ON FINDING OPTIMAL POLICIES IN DISCRETE DYNAMIC PROGRAMMING WITH NO DISCOUNTING[1]

By Arthur F. Veinott, Jr.

Stanford University

1. Introduction. In an elegant paper [1] Blackwell has studied the infinite horizon discrete time parameter Markovian sequential decision problem with finitely many states and actions. He focuses initially on the case where there is a discount factor $\beta, 0 \leqq \beta < 1$. The problem is to choose a policy, termed *β-optimal*, that maximizes the total expected discounted return over an unbounded time horizon. He shows that there is a β-optimal policy which is stationary. He also gives a neat proof that Howard's [2], p. 84, policy improvement method yields a β-optimal stationary policy in finitely many steps.

For the case $\beta = 1$, a policy is called *1-optimal* if the difference between the total expected discounted return with that policy and the β-optimal policy for $0 \leqq \beta < 1$ tends to 0 as $\beta \nearrow 1$.[2] Blackwell established the existence of a 1-optimal policy that is stationary. He also shows that Howard's [2], p. 64, policy improvement method yields an element of the set of stationary policies that maximize the long run average return per unit time. Blackwell shows that this set contains the set of stationary 1-optimal policies. Thus if there is only one stationary policy with maximal average return, then that policy is 1-optimal and will be found by the policy improvement method. If there are two or more stationary policies having maximal average return, the method still yields a 1-optimal policy in certain special cases—e.g., where the chains associated with the stationary policies have a common absorbing state and transient states elsewhere. However, there are situations, e.g., example 2 in [1], p. 726, in which the policy improvement method fails to produce a 1-optimal policy.

Blackwell does not give an algorithm that will always find a 1-optimal policy. The principal purpose of this paper is to fill this gap by generalizing the policy improvement method to solve this problem (Theorem 14 below).

2. Review and extension of Blackwell's results. Following Blackwell [1] consider a system which is observed at each of a sequence of points in time labeled $1, 2, \cdots$. At those points the system is observed to be in one of S states labeled $1, 2, \cdots, S$. Each time the system is observed in state s, an action a is chosen from

Received 8 February 1966.

[1] This work was supported by the National Science Foundation under Grant GP-3739.

[2] Actually Blackwell uses the term "nearly optimal" for what we call 1-optimal policies for the case $\beta = 1$. He reserves the term "optimal" for $\beta = 1$ for a policy that is α-optimal (in our sense) for all $\alpha, 0 \leqq \alpha \leqq \beta$, sufficiently near β $(= 1)$. He does not consider this latter concept for $0 \leqq \beta < 1$ even though it is also meaningful (see Theorem 2 below) in that case. We have changed his terminology to establish what seems to us to be a more natural relationship between the definitions for the case $0 \leqq \beta < 1$ and $\beta = 1$.

a finite set A_s of possible actions and an income $i(s, a)$ is received. The conditional probability that the system is observed in state s' at time $n + 1$ given that it is found in state s at time n, that action a is taken at that time, and given the observed states and actions taken at times $1, 2, \cdots, n - 1$, is assumed to be a function $q(s' \mid s, a)$ depending only on s', s, and a.

Let $F = \mathbf{X}_{s=1}^{S} A_s$. A *policy* π is a sequence $(f_1, f_2, \cdots)$ of elements f_n of F. Using the policy π means that if the system is in state s at time n, the action chosen at that time is $f_n(s)$, the sth coordinate of f_n. Let $f^\infty = (f, f, \cdots)$ and $(g, f^\infty) = (g, f, f, \cdots)$ for f and g in F. f^∞ is called a *stationary policy*. For any $f \varepsilon F$, let $r(f)$ be the $S \times 1$ column vector whose sth element is $i(s, f(s))$, and let $Q(f)$ be the $S \times S$ Markov matrix whose (s, s') element is $q(s' \mid s, f(s))$. If $\pi = (f_1, f_2, \cdots)$, let $Q_n(\pi) = Q(f_1) \cdots Q(f_n)$. Thus the vector of total expected discounted returns starting from each state and using the policy π is

$$V_\beta(\pi) = \sum_{n=0}^{\infty} \beta^n Q_n(\pi) r(f_{n+1})$$

where $0 \leqq \beta < 1$ is a discount factor and $Q_0(\pi) = I$ (the $S \times S$ identity matrix).

For any two S-vectors $u = (u_i)$ and $v = (v_i)$, write $u \geqq v$ if $u_i \geqq v_i$ for all i and write $u > v$ if $u \geqq v$ and $u \neq v$. A policy π^* is called β-optimal if $V_\beta(\pi^*) \geqq V_\beta(\pi)$ for all π where $0 \leqq \beta < 1$.

THEOREM 1 (Blackwell). *Exactly one of the following must occur for each $f \varepsilon F$ and $0 \leqq \beta < 1$:*

(a) $V_\beta(f^\infty) \geqq V_\beta(g, f^\infty)$ *for all $g \varepsilon F$ and f^∞ is β-optimal.*

(b) $V_\beta(f^\infty) < V_\beta(g, f^\infty)$ *for some $g \varepsilon F$ and $V_\beta(f^\infty) < V_\beta(g^\infty)$.*

This theorem describes a finite algorithm (the policy improvement method) for finding a β-optimal policy that is stationary, $0 \leqq \beta < 1$. Let $\pi(\beta)$ be β-optimal and $U(\beta) = V_\beta(\pi(\beta))$, $0 \leqq \beta < 1$. A policy π^* is called *1-optimal* if $\lim_{\beta \nearrow 1} [V_\beta(\pi^*) - U(\beta)] = 0$. The next theorem is a slight generalization of Blackwell's Theorem 5 [1], p. 725. Only slight modifications of the original proof are required.

THEOREM 2. *For each $0 \leqq \beta \leqq 1$, there is a β-optimal policy that is stationary and that is α-optimal for all α, $0 \leqq \alpha \leqq \beta$ $(\beta \leqq \alpha \leqq 1)$, sufficiently near β.*

EXAMPLE 1. *An f which satisfies the hypotheses of part* (a) *of Theorem* 1 *and is not α-optimal for $\alpha \neq \beta$.* There are two states 1, 2. In state 1 there are three actions 1, 2, 3 with action i yielding an income of $\frac{1}{4}(5 - i)$ and the system remaining in state 1 with probability $\frac{1}{2}(i - 1)$. In state 2 there is only one action which yields an income of 0 and the system remains in state 2. Now $F = \{f_1, f_2, f_3\}$ where $f_i(1) = i, f_i(2) = 1$, and

$$V_\beta(f_i^\infty) = \begin{bmatrix} \frac{1}{4}(5 - i) \\ \hline 1 - \frac{1}{2}(i - 1)\beta \\ 0 \end{bmatrix}, \qquad i = 1, 2, 3.$$

Thus f_1^∞ is α-optimal for $0 \leqq \alpha \leqq \frac{1}{2}$, f_2^∞ is $\frac{1}{2}$-optimal only, and f_3^∞ is α-optimal for $\frac{1}{2} \leqq \alpha \leqq 1$. For $\beta = \frac{1}{2}$, f_2^∞ satisfies the hypotheses of part (a) of Theorem 1.

But only f_1^∞ and f_3^∞ are α-optimal for α in a neighborhood of $\frac{1}{2}$ relative to $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$ respectively as described in Theorem 2.

If we replace the income $\frac{1}{4}(5 - i)$ by $3 - i$, the above illustration reduces to Blackwell's example 1 [1], p. 726, in which f_1^∞ is α-optimal for $0 \leqq \alpha \leqq 1$ and f_2^∞ is 1-optimal only. Moreover f_2^∞ satisfies the hypothesis of part (b) of Theorem 14 below. Thus the algorithms given in Theorems 1 and 14 for finding β-optimal policies, $0 \leqq \beta \leqq 1$, do not always yield a policy satisfying the conditions of Theorem 2, i.e., a policy that is α-optimal for an interval to the left or right of β.

THEOREM 3 (Blackwell). *For each $f \in F$, let $Q^*(f) = \lim_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} Q(f)^i$. Then*

$$V_\beta(f^\infty) = x(f)/(1 - \beta) + y(f) + \epsilon(\beta, f), \qquad 0 \leqq \beta < 1, \tag{1}$$

where $x(f)$ is the unique solution of

$$[I - Q(f)]x = 0, \qquad Q^*(f)x = Q^*(f)r(f), \tag{2}$$

$y(f)$ is the unique solution of

$$[I - Q(f)]y = r(f) - x(f), \qquad Q^*(f)y = 0, \tag{3}$$

and $\lim_{\beta\to 1-} \epsilon(\beta, f) = 0$.

By Lemma 1 in [1] we have for $f \in F$ that

$$Q^*(f)Q(f) = Q(f)Q^*(f) = Q^*(f)Q^*(f) = Q^*(f). \tag{4}$$

Thus

$$x(f) = Q^*(f)r(f) \tag{5}$$

solves (2). Also upon premultiplying (1) by $Q^*(f)$ and using (3) and (5) we get

$$x(f)/(1 - \beta) = Q^*(f)V_\beta(f^\infty) + \delta(\beta, f), \qquad 0 \leqq \beta < 1, \tag{6}$$

where $\lim_{\beta\to 1-} \delta(\beta, f) = 0$ so by (1),

$$y(f) = V_\beta(f^\infty) - Q^*(f)V_\beta(f^\infty) + \eta(\beta, f), \qquad 0 \leqq \beta < 1, \tag{7}$$

where $\lim_{\beta\to 1-} \eta(\beta, f) = 0$. The formulas (5)–(7) give interpretations to $x(f)$ and $y(f)$. One may compute $x(f)$ and $y(f)$ by determining $Q^*(f)$, then computing $x(f)$ from (5), and finally solving (3) for $y(f)$.

Let $F' = \{f \mid f \in F, x(f) \geqq x(g) \text{ all } g \in F\}$ and $F'' = \{f \mid f \in F', y(f) \geqq y(g) \text{ all } g \in F'\}$. In view of (5), F' is evidently the set of all $f \in F$ having maximal average return per unit time.

THEOREM 4 (Blackwell). *F'' is the (nonempty) set of all $f \in F$ for which f^∞ is 1-optimal.*

PROOF. Use Theorem 2 and the representation (1).

In the sequel we shall sometimes denote by $[u]_s$ the sth component of a vector u. Consider the following inequalities associated with any $f, g \in F$:

(i) $[Q(g)x(f)]_s \geqq [x(f)]_s$;

(ii) $[r(g) + Q(g)y(f)]_s \geqq [x(f) + y(f)]_s$.

Let $G(f)$ be the set of $g \in F$ such that (i) holds for all s; (ii) holds for each s for which (i) holds with equality; at least one of the inequalities in (i), (ii) is strict; and for each s for which (i) and (ii) hold with equality, $g(s) = f(s)$.

THEOREM 5 (Blackwell). *Suppose $f \in F$.*

(a) *If $G(f)$ is empty, then $f \in F'$.*

(b) *If $g \in G(f)$, then $V_\beta(g^\infty) > V_\beta(f^\infty)$ for all $\beta(<1)$ sufficiently near 1.*

This theorem describes a finite algorithm (Howard's policy improvement method) for finding an $f \in F'$, i.e., an f with maximal average return per unit time.

COROLLARY 1. *Suppose $f \in F$ and $g \in G(f)$. Then either $x(g) > x(f)$, or $x(g) = x(f)$ and $y(g) > y(f)$.*

PROOF. From part (b) of Theorem 5 and the representation (1) we have $x(g) \geqq x(f)$ and $x(g) = x(f)$ implies $y(g) \geqq y(f)$. It remains only to show $(x(g), y(g)) \neq (x(f), y(f))$. If instead $(x(g), y(g)) = (x(f), y(f))$, then $g \notin G(f)$ which is a contradiction and completes the proof.

Let $E(f)$ be the set of $g \in F$ such that (i) and (ii) hold with equality for all s. Since the sth components of the bracketed vectors in (i), (ii) depend on the sth component of g and no others, it follows that $E(f)$ can be expressed in the form

$$E(f) = \mathbf{X}_{s=1}^{S} E(s, f) \tag{8}$$

where $E(s, f) = \{g(s) \mid g \in E(f)\}$.

LEMMA 1 (Blackwell). *If $f \in F$ and $g \in E(f)$, then $x(g) = x(f)$. If in addition $Q^*(g)Q^*(f) = Q^*(g)$, then $y(g) = y(f)$.*

LEMMA 2. *Suppose $f, g \in F$ and the Markov chain defined by $Q(g)$ has no transient states.*

(a) *If $g \in G(f)$, then $x(g) > x(f)$.*

(b) *If $G(f)$ is empty and $g \notin E(f)$, then $x(f) > x(g)$.*

PROOF. From Corollary 1, $g \in G(f)$ implies $x(f) \leqq x(g)$. From part (a) of Theorem 5, $G(f)$ empty implies $x(f) \geqq x(g)$. Thus it suffices to show $x(f) \neq x(g)$ in each case.

Suppose to the contrary that $x(f) = x(g)$. Then by (4),

$$Q^*(g)[Q(g)x(f) - x(f)] = Q^*(g)x(f) - Q^*(g)x(f) = 0.$$

But $Q^*(g) \geqq 0$ has a positive element in every column since there are no transient states; and $Q(g)x(f) - x(f)$ is nonnegative if $g \in G(f)$ and nonpositive if $G(f)$ is empty. Thus

$$Q(g)x(f) = x(f). \tag{9}$$

Also by (2), (4), and (5),

$$Q^*(g)[r(g) + Q(g)y(f) - x(f) - y(f)]$$
$$= x(g) + Q^*(g)y(f) - x(g) - Q^*(g)y(f) = 0.$$

Thus since $Q^*(g)$ has a positive element in each column and by (9) the bracketed term above is nonnegative if $g \in G(f)$ and nonpositive if $G(f)$ is empty, we have

$$r(g) + Q(g)y(f) = x(f) + y(f).$$

It follows from this equation and (9) that $g \in E(f)$. This contradicts the hypothesis $g \in G(f)$ in part (a) and $g \notin E(f)$ in part (b), which completes the proof.

COROLLARY 2. *Suppose* $f \in F'$. *Then*

(a) $E(f) \subset F'$.

(b) *If the Markov chain defined by* $Q(g)$ *has no transient states for each* $g \in F$, *then* $G(f)$ *is empty and* $E(f) = F'$.

PROOF. Part (a) follows from Lemma 1.

To prove part (b), suppose first $g \in G(f)$. Then $x(f) < x(g)$ by part (a) of Lemma 2 which contradicts the hypothesis $f \in F'$. Thus $G(f)$ is empty. Now suppose $g \in F' - E(f)$. Then by part (b) of Lemma 2 $x(g) < x(f)$ which contradicts the hypothesis $g \in F'$. Thus $F' \subset E(f)$. Combining this fact with part (a) completes the proof.

If the Markov chain defined by $Q(g)$ has no transient states for each $g \in F$, then by part (a) of Theorem 5 and part (b) of Corollary 2, the policy improvement method terminates immediately upon finding an element f of F'. Moreover, $E(f)$ is precisely the set of all g with maximal average return per unit time. Thus if the algorithm is initiated with $f \in F' - F''$, the method does not yield a 1-optimal policy. For a numerical illustration, see Example 2 in [1], p. 726.

LEMMA 3. *If* $f, g \in F''$, *then* $E(f) = E(g)$. *Moreover,* $F'' \subset E(f) \subset F'$.

PROOF. The set $E(f)$ depends only on $(x(f), y(f))$, and similarly for g. But $f, g \in F''$ imply $(x(f), y(f)) = (x(g), y(g))$ so $E(f) = E(g)$. Also since $g \in E(g) = E(f)$, $F'' \subset E(f)$. Finally, $E(f) \subset F'$ follows from Corollary 2.

EXAMPLE 2. *An* F'' *that is a proper subset of* $E(f)$, $f \in F''$. Blackwell's Example 2, [1], p. 726.

EXAMPLE 3. *An* $E(f)$, $f \in F''$, *that is a proper subset of* F'. There are two states 1, 2. In state 1 there are two actions 1, 2. Action 1 yields an income of 0 and the system remains in state 1. Action 2 yields an income of -2 and the system moves to state 2. In state 2 there is one action yielding an income of 0 and the system remains in state 2. Now $F = \{f_1, f_2\}$ where $f_i(1) = i$. Also

$$F'' = E(f_1) = \{f_1\} \neq F' = F.$$

3. Computing 1-optimal policies. In this section we develop an algorithm for finding a 1-optimal policy. To this end suppose we have found an $f' \in F'$ with $G(f')$ empty by employing the policy improvement method given in Theorem 5. It would then suffice to have a method for finding an $f'' \in F'$ that maximizes $y(g)$ over $g \in F'$ since by Theorem 4 it would follow that f'' is 1-optimal. Unfortunately it seems to be difficult to characterize F' in general. However, $E(f')$ is immediately available and from Corollary 2 we have $E(f') \subset F'$. This suggests that we consider instead maximizing $y(g)$ over $g \in E(f')$. It turns out that this problem can be solved with the techniques already developed as the next lemma shows.

LEMMA 4. *Suppose $f \varepsilon F$ and $g \varepsilon E(f)$. Let $w(g)$ be the unique solution to*

$$[I - Q(g)]w = 0, \qquad Q^*(g)w = Q^*(g)(-y(f)). \tag{10}$$

Then

$$y(g) = y(f) + w(g). \tag{11}$$

PROOF. The uniqueness of $w(g)$ follows from Theorem 3. Since $g \varepsilon E(f)$ we have from (ii) that

$$[I - Q(g)]y(f) = r(g) - x(f).$$

Adding this equation to the first equation in (10) and then rewriting the second equation in (10) we get

$$[I - Q(g)][y(f) + w(g)] = r(g) - x(f), \qquad Q^*(g)[y(f) + w(g)] = 0.$$

But from Theorem 3 the unique solution to this system is $y(g)$. Thus (11) holds, which completes the proof.

Notice from (2), (8), and Lemma 4 that the problem of maximizing $w(g)$ over $g \varepsilon E(f)$ has the same form as that of maximizing $x(g)$ over $g \varepsilon F$ where we replace A_s by $E(s, f)$, F by $E(f)$, and $r(g)$ by $-y(f)$ for $s = 1, \cdots, S$ and all $g \varepsilon E(f)$. Thus the policy improvement method of Theorem 5 can be used to find an h that maximizes $w(g)$—and hence $y(g)$ in view of (11)—over $g \varepsilon E(f)$.

Returning now to the discussion preceding Lemma 4, suppose we find $f' \varepsilon F'$ with $G(f')$ empty and then $f'' \varepsilon E(f')$ that maximizes $y(g)$ over $g \varepsilon E(f')$. If, as in part (b) of Corollary 2, we have $E(f') = F'$, then $f'' \varepsilon F''$ and by Theorem 4 $(f'')^\infty$ is the desired 1-optimal policy. Unfortunately, as Example 3 shows, $E(f')$ may be a proper subset of F' in which case we need a method for determining if $f'' \varepsilon F''$. Such a method is given in the next lemma.

LEMMA 5. *If $f \varepsilon F$, if $G(f)$ is empty, and if $y(f) \geqq y(g)$ for all $g \varepsilon E(f)$, then $f \varepsilon F''$.*

PROOF. (The proof is essentially a slight modification of Blackwell's proof of part (d) of Theorem 4 in [1].)

Choose $\beta < 1$ so near 1 that for any pair f_0, f_1 with $G(f_0)$ empty we have $V_\beta(f_1, f_0^\infty) \geqq V_\beta(f_0^\infty)$ implies $f_1 \varepsilon E(f_0)$, and $V_\beta(f_1) \geqq V_\beta(f_0)$ and $x(f_1) = x(f_0)$ implies $y(f_1) \geqq y(f_0)$. If f^∞ is not β-optimal, let $f_0 = f$ and let $f_1, f_2, \cdots, f_k$ be a sequence of β-improvements, obtained as in Theorem 1, terminating in a β-optimal f_k^∞. We show by induction on i that $(x(f_i), y(f_i)) = (x(f_0), y(f_0))$ and $G(f_i)$ is empty for $i = 0, 1, \cdots, k$. This is true for $i = 0$. If true for a given i, then $E(f_i) = E(f)$ because $E(f)$ depends only on $(x(f), y(f))$. Thus $f_{i+1} \varepsilon E(f)$ so by Lemma 1, $x(f_{i+1}) = x(f)$. Moreover, $V_\beta(f_{i+1}^\infty) > V_\beta(f^\infty)$ so by the definition of β, $y(f_{i+1}) \geqq y(f)$. But by hypothesis, $y(f_{i+1}) \leqq y(f)$ so necessarily $y(f_{i+1}) = y(f)$. Since $(x(f_{i+1}), y(f_{i+1})) = (x(f_0), y(f_0))$ and $G(f_0)$ is empty, $G(f_{i+1})$ is empty also.

(I am indebted to Bruce Miller for correcting an error in the original proof.)

Consequently, writing $f(\beta)^\infty$ for the β-optimal f_k^∞, we have from (1) that

$$U(\beta) = x(f)/(1 - \beta) + y(f) + \epsilon(\beta, f(\beta))$$

and

$$V_\beta(f^\infty) = x(f)/(1 - \beta) + y(f) + \epsilon(\beta, f),$$

so $U(\beta) - V_\beta(f^\infty) \to 0$ as $\beta \nearrow 1$. Thus f^∞ is 1-optimal and so by Theorem 4, $f \in F''$.

The next result generalizes part (d) of Theorem 4 in [1] slightly.

COROLLARY 3. *If $f \in F$, if $G(f)$ is empty, and if $Q^*(g)Q^*(f) = Q^*(g)$ for all $g \in E(f)$, then $E(f) = F''$.*

PROOF. By part (a) of Theorem 5 and Lemma 1, $(x(f), y(f)) = (x(g), y(g))$ for all $g \in E(f)$. Thus $E(g) = E(f)$ and $G(g) = G(f)$ for $g \in E(f)$ so by Lemma 5, $E(f) \subset F''$. On the other hand, by Lemma 3, $F'' \subset E(f)$ which completes the proof.

For each $f \in F$ let $z(f)$ be the unique solution (by Theorem 3) to

$$[I - Q(f)]z(f) = -y(f), \qquad Q^*(f)z(f) = 0. \tag{12}$$

Consider the following inequalities associated with any $f \in F$ and $g \in E(f)$:

$$[-y(f) + Q(g)z(f)]_s \geqq [z(f)]_s . \tag{iii}$$

Let $H(f)$ be the set of $g \in E(f)$ such that (iii) holds for all s and strictly for some s; and for each s for which (iii) holds with equality, $g(s) = f(s)$. We now come to our main result.

THEOREM 6. *Suppose $f \in F$.*

(a) *If $G(f)$ is empty, then $f \in F'$.*

(b) *If $G(f) \cup H(f)$ is empty, then $f \in F''$.*

(c) *If $g \in G(f)$, then either $x(g) > x(f)$, or $x(g) = x(f)$ and $y(g) > y(f)$.*

(d) *If $g \in H(f)$, then $x(g) = x(f)$; and either $y(g) > y(f)$, or $y(g) = y(f)$ and $z(g) > z(f)$.*

PROOF. Parts (a) and (c) follow from Theorem 5 and Corollary 1.

We now prove part (b). First note from (11) that $w(f) = 0$. Thus since $H(f)$ is empty, it follows from Theorem 5 and Lemma 4 that $y(g) \leqq y(f)$ for all $g \in E(f)$. Part (b) then follows from Lemma 5.

It remains to establish part (d). Now $g \in H(f) \subset E(f)$ so we have from Lemma 1 that $x(g) = x(f)$. It follows from Theorem 5, Lemma 4, and part (c) above that either $w(g) > w(f) = 0$ or $w(g) = w(f) = 0$ and $z(g) > z(f)$. Upon taking account of (11), the proof is complete.

Observe that Theorem 6 provides the following algorithm for finding an $f \in F$ for which $G(f) \cup H(f)$ is empty (and hence f^∞ is 1-optimal). Let $f_1 \in F$ be given. Choose $f_2, f_3, \cdots$ inductively so as to satisfy $f_{i+1} \in G(f_i) \cup H(f_i)$. The sequence $\{f_i\}$ is finite because F is finite and the sequence of triples $\{(x(f_i), y(f_i), z(f_i))\}$ is lexicographically increasing[3] so no f_i can recur. Thus, $G(f_i) \cup H(f_i)$ must be empty for some i.

[3] We say $u \in R^n$ is lexicographically smaller than $v \in R^n$ if $u \neq v$ and the first nonzero component of $v - u$ is positive. The relation "lexicographically smaller than" is transitive and completely orders R^n.

In carrying out the computations one would probably choose $f_{i+1} \varepsilon G(f_i)$ if possible and $f_{i+1} \varepsilon H(f_i)$ only if $G(f_i)$ were empty. Under this rule the first phase of the algorithm, i.e., the calculations leading to an element of F', coincide with the method of Theorem 5. The second phase, i.e., the remaining calculations, leads to an element of F''. Notice that in the first phase one must compute $x(f_i)$ and $y(f_i)$ at the ith stage but not $z(f_i)$. In the second phase $x(f_i)$ remains fixed and it is necessary to compute only $y(f_i)$ and, when $G(f_i)$ is empty, also $z(f_i)$.

Let $E'(f)$ be the set of all $g \varepsilon E(f)$ for which (iii) holds with equality for all s.

COROLLARY 4. *Suppose* $f \varepsilon F''$. *Then*

(a) $E'(f) \subset F'' \subset E(f) \subset F'$.

(b) *If the Markov chain defined by* $Q(g)$ *has no transient states for each* $g \varepsilon F$, *then* $G(f) \cup H(f)$ *is empty and* $E'(f) = F''$.

PROOF. To prove part (a), note from Lemma 3 that we must show only that $E'(f) \subset F''$. This relation follows from part (a) of Corollary 2 and Lemma 4. Part (b) follows from part (b) of Corollary 2 and Lemma 4.

4. An alternative proof. Our proof of Theorem 6 in the preceding section depended heavily upon treating $\beta = 1$ as a limiting case of $\beta < 1$. In this section we give an alternative proof of the theorem that was suggested by the approach of Howard [2], pp. 69–73. Howard proved part (a) of Theorem 6 and that $g \varepsilon G(f)$ implies $x(g) \geqq x(f)$. (We shall repeat his proofs of these facts below.) He did not show, however, that his policy improvement method terminates in finitely many steps—i.e., that "cycling" does not occur. That cycling does not occur follows from part (c) of Theorem 6, however.

PROOF OF THEOREM 6. Suppose $f, g \varepsilon F$. From Markov chain theory by appropriately relabeling the states we may write

$$Q(g) = \left[\begin{array}{cccc|c} Q_{11} & & & 0 & \\ & Q_{22} & & & 0 \\ & & \ddots & & \\ 0 & & & Q_{N-1,N-1} & \\ \hline Q_{N1} & Q_{N2} & \cdots & Q_{N,N-1} & Q_{NN} \end{array}\right]$$

and

$$Q^*(g) = \left[\begin{array}{cccc|c} Q^*_{11} & & & 0 & \\ & Q^*_{22} & & & 0 \\ & & \ddots & & \\ 0 & & & Q^*_{N-1,N-1} & \\ \hline Q^*_{N1} & Q^*_{N2} & \cdots & Q^*_{N,N-1} & 0 \end{array}\right]$$

where Q_{ii}^*, $1 \leqq i < N$, has identical rows of positive numbers and $(I - Q_{NN})^{-1}$ exists and is nonnegative. Let

$$\psi = Q(g)x(f) - x(f);$$
$$\gamma = r(g) + Q(g)y(f) - x(f) - y(f);$$
$$\theta = -y(f) + Q(g)z(f) - z(f).$$

Let $\Delta x = x(g) - x(f)$. Define Δy, Δr, Δz similarly. From the definitions involved we have

$$\Delta x = \psi + Q(g)\Delta x; \tag{13}$$

$$\Delta x + \Delta y = \gamma + Q(g)\Delta y; \tag{14}$$

$$\Delta y + \Delta z = \theta + Q(g)\Delta z. \tag{15}$$

Upon premultiplying (13)–(15) by $Q^*(g)$ and using (4) we get respectively

$$0 = Q^*(g)\psi; \tag{16}$$

$$Q^*(g)\Delta x = Q^*(g)\gamma; \tag{17}$$

$$Q^*(g)\Delta y = Q^*(g)\theta. \tag{18}$$

Now partition the transpose of $x(f)$ by $x(f)^T = (x^1(f)^T, \cdots, x^N(f)^T)$ where the column vector $x^i(f)$ has the same number of components as Q_{ii} has columns. Do the same for $x(g)$, $y(f)$, $y(g)$, $z(f)$, $z(g)$, ψ, γ, θ, Δx, Δy, Δz.

If $\psi^i = 0$ for fixed i, $1 \leqq i < N$, we have from (13) that $\Delta x^i = Q_{ii}\Delta x^i$, so from Markov chain theory Δx^i has identical components. Consequently, it follows from (17) that

$$\psi^i = 0 \quad \text{implies} \quad \Delta x^i = Q_{ii}^*\gamma^i, \qquad 1 \leqq i < N. \tag{19}$$

A similar argument using (14) and (18) gives

$$\Delta x^i = \gamma^i = 0 \quad \text{implies} \quad \Delta y^i = Q_{ii}^*\theta^i, \qquad 1 \leqq i < N. \tag{20}$$

We can solve (13)–(15) respectively for Δx^N, Δy^N, Δz^N as follows:

$$\Delta x^N = (I - Q_{NN})^{-1}(\psi^N + \textstyle\sum_{j=1}^{N-1} Q_{Nj}\Delta x^j); \tag{21}$$

$$\Delta y^N = (I - Q_{NN})^{-1}(\gamma^N - \Delta x^N + \textstyle\sum_{j=1}^{N-1} Q_{Nj}\Delta y^j); \tag{22}$$

$$\Delta z^N = (I - Q_{NN})^{-1}(\theta^N - \Delta y^N + \textstyle\sum_{j=1}^{N-1} Q_{Nj}\Delta z^j). \tag{23}$$

We now prove part (a). Suppose $G(f)$ is empty. Then $\psi \leqq 0$ and $\psi^i = 0$ implies $\gamma^i \leqq 0$, $1 \leqq i \leqq N$. Now by (16) and $\psi^i \leqq 0$, we have $\psi^i = 0$, $1 \leqq i < N$. Thus $\gamma^i \leqq 0$, $1 \leqq i < N$, so from (19) and (21), $\Delta x \leqq 0$ which establishes part (a).

Turning now to part (b), suppose $G(f) \cup H(f)$ is empty. Then $\psi \leqq 0$; $\psi^i = 0$ implies $\gamma^i \leqq 0$, $1 \leqq i \leqq N$; and $\psi^i = \gamma^i = 0$ implies $\theta^i \leqq 0$, $1 \leqq i \leqq N$. From part (a), $f \varepsilon F'$ so $f \varepsilon F''$ if $\Delta x = 0$ implies $\Delta y \leqq 0$. From the preceding paragraph,

(21), and $\Delta x^N = 0$, we have $\psi = 0$. Thus from (19), $\gamma^i = 0$, $1 \leqq i < N$; and $\gamma^N \leqq 0$. Hence $\theta^i \leqq 0$, $1 \leqq i < N$ so from (20) and (22) we have $\Delta y \leqq 0$, which proves part (b).

To prove part (c), suppose $g \varepsilon G(f)$. Then $\psi \geqq 0$ and $\psi^i = 0$ implies $\gamma^i \geqq 0$, $1 \leqq i \leqq N$. Now by (16) and $\psi^i \geqq 0$, we have $\psi^i = 0$, $1 \leqq i < N$. Thus $\gamma^i \geqq 0$, $1 \leqq i < N$, so from (19) and (21), $\Delta x \geqq 0$. To complete the proof it suffices to show $\Delta x = 0$ implies $\Delta y > 0$. It follows from $\Delta x = 0$, $g \varepsilon G(f)$, (19), and (21) that $\psi = 0$; $\gamma^i = 0$, $1 \leqq i < N$; $\gamma^N > 0$. Thus from (14)

$$\Delta y^i = Q_{ii}\Delta y^i, \qquad i = 1, \cdots, N-1. \tag{24}$$

Since $(\psi_s, \gamma_s) = 0$ and $g \varepsilon G(f)$ imply $f(s) = g(s)$, it follows that the first $N - 1$ rows of submatrices of $Q(g)(Q^*(g))$ equal the corresponding rows of submatrices of $Q(f)(Q^*(f))$. Consequently, for $1 \leqq i < N$, we have

$$Q^*_{ii}\Delta y^i = Q^*_{ii}y^i(g) - Q^*_{ii}y^i(f) = 0. \tag{25}$$

Since $\Delta y^i = 0$ satisfies (24), (25), and since (24), (25) have a unique solution, we have $\Delta y^i = 0$, $1 \leqq i < N$. Thus since $\gamma^N > 0$ and $\Delta x^N = 0$, we have from (22) that $\Delta y^N > 0$ so $\Delta y > 0$, which establishes part (c).

We now prove part (d). Suppose $g \varepsilon H(f)$. Then $\psi = \gamma = 0$ so $\theta \geqq 0$. Thus from (19) and (21), $\Delta x = 0$. Hence from (20) and (22), $\Delta y \geqq 0$. It remains to show only that $\Delta y = 0$ implies $\Delta z > 0$. We have from $\Delta y = 0$, (20), and $g \varepsilon H(f)$ that $\theta^i = 0$, $1 \leqq i < N$; also $\theta^N > 0$. Thus from (15)

$$\Delta z^i = Q_{ii}\Delta z^i, \qquad i = 1, \cdots, N-1.$$

Since $(\psi_s, \gamma_s, \theta_s) = 0$ and $g \varepsilon H(f)$ imply $f(s) = g(s)$, it follows by an argument like that of the preceding paragraph that $\Delta z^i = 0$, $1 \leqq i < N$. Thus since $\theta^N > 0$ and $\Delta y^N = 0$, we have from (23) that $\Delta z^N > 0$ so $\Delta z > 0$. This completes the proof.

5. Another criterion for optimality. The vector of total expected returns in periods $1, 2, \cdots, n$ starting from each state and using the policy π is

$$V^n(\pi) = \textstyle\sum_{i=0}^{n-1} Q_i(\pi) r(f_{i+1}).$$

A policy π^* will be called *optimal* if

$$\liminf_{N\to\infty} N^{-1} \textstyle\sum_{n=1}^{N} [V^n(\pi^*) - V^n(\pi)] \geqq 0 \tag{26}$$

for all π.

Observe from the definitions of $x(f)$ and $y(f)$ that

$$\begin{aligned} y(f) &= r(f) - x(f) + Q(f)y(f) \\ &= \cdots = \textstyle\sum_{i=0}^{n-1} Q(f)^i[r(f) - x(f)] + Q(f)^n y(f) \\ &= V^n(f^\infty) - nx(f) + Q(f)^n y(f). \end{aligned}$$

Summing both sides fom 1 to N, dividing by N, letting $N \to \infty$, and using

$Q^*(f)y(f) = 0$, we get

$$y(f) = \lim_{N\to\infty} N^{-1} \sum_{n=1}^{N} [V^n(f^\infty) - nx(f)]. \tag{27}$$

Thus $[y(f)]_s$ may be interpreted as the average amount by which the total expected return for n periods starting from state s exceeds that starting with the stationary probability vector $[Q^*(f)]_s$, the sth row of $Q^*(f)$. We may rewrite (27) in the form

$$N^{-1} \sum_{n=1}^{N} V^n(f^\infty) = [(N+1)/2]x(f) + y(f) + \sigma(N, f) \tag{28}$$

where $\lim_{N\to\infty} \sigma(N, f) = 0$. The next theorem is an immediate consequence of the representation (28).

THEOREM 7. *$f \varepsilon F''$ if and only if*

$$\lim_{N\to\infty} N^{-1} \sum_{n=1}^{N} [V^n(f^\infty) - V^n(g^\infty)] \geqq 0$$

for all $g \varepsilon F$.

It follows from this theorem that if f^∞ is optimal, then f^∞ is 1-optimal. We conjecture that the converse is also true.

REFERENCES

[1] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.
[2] HOWARD, R. (1960). *Dynamic Programming and Markov Processes.* Wiley, New York.