Introduction to Dynamic Programming

by

LEON COOPER

and

MARY W. COOPER

Southern Methodist University, Dallas, Texas, USA

PERGAMON PRESS

OXFORD · NEW YORK · TORONTO · SYDNEY · PARIS · FRANKFURT

Chapter 8

Stochastic Processes and Dynamic Programming

8.1. Introduction

With one or two exceptions (e.g., the stochastic gold mining problem in Chapter 6) all of the processes or systems that have been formulated in the previous chapters have been deterministic. What this means is that when a decision is made at some stage, the state resulting from that decision is completely determined and known. The essential difference between that kind of process and a stochastic process is that the state resulting from a decision is not predetermined. It can only be described by some known probability distribution function which depends upon the initial state, and in some cases the decision that has been made.

We can clearly exhibit the differences, as well as the similarities, between the application of dynamic programming to deterministic and stochastic processes by returning to the general description of a dynamic programming process in Chapter 3. We briefly alluded to this difference in Section 3.2.

Consider a deterministic n-stage process in which the state of the system is λ_s. A decision x_s is made and the system undergoes a transformation which results in the state λ_{s-1}. We have seen many examples of this in previous chapters, e.g., $\lambda_{s-1} = \lambda_s - a_s x_s$. In general terms we can describe this as

$$\lambda_{s-1} = T_s(\lambda_s, x_s(\lambda_s)) \tag{8.1.1}$$

The important property of this change from state to state is that it is completely predetermined once T is specified. For example, if T_s is given by

$$\lambda_{s-1} = T_s(\lambda_s, x_s(\lambda_s)) = \lambda_s - a_s x_s \tag{8.1.2}$$

then once λ_s and x_s are specified, λ_{s-1} is completely determined.

Consider now the case of a stochastic process. It is the process of going from one state to the next that can only be described by a probability distribution of some kind. This means that the specification of the current state λ_s and a decision or action to be taken x_s does not determine the resulting state λ_{s-1}. Therefore, the transformation $T_s(\lambda_s, x_s(\lambda_s))$ does not uniquely determine a new state λ_{s-1}. What happens instead is that the new state is a random variable. We shall designate it $\tilde{\lambda}_{s-1}$ and it has associated with it a distribution function $H(\lambda_s, \tilde{\lambda}_{s-1}, x_s(\lambda_s))$. The distribution function depends upon the known state λ_s, the stochastic state vector $\tilde{\lambda}_{s-1}$, and the decision x_s that is made.

The reason H depends upon λ_s is that we always start with a *known* state. It is the transformation that is uncertain. Hence, the transformation is described by a stochastic state $\tilde{\lambda}_{s-1}$. However, before a decision is made at stage $s-1$, the actual state λ_{s-1} will be observed and hence will be known. Hence our *a priori* description of the state resulting from the transformation is in terms of a set of stochastic states $\tilde{\lambda}_{s-1}$, one of which will be the actual state, λ_{s-1}.

To clarify the difference between deterministic and stochastic processes we shall consider a familiar deterministic problem and then consider how one would deal with a stochastic analog of the problem. Suppose we wish to solve

$$\begin{aligned} &\max \quad z = \sum_{j=1}^{n} \phi_j(x_j) \\ &\text{subject to} \qquad x_j \in X_j \qquad (j = 1, 2, \ldots, n) \end{aligned} \tag{8.1.2a}$$

The sets X_j are specified in some way, the details of which are not important here.

Let us suppose that the state transformation relations are

$$\lambda_{s-1} = T_s(\lambda_s, x_s(\lambda_s)) \qquad (s = 2, 3, \ldots, n) \tag{8.1.3}$$

Then the recurrence relations that solve (8.1.2) are given by

$$g_1(\lambda_1) = \max_{x_1 \in X_1} \phi_1(x_1) \tag{8.1.4}$$

$$g_s(\lambda_s) = \max_{x_s \in X_s} [\phi_s(x_s) + g_{s-1}(T_s(\lambda_s, x_s(\lambda_s)))] \qquad (s = 2, 3, \ldots, n) \tag{8.1.5}$$

We have explained many times in this book how (8.1.4) and (8.1.5) are employed to solve (8.1.2).

Now let us suppose that the state transformations are stochastic in the sense we have previously discussed. Since a unique transformation from state λ_s result not in a unique state but rather in a vector of possible states $\tilde{\lambda}_{s-1}$, we shall assume that the state transformation relations are

$$\tilde{\lambda}_{s-1} = \tilde{T}_s(\lambda_s, x_s(\lambda_s)) \qquad (s = 2, 3, \ldots, n) \tag{8.1.6}$$

Since the $\tilde{\lambda}_{s-1}$ terms are stochastic, the x_s are also stochastic, since the choice of some specific x_s does not yield a determinate result. If we now consider once more the desire to maximize $\sum_{j=1}^{n} \phi_j(x_j)$ as in (8.1.2), it is quite clear that this objective function is now stochastic in view of the uncertainty inherent in $\tilde{\lambda}_s$ and x_s. Hence it is no longer feasible to maximize the return. What then can we do?

In the face of this kind of uncertainty, where a known probability distribution describes the uncertainty in the transformations, the most common substitute resorted to in probability theory is to maximize the *expectation* or *expected* return. Instead of optimal return, we shall have optimal expected return. We then define

$\tilde{g}_s(\lambda_s)$ = maximum expected return over $\tilde{\lambda}_{s-1}$, over the s remaining stages when the system is in state λ_s and using an optimal policy.

Before proceeding it will be recalled that the expectation $E(y)$ of a random variable y whose probability density function is $f(y)$ is, for the continuous case,

$$E(y) = \int_{-\infty}^{\infty} yf(y)\,dy \tag{8.1.7}$$

and for the case of the discrete density distribution, it is

$$E(y) = \sum_{i=1}^{k} y_i f(y_i) \tag{8.1.8}$$

We see that instead of solving (8.1.2) we seek to

$$\max \quad \tilde{z} = E\left[\sum_{j=1}^{n} \phi_j(x_j)\right] \tag{8.1.9}$$

Since the x_j are also stochastic, we shall emphasize this by rewriting (8.1.9) as

$$\max \quad \tilde{z} = E\left[\sum_{j=1}^{n} \phi_j(\tilde{\lambda}_{j-1},\, x_j)\right] \tag{8.1.10}$$

We recall from probability theory that the expectation of a sum is the sum of the expectations, i.e.,

$$E(u+v) = E(u)+E(v) \tag{8.1.11}$$

Using (8.1.10), (8.1.11), and the definition of $\tilde{g}_s(\lambda_s)$ we have for continuous density functions that

$$\tilde{g}_1(\lambda_1) = \max_{x_1 \in X_1} \left[\int_{-\infty}^{\infty} \phi_1(\tilde{\lambda}_0, x_1)\, d\Phi(\lambda_1, \tilde{\lambda}_0, x_1)\right] \tag{8.1.12}^{\dagger}$$

Similarly, the general recurrence relation is

$$\begin{aligned} \tilde{g}_s(\lambda_s) &= \max_{x_s \in X_s} E[\phi_s(\tilde{\lambda}_{s-1}, x_s)+g_{s-1}+(\tilde{\lambda}_{s-1})] \\ &= \max_{x_s \in X_s} \left[\int_{-\infty}^{\infty} [\phi_s(\tilde{\lambda}_{s-1}, x_s)+g_{s-1}(\tilde{\lambda}_{s-1})]\, d\Phi(\tilde{\lambda}_s, \tilde{\lambda}_{s-1}, x_s)\right] \\ &\qquad\qquad (s = 2, 3, \ldots, n) \end{aligned} \tag{8.1.13}$$

In the case of discrete density functions, the distribution $d\Phi(\lambda_s, \tilde{\lambda}_{s-1}, x_s)$ will be replaced by a set of probabilities $\{p_k\}$, so that the equivalent expression for (8.1.13) is

$$\tilde{g}_s(\lambda_s) = \max_{x_s \in X_s} \sum_{k=1}^{K} [\phi_s(\tilde{\lambda}_{s-1}, x_s)+g_{s-1}(\tilde{\lambda}_{s-1})]p_k \tag{8.1.14}$$

where

$$0 \leqslant p_k \leqslant 1 \qquad (k = 1, 2, \ldots, K)$$

$$\sum_{k=1}^{K} p_k = 1$$

† We are using the notation $E(y) = \int_{-\infty}^{\infty} y\, d\Phi(y)$, which is a Riemann-Stieltjes integral, to describe both continuous and discrete expectations.

In this case the probability density function is defined over K possible discrete states that the stochastic state vector $\tilde{\lambda}_{s-1}$ may assume.

In the next several sections we shall discuss, in some detail, problems which illustrate these ideas.

8.2. A Stochastic Allocation Problem—Discrete Case

Consider the problem we examined in Section 7.4 and special cases of which were treated in Section 4.6. We assumed that a certain resource w is to be divided into two parts x and $w-x$. The yields or profits from each of the parts are $f(x)$ and $h(w-x)$, respectively, in some operation or process. In the course of doing this, x is reduced to αx, $\alpha > 0$, and $w-x$ is reduced to $\beta(w-x)$, $0 \leq \beta \leq 1$. We wish to find the value of x_s at each stage of an n-stage process that results in the greatest overall yield from the operation.

The recurrence relations for the dynamic programming solution are

$$\begin{aligned} g_1(\lambda_1) &= \max_{0 \leq x_1 \leq \lambda_1} [f(x_1)+h(\lambda_1-x_1)] \\ g_s(\lambda_s) &= \max_{0 \leq x_s \leq \lambda_s} [f(x_s)+h(\lambda_s-x_s)+g_{s-1}(\alpha x_s+\beta(\lambda_s-x_s))] \\ & \qquad (s = 2, 3, \ldots, n) \end{aligned} \tag{8.2.1}$$

We shall now present a stochastic version of the above problem. We shall assume, when we choose x_s, that $f(x_s)$ is uncertain to the extent that it may take one of two possible values. With probability p_1 it assumes the value $f_1(x_s)$ and x_s becomes $\alpha_1 x_s$. With probability $p_2 = 1-p_1$, it takes the value $f_2(x_s)$ and x_s is reduced to $\alpha_2 x_s$. There are also two possibilities for $h(\lambda_s-x_s)$. With probability q_1 it assumes the value $h_1(\lambda_s-x_s)$ and λ_s-x_s is reduced to $\beta_1(\lambda_s-x_s)$. Similarly, with probability $q_2 = 1-q_1$, the value of $h_2(\lambda_s-x_s)$ is assumed and λ_s-x_s is reduced to $\beta_2(\lambda_s-x_s)$. We regard the choices which take place as being independent. Hence we have the following probabilities and random variables:

$$\begin{aligned} P_1 &= p_1 q_1, \quad \alpha_1 x_s+\beta_1(\lambda_s-x_s) \\ P_2 &= p_1 q_2, \quad \alpha_1 x_s+\beta_2(\lambda_s-x_s) \\ P_3 &= p_2 q_1, \quad \alpha_2 x_s+\beta_1(\lambda_s-x_s) \\ P_4 &= p_2 q_2, \quad \alpha_2 x_s+\beta_2(\lambda_s-x_s) \end{aligned} \tag{8.2.2}$$

We can observe that

$$\begin{aligned} P_1+P_2+P_3+P_4 &= p_1 q_1+p_1 q_2+p_2 q_1+p_2 q_2 \\ &= p_1 q_1+p_1(1-q_1)+(1-p_1)q_1+(1-p_1)(1-q_1) = 1 \end{aligned}$$

and that

$$0 \leq P_i \leq 1$$

Hence this distribution satisfies the requirements for a discrete probability distribution.

In order to formulate the recurrence relations we note that at each stage λ_s is now a stochastic variable $\tilde{\lambda}_s$ and is transformed into $\tilde{\lambda}_{s-1}$ with four different possibilities:

$$\begin{aligned} &\alpha_1 x_s+\beta_1(\lambda_s-x_s), \quad \alpha_2 x_s+\beta_1(\lambda_s-x_s) \\ &\alpha_1 x_s+\beta_2(\lambda_s-x_s), \quad \alpha_2 x_s+\beta_2(\lambda_s-x_s) \end{aligned} \tag{8.2.3}$$

Using the principle of optimality, (8.2.2) and (8.2.3), we obtain the following counterparts of (8.1.14):

$$\tilde{g}_1(\lambda_1) = \max_{0 \le x_1 \le \lambda_1} [p_1q_1\{f_1(x_1)+h_1(\lambda_1-x_1)\}+p_1q_2\{f_1(x_1)+h_2(\lambda_1-x_1)\} + p_2q_1\{f_2(x_1)+h_1(\lambda_1-x_1)\}+p_2q_2\{f_2(x_1)+h_2(\lambda_1-x_1)\}] \tag{8.2.4}$$

$$\begin{aligned}\tilde{g}_s(\lambda_s) = \max_{0 \le x_s \le \lambda_s} [&p_1q_1\{f_1(x_s)+h_1(\lambda_s-x_s)+g_{s-1}(\alpha_1x_s+\beta_1(\lambda_s-x_s))\} \\ &+p_1q_2\{f_1(x_s)+h_2(\lambda_s-x_s)+g_{s-1}(\alpha_1x_s+\beta_2(\lambda_s-x_s))\} \\ &+p_2q_1\{f_2(x_s)+h_1(\lambda_s-x_s)+g_{s-1}(\alpha_2x_s+\beta_1(\lambda_s-x_s))\} \\ &+p_2q_2\{f_2(x_s)+h_2(\lambda_s-x_s)+g_{s-1}(\alpha_2x_s+\beta_2(\lambda_s-x_s))\}]\end{aligned} \tag{8.2.5}$$

We see from (8.2.4) and (8.2.5) that if the p_1, p_2, q_1, q_2 are known, the calculation is no more difficult than the deterministic case, although somewhat lengthier.

8.3. A Stochastic Allocation Problem—Continuous Case

Let us now consider a stochastic version of the grower's problem of Section 4.6. We shall modify our notation slightly for convenience. We assume that in a given year λ_s units of the crop have been grown, some of which will be retained as seed for the succeeding year and the remainder of which will be sold. In the deterministic case, the yield from each unit of the seed crop was $\alpha > 1$ units and the income from y units of crop which are sold is $\phi(y)$. We shall now modify this assumption. Suppose that the grower decides to retain x_s and to sell $\lambda_s - x_s$ in year s. Suppose that the income $\phi(\lambda_s - x_s)$ is deterministic but that there is uncertainty about the amount of crop, $\tilde{\lambda}_{s-1}$ that will be available next year as a result of retaining x_s in year s. This is a common condition in growing, breeding, and investing. Let us assume that this uncertainty can be expressed as

$$\tilde{\lambda}_{s-1} = \tilde{T}_s(\lambda_s, x_s(\lambda_s)) = r_s x_s \tag{8.3.1}$$

The random variable r_s is assumed to have a probability density $p(r_s)$ that is known to the grower. Assuming that the probability density $p(r_s)$ remains constant during the n-year process, we shall assume that

$$p(r_s) = p(r) \tag{8.3.2}$$

The expected value of this density is

$$E(r) = \int_{-\infty}^{\infty} rp(r)\,dr \tag{8.3.3}$$

We assume that $E(r) > 1$. This assumption is equivalent to the assumption of $\alpha > 1$ in the deterministic case. Without such an assumption the grower would sell his entire crop and not preserve some for seed.

The total income from an n-stage process is

$$z = \sum_{j=1}^{n} \phi_j(\lambda_j - x_j) \tag{8.3.4}$$

If the initial amount of crop (for the nth year—the stages are numbered in reverse) was $\lambda_n = q$, the total amount of income from the stochastic process is

$$\tilde{z} = \phi(q-x_n)+\phi(r_n x_n - x_{n-1})+ \ \dots \ +\phi(r_2 x_2 - x_1) \tag{8.3.5}$$

We observe from (8.3.5) that the total income no longer depends only upon the initial amount of crop q and the series of decisions $x_n, x_{n-1}, \dots, x_1$, as in the deterministic case. The total income also depends upon the sequence of random numbers $r_n, r_{n-1}, \dots, r_2$. Hence it is not meaningful to ask what values of $x_1, x_2, \dots, x_n$ maximize z. Instead we shall maximize expected income.

The expected income is calculated by utilizing the known probability distributions of all the random variables that will affect this process. Therefore, at the beginning of the n-year planning period (year 1), the expected total income will be

$$\tilde{z} = \phi(q-x_n)+\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} [\phi(r_n x_n - x_{n-1})+ \ \dots \ +\phi(r_2 x_2 - x_1)] \prod_{j=n}^{2} [p(r_j)\, dr_j] \tag{8.3.6}$$

It must be recognized that the expected value of the total income does not remain constant throughout the process. In year 2 we have already selected some value x_n and so this value is known and the state variable λ_{n-1} is now known, even though we could not predict its value in year 1. Therefore using the values x_n and λ_n in year 2, the total expected income over n years is

$$\begin{aligned}\tilde{z} = \phi(q-x_n)+\phi(\lambda_{n-1}-x_{n-1})+\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} [\phi(r_{n-1}x_{n-1}-x_{n-2}) \\ + \ \dots \ +\phi(r_2 x_2 - x_1)] \prod_{j=n}^{2} [p(r_j)\, dr_j]\end{aligned} \tag{8.3.7}$$

which is not the same as the expected total return at the beginning (year 1) of the process. Similarly, at any stage s, the values of $x_n, x_{n-1}, \dots, x_{n-s+1}$ will have been observed and the states $\lambda_n, \lambda_{n-1}, \dots, \lambda_{n-s+1}$ will be known. Consequently, the expected total income is given by

$$\begin{aligned}\tilde{z} = \phi(q-x_n)+\phi(\lambda_{n-1}-x_{n-1})+ \ \dots \ +\phi(\lambda_{n-s+2}-x_{n-s+2}) \\ +\phi(\lambda_{n-s+1}-x_{n-s+1})+\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} [\phi(r_{n-s+1}x_{n-s+1}-x_{n-s}) \\ +\phi(r_2 x_2 - x_1)] \prod_{j=n-s+1}^{2} [p(r_j)\, dr_j]\end{aligned} \tag{8.3.8}$$

Although the expected value of income from n years of this stochastic process varies from stage to stage, it has an important property, viz., that at any stage s it is a deterministic function of the current state λ_s and the decisions $x_{n-s}, x_{n-s-1}, \dots, x_1$ which remain to be made. This, coupled with the separability that is inherent in (8.3.8), allows a dynamic programming solution to be made to this stochastic multistage decision process. By doing so, we shall obtain a solution such that the expected total income is maximized at each stage.

The optimal return functions $\tilde{g}_s(\lambda_s)$ are obtained as follows. For $s = 1$

$$\tilde{g}_1(\lambda_1) = \max_{0 \leq x_1 \leq \lambda_1} \phi(\lambda_1 - x_1) \tag{8.3.9}$$

Now let us consider the derivation of the general recurrence relation using the principle of optimality. After x_{s-1} is determined, at stage s the state

$$\lambda_s = rx_{s-1} \tag{8.3.10}$$

where r is the value taken by random variable r_{s-1}. When an optimal policy is used the expected return from the previous $s-1$ stages starting from state rx_{s-1} is

$$\tilde{g}_{s-1}(rx_s)$$

However, the return of the current stage is

$$\phi(\lambda_s - x_s)$$

Therefore the total income when an optimal policy is used over s stages is

$$z = \phi(\lambda_s - x_s) + \tilde{g}_{s-1}(rx_s) \tag{8.3.11}$$

When we now take account of the probability density $p(r)$ for the random variable r, the expected total return

$$\tilde{z} = E(z) = \int_{-\infty}^{\infty} zp(r)\,dr = \phi(\lambda_s - x_s) + \int_{-\infty}^{\infty} \tilde{g}_{s-1}(rx_s)\,p(r)\,dr \tag{8.3.12}$$

When the optimal policy is used, this expected value of the total income is maximized with respect to all decisions, including x_s. Therefore we have the general recurrence relation

$$\tilde{g}_s(\lambda_s) = \max_{0 \leq x_s \leq \lambda_s} \left[\phi(\lambda_s - x_s) + \int_{-\infty}^{\infty} \tilde{g}_{s-1}(rx_s)\,p(r)\,dr \right] \qquad (s = 2, 3, \ldots, n) \tag{8.3.13}$$

Hence, we can obtain the dynamic programming solution to the stochastic version of the grower's problem using (8.3.9) and (8.3.13).

Let us consider a particular case of the stochastic grower's problem in which $\phi(y) = ay^b$, $a > 0$, $0 < b < 1$. Then (8.3.9) and (8.3.13) become

$$\tilde{g}_1(\lambda_1) = \max_{0 \leq x_1 \leq \lambda_1} a(\lambda_1 - x_1)^b \tag{8.3.14}$$

$$\tilde{g}_s(\lambda_s) = \max_{0 \leq x_s \leq \lambda_s} \left[a(\lambda_s - x_s)^b + \int_{-\infty}^{\infty} \tilde{g}_{s-1}(rx_s)\,p(r)\,dr \right] \qquad (s = 2, 3, \ldots, n) \tag{8.3.15}$$

From (8.3.14) we see that $x_1 = 0$ will maximize $a(\lambda_1 - x_1)^b$. Therefore

$$\tilde{g}_1(\lambda_1) = a\lambda_1^b, \quad x_1^*(\lambda_1) = 0 \tag{8.3.16}$$

Consider now $s = 2$ in (8.3.15):

$$\tilde{g}_2(\lambda_2) = \max_{0 \leq x_2 \leq \lambda_2} \left[a(\lambda_2 - x_2)^b + \int_{-\infty}^{\infty} \tilde{g}_1(rx_2)\, p(r)\, dr \right] \tag{8.3.17}$$

Substituting (8.3.16) into (8.3.17) we obtain

$$\tilde{g}_2(\lambda_2) = \max_{0 \leq x_2 \leq \lambda_2} \left[a(\lambda_2 - x_2)^b + \int_{-\infty}^{\infty} a(rx_2)^b\, p(r)\, dr \right] \tag{8.3.18}$$

If we specify some particular $p(r)$, we can then evaluate the integral in (8.3.18). Let us assume that

$$\int_{-\infty}^{\infty} r^b p(r)\, dr = \varrho^b \tag{8.3.19}$$

From (8.3.18) and (8.3.19) we then have

$$\tilde{g}_2(\lambda_2) = \max_{0 \leq x_2 \leq \lambda_2} [a(\lambda_2 - x_2)^b + a x_2^b \varrho^b] \tag{8.3.20}$$

We can find the maximum value of x_2 by differentiating the maximand G_2 of $\tilde{g}_2(\lambda_2)$ in (8.3.20). Thus

$$\frac{\partial G_2}{\partial x_2} = ab(\lambda_2 - x_2)^{b-1}(-1) + a\varrho^b b x_2^{b-1} = 0 \tag{8.3.21}$$

Solving (8.3.21) we obtain

$$x_2^*(\lambda_2) = \frac{\lambda_2}{(\varrho^{b/(b-1)} + 1)} \tag{8.3.22}$$

Substitution of (8.3.22) into (8.3.20) yields

$$\tilde{g}_2(\lambda_2) = a\left(\lambda_2 - \frac{\lambda_2}{\varrho^{b/(b-1)} + 1}\right)^b + a\varrho^b\left(\frac{\lambda_2}{\varrho^{b/(b-1)} + 1}\right)^b$$

which simplifies to

$$\tilde{g}_2(\lambda_2) = \frac{a\varrho^b \lambda_2^b}{(\varrho^{b/(b-1)} + 1)^{b-1}} \tag{8.3.23}$$

Continuation of the recursion process will yield

$$x_3^*(\lambda_3) = \frac{\varrho^{b/(b-1)} + 1}{\varrho^{2b/(b-1)} + \varrho^{b/(b-1)} + 1}\, \lambda_3 \tag{8.3.24}$$

$$\tilde{g}_3(\lambda_3) = \frac{a\varrho^{2b}}{(\varrho^{2b/(b-1)} + \varrho^{b/(b-1)} + 1)^{b-1}} \tag{8.3.25}$$

It can be shown by mathematical induction that

$$x_s^*(\lambda_s) = \frac{\varrho^{(s-2)b/(b-1)} + \varrho^{(s-3)b/(b-1)} + \ldots + \varrho^{(0)b}}{\varrho^{(s-1)b/(b-1)} + \varrho^{(s-2)b/(b-1)} + \ldots + \varrho^{(0)b/(b-1)}}\, \lambda_s \tag{8.3.26}$$

$$(s = 2, 3, \ldots, n)$$

and

$$\tilde{g}_s(\lambda_s) = \frac{a\varrho^{(s-1)b}}{\left(\varrho^{(s-1)b/(b-1)} + \varrho^{(s-2)b/(b-1)} + \ldots + \varrho^{(0)b/(b-1)}\right)^{b-1}} \lambda_s^b \qquad (8.3.27)$$

$$(s = 2, 3, \ldots, n)$$

It is important to note how (8.3.16), (8.3.26), and (8.3.27) differ in their use from the deterministic solution of the grower's problem. At the beginning of the n-stage planning period for the grower, we do not know what values will be taken on by $r_1, r_2, \ldots$. We do know, however, that $\lambda_n = q$ and so we can compute the optimal first decision from $\tilde{g}_n(q)$ in the usual way. At any subsequent stage, say stage k, there will be uncertainty about what values the random variables $r_1, r_2, \ldots$, have taken. The optimal decisions are given by (8.3.26) in the same way as in the deterministic case. However, for a stochastic process the sequence of optimal decisions $x_1^*, x_2^*, \ldots, x_n^*$ cannot be calculated in advance because the state transformation relation

$$\lambda_{s-1} = r_s x_s$$

is probabilistic. Hence the result of optimal decision x_s^* is not known until stage $s+1$ after the random variable r_s has assumed some definite value.

8.4. A General Stochastic Inventory Model

Dynamic programming has proven to be of great utility in the solution of inventory problems. A fairly complete treatment of this subject is given in [23]. In this section we shall consider one general formulation based on material in [23].

We consider an inventory system in which some particular item is stocked. In each of n periods we review the inventory and we must decide whether or not to place an order for the item and if an order is placed, how much to order. We shall determine the amounts to be ordered in each of the n periods on the basis of minimizing the sum of all expected costs during the total n periods.

The demand in each period over the total planning period will be considered to be a random variable and will be designated r_j, i.e., the demand in period j. We shall assume that the probability distribution is given by a discrete density function $p_j(r_j)$. The demands in different periods are assumed to be independent of each other.

We assume that in placing an order we incur an ordering cost c_j, independent of the size of the order, if an order is placed in period j. In addition we incur the cost of the number of items we order which is dependent upon the number of items x_j that we order. We shall designate this cost by $\phi_j(x_j)$. In this formulation we shall consider that all demands are filled, i.e., if an order cannot be filled, it is backordered and filled when the stock is replenished. We shall not, however, include the inventory or backorder costs associated with the lead time required when ordering. By the inventory position, which we shall designate λ_j, we mean

λ_j = number of items in inventory + number of items on order − number of items on backorder, at the beginning of period j.

The inventory position in period j, after an order has been placed for x_j items, is λ_j+x_j, if x_j is the quantity ordered. The expected cost of carrying inventory will be designated $f_j(\lambda_j+x_j)$.

In selecting x_j, $j \geq 2$, we wish to take into account the demands which have occurred in periods 1 to $j-1$. This is done implicitly by use of the inventory position λ_j, which is our state variable. Hence, as usual, x_j depends upon λ_j. We then define as our optimal return function

$\tilde{g}_s(\lambda_s) \equiv$ minimum expected discounted cost for periods s through n, when the inventory position at the beginning of period s, before placing an order, is λ_s.

If the discount factor is α, then a mathematical statement of the definition of $g_s(\lambda_s)$, using our previously defined variables, is

$$\tilde{g}_s(\lambda_s) = \min_{x_s \ldots x_n} \sum_{\text{all } r_j} \left[\prod_{j=s}^{n} p_j(r_j)\right] \left\{\sum_{j=1}^{n} \alpha^{j-s}[c_j\delta_j+\phi_j(x_j)]+f_s(\lambda_s+x_s) + \sum_{j=s+1}^{n} \alpha^{j-s} f_j\left(\lambda_s+\sum_{k=s}^{j} x_k-\sum_{k=s}^{j-1} r_k\right)\right\} \qquad (s = 1, 2, \ldots, n) \qquad (8.4.1)$$

where

$$\delta_j = \begin{cases} 0, & x_j = 0 \\ 1, & x_j > 0 \end{cases}$$

It is worthwhile recalling that $x_s = x_s(\lambda_s)$, i.e., that the optimal decisions x_s are functions of the inventory positions λ_s. What (8.4.1) tells us, conceptually, is as follows. Suppose we select some policy, i.e., a set of $x_s(\lambda_s)$ and that we also have a set of demands r_s for each period. Then, beginning with the first period, we can sequentially calculate the λ_s and then the x_j. Using the x_j and r_j we can calculate the discounted cost over the planning period for this one set of r_j. We now repeat this calculation for all possible sets of r_j. Then we weight each of these costs by the probability of obtaining that particular set of r_s, i.e., $p_s(r_s)$. The addition of all of these is the expected cost for this one set of $x_s(\lambda_s)$. If we repeat this entire calculation for every allowable set of $x_s(\lambda_s)$, we can then choose the optimal $x_s^*(\lambda_s)$.

We can derive recurrence relations from the definition given by (8.4.1). First, we note that

$$\sum_{\text{all } r_j} \left[\prod_{j=s}^{n} p_j(r_j)\right] F(r_s, r_{s+1}, \ldots, r_n) = \sum_{r_s=0}^{\infty} p_s(r_s) \left\{\sum_{\text{all } r_j} \left[\prod_{j=s+1}^{n} p_j(r_j)\right] F(r_s, r_{s+1}, \ldots, r_n)\right\} \qquad (8.4.2)$$

Using (8.4.2) and (8.4.1) we obtain

$$\tilde{g}_s(\lambda_s) = \min_{x_s} \left\langle \sum_{\text{all } r_j} \left[\prod_{j=s}^{n} p_j(r_j)\right] [c_s\delta_s+\phi_s(x_s)+f_s(\lambda_s+x_s)] \times \alpha \sum_{r_s=0}^{\infty} p_s(r_s) \min_{x_{s+1}, \ldots, x_n} \sum_{\text{all } r_j} \left[\prod_{j=s+1}^{n} p_j(r_j)\right] \right.$$

$$\times\left\{\sum_{j=s+1}^{n}\alpha^{j-s-1}[c_j\delta_j+\phi_j(x_j)]+f_{s+1}(\lambda_s+x_s+x_{s+1}-r_s)\right.$$

$$\left.+\sum_{j=s+2}^{n}\alpha^{j-s-1}f_j\left(\lambda_s+x_s-r_s+\sum_{k=s+1}^{j}x_k-\sum_{k=s+1}^{j-1}r_k\right)\right\}\Bigg\rangle \qquad (s=1,2,\ldots,n-1) \tag{8.4.3}$$

Using the definition of $\tilde{g}_{s+1}$ and noting that

$$c_s\delta_s+\phi_s(x_s)+f_s(\lambda_s+x_s)$$

is independent of $r_s, r_{s+1}, \ldots, r_n$ and that

$$\sum_{\text{all } r_j}\left[\prod_{j=s}^{n}p_j(r_j)=1\right]$$

we can rewrite (8.4.3) as a recurrence relation

$$\tilde{g}_s(\lambda_s)=\min_{x_s}\left[c_s\delta_s+\phi_s(x_s)+f_s(\lambda_s+x_s)\right.$$

$$\left.+\sum_{r_s=0}^{\infty}p_s(r_s)\,g_{s+1}(\lambda_s+x_s-r_s)\right] \qquad (s=1,2,\ldots,n-1) \tag{8.4.4}$$

For the nth stage we have

$$\tilde{g}_n(\lambda_n)=\min_{x_n}[c_n\delta_n+\phi_n(x_n)+f_n(\lambda_n+x_n)] \tag{8.4.5}$$

In actually solving a problem when some probability distribution is given, it is no more difficult to use (8.4.4) and (8.4.5) than the corresponding relations for a deterministic problem. In order to do so, we must choose some finite limit to replace ∞ on the summation in (8.4.4). We replace the upper limit with some number β_s such that the probability of having a demand greater than β_s in period s is small enough to be ignored. Otherwise the calculation is as usual. We compute the $\tilde{g}_s(\lambda_s)$ starting with $\tilde{g}_n(\lambda_n)$. At the end of this sequence we calculate $\tilde{g}_1(\lambda_1)$, where λ_1 is the known initial inventory position. In doing so, we obtain x_1^*. From this we calculate λ_2 and use the $\tilde{g}_2(\lambda_2)$ function to find x_2^*. We proceed this way to calculate the values of all the remaining variables.

8.5. A Stochastic Production Scheduling and Inventory Control Problem

An important problem in the literature of management science is concerned with scheduling the production of some product over a planning horizon so that the combined expected cost of production and carrying inventory will be minimized. The stochastic nature of the problem resides in the fact that the demand in any period has to be considered to be a random variable. We shall make the following reasonable assumptions:

(1) Decisions on how much to produce are made each period.
(2) Each time production decisions are made we plan ahead for n periods.
(3) The demand r_j for the product in period j is a random variable with continuous density function $p_j(r_j)$.

The first model of such a situation we shall examine assumes that only one decision is made each period, i.e., how much of the product will be made during that period. All other variables such as material requirements, size of labor force, etc., are assumed to be able to be determined from the production decision. (We shall also consider a more complicated and realistic case further on.)

The inventory at the production facility is the inventory we shall consider. Let λ_j be our state variable. It is defined to be the net inventory at the beginning of period j. It is possible for λ_j to be negative if the unfilled orders exceed the inventory on hand. If we define x_j to be the amount produced in period j and r_j to be the demand in period j then we have

$$\lambda_{j+1} = \lambda_j + x_j - r_j \tag{8.5.1}$$

The costs associated with period j are several. There will be the normal costs involved in producing an amount x_j, such as labor, raw materials, etc. We denote this cost as the function $A_j(x_j)$, since it depends only on x_j. There will also be a cost incurred when the level of production, is either increased or decreased. These are costs associated with repair, changeover, hiring, firing, etc. This cost, which depends upon the change in the level of production, will be denoted by the function $B_j(x_j - x_{j-1})$. There is also a cost associated with carrying inventory or a cost associated with backorders. We assume that this cost depends upon $\lambda_j + x_j$ and we shall denote it as $C_j(x_j + \lambda_j)$.

We shall now define a state vector $\tilde{g}_s(\xi_s, \lambda_s)$ which depends upon two state variables λ_s and ξ_s. $\tilde{g}_s$ represents the expected cost for periods s through n, discounted to the beginning of period s, when an optimal decision is made at the beginning of each period $s \ldots n$. λ_s is calculated from (8.5.1) and ξ_s is the current value of x_{s-1}. The recurrence relations are easily found, from the principle of optimality, to be

$$\tilde{g}_s(\xi_s, \lambda_s) = \min_{x_s} \left[A_s(x_s) + B_s(x_s - \xi_s) + C_s(x_s + \lambda_s) \right.$$

$$\left. + \alpha \int_0^\infty \tilde{g}_{s+1}(x_s, \lambda_s + x_s - r_s)\, p_j(r_j)\, dr_j \right] \qquad (s = 1, 2, \ldots, n-1) \tag{8.5.2}$$

α is the discount factor. We also have, for the nth period,

$$\tilde{g}_n(\xi_n, \lambda_n) = \min_{x_n} \left[A_n(x_n) + B_n(x_n - \xi_n) + C_n(x_n + \lambda_n) \right] \tag{8.5.3}$$

The optimal solution is obtained when we compute $\tilde{g}_1(x_0, \lambda_1)$. The computational solution is straightforward providing the integral in (8.5.2) can be evaluated.

We can complicate the model we have just introduced, in the direction of greater practical import, by changing one assumption. We previously assumed that the amount to be produced uniquely determined the work force. This may not be the case, however. For example, several different sizes of work force could be used for any given amount of production by resorting to various amounts of overtime. Hence there could be several alternatives under consideration for the size of the work force. Let y_s be the size of the work force in period s. We now define the following costs:

$P_s(x_s) \equiv$ cost of production, excluding labor costs in period s, when x_s is produced.

$W_s(y_s) \equiv$ cost of labor in period s when y_s is the size of the work force.

$Q_s(y_s - y_{s-1}) \equiv$ cost associated with changing the size of the work force from y_{s-1} to y_s.

$R_s(x_s, y_s) \equiv$ cost of deviating from the ideal or correct size of labor force for producing an amount x_s.

$S_s(x_s + \lambda_s) \equiv$ cost of carrying inventory and of backorders.

We shall again require two state variables but they are different than in the previous case since there is no longer an explicit dependence on $x_s - x_{s-1}$. In addition, we shall require two control variables at each stage, viz., x_s and y_s. The state variables ξ_s and λ_s are associated with y_{s-1} and y_s respectively. We define $\tilde{g}_s(\xi_s, \lambda_s)$ as the expected cost for periods s through n, discounted to the beginning of period s, when an optimal decision is made at the beginning of period s. The recurrence relations are

$$\tilde{g}_s(\xi_s, \lambda_s) = \min_{x_s, y_s} \left[P_s(x_s) + W_s(y_s) + Q_s(y_s - \xi_s) + R_s(x_s, y_s) + S_s(x_s + \lambda_s) + \alpha \int_0^\infty \tilde{g}_{s-1}(y_s, \lambda_s + x_s - r_s)\, \phi_s(r_s)\, dr_s \right] \qquad (s = 1, 2, \ldots, n-1) \tag{8.5.4}$$

$$\tilde{g}_n(\xi_n, \lambda_n) = \min_{x_n, y_n} \left[P_n(x_n) + W_n(y_n) + Q_n(y_n - \xi_n) + R_n(x_n, y_n) + S_n(x_n + \lambda_n) \right] \tag{8.5.5}$$

The computational solution is again straightforward, although it may be quite lengthy because it requires both two state variables and two control variables.

In the stochastic decision problems we have just considered, the costs as well as the density functions for the variables could change from period to period. In such cases, we could only consider solving problems with a finite number of periods, i.e., a finite planning horizon. If one assumes that the costs and density functions do not change, then it is possible to consider an infinite planning horizon, i.e., that the system will continue to operate indefinitely. In such cases, one expects that in some sense future decisions (or the knowledge that the system is to operate indefinitely) affect current decisions. Since most businesses operate (or try to operate) indefinitely, such problems have practical importance. In the next several sections we shall consider the use of Markov processes as models for stochastic decision problems with infinite planning horizons.

8.6. Markov Processes

Consider a system which at any particular time is in one of a finite number of states which we designate $i = 1, 2, \ldots, N$. We further assume that the system undergoes transitions from one state to another state (which may be the same state) at certain discrete intervals of time. If the system is a *simple Markov process*, then the probability of a transition from a state i to a state j during the next time interval depends only upon i and j and not on the previous history of the system before it arrived in state i. Therefore

we may specify a set of conditional probabilities p_{ij} that a system which now occupies state i will occupy state j after its next transition. We shall designate the state transition matrix as $P = [p_{ij}]$. Since the p_{ij} are probabilities it must be true that

$$0 \leqslant p_{ij} \leqslant 1 \quad \text{all } i, j \tag{8.6.1}$$

and since the system must be in some state after its next transition, we have that

$$\sum_{j=1}^{N} p_{ij} = 1 \tag{8.6.2}$$

A simple example of a Markov process of the type we have defined above is a very naive weather model. Suppose that we decide to describe the weather in terms of two states. The weather is in state 1 when it is raining and in state 2 when it is not raining. It has been observed that when it is raining there is a 50% chance that it will also be raining tomorrow and a 50% chance that it will not. If it is not raining on any given day, there is a 25% chance that it will be raining tomorrow and consequently, a 75% chance that it will not be raining. Therefore our state transition probabilities p_{ij} are

$$p_{11} = \tfrac{1}{2}, \quad p_{12} = \tfrac{1}{2}, \quad p_{21} = \tfrac{1}{4}, \quad p_{22} = \tfrac{3}{4}$$

or in matrix form

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

The matrix P, the state transition matrix, is a complete description of a simple Markov process. It can be used to deduce all the important characteristics of a Markov process. For example, if we wished to know the probability that it would be raining 10 days from now, given that it is not raining today, we could compute this probability using P. In order to do so, we need some additional notation.

Let $s_i(n)$ be the probability that the system will occupy state i after n state transitions, assuming that its initial state ($n = 0$) is known. We call $s_i(n)$ the *state probabilities*. Since the $s_i(n)$ are probabilities, it must be true that

$$\sum_{i=1}^{N} s_i(n) = 1 \tag{8.6.3}$$

We determine the state probability for state j for the $(n+1)$st transition by multiplying each state probability $s_i(n)$ by p_{ij} and summing over all states, i.e.,

$$s_j(n+1) = \sum_{i=1}^{N} s_i(n) p_{ij} \qquad (n = 0, 1, 2, \ldots) \tag{8.6.4}$$

We can express (8.6.4) more compactly using vector notation. If $\mathbf{s}(n)$ is the vector whose components are $s_i(n)$, then

$$\mathbf{s}(n+1) = \mathbf{s}(n)P \tag{8.6.5}$$

If we apply (8.6.5) successively, we have

$$\begin{aligned}\mathbf{s}(1) &= \mathbf{s}(0)P\\ \mathbf{s}(2) &= \mathbf{s}(1)P = \mathbf{s}(0)P^2\\ \mathbf{s}(3) &= \mathbf{s}(2)P = \mathbf{s}(0)P^3\\ &\vdots\end{aligned}$$

The general recursive relationship is

$$\mathbf{s}(n) = \mathbf{s}(0)P^n \tag{8.6.6}$$

Equation (8.6.6) enables us to find the probability that the system is in each of its states by multiplying the probability vector of its initial state $\mathbf{s}(0)$ by the nth power of P.

Let us consider our weather model to illustrate the use of (8.6.5) and (8.6.6). Suppose it is raining, which means that $s_1(0) = 1$ and $s_2(0) = 0$. Therefore $\mathbf{s}(0) = [1, 0]$. Hence we have

$$\mathbf{s}(1) = \mathbf{s}(0)P = [1, 0]\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{4} & \frac{3}{4}\end{bmatrix}$$

and

$$\mathbf{s}(1) = \begin{bmatrix}\frac{1}{2} & \frac{1}{2}\end{bmatrix} \tag{8.6.7}$$

After one day it is equally likely to be raining or not raining. After two days

$$\mathbf{s}(2) = \mathbf{s}(1)P = \begin{bmatrix}\frac{1}{2} & \frac{1}{2}\end{bmatrix}\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{4} & \frac{3}{4}\end{bmatrix} \tag{8.6.8}$$

and

$$\mathbf{s}(2) = \begin{bmatrix}\frac{3}{8} & \frac{5}{8}\end{bmatrix}$$

Therefore it is less likely to be raining after two days. We could just as well obtain (8.6.8) by using (8.6.6) directly. Since

$$P^2 = \begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{4} & \frac{3}{4}\end{bmatrix}\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{4} & \frac{3}{4}\end{bmatrix} = \begin{bmatrix}\frac{3}{8} & \frac{5}{8}\\ \frac{5}{16} & \frac{11}{16}\end{bmatrix}$$

$$\mathbf{s}(2) = \mathbf{s}(0)P^2 = [1, 0]\begin{bmatrix}\frac{3}{8} & \frac{5}{8}\\ \frac{5}{16} & \frac{11}{16}\end{bmatrix} = \begin{bmatrix}\frac{3}{8} & \frac{5}{8}\end{bmatrix}$$

If we continue the calculation of $\mathbf{s}(n)$ we obtain the results shown in Table 8.6.1.

TABLE 8.6.1. State probabilities for weather model (rain)

n	0	1	2	3	4	5	6	...
$s_1(n)$	1	0.5	0.375	0.34375	0.33594	0.33398	0.33350	...
$s_2(n)$	0	0.5	0.625	0.65625	0.66406	0.66602	0.66650	...

It can be seen from Table 8.6.1 that as n increases $s_1(n)$ is approaching $\frac{1}{3}$ and $s_2(n)$ appears to be approaching $\frac{2}{3}$. One might wonder what would happen if we had started in initial state 2, i.e., no rain. For this case $\mathbf{s}(0) = [0, 1]$ and successive state probabilities would be as appears in Table 8.6.2.

TABLE 8.6.2. State probabilities for weather model (no rain)

n	0	1	2	3	4	5	6	...
$s_1(n)$	0	0.25	0.3125	0.328125	0.33203	0.33301	0.33325	...
$s_2(n)$	1	0.75	0.6875	0.671875	0.66797	0.66699	0.66675	...

What Tables 8.6.1 and 8.6.2 seem to indicate is that the state probabilities approach $\frac{1}{3}$ and $\frac{2}{3}$ irrespective of the initial state of the system. Whether we started with $\mathbf{s}(0) = [1, 0]$ or $[0, 1]$, we seem to be approaching $s_1(n) = \frac{1}{3}$ and $s_2(n) = \frac{2}{3}$ as n becomes large. This is indeed the case as can be shown by an analysis of the asymptotic form of (8.6.4). A Markov process whose limiting state probability distribution is independent of the initial state of the system is called a *completely ergodic process.* Not all Markov processes are ergodic.

For completely ergodic Markov processes it is possible to find the limiting state probabilities by a direct calculation. If we define s_i as the probability that the system occupies state i after a sufficiently large number of moves, then the vector $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ is the vector of limiting state probabilities. When n is large $\mathbf{s}(n) \to \mathbf{s}$. Therefore, for sufficiently large n we have, from (8.6.5), that

$$\mathbf{s} = \mathbf{s}P \tag{8.6.9}$$

and in addition

$$\sum_{i=1}^{N} s_i = 1 \tag{8.6.10}$$

Equations (8.6.9) and (8.6.10) can be used to solve for the limiting state probabilities. For the weather model example, we have

$$\begin{aligned} s_1 &= \tfrac{1}{2}s_1 + \tfrac{1}{4}s_2 \\ s_2 &= \tfrac{1}{2}s_1 + \tfrac{3}{4}s_2 \\ s_1 + s_2 &= 1 \end{aligned} \tag{8.6.11}$$

Note that (8.6.11) is an overdetermined set of linear equations. Solution of these equations yields $s_1 = \frac{1}{3}$, $s_2 = \frac{2}{3}$, which is what we conjectured from Tables 8.6.1 and 8.6.2.

In the next section we shall add some additional structure and features to a Markov process. The treatment of this subject follows Howard [7], who was responsible for its development.

8.7. Markovian Sequential Decision Processes

We shall now suppose that a Markov process with N states has a reward or economic value structure associated with the transitions from one state to another. We call r_{ij} the reward or profit associated with a transition from state i to state j. The matrix $R = [r_{ij}]$ is the matrix of the rewards associated with all possible state transitions. Negative rewards (profits) indicate a loss on a given transition. The Markov process will generate a sequence of rewards as it moves from state to state. Hence the reward is also a random variable with a probability distribution dependent upon the probabilistic structure of the Markov process.

A Markov process with rewards can be considered to be a crude model of certain kinds of enterprises whose earnings fluctuate because of elements of the operating environment over which the enterprise has no immediate control. Nevertheless, if one has some notion of the probabilities of certain events and factors that may influence them then it is possible to influence or determine longer term profits. A typical desire is to determine, for a Markov process with a reward structure, the expected earnings over the next n transitions given that the system is currently in state i.

We define $v_i(n)$ as the expected total earnings in the next n transitions if the system is currently in state i. We shall now develop a recursive relationship which bears a strong resemblance to, and is one step removed from, a dynamic programming relationship.

Suppose we consider the transitions in reverse. If the system makes a transition from state i to j it will earn an amount r_{ij} plus the amount it expects to earn if it starts in state j with one transition fewer to be made. This total earnings is

$$r_{ij} + v_j(n-1) \tag{8.7.1}$$

However, the probability that the system will go from state i to state j is p_{ij}. Hence the determination of the total expected earnings requires that (8.7.1) be weighted by p_{ij} and summed over all states. Therefore we have that

$$v_i(n) = \sum_{j=1}^{N} p_{ij}[r_{ij} + v_j(n-1)] \qquad \begin{matrix}(i = 1, 2, \ldots, N) \\ (n = 1, 2, \ldots)\end{matrix} \tag{8.7.2}$$

We may rearrange (8.7.2) to yield

$$v_i(n) = \sum_{j=1}^{N} p_{ij} r_{ij} + \sum_{j=1}^{N} p_{ij} v_j(n-1) \qquad \begin{matrix}(i = 1, 2, \ldots, N) \\ (n = 1, 2, \ldots)\end{matrix} \tag{8.7.3}$$

Since the first term on the right-hand side of (8.7.3) does not depend upon n but is a constant for each state i, we may define

$$q_i = \sum_{j=1}^{N} p_{ij} r_{ij} \qquad (i = 1, 2, \ldots, N) \tag{8.7.4}$$

Hence we may rewrite (8.7.3) as

$$v_i(n) = q_i + \sum_{j=1}^{N} p_{ij} v_j(n-1) \qquad \begin{matrix}(i = 1, 2, \ldots, N) \\ (n = 1, 2, \ldots)\end{matrix} \tag{8.7.5}$$

The quantity q_i may be regarded as the amount of earnings to be expected in the next transition from state i. It is called the expected immediate reward for state i. We may also write equation (8.7.5) in vector form as

$$\mathbf{v}(n) = \mathbf{q} + P\mathbf{v}(n-1) \qquad (n = 1, 2, \ldots) \tag{8.7.6}$$

where $v_i(n)$, q_i are the components of $\mathbf{v}(n)$ and $\mathbf{q}$, respectively.

Let us consider as an example of a Markov process with a reward structure, a manufacturer of several brands of cigarettes who sets as his goal that his combined market share for his brands shall exceed 12% which we shall call state 1. When his market share is below 12% we shall call this state 2. We shall refer to the two states as high market share and low market share, respectively. If the manufacturer has a high market share one quarter and also the following quarter, he earns a profit or reward of 100. If he goes from low market share to low market share during the quarter his profit is 21 units. If, on the other hand, he goes from low market share to high or from high market share to low during the quarter, then his profit is 60 units. Therefore his reward matrix R is

$$R = \begin{bmatrix} 100 & 60 \\ 60 & 21 \end{bmatrix}$$

Suppose the probability transition matrix for this process is

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

Then using (8.7.4) we can calculate $\mathbf{q}$ as

$$q_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 100 \\ 60 \end{bmatrix} = 80$$

$$q_2 = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 60 \\ 21 \end{bmatrix} = 34$$

$$\mathbf{q} = (80, 34)$$

Hence the cigarette manufacturer expects to gain 80 units on leaving state 1 and 34 units on leaving state 2.

Using the data in the foregoing paragraph we can calculate the total expected earnings for the cigarette manufacturer if he has n quarters remaining to be in business. Since most companies wish to remain in business indefinitely, this is a decided shortcoming of the present analysis, which we shall correct subsequently. Nevertheless, the analysis is instructive. We shall arbitrarily decide that $\mathbf{v}(0) = [0, 0]$. This is the "salvage" value of the company, i.e., the value of the company to some hypothetical purchaser on the day the company ceases to operate. We calculate successive total earnings using (8.7.6).

$$v_1(1) = q_1 + \sum_{j=1}^{2} p_{1j} v_j(0) = 80 + \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 80$$

$$v_2(1) = q_2 + \sum_{j=1}^{2} p_{2j} v_j(0) = 34 + \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 34$$

$$v_1(2) = q_1 + \sum_{j=1}^{2} p_{1j} v_j(1) = 80 + \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 80 \\ 34 \end{bmatrix} = 137$$

$$v_2(2) = q_2 + \sum_{j=1}^{2} p_{2j} v_j(1) = 34 + \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 80 \\ 34 \end{bmatrix} = 83.33$$

etc.

A tabulation of the total expected earnings is given in Table 8.7.1.

TABLE 8.7.1. Total expected earnings (n = number of quarters remaining)

n	0	1	2	3	4	5	6	7	...
$v_1(n)$	0	80	137	247.17	463.03	889.54	1736.5	3423.3	...
$v_2(n)$	0	34	83.33	184.56	389.99	804.33	1637.1	3307.3	...

An examination of Table 8.7.1 indicates that as n increases, the ratio of $v_1(n)$ to $v_2(n)$ is approaching unity. Hence, if n is large the relative value of being in state 1 over state 2 is small as a percent of the total expected earnings, but is not zero.

We turn now to a consideration of sequential Markovian decision processes. Let us suppose that our cigarette manufacturer is capable of taking different actions when he is in a low or high market share situation and that these actions will change the probabilities and rewards that govern this Markov process. For example, when the company is in state 1 (high market share) it has two courses of action that it may take; it may employ a high or low advertising budget. However, a larger advertising cost will decrease profits somewhat even though it may help prolong the high market share situation. Similarly, when the company is in state 2 (low market share), it may carry out market research studies to determine what would appeal to its potential customers. It may do this at two different budget levels as well, high and low. Each of the alternatives in state 1 or state 2 has its own associated state transition probabilities and reward matrix.

We shall use a superscript k to indicate the alternatives available in each state, i.e., p_{ij} and r_{ij} will now be designated p_{ij}^k and r_{ij}^k. In the case of our cigarette company, if he is in state 1, he has two alternatives. For alternative 1 (high advertising budget) we will have $\mathbf{p}_1^1 = [p_{11}^1, p_{12}^1] = [0.7\ 0.3]$ and for alternative 2 (low advertising) we have $\mathbf{p}_1^2 = [0.5\ 0.5]$. The corresponding reward vectors for each of these alternatives is $\mathbf{r}_1^1 = [r_{11}^1, r_{12}^1] = [80\ 40]$ and $\mathbf{r}_1^2 = [100\ 60]$. For state 2 the two alternatives are high and low market research levels. For these alternatives the probabilities and rewards are: $\mathbf{p}_2^1 = [0.6\ 0.4]$, $\mathbf{p}_2^2 = [0.33\ 0.67]$ and $\mathbf{r}_2^1 = [40\ 10]$, $\mathbf{r}_2^2 = [60\ 21]$. We shall use these data in a subsequent calculation.

In order to clarify the kinds of transitions that may result when there are alternative transition probabilities and rewards, we show the tree of possible transitions in Fig.

8.7.1. In Fig. 8.7.1(a) we have shown the tree of transitions when there is only one alternative between stage transitions. By introducing alternatives in each transition, we have to replace each line in between stages in (a) by K lines, if there are K alternatives in each state transition. As an example, in Fig. 8.7.1(b) we have shown the different transition alternatives between stages 1 and 2 if the system was in state 1 in stage 1 and went to state 2 in stage 2 with the given probabilities and rewards.

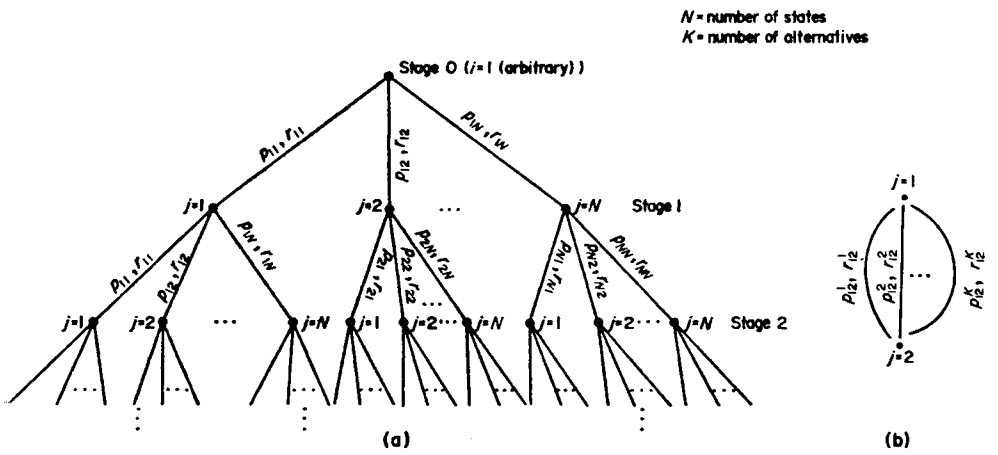

FIG. 8.7.1. State transitions and alternatives.

We now redefine some quantities to take account of the alternatives. Let q_i^k be the expected immediate reward from a single transition from state i under alternative k. Then we have

$$q_i^k = \sum_{j=1}^{N} p_{ij}^k r_{ij}^k \tag{8.7.7}$$

Let n be the number of stages remaining (in the case of the cigarette company, the number of quarters) in the process. Let $d_i(n)$ be the number of the alternative in the ith state that will be used at stage n. Hence $d_i(n)$ is a decision that is made by the policy maker at each stage. A specification of $d_i(n)$ for all i and all n is termed a policy. We define the *optimal policy* as the policy that maximizes total expected earnings or return for each i and n.

We now redefine $v_i(n)$ as the total expected return in n stages starting in state i *under an optimal policy*. This definition and the principle of optimality will allow the derivation of the basic recursion equation. The basic argument is familiar.

Suppose that we have chosen alternatives at stages n, $n-1$, ..., 1, so that we have found $v_j(n)$ for $j = 1, 2, \ldots, N$. We now wish to find for stage $n+1$ the alternative we should follow in state i, in order to maximize $v_i(n+1)$. The choice is $d_i(n+1)$. If we choose alternative k, then the expected return for $n+1$ stages would be

$$\sum_{j=1}^{N} p_{ij}^k [r_{ij}^k + v_j(n)] \tag{8.7.8}$$

We seek the alternative that will maximize (8.7.8). Hence we determine $v_i(n+1)$ from

$$v_i(n+1) = \max_k \sum_{j=1}^{N} p_i^k [r_i^k + v_j(n)] \tag{8.7.9}$$

Equation (8.7.9) can be rewritten as

$$v_i(n+1) = \max_k \left[q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j(n) \right] \tag{8.7.10}$$

Equation (8.7.10) can be used to find which alternative to use in each state at each stage and will also provide the expected future earnings at each stage. We shall not pursue a calculational example because the limitation of the model to processes that will terminate in a finite number of stages is a severe one. Instead, in the next section, we shall develop a technique that will enable us to deal with processes which continue indefinitely, with no prospect of termination in a given number of transitions.

8.8. The Policy Iteration Method of Howard

We shall be dealing with a completely ergodic N-state Markov process, described by a transition probability matrix P and a reward matrix R. As the system makes transition after transition, the total earnings increases as the number of transitions increases. However, if the system undergoes a very large number of transitions, the average earnings of the process will be a more useful measure of the process.

For an ergodic system, the limiting state probabilities s_i are independent of the initial state. Therefore, the *gain* g of the process will be

$$g = \sum_{i=1}^{N} s_i q_i \tag{8.8.1}$$

where q_i is the expected immediate return.

The gain is a useful means of comparison, between different ergodic Markov processes, of their long-term profitability. The one with the largest gain is the most profitable.

We shall now develop the policy iteration method for finding the optimal policy of a Markov process with rewards and alternatives in each state. The method consists of two parts—a value determination step and a policy improvement step.

We consider first the value determination step. Suppose we are operating the system under some given policy. Hence we have a specified Markov process with rewards. We allow the process to operate for n stages or transitions. We now define $v_i(n)$ as the total expected earnings of the system in n transitions if it starts from state i under the assumed policy.

We have already seen, in Section 8.7, that $v_i(n)$ satisfies the recurrence relations

$$v_i(n) = q_i + \sum_{j=1}^{N} p_{ij} v_j(n-1) \qquad \begin{matrix} (i = 1, 2, \ldots, N) \\ (n = 1, 2, \ldots) \end{matrix} \tag{8.8.2}$$

Since we have *assumed* some definite policy there is no need for the superscript k.

For a completely ergodic process an analysis of the limiting form of (8.8.2) as n becomes large gives the result

$$v_i(n) = ng + v_i \quad (n \gg 0) \tag{8.8.3}$$

where g is the gain and v_i is an asymptote of $v_i(n)$.

Since we are concerned with processes that have a very large number of stages we shall use (8.8.3). If we combine (8.8.2) and (8.8.3), we obtain

$$ng + v_i = q_i + \sum_{j=1}^{N} p_{ij}[(n-1)g + v_j] \qquad (i = 1, 2, \ldots, N) \tag{8.8.4}$$

Rewriting (8.8.4) we have

$$ng + v_i = q_i + (n-1)g \sum_{j=1}^{N} p_{ij} + \sum_{j=1}^{N} p_{ij} v_j \qquad (i = 1, 2, \ldots, N)$$

which simplifies to

$$g + v_i = q_i + \sum_{j=1}^{N} p_{ij} v_j \qquad (i = 1, 2, \ldots, N) \tag{8.8.5}$$

Equations (8.8.5) are a set of N linear simultaneous equations in $N+1$ variables (g, v_i, $i = 1, 2, \ldots, N$). Since this is an undetermined system, we may set one of the v_i equal to zero and then determine the values of the remaining v_i as well as g. This will not necessarily be the set of values that satisfy (8.8.3). However, it is a simple matter to show that they will differ from their correct values by some constant value. They will be sufficient for the purpose of the value determination step. The values obtained by setting $v_N = 0$ and solving (8.8.5) will be called *relative values.*

The reason we can use relative values instead of the correct values can be seen by examining (8.8.3). Consider (8.8.3) for two states v_s and v_t. We have from (8.8.3) that for large n

$$v_s(n) = ng + v_s, \quad v_t(n) = ng + v_t \tag{8.8.6}$$

The difference $v_s(n) - v_t(n) = v_s - v_t$. Hence if the correct values of v_s, v_t were $v_s + \alpha$, $v_t + \alpha$, the difference would still be $v_s - v_t$. Since this difference represents the difference in the long-term earnings of the process as a result of starting in state s rather than state t, it will serve our purposes in value determination.

We turn now to the second step of the policy iteration method, viz., the policy improvement step.

We noted in (8.7.10) that if we had an optimal policy up to stage n we could find the best alternative in the ith state at stage $n+1$ by maximizing

$$q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j(n) \tag{8.8.7}$$

over all alternatives k in state i. For large n, we can substitute from (8.8.3) into (8.8.7) to obtain

$$q_i^k + \sum_{j=1}^{N} p_{ij}^k (ng + v_j) \tag{8.8.8}$$

as the quantity that is to be maximized in each state. Rewriting (8.8.8) yields

$$q_i^k + ng \sum_{j=1}^{N} p_{ij}^k + \sum_{j=1}^{N} p_{ij}^k v_j \tag{8.8.9}$$

Since ng is a constant (independent of j) and

$$\sum_{j=1}^{N} p_{ij}^k = 1$$

we need not consider the second term of (8.8.9), when we compare the maximand of (8.8.9) for each alternative k. Hence we may use as a quantity

$$q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j \tag{8.8.10}$$

to be maximized with respect to the alternatives k in state i. The relative values v_i that were determined in the value determination step can be used in (8.8.10) when the maximization is carried out.

The policy improvement step, then, consists in finding, for each state i, the alternative k that maximizes

$$q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j$$

using the relative values v_i, determined under the previous policy. The alternative k becomes d_i, the decision in the state i. The set of d_i then become the new policy. We shall show subsequently that the new policy will have a greater gain than the previous policy.

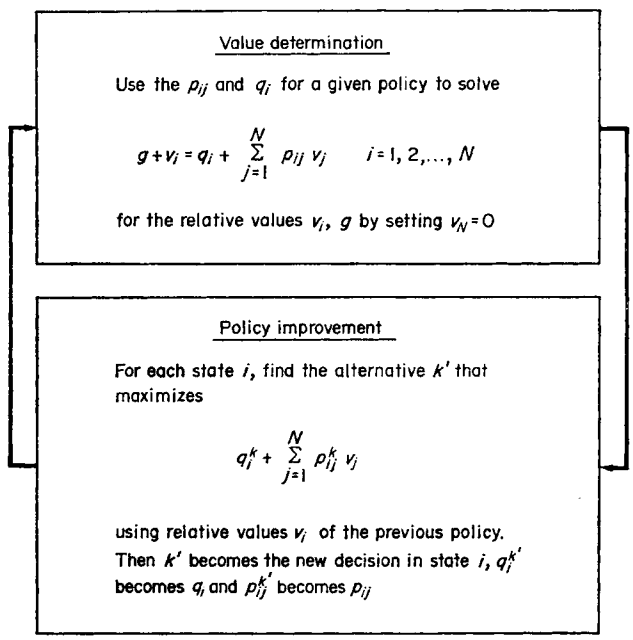

FIG. 8.8.1. Iteration cycle of policy iteration method.

We can put the value determination step and the policy improvement step together to describe the basic iteration cycle of the policy iteration method. Figure 8.8.1 shows the basic iteration cycle.

It is possible to enter the cycle at either step. If we enter at the value determination step, then we must select an initial policy. If we choose to enter the policy improvement step, then we must supply an initial set of relative values. A convenient method of starting is to set all $v_i = 0$ and enter at the policy improvement step.

In order to have a convergent algorithm the policy iteration method must have the following properties:

(1) The successive policies that are found have strictly increasing gains g.
(2) The computation terminates when the policy with the greatest gain has been found. This occurs when the policies on two successive iterations are identical.

We shall now establish that the above properties hold and hence that the policy iteration method will find the optimal policy. This proof is given by Howard [7].

Suppose that our current policy is designated A and that the policy improvement step has produced a new policy B. Then we seek to show, if the gains g^A and g^B designate the gains associated with policies A and B, respectively, that $g^B > g^A$.

Since the test quantity in the policy improvement step is

$$q_i + \sum_{j=1}^{N} p_{ij} v_j \tag{8.8.11}$$

and policy B was chosen as superior to A, then it must be true that

$$q_i^B + \sum_{j=1}^{N} p_{ij}^B v_j^A \geqslant q_i^A + \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad (i = 1, 2, \ldots, N) \tag{8.8.12}$$

Let

$$\gamma_i = q_i^B + \sum_{j=1}^{N} p_{ij}^B v_j^A - q_i^A - \sum_{j=1}^{N} p_{ij}^A v_j^A \tag{8.8.13}$$

Hence $\gamma_i \geqslant 0$ represents the improvement in (8.8.11) for state i. From (8.8.5) we know that for policies A and B we may write

$$g^A + v_i^A = q_i^A + \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad (i = 1, 2, \ldots, N) \tag{8.8.14}$$

$$g^B + v_i^B = q_i^B + \sum_{j=1}^{N} p_{ij}^B v_j^B \qquad (i = 1, 2, \ldots, N) \tag{8.8.15}$$

If we now subtract (8.8.14) from (8.8.15) we obtain

$$g^B - g^A + v_i^B - v_i^A = q_i^B - q_i^A + \sum_{j=1}^{N} p_{ij}^B v_j^B - \sum_{j=1}^{N} p_{ij}^A v_j^A \tag{8.8.16}$$

We now solve (8.8.13) for $q_i^B - q_i^A$ and substitute for $q_i^B - q_i^A$ in (8.8.16). This results in

$$g^B - g^A + v_i^B - v_i^A = \gamma_i - \sum_{j=1}^{N} p_{ij}^B v_j^A + \sum_{j=1}^{N} p_{ij}^A v_j^A + \sum_{j=1}^{N} p_{ij}^B v_j^B - \sum_{j=1}^{N} p_{ij}^A v_j^A$$

which simplifies to

$$g^B - g^A + v_i^B - v_i^A = \gamma_i + \sum_{j=1}^{N} p_{ij}^B (v_j^B - v_j^A) \tag{8.8.17}$$

If we define $$g' = g^B - g^A \quad \text{and} \quad v_i' = v_i^B - v_i^A$$

then we may rewrite (8.8.17) as

$$g' + v_i' = \gamma_i + \sum_{j=1}^{N} p_{ij}^B v_j' \qquad (i = 1, 2, \ldots, N) \tag{8.8.18}$$

Equations (8.8.18) have the same form as (8.8.5). The solution for g from (8.8.5) was

$$g = \sum_{i=1}^{N} s_i q_i$$

Similarly, the solution for g' in (8.8.18) is

$$g' = \sum_{i=1}^{N} s_i^B \gamma_i \tag{8.8.19}$$

where s_i^B is the limiting state probability of state i under policy B. Since all $s_i^B \geqslant 0$ and all $\gamma_i \geqslant 0$, therefore $g' \geqslant 0$ and $g^B \geqslant g^A$. Since we are dealing with a completely ergodic Markov process, g^B will be strictly greater than g^A if an improvement, even in only one state, can be made in policy B over policy A. Hence $g^B > g^A$, when an improvement is possible.

Finally, we show that it is not possible for a better policy to exist and not be determined at some iteration in the policy improvement step. Let us assume that, for two policies A and B, $g^B > g^A$ but the policy improvement step has converged on policy A. Then in each state, $\gamma_i \leqslant 0$, where γ_i is given by (8.8.13). Since $s_i^B \geqslant 0$ for all i, (8.8.19) yields $g' \leqslant 0$ and hence $g^B - g^A \leqslant 0$. However, we assumed that $g^B > g^A$. Hence we have obtained a contradiction. Therefore it is not possible for a superior policy to exist, but not be found, by the policy iteration method.

We shall now solve the problem of the cigarette company, the data for which were presented in Section 8.7. The data are summarized in Table 8.8.1. The last column q_i^k was computed from

$$q_i^k = \sum_{j=1}^{N} p_{ij}^k r_{ij}^k$$

For example, for $i = 1$, $k = 2$, we have

$$q_1^2 = \tfrac{1}{2}(100) + \tfrac{1}{2}(60) = 80$$

We shall begin by choosing $v_1 = v_2 = 0$. Then we enter the policy improvement step. Since $v_1 = v_2 = 0$

$$q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j = q_i^k$$

TABLE 8.8.1. Data for cigarette company

		p_{ij}^k		r_{ij}^k		q_i^k
i	k	$j=1$	2	$j=1$	2	
1	1	$\frac{7}{10}$	$\frac{3}{10}$	80	40	68
	2	$\frac{1}{2}$	$\frac{1}{2}$	100	60	80
2	1	$\frac{3}{5}$	$\frac{2}{5}$	40	10	28
	2	$\frac{1}{3}$	$\frac{2}{3}$	60	21	34

Hence we choose the alternative in each state that maximizes q_i^k. Therefore our first policy is the second alternative in each state, i.e.,

$$\mathbf{d} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

The transition probability matrix and expected immediate rewards corresponding to this policy are

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix} \qquad \mathbf{q} = \begin{bmatrix} 80 \\ 34 \end{bmatrix}$$

We now carry out the value determination step using the P and $\mathbf{q}$ corresponding to th previous policy. Hence we wish to solve

$$\begin{aligned} g + v_1 &= 80 + \tfrac{1}{2}v_1 + \tfrac{1}{2}v_2 \\ g + v_2 &= 34 + \tfrac{1}{3}v_1 + \tfrac{2}{3}v_2 \end{aligned} \tag{8.8.20}$$

We set $v_2 = 0$ and obtain for the solution to (8.8.20)

$$v_1 = 55.2, \quad v_2 = 0, \quad g = 52.4$$

Using the new relative values v_i, we enter the policy improvement step and calculate

$$q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j$$

for each state.

For $i = 1$, $k = 1$ we have

$$q_1^1 + \sum_{j=1}^{N} p_{ij}^1 v_j = 68 + \tfrac{7}{10}(55.2) + \tfrac{3}{10}(0) = 106.64$$

and for $k = 2$ we have

$$80 + \tfrac{1}{2}(55.2) + \tfrac{1}{2}(0) = 107.6$$

Hence we choose alternative 2 in state 1. For $i = 2$, $k = 1$ we have

$$28 + \tfrac{3}{5}(55.2) + \tfrac{2}{5}(0) = 61.12$$

and for $k = 2$

$$34+\tfrac{1}{3}(55.2)+\tfrac{2}{3}(0) = 52.40$$

Hence we choose alternative 1 in state 2. Therefore, our new policy is

$$\mathbf{d} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

The transition probability matrix and expected immediate earnings corresponding to the new policy are

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{3}{5} & \frac{2}{5} \end{bmatrix} \qquad q = \begin{bmatrix} 80 \\ 28 \end{bmatrix}$$

We again enter the value determination step and solve

$$\begin{aligned} g+v_1 &= 80+\tfrac{1}{2}v_1+\tfrac{1}{2}v_2 \\ g+v_2 &= 28+\tfrac{3}{5}v_1+\tfrac{3}{5}v_2 \end{aligned} \tag{8.8.21}$$

Setting $v_2 = 0$, we find the solution to (8.8.21) to be

$$v_1 = 47.27, \quad v_2 = 0, \quad g = 56.36$$

Using the new relative values we calculate

$$g_i^k + \sum_{j=1}^{N} p_{ij}^k v_j$$

for each state. We then have

$$\begin{aligned} i = 1,\ k = 1 \quad & 68+\tfrac{7}{10}(47.27)+\tfrac{3}{10}(0) = 101.09 \\ i = 1,\ k = 2 \quad & 80+\tfrac{1}{2}(47.27)+\tfrac{1}{2}(0) = 103.64 \\ i = 2,\ k = 1 \quad & 28+\tfrac{3}{5}(47.27)+\tfrac{2}{5}(0) = 56.36 \\ i = 2,\ k = 2 \quad & 34+\tfrac{1}{3}(47.27)+\tfrac{2}{3}(0) = 49.76 \end{aligned}$$

The new policy is $\mathbf{d} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ which is the same as the previous policy. Hence we have converged. The optimal policy for the cigarette company is to use alternative 2 (low advertising) when the company is in state 1 (a larger market share) and to use alternative 1 (larger market research activity) when the company is in state 2 (smaller market share).

As the example indicates, the calculation converged very rapidly (2 iterations). In Howard [7], an example is presented with 40 states and 41 alternatives in each state. Hence there are $(41)^{40}$ possible policies, which is approximately 3.25×10^{64} policies, an incredibly large number. However, the policy iteration method found the optimal policy in only seven iterations, which indicates the great efficiency of the method.

Exercises—Chapter 8

8.1. For the problem of Section 8.2, suppose that when we choose x_s, with probability p_1, $f(x)$ assumes the value $f_1(x_s)$ and x_s becomes $\alpha_1 x_s$; with probability p_2, $f(x)$ takes the value $f_2(x_s)$ and x_s becomes $\alpha_2 x_s$; and finally, with probability $p_3 = p - p_1 - p_2$, $f(x)$ takes the value $f_3(x_s)$ and x_s becomes $\alpha_3 x_s$. Similarly, with probability q_1, $h(\lambda_s - x_s)$ becomes $h_1(\lambda_s - x_s)$ and $\lambda_s - x_s$ is reduced to $\beta_1(\lambda_s - x_s)$; with probability q_2, we have $h_2(\lambda_s - x_s)$ and $\beta_2(\lambda_s - x_s)$; with probability $q_3 = 1 - q_1 - q_2$, we have $h_3(\lambda_s - x_s)$ and $\beta_3(\lambda_s - x_s)$. Derive the recurrence relations for the optimal solution.

8.2. Suppose the cigarette manufacturer of Sections 8.7 and 8.8 analyzes his data more closely and allows for three states: 1 (high market share), 2 (average market share), and 3 (low market share) and allows for several alternatives in each state. In state 1, low, average and high advertising are the three alternatives. In state 2 the company may also use two alternatives, i.e., average advertising and low market research. In state 3 we have three alternatives, low, average, or high market research activities. A summary of the data is

State alternative		p_{ij}^k			r_{ij}^k		
i	k	$j=1$	2	3	1	2	3
1	1	.5	.3	.2	80	60	40
	2	.4	.3	.3	85	75	50
	3	.3	.4	.3	100	80	60
2	1	.4	.6		60	20	
	2	.3	.7		40	10	
3	1	.5	.3	.2	40	20	10
	2	.4	.3	.3	50	30	15
	3	.2	.4	.4	60	40	20

Find the optimal policy.

8.3. Derive recurrence relations for the solution of

$$\max \prod_{j=1}^{n} \phi_j(x_j)$$

subject to $x_j \in X_j \qquad (j = 1, 2, \ldots, n)$

where the X_j are given as in Section 8.3.

8.4. Develop a dynamic programming algorithm for the following problem. We are given a network with nodes numbered from 1 through N and arc (ij) connects pairs of nodes i and j. Associated with each arc is a nonnegative distance or cost c_{ij}. We assume there are no loops in the network. Suppose now that once a particular arc is chosen, there is a finite probability that an adjacent arc will be traversed instead. Therefore you cannot be certain of traveling along the arc you desire. Determine the optimal choice of arcs, to minimize the total *expected* distance from node 1 to node N.