The Annals of Mathematical Statistics
1972, Vol. 43, No. 2, 636-644

DISCRETE DYNAMIC PROGRAMMING WITH UNBOUNDED REWARDS[1]

J. MICHAEL HARRISON

Stanford University

Countable state and action Markov decision processes are investigated, the objective being to maximize expected discounted reward. Well-known results of Maitra and Blackwell are generalized, their assumption of bounded rewards being replaced by weaker conditions, the most important of which is as follows. The expected reward to be received at time $n + 1$ minus the actual reward received at time n, viewed as a function of the state at time n, the action at time n and the decision rule to be followed at time $n + 1$, can be bounded. It is shown that there exists an ε-optimal stationary policy for every $\varepsilon > 0$ and that there exists an optimal stationary policy in the finite action case.

1. Introduction. Markov decision processes, first identified by Bellman [1], constitute an important special class of dynamic programming problems. This paper deals with a stationary Markov decision process over a discrete state space S, the set of actions available in each state also being countable. The objective is to maximize expected total discounted rewards over an infinite planning horizon, the discount factor being fixed and less than one. A precise formulation of the problem is contained in Section 2.

Discounted Markov decision processes were first studied by Howard [5], who developed an algorithm for computing an optimal stationary policy in the finite state and action case. Blackwell [2] then showed that in this case an optimal stationary policy is optimal among all (non-randomized Markov) policies. A later paper by Maitra [7] proved that if rewards are bounded then there exists a stationary optimal policy in the countable state and finite action case. Finally Blackwell [3], maintaining the assumption of bounded rewards and allowing an even more general state space, showed that (among all possibly randomized and non-Markov policies) there exists a stationary ε-optimal policy for every $\varepsilon > 0$ in the countable action case and a stationary optimal policy in the finite action case.

This later work of course constitutes a very significant extension of Howard's original treatment, but still the assumption of bounded rewards is quite restrictive. Many problems of stochastic control, notably those arising in

Received April 30, 1971; revised July 7, 1971.

[1] This work was supported by a National Science Foundation traineeship. The original version was distributed as a technical report by the Department of Operations Research, Stanford University, supported by the Office of Naval Research, contract Nonr-225 (NR-042-002).

conjunction with inventory systems, are most naturally formulated with discrete state space and unbounded rewards. (For example, one component of the reward function may be proportional to the state variable.) In our treatment the assumption of bounded rewards is replaced by weaker conditions, the most important of which is as follows. The expected reward to be received at time $n+1$ minus the actual reward received at time n, viewed as a function of the state at time n, the action taken at time n and the decision rule to be followed at time $n+1$, can be bounded. Allowing a general policy to be randomized and non-Markov, our central result is that there exists a stationary ε-optimal policy for every $\varepsilon > 0$ and there exists a stationary optimal policy in the finite action case. (Examples showing that an optimal policy need not exist in the countable action case are easily constructed.) Additional results are presented concerning the policy improvement and successive approximation techniques for computation of optimal policies.

The methods used here are those developed by Blackwell [3] in his elegant treatment of the bounded rewards case, a central role being played by the Banach fixed point theorem for contraction mappings. The basic structure of the argument is roughly as follows. In Section 2 the problem is formulated and the assumptions are presented. In Section 3 it is shown that the expected discounted reward from every policy is finite, and we develop bounds which are satisfied by all such return functions. In Section 4 we define a set B of functions $S \to$ reals which contains the return function for each policy and is a complete metric space with the uniform metric ρ. (Blackwell takes B as the set of all bounded functions $S \to$ reals.) It is then shown that the optimal return operator T is a contraction mapping on (B, ρ), implying by Banach's theorem that T has a unique fixed point in B. Using this basic result, the optimality of stationary policies is then demonstrated in Section 5.

Assuming a familiarity with Blackwell's papers [2], [3], we shall in interest of brevity omit discussion and interpretation of such standard concepts as the optimal return operator.

2. Definitions and assumptions. A system with countable state space S is observed at each of a sequence of points in time labelled $1, 2, \cdots$. Each time that the system is observed in state s an action $a \in A_s$ must be chosen, where A_s is countable for each $s \in S$. A reward $r(s, a)$ is then received. The conditional probability that the system will be observed in state $t \in S$ at time $n+1$, given that it is observed in state $s \in S$ at time n and action $a \in A_s$ is selected, and given all other previous history of the decision process, is $p(t \mid s, a)$. For each $s \in S$ and $a \in A_s$, $p(\cdot \mid s, a)$ is a probability measure over S.

An n-stage history for the decision process is a $(2n-1)$-tuple $h_n = (s_1, a_1, \cdots, s_n)$ with $s_1 \in S$, $a_1 \in A_{s_1}$, $\cdots$, $s_n \in S$, and we let H_n denote the set of all possible

such histories ($n \geqq 1$). A policy π specifies for each $n \geqq 1$ which act will be chosen at time n as a function of the history h_n. More generally, a policy is a sequence $\pi = \{q_1(\cdot \mid \cdot), q_2(\cdot \mid \cdot), \cdots\}$ where, for each $n \geqq 1$ and $h_n = (s_1, a_1, \cdots, s_n) \in H_n$, $q_n(\cdot \mid h_n)$ is a probability measure over A_{s_n}. Here $q_n(a \mid h_n)$ is interpreted as the probability that action a will be chosen at time n if the history h_n is observed.

A policy $\pi = \{q_1, q_2, \cdots\}$ is called non-randomized if $q_n(\cdot \mid h_n)$ is a degenerate measure for each $n \geqq 1$ and $h_n \in H_n$. It is called (non-randomized) Markov if $q_n(\cdot \mid h_n)$ further depends on only the last component of h_n. Thus a Markov policy is of the form $\pi = \{f_1, f_2, \cdots\}$ where f_n is a function which associates with each state $s \in S$ a corresponding action $f(s) \in A_s$ ($n \geqq 1$). Such a function is called a (Markov) decision rule, and we let F denote the set of all possible decision rules. The set of all possible Markov policies will be denoted by F^∞. A Markov policy is called stationary if all of its component decision rules are identical, and we denote by f^∞ the stationary policy all of whose components are f. If $\pi = \{f_1, f_2, \cdots\} \in F^\infty$ and $g \in F$, we use the notation $(g, \pi) = \{g, f_1, f_2, \cdots\}$. Thus policy (g, π) uses decision rule g at time point 1 and then uses the component rules of π in their given order at subsequent time points.

It is assumed that all rewards earned by the system are discounted using a discount factor $\beta (0 \leqq \beta < 1)$ which is fixed. Thus, if history $h_n = (s_1, a_1, \cdots, s_n)$ is observed, the total discounted reward earned through n time periods is

$$V_n = \sum_{i=1}^{n} \beta^{i-1} r(s_i, a_i) .$$

Without some assumptions concerning the reward function $r(\cdot, \cdot)$ or the transition probability function $p(\cdot \mid \cdot, \cdot)$ there is of course no guarantee that under an arbitrary policy V_n has finite expectation. In order to avoid unnecessary clumsiness, we now state our assumptions and prove some of their consequences before stating the optimality criterion.

ASSUMPTIONS.

(i) $\sum_{t \in S} p(t \mid s, a) r(t, f(t))$ is absolutely convergent for all $s \in S$, $a \in A_s$ and $f \in F$.

(ii) There exists a bound $d > 0$ such that

$$\left| \sum_{t \in S} p(t \mid s, a) r(t, f(t)) - r(s, a) \right| \leqq d$$

for all $s \in S$, $a \in A_s$ and $f \in F$.

(iii) For each $s \in S$

$$L(s) = \inf_{a \in A_s} r(s, a) \quad \text{and} \quad U(s) = \sup_{a \in A_s} r(s, a)$$

are finite. Moreover, $U(s) - L(s)$ is bounded over $s \in S$.

Assumption (i) is of course very mild, requiring only that the expected reward at time $n + 1$ be well defined for each possible state and action at time n and

each decision rule which might be used at time $n+1$. Assumptions (ii) and (iii) are much stronger and provide the key to our analysis. They are immediately implied by an assumption of bounded rewards, however.

It is immediate that for any $\varepsilon > 0$ there exist decision rules $f, g \in F$ such that $r(s, f(s)) \geqq L(s) + \varepsilon$ and $r(s, g(s)) \geqq U(s) - \varepsilon$ for all $s \in S$. From this fact and assumption (ii) it follows easily that

$$|\textstyle\sum_{t \in S} p(t \mid s, a)L(t) - r(s, a)| \leqq d \tag{1}$$

and

$$|\textstyle\sum_{t \in S} p(t \mid s, a)U(t) - r(s, a)| \leqq d \tag{2}$$

for all $s \in S$ and $a \in A_s$.

3. The expected discounted reward is finite. Throughout this section let $\pi = \{q_1, q_2, \cdots\}$ be an arbitrary but fixed policy. Given the initial state of the system, this policy together with the transition probability function p determines the probability law of the random process $\{\sigma_1, \alpha_1, \sigma_2, \alpha_2, \cdots\}$, where σ_n is the state of the system at time n and α_n is the action taken at that time. It is understood throughout this section that all expectations are with respect to this probability law. Also, for $n \geqq 1$ let $\eta_n = (\sigma_1, \alpha_1, \cdots, \sigma_n)$, a random element of H_n.

LEMMA 1. *$|E[r(\sigma_n, \alpha_n) \mid \sigma_1 = s] - E[r(\sigma_1, \alpha_1) \mid \sigma_1 = s]| \leqq (n-1)d$ for all $n \geqq 1$ and $s \in S$.*

PROOF. The proposition is trivially true for $n = 1$. Assume it for general $n \geqq 1$. We note first that

$$\begin{aligned}
&E[r(\sigma_{n+1}, \alpha_{n+1}) - r(\sigma_n, \alpha_n) \mid \sigma_1 = s] \\
&\quad = E[\textstyle\sum_{t \in S} p(t \mid \sigma_n, \alpha_n) \sum_{a \in A_t} q(a \mid \eta_n, \alpha_n, t)r(t, a) - r(\sigma_n, \alpha_n) \mid \alpha_1 = s] \\
&\quad \leqq E[\textstyle\sum_{t \in S} p(t \mid \sigma_n, \alpha_n)U(t) - r(\sigma_n, \alpha_n) \mid \sigma_1 = s] \\
&\quad \leqq E[r(\sigma_n, \alpha_n) + d - r(\sigma_n, \alpha_n) \mid \sigma_1 = s] = d\,,
\end{aligned}$$

using (2). Using this inequality and the induction hypothesis,

$$\begin{aligned}
&E[r(\sigma_{n+1}, \alpha_{n+1}) \mid \sigma_1 = s] \\
&\qquad = E[r(\sigma_n, \alpha_n) \mid \sigma_1 = s] + E[r(\sigma_{n+1}, \alpha_{n+1}) - r(\sigma_n, \alpha_n) \mid \sigma_1 = s] \\
&\qquad \leqq E[r(\sigma_1, \alpha_1) \mid \sigma_1 = s] + (n-1)\,d + d.
\end{aligned}$$

That $E[r(\sigma_{n+1}, \alpha_{n+1}) \mid \alpha_1 = s] \geqq E[r(\sigma_1, \alpha_1) \mid \sigma_1 = s] - nd$ follows similarly, using the induction hypothesis and (1) instead of (2).

For each $n \geqq 1$ and $s \in S$ we define

$$V_n(\pi)(s) = \textstyle\sum_{i=1}^{n} \beta^{i-1} E[r(\sigma_i, \alpha_i) \mid \sigma_1 = s]\,,$$

the expected discounted n-period reward under policy π when the initial state

is s. We use the notation $||V|| = \sup_S |V(s)|$ for V a real-valued function over S.

THEOREM 2. *$V(\pi) = \lim V_n(\pi)$ exists and satisfies*

$$||V(\pi) - (1 - \beta)^{-1}V_1(\pi)|| \leqq \beta d(1 - \beta)^{-2}.$$

PROOF. From Lemma 1 we have

$$\sum_{i=1}^{\infty} \beta^{i-1}|E[r(\sigma_i, \alpha_i) \mid \sigma_1 = s] - V_1(\pi)(s)| \leqq \sum_{i=1}^{\infty} \beta^{i-1}(i-1)d = \beta d(1-\beta)^{-2},$$

from which the proposition follows.

THEOREM 3. *For any $\varepsilon > 0$ there exists a Markov policy $\pi^* \in F^\infty$ such that $V(\pi^*) \geqq V(\pi) - \varepsilon$.*

PROOF. We first note that if $\pi' = \{q_1', q_2', \cdots\}$ is another policy such that $q_1' = q_1, \cdots, q_N' = q_N$, then $V_1(\pi) = V_1(\pi')$ and from Lemma 1

$$||V(\pi) - V(\pi')|| \leqq 2\sum_{i=N+1}^{\infty} \beta^{i-1}(i-1)d \to 0 \quad \text{as} \quad N \to \infty.$$

Thus we may assume without loss of generality that π is Markov from some point on, say for all time points $n > N$. We now show that for any $\gamma > 0$ and policy $\pi = \{q_1, \cdots, q_N, f_{N+1}, \cdots\}$ of this form there exists a decision rule $f_N \in F$ such that $V(\pi') \geqq V(\pi) - \gamma$, where $\pi' = \{q_1, \cdots, q_{N-1}, f_N, \cdots\}$. Using this fact N times with $\gamma = \varepsilon/N$ will produce the desired Markov policy π^*.

For each $s \in S$ let

$$V^{N+1}(\pi)(s) = \sum_{i=N+1}^{\infty} \beta^{N+1-i}E[r(\sigma_i, f_i(\sigma_i)) \mid \sigma_{N+1} = s],$$

and choose $f_N(s)$ to be an action in A_s such that

$$r(s, f_N(s)) + \beta \sum_{t \in S} p(t \mid s, f_N(s))V^{B+1}(\pi)(t) \geqq \sup_{a \in A_s}\{r(s, a) + \beta \sum_{t \in S} p(t \mid s, a)V^{N+1}(\pi)(t)\} - \gamma.$$

Since

$$V(\pi)(s) = V_{N-1}(\pi)(s) + \beta^{N-1}E[r(\sigma_N, \alpha_N) + \beta \sum_{t \in S} p(t \mid \sigma_N, \alpha_N)V^{N+1}(\pi)(t) \mid \sigma_1 = s]$$

and

$$V(\pi')(s) = V_{N-1}(\pi)(s) + \beta^{N-1}E[r(\sigma_N, f_N(\sigma_N)) + \beta \sum_{t \in S} p(t \mid \sigma_N, f_N(\sigma_N))V^{N+1}(\pi)(t) \mid \sigma_1 = s]$$

it follows easily that the decision rule f_N has the desired property.

4. Markov plans and contraction mappings. Having shown in the previous section that $V(\pi)$, the expected total discounted reward under policy π as a function of the initial state, is finite for all π, our optimality criterion can now be stated. A policy π^* is called ε-optimal if $V(\pi^*) \geqq V(\pi) - \varepsilon$ for all π, and

it is called optimal if $V(\pi^*) \geqq V(\pi)$ for all π. In order to show that there exists an ε-optimal stationary policy, some preliminary results are needed.

The term "vector" is used hereafter to mean a function $V: S \to$ reals, and "the sth component" of a vector V refers to $V(s)$. Similarly a "matrix" is a function $P: S \times S \to$ reals, and "the (s, t)th element" of a matrix P refers to $P(s, t)$. All vectors should be envisioned as column vectors, and we define matrix multiplication and matrix-vector multiplication in the usual way when all the sums involved are absolutely convergent. For each $f \in F$ we define $r(f)$ to be the vector whose sth component is $r(s, f(s))$ and $P(f)$ to be the matrix whose (s, t)th element is $P(t \mid s, f(s))$. For a Markov policy $\pi = (f_1, f_2, \cdots)$ we define $P^0(\pi) = I$, the identity matrix, and $P^n(\pi) = P(f_1) \cdots P(f_n)$ for $n \geqq 1$. It is immediate then that

$$V_n(\pi) = \textstyle\sum_{i=0}^{n-1} \beta^i P^i(\pi) r(f_{i+1}), \qquad n \geqq 1.$$

We define B to be the set of all vectors V satisfying

$$(1 - \beta)^{-1}L - \beta d(1 - \beta)^{-2} \leqq V \leqq (1 - \beta)^{-1}U + \beta d(1 - \beta)^{-2}.$$

Since clearly $L \leqq V_1(\pi) \leqq U$ for any policy π, it is immediate from Theorem 2 that $V(\pi) \in B$ for each π.

LEMMA 4. *If $V \in B$ and $\pi = \{f_1, f_2, \cdots\} \in F^\infty$, then $\beta^n P^n(\pi) V \to 0$ as $n \to \infty$.*

PROOF. From the definition of B it clearly suffices to show that $\beta^n P^n(\pi) L \to 0$ and $\beta^n P^n(\pi) U \to 0$. From Lemma 1 we have that $||P^n(\pi) r(f) - r(f_1)|| \leqq nd$ for any $f \in F$, and hence $||P^n(\pi) L - r(f_1)|| \leqq nd$. Thus $||\beta^n P^n(\pi) L - \beta^n r(f_1)|| \leqq \beta^n nd \to 0$ as $n \to \infty$. Since $\beta^n r(f_1) \to 0$, we conclude that $\beta^n P^n(\pi) L \to 0$. That $\beta^n P^n(\pi) U \to 0$ follows in identical fashion.

Now for each $f \in F$ we define a mapping T_f on B by letting

$$T_f V = r(f) + \beta P(f) V, \qquad V \in B.$$

LEMMA 5. *For each $f \in F$, T_f maps B into itself.*

PROOF. If $V \in B$, then $V \geqq (1 - \beta)^{-1}L - \beta d(1 - \beta)^{-2}$, implying

$$T_f V = r(f) + \beta P(f) V \geqq r(f) + \beta(1 - \beta)^{-1} P(f) L - \beta^2 d(1 - \beta)^{-2}.$$

But $P(f)L \geqq r(f) - d$ by (1), so

$$\begin{aligned} T_f V &\geqq r(f) + \beta(1 - \beta)^{-1} r(f) - \beta(1 - \beta)^{-1} d - \beta^2 d(1 - \beta)^{-2} \\ &= (1 - \beta)^{-1} r(f) - \beta d(1 - \beta)^{-2} \geqq (1 - \beta)^{-1} L - \beta d(1 - \beta)^{-2}. \end{aligned}$$

That $T_f V \leqq (1 - \beta)^{-1}U + \beta d(1 - \beta)^{-2}$ follows similarly. Hence $T_f V \in B$.

Assumption (iii) states that $||U - L|| < \infty$ and thus

$$\rho(V_1, V_2) = ||V_1 - V_2|| < \infty \qquad \text{if} \quad V_1, V_2 \in B.$$

The following proposition is then elementary.

LEMMA 6. *(B, ρ) is a complete metric space.*

THEOREM 7. *For each $f \in F$, T_f is a contraction of modulus β on (B, ρ). That is, if $V_1, V_2 \in B$ then $||T_f V_1 - T_f V_2|| \leqq \beta ||V_1 - V_2||$.*

PROOF.

$$\begin{aligned} ||T_f V_1 - T_f V_2|| &= ||r(f) + \beta P(f) V_1 - r(f) - \beta P(f) V_2|| \\ &= \beta \sup_{s \in S} |\textstyle\sum_{t \in S} P(t \mid s, f(s))[V_1(t) - V_2(t)]| \\ &\leqq \beta \sup_{s \in S} \textstyle\sum_{t \in S} p(t \mid s, f(s)) ||V_1 - V_2|| = \beta ||V_1 - V_2|| \,. \end{aligned}$$

COROLLARY 8. *For each $f \in F$*

(a) *T_f has a unique fixed point $V_f \in B$,*

(b) *$T_f{}^n V \to V_f$ as $n \to \infty$ and $||T_f{}^n V - V_f|| \leqq \beta^n (1 - \beta)^{-1} ||T_f V - V||$ for each $V \in B$,*

(c) *$V_f = V(f^\infty)$.*

PROOF. Parts (a) and (b) follow directly from Lemma 6, Theorem 7, and the Banach fixed point theorem for contraction mappings (see [6]). Now note that

$$T_f{}^n V = \textstyle\sum_{i=0}^{n-1} \beta^i P^i(f^\infty) r(f) + P^n(f^\infty) V = V_n(f^\infty) + \beta^n P^n(f^\infty) V \,.$$

Since $V_n(f^\infty) \to V(f^\infty)$ and $\beta^n P^n(f^\infty) V \to 0$ by Lemma 4, we have $T_f{}^n V \to V(f^\infty)$. Thus $V_f = V(f^\infty)$ by (b), establishing (c).

We next define the familiar optimal return operator T on B by letting

$$TV = \sup_{f \in F} T_f V = \sup_{f \in F} [r(f) + \beta P(f) V] \,, \qquad V \in B \,.$$

From the structure of the sequential decision process it is immediate that for any $\varepsilon > 0$ and $V \in B$ there exists an $f \in F$ such that $T_f V \geqq TV - \varepsilon$. It then follows from Lemma 5 that T maps B into itself. We now show in the usual way that T inherits the contraction property of the mappings T_f.

THEOREM 9. *T is a contraction of modulus β on (B, ρ).*

PROOF. Let $V_1, V_2, \varepsilon B$ and for $\varepsilon > 0$ choose $f \in F$ such that $T_f V_1 \geqq TV_1 - \varepsilon$. Since $TV_2 \geqq T_f V_2$ we then have $TV_1 - TV_2 \leqq T_f V_1 - T_f V_2 + \varepsilon$, implying that

$$\begin{aligned} \sup_{s \in S} [TV_1(s) - TV_2(s)] &\leqq \sup_{s \in S} [T_f V_1(s) + T_f V_2(s)] + \varepsilon \\ &\leqq ||T_f V_1 - T_f V_2|| + \varepsilon \leqq \beta ||V_1 - V_2|| + \varepsilon \,, \end{aligned}$$

using Theorem 7. Letting $\varepsilon \to 0$ we have $TV_1 - TV_2 \leqq \beta ||V_1 - V_2||$. Similarly $TV_2 - TV_1 \leqq \beta ||V_1 - V_2||$, so $||TV_1 - TV_2|| \leqq \beta ||V_1 - V_2||$.

COROLLARY 10. (a) *T has a unique fixed point $V^* \in B$.* (b) *$T^n V \to V^*$ as $n \to \infty$ and $||T^n V - V^*|| \leqq \beta^n (1 - \beta)^{-1} ||TV - V||$ for all $V \in B$.*

5. The optimality of stationary policies. Since $P(f) \geqq 0$ it is immediate that T_f is monotone for each $f \in F$ and hence T is monotone as well. We shall use these facts frequently and without comment in the proofs of this section.

THEOREM 11. (a) $V(\pi) \leqq V^*$ *for all policies* π. (b) *For each* $\varepsilon > 0$ *there exists an* $f \in F$ *such that* $V(f^\infty) \geqq V^* - \varepsilon$. *The stationary policy* f^∞ *is* ε*-optimal.* (c) *If* $\pi \in F^\infty$ *and* $V(g, \pi) > V(\pi)$ *then* $V(g^\infty) > V(\pi)$. (d) *A policy* π^* *is optimal if and only if* $V(\pi^*) = V^*$.

PROOF. (a) Let $\pi = \{f_1, f_2, \cdots\}$ be an arbitrary Markov policy. Clearly $T_{f_1} \cdots T_{f_n} V^* \leqq T^n V^* = V^*$ for each $n \geqq 1$. But $T_{f_1} \cdots T_{f_n} V^* = V_n(\pi) + \beta^n P^n(\pi) V^*$, so letting $n \to \infty$ and using Lemma 4 we have $V(\pi) = \lim V_n(\pi) \leqq V^*$. Thus the desired result holds for Markov π, and by Theorem 3 it must then hold for all π.

(b) Let $f \in F$ be such that $T_f V^* \geqq TV^* - \varepsilon(1 - \beta) = V^* - \varepsilon(1 - \beta)$. Then using the fact that $T_f(V + c) = T_f V + \beta c$ for any constant c and $V \in B$, it is easily shown by induction that

$$T_f^n V^* \geqq V^* - \varepsilon(1 - \beta)(1 + \beta + \cdots + \beta^{n-1}), \qquad n \geqq 1.$$

Letting $n \to \infty$ and using Corollary 8 we have $V_f = V(f^\infty) \geqq V^* - \varepsilon$. That f^∞ is ε-optimal is then immediate from (a).

(c) Since $V(g, \pi) = T_g V(\pi)$ we have by the monotonicity of T_g that $T_g^n V(\pi) \geqq \cdots \geqq T_g V(\pi) > V(\pi)$. Thus, letting $n \to \infty$ and using Corollary 8,

$$V_g = V(g^\infty) \geqq T_g V(\pi) > V(\pi).$$

(d) If $V(\pi^*) = V^*$ then π^* is optimal by (a). Now note that either $T_g V(\pi^*) \leqq V(\pi^*)$ for all $g \in F$ or else $T_g V(\pi^*) > V(\pi^*)$ for some $g \in F$. Thus if π^* is optimal we have from (c) that $T_g V(\pi^*) \leqq V(\pi^*)$ for all $g \in F$, implying $TV(\pi^*) \leqq V(\pi^*)$ and hence $T^n V(\pi^*) \leqq V(\pi^*)$ for all $n \geqq 1$. Letting $n \to \infty$ and using Corollary 10 then gives $V^* \leqq V(\pi^*)$. Thus $V(\pi^*) = V^*$ by (a).

THEOREM 12. *If* A_s *is finite for each* $s \in S$ *then there exists a stationary policy which is optimal.*

PROOF. If the action sets are finite then, noting that the supremum which defines the operator T can be taken component-wise, there exists an $f \in F$ such that $T_f V^* = TV^* = V^*$ and hence $T_f^n V^* = V^*$. Letting $n \to \infty$ gives $V(f^\infty) = V^*$, implying by Theorem 11 (d) that f^∞ is optimal.

6. Additional remarks. Theorem 11 (c) shows that the policy improvement routine of Howard [5] remains valid in this more general setting. Corollary 10 (b) shows that the method of successive approximations (i.e., repeated application of the optimal return operator T to an arbitrary initial function V) converges uniformly and at a geometric rate. The reader will note that

assumption (iii) is critical in obtaining this uniform convergence, since it allows us to metrize the space B with the usual uniform metric. This is the only sense, however, in which assumption (iii) is important to our results. If it is dropped a more complicated metric must be used, but one can still show that there exists a stationary ε-optimal policy for each $\varepsilon > 0$ and a stationary optimal policy in the case of finite action sets. See Harrison [4] for details in the finite action case. That same paper presents extensions to Markov renewal decision processes under assumptions similar to those used here.

REFERENCES

[1] BELLMAN, R. (1957). *Dynamic Programming*. Princeton Univ. Press.

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

[3] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

[4] HARRISON, J. M. (1970). Countable state discounted Markovian decision processes with unbounded rewards. Technical Report No. 128, Department of Operations Research, Stanford Univ.

[5] HOWARD, R. (1960). *Dynamic Programming and Markov Chains*. M.I.T. Press.

[6] LIUSTERNIK, L. A. and SOBOLEV V. J. (1961). *Elements of Functional Analysis*. Frederick Ungar Publishing Company, New York.

[7] MAITRA, A. (1965). Dynamic programming for countable state systems. *Sankhya Ser. A* **27** 241–248.