The Annals of Mathematical Statistics
1971, Vol. 42, No. 5, 1717–1726

MAXIMAL AVERAGE-REWARD POLICIES FOR SEMI-MARKOV DECISION PROCESSES WITH ARBITRARY STATE AND ACTION SPACE[1]

By Steven A. Lippman
University of California, Los Angeles

We consider the problem of maximizing the long-run average (also the long-run average expected) reward per unit time in a semi-Markov decision processes with arbitrary state and action space. Our main result states that we need only consider the set of stationary policies in that for each $\varepsilon > 0$ there is a stationary policy which is ε-optimal. This result is derived under the assumptions that (roughly) (i) expected rewards and expected transition times are uniformly bounded over all states and actions, and that (ii) there is a state such that the expected length of time until the system returns to this state is uniformly bounded over all policies. The existence of an optimal stationary policy is established under the additional assumption of countable state and finite action space. Applications to queueing reward systems are given.

1. Introduction. We consider the problem of maximizing the long-run average (also the long-run average expected) reward per unit time in a semi-Markov decision process with arbitrary state and action space. Our main result states that we need only consider the set of stationary policies in that for each $\varepsilon > 0$ there is a stationary policy which is ε-optimal. This result is derived under assumptions slightly stronger than (i) expected rewards and expected transition times are uniformly bounded over all states and actions, and (ii) there is a state 0 such that the mean recurrence time until the system returns to state 0 is uniformly bounded over all policies. The existence of an optimal stationary policy is established under the additional assumption of countable state and finite action space. Our method of proof utilizes the strong law of large numbers and a result in positive dynamic programming due to Blackwell (1967).

2. Model description and notation. A *semi-Markov decision process* (SMDP) is specified by five objects: a state space S, an action space A, a law of motion q, a transition time t, and a reward r. Whenever (and however) the system is in state s and you choose action a, three things happen: (i) the system moves to a new state selected according to the probability distribution $q(\cdot \mid s, a)$, (ii) conditional on the event that the new state is s', the length of time it takes the system to move to state s' is a nonnegative random variable with probability distribution $t(\cdot \mid s, a, s')$, and (iii) conditional on the event that the new state is s', immediately

Received October 21, 1970.

[1] This research was partially supported by the National Science Foundation through Grants GS 2041 and GP 26294.

after the transition is completed,[2] you receive a reward whose probability distribution is $r(\cdot \mid s, a, s')$.[3] To ensure the existence of a probability space, we assume that S and A are each Borel subsets of a locally compact, separable, metric space.

A *policy* π is a sequence $\pi_1, \pi_2, \cdots$, of decision rules where the nth decision rule π_n tells you how to select an action in A after completion of the $n-1$st transition. More precisely, π_n is a conditional probability on (the Borel subsets of) A given the history $h^n = (s_1, a_1, r_1, t_1, \cdots, s_{n-1}, a_{n-1}, r_{n-1}, t_{n-1}, s_n)$ of the system up to and including the time of the $n-1$st transition. So given that we have observed the history h^n up to the time of the $n-1$st transition, we choose our nth action according to the distribution $\pi_n(\cdot \mid h^n)$. A policy π is said to be *stationary* if there is a (Borel measurable) map f from S into A and if $f(s_n)$ is the action chosen by π when s_n is the state reached just after completion of the $n-1$st transition. Thus, a stationary policy π always chooses action $f(s)$ whenever and however it reaches state s.

We shall assume throughout this paper, unless stated otherwise, that the initial state is state 0. Given a policy π, we denote by r_n and t_n the nth reward received and the length of the nth transition, and we define the following associated random variables:

$$n(T) = \max\{n: t_1 + \cdots + t_n \leqq T\}, \qquad \text{each } T \geqq 0,$$
$$R_n = r_1 + \cdots + r_n,$$
$$T_n = t_1 + \cdots + t_n.$$

Finally, we define V_π, the long-run average reward per unit time associated with policy π, by

$$V_\pi = \limsup_{T\to\infty} \frac{R_{n(T)}}{T}. \tag{1}$$

Also, we define

$$V_\pi^* = \liminf_{T\to\infty} \frac{R_{n(T)}}{T}. \tag{2}$$

Our goal is to find a policy π^*, termed *optimal*, such that for each policy π we have

$$V_{\pi^*}^* \geqq V_\pi \text{ a.e.}$$

Given $\varepsilon > 0$, π^* is said to be *ε-optimal* if for each policy π we have

$$V_{\pi^*}^* \geqq V_\pi - \varepsilon \text{ a.e.}$$

3. Optimality of stationary policies. If we make no further restrictions on our SMDP, there may be no $\varepsilon > 0$ for which an ε-optimal policy exists, or the return of some policy might be infinite with probability 1 (see Ross (1971) and Examples 2 and 3 of Lippman (1970)).

---

[2] Our results still hold if the reward were received at the beginning of the transition (see proof of Lemma 3).

[3] The immediate reward and the transition time are not assumed to be independent random variables, not even if $S = \{0\}$ and $A = \{0\}$.

In order to ensure the existence of ε-optimal policies for all $\varepsilon > 0$ and to ensure that $P(V_\pi^* = -\infty) = P(V_\pi = +\infty) = 0$ for all π, we make the following assumptions:

ASSUMPTION 1. (all inf's and sup's are over $S \times A \times S$)

(i) $\inf \int \xi \, dr(\xi \mid s, a, s') \geqq -M$,

(ii) $\sup \int \xi \, dr(\xi \mid s, a, s') = M < \infty$,

(iii) $\inf \int \xi \, dt(\xi \mid s, a, s') = L > 0$,

(iv) $\sup \int \xi \, dt(\xi \mid s, a, s') = B < \infty$, and

(v) $\sup [\int \xi^2 \, dr(\xi \mid s, a, s') + \int \xi^2 \, dt(\xi \mid s, a, s')] < \infty$.

ASSUMPTION 2. Let $p(\pi, i)$ be the probability that policy π requires at least i transitions until the first return to state 0, let $N(\pi)$ be the number of transitions that policy π requires until the first return to state 0, and define

$$p_i^* = \sup_\pi \{p(\pi, i)\}.$$

Then

$$\mu_1 \equiv \sum_{i=1}^{\infty} p_i^* < \infty \quad \text{and} \quad \mu_2 \equiv \sup_\pi \operatorname{Var}(N(\pi)) < \infty.$$

Assumption 1 states that expectations and variances of rewards and transition times are uniformly bounded with the lower bound on the expected transition time being strictly positive. The first statement of Assumption 2 is slightly stronger than requiring that the expected number of transitions until state 0 is re-entered is uniformly bounded. Coupled with (iv) of Assumption 1, it implies that the expected length of time until state 0 is re-entered is uniformly bounded over all policies. Also, a uniform bound on the variance of the length of time until the first return to state 0 can be obtained from (v) and $\mu_2 < \infty$. In the context of a queueing reward system in which the decision maker can control the arrival rate and/or the service rate, these assumptions assert that the traffic intensity ρ is bounded away from 1 for all policies. It appears that these assumptions will be easy to verify in practice.

We now present our main result whose proof is given in the next section. This result is particularly useful in the study of queueing optimization problems since, with ingenuity in formulation, Assumption 2 will often be satisfied even when there is no recurrent state. (See Lippman (1970) for several applications of Theorem 1.)

THEOREM 1. *If Assumptions* 1 *and* 2 *hold, then for each* $\varepsilon > 0$ *there is a stationary policy which is* ε*-optimal. Furthermore, there is a stationary policy which is optimal if there is a best policy among the set of stationary policies.*[4]

[4] It can also be shown that if Assumptions 1 and 2 hold, if $q(\cdot \mid s, a)$ has countable support for each pair (s, a), and if there is a policy which maximizes the ratio of the expected reward earned until the first return to state 0 to the expected time until the first return to state 0, then there is a stationary policy which is optimal. The proof of this fact is an immediate consequence of Proposition B of Ornstein (1969) and our Equation (9).

Of course, if the process starts in some state s other than state 0 and if it must return to state 0 in a finite expected amount of time during which a finite expected reward is earned, then Theorem 1 remains true. Also, if we consider distributions rather than random variables and if we say that policy π^* is ε-optimal if the distribution of $V_{\pi^*}+\varepsilon$ is stochastically larger than that of V_π for each π, then we need only assume that S and A are each Borel subsets of a complete, separable, metric space and that the reward distributions are point masses and form a Baire function on $S \times A \times S$.

Next, we define $\bar{V}_\pi$, the long-run expected average reward per unit time associated with policy π, by

$$\bar{V}_\pi = \limsup_{T\to\infty} E\left(\frac{R_{n(T)}}{T}\right). \tag{1'}$$

Also, we define

$$\bar{V}_\pi^* = \liminf_{T\to\infty} E\left(\frac{R_{n(T)}}{T}\right). \tag{2'}$$

A policy π^* is said to be L^1 *optimal* if $\bar{V}_{\pi^*}^* \geqq \sup_\pi \bar{V}_\pi$. Similarly, π^* is said to be ε-L^1 *optimal* if $\bar{V}_{\pi^*}^* \geqq \sup_\pi \bar{V}_\pi - \varepsilon$.

THEOREM 2. *If Assumptions* 1 *and* 2 *hold and if the rewards received are uniformly bounded random variables, then for each* $\varepsilon > 0$ *there is a stationary policy which is* ε-L^1 *optimal. Furthermore, there is a stationary policy which is* L^1 *optimal if there is a best policy among the set of stationary policies.*

Recently, Ross has obtained useful sufficient conditions (see Theorem 3 of Ross (1970b)) which guarantees the optimality of stationary policies for a countable state, finite action SMDP. In the context of queueing reward systems, his conditions essentially require (Ross (1968) Theorem 1.4) that the mean recurrence time to go from state s to state 0 is uniformly bounded over all states and all discount optimal stationary policies. This condition will seldom be met in systems with an infinite queue capacity.

THEOREM 3. *Suppose* S *is countable and* A *is finite. Then under the hypotheses of Theorem* 1 [*Theorem* 2] *there is a stationary policy which is optimal* [L^1 *optimal*].

It is interesting to note that Theorem 3 does not hold if Assumption 2 is replaced by the weaker condition that the mean return times are uniformly bounded (see Fisher and Ross (1968)). Thus, in view of Fisher and Ross (1968), it would appear that our conditions are nearly minimal.

The proofs of Theorems 2 and 3 utilize Theorem 1 and are given in Section 5.

4. Proof of Theorem 1. In view of the length of the proof, we shall break it into several parts.

Part I. Let $\mathscr{X} = \cup_{\lambda\in\Lambda}\mathscr{X}_\lambda$ be a family of sequences of random variables indexed by λ with probability space $(\Omega_1, \mathscr{F}_1, P_1)$ such that (i) the random variables are mutually independent and (ii) for each fixed $\lambda \in \Lambda$, $\mathscr{X}_\lambda$ is a sequence of identically distributed random variables with mean μ_λ and variance $\sigma_\lambda{}^2$.

We will iteratively select a sequence from Λ as follows. Given the first j selections $\lambda_1, \cdots, \lambda_j$ from Λ and the observed values $y_1, \cdots, y_j$ of the associated random variables from the sequences $\mathscr{X}_{\lambda_1}, \cdots, \mathscr{X}_{\lambda_j}$, we refer to $h^j = (\lambda_1, y_1, \cdots, \lambda_j, y_j)$ as the history of the system up to the jth stage. Associated with each history h^j is a random variable D_{j+1} with probability space $(\Omega_2, \mathscr{F}_2, P_2)$ and values in Λ. Thus, given the history h^j up to the jth stage, the $j+1$st index in Λ is chosen according to the distribution of D_{j+1} and the next random variable in the sequence $\mathscr{X}_{\lambda_{j+1}}$ is selected. This random variable is labeled Y_{j+1}. In the following lemma, a.e. refers to the probability space $(\Omega, \mathscr{F}, P)$ where $\Omega = \Omega_1 \times \Omega_2$ and $\mathscr{F}$ and P are the product field and product measure.

LEMMA 1.

If $\sup_{\lambda \in \Lambda} \mu_\lambda = \bar{M} < \infty, \inf_{\lambda \in \Lambda} \mu_\lambda = \underline{M} \geqq -\infty$, *and* $\sup_{\lambda \in \Lambda} \sigma_\lambda^2 < \infty$, *then*

$$\limsup_{n \to \infty} \frac{Y_1 + \cdots + Y_n}{n} \leqq \bar{M} \text{ a.e.},$$

and

$$\liminf_{n \to \infty} \frac{Y_1 + \cdots + Y_n}{n} \geqq \underline{M} \text{ a.e.}$$

PROOF. If $X_{\lambda,j}$ is the jth member of $\mathscr{X}_\lambda$, define $\tilde{X}_{\lambda,j} = X_{\lambda,j} - \mu_\lambda$ so that $\tilde{Y}_i = X_{\lambda,k} - \mu_\lambda$ if $Y_i = X_{\lambda,k}$. Similarly, define $S_n = Y_1 + \cdots + Y_n$ and $\tilde{S}_n = \tilde{Y}_1 + \cdots + \tilde{Y}_n$. (Note that we may not have $\tilde{S}_n = S_n - E(S_n)$.) Now Var $(\tilde{Y}_i) \leqq \sup_{\lambda \in \Lambda} \sigma_\lambda^2$ for each i, and for $i < j$,

$$E(\tilde{Y}_i \tilde{Y}_j) = E\{E(\tilde{Y}_j \mid \tilde{Y}_i) \cdot \tilde{Y}_i\} = E\{0 \cdot \tilde{Y}_i\} = 0$$

so that $\langle \tilde{Y}_i \rangle$ are uncorrelated. Consequently we can conclude from the strong law of large numbers (see Chung (1968) page 97) that $\tilde{S}_n/n \to 0$ a.e.

Because $\tilde{S}_n = S_n - T_n$ where the random variable T_n satisfies $n\underline{M} \leqq T_n \leqq n\bar{M}$, we obtain

$$\limsup_{n \to \infty} \frac{S_n}{n} = \limsup_{n \to \infty} \left(\frac{\tilde{S}_n}{n} + \frac{T_n}{n} \right) \leqq \limsup_{n \to \infty} \frac{\tilde{S}_n}{n} + \limsup_{n \to \infty} \frac{T_n}{n}$$

$$\leqq \bar{M}.$$

Similarly,

$$\liminf_{n \to \infty} \frac{S_n}{n} \geqq \liminf_{n \to \infty} \frac{\tilde{S}_n}{n} + \liminf_{n \to \infty} \frac{T_n}{n} \geqq \underline{M}. \quad \square$$

Part II. On the path towards establishing our main result, we temporarily make the assumption that there is but one state, that is, $S = \{0\}$. With this assumption, we refer to our SMDP as a sequence of identical, time variable games (SITVG).[5] Here, we seek to maximize the average reward per unit time when we repeatedly play the same game, the duration of which depends on our action in each game.

[5] MacQueen (1962) investigated the more subtle situation wherein the games are not identical.

To facilitate the use of the SITVG in our treatment of the SMDP, we will adopt a slightly different notation. We have state space $\{0\}$, action space $\mathscr{A}$, and we denote by r_α and t_α the reward and the transition time associated with action $\alpha \in \mathscr{A}$. Rather than policies $\pi = \langle \pi_n \rangle$, we speak of strategies $\sigma = \langle \sigma_n \rangle$ where each σ_n is a distribution on $\mathscr{A}$.

In establishing our next lemma, we shall need the following assumption:

ASSUMPTION 3. The expectation and variance of both r_α and t_α are uniformly bounded above by $U < \infty$; the expectation of t_α is uniformly bounded below by $L > 0$; and $E(r_\alpha) \geqq -M$ for all $\alpha \in \mathscr{A}$.

Now define

$$L^* = \sup_{\alpha \in \mathscr{A}} \frac{E(r_\alpha)}{E(t_\alpha)}, \tag{3}$$

so $|L^*| < \infty$ by Assumption 3. It follows from the strong law of large numbers that the stationary strategy σ_α which always chooses action α has long-run average reward $V_{\sigma_\alpha} = E(r_\alpha)/E(t_\alpha)$ a.e. (Of course, $V_\sigma{}^* = V_\sigma$ a.e. if σ is stationary.) Moreover, we can choose a sequence $\langle \alpha_n \rangle$ from $\mathscr{A}$ with the property that $E(r_{\alpha_n})/E(t_{\alpha_n}) \uparrow L^*$ so the strategy σ^* which chooses the action α_n after the $n-1$st transition has $V_{\sigma^*} = L^*$ a.e. We now show that L^* is the optimal return so that there is an ε-optimal stationary strategy for each $\varepsilon > 0$ and there is an optimal stationary strategy if there is a best strategy among the set of stationary strategies.

LEMMA 2. *If Assumption* 3 *holds, then for every strategy* σ, *we have* $V_\sigma \leqq L^*$ a.e.

PROOF. Fix σ and let r_n and t_n be the nth reward received and the duration of the nth transition. Now $T_{n(T)}/T \to 1$ a.e., for $T_{n(T)+1}/T_{n(T)} \to 1$ a.e. implies that

$$\lim_{T\to\infty} \frac{T_{n(T)}}{T} = \lim_{T\to\infty} \frac{T_{n(T)}}{T} \frac{T_{n(T)+1}}{T_{n(T)}} \geqq 1 \text{ a.e.}$$

Verifying that $T_{n(T)+1}/T_{n(T)} \to 1$ a.e. is equivalent to verifying that $t_{n+1}/T_n \to 0$ a.e. which, in turn, follows from the Borel–Cantelli Lemma as shown in the proof of Lemma 3. Therefore,

$$V_\sigma = \limsup_{T\to\infty} \frac{R_{n(T)}}{T} = \limsup_{T\to\infty} \frac{R_{n(T)}}{T_{n(T)}} \cdot \frac{T_{n(T)}}{T} = \limsup_{T\to\infty} \frac{R_{n(T)}}{T_{n(T)}}$$

$$= \limsup_{n\to\infty} \frac{R_n}{T_n} \text{ a.e.}$$

If we do not have $V_\sigma \leqq L^*$ a.e., then there is an $\varepsilon > 0$ and a $\delta > 0$ such that

$$P\left(\limsup_{n\to\infty} \frac{R_n}{T_n} > \sup_{\alpha \in \mathscr{A}} \frac{E(r_\alpha)+\varepsilon}{E(t_\alpha)-\varepsilon} \right) \geqq \delta. \tag{4}$$

Now partition $\mathscr{A} = \bigcup_{k=1}^{K} P_k$ in such a way that if α is in the kth subset $k = 1, 2, \cdots, K < \infty$, then there is an α_k in this subset such that

$$E(r_\alpha) \leqq E(r_{\alpha_k}) + \tfrac{1}{2}\varepsilon \quad \text{and} \quad E(t_{\alpha_k}) - \tfrac{1}{2}\varepsilon \leqq E(t_\alpha).$$

Let K_k be the number of $i \leqq n$ such that the observed value of σ_i is in the kth subset of $\mathscr{A}$ (so that K_k is a random function of n), so

$$\frac{R_n}{T_n} = \frac{\sum_{k=1}^{K} \sum_{\alpha_i \in P_k} r_{\alpha_i}}{\sum_{k=1}^{K} \sum_{\alpha_i \in P_k} t_{\alpha_i}} = \frac{\sum_{k=1}^{K} K_k\left(1/K_k \sum_{\alpha_i \in P_k} r_{\alpha_i}\right)}{\sum_{k=1}^{K} K_k\left(1/K_k \sum_{\alpha_i \in P_k} t_{\alpha_i}\right)}.$$

Hence, we can conclude from Lemma 1 that

$$\text{(5)} \qquad \limsup_{n\to\infty} \frac{R_n}{T_n} \leqq \limsup_{n\to\infty} \frac{\sum_{k=1}^{K} K_k(E(r_{\alpha_k})+\varepsilon)}{\sum_{k=1}^{K} K_k(E(t_{\alpha_k})-\varepsilon)} \leqq \sup_{\alpha \in \mathscr{A}} \frac{E(r_\alpha)+\varepsilon}{E(t_\alpha)-\varepsilon} \text{ a.e.,}$$

since, by induction,

$$\frac{\sum_{k=1}^{K} K_k(E(r_{\alpha_k})+\varepsilon)}{\sum_{k=1}^{K} K_k(E(t_{\alpha_k})-\varepsilon)} \leqq \max\left\{\frac{E(r_{\alpha_k})+\varepsilon}{E(t_{\alpha_k})-\varepsilon} : k = 1, 2, \cdots, K\right\} \leqq \sup_{\alpha \in \mathscr{A}} \left\{\frac{E(r_\alpha)+\varepsilon}{E(t_\alpha)-\varepsilon}\right\}.$$

But (5) contradicts (4). □

LEMMA 3. *If our* SMDP *satisfies Assumptions* 1 *and* 2, *then there is a* SITVG *such that it satisfies Assumption* 3 *and there is a* $1-1$ *correspondence between policies* π *and strategies* σ *so that* $V_\pi = V_\sigma$ a.e. *if* π *corresponds to* σ.

PROOF. We will define $\mathscr{A}$ to be the set of actions such that each action α in $\mathscr{A}$ is a Borel measurable rule for deciding what actions in A to select as the system moves in time among the states in S until state 0 is finally reached; the action α can only make use of the history of the SMDP *since* state 0 was last entered. More precisely, let $\mathscr{A}_i$ be the set of probability distributions on A given histories h^i with $s_j \neq 0$ for $1 < j \leqq i$. We define $\mathscr{A}$ to be the set of sequences $\alpha = \langle \hat{\alpha}_i \rangle$ where $\hat{\alpha}_i \in \mathscr{A}_i$ for each i.

Now a strategy σ is a sequence $\langle \sigma_n \rangle$ where σ_n is a distribution on $\mathscr{A}$ (not on A) given the history of the SMDP up to and including the time of the $(n-1)$st transition into state 0. It is clear upon reflection that there is a $1-1$ correspondence between policies for the SMDP and strategies for the SITVG. Furthermore, defining r_α and t_α to be the total reward received and transition time until state 0 is re-entered when action α is selected, it is obvious that if π corresponds to σ, then $V_\pi \geqq V_\sigma$ a.e. To show $V_\pi \leqq V_\sigma$ a.e., let t_n be the length of time between the $n-1$st and nth entry into state 0, and let r_n be the associated reward received so that

$$V_\pi \leqq \limsup_{n\to\infty} \frac{R_n + \bar{r}_{n+1}}{T_n} \leqq \limsup_{n\to\infty} \frac{R_n}{T_n} + \limsup_{n\to\infty} \frac{\bar{r}_{n+1}}{T_n}$$

$$= V_\sigma + \limsup_{n\to\infty} \frac{\bar{r}_{n+1}}{T_n},$$

where $\bar{r}_n$ is the sum of the absolute values of the rewards earned rather than simply the sum of the rewards earned between the time of the $n-1$st and nth entry into state 0. Now by Lemma 1 and (iii) of Assumption 1,

$$\limsup_{n\to\infty} \frac{r_{n+1}/n}{T_n/n} \leqq \frac{1}{L} \limsup_{n\to\infty} \frac{\bar{r}_{n+1}}{n} \text{ a.e.}$$

Since we have a uniform bound on the second moment of the $\bar{r}_n$'s, Chebyshev's inequality together with the Borel–Cantelli Lemma implies that $\bar{r}_n/n \to 0$ a.e.

In verifying that Assumption 3 holds, it suffices to produce a uniform bound on $E(\tau(\pi)^2)$ where $\tau(\pi)$ is the length of time until the first return to state 0 using policy π. Let N be the number of transitions required by policy π to reach state 0, so

$$\begin{aligned} E(\tau(\pi)^2) &= E[(\textstyle\sum_{i=1}^{\infty} t_i 1_{\{N>i\}})^2] \\ &= \textstyle\sum_{i=1}^{\infty} E(t_i^2 \mid 1_{\{N\geqq i\}})P(N \geqq i) + 2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty} E(t_i t_j \mid 1_{\{N\geqq j\}})P(N \geqq j) \\ &\leqq Q\textstyle\sum_{i=1}^{\infty} P(N \geqq i) + 2B^2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty} P(N \geqq j) \\ &= QE(N) + B^2[E(N^2) - E(N)] \end{aligned}$$

where $Q = \sup \int \xi^2 dt(\xi \mid s, a, s')$. □

Part III. Let $\varepsilon > 0$ be given. Since Assumptions 1 and 2 are satisfied, we can use Lemma 3 to find the equivalent SITVG which satisfies Assumption 3. Therefore, Lemma 2 yields the existence of a stationary strategy which is ε-optimal. The long-run average return of the stationary strategy σ which always chooses action α is $E(r_\alpha)/E(t_\alpha)$. Thus, we seek an $\alpha^* \in \mathscr{A}$ such that

$$\frac{E(r_{\alpha^*})}{E(t_{\alpha^*})} + \varepsilon > L^* = \sup_{\alpha \in \mathscr{A}} \frac{E(r_\alpha)}{E(t_\alpha)}. \tag{6}$$

To do so, consider an associated (positive dynamic programming) problem which is such that the process ends whenever it returns to state 0, and $\rho(s, a, s')$, the reward associated with choosing action a from state s and going to state s', is defined by

$$\rho(s, a, s') = \int \xi \, dr(\xi \mid s, a, s') - L^* \int \xi \, dt(\xi \mid s, a, s').$$

A *plan* α is a sequence of decision rules where the nth decision rule tells you how to select an action on the nth day as a function of the previous history of the system. Starting from state 0, a plan α induces an expected nth period reward $\rho_n(\alpha)$ and an expected total reward

$$I(\alpha) \equiv \textstyle\sum_{n=1}^{\infty} \rho_n(\alpha). \tag{7}$$

If $I(\alpha^*) \geqq v - \varepsilon$, where $v \equiv \sup_{\alpha \in \mathscr{A}} I(\alpha)$, then we say that α is ε-optimal.

By a result due to Blackwell (1967) page 416, we need only consider stationary plans if ρ is a bounded nonnegative function and v is finite. That ρ is bounded follows from L^* finite together with (i) through (iv) of Assumption 1, while we can conclude that v is finite from the boundedness of ρ, Assumption 2, and the fact that the process ends when it returns to state 0. But it is not true that ρ is nonnegative. However, if we add $p_n^*(L^*B + M)$ to $\rho_n(\alpha)$, then we can still use Blackwell's result—as $\sum_{n=1}^{\infty} p_n^*(L^*B + M)$ is finite by Assumption 2—to conclude that there is a stationary plan α^* such that

$$I(\alpha^*) \geqq \sup_{\alpha \in \mathscr{A}} I(\alpha) - \varepsilon L. \tag{8}$$

This fact coupled with

$$I(\alpha) = E(r_\alpha) - L^* E(t_\alpha), \qquad \alpha \in \mathscr{A}, \tag{9}$$

and $\sup_{\alpha \in \mathscr{A}} E(t_\alpha) < \infty$ yields the desired result. □

5. Proof of Theorems 2 and 3. Fix π, and denote by r_n and t_n the nth reward received and the length of the nth transition. Also, set $n(T) = \max\{n : t_1 + \cdots + t_n \leqq T\}$. We claim that

$$\bar{V}_\pi \leqq E(V_\pi). \tag{10}$$

Since the rewards are uniformly bounded, say by M, $R_{n(T)}/T \leqq M(n(T)/T)$, and hence (using $V_\pi \geqq -M/L$)

$$\begin{aligned} E\left(\frac{R_{n(T)}}{T} - V_\pi\right) &= \int_{|(R_{n(T)}/T) - V_\pi| < \varepsilon} \left(\frac{R_{n(T)}}{T} - V_\pi\right) dP + \int_{(R_{n(T)}/T) - V_\pi \geqq \varepsilon} \left(\frac{R_{n(T)}}{T} - V_\pi\right) dP \\ &\quad + \int_{(R_{n(T)}/T) - V_\pi \leqq -\varepsilon} \left(\frac{R_{n(T)}}{T} - V_\pi\right) dP \\ &\leqq \varepsilon + M \int_{(R_{n(T)}/T) - V_\pi \geqq \varepsilon} \left(\frac{n(T)}{T} + \frac{1}{L}\right) dP \\ &\leqq \varepsilon + M \left[P\left(\frac{R_{n(T)}}{T} - V_\pi \geqq \varepsilon\right) \int \left(\frac{n(T)}{T} + \frac{1}{L}\right)^2 dP \right]^{\frac{1}{2}}. \end{aligned}$$

As $P((R_{n(T)}/T) - V_\pi \geqq \varepsilon) \to 0$ as $T \to \infty$, our claim is justified upon establishing that the second moment of $n(T)$ is $O(T^2)$.

From (iii) and (v) of Assumption 1, we can conclude that there is an $\varepsilon > 0$ and a $\delta > 0$ such that $P(t_\beta > \varepsilon) \geqq \delta$ for each $\beta \in S \times A \times S$ where t_β has distribution $t(\cdot \mid s, a, s')$ if $\beta = (s, a, s')$. Using this fact, a standard argument (see Lippman (1970) and Chung (1968), page 127 or Ross (1970a), page 88) shows that the second moment of $n(T)$ is $O(T^2)$. This justifies our claim.

Theorem 1 ensures that given $\varepsilon > 0$, there is a stationary policy π^* with

$$V_{\pi^*}^* \geqq V_\pi - \varepsilon \text{ a.e.} \tag{11}$$

Because the stationarity of π^* implies that $\bar{V}_{\pi^*}^* = V_{\pi^*}^*$ a.e., we can combine (10) and (11) to yield the desired result. □

To prove Theorem 3, we utilize (8) to assert that for each k there is a stationary plan f_k such that

$$I(f_k) > \sup_{\alpha \in \mathscr{A}} I(\alpha) - 1/k. \tag{12}$$

Without loss of generality, assume that $S = \{0, 1, 2, \cdots\}$ and let $\langle f_{0,k} \rangle$ be a subsequence of $\langle f_k \rangle$ such that the action chosen at state 0 is the same for each plan $f_{0,k}$; such a subsequence exists since A is finite. Similarly, we continue defining subsequences $\langle f_{j,k} \rangle$ of $\langle f_{j-1,k} \rangle$ such that the action chosen from state j is the same for each plan $f_{j,k}$.

Now consider the stationary plan f which selects the same action at state j (labeled $f(j)$) as does the plan $f_{j,j}$. We claim that f is optimal. To see this, let

$\varepsilon > 0$ be given and define $K \equiv M + BL^* < \infty$ and $I_N(\alpha) \equiv \sum_{n=1}^{N} \rho_n(\alpha)$. Assumptions 1 and 2 imply that

$$|\rho_n(\alpha)| \leqq Kp_n^*, \qquad \alpha \in \mathscr{A}, n = 1, 2, \cdots. \tag{13}$$

Hence, Σp_n^* finite together with (13) implies that the convergence of $I_N(\alpha)$ to $I(\alpha)$ is uniform on $\mathscr{A}$, so that there is an integer N satisfying

$$|I(\alpha) - I_n(\alpha)| < \tfrac{1}{4}\varepsilon \qquad \text{for all } \alpha \text{ and } n \geqq N. \tag{14}$$

In view of (12) and (14), it sufficies to exhibit a plan $f_{k,k}$ with $I_N(f) \geqq I_N(f_{k,k}) - \frac{1}{4}\varepsilon$ and $k > 4/\varepsilon$; for given α, we have

$$\begin{aligned} I(f) - I(\alpha) &= I(f) - I(f_{k,k}) + I(f_{k,k}) - I(\alpha) \\ &\geqq I(f) - I(f_{k,k}) - \tfrac{1}{4}\varepsilon \geqq I_N(f) - I_N(f_{k,k}) - 3\varepsilon/4. \end{aligned}$$

We now exhibit a plan $f_{k,k}$ with $I_N(f) \geqq I_N(f_{k,k}) - \frac{1}{4}\varepsilon$. To start, pick $\delta > 0$ so that $2KN\delta < \frac{1}{4}\varepsilon$, and let K_N be sufficiently large so that the probability that the set $\{0, 1, 2, \cdots, K_N\}$ of states is left by time N is less than δ. It now follows immediately from (13) that

$$I_N(f_{K_N,K_N}) - I_N(f) < (2KN)\delta < \tfrac{1}{4}\varepsilon. \qquad \square$$

REFERENCES

BLACKWELL, DAVID (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

BLACKWELL, DAVID (1967). Positive dynamic programming. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 415–418. Univ. of California Press.

BREIMAN, LEO (1964). Stopping-rule problems. *Applied Combinatorial Mathematics*, ed. E. F. Bechenbach. Wiley, New York.

CHUNG, KAI LAI (1968). *A Course in Probability Theory*. Harcourt, Brace and World, New York.

DERMAN, CYRUS (1966). Denumerable state Markovian decision processes—average cost criterion. *Ann. Math. Statist.* **37** 1545–1554.

DERMAN, CYRUS and VEINOTT, Jr., A. F. (1967). A solution to a countable system of equations arising in Markovian decision processes. *Ann. Math. Statist.* **38** 582–584.

FISHER, L. and ROSS, S. (1968). An example in denumerable decision processes. *Ann. Math. Statist.* **39** 674–676.

LIPPMAN, STEVEN A. (1970). Maximal average-reward policies for a class of semi-Markov decision processes with arbitrary state and action space. Working Paper No. 162, Western Management Science Institute, Univ. of California, Los Angeles.

MACQUEEN, JAMES B. (1962). Sequences of independent time variable games. Paper presented at Western Regional Meetings of the Institute of Mathematical Statistics, April 20, Albuquerque, New Mexico.

ORNSTEIN, DONALD (1969). On the existence of stationary optimal strategies. *Proc. Amer. Math. Soc.* **20** 563–569.

ROSS, SHELDON M. (1968). Now-discounted denumerable Markovian decision models. *Ann. Math. Statist.* **39** 412–423.

ROSS, SHELDON M. (1970a). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.

ROSS, SHELDON M. (1970b). Average cost semi-Markov decision processes. *J. Appl. Probability* **7** 649–656.

ROSS, SHELDON M. (1971). On the nonexistence of ε-optimal randomized stationary policies in average cost Markov decision models. *Ann. Math. Statist.* **42** 1767–1768.