

# Preface to the 2nd Edition

Markov decision process (MDP) models are widely used for modeling sequential decision-making problems that arise in engineering, computer science, operations research, economics, and other social sciences. However, it is well known that many real-world problems modeled by MDPs have huge state and/or action spaces, leading to the well-known curse of dimensionality, which makes solution of the resulting models intractable. In other cases, the system of interest is complex enough that it is not feasible to explicitly specify some of the MDP model parameters, but simulated sample paths can be readily generated (e.g., for random state transitions and rewards), albeit at a non-trivial computational cost. For these settings, we have developed various sampling and population-based numerical algorithms to overcome the computational difficulties of computing an optimal solution in terms of a policy and/or value function. Specific approaches include multi-stage adaptive sampling, evolutionary policy iteration and random policy search, and model reference adaptive search. The first edition of this book brought together these algorithms and presented them in a unified manner accessible to researchers with varying interests and background. In addition to providing numerous specific algorithms, the exposition included both illustrative numerical examples and rigorous theoretical convergence results. This book reflects the latest developments of the theories and the relevant algorithms developed by the authors in the MDP field, integrating them into the first edition, and presents an updated account of the topics that have emerged since the publication of the first edition over six years ago. Specifically, novel approaches include a stochastic approximation framework for a class of simulation-based optimization algorithms and applications into MDPs and a population-based on-line simulation-based algorithm called approximation stochastic annealing. These simulation-based approaches are distinct from but complementary to those computational approaches for solving MDPs based on explicit state-space reduction, such as neuro-dynamic programming or reinforcement learning; in fact, the computational gains achieved through approximations and parameterizations to reduce the size of the state space can be incorporated into most of the algorithms in this book.

Our focus is on *computational* approaches for calculating or estimating optimal value functions and finding optimal policies (possibly in a restricted policy space). As a consequence, our treatment does not include the following topics found in most books on MDPs:

- (i) characterization of fundamental *theoretical* properties of MDPs, such as existence of optimal policies and uniqueness of the optimal value function;
- (ii) paradigms for *modeling* complex real-world problems using MDPs.

In particular, we eschew the technical mathematics associated with defining continuous state and action space MDP models. However, we do provide a rigorous theoretical treatment of convergence properties of the algorithms. Thus, this book is aimed at researchers in MDPs and applied probability modeling with an interest in numerical computation. The mathematical prerequisites are relatively mild: mainly a strong grounding in calculus-based probability theory and some familiarity with Markov decision processes or stochastic dynamic programming; as a result, this book is meant to be accessible to graduate students, particularly those in control, operations research, computer science, and economics.

We begin with a formal description of the discounted reward MDP framework in Chap. 1, including both the finite- and infinite-horizon settings and summarizing the associated optimality equations. We then present the well-known exact solution algorithms, value iteration and policy iteration, and outline a framework of rolling-horizon control (also called receding-horizon control) as an approximate solution methodology for solving MDPs, in conjunction with simulation-based approaches covered later in the book. We conclude with a brief survey of other recently proposed MDP solution techniques designed to break the curse of dimensionality.

In Chap. 2, we present simulation-based algorithms for estimating the optimal value function in finite-horizon MDPs with large (possibly uncountable) state spaces, where the usual techniques of policy iteration and value iteration are either computationally impractical or infeasible to implement. We present two adaptive sampling algorithms that estimate the optimal value function by choosing actions to sample in each state visited on a finite-horizon simulated sample path. The first approach builds upon the expected regret analysis of multi-armed bandit models and uses upper confidence bounds to determine which action to sample next, whereas the second approach uses ideas from learning automata to determine the next sampled action. The first approach is also the predecessor of a closely related approach in artificial intelligence (AI) called Monte Carlo tree search that led to a breakthrough in developing the current best computer Go-playing programs (see Sect. 2.3 Notes).

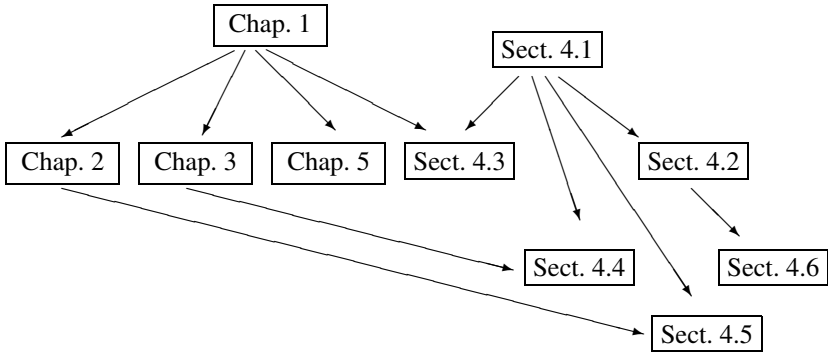
Chapter 3 considers infinite-horizon problems and presents evolutionary approaches for finding an optimal policy. The algorithms in this chapter work with a population of policies—in contrast to the usual policy iteration approach, which updates a single policy—and are targeted at problems with large action spaces (again

possibly uncountable) and relatively small state spaces. Although the algorithms are presented for the case where the distributions on state transitions and rewards are known explicitly, extension to the setting when this is not the case is also discussed, where finite-horizon simulated sample paths would be used to estimate the value function for each policy in the population.

In Chap. 4, we consider a global optimization approach called model reference adaptive search (MRAS), which provides a broad framework for updating a probability distribution over the solution space in a way that ensures convergence to an optimal solution. After introducing the theory and convergence results in a general optimization problem setting, we apply the MRAS approach to various MDP settings. For the finite- and infinite-horizon settings, we show how the approach can be used to perform optimization in policy space. In the setting of Chap. 3, we show how MRAS can be incorporated to further improve the exploration step in the evolutionary algorithms presented there. Moreover, for the finite-horizon setting with both large state and action spaces, we combine the approaches of Chaps. 2 and 4 and propose a method for sampling the state and action spaces. Finally, we present a stochastic approximation framework for studying a class of simulation- and sampling-based optimization algorithms. We illustrate the framework through an algorithm instantiation called model-based annealing random search (MARS) and discuss its application to finite-horizon MDPs.

In Chap. 5, we consider an approximate rolling-horizon control framework for solving infinite-horizon MDPs with large state/action spaces in an on-line manner by simulation. Specifically, we consider policies in which the system (either the actual system itself or a simulation model of the system) evolves to a particular state that is observed, and the action to be taken in that particular state is then computed on-line at the decision time, with a particular emphasis on the use of simulation. We first present an updating scheme involving multiplicative weights for updating a probability distribution over a restricted set of policies; this scheme can be used to estimate the optimal value function over this restricted set by sampling on the (restricted) policy space. The lower-bound estimate of the optimal value function is used for constructing on-line control policies, called (simulated) policy switching and parallel rollout. We also discuss an upper-bound based method, called hindsight optimization. Finally, we present an algorithm, called approximate stochastic annealing, which combines  $Q$ -learning with the MARS algorithm of Section 4.6.1 to directly search the policy space.

The relationship between the chapters and/or sections of the book is shown below. After reading Chap. 1, Chaps. 2, 3, and 5 can pretty much be read independently, although Chap. 5 does allude to algorithms in each of the previous chapters, and the numerical example in Sect. 5.1 is taken from Sect. 2.1. The first two sections of Chap. 4 present a general global optimization approach, which is then applied to MDPs in the subsequent Sects. 4.3, 4.4 and 4.5, where the latter two build upon work in Chaps. 3 and 2, respectively. The last section of Chap. 4 deals with a stochastic approximation framework for a class of optimization algorithms and its applications to MDPs.



Finally, we acknowledge the financial support of several US Federal funding agencies for this work: the National Science Foundation (under Grants DMI-9988867, DMI-0323220, CMMI-0900332, CNS-0926194, CMMI-0856256, EECS-0901543, and CMMI-1130761), the Air Force Office of Scientific Research (under Grants F496200110161, FA95500410210, and FA95501010340), and the Department of Defense.

Seoul, South Korea  
 Stony Brook, NY, USA  
 College Park, MD, USA  
 College Park, MD, USA

Hyeong Soo Chang  
 Jiaqiao Hu  
 Michael Fu  
 Steve Marcus

Simulation-Based Algorithms for Markov Decision  
Processes

Chang, H.S.; Hu, J.; Fu, M.C.; Marcus, S.I.

2013, XVII, 229 p. 49 illus., 1 illus. in color., Hardcover

ISBN: 978-1-4471-5021-3