PROBABILITY AND MATHEMATICAL STATISTICS

A Series of Monographs and Textbooks

INTRODUCTION TO STOCHASTIC DYNAMIC PROGRAMMING

SHELDON M. ROSS

ACADEMIC PRESS

II

Discounted Dynamic Programming

1. Introduction

Consider a process that is observed at time points $n = 0, 1, 2, \ldots$ to be in one of a number of possible states. The set of possible states is assumed to be countable and will be labeled by the nonnegative integers $0, 1, 2, \ldots$. After observing the state of the process, an action must be chosen, and we let A (assumed finite) denote the set of all possible actions.

If the process is in state i at time n and action a is chosen, then, independent of the past, two things occur:

(i) We receive an expected reward $R(i, a)$.

(ii) The next state of the system is chosen according to the transition probabilities $P_{ij}(a)$.

If we let X_n denote the state of the process at time n and a_n the action chosen at that time, then assumption (ii) is equivalent to stating that

$$P\{X_{n+1} = j | X_0, a_0, X_1, a_1, \ldots, X_n = i, a_n = a\} = P_{ij}(a).$$

Thus both the rewards and the transition probabilities are functions only of the last state and the subsequent action. Furthermore, we suppose that the costs are bounded, and we let B be such that $|R(i, a)| < B$ for all i and a.

To choose actions we must follow some policy. We place no restrictions on the class of allowable policies, and we therefore define a *policy* to be any rule for choosing actions. Thus the action chosen by

a policy, for instance, may depend on the history of the process up to that point or it may be randomized in the sense that it chooses action a with some probability P_a, $a \in A$.

An important subclass of the class of all policies is the class of stationary policies. Here, a policy is said to be *stationary* if it is non-randomized and the action it chooses at time t only depends on the state of the process at time t. In other words, a stationary policy is a function f mapping the state space into the action space, with the interpretation that for each state i, $f(i)$ denotes the action the policy chooses when in state i. It follows that if a stationary policy f is employed, then the sequence of states $\{X_n, n = 0, 1, 2, \ldots\}$ forms a Markov chain with transition probabilities $P_{ij} = P_{ij}(f(i))$; and it is for this reason that the process is called a Markov decision process.

To determine policies that are in some sense optimal, we first need to decide on an optimality criterion. In this chapter we use the total expected discounted return as our criterion. This criterion assumes a discount factor α, $0 < \alpha < 1$, and, among all policies π, attempts to maximize

$$V_\pi(i) = E_\pi\left[\sum_{n=0}^{\infty} R(X_n, a_n)\alpha^n \middle| X_0 = i\right], \tag{1.1}$$

where E_π represents the conditional expectation, given that policy π is employed. Because $R(X_n, a_n)$ is just the reward earned at time n, it follows that $V_\pi(i)$ represents the expected total discounted return earned when the policy π is employed and the initial state is i. [Note that (1.1) is well defined because rewards are bounded and $\alpha < 1$, which implies that $|V_\pi(i)| < B/(1 - \alpha)$.]

2. The Optimality Equation and Optimal Policy

The use of a discount factor is economically motivated by the fact that a reward to be earned in the future is less valuable than one earned today. Let

$$V(i) = \sup_\pi V_\pi(i).$$

A policy π^* is said to be *α-optimal* if

$$V_{\pi^*}(i) = V(i) \qquad \text{for all } i \geq 0.$$

Hence, a policy is α-optimal if its expected α-discounted return is maximal for every initial state.

The following theorem yields a functional equation satisfied by the optimal value function V.

Theorem 2.1 The Optimality Equation

$$V(i) = \max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V(j)\right], \qquad i \geq 0. \tag{2.1}$$

Proof: Let π be any arbitrary policy, and suppose that π chooses action a at time 0 with probability P_a, $a \in A$. Then,

$$V_\pi(i) = \sum_{a \in A} P_a\left[R(i, a) + \sum_j P_{ij}(a)W_\pi(j)\right],$$

where $W_\pi(j)$ represents the expected discounted return from time 1 onward, given that policy π is being used and that the state at time 1 is j. However, if the state at time 1 is j, the situation at this time is the same as if the process had started in state j, with the exception that all returns are now multiplied by α. Hence,

$$W_\pi(j) \leq \alpha V(j),$$

and thus

$$\begin{aligned} V_\pi(i) &\leq \sum_{a \in A} P_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V(j)\right] \\ &\leq \sum_{a \in A} P_a \max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V(j)\right] \\ &= \max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V(j)\right]. \end{aligned} \tag{2.2}$$

Because π is arbitrary, (2.2) implies that

$$V(i) \leq \max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V(j)\right]. \tag{2.3}$$

To go the other way, let a_0 be such that

$$R(i, a_0) + \alpha \sum_j P_{ij}(a_0)V(j) = \max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V(j)\right]. \tag{2.4}$$

Let π be the policy that chooses a_0 at time 0 and, if the next state is j,

then views the process as originating in state j, following a policy π_j such that $V_{\pi_j}(j) \geq V(j) - \epsilon$. Hence,

$$V_\pi(i) = R(i, a_0) + \alpha \sum_j P_{ij}(a_0) V_{\pi_j}(j)$$
$$\geq R(i, a_0) + \alpha \sum_j P_{ij}(a_0) V(j) - \alpha\epsilon,$$

which, because $V(i) \geq V_\pi(i)$, implies that

$$V(i) \geq R(i, a_0) + \alpha \sum_j P_{ij}(a_0) V(j) - \alpha\epsilon.$$

Hence, from (2.4) we obtain

$$V(i) \geq \max_a [R(i, a) + \alpha \sum_j P_{ij}(a) V(j)] - \alpha\epsilon, \tag{2.5}$$

and the result follows from (2.3) and (2.5) because ϵ is arbitrary. □

We are now ready to prove the important result that the policy determined by the optimality equation is optimal.

Theorem 2.2 Let f be the stationary policy that, when the process is in state i, selects the action (or an action) maximizing the right side of (2.1), that is, $f(i)$ is such that

$$R(i, f(i)) + \alpha \sum_j P_{ij}(f(i)) V(j) = \max_a [R(i, a) + \alpha \sum_j P_{ij}(a) V(j)], \quad i \geq 0.$$

Then

$$V_f(i) = V(i) \quad \text{for all } i \geq 0,$$

and hence f is α-optimal.

Proof: Because

$$V(i) = \max_a [R(i, a) + \alpha \sum_j P_{ij}(a) V(j)]$$
$$= R(i, f(i)) + \alpha \sum_j P_{ij}(f(i)) V(j), \tag{2.6}$$

we see that V is equal to the expected discounted return of a two-stage problem in which we use f for the first stage and then receive a terminal reward $V(j)$ (if we end in state j). But because this terminal reward has the same value as using f for another stage and then receiving the terminal reward V, we see that V is equal to the expected reward of a

three-stage problem in which we use f for two stages and then receive the terminal reward V. Continuing this argument shows that

$$V(i) = E(n\text{-stage return under } f | X_0 = i) + \alpha^n E(V(X_n) | X_0 = i).$$

Letting $n \to \infty$, we obtain, using $V(j) < B/(1 - \alpha)$ and $0 < \alpha < 1$,

$$V(i) = V_f(i),$$

which proves the theorem. □

Technical Remark: The preceding proof can be stated more formally as follows: For any stationary policy g, define the operator T_g mapping bounded functions on the state space into itself in the following manner. For any bounded function $u(i)$ $(i = 0, 1, \ldots)$, $T_g u$ is defined as that function whose value at i is given by

$$(T_g u)(i) = R(i, g(i)) + \alpha \sum_j P_{ij}(g(i)) u(j).$$

Thus $T_g u$ evaluated at i represents the expected discounted return if the initial state is i and we employ g for one stage and are then terminated with a final return $u(j)$ (if the final state is j).

It is easy to show that for bounded functions u and v

(i) $u \leq v \Rightarrow T_g u \leq T_g v$,
(ii) $T_g^n u \to V_g$ as $n \to \infty$,

where $T_g^1 u = T_g u$, $T_g^n u = T_g(T_g^{n-1} u)$, $n \geq 1$.

If f is the policy chosen by the optimality equation, we have by (2.1) that

$$T_f V = V,$$

implying that

$$T_f^n V = T_f^{n-1}(T_f V) = T_f^{n-1} V = \cdots = V,$$

and, letting $n \to \infty$,

$$V_f = V. \quad \square$$

The following proposition shows that V is the unique bounded solution of the optimality equation.

Proposition 2.3 V is the unique bounded solution of the optimality equation (2.1).

Proof: Suppose that $u(i)$, $i \geq 0$, is a bounded function that satisfies the optimality equation

$$u(i) = \max_a [R(i, a) + \alpha \sum_j P_{ij}(a)u(j)], \qquad i \geq 0.$$

For fixed i let $\bar{a}$ be such that

$$u(i) = R(i, \bar{a}) + \alpha \sum_j P_{ij}(\bar{a})u(j).$$

Hence, because V satisfies the optimality equation, we have

$$\begin{aligned} u(i) - V(i) &= R(i, \bar{a}) + \alpha \sum_j P_{ij}(\bar{a})u(j) - \max_a [R(i, a) + \alpha \sum_j P_{ij}(a)V(j)] \\ &\leq \alpha \sum_j P_{ij}(\bar{a})[u(j) - V(j)] \\ &\leq \alpha \sum_j P_{ij}(\bar{a})|u(j) - V(j)| \\ &\leq \alpha \sum_j P_{ij}(\bar{a}) \sup_j |u(j) - V(j)| \\ &= \alpha \sup_j |u(j) - V(j)|. \end{aligned}$$

By reversing the roles of u and V we can similarly conclude that

$$V(i) - u(i) \leq \alpha \sup_j |V(j) - u(j)|.$$

Therefore

$$|V(i) - u(i)| \leq \alpha \sup_j |V(j) - u(j)|,$$

so

$$\sup_i |V(i) - u(i)| \leq \alpha \sup_j |V(j) - u(j)|,$$

implying (because $\alpha < 1$) that

$$\sup_j |V(j) - u(j)| = 0. \quad \square$$

The following will also be needed later.

Proposition 2.4 For any stationary policy g, V_g is the unique solution of

$$V_g(i) = R(i, g(i)) + \alpha \sum_j P_{ij}(g(i))V_g(j). \tag{2.7}$$

Proof: It is immediate that V_g satisfies (2.7) because $R(i, g(i))$ is the one-stage return and $\alpha \sum_j P_{ij}(g(i))V_g(j)$ is the expected additional return obtained by conditioning on the next state visited. That it is the unique solution follows exactly as in the proof of Proposition 2.3. (In fact, we can use Proposition 2.3 directly by considering a problem in which $g(i)$ is the only action available in state i, so V_g is the optimal value function for that problem.)

Remark: In operator notation, (2.7) states that

$$T_g V_g = V_g.$$

3. Method of Successive Approximations

It follows from Theorem 2.2 that, if we could determine the optimal value function V, then we would know the optimal policy—it would be the stationary policy that, when in state i, chooses an action that maximizes

$$R(i, a) + \alpha \sum_j P_{ij}(a)V(j).$$

In this section we show how V can be obtained as a limit of the n-stage optimal return.

As a prelude, note that for any policy π and initial state j,

$$\begin{aligned} &|E_\pi[\text{return from time } (n+1) \text{ onwards} | X_0 = j]| \\ &\quad = \left| E_\pi \left[\sum_{i=n+1}^{\infty} \alpha^i R(X_i, a_i) | X_0 = j \right] \right| \\ &\quad \leq \frac{\alpha^{n+1} B}{1 - \alpha}. \end{aligned} \tag{3.1}$$

The method of successive approximations is as follows: Let $V_0(i)$ be any arbitrary bounded function, and define V_1 by

$$V_1(i) = \max_a [R(i, a) + \alpha \sum_j P_{ij}(a)V_0(j)].$$

In general, for $n > 1$, let

$$V_n(i) = \max_a [R(i, a) + \alpha \sum_j P_{ij}(a)V_{n-1}(j)].$$

It is worthwhile noting that V_n is the maximal expected discounted return of an n-stage problem that confers a terminal reward $V_0(j)$ if the process ends in state j. The following proposition shows that V_n converges uniformly to V as $n \to \infty$.

Proposition 3.1

(i) If $V_0 \equiv 0$, then $|V(i) - V_n(i)| \leq \alpha^{n+1}B/(1-\alpha)$.
(ii) For any bounded V_0, $V_n(i) \to V(i)$ uniformly in i as $n \to \infty$.

Proof: Suppose $V_0 \equiv 0$, so $V_n(i)$ equals the maximal expected return in an n-stage problem starting in i. Now, for the α-optimal policy f,

$$V(i) = E_f(\text{return during first } n\text{-stages}) + E_f(\text{additional returns})$$
$$\leq V_n(i) + \alpha^{n+1}B/(1-\alpha),$$

where the inequality follows from (3.1) and the definition of V_n. To go the other way, note that V must be larger than the expected return of the policy that uses the n-stage optimal policy for the first n-stages and any arbitrary policy for the remaining time. Hence,

$$V(i) \geq V_n(i) + E[\text{additional return from } (n+1) \text{ onwards}]$$
$$\geq V_n(i) - \alpha^{n+1}B/(1-\alpha),$$

which, together with the preceding, proves (i).

To prove (ii) let V_n^0 denote V_n when $V_0 \equiv 0$. Then for any bounded V_0 we leave it for the reader to show that

$$|V_n(i) - V_n^0(i)| \leq \alpha^n \sup_j |V_0(j)|,$$

which together with (i) proves the result. □

Example 3.1 *A Machine Replacement Model* Suppose that at the beginning of each time period a machine is inspected and its condition or state is noted. After observing this state, a decision as to whether or not to replace the machine must be made. If the decision is to replace, then a cost R is immediately incurred and the state at the beginning of the next time period is 0, the state of a new machine. If the present state is i and a decision not to replace is made, then the state at the beginning of the next time period will be j with probability P_{ij}. In addition, each time the machine is in state i at the beginning of a time period, an operating cost $C(i)$ is incurred.

Let $V(i)$ denote the minimal expected total α-discounted *cost*, given

that the initial state is i. Then V satisfies the optimality equation

$$V(i) = C(i) + \min[R + \alpha V(0), \alpha \sum_j P_{ij}V(j)]. \tag{3.2}$$

Under what conditions on $C(i)$ and the transition probability matrix $\underline{P} = [P_{ij}]$ will $V(i)$ be increasing in i? First, it is clear that the operating costs must be increasing, so let us assume the following condition.

Condition 1: $C(i)$ is increasing in i.

However, Condition 1 by itself is insufficient to imply that $V(i)$ is increasing in i. It is possible for states i and i (where $i \geq i$) that whereas i has a higher operating cost than i it might take the process into a better state than would i. To ensure that this does not occur, we suppose that, under no replacement, the next state from i is stochastically increasing in i. That is, we have the following condition.

Condition 2: For each k, $\sum_{j=k}^{\infty} P_{ij}$ increases in i.

In other words, if T_i is a random variable representing the next state visited after i (assuming no replacement) then $P(T_i = j) = P_{ij}$, so Condition 2 states that†

$$T_i \underset{\text{st}}{\leq} T_{i+1}, \qquad i = 0, 1, \ldots.$$

Because this is equivalent to $E(f(T_i))$ increasing in i for all increasing functions f, Condition 2 is equivalent to the statement that

$$\sum_j P_{ij}f(j) \text{ increases in } i, \qquad \text{for all increasing } f.$$

We now prove by induction that, under Conditions 1 and 2, $V(i)$ increases in i.

Proposition 3.2 Under Conditions 1 and 2, $V(i)$ increases in i.

Proof: Let

$$V_1(i) = C(i),$$

and for $n > 1$

$$V_n(i) = C(i) + \min[R + \alpha V_{n-1}(0), \alpha \sum_j P_{ij}V_{n-1}(j)].$$

† See the Appendix for a discussion on stochastic order relations.

It follows from Condition 1 that $V_1(i)$ increases in i. Hence, assume that $V_{n-1}(j)$ increases in j, so, from Condition 2, $\sum_j P_{ij}V_{n-1}(j)$ increases in i, and thus $V_n(i)$ increases in i. Hence, by induction, $V_n(i)$ increases in i for all n, and because

$$V(i) = \lim_n V_n(i),$$

the result follows. □

The structure of the optimal policy is a simple consequence of Proposition 3.2.

Proposition 3.3 Under Conditions 1 and 2, there exists an $\bar{i}$, $\bar{i} \leq \infty$, such that the α-optimal policy replaces when the state is i if $i \geq \bar{i}$ and does not replace if $i < \bar{i}$.

Proof: It follows from the optimality equation (3.2) that it is optimal to replace in i if

$$\alpha \sum_j P_{ij}V(j) \geq R + \alpha V(0).$$

Because $V(j)$ increasing in j implies that $\sum_j P_{ij}V(j)$ increases in i, the result follows, with $\bar{i}$ being given by

$$\bar{i} = \min\left[i : \alpha \sum_j P_{ij}V(j) \geq R + \alpha V(0)\right],$$

where $\bar{i}$ is taken to be ∞ if the preceding set is empty.

4. Policy Improvement

We have seen that once V is determined the optimal policy is the one that, when in state i, chooses the action a to maximize $R(i, a) + \alpha \sum_j P_{ij}(a)V(j)$. Suppose that for some stationary policy g we have computed V_g, the expected return under g; and suppose that we now define h to be the policy that, when in state i, selects the action that maximizes $R(i, a) + \alpha \sum_j P_{ij}(a)V_g(j)$. How good is h compared with g? We now show that h is at least as good as g, and if it is not strictly better than g for at least one initial state, then g and h are both optimal.

Proposition 4.1 Let g be a stationary policy with expected return V_g and let h be the policy such that

$$R(i, h(i)) + \alpha \sum_j P_{ij}(h(i))V_g(j) = \max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V_g(j)\right]. \quad (4.1)$$

Then

$$V_h(i) \geq V_g(i) \quad \text{for all } i,$$

and if $V_h(i) = V_g(i)$ for all i, then $V_g = V_h = V$.

Proof: Because

$$\max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V_g(j)\right] \geq R(i, g(i)) + \alpha \sum_j P_{ij}(g(i))V_g(j) = V_g(i),$$

it follows from (4.1) that

$$R(i, h(i)) + \alpha \sum_j P_{ij}(h(i))V_g(j) \geq V_g(i) \quad \text{for all } i. \quad (4.2)$$

This inequality states that using h for one stage and then switching to g is better than using g throughout. However, because we can repeat this argument after the first stage (that is, at the moment when the first policy is about to switch to g), we see that using h for two stages and then switching to g is better than using g. Repeating this argument shows that using h for n stages and then switching to g is better than using only g; that is,

$$E_h\left[\sum_{j=0}^{n-1} \alpha^j R(X_j, a_j) | X_0 = i\right] + \alpha^n E_h[V_g(X_n) | X_0 = i] \geq V_g(i).$$

Letting $n \to \infty$ gives

$$V_h(i) \geq V_g(i).$$

Now suppose that $V_h(i) = V_g(i)$ for all i. Then, because

$$R(i, h(i)) + \alpha \sum_j P_{ij}(h(i))V_g(j) = V_h(i),$$

we see from (4.1) (upon substituting V_h for V_g) that

$$V_h(i) = \max_a\left[R(i, a) + \alpha \sum_j P_{ij}(a)V_h(j)\right].$$

Hence, V_h satisfies the optimality equation, and by uniqueness (Proposition 2.3), we conclude that $V_h = V$. □

Remark: In terms of operator notation, we have from (4.2) that

$$T_h V_g \geq V_g,$$

and successively applying T_h to both sides of the preceding inequality gives

$$T_h^n V_g \geq T_h^{n-1} V_g \geq \cdots \geq V_g,$$

implying (letting $n \to \infty$) that

$$V_h \geq V_g.$$

The preceding result gives us a computational approach to obtaining the optimal policy when the state space is finite. For instance, let the states be $1, 2, \ldots, n$. The optimal policy can be obtained by first choosing any stationary policy g. We then compute V_g as the unique solution of the set of equations

$$V_g(i) = R(i, g(i)) + \alpha \sum_j P_{ij}(g(i)) V_g(j), \qquad i = 1, \ldots, n.$$

Once we have solved this set of n equations in n unknowns and have thus obtained V_g, we then improve g by defining h as the policy that, when in state i, selects the action that maximizes $R(i, a) + \alpha \sum_j P_{ij}(a) V_g(j)$. We next solve for V_h and then improve h, and so on. Because there are only a finite number of possible stationary policies when the state space is finite, we shall eventually reach one for which no strict improvement is possible. This will be the optimal policy.

5. Solution by Linear Programming

If u is a bounded function on the state space satisfying

$$u(i) \geq \max_a [R(i, a) + \alpha \sum_j P_{ij}(a) u(j)], \qquad i \geq 0, \tag{5.1}$$

then we shall show that $u \geq V$.

Proposition 5.1 If u satisfies (5.1), then

$$u(i) \geq V(i) \qquad \text{for all } i.$$

Proof: Consider the usual model with the additional proviso of a stop action that, if exercised when in state i, earns one a terminal reward $u(i)$ and ends the problem. Now, (5.1) states that, for any initial state, stopping immediately is better than doing anything else for one stage and then stopping. If something else (aside from stopping) is done at the initial stage, then, also by (5.1), it is better to stop after the initial stage than it is to do anything else and then stop. Hence, stopping immediately is better than doing anything else for two stages and then stopping. Repeating this shows that stopping immediately is better than doing anything for n stages and then stopping. That is, for any policy π,

$$u(i) \geq E_\pi[n\text{-stage return} | X_0 = i] + \alpha^n E_\pi[u(X_n) | X_0 = i],$$

and upon letting $n \to \infty$, we obtain

$$u(i) \geq V_\pi(i),$$

which implies the result. □

Remark: If we define the operator T mapping bounded functions on the state space into itself by $T = \max_g T_g$; that is,

$$(Tu)(i) = \max_a [R(i, a) + \alpha \sum P_{ij}(a) u(j)],$$

then (5.1) states that $u \geq Tu$. Applying T to both sides of this inequality gives $Tu \geq T^2 u$, so $u \geq T^2 u$. Continuing this gives $u \geq T^n u$, and letting $n \to \infty$ and using Proposition 3.1, which states that $T^n u \to V$, gives the result.

Because the optimal value function V satisfies the inequality (as it also satisfies it with equality), it follows from Proposition 5.1 that it is the smallest function that satisfies (5.1). Hence, letting β be such that $0 < \beta < 1$, it follows that V will be the unique solution of the optimization problem

$$\min_u \left[\sum_{i=0}^{\infty} \beta^i u(i) \right],$$

$$\text{subject to} \quad u(i) \geq \max_a [R(i, a) + \alpha \sum_j P_{ij}(a) u(j)], \qquad i \geq 0,$$

or, equivalently,

$$\min_u \left[\sum_{i=0}^{\infty} \beta^i u(i) \right],$$

$$\text{subject to} \quad u(i) \geq R(i, a) + \alpha \sum_j P_{ij}(a) u(j), \qquad i \geq 0, \qquad a \in A.$$

However, this is a linear programming problem and, at least in the case of a finite state space, can be solved by a technique known as the simplex algorithm. In fact, in the case of a finite state space we can let $\beta = 1$ because we only imposed the condition $0 < \beta < 1$ to keep the objective function finite.

Remark: For finite-state problems we thus have two possible computational approaches. The linear programming solution just presented and the policy improvement technique of the previous section.

6. Extension to Unbounded Rewards

To ensure that $V_\pi(i)$, defined by

$$V_\pi(i) = E_\pi \left[\sum_{n=0}^{\infty} R(X_n, a_n) \alpha^n \middle| X_0 = i \right],$$

is well defined, we have assumed up to now that $R(i, a)$ is bounded. This can be generalized. For instance, suppose that for each i there exist numbers B_i and a constant k such that, starting in i, the expected reward at time $n - 1$ is bounded by $B_i n^k$, $n \geq 1$. That is, for $n \geq 1$, given that $X_0 = i$,

$$E_\pi[|R(X_{n-1}, a_{n-1})|] \leq B_i n^k, \qquad \text{for all policies } \pi. \tag{6.1}$$

Under this condition it follows that, conditional on $X_0 = i$,

$$\left| E_\pi \left[\sum_{n=0}^{\infty} R(X_n, a_n) \alpha^n \right] \right| \leq B_i \sum_{n=0}^{\infty} \alpha^n (n + 1)^k < \infty,$$

so V_π remains well defined.

Letting f denote the policy chosen by the optimality equation, then, as in the proof of Theorem 2.2.,

$$V(i) = E_f(n\text{-stage return}) + \alpha^n E_f(V(X_n)), \tag{6.2}$$

where these expectations are conditional on $X_0 = i$. Now, by condition (6.1) we have

$$\begin{aligned} |\alpha^n E_f(V(X_n))| &\leq B_i \alpha^n \sum_{j=0}^{\infty} \alpha^j (n + 1 + j)^k \\ &= B_i \alpha^n \sum_{j=0}^{\infty} \alpha^j \sum_{l=0}^{k} \binom{k}{l} (n + 1)^l j^{k-l} \\ &\to 0 \qquad \text{as} \quad n \to \infty, \end{aligned}$$

where the limit result follows because, for each $l = 0, 1, \ldots, k$,

$$\alpha^n (n + 1)^l \sum_{j=0}^{\infty} \alpha^j j^{k-l} \to 0 \qquad \text{as} \quad n \to \infty.$$

Hence, we see from (6.2), letting $n \to \infty$, that

$$V(i) = V_f(i),$$

so, as in the bounded case, f is optimal.

Policy improvement remains valid in the sense that the improved policy is at least as good as the original. That is, if g is a stationary policy and h is defined so that

$$R(i, h(i)) + \alpha \sum_j P_{ij}(h(i)) V_g(j) = \max_a \left[R(i, a) + \alpha \sum P_{ij}(a) V_g(j) \right],$$

then, exactly as in Proposition 4.1, we have

$$E_h[n\text{-stage return} | X_0 = i] + \alpha^n E_h[V_g(X_n) | X_0 = i] \geq V_g(i).$$

Letting $n \to \infty$ we obtain, using (6.1),

$$V_h(i) \geq V_g(i).$$

Hence, h is at least as good as g. However, it is no longer true that $V_g = V_h$ implies that h and g are optimal. The reason is that it is not necessarily true that V is the unique solution of the optimality equation (in the bounded reward case we could deal with *bounded* functions, and we showed that in the bounded case V is the unique bounded solution of the optimality equation). Of course, if the B_i were bounded, then V would also be bounded, so this result would remain true.

Remark: The preceding remains true even if we suppose that the value of k in (6.1) depends on i; that is, if

$$E_\pi(R(X_{n-1}, a_{n-1})) \leq B_i n^{k(i)}, \qquad n \geq 1,$$

where this expectation is conditional on $X_0 = i$.

Problems

1. Consider a problem in which one is interested in maximizing the total expected return. However, suppose that at the end of each time period there is a probability α that the problem ends. Show that this is equivalent to using an infinite-stage discounted-return criterion.
2. Prove the statements given in the technical remark following Theorem 2.2; that is, show

$$u \leq v \Rightarrow T_g u \leq T_g v,$$

and, for bounded u,

$$T_g^n u \to V_g.$$

3. Prove for any bounded V_0 that, in the successive approximation scheme,

$$|V_n(i) - V_n^0(i)| \leq \alpha^n \sup_j |V_0(j)|.$$

4. A quality control model: Consider a machine that can be in one of two states; good or bad. Suppose that the machine produces an item at the beginning of each day. The item produced is either good (if the machine is good) or bad (if the machine is bad). Suppose that once the machine is in the bad state, it remains in that state until it is replaced. However, if it is in the good state at the beginning of a day, then with probability γ it will be in the bad state at the beginning of the next day.

   We further suppose that after the item is produced we have the option of inspecting the item or not. If the item is inspected and found to be in the bad state, then the machine is instantaneously replaced with a good machine at an additional cost R. Also, the cost of inspecting an item will be denoted by I, and the cost of producing a bad item by C.

   Suppose that the process is in state P at time t if P is the posterior probability at t that the machine in use is in the bad state. If the objective is to minimize the total expected α-discounted cost, set this up as a Markov decision problem and write the optimality equation.

5. Show that $V(P)$, the optimal value function for Problem 4, is an *increasing, concave function of* P.
6. Consider a machine that can be in either of two states, good or bad. At the beginning of each day, the machine produces items that are either defective or nondefective. The probability of a defective item is P_1 when in the good state and P_2 when in the bad state. Once in the bad state, the machine remains in this state until it is replaced. However, if the machine is in the good state at the beginning of one day, then with probability γ it will be in the bad state at the beginning of the next day. A decision as to whether or not to replace the machine must be made each day after observing the item produced. Let R be the cost of replacing the machine and let C be the cost incurred whenever a defective item is produced. Set this problem up as a Markov decision model and determine the functional equation satisfied by V. Assume that at time zero there is a known probability that the machine is in the bad state.
7. Prove that for Problem 6 there is a P^* such that the α-optimal policy replaces whenever the present probability that the process is in the bad state is greater than or equal to P^*.
8. We have two coins, a red one and a green one. When flipped, one lands heads with probability P_1 and the other with probability P_2. However, we do not know which of the coins is the P_1 coin. Suppose that initially we believe that the red coin is the P_1 coin with probability p_0. Suppose we receive one unit for each head that appears, and our objective to to maximize our total expected discounted return.
   (a) Determine the optimality equation.
   (b) If $P_1 > P_2$, guess at the optimal policy.
9. Consider the following inventory problem. At the beginning of each day the amount of goods on hand is noted and a decision is

made as to how much to order. The cost for ordering j additional units is $C(j)$, where

$$C(j) = \begin{cases} K + cj & \text{if } j > 0, \\ 0 & \text{if } j = 0. \end{cases}$$

The order is assumed to be immediately filled. After the order has been filled, the daily demand for the product occurs. The demand will be j with probability P_j, $j \geq 0$. If the demand exceeds the present supply, then a penalty cost of A per unit of unmet demand is incurred. It is also assumed that, if the demand exceeds the supply, then the additional demand is backlogged and is filled when additional inventory becomes available (this can be represented as negative inventory). In addition there is an inventory holding cost of h for each item of remaining inventory at the end of a period.

The objective is to minimize the total expected discounted cost over an infinite time horizon when α is the discount factor.

(a) Set this up as a Markov decision process and write the optimality equation.

Consider now a single-period version of the preceding problem. Let

$$L(j) = A \sum_{k=j}^{\infty} (k - j)P_k + h \sum_{k=0}^{j} (j - k)P_k$$

denote the expected penalty and holding costs if we order to bring inventory up to j.

(b) Show that $L(j)$ is convex. That is, $L(j + 1) - L(j)$ is nondecreasing in j.

(c) Show that the optimal policy when the initial inventory is i is to order

$$\begin{cases} S - i & \text{if } i < s, \\ 0 & \text{if } i \geq s, \end{cases}$$

where S is the value that minimizes $cj + L(j)$ and s is such that $cs + L(s) = K + cS + L(S)$.

10. Assume that an individual has an initial capital of S_0 units. At the beginning of time period n, $n \geq 0$, the individual has S_n units that must be allocated: consuming C_n units, investing I_n units at a sure rate that will return rI_n by the beginning of the next period, and investing J_n units in a risky venture that will return Z_nJ_n units by the next period, where Z_n is a random variable having distribution F. Of course, $C_n + I_n + J_n = S_n$. The utility is in consumption, and consuming c leads to a utility $u(c)$. The objective is to maximize the expectation of $\sum_{n=0}^{\infty} \alpha^n u(C_n)$.

(a) Set this up as a dynamic programming problem. Give the optimality equation.

(b) If $u(c) = c^{\beta}$, $0 < \beta < 1$, show that the optimal policy allocates a fixed proportion of one's current fortune to the three alternatives. Show also, in this case, that $V(s) = Ku(s)$ for some constant K.

(c) Suppose that u is a concave function. If $\mathrm{E}[Z] < r$, do you think that no money would be allocated to the risky venture? If so, prove it.

11. Consider a problem with states 0, 1, and 2 possible actions having rewards

$$R(0, 1) = 1, \qquad R(1, 1) = 0,$$
$$R(0, 2) = 2, \qquad R(1, 2) = 0,$$

and transition probabilities

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

Let $\alpha = \frac{1}{2}$. Starting with $V_0 \equiv 0$, use successive approximations to approximate V by V_3. Then show that the policy obtained by maximizing $R(i, a) + \alpha \sum_j P_{ij}(a)V_3(j)$ is the optimal policy.

Now let f be the policy that chooses action 1 in both states 0 and 1. Show that this improvement of policy f is the optimal policy.

12. Let V_f and V_g denote the return functions for stationary policies f and g, respectively. Let h be the policy that chooses actions to maximize $R(i, a) + \alpha \sum_j P_{ij}(a) \max[V_f(j), V_g(j)]$. Show that $V_h(i) \geq \max[V_f(i), V_g(i)]$ for all i.

13. Let f and g be stationary policies with return functions V_f and V_g. Define the policy h by

$$h(i) = \begin{cases} f(i) & \text{if } V_f(i) \geq V_g(i), \\ g(i) & \text{if } V_f(i) < V_g(i). \end{cases}$$

Show that $V_h(i) \geq \max[V_f(i), V_g(i)]$ for all i.

References

1. Bertsekas, D., *Dynamic Programming and Stochastic Control.* Academic Press, New York, 1976.
2. Blackwell, D., "Discounted dynamic programming," *Ann. Math. Statist.* **36,** 226–235, 1965.
3. Derman, C., "On optimal replacement rules when changes of state are Markovian." In *Mathematical Optimization Techniques,* R. Bellman (ed.). University of California Press, Berkeley, California, 1963.
4. Derman, C., *Finite State Markovian Decision Processes.* Academic Press, New York, 1970.
5. Lippman, S., "On dynamic programming with unbounded rewards," *Management Sci.* **21,** 1225–1233, 1975.

III

Minimizing Costs— Negative Dynamic Programming

1. Introduction and Some Theoretical Results

In this chapter we again assume a countable state space (which, unless otherwise mentioned, will be taken to be the set of nonnegative integers) and a finite action space. However, we now suppose that if action a is taken when in state i, then an expected nonnegative cost $C(i, a)$, $C(i, a) \geq 0$, is incurred. The objective is to minimize the total expected cost incurred. Because this is equivalent to the total expected return for a problem having a reward function $R(i, a)$ $[= -C(i, a)]$ that is nonpositive, we say that we are in the negative case.

For any policy π, let

$$V_\pi(i) = E_\pi\left[\sum_{n=0}^{\infty} C(X_n, a_n) \middle| X_0 = i\right].$$

Because $C(i, a) \geq 0$, $V_\pi(i)$ is well defined, though possibly infinite. Thus we no longer assume a discount factor, and we no longer require the one-stage costs to be bounded. Also, let

$$V(i) = \inf_\pi V_\pi(i),$$

and call the policy π^* optimal if

$$V_{\pi^*}(i) = V(i), \qquad i \geq 0.$$