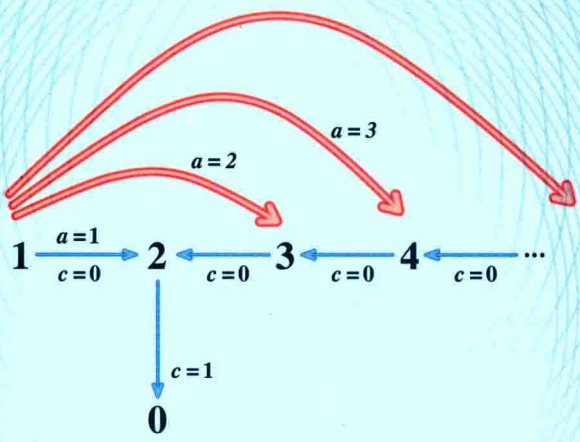
Imperial College Press Optimization Series Vol. 2



Examples in Markov Decision Processes

A. B. Piunovskiy

Imperial College Press

Imperial College Press Optimization Series Vol. 2

Examples in Markov Decision Processes

A. B. Piunovskiy

The University of Liverpool, UK

Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Imperial College Press Optimization Series — Vol. 2
EXAMPLES IN MARKOV DECISION PROCESSES

ISBN 978-1-84816-793-3

Printed in Singapore by Mainland Press Pte Ltd.

Examples in Markov Decision Processes

Imperial College Press Optimization Series

ISSN 2041-1677

Series Editor: Jean Bernard Lasserre *(LAAS-CNRS and Institute of Mathematics, University of Toulouse, France)*

Vol. 1: Moments, Positive Polynomials and Their Applications
by Jean Bernard Lasserre

Vol. 2: Examples in Markov Decision Processes
by A. B. Piunovskiy

Preface

Markov Decision Processes (MDP) is a branch of mathematics based on probability theory, optimal control, and mathematical analysis. Several books with counterexamples/paradoxes in probability [Stoyanov(1997); Szekely(1986)] and in analysis [Gelbaum and Olmsted(1964)] are in existence; it is therefore not surprising that MDP is also replete with unexpected counter-intuitive examples. The main goal of the current book is to collect together such examples. Most of them are based on earlier publications; the remainder are new. This book should be considered as a complement to scientific monographs on MDP [Altman(1999); Bertsekas and Shreve(1978); Hernandez-Lerma and Lasserre(1996a); Hernandez-Lerma and Lasserre(1999); Piunovskiy(1997); Puterman(1994)]. It can also serve as a reference book to which one can turn for answers to curiosities that arise while studying or teaching MDP. All the examples are self-contained and can be read independently of each other. Concerning uncontrolled Markov chains, we mention the illuminating collection of examples in [Suhov and Kelbert(2008)].

A survey of meaningful applications is beyond the scope of the current book. The examples presented either lead to counter-intuitive solutions, or illustrate the importance of conditions in the known theorems. Not all examples are equally simple or complicated. Several examples are aimed at undergraduate students, whilst others will be of interest to professional researchers.

The book has four chapters in line with the four main different types of MDP: the finite-horizon case, infinite horizon with total or discounted loss, and average loss over an infinite time interval. Some basic theoretical statements and proofs of auxiliary assertions are included in the Appendix.

The following notations and conventions will often be used without explanation.

$\stackrel{\triangle}{=}$ means 'equals by definition';
$\mathbf{C}^\infty$ is the space of infinitely differentiable functions;
$\mathbf{C}(\mathbf{X})$ is the space of continuous bounded functions on a (topological) space $\mathbf{X}$;
$\mathbf{B}(\mathbf{X})$ is the space of bounded measurable functions on a (Borel) space $\mathbf{X}$; in discrete (finite or countable) spaces, the discrete topology is usually supposed to be fixed;
$\mathbf{P}(\mathbf{X})$ is the space of probability measures on the (metrizable) space $\mathbf{X}$, equipped with the weak topology;
If Γ is a subset of space $\mathbf{X}$ then Γ^c is the complement;
$\mathbb{N} = \{1, 2, \ldots\}$ is the set of natural numbers; $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$;
$\mathbb{R}^N$ is the N-dimensional Euclidean space; $\mathbb{R} = \mathbb{R}^1$ is the straight line;
$\mathbb{R}^* = [-\infty, +\infty]$ is the extended straight line;
$\mathbb{R}^+ = \{y > 0\}$ is the set of strictly positive real numbers;
$I\{statement\} = \begin{cases} 1, & \text{if the statement is correct;} \\ 0, & \text{if the statement is false;} \end{cases}$ is the indicator function;
$\delta_a(dy)$ is the Dirac measure concentrated at point a: $\delta_a(\Gamma) = I\{\Gamma \ni a\}$;
If $r \in \mathbb{R}^*$ then $r^+ \stackrel{\triangle}{=} \max\{0, r\}$, $r^- \stackrel{\triangle}{=} \min\{0, r\}$;
$\sum_{i=n}^{m} f_i \stackrel{\triangle}{=} 0$ and $\prod_{i=n}^{m} f_i \stackrel{\triangle}{=} 1$ if $m < n$;
$\lfloor r \rfloor$ is the integer part, the maximal integer i such that $i \le r$.

Throughout the current book $\mathbf{X}$ is the state space, $\mathbf{A}$ is the action space, $p_t(dy|x, a)$ is the transition probability, $c_t(x, a)$ and $C(x)$ are the loss functions.

Normally, we denote random variables with capital letters (X), small letters (x) being used just for variables, arguments of functions, etc. Bold case ($\mathbf{X}$) is for spaces. All functions, mappings, and stochastic kernels are assumed to be Borel-measurable unless their properties are explicitly specified.

We say that a function on $\mathbb{R}^1$ with the values in a Borel space $\mathbf{A}$ is *piece-wise continuous* if there exists a sequence y_i such that $\lim_{i\to\infty} y_i = \infty$; $\lim_{i\to-\infty} y_i = -\infty$, this function is continuous on each open interval

(y_i, y_{i+1}) and there exists a right (left) limit as $y \to y_i + 0$ $(y \to y_{i+1} - 0)$, $i = 0, \pm 1, \pm 2 \ldots$. A similar definition is accepted for real-valued piece-wise Lipschitz, continuously differentiable functions.

If $\mathbf{X}$ is a measurable space and ν is a measure on it, then both formulae

$$\int_{\mathbf{X}} f(x) d\nu(x) \qquad \text{and} \qquad \int_{\mathbf{X}} f(x) \nu(dx)$$

denote the same integral of a real-valued function f with respect to ν.

w.r.t. is the abbreviation for 'with respect to', a.s. means 'almost surely', and CDF means 'cumulative distribution function'.

We consider only minimization problems. When formulating theorems and examples published in books (articles) devoted to maximization, we always adjust the statements for our case without any special remarks.

It should be emphasized that the terminology in MDP is not entirely fixed. For example, very often strategies are called policies. There exist several slightly different definitions of a semi-continuous model, and so on.

The author is thankful to Dr.R. Sheen and to Dr.M. Ruck for the proof reading of all the text.

A.B. Piunovskiy

Contents

Chapter 1

Finite-Horizon Models

1.1 Preliminaries

A decision maker is faced with the problem of influencing the behaviour of a probabilistic system as it evolves through time. Decisions are made at discrete points in time referred to as *decision epochs* and denoted as $t = 1, 2, \ldots, T < \infty$. At each time t, the system occupies a *state* $x \in \mathbf{X}$. The *state space* $\mathbf{X}$ can be either discrete (finite or countably infinite) or continuous (non-empty uncountable Borel subset of a complete, separable metric space, e.g. $\mathbb{R}^1$). If the state at time t is considered as a random variable, it is denoted by a capital letter X_t; small letters x_t are just for possible values of X_t. Therefore, the behaviour of the system is described by a *stochastic (controlled) process*

$$X_0, X_1, X_2, \ldots, X_T.$$

In case of uncontrolled systems, the theory of Markov processes is well developed: the initial probability distribution for X_0, $P_0(dx)$, is given, and the dynamics are defined by *transition probabilities* $p_t(dy|x)$. When $\mathbf{X}$ is finite and the process is time-homogeneous, those probabilities form a transition matrix with elements $p(j|i) = P(X_{t+1} = j|X_t = i)$.

In the case of controlled systems, we assume that the *action space* $\mathbf{A}$ is given, which again can be an arbitrary Borel space (including the case of finite or countable $\mathbf{A}$). As soon as the state X_{t-1} becomes known (equals x_{t-1}), the decision maker must choose an action/control $A_t \in \mathbf{A}$; in general this depends on all the realized values of $X_0, X_1, \ldots, X_{t-1}$ along with past actions $A_1, A_2, \ldots, A_{t-1}$. Moreover, that decision can be randomized. The rigorous definition of a control strategy is given in the next section.

As a result of choosing action a at decision epoch t in state x, the decision maker loses $c_t(x, a)$ units, and the system state at the next decision

epoch is determined by the probability distribution $p_t(dy|x,a)$. The function $c_t(x,a)$ is called a *one-step loss*. The *final/terminal* loss equals $C(x)$ when the final state $X_T = x$ is realized.

We assume that the *initial distribution* $P_0(dx)$ for X_0 is given. Suppose a control strategy π is fixed (that is, the rule of choosing actions a_t; see the next section). Then the random sequence

$$X_0, A_1, X_1, A_2, X_2, \ldots, A_T, X_T$$

is well defined: there exists a single probability measure $P^{\pi}_{P_0}$ on the space of trajectories

$$(x_0, a_1, x_1, a_2, x_2, \ldots, a_T, x_T) \in \mathbf{X} \times (\mathbf{A} \times \mathbf{X})^T.$$

For example, if $\mathbf{X}$ is finite and the control strategy is defined by the map $a_t = \varphi_t(x_{t-1})$, then

$$P^{\varphi}_{P_0}\{X_0 = i, A_1 = a_1, X_1 = j, A_2 = a_2, X_2 = k, \ldots, X_{T-1} = l, A_T = a_T, X_T = m\}$$

$$= P_0(i) I\{a_1 = \varphi_1(i)\} p_1(j|i,a_1) I\{a_2 = \varphi_2(j)\} \ldots p_T(m|l,a_T).$$

Here and below, $I\{\cdot\}$ is the indicator function; if $\mathbf{X}$ is discrete then transition probabilities $p_t(\cdot|x,a)$ are defined by the values on singletons $p_t(y|x,a)$. The same is true for the initial distribution.

Therefore, for a fixed control strategy π, the *total expected loss* equals $v^{\pi} = E^{\pi}_{P_0}[W]$, where

$$W = \sum_{t=1}^{T} c_t(X_{t-1}, A_t) + C(X_T)$$

is the *total realized loss*. Here and below, $E^{\pi}_{P_0}$ is the mathematical expectation with respect to probability measure $P^{\pi}_{P_0}$.

The aim is to find an *optimal* control strategy π^* solving the problem

$$v^{\pi} = E^{\pi}_{P_0}\left[\sum_{t=1}^{T} c_t(X_{t-1}, A_t) + C(X_T)\right] \longrightarrow \inf_{\pi}. \tag{1.1}$$

Sometimes we call v^{π} the *performance functional*.

Using the dynamic programming approach, under some technical conditions, one can prove the following statement. Suppose function $v_t(x)$ on $\mathbf{X}$ satisfies the following equation

$$\begin{cases} v_T(x) = C(x); \\ v_{t-1}(x) = \inf\limits_{a \in \mathbf{A}} \left\{ c_t(x,a) + \displaystyle\int_X v_t(y) p_t(dy|x,a) \right\} \\ \qquad\quad = c_t(x, \varphi^*_t(x)) + \displaystyle\int_X v_t(y) p_t(dy|x, \varphi^*_t(x)); \qquad t = T, T-1, \ldots, 1. \end{cases} \tag{1.2}$$

Then, the control strategy defined by the map $a_t = \varphi_t^*(x_{t-1})$ solves problem (1.1), i.e. it is optimal; $\inf_\pi v^\pi = \int_{\mathbf{X}} v_0(x) P_0(dx)$. Therefore, control strategies of the type presented are usually sufficient for solving standard problems. They are called *Markov selectors*.

1.2 Model Description

We now provide more rigorous definitions.

The *Markov Decision Process (MDP)* with a finite horizon is defined by the collection

$$\{\mathbf{X}, \mathbf{A}, T, p, c, C\},$$

where $\mathbf{X}$ and $\mathbf{A}$ are the state and action spaces (Borel); T is the *time horizon*; $p_t(dy|x,a)$, $t = 1, 2, \ldots, T$, are measurable stochastic kernels on $\mathbf{X}$ given $\mathbf{X} \times \mathbf{A}$; $c_t(x,a)$ are measurable functions on $\mathbf{X} \times \mathbf{A}$ with values on the extended straight-line $\mathbb{R}^* = [-\infty, +\infty]$; $C(x)$ is a measurable map $C : \mathbf{X} \to \mathbb{R}^*$. Necessary statements about Borel spaces are presented in Appendix A.

The *space of trajectories* (or *histories*) up to decision epoch t is

$$\mathbf{H}_{t-1} \stackrel{\triangle}{=} \mathbf{X} \times (\mathbf{A} \times \mathbf{X})^{t-1}, \quad t = 1, 2, \ldots, T: \quad \mathbf{H} \stackrel{\triangle}{=} \mathbf{X} \times (\mathbf{A} \times \mathbf{X})^T.$$

A *control strategy* $\pi = \{\pi_t\}_{t=1}^T$ is a sequence of measurable stochastic kernels

$$\pi_t(da|x_0, a_1, x_1, \ldots, a_{t-1}, x_{t-1}) = \pi_t(da|h_{t-1})$$

on $\mathbf{A}$, given $\mathbf{H}_{t-1}$. If a strategy π^{m} is defined by (measurable) stochastic kernels $\pi_t^{\mathrm{m}}(da|x_{t-1})$ then it will be called a *Markov* strategy. It is called *semi-Markov* if it has the form $\pi_t(da|x_0, x_{t-1})$. A Markov strategy π^{ms} is called *stationary* if none of the kernels $\pi^{\mathrm{ms}}(da|x_{t-1})$ depends on the time t. Very often, stationary strategies are denoted as π^{s}. If for any $t = 1, 2, \ldots, T$ there exists a measurable mapping $\varphi_t(h_{t-1}) : \mathbf{H}_{t-1} \to \mathbf{A}$ such that $\pi_t(\Gamma|h_{t-1}) = I\{\Gamma \ni \varphi_t(h_{t-1})\}$ for any $\Gamma \in \mathcal{B}(\mathbf{A})$, then the strategy is denoted by the symbol φ and is called a *selector* or *non-randomized* strategy. Selectors of the form $\varphi_t(x_{t-1})$ and $\varphi(x_{t-1})$ are called *Markov* and *stationary* respectively. Stationary semi-Markov strategies and semi-Markov (stationary) selectors are defined in the same way. In what follows, Δ^{All} is the collection of all strategies, Δ^{M} is the set of all Markov strategies, Δ^{MN} is the set of all Markov selectors. In this connection, letter N

corresponds to non-randomized strategies. Further, Δ^{S} and Δ^{SN} are the sets of all stationary strategies and of all stationary selectors.

We assume that *initial* probability distribution $P_0(dx)$ is fixed. If a control strategy π is fixed too, then there exists a unique probability measure $P^{\pi}_{P_0}$ on $\mathbf{H}$ such that $P^{\pi}_{P_0}(\Gamma^X) = P_0(\Gamma^X)$ for $\Gamma \in \mathcal{B}(\mathbf{H}_0) = \mathcal{B}(\mathbf{X})$ and, for all $t = 1, 2, \ldots, T$, for $\Gamma^G \in \mathcal{B}(\mathbf{H}_{t-1} \times \mathbf{A})$, $\Gamma^X \in \mathcal{B}(\mathbf{X})$

$$P^{\pi}_{P_0}(\Gamma^G \times \Gamma^X) = \int_{\Gamma^G} p_t(\Gamma^X | x_{t-1}) P^{\pi}_{P_0}(dg_t)$$

and

$$P^{\pi}_{P_0}(\Gamma^H \times \Gamma^A) = \int_{\Gamma^H} \pi_t(\Gamma^A | h_{t-1}) P^{\pi}_{P_0}(dh_{t-1})$$

for $\Gamma^H \in \mathcal{B}(\mathbf{H}_{t-1})$, $\Gamma^A \in \mathcal{B}(\mathbf{A})$. Here, with some less-than-rigorous notation, we also denote $P^{\pi}_{P_0}(\cdot)$ the images of $P^{\pi}_{P_0}$ relative to projections of the types

$$\mathbf{H} \to \mathbf{H}_{t-1} \times \mathbf{A} \stackrel{\triangle}{=} \mathbf{G}_t, \quad t = 1, 2, \ldots, T, \text{ and } \mathbf{H} \to \mathbf{H}_t, \quad t = 0, 1, 2, \ldots, T. \tag{1.3}$$

$g_t = (x_0, a_1, x_1, \ldots, a_t)$ and $h_t = (x_0, a_1, x_1, \ldots, a_t, x_t)$ are the generic elements of $\mathbf{G}_t$ and $\mathbf{H}_t$. Where they are considered as random elements on $\mathbf{H}$, we use capital letters G_t and H_t, as usual.

Measures $P^{\pi}_{P_0}(\cdot)$ on $\mathbf{H}$ are called *strategic* measures; they form space $\mathcal{D}$.

One can introduce σ-algebras $\mathcal{G}_t$ and $\mathcal{F}_t$ in $\mathbf{H}$ as the pre-images of $\mathcal{B}(\mathbf{G}_t)$ and $\mathcal{B}(\mathbf{H}_t)$ with respect to (1.3). Now the trivial projections

$$(x_0, a_1, x_1, \ldots, a_T, x_T) \to x_t \text{ and } (x_0, a_1, x_1, \ldots, a_T, x_T) \to a_t$$

define $\mathcal{F}$-adapted and $\mathcal{G}$-adapted stochastic processes $\{X_t\}_{t=0}^T$ and $\{A_t\}_{t=1}^T$ on the *stochastic basis* $(\mathbf{H}, \mathcal{B}(\mathbf{H}), \{\mathcal{F}_0, \mathcal{G}_1, \mathcal{F}_1, \ldots, \mathcal{G}_T, \mathcal{F}_T\}, P^{\pi}_{P_0})$, which is completed as usual. Note that the process A_t is $\mathcal{F}$-predictable, and that this property is natural. That is the main reason for considering sequences $(x_0, a_1, x_1, \ldots, a_T, x_T)$, not $(x_0, a_0, x_1, \ldots, a_{T-1}, x_T)$. The latter notation is also widely used by many authors.

For each $h \in \mathbf{H}$ the (realized) *total loss* equals

$$w(h) = \sum_{t=1}^{T} c_t(x_{t-1}, a_t) + C(x_T),$$

where we put "$+\infty$" + "$-\infty$" $\stackrel{\triangle}{=}$ "$+\infty$". The map $W: \; h \to w(h)$ defines the random total loss, and the performance of control strategy π is given by $v^{\pi} = E^{\pi}_{P_0}[W]$. Here and below,

$$E^{\pi}_{P_0}[W] \stackrel{\triangle}{=} E^{\pi}_{P_0}[W^+] + E^{\pi}_{P_0}[W^-]; \quad \text{"}+\infty\text{"} + \text{"}-\infty\text{"} \stackrel{\triangle}{=} \text{"}+\infty\text{";}$$