

# **SpringerBriefs in Intelligent Systems**

## **Artificial Intelligence, Multiagent Systems, and Cognitive Robotics**

### **Series editors**

Gerhard Weiss, Maastricht University, Maastricht, The Netherlands

Karl Tuyls, University of Liverpool, Liverpool, UK

### **Editorial Board**

Felix Brandt, Technische Universität München, Munich, Germany

Wolfram Burgard, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Marco Dorigo, Université libre de Bruxelles, Brussels, Belgium

Peter Flach, University of Bristol, Bristol, UK

Brian Gerkey, Open Source Robotics Foundation, Bristol, UK

Nicholas R. Jennings, Southampton University, Southampton, UK

Michael Luck, King's College London, London, UK

Simon Parsons, City University of New York, New York, US

Henri Prade, IRIT, Toulouse, France

Jeffrey S. Rosenschein, Hebrew University of Jerusalem, Jerusalem, Israel

Francesca Rossi, University of Padova, Padua, Italy

Carles Sierra, IIIA-CSIC Cerdanyola, Barcelona, Spain

Milind Tambe, USC, Los Angeles, US

Makoto Yokoo, Kyushu University, Fukuoka, Japan

This series covers the entire research and application spectrum of intelligent systems, including artificial intelligence, multiagent systems, and cognitive robotics. Typical texts for publication in the series include, but are not limited to, state-of-the-art reviews, tutorials, summaries, introductions, surveys, and in-depth case and application studies of established or emerging fields and topics in the realm of computational intelligent systems. Essays exploring philosophical and societal issues raised by intelligent systems are also very welcome.

More information about this series at <http://www.springer.com/series/11845>

Frans A. Oliehoek · Christopher Amato

# A Concise Introduction to Decentralized POMDPs

Frans A. Oliehoek  
School of Electrical Engineering, Electronics  
and Computer Science  
University of Liverpool  
Liverpool  
UK

Christopher Amato  
Computer Science and Artificial Intelligence  
Laboratory  
MIT  
Cambridge, MA  
USA

ISSN 2196-548X ISSN 2196-5498 (electronic)  
SpringerBriefs in Intelligent Systems  
ISBN 978-3-319-28927-4 ISBN 978-3-319-28929-8 (eBook)  
DOI 10.1007/978-3-319-28929-8

Library of Congress Control Number: 2016941071

© The Author(s) 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature  
The registered company is Springer International Publishing AG Switzerland

*Dedicated to future generations of intelligent  
decision makers.*

# Preface

This book presents an overview of formal decision making methods for decentralized cooperative systems. It is aimed at graduate students and researchers in the fields of artificial intelligence and related fields that deal with decision making, such as operations research and control theory. While we have tried to make the book relatively self-contained, we do assume some amount of background knowledge.

In particular, we assume that the reader is familiar with the concept of an *agent* as well as search techniques (like depth-first search, A\*, etc.), both of which are standard in the field of artificial intelligence [Russell and Norvig, 2009]. Additionally, we assume that the reader has a basic background in probability theory. Although we give a very concise background in relevant single-agent models (i.e., the ‘MDP’ and ‘POMDP’ frameworks), a more thorough understanding of those frameworks would benefit the reader. A good first introduction to these concepts can be found in the textbook by Russell and Norvig, with additional details in texts by Sutton and Barto [1998], Kaelbling et al. [1998], Spaan [2012] and Kochenderfer et al. [2015]. We also assume that the reader has a basic background in game theory and game-theoretic notations like Nash equilibrium and Pareto efficiency. Even though these concepts are not central to our exposition, we do place the Dec-POMDP model in the more general context they offer. For an explanation of these concepts, the reader could refer to any introduction on game theory, such as those by Binmore [1992], Osborne and Rubinstein [1994] and Leyton-Brown and Shoham [2008].

This book heavily builds upon earlier texts by the authors. In particular, many parts were based on the authors’ previous theses, book chapters and survey articles [Oliehoek, 2010, 2012, Amato, 2010, 2015, Amato et al., 2013]. This also means that, even though we have tried to give a relatively complete overview of the work in the field, the text in some cases is biased towards examples and methods that have been considered by the authors. For the description of further topics in Chapter 8, we have selected those that we consider important and promising for future work. Clearly, there is a necessarily large overlap between these topics and the authors’ recent work in the field.

# Acknowledgments

Writing a book is not a standalone activity; it builds upon all the insights developed in the interactions with peers, reviewers and coauthors. As such, we are grateful for the interaction we have had with the entire research field. We specifically want to thank the attendees and organizers of the workshops on *multiagent sequential decision making (MSDM)* which have provided a unique platform for exchange of thoughts on decision making under uncertainty.

Furthermore, we would like to thank João Messias, Matthijs Spaan, Shimon Whiteson, and Stefan Witwicki for their feedback on sections of the manuscript.

Finally, we are grateful to our former supervisors, in particular Nikos Vlassis and Shlomo Zilberstein, who enabled and stimulated us to go down the path of research on decentralized decision making.

# Contents

<b>1</b>	<b>Multiagent Systems Under Uncertainty</b>	<b>1</b>
1.1	Motivating Examples	2
1.2	Multiagent Systems	4
1.3	Uncertainty	6
1.4	Applications	7
<b>2</b>	<b>The Decentralized POMDP Framework</b>	<b>11</b>
2.1	Single-Agent Decision Frameworks	11
2.1.1	MDPs	12
2.1.2	POMDPs	13
2.2	Multiagent Decision Making: Decentralized POMDPs	14
2.3	Example Domains	17
2.3.1	Dec-Tiger	17
2.3.2	Multirobot Coordination: Recycling and Box-Pushing	19
2.3.3	Network Protocol Optimization	20
2.3.4	Efficient Sensor Networks	20
2.4	Special Cases, Generalizations and Related Models	21
2.4.1	Observability and Dec-MDPs	21
2.4.2	Factored Models	22
2.4.3	Centralized Models: MMDPs and MPOMDPs	24
2.4.4	Multiagent Decision Problems	25
2.4.5	Partially Observable Stochastic Games	30
2.4.6	Interactive POMDPs	30
<b>3</b>	<b>Finite-Horizon Dec-POMDPs</b>	<b>33</b>
3.1	Optimality Criteria	33
3.2	Policy Representations: Histories and Policies	34
3.2.1	Histories	34
3.2.2	Policies	35
3.3	Multiagent Beliefs	37
3.4	Value Functions for Joint Policies	37



3.5	Complexity	39
<b>4</b>	<b>Exact Finite-Horizon Planning Methods</b>	<b>41</b>
4.1	Backwards Approach: Dynamic Programming	41
4.1.1	Growing Policies from Subtree Policies	41
4.1.2	Dynamic Programming for Dec-POMDPs	44
4.2	Forward Approach: Heuristic Search	46
4.2.1	Temporal Structure in Policies: Decision Rules	46
4.2.2	Multiagent A*	47
4.3	Converting to a Non-observable MDP	48
4.3.1	The Plan-Time MDP and Optimal Value Function	49
4.3.2	Plan-Time Sufficient Statistics	49
4.3.3	An NOMDP Formulation	51
4.4	Other Finite-Horizon Methods	52
4.4.1	Point-Based DP	52
4.4.2	Optimization	52
<b>5</b>	<b>Approximate and Heuristic Finite-Horizon Planning Methods</b>	<b>55</b>
5.1	Approximation Methods	56
5.1.1	Bounded Dynamic Programming	56
5.1.2	Early Stopping of Heuristic Search	57
5.1.3	Application of POMDP Approximation Algorithms	57
5.2	Heuristic Methods	58
5.2.1	Alternating Maximization	58
5.2.2	Memory-Bounded Dynamic Programming	59
5.2.3	Approximate Heuristic-Search Methods	61
5.2.4	Evolutionary Methods and Cross-Entropy Optimization	64
<b>6</b>	<b>Infinite-Horizon Dec-POMDPs</b>	<b>69</b>
6.1	Optimality Criteria	69
6.1.1	Discounted Cumulative Reward	69
6.1.2	Average Reward	70
6.2	Policy Representation	70
6.2.1	Finite-State Controllers: Moore and Mealy	71
6.2.2	An Example Solution for DEC-TIGER	73
6.2.3	Randomization	74
6.2.4	Correlation Devices	74
6.3	Value Functions for Joint Policies	75
6.4	Undecidability, Alternative Goals and Their Complexity	76
<b>7</b>	<b>Infinite-Horizon Planning Methods: Discounted Cumulative Reward</b>	<b>79</b>
7.1	Policy Iteration	79
7.2	Optimizing Fixed-Size Controllers	81
7.2.1	Best-First Search	82
7.2.2	Bounded Policy Iteration	83
7.2.3	Nonlinear Programming	85

7.2.4	Expectation Maximization . . . . .	85
7.2.5	Reduction to an NOMDP . . . . .	87
<b>8</b>	<b>Further Topics . . . . .</b>	<b>91</b>
8.1	Exploiting Structure in Factored Models . . . . .	91
8.1.1	Exploiting Constraint Optimization Methods . . . . .	91
8.1.1.1	Coordination (Hyper-)Graphs . . . . .	91
8.1.1.2	ND-POMDPs . . . . .	93
8.1.1.3	Factored Dec-POMDPs . . . . .	95
8.1.2	Exploiting Influence-Based Policy Abstraction . . . . .	100
8.2	Hierarchical Approaches and Macro-Actions . . . . .	102
8.3	Communication . . . . .	104
8.3.1	Implicit Communication and Explicit Communication . . . . .	105
8.3.1.1	Explicit Communication Frameworks . . . . .	106
8.3.1.2	Updating of Information States and Semantics . . . . .	107
8.3.2	Delayed Communication . . . . .	108
8.3.2.1	One-Step Delayed Communication . . . . .	108
8.3.2.2	$k$ -Steps Delayed Communication . . . . .	109
8.3.3	Communication with Costs . . . . .	111
8.3.4	Local Communication . . . . .	112
8.4	Reinforcement Learning . . . . .	113
<b>9</b>	<b>Conclusion . . . . .</b>	<b>115</b>
	<b>References . . . . .</b>	<b>117</b>

# Acronyms

AH	action history
AOH	action-observation history
BG	Bayesian game
CG	coordination graph
CBG	collaborative Bayesian game
COP	constraint optimization problem
DAG	directed acyclic graph
DP	dynamic programming
Dec-MDP	decentralized Markov decision process
Dec-POMDP	decentralized partially observable Markov decision process
DICE	direct cross-entropy optimization
EM	expectation maximization
EXP	deterministic exponential time (complexity class)
FSC	finite-state controller
FSPC	forward-sweep policy computation
GMAA*	generalized multiagent A*
I-POMDP	interactive partially observable Markov decision process
MAS	multiagent system
MARL	multiagent reinforcement learning
MBDP	memory-bounded dynamic programming
MDP	Markov decision process
MILP	mixed integer linear program
NEXP	non-deterministic exponential time (complexity class)
ND-POMDP	networked distributed POMDP
NDP	nonserial dynamic programming
NLP	nonlinear programming
NP	non-deterministic polynomial time (complexity class)
OH	observation history
POMDP	partially observable Markov decision process
PSPACE	polynomial SPACE (complexity class)
PWLC	piecewise linear and convex

RL reinforcement learning

TD-POMDP transition-decoupled POMDP

# List of Symbols

Throughout this text, we tried to make consistent use of typesetting to convey the meaning of used symbols and formulas. In particular, we use blackboard bold fonts ( $\mathbb{A}, \mathbb{B}$ , etc.) to denote sets, and subscripts to denote agents (typically  $i$  or  $j$ ) or groups of agents, as well as time ( $t$  or  $\tau$ ).

For instance  $a$  is the letter used to indicate actions in general,  $a_i$  denotes an action of agent  $i$ , and the set of its actions is denoted  $\mathbb{A}_i$ . The action agent  $i$  takes at a particular time step  $t$  is denoted  $a_{i,t}$ . The profile of actions taken by all agents, a joint action, is denoted  $a$ , and the set of such joint actions is denoted  $\mathbb{A}$ . When referring to the action profile of a subset  $e$  of agents we write  $a_e$ , and for the actions of all agents except agent  $i$ , we write  $a_{-i}$ . On some occasions we will need to indicate the index within a set, for instance the  $k$ -th action of agent  $i$  is written  $a_i^k$ . In the list of symbols below, we have shown all possible uses of notation related to actions (base symbol ‘a’), but have not exhaustively applied such modifiers to all symbols.

$\cdot$	multiplication,
$\times$	Cartesian product,
$\circ$	policy concatenation,
$\Downarrow$	subtree policy consumption operator,
$\triangle(\cdot)$	simplex over $(\cdot)$ ,
$\mathbf{1}_{\{\cdot\}}$	indicator function,
$\beta$	macro-action termination condition,
$\Gamma_j$	mapping from histories to subtree policies,
$\bar{\Gamma}^{\mathcal{H}}$	state factor scope backup operator,
$\bar{\Gamma}^{\mathcal{A}}$	agent scope backup operator,
$\gamma$	discount factor,
$\delta_t$	decision rule for stage $t$ ,
$\hat{\delta}_t$	joint decision rule for stage $t$ ,
$\tilde{\delta}_t$	approximate joint decision rule,
$\delta_{i,t}$	decision rule for agent $i$ for stage $t$ ,
$\Delta t$	length of a stage $ts$ ,

$\varepsilon$	(small) constant,
$\bar{\theta}$	joint action-observation history,
$\bar{\theta}_i$	action-observation history,
$\bar{\Theta}_i$	action-observation history set,
$l_i$	information state, or belief update, function,
$\mu_i$	macro-action policy for agent $i$ ,
$\pi$	joint policy,
$\pi_i$	policy for agent $i$ ,
$\pi_{-i}$	(joint) policy for all agents but $i$ ,
$\pi^*$	optimal joint policy,
$\rho$	number of reward functions,
$\Sigma$	alphabet of communication messages,
$\sigma_t$	plan-time sufficient statistic,
$\tau$	stages-to-go ( $\tau = h - 1$ ),
$v$	domination gap,
$\Phi_{\text{Next}}$	set of next policies,
$\varphi_t$	past joint policy,
$\xi$	parameter vector,
$\psi$	correlation device transition function,
$\mathbb{A}$	set of joint actions,
$\mathbb{A}_i$	set of actions for agent $i$ ,
$\bar{\mathbb{A}}$	joint action history set,
$\bar{\mathbb{A}}_i$	action history set for agent $i$ ,
$a$	joint action,
$a_t$	joint action at stage $t$ ,
$a_i$	action for agent $i$ ,
$a_e$	joint action for subset $e$ of agents,
$a_{-i}$	joint action for all agents except $i$ ,
$\bar{a}_i$	action history of agent $i$ ,
$\bar{a}_{i,t}$	action history of agent $i$ at stage $t$ ,
$\bar{a}$	joint action history,
$\bar{a}_t$	joint action history at stage $t$ ,
$B(b_0, \varphi_t)$	Bayesian game for a stage,
$B(\mathcal{M}_{\text{DecP}}, b_0, \varphi_t)$	CBG for stage $t$ of a Dec-POMDP,
$\mathbb{B}$	set of joint beliefs,
$b_0$	initial state distribution,
$b_i(s_t, q_{-i}^{\tau})$	multiagent belief,
$b$	belief,
$b_i$	belief for agent $i$ (e.g., a multiagent belief),
$\mathbb{C}$	states of a correlation device,
$C_{\Sigma}$	message cost function,
$c$	correlation device state,
$\mathbb{D}$	set of agents,
$\mathbf{E}[\cdot]$	expectation of $\cdot$ ,

$\mathcal{E}$	set of hyper-edges,
$e$	hyper edge, or index of local payoff function (corresponding to a hyper edge),
$f_{\xi}$	probability distribution, parameterized by $\xi$ ,
$f_{\xi}^{(j)}$	distribution over joint policies at iteration $j$ ,
$h$	horizon,
$I_{i \rightarrow j}$	influence of agent $i$ on agent $j$ ,
$I_i$	information states for agent $i$ ,
$\mathbb{I}_i$	set of information states for agent $i$ ,
$\mathcal{M}$	Markov multiagent environment,
$\mathcal{M}_{\text{DecP}}$	Dec-POMDP,
$\mathcal{M}_{\text{MPOMDP}}$	MPOMDP,
$\mathcal{M}_{\text{PT}}$	plan-time NOMDP,
$m_i$	agent model, also finite-state controller,
$m$	agent component (a joint model),
$\mathbf{m}_i$	macro-action for agent $i$ ,
$N_b$	number of best samples,
$N_f$	number of fire levels,
Next	operation constructing next set of partial policies,
NULL	null observation,
$n$	number of agents,
$O$	observation function,
$O_i$	local observation function for agent $i$ ,
$OC$	optimality criterion,
$\mathbb{O}$	set of joint observations,
$\mathbb{O}_i$	set of observations for agent $i$ ,
$\bar{\mathbb{O}}$	joint observation history set,
$\bar{\mathbb{O}}_i$	observation history set for agent $i$ ,
$o$	joint observation,
$o_i$	observation for agent $i$ ,
$o_{i,\emptyset}$	NULL observation for agent $i$ ,
$\bar{o}_i$	observation history of agent $i$ ,
$\bar{o}_t$	joint observation history at stage $t$ ,
$\bar{o}_{t, k }$	joint observation history at stage $t$ of length $k$ ,
$Q^{\pi}$	Q-value function for $\pi$ ,
$Q_i^{\tau}$	set of $q_i^{\tau}$ ,
$Q^{\tau}$	set of $q^{\tau}$ ,
$Q_{e,i}^{\tau+1}$	set of subtree policies resulting from exhaustive backup,
$Q_{m,i}^{\tau+1}$	set of maintained subtree policies,
$q_{t-k}^k$	joint subtree policy of length $k$ to be executed at stage $t - k$ ,
$q_i^{\tau}$	$\tau$ -stage-to-go subtree policy for agent $i$ ,
$q^{\tau}$	$\tau$ -stage-to-go joint subtree policy,
$R$	reward function,
$R_i$	local reward function for agent $i$ ,
$R^e$	local reward function (with index $e$ ),
$\mathbb{R}$	real numbers,

$\mathbb{S}$	state space (the set of all states),
$\mathbb{S}_i$	set of local state for agent $i$ ,
$\tilde{\mathbb{S}}_i$	set of interactive states for agent $i$ ,
$s$	state,
$s_e$	local state of agents participating in $e$ (in agent-wise factored model),
$s_i$	local state for agent $i$ ,
$T$	transition function,
$T_i$	local state transition function for agent $i$ ,
$t$	stage,
$U_{ss}()$	sufficient statistic update,
$u$	payoff function (in context of single-shot game),
$u^e$	local payoff function,
$V$	value function,
$V^*$	optimal value function,
$V^e$	value function for a particular payoff component $e$ ,
$V^\pi$	value function for joint policy $\pi$ ,
$v$	value vector,
$v_\delta$	value vector associated with (meta-level ‘action’) $\delta$ ,
$\mathcal{V}$	set of value vectors,
$\mathcal{X}$	space of candidate solutions (DICE),
$\mathbf{X}$	set of samples (DICE),
$\mathbf{X}_b$	set of best samples (DICE),
$\mathbb{X}$	set of state factors
$\mathbb{X}^j$	set of values for $j$ -th state factor,
$x^j$	value for $j$ -th state factor,
$x^H$	fire level of house $H$ ,
$x_e$	profile of values for state factors in component $e$ ,
$x$	candidate solution (DICE),
$\mathbb{Z}_i$	set of auxiliary observations for agent $i$ ,