Markov Decision Processes

D. J. White

MARKOV DECISION PROCESSES

This page intentionally left blank

MARKOV DECISION PROCESSES

D. J. White
University of Manchester, UK

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

Jacaranda Wiley Ltd, G.P.O. Box 859, Brisbane,
Queensland 4001, Australia

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario M9W 1L1, Canada

John Wiley & Sons (SEA) Pte Ltd, 37 Jalan Pemimpin #05-04,
Block B, Union Industrial Building, Singapore 2057

Library of Congress Cataloging-in-Publication Data

White, D. J. (Douglas John)
Markov decision processes / D. J. White.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-93627-8
1. Markov processes. 2. Statistical decision. I. Title.
QA274.7.W45 1992
519.2'33—dc20 92-1646
CIP

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-471-93627-8

Typeset in 11/13pt Times by Mathematical Composition Setters Ltd, Salisbury, Wiltshire
Printed and bound in Great Britain by Biddles Ltd, Guildford, Surrey

To Solo
who was my study companion

This page intentionally left blank

Contents

Preface

This text is based on a course given for Ph.D. systems engineers at the University of Virginia in the autumn of 1987. The approach, and level of mathematics used, is intended for the mathematically minded post-graduate students in such areas as systems engineering, industrial engineering, management science and operations research, and for final year mathematicians and statisticians.

It is not intended to be a research reference text, although reference will be given to some key texts and papers for those wishing to pursue research in this area. It is intended as a basic text, covering some of the fundamentals involved in the manner in which Markov decision problems may be properly formulated and solutions, or properties of such solutions, determined.

There are three key texts influencing the format of this text, viz. those of Howard [23], van der Wal [53] and Kallenberg [24].

Howard [23], for stationary Markov decision processes, uses what he calls the 'z-transform' method, which is essentially the method of 'generating functions'. This allows expected total discounted, or non-discounted, rewards over a residual n-time unit horizon, to be easily determined from the coefficients of the z-transforms, for any given policy. It also allows one to see how these performance measures depend upon the number of time units, n, of the time horizon, and leads to asymptotic results when n tends to infinity. In principle the z-transforms may be found for each of a set of policies, and the appropriate decision rule selected on the basis of this analysis. However, this can be impracticable, and an alternative approach, based upon so-called 'optimality (functional) equations', is then used. However, the natural insight gained from the use of z-transform analysis is very helpful, particularly when the form of the dependence of the n-time unit performance on n is needed. Thus, Chapter 1 is designed to be an introductory chapter, based on z-transform analysis, without any niceties of random variable structure being included, and restricted strictly to stationary situations.

Van der Wal [53] develops Markov decision processes in terms of the primitive random variable structure governing the process, and does not take for granted, as does Chapter 1 implicitly, that policies may be restricted, in effect, to deterministic Markov policies. Thus, all possible history-remembering policies are initially allowed for and then, for some classes of problem, the validity of this assumption is established. For some classes of problem the assumption is not a valid one, e.g. those where mean-variance analysis is of concern. Thus, various classes of policy have to be considered, and this is the approach of Chapter 2. This chapter leads, for some classes of problem, to the well-known 'optimality (functional) equations', and properties of the solutions of these equations are studied.

Chapter 3 looks at algorithms, and their properties, for solving the optimality equations developed in Chapter 2. There is a close relationship between linear programming and the optimality equations developed in Chapter 3. Kallenberg [24] gives a very complete, and insightful, treatment of this interrelationship. Chapter 4 provides merely the rudiments of this relationship. The policy space iteration algorithm is shown to be a block-pivoting form of linear programming, and a potentially more efficient algorithm, although the existence of efficient linear programming algorithms is a point to be borne in mind when selecting an algorithm.

For many decision situations the interval between successive decision epochs is not constant, and may itself be a random variable. Chapter 5 deals with this departure from standard Markov decision processes, as a natural generalisation of these. The title 'semi-Markov decision processes' is used, although, strictly speaking, the latter refer only to those situations in which decision epochs occur only when there is a change of state. We take some license here in using this terminology. This chapter takes for granted all the corresponding random variable analysis of Chapter 2, and deals solely with the optimality (functional) equations and algorithms.

Up to this point it is assumed that the state of the system is known at any time, and that the governing, fixed, parameters of the process are known. In Chapter 6 we deviate from these assumptions, and consider processes in which the knowledge of the states and/or parameters is encapsulated in the form of probability distributions. The new state space becomes a vector space, and this is a departure from the finite state space framework used up to this point. The random variable analysis is here taken for granted, and the emphasis is on the development of the optimality (functional) equations.

Chapter 7 deals briefly with, to some extent, some more modern developments in the Markov decision process area, covering: structural policy analysis; approximation modelling; post-optimality, parametric and sensitivity analysis; multiple objectives; utility, probabilistic constraints and mean-variance analysis; and Markov games.

In teaching Markov decision process material, the author's experience has been that, whereas algorithms are most readily digested, problem formulation has been least readily digested. Chapter 8 is included as a selection of problem formulation illustrations to assist with this aspect of learning. They are very much simplified, but may be useful as a precursor to much more detailed illustrations which occur in the literature.

Certain points are important to facilitate the reader's ability to cope with the style of presentation.

(i) Throughout, the main objective has been to reduce the analysis to one involving operators of the form T, and others, which transform one function into another. This greatly simplifies the presentation. If u is a real valued function on the state space I, then the value of the transformed function Tu at $i \in I$ is written as $[Tu](i)$.

(ii) Throughout the text, the dual use of 'function' and 'vector' has been used. Thus, sometimes u is to be interpreted as a vector, viz. in those cases where it is expedient to use 'vector-matrix' analysis. Sometimes u is to be interpreted as a function, viz. in those cases where 'optimality (functional) equation' analysis is used. The same u may be used for these two distinct purposes.

(iii) For any matrix M, its (i, j)th element may be written as $[M]_{ij}$, and, for a vector q, its ith element may be written as $[q]_i$, with certain standard exceptions such as p_{ij}^k, r_i^k, etc.

(iv) Because I is used to denote the state space, we use U to denote the identity matrix.

(v) Throughout the text, the format is a 'maximisation' one for Markov decision processes with rewards. Thus, with minor exceptions in exercises, all 'cost minimisation' problems are converted to 'reward maximisation' problems with the use of a 'minus' sign.

(vi) Throughout the text, 'n' or '$n - t$' will be used to denote 'the number of time units (or decision epochs) remaining for the time horizon', or 'n' will be used to denote 'the last time unit in the time horizon', or 'the iteration number in an algorithm', and 't' will be used to denote chronological time.

(vii) Throughout the text we will distinguish between the actual physical values in a Markov process and the solutions to corresponding equations and inequalities. Thus, $\{v, v_n, v^\pi, v_n^\pi\}$ will be used to denote actual discounted reward value functions (vectors), and $\{w, w_n, w^\pi, w_n^\pi\}$, $\{g, g_n, g^\pi, g_n^\pi\}$ will be used to denote 'actual' bias (up to a constant factor difference) and gain functions (vectors), whereas u will be used to denote any 'general solution' to corresponding optimisation problems in the discounted case, and $\{u, h\}$ for the average expected reward case.

Finally, references are given for follow-up purposes, for the reader to obtain further details if required. The references are not necessarily intended to suggest that they are the original sources. Nor do the references necessarily imply that the cited material is identical with the material being discussed. Also, the absence of a reference does not necessarily imply that the material being discussed is original.

CHAPTER 1
Introduction

1.1 AN INTRODUCTORY EXAMPLE

1.1.1 TOYMAKER EXAMPLE—EXPECTED TOTAL REWARD

The following example is taken from Howard [23], which is a useful companion text to this one. It is used in various places in this text to illustrate certain points. We introduce this in a non-rigorous fashion prior to developing the rigour later (see p. 24).

In this problem, a toymaker's business is described as being in one of two conditions (states), $i = 1, 2$, at the beginning of any year (see Table 1.1). He has one of two things he can do (actions), $k = 1, 2$ (see Table 1.2). If he is in state i at the beginning of a year and takes action k then he moves to state j at the beginning of the next year with probability p_{ij}^k (see Table 1.3, where, for the moment, k is restricted to 1 for each state i), with $P = [p_{ij}^k]$ being the transition probability matrix. If he moves from state i to state j in the year, having taken action k, then he receives a reward r_{ij}^k (see Table 1.4, again with $k = 1$, for the moment). Finally, if he is in state i at the beginning of a year and takes action k then his expected reward in the year is r_i^k (see Table 1.5, again with $k = 1$ for the moment) with r being the vector (or function) of expected rewards.

For this problem the action k to be taken is prespecified as $k = 1$ for each state i. The essential objective is to determine which action is actually optimal for each state i, and this is the concern of Markov decision processes to which we will turn later (see p. 20). For the moment we restrict ourselves to the fixed decision rule implicit in Table 1.3, i.e. $k = 1$ for each i.

Let us now look at the manner in which such a system will behave over a number of years, n. In Figure 1.1, n is the number of years to the end of a specific time horizon and not the chronological year. This is the form needed later (see p. 26) for the development of our theory.

Table 1.1 States*

State i	*Physical condition*
1	successful toy
2	unsuccessful toy

Table 1.2 Actions*

State i	*Action k*	*Physical action*
1	1	do not advertise
	2	advertise
2	1	do no research
	2	do research

Table 1.3 Transition probabilities*

		State at beginning of next year j	
		1	2
State at beginning of year i	1	0.5	0.5
	2	0.4	0.6

$[p_{ij}^k] = P$

$k = 1$

Table 1.4 Rewards in year*

		State at beginning of next year j	
		1	2
State at beginning of year i	1	9	3
	2	3	−7

$[r_{ij}^k]$

$k = 1$

* Tables 1.1, 1.2, 1.3, 1.4 and 1.5 are reproduced from [23] Howard (1960), pp. 19 and 26–27, by permission of The MIT Press.

Table 1.5 Expected rewards in year*

State at beginning of year i		$[r_i^k] = r$ $k = 1$
	1	6
	2	−3

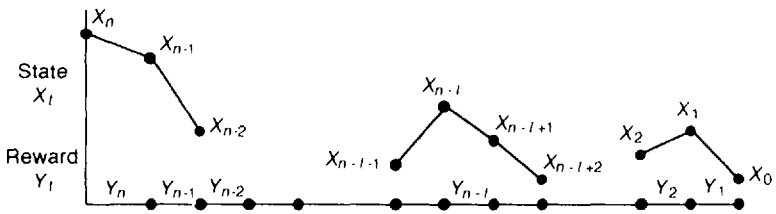

Figure 1.1 System behaviour over n years

On occasion we will have a need to use chronological years also, and we will denote this by t. For the moment let us proceed in terms of n as defined.

In Figure 1.1, in addition to n as defined, X_n is the random state at the beginning of year n from the end of the time horizon, and Y_n is the random reward in this year. In a given realisation (sample path) of the behaviour of the system, X_n will take a particular value i and Y_n will take a particular value r_{ij}^k, if action k is taken and if $X_{n-1} = j$.

Let us now look at the question of deriving the expected rewards for each year over a time horizon of length n years. The expected reward in year n, starting with $X_n = 1$, is 6. The expected reward in year n, starting with $X_n = 2$, is -3. In vector terms we have

$$r_n = \begin{bmatrix} 6 \\ -3 \end{bmatrix} = P^0 r, \tag{1.1}$$

where $[r_n]_i$ is the expected value of Y_n given $X_n = i$, and P^0 is the identity matrix which later (see p. 45) we will label as U generally.

The expected reward in year $n-1$, starting with $X_n = i$ at the beginning of year n, is

$$\text{probability}(X_{n-1} = 1 \mid X_n = i) r_1^k + \text{probability}(X_{n-1} = 2 \mid X_n = i) r_2^k. \tag{1.2}$$

Thus

$$\text{for } i = 1 \text{ we have } 0.5 \times 6 + 0.5 \times (-3) = 1.5,$$
$$\text{for } i = 2 \text{ we have } 0.4 \times 6 + 0.6 \times (-3) = 0.06.$$

For the general i these are equal to $[Pr]_i$ where, for any vector q, $[q]_i$ is the ith component.

In general the expected reward in year $n - l$, starting with $X_n = i$ at the beginning of year n, is

$$\text{probability}(X_{n-l} = 1 \mid X_n = i) r_1^k + \text{probability}(X_{n-l} = 2 \mid X_n = i) r_2^k = [P^l r]_i, \qquad 0 \leqslant l \leqslant n-1. \tag{1.3}$$

Note that if $q_{ij}^{(l)}$ is the probability that, if the process commences in state i, it will be in state j after l years then $q_{ij}^{(l)} = [P^l]_{ij}$ where, for any matrix M, $[M]_{ij}$ is its (i, j)th element.

We will now look at the question of deriving the expected total reward over n years. Let $v_n(i)$ be the expected total reward over n years if we begin with $X_n = i$. This is well determined by i and n, and by the given transition probability matrix P.

Table 1.6 gives a selection of $\{v_n(i)\}$ values (see Howard [23], p. 19), where with E being the expectation operator

$$v_n(i) = E\left(\sum_{l=0}^{l=n-1} Y_{n-l} \mid X_n = i\right) = \sum_{l=0}^{l=n-1} E(Y_{n-l} \mid X_n = i)$$
$$= \left[\left(\sum_{l=0}^{l=n-1} P^l\right) r\right]_i. \tag{1.4}$$

In vector form this is

$$v_n = \left(\sum_{l=0}^{l=n-1} P^l\right) r. \tag{1.5}$$

Formally we define $v_0 = 0$.

Table 1.6 Expected total rewards (reproduced from [23] Howard (1960), p. 19, by permission of The MIT Press)

		Number of years to time horizon n					
		0	1	2	3	4	5
State at beginning of year i	1	0	6	7.5	8.55	9.555	10.5555
	2	0	−3	−2.4	−1.44	−0.444	0.5556

1.1.2 THE USE OF z-TRANSFORMS FOR FINDING EXPECTED TOTAL REWARD

We follow a similar line of development to that of Howard [23]. Define

$$f(z) = \sum_{n=0}^{\infty} v_n z^n, \qquad 0 \leqslant z < 1. \tag{1.6}$$

We note that:

(i) $f(z)$ is a vector for each z;
(ii) $f(z)$ is convergent for some range of z;

and is convergent for $0 \leqslant z < 1$ because, for $n \geqslant 1$, $|v_n(i)| \leqslant 6n$ and $\sum_{n=1}^{\infty} nz^n$ is convergent for $0 \leqslant z < 1$ (D'Alembert's ratio test, see Bromwich [8], p. 39).

We have

$$\underline{n \geqslant 1} \qquad v_n = \left(\sum_{l=0}^{l=n-1} P^l\right) r = r + Pv_{n-1}. \tag{1.7}$$

Thus

$$\sum_{n=1}^{\infty} v_n z^n = r \sum_{n=1}^{\infty} z^n + zP \sum_{n=0}^{\infty} v_n z^n. \tag{1.8}$$

Because $v_0 = 0$ we have

$$f(z) = (U - zP)^{-1}(z/(1-z))r. \tag{1.9}$$

Note that U is the identity matrix.

From Howard [23], p. 21 we have

$$(U - zP)^{-1} = \begin{bmatrix} 1 - 0.5z & -0.5z \\ -0{\cdot}4z & 1 - 0.6z \end{bmatrix}^{-1}$$

$$= \frac{\begin{bmatrix} 1 - 0.6z & 0.5z \\ 0.4z & 1 - 0.5z \end{bmatrix}}{((1 - 0.5z)(1 - 0.6z) - (0.5z)(0.4z))}$$

$$= \frac{\begin{bmatrix} 1 - 0.6z & 0.5z \\ 0.4z & 1 - 0.5z \end{bmatrix}}{(1 - 1.1z + 0.1z^2)}$$

$$= \frac{\begin{bmatrix} 1-0.6z & 0.5z \\ 0.4z & 1-0.5z \end{bmatrix}}{(1-z)(1-0.1z)}$$

$$= \begin{bmatrix} 1-0.6z & 0.5z \\ 1-0.4z & 1-0.5z \end{bmatrix} \left(\frac{\frac{10}{9}}{(1-z)} - \frac{\frac{1}{9}}{(1-0.1z)} \right)$$

= (after some manipulation)

$$(1/(1-z)) \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} + (1/(1-0.1z)) \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix}.$$

Then

$$(U - zP)^{-1} z/(1-z)$$

$$= (z/(1-z)^2) \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} + (z/(1-z)(1-0.1z)) \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix}$$

$$= (z/(1-z)^2) \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} + \left(\frac{\frac{10}{9}}{(1-z)} - \frac{\frac{10}{9}}{(1-0.1z)} \right) \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} \left(\sum_{n=0}^{\infty} n z^n \right) + \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix} \left(\sum_{n=0}^{\infty} (\tfrac{10}{9})(1-(0.1)^n) z^n \right)$$

$$= \sum_{n=0}^{\infty} \left(n \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} + (\tfrac{10}{9})(1-(0.1)^n) \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix} \right) z^n. \tag{1.10}$$

Hence

$$\underline{n \geqslant 0}\ v_n = \left(n \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} + (\tfrac{10}{9})(1-(0.1)^n) \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix} \right) \begin{bmatrix} 6 \\ -3 \end{bmatrix}$$

$$= n \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \frac{50}{9} \\ -\frac{40}{9} \end{bmatrix} - (\tfrac{10}{9})(0.1)^n \begin{bmatrix} 5 \\ -4 \end{bmatrix}. \tag{1.11}$$

Let us now look at the asymptotic behaviour of $\{v_n\}$.

1.1.3 ASYMPTOTIC BEHAVIOUR OF EXPECTED TOTAL REWARD

Equation (1.11) takes the form

$$v_n = ng + w + \varepsilon_n \tag{1.12}$$

where

$$g = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad w = \begin{bmatrix} \frac{50}{9} \\ -\frac{40}{9} \end{bmatrix}, \qquad \varepsilon_n = \begin{bmatrix} -\frac{50}{9} \\ \frac{40}{9} \end{bmatrix} (0.1)^n$$

where g is called the gain function (or vector) and w is called the bias function (or vector). Note that whether $\{v_n, g, w, \varepsilon_n\}$ are seen as functions of the states, or as vectors whose components correspond to the states, is merely a matter of perspective. Then

$$v_n/n = g + w/n + \varepsilon_n/n \tag{1.13}$$

and the function of limiting expected rewards per year is

$$\lim_{n \to \infty} [v_n/n] = g. \tag{1.14}$$

We note the following:

(i) In this example, $\lim_{n \to \infty} [\varepsilon_n] = 0$. This is so for problems where $\lim_{n \to \infty} [P^n]$ exists (see Mine and Osaki [34], Lemma 3.6). It is not so in general.

(ii) In this example, $g(1) = g(2) = 1$. This is so for problems with one ergodic state set (uni-chain case), even with transient states. It is not so in general.

A transient state j is one for which

$$\lim_{n \to \infty} \left[\left[\left(\sum_{l=0}^{n-1} P^l \right) \Big/ n \right] \right]_{ij} = 0, \quad \forall\, i. \tag{1.15}$$

The left-hand side of (1.15) is the limiting average probability of being in state j at the beginning of each year if the initial state is i.

In our example we see that, from (1.10),

$$\left(\sum_{l=0}^{n-1} P^l \right) = n \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} + \left(\tfrac{10}{9}\right)(1 - (0.1)^n) \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix}.$$

Hence

$$\lim_{n \to \infty} \left[\left(\sum_{l=0}^{n-1} P^l \right) \Big/ n \right] = \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix}.$$

This is a uni-chain case, where the limiting probabilities are independent of the starting state i, and there are no transient states. In this case,

the limits for $j = 1$, $j = 2$ are 4/9, 5/9. These are also steady-state probabilities, where a probability vector θ is said to be steady state for P if

$$\theta P = \theta. \tag{1.16}$$

The example we have studied has a single chain and no transient states. Let us look at multiple-chain and transient-state possibilities.

1.2 MULTIPLE CHAINS AND TRANSIENT STATES

The following example is taken from Howard [23], pp. 12, 13:

$$P = \begin{bmatrix} 0.75 & 0.25 \\ 0 & 1 \end{bmatrix}.$$

Here P^n is the coefficient of z^n in $\sum_{n=0}^{\infty} P^n z^n$.

$$\left(\sum_{n=0}^{\infty} P^n z^n\right) = (U - zP)^{-1} = \begin{bmatrix} 1 - 0.75z & -0.25z \\ 0 & 1 - z \end{bmatrix}^{-1}$$

$$= \text{(by using a process similar to that of the toymaker example)}$$

$$(1/(1-z))\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} + (1/(1 - 0.75z))\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}.$$

Thus

$$P^n = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} + (0.75)^n \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$$

and

$$\lim_{n \to \infty} [P^n] = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Thus, whatever the starting state i is, the limiting probability of being in state $j = 1$ is 0. Also

$$\sum_{l=0}^{n-1} P^l = n\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} + 4(1 - (0.75)^n)\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}.$$

Thus

$$\lim_{n\to\infty}\left[\left(\sum_{l=0}^{n-1} P^l\right)\Big/n\right] = \begin{bmatrix}0 & 1\\ 0 & 1\end{bmatrix}.$$

The state $j = 1$ is a transient state.

We make the following notes:

(i) If $\lim_{n\to\infty}[P^n]$ exists then its ith row will give the limiting probabilities of being in each state for the specified initial state i.

(ii) The $\lim_{n\to\infty}[P^n]$ need not exist, e.g. if

$$P = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$$

then

$$P^n = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \text{ if } n \text{ is odd, } P^n = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \text{ if } n \text{ is even.}$$

(iii) The $\lim_{n\to\infty}\left(\sum_{l=0}^{n-1} P^l/n\right)$ always exists, and is the Cesàro limit, P^*.

(iv) When $\lim_{n\to\infty}[P^n]$ exists, it is equal to the Cesàro limit.

(v) The $\lim_{n\to\infty}[P^n]_i$, when it exists, is the vector of steady-state probabilities for initial state i.

$$[P^*]_i = \lim_{n\to\infty}\left[\left(\sum_{l=0}^{n-1} P^l\right)\Big/n\right]_i$$

is the vector of limiting average probabilities of being in various states for initial state i.

(vi) We will use the latter in subsequent work for average expected reward problems.

(vii)
$$PP^* = P^*P = P^*. \tag{1.17}$$

(viii) The following example is taken from Howard [23], pp. 13–15:

$$P = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}.$$

Figure 1.2 gives a graphical representation of the state transitions, from which, intuitively, one can see that the limiting state probabilities, and the Cesàro limits, depend upon the starting state, with states 1

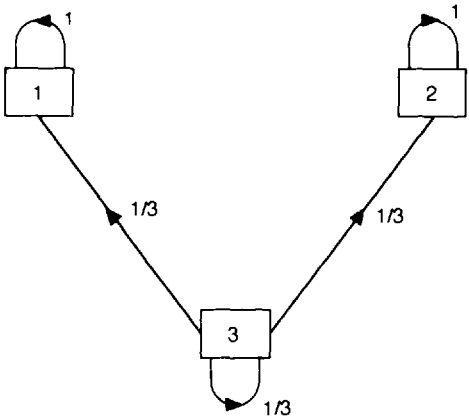

Figure 1.2 State transitions (reproduced from [23] Howard (1960), p. 14, by permission of The MIT Press)

and 2 playing equal roles, distinguishable between themselves and from state 3.

Using z-transform analysis we obtain

$$\underline{n \geqslant 1} \qquad P^n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} + (\tfrac{1}{3})^n \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix},$$

$$\lim_{n \to \infty} [P^n] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = P^*.$$

$[P^*]_{i1}$ depends upon i.
$[P^*]_{i2}$ depends upon i.
$[P^*]_{i3}$ is independent of i, and state 3 is transient.

There are two chains here, each with one member, viz.

State 1 constitutes one chain.
State 2 constitutes a second chain.
State 3 is transient.

$$g = \lim_{n \to \infty} [v_n / n] \text{ (see (1.14))}$$

$$= \lim_{n \to \infty} \left[\left(\sum_{l=0}^{n-1} P^l \right) r / n \right] \text{ (see (1.4))}$$

$$= P^* r \text{ (see note (iii)).} \qquad (1.18)$$

Thus

$$g(1) = r_1, \qquad g(2) = r_2, \qquad g(3) = (r_1 + r_2 + r_3)/3$$

and the gains can be different for each state.

Let us now look at how the general form of v_n in (1.12) may be used to give an asymptotic gain–bias equation in the uni-chain case. We will restrict ourselves throughout this text to this case for ease of exposition. However, much of what we do will carry over to the multiple-chain case with consequential increases in complexity of computations.

1.3 LIMITING GAIN–BIAS EQUATIONS IN THE UNI-CHAIN CASE

From (1.12) we have

$$\underline{n \geqslant 0} \qquad v_n = ng + w + \varepsilon_n. \tag{1.19}$$

From (1.7) we have

$$\underline{n \geqslant 1} \qquad v_n = r + Pv_{n-1}. \tag{1.20}$$

Suppose that

$$\lim_{n \to \infty} [\varepsilon_n] = 0. \tag{1.21}$$

This holds for our initial example, as will be seen from (1.11).

Substitute (1.19) into (1.20) to obtain

$$\underline{n \geqslant 1} \qquad ng + w + \varepsilon_n = r + P((n-1)g + w + \varepsilon_{n-1}). \tag{1.22}$$

Then (1.22) gives

$$w + ng - (n-1)Pg = r + Pw + P\varepsilon_{n-1} - \varepsilon_n. \tag{1.23}$$

In the uni-chain case, for our example, $g(1) = g(2) = g$, say. Thus (1.21) and (1.23) give

$$w(i) + g = r_i + [Pw]_i \quad \text{for each } i, \tag{1.24}$$

i.e. w, g satisfy

$$u(i) + h = r_i + \sum_j p_{ij}u(j) \quad \text{for each } i. \tag{1.25}$$

Equation (1.25) is a fundamental equation which we will use later (see p. 76) for computational purposes.

Note that in (1.19), (1.22) and (1.23), g is a function, whereas in (1.24) and (1.25), $\{g, h\}$ are scalars. We have taken some license in adopting this convention, so that (1.24) and (1.25) will assume the conventional notational form for the uni-chain case on which we will concentrate.

For uni-chain problems, (1.25) will always have a unique solution once we have set $u(i)$ equal to 0 for some value of i, e.g. $u(2) = 0$ (see Mine and Osaki [34], Lemma 3.3, when rank $(P^*) = 1$). Such a u vector will differ from the bias vector

$$w = \begin{bmatrix} \frac{50}{9} \\ -\frac{40}{9} \end{bmatrix}$$

in (1.11) by the same constant for each component.

We make the following notes:

(i) The condition (1.21) holds if limit $[P^n]$ exists (see Mine and Osaki [34], Lemma 3.6).

(ii) If, in (1.7), we let $v_0 = u$ where u is a solution to equation (1.25), instead of $v_0 = 0$ then

$$v_n = \left(\sum_{l=0}^{n-1} P^{n-l} \right) r + P^n u \tag{1.26}$$

and

$$v_n = ng + u. \tag{1.27}$$

This corresponds to a terminal value function equal to u instead of 0 in (1.7).

(iii) For the multiple-chain case the form of (1.19), (1.20), (1.22), (1.23) will still hold. Thereafter there is a departure because $g(i)$ is not necessarily the same for all i.

1.4 EXPECTED TOTAL DISCOUNTED REWARD OVER n YEARS

We have defined Y_n to be the random reward in year n from the end of the time horizon. Let us now introduce a discount factor ρ for each year, so that the value of a reward is reduced by a factor of ρ for each year of delay in its receipt.

In order to discount we need a time origin. This will be the beginning of year n from the end of the time horizon. Note that this will then vary

with n for a fixed terminal chronological year. The reward Y_{n-l} in year $n-l$ (see Figure 1.1), discounted back to the beginning of year n, is

$$\rho^l Y_{n-l}. \tag{1.28}$$

Following (1.1) and (1.3) we see that the expected discounted reward function in year $n-l$ is

$$\rho^l P^l r = (\rho P)^l r. \tag{1.29}$$

The expected total discounted reward function is, analogously to (1.5),

$$v_n = \left(\sum_{l=0}^{n-1} (\rho P)^l\right) r. \tag{1.30}$$

In (1.30) v_n is a function of ρ, but we will not put this in explicitly at this stage to avoid undue notational problems. The analysis in (1.6)–(1.9) follows in exactly the same manner to give (replacing P by ρP in (1.9))

$$f(z) = (U - \rho P z)^{-1}(z/(1-z))r. \tag{1.31}$$

For the toymaker example we have

$$f(z) = \left((1/(1-\rho z))\begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} + (1/(1-0.1\rho z))\begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix}\right)(z/(1-z))r.$$

From Howard [23], pp. 78, 79, when $\rho = 0.5$ we obtain (after some manipulation)

$$f(z) = \left((1/(1-z))\begin{bmatrix} \frac{28}{19} & \frac{10}{19} \\ \frac{8}{19} & \frac{30}{19} \end{bmatrix} + (1/(1-0.5z))\begin{bmatrix} -\frac{8}{9} & -\frac{10}{9} \\ -\frac{8}{9} & -\frac{10}{9} \end{bmatrix}\right.$$

$$\left. + (1/(1-0.05z))\begin{bmatrix} -100/171 & 100/171 \\ 80/171 & -80/171 \end{bmatrix}\right) r.$$

Hence, taking the coefficient of z^n in $f(z)$ we have, with

$$r = \begin{bmatrix} 6 \\ -3 \end{bmatrix},$$

$$\underline{n \geqslant 0} \qquad v_n = \begin{bmatrix} \frac{138}{19} \\ -\frac{42}{19} \end{bmatrix} + (0.5)^n \begin{bmatrix} -2 \\ -2 \end{bmatrix} + (0.05)^n \begin{bmatrix} -\frac{100}{19} \\ \frac{80}{19} \end{bmatrix}.$$

This takes the form, analogously to (1.12),

$$v_n = v + \varepsilon_n \tag{1.32}$$

where

$$v = \begin{bmatrix} \frac{138}{19} \\ -\frac{42}{19} \end{bmatrix}, \qquad \varepsilon_n = (0.05)^n \begin{bmatrix} -2 \\ -2 \end{bmatrix} + (0.05)^n \begin{bmatrix} -\frac{100}{19} \\ \frac{80}{19} \end{bmatrix}.$$

1.5 ASYMPTOTIC BEHAVIOUR OF EXPECTED TOTAL DISCOUNTED REWARD

Clearly in the above example

$$\lim_{n \to \infty} [v_n] = \begin{bmatrix} \frac{138}{19} \\ -\frac{42}{19} \end{bmatrix} = v.$$

We make the following notes:

(i) For discounted problems ($0 \leqslant \rho < 1$) $\lim_{n \to \infty} [v_n] = v$ will always exist (e.g. see White [58] Theorem 2.7, restricted to a single policy).

(ii) Let us look at the expression for $f(z)$ in (1.31) for the toymaker example. We have, when $0 \leqslant \rho < 1$,

$$\frac{1}{(1-\rho z)(1-z)} = \frac{1}{(1-\rho)} \left(\frac{1}{1-z} - \frac{\rho}{(1-\rho z)} \right)$$

and

$$\frac{1}{(1-0.1\rho z)(1-z)} = \frac{1}{(1-0.1\rho)} \left(\frac{1}{(1-z)} - \frac{0.1\rho}{(1-0.1\rho z)} \right).$$

Picking out the coefficient of z^n in $f(z)$ we have, for $n \geqslant 1$,

$$v_n = ((1-\rho^n)/(1-\rho)) \begin{bmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{4}{9} & \frac{5}{9} \end{bmatrix} r$$

$$+ ((1-(0.1\rho)^n)/(1-0.1\rho)) \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ -\frac{4}{9} & \frac{4}{9} \end{bmatrix} r.$$

Table 1.7 Expected total discounted rewards

		Number of years to time horizon n					
		0	1	2	3	4	5
State at beginning of year i	1	0	6	6.75	7.01	7.14	7.20
	2	0	−3	−2.7	−2.46	−2.34	−2.32

If $v = \lim_{n \to \infty} [v_n]$ then, for $0 \leqslant \rho < 1$,

$$v = \frac{1}{(1-\rho)} g + \frac{\frac{9}{10}}{(1-0.1\rho)} w$$

where g, w are as for the non-discounted case $\rho = 1$ (see (1.19) and subsequent calculations). Thus, noting that v is a function of ρ, we have

$$\lim_{\rho \to 1} [(1-\rho)v] = g. \tag{1.33}$$

Equation (1.33) is generally true (see Mine and Osaki [34], Lemmas 3.2 and 3.4) even for the multiple-chain case.

(iii) Replacing P by ρP in (1.7) we obtain

$$\underline{n \geqslant 1} \qquad v_n = r + \rho P v_{n-1}. \tag{1.34}$$

The sequence $\{v_n\}$ may be computed using (1.34). Table 1.7 gives, analogously to Table 1.6, the values for $\rho = 0.5$.

(iv) Using (1.34) we see that $\lim_{n \to \infty} [v_n] = v$ in (1.32), or (1.34), satisfies uniquely the following equation (see White [58], Theorem 2.7):

$$u = r + \rho P u. \tag{1.35}$$

1.6 ABSORBING STATE PROBLEMS

Consider the example on p. 11 with the reward function

$$r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and with no discounting (i.e. $\rho = 1$). As soon as the system enters state $i = 2$ it stays there. State $i = 2$ is an absorbing state. We assume that the process terminates on absorption.

Let us now look at the expected total reward to absorption. The analysis is similar to that of (1.1)–(1.11) with

$$P = \begin{bmatrix} 0.75 & 0.25 \\ 0 & 1 \end{bmatrix}, \qquad r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Here v_n is now the expected total reward function for n years up to absorption at some time during those n years.

Equation (1.9) still holds. Then

$$(U - Pz)^{-1} = \begin{bmatrix} 1 - 0.75z & -0.25z \\ 0 & 1 - z \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 1 - z & 0.25z \\ 0 & 1 - 0.75z \end{bmatrix} \Big/ (1 - 0.75z)(1 - z)$$

$$= \begin{bmatrix} 1 - z & 0.25z \\ 0 & 1 - 0.75z \end{bmatrix} \left(\frac{4}{1 - z} - \frac{3}{1 - 0.75z} \right)$$

$$= \text{(after some manipulation)}$$

$$(1/(1 - z)) \begin{bmatrix} 0 & 3 \\ 0 & 3 \end{bmatrix} + (1/(1 - 0.75z)) \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}.$$

Then, with

$$r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$f(z) = (z/(1 - z)(1 - 0.75z)) \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= (1/(1 - z)) \begin{bmatrix} 4 \\ 0 \end{bmatrix} - (1/(1 - 0.75z)) \begin{bmatrix} 4 \\ 0 \end{bmatrix}.$$

Hence, taking the coefficient of z^n in $f(z)$, we have

$$\underline{n \geqslant 0} \qquad v_n = 4(1 - (0.75)^n) \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

1.7 ASYMPTOTIC BEHAVIOUR OF EXPECTED TOTAL REWARD TO ABSORPTION

We see that

$$\lim_{n \to \infty} [v_n] = v \tag{1.36}$$

where, in this example,

$$v = \begin{bmatrix} 4 \\ 0 \end{bmatrix}.$$

We make the following notes:

(i) For problems which have a limiting probability of 1 to absorption, $\lim_{n \to \infty} [v_n]$ exists (this is implicit in Mine and Osaki [34] pp. 42–43; Derman [15], Theorem 1, gives a proof; White [58], Theorem 1.9, gives a slightly weaker result).

(ii) Using (1.7) and (1.36), v satisfies uniquely (under the conditions in (*i*))

$$u = r + Pu, \tag{1.37}$$

$$u(2) = 0. \tag{1.38}$$

(iii) We have assumed that the process terminates on absorption. If, however, the process continues in the absorbed state, the results are merely those for the expected total reward case and, if the reward in the absorbed state is not 0, (1.36) will not hold (unless we allow $v = \infty$).

(iv) It is possible to use discounting in the absorbing state problem.

1.8 SOME ILLUSTRATIONS

(a) Inventory

We make the following assumptions:

(i) there is a single commodity to be supplied;
(ii) the stock level at the beginning of each time unit is i, $1 \leqslant i \leqslant m$;
(iii) the cost of ordering y units of stock is $c(y)$;
(iv) no backlogs are allowed;

(v) the cost of shortage y is $l(y)$;
(vi) the stockholding cost per unit held per time unit is a;
(vii) the stock reorder rule is, for a specified critical level k,

$$\begin{aligned} &\text{if } i \leqslant k \leqslant m, \text{ order quantity } k - i, \\ &\text{if } i > k, \text{ order quantity } 0; \end{aligned}$$

(viii) the probability that the demand in a time unit is s is $q(s)$, $0 \leqslant s \leqslant \bar{s}$, and the demand is identically and independently distributed in each time unit;
(ix) the demand in any time unit is satisfied from stock on hand plus any stock ordered at the beginning of that time unit;
(x) orders are placed only at the beginning of each time unit;
(xi) it is required to find, for each value of k, the expected total cost function over n time units.

Thus we may use the standard formulations in terms of $\{v_n\}$, where v_n is the expected total reward function for n time units. We simply identify the forms of the parameters $\{p_{ij}^k\}$, $\{r_{ij}^k\}$ and $\{r_i^k\}$, noting that, in order to conform with our definition of rewards, rewards are the negatives of costs.

For the stockholding costs we use an approximation involving one-half of the sum of the stock level immediately after ordering and the stock level immediately preceding the next order point (i.e. the beginning of the next time unit).

For $0 \leqslant i, j \leqslant m$ we have the following identifications:

$$\begin{aligned} j &= \max[k, i] - s \quad \text{if } s \leqslant \max[k, i] \\ &= \quad 0 \qquad\quad \text{if } s > \max[k, i], \end{aligned}$$

$$p_{ij}^k = \begin{Bmatrix} 0 & \text{if } j > \max[k, i] \\ q(\max[k, i] - j) & \text{if } 0 < j \leqslant \max[k, i] \\ \sum_{s \geqslant \max[k, i]} q(s) & \text{if } j = 0 \end{Bmatrix},$$

$$r_i^k = -\left(c(\max[k - i, 0]) + \sum_{s > k}' q(s) l(s - k) + \tfrac{1}{2} a\left(\max[k, i] + \sum_{s < \max[k, i]} q(s)(\max[k, i] - s)\right)\right).$$

The $\{r_{ij}^k\}$ are automatically incorporated into $\{r_i^k\}$. To record $\{r_{ij}^k\}$ explicitly we need to allow j to take nominal negative values. Then

$$r_{ij}^k = -(c(\max[k-i,0]) + l(\max[-j,0]) + \tfrac{1}{2}a(2\max[k,i] + \max[j,0])).$$

(b) Queuing

We make the following assumptions:

(i) there is a single arrival stream and a single service facility;
(ii) at most one customer can arrive in a given time unit, the arrival probability is p, and arrivals are independently and identically distributed in each time unit;
(iii) at most one customer can be served in a given time unit, when the server is occupied, and the probability of the service being completed in that time unit is q, with services being independently and identically distributed in each time unit;
(iv) arriving customers arrive at the ends of time units, and completed services are completed at the ends of time units;
(v) the decision rule is to send all customers, in excess of k in the system, for service elsewhere, where k is a prespecified critical level;
(vi) the cost of sending y customers elsewhere is $c(y)$;
(vii) the cost of one customer waiting in the system for one time unit is a;
(viii) it is required to find, for each value of k, the expected total cost function over n time units.

Again we may use the standard formulations in terms of $\{v_n\}$, where v_n is the expected total reward function for n time units, and again rewards are the negatives of costs.

We now identify $\{p_{ij}^k\}$ and $\{r_{ij}^k\}$, $\{r_i^k\}$. For $1 \leqslant i, j \leqslant k+1$, we have the following identifications. If $\min[k,i] > 0$ then

$$j = \text{minimum}[k,i] + \xi - \eta$$

where

$$\xi = \begin{cases} 1 & \text{if an arrival occurs} \\ 0 & \text{otherwise} \end{cases},$$

$$\eta = \begin{cases} 1 & \text{if a service occurs} \\ 0 & \text{otherwise} \end{cases}.$$

If minimum$[k, i] = 0$ then

$$j = \xi.$$

If $\min[k, i] > 0$ then

$$p_{ij}^k = \begin{Bmatrix} (1-p)q & \text{if } j = \text{minimum}[i,k] - 1 \\ pq + (1-p)(1-q) & \text{if } j = \text{minimum}[i,k] \\ p(1-q) & \text{if } j = \text{minimum}[i,k] + 1 \\ 0 & \text{otherwise} \end{Bmatrix}.$$

If $\min[k, i] = 0$ then

$$p_{ij}^k = \begin{Bmatrix} (1-p) & \text{if } j = 0 \\ p & \text{if } j = 1 \\ 0 & \text{otherwise} \end{Bmatrix}.$$

Finally

$$r_i^k = -(c(\text{maximum}[i-k], 0) + a \text{ minimum}[i, k]).$$

We see that $r_{ij}^k = r_i^k$ independently of j.

1.9 SELECTING DECISION RULES

Throughout this chapter we have considered only one decision rule. The objective of Markov decision process theory is to provide frameworks for finding a best decision rule from among a given set. For our toymaker example, Howard [23], p. 28, gives four decision rules with the $\{p_{ij}^k\}\{r_{ij}^k\}, \{r_i^k\}$ given in Table 1.8.

We have four decision rules, which we call $\delta^1, \delta^2, \delta^3, \delta^4$. These are given in Table 1.9.

Let us now look at the infinite horizon solutions. These are given in Tables 1.10 and 1.11 for the average expected reward per year and for the expected total discounted reward cases respectively.

In Table 1.10 we have replaced w by u and used (1.25) with u in place of w, setting $u(2) = 0$. The actual function w differs, for each decision rule, from u by a function $(w(2) - u(2))e$, where e is the unit function. To find w we would need to use the z-transform analysis for each decision rule $g = h$ in (1.25).

Table 1.11 is computed for $\rho = 0.9$ using (1.35) with $u = v$. In Table 1.10 δ^4 produces the maximal average expected reward per year, with $g = 2$ independently of the initial state i. In Table 1.11 δ^4 produces

Table 1.8 Complete data for toymaker problem (reproduced from [23] Howard (1960), p. 28, by permission of The MIT Press)

State i	*Action* k	*Transition probability* p_{ij}^k 1	2	*Reward* r_{ij}^k 1	2	*Expected reward* r_i^k
1 (successful toy)	1 (no advertising)	0.5	0.5	9	3	6
	2 (advertising)	0.8	0.2	4	4	4
2 (unsuccessful toy)	1 (no research)	0.4	0.6	3	−7	−3
	2 (research)	0.7	0.3	1	−19	−5

Table 1.9 Decision rules

Decision rule	*State* i 1	2
δ^1	$k=1$	$k=1$
δ^2	$k=1$	$k=2$
δ^3	$k=2$	$k=1$
δ^4	$k=2$	$k=2$

Table 1.10 Average expected reward per year

Decision rule	g	$u(1)$	$u(2)$
δ^1	1	10	0
δ^2	$1\frac{5}{12}$	$9\frac{1}{6}$	0
δ^3	$1\frac{2}{3}$	$11\frac{2}{3}$	0
δ^4	2	10	0

Table 1.11 Expected total discounted reward

Decision rule	$v(1)$	$v(2)$
δ^1	15.5	5.6
δ^2	18.3	9.1
δ^3	20.0	8.9
δ^4	22.2	12.3

the maximal expected total discounted reward. The expected total discounted reward depends upon the initial state.

1.10 EXERCISES TO CHAPTER 1

1. Use the z-transform method to find, in general form, v_n, the expected total reward function for n time units, for the following problem data with $v_0 = 0$:

State i	*Transition probability* p_{ij} 1	2	*Reward* r_{ij} 1	2
1	0.8	0.2	4	4
2	0.7	0.3	1	−19

2. Repeat Exercise 1 for the discounted problem with $\rho = 0.9$.
3. In Exercises 1 and 2 it is assumed that there is no terminal value at the end of the nth time unit. Suppose now that there is an expected terminal value function, if the process terminates after n time units, of $\begin{bmatrix} 10 \\ 1 \end{bmatrix}$. Now find v_n for the conditions of Exercises 1 and 2.
4. Construct a simple three-state example for which, in the expression for v_n given by (1.12), g is a function with not all components equal (i.e. the gain depends upon the state), and ε_n does not tend to zero as n tends to ∞. By selecting the data carefully you may use the function form v_n given by (1.5) without having to use z-transforms.

5. Derive the results given in Tables 1.10 and 1.11 for policy δ^4.
6. (a) Prove, first of all, the result for v_n given in (1.26); (b) then assume that $\lim_{n \to \infty} [P^n] = \tilde{P}$ exists, and prove the result for ε_n given in (1.21).

   Note: $\lim_{n \to \infty} [P^n] = \tilde{P}$ means that $P^n = \tilde{P} + E_n$ where E_n is a matrix tending to 0 as n tends to ∞.
7. Assuming that the Cesàro limit $P^* = \lim_{n \to \infty} \left(\sum_{l=0}^{n-1} P^l / n \right)$ exists, prove (1.17).

CHAPTER 2

A general framework for Markov decision processes

2.1 PRELIMINARY REMARKS

The development of a general framework for Markov decision processes may be found in the texts of Derman [15] and of van der Wal [53]. The reader should be aware of notational differences.

In what follows we will assume that the state and action sets are finite. This is no real practical restriction of the use of Markov decision process models, but it does enable simpler proofs to be provided of some fundamental results. We will later (see p. 131) deal with problems with infinite state spaces and/or action spaces, but will, in doing so, take certain results for granted. We will also assume that decisions are taken at the beginning of each time unit in much of what we do. Later (see p. 116) we will look at problems involving variable decision intervals, but again taking some results for granted.

In Chapter 1, for introductory purposes, we have assumed that all the components of the problem are independent of time. In this chapter we will allow for some dependence on time. In addition, in Chapter 1 we have assumed that decision rules are specified in terms of the current state only, and not in terms of the history of the system to date. This is quite correct for what we have done, and is intuitively obvious. However, this requires demonstrating rigorously. In addition, for some situations such as those where the variance of behaviour is important, history-based decision rules become relevant.

2.2 THE GENERAL FRAMEWORK

Our framework is as follows:

(i) A system occupies one of a finite set of states at the beginning of each of a set of time units, which we label $t = 1, 2, ., ., .$ moving forward

from some time origin. The reader should not confuse the chronological time t and time n from the system's time horizon as defined in Chapter 1.

(ii) We designate the state set by I_t to denote the possible dependence on t for each value of t. In the stationary case we will replace I_t by I.

(iii) The random variable X_t will denote the state at the beginning of time unit t, and realised values will be denoted by i_t for each value of t.

(iv) For each state $X_t = i_t$ there is a feasible action space $K_t(i_t)$, where $K_t(i_t)$ is finite for each value of t. In the stationary case we will replace K_t by K.

(v) The random variable Z_t will denote the action taken at the beginning of time unit t, and realised values of Z_t will be denoted by k_t for each value of t.

(vi) In each time unit t there will be a reward. Costs or penalties may be seen as negative rewards.

(vii) The random variable Y_t will be used to denote the reward in time unit t, and realised values will be denoted by l_t, for each value of t. Thus each time unit t will be characterised by a triple (X_t, Z_t, Y_t) with realised values (i_t, k_t, l_t).

(viii) The history of the system up to the beginning of time unit t is a random variable H_t given as follows for each value of t (noting that Y_t is determined by X_t, Z_t, X_{t+1}):

$$\underline{t \geqslant 2} \qquad H_t = (X_1, Z_1, X_2, Z_2, ., ., . X_{t-1}, Z_{t-1}, X_t), \tag{2.1}$$

$$\underline{t = 1} \qquad H_1 = (X_1). \tag{2.2}$$

(ix) A realised history will be denoted by h_t for each value of t. Thus

$$\underline{t \geqslant 2} \qquad h_t = (i_1, k_1, i_2, k_2, ., ., . i_{t-1}, k_{t-1}, i_t), \tag{2.3}$$

$$\underline{t = 1} \qquad h_1 = (i_1). \tag{2.4}$$

We will now define a decision rule and a policy. These are a matter of convention. The definitions differ in different texts, and the reader should be aware of this. The modern convention (e.g. see Derman [15], p. 3) is the one adopted in this text, and is the one which Howard [23] uses. The distinction between policy and decision rule is purely conventional in the case of infinite horizon stationary processes with average expected reward or expected total discounted reward criteria, where a policy may be a repetition of a single decision rule. In such cases (e.g. see White [58], p. xii) the distinction is a little blurred.

(x) A decision rule δ_t, for time unit t, determines a probability distribution of actions over $K_t(i_t)$ when the history $H_t = h_t$ is known.

For almost all of our problems we will only require, for each value of t, that a specific action $Z_t = k_t$ be taken with probability 1 for a given $H_t = h_t$, but we leave the framework general at this stage.

Note that, for each value of t, K_t is specified as a function of i_t and not of h_t. If K_t is required to be a function at h_t this can be incorporated by redefining the concept of a state so that i_t is replaced by h_t. In this case some of the subsequent analysis in this chapter is not needed because decision rules are then automatically Markov in terms of h_t (see (xv)).

(xi) A policy is a sequence of decision rules. A policy is denoted by π and takes the form

$$\pi = (\delta_1, \delta_2, ., ., ., ., .\delta_t, ., .). \tag{2.5}$$

We will use π both for infinite and for finite horizon problems. To obtain a finite horizon policy from an infinite horizon policy we merely restrict its operation to a finite number of time units.

A policy tells us how to determine actions for any time unit of the process.

(xii) Let $\rho(t)$ be the discount factor for time unit t with, by convention, $\rho(0) = 1$, and where $\rho(t) \geqslant 0$.

(xiii) Let R_n be the total discounted random reward over the first n time units, starting at chronological time unit $t = 1$. Then

$$R_n = \sum_{t=1}^{n} \left(\prod_{s=0}^{s=t-1} \rho(s) \right) Y_t. \tag{2.6}$$

It is assumed that the reward in a given time unit is received at the beginning of that time unit. Also, for finite n, the total reward is given by setting $\rho(s) = 1$, $0 \leqslant s \leqslant n-1$, in (2.6).

(xiv) Let $v_n^{\pi}(i)$ be the expected total discounted reward over the next n time units if $X_1 = i$ and we use policy π, again noting that we begin at chronological time $t = 1$; $v_n^{\pi}(\cdot)$ is called a value function (or vector).

Note that, because we begin with $t = 1$, then $H_1 = (X_1) = (h_1) = (i)$. Strictly speaking we should define $v_{in}^{\pi}(h_t)$ to be the expected total discounted reward over the next $(n - t + 1)$ time units, using policy π, beginning with time unit t, and given the history h_t at this time. In this case π would have to be defined as a sequence of decision rules $(\delta_t, \delta_{t+1}, ., ., .)$. We wish to avoid the complications which this would produce but the reader should be aware of this. What we have chosen

to use is all we need for most of our purposes. We are, in effect, starting our process at a point where there is no history except the current state i. It is easy to carry out the subsequent analysis if we begin with $H_1 = h_1$ for a specified h_1. This point is discussed on p. 36. Also, some consideration of the variable starting time unit t will be given on p. 40.

We assume that π is such that $v_n^\pi(i)$ exists. For much of what we do this will clearly be the case. For more general policies we need to involve measurability assumptions. We mention this for housekeeping purposes only. Let us now look at possible policy sets.

(xv) Let Π be the set of all (measurable) policies (in effect this simply requires that all of our specified expected rewards, discounted or otherwise, exist); Π_M be the subset of Π which are Markov, i.e. if $\pi = (\delta_1, ., ., \delta_t, ., .)$ then δ_t is a function of i_t only and not of the whole history h_t; Π_S be the subset of Π_M which are stationary, i.e. if $\pi = (\delta_1, \delta_2, ., ., \delta_t, ., .)$ then $\delta_t = \delta$ for some δ, and we write such a policy as $\pi = (\delta)^\infty$; Π_D be the subset of Π_S which are deterministic, i.e. if $\pi = (\delta)^\infty$ then for each $i_t \in I_t$ $\exists k_t \in K_t(i_t)$ such that probability $(Z_t = k_t \mid X_t = i_t) = 1$; Π_{MD} be the subset of Π_M which are deterministic. Note that $\pi \in \Pi_{MD}$ is not necessarily stationary.

With our definitions we have

$$\begin{aligned} \Pi_D \subseteq \Pi_S \subseteq \Pi_M \subseteq \Pi, \\ \Pi_{MD} \subseteq \Pi_M. \end{aligned} \tag{2.7}$$

We will also let Δ_t^* be the set of all decision rules based on I_t at time unit t, and $K_t^*(i)$ be the set of probability distributions over actions in $K_t(i)$. Finally, we will let Δ_t be the set of all non-probabilistic decision rules based on I_t at time unit t. We will drop the suffix t in the stationary case.

Later (see p. 42) we will show that, for infinite horizon stationary problems (i.e. those for which rewards, transition probabilities and discount factors are independent of t), we need only consider Π_D for certain classes of optimality criteria. For finite horizon problems, optimal policies may not be stationary for these criteria.

Let us now turn to the question of optimality criteria.

(xvi) For finite n we wish to find

$$v_n(i) = \operatorname*{supremum}_{\pi \in \Pi} [v_n^\pi(i)], \quad \forall\ i \in I. \tag{2.8}$$

We make the following notes:

(a) The supremum $\bar{a}$, of a set of scalar quantities $\{a^\pi\}$ is given by

$$a^\pi \leqslant \bar{a}, \quad \forall\ \pi \in \Pi \tag{2.9}$$

and, given $\varepsilon > 0$, $\exists\ \pi \in \Pi$ such that

$$\bar{a} - \varepsilon \leqslant a^\pi \leqslant \bar{a} \tag{2.10}$$

where $\bar{a}$ is a least upper bound of $\{a^\pi\}$.
Eventually we will replace 'supremum' by 'maximum', but not all sequences have maxima in general, e.g. if $\tilde{\Pi} = \{\pi^s\}$, $1 \leqslant s < \infty$, is a sequence of policies with $a^{\pi^s} = 1 - (\frac{1}{2})^s$, then

$$\underset{\pi \in \tilde{\Pi}}{\text{supremum}}\ [a^{\pi^s}] = 1$$

but no policy $\pi \in \tilde{\Pi}$ exists with $a^\pi = 1$.

(b) For some problems we require infimum/minimum instead of supremum/maximum, where infimum is a greatest lower bound defined analogously to (2.9) and (2.10).

(c) Equation (2.8) applies for the non-discounted case ($\rho(t) = 1$, $\forall\ t$), and for cases where we might have $\rho(t) > 1$, when n is finite.

(xvii) For infinite n we differentiate between the discounted case, the non-discounted case and the absorbing states case, which we now discuss.

(a) *Discounting.* In this case we assume that there is a $\rho < 1$ such that

$$\rho(t) \leqslant \rho < 1, \quad \forall\ t \geqslant 1. \tag{2.11}$$

Let

$$v^\pi(i) = \underset{n \to \infty}{\text{limit}}\ [v_n^\pi(i)] \tag{2.12}$$

where v^π depends on $\{\rho(t)\}$ which we will suppress for notational convenience. This limit will exist, given our measurability assumptions. All we need is that $E(Y_t \mid X_1 = i, \text{ policy } \pi)$ exists, where E is the expectation operator.

We now wish to find

$$v(i) = \underset{\pi \in \Pi}{\text{supremum}}\ [v^\pi(i)]. \tag{2.13}$$

Again, later (see p. 41), we will replace 'supremum' by 'maximum'. For stationary problems we write $\rho(t) = \rho$ for all t.

(b) *Non-discounting.* In this case

$$\rho(t) = 1, \quad \forall\, t \geqslant 1. \tag{2.14}$$

We now need to be careful because $\{v_n^\pi(i)\}$ can increase or decrease without bound as n tends to ∞. We therefore replace $\{v_n^\pi(i)\}$ by $\{g_n^\pi(i)\}$ where

$$g_n^\pi(i) = v_n^\pi(i)/n, \quad \forall\, i \in I. \tag{2.15}$$

We also need to be even more careful because $\lim_{n\to\infty} [g_n^\pi(i)]$ need not exist for general Markov decision processes. It will exist for stationary policies in finite action, finite state situations, however, to which we restrict much of this text.

Consider, for example, the following problem where the policy π produces the specified deterministic rewards in each time unit. Figure 2.1 illustrates the problem, where the time units are grouped in ever-increasing groups indicated by q.

There is a single state $i = 1$. The following are the $\{g_n^\pi(1)\}$ for $1 \leqslant n \leqslant 9$:

$$\begin{aligned} &g_1^\pi(1) = \tfrac{1}{2}, \quad && g_2^\pi(1) = -\tfrac{1}{4}, \quad && g_3^\pi(1) = -\tfrac{1}{2} \\ &g_4^\pi(1) = -\tfrac{1}{8}, \quad && g_5^\pi(1) = \tfrac{1}{10}, \quad && g_6^\pi(1) = \tfrac{1}{4} \\ &g_7^\pi(1) = \tfrac{5}{12}, \quad && g_8^\pi(1) = \tfrac{7}{16}, \quad && g_9^\pi(1) = \tfrac{1}{2}. \end{aligned}$$

In the general case, for any t we have the following specification:

$q \geqslant 1$ If $3^{q-1} < t \leqslant 3^q$, then set

$$\begin{aligned} Y_t &= 1 \quad \text{if } q \text{ is even,} \\ &= -1 \quad \text{if } q \text{ is odd.} \end{aligned}$$

Set

$$Y_1 = \tfrac{1}{2} \text{ (set } 3^{q-1} = 0 \text{ for } q = 0).$$

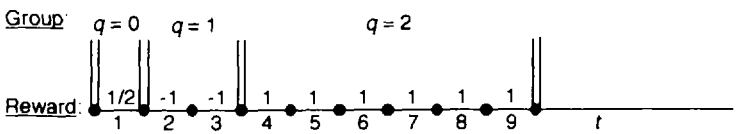

Figure 2.1 Time behaviour for example

Then

$$\text{for } n = 3^q,\ q \text{ even}, \qquad g_n^\pi(1) = \tfrac{1}{2}$$
$$\text{for } n = 3^q,\ q \text{ odd}, \qquad g_n^\pi(1) = -\tfrac{1}{2}$$

and

$$-\tfrac{1}{2} \leqslant g_n^\pi(1) \leqslant \tfrac{1}{2}, \quad \forall\ n.$$

Diagrammatically (drawn as a continuous approximation) $g_n^\pi(1)$ takes the form given in Figure 2.2.

The upper bounding line is the limit supremum line with limit $\underset{n\to\infty}{\text{supremum}}\ [g_n^\pi(1)] = \frac{1}{2}$. The lower bounding line is the limit infimum line with limit $\underset{n\to\infty}{\text{infimum}}\ [g_n^\pi(1)] = -\frac{1}{2}$; $\underset{n\to\infty}{\text{limit}}\ [g_n^\pi(1)]$ does not exist.

Here limit $\underset{n\to\infty}{\text{infimum}}$ is the limiting worst reward per unit time as n tends to ∞. When we want to make this as high as possible we use as our criterion, to be made as large as possible

$$g^\pi(i) = \underset{n\to\infty}{\text{limit infimum}}\ [g_n^\pi(i)], \quad \forall\ i \in I. \tag{2.16}$$

Similarly, if we want to make the best reward per unit time as high as possible, we use as our criterion, to be made as large as possible,

$$g^\pi(i) = \underset{n\to\infty}{\text{limit supremum}}\ [g_n^\pi(i)], \quad \forall\ i \in I. \tag{2.17}$$

For most of what we will do limits will exist and then (see Bromwich [8], p. 18)

$$g^\pi(i) = \underset{n\to\infty}{\text{limit}}\ [g_n^\pi(i)]$$

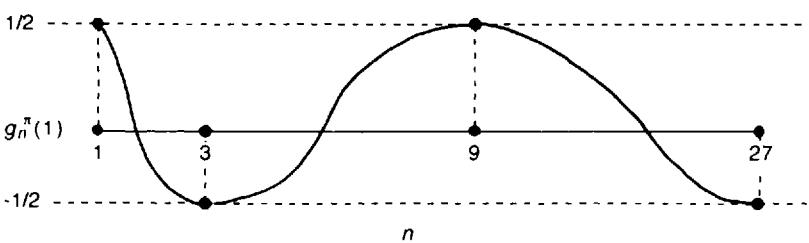

Figure 2.2 Continuous approximation of system behaviour

$$= \operatorname*{limit\ infimum}_{n \to \infty} [g_n^\pi(i)]$$

$$= \operatorname*{limit\ supremum}_{n \to \infty} [g_n^\pi(i)], \quad \forall\, i \in I. \tag{2.18}$$

We then want to find

$$g(i) = \operatorname*{supremum}_{\pi \in \Pi} [g^\pi(i)], \quad \forall\, i \in I. \tag{2.19}$$

In much of what we will do we may replace 'supremum' by 'maximum'.

(c) *Absorbing states.* In this case (see our example on p. 15) the process terminates once a state (say state m) has been entered or, more generally, once a set of states (say I_a) has been entered. Here $v^\pi(i)$ takes the form of (2.12) where we may, if we wish, include discount factors $\rho(t) \leqslant \rho < 1$.

Under certain conditions (e.g. see Mine and Osaki [34], p. 42, Derman [15], pp. 53–54) $v^\pi(i)$ will be well defined and finite for all i. We wish to find

$$v(i) = \operatorname*{supremum}_{\pi \in \Pi} [v^\pi(i)], \quad \forall\, i \in I. \tag{2.20}$$

In most of what we will do we may replace 'supremum' by 'maximum'. We have a boundary condition

$$v(i) = 0, \quad \forall\, i \in I_a. \tag{2.21}$$

It is to be noted that all of our optimisation problems, (2.8), (2.13), (2.19), involve the optimisation of some variant of expected reward in some form. As a result of this we will be able to reduce our problems to optimisation problems over the policy space Π_M, and even further in some cases.

There are, however, some optimisation problems in which history-remembering policies in Π should be used to avoid suboptimal solutions. For example consider the following problem over three time units, where the objective is to minimise the variance of the total reward. The definitions of $\{p_{ij}^k(t)\}$, $\{r_{ij}^k(t)\}$ will be found in (2.22) and (2.24) respectively.

$$I_t = \{1, 2, 3, 4\}, \qquad 1 \leqslant t \leqslant 3,$$
$$K_1(1) = \{1\},\ K_2(2) = K_2(3) = \{1\},$$
$$K_3(4) = \{1, 2\} \text{ (we do not need } K_4(4)).$$

$$p^1_{12}(1) = 0.5, \quad p^1_{13}(1) = 0.5, \quad p^1_{24}(2) = 1, \quad p^1_{34}(2) = 1,$$
$$p^k_{44}(3) = 1, \quad k = 1, 2,$$
$$r^1_{12}(1) = 1, \quad r^1_{13} = -1, \quad r^1_{24}(2) = r^1_{34}(3) = 0,$$
$$r^1_{44}(3) = 1, \quad r^2_{44}(3) = -1.$$

This is shown diagrammatically in Figure 2.3.

δ_1 is fixed, viz. $\delta_1(1) = 1$.
δ_2 is fixed, viz. $\delta_2(2) = \delta_2(3) = 1$.

If δ_3 is Markov (non-history remembering) then

$$\delta_3 = \delta^1_3(\delta^1_3(4) = 1) \quad \text{or} \quad \delta_3 = \delta^2_3(\delta^2_3(4) = 2).$$

If $\pi^1 = (\delta_1, \delta_2, \delta^1_3)$ then variance $(R^{\pi^1}_3) = 1$ (see (2.6) with $\rho(s) = 1$). If $\pi^2 = (\delta_1, \delta_2, \delta^2_3)$ then variance $(R^{\pi^2}_3) = 1$.

Now let $\pi^3 = (\delta_1, \delta_2, \delta^3_3)$ where δ^3_3 is history remembering and $\delta^3_3(i_1, k_1, i_2, k_2, i_3)$ is given by

$$\delta^3_3(1, 1, 2, 1, 4) = 2 \qquad \delta^3_3(1, 1, 3, 1, 4) = 1.$$

Then $R^{\pi^3}_3 = 0$ always and

$$\text{variance}(R^{\pi^3}_3) = 0.$$

Clearly, in general, knowledge of the history of a process will enable future actions to be taken to compensate for earlier rewards if reward variation is of importance. We will return to this later (see p. 64).

With the implicit exception of our introductory example on p. 1 we have not yet formally defined our Markov property. Nor have we defined the terms which we will later use. We will now do this.

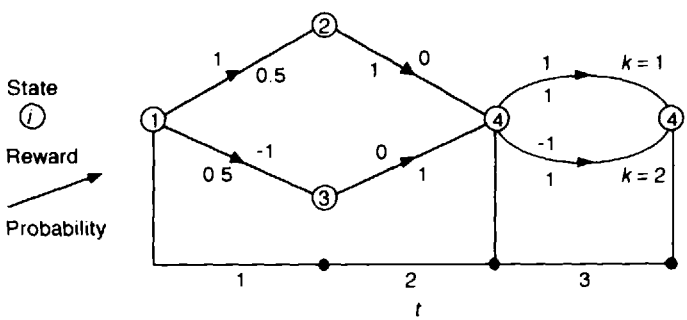

Figure 2.3 Possible time behaviour for example

(xviii) Our system will be said to have the Markov property with respect to the state structure $\{X_t\}$ if

$$\text{probability}\,(X_{t+1} = j \mid H_t = h, Z_t = k \in K_t(i)) = \text{probability}(X_{t+1} = j \mid X_t = i, Z_t = k \in K_t(i)). \quad (2.22)$$

We denote the right-hand side of (2.22) by $p_{ij}^k(t)$. For stationary problems we write this as p_{ij}^k. Thus the state transitions depend upon i_t and k_t only and not on other aspects of the history. The Markov property is relative to the state structure $\{X_t\}$.

Consider the inventory example on p. 17. Suppose now that the random demand Ω_t in time unit t depends upon the demand in the previous time unit so that, for $t \geqslant 2$

$$\text{probability}(\Omega_t = s \mid \Omega_{t-1} = r) = q(s, r). \quad (2.23)$$

Then the process is not Markov in terms of the stock level at the beginning of the time unit but it is Markov in terms of stock level at the beginning of the time unit, and the previous demand level, combined.

Finally we use the following notational conventions (see (viii)):

$$Y_t \mid (X_t = i, \quad Z_t = k \in K_t(i), \quad X_{t+1} = j) = r_{ij}^k(t). \quad (2.24)$$

$$E(Y_t \mid X_t = i, Z_t = k \in K_t(i)) = \sum_{j \in I_{t+1}} p_{ij}^k(t) r_{ij}^k(t) = r_i^k(t). \quad (2.25)$$

In the case of stationary problems we write these as r_{ij}^k and r_i^k respectively.

2.3 FINITE HORIZON MARKOV DECISION PROCESSES

2.3.1 STATIONARY CASE

For the moment let us assume that our processes are stationary with parameters $\{\{K(i)\}, I, \{p_{ij}^k\}, \{r_{ij}^k\}, \{r_i^k\}, \rho\}$ independent of t. We will treat the discounted case ($\rho < 1$) and the non-discounted case ($\rho = 1$) together because, for finite horizon problems, these pose no special problems. Our objective is to find (see (2.8))

$$v_n(i) = \underset{\pi \in \Pi}{\text{supremum}}\, [v_n^\pi(i)], \quad \forall\, i \in I. \quad (2.26)$$

Eventually we will show that we may replace Π by Π_D (see p. 42) without loss. In order to do this we need a result which says that, for the criterion which we are currently considering, we may keep, in the first instance, to Π_M (see p. 36) without loss. This result is a special case of Lemma 2.1, van der Wal [53]. We need to use the concept of randomised actions at this stage.

Result 2.1. Let $X_1 = i_1$ be fixed and, for each $\pi \in \Pi$, let $x_i^{\pi k}(t)$ be the probability that, at the beginning of time unit t, the system will be in state $i \in I$ and action $k \in K(i)$ will be taken (conditional on π and on $X_1 = i_1$, where i_1 is suppressed but understood).

Let

$$x_i^{\pi}(t) = \sum_{k \in K(i)} x_i^{\pi k}(t), \quad \forall\, t \geqslant 1, \quad i \in I. \tag{2.27}$$

Let us define a policy $\tau \in \Pi_M$ as follows, where $t \geqslant 1$ and the argument t is used in conformity with (2.5):

$$\tau = (\tau_1, \tau_2, ., ., ., ., \tau_t, ., .) \tag{2.28}$$

with $\{\tau_t\}$ given by (2.29) following. Also $\{x_i^{\tau k}(t), x_i^{\tau}(t)\}$ are defined as for $\{x_i^{\pi k}(t), x_i^{\pi}(t)\}$ when τ replaces π.

At the beginning of time unit t, if the system is in state $i \in I$ then action $k \in K(i)$ is taken with probability

$$\begin{aligned} \tau_{ik}(t) &= x_i^{\pi k}(t) \Big/ \left(\sum_{a \in K(i)} x_i^{\pi a}(t) \right) \\ &= x_i^{\pi k}(t) / x_i^{\pi}(t) \end{aligned} \tag{2.29}$$

if the denominator in (2.29) $\neq 0$,

$$\tau_{ik}(t) \text{ is arbitrary otherwise.} \tag{2.30}$$

Then

$$x_i^{\tau k}(t) = x_i^{\pi k}(t), \quad \forall\, t \geqslant 1,\ i \in I,\ k \in K(i), \tag{2.31}$$

$$x_i^{\tau}(t) = x_i^{\pi}(t), \quad \forall\, t \geqslant 1,\ i \in I. \tag{2.32}$$

Equation (2.29) simply gives, for policy τ, the conditional probability of taking action k given state i at the beginning of time unit t, and the denominator $x_i^{\pi}(t)$ is the probability of being in state i at the beginning of time unit t, given policy π and $X_1 = i_1$.

Proof. We prove by induction on t. Equation (2.31) is clearly true for $t = 1$ because

$$\tau_{i_1 k}(1) = x_{i_1}^{\pi k}(1)/1 = x_{i_1}^{\pi k}(1), \tag{2.33}$$

$$\tau_{ik}(1) \text{ is arbitrary if } i \neq i_1. \tag{2.34}$$

Then, because probability $(X_1 = i_1) = 1$,

$$x_{i_1}^{\tau k}(1) = \tau_{i_1 k}(1) = x_{i_1}^{\pi k}(1) \tag{2.35}$$

and, because probability $(X_1 \neq i_1) = 0$,

$$x_i^{\tau k}(1) = 0 = x_i^{\pi k}(1) \quad \text{if } i \neq i_1. \tag{2.36}$$

Also (2.32) is then true for $t = 1$.

Assume that (2.31) is true for $1 \leqslant t \leqslant s$. Then (2.32) is also true for $1 \leqslant t \leqslant s$. We now prove that (2.31) (and hence (2.32)) are true for $t = s + 1$. If $x_i^{\pi}(s+1) \neq 0$ (required in (2.38)) then

$$x_i^{\tau k}(s+1) = \sum_{j \in I, a \in K(j)} x_j^{\tau a}(s) p_{ji}^a \tau_{ik}(s+1) \tag{2.37}$$

$$= \sum_{j \in I, a \in K(j)} x_j^{\pi a}(s) p_{ji}^a x_i^{\pi k}(s+1)/x_i^{\pi}(s+1) \tag{2.38}$$

$$= x_i^{\pi}(s+1) x_i^{\pi k}(s+1)/x_i^{\pi}(s+1). \tag{2.39}$$

The latter equality holds because

$$x_i^{\pi}(s+1) = \sum_{j \in I, a \in K(j)} x_j^{\pi a}(s) p_{ji}^a. \tag{2.40}$$

Thus

$$x_i^{\tau k}(s+1) = x_i^{\pi k}(s+1), \quad \forall\ k \in K(i). \tag{2.41}$$

If $x_i^{\pi}(s+1) = 0$ then from (2.37) and (2.40) we have $x_i^{\pi k}(s+1) = 0$ and hence

$$x_i^{\tau k}(s+1) = 0, \quad \forall\ k \in K(i). \tag{2.42}$$

Thus, again we have

$$x_i^{\tau k}(s+1) = x_i^{\pi k}(s+1) = 0, \quad \forall\ k \in K(i). \tag{2.43}$$

○

At this point we refer back to our comment made on p. 27 concerning the definition of v_n in terms of the current state i_1 for $t = 1$, and not in terms of the current history h_1. It is easily seen that τ may be defined in exactly the way it has been in (2.29) and (2.30) and that (2.31) and (2.32), and hence Result 2.1, will follow even if i_1 is replaced by h_1. Then we have $v_n^{\pi}(h_1) = v_n^{\tau}(h_1)$. However, v_n^{τ} is simply determined by i_1. Thus we need only record i_1 in our analysis.

Let us now look at the n time unit total discounted reward (see (2.6)) for policy τ with $\rho(s) = \rho$ for all s.

$$R_n^{\tau} = \sum_{t=1}^{n} \left(\prod_{s=0}^{s=t-1} \rho(s) \right) Y_t^{\tau}$$

$$= \sum_{t=1}^{n} \rho^{t-1} Y_t^{\tau}. \tag{2.44}$$

Superfix τ is used to denote the dependence of R_n on τ. Then, replacing i_1 by a general i, we have

$$v_n^{\tau}(i) = E(R_n^{\tau} \mid X_1 = i) = \sum_{t=1}^{n} \rho^{t-1} E(Y_t^{\tau} \mid X_1 = i)$$

$$= \sum_{t=1}^{n} \rho^{t-1} \sum_{j \in I, k \in K(j)} x_j^{\tau k}(t) r_j^k, \quad \forall\, i \in I. \tag{2.45}$$

Thus, from (2.31) of Result 2.1, $v_n^{\tau} = v_n^{\pi}$ and we need only keep to Π_M to find any v_n^{π}. We thus have Result 2.2.

Result 2.2. For any policy $\pi \in \Pi$ there is a policy $\tau \in \Pi_M$ with $v_n^{\pi} = v_n^{\tau}$. ○

For a given policy $\pi = (\delta_1, \delta_2, ., ., \delta_t, ., .)$ (1.34) gives

$$\underline{n \geq 1} \qquad v_n^{\pi} = r^{\delta_1} + \rho P^{\delta_1} v_{n-1}^{(\delta_2, \ldots, \delta_t, \ldots)}. \tag{2.46}$$

We can write this in the functional operator form

$$v_n^{\pi} = T^{\delta_1} v_{n-1}^{(\delta_2, \ldots, \delta_t, \ldots)} \tag{2.47}$$

where, for any u: $I \to R$, and any decision rule δ

$$[T^{\delta} u](i) = r_i^{\delta(i)} + \rho [P^{\delta} u]_i, \quad \forall\, i \in I. \tag{2.48}$$

The operator T^{δ} will be used generally in Chapter 3 for algorithms.

Now policy π relates to time units $t = 1, 2, 3, ., ., ., ..$ It may be written as

$$\pi = (\delta_1, \tau) \tag{2.49}$$

where policy τ relates to time units $t = 2, 3, ., ., ..$

In (2.5) we have defined a policy π in such a way that the tth decision rule corresponds to time unit t, and $t = 1$ corresponds to the first time unit in the n time unit horizon. In (2.49) τ relates only to $(\delta_2, \delta_3, ., ., .\delta_t, ., .)$. Thus the first decision rule of τ is δ_2 relating to $t = 2$. Thus, in terms of the definition (2.5), formally $\tau \notin \Pi_M$ in general as we have defined Π_M on p. 27. However, in the stationary case, taking $t = 2$ as the first time unit in the remaining sequence of $(n - 1)$ time units, we may consider Π_M as defined on p. 27 to be independent of time and then $\tau \in \Pi_M$.

Equation (2.47) is better written as (using (2.49))

$$v_n^{\pi} = T^{\delta_1} v_{n-1}^{\tau}. \tag{2.50}$$

Let us assume that there exists a policy $\tau \in \Pi_{MD}$ which is optimal for all $i \in I$ for the residual $n - 1$ time units. This is clearly true if $n - 1 = 1$. We proceed inductively.

Now the choice of δ_1 is equivalent to the choice of $k \in K^*(i)$ (see p. 27) for each $i \in I$. Then, using Result 2.1, and with obvious definitions of rewards and transition probabilities for $\delta \in \Delta^*$, $k \in K^*(i)$, we have

$$\underline{n \geqslant 1} \qquad v_n(i) = \operatorname*{supremum}_{\pi \in \Pi_M} [v_n^{\pi}(i)]$$

$$= \operatorname*{supremum}_{\delta_1(i) \in K^*(i)} \operatorname*{supremum}_{\tau \in \Pi_M} [r_i^{\delta_1(i)} + \rho [P^{\delta_1} v_{n-1}^{\tau}]_i]$$

$$= \operatorname*{supremum}_{k \in K^*(i)} \left[r_i^k + \rho \sum_{j \in I} p_{ij}^k \operatorname*{supremum}_{\tau \in \Pi_M} [v_{n-1}^{\tau}(j)] \right]$$

$$= \operatorname*{supremum}_{k \in K^*(i)} \left[r_i^k + \rho \sum_{j \in I} p_{ij}^k v_{n-1}(j) \right], \quad \forall\, i \in I. \tag{2.51}$$

In (2.51) k is a probability distribution over $K(i)$, i.e. k chooses $a \in K(i)$ with probability k_i^a. Then

$$r_i^k = \sum_{a \in K(i)} k_i^a r_i^a \tag{2.52}$$

and

$$p_{ij}^{k} = \sum_{a \in K(i)} k_i^a p_{ij}^k. \tag{2.53}$$

Then (2.51) may be written as

$$v_n(i) = \underset{k \in K^*(i)}{\text{supremum}} \left[\sum_{a \in K(i)} k_i^a \left[r_i^a + \rho \sum_{j \in I} p_{ij}^a v_{n-1}(j) \right] \right]$$

$$= \underset{a \in K(i)}{\text{supremum}} \left[r_i^a + \rho \sum_{j \in I} p_{ij}^a v_{n-1}(j) \right], \quad \forall\, i \in I. \tag{2.54}$$

Equation (2.54) follows because we cannot do better on the right-hand side of (2.54) than take the best of all the values of the terms taken over all $a \in K(i)$. The right-hand side of (2.54) takes only a finite number of values. Hence we may replace 'supremum' by 'maximum'.

We then see that if δ_1 is determined by (2.54), with 'maximum' replacing 'minimum', then $\pi = (\delta_1, \tau)$ is an optimal policy over n time units for all initial states $i \in I$. This establishes our initial inductive hypothesis. It is precisely issues of this kind which make more general state space–action space Markov decision processes a little more difficult to handle, although usually producing essentially similar basic results.

Let us now define, as an extension to (2.48) for any u: $I \to R$, an operator T as follows:

$$[Tu](i) = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k + \rho \sum_{j \in I} p_{ij}^k u(j) \right], \quad \forall\, i \in I \tag{2.55}$$

which may be written as

$$Tu = \underset{\delta \in \Delta}{\text{maximum}}\ [T^\delta u] \tag{2.56}$$

where

$$[T^\delta u](i) = r_i^{\delta(i)} + \rho [P^\delta u]_i, \quad \forall\, i \in I \tag{2.57}$$

and Δ is the set of all deterministic rules for determining actions (see p. 27).

The reader should be aware in (2.48) and in (2.55)–(2.57), that the use of u as both a vector and a function is used throughout. Thus in (2.48) and (2.57) the ith component of u is taken as $[u]_i$ for matrix operation purposes, whereas in (2.55) the ith component of u is $u(i)$ for function purposes. We have now proved the next result.

Result 2.3. If $v_0 = u_0 = 0$ is specified, then $\{v_n\}$ is the unique solution set to the equations

$$\underline{n \geqslant 1} \qquad u_n = Tu_{n-1} \tag{2.58}$$

and, for each $n \geqslant 1$, there exists a policy $\pi \in \Pi_{MD}$ which is optimal for all initial states $i \in I$. ○

The uniqueness is trivial because the sequence is determined completely by u_0.

We have so far only reduced the set of policies needed to the set Π_{MD}, the set of all Markov deterministic policies, but not necessarily to Π_D because, in general, non-stationary optima are involved. We make the following notes:

(i) Any policy obtained by solving (2.58) is optimal, simultaneously, for all initial states.
(ii) Result 2.3, equation (2.58) is discussed in Howard [23], p. 79–81, and in White [58], p. 24. No rigorous proofs are given there, although White gives a partial verbal explanation.

Example (Howard [23], p. 80). The toymaker example, with $\rho = 0.9$, $n = 4$ (see Table 1.8) gives rise to Table 2.1.

As an illustration, given $\{v_0, v_1\}$, let us find $\{v_2(1), \delta_3(1)\}$. We have

$$\begin{aligned} v_2(1) &= [Tv_1](1) \\ &= \text{maximum}\begin{bmatrix} k=1\colon\ 6 + 0.9(0.5v_1(1) + 0.5v_1(2)) \\ k=2\colon\ 4 + 0.9(0.8v_1(1) + 0.2v_1(2)) \end{bmatrix} \\ &= \text{maximum}\begin{bmatrix} k=1\colon\ 6 + 0.45 \times 6 + 0.45 \times (-3) \\ k=2\colon\ 4 + 0.72 \times 6 + 0.18 \times (-3) \end{bmatrix} \\ &= \text{maximum}\begin{bmatrix} k=1\colon\ 7.35 \\ k=2\colon\ 7.78 \end{bmatrix} = 7.78, \end{aligned}$$

$$\delta_3(1) = 2.$$

Note that we want the optimal decision rules, and that π is not stationary. Note that, in accordance with our convention, $t = 1$ corresponds to $n = 4$ and that the process ends after four time units.

Table 2.1 Expected total discounted rewards and optimal decision rules for toymaker problem (reproduced from [23] Howard (1960), p. 80, by permission of The MIT Press)

		$v_n(i)$		$\delta_t(i)$			
		i		i			
		1	2	1	2		
n	0	0	0	—	—	—	t
	1	6	−3	1	1	4	
	2	7.78	−2.03	2	2	3	
	3	9.2362	−0.6467	2	2	2	
	4	10.533658	0.644197	2	2	1	

2.3.2 NON-STATIONARY CASE

In the stationary case, v_n can be defined independently of the actual chronological time at which the nth time unit from the end of the time horizon begins. For the non-stationary case we need to know not only how many time units remain but also where we are in chronological time. We need to redefine v_n.

Let $v_{tn}(i)$ be the supremal expected total discounted reward over the next $(n - t + 1)$ time units, beginning at the beginning of chronological time unit t with $X_t = i$. We could let n denote the number of time units remaining, and the approaches are equivalent, but we will proceed as indicated.

Following a similar line of reasoning as for the stationary case, we may deduce that $\{v_{tn}\}$ is a unique solution to the equations

$$\underline{t \leqslant n} \qquad u_{tn} = T_t u_{t+1,n} \tag{2.59}$$

where, for any u: $I_t \to R$,

$$[T_t u](i) = \underset{k \in K_t(i)}{\text{maximum}} \left[r_i^k(t) + \rho(t) \sum_{j \in I_{t+1}} p_{ij}^k(t) u(j) \right], \quad \forall\, i \in I_t, \tag{2.60}$$

$$\underline{t = n+1} \qquad u_{n+1,n} = 0. \tag{2.61}$$

Computations follow in exactly the same way as for the stationary case. Note that (2.61) may be replaced by

$$\underline{t = n+1} \qquad u_{n+1,n} = u \tag{2.62}$$

if there is a terminal value function u. This applies also to Result 2.3 as

a special case. Also, for each $n \geqslant 1$ there exists a policy $\pi \in \Pi_{MD}$ which is optimal for all initial states $i \in I$.

An alternative approach for the non-stationary case is to use the stationary approach and to define a new state space

$$\tilde{I} = I \times \{1, 2, ., ., .\} \tag{2.63}$$

with generic member

$$\tilde{i} = (i, t). \tag{2.64}$$

The transitions in the t variable are deterministic, viz. $t \to t + 1$. The net result is (2.59) and (2.60) with $u_{tn}(i)$ replaced by $u_n(i, t)$.

Let us now turn to the infinite horizon case.

2.4 INFINITE HORIZON MARKOV DECISION PROCESSES

2.4.1 THE DISCOUNTED STATIONARY CASE

Our objective is to find (see definition (2.13))

$$v(i) = \underset{\pi \in \Pi}{\text{supremum}} \; [v^{\pi}(i)], \quad \forall \; i \in I. \tag{2.65}$$

Real-life problems do not involve an infinite horizon. However, some real-life problems involve large sequences of decisions. Thus infinite horizon models may be useful as an approximation to some real-life problems where it is easier to solve the former, and use a solution of this to obtain a solution to the latter. We will return to this point subsequently. For the moment we study the, hypothetical, infinite horizon case.

We note, first of all, that Results 2.1 and 2.2 still hold, and we may keep to Π_M without loss. With $n = \infty$ (2.45) still holds, because $\rho < 1$ and the series converges. The analysis on pp. 36–38 follows in exactly the same way, replacing both n and $n - 1$ by ∞, and replacing both v_n and v_{n-1} by v. We thus obtain Result 2.4.

Result 2.4. The function v is a solution to the equation

$$u = Tu. \tag{2.66}$$

○

Thus v is a fixed point of the operator T.

Here $v(i)$ is defined by (2.65) and not by $\lim_{n\to\infty}[v_n(i)]$. They happen to be the same, but this has to be proved. Howard [23] implicitly takes this to be so. White [58], Corollary 2.9.1, proves this. Derman [15] and van der Wal [53] give rigorous treatments.

Derman's proof that we need only keep to Π_D is somewhat different (see Chapter 3, Theorem 1) because he goes directly to showing that optimal solutions in Π_D exist. Result 2.1 is slightly more general.

Eventually we will wish to solve (2.66). For this to be relevant we need to show that it has a unique u solution, and that all of the decision rule solutions δ giving rise to stationary policies $\pi = (\delta)^\infty$ are optimal. We establish the following results.

Result 2.5. Equation (2.66) has the unique solution $u = v$.

Proof (this is implicit in White [58], Theorem 2.7). Let u be any solution of (2.66). We will use the T^δ notation of (2.48).

Let δ, σ be decision rule solutions to (2.66) corresponding to v, u respectively. Then

$$v = Tv = T^\delta v \geqslant T^\sigma v = r^\sigma + \rho P^\sigma v, \tag{2.67}$$

$$u = T^\sigma u = r^\sigma + \rho P^\sigma u. \tag{2.68}$$

Thus (see p. 59 for $\geqslant$)

$$v - u \geqslant \rho P^\sigma (v - u). \tag{2.69}$$

Because $P^\sigma \geqslant 0$ we can repeat (2.69) s times to obtain

$$v - u \geqslant \rho^s((P^\sigma)^s(v - u)). \tag{2.70}$$

The right-hand side of (2.70) tends to 0 as s tends to ∞ because $\rho < 1$. Thus

$$v - u \geqslant 0. \tag{2.71}$$

Similarly we can obtain the reverse of (2.71). Thus

$$u = v. \tag{2.72}$$

○

Result 2.6. Let δ be any decision rule solution to (2.66), $\pi = (\delta)^\infty$, and v^π be the expected total discounted reward value function for policy π.

Then $v^{\pi} \geqslant v^{\tau}$ for all $\tau \in \Pi$, and stationary policies exist which are optimal for each state simultaneously.

Proof. Let τ be any policy in Π. Then

$$v^{\tau} \leqslant v. \tag{2.73}$$

We also have

$$v^{\pi} = T^{\delta} v^{\pi}, v = T^{\delta} v. \tag{2.74}$$

The first part of (2.74) is obtained using arguments similar to (2.49) and (2.50). Thus

$$v^{\pi} - v = \rho P^{\delta}(v^{\pi} - v). \tag{2.75}$$

Repeating (2.75) s times we obtain

$$v^{\pi} - v = \rho^{s}((P^{\delta})^{s}(v^{\pi} - v)). \tag{2.76}$$

The right-hand side of (2.76) tends to 0 as s tends to ∞. Thus

$$v^{\pi} = v \geqslant v^{\tau}, \quad \forall\ \tau \in \Pi. \tag{2.77}$$

○

Theorem 1 of Chapter 3 of Derman [15] and Theorem 5.13 of van der Wal [53] establish the existence of an optimal, simultaneously for each state, stationary policy. Result 2.6 is slightly stronger.

Example (Howard [23], p. 85). For the toymaker example of Table 1.8, for $n = \infty$, (2.66) has the solution, with $\rho = 0.9$,

$$\begin{aligned} u(1) &= 22.2, & u(2) &= 12.3, \\ \delta(1) &= 2, & \delta(2) &= 2. \end{aligned}$$

This should be checked, viz.

$$22.2 = \text{maximum}\begin{bmatrix} 6 + 0.9(0.5 \times 22.2 + 0.5 \times 12.3) \\ 4 + 0.9(0.8 \times 22.2 + 0.2 \times 12.3) \end{bmatrix},$$

$$12.3 = \text{maximum}\begin{bmatrix} -3 + 0.9(0.4 \times 22.2 + 0.6 \times 12.3) \\ -5 + 0.9(0.7 \times 22.2 + 0.3 \times 12.3) \end{bmatrix}.$$

We will consider solution algorithms later (see pp. 62 and 71).

2.4.2 THE DISCOUNTED NON-STATIONARY CASE

We follow the format for the finite horizon case on pp. 40–41 and define $v_t(i)$ to be the supremal expected total discounted reward over an infinite time horizon, beginning at the beginning of chronological time unit t with $X_t = i$.

The analysis for the infinite horizon case goes through in much the same way as for pp. 41–43 and we have analogous results to those of Results 2.4–2.6, provided that $\{|r_t^k(t)|\}$ are bounded and $\rho(t) \leqslant \rho < 1$ for some ρ.

Result 2.7. The function v_t is a solution to the equation

$$\underline{t \geqslant 1} \qquad u_t = T_t u_{t+1} \tag{2.78}$$

where T_t is defined in (2.60). ○

Result 2.8. Equation (2.78) has a unique bounded solution. ○

Reference to boundedness is made in Result 2.8 because there is a countable set of $\{u_t\}$ whereas, in Result 2.6, any real u is bounded.

Result 2.9. Let $\{\delta_t\}$ be any decision rule sequence solution to equation (2.78), and $\pi = (\delta_1, \delta_2, ., ., \delta_t ., .,)$ be the associated policy. Then $v_t^\pi \geqslant v_t^\tau$ for all $t \geqslant 1$ and for all $\tau \in \Pi$. ○

We make the following notes:

(i) The π in Result 2.9 is, in general, non-stationary.
(ii) The conditions on the boundedness of $\{r_t^k(t)\}$ and $\{\rho(t)\}$ are not strictly necessary.
(iii) Limited work seems to exist on infinite horizon non-stationary problems.
(iv) In order to solve (2.78), some asymptotic form of $\{p_{ij}^k(t)\}$, $\{r_t^k(t)\}$ and $\{\rho(t)\}$ is required.

2.4.3 THE AVERAGE EXPECTED REWARD PER UNIT TIME. STATIONARY CASE

We will deal only with the stationary case. We wish to find (see (2.19) and (2.15))

$$g(i) = \underset{\pi \in \Pi}{\text{supremum}} \; [g^\pi(i)], \quad \forall\, i \in I \tag{2.79}$$

where (see (2.16))

$$g^{\pi}(i) = \underset{n \to \infty}{\text{limit infimum}} \; [v_n^{\pi}(i)/n], \quad \forall\, i \in I. \tag{2.80}$$

Subsequently we will be able to replace 'limit infimum' by 'limit'.

We will consider only uni-chain cases, i.e. for all $\delta \in \Delta$ the process corresponding to the probability transition matrix P^{δ} has only a single chain plus, perhaps, some transient states. Such processes are sometimes called 'ergodic' (see Bartlett [2], p. 33). Definitions are not always consistent in the literature. Kemeny and Snell [25], p. 37, use the term 'ergodic (chain)' to exclude transient states. A 'transient state' i, is one for which the limiting average probability of being in state i $\left(\text{i.e. } \underset{n \to \infty}{\text{limit}} \left[\left(\sum_{t=0}^{n} q(P^{\delta})^t\right)/n\right]\right)$ is 0 for all initial probability vectors q of states. This is equivalent to definition (1.15).

The term 'completely ergodic' is also in use. Howard [23], p. 6, uses this term for any transition matrix P^{δ}, where $\underset{n \to \infty}{\text{limit}} [(P^{\delta})^n]$ exists (see p. 9). This gives the usual steady-state probability vector, independent of initial state, of each row of the limit. White [58], p. 31, uses the term in the same sense. Bartlett [2], p. 33, uses the term 'regular'. Mine and Osaki [34], p. 27, use the term 'completely ergodic' simply to mean that all matrices P^{δ} are ergodic in the sense of Kemeny and Snell [25]. Kallenberg [24], p. 24, defines 'completely ergodic' as for Howard but excludes transient states.

For an ergodic process in our sense (see Bartlett [2], p. 33), we have

$$\text{rank}(U - P^{\delta}) = m - 1 \tag{2.81}$$

where U is the identity matrix. Equation (2.81) is important in what follows. Multiple-chain Markov decision processes are covered in Derman [15], in Mine and Osaki [34] and in Howard [23].

We will now deal with results analogous to those of Results 2.4–2.6 for the discounted problem. The approach will be slightly different to the proofs in other texts, to avoid a deeper use of 'limit infimum' ideas. Theorem 2 of Chapter 3 of Derman [15], and Theorem 3.4 of Mine and Osaki [34], use a 'limit infimum' result of Widder [78] which we will avoid.

Although our main concern is with the uni-chain case, some generality will be maintained for a while, so that some multiple-chain results may also be seen. Here g^{π} will always be a function (or vector) in this section, with components $\{g^{\pi}(i)\}$. In the uni-chain cases (see Result

2.10 and equations (2.118) and (2.119)), g^π will take the form he, for some h.

Using the $\{x_i^{\pi k}(t)\}$ notation of Result 2.1 we can write $v_n^\pi(i)$ in the form

$$v_n^\pi(i) = \sum_{t=1}^{n} \sum_{j \in I, k \in K(j)} x_j^{\pi k}(t) r_j^k, \quad \forall\, i \in I. \tag{2.82}$$

Using Result 2.1 we see that we may assume that $\pi \in \Pi_M$.

Now consider the following equation, suggested by equation (1.25) for a specific stationary policy. This will be our optimality equation, which we will discuss in more detail when we look at algorithms.

$$u(i) + h = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k + \sum_{j \in I} p_{ij}^k u(j) \right], \quad \forall\, i \in I, \tag{2.83}$$

$$u(m) = 0. \tag{2.84}$$

We write this in function/decision rule form as follows, where e is the unit function:

$$\begin{aligned} u + he &= \underset{\delta \in \Delta}{\text{maximum}} \; [r^\delta + P^\delta u] \\ &= \underset{\delta \in \Delta}{\text{maximum}} \; [T^\delta u] = Tu, \end{aligned} \tag{2.85}$$

$$u(m) = 0. \tag{2.86}$$

In (2.83)–(2.86) (u, h) is any solution to (2.85) and (2.86). We will use (w, g) for (u, h) solutions which correspond to optimal policies for our Markov decision processes. Here w need not be the actual bias function for the policy, and will differ from this by a constant function. Also, because we wish to use matrices, u is interpreted as a vector for matrix operations, although we will still adhere to functional notation $u(i)$, $[Tu](i)$ and $[T^\delta u](i)$ as indicated on p. 38.

As a consequence of (2.81) for all $\delta \in \Delta$, (2.85) and (2.86) will always have a solution. We will deal with this later on (see p. 84) under the topic of 'policy space algorithm'.

The proof of Result 2.10, which we will shortly address, is a different version of the proofs in Theorems 3.5 and 3.6 of White [58]. The difference lies essentially in the specification of v_0. Lemma 3.6 of Mine and Osaki [34] is also related to our proof, but we do not require the regularity condition of Bartlett [2].

Let (u, h) be any solution to (2.85) and (2.86) with an associated decision rule δ. Let $\pi = (\delta)^\infty$. Let us now look at the form of $\tilde{v}_n$ when $\tilde{v}_0 = u$, where $\tilde{v}_n$ is v_n modified to have a terminal value function $\tilde{v}_0 = u$, whereas $v_0 = 0$ by definition; $\tilde{v}_n^\pi$ is a similar modification of v_n^π.

We have (see (2.58) with $\rho = 1$)

$$\tilde{v}_1 = T\tilde{v}_0 = Tu = u + he. \tag{2.87}$$

Now assume that, for $0 \leqslant s \leqslant n-1$,

$$\tilde{v}_s = u + she. \tag{2.88}$$

Then

$$\begin{aligned} \tilde{v}_n &= T\tilde{v}_{n-1} = T(u + (n-1)he) \\ &= (n-1)he + Tu \\ &= (n-1)he + u + he \\ &= u + nhe. \end{aligned} \tag{2.89}$$

Also

$$\tilde{v}_n^\pi = \tilde{v}_n \tag{2.90}$$

because

$$Tu = T^\delta u. \tag{2.91}$$

Now if τ is any policy in Π and if $\tilde{v}_n^\tau$ is as for v_n^τ but with a terminal value function $\tilde{v}_0 = u$, we have

$$\tilde{v}_n^\tau \leqslant \tilde{v}_n^\pi, \quad \forall\ \tau \in \Pi. \tag{2.92}$$

Hence

$$\liminf_{n\to\infty} [\tilde{v}_n^\tau(i)/n] \leqslant \liminf_{n\to\infty} [\tilde{v}_n^\pi(i)/n], \quad \forall\ \tau \in \Pi,\ i \in I. \tag{2.93}$$

Now if $\tau = (\delta_1, \delta_2, ., ., \delta_t, \ldots) = (\delta_1, \eta)$ then

$$\underline{n \geqslant 1} \qquad v_n^\tau = r^{\delta_1} + P^{\delta_1} v_{n-1}^\eta, \tag{2.94}$$

$$\tilde{v}_n^\tau = r^{\delta_1} + P^{\delta_1} \tilde{v}_{n-1}^\eta. \tag{2.95}$$

Thus

$$\begin{aligned} v_n^\tau - \tilde{v}_n^\tau &= P^{\delta_1}(v_{n-1}^\eta - \tilde{v}_{n-1}^\eta) \\ &= \left(\prod_{t=1}^{n} P^{\delta_t}\right)(v_0 - \tilde{v}_0) \\ &= -\left(\prod_{t=1}^{n} P^{\delta_t}\right)u. \end{aligned} \tag{2.96}$$

From (2.96)

$$\| v_n^\tau - \tilde{v}_n^\tau \| \leqslant \| u \|, \tag{2.97}$$

where

$$\| u \| = \max_{i \in I} [| u(i) |]. \tag{2.98}$$

Combining (2.97) and (2.93) we see that

$$\begin{aligned} &\liminf_{n \to \infty} \quad [(v_n^\tau(i) - \| u \|)/n] \\ &\qquad \leqslant \liminf_{n \to \infty} \quad [(v_n^\pi(i) + \| u \|)/n], \quad \forall\ i \in I. \end{aligned} \tag{2.99}$$

Here $\| u \|/n$ tends to 0 as n tends to ∞, and hence (2.99) gives

$$\liminf_{n \to \infty} [v_n^\tau(i)/n] \leqslant \liminf_{n \to \infty} [v_n^\pi(i)/n], \quad \forall\ \tau \in \Pi,\ i \in I. \tag{2.100}$$

We have thus demonstrated the following result for the uni-chain case.

Result 2.10. Let δ be any decision rule solution to (2.85) and (2.86) and let $\pi = (\delta)^\infty$. Then $g^\pi \geqslant g^\tau$ for all $\tau \in \Pi$, and stationary policies exist which are optimal for each state simultaneously. ○

In the discounted case we could replace 'limit infimum' by 'limit' because limits exist in that case. Limits do not always exist for $\{v_n^\pi(i)/n\}$ (see pp. 29–30). However, if (see p. 27) $\pi \in \Pi_D$ (i.e. $\pi = (\delta)^\infty$ for some $\delta \in \Delta$) then limits do exist (e.g. Mine and Osaki [34], p. 25, or Derman [15], Theorem 1, Appendix A). Thus we have Result 2.11.

Result 2.11. In Result (2.10)

$$g^{\pi}(i) = \underset{n \to \infty}{\text{limit}}\ \text{infimum}\ [v_n^{\pi}(i)/n], \quad \forall\ i \in I \tag{2.101}$$

may be replaced by

$$g^{\pi}(i) = \lim_{n \to \infty} [v_n^{\pi}(i)/n], \quad \forall\ i \in I. \tag{2.102}$$

○

Let us now look at one special relationship between average expected reward per unit time and expected total discounted reward, for infinite time horizons. For a specific example (see (1.33)) it was shown for a given policy that

$$\lim_{\rho \to 1^-} [(1-\rho)v_{\rho}] = g \tag{2.103}$$

where, at this point, we put a subscript ρ to v to show its dependence on ρ. Equation (2.103) is true in general for all $\pi \in \Pi_D$ (a consequence of Derman [15], Theorem 1 of Appendix A, plus the Cesàro limit result of Bromwich [8], p. 150, or see Mine and Osaki [34], Lemma 3.2) i.e. for all $\pi \in \Pi_D$

$$\lim_{\rho \to 1^-} [(1-\rho)v_{\rho}^{\pi}] = g^{\pi}. \tag{2.104}$$

Equation (2.104) is true even for multiple-chain cases.

Now select a sequence $\{\rho_s\} \to 1^-$. Let $\pi^s \in \Pi_D$ be an optimal policy for discount factor ρ_s. Because, for all $i \in I$, $K(i)$ is finite, and I is finite, then Π_D contains only a finite number of policies. Hence some policy π^*, say, must repeat indefinitely in the sequence $\{\pi^s\}$. Let S be a subsequence $\{s\}$ for the repeated π^*.

We know that

$$v_{\rho_s}^{\tau} \leqslant v_{\rho_s}^{\pi^*}, \quad \forall\ s \in S,\ \tau \in \Pi_D. \tag{2.105}$$

Thus

$$\lim_{s \to \infty, s \in S} [(1-\rho_s)v_{\rho_s}^{\tau}] \leqslant \lim_{s \to \infty, s \in S} [(1-\rho_s)v_{\rho_s}^{\pi^*}], \quad \forall\ \tau \in \Pi_D. \tag{2.106}$$

Hence using (2.104) we have

$$g^{\tau} \leqslant g^{\pi^*}, \quad \forall\ \tau \in \Pi_D. \tag{2.107}$$

We have thus shown Result 2.12 for the general case.

Result 2.12. There exists a policy $\pi^* \in \Pi_D$ for which g^{π^*} is optimal, and which is also optimal for a sequence $\{\rho_s\}$ of discount factors tending to 1^-. ○

A stronger form of this result may be found in Derman [15], Chapter 3, Corollary 1, or in White [58], Exercise 8 of Chapter 9, where a sequence $\{\rho_s\}$ of discount factors tending to 1^- may be replaced by an interval $[\rho^*, 1)$ of discount factors for some $\rho^* < 1$. The results are, however, effectively the same in that if ρ is large enough (for a given data set) we need not worry about whether or not we discount. However, there is still the problem of finding the appropriate $\{\pi^*, \rho^*\}$. Howard [23], p. 88, tabulates solutions for a range of discount factors for a taxi-cab problem which illustrates the above result.

If $\pi = (\delta)^\infty$ is an optimal policy for the average expected reward case, in order to check if it is also optimal for the expected total discounted reward case for a given ρ, we need to check that (2.66) is satisfied. Now, again at this point using the suffix ρ, for a given policy $\pi = (\delta)^\infty$ we have (see (2.74))

$$v_\rho^\pi = T^\delta v_\rho^\pi = r^\delta + \rho P^\delta v_\rho^\pi. \tag{2.108}$$

Thus

$$v_\rho^\pi = (U - \rho P^\delta)^{-1} r^\delta. \tag{2.109}$$

Hence we need to show that

$$\begin{aligned}(U - \rho P^\delta)^{-1} r^\delta &\geqslant T^\sigma (U - \rho P^\delta)^{-1} r^\delta \\ &= r^\sigma + \rho P^\sigma (U - \rho P^\delta)^{-1} r^\delta, \quad \forall\ \sigma \in \Delta. \end{aligned} \tag{2.110}$$

Establishing (2.110) is the same as establishing that

$$[(U - \rho P^\delta)^{-1} r^\delta]_i \geqslant r_i^k + \rho \sum_{j \in I} p_{ij}^k [(U - \rho P^\delta)^{-1} r^\delta]_j, \quad \forall\ i \in I,\ k \in K(i). \tag{2.111}$$

White [72] gives an example where two policies are optimal for the average expected reward per unit time case, but one of which is not optimal for the total expected discounted reward case for any $\rho \in [0, 1)$. Also, in White [72] the result of Derman [15] is extended by looking at the class Π^{**} of all policies which are optimal for some sequence of discount factors tending to unity in the expected total discounted reward case, and establishes some properties of Π^{**}.

One other point needs mentioning. For any policy $\pi = (\delta)^\infty$, $\delta \in \Delta$ we have

$$v_n^\pi = r^\delta + P^\delta v_{n-1}^\pi \tag{2.112}$$

which by repetition is equal to

$$\left(\sum_{l=1}^{n} (P^\delta)^{l-1}\right) r^\delta. \tag{2.113}$$

Then, noting that at this stage, g^π is a function, we have

$$g^\pi = \lim_{n \to \infty} \left[\sum_{l=1}^{n} (P^\delta)^{l-1} \Big/ n\right] r^\delta \tag{2.114}$$

where (see Mine and Osaki [34], Lemma 3.3) the specified

$$\lim_{n \to \infty} \left[\left(\sum_{l=1}^{n} (P^\delta)^{l-1}\right) \Big/ n\right] = P^{\delta *}$$

(see p. 9) exists and (see (1.17))

$$P^\delta * P^\delta = P^\delta P^{\delta *} = P^{\delta *}. \tag{2.115}$$

In the uni-chain case each row of $P^{\delta *}$ is the same. Let this be θ^δ. In the regular case (see p. 45) θ^δ is also the steady-state probability vector, independent of the starting state (see Bartlett [2], p. 33). We have

$$\sum_{i \in I} \theta_i^\delta = 1, \qquad \theta^\delta \geqslant 0. \tag{2.116}$$

We thus have

$$\theta^\delta P^\delta = \theta^\delta. \tag{2.117}$$

For a policy $\pi = (\delta)^\infty$ generated by any solution (u, h) to (2.85) and (2.86) we have

$$he = g^\pi \tag{2.118}$$

and

$$u + g^\pi = r^\delta + P^\delta u. \tag{2.119}$$

Equation (2.118) holds for the following reason. From (2.89) we have

$$he = \lim_{n \to \infty} [\tilde{v}_n / n]. \tag{2.120}$$

Also from (2.96), (2.90) with $\tau = \pi$ we have

$$g^\pi = \lim_{n\to\infty} [v_n^\pi/n] = \lim_{n\to\infty} [\tilde{v}_n^\pi/n]$$

$$= \text{limit } [\tilde{v}_n/n] = he. \tag{2.121}$$

Hence using (2.117) and (2.121) we have, premultiplying (2.119) by θ^δ,

$$g^\pi = \theta^\delta r^\delta e \tag{2.122}$$

(see Howard [23], pp. 34–36). Equation (2.122) written in full is

$$g^\pi = \left(\sum_{i\in I} \theta_i^\delta r_i^{\delta(i)}\right) e. \tag{2.123}$$

Example (Howard [23], p. 41). For the toymaker example of Table 1.8, for $n = \infty$, $\rho = 1$, (2.85), (2.86) have the solution (setting $u(2) = 0$)

$$h = g^\pi(1) = g^\pi(2) = 2, \qquad \pi = (\delta)^\infty, \qquad \delta = (2, 2),$$
$$u(1) = w(1) = 10, \qquad u(2) = w(2) = 0.$$

Again note that w differs from the true bias function by a constant function.

It is easily checked that

$$10 + 2 = \text{maximum}\begin{bmatrix} 6 + & 0.5 \times 10 & +0.5 \times 0 \\ 4 + & 0.8 \times 10 & +0.2 \times 0 \end{bmatrix},$$

$$0 + 2 = \text{maximum}\begin{bmatrix} -3 & +0.4 \times 10 & +0.6 \times 0 \\ -5 & +0.7 \times 10 & +0.3 \times 0 \end{bmatrix},$$

$$\delta(1) = 2, \qquad \delta(2) = 2.$$

Note that δ is the same as the optimal δ for the discounted problem on p. 22. In fact δ will be optimal for all $\rho \in [0.9, 1)$ and we will have ρ^* (see p. 50) $\leqslant 0.9$.

To find the limiting probability vector θ^δ (see (2.117)) we have to solve

$$\theta^\delta \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix} = \theta^\delta$$

with $\theta_1^\delta + \theta_2^\delta = 1$. Thus $\theta_1^\delta = 0.78$, $\theta_2^\delta = 0.22$. Then (checking with (2.123))

$$g^\pi = (0.78 \times 4 + 0.22 \times -5)e = 2e.$$

2.5 ABSORBING STATE PROBLEMS. THE INFINITE HORIZON STATIONARY CASE

In this section we will give no proofs but merely quote results. The objective is (see (2.20)) to find

$$v(i) = \underset{\pi \in \Pi}{\text{supremum}} \; [v^\pi(i)], \quad \forall\, i \in I \tag{2.124}$$

with

$$v(i) = 0, \quad \forall\, i \in I_a. \tag{2.125}$$

We will assume either that the rewards are discounted or that for each $i \in I$, $\pi \in \Pi$ there is a $t \geqslant 1$ such that

$$\text{probability}(X_t \in I_a \mid X_1 = i) = 1 \tag{2.126}$$

(see p. 25 for $\{X_t\}$, and see Derman [15], p. 30 or Mine and Osaki [34], p. 42 for (2.126)).

In the discounted case all of the discounted results apply automatically. The analysis is exactly the same. The additional point is that (2.66) now takes the form

$$u = Tu, \tag{2.127}$$

$$u(i) = 0, \quad \forall\, i \in I_a. \tag{2.128}$$

Results 2.4 and 2.6 hold in the case when (2.126) is assumed, using (2.128) in addition to (2.66) (see Mine and Osaki [34], Lemmas 3.10, 3.11 and Theorem 12). Result 2.5 also holds in the form of (2.127) and (2.128) (see Derman [15], Chapter 5, Corollary 1, who also establishes Results 2.4 and 2.6 in Theorem 1, Chapter 5 and Theorem 4, Chapter

3). White [58], Theorem 1.10, gives condition under which (2.127) and (2.128) (replacing m by I_a) have a unique solution, and Corollary 1.13.1, together with Theorems 1.10, 1.13 and Corollary 1.13.2, give all the Results 2.4–2.6 (with (2.127) and (2.128) instead of (2.66)) because condition (2.126) implies the prerequisites of those corollaries and theorems.

Let us now look at some illustrations of discounted, average reward and absorbing state problems. In these illustrations the objective is to 'minimise' a specific measure of performance. We will conform with 'maximisation' by using the negatives of the performance measures, but one can simply redefine the functions used and use 'minimisation'.

2.6 SOME ILLUSTRATIONS

(a) Inventory. Infinite horizon expected total discounted reward

In p. 17 we introduced an inventory problem with a stock reorder rule

$$\begin{aligned} &\text{if } i(\text{stock level}) \leqslant k \leqslant m, \text{ order } (k-i), \\ &\text{if } i > k, \text{ order zero.} \end{aligned} \tag{2.129}$$

Here k was fixed. Our problem is to find an optimal k which minimises the expected total discounted cost. Using our maximisation formulation equation (2.66) becomes

$$\begin{aligned} u(i) = \underset{m \geqslant k \geqslant i}{\text{maximum}} \Bigg[&- \Bigg(c(k-i) + \sum_{s>k} q(s) l(s-k) \\ &+ \tfrac{1}{2} a\Bigg(k + \sum_{s<k} q(s)(k-s)\Bigg)\Bigg) \\ &+ \rho\Bigg(\sum_{s<k} q(s) u(k-s) + \Bigg(\sum_{s \geqslant k} q(s)\Bigg) u(0)\Bigg)\Bigg], \qquad 1 \leqslant i \leqslant m. \end{aligned} \tag{2.130}$$

Here k is allowed to depend upon i and an optimal decision rule may not involve a fixed k for all i. In some cases it will (e.g. see Bellman [4], Chapter V).

We will return to the analytic solution of (2.130) later (p. 147).

(b) Queuing. Infinite horizon average expected reward per unit time

In p. 19 we introduced a queuing problem with the following decision rule:

send all customers in the system in excess of k elsewhere. (2.131)

There k was fixed. Our problem is to find an optimal k which minimises the average expected cost per unit time over an infinite time horizon. Using our maximisation formulation (2.85) becomes

$$1 \leqslant i < m \qquad u(i) + h = \underset{0 \leqslant k \leqslant i}{\text{maximum}} \; [-(c(i-k) + ak) + p(1-q)u(i+1) + (pq + (1-p)(1-q))u(i) + (1-p)qu(i-1)], \tag{2.132}$$

$$i = 0 \qquad u(0) + h = pu(1) + (1-p)u(0). \tag{2.133}$$

$$i = m \qquad u(m) + h = \underset{0 \leqslant k \leqslant m}{\text{maximum}} \; [-(c(m-k) + ak) + (1-q)u(m) + qpu(m) + q(1-p)u(m-1)]. \tag{2.134}$$

For the case $i = m$ we only accept an arrival if a service is also completed. We also have (see (2.86))

$$u(m) = 0. \tag{2.135}$$

(c) Defective production. Absorbing state expected total reward

The following is taken from White [57], pp. 114–15.

(i) A decision-maker has to produce a number i of specialised items for a client. The items are of no use to anyone else. Because the job is infrequent and specialised, defective items can be produced.

(ii) If a production run of k is planned then the probability that s good items will be produced in $p(k, s)$.

(iii) Each production run costs a to set up and each unit in the run costs b.

(iv) It is required to find a production-run policy to minimise the expected cost of eventually supplying at least i good items. So $-v(i)$ will be this minimal expected cost given i. The process is absorbing state because it stops when $i = 0$. The maximisation

formulation for (2.127) and (2.128) is, assuming $k \geqslant i$ and with $I_a = \{0\}$,

$$\underline{i > 0} \qquad u(i) = \underset{k \geqslant i}{\text{maximum}} \left[-(a + bx) + \sum_{s=0}^{i-1} p(k, s)u(i - s) \right], \tag{2.136}$$

$$u(0) = 0. \tag{2.137}$$

Equation (2.136) is of a 'directed' form (i.e. the states never increase in terms of i) and may be reduced to (if $p(k, 0) \neq 1$)

$$\underline{i > 0} \qquad u(i) = \underset{k \geqslant i}{\text{maximum}} \left[\frac{-(a + b) + \sum_{s=1}^{i-1} p(k, s)u(i - s)}{(1 - p(k, 0))} \right], \tag{2.138}$$

$$u(0) = 0. \tag{2.139}$$

Equations (2.138) and (2.139) may be solved by working backwards calculating $u(0)(= 0), u(1), u(2) ., ., u(i)$, where $u = v$ is the solution to (2.138) and (2.139).

We have, in all cases, cast the problem into a 'maximisation' form by using the negatives of the costs in order to conform exactly with our maximisation format. However, they may equally well be cast in a cost 'minimisation' form.

2.7 EXERCISES FOR CHAPTER 2

1. On p. 27 we introduced deterministic policies Π_D. Later on we will consider problems involving the variance of the rewards. Let A, B be two independent random variables (i.e. A is a sample from a finite population A with a specified probability $p(r)$ that $A = r$, and B is a sample from a different finite population B with a specified probability $q(r)$ that $B = r$). Suppose that we may (a) select A from A consistent with $p(.)$, or (b) select B from B consistent with $q(.)$, or (c) take the sample 'AαB', which means that 'we take action (a) with probability α and action (b) with probability $(1 - \alpha)$'. Let

$C(a), C(b), C(c)$ be the random variable outcomes of $(a), (b), (c)$ respectively. Show that

$$\text{variance}(C(c)) \geqslant \min[\text{variance}(C(a)), \text{variance}(C(b))].$$

Thus if we wish to minimise variance we need not consider randomised actions.

2. On p. 26 it is assumed that the reward in any time unit is received at the beginning of the time unit.

   (a) How would (2.6) be modified if the reward is received at the end of the time unit?
   (b) How can the form of (2.6) be retained by redefining terms?

3. On p. 17 we introduced an inventory problem where the probability that the demand in any time unit is equal to s is known to be $q(s)$. Suppose now that this probability is $q(s, \alpha)$, where the value of the parameter α is unknown but has prior probabilities $\{p(\alpha)\}$, $\alpha = 1, 2$. Confine yourself to the transition probabilities only and give two ways of representing the new state of the system, in order to keep the Markov property (2.22), which takes into account any actual demands up to the beginning of each time unit, and derive the appropriate transition probabilities.
4. Find $\{v_n\}$ and $\{\delta_t\}$, $0 \leqslant n \leqslant 7$, $1 \leqslant t \leqslant 7$ for the following problem:

$$I = \{1, 2, 3\}, \qquad K(i) = \{1, 2\},\ i \in I, \qquad v_0 = 0, \qquad \rho = 0.75.$$

i	k	r_i^k	p_{i1}^k	p_{i2}	p_{i3}
1	1	4	0	$\frac{1}{3}$	$\frac{2}{3}$
1	2	4	0	$\frac{1}{2}$	$\frac{1}{2}$
2	1	9	$\frac{1}{4}$	0	$\frac{3}{4}$
2	2	10	$\frac{2}{3}$	0	$\frac{1}{3}$
3	1	4	$\frac{2}{3}$	$\frac{1}{3}$	0
3	2	3	$\frac{1}{2}$	$\frac{1}{2}$	0

5. For the infinite horizon version of Exercise 4 show that the policy $\pi = (\delta)^\infty$, where $\delta(1) = \delta(2) = 2$, $\delta(3) = 1$, is an optimal policy among all policies and that it is a unique optimal policy among all stationary deterministic Markov policies.

6. For the average expected reward version of Exercise 4 demonstrate that π is an optimal policy. You may assume that all the transition matrices for the Markov decision rules are uni-chain.
7. In (2.90) it is stated that

$$\tilde{v}_n^{\pi} = \tilde{v}_n.$$

   Prove this.
8. Derive a corresponding expression to (2.59) if n is now the number of time units remaining from the beginning of time unit t.
9. In Exercise 8 how would you handle the infinite horizon case in principle, and what would be the resulting optimality equation?

CHAPTER 3

Algorithms

3.1 INFINITE HORIZON EXPECTED TOTAL DISCOUNTED REWARD

3.1.1 STATIONARY CASE

We need to find a solution (u, δ) to the following equation (see (2.66), (2.55)–(2.57)):

$$u = Tu \tag{3.1}$$

where for u: $I \to R$

$$Tu = \underset{\delta \in \Delta}{\text{maximum}} \; [T^{\delta} u], \tag{3.2}$$

$$[Tu](i) = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k + \rho \sum_{j \in I} p_{ij}^k u(j) \right], \tag{3.3}$$

$$\begin{aligned} [T^{\delta} u](i) &= r_i^{\delta(i)} + \rho \sum_{j \in I} p_{ij}^{\delta(i)} u(j) \\ &= r_i^{\delta(i)} + \rho [P^{\delta} u]_i. \end{aligned} \tag{3.4}$$

Before we tackle the algorithms we need some properties of the operator T. For u, u': $I \to R$ define

$$u \geqslant u' \rightleftarrows u(i) \geqslant u'(i), \quad \forall \; i \in I \tag{3.5}$$

with a similar definition for $\{\leqslant, >, <, =\}$ and the norm $\| \cdot \|$ by

$$\| u \| = \underset{i \in I}{\text{maximum}} \; [| u(i) |]. \tag{3.6}$$

Note again that we will treat u both as a function and as a vector from time to time.

Result 3.1 (Lemma 2.2, Mine and Osaki [34]). If $\delta \in \Delta$, $u \geqslant u'$ then

$$T^{\delta}u \geqslant T^{\delta}u'. \tag{3.7}$$

Proof.

$$T^{\delta}u - T^{\delta}u' = r^{\delta} + \rho P^{\delta}u - (r^{\delta} + \rho P^{\delta}u') = \rho P^{\delta}(u - u') \geqslant 0. \tag{3.8}$$

○

Result 3.2 (Lemma 8.8, Mine and Osaki [34]). If $u \geqslant u'$ then

$$Tu \geqslant Tu'. \tag{3.9}$$

Proof. From Result 3.1, if $u \geqslant u'$ then

$$T^{\delta}u \geqslant T^{\delta}u', \quad \forall\ \delta \in \Delta. \tag{3.10}$$

Then

$$Tu = \underset{\delta \in \Delta}{\text{maximum}}\ [T^{\delta}u] \geqslant \underset{\delta \in \Delta}{\text{maximum}}\ [T^{\delta}u'] = Tu'. \tag{3.11}$$

○

Result 3.3. For any u, u'

$$\| Tu - Tu' \| \leqslant \rho \| u - u' \|. \tag{3.12}$$

Proof. Let $\delta \in \Delta$ be such that

$$Tu = T^{\delta}u. \tag{3.13}$$

Then

$$Tu' \geqslant T^{\delta}u' \tag{3.14}$$

and

$$\begin{aligned} Tu - Tu' &\leqslant T^{\delta}u - T^{\delta}u' \\ &= \rho P^{\delta}(u - u'). \end{aligned} \tag{3.15}$$

Similarly, if $\tau \in \Delta$ is such that

$$Tu' = T^{\tau}u' \tag{3.16}$$

we obtain

$$Tu' - Tu \leqslant \rho P^{\tau}(u' - u). \tag{3.17}$$

Hence

$$\rho P^{\tau}(u - u') \leqslant Tu - Tu' \leqslant \rho P^{\delta}(u - u'). \tag{3.18}$$

Thus

$$-\rho \| u - u' \| e \leqslant \rho P^{\tau}(u - u') \leqslant Tu - Tu' \leqslant \rho P^{\delta}(u - u') \leqslant \rho \| u - u' \| e,$$

i.e.

$$-\rho \| u - u' \| e \leqslant Tu - Tu' \leqslant \rho \| u - u' \| e. \tag{3.19}$$

This establishes the result. ○

Result 3.4 (Mine and Osaki [34], p. 6). Let

$$\bar{r} = \underset{i \in I, k \in K(i)}{\text{maximum}} [r_i^k], \tag{3.20}$$

$$\underline{r} = \underset{i \in I, k \in K(i)}{\text{minimum}} [r_i^k]. \tag{3.21}$$

Then

$$(\underline{r}/(1 - \rho))e \leqslant v^{\pi} \leqslant (\bar{r}/(1 - \rho))e, \quad \forall\ \pi = (\delta)^{\infty}, \quad \delta \in \Delta. \tag{3.22}$$

Proof. From (2.74) restricted to a single policy π

$$v^{\pi} = r^{\delta} + \rho P^{\delta} v^{\pi} \leqslant \bar{r}e + \rho P^{\delta} v^{\pi}. \tag{3.23}$$

Hence (with U as identity matrix)

$$\begin{aligned} v^{\pi} &\leqslant (U - \rho P^{\delta})^{-1} \bar{r} e \\ &= (1 - \rho)^{-1} \bar{r} e. \end{aligned} \tag{3.24}$$

Similarly

$$\begin{aligned} v^{\pi} &\geqslant (U - \rho P^{\delta})^{-1} \underline{r} e \\ &= (1 - \rho)^{-1} \underline{r} e. \end{aligned} \tag{3.25}$$

○

We will now consider two basic methods of solving equation (3.1), viz. value iteration (often called successive approximations, see Bellman [4], p. 14, and White [58], p. 111) and policy space iteration. Value iteration may not give an optimal policy solution unless we carry out enough iterations, the number of which we will not, in general, know in advance. Thus value iteration is used to obtain approximately optimal policies although it may actually give optimal policies. For policy space iteration, the number of iterations is bounded by the number of policies, i.e. $\prod_{i \in I} \# K(i)$, where $\prod$ means product, and # means cardinality, i.e. number of members.

3.1.2 VALUE ITERATION (Bellman [4], p. 14, White [58], p. 24, Howard [23])

This considers the limiting form of finite horizon equations (see (2.58) with $u_0 = u$, arbitrary), viz.

$$\underline{n \geqslant 1} \qquad u_n = Tu_{n-1}. \tag{3.26}$$

We will use v to denote the unique solution to (3.1) (see (2.13) and Result 2.5)). Let

$$\alpha = \max_{i \in I} [v(i) - u(i)], \tag{3.27}$$

$$\beta = \min_{i \in I} [v(i) - u(i)]. \tag{3.28}$$

Result 3.5 (Shapiro [45]).

$$\underline{n \geqslant 0} \qquad u_n + \beta \rho^n e \leqslant v \leqslant u_n + \alpha \rho^n e. \tag{3.29}$$

Proof.

$$\underline{n \geqslant 1} \qquad v = Tv = T^{\delta} v, \tag{3.30}$$

$$u_n = Tu_{n-1} = T^{\sigma_n} u_{n-1}. \tag{3.31}$$

Care should be taken not to confuse σ_n with δ_t as defined in (x) on p. 26. Here σ_n is the decision rule obtained from operating on u_{n-1} by

the operator T in (3.31). If $t = 1$ corresponds to n time units remaining, then $\sigma_n = \delta_1$. Then

$$\begin{aligned} v - u_n &= T^{\delta} v - T^{\sigma_n} u_{n-1} \\ &\leqslant T^{\delta} v - T^{\delta} u_{n-1}. \end{aligned} \tag{3.32}$$

Now the right-hand side of (3.29) is true for $n = 0$ because

$$v - u_0 = v - u \leqslant \alpha e. \tag{3.33}$$

Using Result 3.1, with u being replaced by $u_{n-1} + \alpha\rho^{n-1}e$ and u' being replaced by v, if we assume inductively that the right-hand side inequality at (3.29) is true, with n replaced by s, for all $0 \leqslant s \leqslant n-1$, then using (3.32) we have

$$v - u_n \leqslant T^{\delta}(u_{n-1} + \alpha\rho^{n-1}e) - T^{\delta}u_{n-1} = \rho P^{\delta}\alpha\rho^{n-1}e = \rho^n \alpha e. \tag{3.34}$$

For the left-hand side of (3.29) we have

$$v - u_n \geqslant T^{\sigma_n} v - T^{\sigma_n} v_{n-1}. \tag{3.35}$$

A similar analysis to the above now gives the left-hand side inequality of (3.29). □

From (3.29) we obtain the following convergence result.

Result 3.6 (White [58], Theorem 2.7). The sequence $\{u_n\}$ converges to v as n tends to infinity, with respect to the norm $\| \quad \|$.

Proof. If $\rho = 0$, the result is clearly true. Assume that $\rho \neq 0$. Let

$$\eta = \text{maximum}[|\alpha|, |\beta|] > 0. \tag{3.36}$$

Set $\eta > \varepsilon \geqslant 0$, and

$$n(\varepsilon) = \log(\varepsilon/\eta)/\log(\rho). \tag{3.37}$$

Then if $n \geqslant n(\varepsilon)$ we have

$$\alpha\rho^n \leqslant \varepsilon, \tag{3.38}$$

$$\beta\rho^n \geqslant -\varepsilon. \tag{3.39}$$

Then for $n \geqslant n(\varepsilon)$

$$-\varepsilon e \leqslant v - u_n \leqslant \varepsilon e. \tag{3.40}$$

Because ε is arbitrary we see that $\| v - v_n \|$ tends to zero as n tends to infinity. If $\eta = 0$ the result is trivially true. □

For some situations the convergence of $\{u_n\}$ will be monotone, e.g. if $r_i^k \geqslant 0$ for all $i \in I$, $k \in K(i)$ and $u = 0$ (see White [58], Theorem 2.6). We generalise this result as follows.

Result 3.7. Let $u_0 = u$ and u satisfy

$$u \leqslant Tu. \tag{3.41}$$

Then $\{u_n\}$ is non-decreasing in n.

Proof.

$$\underline{n \geqslant 2} \qquad u_n = Tu_{n-1} = T^{\sigma_n} u_{n-1}, \tag{3.42}$$

$$u_{n-1} = Tu_{n-2} = T^{\sigma_{n-1}} u_{n-2}. \tag{3.43}$$

Then

$$u_n - u_{n-1} \geqslant T^{\sigma_{n-1}} u_{n-1} - T^{\sigma_{n-1}} u_{n-2}. \tag{3.44}$$

Now

$$u_1 - u_0 = Tu_0 - u_0 = Tu - u \geqslant 0. \tag{3.45}$$

Assume inductively that

$$u_s = u_{s-1} \geqslant 0, \qquad 1 \leqslant s \leqslant n-1. \tag{3.46}$$

Then combining (3.44) and (3.46) we have, using Result (3.1)

$$u_n - u_{n-1} \geqslant 0. \tag{3.47}$$

○

As a special case, if $r_i^k \geqslant 0$ for all $i \in I$, $k \in K(i)$ and $u = 0$ then (3.41) holds. If we want monotone convergence we may transform the problem into one in which $r_i^k \geqslant 0$ (see White [58], p. 25) and condition (3.41) is satisfied with $u = 0$. This gives Theorem 2.6 of White [58].

The inequalities given in (3.29) are prior inequalities giving prior bounds, i.e. they are known before any computations are carried out. In practice, convergence may be much quicker (see Scherer and White [41]) and we may wish to assess how good a solution is in terms of the computational performance to date at any stage of the computations. This gives rise to the notion of posterior bounds.

Let

$$\underline{n \geqslant 1} \qquad \alpha_n = \underset{i \in I}{\text{maximum}} \; [u_n(i) - u_{n-1}(i)], \tag{3.48}$$

$$\beta_n = \underset{i \in I}{\text{minimum}} \; [u_n(i) - u_{n-1}(i)]. \tag{3.49}$$

We then have the following result.

Result 3.8 (White [58], p. 11, with $L = 1$, and Hastings and Mello [21]).

$$\underline{n \geqslant 1} \qquad u_n + (\rho/(1-\rho))\beta_n e \leqslant v \leqslant u_n + (\rho/(1-\rho))\alpha_n e. \tag{3.50}$$

Proof.

$$\underline{n \geqslant 1} \qquad u_n = Tu_{n-1} = T^{\sigma_n} u_{n-1}, \tag{3.51}$$

$$v = Tv = T^{\delta} v. \tag{3.52}$$

Then

$$\begin{aligned} v - u_n &\leqslant T^{\delta} v - T^{\delta} u_{n-1} \\ &= T^{\delta} v - T^{\delta} u_n + T^{\delta} u_n - T^{\delta} u_{n-1} \\ &= \rho P^{\delta}(v - u_n) + \rho P^{\delta}(u_n - u_{n-1}). \end{aligned} \tag{3.53}$$

Thus, with U as the identity matrix

$$v - u_n \leqslant (U - \rho P^{\delta})^{-1} \rho P^{\delta}(u_n - u_{n-1}). \tag{3.54}$$

This gives the right-hand side inequality of (3.50). The left-hand side inequality of (3.50) may likewise be determined. ○

Results 3.5 and 3.8 give us results about the behaviour of $\{u_n\}$. What we really want is an optimal, or approximately optimal, policy $\pi = (\delta)^{\infty}$, and we need bounds for the value function v^{π} of this policy. In particular, if we terminate the computations at stage n with $Tu_{n-1} = T^{\sigma_n} u_{n-1}$ we want to know how good the policy $\pi_n = (\sigma_n)^{\infty}$ is. We have the following result.

Result 3.9 (White [58], p. 111, with $L = 1$, and Hastings and Mello [21]).

$$\underline{n \geqslant 1} \qquad u_n + (\rho/(1-\rho))\beta_n e \leqslant v^{\pi_n} \leqslant v. \tag{3.55}$$

Proof. The right-hand side inequality (3.55) is obviously true because v is the optimal value function solution. Now

$$v^{\pi_n} = T^{\sigma_n} v^{\pi_n}, \qquad u_n = T^{\sigma_n} u_{n-1}. \tag{3.56}$$

Hence

$$\begin{aligned} v^{\pi_n} - u_n &= T^{\sigma_n} v^{\pi_n} - T^{\sigma_n} u_{n-1} \\ &= T^{\sigma_n} v^{\pi_n} - T^{\sigma_n} u_n + T^{\sigma_n} u_n - T^{\sigma_n} u_{n-1} \\ &= \rho P^{\sigma_n}(v^{\pi_n} - u_n) + \rho P^{\sigma_n}(u_n - u_{n-1}). \end{aligned} \tag{3.57}$$

Thus

$$\begin{aligned} v^{\pi_n} - u_n &= (U - \rho P^{\sigma_n})^{-1} \rho P^{\sigma_n}(u_n - u_{n-1}) \\ &\geqslant (U - \rho P^{\sigma_n})^{-1} \rho P^{\sigma_n} \beta_n e \\ &= (\rho/(1-\rho))\beta_n e. \end{aligned} \tag{3.58}$$

Result 3.10.

$$\underline{n \geqslant 1} \qquad v + (\rho/(1-\rho))(\beta_n - \alpha_n)e \leqslant v^{\pi_n} \leqslant v. \tag{3.59}$$

Proof. Combine Results 3.8 and 3.9.

Results 3.8, 3.9 and 3.10 are posterior-bound results in the sense that the error terms are evaluated at the current iteration in terms of the known u_n, u_{n-1}. For general use we may use the following result:

Result 3.11.

$$\underline{n \geqslant 2} \qquad \beta_n \geqslant \rho\beta_{n-1} \geqslant \rho^{n-1}\beta_1 = \rho^{n-1} \operatorname*{minimum}_{i \in I} [[Tu - u](i)], \tag{3.60}$$

$$\alpha_n \leqslant \rho\alpha_{n-1} \leqslant \rho^{n-1}\alpha_1 = \rho^{n-1} \operatorname*{maximum}_{i \in I} [[Tu - u](i)]. \tag{3.61}$$

Proof.

$$\underline{n \geqslant 2} \qquad u_n = T^{\sigma_n} u_{n-1}, \tag{3.62}$$

$$u_{n-1} = T^{\sigma_{n-1}} u_{n-2}. \tag{3.63}$$

Hence

$$\begin{aligned} u_n - u_{n-1} &\leqslant T^{\sigma_n} u_{n-1} - T^{\sigma_n} u_{n-2} \\ &= \rho P^{\sigma_n}(u_{n-1} - u_{n-2}). \end{aligned} \tag{3.64}$$

Clearly, (3.61) is true for $n = 1$. Let us assume it is true with n replaced by s for $1 \leqslant s \leqslant n-1$. If we substitute in (3.64) we obtain (3.61) for $s = n$. A similar analysis applies for (3.60). ○

So far we have concentrated on solving equation (3.1) by continued value iteration. It is of interest to see what would happen if we simply did one iteration and used the policy thus obtained. We have the following results the worthwhileness of which depends on how close $u_0 = u$ is to v in the first instance. If we are able to guess at a good u then one iteration will suffice.

For any u: $I \to R$ define

$$\operatorname{span}(u) = \underset{i \in I}{\text{maximum}} \; [u(i)] - \underset{i \in I}{\text{minimum}} \; [u(i)]. \tag{3.65}$$

Result 3.12 (Porteus [37]). Select u: $I \to R$. Let

$$Tu = T^{\delta} u \quad \text{and} \quad \pi = (\delta)^{\infty}. \tag{3.66}$$

Then

$$v \geqslant v^{\pi} \geqslant v - (\rho/(1-\rho))\operatorname{span}(Tu - u)e. \tag{3.67}$$

Proof. Clearly $v \geqslant v^{\pi}$. Then set $n = 1$ in (3.55) and (3.50) to give

$$\begin{aligned} v^{\pi} &\geqslant u_1 + (\rho/(1-\rho))\beta_1 e \\ &\geqslant v + (\rho/(1-\rho))(\beta_1 - \alpha_1)e. \end{aligned} \tag{3.68}$$

Now

$$\beta_1 - \alpha_1 = -\operatorname{span}(Tu - u). \tag{3.69}$$

Thus our result follows. ○

Example. Let us return to the example of Table 1.8 with $\rho = 0.9$. We retabulate in $\{u_n, \sigma_n\}$ form, in Table 3.1, the $\{v_n\}$ given in Table 2.1 with $u_0 = u = 0$.

The solution in equation (3.1) is given (see p. 43) by $u(1) = v(1) = 22.2$, $u(2) = v(2) = 12.3$, $\delta(1) = \delta(2) = 2$. We have $\bar{r} = 6$, $\underline{r} = -5$. Let us now check Results 3.4, 3.5, 3.7–3.12.

(i) *Result 3.4.* This requires that, for an optimal π in particular, $-5/(0.1) \leqslant v^{\pi}(i) \leqslant 6/(0.1)$, i.e. $-50 \leqslant v^{\pi}(i) \leqslant 60$, $\forall\ i \in I$.

(ii) *Result 3.5.* This requires that, with $\alpha = 22.2$, $\beta = 12.3$

$$\underline{n \geqslant 0} \qquad u_n(i) + 12.3(0.9)^n \leqslant v(i) \leqslant u_n(i) + 22.2(0.9)^n, \quad \forall\ i \in I.$$

(iii) *Result 3.7.* With $u_0 = u = 0$, (3.41) requires that

$$0 \leqslant \text{maximum} \begin{bmatrix} 6 \\ 4 \end{bmatrix},$$

$$0 \leqslant \text{maximum} \begin{bmatrix} -3 \\ -5 \end{bmatrix}.$$

This is not satisfied and $\{u_n\}$ is not non-decreasing in n. In fact, $u_0(2) = 0 > u_1(2) = -3$.

Let us tabulate $\{\alpha_n\}$, $\{\beta_n\}$ (see (3.48) and (3.49)) given in Table 3.2.

Table 3.1 Expected total discounted rewards and optimal decision rules for toymaker problem

			$u_n(i)$		$\sigma_n(i)$	
	State		i		i	
			1	2	1	2
		0	0	0	—	—
Iteration		1	6	−3	1	1
number	n	2	7.78	−2.03	2	2
		3	9.2362	−0.6467	2	2
		4	10.533658	0.644197	2	2

Table 3.2 $\{\alpha_n, \beta_n\}$ computations for Table 3.1

		α_n	β_n
Iteration number n	1	6	−3
	2	1.78	0.97
	3	1.4562	1.3562
	4	1.2974	1.290897

(iv) *Result 3.8.* This requires that

$$\underline{n \geqslant 1} \qquad u_n(i) + 9\beta_n \leqslant v(i) \leqslant u_n(i) + 9\alpha_n, \quad \forall\, i \in I.$$

(v) *Result 3.9.* This requires that

$$\underline{n \geqslant 1} \qquad u_n(i) + 9\beta_n \leqslant v^{\pi_n}(i) \leqslant v(i), \quad \forall\, i \in I.$$

From Table 1.11, identifying $\sigma_1 = \delta^1$, $\sigma_2 = \delta^4$, we have

$$v^{\pi_1}(1) = 15.5, \qquad v^{\pi_1}(2) = 5.6, \qquad v^{\pi_2} = v.$$

Policy π_2 is optimal.

(vi) *Result 3.10.* In addition to Result 3.9 this requires that

$$\underline{n \geqslant 1} \qquad v(i) + 9(\beta_n - \alpha_n) \leqslant v^{\pi_n}(i), \quad \forall\, i \in I.$$

(vii) *Result 3.11.* Requires that

$$\begin{aligned} \underline{n \geqslant 2} \qquad & \beta_n \geqslant (0.9)^{n-1}\beta_1 = -3(0.9)^{n-1}, \\ & \alpha_n \leqslant (0.9)^{n-1}\alpha_1 = 6(0.9)^{n-1}, \\ & \alpha_n \leqslant (0.9)\alpha_{n-1}, \\ & \beta_n \geqslant (0.9)\beta_{n-1}. \end{aligned}$$

(viii) *Result 3.12.* This requires that

$$\begin{aligned} v(i) \geqslant v^{\pi_1}(i) &\geqslant v(i) - 9(\alpha_1 - \beta_1) \\ &= v(i) - 81, \quad \forall\, i \in I. \end{aligned}$$

Although $\{u_n\}$ has far from converged in Table 3.1 for $1 \leqslant n \leqslant 4$, we see that $\{\sigma_n\}$ has apparently stabilised at $n = 2$. Such a stabilisation must take place for some finite n, although it is not known how to compute such an n because it depends on some prior knowledge of v. We have the following result.

Result 3.13 (White [58], pp. 103–104, Beckmann [3] and Shapiro [45]). Let σ_n, as defined in (3.31), satisfy $Tu_{n-1} = T^{\sigma_n}u_{n-1}$. Then there is an n_0 such that, for $n \geqslant n_0$, $\pi_n = (\sigma_n)^\infty$ is optimal for the infinite horizon problem, i.e. $v^{\pi_n} = v$.

Proof. For $i \in I$ let $N(i)$ be the set of all $k \in K(i)$ which are not optimal for state i for the infinite horizon problem. If $N(i) = \phi$ for all $i \in I$ then all policies are optimal and the result is trivially true. Assume $N(i) \neq \phi$ for some $i \in I$. Select $q \in K(i) \backslash N(i)$. Define

$$\varepsilon(i) = \underset{k \in N(i)}{\text{minimum}} \left[r_i^q + \rho \sum_{j \in I}' p_{ij}^q v(j) - r_i^k - \rho \sum_{j \in I}' p_{ij}^k v(j) \right]$$

$$= \underset{k \in N(i)}{\text{minimum}} \left[[T^q v](i) - [T^k v](i) \right]. \tag{3.70}$$

For $k \in N(i)$ and $q \in K(i) \backslash N(i)$ we have

$$[T^q v](i) = [Tv](i) > [T^k v](i). \tag{3.71}$$

Hence

$$\varepsilon(i) > 0. \tag{3.72}$$

For $k \in N(i)$, $q \in K(i) \backslash N(i)$

$$[T^q u_{n-1}](i) - [T^k u_{n-1}](i) = ([T^q u_{n-1}](i) - [T^q v](i)) \tag{3.73}$$

$$+ ([T^q v](i) - [T^k v](i)) \tag{3.74}$$

$$+ ([T^k v](i) - [T^k u_{n-1}](i)) \tag{3.75}$$

$$= A + B + C, \text{ say,} \tag{3.76}$$

where A, B, C are respectively the terms in (3.73), (3.74) and (3.75). We have

$$B \geqslant \varepsilon(i). \tag{3.77}$$

Using (3.29) and Result 3.3, restricted to Π_D being as singleton for this purpose, or using (3.15), we have

$$A \geqslant -\alpha\rho^n, \qquad C \geqslant -\beta\rho^n. \tag{3.78}$$

If $n^0(i)$ is any sufficiently large integer we will have

$$A \geqslant -\varepsilon(i)/3, \qquad C \geqslant -\varepsilon(i)/3 \tag{3.79}$$

for $n \geqslant n^0(i)$. Hence

$$[T^q u_{n-1}](i) - [T^k u_{n-1}](i) \geqslant \varepsilon(i)/3 \tag{3.80}$$

if $n \geqslant n_0(i)$. Hence if $n \geqslant n^0(i)$, no matter what $\sigma_n(i)$ is chosen,

$$k \neq \sigma_n(i). \tag{3.81}$$

Now let

$$n_0 = \underset{i \in I: N(i) \neq \emptyset}{\text{maximum}} [n_0(i)]. \tag{3.82}$$

Thus if $n \geqslant n_0$ and $k \in N(i)$ then $\sigma_n(i) = k$ cannot satisfy (3.31). For $n \geqslant n_0$, $\pi_n = (\sigma_n)^\infty$ is optimal for the infinite horizon problem

3.1.3 POLICY SPACE ITERATION (Bellman [4], p. 89, White [58], p. 25, Howard [23])

The policy space method is as follows:

(i) Select an initial policy $\pi^0 = (\sigma^0)^\infty$.
(ii) Solve the equation

$$u^0 = T^{\sigma^0} u^0 \tag{3.83}$$

for u^0.
(iii) Find a new policy $\pi^1 = (\sigma^1)^\infty$ by finding

$$\sigma^1 \in \arg \underset{\delta \in \Delta}{\text{maximum}} [T^\delta u^0]. \tag{3.84}$$

(iv) Replace σ^0 in (i) by σ^1 and repeat the procedure.

The procedure produces two sequences, viz. $\{\sigma^n\}$, $\{u^n\}$. The procedure terminates when at step (iii) we have

$$\sigma^n \in \arg \underset{\delta \in \Delta}{\text{maximum}} \ [T^\delta u^n] \tag{3.85}$$

or alternatively when

$$u^{n+1} = u^n. \tag{3.86}$$

Equations (3.85) and (3.86) are equivalent termination conditions. We have the following result.

Result 3.14 (White [58], Theorem 2.8, Mine and Osaki [34], p. 8, Howard [23]). The policy space method produces a sequence $\{u^n\}$ which converges non-decreasing to v in a finite number of iterations, with an optimal policy given by the terminating policy.

Proof.

$$\underline{n \geqslant 0} \qquad u^n = T^{\sigma^n} u^n, \tag{3.87}$$

$$u^{n+1} = T^{\sigma^{n+1}} u^{n+1}. \tag{3.88}$$

Hence

$$\begin{aligned} u^{n+1} - u^n &= T^{\sigma^{n+1}} u^{n+1} - T^{\sigma^n} u^n \\ &= (T^{\sigma^{n+1}} u^n - T^{\sigma^n} u^n), \end{aligned} \tag{3.89}$$

$$+ (T^{\sigma^{n+1}} u^{n+1} - T^{\sigma^{n+1}} u^n) = A^{n+1} + \rho P^{\sigma^{n+1}} (u^{n+1} - u^n) \tag{3.90}$$

where A^{n+1} is the term in (3.89). Hence

$$u^{n+1} - u^n = (U - \rho P^{\sigma^{n+1}})^{-1} A^{n+1} \tag{3.91}$$

where U is the identity matrix.

From the generalised step (iii) we have

$$\sigma^{n+1} \in \arg \underset{\delta \in \Delta}{\text{maximum}} \ [T^\delta u^n]. \tag{3.92}$$

Hence

$$A^{n+1} \geqslant 0. \tag{3.93}$$

Thus

$$u^{n+1} \geqslant u^n. \tag{3.94}$$

Using the boundedness of $\{u^n\}$ (see Result 3.4 with $v^\pi = u^n$) and the convergence of bounded monotonic sequences (e.g. see Bromwich [8], p. 409) this is enough to show that $\{u^n\}$ converges non-decreasing to some function v^* in norm $\| \quad \|$.

If, for termination, we use (3.86) then $A_{n+1} = 0$ (see (3.91)). Then $\sigma^n \in \arg \underset{\delta \in \Delta}{\text{maximum}} [T^\delta u^n]$ and (3.85) may be used for termination purposes equally well.

If, for termination, we use (3.85) then again $A^{n+1} = 0$ (see (3.89) with $\sigma^n = \sigma^{n+1}$) and then $u^{n+1} = u^n$ (see (3.91)) and we may use (3.86) equally well for termination purposes. Thus, termination rules (3.85) and (3.86) are equivalent, i.e.

$$u^{n+1} = u^n \rightleftarrows A^{n+1} = 0. \tag{3.95}$$

We thus terminate at some n with $u^n = v^*$, $\sigma^n \in \arg \underset{\delta \in \Delta}{\text{maximum}} [T^\delta v^*]$, i.e.

$$v^* = T^{\sigma^n} v^* = Tv^*. \tag{3.96}$$

From Result 2.5 and equation (3.96) we have

$$v^* = v. \tag{3.97}$$

Clearly $\pi^n = (\sigma^n)^\infty$ is optimal. ○

We may, if we wish, terminate before conditions (3.85) or (3.86) are satisfied, with some appropriate stopping rule. In this case we may use Result 3.12 with $u = u^n$, $\pi = (\sigma^{n+1})^\infty$, to determine the potential loss of optimality in doing so.

Example. We will use the example of Table 1.8 (see Howard [23], p. 85) with $\rho = 0.9$.

(i) Select $\sigma^0 = (1, 1)$, i.e. $\sigma^0(1) = \sigma^0(2) = 1$.

(ii) Solve

$$u^0 = \begin{bmatrix} 6 \\ -3 \end{bmatrix} + (0.9) \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix} u^0$$

to give

$$u^0 = \begin{bmatrix} 15.5 \\ 5.6 \end{bmatrix}.$$

(iii) Find

$$\sigma^1(1) \in \text{arg maximum}\begin{bmatrix} k = 1\text{: } 6 + 0.9(0.5 \times 15.5 + 0.5 \times 5.6) \\ k = 2\text{: } 4 + 0.9(0.8 \times 15.5 + 0.2 \times 5.6) \end{bmatrix},$$

i.e.

$$\sigma^1(1) \in \text{arg maximum}\begin{bmatrix} k = 1\text{: } 15.495 \\ k = 2\text{: } 16.168 \end{bmatrix} = \{2\}.$$

Thus

$$\sigma^1(1) = 2.$$

Find

$$\sigma^1(2) \in \text{arg maximum}\begin{bmatrix} k = 1\text{: } -3 + 0.9(0.4 \times 15.5 + 0.6 \times 5.6) \\ k = 2\text{: } -5 + 0.9(0.7 \times 15.5 + 0.3 \times 5.6) \end{bmatrix},$$

i.e.

$$\sigma^1(2) \in \text{arg maximum}\begin{bmatrix} k = 1\text{: } 5.064 \\ k = 2\text{: } 6.277 \end{bmatrix} = \{2\}.$$

Thus

$$\sigma^1(2) = 2.$$

(i) $\sigma^1 = (2, 2)$.

(ii) Solve

$$u^1 = \begin{bmatrix} 4 \\ -5 \end{bmatrix} + (0.9)\begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix} u^1$$

to give

$$u^1 = \begin{bmatrix} 22.2 \\ 12.3 \end{bmatrix}.$$

(iii) Find

$$\sigma^2(1) \in \text{arg maximum}\begin{bmatrix} k = 1\text{: } 6 + 0.9(0.5 \times 22.2 + 0.5 \times 12.3) \\ k = 2\text{: } 4 + 0.9(0.8 \times 22.2 + 0.2 \times 12.3) \end{bmatrix},$$

i.e.

$$\sigma^2(1) \in \text{arg maximum}\begin{bmatrix} k = 1\text{: } 15.525 \\ k = 2\text{: } 22.2 \end{bmatrix} = \{2\}.$$

Thus

$$\sigma^2(1) = 2.$$

Find

$$\sigma^2(2) \in \text{arg maximum} \begin{bmatrix} k = 1: & -3 + 0.9(0.4 \times 22.2 + 0.6 \times 12.3) \\ k = 2: & -5 + 0.9(0.7 \times 22.2 + 0.3 \times 12.3) \end{bmatrix},$$

i.e.

$$\sigma^2(2) \in \text{arg maximum} \begin{bmatrix} k = 1: & 11.635 \\ k = 2: & 12.3 \end{bmatrix} = \{2\}.$$

$\sigma^2(2) = 2$. Thus $\sigma^2 = \sigma^1$ and we terminate using (3.85) giving an optimal solution

$$v^* = v = \begin{bmatrix} 22.2 \\ 12.3 \end{bmatrix}, \qquad \delta = \sigma^2 = (2, 2).$$

3.1.4 VALUE ITERATION AND POLICY SPACE ITERATION INTERRELATIONSHIP

The value iteration scheme (3.26) and the policy space iteration scheme (3.83)–(3.86) are actually extreme forms of a more general scheme (see Puterman [39], pp. 91–130).

For the policy space iteration method we have

$$\underline{n \geq 0} \qquad Tu^n = T^{\sigma^{n+1}} u^n = r^{\sigma^{n+1}} + \rho P^{\sigma^{n+1}} u^n, \tag{3.98}$$

$$u^{n+1} = T^{\sigma^{n+1}} u^{n+1} = r^{\sigma^{n+1}} + \rho P^{\sigma^{n+1}} u^{n+1}. \tag{3.99}$$

Thus, with U as the identity matrix

$$\begin{aligned} u^{n+1} &= (U - \rho P^{\sigma^{n+1}})^{-1} r^{\sigma^{n+1}} \\ &= (U - \rho P^{\sigma^{n+1}})^{-1} (T - \rho P^{\sigma^{n+1}}) u^n \\ &= (U - \rho P^{\sigma^{n+1}})^{-1} ((T - U)u^n + (U - \rho P^{\sigma^{n+1}}) u^n) \\ &= (U + (U - \rho P^{\sigma^{n+1}})^{-1} (T - U)) u^n. \end{aligned} \tag{3.100}$$

If we expand $(U - \rho P^{\sigma^{n+1}})^{-1}$ to s terms we have a scheme

$$u_s^{n+1} = U + \sum_{l=0}^{s-1} \rho^l (P^{\sigma_s^{n+1}})^l (T - U) u_s^n \tag{3.101}$$

with

$$\sigma_s^{n+1} \in \text{arg maximum}_{\delta \in \Delta} [T^\delta u_s^n]. \tag{3.102}$$

Value iteration and policy space iteration correspond to $s = 1$ and to $s = \infty$ respectively.

3.2 INFINITE HORIZON AVERAGE EXPECTED REWARD PER UNIT TIME. STATIONARY CASE

In accordance with our earlier assumption we will only consider the uni-chain case. In this case we wish to find a solution (u, h, δ) to the following equations (see (2.85) and (2.86)):

$$u + he = \underset{\delta \in \Delta}{\text{maximum}} \; [r^\delta + P^\delta u] \tag{3.103}$$

$$= \underset{\delta \in \Delta}{\text{maximum}} \; [T^\delta u]$$

$$= Tu, \tag{3.104}$$

$$u(m) = 0. \tag{3.105}$$

Then an optimal policy is $\pi = (\delta)^\infty$ and the optimal gain g is equal to h.

As before, with the discounted case, we will look at value iteration and policy space iteration. Because we deal only with the uni-chain case, g and g^π will be scalar gains in what follows.

3.2.1 VALUE ITERATION

This takes the form

$$u_0 = u, \text{ arbitrary}, \tag{3.106}$$

$$\underline{n \geqslant 1} \qquad u_n = Tu_{n-1}. \tag{3.107}$$

Because the discount factor is, in effect, $\rho = 1$, $\{u_n\}$ may increase without bound. In (1.12) we showed, for a specific example and policy π, that the value function, gain and bias function took the form

$$v_n^\pi = ng^\pi e + w^\pi + \varepsilon_n^\pi \tag{3.108}$$

with ε_n^π tending to zero as n tended to infinity.

We will make a stronger assumption, viz. that for forms (3.106) and (3.107)

$$\underline{n \geqslant 0} \qquad u_n = nge + w + \varepsilon_n \tag{3.109}$$

where

$$\lim_{n \to \infty} [\varepsilon_n] = 0. \tag{3.110}$$

Conditions for (3.110) to hold are given in White [58], Theorem 3.11, leading to the form (98), p. 44 of White [58]. A more general set of conditions is given by Schweitzer and Federgruen [42].

In particular, the regular case of p. 45 will suffice, viz.

$$\lim_{n \to \infty} [(P^{\delta})^n] \text{ exists}, \quad \forall\, \delta \in \Delta. \tag{3.111}$$

Let us define $\{\alpha_n, \beta_n\}$ as in (3.48) and (3.49) for $\rho = 1$, viz.

$$\underline{n \geqslant 1} \qquad \alpha_n = \max_{i \in I} [u_n(i) - u_{n-1}(i)], \tag{3.112}$$

$$\beta_n = \min_{i \in I} [u_n(i) - u_{n-1}(i)]. \tag{3.113}$$

We first of all prove a monotonicity result for $\{\alpha_n, \beta_n\}$.

Result 3.15. The sequence $\{\alpha_n\}$ converges non-increasing to some limit α as n tends to ∞, and the sequence $\{\beta_n\}$ converges non-decreasing to some limit β as n tends to ∞. Also

$$\underline{n \geqslant 1} \qquad \beta_n \leqslant \beta \leqslant \alpha \leqslant \alpha_n. \tag{3.114}$$

Proof.

$$\begin{aligned} \underline{n \geqslant 2} \qquad u_n - u_{n-1} &= Tu_{n-1} - Tu_{n-2} \\ &= T^{\sigma_n} u_{n-1} - Tu_{n-2} \\ &\leqslant T^{\sigma_n} u_{n-1} - T^{\sigma_n} u_{n-2} \\ &= P^{\sigma_n}(u_{n-1} - u_{n-2}) \\ &\leqslant \alpha_{n-1} e. \end{aligned} \tag{3.115}$$

Thus

$$\alpha_n \leqslant \alpha_{n-1}. \tag{3.116}$$

Similarly

$$\beta_n \geqslant \beta_{n-1}. \tag{3.117}$$

Clearly

$$\beta_n \leqslant \alpha_n. \tag{3.118}$$

Thus

$$\beta_{n-1} \leqslant \beta_n \leqslant \alpha_n \leqslant \alpha_{n-1}. \tag{3.119}$$

Here $\{\alpha_n\}$ is bounded below by β_1, $\{\beta_n\}$ is bounded above by α_1. Thus, using the monotone convergence result of Bromwich [8] (see p. 73), our requisite result follows.

We are now in a position to prove the following result.

Result 3.16 (Odoni ([38]). Under conditions (3.109) and (3.110)

$$\underline{n \geqslant 1} \qquad \beta_n \leqslant g \leqslant \alpha_n, \tag{3.120}$$

$$\alpha = g = \beta. \tag{3.121}$$

Proof. From (3.109) and (3.110) $\{\alpha_n\}, \{\beta_n\}$ tend to g as n tends to infinity. Because the convergence is monotone the requisite result holds.

Result 3.16 only gives us bounds on the value of g. It does not tell us how good the policy $\pi_n = (\sigma_n)^\infty$ might be. We would like a result similar to inequality (3.55) for discounted problems. We can do this but its use depends upon the behaviour of $\{\varepsilon_n\}$ in (3.109). We will give a result, but only in special cases do we know sufficient about how $\{\varepsilon_n\}$ behaves for this result to hold (e.g. see Exercise (22) of Chapter 9 of White [58], under the conditions of Theorem 3.11).

Result 3.17. Under conditions (3.109) and (3.110)

$$\underline{n \geqslant 1} \qquad g - 2\| \varepsilon_{n-1} \| \leqslant g^{\pi_n} \leqslant g. \tag{3.122}$$

Proof. We have

$$\underline{n \geqslant 1} \qquad u_n = Tu_{n-1}. \tag{3.123}$$

Thus, using (3.109) we have

$$nge + w + \varepsilon_n = T((n-1)ge + w + \varepsilon_{n-1}). \tag{3.124}$$

Thus

$$w + ge = T(w + \varepsilon_{n-1}) - \varepsilon_n. \tag{3.125}$$

Using (3.110) we obtain

$$w + ge = Tw. \tag{3.126}$$

For policy π_n we use (2.118) and (2.119), restricting Π_D to a singleton $\{\pi_n\}$, and obtain for some $u = u^{\pi_n}$ with $h = g^{\pi_n}$

$$u^{\pi_n} + g^{\pi_n} e = T^{\sigma_n} u^{\pi_n}. \tag{3.127}$$

Note that we are not assuming that u^{π_n} takes a similar form to (3.109). Now

$$T^{\sigma_n} u_{n-1} = Tu_{n-1}. \tag{3.128}$$

Thus

$$T^{\sigma_n}((n-1)ge + w + \varepsilon_{n-1}) = T((n-1)ge + w + \varepsilon_{n-1}). \tag{3.129}$$

Hence

$$T^{\sigma_n}(w + \varepsilon_{n-1}) = T(w + \varepsilon_{n-1}). \tag{3.130}$$

From (3.130) we have

$$T^{\sigma_n} w \geqslant Tw - 2 \| \varepsilon_{n-1} \| e. \tag{3.131}$$

Combining (3.131) and (3.126) we have

$$w + ge \leqslant T^{\sigma_n} w + 2 \| \varepsilon_{n-1} \| e. \tag{3.132}$$

Combining (3.127) and (3.132) we have

$$(u^{\pi_n} - w) + (g^{\pi_n} - g)e \geqslant P^{\sigma_n}(u^{\pi_n} - w) - 2 \| \varepsilon_{n-1} \| e. \tag{3.133}$$

From (2.117) we have a vector θ^{σ_n} such that

$$\theta^{\sigma_n} P^{\sigma_n} = \theta^{\sigma_n} \tag{3.134}$$

with

$$\theta^{\sigma_n} \geqslant 0 \tag{3.135}$$

and

$$\sum_{i \in I} \theta_i^{\sigma_n} = 1. \tag{3.136}$$

Taking the scalar product of both sides of inequality (3.133) and θ^{σ_n} we obtain

$$g^{\pi_n} - g \geqslant -2 \| \varepsilon_{n-1} \|. \tag{3.137}$$

Thus

$$g^{\pi_n} \geqslant g - 2 \| \varepsilon_{n-1} \|. \tag{3.138}$$

In addition, we automatically have

$$g \geqslant g^{\pi_n}. \tag{3.139}$$

○

An alternative way of avoiding the possible infinite limit of $\{u_n\}$ is to use the relative value approach (see White [58], pp. 39–44).

We will assume that the form (3.109) holds together with (3.110), viz.

$$\underline{n \geqslant 0} \quad u_n = nge + w + \varepsilon_n, \tag{3.140}$$

$$\lim_{n \to \infty} [\varepsilon_n] = 0. \tag{3.141}$$

Then

$$u_n(i) - u_n(m) = w(i) - w(m) + \varepsilon_n(i) - \varepsilon_n(m). \tag{3.142}$$

These relative values $\{u_n(i) - u_n(m)\}$ will be bounded and we transform our equation (3.107).

We make the following transformations which are equivalent to those of White [58], pp. 39–43:

$$\tilde{u}_n(i) = u_n(i) - u_n(m), \quad \forall\, i \in I. \tag{3.143}$$

Equation (3.107) becomes

$$\tilde{u}_n(i) + u_n(m) = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k + \sum_{j \in I} p_{ij}^k \tilde{u}_{n-1}(j) + u_{n-1}(m) \right], \quad \forall\, i \in I. \tag{3.144}$$

Equation (3.144) becomes

$$\tilde{u}_n(i) + (u_n(m) - u_{n-1}(m)) = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k + \sum_{j \in I} p_{ij}^k \tilde{u}_{n-1}(j) \right], \quad \forall\, i \in I. \tag{3.145}$$

This takes the form

$$\underline{n \geq 1} \qquad \tilde{u}_n + g_n e = T\tilde{u}_{n-1}, \tag{3.146}$$

$$g_n = u_n(m) - u_{n-1}(m), \tag{3.147}$$

$$\tilde{u}_n(m) = 0. \tag{3.148}$$

We see that

$$\tilde{u}_n(i) = w(i) - w(m) + \varepsilon_n(i) - \varepsilon_n(m), \quad \forall\, i \in I. \tag{3.149}$$

Hence $\{\tilde{u}_n\}$ converges and (3.146)–(3.148) will solve our problem when used as an iterative procedure. In the limit we have

$$\lim_{n \to \infty} \{\tilde{u}_n\} = \tilde{w}, \tag{3.150}$$

$$\lim_{n \to \infty} \{g_n\} = g, \tag{3.151}$$

$$\tilde{u}(m) = 0 \tag{3.152}$$

where $(\tilde{w}, g)$ satisfy (3.103)–(3.105).

Inequality (3.122) also applies for this procedure.

Example. Let us do some calculations for the problem of Table 1.8, part of which is reproduced in Table 3.3.

Table 3.4 gives the value iterations.

From Table 1.10 we see that $g = 2$, $u(1) = 10$, $u(2) = 0$. We have the following tabulations in Table 3.5 after finding $w(1) = 4.222$, $w(2) = -5.778$ using z-transform analysis:

Table 3.3 Data for toymaker problem (reproduced from [23] Howard (1960), p. 40, by permission of The MIT Press)

State i	*Action* k	*Transition probability* p_{ij}^k		*Expected reward* r_i^k
1	1	0.5	0.5	6
	2	0.8	0.2	4
2	1	0.4	0.6	−3
	2	0.7	0.3	−5

Table 3.4 Value iteration for toymaker problem

			$u_n(i)$		$\sigma_n(i)$	
	State		*i*		*i*	
			1	2	1	2
		0	0	0	—	—
Iteration		1	6	−3	1	1
number	*n*	2	8.2	−1.7	2	2
		3	10.22	0.23	2	2
		4	12.222	2.223	2	2
		5	14.222	4.223	2	2

Table 3.5 Value differences and errors for toymaker problem

	Iteration number *n*				
	1	2	3	4	5
α_n	6	2.2	2.02	2.002	2.000
β_n	−2	1.4	1.93	1.993	2.000
$\varepsilon_n(1)$	−0.222	−0.022	−0.002	−0.000	−0.000
$\varepsilon_n(2)$	1.222	0.122	0.012	0.001	0.000

The results in Table 3.5 are obtained by noting that, with $\delta = (1, 1)$, $\tau = (2, 2)$, then

$$\underline{n \geqslant 1} \qquad u_n = (T^\tau)^{n-1} T^\delta 0$$

$$= \sum_{l=0}^{n-2} (P^\tau)^l r^\tau + (P^\tau)^{n-1} P^\delta r^\delta.$$

Although the optimal decision rules have been only demonstrated for $1 \leqslant n \leqslant 5$, it is also true that τ is optimal for all $n \geqslant 2$.

The general form of $(P^\tau)^l$ is given in Exercise 2 of Chapter 1, viz.

$$(P^\tau)^l = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix} + (0.1)^l \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Then

$$n \geqslant 2 \qquad u_n = (n-1)\begin{bmatrix} \frac{7}{9} & \frac{2}{9} \\ \frac{7}{9} & \frac{2}{9} \end{bmatrix}\begin{bmatrix} 4 \\ -5 \end{bmatrix} + \begin{bmatrix} \frac{7}{9} & \frac{2}{9} \\ \frac{7}{9} & \frac{2}{9} \end{bmatrix}\begin{bmatrix} 6 \\ -3 \end{bmatrix}$$

$$+ \left(\sum_{l=0}^{n-2} (0.1)^l\right)\begin{bmatrix} \frac{2}{9} & -\frac{2}{9} \\ -\frac{7}{9} & \frac{7}{9} \end{bmatrix}\begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

$$+ (0.1)^{n-1}\begin{bmatrix} \frac{2}{9} & -\frac{2}{9} \\ -\frac{7}{9} & \frac{7}{9} \end{bmatrix}\begin{bmatrix} 6 \\ -3 \end{bmatrix}$$

$$= (n-1)\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 4 \\ 4 \end{bmatrix} + (\tfrac{10}{9})(1-(0.1)^{n-1})\begin{bmatrix} 2 \\ -7 \end{bmatrix}$$

$$+ (0.1)^{n-1}\begin{bmatrix} 2 \\ -7 \end{bmatrix}$$

$$= n\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 4\frac{2}{9} \\ -5\frac{7}{9} \end{bmatrix} + (0.1)^{n-1}\begin{bmatrix} -\frac{2}{9} \\ \frac{7}{9} \end{bmatrix}.$$

Thus

$$\varepsilon_n = (0.1)^{n-1}\begin{bmatrix} -\frac{2}{9} \\ \frac{7}{9} \end{bmatrix}, \qquad n \geqslant 2.$$

A check shows that this is true also for $n = 1$.

The actual bias function is

$$w = \begin{bmatrix} 4\frac{2}{9} \\ -5\frac{7}{9} \end{bmatrix},$$

which is

$$\begin{bmatrix} 5\frac{7}{9} \\ 5\frac{7}{9} \end{bmatrix} \text{ less than } u = \begin{bmatrix} 10 \\ 0 \end{bmatrix}.$$

We now check our results.

(i) *Results 3.15 and 3.16.* These hold with

$$\lim_{n\to\infty} [\alpha_n] = \lim_{n\to\infty} [\beta_n] = 2 = g.$$

(ii) *Result 3.17.* This is clearly true for $n \geqslant 2$ because $\sigma_n = (2, 2)$ for $n \geqslant 2$. For $n = 1$ it is also easily checked. We have $\varepsilon_0(1) = -4.222$,

$\varepsilon_0(2) = 5.778$, $g^{\pi_1} = 1$. We have not tabulated the results for relative values (3.146)–(3.148) but they are derivable from the results just obtained.

Under conditions (3.109) and (3.110) we have an analogous result to Result 3.13 for the discounted value iteration procedure, which we will not prove. We simply state it.

Result 3.18 (White [58], p. 104, and Beckmann [3], pp. 51–52). Let conditions (3.109) and (3.110) hold. Let σ_n satisfy $Tu_{n-1} = T^{\sigma_n}u_{n-1}$ or, equivalently, $T\tilde{w}_{n-1} = T^{\sigma_n}\tilde{w}_{n-1}$. Then there is an n_0 such that for $n \geqslant n_0$, $\pi_n = (\sigma_n)^\infty$ is optimal for the infinite horizon problem, i.e. $g^{\pi_n} = g$. ○

3.2.2 POLICY SPACE ITERATION (White [58], p. 38, Howard [23])

This is similar to the one for discounted processes given on pp. 71–73 viz.

(i) Select an initial policy $\pi^0 = (\sigma^0)^\infty$.

(ii) Solve the equations

$$u^0 + h^0 e = T^{\sigma^0}u^0, \tag{3.153}$$

$$u^0(m) = 0 \tag{3.154}$$

for (u^0, h^0).

(iii) Find a new policy $\pi^1 = (\sigma^1)^\infty$ by finding

$$\sigma^1 \in \arg \underset{\delta \in \Delta}{\text{maximum}} \, [T^\delta u^0]. \tag{3.155}$$

(iv) Replace σ^0 in (i) by σ^1 and repeat the procedure.

The procedure produces three sequences $\{\sigma^n\}$, $\{u^n\}$, $\{h^n\}$. The procedure terminates when at step (iii) we have

$$\sigma^n \in \arg \underset{\delta \in \Delta}{\text{maximum}} \, [T^\delta u^n] \tag{3.156}$$

or alternatively when

$$h^{n+1} = h^n. \tag{3.157}$$

Equations (3.156) and (3.157) are equivalent as termination rules when no transient states exist. We have the following result.

Result 3.19 (White [58], Theorem 3.9, Howard [23]). If no transition probability matrix P^{δ}, for any $\delta \in \Delta$, has any transient states then the policy space iteration method produces a sequence $\{h^n\}$ which converges increasing to g in a finite number of iterations, and the terminating policy is optimal.

Proof.

$$u^n + h^n e = T^{\sigma^n} u^n, \tag{3.158}$$

$$u^{n+1} + h^{n+1} e = T^{\sigma^{n+1}} u^{n+1}. \tag{3.159}$$

Hence

$$\begin{aligned}(u^{n+1} - u^n) + (h^{n+1} - h^n)e &= T^{\sigma^{n+1}} u^{n+1} - T^{\sigma_n} u^n \\ &= (T^{\sigma^{n+1}} u^n - T^{\sigma^n} u^n) + T^{\sigma^{n+1}} u^{n+1} - T^{\sigma^{n+1}} u^n \\ &= B^{n+1} + P^{\sigma^{n+1}}(u^{n+1} - u^n)\end{aligned} \tag{3.160}$$

where B^{n+1} is the term in parentheses. Hence, with U as the identity matrix

$$(U - P^{\sigma^{n+1}})(u^{n+1} - u^n) + (h^{n+1} - h^n)e = B^{n+1}. \tag{3.161}$$

From (2.117) we have a vector $\theta^{\sigma^{n+1}}$ such that

$$\theta^{\sigma^{n+1}} P^{\sigma^{n+1}} = \theta^{\sigma^{n+1}} \tag{3.162}$$

with

$$\sum_{i \in I} \theta_i^{\sigma^{n+1}} = 1 \tag{3.163}$$

and, because we have no transient states,

$$\theta_i^{\sigma^{n+1}} > 0, \quad \forall\, i \in I. \tag{3.164}$$

Combining (3.161)–(3.163) we obtain

$$h^{n+1} - h^n = \theta^{\sigma^{n+1}} B^{n+1}. \tag{3.165}$$

Because $\theta^{\sigma^{n+1}} > 0$ (see p. 59 for $>$), we see that

$$h^{n+1} = h^n \rightleftarrows B^{n+1} = 0. \tag{3.166}$$

This gives the equivalence of (3.156) and (3.157) as termination rules.

From (3.165) we see that

$$h^{n+1} > h^n \tag{3.167}$$

and $\{h^n\}$ is increasing until the termination point.

Policies cannot repeat, as a result of this monotonicity property. Hence $\{h^n\}$ converges in a finite number of iterations to some value, h^* say, because we have only a finite number of policies $\pi = (\delta)^\infty$, $\delta \in \Delta$.

At the terminal value of n we have, for some (u^*, h^*)

$$h^n = h^*, \qquad u^n = u^* \tag{3.168}$$

and

$$u^* + h^* e = T^{\sigma^n} u^* = Tu^* \tag{3.169}$$

by virtue of (3.156).

For the optimal gain g we have for some u' (see Result 2.10)

$$u' + ge = Tu'. \tag{3.170}$$

Thus, if $\delta \in \arg \underset{\delta \in \Delta}{\text{maximum}} [T^\delta u']$ then

$$\begin{aligned}(u^* - u') + (h^* - g)e &= Tu^* - Tu' \\ &= Tu^* - T^\delta u' \\ &\geqslant T^\delta u^* - T^\delta u' \\ &= P^\delta (u^* - u').\end{aligned} \tag{3.171}$$

Following an analysis similar to that giving (3.165) and (3.167) we have

$$h^* \geqslant g. \tag{3.172}$$

Then, because $h^* = g^{\pi^*} \leqslant g$ we have

$$h^* = g^{\pi^*} = h^n = g. \tag{3.173}$$

Let us now look again at (3.165), viz.

$$h^{n+1} - h^n = \theta^{\sigma^{n+1}} B^{n+1}. \tag{3.174}$$

We have assumed that $\theta^{\sigma^{n+1}} > 0$.

However, it is possible to have, without this restriction

$$B^{n+1} \neq 0, \qquad \theta_i^{\sigma^{n+1}} \times B^{n+1}(i) = 0, \quad \forall\, i \in I. \tag{3.175}$$

It is then possible to continue the iterations if we use termination rule (3.156), but have $h^{n+1} = h^n$ and $g^{\pi^n} = h^n$ not optimal.

Let us now use only termination procedure (3.156) and, in addition, not allow any policies to be repeated except at a termination iteration. We then have Result 3.20 which we will not prove.

Result 3.20 (White [58], Theorem 3.10). Under the modified algorithm, the results of Result 3.19 hold without the no-transient-state condition, with $\{h^n\}$ non-decreasing rather than increasing.

Result 3.19 uses the no-transient-state condition in order to cater for terminating conditions (3.156) or (3.157) as equivalent conditions. However, Result 3.19 is valid, without the no-transient-state condition, if only terminating condition (3.156) is used. This is given by Blackwell [7], Theorem 4. The analysis is carried out by looking at the behaviour of v_ρ (see p. 49) in the region of $\rho = 1^-$. The proof is a little complicated and hence we have included Result 3.20, which is more easily demonstrated.

We may terminate the procedure at any iteration, and we have an analogous result to that of Result 3.12. If we terminate with $\pi^n = (\sigma^n)^\infty$ and if we set $u^{n-1} = u$ in the following result we obtain a bound on the loss of optimality if π^n is used.

Result 3.21. Select u: $I \to R$. Let

$$Tu = T^\delta u \quad \text{and} \quad \pi = (\delta)^\infty. \tag{3.176}$$

Then if $\arg \underset{\delta \in \Delta}{\text{maximum}} [T^\delta u]$ is a singleton

$$g \geqslant g^\pi \geqslant g - \text{span}(Tu - u). \tag{3.177}$$

Proof. Inequality (3.67) for the discounted problem gives (using suffix ρ for v, T and π)

$$v_\rho \geqslant v_\rho^{\pi_\rho} \geqslant v_\rho - (\rho/(1-\rho))\text{span}(T_\rho u - u)e. \tag{3.178}$$

Now use (2.104) in a similar manner to obtain

$$\lim_{s \to \infty} [(1-\rho_s)v_{\rho_s}] = g, \qquad \lim_{s \to \infty} [(1-\rho_s)v_{\rho_s}^{\pi_{\rho_s}}] = g^\pi \tag{3.179}$$

where $\{\rho_s\}$ is a sequence tending to 1^-. The second equality in (3.179)

arises from the fact that π_{ρ_s} may be taken to be fixed if s is large enough. It then has a limit $\pi^* = (\delta)^\infty$ with $\delta \in \arg\operatorname{maximum}_{\delta \in \Delta} [T^\delta u]$. Because the latter is a singleton by assumption, this limit is the same as π.

We then obtain the requisite result if we combine (3.178) and (3.179). ○

Example. Let us do the example given in Table 3.3 which we repeat as Table 3.6 (Howard [23], p. 40).

(i) Select $\sigma^0 = (1, 1)$, $\pi^0 = (\sigma^0)^\infty$.

(ii) Solve

$$\begin{aligned} u^0(1) + h^0 &= 6 + 0.5u^0(1) + 0.5u^0(2), \\ u^0(2) + h^0 &= -3 + 0.4u^0(1) + 0.6u^0(2), \\ u^0(2) &= 0. \end{aligned}$$

We obtain $h^0 = 1$, $u^0(1) = 10$, $u^0(2) = 0$.

(iii) Find

$$\sigma^1(1) \in \arg\operatorname{maximum} \begin{bmatrix} k = 1\colon\ 6 + 0.5 \times 10 + 0.5 \times 0 \\ k = 2\colon\ 4 + 0.8 \times 10 + 0.2 \times 0 \end{bmatrix},$$

i.e.

$$\sigma^1(1) \in \arg\operatorname{maximum} \begin{bmatrix} k = 1\colon\ 11 \\ k = 2\colon\ 12 \end{bmatrix} = \{2\}.$$

Table 3.6 Data for toymaker problem (reproduced from [23] Howard (1960), p. 40, by permission of The MIT Press)

State i	*Action* k	*Transition probability* p_{ij}^k		*Expected reward* r_i^k
1	1	0.5	0.5	6
	2	0.8	0.2	4
2	1	0.4	0.6	−3
	2	0.7	0.3	−5

Find

$$\sigma^1(2) \in \arg\ \text{maximum} \begin{bmatrix} k=1: & -3 + 0.4 \times 10 + 0.6 \times 0 \\ k=2: & -5 + 0.7 \times 10 + 0.3 \times 0 \end{bmatrix},$$

i.e.

$$\sigma^1(2) \in \arg\ \text{maximum} \begin{bmatrix} k=1: & 1 \\ k=2: & 2 \end{bmatrix} = \{2\}.$$

Thus

$$\sigma^1 = (2, 2).$$

(ii) Solve

$$\begin{aligned} u^1(1) + h^1 &= 4 + 0.8 \times u^1(1) + 0.2 \times u^1(2), \\ u^1(2) + h^1 &= -5 + 0.7 \times u^1(1) + 0.3 \times u^1(2), \\ u^1(2) &= 0. \end{aligned}$$

We obtain $h^1 = 2$, $u^1(1) = 10$, $u^1(2) = 0$.

(iii) Hence $u^1 = u^0$ and we see that

$$\sigma^1 \in \arg \max_{\delta \in \Delta} [T^\delta u^1 (= T^\delta u^0)].$$

Hence from (3.156) we terminate with

$$\begin{aligned} \delta &= (2, 2), \qquad g = h = 2, \\ u(1) &= 10, \qquad u(2) = 0. \end{aligned}$$

3.3 ABSORBING STATE PROBLEMS. STATIONARY CASE

We wish to solve the equations (see (2.127) and (2.128))

$$u = Tu, \tag{3.180}$$

$$u(i) = 0, \quad \forall\ i \in I_a. \tag{3.181}$$

We will not deal in detail with this. Under the condition (2.126) the value iteration and policy space iteration results given for the discounted problem all hold, viz. Results 3.6, 3.7, 3.13 and 3.14.

Detailed analyses may be found in Derman [15], Chapter 5, Mine and Osaki [34], pp. 42–44 and in White [58], pp. 9–16, under various

conditions. The rates of convergence depend upon the nature of the transition matrices. For discounted absorbing state problems all the discounted results apply.

3.4 ELIMINATION OF NON-OPTIMAL ACTIONS. INFINITE HORIZON STATIONARY CASE (White [58], pp. 114–118)

In each of the cases considered we need to solve certain equations. Let us consider the following infinite horizon equations.

(a) Expected total discounted reward (see equation (3.1))

$$u = Tu. \tag{3.182}$$

(b) Average expected reward per unit time (see equations (3.104) and (3.105))

$$u + he = Tu, \tag{3.183}$$

$$u(m) = 0. \tag{3.184}$$

We will assume no transient states for any decision rule.

(c) Expected total reward to absorption (see equations (3.180) and (3.181))

$$u = Tu, \tag{3.185}$$

$$u(i) = 0, \quad \forall\, i \in I_a. \tag{3.186}$$

The T operators in (3.182), (3.183) and (3.185) are essentially the same, differing, in some cases, by a discount factor.

For each state $i \in I$ an action $k \in K(i)$ is optimal if and only if, using u as v or w, as the case may be

$$[T^k u](i) = [Tu](i) \geqslant [T^q u](i), \quad \forall\, q \in K(i). \tag{3.187}$$

Any k which does not satisfy (3.187) cannot be optimal.

Now suppose we have upper and lower bounding functions $\{\bar{u}, \underline{u}\}$ for u, so that

$$\underline{u} \leqslant u \leqslant \bar{u}. \tag{3.188}$$

Suppose now that for some pair $\{k, i\}$, $k \in K(i)$ we have, for some $\{q, i\}$, $q \in K(i)$

$$[T^k \bar{u}](i) < [T^q \underline{u}](i). \tag{3.189}$$

Then using Result 3.1 we have

$$[T^k u](i) \leqslant [T^k \bar{u}](i) < [T^q \underline{u}](i) \leqslant [T^q u](i). \tag{3.190}$$

Thus

$$[T^k u](i) < [T^q u](i). \tag{3.191}$$

Thus k cannot be optimal for i.

Suppose now that we wish to operate a similar scheme to value iteration scheme (3.26) but using action elimination. Let $R_n(i)$ now be the non-eliminated actions for state i with n iterations remaining. Then (3.26) takes a new form

$$\underline{n \geqslant 1} \qquad \hat{u}_n = T_n \hat{u}_{n-1} \tag{3.192}$$

where, for u: $I \to R$.

$$[T^k u](i) = r_i^k + \rho \sum_{j \in I} p_{ij}^k u(j), \quad \forall\, i \in I, \tag{3.193}$$

$$[T_n u](i) = \underset{k \in R_n(i)}{\text{maximum}} \, [[T^k u](i)], \quad \forall\, i \in I. \tag{3.194}$$

We cannot get an exactly analogous result to that of Result 3.11 because of the fact that although $R_{n+1}(i) \subseteq R_n(i)$ we need not have $R_n(i) \subseteq R_{n+1}(i)$, and the corresponding $\{\beta_n\}$ inequality (3.61) may fail. We do, however, have an analogue of inequality (3.50). The analogue of (3.50) becomes

$$\hat{u}_n + (\rho/(1-\rho))\hat{\beta}_n e \leqslant v \leqslant \hat{u}_n + (\rho/(1-\rho))\hat{\alpha}_n e \tag{3.195}$$

where

$$\underline{n \geqslant 1} \qquad \hat{\alpha}_n = \underset{i \in I}{\text{maximum}} \, [\hat{u}_n(i) - \hat{u}_{n-1}(i)], \tag{3.196}$$

$$\hat{\beta}_n = \underset{i \in I}{\text{minimum}} \, [\hat{u}_n(i) - \hat{u}_{n-1}(i)]. \tag{3.197}$$

We also have the obvious analogue of (3.55).

Thus in inequality (3.188) we may set, varying with each iteration n in general in (3.192), using (3.189) for elimination purposes

$$\underline{n \geqslant 2} \qquad \bar{u}_n = \hat{u}_{n-1} + (\rho/(1-\rho))\hat{\alpha}_{n-1}e, \tag{3.198}$$

$$\underline{u}_n = \hat{u}_{n-1} + (\rho/(1-\rho))\hat{\beta}_{n-1}e. \tag{3.199}$$

Example. Let us take the example of Table 3.6 with $\rho = 0.9$, repeated as Table 3.7.

$$\underline{n=0} \quad \hat{u}_0 = 0,$$

$$\underline{n=1} \quad \hat{u}_1(1) = 6, \quad \hat{u}_1(2) = -3, \quad R_1(2) = R_1(1) = \{1,2\},$$

$$\underline{n=2} \quad \bar{u}_2 = (60, 51), \quad \underline{u}_2 = (-21, -30),$$

$$[T\underline{u}_2](1) = \text{maximum}\begin{bmatrix} k=1\colon\ 6 + 0.5 \times (-21) + 0.5 \times (-30) \\ k=2\colon\ 4 + 0.8 \times (-21) + 0.2 \times (-30) \end{bmatrix}$$

$$= -18.8,$$

$$[T^1\bar{u}_2](1) = 6 + 0.5 \times 60 + 0.6 \times 51 = 61.5,$$

$$[T^2\bar{u}_2](1) = 4 + 0.8 \times 60 + 0.2 \times 51 = 62.2.$$

Hence we cannot eliminate $k = 1$ or $k = 2$ for $i = 1$.

$$[T\underline{u}_2](2) = \text{maximum}\begin{bmatrix} k=1\colon\ -3 + 0.4 \times (-21) + 0.6 \times (-30) \\ k=2\colon\ -5 + 0.7 \times (-21) + 0.3 \times (-30) \end{bmatrix}$$

$$= -28.7,$$

$$[T^1\bar{u}_2](2) = -3 + 0.4 \times 60 + 0.6 \times 51 = 51.6,$$

$$[T^2\bar{u}_2](2) = -5 + 0.7 \times 60 + 0.3 \times 51 = 52.3.$$

Table 3.7 Data for toymaker problem

State i	*Action* k	*Transition probability* p_{ij}^k		*Expected reward* r_i^k
1	1	0.5	0.5	6
	2	0.8	0.2	4
2	1	0.4	0.6	−3
	2	0.7	0.3	−5

Hence we cannot eliminate $k = 1$ or $k = 2$ for $i = 2$. Thus $\hat{u}_2(1) = 7.78$, $\hat{u}_2(2) = -2.03$, $R_2(1) = R_2(2) = \{1, 2\}$.

$n = 3$ If the analysis is carried out we see that we get no elimination. Thus $\hat{u}_3(1) = 9.2362$, $\hat{u}_3(2) = -0.6467$, $R_3(1) = R_3(2) = \{1, 2\}$.

$n = 4$ We find that

$$\bar{u}_4(1) = 22.3420, \qquad \bar{u}_4(2) = 12.4591,$$
$$\underline{u}_4(1) = 21.4420, \qquad \underline{u}_4(2) = 11.5591.$$

We find that

$$[T^1\bar{u}_4](1) < [T^2\underline{u}_4](1),$$
$$[T^1\bar{u}_4](2) < [T^2\underline{u}_4](2).$$

Hence $k = 1$ is eliminated for both $i = 1$ and $i = 2$. Thus we need do no more calculations.

In performing computations it is to be noted that (3.189) is equivalent to

$$[T^k\bar{u}](i) < [T\underline{u}](i). \tag{3.200}$$

3.5 FINITE HORIZON MARKOV DECISION PROCESSES

For the general, inclusive of non-stationary, case we have the iteration procedure given by (2.59) and (2.61) viz.

$$t \leqslant n \qquad u_{tn} = T_t u_{t+1,n}, \tag{3.201}$$

$$t = n + 1 \qquad u_{n+1,n} = 0. \tag{3.202}$$

We may, of course, replace $u_{n+1,n} = 0$ by any appropriate terminal value function. Equations (3.201) and (3.202) may be solved by policy space iteration also, by redefining the states (see (2.63) and (2.64)). Action elimination approaches are also possible.

Now let us turn to a fundamental point concerning infinite horizon processes, which we discussed on p. 41. Real-life problems have finite horizons, albeit uncertain in duration. However, if the durations are large (i.e. there is a large number of decision epochs) we may consider the horizon to be effectively infinite.

Now it may be easier to solve infinite horizon problems than it is to solve the corresponding finite horizon problems. The latter involve, in general, non-stationary solutions even for stationary processes, e.g.

see Tables 3.1 and 3.4 for discounted and non-discounted illustrations respectively.

We may be able to solve the infinite horizon case, for stationary parameters, more easily than solving the finite horizon case if we use policy space iteration or if we use linear programming. We will discuss the latter in Chapter 4. Alternatively, the value iteration method with an appropriate initial value function, for the infinite horizon case, may converge satisfactorily in less iterations than the number of time units in the finite horizon case.

Let $\pi = (\delta)^\infty$ be an optimal policy for the stationary infinite horizon case. The question is: how good is π for the finite horizon case? We have the following results.

Result 3.22. Discounted case. If we set $v_0 = v_0^\pi = 0$ then

$$\underline{n \geqslant 0} \qquad v_n^\pi \leqslant v_n \leqslant v_n^\pi + \rho^n \operatorname{span}[v]\, e. \tag{3.203}$$

In (3.203) v is the optimal infinite horizon expected total discounted value function, π is an optimal infinite horizon policy and v_n the optimal n time unit expected total discounted reward function with $v_0 = 0$.

Proof. We have

$$v = v^\pi. \tag{3.204}$$

Then

$$v^\pi = v_n^\pi + \rho^n (P^\delta)^n v^\pi. \tag{3.205}$$

Identity (3.205) just says that, for each $i \in I$, the infinite horizon expected total discounted reward using policy $\pi = (\delta)^\infty$ is the sum of the expected total discounted reward over n time units and the expected total discounted reward for the remaining infinite horizon. It is also clear that v^π is at least as good as one obtains using any policy τ, and hence, in particular, when $\tau = (\sigma_n, \sigma_{n-1}, ., ., \sigma_1, (\delta)^\infty)$ where

$$\sigma_s \in \arg \underset{\delta \in \Delta}{\text{maximum}} \; [T^\delta v_{s-1}], \qquad 1 \leqslant s \leqslant n \tag{3.206}$$

and hence

$$v_n = v_n^\tau. \tag{3.207}$$

Then

$$v^{\pi} \geqslant v_n + \rho^n \left(\prod_{t=1}^{n} P^{\delta_t} \right) v^{\pi}. \tag{3.208}$$

Combining (3.205) and (3.208) we obtain

$$v_n^{\pi} \geqslant v_n + \rho^n \left(\prod_{s=1}^{n} P^{\sigma_s} - (P^{\delta})^n \right) v^{\pi}. \tag{3.209}$$

Now

$$v_n \geqslant v_n^{\pi}. \tag{3.210}$$

Thus we obtain our requisite result. ○

Result 3.23. Average reward case. If $v_0 = 0$ then

$$\underline{n \geqslant 1} \qquad g_n^{\pi} \leqslant g_n \leqslant g_n^{\pi} + (\text{span}(u)/n)e \tag{3.211}$$

where (see (2.15))

$$g_n^{\pi} = v_n^{\pi}/n, \quad g_n = v_n/n \tag{3.212}$$

and u is a solution to (2.85) and (2.86).

Proof. We follow a similar analysis to that of p. 47. Let $\tilde{v}_0 = u$, and $\tilde{v}_n$ be the optimal n time units expected total reward function for this case (see p. 47). Then

$$\underline{n \geqslant s \geqslant 1} \qquad \tilde{v}_s = T\tilde{v}_{s-1} = T^{\delta}\tilde{v}_{s-1}, \tag{3.213}$$

$$v_s = Tv_{s-1} \tag{3.214}$$

where δ is defined on p. 47.

Combining (3.213) and (3.214) we have inductively.

$$\underline{n \geqslant 1} \qquad \tilde{v}_n \geqslant v_n + P^{\sigma_n}(\tilde{v}_{n-1} - v_{n-1})$$

$$\geqslant v_n + \left(\prod_{s=1}^{n} P^{\sigma_s} \right) u. \tag{3.215}$$

Clearly we also have, with $\pi = (\delta)^{\infty}$

$$v_n^{\pi} - \tilde{v}_n = -(P^{\delta})^n u. \tag{3.216}$$

Combining (3.215) and (3.216) we obtain

$$v_n^{\pi} \geqslant v_n + \left(\left(\prod_{s=1}^{n} P^{\sigma_s}\right) - (P^{\delta})^n\right)u. \tag{3.217}$$

Also

$$v_n^{\pi} \leqslant v_n. \tag{3.218}$$

Hence our requisite result follows. ○

Note that (3.211) is a function result.

3.6 EXERCISES FOR CHAPTER 3

1. For the data of Exercise 4 of Chapter 2 state and check Results 3.4, 3.7, 3.8, 3.10–3.12. In Result 3.4 restrict π to optimal policies.
2. For the same problem in Table 1.8, with $\rho = 0.9$ carry out an action-elimination exercise using (3.195)–(3.197), (3.198), (3.199), establishing first of all that (3.195) is valid in general.
3. Explain why, for the action-elimination scheme of (3.192)–(3.199), the analogue of Result 3.11 might not completely hold. Explain which part will hold and which will not.
4. Prove Result 3.18 under the conditions (3.109) and (3.110). Refer to other results needed to prove this.
5. Result 2.10 says that all stationary policy solutions to (2.85) and (2.86) are optimal. Give a simple example of an optimal stationary policy which does not satisfy (2.85) and (2.86) and explain intuitively why this arises.
6. Give an example to show that the converse of Result 3.13 does not hold, i.e. a policy which is optimal for the infinite horizon problem need not be optimal for any finite n in the value iteration scheme (3.26).
7. The optimality equation for a certain infinite horizon expected total discounted cost Markov decision process is as follows:

$$u(1) = \text{minimum}\begin{bmatrix} 90 + 0.45u(1) + 0.45u(2) \\ 100 + 0.45u(1) + 0.45u(3) \end{bmatrix}$$

$$u(2) = \text{minimum}\begin{bmatrix} 90 + 0.63u(1) + 0.27u(2) \\ 60 + 0.63u(1) + 0.27u(3) \end{bmatrix},$$

$$u(3) = \text{minimum}\begin{bmatrix} 90 + 0.27u(1) + 0.63u(2) \\ 140 + 0.27u(1) + 0.63u(3) \end{bmatrix}.$$

Note that this is in minimisation form.

With careful explanation of your calculations, and statements of any theoretical results you use, determine whether or not the policy $\pi = (\delta)^\infty$, where $\delta = (1, 2, 1)$, is a uniquely optimal policy among the stationary deterministic Markov policies.

8. The data for a two-state, two-action Markov decision process is as follows:

State i	*Action* k	*Transition probability* p_{ij}^k		*Reward* r_{ij}^k	
1	1	0.5	0.5	20	10
	2	0.8	0.2	18	8
2	1	0.25	0.75	15	8
	2	0.5	0.5	13	6

Consider the problem of maximising the infinite horizon average expected reward per unit time, and the policy $\pi = (\delta)^\infty$, where $\delta = (2, 2)$. Determine whether this policy is optimal, explaining carefully, with reference to appropriate results, how you obtain your answer and formally stating the equations used. You may assume that the problem is uni-chain.

9. For action elimination determine $\{\hat{u}_n\}$ for $n = 1, 2$, and the optimal decision rules, using the data of Exercise 7 carefully explaining the analysis.

CHAPTER 4

Linear programming formulations for Markov decision processes

4.1 INTRODUCTORY REMARKS

Kallenberg [24] gives the most comprehensive linear programming treatment of Markov decision processes. Derman [15] and Mine and Osaki [34] contain some material. This chapter is based on White [58], Chapter 7. As with our earlier treatment our problems are put in a maximisation form, whereas White [58] uses a minimisation form. The principles are exactly the same. We first of all give the result concerning the minimal elements of a set of functions which we will use.

Let $V \subseteq \{u\colon I \to R\}$. Then $u^* \in V$ is said to be a minimal element in V with respect to $\leqslant$ if whenever $u \in V$ and $u \leqslant u^*$ then $u = u^*$, i.e. no point in V can dominate u^*.

Our conventions are as follows with $u, u' \in V$ (see (3.5)):

$$u \leqslant u' \rightleftarrows u(i) \leqslant u'(i), \quad \forall\, i \in I, \tag{4.1}$$

$$u = u' \rightleftarrows u(i) = u'(i), \quad \forall\, i \in I, \tag{4.2}$$

$$u \geqslant u' \rightleftarrows u(i) \geqslant u'(i), \quad \forall\, i \in I, \tag{4.3}$$

$$u > u' \rightleftarrows u(i) > u'(i), \quad \forall\, i \in I, \tag{4.4}$$

$$u < u' \rightleftarrows u(i) < u'(i), \quad \forall\, i \in I. \tag{4.5}$$

Result 4.1. Let $\lambda \in R^m$, $\lambda > 0$ and let u^* minimise $\left[\lambda u = \sum_{i \in I} \lambda_i u(i)\right]$ over V. Then u^* is a minimal element of X with respect to $\leqslant$.

Proof. If u^* is not a minimal element then there is a $u' \in V$ with

$u' \leqslant u^*$, $u' \neq u^*$. Then, because $\lambda > 0$, we have $\lambda u' < \lambda u^*$ contradicting the assumption that u^* minimises $[\lambda u]$ over V.

4.2 INFINITE HORIZON EXPECTED TOTAL DISCOUNTED REWARD. STATIONARY CASE

We begin with our basic equation which we need to solve (see equation (2.66))

$$u = Tu. \tag{4.6}$$

Let v be the unique solution to equation (4.6). Now consider the set of functions $S \subseteq \{u\colon I \to R\}$ given by

$$S = \{u\colon I \to R \text{ satisfying the following inequality (4.8)}\}\colon \tag{4.7}$$

$$u \geqslant Tu. \tag{4.8}$$

The function $u \in S$ is said to be a superharmonic (Kallenberg [24], p. 52).

Inequality (4.8) is the same as

$$u(i) \geqslant [T^k u](i)$$

$$= r_i^k + \rho \sum_{j \in I} p_{ij}^k u(j), \quad \forall\, i \in I, \quad k \in K(i). \tag{4.9}$$

Treating $\{u(i)\}$ as variables, (4.9) is a set of $\sum_{i \in I} \#K(i)$ inequalities in m variables where $\#I = m$.

We now prove the following result as a precursor to our main result.

Result 4.2. The function v is a unique minimal element of S with respect to $\leqslant$.

Proof. Let u be any member of S. Then

$$u \geqslant Tu. \tag{4.10}$$

Also

$$v = Tv. \tag{4.11}$$

Hence

$$u - v \geqslant Tu - Tv. \tag{4.12}$$

Let

$$Tv = T^{\delta}v. \tag{4.13}$$

Then

$$\begin{aligned} u - v &\geqslant T^{\delta}u - T^{\delta}v \\ &= \rho P^{\delta}(u - v). \end{aligned} \tag{4.14}$$

Repeating (4.14) s times we obtain

$$u - v \geqslant \rho^{s}(P^{\delta})^{s}(u - v), \quad \forall\ s. \tag{4.15}$$

Letting s tend to infinity in (4.15) we obtain

$$u \geqslant v. \tag{4.16}$$

If $u \in S$ and $u \leqslant v$ then

$$u = v. \tag{4.17}$$

Hence v is a minimal element. It is also unique, for let u' be another minimal element. Then from (4.16) and the fact that u' is a minimal element we have $u' = v$. ○

The function v is often called a least element, or a perfect solution, or a feasible ideal solution of S with respect to $\leqslant$.

We may now use Results 4.1 and 4.2 to obtain our first linear programming result.

Result 4.3 (White [58], Theorem 7.4). Let $\lambda \in R^{m}$, $\lambda > 0$, $\sum_{i \in I} \lambda_i = 1$. Consider the following linear programme.

LP1

$$\underset{u}{\text{minimise}}\ [\lambda u] \tag{4.18}$$

subject to

$$u \geqslant Tu. \tag{4.19}$$

Then v is a unique solution to LP1. The optimal actions for each $i \in I$ are given by those $k \in K(i)$ in equalities (4.9) for which, in the optimal solution, the inequalities are realised as equalities.

Proof. With $V = S$ in Result 4.1 any solution to LP1 must be a minimal element. From Result 4.2 this must be v.

Because v satisfies

$$u = Tu \tag{4.20}$$

an optimal δ satisfies, with $k = \delta(i)$

$$u(i) = [T^{\delta(i)}u](i), \quad \forall\, i \in I \tag{4.21}$$

and this gives equality in (4.9) for (k, i), for all $i \in I$, $k = \delta(i)$. ○

We may now dualise LP1 to obtain the following dual linear programme. We first of all lay (4.19) in LP1 out in full for this purpose.

$$\begin{bmatrix} M_1 - & \rho P_1 \\ M_2 - & \rho P_2 \\ \vdots & \vdots \\ M_m - & \rho P_m \end{bmatrix} u \geqslant \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}. \tag{4.22}$$

M_i has ones in the ith column and zeros elsewhere, $\forall\, i \in I$, (4.23)

$$P_i = \begin{bmatrix} p_i^1 \\ p_i^2 \\ \vdots \\ p_i^m \end{bmatrix}, \qquad p_i^k = [p_{i1}^k, p_{i2}^k, \ldots p_{im}^k], \quad \forall\, i \in I, \tag{4.24}$$

$$r_i = \begin{bmatrix} r_i^1 \\ r_i^2 \\ \vdots \\ r_i^k \end{bmatrix}, \quad \forall\, i \in I. \tag{4.25}$$

DLP1 (White [58], p. 80, Mine and Osaki [34], p. 10)

$$\underset{x}{\text{maximise}} \left[\sum_{i \in I, k \in K(i)} r_i^k x_i^k \right] \tag{4.26}$$

subject to

$$\sum_{k \in K(i)} x_i^k - \rho \sum_{j \in I, k \in K(j)} p_{ji}^k x_j^k = \lambda_i, \quad \forall\, i \in I, \tag{4.27}$$

$$x_i^k \geqslant 0, \quad \forall\, i \in I, \quad k \in K(i). \tag{4.28}$$

We have the following result relating basic feasible solutions of DLP 1 and Π_D (the set of stationary deterministic Markov policies (see p. 27)).

Result 4.4 (Mine and Osaki [34], Theorem 2.9). There is a one-one correspondence between Π_D and the basic feasible solution set of DLP1 and the latter are all non-degenerate.

Proof. DLP1 has m main inequalities and hence a basic feasible solution contains m members. Because $\lambda > 0$ and $x \geqslant 0$ we must have

$$\sum_{k \in K(i)}{}' x_i^k > 0, \ \forall i \in I. \tag{4.29}$$

Hence for each $i \in I$ at least one x_i^k is positive for some $k \in K(i)$. However, we can only have m positive $\{x_i^k\}$ values and, for each $i \in I$, exactly one x_i^k is positive for some $k \in K(i)$. Thus basic feasible solutions are non-degenerate.

Now we need to look at the meaning of $\{x_i^k\}$. In order to do this let us use the formulation $\{x_i^{\pi k}(t)\}$ (see Result 2.1) where, for a given policy π, $x_i^{\pi k}(t)$ is the probability, for a given starting state $X_1 = i_1$, that at the beginning of time unit t we will be in state $i \in I$ and will take action $k \in K(i)$.

Let us generalise $\{x_i^{\pi k}(t)\}$ to be conditional on prior probabilities $\{\lambda_j\}$ where

$$\lambda_j = \text{probability}(X_1 = j), \quad \forall j \in I. \tag{4.30}$$

We then have the following equations:

$$\sum_{k \in K(i)}{}' x_i^{\pi k}(1) = \lambda_i, \quad \forall i \in I, \tag{4.31}$$

$$\underline{t \geqslant 2} \qquad \sum_{k \in K(i)}{}' x_i^{\pi k}(t) = \sum_{j \in I, k \in K(j)}{}' p_{ji}^k x_j^{\pi k}(t-1), \quad \forall i \in I. \tag{4.32}$$

Now define

$$x_i^{\pi k} = \sum_{t=1}^{\infty} \rho^{t-1} x_i^{\pi k}(t). \tag{4.33}$$

We then see that $\{x_i^{\pi k}\}$ satisfy the dual constraints (4.29) and (4.30).

Now for any policy $\pi = (\delta)^\infty \in \Pi_D$, for each $i \in I$ only one $x_i^{\pi k}$ for some $k \in K(i)$ can be positive, because a definite k is chosen for each i. Thus $\{x_i^{\pi k}\}$ is a basic feasible solution of (4.27) and (4.28).

Conversely, suppose that $\{x_i^k\}$ is a basic feasible solution of (4.27) and (4.28). Consider the policy $\pi = (\delta)^\infty$ where

$$\delta(i) = k \quad \text{if } x_i^k > 0, \quad \forall\, i \in I. \tag{4.34}$$

Using (4.31) and (4.32), $\{x_i^k\}$ will be given by (4.33) with $\pi = (\delta)^\infty$.

Numerical example. Let us look at our example of Table 1.8 with $\rho = 0.9$, retabulated as Table 4.1.

LP1

$$\underset{u}{\text{minimise}}\ [\lambda_1 u(1) + \lambda_2 u(2)] \tag{4.35}$$

subject to

$$\underline{i=1} \qquad u(1) \geqslant 6 + 0.45u(1) + 0.45u(2), \tag{4.36}$$

$$u(1) \geqslant 4 + 0.72u(1) + 0.18u(2), \tag{4.37}$$

$$\underline{i=2} \qquad u(2) \geqslant -3 + 0.36u(1) + 0.54u(2), \tag{4.38}$$

$$u(2) \geqslant -5 + 0.63u(1) + 0.27u(2). \tag{4.39}$$

The solution is (see p. 43)

$$u(1) = v(1) = 22.2, \qquad u(2) = v(2) = 12.3. \tag{4.40}$$

Table 4.1 Complete data for toymaker problem

State i	*Action* k	*Transition probability* p_{ij}^k		*Reward* r_{ij}^k		*Expected reward* r_i^k
1	1	0.5	0.5	9	3	6
	2	0.8	0.2	4	4	4
2	1	0.4	0.6	3	−7	−3
	2	0.7	0.3	1	−19	−5

Inequalities (4.37) and (4.39) are realised as equalities. Therefore $\delta(1) = \delta(2) = 2$. The solution is independent of (λ_1, λ_2) providing $\lambda > 0$. The optimal objective function value is

$$22.2\lambda_1 + 12.3\lambda_2. \tag{4.41}$$

DLPI

$$\underset{x}{\text{maximise}} \; [6x_1^1 + 4x_1^2 - 3x_2^1 - 5x_2^2] \tag{4.42}$$

subject to

$$\underline{i=1} \qquad x_1^1 + x_1^2 - 0.45x_1^1 - 0.72x_1^2 - 0.36x_2^1 - 0.63x_2^2 = \lambda_1, \tag{4.43}$$

$$\underline{i=2} \qquad x_2^1 + x_2^2 - 0.45x_1^1 - 0.18x_1^2 - 0.54x_2^1 - 0.27x_2^2 = \lambda_2. \tag{4.44}$$

Equations (4.43) and (4.44) become

$$\underline{i=1} \qquad 0.55x_1^1 + 0.28x_1^2 - 0.36x_2^1 - 0.63x_2^2 = \lambda_1, \tag{4.45}$$

$$\underline{i=2} \qquad -0.45x_1^1 - 0.18x_1^2 + 0.46x_2^1 + 0.73x_2^2 = \lambda_2. \tag{4.46}$$

In addition we have

$$x_1^1 \geqslant 0, \qquad x_1^2 \geqslant 0, \qquad x_2^1 \geqslant 0, \qquad x_2^2 \geqslant 0. \tag{4.47}$$

The solution is

$$\begin{aligned} x_1^2 &= 8.02\lambda_1 + 6.90\lambda_2, \\ x_2^2 &= 1.98\lambda_1 + 3.06\lambda_2, \\ x_1^1 &= x_2^1 = 0. \end{aligned}$$

The objective function value is

$$22.2\lambda_1 + 12.3\lambda_2. \tag{4.48}$$

Expressions (4.41) and (4.48) are identical.

4.3 INFINITE HORIZON AVERAGE EXPECTED REWARD PER UNIT TIME. STATIONARY CASE

We will only consider the uni-chain case (see p. 45). Our basic

equation is then (see (3.103)–(3.105))

$$u + he = Tu, \tag{4.49}$$

$$u(m) = 0. \tag{4.50}$$

We will not be able to use a result similar to that of Result 4.2. It is possible to use a least element approach (see p. 100). For example (see Denardo [14] and Kallenberg [24], Chapter 5 on bias optimality) for the optimal value of the gain g, a least normalised (equivalent) bias function w of the linear programming inequalities which we will use does exist. For the multiple-chain case the optimal gain function g will also be a least element of the corresponding inequalities. Our approach will be simply aimed at finding the optimal gain g.

Result 4.5 (White [58], Theorem 7.7). Consider the following linear programme:

LP2

$$\underset{u,h}{\text{minimise}} \; [h] \tag{4.51}$$

subject to

$$u + he \geqslant Tu, \tag{4.52}$$

$$u(m) = 0. \tag{4.53}$$

If no policy $\pi = (\delta)^\infty$ has transient states then LP2 and (4.49) and (4.50) have the same unique solution (w, g). The optimal actions for each $i \in I$ are given by those $k \in K(i)$ in (4.52) for which, in an optimal solution, the inequalities are realised as equalities.

Proof. If (w, g) is a solution to (4.49), (4.50) and (u, h) is any optimal solution of LP2 we have for some δ

$$\begin{aligned}(w - u) + (g - h)e &\leqslant Tw - Tu \\ &= T^\delta w - Tu \\ &\leqslant T^\delta w - T^\delta u \\ &= P^\delta (w - u).\end{aligned} \tag{4.54}$$

If θ^δ (see p. 51) is the limiting average state probability vector for P^δ, from (4.54) we have by premultiplying everything by θ^δ

$$(g - h) \leqslant 0. \tag{4.55}$$

Because (w, g) is feasible for LP2 we have the converse of (4.55). Thus

$$g = h. \tag{4.56}$$

Now let $g = h$. Then (4.54) takes the form

$$w - u = d + P^\delta(w - u) \tag{4.57}$$

where

$$d \leqslant 0.$$

Premultiplying (4.57) by θ^δ we have

$$\theta^\delta d = 0. \tag{4.58}$$

Because P^δ has no transient states we have

$$\theta_i^\delta > 0, \quad \forall\, i \in I. \tag{4.59}$$

Hence

$$d_i = 0, \quad \forall\, i \in I. \tag{4.60}$$

Then (4.57) reduces to

$$w - u = P^\delta(w - u). \tag{4.61}$$

From (2.81) (i.e. $\text{rank}(U - P^\delta) = m - 1$), (4.50) for w and (4.53) for u, the solution to (4.61) is

$$w = u. \tag{4.62}$$

Thus (w, g) is an optimal solution to LP2 and is uniquely defined. Because (w, g) satisfies (4.49) and (4.50), so does (u, h).

An optimal δ satisfies, with $k = \delta(i)$

$$w(i) + ge = [T^{\delta(i)} w](i), \quad \forall\, i \in I. \tag{4.63}$$

Thus with $(u, h) = (w, g)$ the optimal action k for a given i is such that we have equality in the corresponding inequality in (4.52). ○

In Result 4.5 we have assumed that no policy has any transient states. This was only required to prove that $d = 0$ in (4.60) in order to show that (w, g) is also a unique optimal solution to LP2. Result 4.5 holds

even without the no-transient-states condition. If we drop the no-transient states condition we obtain the following more easily proved result.

Result 4.6. Any solution (w, g) to (4.49) and (4.50) is also a solution to LP2. For such a solution the actions are given by those $k \in K(i)$ in inequality (4.52) for which, in this solution, the inequalities are realised as equalities. Any such decision rule solution generated by LP2 is optimal.

Proof. The first part follows up to (4.56) as for the proof of Result 4.5. The second part follows as for Result 4.5 from the fact that (w, g) satisfies (4.49) leading to (4.63). The last part follows after noting that, by setting $\delta(i) = k$ in the places where equalities occur in inequality (4.52), with $(u, h) = (w, g)$, we generate a policy $\pi = (\delta)^\infty$ with an optimal g^π value. ○

Let us now turn to the dual problem for LP2. For ease of use note that (4.52) may be laid out as for (4.19) in (4.22)–(4.25), setting $\rho = 1$ and adding he to the left-hand side of (4.22) for each $i \in I$.

DLP2 (White [58], pp. 83–84, Mine and Osaki [34], p. 32)

$$\underset{x}{\text{maximise}} \left[\sum_{i \in I, k \in K(i)} r_i^k x_i^k \right] \tag{4.64}$$

subject to

$$\sum_{k \in K(i)} x_i^k - \sum_{j \in I, k \in K(j)} p_{ji}^k x_j^k = 0, \quad \forall\, i \in I, \tag{4.65}$$

$$x_m + \sum_{k \in K(m)} x_m^k - \sum_{j \in I, k \in K(j)} p_{jm}^k x_j^k = 0, \tag{4.66}$$

$$\sum_{i \in I, k \in K(i)} x_i^k = 1, \tag{4.67}$$

$$x_i^k \geqslant 0, \quad \forall\, i \in I, \quad k \in K(i), \; x_m \text{ unsigned}. \tag{4.68}$$

Equality (4.66) corresponds to $u(m) = 0$ in LP1 and is redundant.

We have a result corresponding to Result 4.4 for the discounted problem, i.e. there is a one-one correspondence between Π_D and the

basic feasible solutions of DLP2 (see Kallenberg [24], Theorem 4.6.1 plus Remark 4.7.4). We have the following more simply proved result.

Result 4.7. To each $\pi \in \Pi_D$ corresponds to a basic feasible solution x^π of DLP2. To each basic feasible solution x of DLP2 such that for each $i \in I$, $x_i^k > 0$ for exactly one $k \in K(i)$, there corresponds a policy $\pi^x \in \Pi_D$. To each optimal policy $\pi \in \Pi_D$ satisfying (4.49) and (4.50) there corresponds an optimal basic feasible solution x^π of DLP2. Each optimal basic feasible solution x to DLP2 corresponds to an optimal policy $\pi^x \in \Pi_D$.

Proof. Let us define for $\delta \in \Delta$ (see 1.17)

$$P^{\delta *} = \lim_{n \to \infty} \left[\left(\sum_{l=1}^{n} (P^\delta)^{l-1} \right) \Big/ n \right]. \tag{4.69}$$

This limit exists (see Mine and Osaki [34], Lemma 3.3) and all the rows of $P^{\delta *}$ will be the same as θ^δ (see p. 51) where, for $i \in I$, θ_i^δ is the limiting average probability that the system will be in state i if policy $\pi = (\delta)^\infty$ is used.

From (2.117) we have

$$\theta^\delta = \theta^\delta P^\delta. \tag{4.70}$$

Let

$$\begin{aligned} x_i^k &= \theta_i^\delta, \quad \text{if } k = \delta(i),\ i \in I, \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{4.71}$$

Then (4.70) and (4.71) give exactly the set of equalities (4.65) of DLP2.

Because (see (2.116))

$$\sum_{i \in I} \theta_i^\delta = 1, \qquad \theta^\delta \geqslant 0 \tag{4.72}$$

we see that (4.65)–(4.68) are satisfied. Thus, each $\pi \in \Pi_D$ gives a feasible solution x^π of DLP2. It is a basic solution because DLP2 has $(m+1)$ main constraints, excluding the redundant equality (4.66), and the rank of $(U - P^\delta)$ is $(m-1)$ (see (2.81)). Here x can be degenerate, i.e. $x_i^k = 0$, for some $i \in I$ and all $k \in K(i)$ with i corresponding to a transient state.

Now consider any basic feasible solution of DLP2 for which $x_i^k > 0$ for exactly one $k \in K(i)$ for each $i \in I$. If $x_i^k > 0$ set $\delta(i) = k$.

Because $\{x_i^k\}$ satisfies (4.65)–(4.68), if we use (4.70) and (4.71) in reverse we see that θ^δ satisfies (4.70) and (4.72). Thus θ^δ is the limiting average state probability vector for policy $\pi^x = (\delta)^\infty$.

Now let π be any optimal policy in Π_D which satisfies (4.49) and (4.50) (see Result 2.10). Then (4.56) still holds and it gives rise to a solution (w, g) to LP2 and, in turn, to a basic feasible optimal solution x^π to DLP2.

Finally, consider any optimal basic feasible solution x to DLP2 with exactly one $x_i^k > 0$, $k \in K(i)$ for each $i \in I$. By construction, and using duality complementary slackness conditions, we see that its corresponding policy π^x satisfies (4.49) and (4.50) and hence, from Result 2.10, is optimal in Π_D. ○

Example. We consider the toymaker example of Table 1.8, reproduced here as Table 4.2.

LP2

$$\underset{u,h}{\text{minimise}}\ [h] \tag{4.73}$$

subject to

$$\underline{i=1} \qquad u(1) + h \geqslant 6 + 0.5u(1) + 0.5u(2), \tag{4.74}$$

$$u(1) + h \geqslant 4 + 0.8u(1) + 0.2u(2), \tag{4.75}$$

$$\underline{i=2} \qquad u(2) + h \geqslant -3 + 0.4u(1) + 0.6u(2), \tag{4.76}$$

Table 4.2 Complete data for toymaker problem

State i	*Action* k	*Transition probability* p_{ij}^k		*Reward* r_{ij}^k		*Expected reward* r_i^k
1	1	0.5	0.5	9	3	6
	2	0.8	0.2	4	4	4
2	1	0.4	0.6	3	−7	−3
	2	0.7	0.3	1	−19	−5

$$u(2) + h \geqslant -5 + 0.7u(1) + 0.3u(2), \tag{4.77}$$

$$u(2) = 0. \tag{4.78}$$

The solution is (see p. 52) $h = g = 2$, $u(1) = w(1) = 10$, $u(2) = w(2) = 0$,

$$\delta(1) = \delta(2) = 2.$$

The decision rule δ is obtained from the realised equalities in (4.75) and (4.77).

DLP2

$$\text{minimise}[6x_1^1 + 4x_1^2 - 3x_2^1 - 5x_2^2] \tag{4.79}$$

subject to

$$\underline{i=1} \qquad x_1^1 + x_1^2 - 0.5x_1^1 - 0.8x_1^2 - 0.4x_2^1 - 0.7x_2^2 = 0, \tag{4.80}$$

$$\underline{i=2} \qquad x_2^1 + x_2^2 - 0.5x_1^1 - 0.2x_1^2 - 0.6x_2^1 - 0.3x_2^2 = 0, \tag{4.81}$$

$$x_1^1 + x_1^2 + x_2^1 + x_2^2 = 1, \tag{4.82}$$

$$x_1^1 \geqslant 0, \qquad x_1^2 \geqslant 0, \qquad x_2^1 \geqslant 0, \qquad x_2^2 \geqslant 0. \tag{4.83}$$

The solution is

$$x_1^2 = 0.778, \qquad x_2^2 = 0.222,$$
$$x_1^1 = x_2^1 = 0.$$

The objective function value is

$$g = 2.$$

Also $\delta(1) = \delta(2) = 2$ because $x_1^2 > 0$, $x_1^1 = 0$, $x_2^2 > 0$, $x_2^1 = 0$.

4.4 ABSORBING STATE PROBLEMS. STATIONARY CASE

The results for this case are similar to those of the discounted case under the assumptions made in (2.126).

The primal and dual linear programmes are as follows. Let $\lambda \in R^m$, $\lambda > 0$, $\sum_{i \in I} \lambda_i = 1$.

LP3 (White [58], p.78)

$$\underset{u}{\text{minimise}}\,[\lambda u] \tag{4.84}$$

subject to

$$u \geqslant Tu, \tag{4.85}$$

$$u(i) = 0, \quad \forall\, i \in I_a. \tag{4.86}$$

DLP3 (White [58], pp. 78–79, Mine and Osaki [34], p. 43)

$$\underset{x}{\text{maximise}} \left[\sum_{i \in I} \sum_{I_a, k \in K(i)} r_i^k x_i^k \right] \tag{4.87}$$

subject to

$$\sum_{k \in K(i)} x_i^k - \sum_{j \in I, k \in K(j)} p_{ji}^k x_j^k = \lambda_i, \quad \forall\, i \in I \backslash I_a, \tag{4.88}$$

$$x_i - \sum_{j \in I, k \in K(j)} p_{ji}^k x_j^k = \lambda_i, \quad \forall\, i \in I_a, \tag{4.89}$$

$$x_i^k \geqslant 0, \quad \forall\, i \in I \backslash I_a, \quad k \in K(i), \tag{4.90}$$

$$x_i \text{ unsigned}, \ \forall\, i \in I_a. \tag{4.91}$$

Equalities (4.89) are redundant in view of (4.91) which corresponds to $u(i) = 0$, $\forall\, i \in I_a$ in LP3.

4.5 POLICY SPACE ITERATION METHOD AND SIMPLEX BLOCK PIVOTING (Kallenberg [24], p. 132)

The policy space improvement step (see step (iii) in (3.84) for the discounting case, and in (3.155) for the average reward case) are equivalent to block pivoting in the dual problems DLP1 and DLP2 respectively.

Let us look at the discounted problem first.

4.5.1 DISCOUNTED CASE

Let π^n, π^{n+1} be two successive policies with value function u^n, u^{n+1}

respectively. Then from (3.91) we have, for a prior probability vector λ for the initial states

$$\lambda u^{n+1} = \lambda u^n + \lambda(U - \rho P^{\sigma^{n+1}})^{-1} A^{n+1} \tag{4.92}$$

where (see (3.89))

$$A^{n+1} = (T^{\sigma^{n+1}} u^n - T^{\sigma^n} u^n). \tag{4.93}$$

Now

$$\lambda(U - \rho P^{\sigma^{n+1}})^{-1} = \lambda \sum_{l=1}^{\infty} \rho^{l-1} (P^{\sigma^{n+1}})^l. \tag{4.94}$$

The right-hand side of (4.94) is equal to the vector x^π as given in (4.33) where $\pi = (\sigma^{n+1})^\infty$, adjusted for the prior probability vector λ.

For $k \neq \sigma^{n+1}(i)$

$$x_i^{\pi k} = 0. \tag{4.95}$$

At iteration $n+1$, π is to be freely chosen. Thus A^{n+1} may be written as A^π where

$$A_i^{\pi k} = [T^k u^n - T^{\sigma^n} u^n](i), \quad \forall\, i \in I, \quad k \in K(i). \tag{4.96}$$

Thus the right-hand side of (4.92) may be written as

$$\lambda u^n + \sum_{i \in I, k \in K(i)} x_i^{\pi k} A_i^{\pi k}. \tag{4.97}$$

We are free to choose π to maximise the summation in (4.97). We replace $\{x_i^{\pi k}, A_i^{\pi k}\}$ by $\{x_i^k, A_i^k\}$. Then the $\{A_i^k\}$ in (4.97) are just the shadow prices if we use the linear programme DLP1. We apply the usual linear programming improvement step, but we are free to choose k for each i. Hence we can change up to m variables $\{x_i^k\}$ at a time. This is block pivoting.

Let us now look at the average reward case.

4.5.2 AVERAGE REWARD CASE

From (3.165) we have

$$h^{n+1} = h^n + \theta^{\sigma^{n+1}} B^{n+1} \tag{4.98}$$

where (see (3.160))

$$B^{n+1} = T^{\sigma^{n+1}} u^n - T^{\sigma^n} u^n. \tag{4.99}$$

If $\pi = (\sigma^{n+1})^{\infty}$ then

$$x_i^{\pi k} = \theta_i^{\sigma^{n+1}}, \quad \text{if } i \in I, \quad k = \sigma^{n+1}(i), \tag{4.100}$$

$$= 0 \quad \text{otherwise.} \tag{4.101}$$

The block pivoting result now follows as for the discounted case.

4.6 FINITE HORIZON PROBLEMS

The relevant equations are (see (3.201) and (3.202))

$$\underline{t \leqslant n} \qquad u_{tn} = T_t u_{t+1,n}, \tag{4.102}$$

$$\underline{t = n+1} \qquad u_{n+1,n} = 0, \tag{4.103}$$

where the solution u_{tn} is equal to v_{tn}, the maximal expected total discounted reward value function over the next $n - t + 1$ time units beginning at the beginning of time unit t. We allow the parameters to be time dependent.

The easier formal way to handle this is (see (2.63) and (2.64)) to define a new state set $\tilde{I} = I \times \Gamma$, where $\Gamma = \{1, 2, ., ., ., .n + 1\}$, and to consider the problem as an absorbing state problem with an absorbing state set

$$\tilde{I}_a = I \times \{n+1\}. \tag{4.104}$$

We may then use the LP3 and DLP3 formulations, with the same properties derivable from the stationary discounted problem.

Let $\lambda \in R^{m \times n}$, $\lambda > 0$, $\sum_{i \in I, 1 \leqslant t \leqslant n} \lambda_{it} = 1$.

LP4 (see White [58], p. 89)

$$\underset{u}{\text{minimise}} \left[\lambda u = \sum_{t=1}^{n} \lambda_t u_t\right] \tag{4.105}$$

subject to

$$\underline{1 \leqslant t \leqslant n} \qquad u_t \geqslant T_t u_{t+1}, \tag{4.106}$$

$$\underline{t = n+1} \qquad u_{n+1} = 0, \tag{4.107}$$

$$\lambda_t \in R^m, \ 1 \leqslant t \leqslant n, \ u = (u_1, u_2, ., ., ., u_n), \qquad u_t \colon I \to R, \ 1 \leqslant t \leqslant n. \tag{4.108}$$

DLP4 (White [58], p. 89)

$$\underset{x}{\text{maximise}} \left[\sum_{i \in I, k \in K_i(t), 1 \leqslant t \leqslant n} \left(\prod_{s=0}^{t-1} \rho(s) \right) x_i^k(t) r_i^k(t) \right] \tag{4.109}$$

subject to

$$\sum_{k \in K_{t+1}(t)} x_i^k(t+1) - \sum_{j \in I, k \in K_t(j)} p_{ji}^k(t) x_j^k(t) = \lambda_{i,t+1},$$

$$\forall\, i \in I, \quad 1 \leqslant t \leqslant n-1, \tag{4.110}$$

$$\sum_{k \in K_1(i)} x_i^k(1) = \lambda_{i1}, \quad \forall\, i \in I, \tag{4.111}$$

$$x_i^k(t) \geqslant 0, \quad \forall\, i \in I, \quad k \in K_t(i), \qquad 1 \leqslant t \leqslant n-1. \tag{4.112}$$

In (4.112) we do not need $t = n$, because $t = n$ is not in (4.110). However, we can cater for different terminal values.

It is possible to put $\lambda_{it} = 0$ for any $i \in I_t$, $t \geqslant 1$ in LP4 and in DLP4. This will have the effect of possibly producing decision rules for time unit t which are not optimal for that i, but optimal rules will be given for $i \in I_t$, $t \geqslant 1$ for which $\lambda_{it} > 0$. In the dual DLP4, setting $\lambda_{it} = 0$ may result in state $i \in I_t$ for time unit t having zero probability, in which case whether or not a rule is optimal for this i for time unit t is irrelevant.

If we use LP4 or DLP4 to solve our problem, the variables are functions of t, even in the stationary case. However, in the latter case we may consider the infinite horizon case, $n = \infty$, and solve LP1 or DLP1 to produce solutions in terms of the state space I. We may then use Result 3.22 where now $\{v_n^\pi, v_n, v\}$ are defined on the state space $\tilde{I}$ (see p. 113).

Similar comments apply for the average reward case, via Result 3.23.

4.7 EXERCISES FOR CHAPTER 4

1. For the data of Exercise 4 of Chapter 2 for an infinite horizon expected discounted Markov decision process

   (i) state the primal linear programme;
   (ii) state the dual linear programme;
   (iii) using the linear programming formulations in (i) and in (ii), separately show that the policy $\pi = (\delta)^\infty$, where $\delta(1) = \delta(2) = 2$,

$\delta(3) = 1$, is optimal, carefully explaining why the policy is optimal and obtaining the form of the optimal objective functions as functions of the initial state probabilities.

2. Formulate the inventory, queuing and defective product problems, given at the end of Chapter 2, as linear programmes using the dual formulation, making due reference to the appropriate parts of the text, and to defining your decision variables. You may assume a uni-chain structure where needed. The $\{r_i^k\}$, $\{p_{ij}^k\}$ must be explicitly put into your equations as functions of the random variables of the problems.

CHAPTER 5

Semi-Markov decision processes

5.1 INTRODUCTORY REMARKS

So far we have assumed that decisions are taken at each of a sequence of unit time intervals. In this chapter we will allow decisions to be taken at varying integral multiples of the unit time interval. The interval between decisions may be predetermined or random.

We will call these processes 'semi-Markov decision processes'. However, strictly speaking semi-Markov decision processes are more restrictive (e.g. see Mine and Osaki [34], Chapter 6). Strictly speaking they relate to situations where decisions are taken when a change of state occurs. We wish to allow decisions to be taken even if the state does not change at successive decision epochs.

5.1.1 ILLUSTRATION. QUEUING

On p. 55 we discussed a simple queuing problem where decisions were taken at each of a succession of unit time intervals, and this gives us a Markov decision process. However, we could equally well take decisions (a) when a customer arrives, or (b) when a customer departs. In these cases the decision epoch intervals are random variables.

In case (a) if Γ is the random interval between decision epochs we have, independently of state and time

$$\text{probability}(\Gamma = \gamma) = p(1-p)^{\gamma-1}, \quad \forall\ \gamma \geqslant 1. \tag{5.1}$$

In case (b), independently of time, if the system is not empty (i.e. $i \geqslant 1$) then

$$\text{probability}(\Gamma = \gamma) = q(1-q)^{\gamma-1}, \quad \forall\ \gamma \geqslant 1. \tag{5.2}$$

For an empty system (i.e. $i = 0$) if $p \neq q$

$$\text{probability}(\Gamma = \gamma) = pq((1-q)^{\gamma-1} - (1-p)^{\gamma-1})/(p-q), \quad \forall\ \gamma \geqslant 2. \tag{5.3}$$

If $p = q$ (5.3) becomes $(\gamma - 1)p^2$.

We make the following notes:

(i) The following is based upon White [58], Chapter 4.

(ii) In White [58] 'minimisation' is used, but the translation from 'minimisation' to 'maximisation' is straightforward.

(iii) We will consider the infinite horizon stationary case only. Non-stationary and finite horizon cases may be treated as extensions as for Markov decision processes.

(iv) We will be a little less formal than we have been up until now in proving certain results. The analyses given in earlier sections are easily generalised, e.g. z-transforms, bound analysis, and so on. Indeed all the earlier work is a special case of a semi-Markov decision process with

$$\text{probability}(\Gamma = \gamma) = 1, \quad \text{if } \gamma = 1, \tag{5.4}$$

$$= 0, \quad \forall\ \gamma \neq 1. \tag{5.5}$$

(v) For the limiting average expected reward per unit time case we will consider only the uni-chain case (see p. 45).

(vi) For the absorbing state case we will assume that, if non-discounted, then condition (2.126) holds where t now relates to the tth decision epoch as distinct from the tth unit time interval. These definitions are identical for a Markov decision process.

5.1.2 THE FRAMEWORK

Here X_t (state), Y_t (reward, or discounted reward), Z_t (action) are defined as for the Markov decision process (see p. 25) noting, for example, that Y_t is now the reward, or discounted reward (discounted back to the beginning of the tth decision epoch) in the interval between the tth and $(t+1)$th decision epochs.

The policy spaces are defined as for Markov decision processes (see p. 27) noting that histories are still defined in terms of the new $\{X_t, Y_t, Z_t\}$ and make no reference to states and rewards arising at times other than the decision epochs because we will be assuming that the Markov property relates only to successive decision epochs.

We need to generalise the transition probability form of (2.22). We let Γ_t be the random integer number of time units between the tth and $(t+1)$th decision epoch. Here Γ_t will be referred to as the 'decision interval'. Our Markov assumption is then, in the stationary case

$$\begin{aligned}\text{probability}(X_{t+1} = j,\ \Gamma_t = \gamma \mid H_t = h,\ Z_t = k \in K_t(i)) \\ = \text{probability}(X_{t+1} = j,\ \Gamma_t = \gamma \mid X_t = i, \\ Z_t = k \in K_t(i)) \quad (5.6)\end{aligned}$$

$$= p^k_{ij\gamma}. \quad (5.7)$$

We assume that

$$1 \leqslant \gamma \leqslant L < \infty. \quad (5.8)$$

If $r^k_{ij\gamma}$ is the reward in a decision interval, discounted to the beginning of that interval, given i, j, γ, k then let

$$r^k_i = E(Y_t \mid X_t = i,\ Z_t = k) = \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p^k_{ij\gamma} r^k_{ij\gamma}, \quad \forall\ i \in I, \quad k \in K(i). \quad (5.9)$$

We will not repeat the analysis of the Markov decision process results. This is very similar and is easily constructed. We will concentrate on the basic equations, algorithms and some results without proof.

5.2 INFINITE HORIZON EXPECTED TOTAL DISCOUNTED REWARD (White [58], p. 15)

In this case the optimality equation (cf (2.66)) is

$$u = Fu \quad (5.10)$$

where (cf. (2.55)–(2.57)) for u: $I \to R$

$$[Fu](i) = \underset{k \in K(i)}{\text{maximum}} \left[r^k_i + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} \rho^\gamma p^k_{ij\gamma} u(j) \right], \quad \forall\ i \in I, \quad (5.11)$$

$$[F^\delta u](i) = r^{\delta(i)}_i + \sum_{1 \leqslant \gamma \leqslant L} \rho^\gamma [P^\delta_\gamma u]_i, \quad \forall\ i \in I, \quad (5.12)$$

$$Fu = \underset{\delta \in \Delta}{\text{maximum}}\ [F^\delta u] \quad (5.13)$$

and

$$[P_\gamma^\delta]_{ij} = p_{ij\gamma}^{\delta(i)}, \quad \forall\, i, j \in I. \tag{5.14}$$

Results 2.1 and 2.2, with our new interpretation of t, hold and we may keep to Π_M without loss. Similarly, as in Result 2.3 we may reduce Π_M to Π_{MD}. Finally, we obtain Result 2.4 also, viz. our optimal value function v is a solution to (5.10). This is explained on p. 51 of White [58]. Result 2.5, viz. v is the unique solution to our optimality equation (5.10), is easily established. Finally, Result 2.6, holds, viz. any policy $\pi = (\delta)^\infty$ is optimal where δ is a decision rule solution to our optimality equation (5.10).

Thus we have the following result.

Result 5.1. The value function v of maximal expected total discounted rewards is a unique solution to (5.10). Optimal stationary policies exist and they all take the form $\pi = (\delta)^\infty$, where δ is any decision rule solution to (5.10). ○

Let us now look at algorithms.

These are analogous to those given in Chapter 3. We will not deal with bounds or with action elimination. These may be found in White [58], Chapter 9, Section 4. Procedures are very similar to those for the Markov decision process case.

5.2.1 VALUE ITERATION (cf. (3.26) and (3.27) and White [58], p. 70)

$$\underline{n \geqslant 1} \qquad \tilde{u}_n = F\tilde{u}_{n-1}, \tag{5.15}$$

$$\underline{n = 0} \qquad \tilde{u}_0 = u, \quad \text{arbitrary}. \tag{5.16}$$

In White [58], p. 50 F is defined somewhat differently, to cater for finite horizon equations. There, using 'maximisation' instead of 'minimisation' we have, with n being the number of time units remaining

$$\underline{n \geqslant 1} \qquad u_n = G(u_{n-1}, u_{n-2}, ., ., u_{n-\gamma}), \tag{5.17}$$

$$\underline{n \leqslant 0} \qquad u_n = 0 \tag{5.18}$$

where G is defined by

$$G(u_{n-1}, u_{n-2}, ., ., u_{n-\gamma})(i)$$

$$= \underset{k \in K(i)}{\text{maximum}} \left[r_i^k + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} \rho^\gamma p_{ij\gamma}^k u_{n-\gamma}(j) \right], \quad \forall\, i \in I. \quad (5.19)$$

The function G is thus defined in White [58] on $(u_{n-1}, u_{n-2}, ., ., u_{n-\gamma})$ and not just on u_{n-1}.

In (5.15) $\{\tilde{u}_n\}$ is not the same as $\{u_n\}$ in (5.17). We may replace (5.18) without loss for the infinite horizon case by

$$\underline{n \leqslant 0} \qquad u_n = u, \quad \text{arbitrary.} \quad (5.20)$$

In (5.17) $\{u_n\}$ is consistent with $\{u_n\}$ as defined in (2.58) for Markov decision processes with $T = G$ and, for Markov decision processes, $\tilde{u}_n = u_n$. In (5.15), n has the interpretation of n decision epochs remaining. In (5.17), n has the interpretation of n time units remaining.

Our use of F is consistent with the use in White [58], Chapter 6.

Analogously to Result 3.5 we have the following result which we will not prove; α and β are as defined in (3.27) and (3.28) and, if X is the random state at a decision epoch

$$\bar{\rho} = \underset{i \in I}{\text{maximum}} \; [E(\rho^\Gamma \mid X = i)], \quad (5.21)$$

$$\underline{\rho} = \underset{i \in I}{\text{minimum}} \; [E(\rho^\Gamma \mid X = i)]. \quad (5.22)$$

Also, as before

$$e(i) = 1, \quad \forall\, i \in I. \quad (5.23)$$

Result 5.2.

$$\tilde{u}_n + \text{minimum}\,[\bar{\rho}^n\beta, \underline{\rho}^n\beta]\, e \leqslant v \leqslant \tilde{u}_n + \text{maximum}\,[\bar{\rho}^n\alpha, \underline{\rho}^n\alpha]\, e. \quad (5.24)$$

○

As in Result 3.6, $\{\tilde{u}_n\}$ converges to v as n tends to infinity, with respect to the norm $\| \quad \|$.

We will give no more results. There will, for example, be results analogous to Result 3.12 and so on, and these are similarly obtained.

5.2.2 POLICY SPACE ITERATION (cf. (3.83)–(3.86) and White [58], p. 51)

(i) Select an initial policy $\pi^0 = (\sigma^0)^\infty$.
(ii) Solve the equation

$$\tilde{u}^0 = F^{\sigma^0}\tilde{u}^0 \tag{5.25}$$

for $\tilde{u}^0$.
(iii) For a new policy $\pi^1 = (\sigma^1)^\infty$ by finding

$$\sigma^1 \in \arg \underset{\delta \in \Delta}{\text{maximum}} \, [F^\delta \tilde{u}^0]. \tag{5.26}$$

(iv) Replace σ^0 in (i) by σ^1, and repeat the procedure.

We have the following result analogous to Result 3.14.

Result 5.3 (White [58], pp. 52–53). The policy space method produces a sequence $\{\tilde{u}^n\}$ which converges non-decreasing to v in a finite number of iterations, with an optimal policy given by the terminating policy. ○

Stopping rules may be considered using an analogous result to Result 3.12 with $u = \tilde{u}^n$.

5.3 INFINITE HORIZON AVERAGE EXPECTED REWARD PER UNIT TIME (White [58], Chapter 5)

In this case the optimality equation may take one of two forms. The first form is analogous to (3.103)–(3.106) as follows:

$$h = \underset{k \in K(i)}{\text{maximum}} \left[\left(r_i^k + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k u(j) - u(i) \right) \Big/ \left(\sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k \gamma \right) \right], \quad \forall\, i \in I, \tag{5.27}$$

$$u(m) = 0. \tag{5.28}$$

The denominator in (5.27) is non-zero. This denominator is simply the expected number of time units to the next decision epoch, given $X = i$, $Z = k$ at the current decision epoch.

The second form is a rearrangement of the first form as follows:

$$u(i) = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k - h \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k \gamma \right.$$

$$\left. + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k u(j) \right], \quad \forall\, i \in I, \tag{5.29}$$

$$u(m) = 0. \tag{5.30}$$

Equations (5.27) and (5.29) are equivalent forms. They may be put in operator forms

$$he = Hu \tag{5.31}$$

or

$$u = \tilde{H}(h, u) \tag{5.32}$$

with appropriate definitions of $H, \tilde{H}$.

Equation (5.32) is the more natural one on the basis of the following argument. Let us follow (3.109) and assume that n is now the number of time units over which we are optimising and u_n is defined appropriately, with $u_n = 0$, although the latter is not necessary (see (5.17)–(5.19)) and (see (3.109) and (3.110))

$$\underline{n \geqslant 0} \qquad u_n = ng + w + \varepsilon_n, \tag{5.33}$$

$$\underset{n \to \infty}{\text{limit}} \; [\varepsilon_n] = 0. \tag{5.34}$$

Here g will be the optimal average expected reward per time unit.

Using (5.17) and (5.19) with $\rho = 1$ and substituting (5.33) we obtain

$$ng + w(i) + \varepsilon_n(i) = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k ((n - \gamma)g \right.$$

$$\left. + w(j) + \varepsilon_{n-\gamma}(j)) \right], \quad \forall\, i \in I. \tag{5.35}$$

Letting $\varepsilon_n, \varepsilon_{n-\gamma}$ tend to zero as n tends to infinity and, cancelling ng on the left-hand and right-hand sides of (5.35), we see that (w, g) satisfies equation (5.32).

We will have the following result (cf. Result 2.10) which does not require (5.33) to hold.

Result 5.4 (White [58], Theorems 5.9, 5.10). Let δ be any decision rule solution to (5.27) and (5.28), or to (5.29) and (5.30), and let $\pi = (\delta)^\infty$. Then $g^\pi \geqslant g^\tau$ for all $\tau \in \Pi$ and stationary policies exist which are optimal for each state simultaneously. ○

Let us now look briefly at algorithms.

5.3.1 VALUE ITERATION (cf. (3.106) and (3.107) and White [58], p. 50)

We cannot use (5.15) as our iterative scheme because this would result in maximising the expected reward between successive decision epochs and not necessarily the average expected reward per unit time. The scheme to use is as follows (see (5.17), (5.19), (5.20)):

$$\underline{n \geqslant 1} \qquad u_n = G(u_{n-1}, u_{n-2}, ., ., u_{n-\gamma}), \tag{5.36}$$

$$\underline{n \leqslant 0} \qquad u_n = u, \quad \text{arbitrary} \tag{5.37}$$

where G is given by (5.19) with $\rho = 1$. We will not analyse the behaviour of this algorithm. Some results may be found in White [58], Chapter 5. Similar analyses to those given in pp. 76–81 are possible.

5.3.2 POLICY SPACE ITERATION (cf. (3.153)–(3.157) and White [58], p. 59)

(i) Select an initial policy $\pi^0 = (\sigma^0)^\infty$.

(ii) Solve the equations for (h^0, u^0)

$$h^0 e = H^{\sigma^0} u^0 \tag{5.38}$$

or

$$u^0 = \tilde{H}^{\sigma^0}(h^0, u^0) \tag{5.39}$$

(see (5.31) or (5.32) respectively with $H^\delta, \tilde{H}^\delta$ being defined analogously to $H, \tilde{H}$ for $\delta \in \Delta$).

(iii) Find a new policy $\pi^1 = (\sigma^1)^\infty$ by finding

$$\sigma^1 \in \arg \underset{\delta \in \Delta}{\text{maximum}} \; [H^\delta u^0] \tag{5.40}$$

or

$$\sigma^1 \in \arg \underset{\delta \in \Delta}{\text{maximum}} \; [\tilde{H}^\delta(h^0, u^0)]. \tag{5.41}$$

(iv) Replace σ^0 in (i) by σ^1 and repeat the procedure.

Termination is as for (3.156) or (3.157) using H^δ or $\tilde{H}^\delta$ instead at T^δ.

We have the following result analogous to Result 3.19 where, for $\delta \in \Delta$

$$[P^\delta]_{ij} = \sum_{1 \leqslant \gamma \leqslant L} p_{ij\gamma}^{\delta(i)}, \quad \forall\, i, \quad j \in I. \tag{5.42}$$

Result 5.5 (White [58], Theorems 5.3 and 5.4). If no transition probability matrix P^δ for $\delta \in \Delta$ has any transient states, then the policy space iteration method produces a sequence $\{h^n\}$ which converges increasing to g in a finite number of iterations, and the terminating policy is optimal. ○

Where transient states arise, if we use the modified policy iteration (see p. 87) the same results apply as in Result 3.20. As with Result 3.19, the no-transient-states condition is not strictly necessary, providing we use only terminating condition (3.156) with T^δ replaced by H^δ or $\tilde{H}^\delta$.

We will not analyse the behaviour of the policy space iteration algorithm further, but a similar result to that of Result 3.21 is possible.

5.4 ABSORBING STATE PROBLEMS

We merely state that, for discounted problems, and for non-discounted problems where the condition (2.126) holds, Result 5.1 and the value iteration and policy space iteration Results 5.2 and 5.3 hold.

5.5 LINEAR PROGRAMMING

Linear programming approaches, both dual and primal, may be developed in much the same way as for Markov decision processes (cf. Chapter 4) and White [58], pp. 85–89 also gives linear programming formulations.

5.6 FINITE HORIZON PROBLEMS

Finite horizon formulations are possible following the lines of pp. 33–39, with the linear programming approaches following the lines of pp. 113–114.

5.7 SOME ILLUSTRATIONS

(a) Inventory (cf. p. 54)

Suppose that we allow i to be negative and we use the rule: order stock only when $i < 0$. Let $q_\gamma(S)$ be the probability that the demand over γ time units is S.

Suppose that as soon as $i < 0$ we put the stock level up to $k \geqslant 0$, k depending on i. Then $k = i$ if $i \geqslant 0$. We will, for simplicity, assume that the average stock level between decision epochs is

$$\tfrac{1}{2}(k + 0) = \tfrac{1}{2}k. \tag{5.43}$$

Then the semi-Markov analogue of equation (2.130) is as follows with $q_0(0) = 1$, $q_0(S) = 0$, $S \neq 0$. We include the case $i \geqslant 0$ if we happen to begin our process when this is so. However, once i becomes negative it is always negative at subsequent decision epochs.

$$\underline{i \geqslant 0} \qquad u(i) = \sum_{S \leqslant i, s > i - S, 1 \leqslant \gamma \leqslant L} q_{\gamma-1}(S)q(s) \times \{-l(S + s - i) - \tfrac{1}{2}ai\gamma + \rho^{\gamma-1}u(i - S - s)\}, \tag{5.44}$$

$$\underline{i < 0} \qquad u(i) = \underset{k \geqslant 0}{\text{maximum}} \left[\sum_{S \leqslant k, s > k - S, 1 \leqslant \gamma \leqslant L} q_{\gamma-1}(S)q(s) \times \{-c(k - i) - l(S + s - k) - \tfrac{1}{2}ak\gamma + \rho^{\gamma-1}u(k - S - s)\} \right]. \tag{5.45}$$

(b) Queuing (cf. p. 55)

We will use case (a) of p. 116, viz. decisions are made when someone arrives. Then (see (5.1))

$$\text{probability}(\Gamma = \gamma) = p(1 - p)^{\gamma-1}, \quad \forall\ \gamma \geqslant 1. \tag{5.46}$$

We will use optimality equations (5.29) and (5.30). Equation (5.29) then takes the form, where i, by definition, is the number in the system at the time someone arrives, inclusive of the arrival, and $k \leqslant m-1$

$$u(i) = \underset{0 \leqslant k \leqslant i}{\text{maximum}} \Bigg[-c(i-k) + \sum_{0 \leqslant s \leqslant k, 1 \leqslant \gamma \leqslant L} p(1-p)^{\gamma-1} q_\gamma^k(s) \times \{ -\tfrac{1}{2}\gamma(2k+1-s)a + u(k-s+1) \} - h \sum_{1 \leqslant \gamma \leqslant L} p(1-p)^{\gamma-1}\gamma \Bigg], \qquad 1 \leqslant i \leqslant m \tag{5.47}$$

where $q_\gamma^k(s)$ is the probability that, over γ unit time units, s services will be completed given k customers in the system.

We have used the average of initial system size k and the final system size, represented by $k-(s-1)$, to approximate the waiting cost. The term $k-(s-1)$ might equally well be represented by $(k-s)$ if the last completed service is not at the end of the γth time unit, but we are merely using an approximation.

(c) Process overhaul (White [57], pp. 100–103)

Let us consider a single process with the following features:

(i) The process is to be operated over an infinite time horizon.
(ii) If the process is operated for γ time units the reward will be

$$r(a, b, \gamma) \tag{5.48}$$

where a, b are realisation of two random parameters A, B with

$$\text{probability}(A = a, B = b) = p(a, b). \tag{5.49}$$

(iii) From time to time the process may be overhauled at a cost, which we assume is included in the reward $r(a, b, \gamma)$.
(iv) When the process is overhauled the values of (a, b) are known almost instantaneously. In effect, the process will run for a short while and (a, b) then estimated.
(v) The problem, once (a, b) are known, is to decide at how many time units later on the process should next be overhauled.

The state of the system immediately after an overhaul is $i = (a, b)$ which we assume takes a finite number of values. Let I be the state set.

We will consider the limiting average expected reward per unit time case only. The equation corresponding to (5.29) is as follows, with $1 \leqslant \gamma \leqslant L$:

$$u(a,b) = \underset{1 \leqslant \gamma \leqslant L}{\text{maximum}} \left[r(a,b,\gamma) - h\gamma + \sum_{(a',b') \in I} p(a',b')u(a',b') \right], \quad \forall\ (a,b) \in I. \tag{5.50}$$

If for (5.30) we put, instead of $u(a,b) = 0$ for some $(a,b) \in I$

$$\sum_{(a,b) \in I} p(a,b)u(a,b) = 0 \tag{5.51}$$

and set

$$\mu = \sum_{a,b \in I} p(a,b)u(a,b) \tag{5.52}$$

then equation (5.50) transforms to

$$\mu = \sum_{(a,b) \in I} p(a,b) \underset{1 \leqslant \gamma \leqslant L}{\text{maximum}} \left[r(a,b,\gamma) - h\gamma + \mu \right]. \tag{5.53}$$

Equivalently, (5.53) is

$$\sum_{(a,b) \in I} p(a,b) \underset{1 \leqslant \gamma \leqslant L}{\text{maximum}} \left[r(a,b,\gamma) - h\gamma \right] = 0. \tag{5.54}$$

Equation (5.50) may be solved using the policy space iteration method of pp. 123–124. However, it is easier to solve equation (5.54) by varying (γ, h) until equation (5.54) has been satisfied. If $r(a,b,\gamma)$ has a suitable analytic form we could find, for each (a,b,h) triple, the optimal γ as a function of (a,b,h), and then analytically find the optimal $h(=g)$ value (e.g. see White [57] for the continuous (a,b) case).

Once h has been found an optimal decision rule is given by

$$\delta(a,b) \in \arg \underset{1 \leqslant \gamma \leqslant L}{\text{maximum}} \left[r(a,b,\gamma) - g\gamma \right]. \tag{5.55}$$

5.8 EXERCISES FOR CHAPTER 5

1. In a semi-Markov decision process with two states and two actions

in each state the transition probabilities and rewards are given as follows:

State i	Action k	$\gamma = 1$ p_{ij1}^k		$\gamma = 2$ p_{ij2}^k		$\gamma = 1$ r_{ij1}^k		$\gamma = 2$ r_{ij2}^k	
1	1	0.2	0.3	0.3	0.2	9	3	9	3
1	2	0.6	0.1	0.2	0.1	4	4	4	4
2	1	0.3	0.2	0.1	0.4	3	−7	3	−7
2	2	0.4	0.1	0.3	0.2	1	−19	1	−19

(i) For the infinite horizon expected total discounted reward case write down the optimality equations using the data of the problem, but for a general discount factor ρ. Explain your derivation with reference to the appropriate part of the text.

(ii) Determine, with due reference to the appropriate part of the text, whether or not the policy $\pi = (\delta)^\infty$, $\delta = (2, 2)$ is optimal and uniquely so for $\rho = 0.9$.

2. For the infinite horizon average expected reward per unit time case determine whether or not the policy π in Exercise 1 is optimal, explaining your derivations with due reference to the appropriate part of the text.
3. Prove that {(5.27), (5.28)} and {(5.29), (5.30)} will give exactly the same $\{(u, h, \delta)\}$ solutions.
4. A machine may be functioning at performance levels 1, 2 or 3 or be in a failed state. As soon as it is in a failed state this is known. The machine is then restored to performance level 3 at a cost $c(0)$, and it has to be decided after how many time units it should next be inspected. If a machine is not in a failed state its performance level is only known on inspection at a cost a. Once an inspection has taken place, and if the machine is not in a failed state, it has to be decided when the next inspection should take place. A machine operating at performance level l (which is assumed to be the same through a time unit as it is at the beginning of that time unit) produces an income $r(l)$ in that time unit. If on inspection the machine is found to be performing at level $l (= 1, 2, 3)$ at the beginning of a time unit, it has a probability $p(l, m, s)$ that its performance level will be m at the end of the next s time units, with $p(l, m, s) = 0$ if $m \geqslant l$ (i.e. its performance degrades by a minimal

amount in each time unit). We allow the act of inspection to involve changes in the machine, without specifying these, which will influence the transition probabilities. $m = 0$ corresponds to a failed state. Failures, when they occur, occur at the end of a time unit. Formulate as semi-Markov decision process the problem of finding a policy to maximise the infinite horizon average expected reward per unit time.

CHAPTER 6

Partially observable and adaptive Markov decision processes

6.1 INTRODUCTORY REMARKS

In this chapter we will deal with stationary processes only. So far we have assumed that

(i) at any decision epoch the primitive state $i \in I$ is known;
(ii) $\{p_{ij}^k\}$, $\{r_i^k\}$ are known.

In what follows we will deviate from these assumptions.

We will first of all deal with what is known as 'partially observable Markov decision processes' in which $\{p_{ij}^k\}$, $\{r_i^k\}$ are known, but in which, instead of the state i being known at a decision epoch, we have some observation from which we can infer probabilities for the various states in I. Thus we may not know whether a machine is working properly (state 1) or not (state 2) and we may only have external evidence of the quality of its product. We then have to infer from this observation the probabilities of the machine being in either of its two states.

We will base our development on the paper by Smallwood and Sondik [47].

We will then consider 'adaptive Markov decision processes' as they are called, where the $\{p_{ij}^k\}$, $\{r_i^k\}$ depend upon some unknown vector parameter θ, but for which parameter we have an initial prior probability distribution. In this case the system may or may not be moving between specified primitive states $i \in I$. Thus, in sequential sampling, the quality p of a batch becomes the state i of the system, and this remains essentially the same throughout for large batches and small samples. In a queuing problem, whose primitive state i is the number in the system, we may not know the arrival rate, but the primitive state i changes and is known at each decision epoch.

The partially observable and adaptive Markov decision processes are, in essence, Bayesian decision processes but, because of their particular forms, are not referred to as such. For adaptive Markov decision processes we will make use of White [57], Chapter 6, and we make a brief reference to Martin [31].

In what follows all of the problems will take the form of standard Markov decision processes, arrived at by redefining the state of the system, in effect, to reflect the history of observations in different ways. The new state sets will not necessarily be finite and hence this takes us out of the realm, formally, of the class of problems which we have so far studied. None the less with suitable housekeeping requirements similar results are obtainable. We will not deal with the analysis leading up to the equations which we will derive and will leave intuition to take over the role of rigorous analysis.

6.2 PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

We make the same assumptions regarding the behaviour of our system as for standard Markov decision processes (see Section 2.2). We will, however, make the following additional assumptions:

(i) At any decision epoch the state $i \in I$ may not be known.

(ii) If the system is in state $i \in I$ at any decision epoch, if action $k \in K$ (see (vi)) is taken and the transformed state is $j \in I$, we also receive some observation from a finite set D whose realisation will be represented by d.

(iii) There is a probability q_{jd}^k that if action $k \in K$ (see (vi)) has been taken at the current decision epoch and the state at the next decision epoch is $j \in I$, then the realised observation $d \in D$ will be obtained.

(iv) There is a reward r_{ijd}^k given (i, j, k, d), assumed to be received at the beginning of the time unit. We let

$$r_i^k = \sum_{j \in I, d \in D} p_{ij}^k q_{jd}^k r_{ijd}^k. \tag{6.1}$$

(v) In order to model this problem our new state variable will be

$$\mu = (\mu_1, \mu_2, \ldots \mu_i, ., ., \mu_m) \in R_+^m \tag{6.2}$$

with

$$\sum_{i \in I} \mu_i = 1 \tag{6.3}$$

where μ_i is the current probability that the system is in state $i \in I$. The new state space is M.

Initially we will have, for $t = 1$

$$\mu = \mu^1. \tag{6.4}$$

(vi) Because we do not, in general, know $i \in I$ we will assume that $K(i)$ is independent of $i \in I$, i.e.

$$K(i) = K, \quad \forall\, i \in I. \tag{6.5}$$

(vii) A general policy set Π can be defined analogously to that of the Markov decision process case, where the history is now made up of the initial $\mu = \mu^1$ plus all subsequent observations and actions. However, for the purposes of the work in this chapter all we require is the current μ in (6.2).

We can restrict ourselves to deterministic Markov policies. A deterministic Markov decision rule δ is a function of μ, and Δ will be the set of all such decision rules.

If we take action $k \in K$ and at the next decision epoch observation $d \in D$ is obtained then the new posterior probability vector is $Q^{kd}\mu$, where

$$Q_\mu^{kd} = \left(\sum_{I \in I} \mu_I p_{Ii}^k q_{id}^k\right) \Big/ \left(\sum_{I \in I, j \in I} \mu_I p_{Ij}^k q_{jd}^k\right), \quad \forall\, i \in I. \tag{6.6}$$

The operator Q^{kd} is simply the operator transforming a prior probability vector μ to a posterior probability vector given $k \in K$, $d \in D$. This is simply the standard conditional probability result, the numerator in (6.6) being the new probability of being in state $i \in I$ and receiving information $d \in D$, and the denominator in (6.6) being the probability of receiving information $d \in D$, which is non-zero given that $d \in D$ has been received. If the denominator in (6.6) is zero then (6.6) is not relevant.

With $v_n(\mu)$ now being defined in a manner analogous to (2.8) also allowing a discount factor, we obtain formally the analogue of the optimality equation (2.58), viz.

$$\underline{n \geqslant 1} \qquad u_n = Mu_{n-1} \tag{6.7}$$

where, for u: $\mathrm{M} \to R$

$$[M^{\delta}u](\mu) = \sum_{l \in I} \mu_l r_l^{\delta(\mu)} + \rho \sum_{l \in I, j \in I, d \in D} \mu_l p_{lj}^{\delta(\mu)} q_{jd}^{\delta(\mu)} u(Q\mu^{\delta(\mu)d}), \quad (6.8)$$

$$Mu = \underset{\delta \in \Delta}{\text{maximum}} \; [M^{\delta}u], \quad (6.9)$$

$$\mathrm{M} = \left\{ \mu \in R_+^m : \sum_{l \in I} \mu_l = 1 \right\}. \quad (6.10)$$

The infinite horizon discounted version of equation (6.7) will give the following analogue of optimality equation (2.66):

$$u = Mu \quad (6.11)$$

where M is as defined in (6.8) and (6.9). An average expected reward per unit time version parallels equations (2.85) and (2.86). An absorbing state version parallels equations (2.127) and (2.128).

Solving equation (6.7) and its counterparts can be very difficult. We do, however, have one result which gives a piecewise linearity property which is of some use for small n in equation (6.7) (see Smallwood and Sondik [47]).

Result 6.1. In equation (6.7) if $u_0 = 0$ then u_n takes the form

$$\underline{n \geqslant 0} \qquad u_n(\mu) = \underset{1 \leqslant s \leqslant C(n)}{\text{maximum}} \; [\alpha_{ns}\mu] \quad (6.12)$$

where $\{\alpha_{ns}\} \subseteq R^m$ and are given by the recurrence relation (6.15), the members of K are numbered as $b = 1, 2, ., ., \#K$, and

$$C(n) = (\#K)^n. \quad (6.13)$$

Let $n \geqslant 2$ and

$$\underline{1 \leqslant s \leqslant C(n)} \qquad s = \#K(a-1) + b, \quad 0 \leqslant a \leqslant C(n-1), \quad 1 \leqslant b \leqslant \#K. \quad (6.14)$$

Then

$$\underline{n \geqslant 1} \qquad \alpha_{nsi} = r_i^b + \rho \sum_{j \in I, d \in D} p_{ij}^b q_{jd}^b \alpha_{n-1,aj}, \quad \forall\, i \in I, \quad (6.15)$$

where $\{r_i^k\}, \{p_{ij}^k\}, \{q_{jd}^k\}$ are rewritten as $\{r_i^b\}, \{p_{ij}^b\}, \{q_{jd}^b\}$.

Proof. The result is trivially true for $n = 1$ after setting, formally, $C(0) = 1$, $\alpha_{00i} = 0$, $\forall\, i \in I$.

Suppose it is true with $n-1$ replacing n in (6.12), (6.15) for some n. Then substituting on the right-hand side (6.12) for $n-1$ into equation (6.7) we obtain, using b for $\delta(\mu)$, i for l in (6.8), and a for s in (6.12)

$$u_n(\mu) = \underset{1 \leqslant b \leqslant \#K}{\text{maximum}} \left[\sum_{i \in I} \mu_i r_i^b + \rho \sum_{i \in I, j \in I, d \in D} \mu_i p_{ij}^b q_{jd}^b \underset{1 \leqslant a \leqslant C(n-1)}{\text{maximum}} [\alpha_{n-1,a} Q^{bd} \mu] \right] \tag{6.16}$$

$$= \underset{1 \leqslant b \leqslant \#K, 1 \leqslant a \leqslant C(n-1)}{\text{maximum}} [\beta_{na}^b(\mu)] \tag{6.17}$$

where

$$\beta_{na}^b(\mu) = \sum_{i \in I} r_i^b \mu_i + \rho \sum_{i \in I, j \in I, d \in D} \mu_i p_{ij}^b q_{jd}^b \frac{\left(\sum_{i \in I, j \in I} q_{jd}^b p_{ij}^b \alpha_{n-1,aj} \mu_i \right)}{\left(\sum_{i \in I, j \in I} q_{jd}^b p_{ij}^b \mu_i \right)} \tag{6.18}$$

$$= \sum_{i \in I} \left(r_i^b + \rho \sum_{j \in I, d \in D} p_{ij}^b q_{jd}^b \alpha_{n-1,aj} \right) \mu_i. \tag{6.19}$$

Thus if we put

$$\alpha_{nsi} = r_i^b + \rho \sum_{j \in I, d \in D} p_{ij}^b q_{jd}^b \alpha_{n-1,aj} \tag{6.20}$$

and use (6.14) we see that we have our requisite form. ○

In this formulation we have assumed that the observation is obtained after the state at the next decision epoch has been realised. For some problems the observation is obtained at the current decision epoch after the current decision has been taken. For example, in the search problem which we will discuss later (see p. 138), a location is searched and the observation is obtained at that time and then, if the target is not located, it may move to a new location. The only difference to the earlier model is that, in (iii), p. 131, the new state j is replaced by the current state i.

The net effect is as follows:

(a) replace q_{id}^{k} by q_{id}^{k} in the numerator of (6.6);
(b) replace q_{jd}^{k} by q_{id}^{k} in the denominator of (6.6);
(c) replace $q_{jd}^{\delta(\mu)}$ by $q_{id}^{\delta(\mu)}$ in (6.8);
(d) replace q_{jd}^{b} by q_{id}^{b} in (6.15);
(e) replace q_{jd}^{b} by q_{id}^{b} in (6.16);
(f) replace q_{jd}^{b} by q_{id}^{b} in numerator and denominator of (6.18);
(g) replace q_{jd}^{b} by q_{id}^{b} in (6.19);
(h) replace q_{jd}^{b} by q_{id}^{b} in (6.20).

Here q_{id}^{k} in (iii) p. 131 is now interpreted as the probability that if the system is in state $i \in I$ and decision $k \in K$ taken then observation $d \in D$ will be immediately received; r_{ijd}^{k} in (iv) p. 131 is also redefined, with $d \in D$ as the immediate observation for current state $i \in I$, action $k \in K$ and transformed state $j \in I$.

6.2.1 SOME ILLUSTRATIONS

(a) Machine maintenance (Smallwood and Sondik [47])

We make the following assumptions:

(i) A machine has two identical components each of which may be in a failed or in a non-failed state.
(ii) Each component has a probability of 0.1 of failing when it carries out its operation on a unit of product, if it is in a non-failed state prior to this, independently of the number of operations carried out since its last non-failed state.
(iii) Each unit of product has to be operated on by both components.
(iv) A non-failed component does not produce a defective unit of product. A failed component produces a defective unit of product with probability 0.5.
(v) These are four possible actions which can be taken at each decision epoch immediately following the observation, if any, from the result of the previous combined operations after one unit of product has been operated on by the machine, viz.

Action	*Description*
1	no physical action and no observation obtained;
2	examine the unit of product;

3 inspect the machine and replace failed components if any;

4 replace both components, with no inspection.

(vi) There are various costs and profits associated with the quality of the product and the actions taken.

The net result of all the probabilities, costs and profits is given in Table 6.1, where $i = 1, 2, 3$ corresponds to zero, one, or two failed components respectively. Here $d = 0$ will correspond to a known non-defective unit of product or to no examination of the unit of product, and $d = 1$ will correspond to a known defective unit product. Although $d = 0$ means two different things, the meaning of d is determined by k and thus $\{q_{id}^k\}$, $\{r_{id}^k\}$ are well defined.

(vii) The problem is to find a unit product examination and component inspection decision rule to maximise the expected total net profit over a specified number n of units of product. Thus, time is replaced by number of units of product. The theory is exactly the same as that for time.

Table 6.1 Data for the maintenance problem*

State	*Action*	*Transition probability* p_{ij}^k			*Observation probability* q_{id}^k		*Expected reward* r_i^k
		j			*d*		
i	k	1	2	3	0	1	
1	1	0.81	0.18	0.01	1.00	0.00	0.9025
	2	0.81	0.18	0.01	1.00	0.00	0.6525
	3	1.00	0.00	0.00	1.00	0.00	−0.5000
	4	1.00	0.00	0.00	1.00	0.00	−2.0000
2	1	0.00	0.90	0.10	1.00	0.00	0.4750
	2	0.00	0.90	0.10	0.50	0.50	0.2250
	3	1.00	0.00	0.00	1.00	0.00	−1.5000
	4	1.00	0.00	0.00	1.00	0.00	−2.0000
3	1	0.00	0.00	1.00	1.00	0.00	0.2500
	2	0.00	0.00	1.00	0.25	0.75	0.0000
	3	1.00	0.00	0.00	1.00	0.00	−2.5000
	4	1.00	0.00	0.00	1.00	0.00	−2.0000

* Table 6.1 and Figures 6.1 and 6.2 are reproduced from [47] Smallwood and Sondik (1973), *Operations Research*, **21**, by permission of the authors and ORSA.

This problem falls into the class of problems in which the observation d is obtained after the current decision has been taken, prior to the next decision, and thus we need the interpretation given on p. 135. It is a finite horizon Markov decision process with state set

$$\mathrm{M} = \{\mu \in R_+^3 : \mu_1 + \mu_2 + \mu_3 = 1\} \tag{6.21}$$

where μ_i is the probability that the system is in state i, $i \in \{1, 2, 3\} = I$.

We will not undertake all the calculations. For $n = 0, 1$ we have the following:

$\underline{n = 0}$ $\quad u_0(\mu) = 0, \quad \forall\ \mu \in \mathrm{M}$

$\underline{n = 1}$ $\quad u_1(\mu) =$

$$\text{maximum} \begin{bmatrix} k=1 & 0.9025\mu_1 + 0.4750\mu_2 + 0.2500u_3 \\ k=2 & 0.6525\mu_1 + 0.2250\mu_2 + 0.0000\mu_3 \\ k=3 & -0.5000\mu_1 - 1.500\mu_2 - 2.5000\mu_3 \\ k=4 & -2.000\mu_1 - 2.000\mu_2 - 2.000\mu_3 \end{bmatrix}, \quad \forall\ \mu \in \mathrm{M}.$$

For all $\mu \in \mathrm{M}$ the optimal decision rule is, for $n = 1$

$$\sigma_1(\mu) = 1.$$

For $n = 3, 4$, Smallwood and Sondik [47] give the following triangular coordinates representation of optimal decision rules $\{\sigma_3, \sigma_4\}$ given in Figures 6.1 and 6.2.

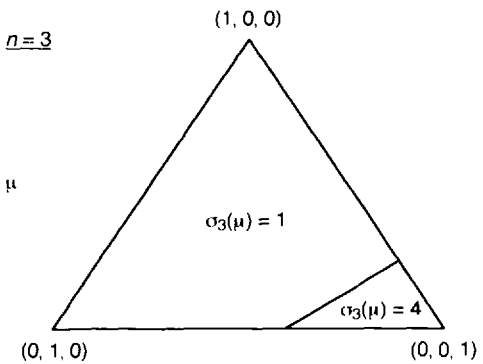

Figure 6.1 Optimal decision rules for maintenance problem*

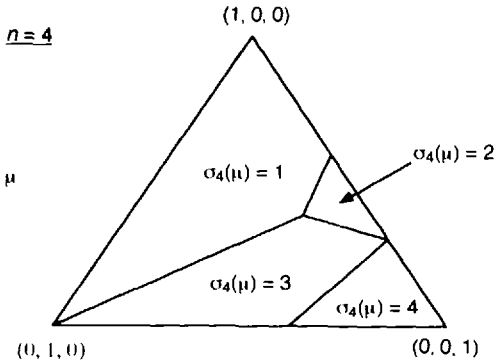

Figure 6.2 Optimal decision rules for maintenance problem*

(b) Search (Ross [40], pp. 67–68)

We make the following assumptions:

(i) A target is in one of m locations $i \in I = \{1, 2, 3, ., ., m\}$, with the temporal location unknown to the searcher until its location is determined.
(ii) The searcher chooses a location $i \in I$ to search at a cost c_i.
(iii) If the target is in location $i \in I$ there is a probability α_i that it will be found when that location is searched.
(iv) The search terminates when the target location has been determined (i.e. actually found by a search or is known to be in a specific location).
(v) If the target is in location $i \in I$ but has not been located, it has a probability p_{ij} of moving to location $j \in I$.
(vi) The problem is to find a search policy which minimises the expected total cost up to the time of locating the target.

This is an infinite horizon absorbing state Markov decision process with the state set

$$\mathrm{M} = \left\{ \mu \in R_+^m : \sum_{i \in I} \mu_i = 1,\ \mu \geqslant 0 \right\} \tag{6.22}$$

where μ is the current probability vector for the primitive states in I. The absorbing states are $\{\mu^i\}$, $i \in I$, where $\mu_j^i = 1$ if $i = j$ and $\mu_j^i = 0$ if $i \neq j$, for all $i, j \in I$. Also $K = I$.

In this case we need to use the q_{id}^k formulation of p. 135, where q_{id}^k is the probability that if we search location $k \in I$ and the target is in location $i \in I$ then the outcome will be d, where $d = 0$ corresponds to not finding the target and $d = 1$ corresponds to finding the target.

Thus

$$q_{i0}^k = 1 \quad \text{if } i \neq k,\ i,\ k \in I,$$
$$= (1 - \alpha_k) \quad \text{if } i = k,\ i,\ k \in I, \tag{6.23}$$

$$q_{i1}^k = 1 - q_{i0}^k,\ i,\ k \in I. \tag{6.24}$$

We also have $r_i^k = -c_k$ for all $i, k \in I$ in order to be consistent with our standard maximisation format.

Noting the comment (a) on $\{q_{id}^k\}$ (see p. 135) we have (see (6.6))

$$[Q^{k0}\mu]_i = \frac{\sum_{l \in I, l \neq k} \mu_l p_{li} + (1 - \alpha_k)\mu_k p_{ki}}{\left(\sum_{l \in I, l \neq k, j \in I} \mu_l p_{lj} + (1 - \alpha_k)\mu_k \sum_{j \in I} p_{kj}\right)}$$

$$= \frac{\sum_{l \in I} \mu_l p_{li} - \alpha_k \mu_k p_{ki}}{(1 - \alpha_k \mu_k)}, \quad \forall\ i, \quad k \in I. \tag{6.25}$$

In the summation on the right-hand side of (6.8) for $d = 0$, extended to the absorbing state case, we have (modifying (6.8) in accordance with (c) on p. 135)

$$\sum_{l \in I, j \in I} \mu_l p_{lj}^k q_{l0}^k = (1 - \alpha_k \mu_k), \quad \forall\ k \in I. \tag{6.26}$$

Then (6.11) (for the absorbing state case) takes the form

$$u(\mu) = \underset{k \in I}{\text{maximum}} \left[-c_k + (1 - \alpha_k \mu_k) u\left(\frac{\sum_{l \in I} \mu_l p_l - \alpha_k \mu_k p_k}{(1 - \alpha_k \mu_k)} \right) \right],$$
$$\forall\ \mu \in \mathrm{M} \backslash \{\mu'\}, \tag{6.27}$$

$$u(\mathrm{M}^i) = 0, \quad \forall\ i \in I, \tag{6.28}$$

where

$$p_k = (p_{k1}, p_{k2}, ., ., p_{km}). \tag{6.29}$$

Equation (6.27) is more easily derived directly than by using (6.8).

For the case of $m = 2$, if we put $\mu_1 = p$, $\mu_2 = 1 - p$ then equations (6.27) and (6.28) take the following form, noting that $p_{i1} + p_{i2} = 1$, for all $i \in I$:

$\underline{p \notin \{0, 1\}}$

$$u(p) = \text{maximum} \left[\begin{array}{l} k = 1: \; -c_1 + (1 - \alpha_1 p)u \\ \qquad \times \left(\dfrac{(1-\alpha_1)}{(1-\alpha_1 p)} \, p p_{11} + \dfrac{1-p}{(1-\alpha_1 p)} \, p_{21} \right) \\ k = 2: \; -c_2 + (1 - \alpha_2(1-p))u \\ \quad \times \left(\dfrac{p}{(1-\alpha_2(1-p))} \, p_{11} + \dfrac{(1-\alpha_2)(1-p)}{(1-\alpha_2(1-p))} \, p_{21} \right) \end{array} \right], \tag{6.30}$$

$$u(1) = u(0) = 0. \tag{6.31}$$

In (iv) (p. 138) we have assumed that the process terminates when the target is physically located by a search, or if $\mu = \mu^i$ for some $i \in I$. This is a slight deviation from the actual problem of Ross [40], where the process terminates only when an actual search finds the target. In this case (6.27) applies for all $\mu \in \mathrm{M}$, and (6.28) is deleted. Similarly (6.30) applies for all $p \in [0, 1]$, and (6.31) is deleted.

6.3 ADAPTIVE MARKOV DECISION PROCESSES (White [57], Martin [31])

We make the following assumptions:

(i) At any decision epoch the primitive state $i \in I$ is known.
(ii) If we are in state $i \in I$ at any decision epoch and take action $k \in K(i)$ the probability of moving to primitive state $j \in I$ and receiving an observation $d \in D$ is $p^k_{i\theta jd}$, where θ is a fixed, but possibly unknown, state of nature belonging to a finite set Θ.
(iii) There is an immediate reward $r^k_{i\theta jd}$ whose expectation over $j \in I$, $d \in D$, given $i \in I$, $k \in K(i)$, $\theta \in \Theta$, is written as $r^k_{i\theta}$.

In order to model this problem our new state will be

$$(i, \mu), \; \mu = (\mu_1, \mu_2, ., \mu_\theta, ., ., \mu_M), \tag{6.32}$$

where μ_θ is the current probability that the state of nature is $\theta \in \Theta$. We

let M be the set of μ in (6.32). Initially we will have a prior probability vector for the states of nature, viz.

$$\mu = \mu^1. \tag{6.33}$$

Following a similar reasoning to that of p. 132, if we take action $k \in K(i)$ and move to primitive state $j \in I$ from $i \in I$ receiving observation $d \in D$ the transformed state takes the form

$$(j, Q^{kijd}\mu) \tag{6.34}$$

where

$$[Q^{kijd}\mu]_\theta = \frac{\mu_\theta p^k_{i\theta jd}}{\sum_{\phi \in \Theta} \mu_\phi p^k_{i\phi jd}}. \tag{6.35}$$

The analogue of equation (6.7) for a finite number of decision epochs n is

$$\underline{n \geqslant 1} \qquad u_n = Nu_{n-1} \tag{6.36}$$

where, for u: $I \times \mathrm{M} \to R$

$$[N^\delta u](i, \mu) = \sum_{\phi \in \Theta} \mu_\phi r_{i\phi}^{\delta(i,\mu)} + \rho \sum_{\phi \in \Theta, j \in I, d \in D} \mu_\phi p_{i\phi jd}^{\delta(i,\mu)} u(j, Q^{\delta(i,\mu)ijd}\mu), \tag{6.37}$$

$$Nu = \underset{\delta \in \Delta}{\text{maximum}} \; [N^\delta u] \tag{6.38}$$

where now Δ is defined on $I \times \mathrm{M}$. Infinite horizon expected total discounted reward and finite and infinite horizon absorbing state formulations follow in a similar manner.

For some classes of problem it is not necessary to use the state form (i, μ) given by (6.32). This applies when μ can be expressed as a function of some parameter ψ which is transformed in a specific manner to a new parameter $S^{kijd}\psi$ given i, j, d, k, and where $\mu_\theta = f(\theta, \psi)$, $\theta \in \Theta$. For this to be possible, putting $\mu_\theta = f(\theta, \psi)$, $\mu_\varphi = f(\varphi, \psi)$ in (6.35) must reduce to $f(\theta, S^{kijd}\psi)$. In this case (i, μ) is replaced by (i, ψ) and (6.36) still holds where, in (6.37), $Q^{\delta(i,\mu)ijd}\mu$ is replaced by $S^{\delta(i,\mu)ijd}\psi$.

In effect, for the above approach to apply, each μ must belong to a parametric set of probability distributions which is closed under the transformations brought about by k, i, j, d. This approach is developed by Martin [31], Chapter 3. The sequential sampling illustration described in the next section falls into this category if we begin with a specific form of prior distribution.

Further information on this approach, and on 'sufficient statistics', can be found in White [57].

6.3.1 SOME ILLUSTRATIONS

(a) Sequential sampling

We make the following assumptions:

(i) We have a large batch of items containing an unknown proportion θ of defective items.
(ii) The possible θ values form a set Θ.
(iii) At any decision epoch we can accept or reject the batch without further sampling, with respective expected costs $a(\theta)$, $b(\theta)$ if $\theta \in \Theta$ is the level of the defectives proportion, or we may take a sample of size s at a cost $c(s)$ with $0 \leqslant s \leqslant \bar{s}$ and then continue the accept/reject/sample sequence in the light of the number of defective items found.
(iv) The problem is to find an optimal accept/reject/sample size policy to minimise the expected total cost up to the ultimate accept/reject decision.

In this case there is no primitive state set I and we just require the set M. The observation received is the number of defective items in a sample. Thus $\{p^k_{i\theta jd}\}, \{r^k_{i\theta}\}$ reduce to $\{p^k_{\theta d}\}, \{r^k_\theta\}$, where $D = \{d\colon 0 \leqslant d \leqslant \bar{s}\}$ and $K = \{1, 2, 3.s\}$ corresponds to accept, reject or take a further sample of size s respectively. Also the number of levels of θ is infinite but we may still apply our theory.

Then we have the following:

$$p^{3.s}_{\theta d} = (s!/(d!(s-d)!))\theta^d(1-\theta)^{s-d}, \quad \forall\ \theta \in \Theta,\ d \in D,\ 0 \leqslant d \leqslant s \leqslant \bar{s}, \tag{6.39}$$

$$p^k_{\theta d} = 0, \quad \text{for } k = 1, 2, \quad \forall\ \theta \in \Theta, \quad d \in D, \tag{6.40}$$

$$r^k_\theta = -a(\theta) \quad \text{for } k = 1, \quad \forall\ \theta \in \Theta, \tag{6.41}$$

$$= -b(\theta) \quad \text{for } k = 2, \quad \forall\ \theta \in \Theta, \tag{6.42}$$

$$= -c(s) \quad \forall\ k = 3.s, \quad 1 \leqslant s \leqslant \bar{s}. \tag{6.43}$$

Equation (6.36) then takes the form, with $u_0 = 0$

$n \geqslant 1$

$$u_n(\mu) = \text{maximum} \left[\begin{array}{l} k = 1: \quad -\sum_{\phi \in \Theta} \mu_\phi a(\phi) \\ \\ k = 2: \quad -\sum_{\phi \in \Theta} \mu_\phi b(\phi) \\ k = 3.s: \; -c(s) \\ + \sum_{\phi \in \Theta, 0 \leqslant d \leqslant s} \mu_\phi (s!/(d!(s-d)!)) \phi^d (1-\phi)^{s-d} \\ \qquad \times u_{n-1}(Q^{3.sd}\mu) \end{array} \right] \tag{6.44}$$

where, dropping the i and j in (6.35)

$$[Q^{3.sd}\mu]_\theta = \frac{\mu_\theta \theta^d (1-\theta)^{s-d}}{\left(\sum_{\phi \in \Theta} \mu_\phi \phi^d (1-\phi)^{s-d} \right)}, \quad \forall\, d \in D, \quad \mu \in \mathrm{M}, \quad 0 \leqslant s \leqslant \bar{s}. \tag{6.45}$$

This problem fits into the closed distribution form of the previous section if μ_θ takes a special form. Thus suppose that

$$\mu_\theta = f(\theta, \psi) = A(\psi)\theta^\alpha (1-\theta)^\beta \tag{6.46}$$

with

$$\psi = (\alpha, \beta). \tag{6.47}$$

Then, from (6.45) we see that

$$S^{3.sd}\psi = (\alpha + d, \beta + s - d). \tag{6.48}$$

(b) Inventory

On p. 54 we discussed an inventory problem where $q(s)$ was the probability that the demand was s, $0 \leqslant s \leqslant \bar{s}$. Suppose now that $q(s)$

is replaced by $q(s, \theta)$ where $\theta \in \Theta$ is an unknown parameter. Then the analogue of equation (2.130) is

$$u(i,\mu) = \underset{m \geqslant k \geqslant i}{\text{maximum}}$$

$$\times \left[\begin{array}{l} -\Big(c(k-i) \\ \\ + \sum\limits_{\phi \in \Theta, s > k} \mu_\phi q(s,\phi) l(s-k) \\ \\ + \frac{1}{2} a\Big(k + \sum\limits_{\phi \in \Theta, 0 \leqslant s \leqslant k} \mu_\phi q(s,\phi)(k-s)\Big)\Big) \\ \\ + \rho \sum\limits_{0 \leqslant s < k, \phi \in \Theta} \mu_\phi q(s,\phi) u\left(k-s, \dfrac{\mu \,.\, q(s,.)}{\left(\sum\limits_{\phi \in \Theta} \mu_\phi q(s,\phi)\right)}\right) \\ \\ + \rho \sum\limits_{s \geqslant k, \phi \in \Theta} \mu_\phi q(s,\phi) u\left(0, \dfrac{\mu \,.\, q(s,.)}{\left(\sum\limits_{\phi \in \Theta} \mu_\theta q(s,\phi)\right)}\right) \end{array} \right], \quad \forall (i,\mu) \in I \times \mathrm{M} \quad (6.49)$$

where, in (6.49), $\mu \,.\, q(s,.)$ is the vector whose θ component is $\mu_\theta q(s,\theta)$, $\theta \in \Theta$, $0 \leqslant s \leqslant \bar{s}$.

6.4 EXERCISES FOR CHAPTER 6

1. In an experiment there are two of states of nature $\theta \in \{1,2\}$. If the experiment is conducted there are two possible experimental results $d \in \{1,2\}$. The probabilities of obtaining $d \in \{1,2\}$ given $\theta \in \{1,2\}$ are given in the following table.

		State of nature θ	
		1	2
Experimental outcome d	1	$\frac{1}{3}$	$\frac{2}{3}$
	2	$\frac{2}{3}$	$\frac{1}{3}$

You make a series of decisions. You can either terminate the process by taking one of two actions $k \in \{1, 2\}$ or obtain a further observation by carrying out the above experiment at a cost of 1 unit. The losses associated with any terminal action are given in the following table:

		State of nature θ	
		1	2
Action k	1	0	20
	2	20	0

(i) If p is the current probability of the state of nature being $\theta = 1$, formulate with careful explanation, as an adaptive Markov decision process, the problem of finding a policy to minimise the expected sum of experimental costs and losses up to and including a terminal action, allowing an indefinite number of experiments if desired, with terminal decisions being possible after each experiment.

(ii) Solve the problem for all $0 \leqslant p \leqslant 1$ if you are allowed to experiment at most once, by finding the optimal actions for each p.

Do both parts in minimisation form without converting to maximisation form.

2. In an experiment there are two states of nature $\theta \in \{1, 2\}$. If the experiment is conducted there are two possible experimental results $d \in \{1, 2\}$. The probabilities of obtaining $d \in \{1, 2\}$ given $\theta \in \{1, 2\}$ are given in the following table:

		State of nature θ	
		1	2
Experimental outcome d	1	0	$\frac{2}{3}$
	2	1	$\frac{1}{3}$

You make a series of decisions. You can either terminate the process by taking one of two actions, $k \in \{1, 2\}$ or obtain a further observation by carrying out the above experiment at a cost of 1 unit.

The losses associated with any terminal action are given in the following table:

		State of nature θ	
		1	2
Action k	1	8	0
	2	0	8

(i) If p is the current probability of the state of nature being $\theta = 1$, formulate as an adaptive Markov decision process the problem of finding a policy to minimise the expected sum of experimental costs and losses up to and including a terminal action, allowing an indefinite number of experiments if desired, with terminal actions being possible after each experiment,

(ii) Solve the problem for all $0 \leqslant p \leqslant 1$ when you are allowed to experiment at most once, by finding the optimal actions for each p.

Do both parts in minimisation form without converting to maximisation form.

CHAPTER 7

Further aspects of Markov decision processes

In this chapter we will briefly cover some aspects of Markov decision processes which supplement the earlier material. We will confine ourselves largely to infinite horizon stationary discounted reward Markov decision processes for ease of exposition, although we will occasionally digress from this, but most of the aspects with which we will deal have their analogues in the various other sorts of Markov decision processes with which we have dealt. A survey of Markov decision processes is given in White and White [56].

Our basic optimality equation when our criterion is that of expected total discounted reward is

$$u = Tu \tag{7.1}$$

(see (2.66) and (2.55)–(2.57)).

7.1 STRUCTURED POLICIES

Consider the inventory problem of p. 54.

It may be shown (see Bellman [4], p. 160 for the continuous demand case) that if $c(.)$ and $l(.)$ are linear functions then, recognising its dependence on the discount factor ρ, an optimal policy π_ρ takes the form $\pi_\rho = (\delta_\rho)^\infty$ where

$$\delta\rho = \max[i, k_\rho], \quad \forall\ i \in I \tag{7.2}$$

for some k_ρ which can be computed as a function of ρ. Thus, k_ρ is a critical stock level, and if $i < k_\rho$ we order a quantity $k_\rho - i$, otherwise ordering zero.

Thus, the policy has a special structural form. In particular we see that

$$i' \geqslant i \rightarrow \delta_\rho(i') \geqslant \delta_\rho(i). \tag{7.3}$$

It may also be shown that

$$\rho' \geqslant \rho \rightarrow k_{\rho'} \leqslant k_\rho. \tag{7.4}$$

The results in (7.3) and (7.4) give rise to the following:

(i) a relationship between the order $\geqslant$ over I and order $\geqslant$ over $K = \cup_{i \in I} K(i)$;
(ii) a relationship between the order $\geqslant$ over $[0,1)$ and the order $\geqslant$ over K.

These are special cases of a more general isotonicity characteristic of some Markov decision processes for which the general paradigm for the result in (7.3) would take the following form, where a quasi-order is any binary relation which is reflexive and transitive:

(i) a quasi-order $\geqslant^*$ over I exists;
(ii) a quasi-order $\geqslant$ over K exists;
(iii) an optimal policy $\pi = (\delta)^\infty$ exists for which

$$i' \geqslant^* i \rightarrow \delta(i') \geqslant \delta(i). \tag{7.5}$$

Here $(\geqslant^*, \geqslant)$ correspond to $(\geqslant, \geqslant)$ for (7.3) in the inventory problem.

Results of the kind given by (7.5) may be obtained in various ways. For example the result in (7.2) is obtainable by a form of convexity analysis. Other results may be obtained using lattice and supermodularity properties (e.g. see White [54] and White [61], [66]). We will not deal with these in general but consider the following example (see White [66]).

Consider our inventory problem but with N commodities instead of one commodity. Thus, our state is $i \in Z_+^N$, the set of integer non-negative vectors of dimension N, and $k \in Z_+^N$ with

$$K(i) = \{k \geqslant i\}, \quad \forall\, i \in I. \tag{7.6}$$

For $i, i' \in I, k, k' \in K$, let

$$\underline{k} = \text{vector minimum}\,[k', k] \tag{7.7}$$

i.e.

$$\underline{k}_\alpha = \text{minimum}\,[k'_\alpha, k_\alpha], \qquad 1 \leqslant \alpha \leqslant N, \tag{7.8}$$

$$\bar{k} = \text{vector maximum}[k', k], \tag{7.9}$$

i.e.

$$\bar{k}_\alpha = \text{maximum}[k'_\alpha, k_\alpha], \qquad 1 \leqslant \alpha \leqslant N \tag{7.10}$$

with similar definitions for $\{\bar{i}, \underline{i}\}$.

Suppose that, using operator T in maximisation form with ρ fixed which we suppress, we have for $u = v$ the solution to (7.1)

$$[T^{\bar{k}}v](i) + [T^{\underline{k}}v](\underline{i}) \geqslant [T^{k'}v](i') + [T^{k}v](i). \tag{7.11}$$

Inequality (7.11) is a 'supermodularity' condition. Now let

$$i' \geqslant i. \tag{7.12}$$

Because $\bar{i} = i'$, $\underline{i} = i$ and because $\bar{k} \in K(i')$, $\underline{k} \in K(i)$ we have $\bar{k}$ feasible for i', $\underline{k}$ feasible for i and from (7.11)

$$[T^{k}v](i') - [T^{\underline{k}}v](i') \geqslant [T^{k}v](i) - [T^{\underline{k}}v](i). \tag{7.13}$$

The right-hand side of (7.13) is non-negative if k is optimal for i. Then we also have $\bar{k}$ optimal for i'.

Also $\bar{k} \geqslant k$. Thus, if we set $\delta(i') = \bar{k}$ we have property (7.5) where $\geqslant^*$, $\geqslant$ are the same as $\geqslant$, i.e. an isotone optimal policy $\pi = (\delta)^\infty$ exists. Conditions for property (7.13) to hold may be developed along the lines of those given in White [66], although that paper specifically relates to an extension of property (7.4).

These results may be used for action elimination purposes in a similar manner to that specified in Chapter 3 where upper and lower bounds on v were used. In the current context, if an optimal action $k \in K(i)$ is known for state $i \in I$ then, given the isotonicity condition (7.5), if $i' \geqslant^* i$ we may restrict $\delta(i')$ to $\delta(i') \geqslant \delta(i)$. In the case of parametric isotonicity, such as that given in (7.4), if an optimal policy is known for a given ρ (e.g. $\rho = 0$) then if $\rho' \geqslant \rho$ we can restrict $k_{\rho'}$ to $k_{\rho'} \leqslant k_\rho$.

Finally (see White [61]) isotonicity ideas can be extended to the value iteration scheme (3.26) giving rise to the existence of optimal $\{\sigma_n\}$ such that if $n' \geqslant n$ then $\sigma_{n'}(i) \geqslant \sigma_n(i)$, $\forall\, i \in I$. This is again useful for action elimination purposes coupled with action elimination scheme (3.192).

There is another form of structure which relates to the form of v rather than to the form of an optimal policy. A simple case arises in the inventory problem of Bellman [4] where for $i \leqslant k_\rho$ we have $v_\rho(i) = v_\rho(0) - \gamma i$. Another structural form is that of separability of v. Mendelssohn [33], in a salmon-harvesting exercise, where

$i = x = (x_1, x_2, ., ., ., x_p)$ with x_α being the size of the fish population in category α, shows that $u(x) = \sum_{i=1}^{p} v_i(x_i)$. This is generalised by Lovejoy [29] for the expected total discounted reward case and by White [69] for the average expected reward per unit time case. Clearly the knowledge of such situations facilitates the computational procedures.

7.2 APPROXIMATION MODELLING

The computation of a solution of equation (7.1) will usually be extremely difficult if the state space I is large, i.e. the state space I has a large number $\# I$ of states in the standard finite form or has a large dimension if the state space I is a subset of R^N for some N, e.g. see the partially observable Markov decision process case of Chapter 6 with $N = m$. The natural procedure is to modify equation (7.1) to make it more tractable.

For the standard Markov decision process with finite I one way is to group states together and represent each group by a superstate in a new state space, $\tilde{I}$. Let $\tilde{I} = \{1, 2, ., ., ., \tilde{m}\}$ with generic member $\tilde{i}$ and let $\tilde{k}$ be the generic action.

We need to devise appropriate rewards, $\{\tilde{r}_i^{\tilde{k}}\}$ and transition probabilities $\{\tilde{p}_i^{\tilde{k}}\}$. For example, given $\tilde{i} \in \tilde{I}$ let $i^* \in \tilde{i}$ be a special state and, noting that $i^* = i^*(\tilde{i})$, let

$$\tilde{K}(\tilde{i}) = K(i^*) = K(i), \quad \forall\, i \in \tilde{i}, \tag{7.14}$$

$$\tilde{r}_{\tilde{i}}^{\tilde{k}} = r_{i^*}^{k}, \ \forall\, \tilde{k} \in \tilde{K}(\tilde{i}), \tag{7.15}$$

$$\tilde{p}_{\tilde{i}\tilde{j}}^{\tilde{k}} = \sum_{j \in \tilde{j}} p_{i^* j}^{k}, \quad \forall\, \tilde{k} \in \tilde{K}(\tilde{i}), \quad \tilde{j} \in \tilde{I}. \tag{7.16}$$

Equation (7.1) is then replaced by

$$\tilde{u} = \tilde{T}\tilde{u} \tag{7.17}$$

over the smaller state space $\tilde{I}$ where $\tilde{T}$ is defined analogously to T in (2.55)–(2.57). The solution to equation (7.17) may then be used to construct a solution to equation (7.1). A natural way to do this is as follows. Suppose an optimal decision rule $\tilde{\delta}$: $\tilde{I} \to \tilde{K} = \cup_{\tilde{i} \in \tilde{I}} \tilde{K}(\tilde{i})$ is found by solving equation (7.17). Then a decision rule δ for the original

Markov decision process might be

$$\delta(i) = \bar{\delta}(\bar{i}), \quad \forall\, i \in \bar{i}, \quad \bar{i} \in \bar{I}. \tag{7.18}$$

Schweitzer et al. [44] use a somewhat more sophisticated extension of (7.14)–(7.17) which allows an interlinking of the solution of (7.14)–(7.17) for a given decision rule $\bar{\delta}$ and solutions to subproblems generated when (7.1) is restricted to $i \in \bar{i}$ with appropriate modifications to cater for $i \notin \bar{i}$ for a given $\bar{i}$. The iterative procedure moves sequentially between the solutions in the state space $\bar{I}$ and the solutions in the state space I restricted to $\bar{i}$ for each $\bar{i} \in \bar{I}$. The procedure leads to a sequence of functions from I to R which converge to the solution of (7.1).

Mendelssohn [32] uses a primal–dual approach with a similar state aggregation procedure and a corresponding convergence result is established.

In each of these cases, however, it is either necessary to store computed values for each state $i \in I$, or these values are computed as required to avoid storing them. The subproblems are, however, somewhat smaller than the original problem.

State aggregation may be achieved in a somewhat different way to that given by (7.14)–(7.17). Thus, suppose that $I \subseteq R^N$, with generic member $i = x = (x_1, x_2, ., ., ., x_N)$. Then I might be transformed to $I^* \subseteq R^{N^*}$, with generic member $x^* = (x_1^*, x_2^*, ., ., ., x_{N^*}^*)$, e.g. $x_1^* = x_1 + x_2$, $x_2^* = x_3 + x_4 + x_5$, ., ., $x_{N^*} = x_{N-1} + x_{N-2}$ with appropriately transformed rewards and transition probabilities and a rule for disaggregation to convert a policy defined on I^* to one defined on I. Terry et al. [49] apply this approach to the control of a hydrothermal system although it is a one-pass application with no iterative aggregation–disaggregation. Buzacott and Callahan [10] apply this approach to a production scheduling problem.

An alternative procedure to aggregation–disaggregation is decomposition–recomposition. In this case a multicomponent system is split into a smaller number of single or multiple-component systems and appropriate Markov decision process models are solved for each of the smaller systems. In each of these smaller systems some allowance is made for the effects of the remaining systems. Turgeon [52] applies this approach to a multiple-reservoir multiple-river system. In one approach each river is treated separately assuming operational rules for all the other rivers. In a second approach each river is treated separately with an aggregation of the remaining rivers. A final value aggregation

approach is used to find solutions to the original problem. Both approaches are used in a one-pass mode. A similar procedure is used by Wijngaard [79] and White [64] for multiple-item inventory control where, in the latter case, a Lagrangian relaxation method is given together with error bounds. We will outline the approach of White [64] for an inventory problem with a single resource constraint, in maximisation form.

Suppose that there is a single resource constraint so that if $\{k_\alpha\}$ are the new inventory levels after a vector purchase order is placed, for some specified total stock constraint level b we have to observe the following inequality:

$$\sum_{\alpha=1}^{N} (k_\alpha - i_\alpha) \leqslant b. \tag{7.19}$$

If we introduce inequality (7.19) via a Lagrange multiplier $\lambda \in R_+$ then in $\{r_i^k\}$ form the new r_i^k takes the form

$$r_i^k(\lambda) = \sum_{\alpha=1}^{N} (r_{i\alpha}^k - \lambda k_\alpha) \tag{7.20}$$

where $\{r_{i\alpha}^k\}$ are suitably defined for each commodity α as on p. 18. The form (7.20) is separable and this allows us to treat each commodity separately to give a value function set $\{v_\alpha^\lambda\}$, $1 \leqslant \alpha \leqslant N$, as functions of λ. These may be used to find an actual decision rule which meets the constraints.

For example we may set

$$v^\lambda(i) = \sum_{\alpha=1}^{N} v_\alpha^\lambda(i), \quad \forall\, i \in I \tag{7.21}$$

and then find a decision rule $\sigma^\lambda \in \Delta$ by solving

$$Tv^\lambda = T^{\sigma^\lambda} v^\lambda \tag{7.22}$$

for σ^λ where

$$[Tv^\lambda](i) = \underset{\substack{k_\alpha \geqslant i_\alpha, 1 \leqslant \alpha \leqslant N \\ \sum_{\alpha=1}^{N} (k_\alpha - i_\alpha) \leqslant b}}{\text{maximum}} \left[r_i^k(0) + \rho \sum_{j \in I} p_{ij}^k v^\lambda(j) \right]. \tag{7.23}$$

For this problem a complete solution would require that λ be made a function of i and, in such a case, a function $\lambda(.)$ for the continuous

demand case exists which will give a true optimal solution, but it is too difficult to solve the problem this way. However, for any given $\lambda \in R_+$, bounds on the loss of optimality for the policy generated may be computed.

An alternative approximation method which does not reduce the number of states, but which reduces computations, is to use expected state transitions (see Norman and White [36] and White [62]). In this case for each $i \in I$ and $k \in K(i)$ the new transition matrix is $[\tilde{p}_{ij}^k]$ where

$$\tilde{p}_{ij}^k = 1 \quad \text{if } j = \mu^k(i) = \sum_{s \in I} sp_{is}^k, \tag{7.24}$$

$$= 0 \quad \text{otherwise.} \tag{7.25}$$

In (7.24) it is possible that $\mu^k(i) \notin I$, and a nearest point j to $\mu^k(i)$ in I may be taken.

Then in equation (7.17), replacing $\tilde{u}$: $\tilde{I} \to R$ by u: $I \to R$, $\tilde{T}\tilde{u}$ is replaced by $\hat{T}u$ where

$$[\hat{T}u](i) = \underset{k \in K(i)}{\text{maximum}} \; [r_i^k + \rho u(\mu^k(i))], \quad \forall\, i \in I. \tag{7.26}$$

If σ_i^k is the variance of j given (k, i) and if δ is a decision rule solution for the modified equation (7.17) it may be shown that, with $\pi = (\delta)^\infty$

$$\| v - v^\pi \| \leqslant \gamma \underset{i \in I, k \in K(i)}{\text{maximum}} \; [\sigma_i^k] \tag{7.27}$$

where γ may be calculated. For the definition of $\{\sigma_i^k\}$ an appropriate metric is needed for I, e.g. $\| \quad \|$ in (7.27) when I is a set of vectors.

Bitran and Yanasse [6] use an approach in which the random variables are replaced by their expectations, resulting in a slightly different approach to that just described. Amström [1] also uses an expected state transition approach for partially observable Markov decision processes. The value iteration process (6.7) is replaced by

$$\underline{n \geqslant 1} \qquad \tilde{u}_n = \tilde{M}\tilde{u}_{n-1} \tag{7.28}$$

where for u: $\mathrm{M} \to R$

$$\tilde{M}u = \underset{\delta \in \Delta}{\text{maximum}} \; [\tilde{M}^\delta u], \tag{7.29}$$

$$[\tilde{M}^\delta u](\mu) = \sum_{l \in I} \mu_l r_l^{\delta(\mu)} + \rho u(\mu P^{\delta(\mu)}), \quad \forall\, \mu \in \mathrm{M}, \tag{7.30}$$

$$[P^{\delta(\mu)}]_{ij} = p_{ij}^{\delta(\mu)}, \quad \forall\, \mu \in \mathrm{M}, \quad i, j \in I. \tag{7.31}$$

Amström shows that, with $\tilde{u}_0 = u_0$

$$\underline{n \geqslant 1} \qquad \tilde{u}_n \leqslant u_n. \tag{7.32}$$

Amström [1] also introduces a variant of the representation approach with which we began, by restricting μ to the case $\mu = \mu^i$, $1 \leqslant i \leqslant m$, where

$$\begin{aligned} &\mu_j^i = 1 \quad \text{if} \quad i = j, i, j \in I, \\ &\mu_j^i = 0, \quad \text{if} \quad i \neq j, i, j \in I. \end{aligned} \tag{7.33}$$

Equations (7.28)–(7.30) are replaced by

$$\underline{n \geqslant 1} \qquad \tilde{\tilde{u}}_n = \tilde{\tilde{M}} \tilde{\tilde{u}}_{n-1} \tag{7.34}$$

where, for u: $\{\mu^i\} \to R$

$$\tilde{\tilde{M}} u = \underset{\delta \in \Delta}{\text{maximum}} \, [\tilde{\tilde{M}}^\delta u], \tag{7.35}$$

$$[\tilde{\tilde{M}}^\delta u](\mu^i) = r_i^{\delta(\mu^i)} + \rho [P^{\delta(\mu^i)} u]_i. \tag{7.36}$$

Also

$$\underline{n \geqslant 1} \qquad \sum_{i \in I} \mu_i \tilde{\tilde{u}}_n(\mu^i) \geqslant u_n(\mu), \quad \forall \, \mu \in \mathrm{M}. \tag{7.37}$$

Equations (7.32) and (7.37) may be used for bounding purposes and for action elimination purposes.

Another way of modifying the state set I to make the problem more tractable in terms of a new set $\tilde{I}$ is given by White [68]. In this case $i = (i_1, i_2, ., ., ., i_N)$, where i_t, $t \geqslant 2$, refers to information gained in the $(t-2)$th last time unit. As t increases i_t becomes less significant and it is possible to replace i by $\tilde{i} = (i_1, i_2, ., ., ., i_m)$ with some specified loss in doing so which can be evaluated in terms of m. This is similar to, but not identical with, the scheme given by (7.14)–(7.17).

Finally let us look at the functional approximation technique introduced, in dynamic programming, by Bellman [4] (see, also White [57]). This technique expresses the value function parametrically as a member of a suitable class of functions, and then finds the parameters which give the best fit in a specified sense within this class.

One functional approximation scheme to approximate the solution to (7.1) is as follows. Let $\{v_a\}$, $a \in A$ be a specified collection of

functions from I to R. The approximating function $\tilde{v}$ to v is then

$$\tilde{v} = \sum_{a \in A} \mu_a v_a \tag{7.38}$$

with $\mu_a \in R$, $\forall\, a \in A$. The functional approximation problem might then be as follows:

$$\underset{\mu \in R^{\#A}}{\text{infimum}} \; [\| \tilde{v} - T\tilde{v} \|]. \tag{7.39}$$

If $\#I$ is very large then, to overcome the storage problem, (7.39) is solved approximately by selecting an appropriate subset of I and restricting $\| \quad \|$ to this subset. If μ is further restricted then 'infimum' in (7.39) may be replaced by 'minimum'.

Ben-Ari and Gal [5] use the value iteration scheme (3.26) with (7.38) replaced by

$$\tilde{u}_n(x) = \mu_{n0} + \sum_{\alpha=1}^{N} \mu_{n\alpha} x_\alpha \tag{7.40}$$

for solving a dairy herd policy problem, where x_α is the number of the herd in condition α. Their procedure computes the $\{\mu_{n\alpha}\}$ given $\{\mu_{n-1,\alpha}\}$ in a special best-fit way, which involves selecting a special subset of I and computing $\{\mu_{n\alpha}\}$ in a special averaging procedure somewhat different from solving (7.39) with $\tilde{v}$ replaced by $\tilde{u}_n$ given $\tilde{u}_{n-1}$.

Deuermeyer and Whinston [16] also suggest the use of the value iteration scheme. For any u: $I \to R$ the functional approximation process is represented by an operator F on u, i.e.

$$\tilde{u} = Fu. \tag{7.41}$$

The scheme (7.38) is replaced by

$$\tilde{u} = \sum_{a \in A} \mu_a v_a \tag{7.42}$$

where $\{v_a\}$ are spline functions and $\{\mu_a\}$ are determined by F.

The combined functional approximation value iteration scheme is then

$$\underline{n \geqslant 1} \qquad \tilde{u}_n = H\tilde{u}_{n-1}, \tag{7.43}$$

$$\underline{n = 0} \qquad \tilde{u}_0 = u \tag{7.44}$$

where

$$H = TF \tag{7.45}$$

and u is chosen arbitrarily within the set given on the right-hand side of (7.42).

If the functional approximation operator F is chosen so that for all relevant u: $I \to R$

$$\| Fu - u \| \leqslant \beta \tag{7.46}$$

error bounds are given for

$$\| \tilde{u}_n - v \| \tag{7.47}$$

in terms of β.

Schweitzer and Seidman [43] combine the functional approximation form (7.38) with the superharmonic set S defined in section 7.4, where v is shown to be the unique minimal element in S.

The functional approximation optimisation problem is, for $\lambda > 0$

$$\underset{\mu \in R^{n-1}}{\text{minimum}} \ [\lambda \tilde{v}] \tag{7.48}$$

subject to (7.38) and $\tilde{v} \in S$. The minimum exists in (7.48) and will give, via (7.38), an approximating function for v. The number of constraints generated by the restriction that $\tilde{v} \in S$ may be large. However, (7.48) may be solved by solving its dual.

Schweitzer and Seidman [43] also suggest a policy space quadratic approximation scheme. For any specific policy $\pi = (\sigma^1)^\infty$ the functional approximation optimisation problem is, for $\lambda > 0$

$$\underset{\mu \in R^{n-1}}{\text{minimum}} \ [\lambda \varepsilon^1] \tag{7.49}$$

subject to $\tilde{u}$ taking the form of the right-hand side of (7.38) and where

$$\varepsilon^1(i) = ([T^{\sigma^1}\tilde{u} - \tilde{u}](i))^2, \quad \forall\, i \in I. \tag{7.50}$$

Solving (7.49)–(7.50), (7.38) leads to

$$\tilde{u} = \tilde{u}^1. \tag{7.51}$$

Then σ^1 is replaced by

$$\sigma^2 \in \arg \underset{\delta \in \Delta}{\text{maximum}} \ [T^\delta \tilde{u}^1]. \tag{7.52}$$

The procedure is repeated with σ^2 replacing σ^1, ε^2 replacing ε^1 and further iterations if required.

An alternative form of functional approximation to that of (7.38) and its value iteration analogues arises when it is known that v takes the form

$$v(x) = \underset{s \in C}{\text{maximum}} \ [\alpha_s x] \tag{7.53}$$

where $\alpha_s \in R^N$, $x \in R^N$. This arises in partially observable Markov decision processes and, in this case, (7.53) is the infinite horizon analogue of (6.12) with x instead of μ. A similar situation arises in adaptive Markov decision processes.

The great difficulty is that $\#C$ might be large. Thus, we may seek a functional approximation of the kind

$$\tilde{v}(x) = \underset{s \in \tilde{C}}{\text{maximum}} \ [\tilde{\alpha}_s x] \tag{7.54}$$

where $\#\tilde{C}$ is much smaller than $\#C$. Both for the finite horizon version of (7.1) and for (7.1) schemes of the form (7.54) are developed by White [73], [74], Lovejoy [30] and others.

A general theory of approximation modelling error analysis is given by Whitt [76], [77].

7.3 POST-OPTIMALITY, PARAMETRIC AND SENSITIVITY ANALYSIS

In general the parameters

$$\{\{r_i^k\}, \{p_{ij}^k\}, \rho\} \tag{7.55}$$

will be estimates of some unknown parameters and it will be necessary to carry out some parametric, sensitivity or post-optimality analysis.

Let

$$\{\{\hat{p}_{ij}^k\}, \{\hat{r}_i^k\}, \hat{\rho}\} \tag{7.56}$$

be the estimates and $\hat{\pi} = (\hat{\delta})^\infty$ be an optimal decision rule for these estimates. For any set of parameters $\{\{p_{ij}^k\}, \{r_i^k\}, \rho\}$, for policy $\hat{\pi}$ we may obtain the value function of expected total discounted rewards (see form (2.109)), using parametric suffixes $\{P, r, \rho\}$

$$v_{Pr\rho}^{\hat{\delta}} = (U - \rho P^{\hat{\delta}})^{-1} r^{\hat{\delta}}. \tag{7.57}$$

If $\hat{\delta}$ is to be optimal for $\{P, r, \rho\}$, with U as the identity matrix then

$$(U - \rho P^{\hat{\delta}})^{-1} r^{\hat{\delta}} \geqslant T^{\sigma}_{P r \rho}[(U - \rho P^{\hat{\delta}})^{-1} r^{\hat{\delta}}], \quad \forall\ \sigma \in \Delta. \tag{7.58}$$

This is a necessary and sufficient condition for $\hat{\delta}$ to be optimal for $\{P, r, \rho\}$.

Now

$$(U - \rho P^{\hat{\delta}})^{-1} \geqslant 0. \tag{7.59}$$

Thus a sufficient condition for inequality (7.58) to hold is

$$r^{\hat{\delta}} \geqslant (U - \rho P^{\hat{\delta}}) r^{\sigma} + \rho (U - \rho P^{\hat{\delta}}) P^{\sigma} (U - \rho P^{\hat{\delta}})^{-1} r^{\hat{\delta}}, \quad \forall\ \sigma \in \Delta. \tag{7.60}$$

Inequalities (7.58) and (7.60) will give regions of $\{P, r, \rho\}$ parameters for which $\hat{\pi}$ will be optimal, and provide a basis for post-optimality analysis.

If $\{P, r, \rho\}$ can be parametrised in terms of some parameter $\theta \in \Theta$ then the region of Θ for which $\hat{\delta}$ is optimal may be found. For such problems as inventory problems (e.g. see p. 143 where an adaptive approach is used) there may be an unknown parameter θ, and, for post-optimality purposes, one may wish to know whether a particular policy is optimal within a given range of this parameter.

White and El-Deib [55] discuss the problem of finding solutions to

$$u_{\theta} = T_{\theta} u_{\theta}, \qquad \theta \in \Theta \tag{7.61}$$

where T_{θ} is the transformation operator with possible parameter set Θ, under certain assumptions on T_{θ} which guarantee that, if, for $\theta \in \Theta$, v_{θ} solves (7.61), then v_{θ} is piecewise affine and convex on Θ. For the purposes of post-optimality analysis this property assists the identification of the θ for which particular policies are optimal.

If the range is small then we may replace θ by $\hat{\theta} + \partial\theta$ (or $\{P, r, \rho\}$ by $\{\hat{P} + \partial\hat{P}, \hat{r} + \partial\hat{r}, \hat{\rho} + \partial\hat{\rho}\}$) and one can derive linear inequalities in $\partial\theta$ (or $(\partial\hat{P}, \delta\hat{r}, \partial\hat{\rho})$) for $\hat{\delta}$ to remain optimal. This will provide a basis for sensitivity analysis.

Smallwood [46] shows how the discount factor range may be partitioned according to which policies are optimal. The procedure operates by looking at the roots of polynomials in ρ. White [70] also looks at the way in which optimal solutions behave in terms of the discount factor. Selecting a distinguished value of ρ the paper looks at the solutions of

$$u_{t\rho} = T u_{t\rho} \tag{7.62}$$

for $t \in [0, 1]$. Several algorithms for finding approximately optimal

solutions are given which use approximate solutions of (7.62) for a given t-value to compute approximate solutions to (7.62) with t increased to $t + \eta$ where η is small. Error bounds associated with the procedures are also given. The latter two papers provide a basis for parametric analysis.

In White [65] both uncertainties in the value of a fixed discount factor and variabilities in the discount factor over time (see (2.6)) are considered.

Hopp [22] considers what are referred to as 'rolling horizons' for Markov decision processes in which the parameters $\{\{r_i^k\}, \{p_{ij}^k\}, \rho\}$ are time dependent. For a specified time horizon limit n, if the parameters are known at the beginning of time unit t for the whole of the time interval $\{t, t+1, .., n\}$ then the optimality equation is given by (2.59), and for the given t an optimal decision rule σ_{tn} may be determined. If t is fixed, Hopp shows that there exists a value of n, say $n(t)$, such that $\sigma_{tn(t)}$ is optimal for all time intervals $\{t, t+1, .., n\}$ with $n \geqslant n(t)$, whatever the parameters $\{\{r_i^k(t')\}, \{p_{ij}^k(t')\}, \{\rho(t')\}\}$ for $t' \geqslant n(t)$. Thus even though these parameters for $t' \geqslant n(t)$ are unknown at the beginning of time unit t, $\sigma_{tn(t)}$ will still be an optimal decision rule for time unit t. At time $t+1$, $\sigma_{tn(t+1)}$ is recomputed and the procedure continues in this way providing, at any time unit t, the parameters are known for the interval $\{t, t+1, .., n(t+1)\}$.

Related work may be found in Morton [35] and Chand [11].

7.4 MULTIPLE-OBJECTIVE MARKOV DECISION PROCESSES

For some problems we may have several sorts of reward, or cost, which we may not be able to optimise simultaneously. For example, in the inventory problem of p. 54 we have order costs, holding costs and shortage costs. Sometimes it is not easy to determine, for example, actual holding or shortage costs. We might then treat the problem as one involving three objectives, viz. order quantities, holding quantities and shortage quantities. Then minimising, for example, expected order quantities will not in general lead to minimal expected shortage quantities. The multiple-objective inventory problem is studied in White [60].

Let us turn now to the general multiple-objective problem and replace the single objective rewards $\{r_i^k\}$ by $\{r_i^{lk}\}$ with $1 \leqslant l \leqslant L$ where l relates to the lth objective. Let

$$\Lambda_+^L = \{\lambda \in R^L : \lambda > 0\}. \tag{7.63}$$

For $\lambda \in \Lambda_+^L$ set

$$r_i^k(\lambda) = \sum_{l=1}^{L} \lambda_l r_i^{lk}. \tag{7.64}$$

We may now replace equation (7.1) by

$$u = T_\lambda u \tag{7.65}$$

where

$$[T_\lambda u](i) = \underset{k \in K(i)}{\text{maximum}} \left[r_i^k(\lambda) + \rho \sum_{j \in I} p_{ij}^k u(j) \right], \quad \forall\, i \in I. \tag{7.66}$$

We may now show, using a similar analysis to that of Result 4.1, that if for each $i \in I$ we consider the value functions of the L expected total discounted rewards for a given policy, and define, for each $i \in I$, a maximal element analogously to a minimal element given in p. 98, but for the multiple-objective form, and if δ_λ is any decision rule solution to (7.65), the policy $\pi_\lambda = (\delta_\lambda)^\infty$ gives a maximal element within the set of all such functions generated by policies in Π. A maximal element is defined relative to the specified $i \in I$. If $\lambda > 0$ then π_λ is a maximal element for all $i \in I$.

The λ-procedure will not generate, via equation (7.65), the set of all maximal elements even within $\Pi_D \subseteq \Pi$ in general, i.e. for any given state $i \in I$ there may be policies in Π_D which lead to maximal elements, relative to Π_D, which will not be found by using the λ-procedure. The reason is that they may be dominated by some randomised combination of other policies in Π_D.

If we allow randomised non-stationary Markov actions, i.e. the set Π_M of policies, then the λ-approach will lead to the set of maximal elements relative to Π, but not via equation (7.65) which allows no randomised actions and leads to stationary policies. Returning to definition (2.65) we define

$$v^l(i) = \underset{\pi \in \Pi}{\text{supremum}} \, [v^{l\pi}(i)], \quad \forall\, i \in I, \quad 1 \leqslant l \leqslant L \tag{7.67}$$

where $v^{l\pi}$ is defined analogously to definition (2.12), $1 \leqslant l \leqslant L$ where, in definition (2.12), $v_n^{l\pi}$ is defined analogously to v_n^π on p. 26.

We may now use Result 2.1 in a similar manner to the way we use it for Result 2.2, together with an extension of multiple-objective linear programming to multiple-objective infinite linear programming (because we have an infinite number of equations and variables in equations (4.32)) to show that the set of policies within Π, which lead for a specific i to a maximal element relative to $\prod$, is equivalent to the set of policies within Π_M which lead to a maximal element relative to Π_M (in the sense that they give the same set of maximal elements in their respective sets) and that the policy subset of Π_M which gives these maximal elements is equal to, for a given $i \in I$

$$\pi \in \Pi_M\colon\ v_\lambda^\pi(i) \geqslant v_\lambda^\tau(i), \quad \text{for some } \lambda \in \Lambda_+^L \quad \text{and} \quad \forall\ \tau \in \Pi_M \tag{7.68}$$

where

$$v_\lambda^\pi = \sum_{l=1}^{L} \lambda_l v^{l\pi}. \tag{7.69}$$

Finding all the maximisers over Π_M (which involves $\{x_l^k(t)\}$) of $v_\lambda^\tau(i)$ is not easy. Finding some of the maximisers by finding maximisers over Π_D using (7.65) and Result 2.6 extended to $\{v_\lambda\}$ is easier, but it will not give the whole set of maximal elements in general.

There is an easier way to find equivalent maximal element sets relative to Π. This comes from a result of Hartley [20]. This says that the equivalent set of policies which lead to the maximal element set relative to Π may be obtained by restriction to Π_D^*, where Π_D^* is the set of randomised policies of Π_D. Thus if

$$\Pi_D = \{\pi^s\}, \qquad 1 \leqslant s \leqslant \bar{s} \tag{7.70}$$

then

$$\Pi_D^* = \left\{(\alpha_1, \pi^1, \alpha_2, \pi^2, ., ., \alpha_s, \pi^s, ., ., \alpha_{\bar{s}}\pi^{\bar{s}})\colon \sum_{s=1}^{\bar{s}} \alpha_s = 1, \alpha \geqslant 0\right\} \tag{7.71}$$

where the part in parentheses means: select pure policy π^s with probability α_s, $1 \leqslant s \leqslant \bar{s}$, initially and continue to use this policy indefinitely. In this case using multiple-objective linear programming the equivalent policy subset of Π which generates, for any specific $i \in I$, the maximal element set relative to Π is

$$\pi \in \Pi_D^*\colon\ v_\lambda^\pi(i) \geqslant v_\lambda^\tau(i) \quad \text{for some } \lambda \in \Lambda_+^L \quad \text{and} \quad \forall\ \tau \in \Pi_D^*. \tag{7.72}$$

The set given by (7.69) and (7.72) are in specific state inequality form. This reflects the fact that in general a policy which leads to a maximal

element for one state does not necessarily lead to a maximal element for other states. However (see White [58], Exercise 8, Chapter 8), for unichain no-transient-state situations, any π policy satisfying inequality (7.72) for some $i \in I$ and some $\lambda \in \Lambda_+^L$ will also satisfy inequality (7.72) for the same λ and for all $i \in I$. Thus the set given by inequality (7.72) can be written in function $\{v^{l\pi}\}$ form in this case, i.e.

$$\pi \in \Pi_D^*: \ \lambda v^{\pi} \geqslant \lambda v^{\tau} \quad \text{for some } \lambda \in \Lambda_+^L \quad \text{and} \quad \forall\, \tau \in \Pi_D^*. \tag{7.73}$$

In (7.73)

$$v^{\pi} = (v^{1\pi}, v^{2\pi}, ., ., ., v^{L\pi}) \in R^{L \times m} \tag{7.74}$$

and $\geqslant$ is the vector order over R^m (see (4.3)). Finding the policy set specified by inequality (7.72) is equivalent to solving the following problem with $\{v^{l\pi}\}$ written as $\{v^{ls}\}$:

$$\underset{\alpha}{\text{maximise}} \left[\sum_{s=1}^{\bar{s}} \alpha_s \left(\sum_{l=1}^{L} \lambda_l v^{ls}(i) \right) \right] \tag{7.75}$$

subject to

$$\sum_{s=1}^{\bar{s}} \alpha_s = 1, \tag{7.76}$$

$$\alpha_s \geqslant 0, \qquad 1 \leqslant s \leqslant \bar{s}. \tag{7.77}$$

Among these feasible policy solutions will be the policy solutions to equation (7.65).

For a given λ the problem (7.75)–(7.77) may be solved by the column generation method of Dantzig and Wolfe [12] bearing in mind the fact that $\{\pi^s\}$ are not known explicitly and have to be generated as required until an optimal solution is found. Some form of parametric programming as λ is varied is also possible.

An alternative approach to finding all of Π_D which lead to maximal elements relative to Π_D, is to use an extended maximal element form of equation (7.1). In effect equation (7.1) is replaced by (for a specific $i \in I$)

$$E = \mathrm{E}E \tag{7.78}$$

where E is the maximal element value function set and E is a maximal element operator replacing the usual scalar operator T. This is detailed in White [59].

The average expected reward per unit time case may be solved in a similar manner. Equation (7.67) is replaced by

$$g^l(i) = \underset{\pi \in \Pi}{\text{supremum}} \; [g^{l\pi}(i)], \quad \forall\, i \in I, \quad 1 \leqslant l \leqslant L. \tag{7.79}$$

Equation (7.69) is replaced by (with $\lambda \in \Lambda_+^L$)

$$g_\lambda^\pi = \sum_{l=1}^{L} \lambda_l g^{l\pi} \tag{7.80}$$

with $g^{l\pi} \in R^m$, $1 \leqslant l \leqslant L$, $\pi \in \Pi$.

Restricting ourselves to the uni-chain case then maximising $[g_\lambda(i)]$ for any i over Π_D will give a maximal element, in terms of $g = (g^1, g^2, ., ., ., g^L)$, for all i. To obtain all maximal elements the maximisation must be carried out over Π_M or alternatively over Π_D^*. Equation (7.75) is replaced by

$$\underset{\alpha}{\text{maximise}} \left[\sum_{s=1}^{\bar{s}} \alpha_s \left(\sum_{l=1}^{L} \lambda_l g^{ls} \right) \right] \tag{7.81}$$

where g^{ls} is the lth objective function gain for policy π^s.

7.5 UTILITY, PROBABILISTIC CONSTRAINTS AND MEAN–VARIANCE CRITERIA

The discounted random reward for a given policy π (as an extension of (2.6)) for the infinite horizon stationary problem is

$$R^\pi = \sum_{t=1}^{\infty} \rho^{t-1} Y_t^\pi. \tag{7.82}$$

As a housekeeping requirement we will assume that π is measurable so that R^π is a proper random variable.

So far we have concentrated on the expected value of R^π. Such a criterion may be quite inadequate to describe a decision-maker's preferences. In general we would wish to consider some utility function of $\{Y_1^\pi, Y_2^\pi, ., Y_t^\pi, ., ...\}$. One possible utility function may be simply a function W of R^π so that our problem would be, instead of the form (2.65)

$$v(i) = \underset{\pi \in \Pi}{\text{supremum}} \; [v^\pi(i)] \tag{7.83}$$

where

$$v^{\pi}(i) = E(W(R^{\pi}) \mid X_1 = i). \tag{7.84}$$

In general form (7.84) does not enable us to set up an equation similar to equation (7.1) because at any point in time the appropriate action to take may depend on the contribution to R^{π} already achieved at the time of that action. We therefore add in an extra state variable, viz. r: the cumulative total discounted reward to date. With a new state definition (i, r) the analogous equation to equation (7.1) is

$$\underline{t \geqslant 1} \qquad u_t(i,r) = \underset{k \in K(i)}{\text{maximum}} \left[\sum_{j \in I} p_{ij}^k u_{t+1}(j, r + \rho^{t-1} r_{ij}^k) \right], \qquad \forall\ (i,r) \in I \times R \tag{7.85}$$

where

$$\lim_{t \to \infty} [\| u_t - W \|] = 0. \tag{7.86}$$

The function W is defined on R, but can also be considered to be a function on $I \times R$ whose value is independent of $i \in I$.

Some housekeeping axioms are needed and this result uses results of Fainberg [17], [18] and of Kreps [26], both of which are used in White [63]. Then $v(i)$ is given by $u_1(i, 0)$.

An alternative formulation is to find (in maximisation form)

$$\underset{\pi \in \Pi}{\text{supremum}} \ [\text{probability}(R^{\pi} \geqslant r_0)] \tag{7.87}$$

with some prespecified value of r_0. These ideas are introduced by Sobel [48] and discussed in Bouakis [9] and White [75] using minimisation forms. With suitable housekeeping requirements they lead to the equation

$$u(i,r) = \underset{k \in K(i)}{\text{maximum}} \left[\sum_{j \in I} p_{ij}^k u(j, (r - r_{ij}^k)/\rho) \right], \quad \forall\ (i,r) \in I \times R \tag{7.88}$$

with

$$v(i,r) = \underset{\pi \in \Pi}{\text{supremum}}$$

[probability that the total discounted reward over an infinite horizon, beginning in state i, is at least r] (7.89)

as the unique solution to (7.88)

Alternatively we may wish to find, using definition (2.13)

$$v(i) = \underset{\pi \in \Pi}{\text{supremum}} \; [v^{\pi}(i)] \tag{7.90}$$

subject to

$$\text{probability}(R^{\pi} \geqslant r_0 \mid X_1 = i) \geqslant \gamma \tag{7.91}$$

where $\{r_0, \gamma\}$ are prespecified in (7.91).

This may be put in infinite linear programming form and optimal policies in Π_{MD} (see p. 27) found which will guarantee optimality over Π. We may equally well work within Π_D^* (see p. 161) and develop a corresponding formulation to that of problem (7.75)–(7.77) in α, the solution of which is facilitated by the column generation method of Dantzig and Wolfe [12]. In the above, Π and all the related policy sets are defined with respect to the state set $I \times R$.

It is possible to formulate problems in terms of mean-variance analysis.

In White [63] it is shown that we may, within ε-approximations, restrict ourselves to Π_M or to Π_{MD}^* (defined analogously to Π_D^*), and again defined with respect to the state set $I \times R$. In this case the additional variable r is the level of the cumulative total discounted reward to date at the current time unit t, measured as from time unit $t = 1$ and discounted to the beginning of time unit $t = 1$.

If $\{\pi^s\}$, $1 \leqslant s \leqslant \infty$, are the policies in Π_{MD}^* and if, beginning in state $(i, 0)$, we wish to minimise the variance of R^{π} subject to its expectation being at least equal to β, and if $\{v_t^s, V_t^s\}$ are the means and variances of $\{R^{\pi^s}\}$ measured from, and discounted to, the beginning of time unit $t = 1$, given the state (i, r) at the beginning of time unit t then our problem may be cast as follows, in maximisation form:

$$\underset{\alpha}{\text{maximise}} \left[\sum_{s=1}^{\infty} - \alpha_s (V_1^s(i, 0) + v_1^s(i, 0)^2) + \left(\sum_{s=1}^{\infty} \alpha_s v_1^s(i, 0) \right)^2 \right] \tag{7.92}$$

subject to

$$\sum_{s=1}^{\infty} \alpha_s v_1^s(i, 0) \geqslant \beta, \tag{7.93}$$

$$\sum_{s=1}^{\infty} \alpha_s = 1, \tag{7.94}$$

$$\alpha_s \geqslant 0, \qquad 1 \leqslant s < \infty. \tag{7.95}$$

If the inequality in (7.93) is required to be an equality the problem reduces to

$$\text{maximise}\left[-\sum_{s=1}^{\infty}\alpha_s M_1^{2s}(i,0)\right] \tag{7.96}$$

subject to

$$\sum_{s=1}^{\infty}\alpha_s v_1^s(i,0)=\beta, \tag{7.97}$$

$$\sum_{s=1}^{\infty}\alpha_s = 1, \tag{7.98}$$

$$\alpha_s \geqslant 0, \qquad 1 \leqslant s < \infty \tag{7.99}$$

where $M_t^{2s}(i,r)$ is the second moment of $R^{\pi'}$ about the origin for policy $\pi = \pi^s$, measured from the beginning of the time unit $t=1$, given the state (i,r) at the beginning time unit t. Following the lines of Sobel [48], and discussed in White [63], it is possible to establish the following recurrence relations for $\{M_t^{2s}\}$, $\{v_t^s\}$ from which $\{V_t^s\}$ are also derivable:

$$\underline{t \geqslant 1} \qquad v_t^s(i,r)=\sum_{j\in I} p_{ij}^{st} v_{t+1}^s(j, r+\rho^{t-1}r_{ij}^{st}),$$

$$\forall\ (i,r)\in I\times R, \quad 1\leqslant s<\infty, \tag{7.100}$$

$$M_t^{2s}(i,r)=\sum_{j\in I} p_{ij}^{st} M_{t+1}^{2s}(j, r+\rho^{t-1}r_{ij}^{st}),$$

$$\forall\ (i,r)\in I\times R, \quad 1\leqslant s<\infty \tag{7.101}$$

where

$$p_{ij}^{st}=p_{ij}^{\delta_t^s}, r_{ij}^{st}=r_{ij}^{\delta_t^s}, \quad \forall\ (i,j)\in I\times R, \quad 1\leqslant s<\infty, \tag{7.102}$$

$$\pi^s=(\delta_1^s, \delta_2^s, ., ., \delta_t^s, ., ., .), \tag{7.103}$$

$$\lim_{t\to\infty}\ [v_t^s(i,r)] = r, \quad \forall\ (i,r)\in I\times R, \quad 1\leqslant s<\infty, \tag{7.104}$$

$$\lim_{t\to\infty}\ [M_t^{2s}(i,r)] = r^2, \quad \forall\ (i,r)\in I\times R, \quad 1\leqslant s<\infty. \tag{7.105}$$

These results are in White [63], Corollary 4.2. The problem given in (7.92)–(7.95) is a semi-infinite linear programme which may be solved using the column generation method of Dantzig and Wolfe [12], so

that $\{v_1^s, M_1^{2s}\}$ need only be calculated as they are required. At most two policies in the set $\{\pi^s\}$ will feature in an optimal vertex solution.

Filar et al. [19] discuss a slightly different mean-variance analysis where the variance is the variance in any time unit about the average expected reward per unit time in the long run, and under the assumption that $r_{ij}^k = r_i^k$, $\forall\, i, j \in I$, $k \in K(i)$. Using the linear programming formulation (4.64)–(4.68) the problem takes the maximisation form, assuming a uni-chain situation and with $\lambda \in R_+$

$$\text{maximise}\left[\sum_{i\in I, k\in K(i)} x_i^k r_i^k - \lambda\left(\sum_{i\in I, k\in K(i)} x_i^k\left(r_i^k - \sum_{j\in I, l\in K(j)} x_j^l r_j^l\right)^2\right)\right] \tag{7.106}$$

subject to (4.65)–(4.68). White [71] provides computational procedures for solving this problem.

A survey of related formulations may be found in White [67].

7.6 MARKOV GAMES

In the general framework in Chapter 2, on which subsequent material is based, it is assumed that there is a single decision-maker with a specified objective to be optimised.

In this section we briefly outline the extension of these ideas to the case when there are two decision-makers taking simultaneous actions at each of a sequence of time units. Much of the framework of Chapter 2 applies. If Π^1, Π^2 are the policy spaces of players 1 and 2 respectively, the analogues of $v_n^\pi(i)$, $\pi \in \Pi$ in (xiv) of Chapter 2 are, restricting ourselves to the stationary case, $v_n^{\pi\tau q}(i)$: the expected total discounted reward over the next n time units for player q if $X_1 = i$, and if players 1 and 2 play policies π, τ respectively, with $\pi \in \Pi^1$, $\tau \in \Pi^2$, $q = 1, 2$.

The following is based on van der Wal [53]. The objective is to obtain analogous results to those of standard two-person game theory. In order to obtain these we may, without loss, restrict ourselves to $\{\Pi_M^1, \Pi_M^2\}$. As in standard game theory, we must allow randomised actions, and cannot necessarily restrict ourselves to $\{\Pi_{MD}^1, \Pi_{MD}^2\}$ the Markov deterministic policies.

Let $K^1(i), K^2(i)$ be the feasible action spaces for players 1 and 2 respectively, and $K^{1*}(i), K^{2*}(i)$ be the corresponding randomised action spaces for $i \in I$.

For zero-sum games we have

$$\underline{n \geqslant 1} \qquad v_n^{\pi\tau 1} = -v_n^{\pi\tau 2}. \tag{7.107}$$

As with standard game theory let us suppose that we seek $\underset{\pi \in \Pi_M^1}{\text{supremum}}$ $\underset{\tau \in \Pi_M^2}{\text{infimum}}$ solutions. Define, analogously to (2.8)

$$v_n^1(i) = \underset{\pi \in \Pi_M^1}{\text{supremum}} \ \underset{\tau \in \Pi_M^2}{\text{infimum}} \ [v_n^{\pi\tau 1}(i)], \quad \forall\, i \in I, \tag{7.108}$$

$$v_n^2(i) = \underset{\tau \in \Pi_M^2}{\text{supremum}} \ \underset{\pi \in \Pi_M^1}{\text{infimum}} \ [v_n^{\pi\tau 2}(i)], \quad \forall\, i \in I. \tag{7.109}$$

Then

$$v_n^1(i) = -v_n^2(i), \quad \forall\, i \in I. \tag{7.110}$$

The analogue of Result 2.3 is that v_n^1, v_n^2 are, respectively, unique solutions of the equations

$$\underline{n \geqslant 1} \qquad u_n^1 = T^1 u_{n-1}^1, \tag{7.111}$$

$$u_n^2 = T^2 u_{n-1}^2 \tag{7.112}$$

where, for $u\colon I \to R$

$$[T^1 u](i) = \underset{k \in K^{1*}(i)}{\text{maximum}} \ \underset{l \in K^{2*}(i)}{\text{minimum}} \left[r_i^{kl} + \rho \sum_{j \in I} p_{ij}^{kl} u(j) \right], \qquad \forall\, i \in I, \tag{7.113}$$

$$[T^2 u](i) = \underset{l \in K^{2*}(i)}{\text{maximum}} \ \underset{k \in K^{1*}(i)}{\text{minimum}} \left[-r_i^{kl} + \rho \sum_{j \in I} p_{ij}^{kl} u(j) \right]. \qquad \forall\, i \in I. \tag{7.114}$$

The quantity r_i^{kl} is the immediate expected return to player 1 if the state is $i \in I$ and actions $k \in K^{1*}(i)$, $l \in K^{2*}(i)$ are taken, and p_{ij}^{kl} is the probability that if we are in state $i \in I$ and actions $k \in K^{1*}(i)$, $l \in K^{2*}(i)$ are taken we move to a new state $j \in I$.

We may use (7.111) to find policies as follows:

There exists an optimal policy π_n for player 1, i.e. for which

$$v_n^{\pi_n \tau 1} \geqslant v_n^1, \qquad \forall\, \tau \in \Pi^2. \tag{7.115}$$

For each $\varepsilon > 0$, using (7.111), there exists an ε-optimal strategy τ_n for player 2, i.e. for which

$$v_n^{\pi \tau_n 2} \geqslant v_n^2 - \varepsilon e, \quad \forall\, \pi \in \Pi^1. \tag{7.116}$$

In (7.115), (7.116) equation (7.111) is used to find $\{\pi_n, \tau_n\}$. Note that (7.116) is equivalent to

$$v_n^{\pi\tau_n 1} \leqslant v_n^1 + \varepsilon e, \quad \forall\ \pi \in \Pi^1 \tag{7.117}$$

and that (7.115) and (7.117) combined imply that

$$v_n^1 \leqslant v^{\pi_n \tau_n 1} \leqslant v_n^1 + \varepsilon e. \tag{7.118}$$

For the infinite horizon problem the analogues of (2.12) are

$$v^{\pi\tau q}(i) = \lim_{n \to \infty} [v_n^{\pi\tau q}(i)], \quad \forall\ i \in I, \quad \pi \in \Pi^1, \quad \tau \in \Pi^2, \quad q = 1, 2. \tag{7.119}$$

The analogues of (2.13) are then

$$v^1(i) = \operatorname*{supremum}_{\pi \in \Pi^1} \operatorname*{infimum}_{\tau \in \Pi^2} [v^{\pi\tau 1}(i)], \quad \forall\ i \in I, \tag{7.120}$$

$$v^2(i) = \operatorname*{supremum}_{\tau \in \Pi^2} \operatorname*{infimum}_{\pi \in \Pi^1} [v^{\pi\tau 2}(i)], \quad \forall\ i \in I. \tag{7.121}$$

We have

$$v^1 = -v^2. \tag{7.122}$$

The analogues of (2.66) are

$$u = T^1 u, \tag{7.123}$$

$$u = T^2 u \tag{7.124}$$

and v^1, v^2 are, respectively, unique solutions of (7.123) and (7.124). The schemes (7.111) or (7.112) may be used to solve (7.123) or (7.124) to a requisite degree of approximation.

The policy solutions for the finite horizon case and for the infinite horizon case are obtained by solving (7.111) or (7.112), (7.123) or (7.124) respectively. In the latter case $\pi \in \Pi_{MS}^1$, $\tau \in \Pi_{MS}^2$ where Π_{MS}^q is the set of Markov stationary policies, $q = 1, 2$.

To solve (7.108), for example, $u_n(i)$ can be computed by linear programming procedures for each $i \in I$ once u_{n-1} has been determined, after noting that in (7.113), for a given $k \in K^{1*}(i)$, the minimum over $K^{2*}(i)$ is attainable in $K^2(i)$.

The linear programme takes the form, $\forall\, i \in I$

$$\underline{n \geqslant 1} \qquad u_n(i) = \text{maximum}\,[z] \tag{7.125}$$

subject to

$$z \leqslant \sum_{k \in K(i)} r_i^{kl} x_{ik} + \rho \sum_{j \in I, k \in K(i)} p_{ij}^{kl} u_{n-1}(j) x_{ik}, \quad \forall\, l \in K^2(i), \tag{7.126}$$

$$\sum_{k \in K(i)} x_{ik} = 1, \tag{7.127}$$

$$x_{ik} \geqslant 0, \quad \forall\, k \in K(i) \tag{7.128}$$

where x_{ik} is the probability of taking action $k \in K(i)$ given state $i \in I$.

The solution for any specific $n-1$ may be used to facilitate the derivation of a solution for n, particularly as the sequence $\{u_n\}$ converges. A solution to, for example, (7.123) is not in general possible via linear programming. Kallenberg [24] shows, however, how a linear programming approach is possible if $\{p_{ij}^{kl}\}$ are independent of l or independent of k.

Further material on Markov games is to be found in van der Wal [53] and in Kallenberg [24].

CHAPTER 8

Some Markov decision process problems, formulations and optimality equations

8.1 SOME ILLUSTRATIONS

8.1.1 ILLUSTRATION 1: OVERHAUL AND REPLACEMENT

Problem statement

An airline classifies the condition of its planes into three categories, viz. excellent, good and poor. The annual running costs for each category are $\$0.25 \times 10^6$, $\$10^6$ and $\$2 \times 10^6$ respectively. At the beginning of each year the airline has to decide whether or not to overhaul each plane individually. With no overhaul a plane in excellent condition has probabilities of 0.75 and 0.25 of its condition being excellent or good, respectively, at the beginning of the next year. A plane in good condition has probabilities of 0.67 and 0.33 of its condition being good or poor, respectively, at the beginning of the next year. A plane in poor condition will remain in a poor condition at the beginning of the next year. An overhaul costs $\$2 \times 10^6$ and takes no significant time to do. It restores a plane in any condition to an excellent condition with probability 0.75, and leaves it in its current condition with probability 0.25. The airline also has an option of scrapping a plane and replacing it with a new one at a cost of $\$5 \times 10^6$. Such a new plane will be in excellent condition initially. There is an annual discount factor of $\rho = 0.5$.

We derive the optimality equation for the problem of maximising the negative of the infinite horizon expected total discounted cost with, as

in the text, all costs incurred in a given year assigned to the beginning of that year.

Formulation

States	$i = 1$:	excellent condition;
	$i = 2$:	good condition;
	$i = 3$:	poor condition.
Actions	$k = 1$:	do nothing;
	$k = 2$:	overhaul;
	$k = 3$:	replace.

Optimality equation (see (2.66))

$$u(1) = \text{maximum} \begin{bmatrix} k=1: & -0.25 \times 10^6 + 0.5(0.75u(1) + 0.25u(2)) \\ k=2: & -2 \times 10^6 + u(1) \\ k=3: & -5 \times 10^6 + u(1) \end{bmatrix}^*,$$

$$u(2) = \text{maximum} \begin{bmatrix} k=1: & -10^6 + 0.5(0.67u(2) + 0.33u(3)) \\ k=2: & -2 \times 10^6 + 0.75u(1) + 0.25u(2) \\ k=3: & -5 \times 10^6 + u(1) \end{bmatrix}^*,$$

$$u(3) = \text{maximum} \begin{bmatrix} k=1: & -2 \times 10^6 + 0.5u(3) \\ k=2: & -2 \times 10^6 + 0.75u(1) + 0.25u(3) \\ k=3: & -5 \times 10^6 + u(1) \end{bmatrix}^*.$$

We have deviated slightly from equation (2.66). In the places marked by * the expressions reflect the immediate conditions after overhaul or purchase, as the case may be, and not the conditions at the beginning of the next year. Thus, no discount factor appears on the right-hand side of the * components.

8.1.2 ILLUSTRATION 2: CROSSING A ROAD

Problem statement

A man is trying to cross a road. Cars pass in such a manner that the time t in seconds between successive cars, combining both directions into one car stream for this purpose, is independently and identically distributed according to a uniform distribution over the interval $0 \leqslant t \leqslant 10$. He can see only one car at a time. He wants to cross the

road within the next 16 seconds, and before two more cars have passed him. He takes, effectively, no time to cross the road. He wishes to maximise the expected length of time he has to cross, between commencing to cross and a car passing him. We derive the optimality equation for this problem.

Formulation

States	t:	the time that the next car will take to pass him, $0 \leqslant t \leqslant 10$;
	s:	the time he has left to cross, $0 \leqslant s \leqslant 16$.
Decision epochs	$n = 1$:	the first car has just passed;
	$n = 2$:	no car has passed.
Actions	$k = 1$:	cross now;
	$k = 2$:	wait for next car to pass.

Optimality equation (see (2.58))

$n = 1$ $k = 1$ because he must cross now.

$$u_1(t, s) = t, \quad \forall\ 0 \leqslant t \leqslant 10, \quad 0 \leqslant s \leqslant 16.$$

Define

$$u_1(t, s) = -\infty, \quad \text{if } s < 0.$$

$n = 2$

$$u_2(t, s) = \text{maximum} \begin{bmatrix} k = 1\text{: } t \\ k = 2\text{: } \int_0^{10} 0.1 u_1(y, s - t)\, \mathrm{d}y \end{bmatrix}, \quad \begin{matrix} 0 \leqslant t \leqslant 10, \\ 0 \leqslant s \leqslant 16. \end{matrix}$$

8.1.3 ILLUSTRATION 3: OYSTER FARMING (Kushner [27])

Problem statement

A pearl-containing oyster is either diseased or not diseased. At each of a succession of unit time intervals it may be checked to see if it is diseased or not diseased at a cost \$$c$. If the oyster is diseased it has value 0 and is disposed of. If it is not diseased it has a value $g(t)$, where t is the age of the oyster. It costs \$$a$ per unit time to maintain an oyster. If the oyster is not diseased the probability that it will remain in a

non-diseased state for the next unit time interval is p. We derive the optimality equation for maximising the net expected revenue from an oyster.

For the first decision it is known whether or not the oyster is diseased.

Formulation

States t: age of oyster when checked, $t \geqslant 0$;
$s = 1$: if the oyster is not diseased;
$s = 2$: if the oyster is diseased.
Absorbing states: $\{(t, 2)\}$, $\forall\ t \geqslant 0$.

Actions $k = 1$: sell;
$k = (2, \gamma)$: do not sell, and inspect after γ time units later.

Optimality equation (see (2.128), (5.10)–(5.14))

$$u(t, 1) = \text{maximum} \times \begin{bmatrix} k = 1: & g(t) \\ k = (2, \gamma): & \max\limits_{y \geqslant 1} \left[-ya - c + p^{\gamma} u(t + y, 1) + (1 - p^{\gamma}) u(t + y, 2) \right] \end{bmatrix}, \quad \forall\ t \geqslant 0,$$

$$u(t, 2) = 0, \quad \forall\ t \geqslant 0.$$

8.1.4 ILLUSTRATION 4: BURGLING (Ross [40])

Problem statement

At the beginning of each night a burgler has to decide whether to retire from burgling or, if not, how many burglaries to attempt in that night. Each successful burglary has a value of \$100. He can do a maximum of two burglaries in one night. Each burglary has a probability of 0.75 of being successful. The first time he is caught he loses all his gains for that night and is put on probation indefinitely. If he is caught a second time, at any time, he goes to jail and retires from his profession. He values this circumstance as equivalent to a loss of \$500.

We derive the optimality equation for maximising his expected net profit.

Formulation

States $i = 1$: he has not been caught to date;
$i = 2$: he has been caught once;
$i = 3$: he has been jailed and has retired.
State 3 is an absorbing state.

Actions $k = 1$: retire;
$k = 2$: plan to carry out one burglary in the given night;
$k = 3$: plan to carry out two burglaries in a given night.

Optimality equation (see equations (2.127) and (2.128))

$$u(1) = \text{maximum}\begin{bmatrix} k = 1\colon 0 \\ k = 2\colon 75 + 0.75u(1) + 0.25u(2) \\ k = 3\colon 112.5 + 0.5875u(1) + 0.4125u(2) \end{bmatrix},$$

$$u(2) = \text{maximum}\begin{bmatrix} k = 1\colon 0 \\ k = 2\colon -50 + 0.75u(2) + 0.25u(3) \\ k = 3\colon -106.25 + 0.5875u(2) + 0.4125u(3) \end{bmatrix},$$

$$u(3) = 0.$$

8.1.5 ILLUSTRATION 5: SEARCH

Problem statement

A set of locations $I = \{1, 2, ., .m\}$ are, in this order, equally spaced in a line. A target is in one of these locations and always remains there. The prior probability of it being in location i is μ_i, $\forall\, i \in I$. It is required to find a policy of search to minimise the expected distance travelled in locating the target where, at any time, any location may be searched at zero cost, and the target will be found if it is in that location.

We derive the appropriate optimality equation which we cast in maximisation form.

Formulation

States $\mu \in R^m$, $\mu \geqslant 0$, $\sum_{j \in I} \mu_j = 1$;
μ_j is the current probability of the target being in location j, $\forall\, j \in I$;

i: the location just searched.
Actions k: the next location to be searched.
(i, μ) is an absorbing state if $\mu_j = 1$ for some $j \in I$.

Optimality equation (see equations (2.127) and (2.128))

$\underline{\mu_j \neq 1, \ \forall \ j \in I}$

$\underline{i \notin \{1, m\}}$

$$u(i,\mu) = \text{maximum}\begin{bmatrix} k = i+1: & -1 + (1-\mu_{i+1})u(i+1, \mu(i+1)) \\ k = i-1: & -1 + (1-\mu_{i-1})u(i-1, \mu(i-1)) \end{bmatrix}, \quad \forall\ i \in I, \quad \mu \in \mathrm{M},$$

where

$$\mu_{i+1}(i+1) = 0, \qquad \mu_j(i+1) = \mu_j/(1-\mu_{i+1}), \quad \forall\ j \neq i+1,$$
$$\mu_{i-1}(i-1) = 0, \qquad \mu_j(i-1) = \mu_j/(1-\mu_{i-1}), \quad \forall\ j \neq i-1,$$

$\underline{i = 1}$ $\qquad u(1,\mu) = -1 + (1-\mu_2)u(2, \mu(2)),$
$\underline{i = m}$ $\qquad u(m,\mu) = -1 + (1-\mu_{m-1})u(m-1, \mu(m-1)),$
$\underline{\mu_j = 1 \text{ for some } j}$ $\qquad u(i,\mu) = 0 \quad \text{if } \mu_j = 1 \quad \text{for some } j, \quad \forall\ i \in I.$

We need only search adjacent locations.

8.1.6 ILLUSTRATION 6: SHUTTLE OPERATIONS (Deb [13])

Problem statement

A shuttle operates between two terminals. In each time unit a passenger arrives just after the beginning of the time unit with probability p at terminal 1. Similarly a passenger arrives with probability q at terminal 2. An operator of a shuttle of infinite capacity decides, at the beginning of each time unit, whether or not to transport the passengers at its current terminal to the other terminal. It takes s units of time to complete a journey. It is required to find a shuttle-dispatching policy to minimise the infinite horizon average expected waiting time per unit time at the terminals for the passengers.

We derive the optimality equation in maximisation form.

Formulation

States x: number of passengers at terminal 1;
y: number of passengers at terminal 2;

$\sigma = 1$: if the shuttle is at terminal 1;
$\sigma = 2$: if the shuttle is at terminal 2.

Actions $k = 1$: transport passengers to the other terminal;
$k = 2$: do nothing

Optimality equation (see (5.29) and (5.30))

$$u(x, y, 1) = \text{maximum}$$

$$\times \left[\begin{array}{l} k = 1: \; -sy - (p+q)s(s+1)/2 - hs + \displaystyle\sum_{\substack{0 \leqslant a \leqslant s \\ 0 \leqslant b \leqslant s}} \frac{s!}{a!(s-a)!} \\ \qquad \times \dfrac{s!}{b!(s-b)!}\, p^a(1-p)^{s-a} q^b (1-q)^{s-b} u(a, y+b, 2) \\ k = 2: \; -(x+y+p+q) - h \\ \qquad + \displaystyle\sum_{\substack{a \in \{0,1\} \\ b \in \{0,1\}}} p^a(1-p)^{1-a} q^b (1-q)^{1-b} u(x+a, y+b, 1) \end{array} \right],$$

$$u(x, y, 2) = \text{maximum}$$

$$\times \left[\begin{array}{l} k = 1: \; -sx - (p+q)s(s+1)/2 - hs + \displaystyle\sum_{\substack{0 \leqslant a \leqslant s \\ 0 \leqslant b \leqslant s}} \frac{s!}{a!(s-a)!} \\ \qquad \times \dfrac{s!}{b!(s-b)!}\, p^a(1-p)^{s-a} q^b (1-q)^{s-b} u(x+a, b, 1) \\ k = 2: \; -(x+y+p+q) - h \\ \qquad + \displaystyle\sum_{\substack{a \in \{0,1\} \\ b \in \{0,1\}}} p^a(1-p)^{1-a} q^b (1-q)^{1-b} u(x+a, y+b, 2) \end{array} \right].$$

8.1.7 ILLUSTRATION 7: CRICKET (Thomas [50])

Problem statement

A well-known cricketer carries a programmed Markov decision process with him at all times when batting. He may receive any one of three types of delivery for the next ball, viz. a bumper (B), a yorker (Y), or

ordinary (O). He also knows that the probabilities of a delivery being in each class depends on his immediately preceding score off the previous ball, where he may have scored 0, 2 or 4 runs. The conditional probabilities of each delivery are given in Table 8.1.

He has only three types of play which he may make to the next delivery, viz. hook (H), drive (D) or forward defensive (F). For each delivery and each play the scores obtained are given in Table 8.2 where T means that his batting innings is terminated. This is assumed to take place only where indicated.

He has to decide which sort of play (H, D, F) to make, before he knows the type of delivery (B, Y, O) which he will receive. He wants to find a policy to maximise his expected total score until his batting terminates.

We will derive the optimality equation.

Formulation

States $i = 1$: 0 runs scored off previous delivery;
$i = 2$: 2 runs scored off previous delivery;

Table 8.1 Cricketing probabilities

Score off previous ball	Next delivery B	Y	O
0	0.00	0.50	0.50
2	0.25	0.25	0.50
4	0.75	0.25	0.00

Table 8.2 Cricketing scores

Play	Delivery B	Y	O
H	4	0(T)	0
D	0(T)	4	2
F	0	2	0

$i = 3$: 4 runs scored off previous delivery;
$i = 4$: out off previous delivery.

State $i = 4$ is an absorbing state.

We assume that we begin the process after one delivery has been made.

Actions $k = 1$: H;
$k = 2$: D;
$k = 3$: F.

Optimality equation (see (2.127) and (2.128))

$$u(1) = \text{maximum}\begin{bmatrix} k=1\colon\ 0 + 0.5u(1) + 0.5u(4) \\ k=2\colon\ 3 + 0.5u(2) + 0.5u(3) \\ k=3\colon\ 1 + 0.5u(1) + 0.5u(2) \end{bmatrix},$$

$$u(2) = \text{maximum}\begin{bmatrix} k=1\colon\ 1 + 0.5u(1) + 0.25u(3) + 0.25u(4) \\ k=2\colon\ 2 + 0.5u(2) + 0.25u(3) + 0.25u(4) \\ k=3\colon\ 0.5 + 0.75u(1) + 0.25u(2) \end{bmatrix},$$

$$u(3) = \text{maximum}\begin{bmatrix} k=1\colon\ 3 + 0.75u(3) + 0.25u(4) \\ k=2\colon\ 1 + 0.25u(3) + 0.75u(4) \\ k=3\colon\ 0.5 + 0.75u(1) + 0.25u(4) \end{bmatrix},$$

$$u(4) = 0.$$

8.1.8 ILLUSTRATION 8: CAPACITY PLANNING (Thomas [51])

Problem statement

A country has to decide how many nuclear power stations to build in each five-year planning time unit. The cost of building a power station is $\$20 \times 10^6$ and it costs $\$5 \times 10^6$ to maintain a condition capable of operation in each five-year time unit. The environmental lobby has enabled a bill to be passed saying that there can be no more than three power stations in total being built and/or capable of operation in any five-year time unit. The output from one power station is required to supply power requirements and, without this, it will cost $\$50 \times 10^6$ in each five-year time unit to provide alternative supplies. It takes five

years to build a power station. If it is capable of operation at the beginning of a five-year time unit it will be capable of operation throughout this time unit, but it will only be capable of operation through the second five-year time unit with probability 0.5. It is required to find a policy to minimise the infinite horizon expected total discounted cost where the discount factor for a five-year time unit is $\rho = 0.5$.

If a power station becomes incapable of operation it remains incapable indefinitely and is not counted in the power station tallies.

Each power station in an operationally fit condition, whether being used or not, will incur the $\$5 \times 10^6$ cost of keeping it capable of operation in each five-year time unit.

We derive the optimality equation in maximisation form.

Formulation

States $i = 1$: no power stations capable of operation;
$i = 2$: 1 power station capable of operation;
$i = 3$: 2 power stations capable of operation;
$i = 4$: 3 power stations capable of operation.

Actions $k \in \{0, 1, 2, 3\}$: the number of power stations to build.

Optimality equation (see (2.66))

$$u(1) = \text{maximum} \begin{bmatrix} k = 0: & -50 + 0.5u(1) \\ k = 1: & -70 + 0.5u(2) \\ k = 2: & -90 + 0.5u(3) \\ k = 3: & -110 + 0.5u(3) \end{bmatrix},$$

$$u(2) = \text{maximum} \begin{bmatrix} k = 0: & -5 + 0.25u(1) + 0.25u(2) \\ k = 1: & -25 + 0.75u(2) + 0.75u(3) \\ k = 2: & -45 + 0.25u(3) + 0.25u(4) \end{bmatrix},$$

$$u(3) = \text{maximum} \begin{bmatrix} k = 0: & -10 + 0.125u(1) + 0.25u(2) + 0.125u(3) \\ k = 1: & -30 + 0.125u(2) + 0.25u(3) + 0.125u(4) \end{bmatrix},$$

$$u(4) = 15 + 0.125u(1) + 0.375u(2) + 0.275u(3) + 0.125u(4).$$

8.2 EXERCISES FOR CHAPTER 8

1. A boxer has to plan his fight programme one year in advance. He can have up to three fights in any one year. If he sustains an injury

in any fight he must wait until next year for a further fight. As soon as he accumulates two injuries in total he must retire from fighting. If he sustains no injuries in a given year he must complete his programme of fights planned for that year. The probability of an injury in any single fight is 1/4 if he has accumulated no injuries prior to this, and 1/2 if he has accumulated one injury prior to this. His earnings for a fight have an expected value of \$5000. If he retires through injury he values this at an equivalent loss of \$20 000. He may, if he wishes, retire at the beginning of any year, in which case he values this at \$0.

With careful explanation derive the optimality equation for maximising his expected net income up to his eventual retirement.

2. In a certain queuing situation we have a single server and two different arrival streams of customers. When customers arrive they arrive at the end of a time unit. When a customer finishes his service this takes place at the end of a time unit. For customers of type 1 there is a probability p_1 that a single customer will arrive in any given time unit and, if a customer of type 1 is being served, there is a probability q_1 that his service will be completed in the current time unit. The corresponding probabilities for customers of type 2 are p_2, q_2, respectively. At the beginning of each time unit it has to be decided which type of customer will be served using a pre-emptive rule, i.e. the current customer being served may be put back into the waiting queue even if his service has not been completed. If any customer arrives at the end of a time unit after any services have been completed, and finds m or more customers in total in the system, he goes elsewhere and does not join the system.

   It is required to find a service policy to minimise the infinite horizon average expected waiting time of customers per unit time. Derive the optimality equation for this problem, using the minimisation form rather than converting to the maximisation form.

3. Consider the following single commodity inventory problem:

   (i) The probability that the demand in any unit time is s is equal to $q(s)$ and the demand is identically and independently distributed in each time unit, with $1 \leqslant s \leqslant \bar{s}$.

   (ii) An order for more stock can only be placed when a shortage is incurred, and then the shortage is immediately satisfied and an additional quantity ordered which is immediately supplied.

(iii) If y is the shortage plus the extra quantity ordered there is a cost $c(y)$ incurred and $y \leqslant L$.
(iv) In addition, for a shortage y there is a cost $l(y)$ incurred.
(v) There is stockholding cost equal to α per unit of stock held for one unit of time, and you may assume that the average stock level between decision epochs is $1/2 \times$ the stock level immediately following the order decision made at any decision epoch.
(vi) The stock level is never allowed to rise above a level L (see (iii)).

It is required to find an ordering policy to minimise the infinite horizon average cost per unit time. Derive the optimality equation for this problem using the minimisation format.

Explain why, if c is a linear function, an optimal policy exists where the total quantity ordered as soon as a shortage occurs is independent of the shortage quantity.

4. Consider the following tax problem (Landsberger and Meilijson [28]). Each year you have to decide whether or not to declare a particular part of your income to the Inland Revenue. If you do declare you pay \$90 in tax. If you do not declare and the Inland Revenue audits your submission you pay \$200. The Inland Revenue splits the taxpayers into three groups, viz. those whom they did not audit in the previous year, those whom they did audit in the previous year and were found to have declared correctly, and those whom they did audit in the previous year and were found to have declared incorrectly. The Inland Revenue adopts a policy of inspecting each of these groups with probabilities 0.5, 0.3 and 0.7, respectively, and these probabilities are known to you. You wish to find a declaration/no declaration policy to minimise your infinite horizon expected total discounted payments to the Inland Revenue with a discount factor $\rho = 0.9$. Derive the optimality equation in minimisation form.
5. Consider the following problem which is an extension of Exercise 4. You can declare the \$90 or not. If you are audited and found to have defaulted you still pay \$200. However, the probability of auditing depends on whether your history leads to you being classified as a reliable or unreliable individual in your tax matters. You do not know how you are classified, but you have your own subjective probabilities. Let p be your probability that you are classified as reliable. The Inland Revenue will audit you with probability 0.3 if they see you as reliable and with probability 0.7 otherwise. Your

subjective probability p of being classified as reliable will change in the following manner, depending on whether or not you submit correctly and whether or not you are audited.

	Audited	*Not audited*
Correct submission	$(1+p)/2$	$(1+4p)/5$
Incorrect submission	$p/2$	$(1+4p)/5$

You wish to minimise your infinite horizon expected total discounted payments when the discount factor is $\rho = 0.9$. Derive the optimality equation in minimisation form.

References

1. K. J. Amström (1965). Optimal control of Markov processes with incomplete state information, *Journal of Mathematical Analysis and Applications*, **10**, 174–205.
2. M. S. Bartlett (1956). *An Introduction to Stochastic Processes*, Cambridge University Press, London.
3. M. J. Beckmann (1968). *Dynamic Programming of Economic Decisions*, Springer-Verlag, Berlin.
4. R. Bellman (1957). *Dynamic Programming*, Princeton University Press, Princeton, New Jersey.
5. Y. Ben-Ari and S. Gal (1986). Optimal replacement policy for multicomponent systems: an application to a dairy herd, *European Journal of Operational Research*, **23**, 213–221.
6. G. R. Bitran and H. H. Yanasse (1984). Deterministic approximations to stochastic production problems, *Operations Research*, **32**, 999–1018.
7. D. Blackwell (1962). Discrete dynamic programming, *Annals of Mathematical Statistics*, **33**, 719–726.
8. T. J. I'A. Bromwich (1926). *An Introduction to the Theory of Infinite Series*, Macmillan, London.
9. M. Bouakis (1986). The target level criterion in Markov decision processes, Francis Marion College, Florence, South Carolina. Short paper based on unpublished PhD thesis.
10. J. A. Buzacott and J. R. Callahan (1973). The pit charging problem in steel production, *Management Science*, **20**, 665–674.
11. S. Chand (1983). Rolling horizon procedures for the facilities in series inventory model with nested schedules, *Management Science*, **29**, 237–249.
12. G. B. Dantzig and P. Wolfe (1961). The decomposition algorithm for linear programming, *Econometrica*, **29**, 767–778.
13. R. K. Deb (1978). Optimal despatching of a finite capacity shuttle, *Management Science*, **24**, 1362–1372.
14. E. V. Denardo (1970). Computing a bias-optimal policy in a discrete time Markov decision problem, *Operations Research*, **18**, 279–289.

15. C. Derman (1970). *Finite State Markov Decision Processes*, Academic Press, New York.
16. B. L. Deuermeyer and A. B. Whinston (1981). On the convergence of polynomial approximation in dynamic programming, Department of Industrial Engineering, University of Texas, and Graduate School of Management, Purdue University. Working Paper.
17. E. A. Fainberg (1982). Non-randomised Markov and semi-Markov strategies in dynamic programming, *Theory and Applied Probability*, **27**, 116–126.
18. E. A. Fainberg (1982). Controlled Markov processes with arbitrary numerical criteria, *Theory and Applied Probability*, **27**, 486–503.
19. J. Filar, L. C. M. Kallenberg and H. M. Lee (1989). Variance-penalised Markov decision processes, *Mathematics of Operations Research*, **14**, 147–161.
20. R. Hartley (1979). Finite, discounted, vector Markov decision processes, Department of Decision Theory, University of Manchester. Working paper.
21. N. Hastings and J. Mello (1973). Tests for suboptimal actions in discounted Markov programming, *Management Science*, **19**, 1019–1022.
22. W. J. Hopp (1989). Identifying forecast horizons in non-homogeneous Markov decision procedures, *Operations Research*, **37**, 339–343.
23. R. A. Howard (1960). *Dynamic Programming and Markov Processes*, Wiley, London.
24. L. C. M. Kallenberg (1983). *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts, **148**, Mathematisch Centrum, Amsterdam.
25. J. G. Kemeny and J. L. Snell (1960). *Finite Markov Chains*, Van Nostrand, New York.
26. D. M. Kreps (1977). Decision problems with expected utility criteria, 1, Upper and lower convergent utility, *Mathematics of Operations Research*, **2**, 45–53.
27. H. Kushner (1971). *Introduction to Stochastic Control*, Holt, Rinehart and Winston, New York.
28. M. Landsberger and I. Meilijson (1982). Incentive generating state dependent penalty system—the case of tax evasion, *Journal of Public Economics*, **19**, 333–352.
29. W. S. Lovejoy (1986). Policy bounds for Markov decision processes, *Operations Research*, **34**, 630–637.

30. W. S. Lovejoy (1988). Computationally feasible bounds for partially observed Markov decision processes, *Research Paper No. 1024*, Graduate School of Business, Stanford University.
31. J. J. Martin (1967). *Bayesian Decision Problems and Markov Chains*, Wiley, London.
32. R. Mendelssohn (1982). An iterative aggregation procedure for Markov decision processes, *Operations Research*, **30**, 62–73.
33. R. Mendelssohn (1980). Managing stochastic multispecies models, *Mathematical Biosciences*, **49**, 249–261.
34. H. Mine and S. Osaki (1970). *Markovian Decision Processes*, Elsevier, New York.
35. T. E. Morton (1979). Infinite horizon dynamic programming models—a planning horizon formulation, *Operations Research*, **27**, 730–742.
36. J. M. Norman and D. J. White (1968). A method for approximate solutions to stochastic dynamic programming problems using expectations, *Operations Research*, **16**, 296–306.
37. E. L. Porteus (1971). Some bounds for discounted sequential decision processes, *Management Science*, **18**, 7–11.
38. A. R. Odani (1969). On finding the maximal gain for Markov decision processes, *Operations Research*, **17**, 857–860.
39. M. Puterman (1978). Theory of policy iteration, in *Dynamic Programming and Its Applications*, pp. 91–130, ed. M. Puterman, Academic Press, New York.
40. S. M. Ross (1983). *Introduction to Stochastic Dynamic Programming*, Academic Press, New York.
41. W. T. Scherer and D. J. White (1991). *The Convergence of Value Iteration in Discounted Markov Decision Processes*, working paper, Universities of Virginia and Manchester.
42. P. J. Schweitzer and A. Federgruen (1977). The asymptotic behaviour of value iteration in Markov decision problems, *Mathematics of Operations Research*, **2**, 360–381.
43. P. J. Schweitzer and A. Seidman (1983). Generalised polynomial approximations in Markov decision processes, *Working Paper No. QM8326*, The Graduate School of Management, The University of Rochester.
44. P. J. Schweitzer, M. L. Puterman and K. W. Kindle (1985). Iterative aggregation–disaggregation procedures for discounted semi-Markov reward processes, *Operations Research*, **33**, 589–605.

45. J. F. Shapiro (1968). Turnpike planning horizons for a Markovian decision model, *Management Science*, **14**, 292–306.
46. R. D. Smallwood (1966). Optimum policy regions for Markov processes with discounting, *Operations Research*, **14**, 658–669.
47. R. D. Smallwood and E. J. Sondik (1973). The optimal control of partially observable Markov processes over a finite horizon, *Operations Research*, **21**, 1071–1088.
48. M. J. Sobel (1982). The variance of discounted Markov decision processes, *Journal of Applied Probability*, **19**, 774–802.
49. L. A. Terry, M. V. F. Pereira, T. A. A. Neto, L. C. F. A. Silva and P. R. H. Sales (1986). Coordinating the energy generation of the Brazilian national hydrothermal electrical generating system, *Interfaces*, **16**, 16–38.
50. L. C. Thomas (1978). Student exercise, Department of Decision Theory, University of Manchester.
51. L. C. Thomas (1978). Student exercise, Department of Decision Theory, University of Manchester.
52. A. Turgeon (1980). Optimal operation of multireservoir power systems with stochastic inflow, *Water Resources Research*, **16**, 275–283.
53. J. van der Wal (1981). *Stochastic Dynamic Programming*, Mathematical Centre Tracts, **139**, Mathematisch Centrum, Amsterdam.
54. C. C. White (1980). The optimality of isotone strategies in Markov decision processes with utility criterion, in *Recent Developments in Markov Decision Processes*, pp. 261–276, eds. R. Hartley, L. C. Thomas, and D. L. White, Academic Press, London.
55. C. C. White and H. El-Deib (1986). Parameter imprecision in finite state, finite action dynamic programs, *Operations Research*, **34**, 120–129.
56. C. C. White and D. J. White (1989). Markov decision processes, invited review, *European Journal of Operational Research*, **39**, 1–16.
57. D. J. White (1969). *Dynamic Programming*, Holden-Day, San Francisco.
58. D. J. White (1978). *Finite Dynamic Programming*, Wiley, Chichester.
59. D. J. White (1982). *Optimality and Efficiency*, Wiley, Chichester.
60. D. J. White (1982). A multi-objective version of Bellman's inventory problem, *Journal of Mathematical Analysis and Applications*, **87**, 219–227.

61. D. J. White (1984). Isotone policies for the value iteration method for Markov decision processes, *OR Spektrum*, **6**, 223–227.
62. D. J. White (1985). Approximating Markov decision processes using expected state transitions, *Journal of Mathematical Analysis and Applications*, **107**, 167–181.
63. D. J. White (1987). Utility, probabilistic constraints, mean and variance of discounted rewards in Markov decision processes, *OR Spektrum*, **9**, 13–22.
64. D. J. White (1987). Decomposition in multi-item inventory control, *Journal of Optimisation Theory and Applications*, **54**, 383–401.
65. D. J. White (1987). Infinite horizon Markov decision processes with unknown or variable discount factors, *European Journal of Operational Research*, **28**, 96–100.
66. D. J. White (1988). Discount-isotone policies for Markov decision processes, *OR Spektrum*, **10**, 13–22.
67. D. J. White (1988). Mean, variance and probabilistic criteria in finite Markov decision processes: A review, *Journal of Optimisation Theory and Applications*, **56**, 1–29.
68. D. J. White (1989). Approximating the Markov property in Markov decision processes, *Information and Decision Technologies*, **15**, 147–162.
69. D. J. White (1989). Separable value functions for infinite horizon average reward Markov decision processes, *Journal of Mathematical Analysis and Applications*, **144**, 450–465.
70. D. J. White (1989). Solving infinite horizon discounted Markov decision process problems for a range of discount factors, *Journal of Mathematical Analysis and Applications*, **14**, 303–317.
71. D. J. White (1992). Computational procedures for variance-penalised Markov decision processes, *OR Spektrum*, **14**, 79–73.
72. D. J. White (1991). Markov decision processes: discounted expected reward or average expected reward? (working paper, to appear in *Journal of Mathematical Analysis and Applications*).
73. D. J. White (1992). Piecewise linear approximations for partially observable Markov decision processes, *Journal of Information and Optimisation Sciences*, **13**, 1–14.
74. D. J. White (1991). A superharmonic approach to partially observable Markov decision processes, University of Manchester. Working paper.
75. D. J. White (1991). Minimising a threshold probability in discounted Markov decision processes (working paper, to appear in *Journal of Mathematical Analysis and Applications*).

76. W. Whitt (1978). Approximations for dynamic programs, I, *Mathematics of Operations Research*, **3**, 231–243.
77. W. Whitt (1979). Approximations of dynamic programs, II, *Mathematics of Operations Research*, **4**, 179–185.
78. D. V. Widder (1946). *The Laplace Transform*, Princeton University Press, Princeton, New Jersey.
79. J. Wijngaard (1979). Decomposition for dynamic programming in production and inventory control, *Engineering and Process Economics*, **4**, 385–388.

Solutions to Exercises

CHAPTER 1

1.

$$r = \begin{bmatrix} 4 \\ -5 \end{bmatrix},$$

$$U - Pz = \begin{bmatrix} 1 - 0.8z & , & -0.2z \\ -0.7z & , & 1 - 0.3z \end{bmatrix},$$

$$(U - Pz)^{-1} = \begin{bmatrix} 1 - 0.3z & , & 0.2z \\ 0.7z & , & 1 - 0.8z \end{bmatrix} \Big/ (1 - 1.1z + 0.1z^2)$$

$$= \begin{bmatrix} 1 - 0.3z & , & 0.2z \\ 0.7z & , & 1 - 0.8z \end{bmatrix} \left(\frac{\frac{10}{9}}{1 - z} - \frac{\frac{1}{9}}{1 - 0.1z} \right)$$

$$= \begin{bmatrix} 0.7 + 0.3(1 - z) & , & 0.2 - 0.2(1 - z) \\ 0.7 - 0.7(1 - z) & , & 0.2 + 0.8(1 - z) \end{bmatrix} \frac{\frac{10}{9}}{(1 - z)}$$

$$+ \begin{bmatrix} 2 - 3(1 - 0.1z) & , & -2 + 2(1 - 0.1z) \\ -7 + 7(1 - 0.1z) & , & 7 - 8(1 - 0.1z) \end{bmatrix} \frac{\frac{1}{9}}{(1 - 0.1z)}$$

= (after some manipulation)

$$(1/(1 - z)) \begin{bmatrix} \frac{7}{9} & \frac{2}{9} \\ \frac{7}{9} & \frac{2}{9} \end{bmatrix} + (1/(1 - 0.1z)) \begin{bmatrix} \frac{2}{9} & -\frac{2}{9} \\ -\frac{7}{9} & \frac{7}{9} \end{bmatrix}.$$

Then

$$f(z) = (z/(1 - z)^2) \begin{bmatrix} \frac{7}{9} & \frac{2}{9} \\ \frac{7}{9} & \frac{2}{9} \end{bmatrix} \begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

$$+ (z/(1 - z)(1 - 0.1z)) \begin{bmatrix} \frac{2}{9} & -\frac{2}{9} \\ -\frac{7}{9} & \frac{7}{9} \end{bmatrix} \begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

$$= (z/(1 - z)^2) \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \left(\frac{\frac{10}{9}}{(1 - z)} - \frac{\frac{10}{9}}{(1 - 0.1z)} \right) \begin{bmatrix} 2 \\ -7 \end{bmatrix}$$

$$= \sum_{n=0}^{\infty} \left(n\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} \frac{20}{9} \\ -\frac{70}{9} \end{bmatrix} - (0.1)^n \begin{bmatrix} \frac{20}{9} \\ -\frac{70}{9} \end{bmatrix} \right) z^n.$$

Hence

$$v_n = n\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} \frac{20}{9} \\ -\frac{70}{9} \end{bmatrix} - (0.1)^n \begin{bmatrix} \frac{20}{9} \\ -\frac{70}{9} \end{bmatrix}.$$

2. From Exercise 1 we have

$$(U - Pz)^{-1} = (1/(1-z)) \begin{bmatrix} \frac{7}{9} & \frac{2}{9} \\ \frac{7}{9} & \frac{2}{9} \end{bmatrix} + (1/(1-0.1z)) \begin{bmatrix} \frac{2}{9} & -\frac{2}{9} \\ -\frac{7}{9} & \frac{7}{9} \end{bmatrix}.$$

Hence substituting $0.9z$ for $z (\rho = 0.9)$ we have

$$(U - 0.9Pz)^{-1} = (1/(1-0.9z)) \begin{bmatrix} \frac{7}{9} & \frac{2}{9} \\ \frac{7}{9} & \frac{2}{9} \end{bmatrix} + (1/(1-0.09z)) \begin{bmatrix} \frac{2}{9} & -\frac{2}{9} \\ -\frac{7}{9} & \frac{7}{9} \end{bmatrix}.$$

Hence

$$f(z) = (z/(1-z)(1-0.9z)) \begin{bmatrix} 2 \\ 2 \end{bmatrix} + (z/(1-z)(1-0.09z)) \begin{bmatrix} 2 \\ -7 \end{bmatrix}.$$

Now

$$\frac{z}{(1-z)(1-0.9z)} = \frac{10}{(1-z)} - \frac{10}{(1-0.9z)},$$

$$\frac{z}{(1-z)(1-0.09z)} = \frac{\frac{100}{91}}{(1-z)} - \frac{\frac{100}{91}}{(1-0.09z)}.$$

With some manipulation we obtain

$$f(z) = (1/(1-z)) \begin{bmatrix} 22 & \frac{18}{91} \\ 12 & \frac{28}{91} \end{bmatrix} + (1/(1-0.9z)) \begin{bmatrix} -20 \\ -20 \end{bmatrix}$$

$$+ (1/(1-0.09z)) \begin{bmatrix} -\frac{200}{91} \\ \frac{700}{91} \end{bmatrix}.$$

Hence

$$v_n = \begin{bmatrix} 22 & \frac{18}{91} \\ 12 & \frac{28}{91} \end{bmatrix} + (0.9)^n \begin{bmatrix} -20 \\ -20 \end{bmatrix} + (0.09)^n \begin{bmatrix} -\frac{200}{91} \\ \frac{700}{91} \end{bmatrix}.$$

3. *Non-discounted case.* We need to add $P^n\begin{bmatrix}10\\1\end{bmatrix}$ to the v_n form of Exercise 1 where, from Exercise 1, P^n is the coefficient of z^n in $(U - Pz)^{-1}$. Thus

$$v_n = n\begin{bmatrix}2\\2\end{bmatrix} + \begin{bmatrix}\frac{20}{9}\\-\frac{70}{9}\end{bmatrix} - (0.1)^n\begin{bmatrix}\frac{20}{9}\\-\frac{70}{9}\end{bmatrix}$$

$$+ (0.1)^n\begin{bmatrix}\frac{2}{9} & -\frac{2}{9}\\-\frac{7}{9} & \frac{7}{9}\end{bmatrix}\begin{bmatrix}10\\1\end{bmatrix} + \begin{bmatrix}\frac{7}{9} & \frac{2}{9}\\\frac{7}{9} & \frac{2}{9}\end{bmatrix}\begin{bmatrix}10\\1\end{bmatrix}$$

$$= n\begin{bmatrix}2\\2\end{bmatrix} + \begin{bmatrix}\frac{92}{9}\\\frac{2}{9}\end{bmatrix} + (0.1)^n\begin{bmatrix}-\frac{2}{9}\\\frac{7}{9}\end{bmatrix}.$$

Discounted case. We need to add $(0.9)^n P^n\begin{bmatrix}10\\1\end{bmatrix}$ to the v_n form of Exercise 2. Thus

$$v_n = \begin{bmatrix}22 & \frac{18}{91}\\12 & \frac{28}{91}\end{bmatrix} + (0.9)^n\begin{bmatrix}-20\\-20\end{bmatrix} + (0.09)^n\begin{bmatrix}-\frac{200}{91}\\\frac{700}{91}\end{bmatrix}$$

$$+ (0.09)^n\begin{bmatrix}\frac{2}{9} & -\frac{2}{9}\\-\frac{7}{9} & \frac{7}{9}\end{bmatrix}\begin{bmatrix}10\\1\end{bmatrix} + (0.9)^n\begin{bmatrix}\frac{7}{9} & \frac{2}{9}\\\frac{7}{9} & \frac{2}{9}\end{bmatrix}\begin{bmatrix}10\\1\end{bmatrix}$$

$$= \begin{bmatrix}22 & \frac{18}{91}\\12 & \frac{28}{91}\end{bmatrix} + (0.9)^n\begin{bmatrix}-12\\-12\end{bmatrix} + (0.09)^n\begin{bmatrix}-\frac{18}{91}\\\frac{63}{91}\end{bmatrix}.$$

4. One possibility is as follows:

$$P = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 1\\0 & 1 & 0\end{bmatrix}, \qquad r = \begin{bmatrix}1\\2\\3\end{bmatrix}.$$

Then

$$P^t = P \text{ if } t \text{ is odd},$$
$$P^t = U \text{ if } t \text{ is even}.$$

Then, using (1.5), *if n is even*

$$v_n = ((U + P) + (P + P^2) + \cdots + (P^{n-2} + P^{n-1}))r$$

$$= \begin{bmatrix}1\\2\\3\end{bmatrix} + \begin{bmatrix}1\\3\\2\end{bmatrix} + \left(\frac{n-2}{2}\right)\begin{bmatrix}2\\5\\5\end{bmatrix} = n\begin{bmatrix}1\\\frac{5}{2}\\\frac{5}{2}\end{bmatrix} + \begin{bmatrix}0\\0\\0\end{bmatrix},$$

if n is odd

$$v_n = (U + (P + P^2) + \cdots + (P^{n-2} + P^{n-1}))r$$

$$= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \left(\frac{n-1}{2}\right)\begin{bmatrix} 2 \\ 5 \\ 5 \end{bmatrix} = n\begin{bmatrix} 1 \\ \frac{5}{2} \\ \frac{5}{2} \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}.$$

Thus

$$g = \begin{bmatrix} 1 \\ \frac{5}{2} \\ \frac{5}{2} \end{bmatrix},$$

w is not well defined, and $w + \varepsilon_n$ does not converge, but oscillates between

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}.$$

5. (a) *Average expected reward.* From (1.25) we have, setting $h = g$,

$$u + ge = r + Pu, \qquad u(2) = 0.$$

Thus

$$u(1) + g = 4 + 0.8u(1),$$
$$g = -5 + 0.7u(1).$$

So $0.9u(1) = 9$, i.e. $u(1) = 10$, $g = 2$. Note that u differs from the true bias function by a constant function $(w(2) - u(2))e$.

(b) *Expected discounted reward.* From (1.35) we have, for $u = v$,

$$v = r + \rho Pv \quad \text{i.e. } v = (I - \rho P)^{-1}r$$

$$= \begin{bmatrix} 0.28 & -0.18 \\ -0.63 & 0.17 \end{bmatrix}^{-1}\begin{bmatrix} 4 \\ -5 \end{bmatrix} = \frac{\begin{bmatrix} 0.73 & 0.18 \\ 0.63 & 0.28 \end{bmatrix}}{(0.28 \times 0.73 - 0.18 \times 0.63)}\begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

$$= \begin{bmatrix} 2.02 \\ 1.12 \end{bmatrix} \Big/ 0.91 = \begin{bmatrix} 22.2 \\ 12.3 \end{bmatrix}.$$

6. (a) The function v_n is the expected total reward function when we have a final pay-off function equal to u. Let us use v_n^u instead of v_n, to avoid confusion with the case when $v_0 = 0$ (so $v_n = v_n^0$). Then from (1.20) we have, for $v_0 = u$

$$\underline{n \geqslant 1} \qquad v_n^u = r + Pv_{n-1}^u,$$
$$\underline{n = 0} \qquad v_0^u = u.$$

Then

$$v_1^u = r + Pu = u + g \text{ (using (1.25) with } h = g).$$

Assume that

$$v_s^u = u + sg, \qquad 0 \leqslant s \leqslant n - 1.$$

Then

$$\begin{aligned} v_n^u &= r + Pv_{n-1}^u = r + P(u + (n-1)g) \\ &= r + Pu + (n-1)g = u + g + (n-1)g \text{ (using (1.25))} \\ &= u + ng. \end{aligned}$$

(b) From (1.26) we have

$$\begin{aligned} v_n^u &= v_n^0 + P^n u \\ &= v_n^0 + \tilde{P}u + E_n u. \end{aligned}$$

Hence

$$\begin{aligned} v_n^0 &= v_n^u - \tilde{P}u - E_n u \\ &= (U - \tilde{P})u + ng - E_n u \\ &= ng + w^0 + \varepsilon_n \end{aligned}$$

where

$$w^0 = (U - \tilde{P})u, \qquad \varepsilon_n = -E_n u.$$

Then

$$\lim_{n \to \infty} [\varepsilon_n] = 0.$$

The matrix $\tilde{P}$ has identical rows, each equal to the steady-state probability vector.

7.

$$\left(\sum_{l=0}^{n-1} P^l \middle/ n \right) = P^* + F_n$$

where F_n tends to the zero matrix as n tends to infinity. Then

$$\begin{aligned} PP^* &= \left(\sum_{l=1}^{n} P^l \middle/ n \right) - PF_n \\ &= \left(\sum_{l=0}^{n-1} P^l \right) n - PF_n + (P^n - U)/n \\ &= P^* + (U - P)F_n + (P^n - U)/n. \end{aligned}$$

Now let n tend to ∞ to obtain the requisite result. The same analysis applies for P^*P.

CHAPTER 2

1.
$$\begin{aligned} E(C(c)) &= \alpha E(C(a)) + (1-\alpha)E(C(b)), \\ V(C(c)) &= E(C^2(c)) - E^2(C(c)) \\ &= \alpha E(C^2(a)) + (1-\alpha)E(C^2(b)) \\ &\quad - (\alpha E(C(a)) + (1-\alpha)E(C(b)))^2 \\ &= f(\alpha), \text{ say.} \end{aligned}$$

The function f is concave on $[0, 1]$. Hence f takes its minimal value at $\alpha = 0$ or $\alpha = 1$ (not necessarily uniquely). Hence

$$\begin{aligned} f(\alpha) &\geqslant \text{minimum}[f(1), f(0)] \\ &= \text{minimum}[V(C(a)), V(C(b))], \quad \forall\, \alpha \in [0, 1]. \end{aligned}$$

2. (a)
$$R_n = \sum_{t=1}^{n} \left(\prod_{s=1}^{t} \rho(s) \right) Y_t.$$

(b)
$$\begin{aligned} R_n &= \sum_{t=1}^{n} \left(\prod_{s=1}^{t-1} \rho(s) \right) (\rho(t) Y_t) \\ &= \sum_{t=1}^{n} \left(\prod_{s=0}^{t-1} \rho(s) \right) \tilde{Y}_t \end{aligned}$$

where $\tilde{Y}_t = \rho(t) Y_t$ and $\rho(0) = 1$.

3. (a) The state of the system, X_t, at the beginning of time unit t may take the form
$$X_t = (S_t, D_1, D_2, ., ., .D_{t-1})$$
where S_t is the stock level at the beginning of time unit t and D_u is the demand in time unit u. Then

probability
$$\begin{aligned} &X_{t+1} = (i_{t+1}, s_1, s_2, ., ., s_t \mid X_t = (i_t, s_1, s_2, ., ., ., s_{t-1}), Z_t = k_t) \\ &\quad = \text{probability } (S_{t+1} = i_{t+1}, \\ &\quad D_t = s_t \mid X_t = (i_t, s_1, s_2, ., ., s_{t-1}), Z_t = k_t) \end{aligned}$$

$$= \frac{\sum_{\alpha=1}^{2} p(\alpha)\left(\prod_{u=1}^{t} q(s_u, \alpha)\right)}{\left(\sum_{\alpha=1}^{2} p(\alpha) \prod_{u=1}^{t-1} q(s_u, \alpha)\right)} \text{ if } i_{t+1} = \text{maximum}[i_t, k_t] - s_t,$$

$$= 0 \text{ otherwise.}$$

Thus the state i is replaced by $(i, s_1, s_2, ., ., s_{t-1})$ whose dimension increases with t.

(b) Alternatively, the state X_t may take the form $X_t = (S_t, \Phi_t(1), \Phi_t(2))$ where S_t is as in (a), and $\Phi_t(\alpha)$ is the probability, at the beginning of time unit t, that the parameter takes a value α. Then

$$\begin{aligned} &\text{probability}(X_{t+1} = (i_{t+1}, x_1(t+1), x_2(t+1)) \\ &\qquad \mid X_t = (i_t, x_1(t), x_2(t)), Z_t = k_t) \\ &\qquad = \text{probability}(D_t = \text{maximum}[i_t, k_t] - i_{t+1} \\ &\qquad \mid \Phi_t(1) = x_1(t), \Phi_t(2) = x_2(t)) \end{aligned} \tag{A}$$

if

$$x_\alpha(t+1) = \frac{x_\alpha(t) q(\text{maximum}[i_t, k_t] - i_{t+1}, 1)}{\sum_{\alpha=1}^{2} x_\alpha(t) q(\text{maximum}[i_t, k_t] - i_{t+1}, \alpha)}, \qquad \alpha = 1, 2,$$

$$= 0 \quad \text{otherwise.} \tag{B}$$

The expression for (A) is the denominator in (B).

4.

$\underline{n = 0}$

$v_0(1) = v_0(2) = v_0(3) = 0.$

$\underline{n = 1}$

$$v_1(1) = \text{maximum}\begin{bmatrix} k = 1\colon 4 \\ k = 2\colon 4 \end{bmatrix} = 4, \qquad \delta_7(1) = 1 \text{ or } 2,$$

$$v_1(2) = \text{maximum}\begin{bmatrix} k = 1\colon 9 \\ k = 2\colon 10 \end{bmatrix} = 10, \qquad \delta_7(2) = 2,$$

$$v_1(3) = \text{maximum}\begin{bmatrix} k = 1\colon 4 \\ k = 2\colon 3 \end{bmatrix} = 4, \qquad \delta_7(3) = 1.$$

$n = 2$

$$v_2(1) = \text{maximum}\begin{bmatrix} k=1\text{: } 4 + 0\times 0 + \frac{1}{4}\times 10 + \frac{1}{2}\times 4 = 8.5 \\ k=2\text{: } 4 + 0\times 0 + \frac{3}{8}\times 10 + \frac{3}{8}\times 4 = 9.25 \end{bmatrix}$$

$$= 9.25, \delta_6(1) = 2,$$

$$v_2(2) = \text{maximum}\begin{bmatrix} k=1\text{: } 9 + \frac{3}{16}\times 4 + 0\times 10 + \frac{9}{16}\times 4 = 12.0 \\ k=2\text{: } 10 + \frac{1}{2}\times 4 + 0\times 10 + \frac{1}{4}\times 4 = 13.0 \end{bmatrix}$$

$$= 13.0, \delta_6(2) = 2,$$

$$v_2(3) = \text{maximum}\begin{bmatrix} k=1\text{: } 4 + \frac{1}{2}\times 4 + \frac{1}{4}\times 10 + 0\times 4 = 8.5 \\ k=2\text{: } 3 + \frac{3}{8}\times 4 + \frac{3}{8}\times 10 + 0\times 4 = 8.25 \end{bmatrix}$$

$$= 8.5, \delta_6(3) = 1,$$

n		i = 1	2	3
0	v_0	0	0	0
	δ_8	—	—	—
1	v_1	4	10	4
	δ_7	1 or 2	2	1
2	v_2	9.25	13.00	8.5
	δ_6	2	2	1
3	v_3	12.06	16.75	11.88
	δ_5	2	2	1
4	v_4	14.74	19.00	14.22
	δ_4	2	2	1
5	v_5	16.46	20.93	16.12
	δ_3	2	2	1
6	v_6	17.89	22.26	17.46
	δ_2	2	2	1
7	v_7	18.89	23.31	18.51
	δ_1	2	2	1

Note that $n(\text{maximum}) = 7$, and $t = 1$ corresponds to $n = 7$ in accordance with our convention.

5. Results 2.4 and 2.5 show that v is the unique solution to $u = Tu$, i.e.

$$u(1) = \text{maximum}\begin{bmatrix} k = 1\colon\ 4 + \frac{1}{4}u(2) + \frac{1}{2}u(3) \\ k = 2\colon\ 4 + \frac{3}{8}u(2) + \frac{3}{8}u(3) \end{bmatrix},$$

$$u(2) = \text{maximum}\begin{bmatrix} k = 1\colon\ 9 + \frac{3}{16}u(1) + \frac{9}{16}u(3) \\ k = 2\colon\ 10 + \frac{1}{2}v(1) + \frac{1}{4}v(3) \end{bmatrix},$$

$$u(3) = \text{maximum}\begin{bmatrix} k = 1\colon\ 4 + \frac{1}{2}u(1) + \frac{1}{4}u(2) \\ k = 2\colon\ 3 + \frac{3}{8}u(1) + \frac{3}{8}u(2) \end{bmatrix}.$$

Result 2.6 shows that if δ is a decision rule solution to the above then the policy $\pi = (\delta)^\infty$ is optimal among all policies. Finally let $\tau = (\sigma)^\infty$ be any other optimal stationary deterministic Markov policy. Then its associated reward value function $v^\tau = v$ satisfies

$$T^\sigma u = u = Tu.$$

Thus σ is also a decision rule solution to $u = Tu$. Hence if we evaluate v^τ for π as specified (using $v^\tau = T^\delta v^\tau$) we need only show that δ is the only decision rule solution to $u = Tu$. To obtain v^π we solve

$$\begin{aligned} u(1) &= 4 + \tfrac{3}{8}u(2) + \tfrac{3}{8}u(3), \\ u(2) &= 10 + \tfrac{1}{2}u(1) + \tfrac{1}{4}u(3), \\ u(3) &= 4 + \tfrac{1}{2}u(1) + \tfrac{1}{4}u(2) \end{aligned}$$

giving

$$v^\pi = u = (22.00, 26.40, 21.60).$$

It is easily established that $v = v^\pi$ satisfies $u = Tu$ and that the only decision rule optimum is δ.

6. From Result 2.10, if we can show that δ is a decision rule solution to $u + he = Tu$, $u(m) = 0$ then $\pi = (\delta)^\infty$ will be an optimal policy. The equations take the form

$$u(1) + h = \text{maximum}\begin{bmatrix} k = 1\colon\ 4 + \frac{1}{3}u(2) + \frac{2}{3}u(3) \\ k = 2\colon\ 4 + \frac{1}{2}u(2) + \frac{1}{2}u(3) \end{bmatrix},$$

$$u(2) + h = \text{maximum}\begin{bmatrix} k = 1\colon\ 9 + \frac{1}{4}u(1) + \frac{3}{4}u(3) \\ k = 2\colon\ 10 + \frac{2}{3}u(1) + \frac{1}{3}u(3) \end{bmatrix},$$

$$u(3) + h = \text{maximum}\begin{bmatrix} k = 1\colon\ 4 + \frac{2}{3}u(1) + \frac{1}{3}u(2) \\ k = 2\colon\ 3 + \frac{1}{2}u(1) + \frac{1}{2}u(2) \end{bmatrix},$$

$$u(3) = 0.$$

For policy $\pi = (\delta)^\infty$ we solve the equations

$$\begin{aligned} u(1) + h &= 4 + \tfrac{1}{2}u(2) + \tfrac{1}{2}u(3), \\ u(2) + h &= 10 + \tfrac{1}{3}u(1) + \tfrac{1}{3}u(3), \\ u(3) + h &= 4 + \tfrac{2}{3}u(1) + \tfrac{1}{3}u(2), \\ u(3) &= 0. \end{aligned}$$

Thus

$$w^\pi(1) = u(1) = \tfrac{9}{20}, \qquad w^\pi(2) = u(2) = \tfrac{9}{2}, \qquad w^\pi(3) = u(3) = 0,$$
$$g^\pi = he = 5\tfrac{4}{5}\, e.$$

It is easily seen that these satisfy the optimality equation.

7. With $\tilde{v}_0 = u$, where u solves (2.85) and (2.86), we have

$$\tilde{v}_1 = T\tilde{v}_0 = Tu = T^\delta u = \tilde{v}_1^\pi.$$

Assume that

$$\tilde{v}_s = \tilde{v}_s^\pi \quad \text{for } 1 \leqslant s \leqslant n-1.$$

Then

$$\begin{aligned} \tilde{v}_n^\pi &= T^\delta \tilde{v}_{n-1}^\pi = T^\delta \tilde{v}_{n-1} \\ &= T^\delta(u + (n-1)he) = T^\delta u + (n-1)he \\ &= Tu + (n-1)he = u + nhe = \tilde{v}_n. \end{aligned}$$

Alternatively we have

$$\tilde{v}_1 = T^\delta u = \tilde{v}_1^\pi.$$

Assume that

$$\tilde{v}_s = (T^\delta)^s u = \tilde{v}_s^\pi, \qquad 1 \leqslant s \leqslant n-1.$$

Then

$$\tilde{v}_n = T\tilde{v}_{n-1} = T(T^\delta)^{n-1}u.$$

Now

$$\begin{aligned} TT^\delta u &= T^2 u, \\ T^\delta Tu &= T^\delta(u + he) = T^\delta u + he \\ &= Tu + he = u + 2he = \tilde{v}_2 = T^2 u. \end{aligned}$$

Thus T and T^δ commute. Hence

$$\tilde{v}_n = T(T^\delta)^{n-1}u = (T^\delta)^{n-1}Tu = (T^\delta)^{n-1}T^\delta u = (T^\delta)^n u = \tilde{v}_n^\pi.$$

8. Define v_{tn} to be the maximal expected total discounted reward value function for the next n time units from the beginning of time unit t. As with the earlier reasoning we may keep within Π_{MD} without loss. Then v_{tn} satisfies

$\underline{t \geqslant 1, n \geqslant 1}$

$$u_{tn}(i) = \underset{k \in K_t(i)}{\text{maximum}} \left[r_i^k(t) + \rho(t) \sum_{j \in I_{t+1}} p_{ij}^k(t) u_{t+1,n-1}(j) \right], \quad \forall\, i \in I_t,$$

$\underline{t \geqslant 1, n = 0}$ $\qquad u_{t0} = 0.$

9. Some conditions need to be placed on $\{r_i^k(t)\}$, $\{\rho(t)\}$ to ensure that $\lim_{n \to \infty} [v_{tn}] = v_t$ exists (e.g. $|r_i^k(t)| \leqslant M < \infty$, $\forall\, i \in I_t$, $k \in K_t(i)$, $t \geqslant 1$, and $0 \leqslant \rho_t \leqslant \rho < 1$ for some ρ). Then v_t satisfies

$$u_t(i) = \underset{k \in K_t(i)}{\text{maximum}} \left[r_i^k(t) + \rho(t) \sum_{j \in I_{t+1}} p_{ij}^k(t) u_{t+1}(j) \right], \quad \forall\, i \in I_t,\ t \geqslant 1.$$

CHAPTER 3

1. $\bar{r} = 10$, $\underline{r} = 3$. We require that

Result 3.4. $3/0.25 \leqslant v(i) \leqslant 10/0.25$ i.e. $12 \leqslant v(i) \leqslant 40$, $1 \leqslant i \leqslant 3$. This is satisfied since $v = (22.00, 26.40, 21.60)$.

Result 3.5. $\alpha = 26.40$, $\beta = 21.60$.

$$u_n(i) + 21.60\rho^n \leqslant v(i) \leqslant u_n(i) + 26.40\rho^n.$$

$\underline{i = 1}$ $\quad u_n(1) + 21.60\rho^n \leqslant 22.00 \leqslant u_n(1) + 26.40\rho^n,$

$\underline{i = 2}$ $\quad u_n(2) + 21.60\rho^n \leqslant 26.40 \leqslant u_n(2) + 26.40\rho^n,$

$\underline{i = 3}$ $\quad u_n(3) + 21.60\rho^n \leqslant 21.60 \leqslant u_n(3) + 26.40\rho^n.$

These should be checked for $1 \leqslant n \leqslant 7$ against the answer to Exercise 4 of Chapter 2.

Result 3.7. The sequence $\{u_n\}$ is non-decreasing. This must hold since with $u = 0$ and $r_i^k \geqslant 0$, $\forall\ i, k$, we must have

$$0 \leqslant T0.$$

Result 3.8.

n	$(\rho/(1-\rho))\alpha_n = U_n$	$(\rho/(1-\rho))\beta_n = L_n$
1	30.00	12.00
2	16.75	9.00
3	11.25	8.43
4	8.04	7.02
5	5.79	5.16
6	4.29	3.99
7	3.15	3.00.

Check that for $1 \leqslant n \leqslant 7$

$$\begin{aligned} &\underline{i=1} \quad u_n(1) + L_n \leqslant 22.00 \leqslant u_n(1) + U_n, \\ &\underline{i=2} \quad u_n(2) + L_n \leqslant 25.40 \leqslant u_n(2) + U_n, \\ &\underline{i=3} \quad u_n(3) + L_n \leqslant 21.60 \leqslant u_n(3) + U_n. \end{aligned}$$

Result 3.10.

$$v + (\rho/(1-\rho))(\beta_n - \alpha_n)e \leqslant v^{\pi_n} \leqslant v.$$

We have

$$\begin{aligned} &\sigma_1^1 = (1, 2, 1), \qquad \sigma_1^2 = (2, 2, 1), \\ &\sigma_n = (2, 2, 1), \quad \forall\ n \geqslant 2. \end{aligned}$$

Policy $\pi = (\delta)^\infty$, where $\delta = (2, 2, 1)$, is optimal. Since $\alpha_n \geqslant \beta_n$ the above inequality is automatically true $\forall\ n \geqslant 2$ and for $n = 1$ when $\sigma_1 = (2, 2, 1)$. We need to check the policy $\pi_1 = (\sigma_1^1)^\infty$ for $n = 1$.

We have

$$\begin{aligned} v^{\pi_1}(1) &= 4 + \tfrac{1}{4} \times v^{\pi_1}(2) + \tfrac{1}{2} \times v^{\pi_1}(3), \\ v^{\pi_1}(2) &= 10 + \tfrac{1}{2} \times v^{\pi_1}(1) + \tfrac{1}{4} \times v^{\pi_1}(3), \\ v^{\pi_1}(3) &= 4 + \tfrac{1}{2} \times v^{\pi_1}(1) + \tfrac{1}{4} \times v^{\pi_1}(2). \end{aligned}$$

Thus $v^{\pi_1} = (20\frac{4}{5}, 25\frac{3}{5}, 20\frac{4}{5})$. Thus we need to check that, for $n = 1$

$$\begin{array}{ll} \underline{i = 1} & 22.00 - 18 \leqslant 20.80 \leqslant 22.00, \\ \underline{i = 2} & 26.40 - 18 \leqslant 25.60 \leqslant 26.40, \\ \underline{i = 3} & 21.60 - 18 \leqslant 20.80 \leqslant 21.60. \end{array}$$

Result 3.11. We have

n	α_n	β_n
1	10.00	4.00
2	5.25	3.00
3	3.75	2.81
4	2.68	2.34
5	1.93	1.72
6	1.43	1.33
7	1.05	1.00.

We need to check that $\beta_n/\beta_{n-1} \geqslant 0.75 \geqslant \alpha_n/\alpha_{n-1}$ for $n \geqslant 2$.

Result 3.12. This is the same as the (modified) Result 3.10.

2. We first of all establish (3.195) following the analysis of Result 3.8. Here $\hat{u}_n = T^{\hat{\sigma}_n}\hat{u}_{n-1}$ with $\hat{\sigma}_n(i) \in R_n(i)$, $\forall\, i \in I$; $v = Tv = T^{\delta}v$, with $\delta \in \underset{\delta \in \Delta}{\text{arg maximum}}\, [T^{\delta}v]$. Note that $\delta(i) \in R_n(i)$, $\forall\, i \in I$. Then $v - \hat{u}_n \leqslant T^{\delta}v - T^{\delta}\hat{u}_{n-1}$ since $\delta(i) \in R_n(i)$, $\forall\, i \in I$.

The analysis follows on as for Result 3.8 to give, analogously to (3.54)

$$v - \hat{u}_n \leqslant (U - \rho P^{\delta})^{-1}\rho P^{\delta}(\hat{u}_n - \hat{u}_{n-1}).$$

This gives the right-hand side of (3.195). The left-hand side follows in a similar manner since $\hat{\sigma}_n \in \Delta$.

$\underline{n = 0}$

For the example we have $\hat{u}_0 = 0$, $R_1(i) = K(i)$, $i = 1, 2$.

$\underline{n = 1}$

Thus $\hat{u}_1 = (4, -3)$, $\hat{\alpha}_1 = 6$, $\hat{\beta}_1 = -3$. Then

$\underline{n = 2}$

$$\begin{array}{ll} \bar{u}_2(1) = 6 + 54 = 60, & \underline{u}_2(1) = 6 - 27 = -21, \\ \bar{u}_2(2) = -3 + 54 = 51, & \underline{u}_2(2) = -3 - 27 = -30. \end{array}$$

Then

$$\underline{i=1} \qquad [T^1\bar{u}_2](1) = 6 + 0.9(0.5 \times 60 + 0.5 \times 51) = 56.00,$$
$$[T^2\underline{u}_2](1) = 4 + 0.9(0.8 \times (-21) + 0.2 \times (-30))$$
$$= -16.52,$$
$$[T^2\bar{u}_2](1) = 4 + 0.9(0.8 \times 60 + 0.5 \times 51) = 56.40,$$
$$[T^1\underline{u}_2](1) = 6 + 0.9(0.5 \times (-21) + 0.5 \times (-30))$$
$$= -16.94.$$

Thus $R_2(1) = K(1)$.

$$\underline{i=2} \qquad [T^1\bar{u}_2](2) = -3 + 0.9(0.4 \times 60 + 0.6 \times 51) = 41.55,$$
$$[T^2\underline{u}_2](2) = -5 + 0.9(0.7 \times (-21) + 0.3 \times (-30))$$
$$= -21.33,$$
$$[T^2\bar{u}_2](2) = -5 + 0.9(0.7 \times 60 + 0.3 \times 51) = 46.75,$$
$$[T^1\underline{u}_2](2) = -3 + 0.9(0.4 \times (-21) + 0.6 \times (-30))$$
$$= -26.76.$$

Thus $R_2(2) = K(2)$.

$$\hat{u}_2 = u_2, \qquad \hat{\alpha}_2 = 1.78, \qquad \hat{\beta}_2 = 0.97.$$

$\underline{n=3}$

$$\bar{u}_3(1) = 7.78 + 16.02 = 23.80, \qquad \underline{u}_3(1) = 7.78 + 8.73 = 16.51,$$
$$\bar{u}_3(2) = -2.03 + 16.02 = 13.97, \qquad \underline{u}_2(2) = -2.03 + 8.73 = 6.70.$$

Then

$$\underline{i=1} \qquad [T^1\bar{u}_3](1) = 6 + 0.9(0.5 \times 23.80 + 0.5 \times 13.97) = 23.00,$$
$$[T^2\underline{u}_3](1) = 4 + 0.9(0.8 \times 16.51 + 0.2 \times 6.70) = 17.09,$$
$$[T^2\bar{u}_3](1) = 4 + 0.9(0.8 \times 23.80 + 0.2 \times 13.97) = 23.65,$$
$$[T^1\underline{u}_3](1) = 6 + 0.9(0.5 \times 16.51 + 0.5 \times 6.70) = 16.05.$$

Hence $R_3(1) = K(1)$.

$$\underline{i=2} \qquad [T^1\bar{u}_3](2) = -3 + 0.9(0.4 \times 23.80 + 0.6 \times 13.97) = 13.11,$$
$$[T^2\underline{u}_3](2) = -5 + 0.9(0.7 \times 16.51 + 0.3 \times 6.70) = 7.21,$$
$$[T^2\bar{u}_3](2) = -5 + 0.9(0.7 \times 23.80 + 0.3 \times 13.97) = 13.77,$$
$$[T^1\underline{u}_3](2) = -3 + 0.9(0.4 \times 16.51 + 0.6 \times 6.70) = 6.56.$$

Hence $R_3(2) = K(2)$.

$$\hat{u}_3 = u_3, \quad \hat{\alpha}_3 = 1.46, \quad \hat{\beta}_3 = 1.38.$$

$n = 4$

$$\bar{u}_4(1) = 9.24 + 13.14 = 22.38, \qquad \underline{u}_4(1) = 9.24 + 12.42 = 21.66,$$
$$\bar{u}_4(2) = -0.65 + 13.14 = 12.49, \qquad \underline{u}_4(2) = -0.65 + 12.42 = 11.77.$$

Then

$i = 1$
$$[T^1\bar{u}_4](1) = 6 + 0.9(0.5 \times 22.38 + 0.5 \times 12.49) = 21.70,$$
$$[T^2\underline{u}_4](1) = 4 + 0.9(0.8 \times 21.66 + 0.2 \times 11.77) = 21.72.$$

Hence eliminate action 1.

$i = 2$
$$[T^1\bar{u}_4](2) = -3 + 0.9(0.4 \times 22.38 + 0.6 \times 12.49) = 11.80,$$
$$[T^2\underline{u}_4](2) = -5 + 0.9(0.7 \times 21.66 + 0.3 \times 11.77) = 11.82.$$

Hence eliminate action 1. Thus $R_4(1) = R_4(2) = \{2\}$ and $\delta = (2, 2)$ is optimal.

3.
$$\hat{u}_n = T^{\hat{\sigma}_n}\hat{u}_{n-1}, \qquad \hat{u}_{n-1} = T^{\hat{\sigma}_{n-1}}\hat{u}_{n-2},$$
$$\hat{u}_n - \hat{u}_{n-1} = T^{\hat{\sigma}}\hat{u}_{n-1} - T^{\hat{\sigma}_{n-1}}\hat{u}_{n-2},$$
$$\hat{\sigma}_n(i) \in \arg \max_{k \in R_n(i)} [[T^k\hat{u}_{n-1}](i)],$$
$$\hat{\sigma}_{n-1}(i) \in \arg \max_{k \in R_{n-1}(i)} [[T^k\hat{u}_{n-2}](i)].$$

Thus $\hat{\sigma}_n(i) \in R_n(i) \subseteq R_{n-1}(i)$ and since $T^{\hat{\sigma}_{n-1}}\hat{u}_{n-2} \leqslant T\hat{u}_{n-2}$ we have

$$\hat{u}_n - \hat{u}_{n-1} \leqslant T^{\hat{\sigma}_n}\hat{u}_{n-1} - T^{\hat{\sigma}_n}\hat{u}_{n-2}$$
$$= \rho P^{\hat{\sigma}_n}(\hat{u}_{n-1} - \hat{u}_{n-2}).$$

Hence for $n \geqslant 2$, $\hat{\alpha}_n \leqslant \rho\hat{\alpha}_{n-1} \leqslant \cdots \rho^{n-1}\hat{\alpha}_1$.

However, for the $\hat{\beta}_n$ result we would require that $\hat{\sigma}_{n-1}(i) \in R_n(i)$ in order to be able to deduce that $T^{\hat{\sigma}_n}\hat{u}_{n-1} \geqslant T^{\hat{\sigma}_{n-1}}\hat{u}_{n-1}$, and hence in order to deduce $\hat{u}_n - \hat{u}_{n-1} \geqslant \rho P^{\hat{\sigma}_n}(\hat{u}_{n-1} - \hat{u}_{n-2})$ and $\hat{\beta}_n \geqslant \rho\hat{\beta}_{n-1}$. However, $\hat{\sigma}_{n-1}(i)$ might be eliminated from $R_{n-1}(i)$ and not be in $R_n(i)$.

4. Let $N(i)$ be the non-optimal actions for state i. Let $k \in N(i)$. Then for the function w in (3.109) in particular, we have for some $q \in K(i)$

$$[T^q w](i) > [T^k w](i).$$

This comes essentially from Result 2.10 since, if δ solves (2.85) and $\delta(i) = q$, then so does τ solve (2.85) where $\tau(j) = \delta(j)$, $\forall\, j \neq i$, and

$\tau(i) = k$ if the above inequality does not hold. Then

$$\begin{aligned}[T^q u_{n-1}](i) - [T^k u_{n-1}](i) &= ([T^q u_{n-1}](i) - [T^q w](i)) \\ &\quad + ([T^q w](i) - [T^k w](i)) \\ &\quad + ([T^k w](i) - [T^k u_{n-1}](i)) \\ &= A + B + C, \text{ say.}\end{aligned}$$

Let $[T^q w](i) - [T^k w](i) = \varepsilon(i) > 0$. Then $B = \varepsilon(i)$.

The quantity A is equal to $(n-1)g + p_i^q \varepsilon_{n-1}$, where p_i^q is the transition probability vector for action q and initial state i.

$$C = -(n-1)g - p_i^k \varepsilon_{n-1}.$$

So

$$A + B + C = \varepsilon(i) + (p_i^q - p_i^k)\varepsilon_{n-1}.$$

Since $\{\varepsilon_n\}$ tends to zero as n tends to infinity we have $A + B + C > 0$ if n is large enough. Hence $[T^q u_{n-1}](i) > [T^k u_{n-1}](i)$ if n is large enough, i.e. k is not optimal for the nth iteration if $n \geqslant n_0(i)$ for some $n_0(i)$. If $n_0 = \underset{i \in I}{\text{maximum}}\,[n_0(i)]$ we obtain the desired result.

5. Let $I = \{1, 2\}$, $K(1) = \{1, 2\}$, $r_1^1 = 0$, $r_1^2 = 1$, $K(2) = \{1\}$, $r_2^1 = 1$, $p_{12}^1 = p_{12}^2 = 1$, $p_{22}^1 = 1$. Then the optimality equations are

$$u(1) + h = \text{maximum}\begin{bmatrix} k = 1\colon\ 0 + u(2) \\ k = 2\colon\ 1 + u(2) \end{bmatrix},$$

$$u(2) + h = 1 + u(2).$$

So $w(2) = u(2) = 0$ (by convention)

$$\begin{aligned} g &= h = 1, \\ w(1) &= u(1) = 0, \\ \pi &= (\delta)^\infty, \quad \delta(1) = 2, \quad \delta(2) = 1 \text{ is optimal.}\end{aligned}$$

However, consider the policy $\tau = (\tilde{\delta})^\infty$ where $\tilde{\delta}(1) = 1$, $\tilde{\delta}(2) = 1$. This gives the same gain $g^\tau = 1$. However, with w as above

$$[Tw](1) = 1 \neq [T^1 w](1) = 0$$

although

$$[Tw](2) = 1 = [T^1 w](2).$$

The reason why this can happen is that $i = 1$ is a transient state, for the optimal policy τ (in particular), with a zero steady state (or

limiting average) probability and it does not matter which action is taken in this state for gain optimality.

6. Let $I = \{1, 2, 3\}$, $K(1) = \{1, 2\}$, $K(2) = \{1\}$, $K(3) = \{1\}$, $p^1_{12} = 1$, $p^1_{11} = p^1_{13} = 0$, $p^2_{13} = 1$, $p^2_{11} = p^2_{12} = 0$, $p^1_{22} = 1$, $p^1_{12} = p^1_{23} = 0$, $p^1_{33} = 1$, $p^1_{31} = p^1_{32} = 0$, $r^1_{12} = 1$, $r^2_{13} = \frac{1}{2}$, $r^1_{22} = -\frac{1}{2}$, $r^1_{33} = \frac{1}{2}$, $\rho = \frac{1}{3}$, $u_0 = 0$. Then

$$\underline{n \geqslant 1} \qquad u_n(2) = -\tfrac{1}{2}(1-\rho^n)/(1-\rho), \qquad \sigma_n(2) = 1,$$
$$u_n(3) = \tfrac{1}{2}(1-\rho^n)/(1-\rho), \qquad \sigma_n(3) = 1.$$

$$\underline{n = 1} \qquad u_1(1) = [Tu_0](1) = 1, \qquad \sigma_1(1) = 1,$$

$$\underline{n \geqslant 2} \qquad u_n(1) = [Tu_{n-1}](1) = \text{maximum}\begin{bmatrix} k=1\colon\ 1 + \rho u_{n-1}(2) \\ k=2\colon\ \frac{1}{2} + \rho u_{n-1}(3) \end{bmatrix}$$

$$= \text{maximum}\begin{bmatrix} k=1\colon\ 1 - (\rho/2)(1-\rho^{n-1})/(1-\rho) \\ k=2\colon\ \frac{1}{2} + (\rho/2)(1-\rho^{n-1})/(1-\rho) \end{bmatrix}$$

$$= \text{maximum}\begin{bmatrix} k=1\colon\ ((2-3\rho)+\rho^n)/2(1-\rho) \\ k=2\colon\ (1-\rho^n)/2(1-\rho) \end{bmatrix}$$

$$= \text{maximum}\begin{bmatrix} k=1\colon\ (1+\rho^n)/2(1-\rho) \\ k=2\colon\ (1-\rho^n)/2(1-\rho) \end{bmatrix}.$$

Thus

$$\sigma_n = (1, 1, 1) \text{ uniquely}, \quad \forall\ n \geqslant 1.$$

However, solving (3.1) we have

$$v(2) = u(2) = -\tfrac{1}{2}(1-\rho),$$
$$v(3) = u(3) = \tfrac{1}{2}(1-\rho),$$

$$v(1) = u(1) = \text{maximum}\begin{bmatrix} k=1\colon\ (2-3\rho)/2(1-\rho) \\ k=2\colon\ 1/2(1-\rho) \end{bmatrix} = \tfrac{3}{4} \quad \text{when } \rho = \tfrac{1}{3},$$

and $\delta = (1, 1, 1)$ or $(2, 1, 1)$ are equally optimal.

7. We first of all find v^π by solving

$$v^\pi(1) = 90 + 0.45v^\pi(1) + 0.45v^\pi(2),$$
$$v^\pi(2) = 60 + 0.63v^\pi(1) + 0.27v^\pi(3),$$
$$v^\pi(3) = 90 + 0.27v^\pi(1) + 0.63v^\pi(3).$$

This gives

$$v^{\pi}(1) = 803,$$
$$v^{\pi}(2) = 782,$$
$$v^{\pi}(3) = 800.$$

$$[T^1 v^{\pi}](1) = 803, \qquad [T^2 v^{\pi}](1) = 822, \tag{1.1}$$

$$[T^1 v^{\pi}](2) = 807, \qquad [T^2 v^{\pi}](2) = 782, \tag{1.2}$$

$$[T^1 v^{\pi}](3) = 800, \qquad [T^2 v^{\pi}](3) = 855. \tag{1.3}$$

Thus v^{π} satisfies $u = Tu$. Thus π is optimal among all the stationary deterministic Markov policies since the policy solutions to $u = Tu$ give exactly the set of such optimal policies. Any stationary deterministic Markov policy $\tau = (\sigma)^{\infty} \neq \pi$ will give, from (1.1) to (1.3), $[T^{\sigma(i)} v^{\pi}](i) > v^{\pi}(i)$ for some i and cannot be optimal.

8. If $\{\delta, w^{\pi}, g^{\pi}\}$ satisfy the optimality equation then $\pi = (\delta)^{\infty}$ is optimal. The $\{r_i^k\}$ tabulations are as follows:

i	k	r_i^k
1	1	15
	2	16
2	1	9.75
	2	9.5

Solving for $\{w^{\pi}, g^{\pi}\}$ with $w^{\pi}(2) = 0$, we need to solve

$$g^{\pi} + w^{\pi}(1) = 16 + 0.8 w^{\pi}(1),$$
$$g^{\pi} = 9.5 + 0.5 w^{\pi}(1).$$

Thus

$$g^{\pi} = 14\tfrac{1}{7}, \qquad w^{\pi}(1) = 9\tfrac{2}{7}, \qquad w^{\pi}(2) = 0.$$

	$[T^k w^{\pi}](i)$
$i = 1, k = 1$	$15 + 0.5 \times 9\tfrac{2}{7} = 19\tfrac{9}{14}$,
$k = 2$	$16 + 0.8 \times 9\tfrac{2}{7} = 23\tfrac{3}{7}$,
$i = 2, k = 1$	$9\tfrac{3}{4} + 0.25 \times 9\tfrac{2}{7} = 12\tfrac{1}{14}$,
$k = 2$	$9\tfrac{1}{2} + 0.5 \times 9\tfrac{2}{7} = 14\tfrac{1}{7}$.

Hence δ is optimal.

9. $n = 1$

$$R_1(i) = \{1, 2\}, \qquad 1 \leqslant i \leqslant 3,$$
$$\hat{u}_1(1) = 90, \qquad \hat{\sigma}_1(1) = 1,$$
$$\hat{u}_1(2) = 60, \qquad \hat{\sigma}_1(2) = 2,$$
$$\hat{u}_1(3) = 90, \qquad \hat{\sigma}_1(3) = 1,$$
$$(\rho/(1-\rho))\hat{\alpha}_1 = 810, \qquad (\rho/(1-\rho))\hat{\beta}_1 = 540,$$
$$\hat{u}_1 = (90, 60, 90), \qquad \hat{\sigma}_1 = (1, 2, 1).$$

$n = 2$

Thus we have

$$\underline{u}_2 = \hat{u}_1 + 540e \leqslant v \leqslant \hat{u}_1 + 810e = \bar{u}_2,$$
$$\underline{u}_2 = (630, 600, 630), \bar{u}_2 = (900, 870, 900).$$

$i = 1$

$$[T\bar{u}_2](1) = \text{minimum}\begin{bmatrix} 90 + 0.45 \times 900 + 0.45 \times 870 = 889 \\ 100 + 0.45 \times 900 + 0.45 \times 900 = 910 \end{bmatrix}$$
$$= 889,$$
$$[T^1\underline{u}_2](1) = 90 + 0.45 \times 630 + 0.45 \times 600 = 644,$$
$$[T^2\underline{u}_2](1) = 100 + 0.45 \times 630 + 45 \times 630 = 667,$$
$$[T^1\underline{u}_2](1) < [T\bar{u}_2](1), [T^2\underline{u}_2](1) < [T\bar{u}_2](1).$$

Hence no elimination for $i = 1$, i.e. $R_2(1) = K(1)$.

$i = 2$

$$[T\bar{u}_2](2) = \text{minimum}\begin{bmatrix} 90 + 0.63 \times 900 + 0.27 \times 870 = 889 \\ 60 + 0.63 \times 900 + 0.27 \times 900 = 870 \end{bmatrix}$$
$$= 870,$$
$$[T^1\underline{u}_2](2) = 90 + 0.63 \times 630 + 0.27 \times 600 = 650,$$
$$[T^2\underline{u}_2](2) = 60 + 0.63 \times 630 + 0.27 \times 630 = 628,$$
$$[T^1\underline{u}_2](2) < [T\bar{u}_2](2), [T^2\underline{u}_2](2) < [T\bar{u}_2](2).$$

Hence no elimination for $i = 2$, i.e. $R_2(2) = K(2)$.

$i = 3$

$$[T\bar{u}_2](3) = \text{minimum}\begin{bmatrix} 90 + 0.27 \times 900 + 0.63 \times 870 = 880 \\ 140 + 0.27 \times 900 + 0.63 \times 900 = 960 \end{bmatrix}$$
$$= 880,$$

$[T^1\underline{u}_2](3) = 90 + 0.27 \times 630 + 0.63 \times 600 = 638,$

$[T^2\underline{u}_2](3) = 140 + 0.27 \times 630 + 0.63 \times 630 = 707,$

$[T^1\underline{u}_2](3) < [T\bar{u}_2](3), [T^2\underline{u}_2](3) < [T\bar{u}_2](3).$

Hence no elimination for $i = 3$, i.e. $R_2(3) = K(3)$. Hence $\hat{u}_n = u_n$, $n = 1, 2$.

CHAPTER 4

1. (i) *LP1*

$$\underset{u}{\text{minimise}} \; [\lambda_1 u(1) + \lambda_2 u(2) + \lambda_3 u(3)] \tag{1}$$

subject to

$$u(1) \geqslant 4 + \tfrac{1}{4}u(2) + \tfrac{1}{2}u(3), \tag{2}$$

$$u(1) \geqslant 4 + \tfrac{3}{8}u(2) + \tfrac{3}{8}u(3), \tag{3}$$

$$u(2) \geqslant 9 + \tfrac{3}{16}u(1) + \tfrac{9}{16}u(3), \tag{4}$$

$$u(2) \geqslant 10 + \tfrac{1}{2}u(1) + \tfrac{1}{4}u(3), \tag{5}$$

$$u(3) \geqslant 4 + \tfrac{1}{2}u(1) + \tfrac{1}{4}u(2), \tag{6}$$

$$u(3) \geqslant 3 + \tfrac{3}{8}u(1) + \tfrac{3}{8}u(2). \tag{7}$$

(ii) *DLP1*

$$\underset{x}{\text{maximise}} \; [4x_1^1 + 4x_1^2 + 9x_2^1 + 10x_2^2 + 4x_3^1 + 3x_3^2] \tag{8}$$

subject to

$$x_1^1 + x_1^2 - \tfrac{3}{16}x_2^1 - \tfrac{1}{2}x_2^2 - \tfrac{1}{2}x_3^1 - \tfrac{3}{8}x_3^2 = \lambda_1, \tag{9}$$

$$x_2^1 + x_2^2 - \tfrac{1}{4}x_1^1 - \tfrac{3}{8}x_1^2 - \tfrac{1}{4}x_3^1 - \tfrac{3}{8}x_3^2 = \lambda_2, \tag{10}$$

$$x_3^1 + x_3^2 - \tfrac{1}{2}x_1^1 - \tfrac{3}{8}x_1^2 - \tfrac{9}{16}x_2^1 - \tfrac{1}{4}x_2^2 = \lambda_3, \tag{11}$$

$$x_1^1 \geqslant 0, x_1^2 \geqslant 0, x_2^1 \geqslant 0, x_2^2 \geqslant 0, x_3^1 \geqslant 0, x_3^2 \geqslant 0, \tag{12}$$

$$x_1^1 + x_1^2 + x_2^1 + x_2^2 + x_3^1 + x_3^2 = 1. \tag{13}$$

(iii) *LP1* Use Result 4.3. The decision rule given corresponds to equalities in (3), (5), (6). So we choose, as basic variables, $\{u(1), u(2), u(3)\}$ and the slack variables $\{s_1^1, s_2^1, s_3^2\}$ in (2), (4), (7).

Putting in all the slack variables $\{s_i^k\}$ and solving we obtain

$$u(1) = 22.00 + 1\tfrac{3}{5}s_1^2 + s_2^2 + s_3^1,$$
$$u(2) = 26.40 + \tfrac{4}{5}s_1^2 + 1\tfrac{11}{15}s_2^2 + \tfrac{14}{15}s_3^1,$$
$$u(3) = 21.60 + \tfrac{4}{5}s_1^2 + \tfrac{14}{15}s_2^2 + 1\tfrac{11}{15}s_3^1.$$

We do not need to find expressions for the other basic variables s_1^1, s_2^1, s_3^2 since the coefficients of s_1^2, s_2^2, s_3^1 in the above are all positive, showing that $\{u(1), u(2), u(3)\}$ are simultaneously optimal. The objective function value is

$$22.00\lambda_1 + 26.40\lambda_2 + 21.60\lambda_3.$$

With λ_i being the probability of initially being in state i, $\forall\, i \in I$, this is the expected total discounted reward for the policy $\pi = (\delta)^\infty$ and for the specified $\{\lambda_i\}$.

DLP1 Use Result 4.4. The decision rule given corresponds to $x_1^2 > 0$, $x_2^2 > 0$, $x_3^1 > 0$, $x_i^k = 0$ for all other (i, k). Thus $\{x_1^2, x_2^2, x_3^1\}$ are the basic variables. Solving (9)–(11) we obtain

$$x_1^2 = 2\lambda_1 + \tfrac{20}{15}\lambda_2 + \tfrac{20}{15}\lambda_3 - x_1^1 - \tfrac{5}{24}x_2^1 - \tfrac{1}{2}x_3^2,$$
$$x_2^2 = \lambda_1 + \tfrac{26}{15}\lambda_2 + \tfrac{14}{15}\lambda_3 - \tfrac{1}{10}x_1^1 - \tfrac{49}{48}x_2^1 + \tfrac{11}{20}x_3^2,$$
$$x_3^1 = \lambda_1 + \tfrac{14}{15}\lambda_2 + \tfrac{26}{15}\lambda_3 + \tfrac{1}{10}x_1^1 + \tfrac{11}{48}x_2^1 - \tfrac{121}{120}x_3^2.$$

The objective function value becomes

$$22\lambda_1 + 26.40\lambda_2 + 21.60\lambda_3 - \tfrac{3}{5}x_1^1 - \tfrac{9}{8}x_2^1 - \tfrac{9}{20}x_3^2.$$

Since the coefficients of x_1^1, x_2^1, x_3^2 are negative, δ is an optimal decision rule.

2. (a) *Inventory* (see (4.26)–(4.28)). Here x_i^k is the infinite horizon discounted probability of having stock level i and topping it up to k if we begin with prior state probabilities $\{\lambda_i\}$. We can set $\lambda_i = 1$, $\lambda_j = 0$, $j \neq i$, if we begin in state i.

DLP1

$$\underset{x}{\text{maximise}} \left[\sum_{0 \leqslant i \leqslant m, 1 \leqslant k \leqslant m} r_i^k x_i^k \right]$$

where

$$r_i^k = -\left(c(k-i) + \sum_{s>k} q(s)l(s-k) + \tfrac{1}{2}a\left(k + \sum_{s<k} q(s)(k-s)\right)\right)$$

subject to

$$\underline{1 \leqslant i \leqslant m} \qquad \sum_{i \leqslant k \leqslant m} x_i^k - \rho \sum_{0 \leqslant j \leqslant m, j \leqslant k \leqslant m} q(k-i)x_j^k = \lambda_i,$$

$$\underline{i = 0} \qquad \sum_{0 \leqslant k \leqslant m} x_0^k - \rho \sum_{0 \leqslant j \leqslant m, j \leqslant k \leqslant m, s \geqslant k} q(s)x_j^k = \lambda_0,$$

$$x_i^k \geqslant 0, \qquad 0 \leqslant i \leqslant m, \qquad i \leqslant k \leqslant m.$$

(b) *Queuing* (see (4.64)–(4.68)). Here x_i^k is the infinite horizon limiting average probability of there being i people in the system and $(i-k)$ of these sent elsewhere for service. We will assume that the transition probability structure is uni-chain.

DLP2

$$\text{maximise}\left[-\sum_{0 \leqslant i \leqslant m, 0 \leqslant k \leqslant i} (c(i-k) + wk)x_i^k\right]$$

subject to

$$\underline{1 \leqslant i < m}$$

$$\sum_{0 \leqslant k \leqslant i} x_i^k - p(1-q) \sum_{i-1 \leqslant j \leqslant m} x_j^{j-1} - (1-p)q \sum_{i+1 \leqslant j \leqslant m} x_j^{j+1}$$

$$-(pq + (1-p)(1-q)) \sum_{i \leqslant j \leqslant m} x_j^i = 0,$$

$$\underline{i = 0} \quad x_0^0 - (1-p)q \sum_{1 \leqslant j \leqslant m} x_j^1 - (1-p)(1-q) \sum_{0 \leqslant j \leqslant m} x_j^0 = 0,$$

$$\underline{i = m}$$

$$\sum_{0 \leqslant k \leqslant m} x_m^k = p(1-q) \sum_{m-1 \leqslant j \leqslant m} x_j^{m-1} - (1-p)(1-q)x_m^m = 0,$$

$$\sum_{i=1}^{m} \sum_{k=0}^{i} x_i^k = 1,$$

$$x_i^k \geqslant 0, \quad 0 \leqslant i \leqslant m, \qquad 0 \leqslant k \leqslant i.$$

(c) *Defective product* (see (4.87) and (4.88), (4.90)). Here x_i^k is the probability of having a deficit i and producing k at any time up to the termination of the process. We let $i \leqslant m$, $k \leqslant m$.

DLP3

$$\text{maximise}\left[-\sum_{0 \leqslant i \leqslant m, i \leqslant k \leqslant m} (a + bk)x_i^k\right]$$

subject to

$$\underline{i \neq 0} \qquad \sum_{i \leqslant k \leqslant m} x_i^k - \sum_{i \leqslant j \leqslant m, j \leqslant k \leqslant m} p(k, j - i)x_j^k = \lambda_i,$$

$$x_i^k \geqslant 0, \quad 0 \leqslant i \leqslant m, \qquad i \leqslant k \leqslant m.$$

CHAPTER 5

1. (i) Refer to (5.10)–(5.14). The equations for general ρ are as follows:

$u(i) =$ maximum

$$\times \begin{bmatrix} k = 1\colon\ 6 + (0.2\rho + 0.3\rho^2)u(1) + (0.3\rho + 0.2\rho^2)u(2) \\ k = 2\colon\ 4 + (0.6\rho + 0.2\rho^2)u(1) + (0.1\rho + 0.1\rho^2)u(2) \end{bmatrix},$$

$u(2) =$ maximum

$$\times \begin{bmatrix} k = 1\colon\ -3 + (0.3\rho + 0.1\rho^2)u(1) + (0.2\rho + 0.4\rho^2)u(2) \\ k = 2\colon\ -5 + (0.4\rho + 0.3\rho^2)u(1) + (0.1\rho + 0.2\rho^2)u(2) \end{bmatrix}.$$

(ii) The equations for $\rho = 0.9$ are as follows:

$$u(1) = \text{maximum}\begin{bmatrix} k = 1\colon\ 6 + 0.423u(1) + 0.439u(2) \\ k = 2\colon\ 4 + 0.702u(1) + 0.171u(2) \end{bmatrix},$$

$$u(2) = \text{maximum}\begin{bmatrix} k = 1\colon\ -3 + 0.351u(1) + 0.508u(2) \\ k = 2\colon\ -5 + 0.603u(1) + 0.252u(2) \end{bmatrix}.$$

For $\pi = (\delta)^\infty$, $\delta = (2, 2)$ we solve the following equations:

$$v^\pi(1) = 4 + 0.702v^\pi(1) + 0.171v^\pi(2),$$
$$v^\pi(2) = -5 + 0.603v^\pi(1) + 0.252v^\pi(2)$$

to give

$$v^{\pi} = (17.9, 7.9).$$

We now check for optimality by using Result 5.1 with $u = v^{\pi}$

$$\underline{i = 1} \qquad \text{maximum}\begin{bmatrix} k = 1\text{: } 6 + 0.423 \times 17.9 + 0.439 \times 7.9 \\ k = 2\text{: } 17.9 \end{bmatrix} = 17.9,$$

$$\underline{i = 2} \qquad \text{maximum}\begin{bmatrix} k = 1\text{: } -3 + 0.351 \times 17.9 + 0.508 \times 7.9 \\ k = 2\text{: } 7.9 \end{bmatrix} = 7.9.$$

Thus π is optimal. It is also uniquely optimal since $\delta = (2, 2)$ is uniquely optimal in the above.

2. We will use optimality equations (5.29) and (5.30). Equations (5.27) and (5.28) may equally well be used. The equations take the following form:

$$u(1) = \text{maximum}\begin{bmatrix} k = 1\text{: } 6 - 1.5h + 0.5u(1) + 0.5u(2) \\ k = 2\text{: } 4 - 1.3h + 0.8u(1) + 0.2u(2) \end{bmatrix},$$

$$u(2) = \text{maximum}\begin{bmatrix} k = 1\text{: } -3 - 1.5h + 0.4u(1) + 0.6u(2) \\ k = 2\text{: } -5 - 1.5h + 0.7u(1) + 0.3u(2) \end{bmatrix},$$

$$u(2) = 0.$$

For the policy π we need to solve the equations

$$w^{\pi}(1) = 4 - 1.3g^{\pi} + 0.8w^{\pi}(1),$$
$$0 = -5 - 1.5g^{\pi} + 0.7w^{\pi}(1).$$

Thus

$$w^{\pi}(1) = 10.3, \qquad w^{\pi}(2) = 0, \qquad g^{\pi} = 1.48.$$

We now use Result 5.4 with $u = w^{\pi}$, $h = h^{\pi}$.

$$\underline{i = 1} \qquad \text{maximum}\begin{bmatrix} k = 1\text{: } 6 - 1.5 \times 1.48 + 0.5 \times 10.3 \\ k = 2\text{: } 10.3 \end{bmatrix} = 10.3,$$

$$\underline{i = 2} \qquad \text{maximum}\begin{bmatrix} k = 1\text{: } -3 - 1.5 \times 1.48 + 0.4 \times 10.3 \\ k = 2\text{: } 0 \end{bmatrix} = 0.$$

Thus π is optimal.

3. Let (u, h, δ) be any solution to (5.27) and (5.28). Then for each $i \in I$

$$h \geqslant \frac{\left(r_i^k + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k u(j) - u(i)\right)}{\left(\sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k \gamma\right)}, \quad \forall\, k \in K(i) \tag{1}$$

and

$$h = \frac{r_i^k + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k u(j) - u(i)}{\left(\sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k \gamma\right)} \tag{2}$$

for some $k \in K(i)$. From (1) by rearranging the terms, we obtain for all $i \in I$

$$u(i) \geqslant r_i^k - h\left(\sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k \gamma\right) + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k u(j), \quad \forall\, k \in K(i) \tag{3}$$

with equality in (3) for some $k \in K(i)$ coming from (2). Thus

$$u(i) = \underset{k \in K(i)}{\text{maximum}} \times \left[r_i^k - h\left(\sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k \gamma\right) + \sum_{j \in I, 1 \leqslant \gamma \leqslant L} p_{ij\gamma}^k u(j)\right], \quad \forall\, i \in I. \tag{4}$$

Equation (4) is the same as (5.29). Thus together with $u(m) = 0$ in (5.28), (u, h, δ) solves (5.29) and (5.30). The converse analysis also applies.

4. The states of the system are $i = 0$ (failure) $1, 2, 3$ (performance levels). We will use optimality equations (5.29) and (5.30).

$\underline{i = 0}$

$$\begin{aligned} u(0) = \underset{k \geqslant 1}{\text{maximum}} \Bigg[& r(3) + \sum_{0 \leqslant j \leqslant 2, 2 \leqslant \gamma \leqslant k} p(3, j, \gamma - 1) r(j) - c(0) \\ & - a(p(3,1,k) + p(3,2,k)) \\ & - h\left(\sum_{1 \leqslant \gamma \leqslant k} p(3,0,\gamma)\gamma + k(p(3,1,k) + p(3,2,k))\right) \\ & + \left(\sum_{1 \leqslant \gamma \leqslant k} p(3,0,\gamma)\right) u(0) + p(3,1,k)u(1) + p(3,2,k)u(2) \Bigg]. \end{aligned}$$

$\underline{3 \geqslant i \geqslant 1}$

$$u(i) = \underset{k \geqslant 1}{\text{maximum}} \Bigg[r(i) + \sum_{0 \leqslant j \leqslant i-1, 2 \leqslant \gamma \leqslant k} p(i, j, \gamma - 1) r(j) - a(p(i, 1, k) + p(i, 2, k)) - h\Big(\sum_{1 \leqslant \gamma \leqslant k} p(i, 0, \gamma)\gamma + k(p(i, 1, k) + p(i, 2, k)) \Big) + \Big(\sum_{1 \leqslant \gamma \leqslant k} p(i, 0, \gamma) \Big) u(0) + p(i, 1, k) u(1) + p(i, 2, k) u(2) \Bigg].$$

CHAPTER 6

1. (i) This is similar to the sequential sampling example of the text. The infinite horizon equation is $u = Nu$ with N given by (6.37) and (6.38). As with the sequential sampling problem, the states are the vectors μ of probabilities over $\Theta = \{1, 2\}$.

We have only two primitive states $\theta = 1$ or 2. Thus μ can be replaced by the scalar p in the equations, dropping the $\{i, j\}$ in the various terms.

$$p^k_{\theta d} = 0 \quad \text{if } k \in \{1, 2\}, \quad \forall\, \theta \in \Theta, \qquad d \in D,$$
$$p^3_{11} = \tfrac{1}{3}, \qquad p^3_{12} = \tfrac{2}{3}, \qquad p^3_{21} = \tfrac{2}{3}, \qquad p^3_{22} = \tfrac{1}{3},$$
$$[Q^{31}\mu]_1 = p/(2 - p), \qquad [Q^{32}\mu]_1 = 2p/(1 + p).$$

Then the infinite horizon optimality equation is retained in minimisation form

$$u(p) = \text{minimum} \begin{bmatrix} k = 1\colon\ 20(1 - p) \\ k = 2\colon\ 20p \\ k = 3\colon\ 1 + ((2 - p)/3) u(p/(2 - p)) \\ \qquad\qquad + ((1 + p)/3) u(2p/(1 + p)) \end{bmatrix},$$
$$\forall\ 0 \leqslant p \leqslant 1.$$

The unique solution to this equation gives the value function v.

(ii) Let $v_n(p)$ be the minimal expected sum of experimentation costs and losses if we allow at most n experiments. Then v_n is a unique solution to the following optimality equation for $n = 0, 1$:

$$u_0(p) = \text{minimum} \begin{bmatrix} k = 1\colon\ 20(1 - p) \\ k = 2\colon\ 20p \end{bmatrix}, \quad \forall\ 0 \leqslant p \leqslant 1.$$

We have

$$k = 2 \quad \text{if } p \leqslant \tfrac{1}{2},$$
$$k = 1 \quad \text{if } p \geqslant \tfrac{1}{2}.$$

$$u_1(p) = \text{minimum}\begin{bmatrix} k=1\colon\ 20(1-p) \\ k=2\colon\ 20p \\ k=3\colon\ 1 + ((2-p)/3)u_0(p/(2-p)) \\ \qquad\qquad + ((1+p)/3)u_0(2p/(1+p)) \end{bmatrix}$$

$$= \text{minimum} \times \begin{bmatrix} k=1\colon\ 20(1-p) \\ k=2\colon\ 20p \\ k=3\colon\ 1 + ((2-p)/3)\text{minimum}\left[\dfrac{40(1-p)}{(2-p)}, \dfrac{20p}{(2-p)}\right] \\ \qquad + ((1+p)/3)\text{minimum}\left[\dfrac{20(1-p)}{(1+p)}, \dfrac{40p}{(1+p)}\right] \end{bmatrix}$$

$$= \text{minimum}\left[20(1-p), 20p, \text{minimum}\left[\frac{43-40p}{3}, \frac{3+20p}{3}\right] + \text{minimum}\left[\frac{20-30p}{3}, \frac{40p}{3}\right]\right]$$

$$= \text{minimum}\left[20 - 20p, 20p, \frac{63-60p}{3}, \frac{23}{3}, \frac{43}{3}, \frac{3+60p}{3}\right],$$

$$\forall\ 0 \leqslant p \leqslant 1.$$

The optimal decision rule for $n = 1$ is given by $k = 3$, $0 \leqslant p \leqslant 1$.

2. (i) This is similar to Exercise 1. We have

$$p^k_{\theta d} = 0 \quad \text{if } k \in \{1,2\}, \quad \forall\ \theta \in \Theta, \quad d \in D,$$
$$p^3_{11} = 0, \qquad p^3_{12} = 1, \qquad p^3_{21} = \tfrac{2}{3}, \qquad p^3_{22} = \tfrac{1}{3},$$
$$[Q^{31}\mu]_1 = 0, \qquad [Q^{32}\mu]_1 = 3p/(1+2p).$$

The infinite horizon optimality equation is, retained in minimisation form

$$u(p) = \text{minimum}\begin{bmatrix} k=1:\ 8p \\ k=2:\ 8(1-p) \\ k=3:\ 1+\frac{2}{3}(1-p)u(0) \\ \qquad\qquad +((1+2p)/3)u(3p/(1+2p)) \end{bmatrix},$$

$$\forall\ 0 \leqslant p \leqslant 1.$$

Clearly $u(0) = 0$ and this simplifies the equation.

(ii)

$$u_0(p) = \text{minimum}\,[8p, 8(1-p)], \quad \forall\ 0 \leqslant p \leqslant 1.$$

We have

$$k = 1 \quad \text{if } p \leqslant \tfrac{1}{2},$$
$$k = 2 \quad \text{if } p \geqslant \tfrac{1}{2},$$

$$u_1(p) = \text{minimum}\left[8p, 8(1-p), 1+((1+2p)/3)u_0\left(\frac{3p}{1+2p}\right)\right]$$

$$= \text{minimum}\left[8p, 8(1-p), 1+((1+2p)/3)\text{minimum} \times \left[\frac{24p}{1+2p}, \frac{8(1-p)}{(1+2p)}\right]\right]$$

$$= \text{minimum}\,[8p, 8(1-p), 1+\text{minimum}\,[8p, 8(1-p)/3]]$$
$$= \text{minimum}\,[8p, 8(1-p), 1+8p, (11-8p)/3]$$
$$= \tfrac{1}{3}\ \text{minimum}\,[24p, 24(1-p), 3+24p, 11-8p].$$

The optimal decision rules for $n = 1$ are given by

$$k = 1 \quad \text{if } 0 \leqslant p \leqslant \tfrac{11}{32},$$
$$k = 3 \quad \text{if } \tfrac{11}{32} \leqslant p \leqslant \tfrac{13}{16},$$
$$k = 2 \quad \text{if } \tfrac{13}{16} \leqslant p \leqslant 1.$$

Optional k values occur at $p = \tfrac{11}{32}, \tfrac{13}{16}$.

CHAPTER 8

1. *States* (at the beginning of a year)
   $i = 1$: no accumulated injuries;
   $i = 2$: one accumulated injury;
   $i = 3$: two accumulated injuries;
   $i = 3$ is an absorbing state.
   Actions $k \in \{0, 1, 2, 3\}$, with k fights planned in the year, and $k = 0$ is retirement.

   Optimality equation

$$u(1) = \text{maximum} \begin{bmatrix} k = 0\text{: } 0 \\ k = 1\text{: } 5000 + \frac{3}{4}u(1) + \frac{1}{4}u(2) \\ k = 2\text{: } 5000 + \frac{3}{4} \times 5000 + \frac{9}{16}u(1) + \frac{7}{16}u(2) \\ k = 3\text{: } 5000 + \frac{3}{4} \times 5000 + \frac{9}{16} \times 5000 \\ \qquad\qquad + \frac{27}{64}u(1) + \frac{37}{64}u(2) \end{bmatrix},$$

$$u(2) = \text{maximum} \begin{bmatrix} k = 0\text{: } 0 \\ k = 1\text{: } 5000 + \frac{1}{2}u(2) + \frac{1}{2}u(3) \\ k = 2\text{: } 5000 + \frac{1}{2} \times 5000 + \frac{1}{4}u(2) + \frac{3}{4}u(3) \\ k = 3\text{: } 5000 + \frac{1}{2} \times 5000 + \frac{1}{4} \times 5000 \\ \qquad\qquad + \frac{1}{8}u(2) + \frac{7}{8}u(3) \end{bmatrix},$$

   $u(3) = -20\,000.$

2. *States*
   i_1: the number of type 1 customers in the system;
   i_2: the number of type 2 customers in the system.
   Actions
   $k = 1$: begin service on type 1 customer or continue service if current customer being served is type 1;
   $k = 2$: begin service on type 2 customer or continue service if current customer being served is type 2.

   Optimality equation

   Whatever policy is adopted there is always a positive probability that the system will eventually become empty. Clearly we will always serve if the system is not empty. Hence the problem is uni-chain.

$\underline{i_1 = 0, i_2 > 0}$

$$u(0, i_1) + h = i_2 + p_2(1 - q_2)p_1u(1, i_2 + 1) + p_2(1 - q_2)(1 - p_1)u(0, i_2 + 1) + (p_2q_2 + (1 - p_2)(1 - q_2))p_1u(1, i_2) + (p_2q_2 + (1 - p_2)(1 - q_2))(1 - p_1)u(0, i_2),$$

$\underline{i_1 = i_2 = 0}$

$$u(0, 0) + h = p_1p_2u(1, 1) + p_1(1 - p_2)u(1, 0) + p_2(1 - p_1)u(0, 1) + (1 - p_1)(1 - p_2)u(0, 0).$$

For $i_1 + i_2 > m$ we have no equations. Also for the transformed states on the right-hand side of the equation, set $u(i_1, i_2) = 0$ when $i_1 + i_2 > m$.

$\underline{i_1 + i_2 \leqslant m}$

$\underline{i_1 > 0, i_2 > 0}$

$$u(i_1, i_2) + h = \text{minimum} \begin{bmatrix} k = 1: i_1 + i_2 + p_1(1 - q_1)p_2u(i_1 + 1, i_2 + 1) \\ \quad + p_1(1 - q_1)(1 - p_2)u(i_1 + 1, i_2) \\ \quad + (p_1q_1 + (1 - p_1)(1 - q_1))p_2u(i_1, i_2 + 1) \\ \quad + (p_1q_1 + (1 - p_1)(1 - q_1))(1 - p_2)u(i_1, i_2) \\ k = 2: i_1 + i_2 + p_2(1 - q_2)p_1u(i_1 + 1, i_2 + 1) \\ \quad + p_2(1 - q_2)(1 - p_1)u(i_1, i_2 + 1) \\ \quad + (p_2q_2 + (1 - p_2)(1 - q_2))p_1u(i_1 + 1, i_2) \\ \quad + (p_2q_2 + (1 - p_2)(1 - q_2))(1 - p_1)u(i_1, i_2) \end{bmatrix},$$

$\underline{i_1 > 0, i_2 = 0}$

$$u(i_1, 0) + h = i_1 + p_1(1 - q_1)p_2u(i_1 + 1, 1) + p_1(1 - q_1)(1 - p_2)u(i_1 + 1, 0) + (p_1q_1 + (1 - p_1)(1 - q_1))p_2u(i_1, 1) + (p_1q_1 + (1 - p_1)(1 - q_1))(1 - p_2)u(i_1, 0).$$

3. *States i*: the stock level prior to a decision, allowing $i < 0$ for a shortage situation

   Actions

   k: the new level to which the stock level is raised at a decision epoch.

Optimality equation

Let $q_\gamma(S)$ be the probability that the total demand is equal to S over γ time units.

$$\underline{L \geqslant i \geqslant 0} \qquad u(i) = \sum_{1 \leqslant \gamma \leqslant L, i > S, s \geqslant i - S} q_{\gamma-1}(S)q(s)(l(S + s - i) + \tfrac{1}{2} ai\gamma - h\gamma + u(i - S - s)),$$

$$\underline{-\bar{s} \leqslant i < 0}$$

$$u(i) = \underset{k \geqslant 0}{\text{minimum}} \left[\sum_{1 \leqslant \gamma \leqslant L, k > S, s \geqslant k - S} q_{\gamma-1}(S)q(s)(c(k - i) + l(S + s - k) + \tfrac{1}{2} ak\gamma - h\gamma + u(k - S - s)) \right].$$

For $i < 0$, when c is linear the right-hand side of the optimality equation is of the form

$$-ci + G(k).$$

Thus, to minimise this is equivalent to minimising $G(k)$ independently of i.

4. *States*
   $i = 1$: not audited in the previous year;
   $i = 2$: audited in the previous year and declaration correct;
   $i = 3$: audited in the previous year and declaration incorrect.
   Actions
   $k = 1$: declare;
   $k = 2$: do not declare.

Optimality equation

$$u(1) = \text{minimum} \begin{bmatrix} k = 1: & 90 + 0.9(0.5u(1) + 0.5u(2)) \\ k = 2: & (0.5 \times 200 + 0.5 \times 0) + 0.9(0.5u(1) + 0.5u(3)) \end{bmatrix}$$

$$= \text{minimum} \begin{bmatrix} k = 1: & 90 + 0.45u(1) + 0.45u(2) \\ k = 2: & 100 + 0.45u(1) + 0.45u(3) \end{bmatrix},$$

$$u(2) = \text{minimum}\begin{bmatrix} k=1:\ 90 + 0.9(0.7u(1) + 0.3u(2)) \\ k=2:\ (0.3 \times 200 + 0.7 \times 0) + 0.9(0.7u(1) + 0.2u(3)) \end{bmatrix}$$

$$= \text{minimum}\begin{bmatrix} k=1:\ 90 + 0.63u(1) + 0.27u(2) \\ k=2:\ 60 + 0.63u(1) + 0.27u(3) \end{bmatrix},$$

$$u(3) = \text{minimum}\begin{bmatrix} k=1:\ 90 + 0.9(0.3u(1) + 0.7u(2)) \\ k=2:\ (0.7 \times 200 + 0.3 \times 0) + 0.9(0.3u(1) + 0.7u(3)) \end{bmatrix}$$

$$= \text{minimum}\begin{bmatrix} k=1:\ 90 + 0.27u(1) + 0.63u(2) \\ k=2:\ 140 + 0.27u(1) + 0.63u(2) \end{bmatrix}.$$

5. *States*
p: the subjective probability of being classified as being reliable.

Actions
$k = 1$: declare;
$k = 2$: do not declare,

Optimality equation

$$u(p) = \text{minimum}$$

$$\times \begin{bmatrix} k=1:\ 90 + 0.9\left(p\left(0.3u\left(\frac{1+p}{2}\right) + 0.7u\left(1 + \frac{4p}{5}\right)\right)\right. \\ \qquad \left. + (1-p)\left(0.7u\left(\frac{1+p}{2}\right) + 0.3u\left(1 - \frac{4p}{5}\right)\right)\right) \\ k=2:\ 200(p \times 0.3 + (1-p) \times 0.7) + 0.9\left(p\left(0.3u(p/2)\right.\right. \\ \qquad \left. + 0.7u\left(\frac{1+4p}{5}\right)\right) + (1-p)\left(0.7u(p/2)\right. \\ \qquad \left.\left. + 0.3u\left(\frac{1+4p}{5}\right)\right)\right) \end{bmatrix}$$

$$= \text{minimum}$$

$$\begin{bmatrix} k=1\text{: } 90+(0.63-0.36p)u\left(\dfrac{1+p}{2}\right)+(0.27+0.36p)u\left(\dfrac{1+4p}{5}\right) \\ k=2\text{: } 140-80p+(0.63-0.36p)u(p/2)+(0.27+0.36p)u\left(\dfrac{1+4p}{5}\right) \end{bmatrix},$$

$$\forall\ 0 \leqslant p \leqslant 1.$$

Index