The Annals of Mathematical Statistics
1969, Vol. 40, No. 2, 366–370

DISCRETE DYNAMIC PROGRAMMING WITH A SMALL INTEREST RATE[1]

BY BRUCE L. MILLER AND ARTHUR F. VEINOTT, JR.

The RAND Corporation and Stanford University

1. Introduction. In a fundamental paper on stationary finite state and action Markovian decision processes, Blackwell [1] defines an optimal policy to be one that maximizes the expected total discounted rewards for all sufficiently small interest rates $\rho > 0$. He also establishes the existence of a stationary optimal policy by a limit process that does not give a finite algorithm. The purpose of this paper is to prove this result constructively by devising a finite policy improvement method for finding stationary optimal policies. The algorithm is based on the representation of the vector of expected discounted returns under a stationary policy as a Laurent series in the interest rate for all small enough $\rho > 0$.

2. Preliminaries. Consider a system which is observed at each of a sequence of points in time labeled $1, 2, \cdots$. At each of these points the system is found to be in one of S states labeled $1, \cdots, S$. Each time the system is observed in state s, an action a is chosen from a finite set A_s of possible actions and a reward $r(s, a)$ is received. The conditional probability that the system is observed in state t 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 $p(t \mid s, a)$ depending only on t, s, and a.

Let $F = \times_{s=1}^{S} A_s$. A *policy* is a sequence $\pi = (f_1, f_2, \cdots)$ of elements f_N of F. Using the policy π means that if the system is observed in state s at time N, the action chosen at that time is $f_N(s)$, the sth component of f_N. We write f^∞ for the *stationary policy* $(f, f, \cdots)$ and (g, f^∞) for the policy $(g, f, f, \cdots)$.

For any $f \in F$, let $r(f)$ be the S component column vector whose sth component is $r(s, f(s))$, and let $P(f)$ be the $S \times S$ Markov matrix whose stth element is $p(t \mid s, f(s))$. If $\pi = (f_1, f_2, \cdots)$, let $P^N(\pi) = P(f_1) \cdots P(f_N)$ for $N > 0$ and $P^0(\pi) = I$.

Denote by $\rho > 0$ the rate of interest and let $\beta = (1 + \rho)^{-1}$ be the associated discount factor. If $\rho = \infty$, $\beta \equiv 0$. We suppress the dependence of β on ρ in the sequel for simplicity.

The vector of expected total discounted rewards starting from each state and using the policy π is

$$V_\rho(\pi) = \sum_{N=0}^{\infty} \beta^N P^N(\pi) r(f_{N+1}).$$

A policy π^* is called *ρ-optimal* if $V_\rho(\pi^*) \geqq V_\rho(\pi)$ for all π, and *optimal* if it is ρ-optimal for all sufficiently small $\rho > 0$.

Received 6 June 1968.

[1] This research was supported by the United States Air Force under project RAND Contract No. F44620-67-C-0045, the National Science Foundation under Grant GK-1420, and the Office of Naval Research under Contract Nonr 225(77) (NR-347-010).

We will need the following result from Kemeny and Snell [4] and Blackwell [1].

LEMMA 1. *Let P be an $S \times S$ Markov matrix.*

(a) *The sequence $(N + 1)^{-1} \sum_{i=0}^{N} P^i$ converges as $N \to \infty$ to a Markov matrix P^* satisfying $PP^* = P^*P = P^*P^* = P^*$.*

(b) *If $0 \leqq \rho \leqq \infty$, the matrix $[I - \beta(P - P^*)]$ is nonsingular and its inverse, denoted Z_ρ, uniquely satisfies*

$$\begin{bmatrix} I - \beta P \\ P^* \end{bmatrix} Z_\rho = Z_\rho \begin{bmatrix} I - \beta P \\ P^* \end{bmatrix} = \begin{bmatrix} I - \beta P^* \\ P^* \end{bmatrix}.$$

(c) *If $0 \leqq \rho \leqq \infty$, the matrix $H_\rho \equiv Z_\rho(I - P^*) = (I - P^*)Z_\rho = Z_\rho - P^*$ uniquely satisfies*

$$\begin{bmatrix} I - \beta P \\ P^* \end{bmatrix} H_\rho = H_\rho \begin{bmatrix} I - \beta P \\ P^* \end{bmatrix} = \begin{bmatrix} I - P^* \\ 0 \end{bmatrix}.$$

(d) *If $0 < \rho \leqq \infty$, the matrix $[I - \beta P]$ is nonsingular and its inverse, denoted M_ρ, satisfies*

$$M_\rho = \sum_{i=0}^{\infty} \beta^i P^i = P^* M_\rho + H_\rho \quad \textit{and} \quad P^* M_\rho = (1 + \rho)\rho^{-1} P^*.$$

The next result provides an expansion of H_ρ in terms of the powers of $-\rho H$ for small $|\rho|$ where $H \equiv H_0$. To describe this it is convenient to define the norm of a (finite) matrix $C = (c_{ij})$ by $\|C\| \equiv \max_i \sum_j |c_{ij}|$.

LEMMA 2. *If $0 \leqq \rho < \|H\|^{-1}$, then*

(a) *$(I + \rho H)$ is nonsingular and*

$$(I + \rho H)^{-1} = \sum_{n=0}^{\infty} \rho^n (-1)^n H^n;$$

(b) $H_\rho = (1 + \rho)H(I + \rho H)^{-1} = (1 + \rho)(I + \rho H)^{-1} H$.

PROOF. Part (a) follows from $\|\rho H\| < 1$ which justifies the Neumann series expansion therein.

For part (b), we have from (c) of Lemma 1 that

$$(1 + \rho)Z_\rho^{-1} H = (I - P)H + \rho H = (I - P^*)(I + \rho H).$$

Postmultiplying by $(I + \rho H)^{-1}$ and premultiplying by Z_ρ gives, using the definition of H_ρ,

$$(1 + \rho)H(I + \rho H)^{-1} = Z_\rho(I - P^*) = H_\rho,$$

establishing the first equality in (b). The second equality in (b) then follows from (a), which completes the proof.

For each $f \varepsilon F$, let $P^*(f)$, $H_\rho(f)$, and $M_\rho(f)$ denote the matrices in Lemma 1 associated with $P(f)$. Then since $V_\rho(f^\infty) = M_\rho(f)r(f)$, we may combine part (d) of Lemma 1 with Lemma 2 to give the Laurent series expansion of $V_\rho(f^\infty)$ for $\rho > 0$ near zero. The first two terms of this expansion were obtained in [1].

THEOREM 1. *If $f \varepsilon F$ and $0 < \rho < \|H(f)\|^{-1}$, then*

$$V_\rho(f^\infty) = (1 + \rho) \sum_{n=-1}^{\infty} \rho^n y_n(f) \tag{1}$$

where $y_{-1}(f) \equiv P^*(f)r(f)$ and $y_n(f) \equiv (-1)^n H(f)^{n+1} r(f)$, $n = 0, 1, \cdots$.

3. Finding optimal policies. Our policy improvement algorithm for finding optimal policies relies on Howard's [3] policy improvement method for finding ρ-optimal policies ($\rho > 0$) as refined and formulated by Blackwell [1] in the following result.

THEOREM 2. *If $f \varepsilon F$ and $0 < \rho \leqq \infty$, then either $V_\rho(g, f^\infty) > V_\rho(f^\infty)$ for some $g \varepsilon F$ or $V_\rho(g, f^\infty) \leqq V_\rho(f^\infty)$ for all $g \varepsilon F$. In the former case $V_\rho(g^\infty) > V_\rho(f^\infty)$, while in the latter event f^∞ is ρ-optimal.*

If C is a matrix, we say C is *lexicographically nonnegative*, written $C \succcurlyeq 0$, if the first nonvanishing element of each row of C is positive. Similarly, C is called *lexicographically positive*, written $C \succ 0$, if $C \succcurlyeq 0$ and $C \neq 0$.

Let $Y(f) = (y_{-1}(f), y_0(f), \cdots)$ and $Y_n(f) = (y_{-1}(f), y_0(f), \cdots, y_n(f))$ for $n \geqq -1$. It is clear from (1) that $V_\rho(f^\infty) - V_\rho(g^\infty) \geqq 0$ for all small enough $\rho > 0$ if and only if $Y(f) - Y(g) \succcurlyeq 0$.

For $f, g \varepsilon F$, let

$$\begin{aligned} \psi_n(g, f) &= \qquad\quad P(g)y_{-1}(f) \qquad\qquad - y_{-1}(f), &\quad n &= -1, \\ &= r(g) + P(g)y_0(f) - y_{-1}(f) \quad - y_0(f), & n &= 0, \\ &= \qquad\quad P(g)y_n(f) - y_{n-1}(f) - y_n(f), & n &= 1, 2, \cdots, \end{aligned}$$

$\Psi(g, f) = (\psi_{-1}(g, f), \psi_0(g, f), \cdots), \Psi_n(g, f) = (\psi_{-1}(g, f), \psi_0(g, f), \cdots, \psi_n(g, f))$ for $n \geqq -1$, and $\Psi_n(g, f) = 0$ for $n < -1$.

LEMMA 3. *If $f, g \varepsilon F$ and $0 < \rho < \|H(f)\|^{-1}$, then*

$$V_\rho(g, f^\infty) - V_\rho(f^\infty) = \textstyle\sum_{n=-1}^{\infty} \rho^n \psi_n(g, f).$$

PROOF. From Theorem 1,

$$\begin{aligned} &V_\rho(g, f^\infty) - V_\rho(f^\infty) \\ &\qquad = r(g) + [(1 + \rho)^{-1}P(g) - I]V_\rho(f^\infty) \\ &\qquad = r(g) + [P(g) - (1 + \rho)I] \textstyle\sum_{n=-1}^{\infty} \rho^n y_n(f) = \textstyle\sum_{n=-1}^{\infty} \rho^n \psi_n(g, f). \end{aligned}$$

REMARK. One consequence of this lemma is that $\Psi(f, f) = 0$ for $f \varepsilon F$.

THEOREM 3. *If $f \varepsilon F$, then either $\Psi(g, f) \succ 0$ for some $g \varepsilon F$ or $\Psi(g, f) \preccurlyeq 0$ for all $g \varepsilon F$. In the former event $V_\rho(g^\infty) - V_\rho(f^\infty) > 0$ for all small enough $\rho > 0$ and $Y(g) - Y(f) \succ 0$, while in the latter case f^∞ is optimal and $Y(f) - Y(g) \succcurlyeq 0$ for all $g \varepsilon F$.*

PROOF. If $\Psi(g, f) \succ 0$ for some $g \varepsilon F$, then from Lemma 3, $V_\rho(g, f^\infty) - V_\rho(f^\infty) > 0$ for all sufficiently small $\rho > 0$. Hence by Theorem 2, $V_\rho(g^\infty) - V_\rho(f^\infty) > 0$ for all small enough $\rho > 0$. Thus by Theorem 1, $Y(g) - Y(f) \succ 0$.

If $\Psi(g, f) \not\succ 0$ for every $g \varepsilon F$, then since $\Psi(f, f) = 0$ we have $\Psi(g, f) \preccurlyeq 0$ for all $g \varepsilon F$. Thus by Lemma 3, $V_\rho(g, f^\infty) - V_\rho(f^\infty) \leqq 0$ for all $g \varepsilon F$ and all small enough $\rho > 0$. Hence by Theorem 2, f^∞ is ρ-optimal for all small enough $\rho > 0$. Therefore f^∞ is optimal and, by Theorem 1, $Y(f) - Y(g) \succcurlyeq 0$ for all $g \varepsilon F$, which completes the proof.

COROLLARY 1. (Blackwell) *There is a stationary optimal policy.*

PROOF. Let $f_0 \varepsilon F$ be arbitrary. Choose $f_1, f_2, \cdots, f_N$ in F inductively so $\Psi(f_i, f_{i-1}) \succ 0$ for $i = 1, 2, \cdots, N$. Since by Theorem 3, $Y(f_i)$ increases lexicographically with i, no element of F can recur. Thus by Theorem 3 and the finiteness of F, there is an integer $N \geqq 0$ for which $\Psi(g, f_N) \preccurlyeq 0$ for all $g \varepsilon F$. Moreover, f_N^∞ is optimal, completing the proof.

The next theorem shows that we can replace $\Psi(g, f)$ by $\Psi_S(g, f)$ in Theorem 3, and so also in the policy improvement algorithm given in the proof of Corollary 1. That is, of course, an important computational simplification. The theorem also implies that $f^\infty(f \varepsilon F)$ is optimal if and only if $Y_S(f) \succcurlyeq Y_S(g)$ for all $g \varepsilon F$. To prove the theorem we will need a preliminary lemma which, as Joel Brenner has pointed out to one of us, is known ([2], p. 203). We repeat the proof for completeness.

LEMMA 4. *Let M be an $S \times S$ matrix and L a linear subspace of R^S. If $M^n x \varepsilon L$ for $n = 0, \cdots, S - 1$, then $M^n x \varepsilon L$ for $n = 0, 1, \cdots$.*

PROOF. The S component vectors $M^0 x, \cdots, M^S x$ are linearly dependent. Hence, there is a positive integer $T \leqq S$ such that $M^{T+1}x$ is a linear combination of $M^0 x, \cdots, M^T x$. We now show by induction on n that $M^n x$ is a linear combination of $M^0 x, \cdots, M^T x$ for all $n \geqq 0$, which will complete the proof. This is so for $0 \leqq n \leqq T + 1$ by construction. Suppose it is so for all positive integers less than $n(> T + 1)$. Thus

$$M^{n-1}x = \sum_{i=0}^{T} \lambda_i M^i x.$$

Premultiplying both sides of this equation by M gives

$$M^n x = \sum_{i=0}^{T} \lambda_i M^{i+1} x.$$

Since $M^{T+1}x$ is a linear combination of $M^0 x, \cdots, M^T x$, the proof is complete.

THEOREM 4. *Suppose $f, g \varepsilon F$. Then*

(a) $\Psi(g,f) \succ (\succcurlyeq)(=)(\preccurlyeq)(\prec)0$ *if and only if* $\Psi_S(g,f) \succ (\succcurlyeq)(=)(\preccurlyeq)(\prec)0$.

(b) $Y(f) = Y(g)$ *if and only if* $Y_S(f) = Y_S(g)$.

PROOF. For part (a) it suffices to show that $\Psi_S(g, f) = 0$ implies $\Psi(g, f) = 0$. To this end observe that since $\psi_n(f, f) = 0$,

$$\psi_n(g, f) = \psi_n(g, f) - \psi_n(f, f) = [P(g) - P(f)]y_n(f), \quad n = 1, 2, \cdots. \tag{2}$$

Because $\Psi_S(g, f) = 0$, it follows from (2) that

$$[P(g) - P(f)]y_n(f) = 0, \qquad n = 1, \cdots, S. \tag{3}$$

In view of (2), it suffices to show that (3) holds for $n = 1, 2, \cdots$. That this is so follows by an application of Lemma 4 with $M = -H(f)$, L the null space of $P(g) - P(f)$, and $x = y_1(f)$.

For part (b) it suffices to show that $Y_S(f) = Y_S(g)$ implies $Y(f) = Y(g)$. This will be so if we can show

$$[H(g) - H(f)]y_n(f) = 0, \qquad n = 0, 1, \cdots. \tag{4}$$

By hypothesis (4) holds for $n = 0, \cdots, S - 1$. That (4) holds for $n = 0, 1, \cdots$, then follows by applying Lemma 4 with $M = -H(f)$, L the null space of $H(g) - H(f)$, and $x = y_0(f)$, which completes the proof.

In a companion paper [8] one of us establishes and interprets several additional properties of the policy improvement algorithm given in the proof of Corollary 1. We mention a few of these results briefly here. For this purpose let $\Psi_{ns}(g, f)$ denote the sth row of $\Psi_n(g, f)$. For each $f \varepsilon F$ and $n \geqq -1$, let $G_n(f) = \{g: g \varepsilon F, \Psi_n(g, f) \succ 0$, and $g(s) = f(s)$ whenever $\Psi_{ns}(g, f) = 0\}$, $F_n = \{f: f \varepsilon F, Y_n(f) - Y_n(g) \succcurlyeq 0$ all $g \varepsilon F\}$, and $F_\infty = \{f: f \varepsilon F, Y(f) - Y(g) \succcurlyeq 0$ all $g \varepsilon F\}$. For $f, g \varepsilon F$ and $n < -1$, let $\Psi_n(g, f) = 0$, $G_n(f) = \phi$, $Y_n(f) = 0$, and $F_n = F$. It is immediate from (b) of Theorem 4 that $F_S = F_{S+1} = \cdots = F_\infty$.

The following results, among others, are established in [8]. If $f, g \varepsilon F$, $n \geqq -2$, and $\Psi_n(g, f) = 0$, then $Y_{n-1}(g) - Y_{n-1}(f) = 0$; if also $g \varepsilon G_{n+1}(f) - G_n(f)$, then $Y_{n+1}(g) - Y_{n+1}(f) \succ 0$. If $f \varepsilon F$, $n \geqq 0$, and $G_n(f)$ is empty, then $f \varepsilon F_{n-1}$. These results give a policy improvement algorithm for finding an element of F_n for $n \geqq -1$ that terminates more rapidly than the one in the proof of Corollary 1 for $n < S - 1$. For $n = -1$ and $n = 0$ the algorithms reduce respectively to those of Blackwell [1] and Veinott [7].

The results given in this paper extend without difficulty to the continuous time parameter case. A simple method of accomplishing this is given in [8] by exploiting results of Howard [3] and Miller [5], [6].

REFERENCES

[1] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

[2] GANTMACHER, F. R. (1959). *The Theory of Matrices* **1** (English Translation). Chelsea, New York.

[3] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes*. Wiley, New York.

[4] KEMENY, J. G. and SNELL, J. L. (1960). *Finite Markov Chains*. Van Nostrand, Princeton.

[5] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with a finite planning horizon. *SIAM J. Control* **6** 266–280.

[6] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with an infinite planning horizon. *J. Math. Anal. Appl.* **22** 552–569.

[7] VEINOTT, JR., A. F. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.* **37** 1284–1294.

[8] VEINOTT, JR., A. F. (1968). Discrete dynamic programming with sensitive discount optimality criteria. Technical Report No. 6, Department of Operations Research, Stanford Univ. California, 58 pp.