CHAPTER 25

Markovian Decision Process

Chapter Guide. This chapter applies dynamic programming to the solution of a stochastic decision process with a finite number of states. The transition probabilities between the states are described by a Markov chain. The reward structure of the process is a matrix representing the revenue (or cost) associated with movement from one state to another. Both the transition and revenue matrices depend on the decision alternatives available to the decision maker. The objective is to determine the optimal policy that maximizes the expected revenue over a finite or infinite number of stages. The prerequisites for this chapter are basic knowledge of Markov chains (Chapter 17), probabilistic dynamic programming (Chapter 24), and linear programming (Chapter 2).

This chapter includes 6 solved examples and 14 end-of-section problems.

25.1 SCOPE OF THE MARKOVIAN DECISION PROBLEM

We use the gardener problem (Example 17.1-1) to present the details of the Markovian decision process. The idea of the example can be adapted to represent important applications in the areas of inventory, replacement, cash flow management, and regulation of water reservoir capacity.

The transition matrices, $\mathbf{P}^1$ and $\mathbf{P}^2$, associated with the no-fertilizer and fertilizer cases are repeated here for convenience. States 1, 2, and 3 correspond, respectively, to good, fair, and poor soil conditions.

$$\mathbf{P}^1 = \begin{pmatrix} .2 & .5 & .3 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} .30 & .60 & .10 \\ .10 & .60 & .30 \\ .05 & .40 & .55 \end{pmatrix}$$

To put the decision problem in perspective, the gardener associates a return function (or a reward structure) with the transition from one state to another. The return function expresses the gain or loss during a 1-year period, depending on the states between which the transition is made. Because the gardener has the option of using or not using fertilizer, gain and loss vary depending on the decision made. The matrices $\mathbf{R}^1$ and $\mathbf{R}^2$ summarize the return functions in hundreds of dollars associated with matrices $\mathbf{P}^1$ and $\mathbf{P}^2$, respectively.

$$\mathbf{R}^1 = \|r_{ij}^1\| = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix} \end{array}$$

$$\mathbf{R}^2 = \|r_{ij}^2\| = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix} \end{array}$$

The elements r_{ij}^2 of $\mathbf{R}^2$ consider the cost of applying fertilizer. For example, if the soil condition was fair last year (state 2) and becomes poor this year (state 3), its gain will be $r_{23}^2 = 0$ compared with $r_{23}^1 = 1$ when no fertilizer is used. Thus, $\mathbf{R}$ gives the net reward after the cost of the fertilizer is factored in.

What kind of a decision problem does the gardener have? First, we must know whether the gardening activity will continue for a limited number of years or indefinitely. These situations are referred to as **finite-stage** and **infinite-stage** decision problems. In both cases, the gardener uses the outcome of the chemical tests (state of the system) to determine the *best* course of action (fertilize or do not fertilize) that maximizes expected revenue.

The gardener may also be interested in evaluating the expected revenue resulting from a prespecified course of action for a given state of the system. For example, fertilizer may be applied whenever the soil condition is poor (state 3). The decision-making process in this case is said to be represented by a **stationary policy**.

Each stationary policy is associated with different transition and return matrices, which are constructed from the matrices $\mathbf{P}^1$, $\mathbf{P}^2$, $\mathbf{R}^1$, and $\mathbf{R}^2$. For example, for the stationary policy calling for applying fertilizer only when the soil condition is poor (state 3), the resulting transition and return matrices are given as

$$\mathbf{P} = \begin{pmatrix} .20 & .50 & .30 \\ .00 & .50 & .50 \\ .05 & .40 & .55 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 6 & 3 & -2 \end{pmatrix}$$

These matrices differ from $\mathbf{P}^1$ and $\mathbf{R}^1$ in the third rows only, which are taken directly from $\mathbf{P}^2$ and $\mathbf{R}^2$, the matrices associated with applying fertilizer.

PROBLEM SET 25.1A

1. In the gardener model, identify the matrices $\mathbf{P}$ and $\mathbf{R}$ associated with the stationary policy that calls for using fertilizer whenever the soil condition is fair or poor.

*__2.__ Identify all the stationary policies for the gardener model.

25.2 FINITE-STAGE DYNAMIC PROGRAMMING MODEL

Suppose that the gardener plans to "retire" from gardening in N years. We are interested in determining the optimal course of action for each year (to fertilize or not to fertilize) that will return the highest expected revenue at the end of N years.

Let $k = 1$ and 2 represent the two courses of action (alternatives) available to the gardener. The matrices $\mathbf{P}^k$ and $\mathbf{R}^k$ representing the transition probabilities and reward function for alternative k were given in Section 25.1 and are summarized here for convenience.

$$\mathbf{P}^1 = \|r_{ij}^1\| = \begin{pmatrix} .2 & .5 & .3 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{R}^1 = \|r_{ij}^1\| = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{P}^2 = \|p_{ij}^2\| = \begin{pmatrix} .30 & .60 & .10 \\ .10 & .60 & .30 \\ .05 & .40 & .55 \end{pmatrix}, \mathbf{R}^2 = \|r_{ij}^2\| = \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix}$$

The gardener problem is expressed as a finite-stage dynamic programming (DP) model as follows. For the sake of generalization, define

m = Number of states at each stage (year) (= 3 in the gardener problem)

$f_n(i)$ = Optimal *expected* revenue of stages $n, n+1, \ldots, N$, given that i is the state of the system (soil condition) at the beginning of year n

The *backward* recursive equation relating f_n and f_{n+1} is

$$f_n(i) = \max_k \left\{ \sum_{j=1}^{m} p_{ij}^k [r_{ij}^k + f_{n+1}(j)] \right\}, n = 1, 2, \ldots, N$$

where $f_{N+1}(j) = 0$ for all j.

A justification for the equation is that the cumulative revenue, $r_{ij}^k + f_{n+1}(j)$, resulting from reaching state j at stage $n+1$ from state i at stage n occurs with probability p_{ij}^k. Let

$$v_i^k = \sum_{j=1}^{m} p_{ij}^k r_{ij}^k$$

The DP recursive equation can be written as

$$f_N(i) = \max_k \{v_i^k\}$$

$$f_n(i) = \max_k \left\{ v_i^k + \sum_{j=1}^{m} p_{ij}^k f_{n+1}(j) \right\}, n = 1, 2, \ldots, N-1$$

To illustrate the computation of v_i^k, consider the case in which no fertilizer is used ($k = 1$).

$$v_1^1 = .2 \times 7 + .5 \times 6 + .3 \times 3 = 5.3$$
$$v_2^1 = 0 \times 0 + .5 \times 5 + .5 \times 1 = 3$$
$$v_3^1 = 0 \times 0 + 0 \times 0 + 1 \times -1 = -1$$

Thus, if the soil condition is good, a single transition yields 5.3 for that year; if it is fair, the yield is 3, and if it is poor, the yield is -1.

Example 25.2-1

In this example, we solve the gardener problem using the data summarized in the matrices $\mathbf{P}^1$, $\mathbf{P}^2$, $\mathbf{R}^1$, and $\mathbf{R}^2$, given a horizon of 3 years ($N = 3$).

Because the values of v_i^k will be used repeatedly in the computations, they are summarized here for convenience. Recall that $k = 1$ represents "do not fertilize" and $k = 2$ represents "fertilize."

i	v_i^1	v_i^2
1	5.3	4.7
2	3	3.1
3	−1	.4

Stage 3

	v_i^k		Optimal solution	
i	$k = 1$	$k = 2$	$f_3(i)$	k^*
1	5.3	4.7	5.3	1
2	3	3.1	3.1	2
3	−1	.4	.4	2

Stage 2

	$v_i^k + p_{i1}^k f_3(1) + p_{i2}^k f_3(2) + p_{i3}^k f_3(3)$		Optimal solution	
i	$k = 1$	$k = 2$	$f_2(i)$	k^*
1	$5.3 + .2 \times 5.3 + .5 \times 3.1 + .3 \times .4 = 8.03$	$4.7 + .3 \times 5.3 + .6 \times 3.1 + .1 \times .4 = 8.19$	8.19	2
2	$3 + 0 \times 5.3 + .5 \times 3.1 + .5 \times .4 = 4.75$	$3.1 + .1 \times 5.3 + .6 \times 3.1 + .3 \times .4 = 5.61$	5.61	2
3	$-1 + 0 \times 5.3 + 0 \times 3.1 + 1 \times .4 = -.6$	$.4 + .05 \times 5.3 + .4 \times 3.1 + .55 \times .4 = 2.13$	2.13	2

Stage 1

	$v_i^k + p_{i1}^k f_2(1) + p_{i2}^k f_2(2) + p_{i3}^k f_2(3)$		Optimal solution	
i	$k = 1$	$k = 2$	$f_1(i)$	k^*
1	$5.3 + .2 \times 8.19 + .5 \times 5.61 + .3 \times 2.13 = 10.38$	$4.7 + .3 \times 8.19 + .6 \times 5.61 + .1 \times 2.13 = 10.74$	10.74	2
2	$3 + 0 \times 8.19 + .5 \times 5.61 + .5 \times 2.13 = 6.87$	$3.1 + .1 \times 8.19 + .6 \times 5.61 + .3 \times 2.13 = 7.92$	7.92	2
3	$-1 + 0 \times 8.19 + 0 \times 5.61 + 1 \times 2.13 = 1.13$	$.4 + .05 \times 8.19 + .4 \times 5.61 + .55 \times 2.13 = 4.23$	4.23	2

The optimal solution shows that for years 1 and 2, the gardener should apply fertilizer ($k^* = 2$) regardless of the state of the system (soil condition, as revealed by the chemical tests). In year 3, fertilizer should be applied only if the system is in state 2 or 3 (fair or poor soil condition). The total expected revenues for the three years are $f_1(1) = 10.74$ if the state of the system in year 1 is good, $f_1(2) = 7.92$ if it is fair, and $f_1(3) = 4.23$ if it is poor.

Remarks. The finite-horizon problem can be generalized in two ways. First, the transition probabilities and their return functions need not be the same for all years. Second, a discounting factor can be applied to the expected revenue of the successive stages so that $f_1(i)$ will equal the *present value* of the expected revenues of all the stages.

The first generalization requires the return values r_{ij}^k and transition probabilities p_{ij}^k to be functions of the stage, n, as the following DP recursive equation shows

$$f_N(i) = \max_k \{v_i^{k,N}\}$$

$$f_n(i) = \max_k \left\{ v_i^{k,n} + \sum_{j=1}^{m} p_{ij}^{k,n} f_{n+1}(j) \right\}, n = 1, 2, \ldots, N-1$$

where

$$v_i^{k,n} = \sum_{j=1}^{m} p_{ij}^{k,n} r_{ij}^{k,n}$$

In the second generalization, given α (< 1) is the discount factor per year such that D dollars a year from now have a present value of αD dollars, the new recursive equation becomes

$$f_N(i) = \max_k \{v_i^k\}$$

$$f_n(i) = \max_k \left\{ v_i^k + \alpha \sum_{j=1}^{m} p_{ij}^k f_{n+1}(j) \right\}, n = 1, 2, \ldots, N-1$$

PROBLEM SET 25.2A

*1. A company reviews the state of one of its important products annually and decides whether it is successful (state 1) or unsuccessful (state 2). The company must decide whether or not to advertise the product to further promote sales. The following matrices, $\mathbf{P}^1$ and $\mathbf{P}^2$, provide the transition probabilities with and without advertising during any year. The associated returns are given by the matrices $\mathbf{R}^1$ and $\mathbf{R}^2$. Find the optimal decisions over the next 3 years.

$$\mathbf{P}^1 = \begin{pmatrix} .9 & .1 \\ .6 & .4 \end{pmatrix}, \mathbf{R}^1 = \begin{pmatrix} 2 & -1 \\ 1 & -3 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} .7 & .3 \\ .2 & .8 \end{pmatrix}, \mathbf{R}^2 = \begin{pmatrix} 4 & 1 \\ 2 & -1 \end{pmatrix}$$

2. A company can advertise through radio, TV, or newspaper. The weekly costs of advertising on the three media are estimated at \$200, \$900, and \$300, respectively. The company can classify its sales volume during each week as (1) fair, (2) good, or (3) excellent. A summary of the transition probabilities associated with each advertising medium follows.

Radio

	1	2	3
1	.4	.5	.1
2	.1	.7	.2
3	.1	.2	.7

TV

	1	2	3
1	.7	.2	.1
2	.3	.6	.1
3	.1	.7	.2

Newspaper

	1	2	3
1	.2	.5	.3
2	0	.7	.3
3	0	.2	.8

The corresponding weekly returns (in dollars) are

Radio

400	520	600
300	400	700
200	250	500

TV

1000	1300	1600
800	1000	1700
600	700	1100

Newspaper

400	530	710
350	450	800
250	400	650

Find the optimal advertising policy over the next 3 weeks.

*3. *Inventory Problem*. An appliance store can place orders for refrigerators at the beginning of each month for immediate delivery. A fixed cost of \$100 is incurred every time an order is placed. The storage cost per refrigerator per month is \$5. The penalty for running out of stock is estimated at \$150 per refrigerator per month. The monthly demand is given by the following pdf:

Demand x	0	1	2
$p(x)$	.2	.5	.3

The store's policy is that the maximum stock level should not exceed two refrigerators in any single month. Determine the following:

(a) The transition probabilities for the different decision alternatives of the problem.

(b) The expected inventory cost per month as a function of the state of the system and the decision alternative.

(c) The optimal ordering policy over the next 3 months.

4. Repeat Problem 3 assuming that the pdf of demand over the next quarter changes according to the following table:

Demand	Month		
x	*1*	*2*	*3*
0	.1	.3	.2
1	.4	.5	.4
2	.5	.2	.4

25.3 INFINITE-STAGE MODEL

There are two methods for solving the infinite-stage problem. The first method calls for evaluating *all* possible stationary policies of the decision problem. This is equivalent to an *exhaustive enumeration* process and can be used only if the number of stationary policies is reasonably small. The second method, called **policy iteration**, is generally more efficient because it determines the optimum policy iteratively.

25.3.1 Exhaustive Enumeration Method

Suppose that the decision problem has S stationary policies, and assume that $\mathbf{P}^s$ and $\mathbf{R}^s$ are the (one-step) transition and revenue matrices associated with the policy, $s = 1, 2, \ldots, S$. The steps of the enumeration method are as follows.

Step 1. Compute v_i^s, the expected one-step (one-period) revenue of policy s given state i, $i = 1, 2, \ldots, m$.

Step 2. Compute π_i^s, the long-run stationary probabilities of the transition matrix $\mathbf{P}^s$ associated with policy s. These probabilities, when they exist, are computed from the equations

$$\boldsymbol{\pi}^s \mathbf{P}^s = \boldsymbol{\pi}^s$$

$$\pi_1^s + \pi_2^s + \cdots + \pi_m^s = 1$$

where $\boldsymbol{\pi}^s = (\pi_1^s, \pi_2^s, \cdots, \pi_m^s)$.

Step 3. Determine E^s, the expected revenue of policy s per transition step (period), by using the formula

$$E^s = \sum_{i=1}^{m} \pi_i^s v_i^s$$

Step 4. The optimal policy s^* is determined such that

$$E^{s^*} = \max_s \{E^s\}$$

We illustrate the method by solving the gardener problem for an infinite-period planning horizon.

Example 25.3-1

The gardener problem has a total of eight stationary policies, as the following table shows:

Stationary policy, s	Action
1	Do not fertilize at all.
2	Fertilize regardless of the state.
3	Fertilize if in state 1.
4	Fertilize if in state 2.
5	Fertilize if in state 3.
6	Fertilize if in state 1 or 2.
7	Fertilize if in state 1 or 3.
8	Fertilize if in state 2 or 3.

The matrices $\mathbf{P}^s$ and $\mathbf{R}^s$ for policies 3 through 8 are derived from those of policies 1 and 2 and are given as

$$\mathbf{P}^1 = \begin{pmatrix} .2 & .5 & .3 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}^1 = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} .3 & .6 & .1 \\ .1 & .6 & .3 \\ .05 & .4 & .55 \end{pmatrix}, \quad \mathbf{R}^2 = \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix}$$

$$\mathbf{P}^3 = \begin{pmatrix} .3 & .6 & .1 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}^3 = \begin{pmatrix} 6 & 5 & -1 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{P}^4 = \begin{pmatrix} .2 & .5 & 3 \\ .1 & .6 & .3 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}^4 = \begin{pmatrix} 7 & 6 & 3 \\ 7 & 4 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{P}^5 = \begin{pmatrix} .2 & .5 & .3 \\ 0 & .5 & .5 \\ .05 & .4 & .55 \end{pmatrix}, \quad \mathbf{R}^5 = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 6 & 3 & -2 \end{pmatrix}$$

$$\mathbf{P}^6 = \begin{pmatrix} .3 & .6 & .1 \\ .1 & .6 & .3 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}^6 = \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{P}^7 = \begin{pmatrix} .3 & .6 & .1 \\ 0 & .5 & .5 \\ .05 & .4 & .55 \end{pmatrix}, \quad \mathbf{R}^7 = \begin{pmatrix} 6 & 5 & -1 \\ 0 & 5 & 1 \\ 6 & 3 & -2 \end{pmatrix}$$

$$\mathbf{P}^8 = \begin{pmatrix} .2 & .5 & .3 \\ .1 & .6 & .3 \\ .05 & .4 & .55 \end{pmatrix}, \quad \mathbf{R}^8 = \begin{pmatrix} 7 & 6 & 3 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix}$$

The values of v_i^k can thus be computed as given in the following table.

	v_i^s		
s	$i = 1$	$i = 2$	$i = 3$
1	5.3	3.0	−1.0
2	4.7	3.1	0.4
3	4.7	3.0	−1.0
4	5.3	3.1	−1.0
5	5.3	3.0	0.4
6	4.7	3.1	−1.0
7	4.7	3.0	0.4
8	5.3	3.1	0.4

The computations of the stationary probabilities are achieved by using the equations

$$\boldsymbol{\pi}^s \mathbf{P}^s = \boldsymbol{\pi}^s$$
$$\pi_1 + \pi_2 + \cdots + \pi_m = 1$$

As an illustration, consider $s = 2$. The associated equations are

$$.3\pi_1 + .1\pi_2 + .05\pi_3 = \pi_1$$
$$.6\pi_1 + .6\pi_2 + .4\pi_3 = \pi_2$$
$$.1\pi_1 + .3\pi_2 + .55\pi_3 = \pi_3$$
$$\pi_1 + \pi_2 + \pi_3 = 1$$

(Notice that one of the first three equations is redundant.) The solution yields

$$\pi_1^2 = \tfrac{6}{59}, \pi_2^2 = \tfrac{31}{59}, \pi_3^2 = \tfrac{22}{59}$$

In this case, the expected yearly revenue is

$$E^2 = \pi_1^2 v_1^2 + \pi_2^2 v_2^2 + \pi_3^2 v_3^2$$
$$= \left(\tfrac{6}{59}\right) \times 4.7 + \left(\tfrac{31}{59}\right) \times 3.1 + \left(\tfrac{22}{59}\right) \times .4 = 2.256$$

The following table summarizes π^s and E^s for all the stationary policies. (Although this will not affect the computations in any way, note that each of policies 1, 3, 4, and 6 has an absorbing state: state 3. This is the reason $\pi_1 = \pi_2 = 0$ and $\pi_3 = 1$ for all these policies.)

s	π_1^s	π_2^s	π_3^s	E^s
1	0	0	1	−1
2	$\frac{6}{59}$	$\frac{31}{59}$	$\frac{22}{59}$	**2.256**
3	0	0	1	0.4
4	0	0	1	−1
5	$\frac{5}{154}$	$\frac{69}{154}$	$\frac{80}{154}$	1.724
6	0	0	1	−1
7	$\frac{5}{137}$	$\frac{62}{137}$	$\frac{70}{137}$	1.734
8	$\frac{12}{135}$	$\frac{69}{135}$	$\frac{54}{135}$	2.216

Policy 2 yields the largest expected yearly revenue. The optimum long-range policy calls for applying fertilizer regardless of the state of the system.

PROBLEM SET 25.3A

1. Solve Problem 2, Set 25.2a, for an infinite number of periods using the exhaustive enumeration method.
2. Solve Problem 2, Set 25.2a, for an infinite planning horizon using the exhaustive enumeration method.
*3. Solve Problem 3, Set 25.2a, by the exhaustive enumeration method assuming an infinite horizon.

25.3.2 Policy Iteration Method without Discounting

To appreciate the difficulty associated with the exhaustive enumeration method, let us assume that the gardener had four courses of action (alternatives) instead of two: (1) do not fertilize, (2) fertilize once during the season, (3) fertilize twice, and (4) fertilize three times. In this case, the gardener would have a total of $4^3 = 256$ stationary policies. By increasing the number of alternatives from 2 to 4, the number of stationary policies "soars" exponentially from 8 to 256. Not only is it difficult to enumerate all the policies explicitly, but the amount of computations may also be prohibitively large. This is the reason we are interested in developing the *policy iteration* method.

In Section 25.2, we have shown that, for any specific policy, the expected total return at stage n is expressed by the recursive equation

$$f_n(i) = v_i + \sum_{j=1}^{m} p_{ij} f_{n+1}(j), \; i = 1, 2, \ldots, m$$

This recursive equation is the basis for the development of the policy iteration method. However, the present form must be modified slightly to allow us to study the asymptotic behavior of the process. We define η as the number of stages *remaining* for consideration. This is in contrast with n in the equation, which defines stage n. The recursive equation is thus written as

$$f_\eta(i) = v_i + \sum_{j=1}^{m} p_{ij} f_{\eta-1}(j), \; i = 1, 2, 3, \ldots, m$$

Note that f_η is the cumulative expected revenue given that η is the number of stages remaining for consideration. With the new definition, the asymptotic behavior of the process can be studied by letting $\eta \to \infty$.

Given that

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$$

is the steady-state probability vector of the transition matrix $\mathbf{P} = \|p_{ij}\|$ and

$$E = \pi_1 v_1 + \pi_2 v_2 + \cdots + \pi_m v_m$$

is the expected revenue per stage as computed in Section 25.3.1, it can be shown that for very large η,

$$f_\eta(i) = \eta E + f(i)$$

where $f(i)$ is a constant term representing the asymptotic intercept of $f_\eta(i)$ given state i.

Because $f_\eta(i)$ is the cumulative optimum return for η remaining stages given state i and E is the expected revenue *per stage*, we can see intuitively why $f_\eta(i)$ equals ηE plus a correction factor $f(i)$ that accounts for the specific state i. This result assumes that $\eta \to \infty$.

Now, using this information, we write the recursive equation as

$$\eta E + f(i) = v_i + \sum_{j=1}^{m} p_{ij}\{(\eta - 1)E + f(j)\}, i = 1, 2, \ldots, m$$

Simplifying this equation, we get

$$E + f(i) - \sum_{j=1}^{m} p_{ij} f(j) = v_i, i = 1, 2, \ldots, m$$

Here, we have m equations in $m + 1$ unknowns, $f(1), f(2), \ldots, f(m)$, and E.

As in Section 25.3.1, our objective is to determine the optimum policy that yields the maximum value of E. Because there are m equations in $m + 1$ unknowns, the optimum value of E cannot be determined in one step. Instead, a two-step iterative approach is utilized which, starting with an arbitrary policy, will determine a new policy that yields a better value of E. The iterative process ends when two successive policies are identical.

1. *Value determination step.* Choose arbitrary policy s. Using its associated matrices $\mathbf{P}^s$ and $\mathbf{R}^s$ and arbitrarily assuming $f^s(m) = 0$, solve the equations

$$E^s + f^s(i) - \sum_{j=1}^{m} p_{ij}^s f^s(j) = v_i^s, i = 1, 2, \ldots, m$$

   in the unknowns $E^s, f^s(1), \ldots,$ and $f^s(m-1)$. Go to the policy improvement step.
2. *Policy improvement step.* For each state i, determine the alternative k that yields

$$\max_k \left\{ v_i^k + \sum_{j=1}^{m} p_{ij}^k f^s(j) \right\}, \quad i = 1, 2, \ldots, m$$

   The values of $f^s(j), j = 1, 2, \ldots, m,$ are those determined in the value determination step. The resulting optimum decisions for states $1, 2, \ldots, m$ constitute the new policy t. If s and t are identical, t is optimum. Otherwise, set $s = t$ and return to the value determination step.

Example 25.3-2

We solve the gardener problem by the policy iteration method.

Let us start with the arbitrary policy that calls for not applying fertilizer. The associated matrices are

$$\mathbf{P} = \begin{pmatrix} .2 & .5 & .3 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

The equations of the value iteration step are

$$\begin{aligned} E + f(1) - .2f(1) - .5f(2) - .3f(3) &= 5.3 \\ E + f(2) - .5f(2) - .5f(3) &= 3 \\ E + f(3) - f(3) &= -1 \end{aligned}$$

If we arbitrarily let $f(3) = 0$, the equations yield the solution

$$E = -1, f(1) = 12.88, f(2) = 8, f(3) = 0$$

Next, we apply the policy improvement step. The associated calculations are shown in the following tableau.

	$v_i^k + p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)$		Optimal solution	
i	$k = 1$	$k = 2$	$f(i)$	k^*
1	$5.3 + .2 \times 12.88 + .5 \times 8 + .3 \times 0 = 11.876$	$4.7 + .3 \times 12.88 + .6 \times 8 + .1 \times 0 = 13.36$	13.36	2
2	$3 + 0 \times 12.88 + .5 \times 8 + .5 \times 0 = 7$	$3.1 + .1 \times 12.88 + .6 \times 8 + .3 \times 0 = 9.19$	9.19	2
3	$-1 + 0 \times 12.88 + 0 \times 8 + 1 \times 0 = -1$	$.4 + .05 \times 12.88 + .4 \times 8 + .55 \times 0 = 4.24$	4.24	2

The new policy calls for applying fertilizer regardless of the state. Because the new policy differs from the preceding one, the value determination step is entered again. The matrices associated with the new policy are

$$\mathbf{P} = \begin{pmatrix} .3 & .6 & .1 \\ .1 & .6 & .3 \\ .05 & .4 & .55 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix}$$

These matrices yield the following equations:

$$\begin{aligned} E + f(1) - .3f(1) - .6f(2) - .1f(3) &= 4.7 \\ E + f(2) - .1f(1) - .6f(2) - .3f(3) &= 3.1 \\ E + f(3) - .05f(1) - .4f(2) - .55f(3) &= .4 \end{aligned}$$

Again, letting $f(3) = 0$, we get the solution

$$E = 2.26, f(1) = 6.75, f(2) = 3.80, f(3) = 0$$

The computations of the policy improvement step are given in the following tableau.

	$v_i^k + p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)$		Optimal solution	
i	$k = 1$	$k = 2$	$f(i)$	k^*
1	$5.3 + .2 \times 6.75 + .5 \times 3.80 + .3 \times 0 = 8.55$	$4.7 + .3 \times 6.75 + .6 \times 3.80 + .1 \times 0 = 9.01$	9.01	2
2	$3 + 0 \times 6.75 + .5 \times 3.80 + .5 \times 0 = 4.90$	$3.1 + .1 \times 6.75 + .6 \times 3.80 + .3 \times 0 = 6.06$	6.06	2
3	$-1 + 0 \times 6.75 + 0 \times 3.80 + 1 \times 0 = -1$	$.4 + .05 \times 6.75 + .4 \times 3.80 + .55 \times 0 = 2.26$	2.26	2

The new policy, which calls for applying fertilizer regardless of the state, is identical with the preceding one. Thus the last policy is optimal and the iterative process ends. This is the same conclusion obtained by the exhaustive enumeration method (Section 25.3.1). Note, however, that the policy iteration method converges quickly to the optimum policy, a typical characteristic of the new method.

PROBLEM SET 25.3B

1. Assume in Problem 1, Set 25.2a that the planning horizon is infinite. Solve the problem by the policy iteration method, and compare the results with those of Problem 1, Set 25.3a.
2. Solve Problem 2, Set 25.2a by the policy iteration method, assuming an infinite planning horizon. Compare the results with those of Problem 2, Set 25.3a.
3. Solve Problem 3, Set 25.2a by the policy iteration method assuming an infinite planning horizon, and compare the results with those of Problem 3, Set 25.3a.

25.3.3 Policy Iteration Method with Discounting

The policy iteration algorithm can be extended to include discounting. Given the discount factor α (<1), the finite-stage recursive equation can be written as (see Section 25.2)

$$f_\eta(i) = \max_k \left\{ v_i^k + \alpha \sum_{j=1}^{m} p_{ij}^k f_{\eta-1}(j) \right\}$$

(Note that η represents the number of stages *to go*.) It can be proved that as $\eta \to \infty$ (infinite stage model), $f_\eta(i) = f(i)$, where $f(i)$ is the expected present-worth (discounted) revenue given that the system is in state i and operating over an infinite horizon. Thus the long-run behavior of $f_\eta(i)$ as $\eta \to \infty$ is independent of the value of η. This is in contrast with the case of no discounting where $f_\eta(i) = \eta E + f(i)$. This result should be expected because in discounting the effect of future revenues will asymptotically diminish to zero. Indeed, the present worth $f(i)$ should approach a constant value as $\eta \to \infty$.

Based on this information, the steps of the policy iterations are modified as follows.

1. *Value determination step.* For an arbitrary policy s with its matrices $\mathbf{P}^s$ and $\mathbf{R}^s$, solve the m equations

$$f^s(i) - \alpha \sum_{j=1}^{m} p_{ij}^s f^s(j) = v_i^s, i = 1, 2, \ldots, m$$

in the m unknowns $f^s(1), f^s(2), \ldots, f^s(m)$.

2. *Policy improvement step.* For each state i, determine the alternative k that yields

$$\max_k \left\{ v_i^k + \alpha \sum_{j=1}^{m} p_{ij}^k f^s(j) \right\}, i = 1, 2, \ldots, m$$

$f^s(j)$ is obtained from the value determination step. If the resulting policy t is the same as s, stop; t is optimum. Otherwise, set $s = t$ and return to the value determination step.

Example 25.3-3

We will solve Example 25.3-2 using the discounting factor $\alpha = .6$.

Starting with the arbitrary policy, $s = \{1, 1, 1\}$, the associated matrices $\mathbf{P}$ and $\mathbf{R}$ ($\mathbf{P}^1$ and $\mathbf{R}^1$ in Example 25.3-1) yield the equations

$$\begin{aligned} f(1) - .6[.2f(1) + .5f(2) + .3f(3)] &= 5.3 \\ f(2) - .6[\quad\quad .5f(2) + .5f(3)] &= 3. \\ f(3) - .6[\quad\quad\quad\quad f(3)] &= -1. \end{aligned}$$

The solution of these equations yields

$$f_1 = 6.61, f_2 = 3.21, f_3 = -2.5$$

A summary of the policy improvement iteration is given in the following tableau:

	$v_i^k + .6[p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)]$		Optimal solution	
i	$k = 1$	$k = 2$	$f(i)$	k^*
1	$5.3 + .6[.2 \times 6.61 + .5 \times 3.21 + .3 \times -2.5] = 6.61$	$4.7 + .6[.3 \times 6.61 + .6 \times .21 + .1 \times -2.5] = 6.90$	6.90	2
2	$3 + .6[0 \times 6.61 + .5 \times 3.21 + .5 \times -2.5] = 3.21$	$3.1 + .6[.1 \times 6.61 + .6 \times 3.21 + .3 \times -2.5] = 4.2$	4.2	2
3	$-1 + .6[0 \times 6.61 + 0 \times 3.21 + 1 \times -2.5] = -2.5$	$.4 + .6[.05 \times 6.61 + .4 \times 3.21 + .55 \times -2.5] = .54$	.54	2

The value determination step using $\mathbf{P}^2$ and $\mathbf{R}^2$ (Example 25.3-1) yields the following equations:

$$f(1) - .6[\ .3f(1) + .6f(2) + \ .1f(3)] = 4.7$$

$$f(2) - .6[\ .1f(1) + .6f(2) + \ .3f(3)] = 3.1$$

$$f(3) - .6[.05f(1) + .4f(2) + .55f(3)] = \ .4$$

The solution of these equations yields

$$f(1) = 8.89, f(2) = 6.62, f(3) = 3.37$$

The policy improvement step yields the following tableau:

	$v_i^k + .6[p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k(3)]$		Optimal solution	
i	$k = 1$	$k = 2$	$f(i)$	k^*
1	$5.3 + .6[.2 \times 8.89 + .5 \times 6.62 + .3 \times 3.37] = 8.96$	$4.7 + .6[.3 \times 8.89 + .6 \times 6.62 + .1 \times 3.37] = 8.89$	8.96	1
2	$3 + .6[0 \times 8.89 + .5 \times 6.62 + .5 \times 3.37] = 6.00$	$3.1 + .6[.1 \times 8.89 + .6 \times 6.62 + .3 \times 3.37] = 6.62$	6.62	2
3	$-1 + .6[0 \times 8.89 + 0 \times 6.62 + 1 \times 3.37] = 1.02$	$.4 + .6[.05 \times 8.89 + .4 \times 6.62 + .55 \times 3.37] = 3.37$	3.37	2

Because the new policy $\{1, 2, 2\}$ differs from the preceding one, the value determination step is entered again using $\mathbf{P}^3$ and $\mathbf{R}^3$ (Example 25.3-1). This results in the following equations:

$$f(1) - .6[\ .2f(1) + .5f(2) + \ .3f(3)] = 5.3$$

$$f(2) - .6[\ .1f(1) + .6f(2) + \ .3f(3)] = 3.1$$

$$f(3) - .6[.05f(1) + .4f(2) + .55f(3)] = \ .4$$

The solution of these equations yields

$$f(1) = 8.97, f(2) = 6.63, f(3) = 3.38$$

The policy improvement step yields the following tableau:

	$v_i^k + .6[p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)]$		Optimal solution	
i	$k = 1$	$k = 2$	$f(i)$	k^*
1	$5.3 + .6[.2 \times 8.97 + .5 \times 6.63 + .3 \times 3.38] = 8.97$	$4.7 + .6[.3 \times 8.97 + .6 \times 6.63 + .1 \times 3.38] = 8.90$	8.98	1
2	$3 + .6[0 \times 8.97 + .5 \times 6.63 + .5 \times 3.38] = 6.00$	$3.1 + .6[.1 \times 8.97 + .6 \times 6.63 + .3 \times 3.38] = 6.63$	6.63	2
3	$-1 + .6[0 \times 8.97 + 0 \times 6.63 + 1 \times 3.38] = 1.03$	$.4 + .6[.05 \times 8.97 + .4 \times 6.63 + .55 \times 3.38] = 3.37$	3.37	2

Because the new policy $\{1, 2, 2\}$ is identical with the preceding one, it is optimal. Note that discounting has resulted in a different optimal policy that calls for not applying fertilizer if the state of the system is good (state 3).

PROBLEM SET 25.3C

1. Repeat the problems listed, assuming the discount factor $\alpha = .9$.

(a) Problem 1, Set 25.3b.

(b) Problem 2, Set 25.3b.

(c) Problem 3, Set 25.3b.

25.4 LINEAR PROGRAMMING SOLUTION

The infinite-state Markovian decision problems, both with discounting and without, can be formulated and solved as linear programs. We consider the no-discounting case first.

Section 25.3.1 shows that the infinite-state Markovian problem with no discounting ultimately reduces to determining the optimal policy, s^*, which corresponds to

$$\max_{s \varepsilon S} \left\{ \sum_{j=1}^{m} \pi_i^s v_i^s \middle| \boldsymbol{\pi}^s \mathbf{P}^s = \boldsymbol{\pi}^s, \quad \pi_1^s + \pi_2^s + \cdots + \pi_m^s = 1, \quad \pi_i^s \geq 0 \; i = 1, 2, \ldots, m \right\}$$

The set S is the collection of all possible policies of the problem. The constraints of the problem ensure that π_i^s, $i = 1, 2, \ldots, m$, represent the steady-state probabilities of the Markov chain $\mathbf{P}^s$.

The problem is solved in Section 25.3.1 by exhaustive enumeration. Specifically, each policy s is specified by a fixed set of actions (as illustrated by the gardener problem in Example 25.3-1). The same problem is the basis for the development of the LP formulation. However, we need to modify the unknowns of the problem such that the optimal solution *automatically* determines the optimal action (alternative) k when the system is in state i. The collection of all the optimal actions will then define s^*, the optimal policy.

Let

$q_i^k =$ Conditional probability of choosing alternative k given that the system is in state i

The problem may thus be expressed as

$$\text{Maximize } E = \sum_{i=1}^{m} \pi_i \left(\sum_{k=1}^{K} q_i^k v_i^k \right)$$

subject to

$$\pi_j = \sum_{i=1}^{m} \pi_i p_{ij}, \quad j = 1, 2, \ldots, m$$

$$\pi_1 + \pi_2 + \cdots + \pi_m = 1$$

$$q_i^1 + q_i^2 + \cdots + q_i^K = 1, \quad i = 1, 2, \ldots, m$$

$$\pi_i \geq 0, q_i^k \geq 0, \text{ for all } i \text{ and } k$$

Note that p_{ij} is a function of the policy selected and hence of the specific alternatives k of the policy.

The problem can be converted into a linear program by making proper substitutions involving q_i^k. Observe that the formulation is equivalent to the original one in Section 25.3.1 only if $q_i^k = 1$ for exactly *one* k for each i, which will reduce the sum $\sum_{k=1}^{K} q_i^k v_i^k$ to $v_i^{k^*}$, where k^* is the optimal alternative chosen. The linear program we develop here does account for this condition automatically.

Define

$$w_{ik} = \pi_i q_i^k, \quad \text{for all } i \text{ and } k$$

By definition w_{ik} represents the *joint* probability of state i making decision k. From probability theory

$$\pi_i = \sum_{k=1}^{K} w_{ik}$$

Hence,

$$q_i^k = \frac{w_{ik}}{\sum_{k=1}^{K} w_{ik}}$$

We thus see that the restriction $\sum_{i=1}^{m} \pi_i = 1$ can be written as

$$\sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} = 1$$

Also, the restriction $\sum_{k=1}^{K} q_i^k = 1$ is automatically implied by the way we defined q_i^k in terms of w_{ik}. (Verify!) Thus the problem can be written as

$$\text{Maximize } E = \sum_{i=1}^{m} \sum_{k=1}^{K} v_i^k w_{ik}$$

subject to

$$\sum_{k=1}^{K} w_{jk} - \sum_{i=1}^{m} \sum_{k=1}^{K} p_{ij}^k w_{ik} = 0, \quad j = 1, 2, \ldots, m$$

$$\sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} = 1$$

$$w_{ij} \geq 0, i = 1, 2, \ldots, m; k = 1, 2, \ldots, K$$

The resulting model is a linear program in w_{ik}. Its optimal solution automatically guarantees that q_i^k for one k for each i. First, note that the linear program has m independent equations (one of the equations associated with $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ is redundant).

Hence, the problem must have m basic variables. It can be shown that w_{ik} must be strictly positive for at least one k for each i. From these two results, we conclude that

$$q_i^k = \frac{w_{ik}}{\sum_{k=1}^{K} w_{ik}}$$

can assume a binary value (0 or 1) only. (As a matter of fact the preceding result also shows that $\pi_i = \sum_{k=1}^{K} w_{ik} = w_{ik^*}$, where k^* is the alternative corresponding to $w_{ik} > 0$.)

Example 25.4-1

The following is an LP formulation of the gardener problem without discounting:

$$\text{Maximize } E = 5.3w_{11} + 4.7w_{12} + 3w_{21} + 3.1w_{22} - w_{31} + .4w_{32}$$

subject to

$$\begin{aligned}
w_{11} + w_{12} - (.2w_{11} + .3w_{12} \quad\quad + .1w_{22} \quad\quad + .05w_{32}) &= 0 \\
w_{21} + w_{22} - (.5w_{11} + .6w_{12} + .5w_{21} + .6w_{22} \quad\quad + \ .4w_{32}) &= 0 \\
w_{31} + w_{32} - (.3w_{11} + .1w_{12} + .5w_{21} + .3w_{22} + w_{31} + .55w_{32}) &= 0 \\
w_{11} + w_{12} + w_{21} + w_{22} + w_{31} + w_{32} &= 1 \\
w_{ik} \geq 0, \text{ for all } i \text{ and } k &
\end{aligned}$$

The optimal solution is $w_{11} = w_{12} = w_{31} = 0$ and $w_{12} = .1017$, $w_{22} = .5254$, and $w_{32} = .3729$. This result means that $q_1^2 = q_2^2 = q_3^2 = 1$. Thus, the optimal policy selects alternative $k = 2$ for $i = 1, 2$, and 3. The optimal value of E is $4.7(.1017) + 3.1(.5254) + .4(.3729) = 2.256$. It is interesting that the positive values of w_{ik} exactly equal the values of π_i associated with the optimal policy in the exhaustive enumeration procedure of Example 25.3-1. This observation demonstrates the direct relationship between the two methods.

We next consider the Markovian decision problem with discounting. In Section 25.3.2 the problem is expressed by the recursive equation

$$f(i) = \max_k \left\{ v_i^k + \alpha \sum_{j=1}^{m} p_{ij}^k f(j) \right\}, i = 1, 2, \ldots, m$$

These equations are equivalent to

$$f(i) \geq \alpha \sum_{j=1}^{m} p_{ij}^k f(j) + v_i^k, \text{ for all } i \text{ and } k$$

provided that *f(i)* achieves its minimum value for each *i*. Now consider the objective function

$$\text{Minimize} \sum_{i=1}^{m} b_i f(i)$$

where b_i (>0 for all i) is an arbitrary constant. It can be shown that the optimization of this function subject to the inequalities given will result in the minimum value of *f(i)*. Thus, the problem can be written as

$$\text{Minimize} \sum_{i=1}^{m} b_i f(i)$$

subject to

$$f(i) - \alpha \sum_{i=1}^{m} p_{ij}^k f(j) \geq v_i^k, \text{ for all } i \text{ and } k$$

$$f(i) \text{ unrestricted in sign for all } i$$

Now the dual of the problem is

$$\text{Maximize} \sum_{i=1}^{m} \sum_{k=1}^{K} v_i^k w_{ik}$$

subject to

$$\sum_{k=1}^{K} w_{jk} - \alpha \sum_{i=1}^{m} \sum_{k=1}^{K} p_{ij}^k w_{ik} = b_j, j = 1, 2, \ldots, m$$

$$w_{ik} \geq 0, \text{ for } i = 1, 2, \ldots, m; k = 1, 2, \ldots, K$$

Example 25.4-2

Consider the gardener problem given the discounting factor $\alpha = .6$. If we let $b_1 = b_2 = b_3 = 1$, the dual LP problem may be written as

$$\text{Maximize } 5.3w_{11} + 4.7w_{12} + 3w_{12} + 3.1w_{22} - w_{31} + .4w_{32}$$

subject to

$$w_{11} + w_{12} - .6[.2w_{11} + .3w_{12} + .1w_{22} + .05w_{32}] = 1$$

$$w_{21} + w_{22} - .6[.5w_{11} + .6w_{12} + .5w_{21} + .6w_{22} + .4w_{32}] = 1$$

$$w_{31} + w_{32} - .6[.3w_{11} + .1w_{12} + .5w_{21} + .3w_{22} + w_{31} + .55w_{32}] = 1$$

$$w_{ik} \geq 0, \quad \text{for all } i \text{ and } k$$

The optimal solution is $w_{12} = w_{21}\, w_{31} = 0$ and $w_{11} = 1.5678, w_{22} = 3.3528$, and $w_{32} = 2.8145$. The solution shows that that optimal policy is $(1, 2, 2)$.

PROBLEM SET 25.4A

1. Formulate the following problems as linear programs.

(a) Problem 1, Set 25.3b.

(b) Problem 2, Set 25.3b.

(c) Problem 3, Set 25.3b.

REFERENCES

Derman, C., *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

Howard, R., *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.

Grimmet, G., and D. Stirzaker, *Probability and Random Processes*, 2nd ed., Oxford University Press, Oxford, England, 1992.

Kallenberg, O., *Foundations of Modern Probability*, Springer-Verlag, New York, 1997.

Stewart, W., *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1995.