The Annals of Statistics
1974, Vol. 2, No. 1, 159-169

ON MAXIMAL REWARDS AND ε-OPTIMAL POLICIES IN CONTINUOUS TIME MARKOV DECISION CHAINS[1]

BY MARK R. LEMBERSKY

Oregon State University

For continuous time Markov decision chains of finite duration, we show that the vector of maximal total rewards, less a linear average-return term, converges as the duration $t \to \infty$. We then show that there are policies which are both simultaneously ε-optimal for all durations t and are stationary except possibly for a final, finite segment. Further, the length of this final segment depends on ε, but not on t for large enough t, while the initial stationary part of the policy is independent of both ε and t.

1. Introduction. In this paper we consider continuous time parameter stationary finite state and action Markov decision chains with finite durations. We first examine the behavior of the vector of maximal total expected rewards V^t as a function of the process duration t. We show that (Theorem 1) V^t, less t times the maximal long-run average return rate U, converges as $t \to \infty$. Thus, while it was well known that V^t does not generally approach a finite limit, we establish that V^t eventually grows essentially *linearly* in the duration t, at the rate U.

Using the convergence of $V^t - tU$, we then show (Theorem 2) there are policies which are both simultaneously ε-optimal for all process durations and are *stationary* except possibly for a *final*, *finite* segment. Further, the length of this final segment depends on ε but *not* on t for large enough t, while the initial stationary part of the policy is *independent of both* ε *and* t.

Results of this type for *discrete* time Markov decision chains have been presented by Brown [3] and Lanery [8]. Instead of convergence and the existence of initially stationary ε-optimal policies, their results deal with asymptotic periodicity (i.e., the existence of an integer period $p \geqq 1$ such that $V^{np+m} - (np + m)U$ converges as $n \to \infty$ for each $m = 0, 1, \cdots$) and the existence of initially periodic (with period p) ε-optimal policies. Incidentally, it appears that the proof of the main lemma in Brown is either incorrect or incomplete. More specifically, when there are more than two states, the argument of the last paragraph of his Lemma 4.7 is not valid as given. The periodicity cannot in general be removed from the discrete time results, as is shown by examples employing stationary

Received August 1972.

[1] This paper is based on parts of the author's Ph. D. dissertation written at Stanford University under the direction of Arthur F. Veinott, Jr., and partially supported by the National Science Foundation through Grant GK-18339 and the Office of Naval Research through Contract N00014-67-A-0112-0050.

AMS 1970 *subject classifications*. Primary 90C40; Secondary 90B99, 93E20.

Key words and phrases. Markov decision chains, maximal rewards, ε-optimal policies, initially stationary policies, dynamic programming.

policies for which the least common multiple of the periods of the recurrent classes is greater than one. In continuous time, however, under any policy the probability of being in a state at a time s units beyond t, given the system is in that state at time t, is strictly positive for all $s > 0$. In other words, "periodic states" do not exist, suggesting the results of this paper described above. We remark that we will not prove these results by incorporating the "aperiodicity" into arguments paralleling the discrete time development. We found instead that, after the preliminaries, different methods were necessary.

1.1. *The decision processes.* For the most part our formulation of continuous time Markov decision chains follows Miller [13]. However, unlike Miller, we find it convenient to use the reversed time orientation common to many dynamic programming studies. More specifically, we take as our time index set the non-negative real line $[0, \infty)$. Then the duration of any particular decision process can be any $t \geqq 0$, and we assume that if the duration is t, then the process *begins* at time t, with the time index *decreasing* as the process evolves, and with the process *terminating* at time zero.

We consider a system that is always in one of N *states* and we let $S = \{1, 2, \cdots, N\}$ denote the *state space*. For each instant that the system is in the state i, an *action* a is selected from the finite set A_i, one such set being defined for each $i \in S$. Note that the action chosen upon entry to a state need not be the one used for the entire stay in that state. Associated with each action $a \in A_i$ is both a set of *state transition rates* $\{q(j \mid i, a), j \in S\}$ and a *reward rate* $r(i, a)$. The transition rates are such that $q(j \mid i, a) \geqq 0$ for all $j \neq i$, with $\sum_{j=1}^{N} q(j \mid i, a) = 0$. For each $j \neq i$, $q(j \mid i, a)$ can be interpreted as the transition probability 'rate' from state i to state j when the action $a \in A_i$ has been chosen. The number $r(i, a)$ represents the reward earned per unit time whenever the system is in state i and a is the selected action.

When the process stops at time 0 a *terminal reward* is received, its value depending on the final state of the system. In other words, there is an N component column vector of terminal rewards such that if a process ends in state i, then the ith component of that vector is the terminal reward earned.

We define $F = \bigtimes_{i=1}^{N} A_i$ and call $f \in F$ a *decision rule*. Each such decision rule is a function assigning to each state i an action in the corresponding set A_i. For each $f \in F$ we let $Q(f)$ be the $N \times N$ *generator* matrix whose ijth element is $q(j \mid i, f(i))$, and let $r(f)$ be the N component column vector whose ith component is $r(i, f(i))$.

A *policy* $\pi : [0, \infty) \to F$ is any measurable function which specifies for each $t \geqq 0$ a decision rule in F. Here π measurable means that for every $f \in F$, the set $\{t \geqq 0 : \pi(t) = f\}$ is a Lebesgue measurable subset of $[0, \infty)$. Using the policy π in the system means that if at time t we are in state i and $\pi(t) = f$, then $f(i)$ is the action in A_i to be chosen. Notice that for convenience each policy π is defined on the entire interval $[0, \infty)$, while in a decision process of duration t

which operates using π, only the decision rules $\pi(s)$, $0 \leqq s \leqq t$, are used, with $\pi(t)$ the *first* decision rule used and $\pi(0)$ the *last*.

If π is such that $\pi(t) = f$ for all $t \geqq 0$ for some $f \in F$, then the policy is called *stationary* and is denoted f^∞. We will say π is *initially stationary* if there is a $0 \leqq s < \infty$ and an $f \in F$ for which $\pi(t) = f$ for all $t \geqq s$.

Miller [12], [13] proved that for any policy π the accompanying set of generators $\{Q(\pi(t)), t \geqq 0\}$ determine a continuous time (in general, *non*-stationary) Markov chain with piecewise constant sample paths. Let $P(\cdot, \cdot; \pi)$ be the $N \times N$ matrix *transition function* associated with this chain; i.e. for each $0 \leqq s \leqq t$, the ijth element of $P(t, s; \pi)$ is the probability that the system is in state j at time s, given it was in state i at time t and that π is being used. In the sequel we will abbreviate $P(t, 0; \pi)$ to $P(t; \pi)$.

For every $t \geqq 0$ we define $V^t(\pi, x)$ to be the N component column vector of *total expected rewards* earned during a process of duration t—or, equivalently, the total expected remaining rewards in a process t time units from termination—when the system is operated under the policy π and the terminal reward vector is $x \in \mathbb{R}^N$. The ith component of $V^t(\pi, x)$ is thus the total expected reward earned given that the process starts from state i at time t. Evidently,

$$V^t(\pi, x) = \int_0^t P(t, s; \pi) r(\pi(s))\, ds + P(t; \pi)x \qquad \text{for all} \quad t \geqq 0 .$$

Note that it follows from the Markovian nature of our system that if π and π' are two policies such that for some $t \geqq 0$ and all $s \geqq 0$, $\pi'(s) = \pi(t + s)$, then $V^{t+s}(\pi, x) = V^s(\pi', V^t(\pi, x))$ for every $s \geqq 0$.

1.2. *Statement of results.* For any function ϕ with range in $\mathbb{R}^N$, we let $|\phi(u)|_\infty = \max_{i \in S} |\phi_i(u)|$ and we define $\sup_u \phi(u)$, and similar operations, component-wise; e.g. $[\sup_u \phi(u)]_i \equiv \sup_u \phi_i(u)$ for all $i \in S$.

When the terminal reward is $x \in \mathbb{R}^N$, the policy π is called *x-optimal* if $V^t(\pi, x) \geqq V^t(\pi', x)$ for all policies π' and all $t \geqq 0$. We let π_x^* denote any such policy. Observe that such a policy maximizes total expected rewards *simultaneously* for every $i \in S$ and $t \geqq 0$. Also note that $V^{t+s}(\pi_x^*, x) = V^s(\pi_z^*, z)$ for all t, $s \geqq 0$, where $z \equiv V^t(\pi_x^*, x)$. We denote by v the actual terminal reward vector of our decision process and call π_v^* *optimal*. For convenience, we let $\pi^* = \pi_v^*$. The policy π is called *ε-optimal* if, for $\varepsilon > 0$,

$$\sup_{t \geqq 0} |V^t(\pi^*, v) - V^t(\pi, v)|_\infty \leqq \varepsilon .$$

Thus an ε-optimal policy must produce total expected rewards within ε of the maximum possible *simultaneously* for each $i \in S$ and every $t \geqq 0$.

We set $U = \limsup_{t \to \infty} t^{-1} V^t(\pi^*, v)$. It is evident that $\sup_\pi \limsup_{t \to \infty} t^{-1} V^t(\pi, v) = U$, so U is the *maximum* possible long-run *average return rate*.

The two main results to be established can now be restated as

THEOREM 1. *$V^t(\pi^*, v) - tU$ converges as $t \to \infty$.*

THEOREM 2. *There is an $f \in F$ such that for every $\varepsilon > 0$, there is a $t(\varepsilon) > 0$ for*

which the initially stationary policy π^ε,

$$\pi^\varepsilon(t) \equiv \pi^*(t) \quad \text{for} \quad t < t(\varepsilon)$$
$$\equiv f \quad \text{for} \quad t \geqq t(\varepsilon),$$

is ε*-optimal.*

2. Preliminaries. We first summarize some useful results on stationary policies and the decision rules they are constructed from. When the stationary policy f^∞ is used, the resultant process is the continuous time, *stationary* Markov chain generated by the matrix $Q(f)$, with $P(t, s; f^\infty) = e^{(t-s)Q(f)}$ for all $t \geqq s \geqq 0$ (where $e^{uQ(\cdot)} \equiv I + \sum_{n=1}^{\infty} (u^n/n!)Q^n(\cdot)$). It follows that $P(t, s; f^\infty) = P(t - s; f^\infty)$ for all $t \geqq s \geqq 0$. Further, for each $f \in F$ there is an $N \times N$ stochastic matrix $P^*(f)$ such that $\lim_{t\to\infty} P(t; f^\infty) = P^*(f)$; Doob ([4] page 236). Additionally, $P^*(f)$ is such that $P^*(f) = P^*(f)P^*(f) = P(t; f^\infty)P^*(f) = P^*(f)P(t; f^\infty)$ for all $t \geqq 0$, so $Q(f)P^*(f) = P^*(f)Q(f) = 0$.

For each $f \in F$, let $y(f) = \int_0^\infty [P(t; f^\infty) - P^*(f)]r(f)\,dt$, which is known to exist ([14] page 561) and be the unique solution of

$$P^*(f)y(f) = 0, \tag{1}$$

$$Q(f)y(f) = P^*(f)r(f) - r(f). \tag{2}$$

LEMMA 1. *For* $f \in F$ *and* $x \in \mathbb{R}^N$,

(i) $V^t(f^\infty, x) = x + \sum_{n=1}^{\infty} (t^n/n!)Q^{n-1}(f)[r(f) + Q(f)x]$, *for all* $t \geqq 0$.
(ii) $(d/dt)V^t(f^\infty, x) = r(f) + Q(f)V^t(f^\infty, x)$ *for all* $t \geqq 0$.
(iii) $V^t(f^\infty, x) - tP^*(f)r(f) \to y(f) + P^*(f)x$ *as* $t \to \infty$.

PROOF. Parts (i) and (ii) are proved in [13], page 271, for the time orientation opposite to ours. We prove part (iii). Since $P(t, s; f^\infty) = P(t - s; f^\infty)$, it follows that for all $t \geqq 0$, $V^t(f^\infty, x) = \int_0^t P(u; f^\infty)r(f)\,du + P(t; f^\infty)x = \int_0^t [P(u; f^\infty) - P^*(f)]r(f)\,du + P(t; f^\infty)x + tP^*(f)r(f)$. The result follows by letting $t \to \infty$ and noting the definition of $y(f)$ and $P^*(f)$.

From part (iii) of this lemma it follows that $\lim_{t\to\infty} t^{-1}V^t(f^\infty, v) = P^*(f)r(f)$. We define $F' = \{f: f \in F, P^*(f)r(f) = U\}$. Thus if $f \in F'$, then f^∞ maximizes the long-run average return rate over all policies π. It is known ([14] Section 5) that F' is not empty.

We say a policy π is piecewise constant and right continuous if the function $\pi(t)$ is piecewise constant in t and if $\pi(t) = \lim_{s\downarrow 0} \pi(t + s)$ for all $t \geqq 0$. An easy extension of a main result of Miller, proven only for $x \equiv 0$ in [13], is that for any terminal reward vector $x \in \mathbb{R}^N$, there exists a piecewise constant, right continuous, x-optimal policy. (Actually, the decision rules could be selected arbitrarily at the discontinuity points of the policy, but we find it convenient to specify right continuity.) Therefore, without loss of generality, we refine our definition of optimality so that every x-optimal policy is piecewise constant and right continuous. We can then infer from [13] the following result.

LEMMA 2. *When the terminal reward is* $x \in \mathbb{R}^N$,

(i) *a necessary and sufficient condition for the piecewise constant, right continuous policy π to be x-optimal is that*

$$r(\pi(t)) + Q(\pi(t))V^t(\pi, x) \geqq r(f) + Q(f)V^t(\pi, x) \quad \text{for all } f \in F \text{ and } t \geqq 0.$$

(ii) *For any x-optimal policy π_x^*, $V^t(\pi_x^*, x)$ is continuously differentiable in t and*

$$\frac{d}{dt} V^t(\pi_x^*, x) = r(\pi_x^*(t)) + Q(\pi_x^*(t))V^t(\pi_x^*, x) \quad \text{for all } t \geqq 0.$$

The remainder of this section is in the same spirit as parts of Brown [3]. The required proofs are provided in Lembersky [10]. The first result insures the existence of a limit point of $\{V^t(\pi^*, v) - tU, t \geqq 0\}$.

LEMMA 3. *$V^t(\pi^*, v) - tU$ is bounded in t.*

Next we introduce a convenient 'condensed' decision process system. Every action, decision rule, and policy of this new system is also a part of our original system, but not conversely. We define for each $i \in S$, $\tilde{A}_i = \{a : a \in A_i, \sum_{j=1}^N q(j \mid i, a)U_j = 0\}$. Thus the reduced set of decision rules is $\tilde{F} \equiv \bigtimes_{i=1}^N \tilde{A}_i = \{f : f \in F, Q(f)U = 0\}$. Also, $\tilde{r}(i, a) \equiv r(i, a) - U_i$ for all $i \in S$ and $a \in \tilde{A}_i$. Thus $\tilde{r}(f) = r(f) - U$ for each $f \in \tilde{F}$. We let $\tilde{V}^t(\tilde{\pi}, x)$ be the total expected reward earned in the condensed system during a process of duration t, when operating under the policy $\tilde{\pi}$ and when the terminal reward is x. We make the obvious definitions for $\tilde{\pi}_x^*$, $\tilde{y}(f)$, $\tilde{F}'$, etc.

The link between the two systems is provided by

LEMMA 4. *There exists $t^* \geqq 0$ such that $\pi^*(t) \in \tilde{F}$ for all $t \geqq t^*$.*

We set $w = V^{t^*}(\pi^*, v)$ and specify that w is to be the terminal reward of the condensed system. We also define $\tilde{\pi}^* = \tilde{\pi}_w^*$. We then can show using Lemma 4 that

$$\tilde{V}^t(\tilde{\pi}^*, w) = V^t(\pi_w^*, w) - tU \quad \text{for all } t \geqq 0. \tag{3}$$

This follows since whenever π is piecewise constant and $\pi(t) \in \tilde{F}$ for all $t \geqq 0$ then $\tilde{V}^t(\pi, x) = V^t(\pi, x) - tU$ for all $t \geqq 0$, and further, if $\pi_x^*(t) \in \tilde{F}$ for all $t \geqq 0$ then $\tilde{\pi}_x^* = \pi_x^*$.

Note that since $\tilde{U} = \limsup_{t\to\infty} t^{-1}\tilde{V}^t(\tilde{\pi}^*, w)$, it follows from (3) that $\tilde{U} = 0$.

3. Proof of Theorem 1. Since $V^t(\pi_w^*, w) - tU - t^*U = V^{t+t^*}(\pi_v^*, v) - (t + t^*)U$ for all $t \geqq 0$, we have by (3) that to prove Theorem 1 it is sufficient to show that $\tilde{V}^t(\tilde{\pi}^*, w)$ converges as $t \to \infty$, which is what we do in this section. Therefore, throughout this section we will consider only the condensed system. For convenience, we drop the tilde notation.

We start with two results giving some conditions for convergence. We omit the proofs, which use standard methods and are in Lembersky [10].

LEMMA 5. *Suppose $\{f^n(x, z), n = 1, 2, \cdots\}$ is a family of functions from $X \times Z$ into $\mathbb{R}^N$ and $g(x, z)$ is a function from $X \times Z$ into $\mathbb{R}^N$ for which*

(i) $f^n(x, z) \to g(x, z)$ *uniformly in* x *and* z *as* $n \to \infty$, *and*
(ii) *for each* $x \in X$ *there exists* $z(x) \in Z$ *such that* $\sup_z g(x, z) = g(x, z(x))$.
Then

$$\sup_z f^n(x, z) \to \sup_z g(x, z) \qquad \text{uniformly in} \quad x \qquad \text{as} \quad n \to \infty .$$

LEMMA 6. *Suppose* $\{f^{n,m}, n, m = 1, 2, \cdots\}$ *and* $\{g^m, m = 1, 2, \cdots\}$ *are collections of elements of* $\mathbb{R}^N$ *and* h *is an element of* $\mathbb{R}^N$ *for which*

(i) $f^{n,m} \to g^m$ *uniformly in* m *as* $n \to \infty$, *and*
(ii) $f^{m,m} \to h$ *as* $m \to \infty$.

Then $g^m \to h$ *as* $m \to \infty$.

There will be several key applications of the following result.

LEMMA 7. *Suppose there is a set* $\{t_n, n = 1, 2, \cdots\}$ *and a* $W \in \mathbb{R}^N$ *for which* $t_n \to \infty$ *and* $V^{t_n}(\pi_x{}^*, x) \to W$ *as* $n \to \infty$. *Then*

$$V^{t_n+s}(\pi_x{}^*, x) \to V^s(\pi_W{}^*, W) \qquad \text{uniformly in} \quad s \geqq 0 \qquad \text{as} \quad n \to \infty .$$

PROOF. Let $\varepsilon > 0$ and set $V^{t_n} = V^{t_n}(\pi_x{}^*, x)$. Then there is an $n(\varepsilon)$ such that $n \geqq n(\varepsilon)$ implies $|V^{t_n} - W|_\infty \leqq \varepsilon$. Thus, for $n \geqq n(\varepsilon)$, $|P(s; \pi)(V^{t_n} - W)|_\infty \leqq \varepsilon$ for all $s \geqq 0$ and policies π. However, since $V^s(\pi, V^{t_n}) - V^s(\pi, W) = P(s; \pi)(V^{t_n} - W)$, if follows that for all $n \geqq n(\varepsilon)$, $s \geqq 0$, and policies π

$$|V^s(\pi, V^{t_n}) - V^s(\pi, W)|_\infty \leqq \varepsilon .$$

Consequently, if we now identify the set $\{s\colon s \geqq 0\}$ with X, the set of policies with Z, $V^s(\pi, V^{t_n})$ with $f^n(x, z)$, and $V^s(\pi, W)$ with $g(x, z)$, we have that (i) of Lemma 5 is satisfied. Clearly (ii) also is satisfied. Therefore we can conclude that $V^s(\pi^*_{Vt_n}, V^{t_n}) \to V^s(\pi_W{}^*, W)$ uniformly in $s \geqq 0$ as $n \to \infty$, from which the desired result follows.

Let $\mathscr{Y}$ be the set of limit points of $\{V^t(\pi^*, w), t \geqq 0\}$. We now establish some properties of $\mathscr{Y}$ and its members.

LEMMA 8. *The set* $\mathscr{Y}$ *is not empty. Suppose* $Y \in \mathscr{Y}$. *Then*

(i) $\{V^t(\pi_Y{}^*, Y), t \geqq 0\} \subset \mathscr{Y}$.
(ii) $\{V^t(\pi_Y{}^*, Y), t \geqq 0\}$ *is a bounded set.*
(iii) Y *is a limit point of* $\{V^t(\pi_Y{}^*, Y), t \geqq 0\}$.

PROOF. By Lemma 3 and (3), it follows that $V^t(\pi^*, w)$ is bounded in t. Therefore, $\mathscr{Y}$ is not empty.

Let $Y \in \mathscr{Y}$. Then there is a set $\{t_n, n = 1, 2, \cdots\}$ such that $t_n \to \infty$ and $V^{t_n}(\pi^*, w) \to Y$ as $n \to \infty$.

By Lemma 7 it follows that for any $t \geqq 0$, $V^{t_n+t}(\pi^*, w) \to V^t(\pi_Y{}^*, Y)$ as $n \to \infty$, establishing (i).

The second result follows from (i) and the fact that, since $V^t(\pi^*, w)$ is bounded in t, it must be that the set $\mathscr{Y}$ is bounded.

Since $t_n \to \infty$ as $n \to \infty$, we can find a subsequence $\{t_n', n = 1, 2, \cdots\}$ such that $t_n' \to \infty$ and $(t_{n+1}' - t_n') \to \infty$ as $n \to \infty$. For $m = 1, 2, \cdots$, we let $s_m = t_{m+1}' - t_m' \geqq 0$. Obviously $V^{t_m'+s_m}(\pi^*, w) = V^{t_{m+1}'}(\pi^*, w)$, so $V^{t_m'+s_m}(\pi^*, w) \to Y$ as $m \to \infty$. Also, by Lemma 7, $V^{t_n'+s_m}(\pi^*, w) \to V^{s_m}(\pi_Y{}^*, Y)$ uniformly in m as $n \to \infty$. Thus, with the proper identification, (iii) follows by Lemma 6, since $s_m \to \infty$ as $m \to \infty$.

For $f \in F$, let $C(f)$ be the set of recurrent states in the Markov chain associated with the policy f^∞. Also, define $C = \bigcup_{f \in F'} C(f)$.

For $x \in \mathbb{R}^N$ we let $D(x)$ be the set of decision rules g for which $g \in F'$ and for which for all $f \in F$, whenever $[P^*(f)r(f)]_i = 0$ for some $i \in S$, then $[y(g) + P^*(g)x]_i \geqq [y(f) + P^*(f)x]_i$. For $g \in D(x)$, let $x^* = y(g) + P^*(g)x$, which is clearly well defined.

LEMMA 9. *For $x \in \mathbb{R}^N$, $D(x)$ is not empty. Further,*

(i) $r(f) + Q(f)x^* \leqq 0$ *for* $f \in F$.

(ii) $r(g) + Q(g)x^* = 0$ *for* $g \in D(x)$.

A more general version ($U \not\equiv 0$) of this result is given in Lembersky ([10] pages 12–13). The discrete time decision process analog of that result is in Veinott [22] for $x \equiv 0$ and is extended to $x \not\equiv 0$ in Lanery [8]. The argument of Veinott ([23] Section 5), which establishes that the discrete time result for $x \equiv 0$ also holds in continuous time, can be used to provide a proof of the result in [10].

For the balance of this section we set $V^t = V^t(\pi_Y{}^*, Y)$ for all $t \geqq 0$, where Y is chosen from $\mathscr{Y}$.

LEMMA 10. *Suppose $Y \in \mathscr{Y}$ and $f \in F'$. Then*

(i) $Y \geqq Y^* \geqq y(f) + P^*(f)Y$.

(ii) $Y_i = [y(f) + P^*(f)Y]_i$ *for all* $i \in C(f)$.

(iii) $Y_i = Y_i{}^*$ *for all* $i \in C$.

PROOF. By Lemma 8 (iii) there is a set $\{s_m, m = 1, 2, \cdots\}$ such that $s_m \to \infty$ and $V^{s_m} \to Y$ as $m \to \infty$. Let $g \in D(Y)$. Then $P^*(g)r(g) = 0$ and so by Lemma 1 (iii), as $m \to \infty$, $V^{s_m}(g^\infty, Y) \to y(g) + P^*(g)Y = Y^*$. Since $V^t \geqq V^t(g^\infty, Y)$ for all $t \geqq 0$, it follows that $Y \geqq Y^*$. But $Y^* \geqq y(f) + P^*(f)Y$ for $f \in F'$; therefore, (i) is established.

Let $\delta = Y - y(f) - P^*(f)Y$. Then by (1), $P^*(f)\delta = 0$. However, if $i \in C(f)$, then $P_{ij}^*(f)$ is strictly positive whenever j is in the same communicating class as i, and is zero otherwise. Therefore, since $\delta \geqq 0$, it follows that $\delta_i = 0$ for all $i \in C(f)$. This immediately establishes (ii) and implies $Y_i = Y_i{}^*$ for all $i \in C(f)$ whenever $f \in F'$, from which (iii) follows.

LEMMA 11. *Suppose $Y \in \mathscr{Y}$. Then*

(i) $0 \leqq V^{t+s} - Y^* \leqq P(t + s, t; \pi_Y{}^*)(V^t - Y^*)$ *for all* $t, s \geqq 0$.

In fact,

(ii) *there exists* $\alpha \geqq 0$ *such that* $\max_{i \in S} [V^t - Y^*]_i = \alpha$ *for all* $t \geqq 0$.

PROOF. Let $g \in D(Y)$. Then it follows from Lemma 1 (i) and Lemma 9 (ii) that $V^t(g^\infty, Y^*) = Y^*$ for all $t \geqq 0$. Consequently, by Lemma 9 (i) and (ii) we have that for every $f \in F$, $r(f) + Q(f)V^t(g^\infty, Y^*) \leqq r(g) + Q(g)V^t(g^\infty, Y^*)$ for all $t \geqq 0$. This implies, by Lemma 2 (i), that g^∞ is Y^*-optimal.

Therefore, for $t, s \geqq 0$,

$$\begin{aligned} V^{t+s} &= \int_t^{t+s} P(t+s, u; \pi_Y^*) r(\pi_Y^*(u))\, du + P(t+s, t; \pi_Y^*) V^t \\ &= \int_t^{t+s} P(t+s, u; \pi_Y^*) r(\pi_Y^*(u))\, du + P(t+s, t; \pi_Y^*) Y^* \\ &\qquad + P(t+s, t; \pi_Y^*)(V^t - Y^*) \\ &\leqq V^s(g^\infty, Y^*) + P(t+s, t; \pi_Y^*)(V^t - Y^*) \\ &= Y^* + P(t+s, t; \pi_Y^*)(V^t - Y^*)\,. \end{aligned}$$

On the other hand, using Lemma 10 (i), $V^u \geqq V^u(g^\infty, Y) = V^u(g^\infty, Y^*) + P(u; g^\infty)(Y - Y^*) = Y^* + P(u; g^\infty)(Y - Y^*) \geqq Y^*$ for all $u \geqq 0$. Thus it follows that $0 \leqq V^{t+s} - Y^* \leqq P(t+s, t; \pi_Y^*)(V^t - Y^*)$ for all $t, s \geqq 0$, which is (i).

This result implies that $[V^{t+s} - Y^*]_j \leqq \max_{i \in S} [V^t - Y^*]_i$ for all $j \in S$ and $t, s \geqq 0$. Thus

$$\max_{i \in S} [V^t - Y^*]_i \quad \text{is non-increasing in} \quad t\,. \tag{4}$$

By setting $t = 0$ we obtain that

$$\max_{i \in S} [V^t - Y^*]_i \leqq \max_{i \in S} [Y - Y^*]_i \qquad \text{for all} \quad t \geqq 0\,. \tag{5}$$

Further, by (i), $\max_{i \in S} [V^t - Y^*]_i$ is bounded below by zero. Therefore, (4) implies that there exists $\alpha \geqq 0$ such that $\max_{i \in S} [V^t - Y^*]_i \downarrow \alpha$ as $t \to \infty$. However, by Lemma 8 (iii), there is a set $\{s_m, m = 1, 2, \cdots\}$ such that $s_m \to \infty$ and $\max_{i \in S} [V^{s_m} - Y^*]_i \to \max_{i \in S} [Y - Y^*]_i$ as $m \to \infty$. Therefore it must be that $\alpha = \max_{i \in S} [Y - Y^*]_i$. This, with (4) and (5), implies (ii).

LEMMA 12. *Suppose $Y \in \mathscr{Y}$. Then $V_i^t = Y_i^*$ for all $i \in C$, $t \geqq 0$.*

PROOF. Let $f \in F'$. The structure of $Q(f)$ is such that $Q_{ij}(f) = 0$ whenever $i \in C(f)$ and $j \notin C(f)$. Therefore, by Lemma 10 (ii) and (2), $[Q(f)Y]_i = [Q(f)(y(f) + P^*(f)Y)]_i = -r_i(f)$ for all $i \in C(f)$. It therefore follows that $[\max_{g \in F} \{r(g) + Q(g)Y\}]_i \geqq 0$ for all $i \in C$. Certainly this argument applies to every limit point in $\mathscr{Y}$. Therefore, by Lemma 8 (i) we have that $[\max_{g \in F} \{r(g) + Q(g)V^t\}]_i \geqq 0$ for all $i \in C$, $t \geqq 0$. Application of Lemma 2 (i) and (ii) then implies that $[(d/dt)V^t]_i \geqq 0$, so V_i^t is non-decreasing in t for each $i \in C$. Consequently, by Lemma 8 (ii), for each $i \in C$ there is a $\bar{y}(i)$ such that $V_i^t \uparrow \bar{y}(i)$ as $t \to \infty$. But Lemma 8 (iii) implies that $\bar{y}(i) = Y_i$ and thus that $V_i^t \uparrow Y_i$ as $t \to \infty$ for all $i \in C$. But $V^0 = Y$; from this and Lemma 10 (iii), the theorem follows.

We can now prove the convergence result.

PROOF OF THEOREM 1. Let $Y \in \mathscr{Y}$ and let α be as defined in Lemma 11. We first show that $\alpha = 0$. Let $f = \pi_Y^*(0)$. Then, since π_Y^* is piecewise constant and right continuous, there is a $t_1 > 0$ such that $\pi_Y^*(t) = f$ for all $0 \leqq t < t_1$.

We shall write $i \rightsquigarrow j$ if state j is accessible from state i when the policy f^∞ is used. It is well known that $i \rightsquigarrow j$ if and only if $P_{ij}(s; f^\infty) > 0$ for all $s > 0$.

Select any $0 < t < t_1$. By Lemma 11 (ii) there is an $i \in S$ for which $[V^t - Y^*]_i = \alpha$. There are two cases to consider.

Case 1. There is a $0 < u < t$ and a $j \in S$ such that $i \rightsquigarrow j$ and $[V^u - Y^*]_j < \alpha$.

Let $s = t - u$. Then, with the aid of both parts of Lemma 11, we obtain the contradiction that $\alpha = [V^{u+s} - Y^*]_i \leqq [P(s; f^\infty)(V^u - Y^*)]_i \leqq P_{ij}(s; f^\infty)[V^u - Y^*]_j + \alpha(1 - P_{ij}(s; f^\infty)) < \alpha$, the last inequality since $P_{ij}(s; f^\infty) > 0$ and $[V^u - Y^*]_j < \alpha$.

Case 2. If $i \rightsquigarrow j$ then $[V^u - Y^*]_j = \alpha$ for all $0 < u < t$.

Let $S' = \{j \in S : i \rightsquigarrow j\}$. S' is never empty since $i \rightsquigarrow i$. Note that for $j \in S'$ there is no $k \notin S'$ for which $j \rightsquigarrow k$, for if there were then $i \rightsquigarrow k$. Thus there exists a nonempty set S'' of recurrent states of the Markov chain resulting from the use of the policy f^∞ which is such that $S'' \subset S'$. Also, since $Q_{jk}(f) = 0$ whenever $j \in S'$ and $k \notin S'$, since each row of $Q(f)$ sums to zero, and since

$$V_j^u = Y_j^* + \alpha \qquad \text{for all} \quad j \in S' \quad \text{and} \quad 0 < u < t, \tag{6}$$

it follows that $[Q(f)V^u]_j = [Q(f)Y^*]_j$ for all $0 < u < t$ and $j \in S'$. Then, using (6) and Lemma 1 (ii), we obtain that

$$0 = [r(f) + Q(f)Y^*]_j \qquad \text{for all} \quad j \in S'.$$

Let $g \in D(Y)$. Using Lemma 9 (ii), if we define the decision rule h so that

$$\begin{aligned} h(k) &= f(k) \qquad \text{for} \quad k \in S' \\ &= g(k) \qquad \text{for} \quad k \notin S', \end{aligned}$$

then $r(h) + Q(h)Y^* = 0$. Pre-multiplication by $P^*(h)$ reveals that $h \in F'$, and therefore $C(h) \subset C$. However, $S'' \subset C(h)$ and so $S'' \subset C$. Since S'' is nonempty, it follows by (6) and Lemma 12 that $\alpha = 0$.

Using this fact, we now finish the proof. Recall that it suffices to show $\mathscr{Y} = \{Y\}$. Let $Z \in \mathscr{Y}$. There is a sequence $\{t_n, n = 1, 2, \cdots\}$ such that $t_n \to \infty$ and $V^{t_n}(\pi^*, w) \to Y$ as $n \to \infty$. Additionally, there is a second sequence $\{t_n', n = 1, 2, \cdots\}$ such that $t_n' \to \infty$ and $V^{t_n'}(\pi^*, w) \to Z$ as $n \to \infty$, and having the further property that $t_n' \geqq t_n$ for all n. Define $s_m = t_m' - t_m \geqq 0$ for each m. Lemma 7 implies that $V^{t_n + s_m}(\pi^*, w) \to V^{s_m}$ uniformly in m as $n \to \infty$. Certainly $V^{t_m + s_m}(\pi^*, w) \to Z$ as $m \to \infty$. Therefore, by Lemma 6, $V^{s_m} \to Z$ as $m \to \infty$. However, by Lemma 11 (i) and (ii), and since $\alpha = 0$, $V^t = Y^*$ for all $t \geqq 0$. Therefore, $Z = Y^*$. But $V^0 = Y$, and consequently $Z = Y^* = Y$, completing the proof.

4. Proof of Theorem 2. We now use Theorem 1 to prove that there are initially stationary ε-optimal policies, with the decision rule of the stationary part independent of ε. Recall that properties of such policies are valid simul-

taneously for every process duration t. We revert to using the tilde notation when referring to the condensed system.

PROOF OF THEOREM 2. Theorem 1 established that there is a Y to which $\tilde{V}^t(\tilde{\pi}^*, w)$ converges and that $Y = Y^*$. Thus, for every $\varepsilon > 0$ there is an $s(\varepsilon)$ such that $t \geqq s(\varepsilon)$ implies $|\tilde{V}^t(\tilde{\pi}^*, w) - Y^*|_\infty \leqq \varepsilon/2$. If we let $x = \tilde{V}^{s(\varepsilon)}(\tilde{\pi}^*, w)$, then

$$|\tilde{V}^s(\tilde{\pi}_x{}^*, x) - Y^*|_\infty \leqq \varepsilon/2 \qquad \text{for all} \quad s \geqq 0 . \tag{7}$$

Let $f \in \tilde{D}(Y)$, then $\tilde{V}^s(f^\infty, Y^*) = Y^*$ for all $s \geqq 0$. Additionally, since by (7) with $s = 0$, $|x - Y^*|_\infty \leqq \varepsilon/2$, and since $\tilde{V}^s(f^\infty, x) - \tilde{V}^s(f^\infty, Y^*) = P(s; f^\infty)(x - Y^*)$, we have that $|\tilde{V}^s(f^\infty, x) - \tilde{V}^s(f^\infty, Y^*)|_\infty \leqq \varepsilon/2$. It thus follows that

$$|\tilde{V}^s(\tilde{\pi}_x{}^*, x) - \tilde{V}^s(f^\infty, x)|_\infty \leqq \varepsilon \qquad \text{for all} \quad s \geqq 0 . \tag{8}$$

Now set $t(\varepsilon) = s(\varepsilon) + t^*$, where t^* is as in Lemma 4, and define π^ε as in the statement of the theorem. Let $t \geqq t(\varepsilon)$ and $s = t - t(\varepsilon)$. Then, using (3), $V^t(\pi^\varepsilon, v) = V^s(f^\infty, V^{t(\varepsilon)}(\pi^*, v)) = V^s(f^\infty, x + s(\varepsilon)U) = V^s(f^\infty, x) + s(\varepsilon)P(s; f^\infty)U$. But since $f \in \tilde{F}$ and $P(s; f^\infty) = e^{sQ(f)}$, it follows that $P(s; f^\infty)U = U$, and thus that $V^s(f^\infty, x) = \tilde{V}^s(f^\infty, x) + sU$. Consequently,

$$V^t(\pi^\varepsilon, v) = \tilde{V}^s(f^\infty, x) + (s + s(\varepsilon))U . \tag{9}$$

Again using (3), we have that $V^t(\pi^*, v) = V^{s+s(\varepsilon)}(\pi_w{}^*, w) = \tilde{V}^{s+s(\varepsilon)}(\tilde{\pi}^*, w) + (s + s(\varepsilon))U = \tilde{V}^s(\tilde{\pi}_x{}^*, x) + (s + s(\varepsilon))U$. Thus, by (8) and (9), $|V^t(\pi^*, v) - V^t(\pi^\varepsilon, v)|_\infty \leqq \varepsilon$ for all $t \geqq t(\varepsilon)$. This is certainly also true for $t < t(\varepsilon)$. The theorem now follows since our choice of f does not depend on ε.

5. Remarks. Let F^* be the set of decision rules which can be used in forming the initial stationary segments of the policies described in Theorem 2. From the nature of that theorem, it appears that determining such a decision rule would be of interest. We call these decision rules the *preferred* rules. In [10] we show, under a variety of hypotheses related to the state recurrence structure produced by stationary policies, how the *policy improvement* method of Howard [5] that computes an $f \in F'$ can be used to find preferred rules. Interestingly, we also show that the sets of often studied and algorithmically obtainable decision rules optimal under various criteria in infinite duration processes with *discounted* rewards (as in Blackwell [2], Miller [14] and Veinott [22, 23]) can, in general, be disjoint from F^*.

6. Acknowledgment. I am indebted to Arthur F. Veinott, Jr. for his guidance and his helpful suggestions.

REFERENCES

[1] BELLMAN, R. (1957). *Dynamic Programming*. Princeton Univ. Press.
[2] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.
[3] BROWN, B. W. (1965). On the iterative method of dynamic programming on a finite space discrete time Markov process. *Ann. Math. Statist.* **36** 1279–1285.

[4] DOOB, J. (1953). *Stochastic Processes.* Wiley, New York.
[5] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes.* Wiley, New York.
[6] KARLIN, S. (1966). *A First Course in Stochastic Processes.* Academic Press, New York.
[7] KEMENY, J. G. and SNELL, J. L. (1961). Finite continuous time Markov chains. *Theor. Probability Appl.* **6** 101–105.
[8] LANERY, E. (1967). Etude asymptotique des systèmes Markoviens à commande. *Rev. Informat. Recherche Opérationnelle* **3** 3–56.
[9] LANERY, E. (1968). Complements a l'etude asymptotique des systèmes Markoviens à commande. Institut de Recherche d'Informatique et d'Automatique, Rocquencourt, France.
[10] LEMBERSKY, M. R. (1972). Initially stationary ε-optimal policies in continuous time Markov decision chains. Technical Report No. 22, Dept. of Operations Research, Stanford Univ.
[11] MARTIN-LOF, A. (1967). Optimal control of a continuous-time Markov chain with periodic transition probabilities. *Operations Res.* **15** 872–881.
[12] MILLER, B. L. (1967). Finite state continuous-time Markov decision processes with application to a class of optimization problems in queueing theory. Technical Report No. 15, Dept. of Operations Research, Stanford Univ.
[13] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with a finite planning horizon. *SIAM J. Control* **6** 266–280.
[14] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with an infinite planning horizon. *J. Math. Anal. Appl.* **22** 552–569.
[15] MILLER, B. L. and VEINOTT, A. J., JR. (1969). Discrete dynamic programming with a small interest rate. *Ann. Math. Statist.* **40** 366–370.
[16] ODONI, A. R. (1969). On finding the maximal gain for Markov decision processes. *Operations Res.* **17** 857–860.
[17] ROMANOVSKII, I. V. (1964). Asymptotic behavior of dynamic programming processes with a continuous set of states. *Soviet Math. Dokl.* **5** 1684–1687 (Translation).
[18] RYKOV, V. V. (1966). Markov decision processes with finite state and decision spaces. *Theor. Probability Appl.* **11** 302–311.
[19] SCHWEITZER, P. J., (1965). Perturbation theory and Markovian decision processes. Technical Report No. 15, Operations Research Center, MIT, Cambridge, Massachusetts.
[20] SHAPIRO, J. F. (1968). Turnpike planning horizons for a Markovian decision model. *Man. Sci.* **14** 292–300.
[21] SHAPIRO, J. F. (1968). Shortest route methods for finite state space deterministic dynamic programming problems. *SIAM J. Appl. Math.* **16** 1232–1250.
[22] VEINOTT, A. F., JR. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.* **37** 1284–1294.
[23] VEINOTT, A. F., JR. (1969). Discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Statist.* **40** 1635–1660.

DEPARTMENT OF STATISTICS
OREGON STATE UNIVERSITY
CORVALLIS, OREGON 97331