Markov Decision Processes in Artificial Intelligence

MDPs, Beyond MDPs and Applications

Edited by
Olivier Sigaud
Olivier Buffet

iSTE

WILEY

Markov Decision Processes in Artificial Intelligence

Markov Decision Processes in Artificial Intelligence

MDPs, Beyond MDPs and Applications

Edited by
Olivier Sigaud
Olivier Buffet

ISTE

WILEY

First published 2008 in France by Hermes Science/Lavoisier in two volumes entitled: *Processus décisionnels de Markov en intelligence artificielle* © LAVOISIER 2008
First published 2010 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Cataloging-in-Publication Data

Markov decision processes in artificial intelligence : MDPs, beyond MDPs and applications / edited by Olivier Sigaud, Olivier Buffet.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-84821-167-4
1. Artificial intelligence--Mathematics. 2. Artificial intelligence--Statistical methods. 3. Markov processes. 4. Statistical decision. I. Sigaud, Olivier. II. Buffet, Olivier.
Q335.M374 2010
006.301'509233--dc22

2009048651

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-84821-167-4

Printed and bound in Great Britain by CPI Antony Rowe, Chippenham and Eastbourne

Table of Contents

Preface

The present book discusses sequential decision-making under uncertainty and reinforcement learning, two classes of problems in artificial intelligence which can be formalized in the framework of Markov decision processes. It has been written for students, engineers and researchers likely to be interested in these fields and models.

The book is organized as follows:

– Part 1 provides an introduction to this domain and to efficient resolution techniques (Markov decision processes, reinforcement learning, approximate representations, factored representations, policy gradients and online resolution).

– Part 2 presents important extensions of Markov decision processes that make it possible to solve more complex sequential decision-making problems (partially observable Markov decision processes, Markov games, multi-agent approaches and non-classical criteria).

– Part 3 completes the book with example applications showing how Markov decision processes can be employed for various problems (micro-object manipulation, biodiversity preservation, high-level control of a helicopter, control of an exploration mission and operations planning).

It was not possible for this book to cover all research directions in this very active field. We give here some references to point the reader to some uncovered aspects. For example, we have decided not to cover continuous time reinforcement learning [MUN 01], relational reinforcement learning [DZE 01], hierarchical reinforcement learning [BAR 03], learning classifier systems [SIG 07] or predictive state representations [LIT 02].

In addition, we endeavor in each chapter to provide the reader with references to related work.

Additional information related to this book (e.g. *errata*) can be found at the following website: http://www.loria.fr/projets/PDMIA/Book/.

Bibliography

[BAR 03] BARTO A. and MAHADEVAN S., "Recent advances in hierarchical reinforcement learning", *Discrete Event Dynamic Systems*, vol. 13, no. 4, pp. 341–379, 2003.

[DZE 01] DZEROSKI S., DE RAEDT L. and DRIESSENS K., "Relational reinforcement learning", *Machine Learning*, vol. 43, no. 1-2, pp. 7–53, 2001.

[LIT 02] LITTMAN M., SUTTON R. and SINGH S., "Predictive representations of state", *Advances in Neural Information Processing Systems 14 (NIPS'01)*, MIT Press, Cambridge, MA, pp. 1555–1561, 2002.

[MUN 01] MUNOS R. and MOORE A., "Variable resolution discretization in optimal control", *Machine Learning*, vol. 49, pp. 291–323, 2001.

[SIG 07] SIGAUD O. and WILSON S. W., "Learning classifier systems: a survey", *Soft Computing*, vol. 11, no. 11, pp. 1065–1078, 2007.

List of Authors

Aurélie Beynier

Aurélie Beynier is currently an associate professor of computer science at the LIP6 Laboratory. Her PhD thesis has been defended in November 2006 at the University of Caen.

Matthieu Boussard

Matthieu Boussard holds a PhD in computer science from the University of Caen. He is a member of the MAD Group of the GREYC Laboratory.

Maroua Bouzid

Maroua Bouzid is an associate professor at the University of Caen. She is a member of the MAD Group of the GREYC Laboratory.

Olivier Buffet

Olivier Buffet is an INRIA junior research scientist. He is member of the Autonomous Intelligent Machines (MAIA) Team of the LORIA Laboratory.

Andriy Burkov

Andriy Burkov is a PhD student working under supervision of Professor Brahim Chaib-Draa at Laval University, Canada. His main research interests include multiagent learning and game theory.

Iadine Chadès

Iadine Chadès is currently a research fellow at CSIRO Sustainable Ecosystems, Australia, on leave from the French National Institute for Agricultural Research (INRA).

Brahim Chaib-Draa

Professor Brahim Chaib-Draa is the leader of the DAMAS Research Group on Agents and Multiagent Systems at the Computer Science and Software Engineering Department of Laval University, Canada.

François Charpillet

François Charpillet is an INRIA senior research scientist. He is the head of the Autonomous Intelligent Machines (MAIA) Team of the LORIA Laboratory.

Thomas Degris

Thomas Degris holds a PhD in computer science. After working in the video game industry, he is now a postdoctoral fellow at the University of Alberta.

Alain Dutech

Alain Dutech is an INRIA experienced research scientist. He is a member of the Autonomous Intelligent Machines (MAIA) Team of the LORIA Laboratory.

Patrick Fabiani

Patrick Fabiani is the director of the Systems Control and Flight Dynamics Department at ONERA. His research interests include models, methods and algorithms for sequential decision making and planning under uncertainty, applied to autonomous aerial robots in the ReSSAC project at ONERA.

Frédérick Garcia

Frédérick Garcia is a researcher in artificial intelligence at the Department of Applied Mathematics and Informatics at INRA (the French National Institute for Agricultural Research).

Guillaume Laurent

Guillaume Laurent is an associate professor at the ENSMM Graduate School at Besançon. He is a member of the Automatic Control and Micro-Mechatronic Systems Department of the FEMTO-ST Research Institute.

Simon Le Gloannec

Simon Le Gloannec holds a PhD from the University of Caen. He has worked with both the MAIA Team (LORIA laboratory) and the MAD Group (GREYC Laboratory). He currently holds a postdoctoral position at GREYC.

Laëtitia Matignon

Laëtitia Matignon holds a PhD from the University of Franche-Comté at Besançon. She is now a postdoctoral fellow of the Cooperative Decision-Theoretic Autonomous Agent System Group (MAD) of the GREYC Laboratory in Caen.

Abdel-Illah Mouaddib

Abdel-Illah Mouaddib is a full professor at the University of Caen. He is the head of the Cooperative Decision-Theoretic Autonomous Agent System Group (MAD) of the GREYC Laboratory.

Rémi Munos

Rémi Munos is senior researcher at INRIA Lille, France. He works in the fields of reinforcement learning, optimal control and decision theory.

Laurent Péret

Laurent Péret holds a PhD in artificial intelligence, focused on search methods for MDPs. Since 2005, he has been involved in various space programs as a flight dynamics engineer.

Emmanuel Rachelson

Emmanuel Rachelson is a postdoctoral fellow at the University of Liège, Belgium, working on reinforcement and statistical learning. He also works with Pr. Lagoudakis in Greece and with the Electricité de France (EDF) Research Center in Paris.

Régis Sabbadin

Régis Sabbadin has been a research scientist at INRA-Toulouse since 1999. His research focuses on planning under uncertainty, applied to natural resources management and agriculture.

Bruno Scherrer

Bruno Scherrer is an INRIA experienced research scientist. He is member of the Autonomous Intelligent Machines (MAIA) Team of the LORIA Laboratory.

Olivier Sigaud

Olivier Sigaud is a professor of computer science at the UPMC-Paris 6 University. He is the head of the “Motion” Group at the Institute of Intelligent Systems and Robotics (ISIR).

Daniel Szer

Daniel Szer obtained a Master's degree in computer science from UMass Dartmouth and a PhD in artificial intelligence from Henri-Poincaré University, Nancy. He works as an IT analyst in Paris.

Florent Teichteil-Königsbuch

Florent Teichteil-Königsbuch is a researcher in probabilistic sequential decision-making at ONERA. He got a PhD in artificial intelligence from SUPAERO in 2005.

Sylvie Thiébaux

Sylvie Thiébaux is an associate professor at the Australian National University and principal researcher at the National ICT Australia, specializing in automated planning in artificial intelligence.

Paul Weng

Paul Weng has been an associate professor at Paris 6 University since September 2007. Before his PhD obtained in 2006, he worked in finance in London.

Part 1

MDPs: Models and Methods

Chapter 1

Markov Decision Processes

1.1. Introduction

This book presents a decision problem type commonly called *sequential decision problems under uncertainty*. The first feature of such problems resides in the relation between the current decision and future decisions. Indeed, these problems do not consist of one, but several decision problems, presented in a sequence. At each step of this sequence, the *agent* (actor or decision-maker) needs to decide on the current action by taking into account its effect on the solution of future problems. This sequential feature is also typical of *planning* problems in artificial intelligence and is often linked with shortest path methods in graph theory. The second characteristic of the problems discussed in these pages is the uncertainty in the consequences of all available decisions (actions). Knowledge of its decision's effects is not available in advance to the agent in a deterministic form. As such, this problem deals with the various theories of decision under uncertainty which suggest different formalization and resolution approaches. Among these approaches, we need to mention specifically the standard theory of expected utility maximization.

Consequently, problems of sequential decision under uncertainty couple the two problematics of sequential decision and decision under uncertainty. *Markov decision problems* (MDPs) are a general mathematical formalism for representing shortest path problems in stochastic environments. This formalism is based on the theory of *Markov decision processes* (also written as MDPs). A Markov decision process relies on the notions of *state*, describing the current situation of the agent, *action* (or decision), affecting the dynamics of the process, and *reward*, observed for each

Chapter written by Frédérick GARCIA and Emmanuel RACHELSON.

transition between states. Such a process describes the probability of triggering a transition to state s' and receiving a certain reward r when taking decision a in state s. Hence, an MDP can be described as a controlled Markov chain, where the control is given at each step by the chosen action. The process then visits a sequence of states and can be evaluated through the observed rewards. Solving an MDP consists of controlling the agent in order to reach an optimal behavior, i.e. to maximize its overall revenue. Because action effects are stochastic and, thus, can result in different possible states at the next stage of the decision process, the optimal control strategy cannot necessarily be represented as a single sequence of actions.[1] Consequently, solutions of an MDP are usually given as *universal plans* or *policies* (*strategies* or *decision rules*) specifying which action to undertake at each step of the decision process and for every possible state reached by the agent. Due to the uncertainty in actions' results, applying a given policy can result in different sequences of states/actions.

EXAMPLE 1.1. Let us illustrate these concepts with a simple car maintenance example. According to the current state of the car (breakdown, wear, age, etc.), an agent wishes to decide which is its best strategy (do nothing, replace parts preventively, repair, change car, etc.) in order to minimize the maintenance cost in the long run. Assuming the agent knows the consequences and the cost of each separate action in every possible state (e.g. we know the failure probability of an engine if the oil leak is not fixed), we can model this problem as an MDP. Solving this MDP will provide the agent with a policy indicating which is the optimal action to undertake in every state of the problem. This way, the sequence of actions performed as the car's state changes will allow the agent to always minimize the expected maintenance cost.

The theory of Markov decision processes and its generalizations will be developed in the next chapters. These models have become the most popular framework for representing and solving problems of sequential decision under uncertainty. This chapter presents the basics of MDP theory and optimization, in the case of an agent having a perfect knowledge of the decision process and of its state at every time step, when the agent's goal is to maximize its global revenue over time.

1.2. Markov decision problems

1.2.1. *Markov decision processes*

Markov decision processes are defined as *controlled stochastic processes* satisfying the Markov property and assigning reward values to state transitions [BER 87, PUT 94]. Formally, they are described by the 5-tuple (S, A, T, p, r) where:

1. Contrary to the deterministic approaches of classical planning.

– S is the state space in which the process' evolution takes place;

– A is the set of all possible actions which control the state dynamics;

– T is the set of time steps where decisions need to be made;

– $p()$ denotes the state transition probability function;

– $r()$ provides the reward function defined on state transitions.

Figure 1.1 represents an MDP, drawn as an influence diagram. At every time step t in T, action a_t is applied in the current state s_t, affecting the process in its transition to the next state s_{t+1}. Reward r_t is then obtained for this transition.

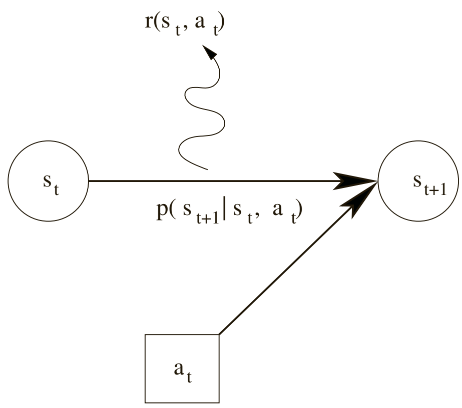

Figure 1.1. *Markov decision process*

The set T of decision epochs is a discrete set, subset of $\mathbb{N}$, which can either be finite or infinite (then we talk, respectively, about finite horizon or infinite horizon). A third case corresponds to the existence of a set of terminal states (or goal states). In this case, the process stops as soon as one of these states is encountered. Then, the horizon is then said to be indefinite. These problems are often related to stochastic shortest path problems. This case, however, can be seen as a specific case of infinite horizon MDPs with absorbing states and will not be presented in detail in this chapter (see Chapter 6, section 6.2.3 and Chapter 15).

In the most general case, the S and A sets are supposed finite, even though many results can be extended to countable or even continuous sets (see [BER 95] for an introduction to the continuous case). Generally, the set A of applicable actions can also depend on the current state: we define a subset A_s of applicable actions in state s. Similarly, S and A can change based on the time step t (S_t and A_t). However, in this chapter, for clarity of presentation, we will restrict ourselves to the standard case where A and S are constant throughout the process.

The transition probabilities $p()$ characterize the state dynamics of the system, i.e. indicate which states are likely to appear after the current state. For a given action a, $p(s' \mid s, a)$ represents the probability for the system to transit to state s' after undertaking action a in state s. Because the $p()$ values are probabilities, we classically have $\forall s, a$, $\sum_{s'} p(s' \mid s, a) = 1$. This $p()$ function is usually represented in matrix form, where we write P_a the $|S| \times |S|$ matrix containing elements $\forall s, s'$ $P_{a,s,s'} = p(s' \mid s, a)$. Consequently, the probability transitions of the decision process are given as $|A|$ matrices P_a. Since each line of these matrices sums to one, the P_a are said to be *stochastic* matrices.

The $p()$ probability distributions over the next state s' follow the fundamental property which gives their name to Markov decision processes. If we write $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$ the history of states and actions until time step t, then the probability of reaching state s_{t+1} consecutively to action a_t is only a function of a_t and s_t, and not of the entire history h_t. Let us write $P(x \mid y)$ the conditional probability of event x, provided that y is true, then we have

$$\forall h_t, a_t, s_{t+1} \quad P\big(s_{t+1} \mid h_t, a_t\big) = P\big(s_{t+1} \mid s_t, a_t\big) = p\big(s_{t+1} \mid s_t, a_t\big).$$

We should note here that the previous condition does not necessarily imply that the resulting stochastic process $(s_t)_{t \in T}$ itself respects the Markov property: this also depends on the action choice policy for a_t.

As a result of choosing action a, in state s, at time t, the deciding agent receives a reward $r_t = r(s, a) \in \mathbb{R}$. We can consider positive values of r_t as gains and negative values as costs. We also sometimes use a cost function $c()$ instead of the reward function $r()$. This reward can be received instantaneously at time t or accumulated between t and $t + 1$. The important feature is that this reward only depends on the simple input of the current state s and the current action a. A vector representation of the $r(s, a)$ reward function consists of $|A|$ vectors r_a of length $|S|$.

A common extension consists of considering random rewards. In this case, we will use their average value for the reward function $r(s, a) = \bar{r}(s, a)$. In particular, the reward obtained at time step t can depend on the final state s' of the transition. We then have a reward specified as $r(s, a, s')$. The value used for the reward vectors is $\bar{r}(s, a) = \sum_{s'} p(s' \mid s, a) r(s, a, s')$. In all cases, r_t is supposed bounded.

Finally, as for S and A, the transition and reward functions can vary across time. In this case they are written, respectively, as p_t and r_t. When these functions do not change from one step to the other, the process is said to be *stationary*: $\forall t \in T$, $p_t() = p()$, $r_t() = r()$. In the rest of this chapter, we will keep this stationarity hypothesis in the study of infinite horizon MDPs.