Part II - Classical Markov Decision Process Models

If I have seen further it is by standing on the shoulders of giants[1].

Sir Isaac Newton, mathematician, physicist, astronomer and much more, 1642-1727.

This part 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

Overview

The chapters in this part describe the classical models, theory, and algorithms of Markov decision processes. The core idea in multi-period Markov decision processes builds directly on the one-period model in Chapter 2: 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 many more decision epochs. Thus, the decision maker

[1] Newton [1959].

needs to consider the possible actions that can be taken in all 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 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.

The best way to learn this material is through analyzing simple examples. Models with a few states and actions, such as the two-state example from Section 2.5, provide excellent vehicles to try out algorithms. By extending analyses to larger models, you will gain an appreciation of the challenges faced when a model has a large number of states and actions. Approaches for such models are the subject of the third part of the book.

Chapter details

Chapter 4 studies finite horizon models, those that end at a fixed future stage that is known at the start of the planning horizon. It introduces the concepts of the value of a policy, an optimal policy, the optimal value function, the optimality (Bellman) equation, and backwards induction. It illustrates these concepts through examples and also shows how to establish the structure of an optimal policy.

Chapters 5-7 focus on infinite horizon models, each with a different optimality criterion. The reason for this separation is that each requires different theoretical considerations.

- In **discounted reward 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.

- In **total reward models**, assumptions on rewards or transition probabilities are needed to ensure values are well defined. These models are most widely applied to *episodic* problems, 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 explicitly on the structure of Markov chains corresponding to stationary policies. Most complete results are available when these Markov chains are aperiodic and in which all states are accessible from each other. Generalizations to models with transient states and multiple recurrent classes require more subtle analyses. Moreover the auxiliary concept of the *bias* of a policy is fundamental to theory and computation.

As closely as possible, each of these chapters evolves as follows. After defining optimality, they establish the existence and optimality of solutions of the appropriate Bellman equation, and that stationary deterministic policies are optimal in the class 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 8 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 is dependent on the true unobservable state of the process. Decisions need to account for this extra source of uncertainty, which leads to analyzing related models with a continuous state space corresponding to probability distributions on the unobservable state. This is the only chapter which considers continuous state spaces.