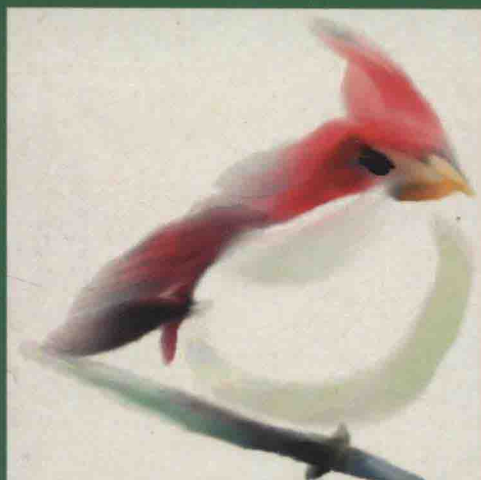
Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

STATISTICAL REINFORCEMENT LEARNING

Modern Machine Learning Approaches



Masashi Sugiyama

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

STATISTICAL REINFORCEMENT LEARNING

Modern Machine Learning Approaches

Masashi Sugiyama
University of Tokyo
Tokyo, Japan

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20150128

International Standard Book Number-13: 978-1-4398-5689-5 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

STATISTICAL REINFORCEMENT LEARNING

Modern Machine Learning Approaches

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

AIMS AND SCOPE

This series reflects the latest advances and applications in machine learning and pattern recognition through the publication of a broad range of reference works, textbooks, and handbooks. The inclusion of concrete examples, applications, and methods is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of machine learning, pattern recognition, computational intelligence, robotics, computational/statistical learning theory, natural language processing, computer vision, game AI, game theory, neural networks, computational neuroscience, and other relevant topics, such as machine learning applied to bioinformatics or cognitive science, which might be proposed by potential contributors.

PUBLISHED TITLES

BAYESIAN PROGRAMMING
Pierre Bessière, Emmanuel Mazer, Juan-Manuel Ahuactzin, and Kamel Mekhnacha

UTILITY-BASED LEARNING FROM DATA
Craig Friedman and Sven Sandow

HANDBOOK OF NATURAL LANGUAGE PROCESSING, SECOND EDITION
Nitin Indurkhya and Fred J. Damerau

COST-SENSITIVE MACHINE LEARNING
Balaji Krishnapuram, Shipeng Yu, and Bharat Rao

COMPUTATIONAL TRUST MODELS AND MACHINE LEARNING
Xin Liu, Anwitaman Datta, and Ee-Peng Lim

MULTILINEAR SUBSPACE LEARNING: DIMENSIONALITY REDUCTION OF MULTIDIMENSIONAL DATA
Haiping Lu, Konstantinos N. Plataniotis, and Anastasios N. Venetsanopoulos

MACHINE LEARNING: An Algorithmic Perspective, Second Edition
Stephen Marsland

SPARSE MODELING: THEORY, ALGORITHMS, AND APPLICATIONS
Irina Rish and Genady Ya. Grabarnik

A FIRST COURSE IN MACHINE LEARNING
Simon Rogers and Mark Girolami

STATISTICAL REINFORCEMENT LEARNING: MODERN MACHINE LEARNING APPROACHES
Masashi Sugiyama

MULTI-LABEL DIMENSIONALITY REDUCTION
Liang Sun, Shuiwang Ji, and Jieping Ye

REGULARIZATION, OPTIMIZATION, KERNELS, AND SUPPORT VECTOR MACHINES
Johan A. K. Suykens, Marco Signoretto, and Andreas Argyriou

ENSEMBLE METHODS: FOUNDATIONS AND ALGORITHMS
Zhi-Hua Zhou

Foreword

How can agents learn from experience without an omniscient teacher explicitly telling them what to do? Reinforcement learning is the area within machine learning that investigates how an agent can learn an optimal behavior by correlating generic reward signals with its past actions. The discipline draws upon and connects key ideas from behavioral psychology, economics, control theory, operations research, and other disparate fields to model the learning process. In reinforcement learning, the environment is typically modeled as a Markov decision process that provides immediate reward and state information to the agent. However, the agent does not have access to the transition structure of the environment and needs to learn how to choose appropriate actions to maximize its overall reward over time.

This book by Prof. Masashi Sugiyama covers the range of reinforcement learning algorithms from a fresh, modern perspective. With a focus on the statistical properties of estimating parameters for reinforcement learning, the book relates a number of different approaches across the gamut of learning scenarios. The algorithms are divided into model-free approaches that do not explicitly model the dynamics of the environment, and model-based approaches that construct descriptive process models for the environment. Within each of these categories, there are policy iteration algorithms which estimate value functions, and policy search algorithms which directly manipulate policy parameters.

For each of these different reinforcement learning scenarios, the book meticulously lays out the associated optimization problems. A careful analysis is given for each of these cases, with an emphasis on understanding the statistical properties of the resulting estimators and learned parameters. Each chapter contains illustrative examples of applications of these algorithms, with quantitative comparisons between the different techniques. These examples are drawn from a variety of practical problems, including robot motion control and Asian brush painting.

In summary, the book provides a thought provoking statistical treatment of reinforcement learning algorithms, reflecting the author's work and sustained research in this area. It is a contemporary and welcome addition to the rapidly growing machine learning literature. Both beginner students and experienced researchers will find it to be an important source for understanding the latest reinforcement learning techniques.

Daniel D. Lee
GRASP Laboratory
School of Engineering and Applied Science
University of Pennsylvania, Philadelphia, PA, USA

Preface

In the coming *big data* era, *statistics* and *machine learning* are becoming indispensable tools for data mining. Depending on the type of data analysis, machine learning methods are categorized into three groups:

- **Supervised learning:** Given input-output paired data, the objective of supervised learning is to analyze the input-output relation behind the data. Typical tasks of supervised learning include *regression* (predicting the real value), *classification* (predicting the category), and *ranking* (predicting the order). Supervised learning is the most common data analysis and has been extensively studied in the statistics community for long time. A recent trend of supervised learning research in the machine learning community is to utilize side information in addition to the input-output paired data to further improve the prediction accuracy. For example, *semi-supervised learning* utilizes additional input-only data, *transfer learning* borrows data from other similar learning tasks, and *multi-task learning* solves multiple related learning tasks simultaneously.
- **Unsupervised learning:** Given input-only data, the objective of unsupervised learning is to find something useful in the data. Due to this ambiguous definition, unsupervised learning research tends to be more ad hoc than supervised learning. Nevertheless, unsupervised learning is regarded as one of the most important tools in data mining because of its automatic and inexpensive nature. Typical tasks of unsupervised learning include *clustering* (grouping the data based on their similarity), *density estimation* (estimating the probability distribution behind the data), *anomaly detection* (removing outliers from the data), *data visualization* (reducing the dimensionality of the data to 1–3 dimensions), and *blind source separation* (extracting the original source signals from their mixtures). Also, unsupervised learning methods are sometimes used as data pre-processing tools in supervised learning.
- **Reinforcement learning:** Supervised learning is a sound approach, but collecting input-output paired data is often too expensive. Unsupervised learning is inexpensive to perform, but it tends to be ad hoc. Reinforcement learning is placed between supervised learning and unsupervised learning — no explicit supervision (output data) is provided, but we still want to learn the input-output relation behind the data. Instead of output data, reinforcement learning utilizes *rewards*, which

evaluate the validity of predicted outputs. Giving implicit supervision such as rewards is usually much easier and less costly than giving explicit supervision, and therefore reinforcement learning can be a vital approach in modern data analysis. Various supervised and unsupervised learning techniques are also utilized in the framework of reinforcement learning.

This book is devoted to introducing fundamental concepts and practical algorithms of statistical reinforcement learning from the modern machine learning viewpoint. Various illustrative examples, mainly in robotics, are also provided to help understand the intuition and usefulness of reinforcement learning techniques. Target readers are graduate-level students in computer science and applied statistics as well as researchers and engineers in related fields. Basic knowledge of probability and statistics, linear algebra, and elementary calculus is assumed.

Machine learning is a rapidly developing area of science, and the author hopes that this book helps the reader grasp various exciting topics in reinforcement learning and stimulate readers' interest in machine learning. Please visit our website at: http://www.ms.k.u-tokyo.ac.jp.

Masashi Sugiyama
University of Tokyo, Japan

Author

Masashi Sugiyama was born in Osaka, Japan, in 1974. He received Bachelor, Master, and Doctor of Engineering degrees in Computer Science from All Tokyo Institute of Technology, Japan in 1997, 1999, and 2001, respectively. In 2001, he was appointed Assistant Professor in the same institute, and he was promoted to Associate Professor in 2003. He moved to the University of Tokyo as Professor in 2014.

He received an Alexander von Humboldt Foundation Research Fellowship and researched at Fraunhofer Institute, Berlin, Germany, from 2003 to 2004. In 2006, he received a European Commission Program Erasmus Mundus Scholarship and researched at the University of Edinburgh, Scotland. He received the Faculty Award from IBM in 2007 for his contribution to machine learning under non-stationarity, the Nagao Special Researcher Award from the Information Processing Society of Japan in 2011 and the Young Scientists' Prize from the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology for his contribution to the density-ratio paradigm of machine learning.

His research interests include theories and algorithms of machine learning and data mining, and a wide range of applications such as signal processing, image processing, and robot control. He published *Density Ratio Estimation in Machine Learning* (Cambridge University Press, 2012) and *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation* (MIT Press, 2012).

The author thanks his collaborators, Hirotaka Hachiya, Sethu Vijayakumar, Jan Peters, Jun Morimoto, Zhao Tingting, Ning Xie, Voot Tangkaratt, Tetsuro Morimura, and Norikazu Sugimoto, for exciting and creative discussions. He acknowledges support from MEXT KAKENHI 17700142, 18300057, 20680007, 23120004, 23300069, 25700022, and 26280054, the Okawa Foundation, EU Erasmus Mundus Fellowship, AOARD, SCAT, the JST PRESTO program, and the FIRST program.

Contents

Foreword ix

Preface xi

Author xiii

I Introduction 1

1 Introduction to Reinforcement Learning 3
1.1 Reinforcement Learning 3
1.2 Mathematical Formulation 8
1.3 Structure of the Book 12
1.3.1 Model-Free Policy Iteration 12
1.3.2 Model-Free Policy Search 13
1.3.3 Model-Based Reinforcement Learning 14

II Model-Free Policy Iteration 15

2 Policy Iteration with Value Function Approximation 17
2.1 Value Functions 17
2.1.1 State Value Functions 17
2.1.2 State-Action Value Functions 18
2.2 Least-Squares Policy Iteration 20
2.2.1 Immediate-Reward Regression 20
2.2.2 Algorithm 21
2.2.3 Regularization 23
2.2.4 Model Selection 25
2.3 Remarks 26

3 Basis Design for Value Function Approximation 27
3.1 Gaussian Kernels on Graphs 27
3.1.1 MDP-Induced Graph 27
3.1.2 Ordinary Gaussian Kernels 29
3.1.3 Geodesic Gaussian Kernels 29
3.1.4 Extension to Continuous State Spaces 30
3.2 Illustration 30
3.2.1 Setup 31

3.2.2 Geodesic Gaussian Kernels 31
3.2.3 Ordinary Gaussian Kernels 33
3.2.4 Graph-Laplacian Eigenbases 34
3.2.5 Diffusion Wavelets 35
3.3 Numerical Examples 36
3.3.1 Robot-Arm Control 36
3.3.2 Robot-Agent Navigation 39
3.4 Remarks 45

4 Sample Reuse in Policy Iteration 47
4.1 Formulation 47
4.2 Off-Policy Value Function Approximation 48
4.2.1 Episodic Importance Weighting 49
4.2.2 Per-Decision Importance Weighting 50
4.2.3 Adaptive Per-Decision Importance Weighting 50
4.2.4 Illustration 51
4.3 Automatic Selection of Flattening Parameter 54
4.3.1 Importance-Weighted Cross-Validation 54
4.3.2 Illustration 55
4.4 Sample-Reuse Policy Iteration 56
4.4.1 Algorithm 56
4.4.2 Illustration 57
4.5 Numerical Examples 58
4.5.1 Inverted Pendulum 58
4.5.2 Mountain Car 60
4.6 Remarks 63

5 Active Learning in Policy Iteration 65
5.1 Efficient Exploration with Active Learning 65
5.1.1 Problem Setup 65
5.1.2 Decomposition of Generalization Error 66
5.1.3 Estimation of Generalization Error 67
5.1.4 Designing Sampling Policies 68
5.1.5 Illustration 69
5.2 Active Policy Iteration 71
5.2.1 Sample-Reuse Policy Iteration with Active Learning . 72
5.2.2 Illustration 73
5.3 Numerical Examples 75
5.4 Remarks 77

6 Robust Policy Iteration 79
6.1 Robustness and Reliability in Policy Iteration 79
6.1.1 Robustness 79
6.1.2 Reliability 80
6.2 Least Absolute Policy Iteration 81

6.2.1 Algorithm 81
6.2.2 Illustration 81
6.2.3 Properties 83
6.3 Numerical Examples 84
6.4 Possible Extensions 88
6.4.1 Huber Loss 88
6.4.2 Pinball Loss 89
6.4.3 Deadzone-Linear Loss 90
6.4.4 Chebyshev Approximation 90
6.4.5 Conditional Value-At-Risk 91
6.5 Remarks 92

III Model-Free Policy Search 93

7 Direct Policy Search by Gradient Ascent 95
7.1 Formulation 95
7.2 Gradient Approach 96
7.2.1 Gradient Ascent 96
7.2.2 Baseline Subtraction for Variance Reduction 98
7.2.3 Variance Analysis of Gradient Estimators 99
7.3 Natural Gradient Approach 101
7.3.1 Natural Gradient Ascent 101
7.3.2 Illustration 103
7.4 Application in Computer Graphics: Artist Agent 104
7.4.1 Sumie Painting 105
7.4.2 Design of States, Actions, and Immediate Rewards . . 105
7.4.3 Experimental Results 112
7.5 Remarks 113

8 Direct Policy Search by Expectation-Maximization 117
8.1 Expectation-Maximization Approach 117
8.2 Sample Reuse 120
8.2.1 Episodic Importance Weighting 120
8.2.2 Per-Decision Importance Weight 122
8.2.3 Adaptive Per-Decision Importance Weighting 123
8.2.4 Automatic Selection of Flattening Parameter 124
8.2.5 Reward-Weighted Regression with Sample Reuse . . . 125
8.3 Numerical Examples 126
8.4 Remarks 132

9 Policy-Prior Search 133
9.1 Formulation 133
9.2 Policy Gradients with Parameter-Based Exploration 134
9.2.1 Policy-Prior Gradient Ascent 135
9.2.2 Baseline Subtraction for Variance Reduction 136
9.2.3 Variance Analysis of Gradient Estimators 136

9.2.4 Numerical Examples 138
9.3 Sample Reuse in Policy-Prior Search 143
9.3.1 Importance Weighting 143
9.3.2 Variance Reduction by Baseline Subtraction 145
9.3.3 Numerical Examples 146
9.4 Remarks 153

IV Model-Based Reinforcement Learning 155

10 Transition Model Estimation 157
10.1 Conditional Density Estimation 157
10.1.1 Regression-Based Approach 157
10.1.2 ϵ-Neighbor Kernel Density Estimation 158
10.1.3 Least-Squares Conditional Density Estimation 159
10.2 Model-Based Reinforcement Learning 161
10.3 Numerical Examples 162
10.3.1 Continuous Chain Walk 162
10.3.2 Humanoid Robot Control 167
10.4 Remarks 172

11 Dimensionality Reduction for Transition Model Estimation 173
11.1 Sufficient Dimensionality Reduction 173
11.2 Squared-Loss Conditional Entropy 174
11.2.1 Conditional Independence 174
11.2.2 Dimensionality Reduction with SCE 175
11.2.3 Relation to Squared-Loss Mutual Information 176
11.3 Numerical Examples 177
11.3.1 Artificial and Benchmark Datasets 177
11.3.2 Humanoid Robot 180
11.4 Remarks 182

References 183

Index 191

Part I

Introduction