# Control Systems and Reinforcement Learning

A high school student can create deep Q-learning code to control her robot, without any understanding of the meaning of "deep" or "Q," or why the code sometimes fails. This book is designed to explain the science behind reinforcement learning and optimal control in a way that is accessible to students with a background in calculus and matrix algebra. A unique focus is algorithm design to obtain the fastest possible speed of convergence for learning algorithms, along with insight into why reinforcement learning sometimes fails. Advanced stochastic process theory is avoided at the start by substituting random exploration with more intuitive deterministic probing for learning. Once these ideas are understood, it is not difficult to master techniques rooted in stochastic control. These topics are covered in the second part of the book, starting with Markov chain theory and ending with a fresh look at actor-critic methods for reinforcement learning.

Sean Meyn is a professor and holds the Robert C. Pittman Eminent Scholar Chair in the Department of Electrical and Computer Engineering, University of Florida. He is well known for his research on stochastic processes and their applications. His award-winning monograph *Markov Chains and Stochastic Stability* with R. L. Tweedie is now a standard reference. In 2015, he and Professor Ana Bušić received a Google Research Award recognizing research on renewable energy integration. He is an IEEE Fellow and IEEE Control Systems Society distinguished lecturer on topics related to both reinforcement learning and energy systems.

# Control Systems and Reinforcement Learning

Sean Meyn
*University of Florida*

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781316511961
DOI: 10.1017/9781009051873



First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

*A catalogue record for this publication is available from the British Library.*

ISBN 978-1-316-51196-1 Hardback

March 9th of 2021 was one of the saddest days on record for me and I'm sure most of my colleagues. On this day, Kishan Baheti was taken from us by the COVID virus.

My dear friend Kishan, this book is dedicated to you.

# Contents

# Preface

During the spring semester of 2020, I was delivering my stochastic control course, for which the final weeks always focus on topics in reinforcement learning (RL). Throughout the semester, I was thinking ahead to crash courses on this topic to be delivered later in the year: two summer courses planned in Paris and Berlin, and another scheduled as part of the Simons Institute program on reinforcement learning.[1] A pandemic altered my travel plans, but it also gave me time to reflect on how to better teach this difficult material.

Soon after the spring semester ended, I was contacted by Diana Gillooly, then an editor at Cambridge University Press. She wrote "someone mentioned that you were scheduled to deliver lectures on reinforcement learning," and asked if I might be interested in writing a book on the topic. Her brief email set in motion the pages that lie in front of you.

There is of course a longer history: the book is a product of handouts I've prepared over more than a decade, and bits and pieces of papers and book chapters prepared over a longer period.

However, I promised my co-organizers of the Simons Institute RL program that I would present a crash course for true beginners, without heavy mathematics. I also promised myself I would write a book that would be accessible to senior undergraduates and graduate students, provided they came with sufficient motivation. These pandemic-induced contemplations led to two themes that I felt a need to spell out:

(i) Within the control systems literature, there are dynamic programming techniques to approximate the Q-function that appears in reinforcement learning. In particular, this "value function" is the solution to a simple convex program (an example is the "DPLP" appearing in Eq. (3.36)). Many of the algorithms in reinforcement learning are designed to approximate the same function but are based on root finding problems that are not well understood outside very special cases.

This is just one example of the need to build better bridges between control and RL. I cannot claim the bridge is fully built. My hope is that the book will provide leads for future discoveries based on insights from each discipline.

(ii) *Stochastic approximation* (SA) is the most common method for analysis of recursive algorithms; this approach is commonly known as the *ordinary differential equations (ODE) method* [136, 229, 301, 357]. The relationship between RL and SA was recognized soon after Watkins introduced Q-learning [169, 352], and ODE methods for

[1] Video and slides from the 2020 Fall program are now available at `https://simons.berkeley.edu/programs/rl20`.

analysis of optimization algorithms have grown in sophistication over the past decade [198, 318, 335, 375]. Related ODE methods are part of a standard modeling framework in statistical mechanics, genetics, epidemiology (e.g., the SIR model), and even voting [24, 122, 225, 276].

The narrative is flipped in this book: rather than treating an ODE as simply an analytical tool, every algorithm in this book begins with an ideal ODE that is regarded as "step 1" in algorithm design. I believe this provides better insight into algorithm synthesis and analysis.

However, justification of this approach using SA is highly technical. In particular, the recent thesis and book chapter [107, 110] (which build on a similar narrative) assume significant background in the theory of stochastic processes. In this book, we lift the veil: there is nothing inherently *stochastic* about *stochastic approximation*, provided you are willing to work with sinusoids or other deterministic probing signals instead of stochastic processes. The ODE methods surveyed in Chapters 4 and 5 make no reference to probability theory.

This is my third book, and like the others the writing came with discoveries. While working on theme (i), my colleague Prashant Mehta and I found that *Convex Q-Learning* could be made much more practical by borrowing batch RL concepts that are currently popular. This led to the new work [246, 247] and new collaborations with Gergely Neu. You will find text and equations from these papers scattered throughout Chapters 3 and 5.

Chapter 4 on ODE methods and *quasi-stochastic approximation* was to be built primarily on [40, 41]. Over the summer of 2020, all of this material was generalized to create a complete theory of convergence and convergence rates for these algorithms, along with a better understanding of their application to both *gradient-free optimization* and *policy gradient* techniques for RL [85–87].

The second part of the book, such as *Zap Zero* for Q-learning, and insights on the rate of convergence for actor-critic methods. Each chapter ends with a "Notes" section that provides an overview of the origins of the main conclusions.

Many newcomers to reinforcement learning may be disappointed to see that the theory and algorithms in this book are far removed from the dream portrayed in the popular media: reinforcement learning is often described as an "agent" interacting in a physical environment, and maturing as it gains experience. Unfortunately, given today's technologies, the process of "learning from scratch while you control" is unlikely to succeed outside very special cases such as online advertising.

The tone of this book is entirely different: we pose an optimal control problem, and show how to obtain an approximate solution based on design of exploration strategies and tuning rules. This is not an eccentricity of the author but a disciplined and accepted approach to derive all of the standard approaches to reinforcement learning. In particular, the Q-learning algorithm of Watkins and its extensions are designed to solve or approximate the "dynamic programming" equations introduced in the 1950s.

The field is young, and its future may look something like the dream you had in mind before you read this preface. I hope that in the near future we will discover new paradigms for RL, perhaps drawing inspiration from intelligent living beings rather than optimality equations from the past century. I have confidence that the fundamental principles in this book will remain valuable without the shackles of the optimal control paradigm!

### *Acknowledgments*

Let's begin three decades back. In the mid-1990s I (figuratively) won the lottery: a Fulbright fellowship that took me and my family, including my young daughters, Sydney and Sophie, to Bangalore, India. Those nine months at the Indian Institute of Science (IISc) with Vivek Borkar were the start of fruitful collaborations and a long-lasting friendship. Vivek's presence can be felt on nearly every page of the second half of this book.

I was also fortunate to have interactions with Ben Van Roy while he was completing his dissertation research at the Massachusetts Institute of Technology (MIT). His work with John Tsitsiklis is an absolute tour-de-force. Many aspects of the book draw from this early RL research. His current research is likely to have similar long-lasting impact.

Prashant Mehta once said to me, "I know how you do it! You surround yourself with amazing people!" Amazing is right, and he is right there at the top. This book is a product of collaborations with Vivek, Prashant, and many others, including Ana Bušić, Ken Duffy, Peter Glynn, Ioannis Kontoyiannis, Eric Moulines, and many old friends at the United Technologies Research Center (UTRC), including Amit Surana and George Mathew. My PhD advisor Peter Caines was my first and one of my all-time best colleagues, who enthusiastically supported my first investigations into Markov chain theory. This set the stage for collaborations with Richard Tweedie that began during my postdoctoral stay at the Australian National University. All amazing people, anyone will agree.

Younger stars that influenced my research include Shuhang Chen, who recently defended his dissertation in the math department and is lead author of [88] on finer ODE methods. Many thanks to current graduate student Fan Lu for comments on early drafts and assistance with numerical experiments.

Prabir Barooah helped to draw me to the University of Florida, from my former home at the University of Illinois. I've benefited from our interactions, and with his students including Naren Raman and Austin Coffman.

Max Raginsky helped me to navigate the literature outside of my usual orbit. In particular, his advice along with Polyak's recent survey [136] helped me to understand the contributions from the USSR in the early days of RL and SA. Max's research is also an inspiration: while references to his work are scattered throughout the book, much of this material is suited for a more advanced monograph.

Much of Chapters 2 and 3 is based on the state space control course created at the Decision and Control Laboratory at the University of Illinois. Many thanks to Bill Perkins, Tamer Basar, and Max Raginsky for allowing me to borrow material from the course manuscript [29], and laboratory manager Daniel Block for leading the design of innovative control experiments.

In 2018, I was fortunate to spend several months at the National Renewable Energy Laboratory (NREL), where I conducted research at the Autonomous Energy Systems laboratory. One outcome of these interactions was research on stochastic approximation, leading to the articles [40, 41, 85–87, 93]. The book would not be the same without collaborations at NREL with Andrey Bernstein, Marcello Colombino, Emiliano Dall'Anese, and my former graduate student Yue Chen.

In reviewing the literature on extremum seeking control for Chapter 4, I was skeptical of the common claim in the research literature that the idea began in the 1920s. The most convincing case for this history was made in [348]. I contacted coauthor Iven Mareels, who

reassured me that this history was accurate. Then, with help from colleagues in France, I found and translated the 1922 document [217] that is considered the source of this optimization technique.

One of the great "bridge builders" at the intersection of RL and control theory is Frank Lewis, who has led the creation of several collected volumes on these topics. I was surprised when he thought of me one decade ago, leading to the contribution [165], and very enthusiastic when he invited me to contribute to a new volume one decade later [110].

Until recently, I have regarded RL as a hobby, as motivation for simplified models for complex systems (such as networks [254]), and as a vehicle to teach control theory. This changed with the arrival of Adithya Devraj to the University of Florida, where he pursued graduate studies with me until he graduated and departed for Stanford in the spring of 2020. His curiosity and intellect were an inspiration in many ways, and in particular drove me to learn more about the evolution of RL over the past decade. Many of the figures and much of the theory in the second part of this book are taken from his dissertation [107]. He also provided suggestions that improved many parts of the book.

I owe the Simons Institute a great debt. During the spring of 2018, I was a long-term visitor during its program on real-time decision making, and was doubly fortunate to be joined by Ana Bušić and Adithya Devraj. We learned a great deal from fellow visitors, and Peter Bartlett (along with other locals). Our discussions back then helped to motivate the 2020 program on RL, which provided a massive crash course on every aspect of the subject, with a strong emphasis on the sort of bridge building I am attempting to pursue with this book. In the fall of 2020, I watched tutorials on recent actor-critic techniques just before finishing Chapter 10 on this very topic. The book benefited from the `rltheory` virtual seminar series, organized by Gergely Neu, Ciara Pike-Burke, and Csaba Szepesvári,[2] which was also inspired by the 2020 RL program at Simons.

Returning to the present: in the spring of 2021, I created a new course based on Part I of this book. Many of the students were hungry for an accessible treatment of both control systems and RL, and survived the sometimes difficult and rocky three months. I appreciate all of the feedback I received over the semester, and did my best to respond. You can thank Arielle Stevens for correcting many confusing passages in the first three chapters, and for the gray boxes used to highlight important concepts. Many more improvements were made in response to input from other students, including Caleb Bowyer, Bo Chen, Austin Coffman, Chetan Dhulipalla, Weihan Shen, Zetong Xuan, Kei-Tai Yu, and Yongxu Zhang. Also on this list is the recent graduate Dr. Bob Moye and current graduate students conducting research with me on RL and related topics: Mario Baquedano Aguilar, Caio Lauand, and Amin Moradi.

I also received substantial feedback from current PhD candidate Vektor Dewanto soon after a draft manuscript was posted on Twitter in August of 2021.

Of course, I cannot forget my sponsors. Bob Bonneau at the Air Force Office for Scientific Research (AFOSR) encouraged funding for early research with Prashant Mehta on Q-learning, mean-field games, and nonlinear filtering. Derya Cansever and Purush Iyer at the Army Research Office (ARO) have funded more recent research on related topics. The National Science Foundation (NSF) has funded my most abstract and seemingly worthless

[2] `https://sites.google.com/view/rltheoryseminars/home`.

research topics, which hopefully led to something of value. My most reliable ally at NSF was Radhakisan (Kishan) Baheti, who recommended funding for my very first grant (on the topic of adaptive control, at the start of the 1990s). He was a fantastic mentor, alert to potentially foolish ideas, and also inspired by new if potentially useless research directions. He knew what everyone in the control community was up to! He also inspired all of us with his marathon runs and mastery of yoga.

– Sean Meyn, August 1, 2021