# Dedication

*Marty: To my wife, Dodie Katzenstein, who made helpful suggestions and encouraged me to write this book, despite claiming she has no idea what it is about.*

*Tim: To my family – mom, dad, Laura, Euclid, Escher – who support me in everything I do, and can now stop asking when the book will be done.*

# Preface

When we set out to write this book, at the onset of the COVID-19 pandemic, our goal was to provide an accessible and logically structured introduction to Markov decision process (MDP) theory, applications and algorithms. As our writing and research progressed, we came to appreciate the vast and rapidly evolving landscape of reinforcement learning as a major sub-field within artificial intelligence (AI)[1]. Consequently, a significant portion of this book now focuses on reinforcement learning methods. The advantage of grounding our treatment in Markov decision processes is that they offer a rigorous and unifying modeling framework, one that applies even in *model-free* reinforcement learning settings in which system dynamics are not fully known.

This book is not meant to be encyclopedic. Rather, we have chosen to focus on the core concepts of these disciplines, with the objective of providing a solid basis for further study, application, and research. Throughout, simple and transparent models illustrate how the fundamental Markov decision process entities – states, actions, transition probabilities and rewards – apply in each context. A sound grasp of these fundamentals is necessary for use in more complex settings. In our experiences reviewing journal submissions, we have observed that model formulations are frequently imprecise, resulting in ambiguity in explaining and interpreting results.

Numerous computational examples in this book illustrate how algorithms work in practice. Commentary accompanying algorithms and in computational examples provides insights into algorithmic details and guidelines for application. Hands-on coding and tuning is a necessary step in learning to use these methods, an aspect that is often missed when learners rely solely on reusing pre-written code found online.

Markov decision processes arise in many fields, most notably operations research, robotics, computer science, engineering, economics, and management. While our approach primarily reflects an operations research perspective, the concepts and tools presented here are broadly applicable. We believe this book will provide a solid foundation for anyone seeking a rigorous introduction to these topics, especially within the growing reinforcement learning community.

A word of caution: the extremely powerful methods in this book are broadly applicable yet have the potential to be misused. We encourage you to apply them responsibly. In the spirit of Google's original guiding principle, "Don't be evil."

[1] Coincidentally, ChatGPT, which leveraged reinforcement learning in its development, has assisted us in editing text, generating code, and producing figures.

### Book objectives

The book aims to provide readers with a solid foundation in Markov decision process principles and algorithms, reinforcement learning methods, and the relationship between the two. Readers will learn to apply Markov decision processes, identify model components, solve models using exact or approximate methods, and interpret solutions.

We strongly believe that working through examples and coding algorithms is the best way to learn this material. Accordingly, through numerous examples, the book includes extensive numerical calculations and discussions of their implications. Additional exercises are provided at the end of each chapter to further illustrate applications and reinforce key concepts.

### Book structure

The book is organized as follows:

- **Introduction:** Chapter 1 describes the foundations and applications of Markov decision processes and reinforcement learning. In addition it provides a historical perspective on the evolution of Markov decision processes in operations research and control theory, and reinforcement learning in behavioural science and artificial intelligence, setting the stage for the integration of these perspectives in contemporary research and applications.

- **Part I - Fundamentals:** Chapters 2-3 develop and apply a rigorous modeling foundation. Chapter 2 introduces basic model components and optimality criteria, illustrates elementary calculations that form the building blocks of more complex algorithms, and includes a numerical example that appears throughout the book to illustrate new ideas and methods. A diverse set of applications in Chapter 3 provides the reader with real-world examples of Markov decision process formulations.

- **Part II - Classical Markov decision process models:** Chapter 4 covers finite horizon MDPs, while Chapters 5-7 address infinite horizon MDPs under several optimality criteria. Chapter 5 focuses on discounted models, which are most widely used and theoretically complete. Chapter 6 characterizes and analyses expected total reward models, also known as episodic models, which are foundational in reinforcement learning. Chapter 7 examines models under the long-run average reward criteria. Finite and infinite horizon partially observable Markov decision processes (POMDPs) are the subject of Chapter 8.

- **Part III - Reinforcement learning:** Chapters 9-11 describe methods for solving large-scale, model-based and model-free sequential decision processes. Chapter 9 combines value function approximation with value iteration, policy iteration and linear programming to find approximate solutions for Markov decision process. This material is often referred to as *approximate dynamic programming.*

Simulation-based methods that apply in both a model-free and model-based environment are the subject of Chapter 10. This chapter considers *tabular models* which are of a size that allows value functions, state-action value functions, and policies to be represented explicitly in look-up tables. Chapter 11 extends these results to large-scale models by combining the approximation methods of Chapters **??** and **??**.

- **Appendices:** Book appendices summarize key mathematical notation and conventions used throughout the book, as well as background on Markov chains and linear programming. Some chapters have their own appendices, which provide supplementary technical details and relevant background. Topics include stochastic approximation, regression, and gradient descent.
- **Bibliographic remarks:** Each chapter includes a brief section that provides historical commentary and references.
- **Exercises:** Exercises at the end of each chapter offer opportunities to formulate models, apply methods, or delve into theoretical details. Many are open-ended. Alternatively, readers are encouraged to choose their own models to formulate and try out the methods in the book, as we have done with the recurring two-state model, queuing control model and coffee-delivering robot model (featured on the book cover). Starred exercises indicate increased difficulty.
- **Starred sections:** Sections with an asterisk (*) contain important, technically demanding or supplementary material. They may be skipped on first reading.
- **Gray boxes:** In addition to theoretical results, algorithms, and examples, important equations and comments are placed in gray boxes throughout the book.

### Relationship to Puterman [1994]

This book is not a revision of [Puterman, 1994] which was written to be a comprehensive and state-of-the-art reference targeted at researchers and advanced graduate students with strong mathematics backgrounds. Our goal in writing this book is to provide a rigorous yet more accessible introduction to Markov decision processes, reinforcement learning and their extensions that will be accessible to advanced undergraduate students, early graduate students, and practitioners. Since the field has rapidly expanded since the early 1990s, especially in the area of reinforcement learning, this book develops these concepts using the common notation and language of Markov decision processes.

Some specific differences with Puterman [1994] include:

- A broad discussion of reinforcement learning methods, many of which were not widely known, especially in the operations research community, in the early 1990s.

- Formulation and analysis of partially observable Markov decision processes (POMDPs).
- Descriptions of state-action value functions permeate the book in anticipation of their fundamental role in reinforcement learning methods.
- Increased emphasis on randomized policies in light of their fundamental role in linear programming, policy gradient and actor-critic methods.
- Illustrating all algorithms with numerical calculations applied to common examples. All computations were done from scratch using R [R Core Team, 2024]. With the rapidly advancing proficiency of AI models to generate useful code and translate code between languages, readers are encouraged to engage hands-on in the coding process (perhaps with AI assistance) in order to truly understand the workings of each algorithm.
- Omission of material on countable state models, multi-chain average reward MDPs and sensitive optimality criteria including Blackwell optimality.
- Limited bibliographic references. With the rapid development of internet resources as well as the presence of some excellent books, historical perspectives are readily available elsewhere.
- Representing component-wise maxima by “c-max” to avoid misinterpretation of “max” when using vector notation.
- Using $r(s,a,j)$ as the basic definition of the reward function instead of $r(s,a)$.
- Emphasizing transient and stochastic shortest path models in the chapter on infinite horizon models with the expected total reward criterion. The advantage of this categorization is that it applies easily to what have become known as *episodic* models. The earlier book adopted a classic approach based on positive and negative models.

### Neural networks

The book intentionally omits the extensive and rapidly evolving body of research and application of *deep reinforcement learning*, a subfield of reinforcement learning that leverages neural network approximations. While neural networks have played a central role in many breakthroughs in reinforcement learning, we have chosen not to cover neural network design or training. Readers interested in learning about and applying neural networks are encouraged to consult dedicated texts on deep learning that describe network architectures, backpropagation, and training strategies.

Our focus is on foundational principles and the core Markov decision process and reinforcement learning theory and algorithms. Understanding these foundations is essential regardless of whether function approximation is through tabular representations, linear functions of features, or deep neural networks. For this reason, we believe a mastery of the basics should precede the study of more complex approximation techniques.

### How to use this book

Our intention in writing this book is that readers find the material self-explanatory and accessible so that independent study is possible. The following description of course structures might provide guidance in working through it. The requisite background includes a solid foundation in probability, linear algebra, real analysis and statistical estimation.

Every course should at a minimum cover Chapters **??**-**??**. Subsequent chapters should be chosen to reflect course and learning objectives.

- An introductory Markov decision process course with an operations research orientation should supplement the above chapters with selected material on optimality equations and algorithms from Chapter **??** on infinite horizon discounted models and Chapter **??** on episodic models. The model formulation in Chapter **??** of partially observable MDPs and the material on simulation of episodic and infinite horizon discounted models in Chapter **??** nicely complements the above material.

- A more advanced MDP course should delve more deeply into theoretical issues in Chapters **??** and **??**, and also include material on average reward models in Chapter **??**, supplemented by material in Chapter 9 of Puterman [1994].

- A reinforcement learning course with a computer science orientation should supplement the material in Chapters **??**-**??** with methods chosen from Chapters **??**-**??**. Such a course should delve more deeply into the topics in Section **??**, especially neural networks. Instructors should add material on neural network approximations of value functions, state-action value functions and policies, as well as focusing on examples arising in robotics, vehicle guidance and generative models.

### Additional resources

Additional resources and errata can be found at `https://www.cambridge.org/9781009098410`.

### Our cover image

Our life-long friend, Flora Gordon, designed our whimsical and aptly-themed cover with assistance from her colleague Sergio Lopez. The robot, who we refer to as “Flow”, appears in the Gridworld model of Section **??**, an application that recurs throughout the book to illustrate Markov decision process and reinforcement learning algorithms. This example is inspired by the Hungarian mathematician Alfréd Rényi who is attributed with the quote:

*A mathematician is a machine for turning coffee into theorems.*

Moreover, robotics has now become one of the principal application areas of reinforcement learning.

### Acknowledgments

We would like to acknowledge many individuals who assisted us on this journey. Significant contributors include Abhijit Gosavi and Nicolas Gast, whose insights helped shape Chapters 10 and 11 (AG) and Chapter 7 (NG). Special thanks go to Antoine Sauré who provided extensive and insightful comments on Chapters 9-11.

We are also grateful to those who have provided suggestions and guidance including Alan Mackworth, Rich Sutton, Reid Swanson, Steven Shechter, Dmitri Bertsekas, John Tsitsiklis, Sergey Levine, Bruno Scherrer and Gergely Neu. Others have offered minor editorial improvements and suggestions, and checked calculations.

Elise Liu, while still in high school and now at The University of St. Andrews, produced many of the book's figures. Discussions with Pritam Dash provided us with insight into a computer science graduate student's perspective on this material. Moreover, he developed a Github repository for code and text. Neal Kaw and Michael Gimelfarb developed exercise solutions and several figures for the early chapters. Ken Wong also contributed to the figures, as well as the bibliography.

In addition, we thank Lauren Cowles and Arman Chowdhury for their timely encouragement, attention to detail, and expert advice on manuscript preparation, and Diana Gillooly who was our initial contact with Cambridge University Press.

# Bibliography

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons., 1994.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.