v)$, we consider a curve in the function space of $k(v)$, written as

$$k(v,t) = k(v) + tb(v), \tag{9.53}$$

where

$$\int b(v)dv = 0, \tag{9.54}$$

because

$$\int k(v,t)dv = 1. \tag{9.55}$$

There are infinitely many curves, each specified by $b(v)$. The tangent vector along a curve (9.53) is defined by

$$\dot{l}_b(\boldsymbol{x},u,k) = \frac{d}{dt}\log p_K\left\{\boldsymbol{x},u,k(v,t)\right\}|_{t=0} = \frac{1}{p_K(\boldsymbol{x},u,k)}\int p(\boldsymbol{x},u,v)b(v)dv. \tag{9.56}$$

Let us denote by $T_K(u,k)$ the space spanned by the tangent vectors of all such curves, called the nuisance tangent subspace at (u,k).

Note that there are tangent vectors not belonging to T_U and T_K, which are not included in the directions of change in u or k. We denote the subspace orthogonal to both of T_U and T_K by T_A, which we call an ancillary tangent subspace (Fig. 9.5). Then, the tangent space is decomposed as

$$T = T_U \oplus T_K \oplus T_A \tag{9.57}$$

at each point (u,k), where $\oplus$ implies the direct sum. T_A is orthogonal to $T_U \oplus T_K$, but T_U and T_K are not orthogonal in general.

We define e-parallel transport and m-parallel transport of a tangent vector $r(\boldsymbol{x})$ along the nuisance submanifold $M_K(u)$. We consider a small change of $\log p_K(\boldsymbol{x},u,k)$ in the direction $r(\boldsymbol{x})$,

$$\delta l_K(\boldsymbol{x},u,k) = \varepsilon r(\boldsymbol{x}), \tag{9.58}$$

where ε is small. Since the e-representation of $p_K(\boldsymbol{x},u,k)$ is $l_K(\boldsymbol{x},u,k)$, it is natural to consider that $r(\boldsymbol{x})$ is e-parallelly transported from k to k' without any change. But when $r(\boldsymbol{x}) \in T_{u,k}$, it does not belong to $T_{u,k'}$, because

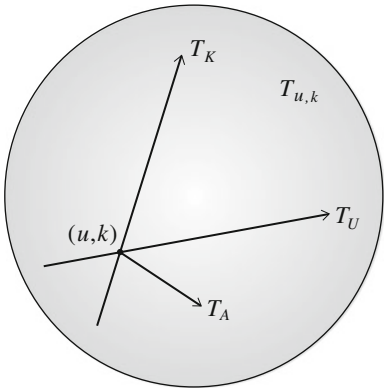


Fig. 9.5 Decomposition of tangent space $\mathrm{T}_{u,k}$

$$\mathrm{E}_{u,k'}\left[r(\boldsymbol{x})\right] \neq 0 \tag{9.59}$$

in general. We subtract the average and define the e-parallel transport of $r(\boldsymbol{x})$ from $p_K(\boldsymbol{x}, u, k)$ to $p_K\left(\boldsymbol{x}, u, k'\right)$ by

$$\prod\nolimits_k^{e\ k'} r(\boldsymbol{x}) = r(\boldsymbol{x}) - \mathrm{E}_{u,k'}[r(\boldsymbol{x})], \tag{9.60}$$

where $\prod_k^{e\ k'}$ is the operator of the e-parallel transport from $k(v)$ to $k'(v)$ in $M_K(u)$. Obviously,

$$\mathrm{E}_{u,k'}\left[\prod\nolimits_k^{e\ k'} r(\boldsymbol{x})\right] = 0. \tag{9.61}$$

We next define the m-parallel transport. Since the m-representation of a deviation of $p(\boldsymbol{x})$ is $\delta p(\boldsymbol{x})$, it is natural to consider that $\delta p(\boldsymbol{x})$ does not change when it is transported in parallel from k to k'. However, its e-representation is

$$\delta l(\boldsymbol{x}) = \frac{\delta p(\boldsymbol{x})}{p_K(\boldsymbol{x}, u, k)}, \tag{9.62}$$

so its e-representation changes at k' as $\delta p(\boldsymbol{x})/p_K\left(\boldsymbol{x}, u, k'\right)$. In order to compensate for this change, we define the m-parallel transport of $r(\boldsymbol{x})$ from k to k' by

$$\prod\nolimits_k^{m\ k'} r(x) = \frac{p_K(\boldsymbol{x}, u, k)}{p_K(\boldsymbol{x}, u, k')} r(\boldsymbol{x}), \tag{9.63}$$

where $\prod_k^{m\ k'}$ is the m-parallel transport operator from k to k'. It satisfies

$$\mathrm{E}_{u,k'}\left[\prod_k^{m\ k'} r(\boldsymbol{x})\right] = 0. \tag{9.64}$$

The two parallel transports are dual, as is shown in the following theorem.

Theorem 9.2 *The e- and m-parallel transports are dual, keeping the inner product invariant:*

$$\langle a(\boldsymbol{x}), b(\boldsymbol{x})\rangle_k = \left\langle \prod_k^{e\ k'} a(\boldsymbol{x}), \prod_k^{m\ k'} b(\boldsymbol{x}) \right\rangle_{k'}. \tag{9.65}$$

The proof is easy from the definitions (9.60) and (9.64).

Lemma *The nuisance tangent space $T_K(u,k)$ is invariant under the m-parallel transport from k to k', where u is fixed.*

Proof Since any tangent vector at k is written in the form of (9.56) by using $b(v)$, it is m-parallelly transported to k' and is written in the same form by using the same $b(v)$, where k is replaced by k'. □

We can now characterize the estimating function in geometrical terms.

Theorem 9.3 *An estimating function is a tangent vector orthogonal to the nuisance tangent space and is invariant under the e-parallel transportation along $M_K(u)$. It includes a non-zero component in the tangent direction T_U of the parameter of interest.*

Proof Because of (9.32),

$$\prod_k^{e\ k'} f(\boldsymbol{x}, u) = f(\boldsymbol{x}, u) \tag{9.66}$$

holds so that it is invariant under the e-parallel transport along the nuisance direction. Let us take a curve $k(v,t)$ and differentiate (9.32) with respect to t. Then we have

$$\int \dot{p}_K(\boldsymbol{x}, u, k(t)) f(\boldsymbol{x}, u) dx = \mathrm{E}\left[\dot{l}_b(\boldsymbol{x}, u, k) f(\boldsymbol{x}, u)\right] = 0. \tag{9.67}$$

Since the nuisance tangent space T_K is spanned by $\dot{l}_b$, f is orthogonal to all the nuisance tangent vectors. We next differentiate (9.32) with respect to u. We then have

$$\mathrm{E}\left[f'(\boldsymbol{x}, u)\right] + \langle \dot{l}_u(\boldsymbol{x}, u, k),\ f(\boldsymbol{x}, u)\rangle = 0. \tag{9.68}$$

Since

$$\mathrm{E}\left[f'(\boldsymbol{x}, u)\right] \neq 0, \tag{9.69}$$

f should include a component in the direction T_U of interest. □

Consider the projection of the score vector $\dot{l}_u(\boldsymbol{x}, u, k)$ to the subspace orthogonal to the tangent space T_K of nuisance parameter and denote it by $\dot{l}_E(\boldsymbol{x}, u, k)$. We call it the efficient score in $\tilde{M}$. Although it depends on $k(v)$, it is an estimating function for any $k(v)$ when it is fixed.

We construct the tangent nuisance space $T_K(\boldsymbol{x}, u, k)$ in terms of the nuisance score

$$\dot{l}_v(\boldsymbol{x}, u, v) = \frac{d}{dv} \log p(\boldsymbol{x}, u, v) \tag{9.70}$$

of M. The tangent nuisance space T_K of $\tilde{M}$ is spanned by the tangent vectors in the directions of $b(v)$ along the curve given by (9.53). Let

$$\delta'_w(v) = \frac{d}{dv}\delta(v - w) \tag{9.71}$$

be the derivative of the delta function. Since $b(v)$ satisfies (9.54), any $b(v)$ is written as a weighted integration of $\delta'_w(v)$,

$$b(v) = \int \delta'_w(v) B(w) dw, \tag{9.72}$$

where the weight is

$$B(w) = -\int_0^w b(v) dv. \tag{9.73}$$

Hence, the tangent vector in the direction of $b(v) = \delta'_w(v)$ is written from (9.56) as

$$\begin{aligned} \dot{l}_{\delta'_w}(\boldsymbol{x}, u, k) &= \frac{-1}{p_K(\boldsymbol{x}, u, k)} \int p(\boldsymbol{x}, u, v) \delta'_w(v) dv \\ &= \frac{p(\boldsymbol{x}, u, w)}{p_K(\boldsymbol{x}, u, k)} \dot{l}_v(\boldsymbol{x}, u, w) \end{aligned} \tag{9.74}$$

by using the nuisance score $l_v(\boldsymbol{x}, u, w)$ of M. Thus, T_K at k is spanned by the m-parallel transports of the elementary tangent scores $l_v(\boldsymbol{x}, u, w)$ for all w and

$$\dot{l}_{\delta'_w}(\boldsymbol{x}, u, k) = \prod_{\delta_w}^{m\ k} \dot{l}_v(\boldsymbol{x}, u, w). \tag{9.75}$$

The following theorem is immediate.

Theorem 9.4 *The nuisance tangent space is m-parallelly invariant,*

$$\prod_{k}^{m\ k'} T_{K,u,k} = T_{K,u,k'} \tag{9.76}$$

and spanned by the m-parallel transports of elementary nuisance scores $\dot{l}_v(\boldsymbol{x}, u, w)$ *for all* w.

Let $f(\boldsymbol{x}, u)$ be an estimating function. It is e-parallelly invariant and orthogonal to T_K. Therefore, because

$$0 = \left\langle f, \prod_{\delta'_w}^{m\ k} \dot{l}_v(\boldsymbol{x}, u, w) \right\rangle = \langle f, \dot{l}_v(\boldsymbol{x}, u, w) \rangle, \tag{9.77}$$

it is orthogonal to the elementary nuisance v-score $\dot{l}_v(\boldsymbol{x}, u, w)$ of M for any $v = w$. In order to obtain the efficient scores in $\tilde{M}$, we consider the tangent vector in the direction of u at a specific point (u, δ_w), where we put $k = \delta_w$. Then, it is the same as the u-score in M,

$$\dot{l}_u(\boldsymbol{x}, u, \delta_w) = \dot{l}_u(\boldsymbol{x}, u, w). \tag{9.78}$$

We construct an efficient score from it, by making it orthogonal to T_K. Since T_K is spanned by all the elementary nuisance scores, we need to project $\dot{l}_u$ to the space orthogonal to all the m-transports of $l_v\left(\boldsymbol{x}, u, w'\right)$ from $\delta_{w'}$ to δ_w for all w'. The projected score is e-invariant, so it is an estimating function. The efficient score $\dot{l}_E(\boldsymbol{x}, u, k)$ at k is constructed by a linear combination with respect to $k(v)$ of these elementary efficient scores.

We have the following theorem from this.

Theorem 9.5 *An estimating function exists when, and only when, the efficient score is non-zero. Any estimating function is written, using an arbitrary nuisance function* $k_0(v)$*, in the form*

$$f(\boldsymbol{x}, u) = \dot{l}_E(\boldsymbol{x}, u, k_0) + a(\boldsymbol{x}), \tag{9.79}$$

where an ancillary tangent vector $a(\boldsymbol{x}) \in T_{A,u,k_0}$ *depends on* k_0.

Proof It is easy to see that $a(\boldsymbol{x})$ is orthogonal to both T_K and T_U. □

Theorem 9.6 *Let* $p_K(\boldsymbol{x}, u, k_0)$ *be the true probability distribution. Then, the best estimating function is* $\dot{l}_E(x, u, k_0)$ *and the asymptotic error covariance is*

$$\mathrm{E}\left(\hat{u} - u\right)^2 = \frac{1}{N} \frac{\mathrm{E}\left[\dot{l}_E^2\right]}{\left\{\mathrm{E}\left[\dot{l}_E\right]\right\}^2}. \tag{9.80}$$

The theorem gives a bound on the asymptotic covariance of error. However, since the true $k_0 = k(v)$ is unknown, we cannot use it. But $\dot{l}_E(\boldsymbol{x}, u, k_1)$ works well even for an approximate value k_1 of k_0. Even when k_1 is quite different from the true one, $\dot{l}_E(\boldsymbol{x}, u, k_1)$ still gives a consistent estimator.

Remark A statistical model in the Neyman–Scott problem is linear in $k(v)$, because it is a mixture model. The nuisance tangent space is invariant under the m-parallel

transport in such a linear model. However, if we study a general semiparametric model where the probability density is not linear with respect to the nuisance function, the tangent nuisance spaces are not invariant by the m-parallel transport. An estimation function is therefore required to be orthogonal to all the tangent nuisance scores at all k. Hence, it is the projection of the u-score vector to the subspace orthogonal to m-transports of the nuisance subspace at all k'. This is called the information score at k. See Amari and Kawanabe (1997).

9.5 Solutions to Neyman–Scott Problems

9.5.1 Estimating Function in the Exponential Case

We consider a typical case where $p(\boldsymbol{x}, u, v)$ is of the exponential type with respect to v, that is,

$$p(\boldsymbol{x}, u, v) = \exp\{vs(\boldsymbol{x}, u) + r(\boldsymbol{x}, u) - \psi(u, v)\}, \tag{9.81}$$

where $s(\boldsymbol{x}, u)$ and $r(\boldsymbol{x}, u)$ are functions of $\boldsymbol{x}$ and u.

Lemma *The u-score at k is given by*

$$\dot{l}_u(\boldsymbol{x}, u, k) = s'(\boldsymbol{x}, u)\mathrm{E}[v|s] + r'(\boldsymbol{x}, u) - \mathrm{E}\left[\psi' | s\right], \tag{9.82}$$

where $\mathrm{E}[\cdot|s]$ *is the conditional expectation conditioned on s.*

Proof We calculate the u-score by differentiating the logarithm of (9.27) with respect to u. By taking (9.81) into account,

$$\begin{aligned}\dot{l}_u(\boldsymbol{x}, u, k) = \frac{1}{p_K(\boldsymbol{x}, u, k)} \int \left\{vs'(\boldsymbol{x}, u) + r'(\boldsymbol{x}, u) - \psi'\right\} \\ \times \exp\{vs + r - \psi\} k(v) dv,\end{aligned} \tag{9.83}$$

where s', r' and ψ' are derivatives of s, r and ψ with respect to u. Since v is a random variable subject to $k(v)$, we consider the joint probability of v and $s(\boldsymbol{x}, u)$. Then, we have the conditional distribution of v conditioned on $s(\boldsymbol{x}, u)$,

$$p(v|s) = \frac{k(v)\exp\{vs + r - \psi\}}{\int k(v)\exp\{vs + r - \psi\} dv} = \frac{k(v)\exp\{vs + r - \psi\}}{p_K(\boldsymbol{x}, u, k)}. \tag{9.84}$$

Hence, we have from (9.83)

$$\dot{l}_u = s'\mathrm{E}[v|s] + r' - \mathrm{E}\left[\psi'|s\right]. \tag{9.85}$$

□

The tangent direction corresponding to a change of k by δk is written as

$$\delta l_K(\boldsymbol{x}, u, k) = \frac{\int p(\boldsymbol{x}, u, v)\delta k(v)}{p_K(\boldsymbol{x}, u, k)} dv. \tag{9.86}$$

Hence, by putting $b(v) = \delta'_w(v)$ and using (9.74), the tangent nuisance space is spanned by

$$\delta_w l(\boldsymbol{x}, u, k) = \varepsilon \frac{p(\boldsymbol{x}, u, w)}{p_K(\boldsymbol{x}, u, k)} \dot{l}_v(\boldsymbol{x}, u, w) \tag{9.87}$$

for all w, which corresponds to a change of $k(v)$ at w. The score corresponding to a change $\delta k(v)$ in the nuisance function $k(v)$ is similarly written in the form of conditional expectation by using (9.84),

$$\dot{l}_K(\boldsymbol{x}, u, k) = \mathrm{E}\left[\frac{\delta k(v)}{k(v)} \,|s \right]. \tag{9.88}$$

This is a function of $s(\boldsymbol{x}, u)$. Hence, the nuisance subspace is generated by $s(\boldsymbol{x}, u)$ and is written as

$$T_K = [h\{s(\boldsymbol{x}, u)\}], \tag{9.89}$$

by using an arbitrary function h of s. We finally have the following theorem.

Theorem 9.7 *The efficient score at k is given by*

$$\dot{l}_E = \mathrm{E}[v|s]\left\{s'(x, u) - \mathrm{E}\left[s' \,|s\right]\right\} + \left\{r'(x, u) - \mathrm{E}\left[r' \,|s\right]\right\}. \tag{9.90}$$

Proof The efficient score is the projection of the score of interest to the subspace orthogonal to the nuisance tangent space. Since, for two random variables s and t, $t - \mathrm{E}[t|s]$ is the projection of t to the subspace orthogonal to the space generated by s, we have the theorem. □

Corollary *When the derivative of s with respect to u is a function of s, we have*

$$\dot{l}_E(\boldsymbol{x}, u, k) = r'(\boldsymbol{x}, u) - \mathrm{E}\left[r'|s\right]. \tag{9.91}$$

Proof In this case,

$$s' - \mathrm{E}\left[s' \,|s\right] = 0, \tag{9.92}$$

which gives (9.91). Since (9.91) does not depend on k, this gives the asymptotically optimal estimating function. □

9.5.2 *Coefficient of Linear Dependence*

After a long journey, we can now solve specific Neyman–Scott problems. The first is the problem of linear dependence. The problem stated in (9.20) is of the exponential type, so it is written in the form of (9.81), where

$$s(\boldsymbol{x}, u) = x + uy \tag{9.93}$$

$$r(\boldsymbol{x}, u) = -\frac{1}{2}\left(x^2 + y^2\right). \tag{9.94}$$

Since r does not depend on u, the efficient score is given as

$$\dot{l}_E = \frac{1}{1+u^2}(y - ux)\mathrm{E}\left[v|s\right]. \tag{9.95}$$

We put

$$\mathrm{E}[v|s] = h(s) = h(uy + x). \tag{9.96}$$

Then, we have a class of estimating functions written as

$$f(\boldsymbol{x}, u) = (y - ux)h(uy + x), \tag{9.97}$$

where h is an arbitrary function.

When the true nuisance function is k, the best $h(s)$ is given by

$$h(s) = \mathrm{E}_{u,k}[v|s], \tag{9.98}$$

which depends on the unknown k. The point is that, even when we do not know k, an estimating function in the class (9.97) gives a consistent estimator of which the error covariance decreases in proportion to $1/N$.

The TLS estimator is obtained by putting

$$h(s) = s. \tag{9.99}$$

The gross average estimator is obtained from

$$h(s) = c, \tag{9.100}$$

where c is a constant. Let us consider a simple linear function

$$h(x) = s + c, \tag{9.101}$$

which will give a better estimator than the two above by choosing c adequately. The estimating equation is

$$\sum (y_i - ux_i)(uy_i + x_i + c) = 0. \tag{9.102}$$

Let $\hat{u}_c$ be the solution of (9.102). Then we have

$$\begin{aligned} &\mathrm{E}\left[(\hat{u}_c - u)^2\right] \\ &= \frac{(1+u^2)\{c + (1+u^2)\bar{v}\}^2 + (1+u^2)^2\left\{\bar{v^2} - (\bar{v})^2\right\} + (1+u^2)}{c\bar{v} + (1+u^2)\bar{v^2}}, \end{aligned} \tag{9.103}$$

where

$$\bar{v} = \frac{1}{N}\sum v_i, \quad \bar{v^2} = \frac{1}{N}\sum v_i^2. \tag{9.104}$$

Therefore, the error is minimized by choosing

$$\hat{c} = \frac{\bar{v}}{\bar{v^2} - (\bar{v})^2}. \tag{9.105}$$

This shows that, when the distribution of $k(v)$ is wide-spread, the TLS is a good estimator, whereas, when the distribution of $k(v)$ is tight, the gross average estimator is better.

9.5.3 Scale Problem

There are two versions in the scale problem. One is to estimate the accuracy of a scale by using N specimens. The other is to estimate the weight of a specimen by using N scales of different accuracies.

1. Accuracy of a scale: We prepare N specimens of which weights $v_1, \ldots, v_N$ are different and unknown. Our aim is to estimate the error variance σ^2 of a scale. When the weight is v and error variance is σ^2, the measurement x is a random variable subject to $N(v, \sigma^2)$. We repeat measurements m times for each specimen. Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ be m measurements by a specimen. The probability density of $\boldsymbol{x}$ is

$$p(\boldsymbol{x}; \mu, \sigma^2) = \exp\left\{-\frac{\sum (x_i - \mu)^2}{2\sigma^2} - \psi\right\}, \tag{9.106}$$

which can be rewritten as

$$p(\boldsymbol{x}, u, v) = \exp\left\{vs(\boldsymbol{x}, u) - \frac{u}{2}r(\boldsymbol{x}, u) - \psi\right\}, \tag{9.107}$$

where we put

$$u = \frac{1}{\sigma^2}, \quad v = \mu \tag{9.108}$$

$$s(\boldsymbol{x}, u) = u\bar{x}, \quad r(\boldsymbol{x}, u) = -\frac{u}{2}\bar{x^2} \tag{9.109}$$

$$\bar{x} = \sum_{i=1}^{m} x_i, \quad \bar{x^2} = \sum_{i=1}^{m} x_i^2. \tag{9.110}$$

Since

$$s'(\boldsymbol{x}, u) = \frac{1}{u} s(\boldsymbol{x}, u) \tag{9.111}$$

is a function of s, the efficient score is

$$\dot{l}_E(\boldsymbol{x}, u) = r' - E\left[r'|s\right]. \tag{9.112}$$

This is the orthogonal projection of $\bar{x^2}$ to the subspace orthogonal to $\bar{x}$. The estimating function is

$$\dot{l}_E(\boldsymbol{x}, u) = \frac{1}{u} - \frac{1}{m-1}\left(\bar{x^2} - \frac{1}{m}\bar{x}^2\right), \tag{9.113}$$

which does not include k. Hence, this gives an efficient estimator,

$$\hat{\sigma}^2 = \frac{1}{N}\sum \hat{\sigma}_i^2, \tag{9.114}$$

$$\hat{\sigma}_i^2 = \frac{1}{m-1}\left[\left(\bar{x^2}\right)_i - \frac{1}{m}(\bar{x})_i^2\right]. \tag{9.115}$$

This is the best estimator different from the MLE. When the numbers m_i of measurements are different, we can solve the problem in a similar way.

2. Weight of a specimen by using N scales: We next consider the case in which we have N scales having different unknown error covariances. In this case, we have only one specimen, the weight of which we want to know. We measure its weight m times by using each scale. In this case, we put

$$u = \mu, \quad v_i = \frac{1}{\sigma_i^2}, \tag{9.116}$$

so, for one scale, the probability density is

$$p(\boldsymbol{x}, u, v) = \exp\left\{-\frac{v}{2}\sum (x_i - u)^2 - \psi\right\}. \tag{9.117}$$

In this case, we have

$$s = -\frac{1}{2}\sum_{i=1}^{k}(x_i - u)^2 = -\frac{1}{2}\left(\bar{x^2} - 2u\bar{x} + u^2\right), \tag{9.118}$$

$$r = 0. \tag{9.119}$$

We can check that s' is orthogonal to s, so the efficient score is

$$\dot{l}_E(\boldsymbol{x}, u) = (\bar{x} - u)\, h(s), \tag{9.120}$$

where h is an arbitrary function. If we fix $h(s)$, then the estimator is

$$\hat{u} = \frac{\sum h(s_i)\,\bar{x}_i}{\sum h(s_i)}. \tag{9.121}$$

The optimum h depends on the unknown $k(v)$,

$$h(s) = \mathrm{E}[v|s], \tag{9.122}$$

but any h will give an asymptotically consistent estimator. It is a surprise that this simple problem has such a complicated structure.

9.5.4 Temporal Firing Pattern of Single Neuron

Let us consider a single neuron which fires stochastically. We assume that it fires at time $t_1, t_2, \ldots, t_{n+1}$, which are random variables. The intervals of spikes are

$$T_i = t_{i+1} - t_i, \quad i = 1, 2, \ldots. \tag{9.123}$$

Obviously, when the firing rate is high, the interval is short. The simplest model of a temporal firing pattern is that all T_i are independent, subject to the exponential distribution

$$q(T, v) = v \exp\{-vT\}, \tag{9.124}$$

where v is the firing rate. The number of spikes is subject to the Poisson distribution. However, T_i are not independent in reality, because of the effect of refractoriness. It is known that the gamma distribution fits well,

$$p(T, v, \kappa) = \frac{(v\kappa)^{\kappa}}{\Gamma(\kappa)} T^{\kappa-1} \exp\{-v\kappa T\}, \tag{9.125}$$

which includes another parameter κ, called the shape parameter. We want to know κ, which is the parameter of interest, so we put $u = \kappa$, whereas v is the nuisance parameter. The average and variance of T_i are

$$E[T] = \frac{1}{v}, \quad \mathrm{Var}[T] = \frac{1}{\kappa v^2}. \tag{9.126}$$

The parameter κ represents the irregularity of spike intervals. When κ is large, the spikes are emitted regularly and have almost the same intervals. When $\kappa = 1$, T_i are independent, and when κ is small, the irregularity increases.

Given observed data $\{T_1, \ldots, T_n\}$, it is easy to estimate the parameters κ and v. This is a simple problem of estimation. However, in a real experimental situation, the firing rate v changes over time but the shape parameter κ is fixed, depending on the type of the neuron. So we regard v as a nuisance parameter changing over time, while κ is the parameter of interest. This is a typical Neyman–Scott problem.

We assume that v takes the same value for two consecutive times. (It can be m consecutive times for $m \geq 2$, but we consider the simplest case.) So we collect two observations T_{2k-1} and T_{2k}, and put them in a box. Hence, the kth observation is $\boldsymbol{T}_k = (T_{2k-1}, T_{2k})$. The two intervals T_{2k-1}, T_{2k} in a box are subject to the same distribution

$$p\left(\boldsymbol{T}_k, v_k, \kappa\right) = \prod_{i=1}^{m} p\left(T_i, v_k, \kappa\right), \tag{9.127}$$

where v_k may change arbitrarily in each box.

We calculate the u-score and v-score as

$$u(\boldsymbol{T}) = \frac{\partial \log p\left(\boldsymbol{T}, v, \kappa\right)}{\partial \kappa}, \quad v(\boldsymbol{T}) = \frac{\partial \log q(\boldsymbol{T}, v, \kappa)}{\partial v}. \tag{9.128}$$

The efficient score is obtained in this case after calculations (see Miura et al. 2006) as

$$u_E(\boldsymbol{T}, \kappa) = \sum \log T_i - m \log \left(\sum T_i \right) + m\phi(mk) - m\phi(\kappa), \tag{9.129}$$

where

$$\phi(\kappa) = \frac{d}{d\kappa} \Gamma(\kappa) \tag{9.130}$$

is the di-gamma function. Since this does not include v, it is the best estimating function, and the estimating equation is

$$\sum u_E\left(\boldsymbol{T}_i, \kappa\right) = 0. \tag{9.131}$$

The statistics used in the estimating function is summarized as

$$S = -\frac{1}{n-1}\sum_i \frac{1}{2}\log\frac{4T_iT_{i+1}}{(T_i+T_{i+1})^2}, \tag{9.132}$$

which includes all the information.

Shinomoto et al. (2003) proposed to use another statistic:

$$L_V = 3 - \frac{12}{n-1}\sum \frac{T_iT_{i+1}}{(T_i+T_{i+1})^2}. \tag{9.133}$$

Interestingly, the two statistics are derived from the same two consecutive time intervals,

$$\frac{4T_iT_{i+1}}{(T_i+T_{i+1})^2}. \tag{9.134}$$

The statistic in (9.132) is the geometric mean, whereas Shinomoto's L_v is the arithmetic mean. From the point of the efficiency of estimation, S is theoretically the best but L_v may be more robust.

Remarks

The Neyman–Scott problem is an interesting estimation problem. It looks simple, but it is very difficult to obtain an optimal solution. Statisticians have struggled with this problem for many years, searching for the optimal solution. In 1984, when Sir David Cox visited Japan and talked about this problem as one of the interesting unsolved problems. I thought it a good challenge for information geometry. It would be wonderful if information geometry could provide a good answer to it.

It is related to a more general semiparametric problem. Since we need a function space of infinite dimensions, it is difficult to construct a mathematically rigorous theory. Bickel et al. (1994) established a rigorous theory of semiparametric estimation by using functional analysis. Information geometry could be more transparent in understanding the structure of the Neyman–Scott problem. We were successful in obtaining a complete set of the estimating functions.

The information-geometric theory is useful, even though the rigorous mathematical foundation is missing. It can solve many famous Neyman–Scott problems. When my official retirement time from the University of Tokyo was approaching, I thought that the results should be publicised even though they include mathematical flaw. So we submitted a paper to Bernoulli. The reviewers pointed out the lack of mathematical justification in the function space. However, the editor Ole Barndorff-Nielsen considered that this was an interesting and useful paper even without rigorous justification being given. So he decided that it was acceptable in the spirit of experimental mathematics, provided that the Theorem–Proof style of statements was replaced by the Proposition and Outline of Proof style.

We did not have many good examples at that time. But later, we found many examples, including neural spike analysis and independent component analysis, the latter of which will be shown in Chap. 13.

Chapter 10
Linear Systems and Time Series

A time series is a sequence of random variables $x_t, t = \ldots, -1, 0, 1, 2, \ldots$, which appears as a function of time. The present chapter deals with an ergodic time series which is generated by a linear system when white noise is applied to its input. We study the geometrical structure of the manifold of the time series. One may identify a time series with a linear system to generate it. Then, the geometry of the time series is identified with the geometry of linear systems, which is important for studying problems of control. For the sake of simplicity, we study only stable systems of discrete time, having one–input and one–output, but generalization is not difficult in principle. The set of all time series has infinite-dimensional degrees of freedom, so our treatment is intuitive and not mathematically rigorous, although it is well-founded in the case of finite-dimensional systems and related time series.

10.1 Stationary Time Series and Linear System

Let us consider a time series $\{x_t\}$, where t denotes discrete time, $t = 0, \pm 1, \pm 2, \ldots$. White Gaussian noise $\{\varepsilon_t\}$ is one of the simplest, which is composed of independent Gaussian random variables with mean 0 and variance 1, so that

$$\mathrm{E}\left[\varepsilon_t \varepsilon_{t'}\right] = 0, \quad t \neq t', \quad \mathrm{E}\left[\varepsilon_t^2\right] = 1. \tag{10.1}$$

We assume that the mean of x_t is equal to 0. A time series is stationary when the probability of $\{x_t\}$ is the same as its time-shifted version $\{x_{t+\tau}\}$ for any τ. More strongly, we consider an ergodic time series.

Ergodic Theorem: For an ergodic time series $\{x_t\}$, the temporal average of a function $f(x_t)$ of x_t converges to the ensemble average with probability 1,

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_10

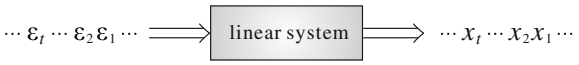


Fig. 10.1 Linear system and generated time series

$$\lim_{T\to\infty} \frac{1}{(2T+1)} \sum_{i=-T}^{T} f(x_i) = \mathrm{E}[f(x_t)]. \tag{10.2}$$

We consider a discrete time linear system, which transforms an input time series into an output time series linearly (Fig. 10.1). When the input is white Gaussian $\{\varepsilon_t\}$, the output $\{x_t\}$ is written as a linear combination of inputs,

$$x_t = \sum_{i=0}^{\infty} h_i \varepsilon_{t-i}. \tag{10.3}$$

A system is characterized by the sequence of parameters,

$$\boldsymbol{h} = (h_0, h_1, h_2, \ldots), \tag{10.4}$$

called the impulse responses of the system. It is assumed that

$$\sum h_i^2 < \infty, \tag{10.5}$$

because

$$E\left[x_t^2\right] = \sum_{i=0}^{\infty} h_i^2. \tag{10.6}$$

The output series is stationary when the input is.

We introduce a time-shift operator z by

$$z\varepsilon_t = \varepsilon_{t+1}, \quad z^{-1}\varepsilon_t = \varepsilon_{t-1}. \tag{10.7}$$

Then, (10.3) is written as

$$x_t = \sum_{i=0}^{\infty} h_i z^{-i} \varepsilon_t. \tag{10.8}$$

By defining

$$H(z) = \sum_{i=0}^{\infty} h_i z^{-i}, \tag{10.9}$$

the output is written as

$$x_t = H(z)\varepsilon_t. \tag{10.10}$$

$H(z)$ is called the transfer function of the system when it is considered as a function of a complex number z, rather than the time shift operator. We assume that $H(z)$ is analytic in the region $|z| \geq 1$.

We define the Fourier transform of an ergodic time series $\{x_t\}$ in the wide sense by

$$X(\omega) = \lim_{T\to\infty} \frac{1}{\sqrt{2T}} \sum_{t=-T}^{T} x_t e^{-i\omega t}. \tag{10.11}$$

Then, $X(\omega)$ is a complex-valued random function of frequency ω. Its absolute value

$$S(\omega) = |X(\omega)|^2 \tag{10.12}$$

is called the power spectrum and is a deterministic function of ω, but the phase of $X(\omega)$ is random, uniformly distributed over $[-\pi, \pi]$. We assume

$$\int_{-\pi}^{\pi} |\log S(\omega)|^2 \, d\omega < \infty. \tag{10.13}$$

The power spectrum of $\{x_t\}$ is written using the transfer function as

$$S(\omega) = \left|H\left(e^{i\omega}\right)\right|^2. \tag{10.14}$$

Conversely, given a time series $\{x_t\}$ having power spectrum $S(\omega)$, we want to identify a system $H(z)$. Such a system exists but is not unique. When $H(z) \neq 0$ outside the unit circle of z (that is $|z| > 1$), such a system is uniquely determined. It is a system of minimal phase. Under this condition, there is one-to-one correspondence among the set of ergodic time series, the set of power spectra $S(\omega)$ and the set of transfer functions $H(z)$. They form an infinite-dimensional manifold L. We will show their coordinates later.

10.2 Typical Finite-Dimensional Manifolds of Time Series

We give typical examples of finite-dimensional systems or time series.

1. AR model

An auto-regressive (AR) model is a time series generated from white noise $\{\varepsilon_t\}$ by

$$a_0 x_t = -\sum_{i=1}^{p} a_i x_{t-i} + \varepsilon_t, \quad a_0 \neq 0. \tag{10.15}$$

This is an AR model of degree p, denoted by $AR(p)$, where x_t is a linear combination (weighted sum) of past p values $x_{t-1}, \ldots, x_{t-p}$ added to by a new Gaussian noise ε_t called innovation. A system is specified by $p+1$ parameters $\boldsymbol{a} = \big(a_0, a_1, \ldots, a_p\big)$.

The transfer function is

$$H(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}} \tag{10.16}$$

and the power spectrum is

$$S(\omega; \boldsymbol{a}) = \left| \sum_{t=0}^{p} a_t e^{i\omega t} \right|^{-2}. \tag{10.17}$$

2. MA model

A moving-average (MA) model of degree q is a time series generated from white noise by

$$x_t = \sum_{i=1}^{q} b_i \varepsilon_{t-i}, \tag{10.18}$$

where $\boldsymbol{b} = \big(b_1, \ldots, b_q\big)$ are the parameters. The present x_t is given by a weighted average of past q noise values. Its transfer function and power spectrum are

$$H(z) = \sum_{i=1}^{q} b_i z^{-i}, \tag{10.19}$$

$$S(\omega) = \left| \sum b_t e^{i\omega t} \right|^2, \tag{10.20}$$

respectively.

3. ARMA model

An ARMA model of degrees (p, q) is the concatenation of AR and MA models, given by

$$x_t = -\sum_{i=0}^{p} a_i x_{t-i} + \sum_{i=1}^{q} b_i \varepsilon_{t-i}. \tag{10.21}$$

Its transfer function and power spectrum are, respectively, given by

$$H(z) = \frac{\sum_{i=1}^{q} b_i z^{-i}}{\sum_{i=0}^{p} a_i z^{-i}}, \tag{10.22}$$

$$S(\omega; \boldsymbol{a}, \boldsymbol{b}) = \left| \frac{\sum b_t e^{i\omega t}}{\sum a_t e^{i\omega t}} \right|^2. \tag{10.23}$$

The above three are of frequent use in time series analysis. The transfer functions are rational functions of z^{-1}.

A continuous-time version of a linear system is used in control systems theory, where time t is continuous and the time-shift operator z is replaced by differential operator $s = d/dt$. The input–output relation of a system is described by

$$x(t) = H(s)u(t) \tag{10.24}$$

for input $u(t)$. Information geometry gives a similar theory to it.

10.3 Dual Geometry of System Manifold

We introduce a Riemannian metric and dually flat affine connections to the manifold L of linear systems. Since L is infinite-dimensional, our theory is intuitive. The Fourier transform $X(\omega)$ of $\{x_t\}$ gives complex-valued Gaussian random variables indexed by frequency ω. We can prove that $X(\omega)$ and $X\left(\omega'\right)$ are independent when $\omega \neq \omega'$ so that we have

$$\mathrm{E}\left[\left|X(\omega)X\left(\omega'\right)\right|\right] = \begin{cases} S(\omega), & \omega' = \omega, \\ 0, & \omega' \neq \omega. \end{cases} \tag{10.25}$$

For complex random variable $X(\omega)$ of (10.11), the phase is uniformly distributed. Therefore, we may write its probability density as

$$p(X; S) \approx \exp\left\{-\frac{1}{2}\int_{-\pi}^{\pi} \frac{|X(\omega)|^2}{S(\omega)} d\omega - \psi(S)\right\}. \tag{10.26}$$

This is an exponential family, where random variables are $X(\omega)$ and the natural parameter indexed by ω is

$$\theta(\omega) = \frac{1}{S(\omega)}. \tag{10.27}$$

This is e-flat coordinates and the expectation parameter is

$$\eta(\omega) = -\frac{1}{2}E\left[|X(\omega)|^2\right] = -\frac{1}{2}S(\omega), \tag{10.28}$$

which is m-flat coordinates. We rewrite the probability density in the form

$$p(X; \theta) = \exp\left\{-\frac{1}{2}\int \theta(\omega)\,|X(\omega)|^2\, d\omega - \psi(\theta)\right\}. \tag{10.29}$$

Two dually coupled potential functions are

$$\psi(\theta) = \frac{1}{2}\int \log\{-\theta(\omega)\}\, d\omega - \frac{\pi}{2} = \frac{1}{2}\int \log S(\omega) d\omega - \frac{\pi}{2}, \tag{10.30}$$

$$\varphi(\eta) = -\frac{1}{2}\int \log\{-2\eta(\omega)\}\, d\omega - \frac{\pi}{2} = -\frac{1}{2}\int \log S(\omega) d\omega - \frac{\pi}{2} \tag{10.31}$$

and they satisfy

$$\psi(\theta) + \varphi(\eta) - \int \theta(\omega)\eta(\omega) d\omega = 0. \tag{10.32}$$

The Riemannian metric is calculated from (10.30) by differentiation,

$$g\left(\omega, \omega'\right) = \frac{\partial^2}{\partial\theta(\omega)\partial\theta\left(\omega'\right)}\psi(\theta), \tag{10.33}$$

so that we have

$$g\left(\omega, \omega'\right) = \begin{cases} \frac{1}{2}S^2(\omega), & \omega' = \omega, \\ 0, & \omega' \neq \omega. \end{cases} \tag{10.34}$$

This is diagonal and hence the squared length of deviation $\delta\theta(\omega)$ is written as

$$\|\delta\theta(\omega)\|^2 = \frac{1}{2}\int S^2(\omega)\left\{\delta\theta(\omega)\right\}^2 d\omega, \tag{10.35}$$

or in terms $S(\omega)$

$$\|\delta S(\omega)\|^2 = \frac{1}{2}\int \frac{\{\delta S(\omega)\}^2}{S^2(\omega)} d\omega = \frac{1}{2}\int \{\delta \log S(\omega)\}^2 d\omega. \tag{10.36}$$

Hence the metric is Euclidean.

The KL-divergence between two systems is written, using their power spectra, as

$$KL\left[S_1 : S_2\right] = D_{-1}\left[S_1 : S_2\right] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{S_1}{S_2} - 1 - \log\frac{S_1}{S_2}\right) d\omega. \tag{10.37}$$

The Shannon entropy is given by

$$H_S = \frac{1}{4\pi}\int \log S(\omega) d\omega + \frac{1}{2}\log(2\pi e). \tag{10.38}$$

We expand the e-affine coordinates $S^{-1}(\omega)$ in Fourier series as

$$S^{-1}(\omega) = \sum_{t=0}^{\infty} r_t e_t(\omega) \tag{10.39}$$

and m-affine coordinates $S(\omega)$ as

$$S(\omega) = \sum_{t=0}^{\infty} r_t^* e_t(\omega), \tag{10.40}$$

where the basis functions are sinusoidal,

$$e_0(\omega) = 1, \quad e_t(\omega) = 2\cos\omega t, \quad t = 1, 2, \ldots. \tag{10.41}$$

Since the resultant coefficients $\{r_t\}$ and $\left\{r_t^*\right\}$ are linear transformations of $\theta(\omega)$ and $\eta(\omega)$, respectively, we can use them as new θ- and η-coordinates.

It is known that the coefficients r_t^* are expressed as

$$r_t^* = E\left[x_s x_{s-t}\right], \tag{10.42}$$

which are called the auto-correlation coefficients. Hence, the m-coordinates are the auto-correlation coefficients.

The inverse system of $H(z)$ is $H^{-1}(z)$, which is obtained by reversing the input and the output. Its power spectrum is $S^{-1}(\omega)$. Hence, r_t are the auto-correlation coefficients of the inverse system. They are called the inverse auto-correlations. The inverse auto-correlations form e-flat coordinate systems.

It is noteworthy that r_t and r_s^* are orthogonal,

$$\langle \boldsymbol{e}_t, \boldsymbol{e}_s^* \rangle = 0, \tag{10.43}$$

where $\boldsymbol{e}_t$ is the tangent vector of r_t coordinate axis and $\boldsymbol{e}_s^*$ is that of r_s^* axis. This implies that r_s, $s > k$ are parameters which are orthogonal to the auto-correlation coefficients $r_1^*, \ldots, r_k^*$. Hence, they represent directions orthogonal to the auto-correlations up to k.

It is easy to introduce the α-connection to L by using the cubic tensor

$$T\left(\omega, \omega', \omega''\right) = \frac{\partial^3}{\partial\theta(\omega)\partial\theta\left(\omega'\right)\partial\theta\left(\omega''\right)}\psi(\theta) \tag{10.44}$$

We can prove the following theorem.

Theorem 10.1 *L is dually flat for any α, having the Euclidean metric. The α-divergence is given by*

$$D^{(\alpha)}\left(S_1 \| S_2\right) = \begin{cases} \dfrac{1}{2\pi\alpha^2}\displaystyle\int \left\{\left(\frac{S_2}{S_1}\right)^{\alpha} - 1 - \alpha\log\frac{S_2}{S_1}\right\} d\omega, & (\alpha \neq 0), \\ \dfrac{1}{4\pi}\displaystyle\int \left(\log S_2 - \log S_1\right)^2 d\omega, & (\alpha = 0). \end{cases} \tag{10.45}$$

To prove the theorem, we introduce the α-representation of the power spectrum as

$$R^{(\alpha)}(\omega) = \begin{cases} -\frac{1}{\alpha} S(\omega)^{-\alpha}, & (\alpha \neq 0), \\ \log S(\omega), & (\alpha = 0). \end{cases} \tag{10.46}$$

Then, its Fourier coefficients are proved to be the α-flat coordinates. The theorem shows that L is like the manifold $\boldsymbol{R}^n_+$ of positive measure rather than the manifold S_n of probability distributions.

We have two dually coupled affine coordinate systems

$$\boldsymbol{r} = (r_0, r_1, r_2, \ldots), \tag{10.47}$$

$$\boldsymbol{r}^* = \left(r_0^*, r_1^*, r_2^*, \ldots\right). \tag{10.48}$$

The AR model of degree p, $AR(p)$, is characterized by

$$\boldsymbol{r} = \left(\boldsymbol{r}_p; 0, \ldots 0\right), \tag{10.49}$$

where $\boldsymbol{r}_p = \left(r_0, r_1, \ldots, r_p\right)$. It is defined by the linear constraints

$$r_{p+1} = r_{p+2} = \cdots = 0 \tag{10.50}$$

in the e-coordinates. Hence, it is an e-flat submanifold. Moreover, the families of all AR models of various degrees form a hierarchical system,

$$AR(0) \subset AR(1) \subset AR(2) \subset \cdots. \tag{10.51}$$

The white noise $S(\omega) = 1$ belongs to $AR(0)$, having the coordinates $\boldsymbol{r} = (1, 0, 0, \ldots)$.

The MA model of degree q, $MA(q)$, is characterized by

$$\boldsymbol{r}^* = \left(\boldsymbol{r}_q^*; 0, \ldots, 0\right), \tag{10.52}$$

where $\boldsymbol{r}_q^* = \left(r_0^*, r_1^*, \ldots, r_q^*\right)$. It is defined by

$$r_{q+1}^* = r_{q+2}^* = \ldots 0 \tag{10.53}$$

in the m-coordinate system. Hence, it is an m-flat submanifold and the MA models of various degrees form a hierarchical system

$$MA(0) \subset MA(1) \subset MA(2) \subset \cdots. \tag{10.54}$$

10.4 Geometry of AR, MA and ARMA Models

AR model: An AR model of degree p, $AR(p)$, is a finite-dimensional model determined by $\boldsymbol{a}$ in (10.15). By expanding the inverse of the its power spectrum $S(\omega;\boldsymbol{a})$, we have

$$S^{-1}(\omega;\boldsymbol{a}) = \sum_{t=0}^{p} r_t e_t(\omega). \tag{10.55}$$

The m-affine coordinates of $AR(p)$ are the auto-correlation coefficients $\boldsymbol{r} = (r_0, r_1, \ldots, r_p)$. However, the higher-order coefficients $r_{p+1}, r_{p+2}, \ldots$ are not 0, although they are not free but determined by $\boldsymbol{r}$. Given a system with power spectrum $S(\omega)$ having auto-correlations $r_0, r_1, r_2, \ldots$, we consider the system in $AR(p)$ of which the auto-correlations are the same as $S(\omega)$ up to $r_1, \ldots, r_p$ and its higher-degree auto-correlations $r_{p+1}, \ldots$ are 0. It is called the p-th order stochastic realization of $S(\omega)$. We denote its power spectrum by $S_p^{AR}(\omega)$. The set of systems having the same auto-correlations up to $r_1, \ldots, r_p$ form an m-flat submanifold, because they have the same values in the first p m-coordinates but the others are free. We denote it by $M_p(\boldsymbol{r})$. The $S_p^{AR}(\omega)$ is the intersection of the m-flat submanifold $M_p(\boldsymbol{r})$ and the submanifold $AR(p)$. The two submanifolds are orthogonal. Hence, $S_p^{AR}(\omega)$ is given by the m-projection of $S(\omega)$ to $AR(p)$. See Fig. 10.2.

Let $S_0(\omega)$ be white noise given by

$$S_0(\omega) = 1. \tag{10.56}$$

It belongs to $AR(p)$ for any p. From the Pythagorean theorem, we have

$$D_{KL}[S : S_0] = D_{KL}\left[S : S_p^{AR}\right] + D_{KL}\left[S_p^{AR} : S_0\right]. \tag{10.57}$$

The stochastic realization is characterized by maximization of entropy.

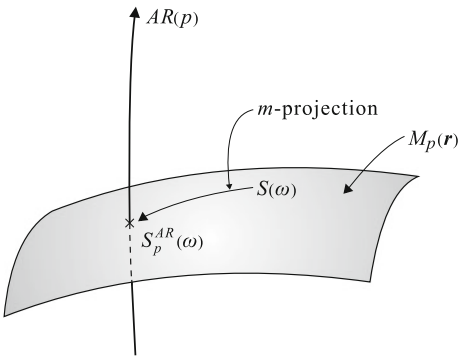


Fig. 10.2 Stochastic realization of $S(\omega)$ up to p-th order auto-correlations

Theorem 10.2 (Maximum Entropy) *The stochastic realization* $S_p^{AR}(\omega)$ *is the one that maximizes entropy among all systems having the same* $\boldsymbol{r} = (r_1, \ldots, r_p)$.

Proof From (10.38), we have

$$D_{KL}\left[S(\omega) : S_0(\omega)\right] = -2H_S(S) + \log(2\pi e) + c_0 - 1. \tag{10.58}$$

From relation (10.57), we see that S_p^{AR} is the minimizer of $D_{KL}\left[\tilde{S} : S_0\right]$ for all $\tilde{S} \in M_p(\boldsymbol{r})$. However, $D_{KL}\left[\tilde{S} : S_0\right]$ is related to the negative of entropy H_S by

$$D_{KL}\left[S : S_0(\omega)\right] = D_{KL}\left[S : S_p^*\right] + D_{KL}\left[S_p^* : S_0\right] = -2H_S + \text{const.} \tag{10.59}$$

Hence, S_p^{AR} is the maximizer of entropy among all systems having the same $r_1, \ldots, r_p$. □

MA model: Similar discussions hold for $MA(q)$ families. They are m-flat and $MA(q)$ is defined by the constraint

$$r_{q+1} = r_{q+2} = \cdots = 0. \tag{10.60}$$

We can define the dual stochastic realization of a system $S(\omega)$ in $MA(q)$, that is the system in $AR(q)$ of which the inverse auto-correlations $r_0^*, r_1^*, \ldots, r_q^*$ are the same as the given $S(\omega)$ up to q. It is interesting to see the following minimum entropy theorem.

Theorem 10.3 (Minimum Entropy) *The dual stochastic realization* $S_q^{MA}(\omega)$ *is the one that minimizes entropy among all systems having the same inverse auto-covariances* $r_1^*, \ldots, r_q^*$.

Proof We have

$$D_{KL}\left[S_0 : S\right] = D_{KL}\left[S_0 : S_q^{MA}\right] + D_{KL}\left[S_q^{MA} : S\right] \tag{10.61}$$

from the Pythagorean theorem. Now we see

$$D_{KL}\left[S_0 : S\right] = H_S + \text{const.} \tag{10.62}$$

Hence minimizing $D_{KL}\left[S_0 : S\right]$ is equivalent to minimizing entropy, proving the theorem. □

One may say that the Pythagorean relation or the projection theorem is more fundamental than the maximum entropy principle.

ARMA model: The ARMA model of degrees p, q is given by (10.21). This is a finite-dimensional subset of L. They form a doubly hierarchical system. However,

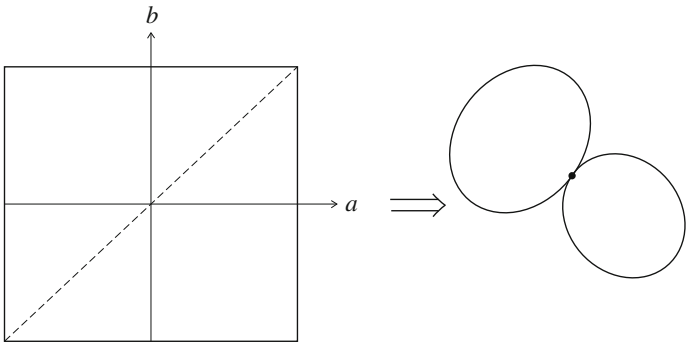


Fig. 10.3 Singularity of (1, 1) ARMA model

they are neither e-flat nor m-flat. Moreover, the set is not a submanifold in the mathematical sense, because it includes singular points. We show this by a simple example. Consider $ARMA(1, 1)$, which is described by

$$x_t = ax_{t-1} + \varepsilon_t + b\varepsilon_{t-1}. \tag{10.63}$$

Its transfer function is

$$H(z) = \frac{1 + bz^{-1}}{1 + az^{-1}}. \tag{10.64}$$

The parameter (a, b) plays the role of coordinates, where $|a| < 1$, $|b| < 1$ should be satisfied because of the stability of the system. However, on the diagonal line $a = b$, all the systems are equivalent, because the nominator and the denominator of (10.64) cancel one another out. Therefore, all the systems satisfying $a = b$ are the same, simply given by $H(z) = 1$.

We reduce the equivalent systems to one point. Then, as is shown in Fig. 10.3, the set $AR(1, 1)$ consists of two subsets (submanifolds) connected by one singular point. This type of reduction happens in any $ARMA(p, q)$ when the denominator and nominator of (10.22) include the same factor which cancels one another out. This fact is pointed out by Brockett (1976). We deal with such singularity later in Chap. 12 where multilayer perceptrons are discussed.

Remarks

Linear systems and time series have long histories of research, having highly organized structures in their fields. Therefore, we only touch upon them from the information geometry point of view, not explaining details. Since we have used Gaussian white noise as inputs, our study includes only systems of minimal phase. We need

non-Gaussian white noise to overcome this difficulty. Finite-dimensional time series and systems are well-founded mathematically, but if we want to treat infinite-dimensional cases, we suffer from a lack of rigorous mathematical foundation. The difficulty is the same as in the case of a manifold of infinite-dimensional probability distributions. The present study will be a starting point for investigating information geometry of systems. See a trial by Ohara and Amari (1994).

There is a statistical problem of estimation from observations of a finite size sample $x_1, x_2, \ldots, x_T$ of time series. We can identify the model which generates the sample by using an adequate degree of AR, MA and ARMA or many other models. The sample is not iid, but we can construct a similar theory of estimation. A higher-order asymptotic theory has been constructed. See Amari and Nagaoka (2000) and Taniguchi (1991) for more details. An AR model is an e-flat manifold, provided we consider time series x_t of infinite length $t = 0, \pm 1, \pm 2, \ldots$. However, it is a curved exponential family when $x_0, x_1, \ldots, x_T$ only are observed, because of the effect of initial and final x_t's. See Ravishanker et al. (1990) and Martin (2000) for applications and Choi and Mullhaupt (2015) for recent developments using Khälerian geometry.

It is interesting to see that an ARMA model includes singularities. Brockett (1976) pointed out that the set of linear systems of which transfer functions are rational functions, nominators with degree p, and denominators with degree q, are split in a number of disjoint components. This is a topological structure of the set of linear systems. When cancellation occurs, the degrees of the nominator and denominator decrease. R. Brockett excluded such low-degree systems from the set. However, a lower degree system is a special case of a higher degree system. Therefore, if we consider rational systems having a nominator degree lower or equal to p and a denominator degree lower than or equal to q, the set splits into multiple components where they are connected by singular points of reduced degrees.

We have considered regular statistical models, which form a manifold. However, not a few important statistical models include this type of singularities. The behavior of an estimator when the true model lies at or close to singularities is interesting. See Fukumizu and Kuriki (2004). We study multilayer perceptrons in Chap. 12, considering how the singularity affects the dynamics of learning.

We did not study multi-input and multi-output systems. The manifold of linear systems having n inputs and m outputs is a Grassman manifold. This is another interesting subject of research from the geometrical point of view.

A Markov chain generates an infinite series of states

$$\{x_t\}, \ t = 0, \ \pm 1, \ \pm 2, \ \ldots, \tag{10.65}$$

where x_t is a state from which x_{t+1} is determined stochastically by a state transition matrix $p\left(x_{t+1} | x_t\right)$. An AR model is regarded as a Markov chain. A Markov chain is an exponential family, so it is dually flat. We can construct a similar geometrical theory (Amari 2001). However, if we consider a finite range $0 \leq t \leq T$ of observations, a Markov chain $\{x_t\}, 0 \leq t \leq T$, is a curved exponential family because of the effects

of initial and final values. Its e-curvature decreases in the order of $1/T$, converging to 0 as T tends to infinity. See Amari (2001), and Hayashi and Watanabe (2014) for information geometry of Markov chains. Takeuchi (2014) used the e-curvature to evaluate the asymptotic error of estimation, which is also related to the minimum regret of a Markov chain (Takeuchi et al. 2013).

Part IV
Applications of Information Geometry

Chapter 11
Machine Learning

11.1 Clustering Patterns

Patterns are categorized into a number of classes. Pattern recognition is the problem of identifying the class to which a given pattern belongs. When a divergence is defined in the manifold of patterns, classification is brought into effect by using the divergence. We begin with the problem of obtaining a representative of a cluster of patterns, called the center of the cluster. When patterns are categorized into clusters, pattern recognition determines the cluster to which a given pattern is supposed to belong, based on the closeness due to the divergence.

Another problem is to divide a non-labeled aggregate of patterns into a set of clusters. This is the problem of clustering. A generalized k-means method is presented by using a divergence. The entire pattern space is divided into regions based on representatives of clusters. Such a division is called a divergence-based Voronoi diagram. When patterns are generated randomly subject to a probability distribution depending on each class, we have a stochastic version of the above problems. Information geometry is useful for understanding these problems in terms of divergence.

11.1.1 Pattern Space and Divergence

Let us consider patterns represented by vector $\boldsymbol{x}$. They belong to a pattern manifold X. We study the case where a divergence $D\left[\boldsymbol{x} : \boldsymbol{x}'\right]$ is defined between two patterns $\boldsymbol{x}$ and $\boldsymbol{x}'$. In a Euclidean space, we have

$$D\left[\boldsymbol{x} : \boldsymbol{x}'\right] = \frac{1}{2}\sum \left(x_i - x_i'\right)^2, \tag{11.1}$$

which is a half of the square of the Euclidean distance. We consider a general dually flat manifold induced by a Bregman divergence. For the sake of notational

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_11

convenience, we suppose that pattern $\boldsymbol{x}$ is represented in the dual affine coordinate system, so that it is represented by the $\boldsymbol{\eta}$-coordinates as

$$\boldsymbol{\eta} = \boldsymbol{x}. \tag{11.2}$$

Then, we use the dual divergence between two patterns $\boldsymbol{x}$ and $\boldsymbol{x}'$

$$D_\phi \left[\boldsymbol{x} : \boldsymbol{x}'\right] = \phi(\boldsymbol{x}) - \phi\left(\boldsymbol{x}'\right) - \nabla\phi\left(\boldsymbol{x}'\right) \cdot \left(\boldsymbol{x} - \boldsymbol{x}'\right), \tag{11.3}$$

which is constructed from a dual convex function ϕ.

We will later use the primal affine coordinate system $\boldsymbol{\theta}$ given by the Legendre transformation

$$\boldsymbol{\theta} = \nabla\phi(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}}\phi(\boldsymbol{\eta}). \tag{11.4}$$

The primal convex function $\psi(\boldsymbol{\theta})$ is given by the Legendre relation

$$\psi(\boldsymbol{\theta}) = -\phi(\boldsymbol{\eta}) + \boldsymbol{\theta} \cdot \boldsymbol{\eta}, \tag{11.5}$$

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}). \tag{11.6}$$

11.1.2 Center of Cluster

Let C be a cluster consisting of k patterns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. We search for the representative of C which should be as close to all the members of C as possible (Fig. 11.1). To obtain such a representative, we calculate the average of the dual divergences from the members of the cluster to a point $\boldsymbol{\eta}$, as

$$D_\phi[C : \boldsymbol{\eta}] = \frac{1}{k} \sum_{\boldsymbol{x}_i \in C} D_\phi\left[\boldsymbol{x}_i : \boldsymbol{\eta}\right]. \tag{11.7}$$

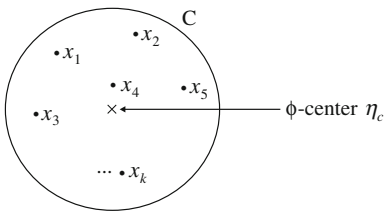


Fig. 11.1 ϕ-center of cluster C

The minimizer of (11.7) is called the ϕ-center of cluster C due to the divergence D_ϕ. If we use the $\boldsymbol{\theta}$-coordinates, this is written as

$$D_\psi[\boldsymbol{\theta} : C] = \frac{1}{k} \sum D_\psi [\boldsymbol{\theta} : \boldsymbol{\theta}_i] , \tag{11.8}$$

where $\boldsymbol{\theta}_i$ is the $\boldsymbol{\theta}$-coordinates of $\boldsymbol{\eta}_i$. The following theorem is due to Banerjee et al. (2005).

Theorem 11.1 *The ϕ-center of cluster C is given by*

$$\boldsymbol{\eta}_C = \frac{1}{k} \sum \boldsymbol{x}_i \tag{11.9}$$

for any ϕ.

Proof By differentiating (11.7) with respect to $\boldsymbol{\eta}$ and using (11.3), we have

$$\frac{\partial}{\partial \boldsymbol{\eta}} D[C : \boldsymbol{\eta}] = \frac{1}{k} \sum \mathbf{G}^{-1}(\boldsymbol{\eta}) (\boldsymbol{x}_i - \boldsymbol{\eta}) , \tag{11.10}$$

where

$$\mathbf{G}^{-1}(\boldsymbol{\eta}) = \nabla \nabla \phi(\boldsymbol{\eta}) \tag{11.11}$$

is a positive-definite matrix. Hence, the minimizer is given by (11.9). □

We can generalize the situation that a probability distribution $p(\boldsymbol{x})$ of $\boldsymbol{x}$ is given instead of cluster C. Then the center of the distribution is defined by the minimizer of

$$D_\phi[p : \boldsymbol{\eta}] = \int D_\phi[\boldsymbol{x} : \boldsymbol{\eta}] p(\boldsymbol{x}) d\boldsymbol{x} . \tag{11.12}$$

The center is merely the expectation of $\boldsymbol{x}$ for any ϕ,

$$\boldsymbol{\eta}_p = \int \boldsymbol{x}\, p(\boldsymbol{x}) d\boldsymbol{x} . \tag{11.13}$$

11.1.3 k-Means: Clustering Algorithm

Assume N points $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ are given, and we categorize them into m clusters such that a cluster includes mutually close points. Let $C_1, \ldots, C_m$ be clusters to be formed and let $\boldsymbol{\eta}_h, h = 1, \ldots, m$, be their centers. It is required that a point $\boldsymbol{x}_i$ belongs to cluster C_h when the divergence $D_\phi \left[\boldsymbol{x}_i : \boldsymbol{\eta}_h\right]$ is the smallest of $D_\phi \left[\boldsymbol{x} : \boldsymbol{\eta}_1\right], \ldots, D_\phi \left[\boldsymbol{x} : \boldsymbol{\eta}_m\right]$. That is, $\boldsymbol{\eta}_h$ is the closest cluster center from $\boldsymbol{x}_i$,

$$h = \arg\min_j D_\phi \left[\boldsymbol{x}_i : \boldsymbol{\eta}_j\right] . \tag{11.14}$$

Let

$$D_\phi[C : D] = \sum_h \sum_{\boldsymbol{x}_i \in C_h} D_\phi\left[\boldsymbol{x}_i : \boldsymbol{\eta}_h\right] \tag{11.15}$$

be the total sum of the divergences from each point $\boldsymbol{x}_i$ to the cluster center $\boldsymbol{\eta}_h$ it belongs to. The best clustering with respect to the ϕ-divergence is the one that minimizes (11.15). We can apply the well-known k-means algorithm, which is usually done by using the Euclidean distance. It is easy to extend it to the general case of a dually flat divergence, because the cluster center is given by the arithmetic mean in the dual coordinates. See Banerjee et al. (2005).

Clustering Algorithm (k-means method)

1. Initial Step: Choose m cluster centers $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m$ arbitrarily such that they are all different.
2. Classification Step: For each $\boldsymbol{x}_i$, calculate the ϕ-divergences to the m cluster centers. Assign $\boldsymbol{x}_i$ to cluster C_h that minimizes the ϕ-divergence,

$$\boldsymbol{x}_i \in C_h : D_\phi\left[\boldsymbol{x}_i : \boldsymbol{\eta}_h\right] = \min_j \left\{D_\phi\left[\boldsymbol{x}_i : \boldsymbol{\eta}_j\right]\right\}. \tag{11.16}$$

   Thus, new clusters $C_1, \ldots, C_m$ are formed.
3. Renewal Step: Calculate the ϕ-centers of the renewed clusters, obtaining new cluster centers $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m$.
4. Termination Step: Repeat the above procedures until convergence.

It is known that the procedures terminate within a finite number of steps, giving a good clustering result, although there is no guarantee that it is optimal. The k-means^{++} method was proposed for choosing good initial values of $\boldsymbol{\eta}_i$ by Arthur and Vassilvitshii (2007).

11.1.4 Voronoi Diagram

Given a point $\boldsymbol{x}$, we need to find the cluster it belongs to. This is information retrieval or pattern classification to decide which category it belongs to. A subset R_h of X is called the region of C_h when it is decided that pattern $\boldsymbol{x} \in R_h$ belongs to C_h. The entire X is partitioned into m regions $R_1, \ldots, R_m$.

We consider a simple case consisting of two clusters C_1 and C_2 for an explanation. The entire X is divided into two regions R_1 and R_2. For $\boldsymbol{x} \in R_1$,

$$D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_1\right] \leq D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_2\right]. \tag{11.17}$$

Therefore, the two regions are separated by the hypersurface B_{12} that is the boundary of the regions,

$$B_{12} = \left\{\boldsymbol{x} \mid D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_1\right] = D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_2\right]\right\}. \tag{11.18}$$

Theorem 11.2 *The hypersurface separating two decision regions is the geodesic hyperplane orthogonal to the dual geodesic connecting the two ϕ-centers of the clusters at the midpoint of the dual geodesic.*

Proof Connect two ϕ-centers $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ by the dual geodesic

$$\boldsymbol{\eta}(t) = (1-t)\boldsymbol{\eta}_1 + t\boldsymbol{\eta}_2. \tag{11.19}$$

The midpoint $\boldsymbol{\eta}_{12}$ is defined by

$$D_\phi\left[\boldsymbol{\eta}_{12} : \boldsymbol{\eta}_1\right] = D_\phi\left[\boldsymbol{\eta}_{12} : \boldsymbol{\eta}_2\right]. \tag{11.20}$$

Let B_{12} be the geodesic hypersurface (that is the linear subspace in the $\boldsymbol{\theta}$ coordinates) passing through $\boldsymbol{\eta}_{12}$ and orthogonal to the dual geodesic (Fig. 11.2). Then, due to the Pythagorean theorem, any point $\boldsymbol{x}$ on the hyperplane satisfies

$$D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_i\right] = D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_{12}\right] + D_\phi\left[\boldsymbol{\eta}_{12} : \boldsymbol{\eta}_i\right], \quad i = 1, 2. \tag{11.21}$$

Hence, we have

$$D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_1\right] = D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_2\right]. \tag{11.22}$$

□

The boundary surface is linear in the $\boldsymbol{\theta}$-coordinates but is nonlinear in the $\boldsymbol{\eta}$-coordinates. When the divergence is the square of the Euclidean distance, $\boldsymbol{\eta}$- and $\boldsymbol{\theta}$-coordinates are the same, so that the boundary is linear in the $\boldsymbol{\eta}$-coordinates. This is a special case.

When there are m clusters $C_1, \ldots, C_m$, X is partitioned into m regions $R_1, \ldots, R_m$, where the boundary of R_i and R_j is the geodesic hypersurface satisfying

$$B_{ij} = \left\{\boldsymbol{x} \mid D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_i\right] = D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_j\right]\right\}. \tag{11.23}$$

Such a partition is called the Voronoi diagram due to the ϕ-divergence (Fig. 11.3). See Nielsen and Nock (2014), Nock and Nielsen (2009), Boissonnat et al. (2010), etc. for details.

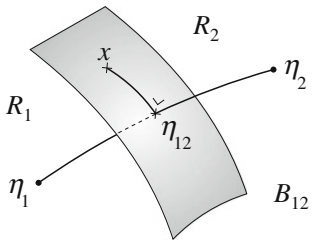


Fig. 11.2 Boundary of two cluster regions

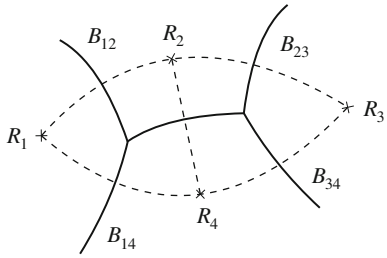


Fig. 11.3 Voronoi diagram

11.1.5 Stochastic Version of Classification and Clustering

11.1.5.1 Probability Distribution Associated with Category

Let us consider a cluster C_h of which ϕ-center is $\boldsymbol{\eta}_h$. We generate a probability distribution

$$p_h(\boldsymbol{x}) = \exp\{\phi(\boldsymbol{x})\} \exp\left\{-D_\phi\left[\boldsymbol{x} : \boldsymbol{\eta}_h\right]\right\}. \tag{11.24}$$

It is centered at $\boldsymbol{\eta}_h$ and the probability density of $\boldsymbol{x}$ decreases exponentially as the divergence between $\boldsymbol{x}$ and $\boldsymbol{\eta}_h$ increases.

As we have shown in Sect. 2.6, it is an exponential family (Banerjee et al. 2005).

Theorem 11.3 *The cluster C_h of which the center is $\boldsymbol{\eta}_h$ defines a probability distribution of patterns $\boldsymbol{x}$, which is an exponential family,*

$$p_h(\boldsymbol{x}) = \exp\{\boldsymbol{\theta}_h \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta}_h)\} \tag{11.25}$$

with respect to the underlying measure

$$d\mu(\boldsymbol{x}) = \exp\{\phi(\boldsymbol{x})\}\, d\boldsymbol{x}. \tag{11.26}$$

The natural parameter $\boldsymbol{\theta}_h$ of the distribution is the Legendre dual of $\boldsymbol{\eta}_h$.

11.1.5.2 Soft Clustering Algorithm

We consider a mixture of probability distributions of exponential families,

$$p(\boldsymbol{x}; \boldsymbol{\xi}) = \sum_h \pi_h \exp\{\boldsymbol{\theta}_h \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta}_h)\}, \tag{11.27}$$

where π_h are the prior probabilities that $\boldsymbol{x}$ is generated from category C_h and is the unknown parameters which we estimate from a number of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. Here, the parameter vector is

$$\boldsymbol{\xi} = (\pi_1, \ldots, \pi_m \,;\, \boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_m) \,. \tag{11.28}$$

The maximum likelihood estimator is given by

$$\hat{\boldsymbol{\xi}} = \arg\max \sum_{i=1}^{N} \log p\,(\boldsymbol{x}_i, \boldsymbol{\xi}) \,. \tag{11.29}$$

Before analyzing the MLE, we consider the conditional distribution of categories given $\boldsymbol{x}$,

$$p\,(C_h\,|\boldsymbol{x}\,) = \frac{\pi_h p\,(\boldsymbol{x}, \boldsymbol{\theta}_h)}{\sum \pi_h p\,(\boldsymbol{x}, \boldsymbol{\theta}_h)}. \tag{11.30}$$

For pattern $\boldsymbol{x}$, the above probabilities show the posterior probabilities of the categories. This is a stochastic classification or soft classification which assigns $\boldsymbol{x}$ to categories according to the posterior probabilities. When we pick up the category that maximizes the probability, it attains hard classification.

Since the distribution (11.29) is a mixture of exponential families, we can use the EM algorithm for estimating $\boldsymbol{\xi}$. The M-step is usually computationally heavy, but in the present case, it is simple because of (11.13).

Soft Clustering Algorithm (soft k-means)

1. Initial Step: Choose prior probabilities π_h and different cluster centers $\boldsymbol{\eta}_h$, $h = 1, \ldots, m$, arbitrarily.
2. Classification Step: For each $\boldsymbol{x}_i$, calculate the conditional probabilities $p\,(C_h|\boldsymbol{x})$ by using the current π_h and $\boldsymbol{\eta}_h$.
3. Renewal Step: By using the conditional probabilities, the new prior π_h of class C_h is calculated as

$$\pi_h = \frac{1}{N} \sum_i p\,(C_h\,|\boldsymbol{x}_i\,) \,. \tag{11.31}$$

Calculate the new cluster center by

$$\boldsymbol{\eta}_h = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i\, p\,(C_h\,|\boldsymbol{x}_i\,) \,. \tag{11.32}$$

4. Termination: Repeat the above procedures until convergence.

The Voronoi diagram is defined in a similar way. When we use hard classification based on the posteriori probabilities, the boundary surface of two categories C_i and C_j is given by

$$p\,(C_i|\boldsymbol{x}) = p\left(C_j|\boldsymbol{x}\right). \tag{11.33}$$

Theorem 11.4 *The boundary of two decision regions is the geodesic hypersurface that is orthogonal to the dual geodesic connecting two cluster centers and passes*

through it at the point satisfying

$$\pi_i D_\phi \left[\boldsymbol{x} \left| \boldsymbol{\eta}_i \right. \right] = \pi_j D_\phi \left[\boldsymbol{x} \left| \boldsymbol{\eta}_j \right. \right]. \tag{11.34}$$

11.1.6 Robust Cluster Center

When a cluster C composed of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ is given, we can calculate the ϕ-center of cluster by (11.9). Assume that a new point $\boldsymbol{x}^*$ is added to C that might be far from the others. By adding this point, the cluster center might deviate largely. If this new point is an outlier, for example given by mistake, it is not desirable that the cluster center is affected heavily by this point. A robust clustering reduces the undesirable influence due to outliers.

More formally, we define the influence function of an outlier $\boldsymbol{x}^*$. Let $\bar{\boldsymbol{\eta}}$ be the center of cluster C, and let $\bar{\boldsymbol{\eta}}^*$ be the center of C^* in which $\boldsymbol{x}^*$ is newly added. We assume that k is large so that the influence of each $\boldsymbol{x}_i$ is only of the order $1/k$. Let us denote the change of $\bar{\boldsymbol{\eta}}$ to $\bar{\boldsymbol{\eta}}^*$ by $\delta\boldsymbol{\eta}$ and define $z\left(\boldsymbol{x}^*\right)$ by

$$\delta\boldsymbol{\eta} = \bar{\boldsymbol{\eta}}^* - \bar{\boldsymbol{\eta}} = \frac{1}{k} z\left(\boldsymbol{x}^*\right) \tag{11.35}$$

as a function of $\boldsymbol{x}^*$. It is called an influence function. When

$$\left| z\left(\boldsymbol{x}^*\right) \right| < c \tag{11.36}$$

holds for a constant c, i.e., $|z\left(\boldsymbol{x}^*\right)|$ is bounded, the cluster center is robust, because even if an infinitely large $\boldsymbol{x}^*$ is merged in C, its effect is bounded and is very small when k is large. A robust center does not seriously suffer from contamination by outliers.

11.1.6.1 Total Bregman Divergence

The Bregman divergence $D_\phi\left[\boldsymbol{\eta}' : \boldsymbol{\eta}\right]$ is measured by the height $\phi\left(\boldsymbol{\eta}'\right)$ at $\boldsymbol{\eta}'$ from the tangent hypersurface of $\phi(\boldsymbol{\eta})$ drawn at $\boldsymbol{\eta}$. This is the vertical length of point $\left(\boldsymbol{\eta}', \phi\left(\boldsymbol{\eta}'\right)\right)$ to the tangent hypersurface (Fig. 11.4a). One may consider the orthogonal projection of $\left(\boldsymbol{\eta}', \phi\left(\boldsymbol{\eta}'\right)\right)$ to the tangent hypersurface (Fig. 11.4b). It defines another measure of divergence from $\boldsymbol{\eta}'$ to $\boldsymbol{\eta}$. This idea, hinted at in the total least squares in regression, was proposed by Vemuri et al. (2011) and called the total Bregman divergence, denoted as tBD.

The length of the orthogonal projection is easily calculated and given by

$$\mathrm{tBD}_\phi\left[\boldsymbol{\eta}' : \boldsymbol{\eta}\right] = \frac{1}{w\left(\boldsymbol{\eta}'\right)} D_\phi\left[\boldsymbol{\eta}' : \boldsymbol{\eta}\right], \tag{11.37}$$

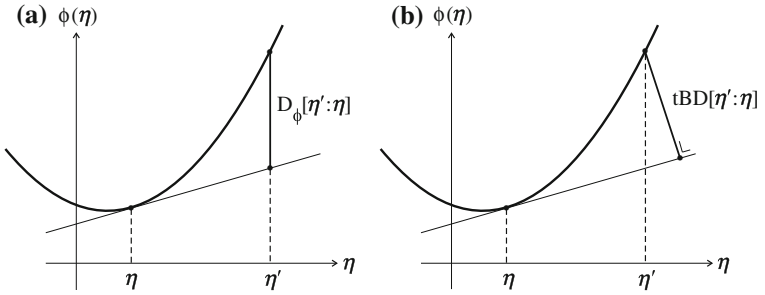


Fig. 11.4 **a** Bregman divergence D; **b** total Bregman divergence tBD

where

$$\boldsymbol{w}\left(\boldsymbol{\eta}'\right) = \sqrt{1 + \|\nabla\phi\left(\boldsymbol{\eta}'\right)\|^2}. \tag{11.38}$$

It is invariant under orthogonal transformations of $(\boldsymbol{\eta}, \phi)$-space. Since the scale of $\phi(\boldsymbol{\eta})$ is arbitrary, we introduce a free parameter c which changes $\phi(\boldsymbol{\eta})$ to $c\phi(\boldsymbol{\eta})$ and define tBD by

$$\mathrm{tBD}_\phi\left[\boldsymbol{\eta}' : \boldsymbol{\eta}\right] = \frac{D_\phi\left[\boldsymbol{\eta}' : \boldsymbol{\eta}\right]}{\sqrt{1 + c^2\|\nabla\phi\left(\boldsymbol{\eta}'\right)\|^2}}. \tag{11.39}$$

This is a conformal transformation of Bregman divergence. The free parameter c controls the degree of conformal modification.

11.1.6.2 Total BD is Robust

The following is one of the remarkable characteristics of tBD, proved in Liu et al. (2012).

Theorem 11.5 *The tBD ϕ-center of a cluster is robust.*

Proof When an outlier $\boldsymbol{x}^*$ is newly added to C of which the previous center is $\bar{\boldsymbol{\eta}}$, the new center $\bar{\boldsymbol{\eta}}^*$ under tBD is the minimizer of

$$\frac{1}{k+1}\sum \frac{D_\phi\left[\boldsymbol{x}_i : \bar{\boldsymbol{\eta}}^*\right]}{w\left(\boldsymbol{x}_i\right)} + \frac{1}{k+1}\frac{D_\phi\left[\boldsymbol{x}^* : \bar{\boldsymbol{\eta}}^*\right]}{w\left(\boldsymbol{x}^*\right)}. \tag{11.40}$$

The influence function $\boldsymbol{z}\left(\boldsymbol{x}^*\right)$ is defined by (11.35). Assuming k is large, we expand the new center in the Taylor series, obtaining

$$\boldsymbol{z}\left(\boldsymbol{x}^*\right) = \frac{1}{w(\boldsymbol{x}^*)}\mathbf{G}^{-1}\left\{\nabla\phi(\bar{\boldsymbol{\eta}}) - \nabla\phi\left(\boldsymbol{x}^*\right)\right\}, \tag{11.41}$$

where

$$\mathbf{G} = \frac{1}{N} \sum \frac{1}{w\left(\boldsymbol{x}_i\right)} \nabla \nabla \phi \left(\bar{\boldsymbol{\eta}}\right). \tag{11.42}$$

Since

$$\frac{1}{w(\boldsymbol{x}^*)} \nabla \phi(\boldsymbol{x}^*) = \frac{\nabla \phi \left(\boldsymbol{x}^*\right)}{\sqrt{1 + c^2 \nabla \phi \left(\boldsymbol{x}^*\right)}} \tag{11.43}$$

is bounded for any large $\boldsymbol{x}^*$, $z\left(\boldsymbol{x}^*\right)$ is bounded, and hence the tBD ϕ-center is robust. □

Vemuri et al. (2011) used tBD to analyze MRI images, obtaining good results. Liu et al. (2012) applied the tBD to the problem of image retrieval, obtaining a state-of-the-art result. Conformal transformations of a Bregman divergence are further developed in Nock et al. (2015).

11.1.7 Asmptotic Evaluation of Error Probability in Pattern Recognition: Chernoff Information

We consider two probability distributions

$$p_i(\boldsymbol{x}) = p_i\left(\boldsymbol{x}, \boldsymbol{\theta}_i\right) = \exp\left\{\boldsymbol{\theta}_i \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\right\}, \quad i = 1, 2 \tag{11.44}$$

in an exponential family. Here, we use $\boldsymbol{\theta}$-coordinates related to $\psi(\boldsymbol{\theta})$ and the KL-divergence $D_\psi = D_{KL}$ instead of the previous $\boldsymbol{\eta}$-coordinates related to $\phi(\boldsymbol{\eta})$ and the dual divergence D_ϕ. When N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are derived, all of which are supposed to be generated from either $p_1(\boldsymbol{x})$ or $p_2(\boldsymbol{x})$, we need to decide which is the true distribution. Let us divide the manifold in two regions R_1 and R_2 such that, when the observed point

$$\bar{\boldsymbol{\eta}} = \frac{1}{N} \sum \boldsymbol{x}_i \tag{11.45}$$

belongs to R_1 (R_2), we decide that the true distribution is $p_1(\boldsymbol{x})$ ($p_2(\boldsymbol{x})$).

When N is large, the probability that $\bar{\boldsymbol{\eta}}$ is generated from $p_i(\boldsymbol{x})$ is given, due to the large deviation theorem in Chap. 3, by

$$P_i\left(\bar{\boldsymbol{\eta}}\right) = \exp\left\{-n D_{KL}\left[\bar{\boldsymbol{\theta}} : \boldsymbol{\theta}_i\right]\right\}, \tag{11.46}$$

where $\bar{\boldsymbol{\theta}}$ is the primal coordinates of $\bar{\boldsymbol{\eta}}$. In order to minimize the probability of misclassification, the regions R_i should be determined as

$$R_i = \left\{\boldsymbol{\theta} \,\middle|\, D_{KL}\left[\boldsymbol{\theta} : \boldsymbol{\theta}_i\right] \leq D_{KL}\left[\boldsymbol{\theta} : \boldsymbol{\theta}_j\right]\right\}. \tag{11.47}$$

That is, the boundary of R_1 and R_2 is the hypersurface satisfying

$$B_{12} = \{\boldsymbol{\theta} \,|\, D_{KL}[\boldsymbol{\theta} : \boldsymbol{\theta}_1] = D_{KL}[\boldsymbol{\theta} : \boldsymbol{\theta}_2]\}. \tag{11.48}$$

Let us consider the e-geodesic connecting $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$,

$$\log p(\boldsymbol{x}, \lambda) = (1-\lambda)\log p_1(\boldsymbol{x}) + \lambda \log p_2(\boldsymbol{x}) - \psi(\lambda) \tag{11.49}$$

or

$$\boldsymbol{\theta}_\lambda = (1-\lambda)\boldsymbol{\theta}_1 + \lambda \boldsymbol{\theta}_2 \tag{11.50}$$

in the $\boldsymbol{\theta}$-coordinates. Its midpoint is defined by $\boldsymbol{\theta}_{\lambda^*}$ satisfying

$$D_{KL}[\boldsymbol{\theta}_{\lambda^*} : \boldsymbol{\theta}_1] = D_{KL}[\boldsymbol{\theta}_{\lambda^*} : \boldsymbol{\theta}_2]. \tag{11.51}$$

Due to the Pythagorean theorem, B_{12} is the m-geodesic hyperplane orthogonal to the e-geodesic, passing through it at $\boldsymbol{\theta}^*_\lambda$. (See Fig. 11.5.)

The midpoint λ^* is given by the minimizer of

$$\psi(\lambda) = \int p_1(x)^\lambda p_2(x)^{1-\lambda} dx, \tag{11.52}$$

$$\lambda^* = \arg\min_\lambda \psi(\lambda). \tag{11.53}$$

The asymptotic error bound is hence given by

$$P_{\text{error}} = \exp\{-N D_{KL}[\boldsymbol{\theta}_{\lambda^*} : \boldsymbol{\theta}_i]\}. \tag{11.54}$$

The negative exponent of error,

$$D_{KL}[\boldsymbol{\theta}_{\lambda^*} : \boldsymbol{\theta}_i] = \psi\left(\lambda^*\right), \tag{11.55}$$

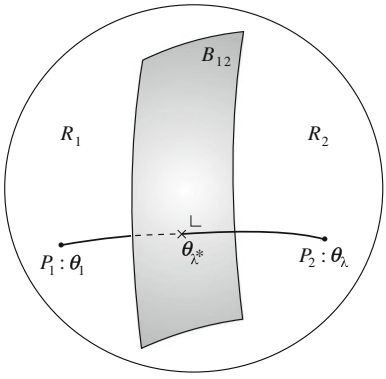


Fig. 11.5 Decision boundary B_{12} and separation midpoint $\boldsymbol{\theta}_{\lambda^*}$

is called the Chernoff information or Chernoff divergence (Chernoff 1952). This is related to the α-divergence $D_\alpha[p_1 : p_2]$. We have

$$\min_\lambda \int p_1(x)^\lambda p_2(x)^{1-\lambda} dx = 1 - \max_\alpha \frac{1-\alpha^2}{4} D_\alpha[p_1 : p_2]. \tag{11.56}$$

Hence, by letting α^* be the maximizer of $\left\{\left(1-\alpha^2\right)/4\right\} D_\alpha[p_1 : p_2]$, we have

$$\lambda^* = \frac{1+\alpha^*}{2}. \tag{11.57}$$

Remark One may use a prior distribution (π_1, π_2) on two classes C_1 and C_2 in the Bayesian standpoint. However, the asymptotic error bound does not depend on it.

11.2 Geometry of Support Vector Machine

The support vector machine (SVM) is one of the powerful learning machines for pattern recognition and regression (Cortes and Vapnik 1995; Vapnik 1998). It embeds pattern signals to a higher-dimensional space, even an infinite-dimensional Hilbert space, and uses a kernel function to calculate outputs. Although the Hilbert space is infinite-dimensional in general, the kernel trick makes it possible to work within a finite regime, avoiding difficulties of infinitely large degrees of freedom. We do not describe the details of the SVM, but focus only on its Riemannian structure. It is used for modifying a given kernel to improve the performance of the machine.

11.2.1 Linear Classifier

We begin with a linear machine for classifying patterns, which is a simple perceptron. Given input pattern $\boldsymbol{x} \in \boldsymbol{R}^n$, consider a linear function

$$f(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{11.58}$$

having parameters $\boldsymbol{\xi} = (\boldsymbol{w}, b)$. The machine classifies patterns into two classes C_+ and C_-, according to the signature of output function $f(\boldsymbol{x}, \boldsymbol{\xi})$. That is, when

$$f(\boldsymbol{x}, \boldsymbol{\xi}) > 0, \tag{11.59}$$

$\boldsymbol{x}$ is classified into C_+, and otherwise into C_-.

Consider a set of training examples $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ which are divided into two classes C_+ and C_-. When $\boldsymbol{x}_i \in C_+$, it is accompanied by teacher signal $y_i = 1$, and when $\boldsymbol{x}_i \in C_-$, it is accompanied by $y_i = -1$. They are linearly separable when

there exists $\boldsymbol{w}$ and b, for which

$$\boldsymbol{w} \cdot \boldsymbol{x} + b > 0, \quad \boldsymbol{x} \in C_+, \quad \boldsymbol{w} \cdot \boldsymbol{x} + b < 0, \quad \boldsymbol{x} \in C_- \tag{11.60}$$

holds. When $(\boldsymbol{w}, b)$ is such a solution, $(c\boldsymbol{w}, cb)$ is also a solution for any $c > 0$. We eliminate this indefiniteness of scale by imposing the constraints

$$|\boldsymbol{w} \cdot \boldsymbol{x}_i + b| \geq 1, \quad \min_i |\boldsymbol{w} \cdot \boldsymbol{x}_i + b| = 1. \tag{11.61}$$

Since the Euclidean distance from point $\boldsymbol{x}$ to the separating hyperplane

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \tag{11.62}$$

is

$$d = \frac{|\boldsymbol{w} \cdot \boldsymbol{x} + b|}{|\boldsymbol{w}|}, \tag{11.63}$$

the distance from $\boldsymbol{x}_i$ to the separating hyperplane is given by

$$d_i = \frac{|\boldsymbol{w} \cdot \boldsymbol{x}_i + b|}{|\boldsymbol{w}|}. \tag{11.64}$$

The minimum of these distances is given by

$$d_{\min} = \frac{1}{|\boldsymbol{w}|} \tag{11.65}$$

and is attained by the points $\boldsymbol{x}_i$ that satisfy

$$y_i \, (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) = 1. \tag{11.66}$$

We call these points $\boldsymbol{x}_i$ the support vectors of the training set D and the minimal distance the margin. There are in general a number of support vectors. See Fig. 11.6. A good machine has a large margin. So the problem of obtaining the optimal linear machine is to minimize

$$C(\boldsymbol{w}) = \frac{1}{2}|\boldsymbol{w}|^2 \tag{11.67}$$

under the constraint

$$y_i \, (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1. \tag{11.68}$$

Let us use Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ for solving the problem. Then, the problem reduces to the unconstrained minimization of

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}|\boldsymbol{w}|^2 - \sum \alpha_i y_i \, (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \, . \tag{11.69}$$

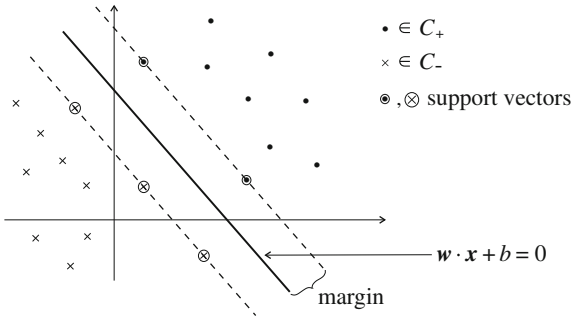


Fig. 11.6 Linear classifier and support vectors

By differentiating it with respect to $\boldsymbol{w}$ and b and making the derivatives equal to 0, we have

$$\sum_i \alpha_i y_i = 0, \quad \boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i. \tag{11.70}$$

Substituting (11.70) in (11.69), the problem is reformulated in the dual form using the dual variables α_i:

$$\text{maximize } L^*(\boldsymbol{\alpha}) = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \tag{11.71}$$

with respect to $\boldsymbol{\alpha}$ under the constraint

$$\alpha_i \geq 0, \quad \sum \alpha_i y_i = 0. \tag{11.72}$$

Since the objective function (11.71) is a quadratic function of α_i, there is a well-known algorithm to solve it. It should be remarked that $\alpha_i = 0$ when $\boldsymbol{x}_i$ is not a support vector.

The optimized output function is written as

$$f(\boldsymbol{x}, \boldsymbol{w}) = \sum \alpha_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x} + b \tag{11.73}$$

in terms of the solution α_i. The function is given by using only the support vectors and the other non-support examples $\boldsymbol{x}_i$ are irrelevant.

The linear output function is useful even when patterns D are not linearly separable. We use slack variables in this case. It can also be used as a regression function, where the output y takes analog values. See textbooks about the support vector machine.

11.2.2 Embedding into High-Dimensional Space

Patterns are not linearly separable in many problems and a linear machine does not work well in many cases. In overcoming this difficulty, it has been known since the early nineteen-sixties that nonlinear embedding of patterns into a high-dimensional space helps. Let us consider a nonlinear transformation of $\boldsymbol{x} \in \boldsymbol{R}^n$ into a high-dimensional space $\boldsymbol{R}^m$ $(m > n)$ by

$$z_i = \varphi_i(\boldsymbol{x}), \quad i = 1, \ldots, m. \tag{11.74}$$

Then pattern $\boldsymbol{x}$ is represented in $\boldsymbol{R}^m$ as

$$z = \boldsymbol{\varphi}(\boldsymbol{x}), \tag{11.75}$$

where

$$\boldsymbol{\varphi}(\boldsymbol{x}) = [\varphi_1(\boldsymbol{x}), \ldots, \varphi_m(\boldsymbol{x})]\,. \tag{11.76}$$

The classification problem is formulated in $\boldsymbol{R}^m$ by using $z = \boldsymbol{\varphi}(\boldsymbol{x})$, where the linear classification function in $\boldsymbol{R}^m$ is written as

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{w} \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + b, \quad \boldsymbol{\xi} = (\boldsymbol{w}, b). \tag{11.77}$$

This was known as the Φ-function method (see Aizerman et al. 1964). The nonlinear embedding improves the linear separability of patterns.

Consider a simple example in which patterns belonging to C_+ are inside a circle and those belonging to C_- are outside the circle (see Fig. 11.7a). The patterns are not linearly separable in $\boldsymbol{R}^2$. However, if we use the following embedding to $\boldsymbol{R}^3$,

$$z_1 = x_1, \quad z_2 = x_2, \quad z_3 = x_1^2 + x_2^2, \tag{11.78}$$

they become linearly separable, as is seen in Fig. 11.7b.

It is expected that patterns become linearly separable when m is large. The multilayer perceptron of Rosenblatt (1961) uses random threshold logic functions in the hidden layer for this purpose. The linear separability is assured when m is sufficiently large. The universality of a three-layer perceptron guarantees that any function $f(\boldsymbol{x})$ can be approximated by a linear function after embedding, provided m is sufficiently large.

However, we need to find good embedding functions for good pattern separation. This is a difficult problem. Moreover, when m is large, in particular infinitely large, calculations of embedded $z = \boldsymbol{\varphi}(\boldsymbol{x})$ are computationally difficult. It is the kernel trick that resolves the difficulty.

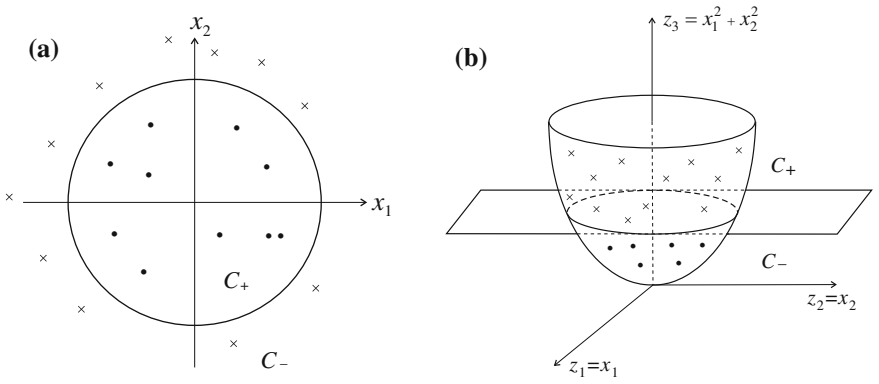


Fig. 11.7 **a** Non-separable in $\boldsymbol{R}_2$; **b** separable in $\boldsymbol{R}^3$

11.2.3 Kernel Method

We consider the inner product of $z = \varphi(\boldsymbol{x})$ and $z' = \varphi\left(\boldsymbol{x}'\right)$ after embedding,

$$K\left(\boldsymbol{x}, \boldsymbol{x}'\right) = z(\boldsymbol{x}) \cdot z\left(\boldsymbol{x}'\right) = \sum z_i(\boldsymbol{x}) z_i\left(\boldsymbol{x}'\right). \tag{11.79}$$

This is a symmetric function of $\boldsymbol{x}$ and $\boldsymbol{x}'$. Moreover, for any coefficients $\boldsymbol{c} = (c_1, \ldots, c_m)$, positivity

$$\sum c_i c_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) > 0 \tag{11.80}$$

is guaranteed for $\boldsymbol{c} \neq 0$, provided $\varphi_1(\boldsymbol{x}), \ldots, \varphi_m(\boldsymbol{x})$ are linearly independent. One may consider that $K\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ is an infinite-dimensional positive-definite matrix in an infinite-dimensional space of $\boldsymbol{z}(\boldsymbol{x})$, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are regarded as indices for specifying the rows and columns of the matrix. That is, $K\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ plays the role of $K(i, j)$, which is a matrix specified by row i and column j.

We consider the eigenvalue problem,

$$\int K\left(\boldsymbol{x}, \boldsymbol{x}'\right) k_i\left(\boldsymbol{x}'\right) d\boldsymbol{x}' = \lambda_i k_i(\boldsymbol{x}), \tag{11.81}$$

where $\lambda_1, \ldots, \lambda_m$ are eigenvalues and $k_1(\boldsymbol{x}), \ldots, k_m(\boldsymbol{x})$ are corresponding eigen-functions. Here, m can be infinite. We call $K\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ the kernel function operating on a function $k(\boldsymbol{x})$ as in the integral (11.81). By using the eigen-functions, the kernel function is expanded as

$$K\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sum \lambda_i k_i(\boldsymbol{x}) k_i\left(\boldsymbol{x}'\right). \tag{11.82}$$

Comparing this with (11.79), we see that the embedding functions are the eigen-functions divided by the square roots of the eigenvalues,

$$z_i(\boldsymbol{x}) = \frac{1}{\sqrt{\lambda_i}} k_i(\boldsymbol{x}). \tag{11.83}$$

The optimal output function (11.73) can be written using the kernel function as

$$f(\boldsymbol{x}, \boldsymbol{w}) = \sum \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{11.84}$$

because of (11.69). This is another expression of (11.77) in terms of the kernel function, where the embedding functions φ are eliminated. Therefore, even when m is infinite, we do not need to calculate $\boldsymbol{z} = \varphi(\boldsymbol{x})$ and the kernel is sufficient to compose the optimal output function. This is called the kernel trick. See Scholkopf (1997) and Shawe-Taylor and Cristianini (2004).

We may start from a kernel function $K(\boldsymbol{x}, \boldsymbol{x}')$, without specifying embedding functions, provided $K(\boldsymbol{x}, \boldsymbol{x}')$ is positive-definite satisfying (11.80), called the Mercer condition.

The Gaussian kernel

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{|\boldsymbol{x} - \boldsymbol{x}'|^2}{\sigma^2}\right\} \tag{11.85}$$

is used frequently, where σ^2 is a free parameter to be adjusted. Its eigen-functions are

$$k_{\omega}(\boldsymbol{x}) = \exp\{-i\boldsymbol{\omega} \cdot \boldsymbol{x}\} \tag{11.86}$$

so that the expansion of a function $f(\boldsymbol{x})$ in terms of the eigen-functions corresponds to the Fourier expansion.

Another kernel of frequent use is the polynomial kernel of order p defined by

$$K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}' + 1)^p. \tag{11.87}$$

The eigen-functions are polynomials of $\boldsymbol{x}$ up to certain degrees and m is finite. The kernel method can be used even when $\boldsymbol{x}$ are discrete symbols, by defining an adequate positive-definite kernel. Therefore, it is a powerful tool in symbol processing and bioinformatics.

11.2.4 Riemannian Metric Induced by Kernel

The kernel method is computationally tractable using a modern computer. However, a good choice of kernel depends on the problem to be solved and no good criteria

exist except for trial-and-error. This section considers the geometry induced by a kernel and proposes a method to improve a given kernel (Amari and Wu 1999; Wu and Amari 2002; Williams et al. 2005).

The original space $\boldsymbol{R}^n$ of patterns is embedded in $\boldsymbol{R}^m$, possibly in $\boldsymbol{R}^\infty$, as a curved n-dimensional submanifold. A Riemannian metric is induced in $\boldsymbol{R}^n$ by this embedding. Two nearby points $\boldsymbol{x}$ and $\boldsymbol{x}+d\boldsymbol{x}$ are embedded to $\boldsymbol{\varphi}(\boldsymbol{x})$ and $\boldsymbol{\varphi}(\boldsymbol{x}+d\boldsymbol{x})$, respectively, and the square of their Euclidean distance in $\boldsymbol{R}^m$ is

$$ds^2 = |\boldsymbol{\varphi}(\boldsymbol{x}+d\boldsymbol{x}) - \boldsymbol{\varphi}(\boldsymbol{x})|^2 = \sum \frac{\partial}{\partial x_i}\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \frac{\partial}{\partial x_j}\boldsymbol{\varphi}(\boldsymbol{x}) dx_i dx_j. \tag{11.88}$$

Therefore, the induced Riemannian metric is given by

$$g_{ij}(\boldsymbol{x}) = \left(\frac{\partial}{\partial x_i}\boldsymbol{\varphi}(\boldsymbol{x})\right) \cdot \left(\frac{\partial}{\partial x_j}\boldsymbol{\varphi}(\boldsymbol{x})\right), \tag{11.89}$$

which is expressed in terms of the kernel as

$$g_{ij}(\boldsymbol{x}) = \frac{\partial^2}{\partial x_i \partial x'_j} K\left(\boldsymbol{x}, \boldsymbol{x}'\right)\big|_{\boldsymbol{x}'=\boldsymbol{x}}. \tag{11.90}$$

The volume element at point $\boldsymbol{x}$ is given by

$$dV(\boldsymbol{x}) = \sqrt{\left|g_{ij}(\boldsymbol{x})\right|} dx_1 \cdots dx_n, \tag{11.91}$$

which shows how the volume is enlarged or contracted at around $\boldsymbol{x}$. Since only the support vectors play a role in the output function, we consider expanding neighborhoods of the support vectors in $\boldsymbol{R}^m$, while other parts remain as they are.

To this end, we modify the current kernel $K\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ to

$$\tilde{K}\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sigma(\boldsymbol{x})\sigma\left(\boldsymbol{x}'\right) K\left(\boldsymbol{x}, \boldsymbol{x}'\right), \tag{11.92}$$

where $\sigma(\boldsymbol{x})$ represents how the volume is enlarged at around $\boldsymbol{x}$. It should be large near the support vectors, so

$$\sigma(\boldsymbol{x}) = \sum_i e^{-\kappa_i |\boldsymbol{x}-\boldsymbol{x}_i^*|} \tag{11.93}$$

was chosen in Amari and Wu (1999), Wu and Amari (2002), where $\boldsymbol{x}_i^*$ are the support vectors and κ_i are adequate constants. Later,

$$\sigma(\boldsymbol{x}) = \exp\left[-\kappa \{f(\boldsymbol{x})\}^2\right] \tag{11.94}$$

was proposed as a more natural choice (Williams et al. 2005).

The transformation (11.92) is called the conformal transformation of a kernel. The Riemannian metric changes to

$$\tilde{g}_{ij}(\boldsymbol{x}) = \sigma^2(\boldsymbol{x}) g_{ij}(\boldsymbol{x}) + \sigma_i(\boldsymbol{x})\sigma_j(\boldsymbol{x}) K(\boldsymbol{x}, \boldsymbol{x}) + \sigma(\boldsymbol{x}) \left\{ \sigma_i(\boldsymbol{x}) K_j(\boldsymbol{x}, \boldsymbol{x}) + \sigma_j(\boldsymbol{x}) K_i(\boldsymbol{x}, \boldsymbol{x}) \right\}, \tag{11.95}$$

where

$$\sigma_i = \frac{\partial}{\partial x_i}\sigma(\boldsymbol{x}), \quad K_i(\boldsymbol{x}, \boldsymbol{x}) = \frac{\partial}{\partial x_i} K\left(\boldsymbol{x}, \boldsymbol{x}'\right) |_{\boldsymbol{x}'=\boldsymbol{x}}. \tag{11.96}$$

When

$$K_i(\boldsymbol{x}, \boldsymbol{x}) = 0, \tag{11.97}$$

which is satisfied by the Gaussian kernel, we have a simplified expression

$$\tilde{g}_{ij}(\boldsymbol{x}) = \{\sigma(\boldsymbol{x})\}^2 g_{ij}(\boldsymbol{x}) + \sigma_i(\boldsymbol{x})\sigma_j(\boldsymbol{x}) K(\boldsymbol{x}, \boldsymbol{x}). \tag{11.98}$$

Computer simulations show that the performance of recognition is improved by up to ten percent by a conformal transformation. This might shed light on the problem of choosing a good kernel.

Recently, Lin and Jiang (2015) proposed another method of choosing $\sigma(\boldsymbol{x})$ adaptively from data.

11.3 Stochastic Reasoning: Belief Propagation and CCCP Algorithms

A graphical model specifies stochastic interactions among a number of random variables. Stochastic reasoning is a procedure to estimate the values of unobserved random variables from those of observed variables based on its graphical structure. Belief propagation (BP) (Pearl 1988) and convex-concave computational procedure (CCCP) (Yuille 2002) are methods in frequent use to obtain good estimates in artificial intelligence and machine learning.

The joint probability distributions of random variables in a graphical model form an exponential family. It has a dually flat Riemannian structure, so these algorithms are well understood from the point of view of dual geometry. The present section studies the BP and CCCP algorithms based on the dually flat structure, based on Ikeda et al. (2004a, b). The belief of each node about the value of its variable is propagated through e- and m-projections to obtain a harmonized consensus in BP. It is a merit of dual geometry that a new simplified version of CCCP is derived naturally.

11.3.1 Graphical Model

Let us consider a set of mutually interacting random variables $x_1, \ldots, x_n$. That is, x_i is a random variable of which the value is determined stochastically under the influence of other variables

$$X_i = \left\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right\}. \tag{11.99}$$

A random variable x_j is called a parent of x_i when it is an element of X_i. We study joint probability distributions of $x_1, \ldots, x_n$. The probability of x_i is given by the conditional probability distribution $p\left(x_i \,|X_i\right)$ conditioned on the values of its parents. We use a graph to represent the parent–child relation (Fig. 11.8). The graph is composed of n nodes corresponding to the random variables $x_1, \ldots, x_n$. There is a branch between nodes x_i and x_j when x_j is a parent of x_i. The branches are oriented in this case, but we consider a non-oriented graph by disregarding the direction of a branch. This is called a graphical model of random variables. See Wainright and Jordan (2008) and Lauritzen (1996), for example.

The joint probability distribution is written using the product of the conditional distributions as

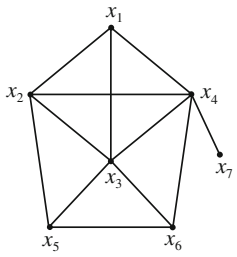
$$p\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} p\left(x_i \,|X_i\right). \tag{11.100}$$

A graphical model is also called a random Markov field. It is an extension of the Markov chain, representing the stochastic causality.

A subgraph composed of nodes $C = \left\{x_{i_1}, \ldots, x_{i_k}\right\}$ is called a clique when it is a complete graph. A graph is complete when any two nodes in it are connected by a branch. See Fig. 11.7, where $\{x_1, x_2, x_3, x_4\}$, $\{x_4, x_7\}$ and $\{x_3, x_5, x_6\}$ are examples of cliques, but $\{x_1, x_2, x_3, x_5\}$ and $\{x_3, x_7\}$ are not. Assume that a graphical model has L cliques $C_1, \ldots, C_L$. Then, it is known that the joint probability distribution (11.100) is decomposed as

$$p\left(x_1, \ldots, x_n\right) = c \prod_{i,r} \tilde{\phi}_i\left(x_i\right) \phi_r\left(C_r\right), \tag{11.101}$$

Fig. 11.8 Graphical model



where c is a normalization constant, $\tilde{\phi}_i, i = 1, \ldots, n$, is a function of x_i and $\phi_r(C_r), r = 1, \ldots, L$, is a function of the variables in clique C_r. The decomposition is not unique in general, but is unique when we use only maximal cliques. A clique is maximal when it is not included in any complete subgraphs.

Divide the nodes of a graphical model into two parts, X_o and X_u. Assume that values of the variables in X_o are observed but those in X_u are not. Stochastic reasoning is the problem of estimating the values of unobserved variables in X_u, under the condition that the variables in X_o are observed. We use the conditional probability of X_u conditioned on X_o to estimate the unknown values of X_u.

Let us fix the values of X_o and consider the conditional distribution of X_u,

$$q(X_u) = p(X_u | X_o), \tag{11.102}$$

where X_o is omitted in the notation of $q(X_u)$. It is again represented by a graphical model consisting of nodes of X_u. So the problem is the estimation of the values of X_u in the reduced graphical model, where the values of X_o are fixed and omitted from the notation. We hereafter denote X_u simply as X and use the vector notation

$$\boldsymbol{x} = (x_1, \ldots, x_n). \tag{11.103}$$

We consider the simple binary case where each x_i takes binary values 1 and -1. The maximum likelihood estimate $\boldsymbol{x}$ based on $q(\boldsymbol{x})$ is the maximizer of $q(\boldsymbol{x})$. However, this is computationally heavy when n is large, because there are 2^n $\boldsymbol{x}$'s and we need to compare the values of $q(\boldsymbol{x})$ for all of them. We use the following simple estimate that the estimated value of x_i is 1 when the probability of $x_i = 1$ is larger than that of $x_i = -1$, and otherwise -1. In other words, let us calculate the expectation of x_i,

$$\eta_i = \mathrm{E}[x_i] = \mathrm{Prob}\{x_i > 0\} - \mathrm{Prob}\{x_i < 0\} \tag{11.104}$$

and $x_i = 1$ when η_i is positive and $x_i = -1$, when η_i is negative. That is, the estimate is given by

$$x_i = \mathrm{sign}\ \eta_i. \tag{11.105}$$

This minimizes the sum of the error probabilities of all the variables.

The problem reduces to the calculation of the expected value of x_i. However, this is again computationally heavy, because

$$\eta_i = \mathrm{E}[x_i] = \sum_{x_1, x_2, \ldots, x_n} x_i q(x_1, \ldots, x_n) \tag{11.106}$$

includes 2^n terms.

We need a computationally tractable algorithm of obtaining a good approximation of the mean values. This problem appears in physics, too, and the mean field approximation is well known to obtain such an approximate solution.

11.3.2 Mean Field Approximation and m-Projection

The probability distribution (11.101) of a graphical model can be written as

$$q(\boldsymbol{x}) = \exp\left\{\sum h_i x_i + \sum_r c_r(\boldsymbol{x}) - \psi\right\}, \tag{11.107}$$

where ψ is the normalization constant, called free energy in physics, with

$$h_i = \frac{1}{2}\log\frac{\tilde{\phi}_i(x_i = 1)}{\tilde{\phi}_i(x_i = -1)} \tag{11.108}$$

and

$$c_r(\boldsymbol{x}) = \log\phi_r\left(x_{r_1}, \ldots, x_{r_s}\right) \tag{11.109}$$

is the term due to clique $C_r = \left\{x_{r_1}, \ldots, x_{r_s}\right\}$.

We consider a new expanded exponential family

$$\tilde{M} = \{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{v})\}, \tag{11.110}$$

$$p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{v}) = \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x} + \sum v_r c_r(\boldsymbol{x}_r) - \psi(\boldsymbol{\theta}, \boldsymbol{v})\right\}, \tag{11.111}$$

which includes two e-affine parameters, namely $\boldsymbol{\theta}$ and $\boldsymbol{v} = (v_1, \ldots, v_L)$. The original $q(\boldsymbol{x})$ is a member of this family and is given by

$$\boldsymbol{\theta} = \boldsymbol{h},\ \boldsymbol{v} = (1, 1, \ldots, 1). \tag{11.112}$$

When $\boldsymbol{v} = 0$, the distributions do not include interaction terms so that the submanifold specified by $\boldsymbol{v} = 0$ is the family of independent distributions of $\boldsymbol{x}$. We denote it by

$$M_0 = \{p_0(\boldsymbol{x}, \boldsymbol{\theta})\} = \{\exp\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \psi(\boldsymbol{\theta})\}\}. \tag{11.113}$$

Figure 11.9 shows the expanded model $\tilde{M}$ and the independent model M_0. The expectation of $\boldsymbol{x}$ is easily calculated for a distribution in M_0, because all $x_1, \ldots, x_n$ are independent. It is given by

$$\eta_i = \mathrm{E}[x_i] = \frac{e^{\theta_i} - e^{-\theta_i}}{e^{\theta_i} + e^{-\theta_i}} = \tanh(\theta_i). \tag{11.114}$$

Given $q(\boldsymbol{x})$, we consider the independent distribution $p^*(\boldsymbol{x}) \in M_0$ that has the same expected value of $\boldsymbol{x}$ as $q(\boldsymbol{x})$. The following theorem shows the relation between $q(\boldsymbol{x})$ and $p^*(\boldsymbol{x})$.

Theorem 11.6 *The m-projection of $q(\boldsymbol{x})$ to M_0 keeps the expectation of $\boldsymbol{x}$ invariant.*

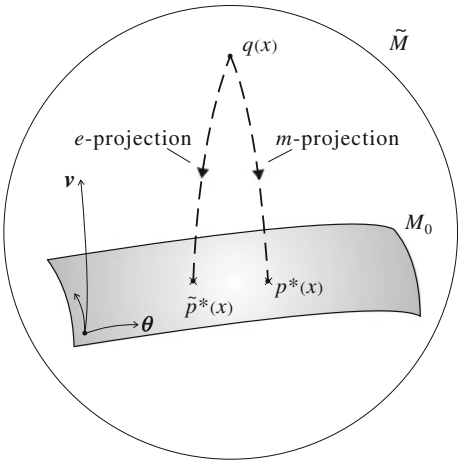


Fig. 11.9 m-projection and e-projection of $\boldsymbol{v}(x)$ to M_0

Proof Let us put

$$p^*(\boldsymbol{x}) = \prod_0^m q(\boldsymbol{x}), \tag{11.115}$$

where $\prod_0^m$ is the operator of m-projection to M_0 and let the e-coordinates of $p^*(\boldsymbol{x})$ be $\boldsymbol{\theta}^*$. The m-coordinates are

$$\boldsymbol{\eta}^* = \mathrm{E}_{p^*}[\boldsymbol{x}]. \tag{11.116}$$

The tangent vector of M_0 at p^* is represented by

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{\theta}^*) = \boldsymbol{x} - \boldsymbol{\eta}^*. \tag{11.117}$$

On the other hand, the tangent vector of the m-geodesic connecting q and p^* is given by

$$t(\boldsymbol{x}) = \frac{q(\boldsymbol{x}) - \boldsymbol{p}^*(\boldsymbol{x})}{p^*(\boldsymbol{x})}. \tag{11.118}$$

They are orthogonal because of the m-projection, so we have

$$\langle t(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{\eta}^* \rangle = \sum \left(\boldsymbol{x} - \boldsymbol{\eta}^*\right) \left\{q(\boldsymbol{x}) - p^*(\boldsymbol{x})\right\} = 0. \tag{11.119}$$

This shows

$$\mathrm{E}_q[\boldsymbol{x}] = \mathrm{E}_{p^*}[\boldsymbol{x}], \tag{11.120}$$

proving the theorem. □

However, the m-projection of $q(\boldsymbol{x})$ is not computationally easy. Statistical physics uses the mean field approximation, which replaces the m-projection by the e-projection (Tanaka 2000, see Amari et al. 2001 for the α-projection). The m-projection is given by the minimizer of the KL-divergence $KL[q : p]$, $p \in M_0$. The mean field approximation uses the dual KL-divergence $KL[p : q]$ and minimizes it with respect to $p \in M_0$. The minimizer is given by the e-projection of q to M_0. This is computationally tractable so it can be used as an approximate solution. See Fujiwara and Shuto (2010) for higher-order mean-field approximation.

We consider

$$q(\boldsymbol{x}) = \exp\left\{\boldsymbol{h}\cdot\boldsymbol{x} + \sum w_{ij}x_i x_j - \psi(\boldsymbol{h}, \boldsymbol{W})\right\} \tag{11.121}$$

as a specific example, which represents a spin system where the interaction of two spins x_i and x_j are given by w_{ij}. The cliques consist of branches (i, j), $w_{ij} \neq 0$. It does not include interactions of more than two nodes and is known as the Boltzmann machine in neural networks, where w_{ij} represents the strength of the synaptic weights of connection between two neurons x_i and x_j.

The KL-divergence from $p \in M_0$ to q is given by

$$\begin{aligned} &KL\left[p(\boldsymbol{x}, \boldsymbol{\theta}) : q(\boldsymbol{x})\right] \\ &\quad = \mathrm{E}_p\left[\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \psi(\boldsymbol{\theta})\} - \left\{\boldsymbol{h}\cdot\boldsymbol{x} + \sum w_{ij}x_i x_j - \psi(\boldsymbol{h}, W)\right\}\right]. \end{aligned} \tag{11.122}$$

It is easy to see

$$\mathrm{E}_p\left[x_i x_j\right] = \eta_i \eta_j, \tag{11.123}$$

because x_i and x_j are independent under p and hence, we have

$$KL[p : q] = \boldsymbol{\theta}\cdot\boldsymbol{\eta} - \psi(\boldsymbol{\theta}) - \boldsymbol{h}\cdot\boldsymbol{\eta} - \sum w_{ij}\eta_i\eta_j + \psi(\boldsymbol{h}, W). \tag{11.124}$$

By differentiating it with respect to η_i and making the derivatives equal to 0, we obtain

$$\eta_i = \tanh\left(\sum w_{ij}\eta_j + h_i\right), \tag{11.125}$$

where

$$\frac{\partial}{\partial\eta_i}\{\boldsymbol{\theta}\cdot\boldsymbol{\eta} - \psi(\boldsymbol{\theta})\} = \tanh^{-1}(\eta_i) \tag{11.126}$$

is taken into account. This is the equation to obtain the minimizer $\tilde{\boldsymbol{\eta}}^*$ of (11.122).

This is a well-known equation. The solution $\tilde{p}^*(\boldsymbol{x})$ is different from the m-projection so that it is an approximation. M_0 is e-flat but not m-flat. Therefore, the m-projection is unique but the e-projection is not necessarily unique. Hence, the solution of (11.125) is not necessarily unique. Moreover, the solution can be a maximum or a saddle point of $KL[p : q]$.

11.3.3 Belief Propagation

Belief propagation is an algorithm, proposed by Pearl (1988), to obtain an approximate value of the expectation of $\boldsymbol{x}$ efficiently. This is a cooperative procedure, where each node exchanges its belief about the expected value through branches. The belief is renewed by taking the beliefs of the other nodes into account. The procedure terminates when a consensus is reached. The information geometry of BP was formulated by Ikeda et al. (2004a, b). We here present a simplified version of it.

Corresponding to each clique C_r, we construct a submodel M_r of $\tilde{M}$,

$$M_r = p\left(\boldsymbol{x}, \boldsymbol{\theta}_r\right) = \exp\left\{\left(\boldsymbol{h} + \boldsymbol{\theta}_r\right) \cdot \boldsymbol{x} + c_r(\boldsymbol{x}) - \psi\left(\boldsymbol{\theta}_r\right)\right\}. \tag{11.127}$$

It includes only one nonlinear term $c_r(\boldsymbol{x})$ corresponding to clique C_r. The sum of all the other interactions, $c_{r'}(\boldsymbol{x})$ of $C_{r'}$, $r' \neq r$, is replaced by a linear term $\boldsymbol{\theta}_r \cdot \boldsymbol{x}$. It is an exponential family, having e-parameter $\boldsymbol{\theta}_r$. This is a submanifold of $\tilde{M}$ obtained by putting

$$\boldsymbol{\theta} = \boldsymbol{h} + \boldsymbol{\theta}_r, \quad v_r = 1, \quad v_{r'} = 0 \text{ for } r' \neq r. \tag{11.128}$$

In addition to the independent submodel M_0, there are L such submodels M_r, $r = 1, \ldots, L$. Since M_r includes only one nonlinear term, it is computationally easy to m-project a member of M_r to M_0.

To avoid complications, we use notational simplification. Since all the probability distributions have the term $\exp(\boldsymbol{h} \cdot \boldsymbol{x})$ in common, we neglect it in the following. This term should be added to the final solution. Mathematically, this corresponds to defining probability densities with respect to the common measure $\exp\{\boldsymbol{h} \cdot \boldsymbol{x}\}$. By this simplification, our target distribution (11.107) is

$$q(\boldsymbol{x}) = \exp\left\{\sum c_r(\boldsymbol{x}) - \psi\right\}, \tag{11.129}$$

and submodels are

$$M_r : p\left(\boldsymbol{x}, \boldsymbol{\theta}_r\right) = \exp\left\{\boldsymbol{\theta}_r \cdot \boldsymbol{x} + c_r(\boldsymbol{x}) - \psi_r\left(\boldsymbol{\theta}_r\right)\right\}, \tag{11.130}$$

$$M_0 : p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}_0 \cdot \boldsymbol{x} - \psi_0(\boldsymbol{\theta}_0)\right\}. \tag{11.131}$$

All the submodels are e-flat in $\tilde{M}$.

Each submodel tries to approximate $q(\boldsymbol{x})$ such that the expectation of $\boldsymbol{x}$ becomes close to $\mathrm{E}_q[\boldsymbol{x}]$. Since M_r includes only one nonlinear term, all the other interaction terms are replaced by the linear term $\boldsymbol{\theta}_r$. They exchange their results concerning the expectation, and eventually reach a consensus satisfying

$$\mathrm{E}_r[\boldsymbol{x}] = \mathrm{E}_0[\boldsymbol{x}], \quad r = 1, \ldots, L, \tag{11.132}$$

where E_r is the expectation with respect to $p_r(\boldsymbol{x}, \boldsymbol{\theta}_r)$. If the consensus is equal to the expectation of $\boldsymbol{x}$ with respect to $q(\boldsymbol{x})$, it is the true solution. But this does not occur in general. However, it would give a good approximation.

We consider the following procedure for reaching the consensus:

1. Initial step: Assign arbitrary initial values $\boldsymbol{\theta}_r^0$ to submodels M_r. They can be 0. Continue the following steps $t = 0, 1, \ldots$ until convergence.
2. $\boldsymbol{m}$-projection step: m-project $p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^t\right)$ at time t of M_r to M_0. Denote the resultant distribution in M_0 by $\tilde{\boldsymbol{\theta}}_{0r}^t$,

$$p_0\left(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}_{0r}^t\right) = \prod\nolimits_0^m p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^t\right). \tag{11.133}$$

3. Calculation of belief of $\boldsymbol{M_r}$: Calculate

$$\boldsymbol{\xi}_r^t = \tilde{\boldsymbol{\theta}}_{0r}^t - \boldsymbol{\theta}_r^t. \tag{11.134}$$

Since the m-projection of $p\left(\boldsymbol{x}, \boldsymbol{\theta}_r^t\right)$ to M_0 is $\tilde{\boldsymbol{\theta}}_{0r}^t$, it includes not only $\boldsymbol{\theta}_r^t$ but also the linearization of the $c_r(\boldsymbol{x})$. Hence, $\boldsymbol{\xi}_r^t$ in (11.134) corresponds to the linearization of the single nonlinear term $c_r(\boldsymbol{x})$. It represents the linearized version of $c_r(\boldsymbol{x})$ in M_0. It is regarded as the belief of M_r that its nonlinear term $c_r(\boldsymbol{x})$ is effectively given by $\boldsymbol{\xi}_r^t$ in M_0.
4. Renewal of the candidate in $\boldsymbol{M_0}$ at $\boldsymbol{t+1}$: Add all the beliefs $\boldsymbol{\xi}_r^t$ of c_r of M_r to give a distribution of M_0 at $t+1$,

$$\boldsymbol{\theta}_0^{t+1} = \sum \boldsymbol{\xi}_r^t. \tag{11.135}$$

5. Renewal of $\boldsymbol{M_r}$ at $\boldsymbol{t+1}$: Construct a new candidate $\boldsymbol{\theta}_r^{t+1}$ of M_r, where the nonlinear terms c_r' $\left(r' \neq r\right)$ other than c_r are replaced by the sum of the beliefs $\boldsymbol{\xi}_{r'}^t$ of $M_{r'}$, but c_r is used as it is. Therefore,

$$\boldsymbol{\theta}_r^{t+1} = \sum_{r' \neq r} \boldsymbol{\xi}_{r'}^t = \boldsymbol{\theta}_0^{t+1} - \boldsymbol{\xi}_r^t. \tag{11.136}$$

When the procedure converges, the converged $\boldsymbol{\theta}_0^*$ and $\boldsymbol{\theta}_r^*$ satisfy

$$p_0\left(\boldsymbol{x}, \boldsymbol{\theta}_0^*\right) = \prod\nolimits_0^m p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^*\right), \tag{11.137}$$

so all the models reach a consensus, having the same expectation of $\boldsymbol{x}$.

11.3.4 Solution of BP Algorithm

We study the solution to which the BP algorithm converges from the geometrical point of view. It should be remarked that there is no guarantee of convergence for the BP algorithm. Note that the CCCP algorithm in the next section always converges.

Theorem 11.7 *When the BP algorithm converges, the following two conditions are satisfied:*

m-condition: $\prod_0^m p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^*\right) = p_0\left(\boldsymbol{x}, \boldsymbol{\theta}_0^*\right)$,

e-condition: $(L-1)\boldsymbol{\theta}_0^* = \sum \boldsymbol{\theta}_r^*$.

Proof The m-condition is the consequence of consensus (11.137). The e-condition is derived by using (11.134) and (11.135). □

We remark that the e-condition is always satisfied for $\boldsymbol{\theta}_0^t$ and $\boldsymbol{\theta}_r^t$ after step 5 of the procedure, but the m-condition is not. The procedure terminates when the m-condition is satisfied. The implications of the two conditions are as follows. See Figs. 11.10 and 11.11. Let M^* be the m-flat submanifold connecting all of $p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^*\right)$ and $p_0\left(\boldsymbol{x}, \boldsymbol{\theta}_0^*\right)$,

$$M^* = \left\{ p(\boldsymbol{x}) \,\middle|\, p(\boldsymbol{x}) = \sum t_r p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^*\right) + \left(1 - \sum t_r\right) p_0\left(\boldsymbol{x}, \boldsymbol{\theta}_0^*\right) \right\}. \tag{11.138}$$

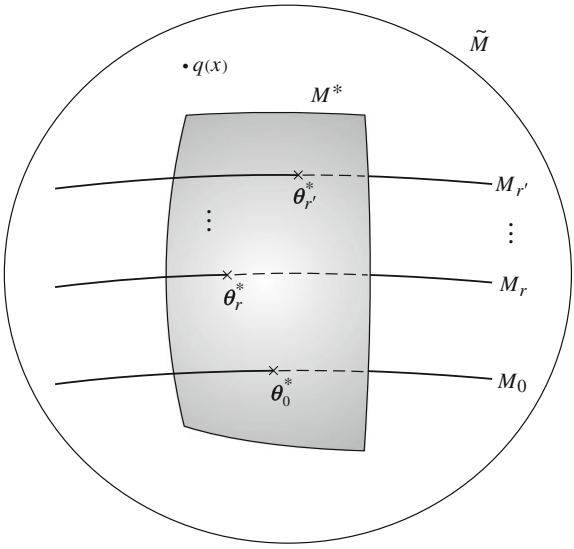


Fig. 11.10 m-condition

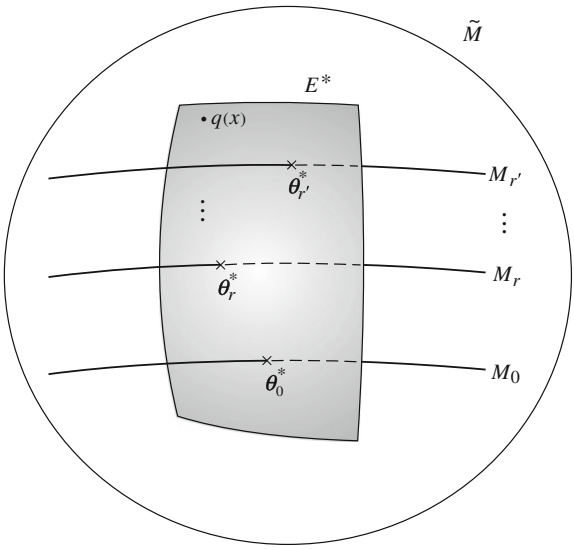


Fig. 11.11 *e*-condition

Let E^* be the e-flat submanifold connecting all of them,

$$E^* = \left\{ \log p(\boldsymbol{x}) = \sum t_r \log p\left(\boldsymbol{x}, \boldsymbol{\theta}^*_r\right) + \left(1 - \sum t_r\right) \log p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*_0\right) - \psi \right\}. \tag{11.139}$$

Corollary *The m- and e-conditions are equivalent to the following two, respectively:*

m*-condition: M^* is orthogonal to M_0.*
e*-condition: E^* includes the true distribution $q(\boldsymbol{x})$.*

If M^* includes $q(\boldsymbol{x})$, its m-projection to M_0 is $\boldsymbol{\theta}^*_0$. The solution is exact in such a case. The following theorem is known.

Theorem 11.8 *When the underlying graph is acyclic, that is, it does not include cycles, M^* includes $q(\boldsymbol{x})$ and the solution gives the exact answer.*

The BP algorithm is stated in geometrical terms in the above explanation. It is beneficial to show the relation between the geometrical algorithm and the conventional BP algorithm written in textbooks. The two are essentially the same. We show only the case where interactions exist between pairs of nodes and no higher-order interactions exist. The conventional algorithm calculates the belief $b(x_i)$ at node x_i and message $m_{ki}(x_i)$, which is transmitted from node x_k to node x_i through branch (i, k). The belief is constructed from the messages by

$$b^t_i(x_i) = \frac{1}{Z} \tilde{\phi}_i(x_i) \prod_{k \in N(i)} m^t_{ki}(x_i), \tag{11.140}$$

where Z is the normalization constant and $N(i)$ is the set of nodes which are connected with node x_i. The messages at $t+1$ are updated by

$$m_{ij}^{t+1}\left(x_j\right) = \frac{1}{Z}\sum_{x_i}\tilde{\phi}_i\left(x_i\right)\phi_{ij}\left(x_i, x_j\right)\prod_{k\in N(i)-j} m_{ki}^{t}\left(x_i\right). \tag{11.141}$$

The correspondence of the quantities appearing in the two approaches are given by

$$\theta_0^i = \frac{1}{2}\log\prod_{k\in N(i)}\frac{m_{ki}\left(x_i=1\right)}{m_{ki}\left(x_i=-1\right)}, \tag{11.142}$$

$$\theta_r^i = \frac{1}{2}\log\prod_{k\in N(i)-j}\frac{m_{ki}\left(x_i=1\right)}{m_{ki}\left(x_i=-1\right)}, \tag{11.143}$$

where r is the branch (clique) connecting i and j.

11.3.5 CCCP (Convex–Concave Computational Procedure)

A new algorithm called CCCP was proposed by Yuille (2002), see also Yuille and Rangarajan (2003). We show a new version of it based on information geometry, which is much simpler than the original one, because the new one does not include double loops in the procedure.

The BP algorithm chooses a set $(\boldsymbol{\theta}_r, \boldsymbol{\theta}_0)$ at each step that satisfies the e-condition and m-projects this set to M_0. It modifies the results toward the satisfaction of the m-condition in the renewal steps. Contrary to this, we may choose $(\boldsymbol{\theta}_r, \boldsymbol{\theta}_0)$ at each step that satisfies the m-condition. Then, we modify them in the renewal steps toward satisfying the e-condition.

This gives a new algorithm (Ikeda et al. 2004a):

1. Initial step: Assign an initial value $\boldsymbol{\theta}_0^0$. It can be $\boldsymbol{\theta}_0^0 = 0$. Do the following iterations until convergence, for $t = 0, 1, 2, \ldots$

2. *m*-condition step: Inversely m-project $p_0\left(\boldsymbol{x}, \boldsymbol{\theta}_0^t\right) \in M_0$ to M_r, that is, to find $p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^t\right) \in M_r$ such that

$$\prod\nolimits_0^m p_r\left(\boldsymbol{x}, \boldsymbol{\theta}_r^t\right) = p_0\left(\boldsymbol{x}, \boldsymbol{\theta}_0^t\right). \tag{11.144}$$

Then, $\left(\boldsymbol{\theta}_0^t, \boldsymbol{\theta}_r^t\right)$ satisfies the m-condition.

3. Renew the $\boldsymbol{\theta}_0^t$ by

$$\boldsymbol{\theta}_0^{t+1} = \sum_r\left(\boldsymbol{\theta}_0^t - \boldsymbol{\theta}_r^t\right) = L\boldsymbol{\theta}_0^t - \sum_r\boldsymbol{\theta}_r^t. \tag{11.145}$$

The e-condition is satisfied when the algorithm converges.

The original form proposed by Yuille (2002) is based on a different idea. In analogy with physics, the BP algorithm is proved to search for the critical point of a function $F(z)$ called free energy where z is the state variables, which in our case is $z = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r)$ (Yedidia et al. 2001). This function is not convex, so there is no guarantee that the gradient descent method converges. Yuille (2002) proved that a function $F(z)$ of z is always decomposed into a sum of a convex function and a concave function,

$$F(z) = F_{\text{convex}}(z) + F_{\text{concave}}(z). \tag{11.146}$$

The decomposition is not unique. The CCCP is an iterative algorithm for obtaining the critical point of F by

$$\nabla E_{\text{convex}}\left(z^{t+1}\right) = -\nabla E_{\text{concave}}\left(z^{t}\right). \tag{11.147}$$

It always converges, whereas BP does not necessarily do so. When it converges, the convergent point satisfies both the m-condition and e-condition.

The original CCCP algorithm by Yuille is written in our geometrical terminology as follows:

1. Calculate $\boldsymbol{\theta}_0^{t+1}$ from
$$\boldsymbol{\theta}_0^{t+1} = L\boldsymbol{\theta}_0^{t} - \sum \boldsymbol{\theta}_r^{t+1}, \tag{11.148}$$
where $\boldsymbol{\theta}_r^{t+1}$ is given by solving
2.
$$p_0\left(\boldsymbol{x}, \boldsymbol{\theta}_0^{t+1}\right) = \prod_0^m \exp\left\{\boldsymbol{\theta}_r^{t+1} \cdot \boldsymbol{x} + c_r(\boldsymbol{x}) - \psi_r\right\}. \tag{11.149}$$

When comparing these with (11.144) and (11.145), $\boldsymbol{\theta}_r^{t+1}$ is used in (11.148) instead of $\boldsymbol{\theta}_r^{t}$ in (11.145). Hence, we need to solve the nonlinear equations to obtain $\boldsymbol{\theta}_0^{t+1}$ and $\boldsymbol{\theta}_r^{t+1}$ in one step. After that, we proceed to the next iteration step increasing t by 1. So it includes double loops and is computationally expensive. Our geometrical algorithm is simpler, and does not include the double loops. The approximation errors due to BP or CCCP are analyzed in Ikeda et al. (2004a) by using the curvature.

11.4 Information Geometry of Boosting

A single learning machine might not be powerful. There is an idea due to M. Kearns and L. Valiant: A powerful machine might be constructed from a number of weak learning machines by integration. This idea was realized by Freund and Schapire (1997) and Schapire et al. (1998) under the name of "boosting". It was shown by Lebanon and Lafferty (2001) that information geometry is useful for understanding

the boosting algorithm. The idea was expanded further by Japanese researchers (including Murata et al. 2004; Takenouchi and Eguchi 2004; Kanamori et al. 2007; and Takenouchi et al. 2008).

11.4.1 Boosting: Integration of Weak Machines

Consider a pattern classifier, which learns from training examples $D = \left\{ \left(\boldsymbol{x}_1, y_1^*\right), \ldots, \left(\boldsymbol{x}_N, y_N^*\right) \right\}$. Here $\boldsymbol{x}_t$ is an input pattern at time t and y_t^* is the correct answer corresponding to $\boldsymbol{x}_t$, which takes binary values 1 and -1. A classifier uses an analog-valued output function $F(\boldsymbol{x})$ and the output y is decided by the decision function $h(\boldsymbol{x})$ which is the signature of $F(\boldsymbol{x})$,

$$y = h(\boldsymbol{x}) = \text{sgn } F(\boldsymbol{x}). \tag{11.150}$$

Assume that we have T weak machines of which the decision functions are

$$h_a(x) = \text{sgn } F_a(x), \quad a = 1, 2, \ldots, T. \tag{11.151}$$

The performance of a weak machine may be very weak, although its error probability should be less than 0.5. By integrating them, we construct a machine of which the output function is

$$F(\boldsymbol{x}) = \sum_{a=1}^{T} \alpha_a h_a(\boldsymbol{x}), \tag{11.152}$$

where α_a are parameters to be determined from the data. See Fig. 11.12. We begin with a weak machine and add new weak machines one by one. The weights α_a are also determined sequentially.

There are two problems to be solved. One is how to compose the next weak machine $h_t(\boldsymbol{x})$ at time t, and the other is how to determine the weight α_t.

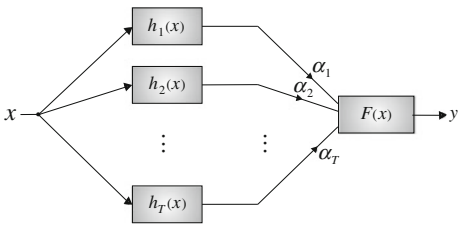


Fig. 11.12 Integration of weak machines

11.4.2 Stochastic Interpretation of Machine

Although a weak learning machine is deterministic, we introduce a stochastic interpretation to evaluate its performance. We consider it as if it were a stochastic machine such that the probability of emitting y is given by

$$q(y|\boldsymbol{x}) = c \exp\left\{\frac{1}{2} y F(\boldsymbol{x})\right\}, \tag{11.153}$$

where c is a normalization constant. Obviously, when $F(\boldsymbol{x})$ takes a large positive value, the probability of $y = 1$ is large and when it takes a negative value with a large magnitude, the probability of $y = -1$ is large. We rewrite (11.153) as

$$q(y|\boldsymbol{x}) = c' \exp\left[\frac{1}{2}\left\{y - y^*(\boldsymbol{x})\right\} F(\boldsymbol{x})\right], \tag{11.154}$$

where $y^*(x)$ is the true output value to $\boldsymbol{x}$ and

$$c' = c \exp\left\{\frac{1}{2} y^*(\boldsymbol{x}) F(\boldsymbol{x})\right\}. \tag{11.155}$$

Note that c' does not depend on y. Since an error occurs when $y = -y^*$, the probability of error for $\boldsymbol{x}$ is

$$q\left(-y_i^* | \boldsymbol{x}_i\right) = c' \exp\left\{-y_i^* F\left(\boldsymbol{x}_i\right)\right\}. \tag{11.156}$$

We define the loss caused by a machine for input $\boldsymbol{x}$

$$\tilde{W}\left(\boldsymbol{x}_i\right) = \exp\left\{-y_i^* F\left(\boldsymbol{x}_i\right)\right\} \tag{11.157}$$

by neglecting the constant c'. We normalize the losses for all the data as

$$W\left(\boldsymbol{x}_i\right) = \frac{1}{Z} \tilde{W}\left(\boldsymbol{x}_i\right), \tag{11.158}$$

where

$$Z = \sum_i \tilde{W}\left(\boldsymbol{x}_i\right). \tag{11.159}$$

Then, $W\left(\boldsymbol{x}_i\right)$ is a distribution of losses over the training examples such that their sum is normalized to 1.

Let I_- be the set of indices i such that $\boldsymbol{x}_i$ are erroneously answered by machine $F(\boldsymbol{x})$. The performance of the machine is evaluated by the error probability

$$\varepsilon_F = \sum_{i \in I_-} W(\boldsymbol{x}_i). \tag{11.160}$$

11.4.3 Construction of New Weak Machines

The weak machines are constructed one by one. Assume that we have constructed t weak machines $h_1(\boldsymbol{x}), \ldots, h_t(\boldsymbol{x})$ and integrated them into the current machine

$$F_t(\boldsymbol{x}) = \sum_{a=1}^{t} \alpha_a h_a(\boldsymbol{x}). \tag{11.161}$$

The performance of a machine is evaluated by the error distribution given by

$$W_t(\boldsymbol{x}_i) = \frac{1}{Z_t} \exp\left\{-y_i^* F_t(\boldsymbol{x}_i)\right\}. \tag{11.162}$$

It is reasonable to add a new machine of which the performance is good for those examples that are bad in the current machine.

To this end, we set up a new machine and train it using the training examples D, but patterns $\boldsymbol{x}_i \in D$ are applied not equally, but with frequency $W_t(\boldsymbol{x}_i)$. This implies that the new training examples are generated from D by resampling such that those which are difficult for the current machine appear frequently. Any type of machine can be used as a new weak machine to be trained, a simple or multilayer perceptron, a support vector machine, a decision tree, and others.

11.4.4 Determination of the Weights of Weak Machines

We add a newly trained weak machine $h_{t+1}(\boldsymbol{x})$ to the previous weak machines, forming a new machine

$$F_{t+1}(\boldsymbol{x}) = \sum_{a=1}^{t} \alpha_a h_a(\boldsymbol{x}) + \alpha h_{t+1}(\boldsymbol{x}). \tag{11.163}$$

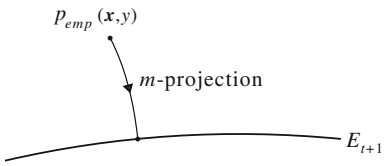


Fig. 11.13 Determination of weight α

Here, α is the parameter to be decided. The conditional probability of y by the new machine is

$$q\left(y\,|\boldsymbol{x},\alpha\right)=c\exp\left\{\frac{1}{2}yF_{t+1}(\boldsymbol{x})\right\}=c\exp\left\{\frac{1}{2}yF_t(\boldsymbol{x})+\frac{1}{2}\alpha yh_{t+1}(\boldsymbol{x})\right\}. \quad (11.164)$$

This forms a one-dimensional exponential family E_{t+1} where the e-coordinate is α. Therefore, given the training data D, the best distribution to fit the training data is given by the m-projection of the empirical distribution of data to the exponential family E_{t+1}. See Fig. 11.13.

The coefficient c in (11.164) is a complicated function of α and D. We ignore this term, considering E_{t+1} as a family of unnormalized positive measures,

$$M=\left\{c\exp\left\{\frac{1}{2}yF_{t+1}(\boldsymbol{x})\right\};\ c>0 \text{ is arbitrary}\right\}. \quad (11.165)$$

Then, the optimum solution is obtained by m-projecting

$$p_{\text{emp}}(y,\boldsymbol{x})=\frac{1}{N}\sum\delta\left(y-y_i^*\right)\delta\left(\boldsymbol{x}-\boldsymbol{x}_i\right) \quad (11.166)$$

to E_{t+1}, that is, by minimizing $KL\left[p_{\text{emp}}:q(y|\boldsymbol{x},\alpha)\right]$. From

$$\tilde{q}(y\,|\boldsymbol{x},\boldsymbol{\alpha})=\exp\left\{\frac{1}{2}\left(y-y^*(\boldsymbol{x})\right)\sum_{i=1}^{t}\alpha_ih_i(\boldsymbol{x})\right\}, \quad (11.167)$$

where c is ignored, the KL-divergence is written as

$$\begin{aligned}
&KL\left[\tilde{p}_{\text{emp}}(y,\boldsymbol{x}):\tilde{q}(y\,|\boldsymbol{x},\boldsymbol{\alpha})\right]\\
&\quad=C-\sum_i\log\tilde{q}\left(y_i\,|\boldsymbol{x}_i,\boldsymbol{\alpha}\right)+\sum_{i,y}q\left(y\,|\boldsymbol{x}_i\right)\\
&\quad=C-\frac{1}{2}\sum_i\left\{y_i^*-y^*\left(\boldsymbol{x}_i\right)\right\}\sum\alpha_jh_j\left(\boldsymbol{x}_i\right)\\
&\qquad+\sum_i\sum_{y_i=1,-1}\exp\left\{\frac{1}{2}\left(y_i-y_i^*\right)\sum\alpha_jh_j\left(\boldsymbol{x}_i\right)\right\},
\end{aligned} \quad (11.168)$$

where C is a term not depending on α. Since $y_i^* = y^*(\boldsymbol{x}_i)$ and the objective function to be minimized is

$$L(D, \alpha) = \sum_i \exp\left\{-y_i^* \left[\sum \alpha_j h_j(\boldsymbol{x}_i) + \alpha h_{t+1}(\boldsymbol{x}_i)\right]\right\}, \tag{11.169}$$

by differentiating it with respect to α, we have

$$\sum_i W_t(\boldsymbol{x}_i)\, y_i^* h_{t+1}(\boldsymbol{x}_i)\, e^{-\alpha\{y_i^* h_{t+1}(\boldsymbol{x}_i)\}} = 0. \tag{11.170}$$

We introduce a new index set I_-^{t+1} such that $i \in I_-^{t+1}$ implies that pattern $\boldsymbol{x}_i$ is wrongly classified by the new machine h_{t+1}, that is,

$$y_i^* h_{t+1}(\boldsymbol{x}_i) = -1. \tag{11.171}$$

Let us put

$$\varepsilon_{t+1} = \sum_{i \in I_-} W_t(\boldsymbol{x}_i), \quad 1 - \varepsilon_{t+1} = \sum_{i \in I_+} W_t(\boldsymbol{x}_i). \tag{11.172}$$

Then, (11.170) reduces to

$$-\varepsilon_{t+1} e^{\alpha} + (1 - \varepsilon_{t+1})\, e^{-\alpha} = 0. \tag{11.173}$$

We obtain the solution

$$\alpha = \frac{1}{2} \log \frac{1 - \varepsilon_{t+1}}{\varepsilon_{t+1}}. \tag{11.174}$$

The weight of example $\boldsymbol{x}_i$ is renewed as

$$W_{t+1}(\boldsymbol{x}_i) = \frac{1}{Z_{t+1}} W_t(\boldsymbol{x}_i) \exp\left\{-\alpha_{t+1} y_i^* h_{t+1}(\boldsymbol{x}_i)\right\}. \tag{11.175}$$

11.5 Bayesian Inference and Deep Learning

Information geometry of Bayesian statistics has not yet been well developed except for preliminary studies (e.g., Zhu and Rohwer 1995). Bayesian theory regards data and parameters as random variables at the same time. Hence, information geometry is applied to their joint probability distributions. It is hoped to construct a deeper structure beyond superficial Bayesian information geometry, which would be useful for machine learning, in particular for deep learning. This section proposes a preliminary trial concerning information geometry of Bayesian statistics. We use the restricted Boltzmann machine (RBM) for this purpose, which is an important constituent in deep learning.

11.5.1 Bayesian Duality in Exponential Family

An exponential family of probability distributions is represented by

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \bar{k}(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{11.176}$$

where $\boldsymbol{x}$ is a vector random variable, $\boldsymbol{\theta}$ is a vector parameter and $\bar{k}(\boldsymbol{x})$ corresponds to the underlying measure of $\boldsymbol{x}$,

$$d\mu(\boldsymbol{x}) = \exp\left\{-\bar{k}(\boldsymbol{x})\right\} d\boldsymbol{x}. \tag{11.177}$$

Bayesian statistics assumes that the parameter $\boldsymbol{\theta}$ is also a random variable subject to a prior distribution $\pi(\boldsymbol{\theta})$. Then, the joint probability of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ is

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \bar{k}(\boldsymbol{x}) - \bar{\psi}(\boldsymbol{\theta})\right\}, \tag{11.178}$$

where

$$\bar{\psi}(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}) - \log\pi(\boldsymbol{\theta}). \tag{11.179}$$

The Bayesian posterior distribution is the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ and is written as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \bar{\psi}(\boldsymbol{\theta}) - k(\boldsymbol{x})\right\}, \tag{11.180}$$

where

$$k(\boldsymbol{x}) = \bar{k}(\boldsymbol{x}) + \log p(\boldsymbol{x}), \tag{11.181}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{11.182}$$

It is an exponential family, where the random variable is $\boldsymbol{\theta}$ and the natural parameter to specify a distribution is $\boldsymbol{x}$. Although the roles of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are different, the conditional distributions have the same exponential form shown in (11.176) and (11.180). We call it the Bayesian duality.

The e-affine parameter is $\boldsymbol{\theta}$ in the manifold of probability distributions (11.176) and hence, the dual m-affine parameter is

$$\boldsymbol{\eta} = \mathrm{E}_{\theta}[\boldsymbol{x}] = \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}. \tag{11.183}$$

whereas, the e-affine parameter is $\boldsymbol{x}$ in the manifold of the posterior probability distributions (11.180) and hence the m-affine parameter is the conditional posterior expectation of $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^* = \mathrm{E}_{x}[\boldsymbol{\theta}] = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}. \tag{11.184}$$

We extend (11.178) to a family of joint probability distributions parameterized by a hyper parameter $\boldsymbol{\zeta}$. Then, $M = \{p(\boldsymbol{x}, \boldsymbol{\theta}; \boldsymbol{\zeta})\}$ forms a manifold consisting of exponential families. Its simple example is the case when a prior distribution $\pi(\boldsymbol{\theta})$ is given in a parametric form as $\pi(\boldsymbol{\theta}, \boldsymbol{\zeta})$. Here, the extra parameter $\boldsymbol{\zeta}$ is called a hyper parameter. A family of prior distributions called conjugate priors is used sometimes, because of its simplicity. A conjugate prior $\pi(\boldsymbol{\theta}, \boldsymbol{\zeta})$ has the same form as the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$. In our exponential case, because of (11.180), the conjugate prior is written as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \exp\{\boldsymbol{\alpha} \cdot \boldsymbol{\theta} - \beta\psi(\boldsymbol{\theta}) - \chi(\boldsymbol{\alpha}, \beta)\}, \tag{11.185}$$

where $\boldsymbol{\zeta} = (\boldsymbol{\alpha}, \beta)$ is the hyper parameter and $\chi(\boldsymbol{\alpha}, \beta)$ is a normalization factor. When we use N independent observations $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, the posterior distribution under prior $\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}, \beta)$ is explicitly given by

$$p(\boldsymbol{\theta}|D, \boldsymbol{\alpha}, \beta) = \exp\{\boldsymbol{\theta} \cdot (\boldsymbol{\alpha} + N\bar{\boldsymbol{x}}) - (N + \beta)\psi(\boldsymbol{\theta}) - \chi(\boldsymbol{\alpha}, \boldsymbol{\beta})\}, \tag{11.186}$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum \boldsymbol{x}_i \tag{11.187}$$

is the observed point. This makes the role of the conjugate prior clear: The conjugate prior has the effect of shifting the observed point from $\bar{\boldsymbol{x}}$ to $\bar{\boldsymbol{x}} + \boldsymbol{\alpha}/N$, that is, of adding β additional pseudo-observations of which the observed value is $\boldsymbol{\alpha}/\beta$ to the previous $N\bar{\boldsymbol{x}}$. Alternatively, observed data D change the parameter of the conjugate prior as follows:

$$\boldsymbol{\alpha} \to \boldsymbol{\alpha} + N\bar{\boldsymbol{x}}, \quad \beta \to \beta + N. \tag{11.188}$$

The geometry of the conjugate prior is studied by Agarwal and Daumé III (2010).

We can enlarge our framework by considering a curved exponential family, where $\boldsymbol{\theta}$ is specified by a low-dimensional parameter $\boldsymbol{u}$ such that

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{u}). \tag{11.189}$$

The random variable $\boldsymbol{x}$ may be an embedded version of low-dimensional signals $\boldsymbol{v}$,

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{v}). \tag{11.190}$$

Then, probability distributions of $\boldsymbol{u}$ and $\boldsymbol{v}$ form a curved exponential family.

We may further consider an extended family of distributions such that a joint distribution (11.178) is specified by an additional parameter W as $p(\boldsymbol{x}, \boldsymbol{\theta}; W)$. We use this as a model of machine learning or the brain. Here, $\boldsymbol{x}$ or $\boldsymbol{x}(\boldsymbol{v})$ is information given from the environment. $\boldsymbol{\theta}$ or $\boldsymbol{\theta}(\boldsymbol{u})$ represents a higher-order concept which specifies the distribution of $\boldsymbol{x}$. An inference system guesses $\boldsymbol{\theta}$ from $\boldsymbol{x}$ such that $\boldsymbol{x}$ is generated from $p(\boldsymbol{x}|\boldsymbol{\theta})$. See Fig. 11.14. This is a simple layered model of the brain,

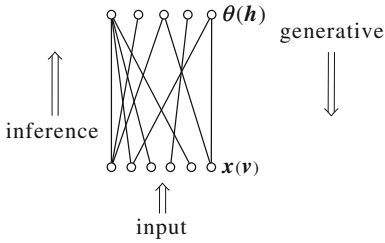


Fig. 11.14 Bayesian inference of higher information $\boldsymbol{\theta}$ from $\boldsymbol{x}$

where $\boldsymbol{x}$ is given to an input layer and $\boldsymbol{\theta}$ is generated in the next layer by Bayesian inference. There may be feedback connections from the higher-order layer to the lower-order layer so that a dynamical process takes place between them. The RBM is its stochastic model.

11.5.2 Restricted Boltzmann Machine

The Boltzmann machine was proposed by Ackley et al. (1985). It is a Markov chain over state $\boldsymbol{x}$, of which the stable distribution is given by

$$p(\boldsymbol{x}, \boldsymbol{c}, \mathbf{W}) = \exp\left\{\boldsymbol{c} \cdot \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \mathbf{W} \boldsymbol{x} - \psi\right\}, \tag{11.191}$$

where $\boldsymbol{c}$ is a vector and $\mathbf{W}$ is a symmetric matrix.

The restricted Boltzmann machine (RBM) is a layered machine consisting of two layers and there are no interactions among elements (we call them neurons) within each layer. Interactions (connections) exist only between neurons of different layers. This was proposed by Smolensky (1986) and has been extensively used in deep learning (Hinton and Salakhutdinov 2006 and others).

We divide $\boldsymbol{x}$ into two parts, $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{h})$, where $\boldsymbol{v}$ and $\boldsymbol{h}$ are binary vector random variables representing activities of neurons of the two layers in the RBM (Fig. 11.15). The first layer is called an input layer or visible layer, to which a signal $\boldsymbol{v}$ is applied from the environment. The second layer is called a hidden layer of which activity pattern $\boldsymbol{h}$ is generated from input $\boldsymbol{v}$ in the first layer.

The stable probability distribution of an RBM is written as

$$p(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W}) = \exp\left\{\boldsymbol{a} \cdot \boldsymbol{v} + \boldsymbol{b} \cdot \boldsymbol{h} + \boldsymbol{h}^T \mathbf{W} \boldsymbol{v} - \psi(\boldsymbol{a}, \boldsymbol{b}, \mathbf{W})\right\}, \tag{11.192}$$

since there are no connections among the neurons in each layer. This is an exponential family of distributions. The stable probabilities of $\boldsymbol{v}$ and $\boldsymbol{h}$ are given by its marginal probability distributions,

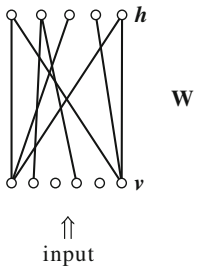


Fig. 11.15 RBM (restricted Boltzmann machine)

$$p_V(\boldsymbol{v}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}), \tag{11.193}$$

$$p_H(\boldsymbol{h}) = \sum_{\boldsymbol{v}} p(\boldsymbol{v}, \boldsymbol{h}), \tag{11.194}$$

and they are not of the exponential type.

We compare the RBM with the Bayesian scheme in the previous section. When the number m of neurons in the hidden layer is smaller than the number n in the visible layer, we introduce new random variables by

$$\boldsymbol{\theta} = \boldsymbol{h}, \tag{11.195}$$

$$\boldsymbol{x} = \mathbf{W}\boldsymbol{v}. \tag{11.196}$$

In the opposite case, we introduce

$$\boldsymbol{\theta} = \boldsymbol{h}^T\mathbf{W}, \tag{11.197}$$

$$\boldsymbol{x} = \boldsymbol{v}. \tag{11.198}$$

In either case, the stationary probability distribution is written in the standard form (11.178) of Bayesian joint distribution. Therefore, we may consider an RBM as representing the Bayesian mechanism of statistical inference.

11.5.3 Unsupervised Learning of RBM

For an RBM having the stationary joint probability (11.192), we have the two conditional distributions

$$p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W}) = \frac{p(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W})}{p_V(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W})}, \tag{11.199}$$

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W}) = \frac{p(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W})}{p_H(\boldsymbol{h}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W})}. \tag{11.200}$$

They show the probabilities of activities of one layer given the activities of the other layer. Let $q(\boldsymbol{v})$ be a probability distribution of $\boldsymbol{v}$ given from the environment, subject to which input $\boldsymbol{v}$ is generated. An RBM is trained by receiving $\boldsymbol{v}$ such that its stationary marginal distribution $p_V(\boldsymbol{v}; \boldsymbol{a}, \boldsymbol{b}, \mathbf{W})$ approximates $q(\boldsymbol{v})$. This is done by modifying $\mathbf{W}$, $\boldsymbol{a}$ and $\mathbf{b}$ so that the KL-divergence $D_{KL}[q(\boldsymbol{v}) : p_V(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{b}, \mathbf{W})]$ is minimized. The minimizing $\mathbf{W}$, $\boldsymbol{a}$, $\boldsymbol{b}$ are the maximum likelihood estimator. For the sake of notational simplicity, we hereafter neglect the bias terms $\boldsymbol{a}$ and $\boldsymbol{b}$ by making them equal to 0, but they can be treated in a similar manner. This is only for the purpose of avoiding unnecessary complication.

Let M_V be a submanifold consisting of the marginal probability distributions of $\boldsymbol{v}$ of the RBM,

$$M_V = \{p(\boldsymbol{v}, \mathbf{W})\} \tag{11.201}$$

in the entire manifold S_V of probability distributions of $\boldsymbol{v}$. The minimizer $\mathbf{W}$ of the KL-divergence $D_{KL}[q(\boldsymbol{v}) : p(\boldsymbol{v}, \mathbf{W})]$ is given by the m-projection of $q(\boldsymbol{v})$ to the submanifold M_V (Fig. 11.16). However, it is simpler to treat the manifold of joint distributions of $(\boldsymbol{v}, \boldsymbol{h})$ rather than the marginal distributions of $\boldsymbol{v}$. To this end, we consider a manifold $S_{V,H}$ consisting of all joint probability distributions of $\boldsymbol{v}$ and $\boldsymbol{h}$. We study two submanifolds in it. One is the submanifold of the RBM,

$$M_{V,H} = \{p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})\}, \tag{11.202}$$

parameterized by $\mathbf{W}$. The other is the data submanifold $M_{V,H|q}$ given by

$$M_{V,H|q} = \{q(\boldsymbol{v})r(\boldsymbol{h}|\boldsymbol{v})\}, \tag{11.203}$$

where $q(\boldsymbol{v})$ is fixed and $r(\boldsymbol{h}|\boldsymbol{v})$ is an arbitrary conditional distribution of $\boldsymbol{h}$ conditioned on $\boldsymbol{v}$. The marginal distribution of any member of $M_{V,H|q}$ is $q(\boldsymbol{v})$. Consider the KL-divergence between the two submanifolds,

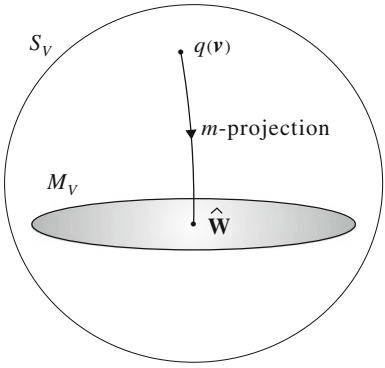


Fig. 11.16 m-projection of $q(\boldsymbol{v})$ to M_V

$$D_{KL}\left[M_{V,H|q} : M_{V,H}\right] = \min_{r,\mathbf{W}} D_{KL}\left[q(\boldsymbol{v})r(\boldsymbol{h}|\boldsymbol{v}) : p(\boldsymbol{v},\boldsymbol{h},\mathbf{W})\right]. \tag{11.204}$$

Theorem 11.9 *The minimizers of the KL-divergence $D_{KL}\left[M_{V,H|q} : M_{V,H}\right]$ between two submanifolds are given by $r(\boldsymbol{h}|\boldsymbol{v}) = p(\boldsymbol{h}|\boldsymbol{v},\hat{\mathbf{W}})$ and $p(\boldsymbol{v},\boldsymbol{h},\hat{\mathbf{W}})$, where $\hat{\mathbf{W}}$ is the MLE of $p_V(\boldsymbol{v},\mathbf{W})$ for data $\boldsymbol{v}$ generated from $q(\boldsymbol{v})$.*

Proof We can decompose D_{KL} as follows:

$$\begin{aligned}
&D_{KL}\left[q(\boldsymbol{v})r(\boldsymbol{h}|\boldsymbol{v}) : p(\boldsymbol{v},\boldsymbol{h},\mathbf{W})\right] \\
&= \int q(\boldsymbol{v})r(\boldsymbol{h}|\boldsymbol{v}) \log \frac{q(\boldsymbol{v})r(\boldsymbol{h}|\boldsymbol{v})}{p(\boldsymbol{v},\boldsymbol{h},\mathbf{W})} d\boldsymbol{v}d\boldsymbol{h} \\
&= \int q(\boldsymbol{v})r(\boldsymbol{h}|\boldsymbol{v}) \left\{ \log \frac{q(\boldsymbol{v})}{p(\boldsymbol{v},\mathbf{W})} + \log \frac{r(\boldsymbol{h}|\boldsymbol{v})}{p(\boldsymbol{h}|\boldsymbol{v},\mathbf{W})} \right\} d\boldsymbol{h}d\boldsymbol{v} \\
&= D_{KL}\left[q(\boldsymbol{v}) : p(\boldsymbol{v},\mathbf{W})\right] + \int q(\boldsymbol{v}) D_{KL}\left[r(\boldsymbol{h}|\boldsymbol{v}) : p(\boldsymbol{h}|\boldsymbol{v},\mathbf{W})\right] d\boldsymbol{v}.
\end{aligned} \tag{11.205}$$

Therefore, the minimum of the D_{KL} with respect to $r(\boldsymbol{h}|\boldsymbol{v})$ is attained by $p(\boldsymbol{h}|\boldsymbol{v},\mathbf{W})$ and the minimum with respect to $\mathbf{W}$ is attained by the minimizer of $D_{KL}[q(\boldsymbol{v}) : p(\boldsymbol{v},\mathbf{W})]$. □

Let $\hat{q} = q(\boldsymbol{v})\hat{r}(\boldsymbol{v}|\boldsymbol{h})$ and $\hat{p} = p\left(\boldsymbol{v},\boldsymbol{h},\hat{\mathbf{W}}\right)$ be the closest pair of $D_{KL}\left[M_{V,H|q} : M_{V,H}\right]$. Then, $\hat{p}$ is given by the m-projection of $\hat{q}$ and $\hat{q}$ is the e-projection of $\hat{p}$. This is clear from the *em* (EM) algorithm in the presence of hidden variable $\boldsymbol{h}$, since the e-projection keeps the conditional probability $p(\boldsymbol{h}|\boldsymbol{v},\mathbf{W})$ and the m-projection maximizes the log likelihood. See Fig. 11.17, where the minimization problem in $S_{V,H}$ is mapped onto that in S_V.

We now give the learning algorithm established by Ackley et al. (1985). This is the stochastic descent method of D_{KL}.

Theorem 11.10 *The averaged learning rule of RBM is given by*

$$\Delta W_{ij} = \varepsilon\left(\langle h_i v_j\rangle_q - \langle h_i v_j\rangle_p\right), \tag{11.206}$$

where ε is a learning constant, $\langle h_i v_j\rangle_q$ is the average of $h_i v_j$ subject to the joint probability distribution

$$q(\boldsymbol{v},\boldsymbol{h};\mathbf{W}) = q(\boldsymbol{v})p(\boldsymbol{h}\,|\boldsymbol{v},\mathbf{W}) \tag{11.207}$$

and $\langle h_i v_j\rangle_p$ is the average over the stationary distribution $p(\boldsymbol{v},\boldsymbol{h},\mathbf{W})$ of the RBM.

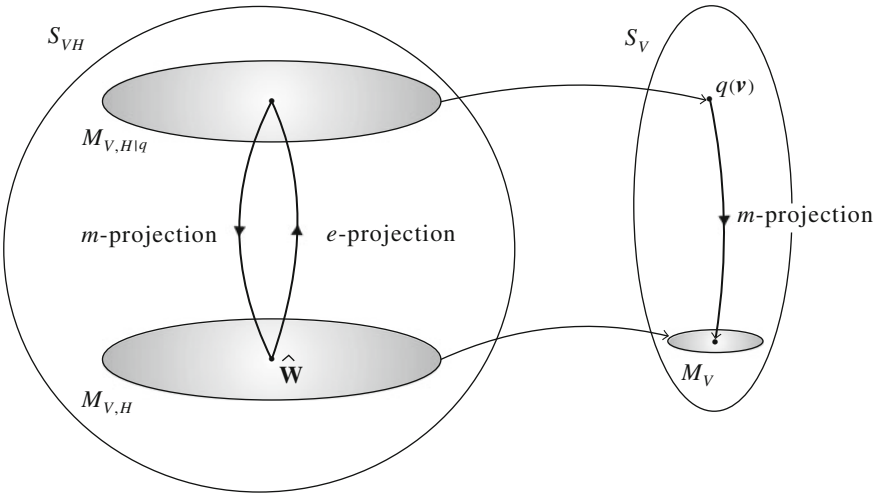


Fig. 11.17 Minimizer of $D_{KL}\left[M_{V,H|V} : H_{V,H}\right]$

Proof Since we have

$$D_{KL}\left[q(\boldsymbol{v}) : p(\boldsymbol{v}, \mathbf{W})\right] = \int q(\boldsymbol{v}) \log q(\boldsymbol{v}) d\boldsymbol{v} - \int q(\boldsymbol{v}) \left(\log \int \exp\left\{\boldsymbol{h}^T \mathbf{W} \boldsymbol{v} - \psi\right\} d\boldsymbol{h} \right) d\boldsymbol{v}, \tag{11.208}$$

we have

$$\begin{aligned} \frac{\partial D_{KL}}{\partial \mathbf{W}} &= -\int q(\boldsymbol{v}) \left(\boldsymbol{h}\boldsymbol{v}^T - \frac{\partial}{\partial \mathbf{W}} \psi \right) \frac{p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})}{p(\boldsymbol{v}, \mathbf{W})} d\boldsymbol{h} d\boldsymbol{v} \\ &= -\int \boldsymbol{h}\boldsymbol{v}^T q(\boldsymbol{v}) p(\boldsymbol{h}|\boldsymbol{v}, \mathbf{W}) d\boldsymbol{v} d\boldsymbol{h} + \frac{\partial}{\partial \mathbf{W}} \psi \\ &= -\langle \boldsymbol{h}\boldsymbol{v}^T \rangle_{q(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v})} + \langle \boldsymbol{h}\boldsymbol{v}^T \rangle_{p(\boldsymbol{v},\boldsymbol{h})}, \end{aligned} \tag{11.209}$$

because

$$\frac{\partial \psi}{\partial \mathbf{W}} = E_p\left[\boldsymbol{h}\boldsymbol{v}^T\right]. \tag{11.210}$$

□

This is the ordinary gradient descent method. The natural gradient method would work better, if we had its computational algorithm. Since the learning rule (11.206) includes only the expectation of the cross term of $\boldsymbol{v}$ and $\boldsymbol{h}$ with respect to $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ and $q(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$, all the other higher-order interaction terms are irrelevant. Therefore, this suggests the use of a mixed coordinate system, which separates the second-order terms from higher-order terms of interactions (see Akaho and Takabatake 2008).

11.5.4 Geometry of Contrastive Divergence

The learning algorithm (11.206) is computationally heavy. This is because, in order to calculate the expectation of $\langle \boldsymbol{h}\boldsymbol{v}^T \rangle_p$, we need a long run of MCMC procedures for obtaining samples from the stable distribution $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$. The MCMC procedures work as follows:

1. Begin with an arbitrary $\boldsymbol{v}_t$ and generate $\boldsymbol{h}_t$ by using the conditional distribution $p(\boldsymbol{h}|\boldsymbol{v}, \mathbf{W})$.
2. Generate $\boldsymbol{v}_{t+1}$ from the current $\boldsymbol{h}_t$ by using the conditional distribution $p(\boldsymbol{v}|\boldsymbol{h}, \mathbf{W})$.
3. Repeat the procedures, $t = 0, 1, 2, \ldots$.

We then have a sequence of $(\boldsymbol{v}_t, \boldsymbol{h}_t)$ of which the empirical distribution converges to $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$. These data can be used to calculate the average $\langle \boldsymbol{h}\boldsymbol{v}^T \rangle_p$ in (11.206) or (11.209).

The contrastive divergence is an approximation of the KL-divergence, proposed by Hinton (2002). This has been used frequently in deep learning. It runs a finite number, say k, of iterations of the above procedures. The order k contrastive divergence (CD_k) uses a pair of $(\boldsymbol{v}_k, \boldsymbol{h}_k)$, where $\boldsymbol{v}_0$ is derived from $q(\boldsymbol{v})$ as an initial value, $\boldsymbol{h}_t$ is derived from $p(\boldsymbol{h}|\boldsymbol{v}_t; \mathbf{W})$ and $\boldsymbol{v}_{t+1}$ is derived from $p(\boldsymbol{v}|\boldsymbol{h}_t; \mathbf{W})$. Repeating the procedures up to $t = k$ from many initial $\boldsymbol{v}$'s, the derived empirical distribution of $(\boldsymbol{v}_k, \boldsymbol{h}_k)$ is used to obtain an approximation of $\langle \boldsymbol{h}\boldsymbol{v}^T \rangle_p$.

We study the probability distribution $p_k(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ of $(\boldsymbol{v}_k, \boldsymbol{h}_k)$, which we call the CD_k distribution, following Karakida et al. (2014). Let its marginal distributions be $p_{Vk}(\boldsymbol{v}, \mathbf{W})$ and $p_{Hk}(\boldsymbol{h}, \mathbf{W})$. They are

$$p_{Vk}(\boldsymbol{v}, \mathbf{W}) = \int p_k(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})\, d\boldsymbol{h}, \tag{11.211}$$

$$p_{Hk}(\boldsymbol{h}, \mathbf{W}) = \int p_k(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})\, d\boldsymbol{v}. \tag{11.212}$$

Then, the CD_j distributions are calculated recursively by

$$p_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W}) = p_{Vj}(\boldsymbol{v}) p(\boldsymbol{h}|\boldsymbol{v}, \mathbf{W}), \quad j = 0, \ldots, k, \tag{11.213}$$

$$\tilde{p}_{Hj+1}(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W}) = p_{Hj}(\boldsymbol{h}, \mathbf{W}) p(\boldsymbol{v}|\boldsymbol{h}, \mathbf{W}). \tag{11.214}$$

In order to understand the CD_k distributions, we consider two submanifolds $M_{H|V}(\mathbf{W})$ and $\tilde{M}_{V|H}(\mathbf{W})$ in $S_{V,H}$. They are defined by

$$M_{H|V}(\mathbf{W}) = \{r(\boldsymbol{v}) p(\boldsymbol{h}|\boldsymbol{v}, \mathbf{W})\}, \tag{11.215}$$

$$\tilde{M}_{V|H}(\mathbf{W}) = \{\tilde{r}(\boldsymbol{h}) p(\boldsymbol{v}|\boldsymbol{h}, \mathbf{W})\}, \tag{11.216}$$

where $r(\boldsymbol{v})$ and $\tilde{r}(\boldsymbol{h})$ are arbitrary distributions. They intersect at $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$, because, when $r(\boldsymbol{v}) = p_V(\boldsymbol{v}, \mathbf{W})$ and when $\tilde{r}(\boldsymbol{h}) = p_H(\boldsymbol{h}, \mathbf{W})$, both the distributions are equal to $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$. Moreover, both $M_{H|V}$ and $\tilde{M}_{V|H}$ are e-flat, because the e-geodesic

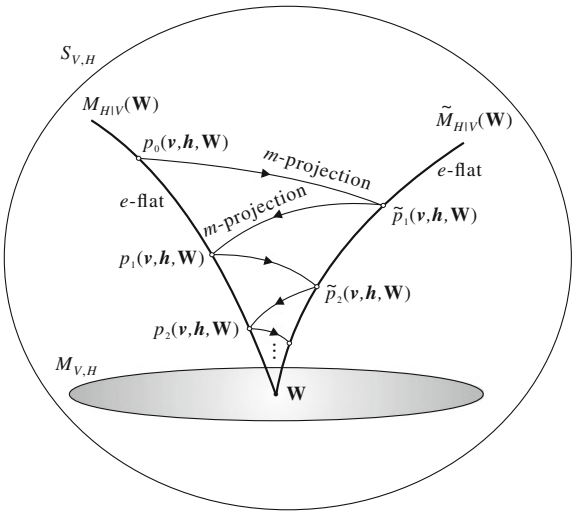


Fig. 11.18 CDR distribution $P_k(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$

connecting $r_1(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v})$ and $r_2(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v})$,

$$t\left\{\log r_1(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v})\right\} + (1-t)\left\{\log r_2(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v})\right\} = \left\{\log r_1(\boldsymbol{v})^t r_2(\boldsymbol{v})^{1-t} p(\boldsymbol{h}|\boldsymbol{v})\right\}, \tag{11.217}$$

is included in $M_{H|V}$, where we have omitted the normalization factor $c(t)$. The same situation holds for $\tilde{M}_{V|H}$. See Fig. 11.18.

The initial distribution is given by putting $p_0(\boldsymbol{v}) = q(\boldsymbol{v})$ as

$$p_0(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W}) = p_0(\boldsymbol{v})p\left(\boldsymbol{h}|\boldsymbol{v}, \mathbf{W}\right). \tag{11.218}$$

Then, the sequence of CD_k distributions is given by the geometrical procedures in the following theorem, due to R. Karakida.

Theorem 11.11 *$\tilde{p}_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ is the m-projection of $p_{j-1}(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ to $\tilde{M}_{\mathbf{H}|\mathbf{V}}(\mathbf{W})$ and $p_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ is the m-projection of $\tilde{p}_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ to $M_{\mathbf{V}|\mathbf{H}}(\mathbf{W})$.*

Proof Given $\tilde{p}_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$, its m-projection to $M_{\mathbf{H}|\mathbf{V}}$ is given by the minimizer of

$$D_{KL}\left[\tilde{p}_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W}) : r(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v})\right] = -\int \tilde{p}_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W}) \log r(\boldsymbol{v}) d\boldsymbol{v} d\boldsymbol{h} + c \tag{11.219}$$

with respect to $r(\boldsymbol{v})$, where c is a term not depending on $r(\boldsymbol{v})$. By adding the constraint

$$\int r(\boldsymbol{v}) d\boldsymbol{v} = 1, \tag{11.220}$$

the variation of D_{KL} gives

$$r(\boldsymbol{v}) = p_{Vj}(\boldsymbol{v}, \mathbf{W}). \tag{11.221}$$

The other case is proved similarly. □

The theorem shows that $p_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ converges to $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ as j increases. Hence, $p_j(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ may be used as an approximation of $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ in calculations of $\langle \boldsymbol{h}\boldsymbol{v}^T \rangle_p$.

The following is an interesting observation based on the Pythagorean theorem.

Theorem 11.12 *The KL-divergence from* $p_0(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ *to* $p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W})$ *is decomposed as*

$$D_{KL}[p_0 : p] = \sum_{j=0} D_{KL}\left[p_j : \tilde{p}_{j+1}\right] + \sum_{j=1} D_{KL}\left[\tilde{p}_j : p_j\right]. \tag{11.222}$$

Proof Since $\tilde{p}_j p_j p$ is an orthogonal triangle in which the m-geodesic $\tilde{p}_j p_j$ is orthogonal to the e-geodesic $p_j p$, we can apply the Pythagorean theorem to decompose $D_{KL}\left[\tilde{p}_j : p\right]$ (Fig. 11.17). Similar decomposition holds for $D_{KL}\left[p_j : p\right]$. Hence, repeating the decomposition recursively, we have the theorem. □

11.5.5 Gaussian RBM

We may consider an analog RBM in which both $\boldsymbol{v}$ and $\boldsymbol{h}$ take analog values. A typical one is a Gaussian RBM in which both $\boldsymbol{v}$ and $\boldsymbol{h}$ are Gaussian random variables. The stationary distribution is written as

$$p(\boldsymbol{v}, \boldsymbol{h}, \mathbf{W}) = \exp\left\{-\frac{1}{2\sigma_v^2}|\boldsymbol{v}|^2 - \frac{1}{2\sigma_h^2}|\boldsymbol{h}|^2 + \frac{\boldsymbol{h}^T \mathbf{W} \boldsymbol{v}}{\sigma_v \sigma_h} - \psi\right\}. \tag{11.223}$$

Here, the quadratic terms of $\boldsymbol{v}$ and $\boldsymbol{h}$ exist but they do not include cross terms such as $v_i v_j$ $(i \neq j)$, so that there are no mutual connections among the neurons in each layer.

The Gaussian RBM is simple and hence tractable, because all related distributions are described in the framework of Gaussian distributions. The conditional distributions are Gaussian given by

$$p(\boldsymbol{h}|\boldsymbol{v}, \mathbf{W}) = c \exp\left\{-\frac{1}{2\sigma_n^2}|\boldsymbol{h} - \frac{\sigma_h}{\sigma_v}\mathbf{W}\boldsymbol{v}|^2\right\}, \tag{11.224}$$

$$p(\boldsymbol{v}|\boldsymbol{h}, \mathbf{W}) = c' \exp\left\{-\frac{1}{2\sigma_v^2}\left|\boldsymbol{v} - \frac{\sigma_v}{\sigma_h}\mathbf{W}^T\boldsymbol{h}\right|^2\right\}, \tag{11.225}$$

and the marginal distribution is also Gaussian,

$$p_V(\boldsymbol{v}, \mathbf{W}) = c'' \exp\left\{-\frac{1}{2\sigma_v^2}\boldsymbol{v}^T\left(I - \mathbf{W}^T\mathbf{W}\right)\boldsymbol{v}\right\}, \tag{11.226}$$

where c, c' and c'' are adequate constants.

Karakida et al. (2014) analyzed the behavior of the Gaussian RBM when the distribution $q(\boldsymbol{v})$ of $\boldsymbol{v}$ given from the outside is mean 0 and its covariance matrix is $\mathbf{C}$. Since

$$\langle \boldsymbol{h}\boldsymbol{v}^T\rangle_q = \frac{1}{\sigma_v^2}\mathbf{W}\mathbf{C}, \tag{11.227}$$

$$\langle \boldsymbol{h}\boldsymbol{v}^T\rangle_p = \mathbf{W}\left(\mathbf{I} - \mathbf{W}^T\mathbf{W}\right)^{-1} \tag{11.228}$$

hold, the equation of learning (11.206) is written as

$$\varepsilon\frac{d\mathbf{W}}{dt} = \frac{1}{\sigma_v^2}\mathbf{W}\mathbf{C} - \left(\mathbf{I} - \mathbf{W}^T\mathbf{W}\right)^{-1}, \tag{11.229}$$

where we use continuous time. They also calculated the equation of learning for CD_k, obtaining

$$\varepsilon\frac{d\mathbf{W}}{dt} = \frac{1}{\sigma_v^2}\mathbf{W}\mathbf{C} - \left\{\frac{1}{\sigma_v^2}\mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^k\mathbf{C}\left(\mathbf{W}^T\mathbf{W}\right)^k + \sum_{i=0}^{2k-1}\left(\mathbf{W}^T\mathbf{W}\right)^i\right\}. \tag{11.230}$$

We can easily see that (11.230) converges to (11.229) as k tends to infinity.

We study the equilibrium solutions and their stability for the above equations. The following theorem shows that a Gaussian RBM performs a PCA-like analysis. To this end, let $\lambda_1, \ldots, \lambda_n$ be n eigenvalues of $\mathbf{C}$ (where we assume that they are all distinct) and let $\mathbf{O}$ be the orthogonal matrix that diagonalizes $\mathbf{C}$,

$$\mathbf{C} = \mathbf{O}^T\Lambda\mathbf{O}. \tag{11.231}$$

Theorem 11.13 *Assume that there are r eigenvalues which are larger than σ_v^2. Then, the equilibrium solutions of (11.229) and (11.230) are the same, given by*

$$\mathbf{W} = \mathbf{U}\tilde{\Lambda}\mathbf{O}, \tag{11.232}$$

where $\mathbf{U}$ is an arbitrary $m \times m$ orthogonal matrix and

$$\tilde{\Lambda} = diag\left(\sqrt{1 - \frac{1}{\lambda}}, \sqrt{1 - \frac{1}{\lambda_2}}, \ldots \sqrt{1 - \frac{1}{\lambda_r}}, 0, \ldots, 0\right). \tag{11.233}$$

The proof is technical and is omitted (see Karakida et al. 2014). The stability of solutions is also analyzed.

By choosing the coordinate axes of $\boldsymbol{v}$ adequately, we see that the marginal distribution of RBM is given:

$$p_V(\boldsymbol{v}, \mathbf{W}) = c \exp\left\{ -\sum_{i=1}^{r} \frac{v_i^2}{2\lambda_i} - \sum_{i=r+1}^{n} \frac{v_i^2}{2\sigma_v^2} \right\}. \tag{11.234}$$

This shows that the Gaussian RBM performs the PCA analysis, neglecting smaller eigenvalues. It is also shown that the CD_1 learning method has a sufficiently good performance compared to the original RBM learning method (maximum likelihood method).

Remarks

We have glanced at topics of machine learning from the information geometry point of view. Since stochastic uncertainty is involved in the real world, it is expected that information geometry will provide good ideas, useful suggestions and clear understanding of aspects of machine learning. Clustering techniques are the main tools of information retrieval, where divergence functions are used. They are connected with information geometry. We have demonstrated that robust clustering is achieved by tBD. This field is developing quickly. See Nock et al. (2015).

Support vector machines are useful tools in pattern recognition and regression. We have avoided following the main stream of the kernel method and instead touched upon how the performance of a kernel is improved by a conformal transformation. This might give a hint for a good choice of kernels.

Stochastic reasoning is an important procedure, where belief propagation (BP) plays a key role. We can reformulate the BP algorithm by using information geometry. This gives a more transparent understanding of the algorithm than the conventional one. Moreover, it provides an efficient algorithm of stochastic inference, which is a new version of the convex–concave computational procedure (CCCP). The boosting of weak learners is also outlined.

Deep learning is a hot topic, for which we still lack convincing theories. We have proposed a way to understand it from information geometry of Bayesian statistics. The restricted Boltzmann machine (RBM) is understood in the framework of Bayesian information geometry. Karakida et al. (2014; 2016) studied the performance of the Gaussian–Bernoulli RBM and showed that it performs ICA in restricted situations. However, this still remains as a half-baked idea, emerging in the last stage of completing this monograph. The geometry of contrast divergences is mostly due to on-going research by R. Karakida (PhD student at the University of Tokyo) and it might be too early to be included here. In order to understand deep learning, we

need to construct a good model of $q(\boldsymbol{v})$ which involves hierarchical structure. Hierarchies of hidden layers unveil their hidden structure one layer at a time. This is unsupervised learning. The supervised aspect of deep learning is related to singularities existing ubiquitously in a neuromanifold, and will be one of the main topics of the next chapter.

Chapter 12
Natural Gradient Learning and Its Dynamics in Singular Regions

Learning takes place in a parameter space, which is not Euclidean in general but Riemannian. Therefore, we need to take the Riemannian structure into account when designing a learning method. The natural gradient method, which is a version of stochastic descent learning, is proposed for this purpose, using the Riemannian gradient. It is a Fisher efficient on-line method of estimation. Its performance is excellent in general and it has been used in various types of learning problems such as neural learning, policy gradient in reinforcement learning, optimization by means of stochastic relaxation, independent component analysis, Monte Carlo Markov Chain (MCMC) in a Riemannian manifold and others.

Some statistical models are singular, implying that its parameter space includes singular regions. The multilayer perceptron (MLP) is a typical singular model. Since supervised learning of MLP is involved in deep learning, it is important to study the dynamical behavior of learning in singular regions, in which learning is very slow. This is known as plateau phenomena. The natural gradient method overcomes this difficulty.

12.1 Natural Gradient Stochastic Descent Learning

12.1.1 On-Line Learning and Batch Learning

Huge amounts of data exist in the real world. Consider a set of data which are generated randomly subject to a fixed but unknown probability distribution. A typical example is shown in the regression problem, where input signal $\boldsymbol{x}$ is generated randomly, accompanied by a desired response $f(\boldsymbol{x})$. A teacher signal y, which is a noisy version of the desired output $f(\boldsymbol{x})$,

$$y = f(\boldsymbol{x}) + \varepsilon, \tag{12.1}$$

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_12

is given together with $\boldsymbol{x}$, where ε is random noise. The task of a learning machine is, in this case, to estimate the desired output mapping $f(\boldsymbol{x})$ by using the available examples of input–output pairs $D = \{(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, T\}$, called training examples. They are subject to an unknown joint probability distribution,

$$p(\boldsymbol{x}, y) = q(\boldsymbol{x})\text{Prob}\{y|\boldsymbol{x}\} = q(\boldsymbol{x})p_{\varepsilon}\{y - f(\boldsymbol{x})\}, \tag{12.2}$$

where $q(\boldsymbol{x})$ is the probability distribution of $\boldsymbol{x}$ and $p_{\varepsilon}(\varepsilon)$ is the probability distribution of noise ε, typically Gaussian. This is a usual scheme of supervised learning.

We use a parameterized family $f(\boldsymbol{x}, \boldsymbol{\xi})$ of functions as candidates for the desired output, where $\boldsymbol{\xi}$ is a vector parameter. The set of $\boldsymbol{\xi}$ is a parameter space and we search for the optimal $\hat{\boldsymbol{\xi}}$ that approximates the true $f(\boldsymbol{x})$ by using training examples D. When y takes an analog value, this is a regression problem. When y is discrete, say binary, this is pattern recognition.

In order to evaluate the performance of machine $f(\boldsymbol{x}, \boldsymbol{\xi})$, we define a loss function or cost function. The instantaneous loss of processing $\boldsymbol{x}$ by machine $f(\boldsymbol{x}, \boldsymbol{\xi})$ is typically given by

$$l(\boldsymbol{x}, y; \boldsymbol{\xi}) = \frac{1}{2}\{y - f(\boldsymbol{x}, \boldsymbol{\xi})\}^2, \tag{12.3}$$

in the case of regression, which is a half of the square of the difference between the teacher output y and machine output $f(\boldsymbol{x}, \boldsymbol{\xi})$.

The loss function of machine $\boldsymbol{\xi}$ is the expectation of the instantaneous loss over all possible pairs $(\boldsymbol{x}, y)$,

$$L(\boldsymbol{\xi}) = \text{E}_p[l(\boldsymbol{x}, y; \boldsymbol{\xi})], \tag{12.4}$$

where the expectation is taken with respect to the unknown joint probability distribution $p(\boldsymbol{x}, y)$. However, since we do not know $p(\boldsymbol{x}, y)$, we use the average over the training data,

$$L_{\text{train}}(\boldsymbol{\xi}) = \frac{1}{T}\sum_{t=1}^{T} l(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}). \tag{12.5}$$

This is called the training error, since the average loss is evaluated by using the data that we used for training. In contrast, (12.4) is called the generalization error, since it evaluates the performance over all possible data $(\boldsymbol{x}, y)$ not used in the process of training. Since we do not know L, we minimize the training error L_{train} to obtain $\hat{\boldsymbol{\xi}}$. A regularization term may be added to L_{train} in order to obtain a regularized optimal solution $\hat{\boldsymbol{\xi}}$ by learning.

A loss function is defined similarly in the case of pattern recognition by the expectation of an instantaneous loss. Even in the case of binary y, $y = 0$ or 1, we can use (12.3) as a loss. However, it is more natural to formulate the problem in terms of logistic regression such that the probability of y is given as a function of $\boldsymbol{\xi} \cdot \boldsymbol{x}$ by

$$\text{Prob}\{y|\ \boldsymbol{\xi} \cdot \boldsymbol{x}\} = \exp\{y\boldsymbol{\xi} \cdot \boldsymbol{x} - \psi(\boldsymbol{\xi} \cdot \boldsymbol{x})\}, \tag{12.6}$$

where the normalization factor ψ is

$$\psi(\boldsymbol{\xi} \cdot \boldsymbol{x}) = \log \{1 + \exp(\boldsymbol{\xi} \cdot \boldsymbol{x})\}. \tag{12.7}$$

This implies

$$\text{Prob}\{y = 1 \mid \boldsymbol{x}; \boldsymbol{\xi}\} = \frac{\exp(\boldsymbol{\xi} \cdot \boldsymbol{x})}{1 + \exp(\boldsymbol{\xi} \cdot \boldsymbol{x})}. \tag{12.8}$$

The instantaneous loss function is the negative of log Prob $\{y \mid \boldsymbol{\xi} \cdot \boldsymbol{x}\}$,

$$l(\boldsymbol{x}, y; \boldsymbol{\xi}) = -y\boldsymbol{x} \cdot \boldsymbol{\xi} + \psi(\boldsymbol{\xi} \cdot \boldsymbol{x}). \tag{12.9}$$

In the problem of estimation of parameters $\boldsymbol{\xi}$ in a statistical model $\{p(\boldsymbol{x}, \boldsymbol{\xi})\}$, we use

$$l(\boldsymbol{x}; \boldsymbol{\xi}) = -\log p(\boldsymbol{x}, \boldsymbol{\xi}), \tag{12.10}$$

the negative of log likelihood, where only $\boldsymbol{x}$'s are observed. The generalization error is

$$L(\boldsymbol{\xi}) = -\text{E}_{\xi_0}\left[\log p(\boldsymbol{x}, \boldsymbol{\xi})\right]. \tag{12.11}$$

where $\boldsymbol{\xi}_0$ is the true parameter, such that $\boldsymbol{x}$ is generated from $p\left(\boldsymbol{x}, \boldsymbol{\xi}_0\right)$. The regression problem is regarded as an estimation problem to estimate $\boldsymbol{\xi}$ of $p(\boldsymbol{x}, y; \boldsymbol{\xi})$, where random variables are $(\boldsymbol{x}, y)$ and we do not care about $q(\boldsymbol{x})$.

An on-line learning procedure modifies the current candidate $\boldsymbol{\xi}_t$ at time t to obtain $\boldsymbol{\xi}_{t+1}$ at the next time based on the current training example $(\boldsymbol{x}_t, y_t)$ so as to decrease the instantaneous loss (Rumelhart et al. 1986). Usually, the negative of the gradient is used to update $\boldsymbol{\xi}_t$,

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_t\right), \tag{12.12}$$

where ∇ is the gradient with respect to $\boldsymbol{\xi}$ and coefficient η_t is called a learning constant, which may depend on t. Since training data are given one by one, the change

$$\Delta\boldsymbol{\xi}_t = -\eta_t \nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_t\right) \tag{12.13}$$

is a random variable depending on $(\boldsymbol{x}_t, y_t)$. The expectation of ∇l is equal to $\nabla L(\boldsymbol{\xi})$. Therefore, the change $\Delta\boldsymbol{\xi}_t$ is random but its expectation is in the direction of $-\nabla L\left(\boldsymbol{\xi}_t\right)$. See Fig. 12.1. Hence, (12.12) is called a stochastic descent learning method. Amari (1967) might be the first to have used this idea for training a multilayer perceptron. The method is now well established as the back-propagation learning method.

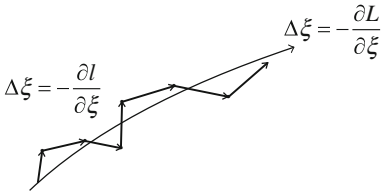


Fig. 12.1 Gradient descent of expected loss L and stochastic gradient descent of l

A batch learning procedure is an iterative method which uses all the training data for modifying $\boldsymbol{\xi}_t$ at one step, such that $\boldsymbol{\xi}_t$ is modified to $\boldsymbol{\xi}_{t+1}$ by

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \frac{1}{T} \sum_{i=1}^{T} l\left(\boldsymbol{x}_i, y_i; \boldsymbol{\xi}_t\right). \tag{12.14}$$

The two types of learning, batch and on-line, have different merits and demerits.

12.1.2 Natural Gradient: Steepest Descent Direction in Riemannian Manifold

Given a function $L(\boldsymbol{\xi})$ in a manifold, it is widely believed that the gradient

$$\nabla L(\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} L(\boldsymbol{\xi}) \tag{12.15}$$

is the direction of the steepest change of $L(\boldsymbol{\xi})$. In a geographical map with contour lines, the steepest direction is given by the gradient of the height function $H(\boldsymbol{\xi})$, that is $\nabla H(\boldsymbol{\xi})$, which is orthogonal to contour lines. However, this is true only when an orthonormal coordinate system is used in a Euclidean space.

In a Riemannian manifold, the square of local distance between two nearby points $\boldsymbol{\xi}$ and $\boldsymbol{\xi} + d\boldsymbol{\xi}$ is given by the quadratic form

$$ds^2 = g_{ij} d\xi^i d\xi^j, \tag{12.16}$$

where $\mathbf{G} = \left(g_{ij}\right)$ is a Riemannian metric tensor. Note that we use the Einstein convention so that the summation symbol $\sum$ is omitted in (12.16). Let us change the current point $\boldsymbol{\xi}$ to $\boldsymbol{\xi} + d\boldsymbol{\xi}$, and see how the value of $L(\boldsymbol{\xi})$ changes, depending on the direction $d\boldsymbol{\xi}$. We search for the direction in which L changes most rapidly. In order to make a fair comparison, the step-size of $d\boldsymbol{\xi}$ should have the same magnitude in all directions, so that the length of $d\boldsymbol{\xi}$ should be the same,

$$g_{ij}(\boldsymbol{\xi}) d\xi^i d\xi^j = \varepsilon^2, \tag{12.17}$$

where ε is a small constant. We put $d\boldsymbol{\xi} = \varepsilon \boldsymbol{a}$ and require that

$$|\boldsymbol{a}|^2 = g_{ij} a^i a^j = 1. \tag{12.18}$$

Then, the steepest direction of L is the maximizer of

$$L(\boldsymbol{\xi} + d\boldsymbol{\xi}) - L(\boldsymbol{\xi}) = \varepsilon \nabla L(\boldsymbol{\xi}) \cdot \boldsymbol{a} \tag{12.19}$$

under the constraint (12.18). See Fig. 12.2. By using the variational method of maximizing (12.19) under the constraint (12.18), we easily obtain the following formulation:

$$\underset{\boldsymbol{a}}{\text{maximize}}\ \nabla L(\boldsymbol{\xi}) \cdot \boldsymbol{a} - \lambda g_{ij} a^i a^j. \tag{12.20}$$

This is a quadratic problem and the steepest direction is obtained as

$$\boldsymbol{a} \propto \mathbf{G}^{-1} \nabla L(\boldsymbol{\xi}). \tag{12.21}$$

We call

$$\tilde{\nabla} L(\boldsymbol{\xi}) = \mathbf{G}^{-1}(\boldsymbol{\xi}) \nabla L(\boldsymbol{\xi}) \tag{12.22}$$

the Riemannian gradient or natural gradient of L, where

$$\tilde{\nabla} = G^{-1} \nabla \tag{12.23}$$

is the natural gradient operator.

From the point of view of geometry, the natural gradient is a contravariant vector

$$A^i = g^{ij}(\xi) \partial_j L, \tag{12.24}$$

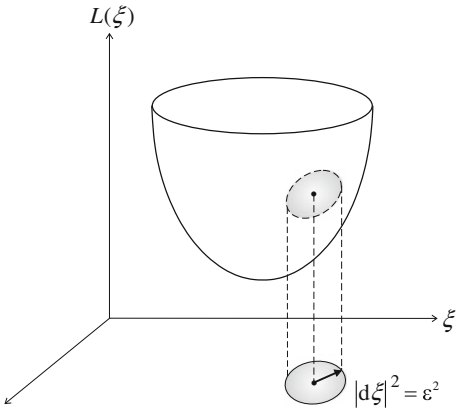


Fig. 12.2 Natural gradient $\tilde{\nabla} L$ of L

and the ordinary gradient is a covariant vector

$$A_j = \partial_j L(\boldsymbol{\xi}) \tag{12.25}$$

in the index notation. They are equal when and only when

$$g_{ij}(\boldsymbol{\xi}) = \delta_{ij}, \tag{12.26}$$

that is, when an orthonormal coordinate system is used in a Euclidean space.

The natural gradient learning method, which was suggested in Amari (1967), was formally introduced in Amari (1998) and defined by

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \tilde{\nabla} l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_t\right). \tag{12.27}$$

In the batch mode, it is

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \frac{1}{T} \sum_{i=1}^{T} \tilde{\nabla} l\left(\boldsymbol{x}_i, y_i, \boldsymbol{\xi}_t\right). \tag{12.28}$$

In the case of statistical estimation where the Fisher information is a Riemannian metric, the loss function L and the Riemannian metric $\mathbf{G}$ is defined by using the same log likelihood function $\log p(\boldsymbol{x}, \boldsymbol{\xi})$. In this case, the natural gradient method is regarded as a version of the Gauss–Newton method. However, there are many other cases where the loss function and the Riemannian metric are not related. The natural gradient learning method is useful in such cases, too. Independent component analysis (ICA) is such an example, where the parameter space is a set of mixing matrices and the Riemannian metric is given by the invariant metric of the underlying Lie group, but the loss is measured by the degree of independence of unmixed signals. In the next subsection, we show an interesting new idea of natural gradient using the "absolute value" of the Hessian as a Riemannian metric (Daupin et al. 2014).

The natural gradient is also used in deep learning (Roux et al. 2007; Ollivier 2015) and in reinforcement learning as a policy natural gradient (e.g., Kakade 2002; Peters and Schaal 2008; Morimura et al. 2009). Another application is found in the optimization problem with stochastic relaxation technique (Malagò and Pistone 2014; Malagò et al. 2013; Yi et al. 2009; see also Hansen and Ostermeier 2001).

12.1.3 Riemannian Metric, Hessian and Absolute Hessian

The Newton method uses the Hessian of $L(\boldsymbol{\xi})$ for obtaining the minimizer of $L(\boldsymbol{\xi})$ by solving $\nabla L(\boldsymbol{\xi}) = 0$ recursively. It updates the current $\boldsymbol{\xi}_t$ to give

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \mathbf{H}^{-1}\left(\boldsymbol{\xi}_t\right) \nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_t\right), \tag{12.29}$$

where

$$\mathbf{H}(\boldsymbol{\xi}) = \nabla\nabla L(\boldsymbol{\xi}). \tag{12.30}$$

The natural gradient replaces $\mathbf{H}$ by the Riemannian metric $\mathbf{G}$. Therefore, it is interesting to see the relation between $\mathbf{G}$ and $\mathbf{H}$.

We study the case where the noise is Gaussian with mean 0 and variance σ^2. The joint probability distribution is written as

$$p(\boldsymbol{x}, y; \boldsymbol{\xi}) = \frac{q(\boldsymbol{x})}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\{y - f(\boldsymbol{x}, \boldsymbol{\xi})\}^2\right]. \tag{12.31}$$

Hence, the loss function is the same as the negative of the log likelihood except for the constant. Minimizing $L(\boldsymbol{\xi})$ is equivalent to maximizing the likelihood of the unknown parameter $\boldsymbol{\xi}$. The on-line learning algorithm (12.27) is regarded as a sequential estimation procedure, and the batch learning algorithm is an iteration procedure of obtaining the maximum likelihood estimator.

The Fisher information in this case is given by

$$\mathbf{G}(\boldsymbol{\xi}) = \nabla\nabla L(\boldsymbol{\xi}) = \mathrm{E}_{p(\boldsymbol{x},y,\boldsymbol{\xi})}[\nabla\nabla l(\boldsymbol{x}, y, \boldsymbol{\xi})]. \tag{12.32}$$

On the other hand, the Hessian of the loss function $L(\boldsymbol{\xi})$ is

$$\mathbf{H}(\boldsymbol{\xi}) = \nabla\nabla L(\boldsymbol{\xi}) = \mathrm{E}_{p(\boldsymbol{x},y,\boldsymbol{\xi}_0)}\left[\nabla\nabla l(\boldsymbol{x}, y, \boldsymbol{\xi})\right], \tag{12.33}$$

where the expectation is taken with respect to the true distribution $p\left(\boldsymbol{x}, y, \boldsymbol{\xi}_0\right)$ from which teacher signal y is generated.

By using (12.3) or by assuming $\sigma^2 = 1$ in (12.31), we easily have

$$\mathbf{G}(\boldsymbol{\xi}) = \mathrm{E}_{\boldsymbol{x}}\left[\nabla f(\boldsymbol{x}, \boldsymbol{\xi}) \nabla f(\boldsymbol{x}, \boldsymbol{\xi})^T\right] \tag{12.34}$$

$$\mathbf{H}(\boldsymbol{\xi}) = \mathbf{G}(\boldsymbol{\xi}) - \mathrm{E}_{\boldsymbol{x}}\left[\left\{f\left(\boldsymbol{x}, \boldsymbol{\xi}_0\right) - f(\boldsymbol{x}, \boldsymbol{\xi})\right\} \nabla\nabla f(\boldsymbol{x}, \boldsymbol{\xi})\right], \tag{12.35}$$

where $\mathrm{E}_{\boldsymbol{x}}$ is the expectation with respect to $q(\boldsymbol{x})$. $\mathbf{G}$ is in general positive-definite, but $\mathbf{H}$ is not necessarily so. (We discuss the singular case later where $\mathbf{G}$ and $\mathbf{H}$ degenerate.) However, $\mathbf{H}$ and $\mathbf{G}$ are exactly equal at $\boldsymbol{\xi} = \boldsymbol{\xi}_0$. Moreover, they are equal when $f(\boldsymbol{x}, \boldsymbol{\xi}) = f\left(\boldsymbol{x}, \boldsymbol{\xi}_0\right)$ holds. We show later that they are equal at critical or singular regions in MLP.

Recently, an interesting new idea of defining a Riemannian metric by the "absolute value" of the Hessian matrix was proposed (Dauphin et al. 2014). The Hessian is decomposed as

$$\mathbf{H} = \mathbf{O}^T \Lambda \mathbf{O}, \tag{12.36}$$

where $\mathbf{O}$ is an orthogonal matrix and $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix having eigenvalues of $\mathbf{H}$ as the diagonal elements. The matrix of the absolute value of $\mathbf{H}$ is defined by

$$|\mathbf{H}| = \mathbf{O}^T \operatorname{diag}\left(|\lambda_1|, \ldots, |\lambda_n|\right) \mathbf{O}. \tag{12.37}$$

When $|\mathbf{H}|$ is used as a Riemannian metric, the natural gradient method becomes

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \left|\mathbf{H}\left(\boldsymbol{\xi}_t\right)\right|^{-1} \nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_t\right). \tag{12.38}$$

The method is called the saddle-free Newton method (SFN) and its good performance is demonstrated. When $\boldsymbol{\xi}'$ is a saddle point, the Newton method stabilizes the saddle and converges to it. Hence, the Newton method does not work well. It is shown that most critical points of L are saddles in high dimensions (Dauphin et al. 2014). Hence, the new idea is introduced as a method of avoiding saddle points, but keeping the good performance of the Newton method. Any natural gradient method is not trapped in a saddle whereas the Newton method is. Moreover, the behaviors of the Fisher information-based natural gradient and the absolute-value-based Hessian natural gradient are the same at around the optimal point $\boldsymbol{\xi}_0$, both enjoying the Fisher efficiency. It is also interesting to see that their behaviors are the same in the critical or singular regions studied later, which are the main source of plateau phenomena (retardation of learning).

12.1.4 Stochastic Relaxation of Optimization Problem

We show a problem in which the natural gradient plays an important role. Let us consider the problem of searching for the minimizer of $f(\boldsymbol{x})$ over $\boldsymbol{x} \in X$. The problem is difficult to solve when f is not convex, in particular when $\boldsymbol{x}$ is discrete. The integer programming is a typical example of the discrete type.

Let us introduce a family of probability distribution $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$ and consider the expectation

$$L(\boldsymbol{\xi}) = E_{p(\boldsymbol{x},\boldsymbol{\xi})}[f(\boldsymbol{x})]. \tag{12.39}$$

The problem of searching for the minimizer of $L(\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ is called the stochastic relaxation of the original problem (Malagò and Pistone 2014; see also Hansen and Ostermeier 2001). It changes the problem of a search in X to a search in M, so the gradient descent method is applicable even when X is discrete. Since M is a Riemannian manifold, we can apply the natural gradient method,

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \mathbf{G}\left(\boldsymbol{\xi}_t\right)^{-1} \nabla L\left(\boldsymbol{\xi}_t\right). \tag{12.40}$$

By choosing model M carefully, it works well. Yi et al. (2009) proposed an efficient way of implementing the natural gradient.

12.1.5 Natural Policy Gradient in Reinforcement Learning

We summarize the natural gradient method in reinforcement learning, following Peters and Schaal (2008). It is called the natural policy gradient method, formulated in the framework of the Markov decision process. See a survey paper by Grondman et al. (2012). Let us consider a system having state space $X = \{\boldsymbol{x}\}$ and action space $U = \{\boldsymbol{u}\}$. At each discrete time t, an action is chosen, depending on the current state $\boldsymbol{x}_t$, subject to policy $\pi(\boldsymbol{u}|\boldsymbol{x}_t)$, which specifies the probability (density) of action $\boldsymbol{u}_t$. We assume that it is a parameterized family of conditional probabilities specified by a vector parameter $\boldsymbol{\theta}$, denoted as $\pi(\boldsymbol{u}|\boldsymbol{x};\boldsymbol{\theta})$. The state transition takes place stochastically depending on the current $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$, and its probability (density) function is given by $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{u}_t)$. While a state transition takes place, an instantaneous reward is derived, which is a function of the current $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$, written as $r = r(\boldsymbol{x}_t, \boldsymbol{u}_t)$. See Fig. 12.3.

The expected reward at time t is a sum of the current reward r_t and future rewards $r_{t+1}, r_{t+2}, \ldots$, but future rewards are discounted. Hence, the expected reward at state $\boldsymbol{x}$, including future rewards, is written as

$$V^{\pi}(\boldsymbol{x}) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \boldsymbol{x}_0 = \boldsymbol{x}\right], \tag{12.41}$$

where $\gamma < 1$ is a discount factor. It depends on policy π or its parameter $\boldsymbol{\theta}$. This is called the state-value function. We also define

$$Q^{\pi}(\boldsymbol{x}, \boldsymbol{u}) = \mathrm{E}\left[\sum \gamma^t r_t \mid \boldsymbol{x}, \boldsymbol{u}\right], \tag{12.42}$$

which is the expected reward when the state is at $\boldsymbol{x}$ and action $\boldsymbol{u}$ is chosen. The expectation is taken throughout all the possible trajectories of $(\boldsymbol{x}_t, \boldsymbol{u}_t)$ pairs.

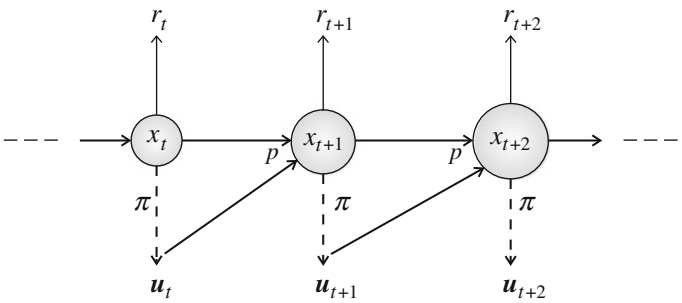


Fig. 12.3 Markov decision process, reward and action

Let us fix an initial state $\boldsymbol{x} = \boldsymbol{x}_0$. The expected reward by taking the policy $\pi(\boldsymbol{u}\,|\boldsymbol{x};\boldsymbol{\theta})$ is

$$J(\boldsymbol{\theta}) = \mathrm{E}\left[\sum \gamma^t r_t |\ \boldsymbol{\theta}\right], \tag{12.43}$$

which is rewritten as

$$J(\boldsymbol{\theta}) = \mathrm{E}\left[d^{\pi}(\boldsymbol{x}) \int \pi(\boldsymbol{u}|\boldsymbol{x};\boldsymbol{\theta}) r(\boldsymbol{x},\boldsymbol{u}) d\boldsymbol{x} d\boldsymbol{u}\right], \tag{12.44}$$

where

$$d^{\pi}(\boldsymbol{x}) = \sum_t \gamma^t p\left(\boldsymbol{x}_t\right) \delta\left(\boldsymbol{x} - \boldsymbol{x}_t\right) \tag{12.45}$$

is the discounted probability of a sequence of states.

We define the Fisher information matrix at the current state $\boldsymbol{x}$ by

$$\mathbf{F}(\boldsymbol{\theta}|\boldsymbol{x}) = \int \pi(\boldsymbol{u}|\boldsymbol{x}) \nabla_{\theta} \log \pi(\boldsymbol{u}|\boldsymbol{x}) \left\{\nabla_{\theta} \log \pi(\boldsymbol{u}|\boldsymbol{x})\right\}^T d\boldsymbol{u}. \tag{12.46}$$

The entire Fisher information matrix is its expectation along all the trajectories,

$$\mathbf{G}(\boldsymbol{\theta}) = \int d^{\pi}(\boldsymbol{x}) \mathbf{F}(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{x}. \tag{12.47}$$

See Kakade (2001), Peters and Schaal (2008).

The natural gradient method, called the natural policy gradient or natural actor-critic, is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta \mathbf{G}^{-1}\left(\boldsymbol{\theta}_t\right) \nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\theta}_t\right). \tag{12.48}$$

However, this is computationally heavy. A good idea is to approximate the state-action value function by a linear combination of adequate basis functions $\{a_i(\boldsymbol{x},\boldsymbol{u})\}$ as

$$Q^{\pi}(\boldsymbol{x},\boldsymbol{u}) = \sum a_i(\boldsymbol{x},\boldsymbol{u}) w_i = \boldsymbol{a}(\boldsymbol{x},\boldsymbol{u}) \cdot \boldsymbol{w}, \tag{12.49}$$

where $\boldsymbol{w}$ is the parameters of weight to be adjusted. We choose

$$\boldsymbol{a}(\boldsymbol{x},\boldsymbol{u}) = \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{u}|\boldsymbol{x};\boldsymbol{\theta}) \tag{12.50}$$

as basis functions. Since the gradient of the expected reward is written as

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \int d^{\pi}(\boldsymbol{x}) \int \nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{\theta}) Q^{\pi}(\boldsymbol{x},\boldsymbol{u}) d\boldsymbol{u} d\boldsymbol{x}, \tag{12.51}$$

its gradient becomes

$$\nabla_{\boldsymbol{\theta}} J = \mathbf{G}\boldsymbol{w}. \tag{12.52}$$

Therefore, the natural gradient takes a very simple form as

$$\tilde{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbf{G}^{-1} \nabla_{\boldsymbol{\theta}} J = \boldsymbol{w}. \tag{12.53}$$

In order to implement the natural policy gradient, we need to evaluate $\boldsymbol{w}$ which gives the best approximation of Q. We use the TD error

$$\delta_t = r_t + \gamma V^{\pi}(\boldsymbol{x}_{t+1}) - V^{\pi}(\boldsymbol{x}_t) \tag{12.54}$$

and solve the linear regression problem recursively as

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha \delta_t \boldsymbol{a}(\boldsymbol{x}_t), \tag{12.55}$$

where the basis function $\boldsymbol{a}(\boldsymbol{x})$ is

$$\boldsymbol{a}(\boldsymbol{x}) = \int \pi(\boldsymbol{u}|\boldsymbol{x}) \boldsymbol{a}(\boldsymbol{x}, \boldsymbol{u}) d\boldsymbol{u}. \tag{12.56}$$

It is reported that the natural policy gradient demonstrates excellent performance in many cases.

12.1.6 Mirror Descent and Natural Gradient

The mirror descent method was introduced by Nemirovski and Yudin (1983) (see also Beck and Teboulle 2003) as a tool to search for the minimum of a convex function $f(\boldsymbol{\theta})$. It is used in convex optimization problems with a constrained region. It uses another convex function $\psi(\boldsymbol{\theta})$ together with its Legendre dual $\varphi(\boldsymbol{\eta})$. They implicitly use a dually flat structure together with a Riemannian metric

$$\mathbf{G}(\boldsymbol{\theta}) = \nabla \nabla \psi(\boldsymbol{\theta}). \tag{12.57}$$

The dual coordinates

$$\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta}) \tag{12.58}$$

are used to update the current $\boldsymbol{\eta}_t$ as

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t - \varepsilon \nabla f(\boldsymbol{\theta}_t), \tag{12.59}$$

where ε is a learning rate. Since both $\boldsymbol{\eta}$ and ∇f are covariant quantities, it is invariant. The result is transformed back to the primal coordinates by

$$\boldsymbol{\theta}_{t+1} = \nabla \varphi\left(\boldsymbol{\eta}_{t+1}\right). \tag{12.60}$$

Since

$$\Delta\boldsymbol{\theta}_t = \mathbf{G}^{-1}\Delta\boldsymbol{\eta}_t, \tag{12.61}$$

we have

$$\Delta\boldsymbol{\theta}_t = -\varepsilon\mathbf{G}^{-1}\nabla f\left(\boldsymbol{\theta}_t\right). \tag{12.62}$$

This is the natural gradient method with the Riemannian metric $\mathbf{G}(\boldsymbol{\theta})$. See Raskutti and Mukherjee (2015).

Since the underlying manifold is dually flat, e- and m-projections can be used to project a point on the restricted region. See sparse signal processing in the next chapter.

12.1.7 Properties of Natural Gradient Learning

12.1.7.1 Natural Gradient Learning is Fisher Efficient

On-line learning is a sequential procedure of modifying the current estimator $\boldsymbol{\xi}_t$ by using one example $(\boldsymbol{x}_t, y_t)$ at a time. Once an example has been used, it is discarded and not used again. This is useful for the estimator $\hat{\boldsymbol{\xi}}_t$ to trace the change when the optimal $\boldsymbol{\xi}_0$ is slowly changing over time or suddenly changes at certain times. However, when the true target is fixed, this might cause loss of efficiency compared with the maximum likelihood estimator which is obtained by batch learning using all the data. This would be a cost to be paid for the benefit of traceability. To our surprise, this is not true. On-line learning can attain Fisher efficient estimation asymptotically, provided the learning constant is chosen adequately. The following theorem shows this (Amari 1998).

Theorem 12.1 *The estimator obtained by on-line natural gradient learning*

$$\tilde{\boldsymbol{\xi}}_{t+1} = \tilde{\boldsymbol{\xi}}_t - \frac{1}{t}\tilde{\nabla} l\left(\boldsymbol{x}_t, y_t, \tilde{\boldsymbol{\xi}}_t\right) \tag{12.63}$$

is Fisher efficient, attaining the Cramér–Rao bound asymptotically.

Proof Let us denote the error covariance matrix of the estimator at time t by

$$\tilde{\mathbf{V}}_{t+1} = \mathrm{E}\left[\left(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_0\right)\left(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_0\right)^T\right], \tag{12.64}$$

where $\boldsymbol{\xi}_0$ is the true value of $\boldsymbol{\xi}$. We expand the loss at $\boldsymbol{\xi}_t$ as

$$\nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_t\right) = \nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_0\right) + \nabla\nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_0\right)\cdot\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_0\right). \tag{12.65}$$

Then, subtracting $\boldsymbol{\xi}_0$ from both sides of (12.63) and substituting it in (12.64), we have

$$\tilde{\mathbf{V}}_{t+1} = \tilde{\mathbf{V}}_t - \frac{2}{t}\tilde{\mathbf{V}}_t + \frac{1}{t^2}\mathbf{G}^{-1} + O\left(\frac{1}{t^3}\right), \tag{12.66}$$

where

$$\mathrm{E}\left[\nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_0\right)\right] = 0, \tag{12.67}$$
$$\mathrm{E}\left[\nabla\nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_0\right)\right] = \mathbf{G}\left(\boldsymbol{\xi}_0\right) \tag{12.68}$$

are taken into account. We also note that

$$\mathbf{G}\left(\boldsymbol{\xi}_t\right) = \mathbf{G}\left(\boldsymbol{\xi}_0\right) + O\left(\frac{1}{t}\right). \tag{12.69}$$

Then the solution of (12.66) is asymptotically

$$\mathbf{V}_t = \frac{1}{t}\mathbf{G}^{-1}, \tag{12.70}$$

which proves the theorem. □

12.1.7.2 Natural Gradient is Saturation Free

Consider a regression problem, where the output is written as

$$y = f(\boldsymbol{x}, \boldsymbol{\xi}) + \varepsilon. \tag{12.71}$$

First we explain a simple perceptron, where f is written as

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = \varphi(\boldsymbol{w} \cdot \boldsymbol{x}). \tag{12.72}$$

Here, we neglect the bias term for simplicity. The parameter is a vector $\boldsymbol{\xi} = \boldsymbol{w}$ and the activation function φ is a sigmoid function, for example,

$$\varphi(u) = \tanh u. \tag{12.73}$$

The gradient is written as

$$\nabla l(\boldsymbol{x}, y, \boldsymbol{w}) = -(y - f)\varphi'(\boldsymbol{w} \cdot \boldsymbol{x})\boldsymbol{x}. \tag{12.74}$$

When the absolute value of $\boldsymbol{w}$ is large, function $\varphi(\boldsymbol{w} \cdot \boldsymbol{x})$ saturates for most $\boldsymbol{x}$, becoming nearly equal 1 or -1. This is the saturation problem, where the gradient becomes

almost equal to 0 because $\varphi' \approx 0$, and ordinary stochastic gradient descent learning becomes slow.

This is not serious in the case of a simple perceptron, but is serious in the case of multilayer perceptrons used in deep learning, where $f(\boldsymbol{x}, \boldsymbol{\xi})$ is composed of a concatenation of many f's. We may write the output as

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = \varphi\left(\mathbf{W}_k \varphi\left(\mathbf{W}_{k-1} \varphi \ldots \varphi\left(\mathbf{W}_1 \boldsymbol{x}\right)\right)\right), \tag{12.75}$$

in the case of MLP, where $\mathbf{W}_j$ is the connection weight matrix of the jth layer to the $(j+1)$th layer, $\boldsymbol{\xi} = (\mathbf{W}_1, \ldots, \mathbf{W}_k)$. Its derivative with respect to $\mathbf{W}_1$, for example, includes the product of many φ''s. Hence, it is almost vanishing in many cases. This is considered as a flaw of back-propagation in deep learning.

The natural gradient learning method is free of such a saturation problem. The gradient is written as

$$\nabla l(\boldsymbol{x}, y, \boldsymbol{\xi}) = -(y - f) \nabla f(\boldsymbol{x}, \boldsymbol{\xi}). \tag{12.76}$$

The Fisher information is given by

$$\mathbf{G}(\boldsymbol{\xi}) = \mathrm{E}\left[\nabla f(\boldsymbol{x}, \boldsymbol{\xi}) \nabla f(\boldsymbol{x}, \boldsymbol{\xi})^T\right]. \tag{12.77}$$

The magnitude of the ordinary gradient would be very small in many cases but the natural gradient is different. We evaluate the magnitude of the natural gradient vector

$$\tilde{\nabla} l(\boldsymbol{x}, \boldsymbol{\xi}) = \mathbf{G}(\boldsymbol{x}, \boldsymbol{\xi})^{-1} \nabla l(\boldsymbol{x}, \boldsymbol{\xi}) \tag{12.78}$$

by its Riemannian magnitude,

$$\mathrm{E}\left[\|\tilde{\nabla} l\|^2\right] = \mathrm{E}\left[\tilde{\nabla} l^T \mathbf{G} \tilde{\nabla} l\right]. \tag{12.79}$$

Theorem 12.2 *The magnitude of the natural gradient is given by*

$$E\left[\|\tilde{\nabla} l\|^2\right] = tr\left(\bar{\boldsymbol{G}}(\boldsymbol{\xi}) \boldsymbol{G}^{-1}(\boldsymbol{\xi})\right), \tag{12.80}$$

where

$$\bar{G}(\boldsymbol{\xi}) = \mathrm{E}_{p(\boldsymbol{x}, y, \boldsymbol{\xi}_0)}\left[\nabla l(\boldsymbol{x}, \boldsymbol{\xi}) \nabla l(\boldsymbol{x}, \boldsymbol{\xi})^T\right]. \tag{12.81}$$

It does not vanish even when φ' is small. Moreover,

$$\mathrm{E}\left[\|\tilde{\nabla} l\|^2\right] \approx k \tag{12.82}$$

in a neighborhood of the optimal $\boldsymbol{\xi}_0$, where k is the dimension of $\boldsymbol{\xi}$.

Proof From (12.78), we have

$$E\left[\|\tilde{\nabla} l\|^2\right] = \mathrm{E}_{p(\boldsymbol{x},y,\boldsymbol{\xi}_0)}\left[\mathrm{tr}\,\mathbf{G}(\boldsymbol{\xi})\mathbf{G}^{-1}(\boldsymbol{\xi})\nabla l(\boldsymbol{x},\boldsymbol{\xi})\nabla l(\boldsymbol{x},\boldsymbol{\xi})^T\mathbf{G}^{-1}(\boldsymbol{\xi})\right], \tag{12.83}$$

which proves (12.80). When $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, we easily have (12.82). □

12.1.7.3 Adaptive Natural Gradient Learning

The natural gradient method uses $\mathbf{G}^{-1}\left(\boldsymbol{\xi}_t\right)$, so that we need to calculate the inverse of $\mathbf{G}\left(\boldsymbol{\xi}_t\right)$ at each step. When the number of parameters is large, this is computationally intractable. Moreover, calculation of $\mathbf{G}(\boldsymbol{\xi}_t)$ is not easy in the case when the distribution $q(\boldsymbol{x})$ of $\boldsymbol{x}$ is unknown. To avoid this situation, an adaptive method of obtaining $\mathbf{G}^{-1}\left(\boldsymbol{\xi}_t\right)$ recursively has been proposed (Amari et al. 2000). By using the Taylor expansion of

$$\mathbf{G}\left(\boldsymbol{\xi}_{t+1}\right) = \mathbf{G}\left(\boldsymbol{\xi}_t - \eta_t\mathbf{G}^{-1}\nabla l\right) \tag{12.84}$$

and inverting it, we have an adaptive method of calculating $\mathbf{G}_t^{-1} = \mathbf{G}^{-1}\left(\boldsymbol{\xi}_t\right)$ recursively by

$$\mathbf{G}_{t+1}^{-1} = (1+\varepsilon_t)\,\mathbf{G}_t^{-1}\left(\boldsymbol{\xi}_t\right) - \varepsilon_t\mathbf{G}_t^{-1}\nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_t\right)\nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_t\right)^T\mathbf{G}_t^{-1}, \tag{12.85}$$

where ε_t is another learning constant.

Park et al. (2000) demonstrated performance of adaptive natural gradient learning using a number of simple examples, and confirmed that its performance is excellent. See also Zhao et al. (2015). The adaptive method can be used to calculate the inverse of the Hessian,

$$\mathbf{H}_{t+1}^{-1} = (1+\varepsilon_t)\,\mathbf{H}_t^{-1} - \varepsilon_t\mathbf{H}_t^{-1}\nabla\nabla l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\xi}_t\right)\mathbf{H}_t^{-1}. \tag{12.86}$$

12.1.7.4 Approximation and Practical Implementation of Natural Gradient

It is not easy to implement the natural gradient in a large network because of a large computational cost. There are many trials to overcome the difficulty and to give a good approximate solution. See Martens (2015) for the perspectives of the natural gradient method.

Martens and Grosse (2015) proposed an efficient method of approximating natural gradient descent in deep neural networks, called the Kronecker-factored approximate curvature (K-FAC). It uses two stages for the approximation of the Fisher information. One is to use the Kronecker product of the matrices due to error terms and activation terms, and the expectation is taken separately for calculating the Fisher information. The other is to use the tridiagonal approximation for the inverse of

the Fisher information matrix (the Riemannian metric). A deep network consists of a concatenation of many layers, and the Fisher information matrix has a block structure. The tridiagonal approximation neglects off-diagonal blocks except for the blocks corresponding to consecutive $(i-1, i, i+1)$ layers. It is demonstrated that this is not only computationally tractable but its performance is excellent.

We remark that the two approximations do not destroy most of the singular structure of the original Fisher information, studied in the next section. Since the singular regions are the main cause of retardation in learning, the K-FAC works well, getting rid of the plateau phenomena.

12.1.7.5 Adaptive Learning Constant

The dynamical behavior of learning depends on the learning constant η_t. When the current $\boldsymbol{\xi}_t$ is far from the optimal value $\boldsymbol{\xi}_0$, it is desirable to use large η_t, because we need to shift $\boldsymbol{\xi}_t$ toward $\boldsymbol{\xi}_0$ with a large step-size. On the other hand, when $\boldsymbol{\xi}_t$ is near the optimal value, if η_t is large, the stochastic fluctuation of ∇l dominates so that it is better to choose a small η_t. When the optimal value of the target is fixed, a good choice of learning constant is given by stochastic approximation,

$$\sum_{t=1}^{\infty} \eta_t > \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty. \tag{12.87}$$

When η_t satisfies (12.87), the estimator $\boldsymbol{\xi}_t$ converges to the optimal $\boldsymbol{\xi}_0$ with probability one. A typical case is given by

$$\eta_t = \frac{c}{t}. \tag{12.88}$$

When the target does not move, the trade-off between the speed of convergence and the accuracy of estimation is given in Amari (1967) for a fixed η. For the cases when the target moves, the idea of modifying η_t adaptively depending on the current situation of the estimator was considered from the early time. An excellent idea of modifying the learning constant was proposed by Barkai et al. (1995) in the case when y is binary. Amari (1998) generalized it and analyzed its behavior. A new adaptive learning method is given by

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \tilde{\nabla} l\left(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\xi}_t\right), \tag{12.89}$$

$$\eta_{t+1} = \eta_t \exp\left\{\alpha\left[\beta l\left(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\xi}_t\right) - \eta_t\right]\right\}, \tag{12.90}$$

where α, β are constants. Here, the natural gradient method is fortified by a learning rule of learning constant (12.90). The learning rate η_t increases, roughly speaking, when the instantaneous loss $l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_t\right)$ is large, which implies that the target lies far away and η_t decreases when the target is closer.

In order to analyze its behavior mathematically, we use the continuous-time version of the learning equation,

$$\frac{d}{dt}\boldsymbol{\xi}_t = -\eta_t \mathbf{G}^{-1}\left(\boldsymbol{\xi}_t\right) \langle \nabla l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi}_t) \rangle, \tag{12.91}$$

$$\frac{d}{dt}\eta_t = \alpha\eta_t \left\{ \beta \langle l\left(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi}_t\right) \rangle - \eta_t \right\}, \tag{12.92}$$

where the equations are averaged over possible input–output pairs $(\boldsymbol{x}_t, y_t)$, $\langle \ \ \rangle$ representing the average with respect to $p(\boldsymbol{x}, y)$.

By using the Taylor expansion

$$\begin{aligned} \langle \nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_t\right) \rangle &= \langle \nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_0\right) \rangle + \langle \nabla\nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\xi}_0\right) \cdot \left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_0\right) \rangle \\ &= \mathbf{G}_0\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_0\right), \end{aligned} \tag{12.93}$$

where we put $\mathbf{G}_0 = \mathbf{G}\left(\boldsymbol{\xi}_0\right)$, we have

$$\frac{d}{dt}\boldsymbol{\xi}_t = -\eta_t\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_0\right), \tag{12.94}$$

$$\frac{d}{dt}\eta_t = \alpha\eta_t \left\{ \frac{\beta}{2}\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_0\right)^T \mathbf{G}_0\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_0\right) - \eta_t \right\}. \tag{12.95}$$

We introduce the squared error at time t by

$$e_t = \frac{1}{2}\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}\right)^T \mathbf{G}_0\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_0\right). \tag{12.96}$$

Then, the equations reduce to

$$\frac{d}{dt}e_t = -2\eta_t e_t, \tag{12.97}$$

$$\frac{d}{dt}\eta_t = \alpha\beta\eta_t e_t - \alpha\eta_t^2, \tag{12.98}$$

when $\boldsymbol{\xi}_0$ is fixed. The behaviors of the error e_t and learning constant η_t described by (12.97) and (12.98) are interesting. The origin $(0, 0)$ is its stable equilibrium, so both e_t and η_t converge to 0. The solution is written approximately as

$$e_t = \frac{1}{\beta}\left(\frac{1}{2} - \frac{1}{\alpha}\right)\frac{1}{t}, \tag{12.99}$$

$$\eta_t = \frac{1}{2t}, \tag{12.100}$$

for large t. This shows that the error converges to 0 in the order of $1/t$ as t goes to infinity when $\boldsymbol{\xi}_0$ is fixed. When the target changes over time, $\boldsymbol{\xi}_t$ traces its change nicely by modifying η_t.

12.2 Singularity in Learning: Multilayer Perceptron

The multilayer perceptron (MLP), proposed by Rosenblatt (1961), is a universal machine that can approximate any input–output function, provided it includes a sufficiently large number of hidden neurons. Although it seemed to be gradually being replaced by new powerful learning machines such as the support vector machine (SVM), MLP has been revived in the 21st century in "deep learning", where a network has a considerably large number of layers. Lots of new tricks are proposed to facilitate deep learning, including unsupervised learning (self-organization) as preprocessing, the convolutional structure, and the drop-out technique in supervised learning. Deep learning has recorded benchmark performances, winning most competitions on pattern recognition. See Schmidhuber (2015) for example. Researchers are astonished by the reincarnation of the multilayer perceptron. The back-propagation learning method is used at the final stage.

There is, however, a serious problem in the parameter space of a multilayer perceptron. It includes singularities, in the sense that the same output function is realized by continuously many parameters in a specific region. One cannot determine the parameter uniquely in such a region, and so the parameter is not identifiable. The Fisher information matrix degenerates in this region. This causes the dynamics of learning to become extremely slow, which is known as a critical slowdown or the plateau phenomena.

The present section studies typical singular structure in the manifold of multilayer perceptrons and clarifies its implications for statistical inference. The dynamical behavior of learning near singularities is studied in detail. Finally, it is shown that the natural gradient learning method, including SFN, overcomes these difficulties.

12.2.1 Multilayer Perceptron

The multilayer perceptron is a layered machine composed of artificial neurons, which receives input $\boldsymbol{x}$ and emits output y. The behavior of an analog artificial neuron is described as follows: It receives a vector input signal $\boldsymbol{x}$, calculates a weighted sum of inputs and subtracts a threshold as

$$u = \sum w_i x_i - h = \boldsymbol{w} \cdot \boldsymbol{x} - h, \tag{12.101}$$

where $\boldsymbol{w} = (w_1, \ldots, w_n)$. It emits an output

$$y = \varphi(u), \tag{12.102}$$

where φ is a sigmoidal function. We use

$$\varphi(u) = \sqrt{\frac{2}{\pi}} \int_0^u \exp\left\{-\frac{s^2}{2}\right\} ds, \tag{12.103}$$

because this is convenient for obtaining explicit analytical solutions. The coefficients w_i are called the synaptic weights. In order to make descriptions simpler, we put $h = 0$ in the following.

A multilayer perceptron consists of many layers in deep learning, but we consider here only three layers, an input layer, a hidden layer and an output layer (Fig. 12.4). The ith neuron of the hidden layer calculates the weighted sum of input $\boldsymbol{x}$ as

$$u_i = \boldsymbol{w}_i \cdot \boldsymbol{x} \tag{12.104}$$

and emits output $\varphi(u_i)$, where $\boldsymbol{w}_i$ is the weight vector of the ith hidden neuron. We consider a simple case that the output layer consists of only one output neuron. It calculates a weighted sum of the outputs of the hidden neurons and the final output is written as

$$y = \sum v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}), \tag{12.105}$$

where v_i are the weights of the output neuron. We may apply a sigmoidal nonlinear function to y, but it is only a nonlinear scale change. So we use a linear output neuron, but a nonlinear function is used when the output neurons are connected to the next layer as its input.

A multilayer perceptron is specified by synaptic weights

$$\boldsymbol{\xi} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m; v_1, \ldots, v_m). \tag{12.106}$$

Let M be the parameter space of perceptrons. Then, it is an N-dimensional manifold, where $\boldsymbol{\xi}$ is a coordinate system including $N = (n+1)m$ components. We write the input–output relation of the perceptron specified by $\boldsymbol{\xi}$ as

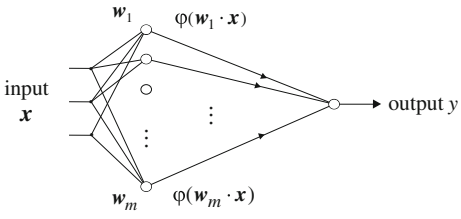


Fig. 12.4 Multilayer perception

$$y = f(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{i=1}^{m} v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}) . \tag{12.107}$$

Learning takes place in the manifold M, where the current value $\boldsymbol{\xi}$ is modified by a stochastic gradient descent method using the current input–output example $(\boldsymbol{x}_t, y_t)$.

12.2.2 Singularities in M

The manifold M includes a set of points which have the same output functions

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = f\left(\boldsymbol{x}, \boldsymbol{\xi}'\right), \tag{12.108}$$

for $\boldsymbol{\xi} \neq \boldsymbol{\xi}'$. Two such points $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ are said to be equivalent and are denoted by

$$\boldsymbol{\xi} \approx \boldsymbol{\xi}', \tag{12.109}$$

since their output functions are the same. When $\boldsymbol{\xi}$ has an equivalent point in M other than itself, we cannot identify $\boldsymbol{\xi}$ uniquely from the output function. There are two types of unidentifiability, originating from the invariance under the following transformations of parameters:

1. Sign change: $\boldsymbol{\xi} \approx -\boldsymbol{\xi}$: This is because φ is an odd function, $\varphi(-u) = -\varphi(u)$, so that $f(\boldsymbol{x}, \boldsymbol{\xi}) = f(\boldsymbol{x}, -\boldsymbol{\xi})$. The unidentifiability due to the sign change is simple, and we may eliminate the unidentifiability by restricting the region within $v_i \geq 0, i = 1, \ldots, m$. However, the boundary $v_i = 0$ causes singularities, as will be shown soon.
2. Permutation: Let $\prod$ be a permutation of indices and i be transformed to i' as $i' = \prod i$. Then,

$$\boldsymbol{\xi} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m;\ v_1, \ldots, v_m) \approx \boldsymbol{\xi}' = (\boldsymbol{w}_{1'}, \ldots, \boldsymbol{w}_{m'};\ v_{1'}, \ldots v_{m'}) . \tag{12.110}$$

We divide M by the equivalence relation $\approx$ and put

$$\tilde{M} = M/\approx . \tag{12.111}$$

Equivalent points in M are reduced to one point in $\tilde{M}$, the space of the output functions of multilayer perceptrons. $\tilde{M}$ is not a manifold in the exact mathematical sense, as will be shown in the following, because it includes singular points due to unidentifiability. It is a manifold if we simply remove the singular points. $\tilde{M}$ is called a behavior manifold or neuromanifold, although it is not a manifold in the exact sense.

We explain the singularity by using simple examples. Consider a very simple perceptron consisting of one hidden neuron, which is included in a larger model as a subnetwork. Its output function is

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}) \tag{12.112}$$

and the parameter space M is $\boldsymbol{\xi} = (\boldsymbol{w}, v)$. When $v = 0$, whatever $\boldsymbol{w}$ is, the output function is 0. On the other hand, when $\boldsymbol{w} = 0$, whatever v is, the output function is also 0, because $\varphi(0) = 0$. We call the set of these points a critical or singular region R of M, that is,

$$R = \{\boldsymbol{\xi} \,|\, v = 0 \text{ or } \boldsymbol{w} = 0\}. \tag{12.113}$$

All the points in R are equivalent. By dividing M by the equivalence relation, $\tilde{M}$ consists of two parts (not four because $(\boldsymbol{w}, v)$ and $(-\boldsymbol{w}, -v)$ are equivalent), which are connected by a single point corresponding to $v = 0$ or $\boldsymbol{w} = 0$. It is a singular point in $\tilde{M}$. See Fig. 12.5. More generally, we consider the following eliminating singularity.

(1) Eliminating singularity: When $v_i = 0$, whatever the value of $\boldsymbol{w}_i$ is, any $\boldsymbol{w}_i$ gives the same output function. Hence, $\boldsymbol{w}_i$ is not identifiable in this case. When $\boldsymbol{w}_i = 0$, whatever v_i is, the output of the neuron is 0. Such a neuron has no effect on the output and it can be eliminated.

Consider a subnetwork consisting of two hidden neurons i and j. Their output function is

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = v_i\varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}) + v_j\varphi\left(\boldsymbol{w}_j \cdot \boldsymbol{x}\right). \tag{12.114}$$

(2) Overlapping singularity: When two neurons i and j in the hidden layer have identical weight vectors,

$$\boldsymbol{w}_i = \boldsymbol{w}_j = \boldsymbol{w}, \tag{12.115}$$

their contribution to the output is

$$v_i\varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}) + v_j\varphi\left(\boldsymbol{w}_j \cdot \boldsymbol{x}\right) = \left(v_i + v_j\right)\varphi(\boldsymbol{w} \cdot \boldsymbol{x}). \tag{12.116}$$

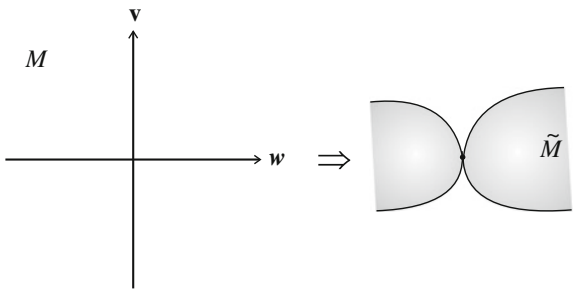


Fig. 12.5 Eliminating singularity

Therefore, the output is the same whatever values v_i and v_j take, as long as $v_i + v_j$ is equal to a fixed value v. That is, the output is the same on the line satisfying

$$v_i + v_j = v \tag{12.117}$$

for any constant v. Hence, v_i and v_j themselves are not identifiable. This occurs when two neurons have the same weight vector $\boldsymbol{w}_i = \boldsymbol{w}_j = \boldsymbol{w}$, with their weight vectors overlapping completely. A similar situation holds when $\boldsymbol{w}_i = -\boldsymbol{w}_j$, but we omit this case for simplicity's sake.

The critical region due to the overlapping singularity is given by

$$R_{oij}(\boldsymbol{w}, v) = \left\{ \boldsymbol{\xi} \,\middle|\, \boldsymbol{w}_i = \boldsymbol{w}_j = \boldsymbol{w},\ v_i + v_j = v \right\}. \tag{12.118}$$

See Fig. 12.6, where $R_{oij}(\boldsymbol{w}, v)$ is mapped to a single point in $\tilde{M}$. The images of the $R_{oij}(\boldsymbol{w}, v)$ form a continuous submanifold as $\boldsymbol{w}$ and v vary. The critical region in M is written as

$$R = \left\{ \boldsymbol{\xi} \,\middle|\, \prod_i v_i \, |\boldsymbol{w}_i| \prod_{i \neq j} \left| \boldsymbol{w}_i - \boldsymbol{w}_j \right| = 0 \right\}, \tag{12.119}$$

which is a union of critical submanifolds (12.118).

We consider an equivalence class $R_{ij}(\boldsymbol{w}, v)$ specified by two parameters $\boldsymbol{w}$ and v, such that any networks in this class have the same output function

$$f(\boldsymbol{x}; \boldsymbol{w}, v) = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}). \tag{12.120}$$

It consists of three parts, R_o,R_{ei} and R_{ej},

$$R_{ij}(\boldsymbol{w}, v) = R_o \cup R_{ei} \cup R_{ej}, \tag{12.121}$$

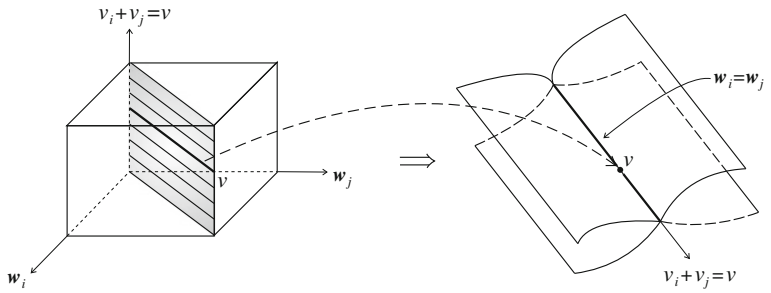


Fig. 12.6 Overlapping singularity

where

$$R_o = \left\{ \boldsymbol{\xi} \,\middle|\, v_i + v_j = v,\ \boldsymbol{w}_i = \boldsymbol{w}_j = \boldsymbol{w} \right\}, \tag{12.122}$$

$$R_{ei} = \left\{ \boldsymbol{\xi} \,\middle|\, v_i = 0,\ v_j = v,\ \boldsymbol{w}_i \text{ is arbitrary}, \boldsymbol{w}_j = \boldsymbol{w} \right\}, \tag{12.123}$$

$$R_{ej} = \left\{ \boldsymbol{\xi} \,\middle|\, v_j = 0,\ v_i = v,\ \boldsymbol{w}_j \text{ is arbitrary}, \boldsymbol{w}_i = \boldsymbol{w} \right\}. \tag{12.124}$$

R_o is a one-dimensional subspace corresponding to the overlapping singularity, where $z = v_i - v_j$ is a free parameter in it, keeping the sum $v_i + v_j = v$ constant. R_{ei} and R_{ej} correspond to the eliminating singularity. They are n-dimensional, since $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, respectively, can take any values. $R_{ij}(\boldsymbol{w}, v)$ is an elementary critical region which is a union of three parts, as is shown in Fig. 12.7. All the points in it are mapped to a single point $f = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x})$ in the behavior manifold $\tilde{M}$. This is a singular point in $\tilde{M}$.

There are infinitely many such critical regions, because we have an elementary critical region for each $\boldsymbol{w}$ and v and they are distributed continuously. So they form a continuum of singular points in the behavior manifold $\tilde{M}$ where $\boldsymbol{w}$ and v are parameters. The region is further contracted when

$$v|\boldsymbol{w}| = 0 \tag{12.125}$$

holds. Such critical regions exist for each pair (i, j) in a larger network and they intersect. So M includes a rich net of critical regions spreading over M.

The trajectory of learning is given by (12.12) in M. It is mapped to $\tilde{M}$ and it may pass through a critical region in M or a singular point in $\tilde{M}$. We study the dynamical behavior of learning near singularities.

The loss function takes the same value in a critical region $R_{ij}(\boldsymbol{w}, v)$, so that its derivative in the tangent directions of $R_{ij}(\boldsymbol{w}, v)$ is always 0. This also implies that the Fisher information degenerates in the critical region R_{ij} of M, because there are directions $\boldsymbol{a}$ in R_{ij} such that

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = f(\boldsymbol{x}, \boldsymbol{\xi} + c\boldsymbol{a}) \tag{12.126}$$

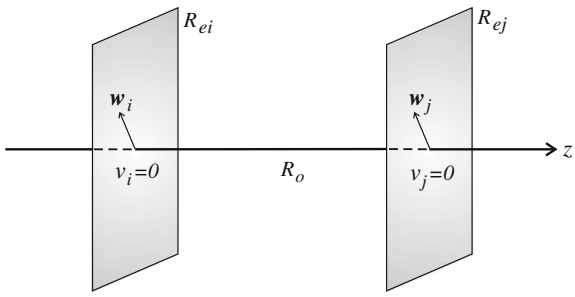


Fig. 12.7 Critical region $R_{ij}(\boldsymbol{w}, v)$

holds for any c, as is derived from (12.118). $\boldsymbol{a}$ is one-dimensional in region R_o and n-dimensional in regions R_{ei} and R_{ej} (Fig. 12.7). Hence, the score function, that is the derivative of log-likelihood, becomes 0 in these directions. This implies that the Fisher information matrix has null directions in which

$$\boldsymbol{a}^T \mathbf{G}(\boldsymbol{\xi})\boldsymbol{a} = 0. \tag{12.127}$$

So it degenerates and $\mathbf{G}^{-1}$ diverges on the critical region. The Fisher information exists and is non-degenerate in $\tilde{M}$ except for singular points. No tangent space exists at a singular point of $\tilde{M}$. This is the same for the absolute Hessian metric and

$$\boldsymbol{a}^T |\boldsymbol{H}(\boldsymbol{\xi})| \boldsymbol{a} = 0 \tag{12.128}$$

holds in R in the direction satisfying $\boldsymbol{\xi} \approx \boldsymbol{\xi} + \boldsymbol{a}$.

A probability distribution $p(\boldsymbol{x}, y, \boldsymbol{\xi})$ accompanies each point of M and $\tilde{M}$, but these probability distributions do not form a regular statistical model, because the non-degenerate Fisher information does not exist in critical regions or at singular points. We will discuss how the singularity affects statistical inference in a later subsection.

12.2.3 Dynamics of Learning in M

Multilayer perceptrons suffer from two types of flaw in their learning behavior. One is local minima such that the global minimum might not be attained by the gradient method. The second is the slowness of convergence, because the trajectory of learning is often trapped on a plateau, staying there for a long time before escaping from it (Amari et al. 2006). This is mostly due to the symmetric structure, such that its behavior is invariant under sign changes and permutations of hidden neurons.

Geometrically speaking, the plateau phenomena are given rise to by the singular structure. A critical region forms a plateau. We will analyze the dynamics of vanilla stochastic gradient learning in the neighborhood of a critical region. We will also show that the natural gradient is free of the plateau phenomena.

In order to analyze the dynamics, we use a very simple model consisting of two hidden neurons described in (12.114). Such simple models are embedded in a general perceptron as parts and cause a serious slowdown in learning. Instead of the difference Eq. (12.12) of stochastic descent learning, we use the averaged version in the continuous time,

$$\dot{\boldsymbol{\xi}}(t) = -\eta \langle \frac{\partial l\,(\boldsymbol{x}, y, \boldsymbol{\xi}(t))}{\partial \boldsymbol{\xi}} \rangle, \tag{12.129}$$

where $\langle \quad \rangle$ is the average with respect to the joint probability distribution $p(\boldsymbol{x}, y, \boldsymbol{\xi}_0)$ of the true or teacher system from which training examples are generated. We further assume that the probability distribution of input $\boldsymbol{x}$ is subject to the Gaussian

distribution $N(0, \mathbf{I})$ with mean 0 and covariance matrix $\mathbf{I}$, the identity matrix. These assumptions are useful for obtaining explicit solutions.

In order to analyze the behavior of dynamics (12.129) consisting of two hidden neurons, we use a new coordinate system $\boldsymbol{\zeta}$ (Wei et al. 2008),

$$\boldsymbol{\zeta} = (\boldsymbol{u}, z, \boldsymbol{s}, r), \tag{12.130}$$

where

$$\boldsymbol{u} = \boldsymbol{w}_2 - \boldsymbol{w}_1, \quad \boldsymbol{s} = \frac{v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2}{v_1 + v_2}, \tag{12.131}$$

$$z = \frac{v_1 - v_2}{v_1 + v_2}, \quad r = v_1 + v_2 \tag{12.132}$$

and we use suffixes 1, 2 instead of i, j. The critical region $R = R_{12}(\boldsymbol{w}, v)$ is given in this new coordinate system by

$$R = \{\boldsymbol{u} = 0 \ \text{ or } \ z = \pm 1\}, \tag{12.133}$$

in which $\boldsymbol{s} = \boldsymbol{w}$ and $r = v$ hold. We divide it into two parts $R = R_o \cup R_e$,

$$R_o = \{\boldsymbol{\zeta} |\ \boldsymbol{u} = 0\}, \tag{12.134}$$

$$R_e = \{\boldsymbol{\zeta} |\ z = \pm 1\}, \tag{12.135}$$

where R_o is the overlapping singularity and $R_e = R_{e1} \cup R_{e2}$ is the eliminating singularity.

The dynamics (12.129) are described in the new coordinate system as

$$\dot{\boldsymbol{\zeta}} = -\eta \mathbf{T}\mathbf{T}^T \langle \frac{\partial l(\boldsymbol{x}, y, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} \rangle, \tag{12.136}$$

where $\mathbf{T}$ is the Jacobian matrix of the coordinate transformation from $\boldsymbol{\xi}$ to $\boldsymbol{\zeta}$,

$$\mathbf{T} = \frac{\partial \boldsymbol{\zeta}}{\partial \boldsymbol{\xi}}. \tag{12.137}$$

The output function f is written as

$$f(\boldsymbol{x}, \boldsymbol{\zeta}) = \frac{1}{2} r(1+z) \varphi \left[\left\{ \boldsymbol{s} + \frac{1}{2}(z-1)\boldsymbol{u} \right\} \cdot \boldsymbol{x} \right] + \frac{1}{2} r(1-z) \varphi \left[\left\{ \boldsymbol{s} + \frac{1}{2}(z+1)\boldsymbol{u} \right\} \cdot \boldsymbol{x} \right] \tag{12.138}$$

in terms of the new coordinates. We expand it in the Taylor series in the neighborhood of R_o,

$$f(\boldsymbol{x}, \boldsymbol{\zeta}) = r\varphi(\boldsymbol{s} \cdot \boldsymbol{x}) + \frac{1}{8}\left(1 - z^2\right)\boldsymbol{u}'\mathbf{J}\boldsymbol{u}, \tag{12.139}$$

$$\mathbf{J} = \frac{\partial^2 \varphi(\boldsymbol{s} \cdot \boldsymbol{x})}{\partial \boldsymbol{s} \partial \boldsymbol{s}}, \tag{12.140}$$

where higher-order terms of $\boldsymbol{u}$ are neglected. We then have the learning dynamics in terms of $\boldsymbol{\zeta} = (\boldsymbol{u}, z)$ in the neighborhood of R_o. The dynamics concerning variables $\boldsymbol{s}$ and r are subject to the usual differential equations (fast dynamics) and their values converge rapidly to their equilibrium values, even when the behaviors of $\boldsymbol{u}$ and z are suffering from a critical slowdown (slow dynamics). Hence, we analyze the equations concerning $\boldsymbol{u}$ and z, where $\boldsymbol{s}$ and r are assumed to have converged to their equilibrium values $\boldsymbol{w}$ and v. The resultant dynamics are

$$\dot{\boldsymbol{u}} = 2\left(1 - z^2\right)\mathbf{K}\boldsymbol{u}, \tag{12.141}$$

$$\dot{z} = -\frac{z\left(z^2 + 3\right)}{r^2}\boldsymbol{u}'\mathbf{K}\boldsymbol{u}, \tag{12.142}$$

where

$$\mathbf{K} = \frac{r}{4}\langle\{y - f(\boldsymbol{x}, \boldsymbol{\zeta})\}\,\mathbf{J}\rangle. \tag{12.143}$$

It is clear that

$$\frac{d\boldsymbol{\zeta}}{dt} = 0 \tag{12.144}$$

in the region $R = R_o \cup R_e$, so any points in R are equilibria. The stability of the equilibria depends on $\mathbf{K}$. We show the results without proofs (which are technical and complicated but not difficult, see Wei et al. 2008; Wei and Amari 2008).

Theorem 12.3 *When the teacher output function is in the critical region, the equilibria are stable.*

This case occurs when the system is over-realizable, having redundant parameters.

Theorem 12.4 *When the teacher output function is outside the critical region, we have three cases, depending on the eigenvalues of* $\mathbf{K}$*:*

(1) The equilibrium solutions on R_o satisfying $|z| > 1$ are stable and those satisfying $|z| < 1$ are unstable when $\mathbf{K}$ is positive-definite.
(2) The equilibrium solutions on R_o satisfying $|z| < 1$ are stable and those satisfying $|z| > 1$ are unstable when $\mathbf{K}$ is negative-definite.
(3) The solutions on R_o are unstable when some eigenvalues are positive and some negative.

We further analyze the trajectories of the solutions in the neighborhood of R_o. Let us introduce a function

$$h(\boldsymbol{u}) = \frac{1}{2}|\boldsymbol{u}|^2, \tag{12.145}$$

which shows how far the current $\boldsymbol{\zeta}$ is from R_o. Its time derivative is given, from (12.141) and (12.142), as

$$\dot{h}(\boldsymbol{u}) = \boldsymbol{u}^T \dot{\boldsymbol{u}} = \frac{2r^2\left(z^2-1\right)}{z\left(z^2+3\right)}\dot{z}. \tag{12.146}$$

The equation is integrable, and the solution is

$$h(\boldsymbol{u}) = \frac{2r^2}{3}\log\frac{\left(z^2+3\right)^3}{|z|} + c, \tag{12.147}$$

where c is an arbitrary constant that specifies a trajectory.

Theorem 12.5 *The trajectories of learning are*

$$h(\boldsymbol{u}) = \frac{2r^2}{3}\log\frac{\left(z^2+3\right)^2}{|z|} + c \tag{12.148}$$

in the neighborhood of R_o.

The family of trajectories shows how the dynamics proceed in the neighborhood of R_o. The behaviors are the same for any $\boldsymbol{\xi} \in R_o$, but their stabilities depend on $\boldsymbol{\xi}$ and $\mathbf{K}$. See Fig. 12.8. When $\boldsymbol{\xi}_0$ is in R, R is stable. When $\mathbf{K}$ is positive-definite or negative-definite, the trajectory starting from the basin of attraction reaches a stable point in R_o and is trapped in it, fluctuating in it randomly before escaping from it.

12.2.4 Critical Slowdown of Dynamics

We consider the two cases separately.

Case 1: **The teacher function is in R**. When the number of hidden neurons is larger in the model network (student network) to be trained than in the teacher network (true network), some neurons are redundant because the optimal solution is realized by using a smaller number of neurons. This is the over-realizable case. In this case, elimination of neurons or overlap of synaptic weight vectors occurs, implying that the optimal solution is in R.

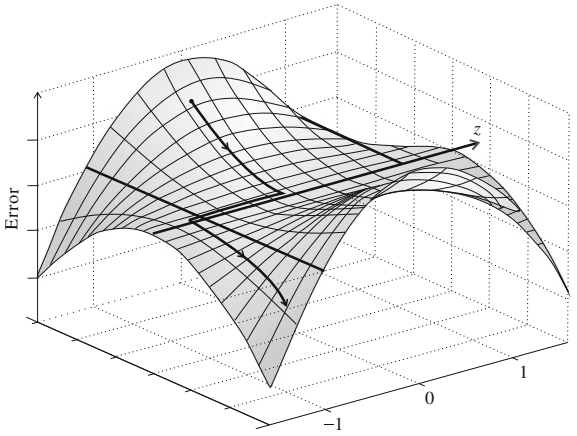


Fig. 12.8 Landscape of error function and learning trajectory

When the teacher network is $\boldsymbol{\xi}_o \in R$, (12.143) is written as

$$\mathbf{K} = \frac{r}{4}\langle e \frac{\partial^2 \varphi}{\partial \boldsymbol{s} \partial \boldsymbol{s}} \rangle, \tag{12.149}$$

where

$$e = \langle f(\boldsymbol{x}, \boldsymbol{\zeta}) - f\left(\boldsymbol{x}, \boldsymbol{\zeta}_0\right) \rangle \tag{12.150}$$

is the error term and is 0 when $\boldsymbol{\zeta} \in R$, in particular, when $\boldsymbol{u} = 0$. By expanding the error term, we can easily obtain

$$\mathbf{K} = O\left(|\boldsymbol{u}|^2\right). \tag{12.151}$$

This implies that the dynamics of $\boldsymbol{u}$ are

$$\frac{d\boldsymbol{u}}{dt} = O\left(|\boldsymbol{u}|^3\right). \tag{12.152}$$

Hence, the speed of convergence of $\boldsymbol{u}$ to 0 is extremely slow, taking a long time for training (Fig. 12.9a). This is frequently observed in simulations.

Case 2: **The optimal solution lies outside R**. Points in R are equilibrium solutions. $\mathbf{K}$ is not small in this case, because the error term is not small at R. When $\mathbf{K}$ is positive-definite or negative-definite, the part of R_o, $|z| > 1$ or $|z| < 1$, respectively, is stable but the other part is unstable. The landscape of the loss function is shown in this case in Fig. 12.8, where R_o is shown by the solid line. Starting $\boldsymbol{\zeta}$ at some initial point belonging to the basin of attraction, the state is attracted to the stable part of R_o. See Fig. 12.9b, c. The value of the loss function is the same and its derivative is 0 on R_o since all points in R_o are equivalent. However, this is not the optimal point. The

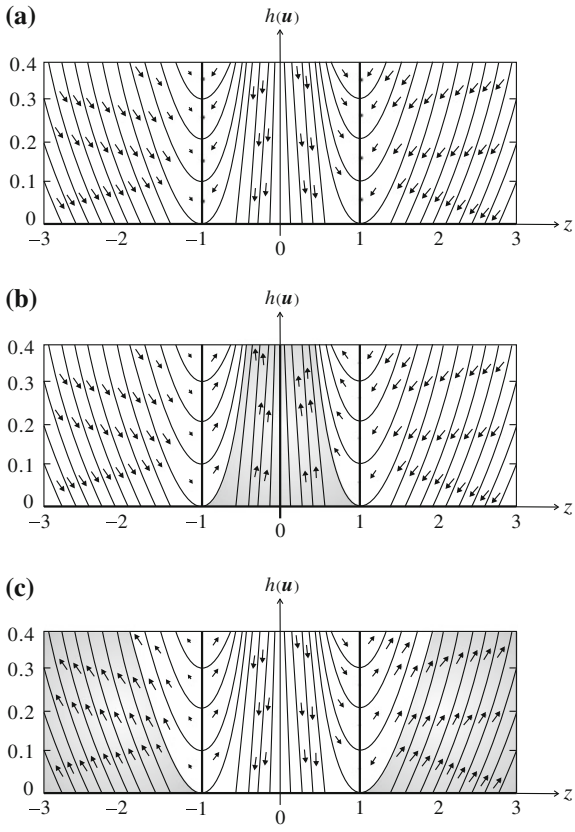


Fig. 12.9 Trajectories of learning near singularity: **a** Teacher is at singularity; **b** $|z| < 1$ is stable; **c** $|z| > 1$ is stable

state fluctuates in the neighborhood of R_o by stochastic dynamics due to randomly selected input $\boldsymbol{x}$. Thus, a random walk of the state takes place in the neighborhood of R_o and the state eventually reaches the boundary $|z| = 1$ of the stable region. It thus enters the unstable region and then escapes from R_o immediately, moving toward the true optimal point. However, it takes a long time before leaving the stable critical region. See Fig. 12.10. Precisely speaking, the fluctuation around R_o is not a random walk, because there are systematic flows out of the stable region in the neighborhood of R_o, but the flow is very small when $\boldsymbol{u}$ is small.

Although the trajectories passing through R have incoming flows and outgoing flows at R, this is completely different from those at a saddle point. The basin of attraction has measure 0 in the case of a saddle. Therefore, it is at measure 0 that the state reaches the saddle. Moreover, the state escapes from the saddle quickly by a small perturbation. On the other hand, the basin of attraction of R has a finite measure and the trajectory exactly reaches R in this case. A small perturbation moves the state but it again reaches R. This does not prevent a trajectory reaching R. A saddle does not cause any serious effect on the slowdown of dynamics. It is a critical region that causes a critical slowdown.

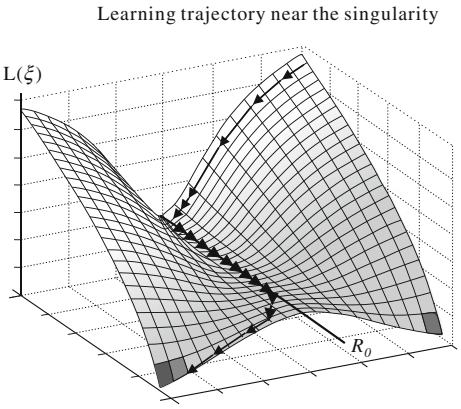


Fig. 12.10 Trajectory of learning near the singularity

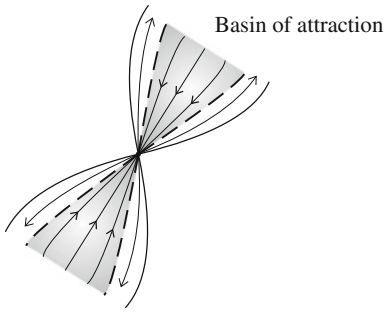


Fig. 12.11 Trajectories in $\tilde{M}$

We can consider the same dynamics in $\tilde{M}$ where R is reduced to one point by the equivalence. The point corresponding to R is a singular point. It is a Milnor attractor in $\tilde{M}$, of which the basin of attraction has a finite measure (Milnor 1985). The trajectories enter it and then emerge from it (Fig. 12.11). A general multilayer perceptron includes a net of such critical regions within it. The trajectory of vanilla stochastic gradient learning is trapped in such critical regions many times before it reaches the optimum solution. This is known as the plateau phenomena. See Fig. 12.12 for an example of learning curves.

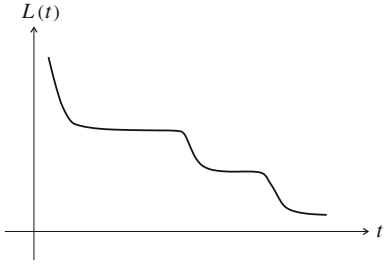


Fig. 12.12 Plateaus

12.2.5 Natural Gradient Learning Is Free of Plateaus

The plateau phenomena are given rise to by the singularities. Let us consider a simple case of (12.114), where the horizontal line (z-axis) in Fig. 12.7 is the critical region and all the points in this line are equivalent. The Riemannian length is 0 along this line and the Riemannian metric degenerates in this direction. The inverse of the Fisher metric diverges in this direction to infinity at R. The gradient of the cost function is also 0 in this direction because all the points in R are equivalent. Therefore, the natural gradient, $\tilde{\nabla} l = G^{-1}\nabla l$, is 0 multiplied by infinity at the singular points. Because of this, the natural gradient takes an ordinary value even in a very small neighborhood of R.

Cousseau et al. (2008) analyzed the dynamics of natural gradient learning near singularity when the teacher $\boldsymbol{\zeta}_0$ is in R. After complicated calculations,

$$\dot{u} = \frac{-\eta}{2}\left(1 - z^2\right) u, \tag{12.153}$$

$$\dot{z} = \frac{\eta}{2}\left(1 - z^2\right) z \tag{12.154}$$

is derived in the one-dimensional case. This shows that the dynamics converges to R in the linear order. Hence, no retardation takes place.

When $\boldsymbol{\zeta}_0$ is outside R, the trajectory is trapped in plateaus in the case of ordinary stochastic gradient learning. However, in the case of natural gradient learning, no retardation takes place, because the Riemannian metric is 0 along the R_o-direction so that all the points are reduced to a single point. That is, the trajectory enters a point in R and goes out immediately not staying within it. This is well understood by considering the trajectory in $\tilde{M}$.

In $\tilde{M}$, R reduces to the single singular point, and all the other points in $\tilde{M}$ are regular, having a non-degenerate Riemannian metric. Even in a very small neighborhood of R, $\mathbf{G}^{-1}\nabla l$ takes ordinary values. Hence, a critical slowdown does not occur. To show this, Cousseau et al. (2008) used the blow-down technique of algebraic geometry. They introduced a new coordinate system $\boldsymbol{\mu} = (\delta, \gamma)$,

$$\delta = \left(1 - z^2\right) u^2, \tag{12.155}$$

$$\gamma = z\left(1 - z^2\right) u^3, \tag{12.156}$$

when $\boldsymbol{u}$ is one-dimensional. All the points in singular region R is mapped to a single point $\boldsymbol{\mu} = (0, 0)$. The Fisher information $\mathbf{G}$ takes ordinary values even in a small neighborhood of R except for $(0, 0)$ at which it is not defined. They showed that

$$\langle \nabla l \rangle = \mathbf{G}\boldsymbol{\mu} \tag{12.157}$$

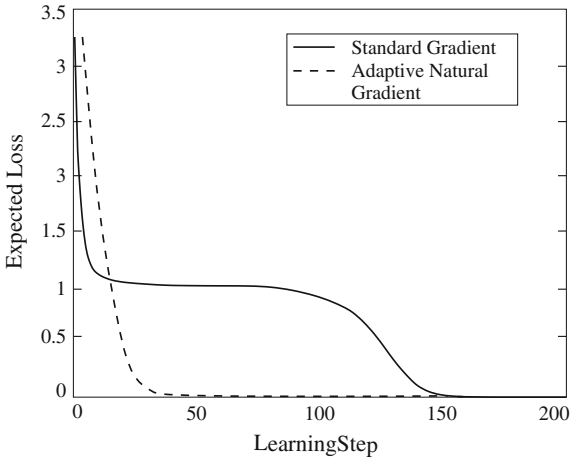


Fig. 12.13 Learning *curves*

holds in this coordinate system when the teacher is in R. Hence, the natural gradient learning dynamics becomes very simple,

$$\dot{\boldsymbol{\mu}} = -\eta \boldsymbol{\mu}, \tag{12.158}$$

in a neighborhood of R, when the teacher is inside R. When the teacher is outside R, the trajectory enters R, that is, $\boldsymbol{\mu} = 0$ without retardation, and then escapes from it immediately. It is interesting to see that, starting from various initial points, the trajectories once enter R and then go out. The basin of attraction of R has a finite major, although the trajectories leave it immediately (see Fig. 12.11). This is a typical Milnor attractor. The new coordinate system $\boldsymbol{\mu}$, using the blow-down technique, is useful. It should be remarked that absolute Hessian dynamics have the same characteristics.

See Fig. 12.13 for examples of the learning curves of the adaptive natural gradient learning method compared to the ordinary back-propagation method.

12.2.6 Singular Statistical Models

A statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$ is regular when it satisfies the two conditions:

(1) The parameter $\boldsymbol{\xi}$ belongs to an open set in a Euclidean space.
(2) The Fisher information matrix exists and is non-singular.

In this case, n score functions

$$u_i(\boldsymbol{x}, \boldsymbol{\xi}) = \frac{\partial \log p(\boldsymbol{x}, \boldsymbol{\xi})}{\partial \xi^i}, \quad i = 1, \ldots, n \tag{12.159}$$

are linearly independent and the tangent space T_{ξ} is spanned by them. The standard asymptotic theory of statistics holds, as is highlighted by the Cramér–Rao theorem. However, the theory is violated in a singular statistical model.

There are many singular statistical models. One type is the case in which the Fisher information matrix degenerates at singularities. A mixture model

$$p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{v}) = \sum v_i p(\boldsymbol{x}, \boldsymbol{w}_i), \quad \sum v_i = 1, \quad v_i \geq 0, \tag{12.160}$$

where $p(\boldsymbol{x}, \boldsymbol{w})$ is a regular statistical model specified by $\boldsymbol{w}$, belongs to this class. The MLP belongs to this class. When $p(\boldsymbol{x}, \boldsymbol{w})$ is a Gaussian distribution with varying mean and variance, it is called a Gaussian mixture model. The changing time model (sometimes called the Nile River model) and the ARMA model in time series also belong to this type.

Another type deals with the case where the Fisher information matrix diverges to infinity. A typical example is the location model written as

$$p(x, \xi) = f(x - \xi), \tag{12.161}$$

where $f(x)$ is a function having a finite support and its derivative is not 0 at the boundaries. The unknown parameter is the mean value ξ. A typical example is the uniform distribution over $[\xi, 1 + \xi]$. We do not discuss this case, although its geometry is interesting, because its metric is not Riemannian but Finslerian. We do not have a good geometrical theory yet. See a preliminary study by Amari (1984).

For N observations from a probability distribution $p(\boldsymbol{x}, \boldsymbol{\xi})$, consider the log likelihood ratio divided by $\sqrt{N}$,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \log \frac{p(\boldsymbol{x}_i, \boldsymbol{\xi})}{p(\boldsymbol{x}_i, \boldsymbol{\xi}_0)}. \tag{12.162}$$

It is asymptotically subject to the χ^2-distribution with n degrees of freedom, where n is the dimension number of $\boldsymbol{\xi}$, when M is regular. By analyzing its behavior, we can prove that the maximum likelihood estimator is asymptotically best, unbiased and Gaussian, the error covariance matrix of which is the inverse of the Fisher information matrix divided by N asymptotically.

The maximum likelihood estimator is no more subject to the Gaussian distribution even asymptotically in a singular statistical model of the first type when the true distribution is at a singular point. However, it is asymptotically consistent and its convergence speed is in the order of $1/\sqrt{N}$. It has been known for many years that some statistical models are singular. Fukumizu (2003) proved that the log likelihood (12.162) diverges to infinity in the order of $\log N$ and $\log \log N$ in the cases of multilayer perceptrons and mixture models, respectively. There is a Japanese monograph by Fukumizu and Kuriki (2004), which studies singular statistical models in detail.

Model selection is an important problem, which decides the number of hidden neurons from observed data in the case of the multilayer perceptron. As is well

known, a model having a large number of free parameters fits the observed data well. The training error decreases as the number of parameters increases. However, the estimated parameters overfit and are not useful for predicting the behavior of future data, because the generalization error increases as the number of parameters increases beyond a certain value. There is an adequate number of parameters, which should be decided from the observed data.

The Akaike Information Criterion (AIC) and Minimum Description Length (MDL) are two well-known criteria for model selection. The Baysian Information Criterion (BIC) is the same as MDL, although their underlying philosophies are different.

Multilayer perceptrons and Gaussian mixtures are models of frequent use in applications. They are hierarchical singular models in which a lower degree model is included in the critical region of a higher degree model. We need to decide an adequate degree, that is, the number of parameters from the observed data. AIC and MDL are frequently used for this purpose without the singular structure being taken into account. There have been many discussions concerning which criteria are to be used, AIC or MDL. Both AIC and MDL are derived by using the maximum likelihood estimator, assuming that it is asymptotically Gaussian with covariance matrix $1/N$ times the inverse of the Fisher information. However, it is not Gaussian when the true parameter is in the critical region. When the true distribution is in a smaller model, it is in a critical region of a larger model. So neither MDL nor AIC are valid in such hierarchical models. They need to be modified. We should take account of corrections due to the singularity. In the case of multilayer perceptrons, the penalty term of AIC should be $\log N$ times the number of parameters, instead of twice the number of parameters. This comes from the asymptotic property of log likelihood. Watanabe (2010) proposed a new information criterion taking the singular structure into account.

12.2.7 Bayesian Inference and Singular Model

Bayesian inference presumes a prior distribution $\pi(\boldsymbol{\xi})$ on the parameters $\boldsymbol{\xi}$ of a statistical model. For a family of probability distributions $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$, the joint probability of $\boldsymbol{\xi}$ and $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}, \boldsymbol{\xi}) = \pi(\boldsymbol{\xi}) p(\boldsymbol{x}|\boldsymbol{\xi}). \tag{12.163}$$

Therefore, the conditional distribution of $\boldsymbol{\xi}$, conditioned on the observed training data is

$$p\left(\boldsymbol{\xi} \,|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) = \frac{\pi(\boldsymbol{\xi}) \prod_{i=1}^{N} p\left(\boldsymbol{x}_i \,|\boldsymbol{\xi}\right)}{\int \pi(\boldsymbol{\xi}) \prod p\left(\boldsymbol{x}_i|\boldsymbol{\xi}\right) d\boldsymbol{\xi}}. \tag{12.164}$$

Its logarithm divided by N is

$$\frac{1}{N}\log p\left(\boldsymbol{\xi}\,|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\right)=\frac{1}{N}\log\pi(\boldsymbol{\xi})+\frac{1}{N}\sum\log p\left(\boldsymbol{x}_i\,|\boldsymbol{\xi}\right)+c, \tag{12.165}$$

where c is a term not depending on $\boldsymbol{\xi}$. The maximum posterior estimate is its maximizer,

$$\hat{\boldsymbol{\xi}}_{MAP}=\arg\max\frac{1}{N}\log p\left(\boldsymbol{\xi}\,|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\right). \tag{12.166}$$

As is shown in (12.165), a penalty term due to the Bayesian prior distribution is added to the loss function, which is the negative of the log prior probability.

The effect of the prior distribution decreases as the number of the training examples N increases in a regular statistical model, as is seen from (12.165). The maximum a posteriori estimator (MAP) converges to the maximum likelihood estimator in this case. However, a singular statistical model has different characteristics.

Let us consider a smooth non-zero prior in a singular model like the multilayer perceptron. It includes the critical region R which is a union of subspaces, including an infinite number of points. Such a region is reduced to one point in the space of the outputs functions $\tilde{M}$. Hence, a uniform prior (improper prior) on the parameter space M is not uniform on $\tilde{M}$. The prior of a singular point is an integration of prior probabilities over an equivalence class R, so that the prior distribution of $\tilde{M}$ is singular, because singular points in $\tilde{M}$ have an infinitely large prior probability measure compared to a regular point.

The parameter space M_n of a perceptron including n hidden neurons is included in M_{n+1} as a submanifold. But M_n is included in M_{n+1} as a critical region, because it is given by $v_i=0$, $|\boldsymbol{w}_i|=0$ or $\boldsymbol{w}_i=\boldsymbol{w}_j$ in M_{n+1}. Hence, when we consider a smooth non-zero prior in M_{n+1}, a singular point $\tilde{M}_{n+1}$ collects prior probabilities of infinitely many points in a critical region of M_{n+1}.

When we take the maximum a posteriori estimator, a model having a smaller number of parameters is advantageous because of the singular prior. Hence, the Bayesian MAP has a tendency to select a smaller model, automatically selecting an adequate model, although there is no guarantee that this is optimal.

Watanabe and his school (Watanabe 2001, 2009) have studied the effects of singularity in Bayesian inference by using modern algebraic geometry. The theory uses deeper knowledge of mathematics and is beyond the scope of the present monograph.

Remarks

The present chapter focuses on the natural gradient method in a Riemannian manifold. Since many engineering problems are formulated in a Riemannian manifold, the natural gradient is useful. We have treated on-line and batch learning procedures and shown that the natural gradient method demonstrates excellent performance.

The multilayer perceptron uses the gradient method (back-propagation) in a Riemannian manifold of parameters. It is a constituent of deep learning, so its dynamical performance should be studied carefully. However, the parameter space includes

widely spread singular regions in which the Fisher metric degenerates. Hence it is not a regular statistical model but is a singular statistical model. We have studied the dynamics of back-propagation learning based on the vanilla gradient, showing its bad performance due to singularities. The natural gradient method is free from such flaws both for the Fisher metric and the absolute Hessian metric. This characteristic is retained in the K-FAC approximation (Martens and Grosse 2015). However, it remains as a problem to be studied how the dynamics of learning behaves in a neighborhood of singularity when the true model is not in the singular region. We will be able to show by using the blow-down technique that the trajectory is not trapped in the singularity. We have also studied the statistical problem related to singularities.

There are other interesting topics related to the natural gradient in a Riemannian manifold. One may use any Riemannian metric, such as the Killing metric in $Gl(n)$ and the absolute Hessian metric (Dauphin et al. 2014). Girolami and Calderhead (2011) presented the MCMC method in a Riemannian manifold by using the natural gradient. Reinforcement learning also uses the natural gradient in a policy manifold which is Riemannian. See, e.g., Kakade (2001), Kim et al. (2010), Roux et al. (2014), Peters and Schaal (2008), Thomas et al. (2013). Optimization in the stochastic relaxation regime is another area where natural gradient learning is effective (Malagò and Pistone 2014; Malagò et al. 2013, Hansen and Ostermeier 2001). One important problem is to evaluate the inverse of the Fisher information or its approximation effectively. See Martens (2015) and Martens and Grosse (2015). The adaptive natural gradient method is one solution.

The natural gradient method is a first-order gradient method in a Riemannian manifold and is different from a second-order method such as the Newton method. We can further extend the natural gradient method to the natural Newton method, natural conjugate gradient method, etc. in a Riemannian manifold. See Edelman et al. (1998), Honkela et al. (2010) and Malago and Pistone (2014).

Chapter 13
Signal Processing and Optimization

In the real world, signals are mostly stochastic. Signal processing makes use of stochastic properties to find the hidden structure we want to know about. The present chapter begins with principal component analysis (PCA), by studying the correlational structure of signals to find principal components in which the directions of signals are widely spread. Orthogonal transformations are used to decompose signals into non-correlated principal components. However, "no correlation" does not mean "independence" except in the special case of Gaussian distributions. Independent component analysis (ICA) is a technique of decomposing signals into independent components. Information geometry, in particular semi-parametrics, plays a fundamental role in this. It has stimulated the rise of new techniques of positive matrix decomposition and sparse component analysis, which we also touch upon. The optimization problem under convex constraints and a game theory approach are briefly discussed in this chapter from the information geometry point of view. The Hyvärinen scoring method shows an attractive direction to be studied further from information geometry.

13.1 Principal Component Analysis

13.1.1 Eigenvalue Analysis

Let $\boldsymbol{x}$ be a vector random variable, which has already been preprocessed such that its expectation is 0,

$$E[\boldsymbol{x}] = 0. \tag{13.1}$$

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_13

Then, its covariance matrix is

$$\mathbf{V}_X = E\left[\boldsymbol{x}\boldsymbol{x}^T\right]. \tag{13.2}$$

If we transform $\boldsymbol{x}$ into $\boldsymbol{s}$ by using an orthogonal matrix $\mathbf{O}$,

$$\boldsymbol{s} = \mathbf{O}^T\boldsymbol{x}, \tag{13.3}$$

the covariance matrix of $\boldsymbol{s}$ is given by

$$\mathbf{V}_S = E\left[\boldsymbol{s}\boldsymbol{s}^T\right] = \mathbf{O}^T\mathbf{V}_X\mathbf{O}. \tag{13.4}$$

Let us consider the eigenvalue problem of $\mathbf{V}_X$,

$$\mathbf{V}_X\boldsymbol{o} = \lambda\boldsymbol{o}. \tag{13.5}$$

Then, we have n eigenvalues $\lambda_1, \ldots, \lambda_n$, $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$ and corresponding n unit eigenvectors $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n$, where we assume that there are no multiple eigenvalues. (When there exist multiple eigenvalues, rotational indefiniteness appears. We do not treat such a case here.) Let $\mathbf{O}$ be the orthogonal matrix consisting of the eigenvectors

$$\mathbf{O} = [\boldsymbol{o}_1 \ldots \boldsymbol{o}_n]. \tag{13.6}$$

Then, $\mathbf{V}_S$ is a diagonal matrix

$$V_S = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \tag{13.7}$$

and the components of $\boldsymbol{s}$ are uncorrelated,

$$E\left[s_i s_j\right] = 0, \quad i \neq j. \tag{13.8}$$

13.1.2 Principal Components, Minor Components and Whitening

Signal $\boldsymbol{x}$ is decomposed into a sum of uncorrelated components as

$$\boldsymbol{x} = \sum_{i=1}^{n} s_i \boldsymbol{o}_i. \tag{13.9}$$

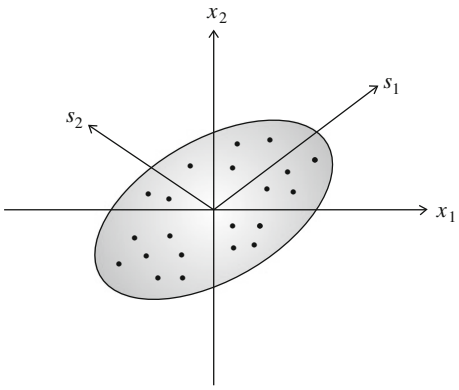


Fig. 13.1 Principal components $s_1, s_2, \ldots$

Since the variance of s_i is λ_i, s_1 has the largest magnitude on average, s_2 the second, and finally s_n has the smallest magnitude. See Fig. 13.1. We call s_1 the (first) principal component of $\boldsymbol{x}$, which is obtained by projecting $\boldsymbol{x}$ to $\boldsymbol{o}_1$. The first k largest components are given by $s_1, \ldots, s_k$. We call the subspace spanned by k eigenvectors $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_k$ the k-dimensional principal subspace. The vector

$$\tilde{\boldsymbol{x}} = \sum_{i=1}^{k} s_i \boldsymbol{o}_i \tag{13.10}$$

is the projection of $\boldsymbol{x}$ to the principal subspace.

The dimensions of $\boldsymbol{x}$ are reduced by the projection, keeping the resultant vector as close to the original one as possible in the sense that the magnitude of the lost part

$$L = \frac{1}{2}\mathrm{E}\left[\left|\boldsymbol{x} - \sum_{i=1}^{k} s_i \boldsymbol{o}_i\right|^2\right] \tag{13.11}$$

is minimized. So the principal components are used for approximating $\boldsymbol{x}$ with a small number of components, reducing the dimensions.

Similarly, the k minor components are given by $s_{n-k+1}, \ldots, s_n$, which are projections of $\boldsymbol{x}$ to $\boldsymbol{o}_i, i = n-k+1, \ldots, n$. The subspace spanned by $\boldsymbol{o}_{n-k+1}, \ldots, \boldsymbol{o}_n$, is called the k-dimensional minor subspace. The projection of $\boldsymbol{x}$ to the minor subspace is given by

$$\tilde{\tilde{\boldsymbol{x}}} = \sum_{i=n-k+1}^{n} s_i \boldsymbol{o}_i. \tag{13.12}$$

This is the maximizer of

$$L = \frac{1}{2}\mathrm{E}\left[\left|\boldsymbol{x} - \sum_{i=1}^{k} s_i \boldsymbol{o}_i\right|^2\right]. \tag{13.13}$$

Note that the minor components of $\mathbf{V}_X$ are the principal components of $\mathbf{V}_X^{-1}$, because the eigenvalues of $\mathbf{V}_X^{-1}$ are $1/\lambda_n, 1/\lambda_{n-1}, \ldots, 1/\lambda_1$. The eigenvectors of $\mathbf{V}_X^{-1}$ are the same as those of $\mathbf{V}_X$, but the order is reversed as $\boldsymbol{o}_n, \ldots, \boldsymbol{o}_1$.

Let us rescale the magnitudes of n eigenvectors to give a new set of basis vectors

$$\tilde{\boldsymbol{o}}_i = \sqrt{\lambda_i}\boldsymbol{o}_i, \quad i = 1, \ldots, n. \tag{13.14}$$

Then, $\boldsymbol{x}$ is written in the new basis as

$$\boldsymbol{x} = \sum \tilde{s}_i \tilde{\boldsymbol{o}}_i, \tag{13.15}$$

where

$$\tilde{s}_i = \frac{1}{\sqrt{\lambda_i}} s_i, \tag{13.16}$$

so that

$$\mathrm{E}\left[\tilde{s}_i \tilde{s}_j\right] = \delta_{ij}. \tag{13.17}$$

This implies that the covariance matrix of $\tilde{\boldsymbol{s}}$ is the identity matrix

$$\mathbf{V}_{\tilde{S}} = \mathbf{I}. \tag{13.18}$$

The transformation of $\boldsymbol{x}$ to $\tilde{\boldsymbol{s}}$ is called whitening of $\boldsymbol{x}$. This naming originates from the fact that, when we deal with time series $x(t), t = 1, 2, 3, \ldots$, the transformation (13.15) changes the time series $x(t)$ into white noise series $\tilde{s}(t)$.

Since $\mathbf{V}_{\tilde{S}}$ is the identity matrix, it is invariant if we further transform $\tilde{\boldsymbol{s}}$ by using an arbitrary orthogonal matrix $\mathbf{U}$ as

$$\tilde{\tilde{\boldsymbol{s}}} = \mathbf{U}\tilde{\boldsymbol{s}}. \tag{13.19}$$

Hence, whitening is not unique and there remains the indefiniteness of rotation, i.e., a further transformation by $\mathbf{U}$. In factor analysis, this fact is known as the indefiniteness of rotation. In order to dissolve the indefiniteness, we need to use higher-order statistics by assuming that the signals are not Gaussian. This is the motivation for discussing independent component analysis (ICA) in the next section.

13.1.3 Dynamics of Learning of Principal and Minor Components

When N examples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are observed as data D, we estimate the covariance matrix by

$$\hat{\mathbf{V}}_X = \frac{1}{N} \sum \boldsymbol{x}_i^T \boldsymbol{x}_i \tag{13.20}$$

and find the principal components by calculating its eigenvalues and eigenvectors. When examples are given one by one, we use a learning algorithm. We begin with a simple case of deriving the first principal component $\boldsymbol{o}_1$. Let $\boldsymbol{w}$ be the candidate of the first principal eigenvector, satisfying

$$|\boldsymbol{w}|^2 = 1. \tag{13.21}$$

Let

$$y = \boldsymbol{w} \cdot \boldsymbol{x} \tag{13.22}$$

be the projection of $\boldsymbol{x}$ to $\boldsymbol{w}$. Then the loss function to be minimized is

$$L = \frac{1}{2} |\boldsymbol{x} - y\boldsymbol{w}|^2 \tag{13.23}$$

under the constraint (13.21). By using the Lagrangian multiplier, the stochastic gradient method of obtaining the principal component is given by

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + y(t)\boldsymbol{x}(t) - \{y(t)\}^2 \boldsymbol{w}(t). \tag{13.24}$$

This was derived by Amari (1977) as a special case of neural learning, because the relation (13.22) is regarded as the output of a linear neuron. The same algorithm was discovered by Oja (1982) and was generalized to obtain the k-dimensional principal subspace (Oja 1992).

Let $\mathbf{W}$ be an $n \times k$ matrix consisting of k orthogonal unit column vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$,

$$\mathbf{W} = [\boldsymbol{w}_1 \boldsymbol{w}_2 \ldots \boldsymbol{w}_k], \tag{13.25}$$

satisfying

$$\mathbf{W}^T \mathbf{W} = \mathbf{I}_k, \tag{13.26}$$

where $\mathbf{I}_k$ is the $k \times k$ unit matrix. The set of all such matrices forms a manifold $S_{n,k}$, called the Stiefel manifold. The projection of $\boldsymbol{x}$ to the subspace spanned by $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ is

$$\tilde{\boldsymbol{x}} = \mathbf{W}\mathbf{W}^T \boldsymbol{x} = \sum y_i \boldsymbol{w}_i, \tag{13.27}$$

where

$$y_i = \boldsymbol{w}_i \cdot \boldsymbol{x}. \tag{13.28}$$

For obtaining the k-dimensional principal subspace spanned by the column vectors of $\mathbf{W}$, the loss function to be minimized is

$$L(\mathbf{W}) = \frac{1}{2}\mathrm{E}\left[\left|\boldsymbol{x} - \mathbf{W}\mathbf{W}^T\boldsymbol{x}\right|^2\right]. \tag{13.29}$$

The gradient descent learning equation for $\mathbf{W}$ is

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + y_i(t)\boldsymbol{x}(t) - \sum_j y_i(t)y_j(t)\boldsymbol{w}(t), \quad i = 1, \ldots, k. \tag{13.30}$$

Its averaged version in continuous time is

$$\dot{\mathbf{W}}(t) = \mathbf{V}_X\mathbf{W}(t) - \mathbf{W}\mathbf{W}^T\mathbf{V}_X\mathbf{W}, \tag{13.31}$$

where $\dot{}$ denotes the time derivative d/dt.

The solution $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ of learning Eqs. (13.30) or (13.31) converges to the subspace spanned by k principal eigenvectors. However, each $\boldsymbol{w}_i$ does not correspond to the eigenvectors $\boldsymbol{o}_i$, although the principal subspace is spanned by $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$.

In order to obtain the k principal eigenvectors, Xu (1993) introduced a diagonal matrix

$$\mathbf{D} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_k \end{bmatrix}, \tag{13.32}$$

satisfying $d_1 > \ldots > d_k$ and modified (13.31) as

$$\dot{\mathbf{W}}(t) = \mathbf{V}_X\mathbf{W}(t)\mathbf{D} - \mathbf{W}\,\mathbf{D}\mathbf{W}^T\mathbf{V}_X\mathbf{W}. \tag{13.33}$$

This algorithm gives the principal eigenvectors $\boldsymbol{w}_i = \boldsymbol{o}_i$.

It appears that a similar algorithm would be applicable to the problem of obtaining the minor component subspace. We need to find $\mathbf{W}$ that maximizes (13.29). If we use gradient ascent instead of gradient descent, the algorithm would be

$$\dot{\mathbf{W}} = -\mathbf{V}_X\mathbf{W} + \mathbf{W}\mathbf{W}^T\mathbf{V}_X\mathbf{W}. \tag{13.34}$$

However, this does not work. Why (13.34) does not work had been a puzzle.

Both algorithms (13.31) and (13.34) work well when $\mathbf{W}$ is limited in the Stiefel manifold $S_{n,k}$. The manifold $S_{n,k}$ is a submanifold of $M_{n,k}$, which is the manifold of all $n \times k$ matrices. When we solve (13.34) or its stochastic version numerically, $\mathbf{W}(t)$ deviates from $S_{n,k}$ because of numerical errors. Algorithms (13.31) and (13.34) define flows $\dot{\mathbf{M}}$ in the entire $M_{n,k}$, where $\mathbf{M}(t) \in M_{n,k}$, when $\mathbf{W}$ is replaced by $\mathbf{M}$. The

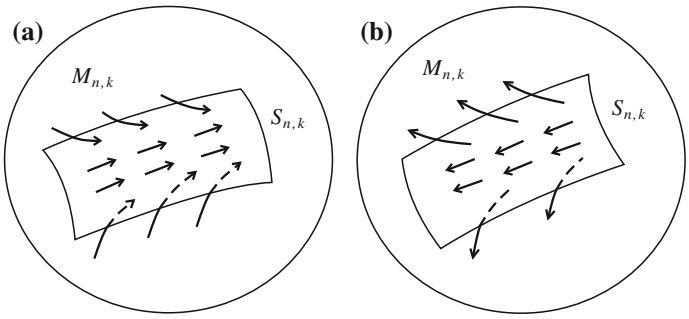


Fig. 13.2 Flow in **a** principal subspace, **b** minor subspace

flow is closed in $S_{n,k}$, that is, $\dot{\mathbf{M}} \in S_{n,k}$ when $\mathbf{M} \in S_{n,k}$. $S_{n,k}$ is a stable submanifold of the flow (13.31) in $M_{n,k}$. Hence, when a small fluctuation occurs in $\mathbf{W}$ and it deviates from $S_{n,k}$ into $M_{n,k}$, it automatically returns to $S_{n,k}$ (Fig. 13.2a). However, in the case of the flow (13.34) for minor components, $S_{n,k}$ is not stable in $M_{n,k}$ and $\mathbf{W}$ leaves $S_{n,k}$ due to the small deviation (Fig. 13.2b). This is the reason why the algorithm (13.34) does not work.

Consider two modified differential equations in $M_{n,k}$ due to Chen et al. (1998),

$$\dot{\mathbf{M}}(t) = \mathbf{V}_X \mathbf{M}\mathbf{M}^T\mathbf{M} - \mathbf{M}\mathbf{M}^T\mathbf{V}_X\mathbf{M}, \tag{13.35}$$

$$\dot{\mathbf{M}}(t) = -\mathbf{V}_X \mathbf{M}\mathbf{M}^T\mathbf{M} - \mathbf{M}\mathbf{M}^T\mathbf{V}_X\mathbf{M}. \tag{13.36}$$

Then, we can prove that the submanifold $S_{n,k}$ is neutrally stable with regard to both of the flows. Therefore, we can use (13.35) to obtain the principal components and (13.36) to obtain the minimal components. The on-line learning versions of (13.35) and (13.36) are

$$\dot{\boldsymbol{m}}_i(t) = \pm \sum_j \left\{ \left(\boldsymbol{m}_i \cdot \boldsymbol{m}_j\right) \left(\boldsymbol{m}_j \cdot \boldsymbol{x}\right) \boldsymbol{x} - \left(\boldsymbol{m}_i \cdot \boldsymbol{x}\right) \left(\boldsymbol{m}_j \cdot \boldsymbol{x}\right) \boldsymbol{m}_j \right\}, \tag{13.37}$$

where $\boldsymbol{m}_i$ is the ith column vector of $\mathbf{M}$.

The dynamics (13.35) and (13.36) possess interesting invariants. Let

$$\mathbf{M}(t) = \mathbf{W}(t)\mathbf{D}(t)\mathbf{U}(t) \tag{13.38}$$

be the singular decomposition of $\mathbf{M}(t)$, where $\mathbf{W}(t)$ is an element of $S_{n,k}$ consisting of k orthogonal unit vectors, $\mathbf{U}(t)$ is a $k \times k$ orthogonal matrix and $\mathbf{D}$ is a $k \times k$ diagonal matrix with diagonal entries $d_1, \ldots, d_k$.

Lemma 13.1 *(1)* $\boldsymbol{M}^T(t)\boldsymbol{M}(t)$ *is an invariant of (13.35) and (13.36),* $\boldsymbol{M}^T(t)\boldsymbol{M}(t) = \boldsymbol{M}^T(0)\boldsymbol{M}(0)$.

(2) $\boldsymbol{D}(t)$ *is an invariant of (13.35) and (13.36),* $\boldsymbol{D}(t) = \boldsymbol{D}(0)$.

(3) $\boldsymbol{U}(t)$ *is an invariant of (13.35) and (13.36),* $\boldsymbol{U}(t) = \boldsymbol{U}(0)$.

We omit the proof (see Chen et al. 1998). We immediately obtain the algorithm of Xu (1993) by using an initial condition $\mathbf{D}(0) = \text{diag}\{d_1, \ldots, d_k\}$ and rewriting (13.35) in terms of $\mathbf{W}(t)$. When $k = n$, both (13.35) and (13.36) give the Brockett flow (Brockett 1991), where the cost function is

$$L(\mathbf{M}) = \pm \text{tr}\left(\mathbf{M}\mathbf{M}^T\mathbf{V}\right). \tag{13.39}$$

This is the natural gradient flow in the manifold of the orthogonal matrices (see Chen et al. 1998).

Since $S_{n,k}$ is neutrally stable in Eqs. (13.35) and (13.36), numerical errors may accumulate. Chen and Amari (2001) proposed the following equations

$$\dot{\mathbf{M}}(t) = \left(\mathbf{V}_X\mathbf{M}\mathbf{M}^T\mathbf{M} - \mathbf{M}\mathbf{M}^T\mathbf{V}_X\mathbf{M}\right) + \mathbf{M}\left(\mathbf{D}^2 - \mathbf{M}^T\mathbf{M}\right), \tag{13.40}$$

$$\dot{\mathbf{M}}(t) = -\left(\mathbf{V}_X\mathbf{M}\mathbf{M}^T\mathbf{M} - \mathbf{M}\mathbf{M}^T\mathbf{V}_X\mathbf{M}\right) + \mathbf{M}\left(\mathbf{D}^2 - \mathbf{M}^T\mathbf{M}\right), \tag{13.41}$$

where $\mathbf{D}$ is a positive diagonal matrix related to the initial value of $\mathbf{M}$,

$$\mathbf{M}(0) = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_k \\ & 0 & \\ & \vdots & \\ & 0 & \end{bmatrix}. \tag{13.42}$$

$S_{n,k}$ is stable both under (13.40) and (13.41), so both the principal eigenvectors and minor eigenvectors are extracted stably by the respective equations, which differ only in signature.

13.2 Independent Component Analysis

Consider the problem of decomposing vector random variable $\boldsymbol{x}$ into n independent components,

$$\boldsymbol{x} = \sum_{i=1}^{n} s_i \boldsymbol{a}_i, \tag{13.43}$$

such that s_i are independent random variables and $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$ is a new set of basis vectors. We consider the case where n independent component signals $s_1, \ldots, s_n$ exist under an adequate basis. When $\boldsymbol{x}$ is Gaussian, PCA is successful for performing this job. However, there are infinitely many such decompositions due to rotational indefiniteness, as stated in the previous section. Moreover, when $\boldsymbol{x}$ is non-Gaussian,

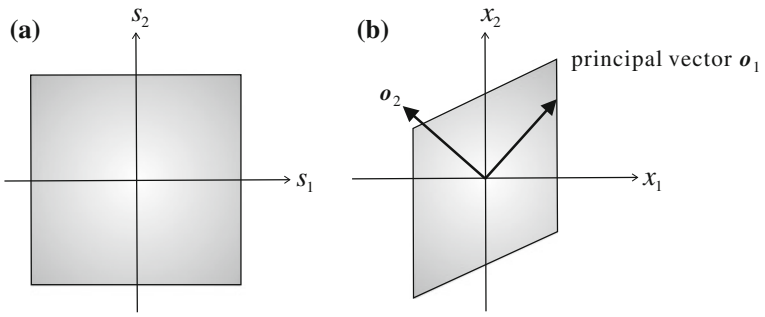


Fig. 13.3 **a** Uniform distribution $P(s)$; **b** Its linear transformation $p(x)$

PCA does not work for this purpose. This is because, even if no correlations exist among n signals $s_1, \ldots, s_n$, this does not imply that they are independent.

We give a simple example. Let s_1 and s_2 be two independent signals, where both s_1 and s_2 are subject to the uniform distribution over $[-0.5, 0.5]$. They are distributed uniformly over the square (Fig. 13.3a). We construct their mixtures $\boldsymbol{x} = (x_1, x_2)^T$ by

$$x_1 = s_1, \tag{13.44}$$

$$x_2 = s_1 + 2s_2. \tag{13.45}$$

Then, $\boldsymbol{x}$ is uniformly distributed in a parallelepiped (see Fig. 13.3b). Its covariance matrix is

$$\mathbf{V}_X = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}, \tag{13.46}$$

of which the eigenvectors are different from the original s_1 and s_2 axes. The PCA solution gives non-correlated components but they are not independent. So we need other methods to decompose $\boldsymbol{x}$ into independent components. Higher-order statistics beyond the covariance is useful for solving the problem.

An illustrative example of ICA is the cocktail party problem. There are n persons in a cocktail party room who are speaking independently. Let $s_i(t)$ be the voice of person i at time t. m microphones are placed in the party room, so that each microphone records a mixture of voices of n persons. Let $x_j(t)$ be the sound recorded by microphone j at time t. See Fig. 13.4. They are written as

$$x_j(t) = \sum A_{ji} s_i(t), \quad \boldsymbol{x}(t) = \mathbf{A}\boldsymbol{s}(t), \tag{13.47}$$

where A_{ji} is a coefficient of mixing depending on the distance between person i and microphone j. The problem is to recover the sounds $\boldsymbol{s}(1), \boldsymbol{s}(2), \ldots$ of all the persons from the recorded mixtures $\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots$, without any knowledge of A_{ji}. Here we assume that the numbers of persons and microphones are the same, $n = m$. When $n < m$, we first apply PCA to $\boldsymbol{x}$, projecting it to the n-dimensional principal

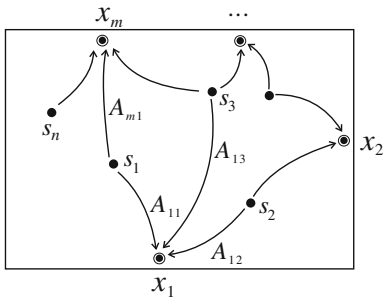


Fig. 13.4 n persons and m microphones in a room

subspace. Then, the problem reduces to the case of $m = n$. When $m < n$, we need techniques of sparse signal processing.

We assume that $\mathbf{A}$ is a regular $n \times n$ matrix. When $\mathbf{A}$ is known, the problem is trivially solved by

$$\boldsymbol{y}(t) = \mathbf{A}^{-1}\boldsymbol{x}(t), \tag{13.48}$$

and $\boldsymbol{y}(t)$ is equal to the original $\boldsymbol{s}(t)$. However, $\mathbf{A}$ or $\mathbf{A}^{-1}$ is unknown. We transform $\boldsymbol{x}$ by using a matrix $\mathbf{W}$ as

$$\boldsymbol{y}(t) = \mathbf{W}\boldsymbol{x}(t), \tag{13.49}$$

and check if n components of $\boldsymbol{y}$ in time series $\boldsymbol{y}(1), \ldots, \boldsymbol{y}(T)$ are independently distributed or not. If they are not independent, we modify $\mathbf{W}$ such that the degree of non-independence decreases. To this end, we need to define the degree of non-independence of n random variables $y_1, \ldots, y_n$. Since it is a function of $\mathbf{W}$, we can apply the stochastic gradient descent or the natural gradient descent method to obtain $\mathbf{W}$ that recovers the independent signals.

Before defining the degree of non-independence, we note the indefiniteness of the solution. As is known, the independent components are recovered only when all the components of $\boldsymbol{s}$ except for one are non-Gaussian. Further, the order of signals $s_1, \ldots, s_n$ is not recovered, since any permutation of n independent signals keeps their independence. Moreover, the magnitude of s_i is not recovered, because, when $s_1, \ldots, s_n$ are independent, $c_1 s_1, c_2 s_2, \ldots, c_n s_n$ are independent for any constants $c_1, \ldots, c_n$. Hence, the independent components are recovered to within the scales and order.

We formulate the problem mathematically. Let $k_i\,(s_i)$ be the probability density function of the ith independent component s_i, where we assume that

$$E\,[s_i] = 0. \tag{13.50}$$

Then, the joint probability density of $\boldsymbol{s}$ is

$$k(\boldsymbol{s}) = \prod k_i\,(s_i)\,. \tag{13.51}$$

For $\boldsymbol{y}$ determined from (13.49), the joint probability density is written as

$$p_Y(\boldsymbol{y}; \mathbf{W}) = |\mathbf{WA}|^{-1} k\left(\mathbf{A}^{-1}\mathbf{W}^{-1}\boldsymbol{y}\right). \tag{13.52}$$

Here, we used the general formula that probability density function $p(\boldsymbol{x})$ changes to

$$p_Y(\boldsymbol{y}) = \left|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|^{-1} p(\boldsymbol{x}), \tag{13.53}$$

when $\boldsymbol{x}$ is transformed to $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}). \tag{13.54}$$

The KL-divergence from $p_Y(\boldsymbol{y})$ to $k(\boldsymbol{y})$,

$$D_{KL}[p_Y : k] = \int p_Y(\boldsymbol{y}) \log \frac{p_Y(\boldsymbol{y})}{k(\boldsymbol{y})} d\boldsymbol{y}, \tag{13.55}$$

would be used as a degree of non-independence. This would be a good choice if we knew $k(\boldsymbol{s})$. However, we do not know $k(\boldsymbol{s})$ and what we know is only the fact that $k(\boldsymbol{s})$ is decomposed into the product of unknown $k_i(s_i)$. We use n arbitrary independent distributions,

$$q(\boldsymbol{y}) = \prod q_i(y_i) \tag{13.56}$$

and define

$$D[p_Y : q] = \int p_Y(\boldsymbol{y}) \log \frac{p_Y(\boldsymbol{y})}{q(\boldsymbol{y})} d\boldsymbol{y} \tag{13.57}$$

as a function to show the degree of non-independence. This choice is reasonable as follows.

We consider the manifold of all the probability distribution

$$S = \{p(\boldsymbol{y})\} \tag{13.58}$$

to understand the situation geometrically. We define the submanifold S_I of all the independent distributions

$$S_I = \left\{ p(\boldsymbol{y}) \,\middle|\, p(\boldsymbol{y}) = \prod p_i(y_i),\ p_i \text{ are arbitrary density functions} \right\}, \tag{13.59}$$

which is an e-flat submanifold of S. It includes both $k(\boldsymbol{y})$ and $q(\boldsymbol{y})$. Another submanifold we consider is

$$S_W = \{p_Y(\boldsymbol{y}; \mathbf{W})\}, \tag{13.60}$$

which is parameterized by $\mathbf{W}$. For each $\mathbf{W}$, we have a distribution $p_Y(\boldsymbol{y}; \mathbf{W})$ given by the transformation of $\boldsymbol{y} = \mathbf{W}\boldsymbol{x}$. It is not a flat submanifold. See Fig. 13.5.

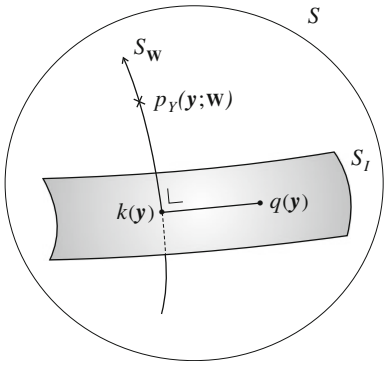


Fig. 13.5 S_I e-flat submanifold of independent distributions, S_N submanifold generated by $\boldsymbol{W}$. They are orthogonal

We use a loss function

$$L_k(\mathbf{W}) = D_{KL}\left[p_Y(\boldsymbol{y};\mathbf{W}) : k(\boldsymbol{y})\right], \tag{13.61}$$

when we know $k(\boldsymbol{s})$. S_W and S_I intersect at $\mathbf{W} = \mathbf{A}^{-1}$ and the loss function L is 0 at this point. However, we do not know $k(\boldsymbol{s})$, so we use

$$L(\boldsymbol{W}) = D_{KL}\left[p_Y(\boldsymbol{y};\mathbf{W}) : q(\boldsymbol{y})\right] \tag{13.62}$$

by using an adequately chosen q (Bell and Sejnowski 1995). We can show that S_W and S_I intersect orthogonally. In spite of this, we cannot apply the Pythagorean theorem, because S_W is not m-flat. However, because of the orthogonality, we show that $\mathbf{W} = \mathbf{A}^{-1}$ is a critical point of L. It is a local minimum, saddle or local maximum depending on the choice of q. The stability of the critical point depends on q and the m-embedding curvature of S_W at $q = k$. When q is close to k, $\mathbf{A}^{-1}$ is certainly a global minimum. We neglect the indefiniteness of $\mathbf{W}$ concerning scales and permutations in the present discussions, but the situation is the same for all equivalent $\mathbf{W}$.

We should remark that there are many loss functions other than (13.62). By mixing independent $s_1, \ldots, s_n$, the central limit theorem suggests that the distribution of $\boldsymbol{x}$ approaches a jointly Gaussian distribution. Hence, the degree of non-Gaussianity can be used as a loss function. The higher-order cumulants of $\boldsymbol{y}$ vanish when $\boldsymbol{y}$ is Gaussian, so that the sum of the absolute values of the third- and fourth-order cumulants play the role of a loss function. We may use other measures of non-Gaussianity as a loss function. See Hyvärinen et al. (2001) and Cichocki and Amari (2002). The following analysis is common to all such loss functions.

The stochastic descent on-line learning algorithm is given by

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \varepsilon \frac{\partial}{\partial \mathbf{W}} D_{KL}\left[p_Y(\boldsymbol{y}_t) : q(\boldsymbol{y}_t)\right]. \tag{13.63}$$

The loss function is written as

$$D_{KL}\left[p_Y(\boldsymbol{y}) : q(\boldsymbol{y})\right] = \int p_Y(\boldsymbol{y}) \log \frac{p_Y(\boldsymbol{y})}{q(\boldsymbol{y})} d\boldsymbol{y} = -H(Y) - \mathrm{E}\left[\log q(\boldsymbol{y})\right], \tag{13.64}$$

where

$$H(Y) = -\int p_W(\boldsymbol{y}) \log p_W(\boldsymbol{y}) d\boldsymbol{y} \tag{13.65}$$

is the entropy of $\boldsymbol{y}$ expressed as a function of $\mathbf{W}$. We see

$$H(Y) = H(X) + \log |\mathbf{W}|. \tag{13.66}$$

In order to calculate the gradient of the instantaneous loss

$$l(\boldsymbol{y}, \mathbf{W}) = -\log |\mathbf{W}| - \log q(\boldsymbol{y}; \mathbf{W}) \tag{13.67}$$

with respect to $\mathbf{W}$, where $H(X)$ is neglected because it does not depend on $\mathbf{W}$, we consider a small change of $l(\boldsymbol{y}, \mathbf{W})$ due to a small change of $\mathbf{W}$, from $\mathbf{W}$ to $\mathbf{W} + d\mathbf{W}$.

We have

$$d \log |\mathbf{W}| = \log |\mathbf{W} + d\mathbf{W}| - \log |\mathbf{W}| = \mathrm{tr}\left(d\mathbf{W}\mathbf{W}^{-1}\right). \tag{13.68}$$

Similarly, we have

$$d \log q_i(\boldsymbol{y}) = \frac{q_i'(y_i)}{q_i(y_i)} dy_i. \tag{13.69}$$

We put

$$\varphi_i(y_i) = \frac{-q_i'(y_i)}{q_i(y_i)}. \tag{13.70}$$

Further, from

$$d\boldsymbol{y} = (d\mathbf{W})\boldsymbol{x}, \tag{13.71}$$

we have, for $\boldsymbol{\varphi}(\boldsymbol{y}) = [\varphi_1(y_1), \ldots, \varphi_n(y_n)]^T$,

$$d \log q(\boldsymbol{y}) = -\boldsymbol{\varphi}(\boldsymbol{y})^T d\mathbf{W}\mathbf{W}^{-1}\boldsymbol{y}. \tag{13.72}$$

Hence, we have

$$dl(\boldsymbol{y}, \mathbf{W}) = -\mathrm{tr}\left(d\mathbf{W}\mathbf{W}^{-1}\right) + \boldsymbol{\varphi}(\boldsymbol{y})^T d\mathbf{W}\mathbf{W}^{-1}\boldsymbol{y}, \tag{13.73}$$

from which the gradient of the instantaneous loss l with respect to $\mathbf{W}$, $\partial D/\partial W_{ij}$, is calculated by using the component form.

In order to obtain the natural gradient, we need to introduce a Riemannian metric in the manifold $Gl(n)$ of matrices. Let $d\mathbf{W}$ be a small line element, which is written as

$$d\mathbf{W} = \sum dW_{ij}\mathbf{E}_{ij}, \tag{13.74}$$

where $\mathbf{E}_{ij}$ is a matrix whose (i, j) element is 1 and all the other elements are 0. They form a basis in the tangent space. We consider the Lie group structure of $Gl(n)$. $\mathbf{W}$ is mapped to the identity matrix by multiplying $\mathbf{W}^{-1}$ from the right,

$$\mathbf{W}\mathbf{W}^{-1} = \mathbf{I}. \tag{13.75}$$

We also map a nearby point $\mathbf{W} + d\mathbf{W}$ by multiplying $\mathbf{W}^{-1}$ from the right, giving

$$(\mathbf{W} + d\mathbf{W})\mathbf{W}^{-1} = \mathbf{I} + d\mathbf{W}\mathbf{W}^{-1}. \tag{13.76}$$

Hence, a small line element $d\mathbf{W}$ in the tangent space of $Gl(n)$ at $\mathbf{W}$ is mapped to

$$d\mathbf{X} = d\mathbf{W}\mathbf{W}^{-1} \tag{13.77}$$

in the tangent space at $\mathbf{I}$. See Fig. 13.6.

We define the magnitude of $d\mathbf{X}$ at $\mathbf{I}$ simply by

$$\langle d\mathbf{X}, d\mathbf{X}\rangle = \text{tr}\ \left(d\mathbf{X}d\mathbf{X}^T\right) = \sum \left(d\mathbf{X}_{ij}\right)^2. \tag{13.78}$$

A Riemannian metric is defined by defining the magnitude of $d\mathbf{W}$ at the tangent space at $\mathbf{W}$. We use the Lie group invariance such that the magnitude does not change by

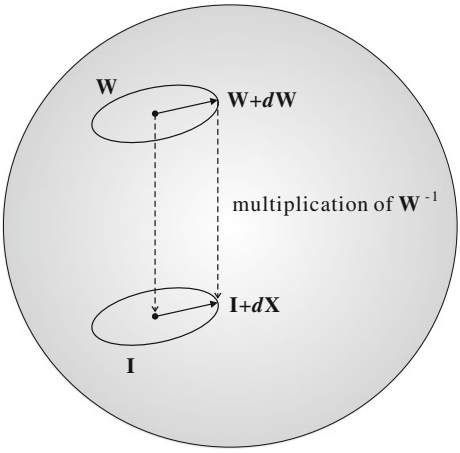


Fig. 13.6 Mapping of $T_{\mathbf{W}}$ to $T_{\mathbf{I}}$

the right multiplication of $\mathbf{W}^{-1}$. Then, the magnitude of $d\mathbf{W}$ is defined by that of the corresponding $d\mathbf{X}$,

$$\langle d\mathbf{W}, d\mathbf{W}\rangle_{\mathbf{W}} = \langle d\mathbf{X}, d\mathbf{X}\rangle_{\mathbf{I}}. \tag{13.79}$$

This is rewritten as

$$\langle d\mathbf{W}, d\mathbf{W}\rangle = \mathrm{tr}\left\{d\mathbf{W}\mathbf{W}^{-1}\left(d\mathbf{W}\mathbf{W}^{-1}\right)^T\right\}, \tag{13.80}$$

which is called the Killing metric. The length of a tangent vector is invariant by multiplying a matrix from the right.

One may wonder if there is a coordinate transformation of $\mathbf{W}$,

$$\mathbf{X} = \mathbf{X}(\mathbf{W}) \tag{13.81}$$

from which $d\mathbf{X}$ is derived by

$$d\mathbf{X} = \frac{\partial \mathbf{X}}{\partial \mathbf{W}} \cdot d\mathbf{W}. \tag{13.82}$$

Unfortunately, there is no such coordinate transformation. We can define $d\mathbf{X}$ but it is not integrable, that is, the integration of $d\mathbf{X}$

$$\mathbf{X}\left(\mathbf{W}'\right) - \mathbf{X}(\mathbf{W}) = \int_{\mathbf{W}}^{\mathbf{W}'} d\mathbf{X} \tag{13.83}$$

from $\mathbf{W}$ to $\mathbf{W}'$ depends on the path connecting $\mathbf{W}$ and $\mathbf{W}'$. So we do not have a coordinate system $\mathbf{X}$ in $Gl(n)$ such that dX_{ij} are increments along new coordinate curves. Such virtual coordinates $\mathbf{X}$ are called a non-holonomic coordinate system, in which only $d\mathbf{X}$ is defined. This non-holonomic basis of the tangent space is convenient for introducing a Riemannian metric to $Gl(n)$ and defining the natural gradient.

The small change (13.73) of l is written in terms of $d\mathbf{X}$ as

$$dl = -tr(d\mathbf{X}) + \boldsymbol{\varphi}(\boldsymbol{y})^T d\mathbf{X}\boldsymbol{y}. \tag{13.84}$$

This is written in the components as

$$\frac{dl}{dX_{ij}} = -\delta_{ij} + \varphi_i\left(y_i\right) y_j. \tag{13.85}$$

Since the inner product $\langle d\mathbf{X}, d\mathbf{X}\rangle$ is Euclidean, as is seen from (13.78), it is the natural gradient due to the Killing metric. The increment of $\mathbf{W}$ is written as

$$\Delta X_{ij} = -\varepsilon\left\{\delta_{ij} - \varphi_i\left(y_i\right) y_j\right\}, \quad \nabla_X l = -\varepsilon\left(\mathbf{I} - \boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T\right) \tag{13.86}$$

by using $d\mathbf{X}$, where ∇_X is the gradient with respect to $\mathbf{X}$. By using (13.77), this is rewritten in terms of the gradient with respect to $\mathbf{W}$ as

$$\nabla_W l = -\varepsilon \left(I - \varphi(\boldsymbol{y}) \boldsymbol{y}^T \right) \mathbf{W}. \tag{13.87}$$

Because of this, the natural gradient has an invariant property that the convergence of learning dynamics is the same whatever the true $\mathbf{W}$ is. The stability does not depend on $\mathbf{W}$, either. These are desirable properties given by Cardoso and Laheld (1996) and Amari et al. (1996).

13.2.3 Estimating Function of ICA: Semiparametric Approach

The probability density function of observed $\boldsymbol{x}$ can be written as

$$p\left(\boldsymbol{x}, W, k\right) = \prod_i k_i \left(\sum_j W_{ij} x_j \right). \tag{13.88}$$

In this statistical model, the unknown parameters include not only $\mathbf{W}$ but also n functions $k_1\left(s_1\right), \ldots k_n\left(s_n\right)$, which are the probability densities of the independent source signals. The probability distribution of $\boldsymbol{x}$ is specified by $n \times n$ matrix $\mathbf{W}$, which are the parameters of interest to be estimated, and also by n functions $k_1\left(s_1\right), \ldots k_n\left(s_n\right)$, which are nuisance parameters of function-degrees of freedom. Therefore, ICA is a semi-parametric statistical problem (Amari and Cardoso 1997).

An estimation function is a matrix $\mathbf{F}(\boldsymbol{x}, \mathbf{W})$ which satisfies

$$\mathrm{E}_{\mathbf{W}}\left[\mathbf{F}\left(\boldsymbol{x}, \mathbf{W}'\right)\right] \begin{cases} = 0, \ \mathbf{W}' \approx \mathbf{W}, \\ \neq 0, \ \mathbf{W}' \not\approx \mathbf{W}. \end{cases} \tag{13.89}$$

Here, the expectation is taken with respect to $p(\boldsymbol{x}, \mathbf{W})$, and $\mathbf{W} \approx \mathbf{W}'$ implies that $\mathbf{W}$ and $\mathbf{W}'$ are equivalent to within the scales and permutations. The estimating equation is given by

$$\sum_{t=1}^{T} \mathbf{F}(\boldsymbol{y}(t)) = \sum_{t=1}^{T} \mathbf{F}\left\{\mathbf{W}\boldsymbol{x}(t)\right\} = 0. \tag{13.90}$$

A sequential estimation is realized by the learning equation

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \varepsilon_t \mathbf{F}\left(\boldsymbol{x}_t, \mathbf{W}_t\right), \tag{13.91}$$

which is expected converge to the solution of (13.90), although the convergence is not necessarily guaranteed.

Information geometry gives a general class of estimating functions. See Amari (1999) for details. Let $\varphi(\boldsymbol{y})$ be an arbitrary vector function of $\boldsymbol{y}$. Then, an effective

class of estimating functions is generated from

$$\mathbf{F}(\boldsymbol{x}, \mathbf{W}) = \mathbf{F}(\boldsymbol{y}) = \mathbf{I} - \boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T, \tag{13.92}$$

including arbitrary vector function $\boldsymbol{\varphi}$. Let $\mathbf{R}(\mathbf{W})$ be a linear reversible transformation of matrices acting on $\mathbf{F}$ as

$$\tilde{\mathbf{F}}(\boldsymbol{x}, \mathbf{W}) = \mathbf{R}(\mathbf{W})\mathbf{F}(\boldsymbol{x}, \mathbf{W}). \tag{13.93}$$

$\mathbf{R}$ is a tensor having four indices and written in the component form as

$$\tilde{F}_{ij}(\boldsymbol{x}, \mathbf{W}) = \sum_{k,l} R_{ij}{}^{kl} F_{kl}(\boldsymbol{x}, \mathbf{W}). \tag{13.94}$$

The estimating equation is the same for $\mathbf{F}$ and $\mathbf{RF}$, because

$$\sum_t \mathbf{RF}\left(\boldsymbol{x}(t), \mathbf{W}\right) = 0 \tag{13.95}$$

is equivalent to (13.90).

The on-line learning equation using $\mathbf{RF}$ is

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \varepsilon_t \mathbf{R}(\mathbf{W}_t)\mathbf{F}. \tag{13.96}$$

Although, the equilibrium point does not depend on $\mathbf{R}$, its stability depends on $\mathbf{R}$ and so does the speed of convergence. Therefore, we need to choose $\boldsymbol{\varphi}(\boldsymbol{y})$ and $\mathbf{R}(\mathbf{W})$ carefully.

Once $\boldsymbol{\varphi}(\boldsymbol{y})$ is chosen, the Newton method is applicable to solve the iterative procedure. From the estimating Eq. (13.90), we have

$$\sum_t \mathbf{F}\left(\boldsymbol{x}_t, \mathbf{W} + \Delta\mathbf{W}\right) = \sum \mathbf{F}\left(\boldsymbol{x}_t, \mathbf{W}\right) + \sum \frac{\partial \mathbf{F}}{\partial \mathbf{W}} \circ \Delta\mathbf{W} = 0, \tag{13.97}$$

where $\boldsymbol{x}_t = \boldsymbol{x}(t)$ and $\circ$ is used for taking the trace of matrix multiplication. Using

$$\Delta\mathbf{X}_t = \Delta\mathbf{W}\mathbf{W}_t^{-1}, \tag{13.98}$$

we define the operator

$$\mathbf{J} = \mathrm{E}\left[\frac{\partial \mathbf{F}}{\partial \mathbf{X}}\right] = \mathrm{E}\left[\frac{\partial \mathbf{F}}{\partial \mathbf{W}}\right]\mathbf{W}^T. \tag{13.99}$$

The Newton method is written as

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \varepsilon_t \mathbf{J}^{-1}\left(\mathbf{W}_t\right)\mathbf{F}\left(\boldsymbol{y}_t\right). \tag{13.100}$$

Therefore, the Newton method is derived by choosing $\mathbf{R}$ in the following way:

$$\mathbf{R}(\mathbf{W}) = \mathbf{J}^{-1}(\mathbf{W}). \tag{13.101}$$

The operator $\mathbf{J}$ is a fourth-order tensor, and we can calculate it explicitly, but it depends on the true $k(\boldsymbol{s})$, which we do not know.

An estimating function $\tilde{\mathbf{F}}(\boldsymbol{x}, \mathbf{W})$ is said to be standard when it satisfies

$$\tilde{\mathbf{J}} = \mathrm{E}\left[\frac{\partial \tilde{\mathbf{F}}}{\partial \mathbf{W}}\right]\mathbf{W}^T = \text{identity operator}. \tag{13.102}$$

Given an estimating function $\mathbf{F}$, we have its standard version by

$$\tilde{\mathbf{F}}(\boldsymbol{x}, \mathbf{W}) = \mathbf{J}^{-1}\mathbf{F}(\boldsymbol{x}, \mathbf{W}). \tag{13.103}$$

The learning equation using a standard estimating function corresponds to the Newton method. The Hyvärinen fast algorithm (Hyvärinen 2005) uses a standard estimating function.

Since the standard estimating function using $\boldsymbol{\varphi}(\boldsymbol{y})$ is written in the form of

$$\tilde{\mathbf{F}} = \mathbf{I} - \alpha\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T + \beta\boldsymbol{y}\boldsymbol{\varphi}^T(\boldsymbol{y}), \tag{13.104}$$

where α and β are adequate parameters, we can use an adaptive method of choosing them from the data. The separating $\mathbf{W}$ is stable when we use a standard estimating function, because the Newton method is applied.

One of surprising results is the following "super efficiency". We define the covariance of the recovered signal at t by

$$V_{ij}(t) = \mathrm{E}\left[y_i(t)y_j(t)\right], \quad i \neq j. \tag{13.105}$$

Then, it converges to 0 when the source separation is successful.

We have the following super efficiency results:

Theorem 13.1 *When*

$$\mathrm{E}\left[\varphi_i\left(s_i\right)\right] = 0, \tag{13.106}$$

by using the standard estimating function $\boldsymbol{F}$*, the covariances decrease in the order of* $1/t^2$ *for the natural gradient learning,*

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \frac{1}{t}\boldsymbol{F}\left(\boldsymbol{x}_t, \boldsymbol{W}_t\right) \tag{13.107}$$

and in the order of η^2 *when the learning constant* η *is fixed,*

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta\boldsymbol{F}\left(\boldsymbol{x}_t, \boldsymbol{W}_t\right). \tag{13.108}$$

The condition (13.106) is satisfied in the following two cases:

$$(1) \quad \varphi_i(s_i) = -\frac{d}{ds_i} \log k_i(s_i), \tag{13.109}$$

$$(2) \quad k_i(s_i) \text{ is an even function and } \varphi_i(s_i) \text{ is an odd function.} \tag{13.110}$$

See Amari (1999) for detailed discussions and proofs.

Remark When independent source signals $s_i(t)$ have temporal correlations such that

$$E[s_i(t)s_i(t-\tau)] = c_i(\tau), \tag{13.111}$$

which are not 0 for some $\tau > 0$, we can use this information even if we do not know $c_i(\tau)$ explicitly. The previous results are valid even in this case, but we have more efficient methods by taking the existence of temporal correlation into account. The joint diagonalization of the delayed covariance matrices is one good idea. See Cardoso and Souloumiac (1996). The method works well even when the source signals are Gaussian.

It is possible to develop a method of estimating functions even in this case. We obtain a general form of estimating functions, which includes arbitrary temporal filters to be applied to the observed signals $\boldsymbol{x}(t)$. The joint diagonalization is a special example of the estimating function method. See Amari (2000) for details.

13.3 Non-negative Matrix Factorization

Given a series of observed signals $\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)$, let us arrange all of them in an $n \times T$ matrix form,

$$\mathbf{X} = [\boldsymbol{x}(1) \ldots \boldsymbol{x}(T)]. \tag{13.112}$$

ICA searches for the basis vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$, which form an $n \times n$ mixing matrix

$$\mathbf{A} = [\boldsymbol{a}_1 \ldots \boldsymbol{a}_n] \tag{13.113}$$

and $\boldsymbol{x}$ is decomposed as

$$\boldsymbol{x}(t) = \mathbf{A}\boldsymbol{s}(t) = \sum s_i(t)\boldsymbol{a}_i, \tag{13.114}$$

such that $s_1, \ldots, s_n$ are independent. (13.114) is represented as

$$\mathbf{X} = \mathbf{AS} \tag{13.115}$$

in the matrix notation.

There are many cases where $\boldsymbol{x}$ is not a mixture of independent sources. ICA does not work in such cases. On the other hand, there are cases where the components s_i are all non-negative. Visual images are such signals, where $s(i, j)$ are the brightness of an image at pixel (i, j).

When all the components of $\boldsymbol{s}$ are non-negative, they are distributed on the first quadrant of the signal space, which is a cone. When signals are transformed linearly by $\mathbf{A}$ as

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{s}, \tag{13.116}$$

$\boldsymbol{x}$'s are distributed in another cone, because linear transformation $\mathbf{A}$ transforms one cone to another cone. Hence, from a number of observations $\boldsymbol{x}(t)$, we can find the cone in which the $\boldsymbol{x}$'s sit (Fig. 13.7). The mixture matrix $\mathbf{A}$ is recovered from the cone of $\mathbf{X}$. When the elements of $\mathbf{A}$ are also non-negative, those of $\mathbf{X}$ are non-negative. Therefore, the problem is formulated as follows:

Non-negative matrix factorization (NMF): Given non-negative matrix $\mathbf{X}$, factorize it as the product of two non-negative matrices $\mathbf{A}$ and $\mathbf{S}$,

$$\mathbf{X} = \mathbf{A}\mathbf{S}. \tag{13.117}$$

We define a divergence $D[\mathbf{M} : \mathbf{N}]$ between two non-negative matrices $\mathbf{M}$ and $\mathbf{N}$. Then, the loss function of decomposition is given by

$$L(\mathbf{A}, \mathbf{S}) = D[\mathbf{X} : \mathbf{A}\mathbf{S}]. \tag{13.118}$$

The Frobenius matrix norm

$$D(\mathbf{A}, \mathbf{B}) = \frac{1}{2} \sum_{i,t} |a_{it} - b_{it}|^2 \tag{13.119}$$

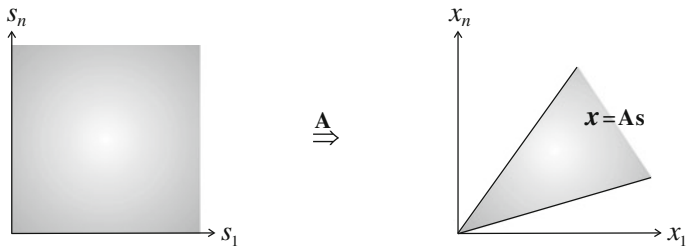


Fig. 13.7 A transforms the positive quadrant to a positive cone

is a divergence of frequent use. This is the square of the Euclidean norm and is symmetric with respect to $\mathbf{A}$ and $\mathbf{B}$. Another divergence is the KL-divergence defined by

$$D_{KL}[\mathbf{A}, \mathbf{B}] = \sum_{i,t} \left\{ a_{it} \log \frac{a_{it}}{b_{it}} - a_{it} - b_{it} \right\} \tag{13.120}$$

Other divergences such as α-, β- and (α, β)-divergences are also used on their own merits. See Cichocki et al. (2011).

The alternating minimization is a useful procedure to find the minimum of two variables $L(\mathbf{A}, \mathbf{S})$. We fix $\mathbf{A}^{(t)}$ at time t, $t = 1, 2, \ldots$, and minimize $L(\mathbf{A}^{(t)}, \mathbf{S})$ with respect to $\mathbf{S}$. Let the minimizer be $\mathbf{S}^{(t)}$. We then fix $\mathbf{S}^{(t)}$ and minimize $L(\mathbf{A}, S^{(t)})$ with respect to $\mathbf{A}$. The minimizer is written as $\mathbf{A}^{(t+1)}$. We repeat this procedure until convergence.

The gradient descent method is used to obtain the minimizer of the loss function. However, we need to take the non-negativity of $\mathbf{A}$ and $\mathbf{S}$ into account. The conventional gradient descent method does not satisfy this requirement and components of matrices would become negative in the procedure.

The exponential gradient descent (Kivinen and Warmuth 1997) is proposed to overcome this difficulty. Its procedure is as follows:

$$\mathbf{S}^{(t+1)} = \mathbf{S}^{(t)} \exp\left\{ -\eta \frac{\partial L}{\partial \mathbf{S}} \right\}, \quad \mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} \exp\left\{ -\eta \frac{\partial L}{\partial \mathbf{A}} \right\}, \tag{13.121}$$

where η is a learning constant. By using the logarithm, we have

$$\log S_{it}^{(t+1)} = \log S_{it}^{(t)} - \eta \frac{\partial L}{\partial S_{it}}, \quad \log A_{it}^{(t+1)} = \log A_{it}^{(t)} - \eta \frac{\partial L}{\partial A_{it}}. \tag{13.122}$$

Hence, (13.121) is the gradient descent applied to $\log \mathbf{S}$ and $\log \mathbf{A}$. When D is the Frobenius norm (13.119), we have

$$\frac{\partial L}{\partial A_{it}} = \left[-\mathbf{X}\mathbf{S}^T + \mathbf{A}\mathbf{S}\mathbf{S}^T \right]_{it}, \quad \frac{\partial L}{\partial S_{it}} = \left[-\mathbf{A}^T\mathbf{X} + \mathbf{A}^T\mathbf{A}\mathbf{S} \right]_{it}. \tag{13.123}$$

In this analogy, we have the following algorithm, originally proposed by Lee and Seung (1999):

$$\log A_{it}^{(t+1)} = \log A_{it}^{(t)} + \log \left(\mathbf{X}\mathbf{S}^T \right)_{it} - \log \left(\mathbf{A}\mathbf{S}\mathbf{S}^T \right)_{it}, \tag{13.124}$$

$$\log S_{it}^{(t+1)} = \log S_{it}^{(t)} + \log \left(\mathbf{A}^T\mathbf{X} \right)_{it} - \log \left(\mathbf{A}^T\mathbf{A}\mathbf{S} \right)_{it}. \tag{13.125}$$

There are many algorithms for NMF. See Cichocki et al. (2011), for example. NMF is further generalized to non-negative tensor factorization (NTF), where tensors are quantities having more than two indices.

13.4 Sparse Signal Processing

We have studied linear signal decomposition from $\boldsymbol{x}$ to $\boldsymbol{s}$,

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{s} = \sum_{i=1}^{n} s_i \boldsymbol{a}_i . \tag{13.126}$$

This section deals with the case that they are mixtures of a very few non-zero components, that is, a vector signal $\boldsymbol{s}$ is sparse. A signal $\boldsymbol{s}$ is said to be k-sparse when the components of $\boldsymbol{s}$ are zero except for at most k components. When k is much smaller than the dimension number n of $\boldsymbol{s}$, it is called a sparse vector. We consider a typical case that k is of the order $\log n$ or smaller, when n is large.

We interpret (13.126) such that $\boldsymbol{x}$ is a linear combination of n basis vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ and a basis $\boldsymbol{a}_i$ is activated when s_i is non-zero. Only a small number of basis vectors are activated in the sparse case. We assume that $\boldsymbol{x}$ is generated sparsely but do not know which basis vectors are activated. Let m be the dimension number of vector $\boldsymbol{x}$. We regard the m components of $\boldsymbol{x}$ as m measurements concerning an unknown signal $\boldsymbol{s}$, where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are known. When $m > n$, (13.126) is overdetermined, that is, the number m of equations is larger than the number n of unknowns. We usually assume that the observations are contaminated by noise, such that

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{s} + \varepsilon, \tag{13.127}$$

where ε is a noise vector, and we search for the least–squares solution.

When $m < n$, the equation is underdetermined. There are infinitely many solutions satisfying (13.126) even when it is noise contaminated. A conventional solution is the generalized inverse that minimizes the Euclidean norm among all possible solutions. When we know that $\boldsymbol{s}$ is sparse, we have a different solution. This was first noted by Chen et al. (1998). The following surprising theorem is known (Donoho 2006; Candes et al. 2006).

Theorem 13.2 *When n and m are large, $\boldsymbol{s}$ is recovered correctly in most cases, provided $\boldsymbol{s}$ is k-sparse and*

$$m > 2k \log n . \tag{13.128}$$

Roughly speaking, when k is a constant, a constant multiple of $\log n$ observations are enough to recover the n-dimensional $\boldsymbol{s}$. Since a very small number of sensors are enough, provided the original signal is sparse, the paradigm is called compressed sensing (Donoho 2006; Candes and Walkin 2008). Such a paradigm has emerged from statistics, ICA, signal processing and many related fields. It has grown to form a very hot field. There are many monographs and papers on this topic, see, e.g., Elad (2010), Eldar and Kutyniok (2012) and Bruckstein et al. (2009).

13.4.1 Linear Regression and Sparse Solution

Let us formulate the linear regression problem

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{\theta} + \varepsilon, \tag{13.129}$$

where $\boldsymbol{x}$ is the observation vector, $\mathbf{A} = \left(A_{ij}\right)$ is a known design matrix, $\boldsymbol{\theta}$ is the factor or explanatory vector to be determined and ε is a noise vector. We use $\boldsymbol{\theta}$ instead of $\boldsymbol{s}$ for the purpose of emphasizing that $\boldsymbol{\theta}$ is an e-affine coordinate system. The loss function to be minimized is

$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\sum |\boldsymbol{x} - \mathbf{A}\boldsymbol{\theta}|^2 . \tag{13.130}$$

We use $\psi(\boldsymbol{\theta})$ for the loss function in this subsection, because it plays the role of a convex function defining dually flat structure. This is the negative of the log likelihood when the noises are independent Gaussian. Since ψ is a quadratic function in the case of (13.130), by defining

$$\mathbf{G} = \mathbf{A}^T\mathbf{A}, \tag{13.131}$$

we have

$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^T\mathbf{G}\boldsymbol{\theta} - \boldsymbol{x}^T\mathbf{A}\boldsymbol{\theta} + c, \tag{13.132}$$

where c is a constant. When $m > n$, $\mathbf{G}$ is regular in general and the optimal solution is

$$\boldsymbol{\theta}^* = \mathbf{G}^{-1}\mathbf{A}^T\boldsymbol{x}. \tag{13.133}$$

When $m < n$, $\mathbf{G}$ is singular and there are infinitely many solutions in this underdetermined case. Let $\boldsymbol{s}_0$ be a solution. Then, for any null vector satisfying

$$\mathbf{G}\boldsymbol{n} = 0, \tag{13.134}$$

$\boldsymbol{s}_0 + \boldsymbol{n}$ is a solution. The solution that minimizes the L_2-norm is given by

$$\boldsymbol{\theta}^* = \mathbf{A}^\dagger\boldsymbol{x}, \tag{13.135}$$

where $\mathbf{A}^\dagger$ is the generalized inverse defined by

$$\mathbf{A}^\dagger = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}. \tag{13.136}$$

However, this solution is not sparse and almost all components are non-zero.

The sparsest solution is the one that minimizes the number of non-zero components

$$L_0(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\theta^i\right)^0 . \tag{13.137}$$

However, this is a combinatorial problem and computationally difficult to solve for large n. One may use the L_1-norm instead of L_0-norm,

$$L_1(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left|\theta^i\right| , \tag{13.138}$$

to obtain a sparse solution (Ishikawa 1996). There are many studies concerning when the minimum L_1-norm solution is identical to the minimum L_0-norm solution. It is now known that the solutions of the two problems coincide when

$$m \approx 2k \log n \tag{13.139}$$

for a randomly generated $\mathbf{A}$ with high probability. See, e.g., Candes et al. (2006).

13.4.2 Minimization of Convex Function Under L_1 Constraint

We generalize the linear regression problem and study the problem of minimizing a general convex function $\psi(\boldsymbol{\theta})$ under the L_1-constraint. See Hirose and Komaki (2010). The constraint is given by

$$L(\boldsymbol{\theta}) = \sum \left|\theta^i\right| = c. \tag{13.140}$$

We define a region of $\boldsymbol{\theta}$ by

$$B_c = \left\{ \boldsymbol{\theta} \,\middle|\, \sum |\theta^i| \leq c \right\} . \tag{13.141}$$

As c decreases, the constraint becomes stronger and finally when $c = 0$, it includes only $\boldsymbol{\theta} = 0$, the extremely sparse solution. See Fig. 13.8.

We have assumed in (13.129) that the noise is Gaussian. When it is not Gaussian, the negative of the log likelihood function, $\psi(\boldsymbol{\theta})$, is convex but is not a quadratic function. Another typical example is the logistic regression. In this case, given input $\boldsymbol{x}_i$, the response y_i is binary, taking values 0 and 1. Its probability is given by

$$\text{Prob}\{y_i = y\} = \exp\left\{ y\boldsymbol{\theta} \cdot \boldsymbol{x}_i - \tilde{\psi}\left(\boldsymbol{\theta} \cdot \boldsymbol{x}_i\right) \right\} , \tag{13.142}$$

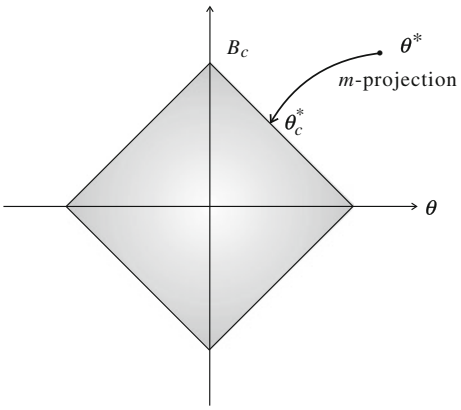


Fig. 13.8 Convex set B_c and m-projection of $\boldsymbol{\theta}^*$ to it

where

$$\tilde{\psi} = 1 + \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x}_i\}. \tag{13.143}$$

The loss function is the negative of log probability of the correct answer,

$$\psi(\boldsymbol{\theta}) = -\sum y_i \boldsymbol{x}_i \cdot \boldsymbol{\theta} + \sum \tilde{\psi}(\boldsymbol{\theta} \cdot \boldsymbol{x}_i). \tag{13.144}$$

This is convex and is strictly convex when $m > n$.

The problem is the minimization of

$$f(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}) + \lambda L(\boldsymbol{\theta}), \tag{13.145}$$

where λ is the Lagrange multiplier. We begin with the overdetermined case because it is simpler. The underdetermined case can be treated similarly, as will be stated later (see Donoho and Tsaig 2008). In the overdetermined case, there is a unique optimum $\boldsymbol{\theta}^*$ minimizing $L(\boldsymbol{\theta})$, that satisfies

$$\nabla\psi\left(\boldsymbol{\theta}^*\right) = 0. \tag{13.146}$$

This is the solution corresponding to a large enough c and is not sparse.

We introduce the dually flat geometry, where the e-affine coordinates are $\boldsymbol{\theta}$ and the dual coordinates (m-flat coordinates) are given by

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}). \tag{13.147}$$

The Riemannian metric is

$$\mathbf{G}(\boldsymbol{\theta}) = \nabla\nabla\psi(\boldsymbol{\theta}). \tag{13.148}$$

The divergence from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, derived from ψ, is

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] = \psi(\boldsymbol{\theta}) - \psi\left(\boldsymbol{\theta}'\right) - \nabla\psi\left(\boldsymbol{\theta}'\right) \cdot \left(\boldsymbol{\theta} - \boldsymbol{\theta}'\right). \tag{13.149}$$

Therefore, from (13.146), we see that

$$D\left[\boldsymbol{\theta} : \boldsymbol{\theta}^*\right] = \psi(\boldsymbol{\theta}) - \text{const.} \tag{13.150}$$

Hence, minimizing $\psi(\boldsymbol{\theta})$ is equivalent to minimizing the divergence from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$, that is the dual divergence from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}$. Since the area B_c defined by the constraint (13.141) is e-convex, the following is immediate from the projection theorem.

Theorem 13.3 *The solution $\boldsymbol{\theta}_c^*$ that minimizes $\psi(\boldsymbol{\theta})$ in the area B_c is given by the m-projection of $\boldsymbol{\theta}^*$ to B_c. The projection is unique.*

The analytical equation for $\boldsymbol{\theta}_c^*$ is obtained, by differentiating (13.145) with respect to $\boldsymbol{\theta}$,

$$\nabla\psi\left(\boldsymbol{\theta}_c^*\right) = -\lambda\nabla L\left(\boldsymbol{\theta}_c^*\right). \tag{13.151}$$

Since the solution is the m-projection of $\boldsymbol{\theta}^*$ to B_c, the m-geodesic connecting $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_c^*$ is orthogonal to the boundary of B_c if it lies on a smooth surface of B_c (Fig. 13.8). The gradient $\nabla L(\boldsymbol{\theta})$ is the normal vector of the surface of L, which is the supporting hypersurface of B_c at this point. However, since convex set B_c is a polyhedron, it is not differentiable at low-dimensional faces, such as vertices, edges, etc., where some components satisfy

$$\theta^i = 0. \tag{13.152}$$

There are infinitely many supporting hypersurfaces at a non-differentiable point. The set of the normal vectors of the supporting hypersurfaces is called the subgradient of L at that point (Fig. 13.9).

We give an explicit form of the subgradient. Let $A(\boldsymbol{\theta})$ be the set of indices for which $\theta^i \neq 0$,

$$A(\boldsymbol{\theta}) = \left\{i \,\middle|\, \theta^i \neq 0\right\}. \tag{13.153}$$

It is called the active set of $\boldsymbol{\theta}$, because θ^i is active, that is, not 0, for $i \in A(\boldsymbol{\theta})$. Then, the subgradient is written as

$$(\nabla L)_i = \begin{cases} \partial_i L(\boldsymbol{\theta}) = \operatorname{sgn}\theta^i, & i \in A, \\ \varepsilon_i, \quad \varepsilon_i \in [-1, 1], & i \in \bar{A}, \end{cases} \tag{13.154}$$

where ε_i may take an arbitrary value in $[-1, 1]$.

There is only one m-geodesic passing through a regular boundary point of B_c orthogonally. On the other hand, there are infinitely many m-geodesics which pass through a non-regular point and their tangent directions belong to the subgradient.

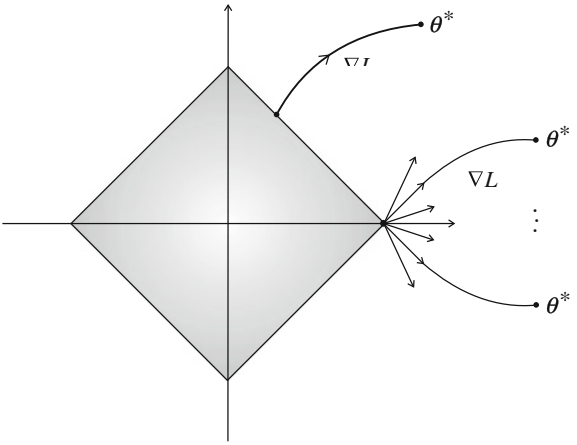


Fig. 13.9 Gradient and subgradient of L

Therefore, there exist a larger number of points $\boldsymbol{\theta}^*$ that are mapped to a non-regular point by the m-projection as the sparsity becomes large. This explains why a sparse solution is obtained by the L_1 regularization. See Fig. 13.9.

13.4.3 Analysis of Solution Path

Let us call $\boldsymbol{\theta}_c^*$ the solution path, considering c as a parameter along the path. It connects the origin 0 and the optimal point $\boldsymbol{\theta}^*$ as c changes from 0 to a large value. Hence, the solution path gives sparse solutions of which the sparsity is specified by c. LASSO is proposed for this purpose (Tibshirani 1996). Since the Lagrangian multiplier λ is determined as a monotone function $\lambda(c)$ of c, we may also regard λ as another parameter of the path (Efron et al. 2004). The dual coordinates of the optimal solution satisfy

$$\boldsymbol{\eta}_c^* = -\lambda \nabla L\left(\boldsymbol{\theta}_c^*\right). \tag{13.155}$$

By differentiating it with respect to c, the path satisfies

$$\mathbf{G}\left(\boldsymbol{\theta}_c^*\right)\dot{\boldsymbol{\theta}}_c^* = -\dot{\lambda}_c \nabla L\left(\boldsymbol{\theta}_c^*\right), \tag{13.156}$$

which is the equation to show the direction $\dot{\boldsymbol{\theta}}_c^*$ of the solution path. See Amari and Yukawa (2013) and Yukawa and Amari (2015).

Let us trace the path $\boldsymbol{\theta}_c^*$ starting from a sufficiently large c, where $\boldsymbol{\theta}_c^* = \boldsymbol{\theta}^*$. As c decreases, the path follows (13.156) as long as the active set $A\left(\boldsymbol{\theta}_c^*\right)$ does not change. But at a point where some θ_c^{*i} becomes newly 0, the active set A changes and the

direction $\dot{\boldsymbol{\theta}}_c^*$ of the path changes discontinuously, because ∇L of (13.154) changes, although the path itself is continuous.

We divide the indices into two parts, one belonging to the active set A and the other to its complement (inactive set) $\bar{A}$, and use the mixed coordinates

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}^A, \boldsymbol{\theta}^{\bar{A}}\right) \tag{13.157}$$

$$\boldsymbol{\eta} = \left(\boldsymbol{\eta}^A, \boldsymbol{\eta}^{\bar{A}}\right). \tag{13.158}$$

Then, we have the following lemma.

Lemma 13.2 *The solution path satisfies*

$$\boldsymbol{\eta}_c^{*A} = -\lambda(c)\boldsymbol{s}^A, \quad \boldsymbol{\theta}_\lambda^{*\bar{A}} = 0, \tag{13.159}$$

while the active set does not change, where $\boldsymbol{s} = \nabla L\left(\boldsymbol{\theta}_c\right)$ *is the vector of which the components are sgn* $\theta_c^{*^i}$.

The following least equiangle theorem of Efron et al. (2004) holds even in our general case.

Theorem 13.4 (Least Equiangle Property) *The direction* $\dot{\boldsymbol{\theta}}_c^*$ *of the solution path has the following properties:*

(1) For any coordinate axis belonging to the active set A, the angle between $\dot{\boldsymbol{\theta}}_\lambda^*$ *and the coordinate axis is the same,*

$$\left|\langle \dot{\boldsymbol{\theta}}_\lambda^*, \boldsymbol{e}_i \rangle\right| = \left|\langle \dot{\boldsymbol{\theta}}_\lambda^*, \boldsymbol{e}_j \rangle\right|, \quad i, j \in A, \tag{13.160}$$

where $\boldsymbol{e}_i$ *is the tangent vector along the coordinate* θ^i.

(2) For any axis belonging to $\bar{A}$*, the angle between* $\dot{\boldsymbol{\theta}}_\lambda^*$ *and the coordinate axis is larger than that of the axis belonging to A,*

$$\left|\langle \dot{\boldsymbol{\theta}}_\lambda^*, \boldsymbol{e}_i \rangle\right| < \left|\langle \dot{\boldsymbol{\theta}}_\lambda^*, \boldsymbol{e}_j \rangle\right|, \quad i \in A, j \in \bar{A}. \tag{13.161}$$

Proof The angle between $\dot{\boldsymbol{\theta}}_\lambda^*$ and any coordinate axis $\boldsymbol{e}_i$ is calculated by the inner product,

$$\langle \dot{\boldsymbol{\theta}}_\lambda^*, \boldsymbol{e}_i \rangle = \dot{\boldsymbol{\eta}}_\lambda^* \cdot \boldsymbol{e}_i = \dot{\eta}_{\lambda,i}^*. \tag{13.162}$$

Since $\dot{\boldsymbol{\eta}}_\lambda^*$ is proportional to $\nabla L\left(\boldsymbol{\theta}_\lambda^*\right)$, whereas

$$\left|\nabla L\left(\boldsymbol{\theta}_\lambda^*\right)\right|_i = 1 \tag{13.163}$$

for $i \in A$ and

$$\left|\nabla L\right|_i < 1 \tag{13.164}$$

for $i \in \bar{A}$, (13.160) and (13.162) hold. The direction of $\dot{\boldsymbol{\theta}}_{\lambda}^{*}$ changes only when i changes from $\bar{A}$ to A. □

This is the principle of Least Angle Regressions (LARS) of Efron et al. (2004), extended to the general class of convex optimization.

13.4.4 Minkovskian Gradient Flow

A gradient flow is the set of paths satisfying

$$\dot{\boldsymbol{\theta}}_c = -\nabla f\left(\boldsymbol{\theta}_c\right) \tag{13.165}$$

for some function $f(\boldsymbol{\theta})$. A gradient flow converges to a minimum of ψ when ψ is bounded, and no oscillation occurs. We show that the solution path of the extended LARS is a gradient flow under the Minkovskian gradient, which is defined in the following (Amari and Yukawa 2013). The natural gradient of $f(\boldsymbol{\theta})$ is the direction $\boldsymbol{a}$ in which the change of f is the largest. We define it by

$$\tilde{\nabla} f(\boldsymbol{\theta}) = \lim_{\varepsilon \to 0} \arg\max_{\boldsymbol{a}} f\left(\boldsymbol{\theta} + \varepsilon \boldsymbol{a}\right) \tag{13.166}$$

under the condition that the norm of $\boldsymbol{a}$ is kept constant. The natural gradient uses the Riemannian norm. We consider the L_q-norm

$$\|\boldsymbol{a}\|_q = \sum |a_i|^q , \tag{13.167}$$

which is a Minkovskian norm. The L_2-norm is a special case of the Minkovskian norm. It is easy to see that the steepest direction is given by

$$a_i = c\, |\partial_i f(\boldsymbol{\theta})|^{\frac{1}{q-1}} \operatorname{sgn}\left\{\partial_i f(\boldsymbol{\theta})\right\}, \tag{13.168}$$

where c is a constant.

Since we are dealing with the L_1-constraint, we define the Minkovskian gradient with respect to the L_1-norm by taking the limit of q approaching to 1 from the above. We take the constant c as

$$c = \frac{1}{\max\left|\frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta})\right|}. \tag{13.169}$$

Then, the limit is

$$a_i = \begin{cases} \operatorname{sgn}\left\{\partial_i f(\boldsymbol{\theta})\right\}, & |\partial_i f| = \max\left\{|\partial_1 f|, \ldots, |\partial_n f|\right\}, \\ 0, & |\partial_i f| \text{ is not the largest among all } |\partial_1 f|, \ldots, |\partial_n f|. \end{cases} \tag{13.170}$$

This is the Minkovskian gradient corresponding to the L_1-norm. The Minkovskian gradient of f is written as

$$\boldsymbol{a} = \tilde{\nabla}_M f(\boldsymbol{\theta}). \tag{13.171}$$

Its components are ± 1 when the absolute values of $\partial_i f$ are maximal and 0 for all the other components. See Amari and Yukawa (2013).

Consider the Minkovskian gradient flow,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \varepsilon \tilde{\nabla}_M \psi\left(\boldsymbol{\theta}^t\right), \tag{13.172}$$

starting from the origin in terms of the dual coordinates. This is the solution path of our problem. The components of $\tilde{\nabla}_M \psi\left(\boldsymbol{\theta}^*\right)$ are zero except for those indices that give the maximal values of $\left|\eta_i^*\right|$, since η_i^* is the derivative of $f(\boldsymbol{\theta})$ with respect to i. Hence, along the Minkovskian gradient flow, only the components η_i^*, which have the largest absolute values change. We need to solve the equation in terms of the primal coordinate system $\boldsymbol{\theta}_c^*$. Any components of $\boldsymbol{\theta}_c^*$ will change subject to the equiangle property.

We restate the LARS algorithm. Starting from the origin 0, we calculate the Minkovskian gradient of $\tilde{\nabla}_M \psi\left(\boldsymbol{\theta}^*\right)$ and pick up the index i^*,

$$i^* = \arg\max_j \left|\eta_j^*\right|. \tag{13.173}$$

The active set consists of a single i^*. (We ignore cases where two or more indices become the maximizer, but it is easy to consider such cases.) The path $\boldsymbol{\eta}_c^*$ proceeds in this direction of the Minkovskian gradient as c increases, while $\left|\eta_{ci^*}^*\right|$ is the smallest. As c becomes larger, another index j^* joins the set of the indices of the maximizer, satisfying

$$\left|\eta_{ci^*}^*\right| = \left|\eta_{cj^*}^*\right|. \tag{13.174}$$

We then add this to the active set, and the Minkovskian gradient is calculated for the new active set. In this way, the active set increases stepwise, until the path converges to $\boldsymbol{\theta}^*$. The Minkovskian gradient flow explains the properties of LARS in terms of the geometry of the gradient flow.

13.4.5 Underdetermined Case

We have so far studied the overdetermined case, where the unique unconstrained optimum $\boldsymbol{\theta}^*$ exists. In the underdetermined case of $m < n$, $\psi(\boldsymbol{\theta})$ is not strictly convex and the solution of

$$\nabla \psi(\boldsymbol{\theta}) = 0 \tag{13.175}$$

is not unique. The solutions form a submanifold. The problem is to obtain the one that has the minimum L_1-norm. The Hessian $\mathbf{G}$ is not strictly positive-definite in this case. Hence, the Riemannian metric does not exists. The transformation (13.147) from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$ exists but is not bijective and the inverse transformation is not necessarily unique.

In spite of these differences, the Eq. (13.151) obtained from the Lagrangian holds. Hence, the equation of the solution path (13.156) holds as well. We can prove the least–angle theorem in a similar way. Therefore, the solution path is given by a Minkovskian gradient flow starting at the origin $\boldsymbol{\theta}_c = 0$. We can use the same algorithm for solving the problem in the underdetermined case. See Donoho and Tsaig (2008) in the regression case.

13.5 Optimization in Convex Programming

Mathematical programming is a problem of finding the optimum solution under various constraints. A typical example is linear programming (LP), which minimizes a linear function under constraints given by linear inequalities. More generally, there is a problem of minimizing a linear loss function in a convex region. See Nesterov and Nemirovski (1993). This is called convex programming. A typical example of it is positive-semidefinite programming. An inner point method searches for the optimum solution sequentially inside the convex region. Since a convex region defines a dually flat structure, information geometry is useful in understanding these problems.

13.5.1 Convex Programming

Let us consider a manifold M having a coordinate system $\boldsymbol{\theta}$ and a bounded convex region Ω. A differentiable function $\psi(\boldsymbol{\theta})$ is called a barrier function when it is convex and diverges to infinity at the boundary $\partial\Omega$ of the region Ω. Let

$$\sum_i A_i(\omega)\theta^i - b(\omega) = 0 \tag{13.176}$$

be the supporting hypersurface of Ω at point $\omega \in \partial\Omega$ (Fig. 13.10). The convex region Ω is defined by

$$\Omega = \left\{ \boldsymbol{\theta} \,\middle|\, \sum_i A_i(\omega)\theta^i - b(\omega) \geq 0 \text{ for all } \omega \in \partial\Omega \right\}. \tag{13.177}$$

Since

$$-\log\left\{\sum A_i(\omega)\theta^i - b(\omega)\right\} \tag{13.178}$$

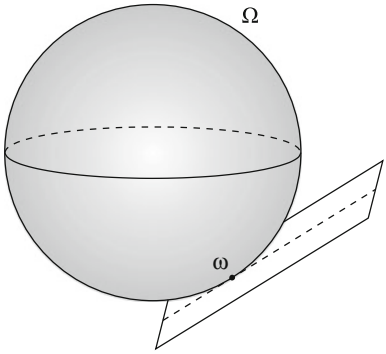


Fig. 13.10 Convex region Ω and supporting hyperplane at ω

diverges to infinity at the boundary, the convex function

$$\psi(\boldsymbol{\theta}) = -\int_{\partial\Omega} w(\omega) \log\left\{\sum A_i(\omega)\theta^i - b(\omega)\right\} d\omega \tag{13.179}$$

is a barrier function.

The supporting hypersurfaces in the case of LP are

$$\sum A_{\kappa i}\theta^i - b_\kappa \geq 0, \quad \kappa = 1, \ldots, m. \tag{13.180}$$

Hence, Ω is a polyhedron and the convex function is

$$\psi(\boldsymbol{\theta}) = -\sum_\kappa \log\left(\sum A_{\kappa i}\theta^i - b_\kappa\right). \tag{13.181}$$

The cost function to be minimized is

$$C(\boldsymbol{\theta}) = \sum c_i\theta^i. \tag{13.182}$$

The positive semi-definite programming is the problem of obtaining the positive semi-definite matrix $\mathbf{X}$ that minimizes the linear function

$$C(\mathbf{X}) = \mathrm{tr}(\mathbf{CX}), \tag{13.183}$$

where $\mathbf{C}$ is a constant matrix. The set of all positive semi-definite matrices forms a cone. We impose the constraints which $\mathbf{X}$ must satisfy:

$$\mathrm{tr}\,(\mathbf{A}_\kappa\mathbf{X}) - b_\kappa = 0, \quad \kappa = 1, \ldots, m, \tag{13.184}$$

where $\mathbf{A}_\kappa$ are constant matrices. The region defined by (13.184) is convex. This type of problem is also called the cone programming problem, appearing in many fields of research, e.g., in control theory. See Ohara (1999).

The barrier function for positive-definite matrices is given by

$$\psi(\mathbf{X}) = -\log\det|\mathbf{X}|. \tag{13.185}$$

The geometrical structure is the same as the invariant geometry of Gaussian distributions with mean 0 and covariance matrix $\mathbf{X}$.

13.5.2 Dually Flat Structure Derived from Barrier Function

Since a barrier function $\psi(\boldsymbol{\theta})$ is convex, it gives a dually flat structure to the manifold M, where $\boldsymbol{\theta}$ is e-affine coordinates and its Legendre transform

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}) \tag{13.186}$$

is m-affine coordinates.

The Riemannian metric $\mathbf{G}$ is given by

$$g_{ij}(\boldsymbol{\theta}) = \partial_i\partial_j\psi(\boldsymbol{\theta}) \tag{13.187}$$

(Nesterov and Todd 2002). Hence,

$$\eta_i = -\int \frac{A_i(\omega)}{\sum A_k(\omega)\theta^k - b(\omega)}d\omega, \tag{13.188}$$

$$g_{ij}(\boldsymbol{\theta}) = \int \frac{A_i(\omega)A_j(\omega)}{\left\{\sum A_k(\omega)\theta^k - b(\omega)\right\}^2}d\omega \tag{13.189}$$

in the case of (13.181).

The interior point method is a sequential search for the solution that minimizes $C(\boldsymbol{\theta})$, by changing $\boldsymbol{\theta}$ in the decreasing direction of C inside Ω. The natural gradient gives the steepest direction of C and is given by

$$\tilde{\nabla}C(\boldsymbol{\theta}) = \mathbf{G}^{-1}(\boldsymbol{\theta})\nabla C(\boldsymbol{\theta}). \tag{13.190}$$

The LP problem uses a linear function

$$C(\boldsymbol{\theta}) = \boldsymbol{c}\cdot\boldsymbol{\theta} \tag{13.191}$$

as a cost function. By using continuous time, the natural gradient flow is

$$\dot{\boldsymbol{\theta}}(t) = -\varepsilon \mathbf{G}(\boldsymbol{\theta})^{-1}\boldsymbol{c}, \tag{13.192}$$

where ε is a constant. The affine-projection method of Karmarkar solves this by using the discrete time step,

$$\Delta\boldsymbol{\theta} = -\varepsilon\tilde{\nabla}C(\boldsymbol{\theta}) = -\varepsilon\mathbf{G}^{-1}(\boldsymbol{\theta})\boldsymbol{c}. \tag{13.193}$$

It is known that this gives an algorithm of polynomial-time complexity. See Tanabe (1980).

The dynamic equation (13.192) reduces to the simple equation given by

$$\dot{\boldsymbol{\eta}}(t) = -\varepsilon\boldsymbol{c} \tag{13.194}$$

in the dual coordinates. The solution is a m-geodesic,

$$\boldsymbol{\eta}(t) = -\varepsilon t\boldsymbol{c} + \boldsymbol{c}_0. \tag{13.195}$$

Although the solution is very simple in the dual coordinates, we need the solution in the $\boldsymbol{\theta}$ coordinate system. Hence, the algorithm is not simple in the $\boldsymbol{\theta}$ coordinates and the transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is expensive. It is popular to solve the problem in the primal-dual formulation by using the Newton method.

13.5.3 Computational Complexity and m-curvature

In order to evaluate the number of steps to reach the optimal solution, we analyze the solution path. To this end, consider the following loss function parameterized by t:

$$L(\boldsymbol{\theta}, t) = tC(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}), \tag{13.196}$$

where the barrier function is added to the cost function. Let $\boldsymbol{\theta}^*(t)$ be the minimizer of $L(\boldsymbol{\theta}, t)$. This defines a path inside Ω parameterized by t, which cannot cross the boundary of Ω. As $t \to \infty$, the effect of the barrier function disappears, so $\boldsymbol{\theta}^*(t)$ converges to the optimum solution $\boldsymbol{\theta}^*$ of the original problem.

By differentiating (13.196) with respect to $\boldsymbol{\theta}$, we obtain the solution path in the dual coordinates,

$$\boldsymbol{\eta}^*(t) = -t\boldsymbol{c}. \tag{13.197}$$

We call the point $\boldsymbol{\eta}^*(0) = 0$ the center of Ω. The solution path is a dual geodesic connecting the center and the optimum solution $\boldsymbol{\eta}^*$. This is the steepest descent path starting at the center by using the natural gradient.

The path is an m-geodesic but is curved in the e-coordinates $\boldsymbol{\theta}$. When the curvature of the path is small, we can solve the discretized path equation by taking a large step size, but when the curvature is large, we need to use a small step size. Therefore, the number of steps depends on the curvature of the path. Kakihara, Ohara and Tsuchiya (2012) evaluated the necessary number of steps to obtain the optimum solution within a preassigned accuracy in terms of the embedding curvature of the path.

13.6 Dual Geometry Derived from Game Theory

13.6.1 Minimization of Game-Score

Statistical inference can be regarded as a game against Nature, where the player estimates the probability distribution Nature has assigned. Nature shows a realized value of random variable x subject to the true probability distribution $p(x)$. The player chooses an action a from the set A of actions. Let $l(x, a)$ be the instantaneous loss when a is chosen for x. The expected loss is

$$L(p, a) = \mathrm{E}[l(x, a)] = \int p(x) l(x, a) dx. \tag{13.198}$$

See Topsoe (1979), Grünwald and Dawid (2004), Dawid (2007) and Dawid et al. (2012) for a detailed formulation.

In the case of estimation, the player's action is to choose a probability distribution $q(x)$ from a set of actions consisting of probability distributions, $A = \{q(x)\}$. We call the loss $l(x, q)$ a game-score in the case of probability distributions and denote it by $S(x, q)$,

$$S(x, q) = l(x, q). \tag{13.199}$$

When N independent observations $x_1, \ldots, x_N$ are available, the game-score is written as

$$S(x_1, \ldots, x_N, q) = \mathrm{E}_{\hat{p}}[S(x, q)] = \frac{1}{N} \sum_{i=1}^{N} S(x_i, q), \tag{13.200}$$

where $\hat{p}(x)$ is the empirical distribution

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i). \tag{13.201}$$

The conventional loss used in statistics is the log loss, so the corresponding game-score is

$$S(x, q) = -\log q(x). \tag{13.202}$$

Minimization of the game-score (13.200) under log loss (13.202) gives the maximum likelihood estimator. We study another type of the game-score in the next subsection, called the Hyvärinen score (Hyvärinen 2005) given by

$$S(x,q) = \ddot{l}(x) + \frac{1}{2}\left\{\dot{l}(x)\right\}^2, \tag{13.203}$$

where

$$l(x,q) = \log q(x) \tag{13.204}$$

and $\dot{l}$ etc. are differentiations with respect to x.

For two probability distributions $p(x)$ and $q(x)$, let us define the game-relative-entropy by

$$H_S[p:q] = \mathrm{E}_p[S\{x,q(x)\}]. \tag{13.205}$$

The game-entropy of $p(x)$ is given by $H_S[p:p]$. When the game-score is given by (13.202), it is the Shannon entropy.

A game-score is proper when

$$H_S[p:q] \geq H_S[p:p] \tag{13.206}$$

holds for any p and q. It is strictly proper when the equality holds only for $q = p$. We study a strictly proper game-score. In this case, we define the game-divergence between $p(x)$ and $q(x)$ by

$$D_S[p:q] = -H_S[p:p] + H_S[p:q]. \tag{13.207}$$

This is the KL-divergence when the game-score is given by (13.202). We can derive a dual geometrical structure $\{g, \nabla, \nabla^*\}$ induced from the game-divergence (Dawid 2007) for any strictly positive game-score $S(x,q)$. We call it the S-geometry, which includes the invariant geometry as a special case of log loss.

Let us consider a parametric form of statistical model $M = \{p(x,\boldsymbol{\xi})\}$, where x is a scalar or a vector. We show only a scalar case, but it is easy to generalize results to the vector case. For a strictly proper game-score

$$S(x,\boldsymbol{\xi}) = S\left\{x, q(x,\boldsymbol{\xi})\right\}, \tag{13.208}$$

the divergence is written as a function of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ as

$$D_S\left[\boldsymbol{\xi}:\boldsymbol{\xi}'\right] = D_S\left[p(x,\boldsymbol{\xi}) : p\left(x,\boldsymbol{\xi}'\right)\right] = E_{p(x,\boldsymbol{\xi})}\left[S(x,\boldsymbol{\xi}') - S\left(x,\boldsymbol{\xi}\right)\right]. \tag{13.209}$$

Hence, from

$$\frac{\partial}{\partial \boldsymbol{\xi}'} D_S\left[\boldsymbol{\xi}:\boldsymbol{\xi}'\right]_{|\boldsymbol{\xi}'=\boldsymbol{\xi}} = 0, \tag{13.210}$$

we have

$$\mathrm{E}_{p(x,\boldsymbol{\xi})}\left[\frac{\partial}{\partial \boldsymbol{\xi}} S(x,\boldsymbol{\xi})\right] = 0. \tag{13.211}$$

This shows that

$$\boldsymbol{s}(x,\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} S(x,\boldsymbol{\xi}), \tag{13.212}$$

is an estimating function derived from game-score S. The estimating equation is

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s}\left(\boldsymbol{x}_i,\boldsymbol{\xi}\right) = 0. \tag{13.213}$$

This is equivalent to minimizing $D_S\left[\hat{p}(x) : p(x,\boldsymbol{\xi})\right]$ for the empirical distribution $\hat{p}(x)$.

We show that there are strict proper game-scores other than $l(x,\boldsymbol{\xi}) = -\log p(x,\boldsymbol{\xi})$. One type is derived from a Bregman divergence $D_\psi[p(x) : q(x)]$ given by

$$D_\psi[p:q] = \int \left[\psi\left\{q(x)\right\} - \psi\left\{p(x)\right\} + \left\{p(x) - q(x)\right\}\psi'\left\{q(x)\right\}\right] dx, \tag{13.214}$$

where $\psi(q)$ is a strictly convex function. It is easy to see that this is a Bregman divergence, and the related game-score is

$$S\left\{x, q(x)\right\} = \psi'\left\{q(x)\right\} + \int \left[\psi\left\{q(y)\right\} - q(y)\psi'\left\{q(y)\right\}\right] dy. \tag{13.215}$$

It reduces to the log score when

$$\psi(u) = -u\log u. \tag{13.216}$$

The estimating function in this case is

$$\boldsymbol{s}(x,\boldsymbol{\xi}) = \psi''\left\{p(x,\boldsymbol{\xi})\right\}\partial_\xi p(x,\boldsymbol{\xi}) - c(\boldsymbol{\xi}), \tag{13.217}$$

where

$$c(\boldsymbol{\xi}) = E\left[\psi''\left\{p(x,\boldsymbol{\xi})\right\}\partial_\xi p(x,\boldsymbol{\xi})\right]. \tag{13.218}$$

Since D_ψ is a Bregman divergence, a dually flat structure is introduced in the manifold $M = \{p(x)\}$. As is seen from (13.214), the convex function is $\psi(q)$, where the θ-coordinates of $q \in M$ are of function degrees of freedom,

$$\theta_x(q) = q(x), \tag{13.219}$$

and the η-coordinates are

$$\eta_x(q) = \psi' \{q(x)\}. \tag{13.220}$$

The Riemannian metric and cubic tensor are derived from ψ.

The estimator $\hat{\boldsymbol{\xi}}$ derived from a game-score is consistent, because $\boldsymbol{s}(x, \boldsymbol{\xi})$ is an estimating function. We study its efficiency. Let $\boldsymbol{\xi}$ be the true value and let us put

$$\hat{\boldsymbol{\xi}} = \boldsymbol{\xi} + \Delta\boldsymbol{\xi}, \tag{13.221}$$

where $\hat{\boldsymbol{\xi}}$ is the estimator satisfying the estimating equation,

$$\frac{1}{N} \sum \boldsymbol{s}\left(x_t, \hat{\boldsymbol{\xi}}\right) = 0. \tag{13.222}$$

By the Taylor expansion, we have

$$\frac{1}{N} \sum \boldsymbol{s}\left(x_t, \boldsymbol{\xi} + \Delta\boldsymbol{\xi}\right) = \frac{1}{\sqrt{N}} \frac{1}{\sqrt{N}} \sum \boldsymbol{s}\left(x_t, \boldsymbol{\xi}\right) + \frac{1}{N} \sum \partial_{\boldsymbol{\xi}} \boldsymbol{s}\left(x_t, \boldsymbol{\xi}\right) \Delta\boldsymbol{\xi}. \tag{13.223}$$

Due to the central limit theorem, $1/\sqrt{N}$ of the first term of (13.223) converges to a Gaussian random variable $\boldsymbol{\varepsilon}$, the mean of which is 0 and the covariance is

$$\mathbf{V} = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \mathrm{E}\left[\boldsymbol{s}(x, \boldsymbol{\xi})\boldsymbol{s}(x, \boldsymbol{\xi})^T\right]. \tag{13.224}$$

The coefficient of the second term converges, due to the law of large numbers, to

$$\mathbf{K}(\boldsymbol{\xi}) = \mathrm{E}\left[\partial_{\boldsymbol{\xi}} \boldsymbol{s}(x, \boldsymbol{\xi})\right]. \tag{13.225}$$

Therefore, the estimation error is

$$\Delta\boldsymbol{\xi} = -\frac{1}{\sqrt{N}} \mathbf{K}^{-1} \boldsymbol{\varepsilon}. \tag{13.226}$$

The asymptotic error covariance of $\hat{\boldsymbol{\xi}}$ is

$$\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \frac{1}{N} \mathbf{K}^{-1} \mathbf{V} \left(\mathbf{K}^{-1}\right)^T, \tag{13.227}$$

which is larger than the inverse $\boldsymbol{G}^{-1}$ of the Fisher information matrix in general.

The loss of information or efficiency is analyzed as follows. Let us decompose random variable $\boldsymbol{s}(x, \boldsymbol{\xi})$ in the direction of the score vector $\partial_{\boldsymbol{\xi}} l(x, \boldsymbol{\xi})$, which consists of random variables representing the tangent vectors along the coordinate curves $\boldsymbol{\xi}$, and orthogonal to it,

$$\boldsymbol{s}(x,\boldsymbol{\xi}) = c(\boldsymbol{\xi})\left\{\partial_{\boldsymbol{\xi}} l(x,\boldsymbol{\xi}) + \boldsymbol{a}(x,\boldsymbol{\xi})\right\}, \tag{13.228}$$

$$\mathrm{E}\left[\boldsymbol{a}(x,\boldsymbol{\xi})\partial_{\boldsymbol{\xi}} l(x,\boldsymbol{\xi})^T\right] = 0. \tag{13.229}$$

We may put $c(\boldsymbol{\xi}) = 1$, since the estimating equation is the same for any $c(\boldsymbol{\xi})$. Then, we have

$$\begin{aligned}\mathbf{K}(\boldsymbol{\xi}) &= \mathrm{E}\left[\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}} l(x,\boldsymbol{\xi}) + \partial_{\boldsymbol{\xi}}\boldsymbol{a}(x,\boldsymbol{\xi})\right] \\ &= -\mathbf{G}(\boldsymbol{\xi}),\end{aligned} \tag{13.230}$$

because

$$0 = \mathrm{E}\left[\boldsymbol{s}(x,\boldsymbol{\xi})\right] = \mathrm{E}\left[\partial_{\boldsymbol{\xi}} l(x,\boldsymbol{\xi}) + \partial_{\boldsymbol{\xi}}\boldsymbol{a}(x,\boldsymbol{\xi})\right] \tag{13.231}$$

and

$$\mathrm{E}\left[\partial_{\boldsymbol{\xi}} l(x,\boldsymbol{\xi})\right] = 0. \tag{13.232}$$

The term $\boldsymbol{a}$ is explicitly given by

$$\boldsymbol{a}(x,\boldsymbol{\xi}) = \boldsymbol{s}(x,\boldsymbol{\xi}) - \boldsymbol{G}(\boldsymbol{\xi})^{-1}\mathrm{E}\left[\boldsymbol{s}(x,\boldsymbol{\xi})\partial_{\boldsymbol{\xi}} l(x,\boldsymbol{\xi})^T\right]\partial_{\boldsymbol{\xi}} l(x,\boldsymbol{\xi}). \tag{13.233}$$

Hence, we have

$$\mathbf{V} = \mathbf{G} + \mathrm{E}\left[\boldsymbol{a}(x,\boldsymbol{\xi})\boldsymbol{a}(x,\boldsymbol{\xi})^T\right] \tag{13.234}$$

and

$$\mathrm{E}\left[\varepsilon\varepsilon^T\right] = \mathbf{G}^{-1} + \mathbf{G}^{-1}\mathbf{A}\mathbf{G}^{-1}, \tag{13.235}$$

where

$$\mathbf{A} = \mathrm{E}\left[\boldsymbol{a}(x,\boldsymbol{\xi})\boldsymbol{a}(x,\boldsymbol{\xi})^T\right]. \tag{13.236}$$

Therefore, the asymptotic error covariance increases by $\mathbf{G}\mathbf{A}\mathbf{G}^{-1}$. The estimator is Fisher efficient when and only when $\boldsymbol{a}(x,\boldsymbol{\xi}) = 0$.

13.6.2 Hyvärinen Score

Hyvärinen (2005, 2007) proposed an interesting game-score given by

$$S(x,q) = \ddot{l}(x) + \frac{1}{2}\left\{\dot{l}(x)\right\}^2, \tag{13.237}$$

where $l(x) = \log q(x)$ and $\dot{}$ denotes the differentiation with respect to x. When $\boldsymbol{x}$ is a vector, it is

$$S(\boldsymbol{x},q) = \Delta l(\boldsymbol{x},\boldsymbol{\xi}) + \frac{1}{2}\left|\nabla l(\boldsymbol{x},\boldsymbol{\xi})\right|^2, \tag{13.238}$$

where Δ is the Laplacian and ∇ is the gradient with respect to $\boldsymbol{x}$. The related game-entropy is

$$H[p(x)] = -\frac{1}{2}\int p(x)\left\{\dot{l}(x)\right\}^2 dx \tag{13.239}$$

and the divergence is

$$D\left[p(x) : q(x)\right] = \mathrm{E}_p\left[S(x,q) - S(x,p)\right]. \tag{13.240}$$

Lemma 13.3 *The Hyvärinen divergence is rewritten as*

$$D\left[p(x) : q(x)\right] = \frac{1}{2}\int p(x)\left\{\frac{d}{dx}\log p(x) - \frac{d}{dx}\log q(x)\right\}^2 dx. \tag{13.241}$$

Proof We calculate $\mathrm{E}_p\left[S(x,q)\right]$ by putting

$$l_p(x) = \log p(x), \quad l_q(x) = \log q(x). \tag{13.242}$$

Then

$$\begin{aligned}\mathrm{E}_p\left[S(x,q)\right] &= \int p(x)\left\{\ddot{l}_q(x) + \frac{1}{2}\left\{\dot{l}_q(x)\right\}^2\right\} dx \\ &= \int \left\{-\dot{p}(x)\dot{l}_q(x) + \frac{1}{2}p(x)\left\{\dot{l}_q(x)\right\}^2\right\} dx \\ &= \frac{1}{2}\mathrm{E}_p\left[\left\{\dot{l}_q(x)\right\}^2 - 2\dot{l}_q(x)\dot{l}_p(x)\right],\end{aligned} \tag{13.243}$$

where the formula of partial integration is used. $\mathrm{E}_p\left[S(x,p)\right]$ is calculated similarly, and we have (13.241).

The Hyvärinen divergence is not a Bregman divergence and hence the geometry derived from it is not dually flat. Note that it does not depend on the normalizing constant of q, because

$$D[p(x) : cq(x)] = D[p(x) : q(x)] \tag{13.244}$$

for any c. Hence, it can be used for estimation when the normalization factor is difficult to calculate.

For parametric family of probability distributions $p(\boldsymbol{x}, \boldsymbol{\xi})$, the Hyvärinen estimating function is given by

$$\boldsymbol{s}(x,\boldsymbol{\xi}) = \nabla S\left\{x, p(x,\boldsymbol{\xi})\right\} = \partial_{\xi}\dot{l}(x,\boldsymbol{\xi}) + \dot{l}(x,\boldsymbol{\xi})\partial_{\xi}\dot{l}(x,\boldsymbol{\xi}). \tag{13.245}$$

It is a homogeneous estimating function, because it does not depend on the normalization factor $c(\boldsymbol{\xi})$ of a probability distribution. For example, an exponential family

is written as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} + k(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\}, \tag{13.246}$$

but $\dot{l}$ and $\ddot{l}$ do not include $\psi(\boldsymbol{\theta})$. Hence, one can easily obtain an estimator without calculating $\psi(\boldsymbol{\theta})$. Calculation of the normalization factor ψ is computationally heavy in Bayesian inference, so the Hyvärinen score is useful in such a case.

We give a simple illustrative example.

Example 13.1 Consider a simple exponential family,

$$p(x, \theta) = \exp\left\{-\theta x^3 - \psi(\theta)\right\}, \quad \theta > 0, x > 0. \tag{13.247}$$

We can calculate ψ in this case as

$$\psi(\theta) = \frac{1}{3}\log\theta + c. \tag{13.248}$$

Therefore, the η-coordinate is

$$\eta = \frac{1}{3}\frac{1}{\theta}. \tag{13.249}$$

The MLE is given by

$$\hat{\theta}_{\text{mle}} = \frac{N}{3\sum x_i^3}. \tag{13.250}$$

The Hyvärinen score is

$$s(x, \theta) = -6x - 9\theta x^4. \tag{13.251}$$

Hence, the related estimator is

$$\hat{\theta} = \frac{2}{3}\frac{\sum x_i}{\sum x_i^4}, \tag{13.252}$$

which is asymptotically unbiased but is not efficient, because the score $s(x, \theta)$ is not included in the space of

$$\partial_\theta \log p(x, \theta) = x^3 - \psi'(\boldsymbol{\xi}). \tag{13.253}$$

The following theorem shows the case when the Hyvärinen estimator is Fisher efficient. See Hyvärinen (2005).

Theorem 13.5 *The Hyvärinen estimator is Fisher efficient for multivariate Gaussian distributions and is not efficient for other distributions.*

Proof Both $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\xi})$ and $\partial_{\boldsymbol{\xi}} l(\boldsymbol{x}, \boldsymbol{\xi})$ are quadratic functions of $\boldsymbol{x}$ in the multi-variate Gaussian case, $\partial_{\boldsymbol{\xi}} l(\boldsymbol{x}, \boldsymbol{\xi})$ spanning all the quadratic functions of $\boldsymbol{x}$. Hence, $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\xi})$ is included in the space spanned by $\partial_{\boldsymbol{\xi}} l(\boldsymbol{x}, \boldsymbol{\xi})$ and $\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{\xi}) = 0$. On the other hand, this occurs only for multivariate Gaussian distributions. □

Parry et al. (2012) and Hyvärinen (2007) extend the Hyvärinen score applicable to the case of discrete $\boldsymbol{x}$ such as a graphical model. We show another new idea.

Consider the case where $\boldsymbol{x}$ is a discrete random variable having a graphical structure. When $\boldsymbol{x}'$ and $\boldsymbol{x}$ are connected by a branch, $\boldsymbol{x}'$ is a neighbor of $\boldsymbol{x}$, $\boldsymbol{x}' \in N_{\boldsymbol{x}}$, where $N_{\boldsymbol{x}}$ is the set of neighbors of $\boldsymbol{x}$. A typical example is a Boltzmann machine, where $\boldsymbol{x}'$ is a neighbor of $\boldsymbol{x}$ when one and only one component of $\boldsymbol{x}'$ is different from $\boldsymbol{x}$. Hence, the graph is represented by an n-cube.

The graph Laplacian Δ is an operator, acting on function $f(\boldsymbol{x})$ as

$$\Delta f(\boldsymbol{x}) = \frac{1}{|N_{\boldsymbol{x}}|} \sum_{\boldsymbol{x}' \in N_{\boldsymbol{x}}} \left\{ f(\boldsymbol{x}) - f\left(\boldsymbol{x}'\right) \right\}, \tag{13.254}$$

where $|N_{\boldsymbol{x}}|$ is the cardinality of $N_{\boldsymbol{x}}$. It can be rewritten as

$$\Delta f(\boldsymbol{x}) = \sum_{\boldsymbol{x}'} C\left(\boldsymbol{x}, \boldsymbol{x}'\right) f\left(\boldsymbol{x}'\right), \tag{13.255}$$

where

$$C\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \begin{cases} \frac{1}{|N_{\boldsymbol{x}}|}, & \boldsymbol{x}' \in N_{\boldsymbol{x}}, \\ -1, & \boldsymbol{x}' = \boldsymbol{x}, \\ 0, & \text{otherwise}. \end{cases} \tag{13.256}$$

An interesting property is shown in the following lemma.

Lemma

$$\sum_{\boldsymbol{x}} \Delta f(\boldsymbol{x}) h(\boldsymbol{x}) = \sum f(\boldsymbol{x}) \Delta' h(\boldsymbol{x}), \tag{13.257}$$

where

$$\Delta' f(\boldsymbol{x}) = \sum_{\boldsymbol{x}'} C\left(\boldsymbol{x}', \boldsymbol{x}\right) f\left(\boldsymbol{x}'\right). \tag{13.258}$$

When the graph is homogeneous, having constant $|N_{\boldsymbol{x}}|$,

$$\Delta' = \Delta. \tag{13.259}$$

Proof From

$$\sum_{\boldsymbol{x}} \Delta f(\boldsymbol{x}) h(\boldsymbol{x}) = \sum_{\boldsymbol{x}, \boldsymbol{x}'} f\left(\boldsymbol{x}'\right) C\left(\boldsymbol{x}, \boldsymbol{x}'\right) h(\boldsymbol{x}), \tag{13.260}$$

(13.257) follows immediately.

We define a new game score when $\boldsymbol{x}$ is discrete and the graph is homogenous, that is, $\Delta = \Delta'$, by

$$S(\boldsymbol{x}, p) = \left\{ \frac{\Delta p(\boldsymbol{x})}{p(\boldsymbol{x})} \right\}^2 - 2\Delta \left\{ \frac{\Delta p(\boldsymbol{x})}{p(\boldsymbol{x})} \right\}. \tag{13.261}$$

This does not depend on the normalization factor of $p(\boldsymbol{x})$. The estimating function $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\xi})$ is defined in the parametric case as

$$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\xi}) = \nabla_{\xi} S\left\{\boldsymbol{x}, p(\boldsymbol{x}, \boldsymbol{\xi})\right\}. \tag{13.262}$$

This gives the estimating equation

$$\sum \boldsymbol{s}\left(\boldsymbol{x}_i, \boldsymbol{\xi}\right) = 0 \tag{13.263}$$

not depending on the normalization factor.

The meaning of this score is given by the following theorem.

Theorem 13.6 *The divergence derived from the score (13.261) is*

$$D\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] = E_{\xi}\left[\left\{ \frac{\Delta p(\boldsymbol{x}, \boldsymbol{\xi})}{p(\boldsymbol{x}, \boldsymbol{\xi})} - \frac{\Delta p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)}{p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)} \right\}^2 \right]. \tag{13.264}$$

Proof We calculate $E_{\xi}\left[S\left\{\boldsymbol{x}, p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)\right\}\right]$ as before. However we use

$$\begin{aligned} \sum_{x} p(\boldsymbol{x}, \boldsymbol{\xi}) \Delta \left\{ \frac{\Delta p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)}{p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)} \right\} &= \sum_{x} \Delta p(\boldsymbol{x}, \boldsymbol{\xi}) \frac{\Delta p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)}{p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)} \\ &= E_{\xi}\left[\frac{\Delta p(\boldsymbol{x}, \boldsymbol{\xi})}{p(\boldsymbol{x}, \boldsymbol{\xi})} \frac{\Delta p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)}{p\left(\boldsymbol{x}, \boldsymbol{\xi}'\right)} \right], \end{aligned} \tag{13.265}$$

instead of the formula of partial integration used in the continuous case. We then have the theorem.

We can calculate the efficiency of the derived estimator by calculating $\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{\xi})$.

Remarks

The last chapter deals with miscellaneous subjects concerning signal processing. PCA is an old subject but is still active. We have focused on the dynamics of learning for PCA from the point of view of geometry. ICA is a relatively newly developed subject, in which non-Gaussianity of distributions plays an important role. Information geometry elucidates its structure. The natural gradient in the manifold of matrices is useful for this purpose. Moreover, it is formulated as a semi-parametric

statistical problem, so that a general form of estimating functions is given by information geometry. We can stabilize and accelerate its learning dynamics by using the Newton method in the manifold of matrices. We have further touched upon the NMF problem.

Sparse signal processing is a hot topic on which many researchers are working. We are not able to overview most of the excellent results in this field. Instead, we have touched upon the minimization problem from the information geometry point of view. The Minkovskian gradient is a new topic, reinterpreting the L_1-constrained minimization. The problem of minimization under $L_p (0 < p < 1)$ is another interesting subject. See Xu et al. (2012), Yukawa and Amari (2015) and Jeong et al. (2015), for example.

Convex programming is a big field in operations research. We discussed only the interior point method, in which information geometry plays an interesting role. Another important topic related to optimization is the stochastic relaxation framework which is useful even for discrete optimization (Malagò et al. 2013), touched upon in the previous chapter. We also touched upon an information geometry framework given by game theory (Dawid 2007). The Hyvärinen score $S(\boldsymbol{x}, p)$ when $\boldsymbol{x}$ is discrete is a new idea emerged at the last stage in preparing the monograph. The dual geometry derived from the Hyvärinen score is an interesting subject in future research.

References

D. Ackley, G. E. Hinton and J. Sejnowski, A learning algorithm for Boltzmann machines. Cognitive Science, 9, 147–169, 1985.

A. Agarwal and H. Daumé III, A geometric view of conjugate priors. Machine Learning, 81, 99–113, 2010.

M. Aizerman, E. Braverman and L. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25, 821–837, 1964.

M. Akahira and K. Takeuchi, Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency, Springer LN in Statistics, vol. 7, 1981.

S. Akaho and K. Takabatake, Information geometry of contrastive divergence. In Information Theory and Statistical Learning, 3–9, 2008.

M. S. Ali and S. D. Silvey, A general class of coefficients of divergence of one distribution from another. Journal of Royal Statistical Society, B, 28, 131–142, 1966.

S. Amari, On some primary structures of non-Riemannian plasticity theory. RAAG Memoirs, 3, D-IX, 99–108, 1962.

S. Amari, A geometrical theory of moving dislocations. RAAG Memoirs, 4, D-XVII, 153–161, 1968.

S. Amari, Theory of adaptive pattern classifiers. IEEE Transactions on Electronic Computers, 16, 299–307, 1967.

S. Amari, Neural theory of association and concept-formation. Biological Cybernetics, 26, 175–185, 1977.

S. Amari, Differential geometry of curved exponential families—curvature and information loss. Annals of Statistics, 10, 357–385, 1982.

S. Amari, Finsler geometry of non-regular statistical models. RIMS Kokyuroku (in Japanese), Non-Regular Statistical Estimation, Ed. M. Akahira, 538, 81–95, 1984.

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

S. Amari, Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections and divergence. Mathematical Systems Theory, 20, 53–82, 1987.

S. Amari, Information geometry of the EM and em algorithms for neural networks. Neural Networks, 8, 1379–1408, 1995.

S. Amari, Natural gradient works efficiently in learning. Neural Computation, 10, 251–276, 1998.

S. Amari, Superefficiency in blind source separation. IEEE Transactions on Signal Processing, 47, 936–944, 1999.

S. Amari, Estimating functions of independent component analysis for temporally correlated signals. Neural Computation, 12, 2083–2107, 2000.

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

S. Amari, Information geometry on hierarchy of probability distributions. IEEE Transactions on Information Theory, 47, 1701–1711, 2001.
S. Amari, Integration of stochastic models by minimizing α-divergence. Neural Computation, 19, 2780–2796, 2007.
S. Amari, α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Transactions on Information Theory, 55, 11, 4925–4931, 2009.
S. Amari, Information geometry of positive measures and positive-definite matrices: Decomposable dually flat structure. Entropy, 16, 2131–2145, 2014.
S. Amari and J. Armstrong, Curvature of Hessian manifolds, Differential Geometry and its Applications 33, 1–12, 2014.
S. Amari and J-F. Cardoso, Blind source separation—Semiparametric statistical approach. IEEE Transactions on Signal Processing, 45, 2692–2700, 1997.
S. Amari, A. Cichocki and H. Yang, A new learning algorithm for blind signal separation. In Advances in Neural Information Processing Systems (Eds. M. Mozer et al.), 8, 757–763, 1996.
S. Amari and M. Kawanabe, Information geometry of estimating functions in semi-parametric statistical models. Bernoulli, 3, 29–54, 1997.
S. Amari, S. Ikeda and H. Shimokawa, Information geometry of α-projection in mean field approximation. In M. Opper and D. Saad (Eds), Advanced Mean Field Methods: Theory and Practice, 241–257. MIT Press, 2001.
S. Amari, K. Kurata and H. Nagaoka, Information geometry of Boltzmann machines. IEEE Transactions on Neural Networks, 3, 260–271, 1992.
S. Amari and H. Nagaoka, Methods of Information Geometry. American Mathematical Society and Oxford University Press, 2000.
S. Amari, H. Nakahara, S. Wu and Y. Sakai, Synchronous firing and higher-order interactions in neuron pool. Neural Computation, 15, 127–142, 2003.
S. Amari and A. Ohara, Geometry of q-exponential family of probability distributions. Entropy, 13, 1170–1185, 2011.
S. Amari, A. Ohara and H. Matsuzoe, Geometry of deformed exponential families: Invariant, dually flat and conformal geometry. Physica A, 391, 4308–4319, 2012.
S. Amari, H. Park and K. Fukumizu, Adaptive method of realizing natural gradient learning for multilayer perceptrons. Neural Computation, 12, 1399–1409, 2000.
S. Amari, H. Park and T. Ozeki, Singularities affect dynamics of learning in neuromanifolds. Neural Computation, 18, 1007–1065, 2006.
S. Amari and S. Wu, Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 12, 783-789, 1999.
S. Amari and M. Yukawa, Minkovskian gradient for sparse optimization. IEEE Journal of Selected Topics in Signal Processing, 7, 576–585, 2013.
D. Arthur and S. Vassilvitskii, k-means++: The advantages of careful seeding, Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035, 2007.
K. Arwini and C. T. J. Dodson, Information Geometry. Springer, 2008.
N. Ay, An information-geometric approach to a theory of pragmatic structuring. Annals of Probability, 30, 416–436, 2002.
N. Ay, Information geometry on complexity and stochastic interaction. Entropy, 17, 2432–2458, 2015.
N. Ay, J. Jost, H. V. Lê and L. Schwachhöfer, Information Geometry and Sufficient Statistics. arXiv:1207.6736, 2013.
N. Ay and S. Amari, A novel approach to canonical divergences within information geometry. Entropy, 17, 8111–8129, 2015.
N. Ay and A. Knauf, Maximizing multi-information. Kybernetika, 42, 517–538, 2006.
N. Ay, E. Olbrich, N. Bertschinger and. J. Jost, A geometric approach to complexity. Chaos, 21, 037103, 2011.
D. Balduzzi and G. Tononi, Integrated information in discrete dynamical systems: Motivation and theoretical framework. PLos Computational Biology, 4, e1000091, 2008.

A. Banerjee, S. Merugu, I. Dhillon and J. Ghosh, Clustering with Bregman divergences. Journal of Machine Learning Research, 6, 1705–1749, 2005.
N. Barkai, H. S. Seung, and H. Sompolinsky. On-line learning of dichotomies. Advances in Neural Information Processing Systems, 7, 303–310, 1995.
O. E. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory. Wiley, 1978.
A. B. Barrett and A. K. Seth, Practical measures of integrated information for time-series data. PLoS Computational Biology, 7, e1001052, 2011.
M. Basseville, Divergence measures for statistical data processing—An annotated bibliography. Signal Processing, 93, 621–633, 2013.
A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 31, 167–175, 2003.
J. M. Begun, W. J. Hall, W. M. Huang and J. A. Wellner, Information and asymptotic efficiency in parametric-nonparametric models. Annals of Statistics, 11, 432-452, 1983.
A. J. Bell and T. Sejnowski, An information maximization approach to blind separation and blind deconvolution. Neural Computation, 7, 1129–1159, 1995.
P. J. Bickel, C. A. J. Ritov, and J. A. Wellner, Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins University Press, 1994.
J.-D. Boissonnat, F. Nielsen and R. Nock, Bregman Voronoi diagrams. Discrete and Computational Geometry, 44, 281–307, 2010.
L. Bregman, The relaxation method of finding a common point of convex sets and its applications to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 7, 200–217, 1967.
R. Brockett, Some geometric questions in the theory of linear systems. IEEE Transactions on Automatic Control, 21, 449–455, 1976.
R. Brockett, Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems. Linear Algebra and its Applications, 146, 79–91, 1991.
A. Bruckstein, D. Donoho and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Review, 51, 34–81, 2009.
J. Burbea and C. R. Rao, On the convexity of some divergence measures based on entropy functions. IEEE Transactions on Information Theory, 28, 489–495, 1982.
W. Byrne, Alternating minimization and Boltzmann machine learning. IEEE Transactions on Neural Networks, 3, 612–620, 1992.
O. Calin and C. Udriste, Geometric Modeling in Probability and Statistics. Springer, 2013.
L. L. Campbell, An extended Chentsov characterization of a Riemannian metric. Proceedings of American Mathematical Society, 98, 135–141, 1986.
E. J. Candes, J. Romberg and T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics 59, 1207–1223, 2006.
E. J. Candes and M. B. Walkin, An introduction to compressive sampling. IEEE Signal Processing Magazine, 25, 21–30, 2008.
J.-F. Cardoso and B. H. Laheld, Equivariant adaptive source separation. IEEE Transactions on Signal Processing, 44, 3017–3030, 1996.
J.-F. Cardoso and A. Souloumiac, Jacobi angles for simultaneous diagonalization. SIAM Journal on Mathematical Analysis and Applications, 17, 161–164, 1996.
A. Cena and G. Pistone, Exponential statistical manifold. Annals of Institute of Statistical Mathematics, 59, 27–56, 2007.
T. Chen and S. Amari, Unified stabilization approach to principal and minor components extraction algorithms. Neural Networks, 14, 1377–1387, 2001.
T. P. Chen, S. Amari and Q. Lin, A unified algorithm for principal and minor components extraction. Neural Networks, 11, 3, 385–390, 1998.
S. S. Chen, D. L. Donoho and M. A. Saunders, Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computation, 20, 33–61, 1998.
N. N. Chentsov, Statistical Decision Rules and Optimal Inference, AMS, 1982 (originally published in Russian, Nauka, 1972).

H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. Annals of Mathematical Statistics, 23, 493–507, 1952.
H. Choi, S. Choi, A. Katake and Y. Choe, Parameter learning for α-integration. Neural Computation, 25, 1585–1604, 2013.
J. Choi and A. P. Mullhaupt, Kahlerian information geometry for signal processing. Entropy, 17, 1581–1605, 2015.
A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing. John Wiley, 2002.
A. Cichocki and S. Amari, Families of α-, β- and γ-divergences: flexible and robust measures of similarities. Entropy, 12, 1532–1568, 2010.
A. Cichocki, S. Cruces and S. Amari, Generalized α-β divergences and their application to robust nonnegative matrix factorization. Entropy, 13, 134–170, 2011.
A. Cichocki, S. Cruces and S. Amari, Log-determinant divergences revisited: α-β and γ log-det divergences. Entropy, 17, 2988–3034, 2015.
A. Cichocki, R. Zdunek, A. H. Phan and S. Amari, Nonnegative Matrix and Tensor Factorizations. John Wiley and Sons, UK, 2009.
C. Cortes and V. Vapnik, Support-vector networks. Machine Learning, 20, 273–297, 1995.
F. Cousseau, T. Ozeki and S. Amari, Dynamics of learning in multilayer perceptrons near singularities. IEEE Transactions on Neural Networks, 19, 1313–1328, 2008.
F. Critchley, P. K. Marriott and M. Salmon, Preferred point geometry and statistical manifolds. Annals of Statistics, 21, 1197–1224, 1993.
I. Csiszár, Information-type measures of difference of probability distributions and indirect observation. Studia Scientiarum Mathematicarum Hungarica, 2, 229–318, 1967.
I. Csiszár, Information measures: A critical survey. in Proceedings of 7th Conference on Information Theory, Prague, Czech Republic, 83–86, 1974.
I. Csiszár, Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Annals of Statistics, 19, 2032–2066, 1991.
I. Csiszár and G. Tusnady, Information geometry and alternating minimization procedure. In E. F. Dedewicz, et. al. (Eds.), Statistics and Decision, 205–237, Oldenburg Verlag, 1984.
Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli and Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. arXiv:1406.2572, NIPS, 2014.
A. P. Dawid, The geometry of proper scoring rules. Annals of Institute of Statistical Mathematics, 59, 77–93, 2007.
A. P. Dawid, S. Lauritzen and M. Parry, Proper local scoring rules on discrete sample spaces. Annals of Statistics, 40, 593–608, 2012.
A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal Royal Statistical Society, B, 39, 1–38, 1977.
S. Dhillon and J. A. Tropp, Matrix nearness problems with Bregman divergences. SIAM Journal on Matrix Analysis and Applications, 29, 1120–1146, 2007.
D. L. Donoho, Compressed sensing. IEEE Transactions on Information Theory, 52, 1289–1306, 2006.
D. L. Donoho and Y. Tsaig, Fast solution of L_1-norm minimization problems when the solution may be sparse. IEEE Transaction on Information Theory, 54, 4789–4812, 2008.
A. Edelman, A. A. Arias and S. T. Smith, The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications, 20, 303–353, 1998.
B. Efron, Defining the curvature of a statistical problem (with application to second order efficiency). Annals of Statistics, 3, 1189–1242, 1975.
B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression. Annals of Statistics, 32, 407–499, 2004.
S. Eguchi, Second order efficiency of minimum contrast estimators in a curved exponential family. Annals of Statistics, 11, 793–803, 1983.
S. Eguchi, O. Komori and A. Ohara, Duality of maximum entropy and minimum divergence. Entropy, 16, 3552–3572, 2014.

M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, 2010.
Y. Eldar and G. Kutyniok, Compressed Sensing. Cambridge University Press, 2012.
Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. Journal Computer and Systems Sciences, 55, 119–139, 1997.
H. Fujisawa and S. Eguchi, Robust parameter estimation with a small bias against heavy contamination. Journal Multivariate Analysis, 99, 2053–2081, 2008.
A. Fujiwara and S. Shuto, Hereditary structure in Hamiltonians: Information geometry of Ising spin chains. Physics Letters A, 374, 911–914, 2010.
K. Fukumizu, Likelihood ratio of unidentifiable models and multilayer neural networks. Annals of Statistics, 31, 833–851, 2003.
K. Fukumizu, Exponential manifold by reproducing kernel Hilbert spaces. In Algebraic and Geometric Methods in Statistics (P. Gibilisco, E. Riccomagno, M.-P. Rogantin and H. Winn Eds.), 291–306, Cambridge University Press, 2009.
K. Fukumizu and S. Kuriki, Statistics of Singular Models. Frontiers in Statistical Sciences, 7, Iwanami, 2004 (in Japanese).
S. Furuichi, An axiomatic characterization of a two-parameter extended relative entropy. Journal of Mathematical Physics, 51, 2010.
P. Gibilisco and G. Pistone, Connections on non-parametric statistical manifolds by Orlicz space geometry: infinite-dimensional analysis. Quantum Probabilities and Related Topics, 1, 325–347, 1998.
M. Girolami and B. Calderhead, Riemannian manifold Langevin and Hamiltonian Monte Carlo methods. Journal of Royal Statistical Society, B-73, 123–214, 2011.
V. P. Godambe, Estimating Functions. Oxford University Press, 1991.
M. Grasselli, Dual connections in nonparametric classical information geometry. Annals of Institute of Statistical Mathematics, 62, 873–896, 2010.
I. Grondman, L. Buşoniu, G.A.D. Lopes and R. Babuška, A survey of actor-critic reinforcement learning: Standard and natural policy gradients. IEEE Transactions on Systems, Man, and Cybernetics–Part C: Applications and Reviews, 42, 1291–1307, 2012.
P. D. Grünwald and A. P. Dawid, Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Annals of Statistics, 32, 1367–1433, 2004.
N. Hansen and A Ostermeier, Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 9, 159–195, 2001.
G. H. Hardy, J. E. Littlewood and G. Polya, Inequalities (2nd ed.). Cambridge: Cambridge University Press, 1952.
K. V. Harsha and K. S. S. Moosath, F-geometry and Amari's α-geometry on a statistical manifold. Entropy, 16, 2472–2487, 2014.
M. Hayashi and S. Watanabe, Information geometry approach to parameter estimation in Markov chains. IEEE Transactions on Information Theory, 2014.
M. Henmi and R. Kobayashi, Hooke's law in statistical manifolds and divergence. Journal Nagoya Mathematical, 159, 1–24, 2000.
G. E. Hinton, Training products of experts by minimizing contrastive divergence. Neural Computation, 14, 1771–1800, 2002.
G. E. Hinton and E. R. Salakhutdinov, Reducing the dimensionality of data with neural networks. Science, 313, 504–507, 2006.
Y. Hirose and F. Komaki, An extension of least angle regression based on the information geometry of dually flat spaces. Journal of Computational and Graphical Statistics, 19, 1007–1023, 2010.
S. W. Ho and R. W. Yeung, On the discontinuity of the Shannon information measures. IEEE Transactions on Information Theory, 55, 5362–5374, 2009.
A. Honkela, T. Raiko, M. Kuusela, M. Tornio and J. Karhunen, Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. Journal of Machine Learning Research, 11, 3235–3268, 2010.

A. Hyvärinen, Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6, 695–709, 2005.
A. Hyvärinen, Some extensions of score matching. Computational Statistics & Data Analysis, 51:2499–2512, 2007.
A. Hyvärinen, J. Karhunen and E. Oja, Independent Component Analysis. John Wiley, 2001.
T. Ichimori, On rounding off quotas to the nearest integers in the problem of apportionment methods. JSIAM Letters, 3, 21–24, 2011.
S. Ikeda, T. Tanaka and S. Amari, Stochastic reasoning, free energy, and information geometry. Neural Computation, 16, 1779–1810, 2004a.
S. Ikeda, T. Tanaka and S. Amari, Information geometry of turbo and low-density parity-check codes. IEEE Transactions on Information Theory, 50, 1097–1114, 2004b.
M. Ishikawa, Structural learning with forgetting. Neural Networks, 9, 509–521, 1996.
R. A. Jacobs, M. I. Jordan, S. J. Nolwan and G. E. Hinton, Adaptive mixtures of local experts. Neural Computation, 3, 79–87, 1991.
A. T. James, The variance information manifold and the function on it. Multivariate Statistical Analysis, Ed. P. K. Krishnaiah, Academic Press, 157–169, 1973.
H. Jeffreys, Theory of Probability, 1st ed. Clarendon Press, 1939.
H. Jeffreys, An invariant form for the prior probability in estimation problems. Proceedings of Royal Society of London, Series A, Mathematical and Physical Sciences, 186, 453–461, 1946.
H. Jeffreys, Theory of Probability, 2nd ed. Oxford University Press, 1948.
K. Jeong, M. Yukawa and S. Amari, Can critical-point paths under lp-regularization $(0\langle p\langle 1)$ reach the sparsest least square solutions?. IEEE Transactions on Information Theory, 60, 2960–2968, 2014.
J. Jiao, T. M. Courtade, A. No, K. Venkat and T. Weissman, Information measure: The curious case of the binary alphabet. IEEE Transactions on Information Theory, 60, 7616–7626, 2015.
S. Kakade, A natural policy gradient. In Advances in Neural Information Processing, 14, 1531–1538, 2001.
S. Kakihara, A. Ohara and T. Tsuchiya, Information geometry and interior-point algorithms in semi-definite programs and symmetric cone programs. Journal of Optimization Theory and Applications, DOI 10.1007/s10957-012-0189-9, 2012.
T. Kanamori, T. Takenouchi, S. Eguchi and N. Murata, Robust loss function for boosting. Neural Computation, 19, 2183–2244, 2007.
K. Kanatani, Statistical optimization and geometric inference in computer vision. Philosophical Transactions of Royal Society of London, Ser. A, 356, 1303–1320, 1998.
Y. Kano, Beyond third-order efficiency. Sankhya, 59, 179–197, 1997.
Y. Kano, More higher order efficiency. Journal of Multivariate Analysis. 67, 349–366, 1998.
R. Karakida, M. Okada and S. Amari, Analyzing feature extraction by contrastive divergence learning in RBM. NIPS Workshop on Deep Learning, 2014.
R. Karakida, M. Okada and S. Amari, Dynamical analysis of contrastive divergence learning. Restricted Boltzmann machines with Gaussian visible units, To appear, 2016.
R. E. Kass and P. Vos, Geometrical Foundations of Asymptotic Inference. Wiley, 1997.
A. Kim, J. Park S. Park and S. Kang, Impedance learning for robotic contact tasks using natural actor-critic algorithm. IEEE Transactions on Systems, Man and Machine, B39, 433–443, 2010.
J. Kivinen and M. K. Warmuth, Exponentiated gradient versus gradient descent for linear predictors. Information and Computation, 132, 1–63, 1997.
G. Kniadakis and A. Scarfone, A new one parameter deformation of the exponential function. Physica A, 305, 69–75, 2002.
K. Kumon and S. Amari, Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference. Proceedings of Royal Society of London, A 387, 429–458, 1983.
M. Kumon, A. Takemura and K. Takeuchi, Conformal geometry of statistical manifold with application to sequential estimation. Sequential Analysis, 30, 308–337, 2011.
T. Kurose, Dual connections and affine geometry. Mathematische Zeitschrift, 203, 115–121, 1990.

T. Kurose, On the divergence of 1-conformally flat statistical manifolds. Tohoku Mathematical Journal, 46, 427–433, 1994.
T. Kurose, Conformal-projective geometry of statistical manifolds. Interdisciplinary Information Sciences, 8, 89–100, 2002.
S. Lauritzen, Graphical Models. Oxford University Press, 1996.
H. V. Lê, Statistical manifolds are statistical models. Journal of Geometry, 84, 83–93, 2005.
G. Lebanon and J. Lafferty, Boosting and maximum likelihood for exponential models. In Advances in Neural Information Processing Systems (NIPS), 14, 2001.
D. D. Lee and S. Seung, Algorithms for nonnegative matrix factorization. Nature, 401, 788–791, 1999.
C. Lin and J. Jiang, Supervised optimizing kernel locality preserving projection with its application to face recognition and palm biometrics. Submitted, 2015.
M. Liu, B. C. Vemuri, S. Amari and F. Nielsen, Shape retrieval using hierarchical total Bregman soft clustering. IEEE Transactions on Pattern Analysis and Machine Learning, 34, 2407–2419, 2012.
L. Malagò, M. Matteucci and G. Pistone, Natural gradient, fitness modelling and model selection: A unifying perspective. IEEE Congress on Evolutionary Computation, 486–493, 2013.
L. Malagò and G. Pistone, Combinatorial optimization with information geometry: Newton method. Entropy 16, 4260–4289, 2014.
P. Marriott, On the local geometry of mixture models. Biometrika, 89, 77–93, 2002.
P. Marriott and M. Salmon, Applications of Differential Geometry to Econometrics. Academic Press, 2011.
J. Martens, New perspectives on the natural gradient method. arXiv:1412.1193, 2015.
J. Martens and R. Grosse, Optimizing neural networks with Kronecker-factored approximate curvature. arXiv:1503.05671, 2015.
R. J. Martin, A metric for ARMA processes. IEEE Transactions on Signal Processing, 48, 1164–1170, 2000.
T. Matumoto, Any statistical manifold has a contrast function—On the C3-functions taking the minimum at the diagonal of the product manifold. Hiroshima Mathematical Journal, 23, 327–332, 1993.
Y. Matsuyama, The α-EM algorithm: Surrogate likelihood maximization using α-logarithmic information measures. IEEE Transactions on Information Theory, 49, 692–706, 2003.
H. Matsuzoe, On realization of conformally-projectively flat statistical manifolds. Hokkaido Mathematical Journal, 27, 409–421, 1998.
H. Matsuzoe, Geometry of contrast functions and conformal geometry. Hokkaido Mathematical Journal, 29, 175–191, 1999.
H. Matsuzoe, J. Takeuchi and S. Amari, Equiaffine structures on statistical manifolds and Bayesian statistics. Differential Geometry and Its Applications, 24, 567–578, 2006.
J. Milnor, On the concept of attractor. Communications of Mathematical Physics, 99, 177–195, 1985.
M. Minami and S. Eguchi, Robust blind source separation by β-divergence. Neural Computation, 14, 1859–1886, 2004.
K. Miura, M. Okada and S. Amari, Estimating spiking irregularities under changing environments. Neural Computation, 18, 2359–2386, 2006.
T. Morimoto, Markov processes and the H-theorem. Journal of Physical Society of Japan, 12, 328–331, 1963.
T. Morimura, E. Uchibe, J. Yoshimoto and K. Doya, A generalized natural actor-critic algorithm. In Advances in Neural Information Processing Systems, 22, MIT Press, 1312–1320, 2009.
R. Morioka and K. Tsuda, Information geometry of input-output table. Technical Report IEICE, 110, 161–168, 2011 (in Japanese).
N. Murata, T. Takenouchi, T. Kanamori and S. Eguchi, Information geometry of U-boost and Bregman divergence. Neural Computation, 16, 1432–1481, 2004.
M. K. Murray and J. W. Rice, Differential Geometry and Statistics. Chapman Hall, 1993.

H. Nagaoka and S. Amari, Differential geometry of smooth families of probability distributions. Technical Report METR 82–7, University of Tokyo, 1982.
H. Nakahara and S. Amari, Information-geometric measure for neural spikes. Neural Computation, 14, 2269–2316, 2002.
H. Nakahara, S. Amari and B. Richmond, A comparison of descriptive models of a single spike train by information-geometric measure. Neural Computation, 18, 545–568, 2006.
J. Naudts, Generalized Thermostatistics. Springer, 2011.
A. Nemirovski and D. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley, 1983.
Y. Nesterov and A. Nemirovski, Interior Point Polynomial Methods in Convex Programming: Theory and Algorithms. SIAM Publications, 1993.
Y. Nesterov and M. Todd, On the Riemannian geometry defined by self-concordant barriers and interior-point methods. Foundations of Computational Mathematics, 2, 333–361, 2002.
N. J. Newton, An infinite-dimensional statistical manifold modeled on Hilbert space. Journal of Functional Analysis, 263, 1661–1681, 2012.
J. Neyman and E. L. Scott, Consistent estimates based on partially consistent observation. Econometrica, 16, 1–32, 1948.
F. Nielsen and R. Nock, On the χ-square and higher-order χ-distances for approximating f-divergences. IEEE Signal Processing Letters, 21, 10–13, 2014.
R. Nock and F. Nielsen, Bregman divergences and surrogates for learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, 2048–2059, 2009.
R. Nock, F. Nielsen and S. Amari, On conformal divergences and their population minimizers. IEEE Transactions on Information Theory, accepted, 2015.
K. Nomizu and T. Sasaki, Affine Differential Geometry. Oxford University Press, 1994.
A. Ohara, Information geometric analysis of an interior point method for semidefinite programming. In O. Barndorff-Nielsen and E. Jensen Eds, Geometry in Present Day Science, World Scientific, 49–74, 1999.
A. Ohara, Geometry of distributions associated with Tsallis statistics and properties of relative entropy minimization. Physics Letters, A, 370, 184–193, 2007.
A. Ohara and S. Amari, Differential geometric structures of stable state feedback systems with dual connections. Kybernetika, 30, 369–386, 1994.
A. Ohara and S. Eguchi, Group invariance of information geometry on q-Gaussian distributions induced by beta-divergence. Entropy, 15, 4732–4747, 2013.
A. Ohara, H. Matsuzoe and S. Amari, Conformal geometry of escort probability and its applications. Modern Physics Letters B, 26, 10, 1250063, 2012.
M. Oizumi, L. Albantakis and G. Tononi, From phenomenology to the mechanism of consciousness: Integrated information theory 3.0. PLoS Computational Biology, 10, e1003588, 2014.
M. Oizumi, S. Amari, T. Yanagawa, N. Fujii and N. Tsuchiya, Measuring integrated information from the decoding perspective. arXiv:1505.04368 [q-bio.NC], To appear in PLoS Computational Biology, 2015.
M. Oizumi, M. Okada and S. Amari, Information loss associated with imperfect observation and mismatched decoding. Frontiers in Computational Neuroscience, 5, 1–13, 2011.
M. Oizumi, N. Tsuchiya and S. Amari, A unified framework for information integration based on information geometry. Submitted, 2016.
E. Oja, A simplified neuron model as a principal component analyzer. Journal of Mathematical Biology, 15, 267–273, 1982.
E. Oja, Principal components, minor components, and linear neural networks. Neural Networks, 5, 927–935, 1992.
I. Okamoto, S. Amari and K. Takeuchi, Asymptotic theory of sequential estimation: Differential-geometrical approach. Annals of Statistics, 19, 961–981, 1991.
T. Okatani and K. Deguchi, Easy calibration of a multi-projector display system. International Journal of Computer Vision, 2009.

Y. Ollivier, Riemannian metric for neural networks I: Feedforward networks. Information and Inference, 4, 108-153, 2015, DOI 10.1093/imaiai/iav006.
H. Park, S. Amari and K. Fukumizu, Adaptive natural gradient learning algorithms for various stochastic models. Neural Networks, 13, 755–764, 2000.
M. Parry, A. P. Dawid and S. Lauritzen, Proper local scoring rule. Annals of Statistics, 40, 561–592, 2012.
J. Pearl, Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988.
J. Peters and S. Schaal, Natural actor-critic. Neurocomputing, 71, 1180–1190, 2008.
G. Pistone, Examples of the application of nonparametric information geometry to statistical physics. Entropy, 15, 4042–4065, 2013.
G. Pistone and M. P. Rogantin, The exponential statistical manifold: mean parameters, orthogonality and space transformations. Bernoulli, 5, 721–760, 1999.
G. Pistone and C. Sempi, An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Annals of Statistics, 23, 1543–1561, 1995.
C. R. Rao, Information and accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society, 37, 81–91, 1945.
C. R. Rao, Efficient estimates and optimum inference procedures in large samples. Journal of Royal Statistical Society, B, 24, 46–72, 1962.
G. Raskutti and S. Mukherjee, The information geometry of mirror descent. IEEE Transactions on Information Theory, 61, 1451–1457, 2015.
J. Rauh, Finding the maximizers of the information divergence from an exponential family. IEEE Transactions on Information Theory, 57, 3236–3247, 2011.
N. Ravishanker, E. L. Melnik and C. Tsai, Differential geometry of ARMA models. Journal of Time Series Analysis, 11, 259–274, 1990.
A. Rényi, On measures of entropy and information, in Proc. 4th Symposium on Mathematical Statistics and Probability Theory, Berkeley, CA, 1, 547–561, 1961.
F. Rosenblatt, Principles of Neurodynamics. Spartan, 1962.
N. L. Roux, P.-A. Manzagol and Y. Bengio, Topmoumoute online natural gradient algorithm. In Advances in Neural Information Processing Systems, 17, 849–856, 2007.
D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors. Nature, 323, 533–536, 1986.
R. E. Schapire, Y. Freend, P. Bartlett and W. S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 26, 1651–1686, 1998.
J. Schmidhuber, Deep Learning in Neural Networks: An Overview. Neural Networks, 61, 85–117, 2015.
B. Scholkopf, Support Vector Learning. Oldenbourg, 1997.
J. A. Schouten, Ricci Calculus. Springer, 1954.
J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
H. Shima, The Geometry of Hessian Structures. World Scientific, 2007.
S. Shinomoto, K. Shima and J. Tanji, Differences in spiking patterns among cortical neurons. Neural Computation, 15, 2823–2842, 2003.
P. Smolensky, Information processing in dynamical systems: Foundations of harmony theory, In D. E. Rumelhart and J. L. McLelland (Eds.), Parallel Distributed Processing, 1, 194–281, MIT Press, 1986.
A. Soriano and L. Vergara, Fusion of scores in a detection context based on alpha integration. Neural Computation, 27, 1983–2010, 2015.
S. M. Stigler, The epic story of maximum likelihood. Statistical Science, 22, 598–620, 2007.
T. Takenouchi and S. Eguchi, Robustifying AdaBoost by adding the naive error rate. Neural Computation 16, 767–787, 2004.
T. Takenouchi, S. Eguchi, N. Murata and T. Kanamori, Robust boosting algorithm against mislabeling in multiclass problems. Neural Computation, 20, 1596–1630, 2008.

J. Takeuchi, Geometry of Markov chains, finite state machines and tree models. Technical Report of IEICE, 2014.
J. Takeuchi, T. Kawabata and A. Barron, Properties of Jeffreys mixture for Markov sources. IEEE Transactions on Information Theory, 41, 643–652, 2013.
K. Tanabe, Geometric method in nonlinear programming. Journal of Optimization Theory and Applications, 30, 181–210, 1980.
T. Tanaka, Information geometry of mean field approximation. Neural Computation, 12, 1951–1968, 2000.
M. Taniguchi, Higher-Order Asymptotic Theory for Time Series Analysis. Springer Lecture Notes in Statistics, 68, 1991.
P. S. Thomas, W. Dabney, S. Mahadeven and S. Giguere, Projected natural actor-critic. In Advances in Neural Information Processing Systems, 26, 2013.
R. Tibshirani, Regression shrinkage and selection via the LASSO. Journal of Royal Statistical Society, Series B, 58, 267–288, 1996.
G. Tononi, Consciousness as integrated information: a provisional manifest. Biological Bulletin, 215, 216–242, 2008.
F. Topsoe, Information-theoretical optimization techniques. Kybernetika, 15, 8–27, 1979.
C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics. Journal of Statistical Physics, 52, 479–487, 1988.
C. Tsallis, Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World, Springer, 2009.
K. Uohashi, α-conformal equivalence of statistical submanifolds. Journal of Geometry, 75, 179–184, 2002.
V. N. Vapnik, Statistical Learning Theory. John Wiley, 1998.
B. C. Vemuri, M. Liu, S. Amari and F. Nielsen, Total Bregman divergence and its applications to DTI analysis. IEEE Transactions on Medical Imaging, 30, 475–483, 2011.
R. F. Vigelis, and C. C. Cavalcante, On ϕ-families of probability distributions. Journal of Theoretical Probabilities, 26, 870–884, 2013.
P. Vos, A geometric approach to detecting influential cases. Annals of Statistics, 19, 1570–1581, 1991.
J. Wada, A divisor apportionment method based on the Kolm-Atkinson social welfare function and generalized entropy. Mathematical Social Sciences, 63, 243–247, 2012.
M. J. Wainwright and M. I. Jordan, Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1, 1–305, 2008.
S. Watanabe, Algebraic geometrical methods for hierarchical learning machines. Neural Networks, 14, 1409–1060, 2001.
S. Watanabe, Algebraic Geometry and Statistical Learning Theory. Cambridge University Press, 2009.
S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular statistical learning theory. Journal of Machine Learning Research, 11, 3571–3591, 2010.
H. Wei and S. Amari, Dynamics of learning near singularities in radial basis function networks. Neural Networks, 21, 989–1005, 2008.
H. Wei, J. Zhang, F. Cousseau, T. Ozeki and S. Amari, Dynamics of learning near singularities in layered networks. Neural Computation, 20, 813–843, 2008.
P. Williams, S. Wu and J. Feng, Two scaling methods to improve performance of the support vector machine. In Support Vector Machines: Theory and Applications, Ed. L. Wang, 205–218, Springer, 2005.
D. Wu, Parameter estimation for α-GMM based on maximum likelihood criterion. Neural Computation, 21, 1776–1795, 2009.
S. Wu and S. Amari, Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers. Neural Processing Letters, 15, 59–67, 2002.

S. Wu, S. Amari and H. Nakahara, Population coding and decoding in a neural field: A computational study. Neural Computation, 14, 999–1026, 2002.
L. Xu, Least mean square error reconstruction principle for self-organizing neural nets. Neural Networks, 6, 627–648, 1993.
Z. Xu, X. Chang, F. Xu and H. Zhang, $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. IEEE Transactions on Neural Networks and Learning Systems, 23, 1013–1027, 2012.
S. Yi, D. Wierstra, T. Schaul and J. Schmidhuber, Stochastic search using the natural gradient. ICML Proceedings of the 26th Annual Internatinal Conference on Machine Learning, 1161–1168, 2009.
J. S. Yedidia, W. T. Freeman and Y. Weiss, Generalized belief propagation. In T. K. Leen, T. G. Dietrich and V. Tresp (Eds.), Advances in Neural Information Processing Systems, 13, 689–695, MIT Press, 2001.
A. Yuille, CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. Neural Computation, 14, 1691–1722, 2002.
A. L. Yuille and A. Rangarajan, The concave-convex procedure. Neural Computation, 15, 915–936, 2003.
M. Yukawa and S. Amari, l_p-regularized least squares $(0 < p < 1)$ and critical path. IEEE Transactions on Information Theory, 62, 1–15, 2016.
J. Zhang, Divergence function, duality and convex analysis. Neural Computation, 16, 159–195, 2004.
J. Zhang, From divergence function to information geometry: Metric, equiaffine and symplectic structures. Geometry Symposium, Japan Mathematical Society, Proceedings, 47–62, 2011.
J. Zhang, Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. Entropy, 15, 5384–5418, 2013.
J. Zhang, On monotone embedding in information geometry. Entropy, 17, 4485–4499, 2015.
J. Zhao, H. Wei, C. Zhang, W. Li, W. Guo and K. Zhang, Natural gradient learning algorithms for RBF networks. Neural Computation, 27, 481–505, 2015.
H. Y. Zhu and R. Rohwer, Bayesian invariant measurements of generalization. Neural Processing Letters, 2, 28–31, 1995.

Index

Symbols
(α-β)-divergence, 100
(α, β)-divergence, 94
(α-β)-log-det divergence, 101
α-divergence, 58, 67, 72
α-expert machine, 84
α-family of probability distributions, 81
α-function, 57
α-geodesic, 75
α-geometry, 136
α-integration, 82
α-mean, 77
α-projection theorem, 76
α-Pythagorean theorem, 76
β-divergence, 95
χ-divergence, 91
χ-escort distribution, 91
χ-exponential family, 90
e-affine parameter, 38
e-condition, 257
e-flat, 38
e-geodesic, 38
e-parallel transport, 202
em algorithm, 28
(F, G, H)-structure, 104
f-divergence, 54
γ-divergence, 102
k-cut, 144
k-means, 234
k-sparse, 336
κ-exponential family, 89
L_0-norm, 338
L_1-norm, 338
m-affine parameter, 38
m-condition, 257
m-flat, 38
m-geodesic, 38
m-parallel transport, 202
m-projection, 46, 252
ϕ-center of cluster, 233
Φ-function method, 245
q-divergence, 85
q-entropy, 85
q-escort geometry, 92
q-exponential, 85
q-exponential family, 86, 89
q-free energy, 87
q-logarithm, 85
q-metric, 88
(ρ, τ)-structure, 104
U-divergence, 95
(u, v)-divergence, 92
(u, v)-structure, 99

A
Absolute-value-based Hessian natural gradient, 286
Active set, 340
Adaptive learning method, 294
Adaptive natural gradient learning, 293
Affine connection, 112
Affine coordinate system, 18
Affine flat structure, 19
Akaike information criterion, 312
Alternating minimization algorithm, 27
Amari–Chentsov structure, 134
Amari–Chentsov tensor, 134
Ancillary submanifold, 169
Ancillary tangent subspace, 201
ARMA model, 218
AR model, 217
Asymptotic theory of hypothesis testing, 175
Auto-correlation coefficients, 221

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

Auto-regression model, 217

B
Back-propagation learning, 281
Barrier function, 345
Basis vectors, 20
Bayesian duality, 266
Bayesian posterior distribution, 266
Belief propagation, 249
Blow-down technique, 309
Boltzmann machine, 181, 268
Boosting, 261
Bregman divergence, 13

C
Canonical divergence, 138
Canonical parameter, 32
Central limit theorem, 60
Chernoff divergence, 242
Chernoff information, 242
Clique, 250
Clustering, 231
Clustering algorithm, 234
Coarse graining, 53
Cocktail party problem, 323
Coefficient of proportionality, 191
Coefficients of affine connection, 113
Conformal transformation, 91
Conformal transformation of a kernel, 249
Conjugate priors, 267
Consistent estimator, 168
Contrastive divergence, 273
Convex-concave computational procedure, 249
Convex function, 12
Convex programming, 345
Coordinate system, 4
Coordinate transformation, 4
Covariant derivative, 117
Cramér–Rao bound, 166
Cramér–Rao theorem, 166
Critical region, 300
Critical slowdown, 305
Cubic tensor, 115
Cumulant generating function, 32

D
Decomposable divergence, 55
Deep learning, 292, 296
Deformed exponential family, 89
Divergence, 9
Dual affine structure, 19
Dual connections, 131
Dual convex function, 17
Dual geodesic, 19
Dually flat manifold, 137

E
Efficient, 173
Efficient score, 194
Einstein summation convention, 20
Eliminating singularity, 299
EM algorithm, 179
Embedding curvature, 129, 349
Ergodic time series, 215
Error covariance matrix, 166
Escort probability distribution, 88
Estimating function, 197
Estimator, 165
Euler–Schouten curvature, 129
Exponential family, 31

F
First-order asymptotic theory, 173
Fisher information matrix, 33
Foliation, 145
Free energy, 32

G
Game, 349
Game-divergence, 350
Game-score, 349
Gaussian kernel, 247
Gaussian mixture model, 180
Gaussian RBM, 275
Generalization error, 280
Generalized inverse, 337
Generalized Pythagorean theorem, 24
Geodesic, 19, 117
Graph Laplacian, 356
Graphical model, 250

H
Hidden variable, 179
Higher-order asymptotic theory of estimation, 173
Higher-order correlations, 149
Higher-order cumulants, 326
Hyvärinen divergence, 354
Hyvärinen score, 353

I
Independent component analysis, 322
Information integration, 150
Information monotonicity, 52
Inner product, 23
Input–output analysis, 157
Instantaneous loss, 280
Integrated information, 152
Integration of weak machines, 261
Invariance criterion, 51
Invariant divergences, 52
Invariant Riemannian metrics, 52

K
Kernel exponential family, 42
Kernel function, 246
Killing metric, 329
KL-divergence, 71, 220
Kronecker-factored approximate curvature, 293
Kullback–Leibler (KL) divergence, 11

L
Large deviation, 60
Large deviation theorem, 61
Learning constant, 294
Least angle regressions, 343
Least equiangle theorem, 342
Legendre transformation, 16
Levi–Civita connection, 113, 125
Linear machine, 242
Linear system, 215
Loss of information by data reduction, 185

M
Machine learning, 231
MA model, 218
Manifold, 3
Margin, 243
Maximum entropy, 223
Maximum entropy principle, 45
Maximum likelihood estimator, 48
Mean field approximation, 254
Metric affine connection, 125
Milnor attractor, 310
Minimum description length, 312
Minimum entropy, 224
Minkovskian gradient, 343
Minor subspace, 317
Mirror descent method, 289
Misspecified model, 186
Mixed coordinate system, 144
Mixture family, 37
Moving-average model, 218
Multilayer perceptron, 292, 296

N
Natural gradient, 283
Natural gradient learning method, 284
Natural parameter, 32
Natural policy gradient, 288
Negative entropy, 33
Neyman–Scott problem, 191
Non-holonomic coordinate system, 329
Non-negative matrix factorization, 333
Nuisance parameter, 191
Nuisance tangent subspace, 201

O
Observed point, 47
Observed submanifold, 180
On-line learning, 281
Overlapping singularity, 299

P
Parallel transport, 22, 118
Parameter of interest, 191
Plateau, 302
Plateau phenomena, 308
Policy natural gradient, 284
Polynomial kernel, 247
Positive-definite symmetric matrix, 96
Power spectrum, 217
Principal component, 317
Principal component analysis, 315
Principal subspace, 317
Prior distribution, 266
Projection theorem, 25, 143

R
RAS transformation, 160
RC curvature, 119
Reinforcement learning, 287
Restricted Boltzmann machine, 268
Riemann–Christoffel curvature tensor, 119
Riemannian connection, 113
Riemannian geometry, 109
Riemannian gradient, 283
Riemannian metric, 19
Riemannian structure, 10
Robust cluster center, 238

S
Saddle-free Newton method, 286
Scale problem, 191
Score function, 111
Semi-definite programming, 346
Shape parameter, 212
Singular point, 301
Singular prior, 313
Singular statistical models, 311
Singular structure, 296
Soft clustering, 236
Solution path, 341
Sparse vector, 336
Standard estimating function, 332
Standard f-divergence, 56
Stiefel manifold, 320
Stochastic descent learning method, 281
Stochastic relaxation, 286
Submanifold, 126
Sufficient statistic, 52
Super efficiency, 332
Support vector, 244
Support vector machine, 242
System complexity, 152

T
Tangent space, 19, 109
Tangent subspace of interest, 201
Temporal firing pattern, 211
Tensor, 114
Time series, 215
Total Bregman divergence, 238
Total least squares, 196
Training error, 280
Transfer function, 217

U
Unidentifiability, 298

V
Voronoi diagram, 234