DENUMERABLE STATE MARKOVIAN DECISION PROCESSES--AVERAGE COST CRITERION

by

Cyrus Derman

TECHNICAL REPORT NO. 86

February 28, 1966

Supported by the Army, Navy, Air Force, and NASA under

Contract Nonr-225(53) (NR-042-002)

with the Office of Naval Research

Gerald J. Lieberman, Project Director

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

DENUMERABLE STATE MARKOVIAN DECISION PROCESSES--AVERAGE COST CRITERION

by

Cyrus Derman

1. Introduction

We are concerned with the optimal control of certain types of dynamic systems. We assume such a system is observed periodically at times $t = 0, 1, 2, \ldots$. After each observation the system is classified into one of a possible number of states. Let I denote the space of possible states. We assume I to be denumerable. After each classification one of a possible number of decisions is made. Let K_i denote the number of possible decisions when the system is in state i, $i \in I$. The decisions interact with the chance environment in the evolution of the system.

Let $\{Y_t\}$ and $\{\Delta_t\}$, $t = 0, 1, \ldots$, denote the sequences of states and decisions. A basic assumption concerning the type of systems under consideration is that

$$P\{Y_{t+1} = j | Y_o, \Delta_o, \ldots, Y_t = i, \Delta_t = k\} = q_{ij}(k) ,$$

for every i, j, k and t; i.e., the transition probabilities from one state to another are functions only of the last observed state and the subsequently made decision. It is assumed that the $q_{ij}(k)$'s are known.

A rule or policy R for controlling the system is a set of functions $\{D_k(Y_o, \Delta_o, \ldots, Y_t)\}$ satisfying

$$0 \leq D_k(Y_o, \Delta_o, \ldots, Y_t) \leq 1 ,$$

for every k, and

$$\sum_{k=1}^{K_i} D_k(Y_o, \Delta_o, \ldots, Y_t = i) = 1 ,$$

for every history $Y_o, \Delta_o, \ldots, Y_t$ $(t = 0, 1, \ldots)$. As part of a controlling rule, $D_k(Y_o, \Delta_o, \ldots, Y_t)$ is the instruction at time t to make decision k with probability $D_k(Y_o, \Delta_o, \ldots, Y_t)$ if the particular history $Y_o, \Delta_o, \ldots, Y_t$ has occurred. We remark that although we have assumed a kind of Markovian property regarding the behavior of the system, the process $\{Y_t\}$, or even the joint process $\{Y_t, \Delta_t\}$, is not necessarily a Markov process; for a rule may or may not depend upon the complete history of the system.

We further assume that there is a known cost (or expected cost) w_{ik} incurred each time the system is in state i and decision k is made. Thus, we can define a sequence of random variables $\{W_t\}$, $t = 0, 1, 2, \ldots$ by $W_t = w_{ik}$ if $Y_t = i$, $\Delta_t = k$, $t = 0, 1, \ldots$. For a given $Y_o = i$ and rule R we can talk about $E_R W_t$, provided it exists. Let

$$Q_{T,R}(i) = \frac{1}{T+1} \sum_{t=0}^{T} E_R W_t , \text{ when } Y_o = i ;$$

thus, $Q_{T,R}(i)$ is the expected average cost per unit time up to time period T. Let $Q_R(i) = \lim_{T\to\infty} Q_{T,R}(i)$, if the limit exists; otherwise, let $Q_R(i) = \limsup_{T\to\infty} Q_{T,R}(i)$.

In this paper we are concerned with the problem of finding an optimal rule R; explicitly, a rule R, for a given i, which minimizes $Q_R(i)$ over <u>all</u> possible rules.

It is convenient to consider sub-classes of the class of all possible rules. Let C denote the entire class of rules. Let C' denote the sub-class of stationary Markovian rules; i.e., a rule R is a member of C' if $D_k(Y_o, \Delta_o, \ldots, Y_t = i) = D_{ik}$, independent of $Y_o, \Delta_o, \ldots, \Delta_{t-1}$ and t. A rule $R \in C'$ is completely defined by the set of numbers $\{D_{ik}\}$, $k = 1, \ldots, K_i$, $i \in I$; i.e., a fixed randomized decision-making procedure is associated with each state. Let C'' denote the sub-class of C' for which $D_{ik} = 0$ or 1. The rules in C'' are stationary Markovian, but non-randomized.

We point out that if $R \in C'$, the resulting stochastic process $\{Y_t\}$, $t = 0, 1, \ldots,$ is a Markov chain with transition probabilities

$$p_{ij} = \sum_{k=1}^{K_i} D_{ik} q_{ij}(k) \ , \ (i, j \in I) \ .$$

If the state space I is finite it is known (see Gillette [8] and Derman [5]) that $Q_R(i)$ can be minimized over C by a rule $R \in C''$. Computing methods using dynamic programming (Blackwell [1], Howard [9]) or linear programming (Manne [12]) exist for obtaining solutions.

For I infinite, and specifically denumerable, <u>little has been</u> published regarding existence and the nature of optimal rules. Iglehart

[10] and Taylor [14] have considered the average cost criterion for the special cases of inventory and replacement systems allowing for an infinite state space. Blackwell [2], [3], Derman [6], Maitra [11], Strauch [13] have considered infinite state spaces in dealing with a discounted cost criterion (Blackwell and Strauch also consider a total expected cost criterion).

Of some related interest is the result (Blackwell [3] and Derman [6]) that for a discounted cost criterion (discount factor strictly less than one) and $K_i < \infty$, $i \in I$, and $\{w_{ik}\}$ bounded, an optimal rule always exists and is a member of C''. If either condition is violated, an optimal rule may not exist. A specific question then arises: Under the same conditions, does an optimal rule always exist for the average cost criterion, and, if it does, is there always an optimal rule in C''? In section 2 we present counterexamples showing that this is not the case. One example shows that no optimal solution exists; another, that an optimal solution exists but is not a member of C''--it is a member of $C' - C$. In the remaining sections we are concerned with obtaining conditions under which a rule in C'' is optimal and for the convergence of an infinite state version of the policy improvement (Howard [9]) computational procedure to the optimal rule.

2. Counterexamples

The first example, due to Maitra [11], shows that under the assumptions

(A) $K_i < \infty$, $i \in I$

and

(B) $\{w_{ik}\}$ is a bounded set of numbers,

an optimal rule need not exist.

Let I consist of the states $0, 0', 1, 1', \ldots$. Suppose $K_i = 2$, $i = 0, 1, 2, \ldots$ and $K_i = 1$, $i = 0', 1', 2', \ldots$ where $q_{i,i+1}(1) = 1$, $q_{i,i'}(2) = 1$, and $q_{i'i'}(1) = 1$ for $i = 0, 1, \ldots$. Assume $w_{ik} = 1$ for $i = 0, 1, \ldots$ and $k = 1, 2$; $w_{i'1} = w_i$ for $i = 0, 1, \ldots$ where $\{w_i\}$ is a decreasing sequence of positive real numbers converging to zero. In words, the system, when in state i, either proceeds to state $i+1$ or i' depending on the decision made; the cost is one unit. When the system is in state i', it remains there at a cost of w_i units per time period.

Assume $Y_o = 0$. Without entering into the details it is clear that we can choose an R such that $Q_R(0)$ is as close to zero as desired. However, any rule R for which there is <u>some</u> positive probability that decision 2 will be made at some state i yields a positive expected average cost. On the other hand, the rule R prescribing decision 1 at all states has $Q_R(0) = 1$. Thus, no rule can achieve a zero expected average cost and, consequently, no optimal rule exists.

The second counterexample shows that, even under conditions (A) and (B), an optimal rule need not be a member of C". By resorting to a randomized stationary Markovian rule one can do better than remaining in the class of deterministic stationary Markovian rules.

Let I be the state space consisting of the non-negative integers. Suppose $K_o = 1$, $K_i = 2$, $i = 1, 2, \ldots$, with $q_{oo}(1) = 0$, $q_{oi}(1) = g_i > 0$, $i = 1, 2, \ldots$; $q_{ii}(1) = 1$, $q_{io}(2) = 1$, $i = 1, 2, \ldots$.

Let $w_{ik} = w_i$, $i = 0, 1, \ldots$ where $\{w_i\}$ is a decreasing of positive real numbers converging to zero. Thus, the system, when in state 0, progresses to state i with probability $g_i > 0$; when in state $i \neq 0$, it either remains in state i (if decision 1 is made) or it reverts to state 0 (if decision 2 is made). The further the system is away from state 0 (i.e., the larger the value of i) the less the cost.

Assume $Y_o = 0$. Let R be any rule in C''; let S_R be the set of states for which $D_{i1} = 1$. If $i \in S_R$, then $Y_t = i$ implies $Y_{t'} = i$ for all $t' > t$; if $i \notin S_R$, then $Y_t = i$ implies $Y_{t+1} = 0$. Suppose S_R is non-empty; then it can be shown that

$$Q_R(0) = \left\{ \sum_{i \in S_R} g_i w_i \Big/ \sum_{i \in S_R} g_i \right\} > 0 .$$

If S_R is empty, then

$$Q_R(0) = \frac{w_o + \sum_{i=1}^{\infty} g_i w_i}{2} > 0 .$$

In either case $Q_R(0) > 0$. Thus, for every $R \in C''$, $Q_R(0) > 0$. Let $R \in C'$ be such that $0 < D_{i2} < 1$, $i \in I$, and $\sum_{i=1}^{\infty} g_i / D_{i2} = \infty$. State 0 is a recurrent state of the resulting Markov chain $\{Y_t\}$ since $P\{Y_t = 0 \text{ for some } t > 0 | Y_o = 0\}$ is equal to one. However, the mean recurrence time of state 0 is $1 + \sum_{i=1}^{\infty} g_i / D_{i2} = \infty$; hence, 0 is a null recurrent state. From Markov chain theory (see Chung [4]) it follows that all states are null recurrent states. Then, for any state i_o,

$$Q_R(0) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{i=0}^{\infty} \sum_{t=1}^{T} w_i P\{Y_t = i \mid Y_o = 0\}$$

$$= \lim_{T\to\infty} \frac{1}{T+1} \left\{ \sum_{i=0}^{i_o} \sum_{t=1}^{T} w_i P\{Y_t = i \mid Y_o = 0\} + \sum_{i=i_o+1}^{\infty} \sum w_i P\{Y_t = i \mid Y_o = 0\} \right\}$$

$$\leq w_o \sum_{i=0}^{i_o} \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P\{Y_t = i \mid Y_o = 0\}$$

$$+ w_{i_o} \lim_{T\to\infty} \sum_{i=i_o+1}^{\infty} \frac{1}{T+1} \sum_{t=0}^{T} P\{Y_t = i \mid Y_o = 0\}$$

$$= w_{i_o} \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P\{Y_t > i_o \mid Y_o = 0\}$$

$$= w_{i_o} ,$$

since i, being null recurrent implies $\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P\{Y_t = i \mid Y_o = 0\} = 0$. However, i_o is arbitrary and $\{w_i\}$ decreases to zero; hence, $Q_R(0) = 0$.

The question as to whether, under assumptions (A) and (B), there may exist a rule $R^* \in C - C'$ such that $Q_{R^*}(i) \leq Q_R(i)$ for all $R \in C'$ remains to be answered.

3. Sufficient Conditions

In this section we arrive at sufficient conditions for the existence of an optimal rule and for it to be a member of C''. Our conditions are motivated by the policy improvement procedure and our proof follows that of Iglehart [10]. An alternative proof of the same (slightly stronger)

result appears, as well as an application of the results of this paper, in Derman and Lieberman [7]. The conditions are summarized in

Theorem 1. *If Conditions (A) and (B) hold and if there exists a bounded set of numbers* $\{g, v_j\}$, $j \in I$, *satisfying*

$$(1) \qquad g + v_i = \min_k \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) v_j \right\}, \quad i \in I,$$

then there exists an $R^* \in C''$ *such that for any* i *and every* $R \in C$

$$g = Q_{R^*}(i) \leq Q_R(i) .$$

R^* *is the rule which, for each* i, *prescribes the decision that minimizes the right side of* (1).

Proof: Let k_i, $i \in I$, denote the decision that minimizes the right side of (1) (or, if there are several minimizing decisions, let k_i be any one of them). Let R^* denote the rule which prescribes decision k_i when in state i, $i \in I$. Let $p_{ij} = q_{ij}(k_i)$ for every i, $j \in I$. Then (1) becomes

$$(2) \qquad g + v_i = w_{ik_i} + \sum_{j \in I} p_{ij} v_j , \quad i \in I .$$

On multiplying (2) by $p_{i'i}^{(t)}$, the t-step transition probability from i' to i calculated from $\{p_{ij}\}$, and summing over i, we get

$$g + \sum_{i\in I} p^{(t)}_{i'i} v_i = \sum_{i\in I} p^{(t)}_{i'i} w_{ik} + \sum_{i\in I} p^{(t)}_{i'i} \sum_{j\in I} p_{ik} v_j$$

$$= \sum_{i\in I} p^{(t)}_{i'i} w_{ik} + \sum_{j\in I} p^{(t+1)}_{i'j} v_j \ , \ i' \in I \ .$$

The latter equality involves an interchange of the order of summation justified by virtue of the assumption that the sequence $\{v_j\}$ is bounded. On averaging over t in (3) and canceling in the limit, we get

$$g = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i\in I} p^{(t)}_{i'i} w_{ik_i} \tag{4}$$

$$= Q_{R^*}(i') \ , \quad i' \in I \ .$$

Thus, g is the expected average cost per unit time under R^*. We now show that R^* is optimal. Let $g_n(i)$, $n = 0, 1, \ldots$ satisfy

$$g_o(i) = \min_k w_{ik} \ , \quad i \in I \tag{5}$$

$$g_{n+1}(i) = \min_k \left\{ w_{ik} + \sum_{j\in I} q_{ij}(k) g_n(j) \right\} \ , \quad i \in I \ ;$$

that is, $g_n(i)$ denotes the total expected cost incurred over the periods $0, 1, \ldots, n$ operating optimally. Because of assumption (A), $g_n(i)$ is well defined. We shall show that there exists an M satisfying

$$ng + v_i - M \le g_n(i) \le ng + v_i + M \ , \quad i \in I \tag{6}$$

for $n = 0, 1, 2, \ldots$. For $n = 0$ and 1 (6) holds since $\{v_i\}$ and $\{w_{ik}\}$ are bounded sequences. Assume (6) holds for $n \leq N$. Then by (5), (6), and (1) we have

$$g_{N+1}(i) \leq \min_k \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) \left(Ng + v_j + M \right) \right\}$$

$$= \min_k \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) v_j \right\} + Ng + M$$

$$= (N + 1)g + v_i + M , \quad i \in I ,$$

the right inequality of (6). The left follows in the same way. Thus (6) holds.

Let R be any $R \in C$ and let $h_n(i)$ be the total expected cost incurred over the periods $0, 1, \ldots,$ under R. Since $g_n(i)$ is the result of an optimal rule for those periods, we have, using (6), that

$$\lim_{n \to \infty} \inf \frac{h_n(i)}{n} \geq \lim_{n \to \infty} \frac{g_n(i)}{n}$$

$$= g , \quad i \in I .$$

This proves the theorem since $Q_R(i) \geq \lim_{n \to \infty} \inf \frac{h_n(i)}{n}$.

We point out the following

Corollary: Under the conditions of theorem 1, $|g_n(i) - ng| \leq 2M$ for every n.

4. Improvement and Convergence

This section is devoted to seeking conditions under which a policy improvement procedure can be effectively used. A condition that we shall need to assume is

(C) For every $R \in C''$ the resulting Markov chain is positive recurrent; i.e., all states belong to one communicating class and are positive recurrent states (see Chung [4]).

Let R (make decision k_i at state i) be any rule in C''. Suppose

(D) There exists a bounded set of numbers $\{g, v_j\}$, $j \in I$, satisfying (2).

Let R' (make decision k_i' at state i) be defined as follows: Set $k_i' = k_i$ for each i such that

$$w_{ik_i} + \sum_{j \in I} q_{ij}(k_i) v_j = \min_k \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k_i) v_j \right\} \tag{7}$$

holds. Assume the set of states such that (7) does not hold is non-empty (otherwise the conditions of theorem 1 would be satisfied). For at least one state i not satisfying (7) let k_i' be such that

$$w_{ik_i'} + \sum_{j \in I} q_{ij}(k_i') v_j < w_{ik_i} + \sum_{j \in I} q_{ij}(k) v_j \, . \tag{8}$$

Denote by I' the set of states where (7) does not hold and for which k_i' is chosen to satisfy (8). For all states $i \notin I'$, let $k_i' = k_i$. (Here, we allow that $k_i' = k_i$ even through (7) does not hold. Later we shall not allow this.) We can assert

Lemma 1. If (A), (B), (C) and (D) hold, then for any initial state i,

$$Q_{R'}(i) < Q_R(i) .$$

Proof: Let $p_{ij} = q_{ij}(k'_i)$ $(i, j \in I)$. Let ϵ_i, $i \in I$, be the difference between the right side and left side of (8); thus, $\epsilon_i > 0$ if $i \in I'$ and $\epsilon_i = 0$ if $i \notin I'$. For any $1 \in I$ and t we get, using (2), that

$$\sum_{i\in I} p_{1i}^{(t)}\epsilon_i = \sum_{i\in I} p_{1i}^{(t)} \left\{g + v_i - (w_{ik_{i'}} + \sum_{j\in I} p_{ij}v_j)\right\}$$

$$= g + \sum_{i\in I} p_{1i}^{(t)}v_i - \sum p_{1i}^{(t)}w_{ik_{i'}} - \sum_{j\in I} p_{1j}^{(t+1)}v_j .$$

On averaging over $t = 0, \ldots, T$ and letting $T \to \infty$ we get (since the ϵ_i's are also bounded)

$$(9) \quad \sum_{i\in I} \epsilon_i \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} p_{1i}^{(t)} = g - \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{i\in I} p_{1i}^{(t)} w_{ik_{i'}}$$

$$= g - Q_{R'}(1) .$$

However, under assumption (C), $\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} p_{1i}^{(t)} > 0$ for every $i \in I$. Therefore, the left side of (9) is strictly positive since at least one ϵ_i is positive. Thus

$$Q_{R'}(1) < g = Q_R(1) , \quad 1 \in I ,$$

and the lemma is proved.

We remark that the amount of improvement obtained in changing from R to R' is precisely $\sum_{i \in I} \pi_i \epsilon_i$ where $\{\pi_i\}$, $i \in I$, are the steady state probabilities of the Markov chain with transition probabilities $\{p_{ij}\}$.

We have directly

Theorem 2: Under the conditions of lemma 1, if $R \in C''$ is optimal over C'', then it is optimal over C.

Proof: If R is optimal over C'', then I' must be empty by lemma 1. Therefore (1) holds and theorem 1 applies.

We shall make use of a further condition.

(E) For every $R \in C''$ there exists a set of real numbers $\{g^R, v_j^R\}$, $j \in I$ satisfying condition (D). The numbers $\{g^R, v_j^R\}$, are bounded uniformly over $j \in I$, $R \in C''$.

We then have the following existence

Theorem 3: Suppose (A), (B), (C), (D) and (E) hold, then there exists a rule $R^* \in C''$ which is optimal over C.

Proof: For any $R \in C''$ let w_{iR} and $q_{ij}(R)$ denote the values w_{ik} and $q_{ij}(k)$ under R for each $i \in I$. Then with this notation (2) becomes

$$g^R + v_i^R = w_{iR} + \sum_{j \in I} q_{ij}(R) v_j^R , \quad i \in I . \tag{10}$$

Let g^* be the greatest lower bound of all g^R, $R \in C''$. Let $\{R_n\}$, $n = 1, \ldots$ be a sequence of rules in C'' such that $\lim_{n \to \infty} g^{R_n} = g^*$.

Because of the uniform boundedness condition on $\{v_j^R\}$ and because C" is compact (Tychonov's theorem) there exists a convergent subsequence $\{R_{n_\nu}\}$, $\nu = 1, 2, \ldots$ such that $\lim_{\nu\to\infty} v_j^{R_{n_\nu}} = v_j^*$, $j \in I$, where $\{v_j^*\}$ is a bounded sequence. Let $R^* = \lim_{\nu\to\infty} R_{n_\nu}$ (Note: Since $k_i < \infty$, $\{R_{n_\nu}\}$ converges to R^* means that $q_{ij}(R_{n_\nu}) = q_{ij}(R^*)$ for sufficiently large ν). On letting $\nu \to \infty$, from (10) we get

$$(11) \qquad g^* + v_i^* = \lim_{\nu\to\infty}\left\{g^{R_{n_\nu}} + v_i^{R_{n_\nu}}\right\}$$

$$= \lim_{\nu\to\infty}\left\{w_{iR_{n_\nu}} + \sum_{j\in I} q_{ij}(R_{n_\nu})v_j^{R_{n_\nu}}\right\}$$

$$= w_{iR^*} + \sum_{j\in I} q_{ij}(R^*)v_j^*\,, \quad i \in I\,.$$

(The fact that $\lim_{\nu\to\infty} \sum_{j\in I} q_{ij}(R_{n_\nu})v_j^{R_{n_\nu}} = \sum_{j\in I} q_{ij}(R^*)v_j^*$ is easily shown.) Thus $\{g^*, v_j^*\}$, $j \in I$ is a bounded set of numbers satisfying (2) (or (10)) for $R = R^* \in C"$. That is, $g^* = g^{R^*}$, $v_j^* = v_j^{R^*}$, $j \in I$. Now suppose (1) does not hold when R^* is the rule. Then from lemma 1 an improvement is possible, contradicting the fact that g^* is the greatest lower bound of all g^R, $R \in C"$. Thus (1) must hold and by theorem 1, R^* is optimal over C.

Since the policy improvement procedure [9] involves solutions to (1) and (2) and converges to an optimal rule in the finite state case, it is of interest to provide a procedure and conditions for convergence in the denumerable state case. Let R (make decision k_i at state i, $i \in I$) be any rule in C". We define an iteration of the <u>policy</u>

improvement procedure for denumerable states as the transformation from R to R' where the decisions $\{k_i'\}$ of R' are decisions for which $\left\{w_{ik} + \sum_{j\in I} q_{ij}(k)v_j^R\right\}$ are minimized. The term "improvement" is justified by lemma 1. Note, that in our definition we now insist upon all possible improvements to be made in each iteration. The policy improvement procedure is a sequence of policy improvement iterations starting from any initial rule $R \in C''$. Before stating conditions under which a sequence of policy improvement iterations converges to an optimal rule we prove another lemma.

Let $\pi_{ij}(R) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P\{Y_t = j | Y_0 = i\}$, $i, j \in I$, for each $R \in C''$. We shall utilize the following condition:

(F) For every $j \in I$, $\inf_{R\in C'', i\in I} \pi_{ij}(r) > 0$.

For any $R \in C''$, let

$$\epsilon_i^R = \left(w_{ik_i} + \sum_{j\in I} q_{ij}(k_i)v_j^R\right) - \left(w_{ik_i'} + \sum_{j\in I} q_{ij}(k_i')v_j^R\right)$$

$$= g^R + v_i^R - \left(w_{ik_i'} + \sum q_{ij}(k_i')v_j^R\right), \quad i \in I,$$

where $\{k_i\}$, $i \in I$, are the decisions of R and $\{k_i'\}$, $i \in I$, are the decisions obtained from R by a policy improvement iteration.

Lemma 2: Assume conditions (A), (B), (C), (D), (E), and (F). Let $R_1 = R \in C''$ be arbitrary, and $\{R_n\}$ be a sequence of policy improvement iterations; then, for each $i \in I$, $\lim_{n\to\infty} \epsilon_i^{R_n} = 0$.

Proof: Under assumption (C), for each $n = 1, 2, \ldots$ $\pi_{ij}(R_n) = \pi_i(R_n)$, the steady-state probability of state i under rule R_n. From the remark following lemma 1 we can write, for each n,

$$g^{R_n} - g^{R_{n+1}} = \sum_{i \in I} \pi_i(R_{n+1}) \epsilon_i^{R_n} .$$

Since the left side tends to zero as $n \to \infty$ ($\lim_{n \to \infty} g^{R_n}$ exists since $\{g^{R_n}\}$ is a decreasing sequence), so must the right side. However, since $\epsilon_i^{R_n} \geq 0$, it follows from condition (F) that $\lim \epsilon_i^{R_n} = 0$, $i \in I$.

We can now state

Theorem 4: If conditions (A), (B), (C), (D), (E) and (F) hold, then given any $R_1 = R \in C''$, the policy improvement procedure converges to a rule $R^* \in C$ which is optimal over C.

Proof: Let $\{R_n\}$ be any sequence of rules obtained under the policy improvement procedure with $R_1 \in C''$ arbitrary. From compactness considerations it is possible to choose a subsequence of rules $\{R_{n_\nu}\}$, $\nu = 1, 2, \ldots$ such that $\lim_{\nu \to \infty} g^{R_{n_\nu}} = g^*$, $\lim_{\nu \to \infty} v_i^{R_{n_\nu}} = v_i^*$ $(i \in I)$, $\lim_{\nu \to \infty} \epsilon_i^{R_{n_\nu}} = 0$ $(i \in I)$, and $\lim_{\nu \to \infty} R_{n_\nu} = R^*$. For any R_{n_ν}, equation (10) holds. On letting $\nu \to \infty$ we get

$$g^* + v_i^* = w_{iR^*} + \lim_{\nu \to \infty} \sum_{j \in I} q_{ij}(R_{n_\nu}) v_j^{R_{n_\nu}}, \quad i \in I . \tag{12}$$

For a given i, for ν large enough, $q_{ij}(R_{n_\nu}) = q_{ij}(R^*)$; thus, from (12) we get that $g^* = g^{R^*}$ and $v_i^* = v_i^{R^*}$, $i \in I$. Clearly,

$$(13) \qquad g^* + v_i^* \geq \lim_{\nu \to \infty} \sup \min_k \left\{ w_{ik} + \sum q_{ij}(k) v_j^{R_{n_\nu}} \right\}, \quad i \in I .$$

However, by definition of ϵ_i^R, we have, for each ν,

$$(14) \qquad g^{R_{n_\nu}} + v_i^{R_{n_\nu}} \leq \min_k \left\{ w_{ik} + \sum q_{ij}(k) v_j^{R_{n_\nu}} \right\} + \epsilon_i^{R_{n_\nu}}, \quad i \in I .$$

Therefore, from (13) and (14), it follows, using lemma 2, that

$$(15) \qquad g^* + v_i^* = \lim_{\nu \to \infty} \min_k \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) v_j^{R_{n_\nu}} \right\}, \quad i \in I .$$

However, for each $i \in I$ and k

$$\lim_{\nu \to \infty} \min \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) v_j^{R_{n_\nu}} \right\}$$

$$\leq \lim_{\nu \to \infty} \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) v_j^{R_{n_\nu}} \right\},$$

so that from (15)

$$g^* + v_i^* \leq \min_k \lim_{\nu \to \infty} \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) v_j^{R_{n_\nu}} \right\}$$

$$= \min_k \left\{ w_{ik} + \sum_{j \in I} q_{ij}(k) v_j^* \right\} .$$

But, for k chosen in accordance with rule R*, equality holds; hence, (1) must hold and theorem 1 applies. This proves the theorem.

REFERENCES

[1] Blackwell, David (1962). Discrete dynamic programming. Ann. Math. Statist. 33, 719-726.

[2] Blackwell, David (1964). Positive bounded dynamic programming. (Mimeographed)

[3] Blackwell, David (1965). Discounted dynamic programming. Ann. Math. Statist. 36, 226-235.

[4] Chung, Kai Lai (1960). Markov chains with stationary transition probabilities, Springer, Berlin.

[5] Derman, Cyrus (1962). On sequential decisions and Markov chains. Management Sci. 9, 16-24.

[6] Derman, Cyrus (1965). Markovian Sequential Control Processes--Denumerable State Space. J. Math. Anal. Appl., 10, 295-302.

[7] Derman, Cyrus and Lieberman, Gerald J. On Machine Setting and Maintenance Problems (in preparation)

[8] Gillette, Dean (1957). Stochastic games with zero stop probabilities. Ann. Math. Studies, 39, Vol III, 179-187.

[9] Howard, Ronald (1960). Dynamic programming and Markov processes. John Wiley, New York.

[10] Iglehart, Donald (1963). Dynamic programming and stationary analysis of inventory problems, chapter 1 of Multi-stage Inventory Models and Techniques (Edited by H. Scarf, D. Gilford, and M. Shelly) Stanford U. Press, Stanford, Calif.

[11] Maitra, Ashok (1964). Dynamic Programming for Countable State Systems, Doctoral thesis, U. of California, Berkeley.

[12] Manne, Alan (1960). Linear programming and sequential decisions, Management Sci., 6, 259-267.

[13] Strauch, Ralph E. (1965). Negative Dynamic Programming. Doctoral thesis, U. of California, Berkeley.

[14] Taylor, Howard. Markovian Sequential Replacement Processes (to appear in Ann. Math. Stat.).

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Stanford University Department of Statistics Stanford, California	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP

3. REPORT TITLE

Denumerable State Markovian Decision Processes--Average Cost Criterion

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report, February 1966

5. AUTHOR(S) (Last name, first name, initial)

Derman, Cyrus

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
28 February 1966	19	14

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Nonr-225(53)(FBM) b. PROJECT NO. NR 042-002 c. Nonr-225(53)(FBM) d.	Technical Report No. 86 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Logistics and Mathematical Statistics Branch Office of Naval Research Washington, D. C. 20360

13. ABSTRACT

Markovian decision processes with a countable number of states and average cost criterion are considered. Counterexamples are presented to show that optimal control rules need not exist or if they do exist they may not be of a deterministic stationary Markovian character. Conditions are presented under which optimal rules do exist and are stationary deterministic. Further conditions are presented under which a policy improvement procedure converges to an optimal rule.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Markovian sequential control processes Markovian decision processes discrete dynamic programming						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

DD FORM 1 JAN 64 1473 (BACK)

UNCLASSIFIED

Security Classification