

# Applied Mathematical Sciences

Volume 194

## Editors

S.S. Antman, Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA

ssa@math.umd.edu

Leslie Greengard, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Greengard@cims.nyu.edu

P.J. Holmes, Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

pholmes@math.princeton.edu

## Advisors

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

P. Constantin, Department of Mathematics, Princeton University, Princeton, NJ, USA

J. Keller, Department of Mathematics, Stanford University, Stanford, CA, USA

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

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

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

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

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

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

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

## Founding Editors

Fritz John, Joseph P. LaSalle and Lawrence Sirovich

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

Shun-ichi Amari

# Information Geometry and Its Applications

 Springer

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

ISSN 0066-5452

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

© Springer Japan 2016

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

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

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

Printed on acid-free paper

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

# Preface

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

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

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

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

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

Allow me to review my own personal history in information geometry. It was in 1958, when I was a graduate student on a master's course, that I followed a seminar on statistics. The text was "Information Theory and Statistics" by S. Kullback, and

a professor suggested to me that the Fisher information might be regarded as a Riemannian metric. I calculated the Riemannian metric and curvature of the manifold of Gaussian distributions and found that it is a manifold of constant curvature, which is no different from the famous Poincaré half-plane in non-Euclidean geometry. I was enchanted by its beauty. I believed that a beautiful structure must have important practical significance, but I was not able to pursue its consequences further.

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

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

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

Last but not least, I would like to thank Dr. M. Kumon and Prof. H. Nagaoka, who collaborated in the early period of the infancy of information geometry. I also thank the many researchers who have supported me in the process of construction of information geometry, Profs. D. Cox, C.R. Rao, O. Barndorff-Nielsen, S. Lauritzen, B. Efron, A.P. Dawid, K. Takeuchi, and the late N.N. Chentsov, among many many others. Finally, I would like to thank Ms. Emi Namioka who arranged my handwritten manuscripts in the beautiful  $\text{\TeX}$  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

<b>1</b>	<b>Manifold, Divergence and Dually Flat Structure</b> . . . . .	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
<b>2</b>	<b>Exponential Families and Mixture Families of Probability Distributions</b> . . . . .	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
<b>3</b>	<b>Invariant Geometry of Manifold of Probability Distributions . . . . .</b>	<b>51</b>
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	$\chi^2$ -Divergence . . . . .	57
3.3.3	$\alpha$ -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
<b>4</b>	<b><math>\alpha</math>-Geometry, Tsallis <math>q</math>-Entropy and Positive-Definite Matrices . . . . .</b>	<b>71</b>
4.1	Invariant and Flat Divergence . . . . .	71
4.1.1	KL-Divergence Is Unique . . . . .	71
4.1.2	$\alpha$ -Divergence Is Unique in $R_+^n$ . . . . .	72
4.2	$\alpha$ -Geometry in $S_n$ and $R_+^n$ . . . . .	75
4.2.1	$\alpha$ -Geodesic and $\alpha$ -Pythagorean Theorem in $R_+^n$ . . . . .	75
4.2.2	$\alpha$ -Geodesic in $S_n$ . . . . .	76
4.2.3	$\alpha$ -Pythagorean Theorem and $\alpha$ -Projection Theorem in $S_n$ . . . . .	76
4.2.4	Apportionment Due to $\alpha$ -Divergence . . . . .	77
4.2.5	$\alpha$ -Mean . . . . .	77
4.2.6	$\alpha$ -Families of Probability Distributions . . . . .	80
4.2.7	Optimality of $\alpha$ -Integration . . . . .	82
4.2.8	Application to $\alpha$ -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 ( $\alpha$ -Family) of Probability Distributions . . . . . 86
  - 4.3.3  $q$ -Escort Geometry . . . . . 87
  - 4.3.4 Deformed Exponential Family:  $\chi$ -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  $R^p_+$  . . . . . 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  $\gamma$ -Divergence . . . . . 102
  - 4.6.2 Other Types of  $(\alpha, \beta)$ -Divergences . . . . . 102
  - 4.6.3 Burbea–Rao Divergence and Jensen–Shannon Divergence . . . . . 103
  - 4.6.4  $(\rho, \tau)$ -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

<b>6</b>	<b>Dual Affine Connections and Dually Flat Manifold</b> . . . . .	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	$\alpha$ -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
<b>Part III Information Geometry of Statistical Inference</b>		
<b>7</b>	<b>Asymptotic Theory of Statistical Inference</b> . . . . .	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
<b>8</b>	<b>Estimation in the Presence of Hidden Variables.</b> . . . . .	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
<b>9</b>	<b>Neyman-Scott Problem: Estimating Function and Semiparametric Statistical Model</b> . . . . .	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
	<b>References</b> . . . . .	359
	<b>Index</b> . . . . .	371