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

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

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, Bušic', 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 — Kyriakos G. Vamvoudakis
Arlington, USA — Yan Wan
Arlington, USA — Frank L. Lewis
Durham, USA — Derya Cansever

Contents

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

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_1

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 AlpaZero, 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_∞ 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 optimally. 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

H. Modares
e-mail: modaresh@msu.edu

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

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

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_∞ 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.