

Hyeong Soo Chang, Michael C. Fu, Jiaqiao Hu,
and Steven I. Marcus

Simulation-Based Algorithms for Markov Decision Processes

SPIN Springer's internal project number, if known

– Monograph –

September 1, 2006

Springer

Berlin Heidelberg New York

Hong Kong London

Milan Paris Tokyo

Preface

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 that 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. This book brings together these algorithms and presents them in a unified manner accessible to researchers with varying interests and background. In addition to providing numerous specific algorithms, the exposition includes both illustrative numerical examples and rigorous theoretical convergence results. These 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 our proposed algorithms.

Our focus is on computational approaches for calculating/estimating optimal value functions and finding optimal policies (possibly in a restricted policy space). What this book does not make any attempts at doing:

- (1) providing a foundation for determining important theoretical and fundamental properties of the underlying MDP models, such as existence of optimal policies and uniqueness of the optimal value function;
- (2) teaching the use of MDPs for modeling real-world problems.

In particular, the technical mathematics that goes along with defining continuous state and action space MDP models is not addressed at all. However, we do provide a rigorous theoretical treatment of convergence properties of the algorithms. In sum, this book is aimed mainly at researchers in MDPs and applied probability modeling with an interest in numerical computation. The prerequisites for understanding the technical content are relatively mild: mainly a strong grounding in calculus-based probability theory and some familiarity with Markov decision processes or stochastic dynamic programming; thus, 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 general discounted cost MDP framework in Chapter 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 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 Chapter 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.

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 Chapter 4, we consider a global optimization approach called model reference adaptive search, 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 it to various MDP settings. For the finite and infinite horizon settings, we show how it can be

used to perform optimization in policy space. In the setting of Chapter 3, we show how it can be incorporated to further improve the exploration step in the evolutionary algorithms presented there. Finally, for the finite horizon setting with both large state and action spaces, we propose a method for combining the approaches of Chapters 2 and 4 in order to sample the state and action spaces, respectively.

In Chapter 5, we consider an approximate receding 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, which 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.

Lastly, we would like to acknowledge the financial support of several US Federal funding agencies for this work, specifically the National Science Foundation (under Grants DMI-9988867 and DMI-0323220), the Air Force Office of Scientific Research (under Grants F496200110161 and FA95500410210), and the Department of Defense.

Seoul, Korea,
College Park, Maryland USA,
Stony Brook, New York USA,
September 2006

Hyeong Soo Chang
Michael C. Fu
Jiaqiao Hu
Steven I. Marcus

Contents

Selected Notation and Abbreviations	xv
1 Markov Decision Processes	1
1.1 Optimality Equations	3
1.2 Policy Iteration and Value Iteration	5
1.3 Receding Horizon Control	6
1.4 Survey of Previous Work on Computational Methods	7
1.5 Simulation	9
1.6 Preview of Coming Attractions	12
1.7 Notes	13
2 Multi-stage Adaptive Sampling Algorithms	17
2.1 Upper Confidence Bound Sampling	19
2.1.1 Regret Analysis in Multi-Armed Bandits	19
2.1.2 Algorithm Description	20
2.1.3 Alternative Estimators	21
2.1.4 Convergence Analysis	24
2.1.5 Numerical Example	31
2.2 Pursuit Learning Automata Sampling	40
2.2.1 Algorithm Description	41
2.2.2 Convergence Analysis	42
2.2.3 Application to POMDPs	51
2.3 Notes	51
3 Population-based Evolutionary Approaches	55
3.1 Evolutionary Policy Iteration	56
3.1.1 Policy Switching	56
3.1.2 Policy Mutation and Population Generation	58
3.1.3 Convergence Analysis	59
3.1.4 Parallelization	60
3.2 Evolutionary Random Policy Search	62

3.2.1	Initialization	62
3.2.2	Policy Improvement with Reward Swapping	64
3.2.3	Sub-MDP Generation	66
3.2.4	Stopping Rule	69
3.2.5	Convergence Analysis	69
3.3	Numerical Examples	72
3.3.1	A One-Dimensional Queueing Example	73
3.3.2	A Two-Dimensional Queueing Example	81
3.4	Extension to Simulation-based Setting	83
3.5	Notes	83
4	Model Reference Adaptive Search	85
4.1	The Model Reference Adaptive Search Method	87
4.1.1	The MRAS ₀ Algorithm (Exact Version)	88
4.1.2	The MRAS ₁ Algorithm (Adaptive Monte Carlo Version)	98
4.2	Extension of MRAS to Stochastic Optimization	110
4.2.1	The MRAS ₂ Algorithm (Stochastic Optimization)	111
4.3	Application to MDPs via Direct Policy Learning	126
4.3.1	Finite Horizon MDPs	126
4.3.2	Infinite Horizon MDPs	128
4.3.3	MDPs with Large State Spaces	128
4.3.4	Numerical Examples	129
4.4	Application to Infinite Horizon MDPs in Population-Based Evolutionary Approaches	135
4.4.1	Algorithm Description	137
4.4.2	Numerical Examples	139
4.5	Application to Finite Horizon MDPs using Adaptive Sampling	141
4.6	Notes	143
5	On-line Control Methods via Simulation	145
5.1	Simulated Annealing Multiplicative Weights Algorithm	149
5.1.1	Basic Algorithm Description	150
5.1.2	Convergence Analysis	151
5.1.3	Convergence of the Sampling Version of the Algorithm	154
5.1.4	A Numerical Example	156
5.1.5	Simulated Policy Switching	160
5.2	Rollout	161
5.2.1	Parallel Rollout	162
5.3	Hindsight Optimization	164
5.3.1	Numerical Example	165
5.4	Notes	170
	References	173
	Index	183