Introduction to Markov Decision Processes

Martin L. Puterman and Timothy C. Y. Chan

July 17, 2024

Part II - Classical Markov Decision Process Models

If I have seen further, it is by standing on the shoulder of giants.

Sir Isaac Newton, Mathematician, physicist, astronomer and ... ,
1642-1727

This section of the book contains five chapters:

- Chapter 4: Finite horizon models
- Chapter 5: Infinite horizon models: Expected discounted reward
- Chapter 6: Infinite horizon models: Expected total reward
- Chapter 7: Infinite horizon models: Long-run average reward
- Chapter 8: Partially observable Markov decision processes

II.1 Overview

The chapters in this part of books describe the models, theory, and algorithms of Markov decision processes. Markov decision processes provide a rigorous mathematical framework for structuring reinforcement learningproblems. The work of the giants of Markov decision processes ; Bellman, Howard, Blackwell, Derman and Veinott have in many ways contributed to the wide impact of reinforcement learning.

The core idea in multi-period Markov decision processes builds directly from the one-period model from Chapter **??**: a decision maker aims to identify an action that balances the immediate reward from choosing that action and the future reward that can be gained from the new state to which the process transitions. However, instead of the process ending in the new state as in the one-period model, the process will continue for possibly many more stages. Thus, the decision maker needs to consider the possible actions that can be taken in that (and all other) subsequent states, and the corresponding rewards that can be generated.

For example, consider a driver in traffic trying to get to their destination as quickly as possible. At an intersection, there is an option to take a side street that appears less busy. Choosing this action might lead to more immediate progress, but comes with uncertainty about whether subsequent streets at the next intersection will be busy or whether it will be possible to merge back on to the main street easily. Alternatively, the driver may choose to stay the course, going less slowly in the near term, but possibly making up time later by staying on the main street.

II.2 Chapter details

Chapter **??** studies finite horizon models, those that end at a fixed future stage that is known at the start of the decision horizon. It introduces the concepts of the value of a policy, an optimal policy, the optimal value function, optimality (Bellman) equations and backwards induction. It illustrates these concepts through examples and also shows how to establish the structure of an optimal policy. Early dynamic programming research focused on establishing *structural results* especially in inventory control. Scarf [1960] provides a quintessential example by providing conditions under which an (s, S) is optimal.

Chapters **??**-**??** focus on infinite horizon models, each with a different optimality criterion. The reason for separating these chapters on the basis of optimality criteria is that each requires different theoretical considerations.

- In **discounted models** the inclusion of a non-negative discount factor less than one ensures under boundedness of rewards that values are well-defined and bounded. Moreover this enables analysis based on properties of contraction mappings.

- **Expected total reward models** require assumptions on rewards or transition probability to ensure values are well defined. They are most widely applied to *episodic* models which reach a reward-free absorbing state in a finite but variable time under some or all policies. A special class of these models referred to as *transient* inherit some of the contraction properties of discounted models.

- Analysis of **average reward models** depends strongly on the structure of underlying Markov chains. Most complete results are available for models in which Markov chains generated by stationary policies are aperiodic and result in all states be accessible from each other. Generalizations to models with transient states and multiple recurrent classes require more subtle analyses. Moreover the concept of the *bias* of a policy is fundamental to theory and computation.

As closely as possible, each of these chapters evolves are follows. After defining optimality, they establish the existence and optimality of solutions of appropriate Bellman equations and that stationary deterministic policies are optimal in the set of all policies. They then show how to find optimal values and policies using value iteration, policy iteration and its variants, and linear programming.

Chapter **??** is rather distinct. It studies finite and infinite-horizon partially observable Markov decision processes (POMDPS), where instead of knowing the state with certainty, the decision maker only observes a signal that varies with of the true unobservable state of the the process. Decisions need to account for this extra source of uncertainty and require methods that apply to continuous state spaces.

II.3 Learning

Learners should focus on developing a fundamental understanding of the Markov decision process modeling framework, the Bellman equations, algorithms that find optimal policies and how the model applies to real problems.

The best way to learn this material is through analyzing simple examples. Models with a few states and actions, such as that in Section **??**, provide excellent vehicles to try out algorithms. With the exception of linear programming we encourage you to code these algorithms yourself to gain a deep understanding of how they work.

By extending analyses to larger models, one should gain an appreciation of the challenges faced when a model has a large number of states and/or actions.

Bibliography

H. Scarf. The optimality of (s, s) policies in the dynamic inventory problem. In S.Karlin K. Arrow and P. Suppes, editors, *Studies in the Mathematical Theory of Inventory and Production*, pages 196–202. Stanford University Press, Stanford,CA, 1960.