Communications and Control Engineering

Series Editors

E.D. Sontag · M. Thoma · A. Isidori · J.H. van Schuppen

Published titles include:

Stability and Stabilization of Infinite Dimensional Systems with Applications
Zheng-Hua Luo, Bao-Zhu Guo and Omer Morgul

Nonsmooth Mechanics (Second edition)
Bernard Brogliato

Nonlinear Control Systems II
Alberto Isidori

L_2-Gain and Passivity Techniques in Nonlinear Control
Arjan van der Schaft

Control of Linear Systems with Regulation and Input Constraints
Ali Saberi, Anton A. Stoorvogel and Peddapullaiah Sannuti

Robust and H_∞ Control
Ben M. Chen

Computer Controlled Systems
Efim N. Rosenwasser and Bernhard P. Lampe

Control of Complex and Uncertain Systems
Stanislav V. Emelyanov and Sergey K. Korovin

Robust Control Design Using H_∞ Methods
Ian R. Petersen, Valery A. Ugrinovski and Andrey V. Savkin

Model Reduction for Control System Design
Goro Obinata and Brian D.O. Anderson

Control Theory for Linear Systems
Harry L. Trentelman, Anton Stoorvogel and Malo Hautus

Functional Adaptive Control
Simon G. Fabri and Visakan Kadirkamanathan

Positive 1D and 2D Systems
Tadeusz Kaczorek

Identification and Control Using Volterra Models
Francis J. Doyle III, Ronald K. Pearson and Bobatunde A. Ogunnaike

Non-linear Control for Underactuated Mechanical Systems
Isabelle Fantoni and Rogelio Lozano

Robust Control (Second edition)
Jürgen Ackermann

Flow Control by Feedback
Ole Morten Aamo and Miroslav Krstić

Learning and Generalization (Second edition)
Mathukumalli Vidyasagar

Constrained Control and Estimation
Graham C. Goodwin, María M. Seron and José A. De Doná

Randomized Algorithms for Analysis and Control of Uncertain Systems
Roberto Tempo, Giuseppe Calafiore and Fabrizio Dabbene

Switched Linear Systems
Zhendong Sun and Shuzhi S. Ge

Subspace Methods for System Identification
Tohru Katayama

Digital Control Systems
Ioan D. Landau and Gianluca Zito

Multivariable Computer-controlled Systems
Efim N. Rosenwasser and Bernhard P. Lampe

Dissipative Systems Analysis and Control (Second edition)
Bernard Brogliato, Rogelio Lozano, Bernhard Maschke and Olav Egeland

Algebraic Methods for Nonlinear Control Systems (Second edition)
Giuseppe Conte, Claude H. Moog and Anna Maria Perdon

Polynomial and Rational Matrices
Tadeusz Kaczorek

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

Simulation-based Algorithms for Markov Decision Processes

Springer

Hyeong Soo Chang
Department of Computer Science
and Engineering
Sogang University
Seoul 121-742
Republic of Korea

Jiaqiao Hu
Department of Applied Mathematics
and Statistics
State University of New York
at Stony Brook
Stony Brook
NY 11794
USA

Michael C. Fu
Smith School of Business
and Institute for Systems Research
University of Maryland
College Park
MD 20742
USA

Steven I. Marcus
Department of Electrical
and Computer Engineering
and Institute for Systems Research
University of Maryland
College Park
MD 20742
USA

British Library Cataloguing in Publication Data
Simulation-based algorithms for Markov decision processes.
- (Communications and control engineering)
1. Decision making - Mathematical models 2. Markov
processes
I. Chang, Hyeong Soo
engineering - Mathematics 3. Matrices 4. Linear systems
658.4'033
ISBN-13: 9781846286896

Library of Congress Control Number: 2007920840

Communications and Control Engineering Series ISSN 0178-5354
ISBN 978-1-84628-689-6 e-ISBN 978-1-84628-690-2 Printed on acid-free paper

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

To my father, You Chan Chang, my mother, Dong Sook Yoon, and two sisters, Gie Hee & Gie Eun – H. S. Chang

To my mother, for continuous support, to Fan, for discrete labors, and to Lara & David, for mixtures of smiles & laughter – M. Fu

To my family – J. Hu

To Jeanne, Jeremy, and Tobin – S. Marcus

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 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 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 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 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.

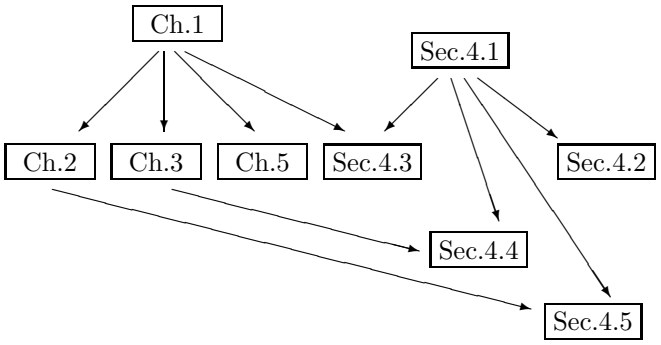
Chapter 3 considers infinite-horizon problems and presents evolutionary approaches for finding an optimal policy. The algorithms in this chapter work with a population of policies — in contrast to the usual policy iteration approach, which updates a single policy — and are targeted at problems with large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

The relationship between the chapters and/or sections of the book is shown below. After reading Chapter 1, Chapters 2, 3, and 5 can pretty much be read independently, although Chapter 5 does allude to algorithms in each of the previous chapters, and the numerical example in Section 5.1 is taken from Section 2.1. The first two sections of Chapter 4 present a general global optimization approach, which is then applied to MDPs in the subsequent Sections 4.3, 4.4 and 4.5, where the latter two build upon work in Chapters 3 and 2, respectively.



We thank Chunyue Song, Yongqiang Wang, Enlu Zhou, Jeff Heath, Scott Nestler, and Huiju Zhang for their comments on various portions and drafts of the manuscript, as well as Oliver Jackson and Sorina Moosdorf for their editorial support. Finally, we acknowledge the financial support of several US Federal funding agencies for this work: 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.

A current list of errata for the book will be kept online at
`http://www.springer.com/978-1-84628-689-6`.

Hyeong Soo Chang, Seoul, Korea
Michael Fu, College Park, Maryland
Jiaqiao Hu, Stony Brook, New York
Steve Marcus, College Park, Maryland
September 2006

Contents

Selected Notation and Abbreviations

Notation specific to a particular chapter is noted parenthetically.
Equation numbers indicate where the quantity is defined.

$\Re$ ($\Re^+$)	set of (non-negative) real numbers
$\mathcal{Z}$ ($\mathcal{Z}^+$)	set of (positive) integers
H	horizon length (number of stages or periods)
X	state space
A	action space
$A(x)$	admissible action space in state x
$P(x,a)(y)$	probability of transitioning to state y from state x when taking action a
$f(x,a,u)$	next state reached from state x when taking action a for random number u
$R(x,a)$	non-negative bounded reward obtained in state x when taking action a
$R'(x,a,w)$	non-negative bounded reward obtained in state x when taking action a for random number w
$R_{\max}$	upper bound on one-period reward
γ	discount factor $\in (0,1]$
π	policy (a sequence of mappings prescribing an action to take for each state)
$\pi_i(x)$	action prescribed for state x in stage i under policy π
$\pi(x)$	action prescribed for state x (under stationary policy π)
π^*	an optimal policy
$\hat{\pi}^k$	an estimated optimal policy at kth iteration
$\mathit{\Pi}$	set of all nonstationary Markovian policies
$\mathit{\Pi}_s$	set of all stationary Markovian policies: (1.10)
$V_i^*(x)$	optimal reward-to-go value from stage i in state x: (1.5)
V_i^*	optimal reward-to-go value function from stage i
$\hat{V}_i^{N_i}$	estimated optimal reward-to-go value function from stage i based on N_i simulation replications in that stage

$V^*(x)$	optimal value for starting state x: (1.2)
V^*	optimal value function
V_i^π	reward-to-go value function for policy π from stage i: (1.6)
V^π	value function for policy π: (1.11)
$\mathcal{V}_H^\pi(x)$	expected total discounted reward over horizon length H under policy π, starting from state x $(= V_0^\pi(x))$
$Q_i^*(x,a)$	Q-function value giving expected reward for taking action a from state x in stage i, plus expected total discounted optimal reward-to-go value from next state reached in stage i+1: (1.9)
$Q^*(x,a)$	infinite-horizon Q-function value: (1.14)
$\hat{Q}_i^{N_i}(x,a)$	estimate for $Q_i^*(x,a)$ based on N_i samples
$\hat{Q}(x,a)$	estimate for $Q^*(x,a)$
$\mathcal{P}_x$	action selection distribution over $A(x)$
a.s.	almost sure(ly)
c.d.f.	cumulative distribution function
i.i.d.	independent and identically distributed
p.d.f.	probability density function
p.m.f.	probability mass function
s.t.	such that (or subject to)
w.p.	with probability
w.r.t.	with respect to
$U(a,b)$	(continuous) uniform distribution with support on $[a,b]$
$DU(a,b)$	discrete uniform distribution on $\{a, a+1, \ldots, b-1, b\}$
$N(\mu, \sigma^2)$	normal (Gaussian) distribution with mean (vector) μ and variance σ^2 (covariance matrix Σ)
E_f	expectation under p.d.f. f (Ch. 4)
E_θ, P_θ	expectation/probability under p.d.f./p.m.f. $f(\cdot,\theta)$ (Ch. 4)
$\tilde{E}_\theta, \tilde{P}_\theta$	expectation/probability under p.d.f./p.m.f. $\tilde{f}(\cdot,\theta)$ (Ch. 4)
$\forall$	for all
$\exists$	there exists
$\mathcal{D}(\cdot,\cdot)$	Kullback–Leibler (KL) divergence between two p.d.f.s/p.m.f.s (Chs. 4, 5)
$d(\cdot,\cdot)$	distance metric (Ch. 3)
$d_\infty(\cdot,\cdot)$	infinity-norm distance between two policies (Ch. 3)
$d_T(\cdot,\cdot)$	total variation distance between two p.m.f.s (Ch. 5)
NEF	natural exponential family (Ch. 4)
$:=$	equal by definition
$\stackrel{d}{=}$	equal in distribution
$\Longleftrightarrow$	if and only if
$\Longrightarrow$	implies (or weak convergence)
$I\{\cdot\}$	indicator function of the set $\{\cdot\}$
$\lvert X\rvert$	cardinality (number of elements) of set X
$\lVert\cdot\rVert$	norm of a function or vector, or induced norm of a matrix

$x \vee y$	$\max(x, y)$
$x \wedge y$	$\min(x, y)$
x^+	$\max(x, 0)$
x^-	$\min(-x, 0)$
$\lceil x \rceil$	least integer greater than or equal to x
$\lfloor x \rfloor$	greatest integer less than or equal to x
$f(n) = O(g(n))$	$\limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty$
$f(n) = \Theta(g(n))$	$f(n) = O(g(n))$ and $g(n) = O(f(n))$