Chapter 6

CONTROL OF MARKOV CHAINS OVER AN INFINITE HORIZON

6.1 Infinite Horizon Stochastic Control Problems for Markov Chains

In this chapter, we shall examine some infinite time stochastic control problems. We have already discussed very briefly the infinite time linear regulator problem in Chapter 5. That problem, however, is very special in the sense that an analytical solution is possible. Few other interesting infinite time stochastic control problems admit such simple analytical solutions. Nevertheless, it is possible to characterize the solution of the stochastic control problem in certain cases. These concern the control of Markov chains with finite control sets.

The systems we shall be interested in have a countable state space. We know such systems can always be represented as a Markov chain. Since the control affects the future evolution of the system, we may view it as a parameter governing the transition probabilities of the chain. This leads us to the following description of the process to be controlled:

The process x_t is a homogeneous Markov chain with state space $\mathcal{X}$ the nonnegative integers. Its evolution is governed by the time-invariant transition probabilities $P_{ij}(u) = P\{x_{t+1} = j | x_t = i \text{ and control } u \text{ was used}\}$. The set of control values $\mathcal{U}$ is usually assumed to be *finite*. This, we shall see, is a crucial simplifying assumption which fortunately is often satisfied in applications. Note that since the state space is the set of nonnegative integers, sometimes k is used to denote a state value. We therefore use t to denote time in this chapter.

It is possible to express any Markov chain in the form of a stochastic difference equation. To see this, suppose we have a countable set of nonnegative numbers $\{a_k, k \geq 0\}$, with $0 \leq a_k \leq 1$, and $\sum a_k = 1$. The a_k's represent a probability distribution. Suppose we now want to construct a random variable X with the probability distribution

$$P(X = k) = a_k$$

We can do this through the use of a random variable w uniformly distributed on [0 1]. Set

$$X = \sum_{k=0}^{\infty} k I_{(\sum_{i=0}^{k-1} a_i,\ \sum_{i=0}^{k} a_i]}(w)$$

where $I_A(x)$ is the indicator function of the set A, i.e.,

$$\begin{aligned} I_A(x) &= 1 \quad \text{if } x \in A \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Then

$$P(X=k) = P(\sum_{i=0}^{k-1} a_i < w \leq \sum_{i=0}^{k} a_i) = a_k$$

as desired. Now consider a homogeneous Markov chain with transition probabilities $P_{ij}(u)$. Define

$$f(i,u,w) = \sum_{k=0}^{\infty} k I_{(\sum_{j=0}^{k-1} P_{ij}(u),\ \sum_{j=0}^{k} P_{ij}(u)]}(w)$$

Then we have

$$f(i,u,w) = k \Longleftrightarrow w \in (\sum_{j=0}^{k-1} P_{ij}(u),\ \sum_{j=0}^{k} P_{ij}(u)]$$

so that

$$P[f(i,u,w)=k] = P_{ik}(u)$$

Finally, define the stochastic process x_t through the stochastic difference equation

$$x_{t+1} = f(x_t, u_t, w_t)$$

where w_t is an independent identically distributed (i.i.d) random sequence with a uniform distribution on [0 1]. The process x_t is the desired Markov chain with the required transition probability distribution $P_{ij}(u_t)$. Owing to this construction, we can interpret stochastic control problems for Markov chains in the same way as we did in the previous chapter.

To describe the stochastic control problem completely, we need to specify the cost criterion. We shall define the per stage cost by $L(\cdot,\cdot) : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. The function L has the interpretation that if $x_t = i$, and the action u is chosen, then the cost incurred is $L(i,u)$. There are 3 main forms of the cost criterion for an infinite time problem:

(i) discounted cost: $E\sum_{t=0}^{\infty} \alpha^t L(x_t, u_t), \quad \alpha \in (0,1)$

(ii) positive cost: $E\sum_{t=0}^{\infty} L(x_t, u_t)$ with $L(i,u) \geq 0$, all i and u

(iii) average cost: $\lim_{T\to\infty} \frac{1}{T} E\sum_{t=0}^{T-1} L(x_t, u_t)$

The simplest cast is the discounted cost case, which we shall first discuss.

6.2 The Discounted Cost Criterion

Assume that $|L(i,u)| < M$ is a finite constant. Since the control set is finite, and the state space is countable, we can allow the admissible control laws to be any function of the state

$$u_t = \phi_t(x_t)$$

If the policy is time varying, that is, if $\Phi = \{\phi_0, \phi_1, ...\}$ consists of different functions, the transition probabilities, being given by $P_{ij}(\phi_t(i))$ at time t, will no longer be stationary. We will only get a Markov chain with stationary transition probabilities if a stationary policy $\Phi = \{\phi, \phi ...\}$ is used.

Let α be any number in (0,1). For any policy Φ, define

$$V_\Phi(i) = E_i^\Phi \left[\sum_{t=0}^{\infty} \alpha^t L(x_t, u_t) \right] \tag{6.1}$$

where E_i^Φ is the expectation operation conditioned on $x_0 = i$ and the policy Φ being used. Since $L(i,u)$ is bounded, every policy Φ gives a bounded $V_\Phi(i)$.

The interpretation of the discount factor α is that costs incurred in the future are less important than those incurred in the present. This is reflected by the fact that future costs are discounted at the rate of α per unit time.

The control problem is to find a policy Φ^* such that $V_\Phi(i)$ is minimized for all i. More precisely, let

$$V_\alpha(i) = \inf_\Phi V_\Phi(i)$$

Φ^* is optimal if $V_{\Phi^*}(i) = V_\alpha(i)$ for each i.

The following theorem gives the dynamic programming equation which characterizes the optimal cost $V_\alpha(i)$.

Theorem 6.1 *The optimal cost V_α satisfies the equation*

$$V_\alpha(i) = \min_u \left\{ L(i,u) + \alpha \sum_{j=0}^{\infty} P_{ij}(u) V_\alpha(j) \right\} \tag{6.2}$$

Proof: Let Φ be an arbitrary policy and suppose Φ chooses control u at time $t = 0$. Then

$$V_\Phi(i) = L(i,u) + \sum_{j=0}^{\infty} P_{ij}(u) W_\Phi(j)$$

where $W_\Phi(j)$ is the expected discounted cost incurred from time 1 onwards, given that the policy Φ is used and $x_1 = j$. But from the form of the cost function

$$W_\Phi(j) \geq \alpha V_\alpha(j)$$

Hence

$$\begin{aligned} V_\Phi(i) &\geq L(i,u) + \alpha \sum_{j=0}^{\infty} P_{ij}(u) V_\alpha(j) \\ &\geq \min_u \left\{ L(i,u) + \alpha \sum_{j=0}^{\infty} P_{ij}(u) V_\alpha(j) \right\} \end{aligned}$$

Since Φ is arbitrary, we obtain

$$V_\alpha(i) \geq \min_u \left\{ L(i,u) + \alpha \sum_{j=0}^{\infty} P_{ij}(u) V_\alpha(j) \right\} \tag{6.3}$$

To go the other way, suppose u_0 is the control which minimizes the right hand side of (6.2). Such a control exists since the control set is finite. Now let Φ be the policy which chooses u_0 at time 0 and if the next state is j, to apply Φ_j, a policy satisfying

$$V_{\Phi_j}(j) \leq V_\alpha(j) + \varepsilon \qquad \varepsilon > 0$$

Hence

$$\begin{aligned} V_\Phi(i) &= L(i,u_0) + \alpha \sum_{j=0}^{\infty} P_{ij}(u_0) V_{\Phi_j}(j) \\ &\leq L(i,u_0) + \alpha \sum_{j=0}^{\infty} P_{ij}(u_0)[V_\alpha(j) + \varepsilon] \\ &= \min_u \left[L(i,u) + \alpha \sum_{j=0}^{\infty} P_{ij}(u) V_\alpha(j) \right] + \alpha\varepsilon \end{aligned}$$

Hence

$$V_\alpha(i) \leq \min_u [L(i,u) + \alpha \sum_{j=0}^{\infty} P_{ij}(u) V_\alpha(j)] \tag{6.4}$$

since ε is arbitrary. Equations (6.3) and (6.4) combine to give the desired result.

Remark 6.2.1: We can interpret Theorem 6.1 using the Principle of Optimality. Suppose we start in state i and choose an arbitrary decision u which may or may not be optimal. This incurs an initial cost of $L(i,u)$ and the system makes a random transition to state j with probability $P_{ij}(u)$. By the Principle of Optimality, for the overall behaviour to be optimal, we must behave optimally from stage 1 onwards. However, starting with stage 1, we apply a discount factor of α. So if we land in state j and behave optimally from stage 1 to infinity, we will incur the cost $\alpha V_\alpha(j)$. Averaging over all random transitions, the total expected cost becomes

$$L(i,u) + \alpha \sum_{j=0}^{\infty} P_{ij}(u) V_\alpha(j)$$

Since the first decision u is arbitrary, the optimal choice at stage 0 must be to minimize the above total expected cost to give the optimal cost. This is precisely the dynamic programming equation (6.2).

Remark 6.2.2: If the state transition is described using a stochastic difference equation (which as shown above can be done for Markov chains),

$$x_{t+1} = f(x_t, u_t, w_t)$$

where w_t is an i.i.d. sequence, then it is readily seen that the optimal value function satisfies the equation

$$V_\alpha(x) = \min[L(x,u) + \alpha E_w V_\alpha(f(x,u,w))] \tag{6.5}$$

Combining with the results of Problem 5.6 in Chapter 5, we see that if we can solve (6.5) for $V_\alpha(x)$ with $\alpha^t E V_\alpha(x_t) \underset{t\to\infty}{\longrightarrow} 0$, $V_\alpha(x)$ is the optimal value function even when x_t is a more general Markov process with a state space which may be uncountable, and when the per stage costs may not be bounded. The policy resulting from minimizing the R.H.S. of (6.5) is then optimal.

We shall defer the discussion of the existence of solutions to (6.2) later. Our present task is to establish the form of the optimal policy. To this end, we introduce the space $B(I)$, which is the space of bounded functions with domain the nonnegative integers, I. For any function $f : I \to U$, define

$$T_f : B(I) \to B(I)$$

by

$$[T_f \gamma](i) = L(i, f(i)] + \alpha \sum_{j=0}^{\infty} P_{ij}[f(i)] \gamma(j)$$

The mappings T_f have the following properties:

(1) T_f is monotone, i.e. if $\gamma_1(i) \geq \gamma_2(i)$, all i, then $(T_f\gamma_1)(i) \geq (T_f\gamma_2)(i)$, all i.

(2) If V_f is the cost incurred using the stationary policy $\{f, f, ...\}$, $T_f V_f = V_f$, i.e., V_f is a fixed point of T_f.

(3) $T_f^n\gamma \underset{n\to\infty}{\longrightarrow} V_f$ all $\gamma \in B(I)$.

Properties (1) and (2) are obvious.
Property (3) may be proved as follows: For any $\gamma \in B(I)$,

$$\begin{aligned}(T_f^2\gamma)(i) &= L(i, f(i)) + \alpha\sum_{j=0}^{\infty} P_{ij}(f(i))(T_f\gamma)(j) \\ &= L(i, f(i)) + \alpha\sum_{j=0}^{\infty} P_{ij}[f(i)]L(j, f(j)) \\ &\quad +\alpha^2\sum_j\sum_k P_{ij}[f(i)]P_{jk}[f(j)]\gamma(k)\end{aligned}$$

So $T_f^2\gamma$ has the interpretation of being the cost incurred by applying the policy f for two periods and then incurring a cost $\alpha^2\gamma$ from then on. By induction, it is not difficult to show that $T_f^n\gamma$ corresponds to the cost incurred after n steps of applying the policy f and incurring a final cost of $\alpha^n\gamma$. Since γ is bounded, $\alpha^n\gamma \underset{n\to\infty}{\longrightarrow} 0$, and the result follows.

We can now prove:

Theorem 6.2 *(Optimality Theorem). Let ϕ_α be the stationary policy which, if the state of the process is i, selects the control which minimizes the right hand side of (6.2), i.e.,*

$$L(i, \phi_\alpha(i)) + \alpha\sum_j P_{ij}[\phi_\alpha(i)]V_\alpha(j) = \min_u[L(i, u) + \alpha\sum_j P_{ij}(u)V_\alpha(j)]$$

Then $V_{\phi_\alpha}(i) = V_\alpha(i)$ for all i and hence ϕ_α is optimal.

Proof: It is easily seen that in terms of the mapping T_{ϕ_α} introduced in the above, (6.2) can simply be written as

$$T_{\phi_\alpha}V_\alpha = V_\alpha$$

Hence

$$T_{\phi_\alpha}^n V_\alpha = V_\alpha$$

But since $V_\alpha \in B(I)$, we have from property (3) then $T_{\phi_\alpha}^n V_\alpha \to V_{\phi_\alpha}$.

So $V_{\phi_\alpha} = V_\alpha$ as claimed, and ϕ_α is optimal.

From Theorem 6.2, we see that once V_α is found, the optimal policy is easily obtained. To find V_α, we introduce the following:

For any function $g \in B(I)$, we define $\|g\| = \sup|g(i)|$.

A mapping $T : B(I) \to B(I)$ is called a contraction if there exists a constant $\beta < 1$, such that

$$\|Tg - Th\| \leq \beta\|g - h\| \qquad \text{for all } g, h \in B(I)$$

The following result, the Contraction Mapping Theorem, is well-known.

If $T : B(I) \to B(I)$ is a contraction mapping, then there exists a unique function $g \in B(I)$ such that $Tg = g$. Furthermore, for any $h \in B(I)$, $T^n h \to g$ as $n \to \infty$. For a proof, see for example, D.G. Luenberger, *Optimization by Vector Space Methods*.

Now define the mapping $T_\alpha : B(I) \to B(I)$ by

$$(T_\alpha g)(i) = \min_u \left[L(i,u) + \alpha \sum_j P_{ij}(u) g(j) \right] \tag{6.6}$$

Equation (6.2) is then equivalent to

$$T_\alpha V_\alpha = V_\alpha \tag{6.7}$$

We show that T_α is a contraction. For any $g, h \in B(I)$,

$$\begin{aligned}
(T_\alpha g)(i) - (T_\alpha h)(i) &= \min_u \left\{ L(i,u) + \alpha \sum_j P_{ij}(u) g(j) \right\} \\
&\quad - \min_u \left\{ L(i,u) + \alpha \sum_j P_{ij}(u) h(j) \right\} \\
&= \min_u \left\{ L(i,u) + \alpha \sum_j P_{ij}(u) g(j) \right\} - L(i,u_0) - \alpha \sum P_{ij}(u_0) h(j)
\end{aligned}$$

where u_0 is the control at which the minimum is attained. Thus

$$\begin{aligned}
(T_\alpha g)(i) - (T_\alpha h)(i) &\leq \alpha \sum_j P_{ij}(u_0)[g(j) - h(j)] \\
&\leq \alpha \sum_j P_{ij}(u) \sup_j |g(j) - h(j)| \\
&= \alpha \|g - h\|
\end{aligned}$$

Since we can clearly reverse the roles of g and h, we also get

$$(T_\alpha h)(i) - (T_\alpha g)(i) \leq \alpha \|g - h\|$$

Hence

$$\sup_i |(T_\alpha g)(i) - (T_\alpha h)(i)| \leq \alpha \|g - h\|$$

Thus

$$\|T_\alpha g - T_\alpha h\| \leq \alpha \|g - h\|$$

so that T_α is indeed a contraction.

By the contraction mapping theorem, (6.7) has a unique solution V_α, which may be obtained by iterating T_α on any function in $B(I)$. A particularly convenient choice is to use the iteration $T_\alpha^n \theta$ where θ is the identically zero function. The method of finding V_α by iteratively applying T_α to any bounded function is called successive approximation.

While the method of successive approximations may be used in principle to compute V_α, in practice we will only get an approximation to V_α after a finite number of iterations. However, if $T_\alpha^n \theta$ is close enough to V_α, then the policy obtained by minimizing

$$L(i,u) + \alpha \sum_j P_{ij}(n)(T_\alpha^n \theta)(j)$$

may be the optimal policy. If we denote the policy obtained by ϕ_n, we can evaluate V_{ϕ_n} by applying the policy ϕ_n in the equation

$$T_{\phi_n} V_{\phi_n} = V_{\phi_n}$$

If V_{ϕ_n} so obtained in fact also satisfies the dynamic programming equation (6.2), then by uniqueness $V_{\phi_n} = V_\alpha$ and ϕ_n is optimal. Later now, we shall illustrate this method with an example.

Another method of finding the optimal policy and the optimal cost is obtained by the following construction.

Suppose for any stationary policy f, we have evaluated V_f. Define f^* as the policy which selects the control that minimizes $[L(i,u) + \alpha \sum_j P_{ij}(u) V_f(j)]$ for each i. We claim that f^* is at least as good as f. For,

$$\begin{aligned}(T_{f^*} V_f)(i) &= L[i, f^*(i)] + \alpha \sum_j P_{ij}[f^*(i)] V_f(j) \\ &\leq L[i, f(i)] + \alpha \sum_j P_{ij}[f(i)] V_f(j) \\ &= V_f(i)\end{aligned}$$

By monotonicity, $(T^n_{f^*} V_f)(i) \leq V_f(i)$. Letting $n \to \infty$, we obtain $V_{f^*}(i) \leq V_f(i)$.

We can in fact make a stronger statement: f^* will either be a policy which is strictly better than f, i.e., $V_{f^*}(i) < V_f(i)$ for some i, or else f^* and f are both optimal. For, if $V_{f^*} = V_f$, then the policy f^* is characterized by selecting that u which minimizes

$$L(i,u) + \alpha \sum_j P_{ij}(u) V_{f^*}(j)$$

Since $T_{f^*} V_{f^*} = V_{f^*}$, we have

$$\begin{aligned}V_{f^*}(i) &= L(i, f^*(i)) + \alpha \sum_j P_{ij}(f^*(i)) V_{f^*}(j) \\ &= \min_u \left[L(i,u) + \alpha \sum_j P_{ij}(u) V_{f^*}(j) \right]\end{aligned}$$

Hence V_{f^*} satisfies (6.2) so that by uniqueness

$$V_{f^*} = V_f = V_\alpha$$

This method of successively improving the policy until the optimal one is obtained is called approximation in policy space, or the policy improvement algorithm. Note that if the state space is finite rather than countable, then the policy improvement algorithm leads to an optimal policy after a finite number of iterations. This is due to the fact that there are only a *finite* number of policies in the case where the state space is finite. Since the above results show that a strict improvement is obtained with each iteration, no repetitions will occur. At some point, no improvements will be possible and the optimal policy will be obtained.

6.3 An Example

We give an example which illustrates the steps involved in obtaining the optimal policy using either successive approximation or the policy improvement algorithm. Let the state space by $\{0, 1\}$ and the

control space be $\{1, 2\}$. The per stage costs are given by

$$\begin{bmatrix} L(0,1) & L(0,2) \\ L(1,1) & L(1,2) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix}$$

The transition probabilities are given by

$$\begin{bmatrix} P_{00}(1) & P_{00}(2) & P_{01}(1) & P_{01}(2) \\ P_{10}(1) & P_{10}(2) & P_{11}(1) & P_{11}(2) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} \\ \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

α is taken to be $\frac{1}{2}$.

Let us define the successive approximations by v_i, $i = 0, 1, 2, ...$ Then we have the following recurrence relations

$$v_{n+1} = T_\alpha v_n \tag{6.8}$$

Letting $v_0(0) = v_0(1) = 0$, we have

$$\begin{aligned} v_1(0) &= \min_{u\in\{1,2\}} \left[L(0,u) + \frac{1}{2}\sum_{j=0}^{1} P_{0j}(u) v_0(j) \right] \\ &= \min_{u\in\{1,2\}} [L(0,u)] = 0 \end{aligned}$$

Similarly, $v_1(1) = \min[L(1,u)] = 2$.

Applying (6.8) to v_1 yields

$$\begin{aligned} v_2(0) &= \min_{u\in\{1,2\}} \left[L(0,u) + \frac{1}{2}\sum_{j} P_{0j}(u) v_1(j) \right] \\ &= \min\left\{ 1 + \frac{1}{2}\left(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 2\right),\ 0 + \frac{1}{2}\left(\frac{1}{4}\cdot 0 + \frac{3}{4}\cdot 2\right) \right\} \\ &= \min\left\{ \frac{3}{2}, \frac{3}{4} \right\} = \frac{3}{4} \end{aligned}$$

and

$$\begin{aligned} v_2(1) &= \min_{u\in\{1,2\}} \left[L(1,u) + \frac{1}{2}\sum_{j} P_{1j}(u) v_1(j) \right] \\ &= \min\left\{ 2 + \frac{1}{2}\left(\frac{1}{3}\cdot 2\right), 2 + \frac{1}{2}\cdot\frac{2}{3}\cdot 2 \right\} \\ &= \min\left\{ \frac{7}{3}, \frac{8}{3} \right\} = \frac{7}{3} \end{aligned}$$

Iterating again gives

$$v_3(0) = \frac{31}{32} \qquad \text{and} \qquad v_3(1) = \frac{95}{36}$$

Noting that v_3 is reasonably close to v_2, we use v_3 to compute a possible candidate ϕ_3 for the optimal policy. To do so, we apply Theorem 6.2 but with v_3 replacing V_α. We then get, if the state is 0,

$$\begin{aligned} L(0,1) + \alpha\sum_j P_{0j}(1) v_3(j) &= 1 + \frac{1}{2}\left(\frac{1}{2}\cdot\frac{31}{32} + \frac{1}{2}\cdot\frac{95}{36}\right) \\ > L(0,2) + \alpha\sum_j P_{0j}(2) v_3(j) &= \frac{1}{2}\left(\frac{1}{4}\cdot\frac{31}{32} + \frac{3}{4}\cdot\frac{95}{36}\right) \end{aligned}$$

so that

$$\phi_3(0) = 2$$

Similarly,

$$\phi_3(1) = 1$$

We check to see if ϕ_3 is in fact optimal. This requires the verification that V_{ϕ_3} satisfies (6.2). First, we compute V_{ϕ_3}

$$\begin{aligned} V_{\phi_3}(0) &= L[0, \phi_3(0)] + \alpha \sum_j P_{0j}[\phi_3(0)] V_{\phi_3}(j) \\ &= \frac{1}{2}\left[\frac{1}{4} V_{\phi_3}(0) + \frac{3}{4} V_{\phi_3}(1)\right] \qquad (6.9) \\ & \qquad (6.10) \end{aligned}$$

and

$$V_{\phi_3}(1) = 2 + \frac{1}{2}\left[\frac{2}{3} V_{\phi_3}(0) + \frac{1}{3} V_{\phi_3}(1)\right] \qquad (6.11)$$

Solving these equations give

$$V_{\phi_3}(0) = \frac{36}{29}, \qquad V_{\phi_3}(1) = \frac{84}{29}$$

On putting into (6.2), we see indeed that V_{ϕ_3} satisfies (6.2). Hence by uniqueness $V_{\phi_3} = V_\alpha$ so that ϕ_3 is the optimal policy ϕ_α.

Let us compute the optimal policy using policy improvement. We begin with the arbitrary choice of $\phi_0(0) = 1$, $\phi_0(1) = 1$. We then get

$$\begin{aligned} V_{\phi_0}(0) &= 1 + \frac{1}{2}\left(\frac{1}{2} V_{\phi_0}(0) + \frac{1}{2} V_{\phi_0}(1)\right) \qquad (6.12) \\ & \qquad (6.13) \\ V_{\phi_0}(1) &= 2 + \frac{1}{2}\left[\frac{2}{3} V_{\phi_0}(0) + \frac{1}{3} V_{\phi_0}(1)\right] \qquad (6.14) \\ & \qquad (6.15) \end{aligned}$$

Solving (6.12) and (6.14) yields

$$V_{\phi_0}(0) = \frac{32}{13}, \qquad V_{\phi_0}(1) = \frac{44}{13}$$

We now find ϕ_1 by minimizing

$$\left[L(i,u) + \frac{1}{2}\sum_{j=0}^{1} P_{ij}(u) V_{\phi_0}(j)\right] \qquad \text{for each } i$$

Now

$$\begin{aligned} L(0,1) + \frac{1}{2}\sum_{j=0}^{1} P_{0j}(1) V_{\phi_0}(j) &= 1 + \frac{1}{2}\left[\frac{1}{2}\cdot\frac{32}{13} + \frac{1}{2}\cdot\frac{44}{13}\right] = \frac{32}{13} \\ > L(0,2) + \frac{1}{2}\sum_{j=0}^{1} P_{0j}(u) V_{\phi_0}(j) &= 0 + \frac{1}{2}\left[\frac{1}{4}\cdot\frac{32}{13} + \frac{3}{4}\cdot\frac{44}{13}\right] = \frac{4+\frac{33}{2}}{13}\ , \end{aligned}$$

while

$$\begin{aligned} L(1,2) \quad + \quad & \frac{1}{2}\sum_{j=0}^{1} P_{1j}(2)V_{\phi_0}(j) = 2 + \frac{1}{2}\left[\frac{1}{3}\cdot\frac{32}{13} + \frac{2}{3}\cdot\frac{44}{13}\right] \\ = \quad & \frac{46}{13} > L(1,1) + \frac{1}{2}\sum_{j=0}^{1} P_{1j}(1)V_{\phi_0}(j) = \frac{44}{13}\ . \end{aligned}$$

Hence $\phi_1(0) = 2$, $\phi_1(1) = 1$ is an improved policy.

Solving for V_{ϕ_1} gives $V_{\phi_1}(0) = \frac{36}{29}$, $V_{\phi_1}(1) = \frac{84}{29}$ and a second iteration on the policy improvement algorithm yields the same policy as ϕ_1. Hence $\phi_1 = \phi_\alpha$ is optimal, a conclusion we had obtained previously using successive approximation.

6.4 Positive Cost Criterion

In this section, we consider the positive cost criterion in which all per stage costs are nonnegative, i.e. $L(i,u) \geq 0$, all i and u. Also, we do not assume $L(i,u)$ to be bounded for all i, u.

As before, let

$$V_\Phi(i) = E^i_\Phi\left[\sum_{t=0}^{\infty} L(x_t, u_t)\right]$$

and

$$V_p(i) = \inf_\Phi V_\Phi(i) \qquad i \geq 0\ .$$

A policy Φ^* is optimal if $V_{\Phi^*}(i) = V_p(i)$, $\forall i$. Of course, $V_p(i)$ may be infinite even if the per stage costs are finite. Thus the above control problem is only meaningful if $V_p(i) < \infty$ for some initial states i.

The following results may be proved in the same way as in the discounted cost problem.

Theorem 6.3 *The optimal cost satisfies the equation*

$$V_p(i) = \min_u[L(i,u) + \sum_j P_{ij}(u)V_p(j)] \quad \forall i \tag{6.16}$$

Denote by $N(I)$ the set of all nonnegative functions on the state space, and define, for any stationary policy $\Phi = \{f, f, ...\}$

$$T_f : N(I) \to N(I)$$

by

$$(T_f\gamma)(i) = L[i, f(i)] + \sum_j P_{ij}(f(i))\gamma(j)$$

Then we have

(i) T_f is monotone

(ii) $T_f V_f = V_f$

(iii) $(T_f^n\theta)(i) \underset{n\to\infty}{\longrightarrow} V_f(i)$ for each i, where θ denotes the identically zero function.

The only difference between the above 3 properties and the analogous results in the discounted cost case is that $(T_f^n\gamma)(i) \to V_f(i)$ only if $\gamma = \theta$. This is due to the fact that in the discounted cost case, we incur final costs of the form $\alpha^n\gamma$. In this case, the discount factor α is absent, and we cannot guarantee the final cost to go to zero unless it is zero.

Theorem 6.4 *(Optimality Theorem) Let ϕ_p be the stationary policy which, if the state is i, selects the control which minimizes the right hand side of (6.16), then $V_{\phi_p}(i) = V_p(i)$, and hence ϕ_p is optimal.*

Proof: By applying $_{\phi_p}$ to V_p, we get

$$T_{\phi_p} V_p = V_p$$

But

$$T_{\phi_p}\theta \le T_{\phi_p} V_p = V_p$$

Hence

$$T_{\phi_p}^n \theta \le V_p$$

On letting $n \to \infty$, we get $V_{\phi_p} \le V_p$. But since $V_p \le V_{\phi_p}$, this proves the desired result.

Although Theorems 6.3 and 6.4 are similar to Theorems 6.1 and 6.2, they are more difficult to use in practice. The successive approximation method does converge to the optimal cost function for the finite control set case. However, without the discount factor, we do not in general have a geometric rate of convergence. Furthermore, the policy improvement method does not necessarily converge to the optimal control law. This is because without the discount factor, uniqueness of solution to the dynamic programming equation (6.16) is not guranateed. For additional details on this problem, see D.P. Bertsekas, *Dynamic Programming and Stochastic Control*.

6.5 Average Cost per Unit Time Criterion

The cost criterion we now consider is the average cost per unit time. For any policy Φ, we define

$$V_\Phi(i) = \lim_{T\to\infty} \frac{1}{T} E_i^\Phi \sum_{t=0}^{T-1} L(x_t, u_t)$$

This problem turns out to be more difficult than the discounted or positive cost problems. It is not necessarily true that an optimal policy exists, and it may also happen that a nonstationary policy is strictly better than a stationary policy (see the counterexamples in S.M. Ross, *Applied Probability Models with Optimization Applications*). We shall give a sufficient condition for the existence of an optimal stationary policy.

Theorem 6.5 *If there exists a bounded function h defined on the nonnegative integers and a constant λ such that*

$$\lambda + h(i) = \min_u \left[L(i,u) + \sum_j P_{ij}(u) h(j) \right] \tag{6.17}$$

then there exists a stationary policy Φ_a such that

$$\lambda = V_{\Phi_a}(i) = \inf_\Phi V_\Phi(i) \qquad \text{for all } i \ge 0,$$

where Φ_a is the policy which, for each i, selects an action which minimizes the right hand side of (6.17).

Proof: For any policy Φ, we have

$$E_i^\Phi \left\{ \sum_{t=1}^{T} [h(x_t) - E^\Phi(h(x_t)|x_{t-1}, u_{t-1})] \right\} = 0 \tag{6.18}$$

But

$$\begin{aligned} E^{\Phi}[h(x_t)|x_{t-1}, u_{t-1}] &= \sum_{j=0}^{\infty} h(j) P_{x_{t-1},j}(u_{t-1}) \\ &= \sum_{j=0}^{\infty} h(j) P_{x_{t-1},j}(u_{t-1}) + L(x_{t-1}, u_{t-1}) - L(x_{t-1}, u_{t-1}) \\ &\geq \min_{u} \left\{ L(x_{t-1}, u) + \sum_{j} P_{x_{t-1},j}(u) h(j) \right\} - L(x_{t-1}, u_{t-1}) \\ &= \lambda + h(x_{t-1}) - L(x_{t-1}, u_{t-1}) \qquad (6.19) \\ & \qquad (6.20) \end{aligned}$$

with equality if $\Phi = \Phi_a$.

Hence from (6.18) and (6.19) we get

$$0 \leq E_i^{\Phi} \left\{ \sum_{t=1}^{T} [h(x_t) - \lambda - h(x_{t-1}) + L(x_{t-1}, u_{t-1})] \right\}$$

or

$$\lambda \leq E_i^{\Phi} \frac{h(x_T)}{T} - E_i^{\Phi} \frac{h(x_0)}{T} + \frac{1}{T} E_i^{\Phi} \sum_{t=1}^{T} L(x_{t-1}, u_{t-1})$$

On letting $T \to \infty$, we obtain

$$\lambda \leq \lim_{T \to \infty} \frac{1}{T} E_i^{\Phi} \sum_{t=1}^{T} L(x_{t-1}, u_{t-1}) \qquad (6.21)$$

Since equality is achieved in (6.21) with Φ_a, the theorem is proved.

An illustration of Theorem 6.5 is given in Problem 5.6 for the linear stochastic regulator problem.

It is possible to relate the average cost per unit time problem to the discounted cost problem. It is also possible to formulate a successive approximation and a policy improvement algorithm. We shall not go into the details but refer the reader to Bertsekas.

6.6 Exercises

1. A Toymaker can be in one of two states: state 1 corresponds to having a successful toy, state 2 an unsuccessful one. If he is in state 1, he can choose to advertise (1) or not advertise (2). If he is in state 2, he can choose to (1) or not to (2) do research. The transition probabilities and rewards are given as follows:

State i	Action u	Transition $P_{i1}(u)$	Probabilities $P_{i2}(u)$	Rewards $L(i,u)$
1	1	0.5	0.5	6
	2	0.8	0.2	4
2	1	0.4	0.6	-3
	2	0.7	0.3	-5

Suppose the objective of the toymaker is to *maximize* the discounted reward criterion $E\sum_{k=0}^{\infty}\alpha^k L(x_k,u_k)$ for $\alpha = 0.9$.

Determine the optimal policy using

(a) Successive approximation starting with $v_0(i) = 0$, $i = 1, 2$.

(b) Policy improvement starting with the policy $\phi(i) = 1$, $i = 1, 2$.

2. The Markov chain x_k under control has two states 1,2. In state 1, there are 3 possible actions: 1, 2, 3. For $n = 1, 2, 3$, action n yields a return of $\frac{5-n}{4}$ and x_k remains in state 1 with probability $\frac{n-1}{2}$. In state 2, there is only one action, 1, that yields a return of 0 and the system stays in state 2 with probability 1. Suppose the criterion is to *maximize* the average discounted return

$$J = E\sum_{k=0}^{\infty}\alpha^k L(x_k,u_k)$$

where $L(x_k,u_k)$ is the return at state x_k with action u_k. Show that the return function under the 3 possible stationary policies ϕ_1, ϕ_2, ϕ_3 where

$$\phi_1(1) = 1, \quad \phi_2(1) = 2, \quad \phi_3(1) = 3$$

$$\phi_1(2) = 1, \quad \phi_2(2) = 1, \quad \phi_3(2) = 1$$

are given respectively by $V_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,

$$V_2 = \begin{bmatrix} \frac{3/4}{1-\alpha/2} \\ 0 \end{bmatrix} \quad \text{and} \quad V_3 = \begin{bmatrix} \frac{1}{2(1-\alpha)} \\ 0 \end{bmatrix}$$

Obtain the optimal control law as a function of α, $0 < \alpha < 1$.

3. Suppose the process x_k satisfies the equation

$$x_{k+1} = w_k u_k$$

where the control u_k is constrained to lie in $[0, x_k]$, is w_k is a nonnegative, independent, identically distributed sequence. The control problem is to find a policy which maximizes the discounted cost criterion $E\sum_{k=0}^{\infty} \alpha^k (x_k - u_k)^{\frac{1}{2}}$ where $0 < \alpha < 1$. The distribution of w is assumed to be such that $\alpha E w_k^{\frac{1}{2}} < 1$.

(a) Derive the dynamic programming equation for the optimal value function for this problem.

(b) The equation in (a) can be expressed in the form

$$V = TV \qquad (DP)$$

The successive approximation algorithm attempts to solve (DP) by computing $v_n = T^n 0$ and taking the limit. Compute $v_1(x)$ and $v_2(x)$.

(c) From the form of v_1 and v_2, guess the form of the function $V(x)$ and solve (DP) explicitly.

(d) Show that the function $V(x)$ determined in (c) is the optimal value function and hence determine the optimal control law.