

Studies in Systems, Decision and Control 325

Kyriakos G. Vamvoudakis

Yan Wan

Frank L. Lewis

Derya Cansever *Editors*

# Handbook of Reinforcement Learning and Control

 Springer

# **Studies in Systems, Decision and Control**

Volume 325

## **Series Editor**

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,  
Warsaw, Poland

The series “Studies in Systems, Decision and Control” (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control—quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at <http://www.springer.com/series/13304>

Kyriakos G. Vamvoudakis · Yan Wan ·  
Frank L. Lewis · Derya Cansever  
Editors

# Handbook of Reinforcement Learning and Control

 Springer

*Editors*

Kyriakos G. Vamvoudakis  
The Daniel Guggenheim School  
of Aerospace Engineering  
Georgia Institute of Technology  
Atlanta, GA, USA

Frank L. Lewis  
Department of Electrical Engineering  
The University of Texas at Arlington  
Arlington, TX, USA

Yan Wan  
Department of Electrical Engineering  
The University of Texas at Arlington  
Arlington, TX, USA

Derya Cansever  
Army Research Office  
Durham, NC, USA

ISSN 2198-4182

ISSN 2198-4190 (electronic)

Studies in Systems, Decision and Control

ISBN 978-3-030-60989-4

ISBN 978-3-030-60990-0 (eBook)

<https://doi.org/10.1007/978-3-030-60990-0>

© Springer Nature Switzerland AG 2021

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, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Teams of autonomous systems operate in complex dynamic and networked environments with spatiotemporal diverse threats generated by malicious attacks, functional failures, and human errors. With the wide availability of data, Reinforcement Learning (RL) enables adaptive autonomy where the agents automatically learn policies to optimize a reward through interactions with the environment. RL has demonstrated the capability to deal with incomplete information and uncertain environment, and realize full autonomy. However, there are still some critical challenges in the application of such learning and adaptation methods to solve academic and industrial problems, for instance, the curse of dimensionality; optimization in dynamic environments; convergence and performance analysis; safety; nonequilibrium learning; and online implementation. On the other hand, some emerging technologies such as deep learning and multiagent systems will also provide a potential opportunity to further tackle these challenges.

This book presents a variety of state-of-the-art methods for the RL and games in multiagent settings by bringing the leading researchers in the field. These methods cover the theory, future perspectives, and applications of learning-based systems. This book has seven sections. The first part provides an introduction to RL. The second part provides some theoretic perspectives for model-free and model-based learning-based control techniques. The third part incorporates constraints and discusses techniques for verification. A multiagent perspective is presented in the fourth part. Part Five discusses bounded rationality in games and the value of shared information. Finally, applications of RL and multidisciplinary connections are presented in part six and seven, respectively.

We summarize in this Preface the main contributions from each chapter to put them in context.

Part I starts with the Chap. 1, where Derya Cansever provides directions on the future of RL. Then Kiumarsi, Modares, and Lewis present “2” a variety of RL methods to successfully learn the solution to the optimal control and game problems online and using measured data along the system trajectories. Powell in Chap. 3 makes the case that the modeling framework of RL, inherited from discrete Markov decision processes, is quite limited and suggests a stochastic control framework that based on the core problem of optimizing over policies. Chapter 4 by Devraj, Bušić’, and Meyn contains a survey of the new class of Zap RL algorithms that can achieve convergence almost universally while also guaranteeing the optimal rate of convergence. Greene, Deptula, Kamalapurkar, and Dixon discuss in Chap. 5, mixed density RL-based approximate optimal control methods applied to deterministic systems. Chapter 6 and Mohammadi, Soltanolkotabi, and Jovanović review recent results on the convergence and sample complexity of the random search method for the infinite horizon continuous-time linear quadratic regulator problem with unknown model parameters.

Part II starts with Yang, Wunsch II, and Yin in Chap. 7, that present a Hamiltonian-driven framework of adaptive dynamic programming for continuous-time nonlinear systems. Rizvi, Wei, and Lin, in Chap. 8 provides RL-based methods for solving the optimal stabilization problem for time delay systems with unknown delays and unknown system dynamics. Chapter 9, written by Moghadam, Jagannathan, Narayanan, and Raghavan, presents a RL method for partially unknown linear continuous-time systems with state delays. Kanellopoulos, Vamvoudakis, Gupta, and Antsaklis in Chap. 10 investigate the problem of verifying desired properties of large-scale RL systems that operate in uncertain and adversarial environments. Chapter 11 and Benosman, Chakrabarty, and Borggaard show a method for RL-based model reduction for partial differential equations and specifically to Burgers equations.

Part III starts with Chap. 12 by Zhang, Yang, and Başar review the theoretical results of multiagent RL algorithms mainly within two representative frameworks, Markov/stochastic games and extensive-form games. Chapter 13 written by Liu, Wan, Lin, Lewis, Xie, and Jalaian describes the use of computationally effective uncertainty evaluation methods for adaptive optimal control, including learning control and differential games. Lin, Montúfar, and Osher in Chap. 14 demonstrate a method, to obtain decentralized multiagents through a top-down approach: first by obtaining a solution with a centralized controller, and then decentralizing using imitation learning. Chapter 15 from Yang and Hespanha provides a security perspective in Stackelberg games and for the case that the attack objective and capacity are unknown, they propose a learning-based approach that predicts the routing cost using a neural network and minimizes the predicted cost via projected gradient descent.

Part IV starts with Chap. 16 from Kokolakis, Kanellopoulos, and Vamvoudakis that present a unified framework of bounded rationality for control systems as this can be employed in a coordinated unmanned aerial vehicle tracking RL problem. Tsiotras in Chap. 17 utilizes bounded-rationality ideas for generating suitable hierarchical abstractions to handle demanding tasks under time and other resource constraints, when exact optimality/rationality may be elusive. Chapter 18 and authors, Zhang

and Liu, review existing literature on the fairness of data-driven sequential decision-making. Sledge and Príncipe in Chap. 19 provide a way to resolve the exploration-exploitation dilemma in RL.

Part V starts with Chap. 20 from Castagno and Atkins, that present the offline construction of a landing site database using publicly available data sources with a focus on rooftop sites. Surana in Chap. 21 presents an industrial perspective of RL. Chapter 22 with authors, Huang, Jiang, Malisoff, and Cui develop RL-based shared control designs for semi-autonomous vehicles with a human in the loop. Wang and Wang in Chap. 23 present a decision-making framework subject to uncertainties that are represented by a set of random variables injected to the system as a group of inputs in both performance and constraint equations.

Finally, Part VI starts with Chap. 24 from Poveda and Teel, that provided a framework for the analysis of RL-based controllers that can safely and systematically integrate the intrinsic continuous-time dynamics and discrete-time dynamics that emerge in cyber-physical systems. Haddad in Chap. 25 looks to systems biology, neurophysiology, and thermodynamics for inspiration in developing innovative architectures for control and learning. The last chapter of the book, written by Rajagopal, Zhang, Balakrishnan, Fakhari, and Busemeyer, reviews work on a new choice rule based on an amplitude amplification algorithm originally developed in quantum computing.

In summary, this book presents a variety of challenging theoretical problems coupled with real practical applications for RL systems. The future directions are also covered in individual chapters.

Atlanta, USA  
Arlington, USA  
Arlington, USA  
Durham, USA

Kyriakos G. Vamvoudakis  
Yan Wan  
Frank L. Lewis  
Derya Cansever

# Contents

<b>Part I Theory of Reinforcement Learning for Model-Free and Model-Based Control and Games</b>	
<b>1</b>	<b>What May Lie Ahead in Reinforcement Learning</b> ..... 3
	Derya Cansever
	References ..... 5
<b>2</b>	<b>Reinforcement Learning for Distributed Control and Multi-player Games</b> ..... 7
	Bahare Kiumarsi, Hamidreza Modares, and Frank Lewis
2.1	Introduction ..... 7
2.2	Optimal Control of Continuous-Time Systems ..... 9
2.2.1	IRL with Experience Replay Learning Technique [12, 13] ..... 11
2.2.1.1	Off-Policy IRL Algorithm [14–16] ..... 12
2.2.2	$\mathcal{H}_\infty$ Control of CT Systems ..... 14
2.2.2.1	Off-Policy IRL [15, 21] ..... 16
2.3	Nash Games ..... 17
2.4	Graphical Games ..... 20
2.4.1	Off-Policy RL for Graphical Games ..... 22
2.5	Output Synchronization of Multi-agent Systems ..... 23
2.6	Conclusion and Open Research Directions ..... 26
	References ..... 26
<b>3</b>	<b>From Reinforcement Learning to Optimal Control: A Unified Framework for Sequential Decisions</b> ..... 29
	Warren B. Powell
3.1	Introduction ..... 29
3.2	The Communities of Sequential Decisions ..... 31
3.3	Stochastic Optimal Control Versus Reinforcement Learning .... 33
3.3.1	Stochastic Control ..... 34
3.3.2	Reinforcement Learning ..... 37
3.3.3	A Critique of the MDP Modeling Framework ..... 41

3.3.4	Bridging Optimal Control and Reinforcement Learning	42
3.4	The Universal Modeling Framework	44
3.4.1	Dimensions of a Sequential Decision Model	45
3.4.2	State Variables	47
3.4.3	Objective Functions	49
3.4.4	Notes	51
3.5	Energy Storage Illustration	52
3.5.1	A Basic Energy Storage Problem	53
3.5.2	With a Time-Series Price Model	55
3.5.3	With Passive Learning	55
3.5.4	With Active Learning	56
3.5.5	With Rolling Forecasts	56
3.5.6	Remarks	57
3.6	Designing Policies	58
3.6.1	Policy Search	58
3.6.2	Lookahead Approximations	60
3.6.3	Hybrid Policies	63
3.6.4	Remarks	64
3.6.5	Stochastic Control, Reinforcement Learning, and the Four Classes of Policies	65
3.7	Policies for Energy Storage	67
3.8	Extension to Multi-agent Systems	69
3.9	Observations	71
	References	72
<b>4</b>	<b>Fundamental Design Principles for Reinforcement Learning Algorithms</b>	<b>75</b>
	Adithya M. Devraj, Ana Bušić, and Sean Meyn	
4.1	Introduction	76
4.1.1	Stochastic Approximation and Reinforcement Learning	77
4.1.2	Sample Complexity Bounds	78
4.1.3	What Will You Find in This Chapter?	79
4.1.4	Literature Survey	80
4.2	Stochastic Approximation: New and Old Tricks	81
4.2.1	What is Stochastic Approximation?	82
4.2.2	Stochastic Approximation and Learning	84
4.2.2.1	Monte Carlo	84
4.2.2.2	Temporal Difference Learning	86
4.2.3	Stability and Convergence	88
4.2.4	Zap–Stochastic Approximation	89
4.2.5	Rates of Convergence	91
4.2.5.1	White Noise Model	92
4.2.5.2	Markovian Model	93

- 4.2.5.3 Implications and Matrix Gain Stochastic Approximation ..... 95
- 4.2.6 Optimal Convergence Rate ..... 96
  - 4.2.6.1 Stochastic Newton–Raphson ..... 96
  - 4.2.6.2 Zap Stochastic Approximation ..... 97
  - 4.2.6.3 Other Optimal Algorithms ..... 98
- 4.2.7 TD and LSTD Algorithms ..... 99
- 4.3 Zap Q-Learning: Fastest Convergent Q-Learning ..... 101
  - 4.3.1 Markov Decision Processes ..... 101
  - 4.3.2 Value Functions and the Bellman Equation ..... 102
  - 4.3.3 Q-Learning ..... 104
  - 4.3.4 Tabular Q-Learning ..... 105
  - 4.3.5 Convergence and Rate of Convergence ..... 108
  - 4.3.6 Zap Q-Learning ..... 112
    - 4.3.6.1 Main Results ..... 113
    - 4.3.6.2 Zap ODE and Policy Iteration ..... 114
    - 4.3.6.3 Overview of Proofs ..... 115
- 4.4 Numerical Results ..... 118
  - 4.4.1 Finite State-Action MDP ..... 119
  - 4.4.2 Optimal Stopping in Finance ..... 124
    - 4.4.2.1 Approximations to the Optimal Stopping Time Problem ..... 125
    - 4.4.2.2 Experimental Results ..... 126
    - 4.4.2.3 Asymptotic Variance of the Discounted Reward ..... 128
- 4.5 Zap-Q with Nonlinear Function Approximation ..... 128
  - 4.5.1 Choosing the Eligibility Vectors ..... 130
  - 4.5.2 Theory and Challenges ..... 131
  - 4.5.3 Regularized Zap-Q ..... 131
- 4.6 Conclusions and Future Work ..... 132
- References ..... 134
- 5 Mixed Density Methods for Approximate Dynamic Programming ..... 139**

Max L. Greene, Patryk Deptula, Rushikesh Kamalapurkar, and Warren E. Dixon

  - 5.1 Introduction ..... 140
  - 5.2 Unconstrained Affine-Quadratic Regulator ..... 142
  - 5.3 Regional Model-Based Reinforcement Learning ..... 150
    - 5.3.1 Preliminaries ..... 151
    - 5.3.2 Regional Value Function Approximation ..... 151
    - 5.3.3 Bellman Error ..... 152
      - 5.3.3.1 Extension to Unknown Dynamics ..... 153
    - 5.3.4 Actor and Critic Update Laws ..... 155
    - 5.3.5 Stability Analysis ..... 156

5.3.6	Summary .....	158
5.4	Local (State-Following) Model-Based Reinforcement Learning .....	159
5.4.1	StaF Kernel Functions .....	160
5.4.2	Local Value Function Approximation .....	161
5.4.3	Actor and Critic Update Laws .....	162
5.4.4	Analysis .....	163
5.4.5	Stability Analysis .....	165
5.4.6	Summary .....	166
5.5	Combining Regional and Local State-Following Approximations .....	167
5.6	Reinforcement Learning with Sparse Bellman Error Extrapolation .....	168
5.7	Conclusion .....	168
	References .....	169
<b>6</b>	<b>Model-Free Linear Quadratic Regulator</b> .....	<b>173</b>
	Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R. Jovanović	
6.1	Introduction to a Model-Free LQR Problem .....	173
6.2	A Gradient-Based Random Search Method .....	175
6.3	Main Results .....	176
6.4	Proof Sketch .....	177
6.4.1	Controlling the Bias .....	179
6.4.2	Correlation of $\widehat{\nabla} f(K)$ and $\nabla f(K)$ .....	181
6.5	An Example .....	182
6.6	Thoughts and Outlook .....	183
	References .....	185
 <b>Part II Constraint-Driven and Verified RL</b>		
<b>7</b>	<b>Adaptive Dynamic Programming in the Hamiltonian-Driven Framework</b> .....	<b>189</b>
	Yongliang Yang, Donald C. Wunsch II, and Yixin Yin	
7.1	Introduction .....	190
7.1.1	Literature Review .....	190
7.1.2	Motivation .....	191
7.1.3	Structure .....	192
7.2	Problem Statement .....	192
7.3	Hamiltonian-Driven Framework .....	195
7.3.1	Policy Evaluation .....	195
7.3.2	Policy Comparison .....	198
7.3.3	Policy Improvement .....	201
7.4	Discussions on the Hamiltonian-Driven ADP .....	206
7.4.1	Implementation with Critic-Only Structure .....	206

7.4.2	Connection to Temporal Difference Learning	209
7.4.2.1	Continuous-Time Integral Temporal Difference	209
7.4.2.2	Continuous-Time Least Squares Temporal Difference	209
7.4.3	Connection to Value Gradient Learning	210
7.5	Simulation Study	210
7.6	Conclusion	213
	References	213
<b>8</b>	<b>Reinforcement Learning for Optimal Adaptive Control of Time Delay Systems</b>	<b>215</b>
	Syed Ali Asad Rizvi, Yusheng Wei, and Zongli Lin	
8.1	Introduction	216
8.2	Problem Description	218
8.3	Extended State Augmentation	219
8.4	State Feedback Q-Learning Control of Time Delay Systems	228
8.5	Output Feedback Q-Learning Control of Time Delay Systems	232
8.6	Simulation Results	238
8.7	Conclusions	240
	References	242
<b>9</b>	<b>Optimal Adaptive Control of Partially Uncertain Linear Continuous-Time Systems with State Delay</b>	<b>243</b>
	Rohollah Moghadam, S. Jagannathan, Vignesh Narayanan, and Krishnan Raghavan	
9.1	Introduction	244
9.2	Problem Statement	245
9.3	Linear Quadratic Regulator Design	246
9.3.1	Periodic Sampled Feedback	247
9.3.2	Event Sampled Feedback	249
9.4	Optimal Adaptive Control	252
9.4.1	Periodic Sampled Feedback	252
9.4.2	Event Sampled Feedback	257
9.4.3	Hybrid Reinforcement Learning Scheme	259
9.5	Perspectives on Controller Design with Image Feedback	260
9.6	Simulation Results	264
9.6.1	Linear Quadratic Regulator with Known Internal Dynamics	265
9.6.2	Optimal Adaptive Control with Unknown Drift Dynamics	265
9.7	Conclusion	267
	References	270

<b>10</b>	<b>Dissipativity-Based Verification for Autonomous Systems in Adversarial Environments</b> .....	273
	Aris Kanellopoulos, Kyriakos G. Vamvoudakis, Vijay Gupta, and Panos Antsaklis	
10.1	Introduction .....	273
10.1.1	Related Work .....	275
10.1.2	Contributions .....	276
10.1.3	Structure .....	276
10.1.4	Notation .....	276
10.2	Problem Formulation .....	277
10.2.1	$(Q, S, R)$ -Dissipative and $L_2$ -Gain Stable Systems .....	278
10.3	Learning-Based Distributed Cascade Interconnection .....	279
10.4	Learning-Based $L_2$ -Gain Composition .....	281
10.4.1	Q-Learning for $L_2$ -Gain Verification .....	281
10.4.2	$L_2$ -Gain Model-Free Composition .....	285
10.5	Learning-Based Lossless Composition .....	286
10.6	Discussion .....	288
10.7	Conclusion and Future Work .....	289
	References .....	290
<b>11</b>	<b>Reinforcement Learning-Based Model Reduction for Partial Differential Equations: Application to the Burgers Equation</b> .....	293
	Mouhacine Benosman, Ankush Chakrabarty, and Jeff Borggaard	
11.1	Introduction .....	293
11.2	Basic Notation and Definitions .....	295
11.3	RL-Based Model Reduction of PDEs .....	296
11.3.1	Reduced-Order PDE Approximation .....	296
11.3.2	Proper Orthogonal Decomposition for ROMs .....	297
11.3.3	Closure Models for ROM Stabilization .....	298
11.3.4	Main Result: RL-Based Closure Model .....	299
11.4	Extremum Seeking Based Closure Model Auto-Tuning .....	304
11.5	The Case of the Burgers Equation .....	305
11.6	Conclusion .....	312
	References .....	316
<b>Part III Multi-agent Systems and RL</b>		
<b>12</b>	<b>Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms</b> .....	321
	Kaiqing Zhang, Zhuoran Yang, and Tamer Başar	
12.1	Introduction .....	322
12.2	Background .....	324
12.2.1	Single-Agent RL .....	324
12.2.1.1	Value-Based Methods .....	325
12.2.1.2	Policy-Based Methods .....	326

- 12.2.2 Multi-Agent RL Framework ..... 327
  - 12.2.2.1 Markov/Stochastic Games ..... 327
  - 12.2.2.2 Extensive-Form Games ..... 330
- 12.3 Challenges in MARL Theory ..... 332
  - 12.3.1 Non-unique Learning Goals ..... 332
  - 12.3.2 Non-stationarity ..... 333
  - 12.3.3 Scalability Issue ..... 334
  - 12.3.4 Various Information Structures ..... 334
- 12.4 MARL Algorithms with Theory ..... 336
  - 12.4.1 Cooperative Setting ..... 336
    - 12.4.1.1 Homogeneous Agents ..... 336
    - 12.4.1.2 Decentralized Paradigm  
with Networked Agents ..... 339
    - 12.4.1.3 Partially Observed Model ..... 344
  - 12.4.2 Competitive Setting ..... 345
    - 12.4.2.1 Value-Based Methods ..... 346
    - 12.4.2.2 Policy-Based Methods ..... 349
  - 12.4.3 Mixed Setting ..... 355
- 12.5 Application Highlights ..... 358
  - 12.5.1 Cooperative Setting ..... 358
  - 12.5.2 Competitive Setting ..... 361
  - 12.5.3 Mixed Settings ..... 365
- 12.6 Conclusions and Future Directions ..... 366
- References ..... 368

**13 Computational Intelligence in Uncertainty Quantification  
for Learning Control and Differential Games ..... 385**

Mushuang Liu, Yan Wan, Zongli Lin, Frank L. Lewis, Junfei Xie,  
and Brian A. Jalaian

- 13.1 Introduction ..... 386
- 13.2 Problem Formulation of Optimal Control for Uncertain  
Systems ..... 387
  - 13.2.1 Optimal Control for Systems with Parameters  
Modulated by Multi-dimensional Uncertainties ..... 387
    - 13.2.1.1 Systems with Parameters Modulated  
by Multi-dimensional Uncertainties ..... 387
    - 13.2.1.2 Optimal Control for Systems  
with Parameters Modulated  
by Multi-dimensional Uncertainties ..... 388
  - 13.2.2 Optimal Control for Random Switching Systems ..... 389
    - 13.2.2.1 Random Switching Models ..... 389
    - 13.2.2.2 Optimal Control for Random Switching  
Systems ..... 390
- 13.3 Effective Uncertainty Evaluation Methods ..... 391
  - 13.3.1 Problem Formulation ..... 391

- 13.3.2 The MPCM ..... 391
- 13.3.3 The MPCM-OFFD ..... 393
- 13.4 Optimal Control Solutions for Systems with Parameter Modulated by Multi-dimensional Uncertainties ..... 394
  - 13.4.1 Reinforcement Learning-Based Stochastic Optimal Control ..... 394
  - 13.4.2 Q-Learning-Based Stochastic Optimal Control ..... 396
- 13.5 Optimal Control Solutions for Random Switching Systems ..... 397
  - 13.5.1 Optimal Controller for Random Switching Systems ..... 397
  - 13.5.2 Effective Estimator for Random Switching Systems ..... 399
- 13.6 Differential Games for Systems with Parameters Modulated by Multi-dimensional Uncertainties ..... 401
  - 13.6.1 Stochastic Two-Player Zero-Sum Game ..... 401
    - 13.6.1.1 On-Policy IRL ..... 403
    - 13.6.1.2 Off-Policy IRL ..... 404
  - 13.6.2 Multi-player Nonzero-Sum Game ..... 405
    - 13.6.2.1 On-Policy IRL ..... 407
    - 13.6.2.2 Off-Policy IRL ..... 408
- 13.7 Applications ..... 409
  - 13.7.1 Traffic Flow Management Under Uncertain Weather ..... 409
  - 13.7.2 Learning Control for Aerial Communication Using Directional Antennas (ACDA) Systems ..... 411
- 13.8 Summary ..... 415
- References ..... 416
- 14 A Top-Down Approach to Attain Decentralized Multi-agents ..... 419**
  - Alex Tong Lin, Guido Montúfar, and Stanley J. Osher
  - 14.1 Introduction ..... 420
  - 14.2 Background ..... 421
    - 14.2.1 Reinforcement Learning ..... 421
    - 14.2.2 Multi-agent Reinforcement Learning ..... 423
  - 14.3 Centralized Learning, But Decentralized Execution ..... 424
    - 14.3.1 A Bottom-Up Approach ..... 425
    - 14.3.2 A Top-Down Approach ..... 425
  - 14.4 Centralized Expert Supervises Multi-agents ..... 425
    - 14.4.1 Imitation Learning ..... 425
    - 14.4.2 CESMA ..... 426
  - 14.5 Experiments ..... 427
    - 14.5.1 Decentralization Can Achieve Centralized Optimality ..... 428
    - 14.5.2 Expert Trajectories Versus Multi-agent Trajectories ..... 428

14.6 Conclusion ..... 429

References ..... 430

**15 Modeling and Mitigating Link-Flooding Distributed Denial-of-Service Attacks via Learning in Stackelberg Games ..... 433**

Guosong Yang and João P. Hespanha

15.1 Introduction ..... 433

15.2 Routing and Attack in Communication Network ..... 436

15.3 Stackelberg Game Model ..... 438

15.4 Optimal Attack and Stackelberg Equilibria for Malicious Adversaries ..... 439

15.4.1 Optimal Attack and Stackelberg Equilibria for Networks with Identical Links ..... 442

15.5 Mitigating Attacks via Learning ..... 450

15.5.1 Predicting the Routing Cost ..... 451

15.5.2 Minimizing the Predicted Routing Cost ..... 451

15.6 Simulation Study ..... 452

15.6.1 Discussion ..... 459

15.7 Conclusion ..... 461

References ..... 461

**Part IV Bounded Rationality and Value of Information in RL and Games**

**16 Bounded Rationality in Differential Games: A Reinforcement Learning-Based Approach ..... 467**

Nick-Marios T. Kokolakis, Aris Kanellopoulos, and Kyriakos G. Vamvoudakis

16.1 Introduction ..... 467

16.1.1 Related Work ..... 468

16.2 Problem Formulation ..... 469

16.2.1 Nash Equilibrium Solutions for Differential Games .... 469

16.3 Boundedly Rational Game Solution Concepts ..... 472

16.4 Cognitive Hierarchy for Adversarial Target Tracking ..... 474

16.4.1 Problem Formulation ..... 474

16.4.1.1 Vehicle Dynamics ..... 475

16.4.1.2 Relative Kinematics ..... 475

16.4.1.3 Differential Game Formulation ..... 476

16.4.2 Zero-Sum Game ..... 478

16.4.3 Cognitive Hierarchy ..... 479

16.4.3.1 Level-0 (Anchor) Policy ..... 479

16.4.3.2 Level- $k$  Policies ..... 480

16.4.4 Coordination with Nonequilibrium Game-Theoretic Learning ..... 481

16.4.5 Simulation ..... 485

16.5 Conclusion and Future Work ..... 486

References ..... 488

**17 Bounded Rationality in Learning, Perception, Decision-Making, and Stochastic Games** ..... 491  
Panagiotis Tsiotras

- 17.1 The Autonomy Challenge ..... 491
  - 17.1.1 The Case of Actionable Data ..... 492
  - 17.1.2 The Curse of Optimality ..... 493
- 17.2 How to Move Forward ..... 494
  - 17.2.1 Bounded Rationality for Human-Like Decision-Making ..... 495
  - 17.2.2 Hierarchical Abstractions for Scalability ..... 496
- 17.3 Sequential Decision-Making Subject to Resource Constraints ..... 497
  - 17.3.1 Standard Markov Decision Processes ..... 498
  - 17.3.2 Information-Limited Markov Decision Processes ..... 500
- 17.4 An Information-Theoretic Approach for Hierarchical Decision-Making ..... 505
  - 17.4.1 Agglomerative Information Bottleneck for Quadtree Compression ..... 506
  - 17.4.2 Optimal Compression of Quadtrees ..... 508
  - 17.4.3 The Q-Tree Search Algorithm ..... 509
- 17.5 Stochastic Games and Bounded Rationality ..... 511
  - 17.5.1 Stochastic Pursuit–Evasion ..... 513
  - 17.5.2 Level-k Thinking ..... 515
  - 17.5.3 A Pursuit–Evasion Game in a Stochastic Environment ..... 517
- 17.6 Conclusions ..... 519
- References ..... 520

**18 Fairness in Learning-Based Sequential Decision Algorithms: A Survey** ..... 525  
Xueru Zhang and Mingyan Liu

- 18.1 Introduction ..... 525
- 18.2 Preliminaries ..... 528
  - 18.2.1 Sequential Decision Algorithms ..... 528
  - 18.2.2 Notions of Fairness ..... 528
- 18.3 (Fair) Sequential Decision When Decisions Do Not Affect Underlying Population ..... 529
  - 18.3.1 Bandits, Regret, and Fair Regret ..... 529
  - 18.3.2 Fair Experts and Expert Opinions ..... 533
  - 18.3.3 Fair Policing ..... 535
- 18.4 (Fair) Sequential Decision When Decisions Affect Underlying Population ..... 535
  - 18.4.1 Two-Stage Models ..... 536
    - 18.4.1.1 Effort-Based Fairness ..... 538

- 18.4.1.2 A Two-Stage Model in College Admissions ..... 540
- 18.4.2 Long-Term Impacts on the Underlying Population ..... 542
  - 18.4.2.1 Effects of Decisions on the Evolution of Features ..... 542
  - 18.4.2.2 Fairness Intervention on Labor Market ..... 547
  - 18.4.2.3 Effects of Decisions on Group Representation ..... 549
  - 18.4.2.4 Combined Effects on Group Representation and Features ..... 552
  - 18.4.2.5 Fairness in Reinforcement Learning Problems ..... 553
- References ..... 553

**19 Trading Utility and Uncertainty: Applying the Value of Information to Resolve the Exploration–Exploitation Dilemma in Reinforcement Learning ..... 557**

Isaac J. Sledge and José C. Príncipe

- 19.1 Introduction ..... 558
- 19.2 Exploring Single-State, Multiple-Action Markov Decision Processes ..... 564
  - 19.2.1 Literature Survey ..... 564
  - 19.2.2 Methodology ..... 567
    - 19.2.2.1 Value of Information ..... 568
    - 19.2.2.2 Value of Information Optimization ..... 571
  - 19.2.3 Simulations and Analyses ..... 575
    - 19.2.3.1 Simulation Preliminaries ..... 575
    - 19.2.3.2 Value of Information Results and Analysis ..... 576
    - 19.2.3.3 Methodological Comparisons ..... 581
  - 19.2.4 Conclusions ..... 586
- 19.3 Exploring Multiple-state, Multiple-Action Markov Decision Processes ..... 587
  - 19.3.1 Literature Survey ..... 587
  - 19.3.2 Methodology ..... 590
    - 19.3.2.1 Value of Information ..... 592
    - 19.3.2.2 Value of Information Optimization ..... 595
  - 19.3.3 Simulations and Analyses ..... 598
    - 19.3.3.1 Simulation Preliminaries ..... 599
    - 19.3.3.2 Value of Information Results and Analyses ..... 600
    - 19.3.3.3 Methodological Comparisons ..... 605
  - 19.3.4 Conclusions ..... 606
- References ..... 607

**Part V Applications of RL**

**20 Map-Based Planning for Small Unmanned Aircraft Rooftop**

**Landing** ..... 613

J. Castagno and E. Atkins

20.1 Introduction ..... 613

20.2 Background ..... 615

20.2.1 Sensor-Based Planning ..... 615

20.2.2 Map-Based Planning ..... 616

20.2.3 Multi-goal Planning ..... 617

20.2.4 Urban Landscape and Rooftop Landings ..... 618

20.3 Preliminaries ..... 619

20.3.1 Coordinates and Landing Sites ..... 619

20.3.2 3D Path Planning with Mapped Obstacles ..... 620

20.4 Landing Site Database ..... 620

20.4.1 Flat-Like Roof Identification ..... 621

20.4.2 Flat Surface Extraction for Usable Landing Area ..... 622

20.4.3 Touchdown Points ..... 623

20.4.4 Landing Site Risk Model ..... 625

20.4.4.1 Vehicle Cost ..... 625

20.4.4.2 Terrain Cost ..... 625

20.4.4.3 Area Cost ..... 626

20.4.4.4 Cumulative Area Cost ..... 626

20.4.4.5 Property Cost ..... 627

20.4.4.6 Human Occupancy Risk Mapping ..... 627

20.5 Three-Dimensional Maps for Path Planning ..... 628

20.6 Planning Risk Metric Analysis and Integration ..... 630

20.6.1 Real-Time Map-Based Planner Architecture ..... 630

20.6.2 Trade-Off Between Landing Site and Path Risk ..... 631

20.6.3 Multi-goal Planner ..... 632

20.6.3.1 Theory ..... 632

20.6.3.2 Multi-goal Path Planning Algorithm ..... 634

20.7 Maps and Simulation Results ..... 635

20.7.1 Landing Sites and Risk Maps ..... 636

20.7.2 Case Studies ..... 637

20.7.3 Urgent Landing Statistical Analysis ..... 641

20.7.3.1 Minimum Radius Footprint ..... 641

20.7.3.2 Performance Benchmarks ..... 641

20.8 Conclusion ..... 644

References ..... 644

**21 Reinforcement Learning: An Industrial Perspective** ..... 647

Amit Surana

21.1 Introduction ..... 647

21.2 RL Applications ..... 648

- 21.2.1 Sensor Management in Intelligence, Surveillance, and Reconnaissance ..... 649
- 21.2.2 High Level Reasoning in Autonomous Navigation ..... 649
- 21.2.3 Advanced Manufacturing Process Control ..... 650
- 21.2.4 Maintenance, Repair, and Overhaul Operations ..... 652
- 21.2.5 Human–Robot Collaboration ..... 652
- 21.3 Case Study I: Optimal Sensor Tasking ..... 653
  - 21.3.1 Sensor Tasking as a Stochastic Optimal Control Problem ..... 653
  - 21.3.2 Multi-Arm Bandit Problem Approximation ..... 654
    - 21.3.2.1 Gittin’s Index in a Simplified Setting ..... 656
    - 21.3.2.2 Tracking with Multiple Sensors Combined with Search ..... 656
  - 21.3.3 Numerical Study ..... 658
- 21.4 Case Study II: Deep Reinforcement Learning for Advanced Manufacturing Control ..... 659
  - 21.4.1 Cold Spray Control Problem ..... 660
  - 21.4.2 Guided Policy Search ..... 662
  - 21.4.3 Simulation Results ..... 665
- 21.5 Future Outlook ..... 667
- References ..... 668
- 22 Robust Autonomous Driving with Human in the Loop ..... 673**  
 Mengzhe Huang, Zhong-Ping Jiang, Michael Malisoff, and Leilei Cui
  - 22.1 Introduction ..... 673
  - 22.2 Mathematical Modeling of Human–Vehicle Interaction ..... 676
    - 22.2.1 Vehicle Lateral Dynamics ..... 676
    - 22.2.2 Interconnected Human–Vehicle Model ..... 678
  - 22.3 Model-Based Control Design ..... 679
    - 22.3.1 Discretization of Differential-Difference Equations ..... 679
    - 22.3.2 Formulation of the Shared Control Problem ..... 681
    - 22.3.3 Model-Based Optimal Control Design ..... 682
  - 22.4 Learning-Based Optimal Control for Cooperative Driving ..... 683
  - 22.5 Numerical Results ..... 686
    - 22.5.1 Algorithmic Implementation ..... 686
    - 22.5.2 Comparisons and Discussions for ADP-Based Shared Control Design ..... 688
  - 22.6 Conclusions and Future Work ..... 690
  - References ..... 690
- 23 Decision-Making for Complex Systems Subjected to Uncertainties—A Probability Density Function Control Approach ..... 693**  
 Aiping Wang and Hong Wang
  - 23.1 Introduction ..... 693

- 23.2 Integrated Modeling Perspectives—Ordinary Algebra Versus  $\{Max, +\}$  Algebra ..... 696
  - 23.2.1 Process Level Modeling via Ordinary Algebra Systems ..... 697
  - 23.2.2  $\{Max, +\}$  Algebra-Based Modeling ..... 697
  - 23.2.3 Learning Under Uncertainties—PDF Shaping of Modeling Error-Based Approach ..... 699
- 23.3 Human-in-the-Loop Consideration: Impact of Uncertainties in Decision-Making Phase ..... 702
- 23.4 Optimization Under Uncertainties Impacts ..... 704
  - 23.4.1 Formulation of Optimization as a Feedback Control Design Problem—Optimization is a Special Case of Feedback Control System Design .... 704
    - 23.4.1.1 Source of Uncertainties in Decision-Making Phase ..... 707
- 23.5 A Generalized Framework for Decision-Making Using PDF Shaping Approach ..... 709
  - 23.5.1 PDF Shaping for the Performance Function ..... 709
  - 23.5.2 Dealing with the Constraint ..... 710
  - 23.5.3 Dealing with Dynamic Constraint ..... 713
  - 23.5.4 A Total Probabilistic Solution ..... 714
    - 23.5.4.1 Relations to Chance Constrained Optimization ..... 714
  - 23.5.5 Uncertainties in Performance Function and Constraints ..... 716
- 23.6 System Analysis: Square Impact Principle as a Mathematical Principle for Integrated IT with Infrastructure Design ..... 717
  - 23.6.1 Description of Operational Optimal Control ..... 718
  - 23.6.2 Square Impact Principle (SIP): Infrastructure Versus Control Performance ..... 719
- 23.7 Conclusions ..... 722
- References ..... 723

**Part VI Multi-Disciplinary Connections**

- 24 A Hybrid Dynamical Systems Perspective on Reinforcement Learning for Cyber-Physical Systems: Vistas, Open Problems, and Challenges ..... 727**
  - Jorge I. Poveda and Andrew R. Teel
  - 24.1 Introduction ..... 727
  - 24.2 Hybrid Dynamical Systems ..... 730
    - 24.2.1 Non-uniqueness of Solutions and Set-Valued Dynamics ..... 734

24.2.2	Hybrid Time Domains and Solutions of Hybrid Dynamical Systems	737
24.2.3	Graphical Convergence, Basic Assumptions and Sequential Compactness	738
24.2.4	Stability and Robustness	740
24.3	Reinforcement Learning via Dynamic Policy Gradient	742
24.3.1	Asynchronous Policy Iteration	744
24.3.2	Synchronous Policy Iteration: Online Training of Actor–Critic Structures	745
24.3.2.1	Learning Dynamics for the Critic	746
24.3.2.2	Learning Dynamics for the Actor	746
24.3.2.3	Closed-Loop System and Extensions to Other Settings	747
24.3.2.4	Extensions to Other Settings	749
24.4	Reinforcement Learning in Hybrid Dynamical Systems	749
24.4.1	Hybrid Learning Algorithms	750
24.4.1.1	Sampled-Data and Event-Triggered Architectures	750
24.4.1.2	Hybrid Coordination of Multi-agent Reinforcement Learning Algorithms	751
24.4.1.3	Hybrid Optimization and Estimation Algorithms	752
24.4.2	Hybrid Dynamic Environments	753
24.5	Conclusions	757
	References	757
<b>25</b>	<b>The Role of Systems Biology, Neuroscience, and Thermodynamics in Network Control and Learning</b>	<b>763</b>
	Wassim M. Haddad	
25.1	Introduction	763
25.2	Large-Scale Networks and Hybrid Thermodynamics	768
25.3	Multiagent Systems with Uncertain Interagent Communication	779
25.4	Systems Biology, Neurophysiology, Thermodynamics, and Dynamic Switching Communication Topologies for Large-Scale Multilayered Networks	789
25.5	Nonlinear Stochastic Optimal Control and Learning	792
25.6	Complexity, Thermodynamics, Information Theory, and Swarm Dynamics	798
25.7	Thermodynamic Entropy, Shannon Entropy, Bode Integrals, and Performance Limitations in Nonlinear Systems	801
25.8	Conclusion	809
	References	810

- 26 Quantum Amplitude Amplification for Reinforcement Learning** ..... 819
- K. Rajagopal, Q. Zhang, S. N. Balakrishnan, P. Fakhari, and J. R. Busemeyer
- 26.1 Exploration and Exploitation in Reinforcement Learning ..... 819
- 26.2 Quantum Probability Theory ..... 820
- 26.3 The Original Quantum Reinforcement Learning (QRL) Algorithm ..... 822
- 26.4 The Revised Quantum Reinforcement Learning Algorithm ..... 823
- 26.5 Learning Rate and Performance Comparisons ..... 824
- 26.6 Other Applications of QRL ..... 827
  - 26.6.1 Example ..... 830
- 26.7 Application to Human Learning ..... 831
- 26.8 Concluding Comments ..... 832
- References ..... 832

**Part I**  
**Theory of Reinforcement Learning**  
**for Model-Free and Model-Based Control**  
**and Games**

# Chapter 1

## What May Lie Ahead in Reinforcement Learning



Derya Cansever

The spectacular success enjoyed by machine learning (ML), primarily driven by deep neural networks can arguably be interpreted as only the tip of the iceberg. As neural network architectures, algorithmic methods, and computational power evolve, the scope of the problems that ML addresses will continue to grow. One such direction for growth materializes in reinforcement learning (RL) [1]. RL involves optimal sequential decision-making in uncertain/unknown environments where decisions at a given time take into account its impact on future events and decisions. As such, RL is akin to optimal adaptive control. In other words, it is a synthesis of dynamic programming and stochastic approximation methods [2]. Due to its sequential nature, RL is fundamentally different and broader than more commonly known deep neural network instantiations of ML, where the principal goal is to match the data with alternative hypotheses. By and large, classical neural networks tend to focus on one-shot, static problems. Ability to design deep neural networks to solve problems such as deciding whether an X-ray image corresponds to cancerous cells with previously unattainable accuracy is a glorious achievement, both in intellect and in consequence. However, it would fade in comparison with the challenges and potential rewards that could be attainable in the field of RL. The “curse of dimensionality” associated with sequential decision-making is well documented in the control literature. Having to deal with system uncertainties make the computational challenges even more daunting. Not only it geometrically increases the scope of the estimation problem but it also introduces a conflict between the need to exploit and the need to optimize, the solution of which is not well understood. Despite formidable computational challenges, the AlphaZero program achieved superhuman performance in the games of chess, shogi, and Go by reinforcement learning from self-play [3]. AlphaZero has shown that machines driven by RL can be the experts, not merely expert tools [4]. It appears

---

D. Cansever (✉)  
US Army Research Office, Adelphi, MD, USA  
e-mail: [derya.h.cansever.civ@mail.mil](mailto:derya.h.cansever.civ@mail.mil)

to have discovered novel fundamental principles about optimal strategies in chess, but it can't explicitly share that understanding with us as of yet [5]. Given the broad domain of applicability of optimal sequential decision making under uncertainty paradigm, it is tempting, and in fact plausible to imagine how RL in the future can radically shape fields that are instrumental in the functioning of society, including economics, business, politics, scientific investigations, etc. It is worthy to note that the first games (chess, shogi, and Go) that AlphaZero mastered occur in the open, fully observed by players. To an extent, this simplifies the computation of optimal strategies in RL. Subsequently, AlphaZero was extended to the game of poker, with uncertainty in the state observations [6]. This increases the dimensions of the affective state and strategy spaces and thus makes the problem even more challenging. The rules of the game, i.e., the system specification, are known and followed by all parties, which makes the problem more tractable than some other real-world problems where the rules of the evolution of the underlying system are not necessarily known in advance. In AlphaZero games, there are more than one decision-makers, which creates additional difficulties for the RL algorithm. For every fixed strategy of the opponent(s), AlphaZero needs to solve an RL induced optimization problem. As players need to take into account others' actions, and since other players' actions may take a plethora of possible values, the resulting set of optimization problems will be very challenging. The fact that players may have private information, e.g., closed cards, could induce more elaborate strategies such as deliberately missignaling their private information for better gains. Analysis and derivation of optimal strategies in games with private information is in general very difficult. Reinforcement learning-based approaches might be helpful with sufficient training data. AlphaZero focuses on zero-sum games, representing total conflict among the players. When all players share the same objective, i.e., a Team problem, players can collaborate to solve a common RL problem. Distributed RL and its corollary transfer learning are at the heart of the emerging topic of autonomy and will be instrumental in its realization. Coming back to the problems attacked by AlphaZero, the fact that the strategy space induced by the rules of the game is finite and reducible to smaller feasible subsets makes them relatively more tractable than other real-world multi-stage decision problems where the relevant strategy space may possibly take an unlimited number of values, and in fact may be infinite-dimensional, e.g., function spaces. AlphaZero enjoyed easily generated and virtually unbounded amounts of training data that are crucial for learning systems using self-play. Even with these advantages, AlphaZero requires considerable amounts of computational power, thus extending RL to more general problems remains a significant challenge. RL is built on a very solid foundation of optimal control theory [7] and stochastic approximation (Borkar [8]), so it should benefit from advances in these active fields. Some of the RL algorithms make use of deep neural network techniques as a component, a very popular research area of rapid advances. This robust mathematical and algorithmic foundation, along with Moore's Law, is expected to continue to fuel advances in RL research in the framework of optimal adaptive control theory.

For many problems of practical interest, computational power and training data requirements of RL can be formidable, and its convergence rate may be unacceptably

slow. In contrast, biological learning and decision making are in many ways more complex but can be very efficient. As AlphaZero demonstrates, in many ways, human brain is no match to sophisticated RL algorithms backed with enormous computing systems. But humans can decide quickly, discard irrelevant data, or defer the decision process to a slower time scale for more thorough processing [9]. The brain explicitly samples raw data from past memory episodes in decision making, which makes the state non-Markovian. The use of raw memory can accelerate decision-making, and also reduce processing burden by invoking strategically placed shortcuts. In contrast, many RL systems are formulated using Markovian assumptions, which may result in states that are too large to allow for efficient computations. Biological learning may involve building representations of the world from a few examples, filtering out superfluous data, and predicting events based on the history. A plausible alternate, or perhaps the complementary path to the optimal adaptive control formulation could emerge as being inspired by human brain, resulting in non-Markovian, hierarchical, and multiple time scales models. Obviously, this is a two-edged sword. It could motivate novel approaches and architectures for RL, but it could lead to some quirky features resulting from the long and windy road of evolution that are no longer relevant. Furthermore, some of the shortcut decision strategies in the brain may not necessarily be optimal in a high-powered computation environment for RL. In any case, it would be safe to postulate that one of the likely beneficiaries of the study of human brain is RL research.

## References

1. Barto, R.S.: Reinforcement Learning: An Introduction. The MIT Press, Cambridge, MA (2018)
2. Sutton, R., Barto, A., Williams, R.: Reinforcement learning is direct adaptive optimal control. *IEEE Control Syst. Mag.* **12**(2), 19–22 (1992)
3. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al.: A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* **362**(6419), 1140–1144 (2018)
4. Kasparov, G.: Chess, a *Drosophila* of reasoning. *Science* **362**(6419), 1087 (2018)
5. Strogatz, S.: One Giant Step for a Chess-Playing Machine. *New York Times*, New York. (2018, December 26)
6. Noam Brown, T.S.: Superhuman AI for multiplayer poker. *Science* **365**(6456), 885–890 (2019)
7. Bertsekas, D.: Reinforcement Learning and Optimal Control. Athena Scientific, Nashua, NH (2019)
8. Borkar, V., Meyn, P.: The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.* **38**(2), 447–469 (2000)
9. Neftci, E., Averbeck, B.: Reinforcement learning in artificial and biological systems. *Nat. Mach. Intell.* **1**, 133–143 (2019)

# Chapter 2

## Reinforcement Learning for Distributed Control and Multi-player Games



Bahare Kiumarsi, Hamidreza Modares, and Frank Lewis

**Abstract** This chapter presents the optimal control solution using reinforcement learning (RL). RL methods can successfully learn the solution to the optimal control and game problems online and using measured data along the system trajectories. However, one major challenge is that standard RL algorithms are data hungry in the sense that they must obtain a large number of samples from the interaction with the system to learn about the optimal policy. We discuss data-efficient RL algorithms using concepts of off-policy learning and experience replay and show how to solve  $H_2$  and  $H_\infty$  control problems, as well as graphical games using these approaches. Off-policy and experience replay-based RL algorithms allow reuse of data for learning and consequently lead to data-efficient RL algorithms.

### 2.1 Introduction

Optimal feedback control design has significantly contributed to the successful performance of engineered systems in aerospace, manufacturing, industrial processes, vehicles, ships, robotics, and elsewhere. The classical optimal control methods rely on a high-fidelity model of the system under control to solve Hamilton–Jacobi equations that are partial differential equations giving necessary and sufficient condition for optimality. However, a high-fidelity model of the system is not available for many real-world applications. Therefore, standard solutions to the optimal control problems require complete and accurate knowledge of the system dynamics and cannot

---

B. Kiumarsi (✉) · H. Modares  
Michigan State University, 428 S Shaw Ln, East Lansing, MI 48824, USA  
e-mail: [kiumarsi@msu.edu](mailto:kiumarsi@msu.edu)

H. Modares  
e-mail: [modaresh@msu.edu](mailto:modaresh@msu.edu)

F. Lewis  
University of Texas at Arlington, 701 S Nedderman Dr, Arlington, TX 76019, USA  
e-mail: [lewis@uta.edu](mailto:lewis@uta.edu)

© Springer Nature Switzerland AG 2021  
K. G. Vamvoudakis et al. (eds.), *Handbook of Reinforcement Learning and Control*,  
Studies in Systems, Decision and Control 325,  
[https://doi.org/10.1007/978-3-030-60990-0\\_2](https://doi.org/10.1007/978-3-030-60990-0_2)

cope with uncertainties in dynamics. While adaptive control techniques can successfully cope with system uncertainties, they are generally far from optimal.

Reinforcement learning (RL) [1–5], a branch of machine learning (ML) inspired by learning in animals, bridges the gap between traditional optimal control and adaptive control algorithms by finding the solutions to the Hamilton–Jacobi equations online in real time for uncertain physical systems. RL algorithms are mainly based on iterating on two steps, namely policy evaluation and policy improvement [1]. The policy evaluation step evaluates the value of a given control policy by solving Bellman equations, and the policy improvement step finds an improved policy based on the evaluated value function. RL algorithms can be categorized into on-policy RL and off-policy RL algorithms depending on whether the policy under evaluation and the policy that is applied to the system are the same or not. That is, in on-policy RL algorithms, the policy that is applied to the system to collect data for learning, called behavior policy, must be the same as the policy under evaluation. This makes on-policy RL algorithms data hungry as for each policy to be evaluated a new set of data must be collected and the algorithm does not use any past data collected.

Off-policy and experience replay-based RL for learning feedback control policies take advantage of episodic memory. In off-policy RL algorithms, the data collected by applying an appropriate behavior policy to the system dynamics can be reused to evaluate as many policies as required, until an optimal policy is found. That is, the policy under evaluation, called target policy, reuses the data collected from applying the behavior policy to the system. This makes off-policy learning data-efficient and fast since a stream of experiences obtained from executing a behavior policy is reused to update several value functions corresponding to different target policies. Moreover, off-policy algorithms are more practical compare to on-policy RL approaches in the presence of disturbance since the disturbance input does not need to be adjusted in a specific manner as required in on-policy RL algorithms. They also take into account the effect of probing noise needed for exploration. On the other hand, experience replay-based RL algorithms have memories of high-value sample (past positive and negative experiences with large Bellman errors) and reuse them to not only achieve faster convergence but also relax the requirement on convergence to optimal feedback solutions. That is, instead of requiring standard persistence of excitation (PE) condition to guarantee convergence, which is generally impossible to check online, a condition on a rank of a matrix can guarantee convergence, which is easy to check.

In this chapter, we discuss how RL has been applied to find the optimal solutions for both single-agent and multi-agent problems without requiring complete knowledge about the system dynamics. In Sect. 2.2, we present the optimal control problems for continuous-time (CT) dynamical systems and its online solutions using both on-policy and off-policy RL algorithms. This includes optimal regulation problem and  $H_\infty$  problem as well as experience replay technique. In Sect. 2.3, we discuss Nash game problems and their online solutions. The RL solution to games on graphs is presented in Sect. 2.4. Output synchronization of distributed multi-agent systems using off-policy RL approach is discussed in Sect. 2.5. Finally, we provide open research directions in Sect. 2.6.