DENUMERABLE STATE MARKOVIAN DECISION PROCESSES—AVERAGE COST CRITERION[1]

By Cyrus Derman

Columbia University and Stanford University

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 $t = 0, 1, 2, \cdots$. 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, $i \varepsilon I$. The decisions interact with the chance environment in the evolution of the system.

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

$$P\{Y_{t+1} = j \mid Y_0, \Delta_0, \cdots, Y_t = i, \Delta_t = k\} = q_{ij}(k),$$

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 $\{D_k(Y_0, \Delta_0, \cdots, Y_t)\}$ satisfying $0 \leqq D_k(Y_0, \Delta_0, \cdots, Y_t) \leqq 1$, for every k, and $\sum_{k=1}^{K_i} D_k(Y_0, \Delta_0, \cdots, Y_t = i) = 1$, for every history $Y_0, \Delta_0, \cdots, Y_t$ $(t = 0, 1, \cdots)$. As part of a controlling rule, $D_k(Y_0, \Delta_0, \cdots, Y_t)$ is the instruction at time t to make decision k with probability $D_k(Y_0, \Delta_0, \cdots, Y_t)$ if the particular history $Y_0, \Delta_0, \cdots, Y_t$ 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 $\{W_t\}$, $t = 0, 1, 2, \cdots$ by $W_t = w_{ik}$ if $Y_t = i$, $\Delta_t = k$, $t = 0, 1, \cdots$. For a given $Y_0 = i$ and rule R we can talk about $E_R W_t$, provided it exists. Let

$$Q_{T,R}(i) = (T + 1)^{-1} \sum_{t=0}^{T} E_R W_t, \quad \text{when} \quad Y_0 = i;$$

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 $Q_R(i) = \limsup_{T\to\infty} Q_{T,R}(i)$.

Received 3 March 1966.

[1] This work was supported in part by an Office of Naval Research contract (Nonr-225(53)-(NR-042-002)) at Stanford University.

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 *all* 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 $D_k(Y_0, \Delta_0, \cdots, Y_t = i) = D_{ik}$, independent of $Y_0, \Delta_0, \cdots, \Delta_{t-1}$ and t. A rule $R \varepsilon C'$ is completely defined by the set of numbers $\{D_{ik}\}, k = 1, \cdots, K_i, i \varepsilon I$; 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 $R \varepsilon C'$, the resulting stochastic process $\{Y_t\}, t = 0, 1, \cdots$, is a Markov chain with transtion probabilities

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

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 $R \varepsilon C''$. Computing methods using dynamic programming (Blackwell [1], Howard [9]) or linear programming (Manne [12]) exist for obtaining solutions.

For I infinite, and specifically denumerable, little has been 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 $K_i < \infty$, $i \varepsilon I$, 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) $K_i < \infty, i \varepsilon I$,

and

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

an optimal rule need not exist.

Let I consist of the states $0, 0', 1, 1', \cdots$. Suppose $K_i = 2, i = 0, 1, 2, \cdots$ and $K_i = 1, i = 0', 1', 2', \cdots$ where $q_{i,i+1}(1) = 1, q_{i,i'}(2) = 1$, and $q_{i'i'}(1) = 1$ for $i' = 0', 1', \cdots$. Assume $w_{ik} = 1$ for $i = 0, 1, \cdots$ and $k = 1, 2$; $w_{i'1} = w_{i'}$ for $i' = 0', 1', \cdots$ 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 $Y_0 = 0$. 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 *some* 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 $K_0 = 1, K_i = 2, i = 1, 2, \cdots$, with $q_{00}(1) = 0, q_{0i}(1) = g_i > 0, i = 1, 2, \cdots$; $q_{ii}(1) = 1, q_{i0}(2) = 1, i = 1, 2, \cdots$.

Let $w_{ik} = w_i$, $i = 0, 1, \cdots$, where $\{w_i\}$ is a decreasing sequence 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 $Y_0 = 0$. Let R be any rule in C''; let S_R be the set of states for which $D_{i1} = 1$. If $i \varepsilon S_R$, then $Y_t = i$ implies $Y_{t'} = i$ for all $t' > t$; $i \not\varepsilon S_R$, then $Y_t = i$ implies $Y_{t+1} = 0$. Suppose S_R is non-empty; then it can be shown that

$$Q_R(0) = \{\textstyle\sum_{i \varepsilon S_R} g_i w_i / \sum_{i \varepsilon S_R} g_i\} > 0.$$

If S_R is empty, then

$$Q_R(0) = (w_0 + \textstyle\sum_{i=1}^{\infty} g_i w_i)/2 > 0.$$

In either case $Q_R(0) > 0$. Thus, for every $R \varepsilon C''$, $Q_R(0) > 0$. Let $R \varepsilon C'$ be such that $0 < D_{i2} < 1, i \varepsilon I$, and $\sum_{i=1}^{\infty} g_i/D_{i2} = \infty$. State 0 is a recurrent state of the resulting Markov chain $\{Y_t\}$ since $P\{Y_t = 0 \text{ for some } t > 0 \mid Y_0 = 0\}$ 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_0 ,

$$\begin{aligned} Q_R(0) &= \lim_{T\to\infty} (T + 1)^{-1} \textstyle\sum_{i=0}^{\infty} \sum_{t=1}^{T} w_i P\{Y_t = i \mid Y_0 = 0\} \\ &= \lim_{T\to\infty} (T + 1)^{-1}\{\textstyle\sum_{i=0}^{i_0} \sum_{t=1}^{T} w_i P\{Y_t = i \mid Y_0 = 0\} \\ &\quad + \textstyle\sum_{i=i_0+1}^{\infty} \sum_{t=1}^{T} w_i P\{Y_t = i \mid Y_0 = 0\}\} \end{aligned}$$

$$\begin{aligned} &\leqq w_0 \textstyle\sum_{i=0}^{i_0} \lim_{T\to\infty} (T+1)^{-1} \sum_{t=0}^{T} P\{Y_t = i \mid Y_0 = 0\} \\ &\quad + w_{i_0} \lim_{T\to\infty} \textstyle\sum_{i=i_0+1}^{\infty} (T+1)^{-1} \sum_{t=0}^{T} P\{Y_t = i \mid Y_0 = 0\} \\ &= w_{i_0} \lim_{T\to\infty} (T+1)^{-1} \textstyle\sum_{t=0}^{T} P\{Y_t > i_0 \mid Y_0 = 0\} \\ &= w_{i_0}\,, \end{aligned}$$

since i, being null recurrent implies

$$\lim_{T\to\infty} (T+1)^{-1} \textstyle\sum_{t=0}^{T} P\{Y_t = i \mid Y_0 = 0\} = 0.$$

However, i_0 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 $R^* \varepsilon C - C'$ such that $Q_{R^*}(i) < Q_R(i)$ for all $R \varepsilon C'$ 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 part of 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* $\{g, v_j\}, j \varepsilon I$, *satisfying*

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

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

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

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

PROOF. Let k_i, $i \varepsilon I$, 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, $i \varepsilon I$. Let $p_{ij} = q_{ij}(k_i)$ for every $i, j \varepsilon I$. Then (1) becomes

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

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

$$\begin{aligned} (3) \qquad g + \textstyle\sum_{i\varepsilon I} p_{i'i}^{(t)} v_i &= \textstyle\sum_{i\varepsilon I} p_{i'i}^{(t)} w_{ik_i} + \sum_{i\varepsilon I} p_{i'i}^{(t)} \sum_{j\varepsilon I} p_{ik} v_j \\ &= \textstyle\sum_{i\varepsilon I} p_{i'i}^{(t)} w_{ik_i} + \sum_{j\varepsilon I} p_{i'j}^{(t+1)} v_j\,, \qquad i' \varepsilon I. \end{aligned}$$

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

$$(4)\qquad \begin{aligned} g &= \lim_{T\to\infty} (T+1)^{-1} \textstyle\sum_{t=0}^{T} \sum_{i\varepsilon I} p_{i'i}^{(t)} w_{ik_i} \\ &= Q_{R^*}(i'), \qquad i' \varepsilon I. \end{aligned}$$

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

$$(5)\qquad \begin{aligned} g_0(i) &= \min_k w_{ik}, \qquad i \varepsilon I, \\ g_{n+1}(i) &= \min_k \{w_{ik} + \textstyle\sum_{j\varepsilon I} q_{ij}(k) g_n(j)\}, \qquad i \varepsilon I; \end{aligned}$$

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

$$(6)\qquad ng + v_i - M \leqq g_n(i) \leqq ng + v_i + M, \qquad i \varepsilon I,$$

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

$$\begin{aligned} g_{N+1}(i) &\leqq \min_k \{w_{ik} + \textstyle\sum_{j\varepsilon I} q_{ij}(k)(Ng + v_j + M)\} \\ &= \min_k \{w_{ik} + \textstyle\sum_{j\varepsilon I} q_{ij}(k) v_j\} + Ng + M \\ &= (N+1)g + v_i + M, \qquad i \varepsilon I, \end{aligned}$$

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

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

$$\begin{aligned} \liminf_{n\to\infty} [h_n(i)/n] &\geqq \lim_{n\to\infty} [g_n(i)/n] \\ &= g, \qquad i \varepsilon I. \end{aligned}$$

This proves the theorem since $Q_R(i) \geqq \liminf_{n\to\infty} [h_n(i)/n]$.

We point out the following:

COROLLARY. *Under the conditions of Theorem* 1, $|g_n(i) - ng| \leqq 2M$ *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 $R \varepsilon C''$ 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 $\{g, v_j\}$, $j \varepsilon I$, 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

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

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

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

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 $i \not\varepsilon I'$, let $k_i' = k_i$. (Here, we allow that $k_i' = k_i$ even though (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 $p_{ij} = q_{ij}(k_i')$ $(i, j \varepsilon I)$. Let ϵ_i, $i \varepsilon I$, be the difference between the right side and left side of (8); thus, $\epsilon_i > 0$ if $i \varepsilon I'$ and $\epsilon_i = 0$ if $i \not\varepsilon I'$. For any $l \varepsilon I$ and t we get, using (2), that

$$\begin{aligned}\sum_{i \varepsilon I} p_{li}^{(t)} \epsilon_i &= \sum_{i \varepsilon I} p_{li}^{(t)} \{g + v_i - (w_{ik_i'} + \sum_{j \varepsilon I} p_{ij}v_j)\} \\ &= g + \sum_{i \varepsilon I} p_{li}^{(t)} v_i - \sum_{i \varepsilon I} p_{li}^{(t)} w_{ik_i'} - \sum_{j \varepsilon I} p_{lj}^{(t+1)} v_j\,.\end{aligned}$$

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

$$\begin{aligned}\sum_{i \varepsilon I} \epsilon_i \lim_{T\to\infty} (T+1)^{-1} \sum_{t=0}^{T} p_{li}^{(t)} & \\ &= g - \lim_{T\to\infty} (T+1)^{-1} \sum_{t=0}^{T} \sum_{i \varepsilon I} p_{li}^{(t)} w_{ik_i'} \\ &= g - Q_{R'}(l).\end{aligned} \tag{9}$$

However, under assumption (C), $\lim_{T\to\infty} (T+1)^{-1} \sum_{t=0}^{T} p_{li}^{(t)} > 0$ for every $i \varepsilon I$. Therefore, the left side of (9) is strictly positive since at least one ϵ_i is positive. Thus $Q_{R'}(l) < g = Q_R(l)$, $l \varepsilon I$, and the lemma is proved.

We remark that the amount of improvement obtained in changing from R to R' is precisely $\sum_{i \varepsilon I} \pi_i \epsilon_i$ where $\{\pi_i\}$, $i \varepsilon I$, 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 $R \varepsilon C''$ 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 $R \varepsilon C''$ there exists a set of real numbers $\{g^R, v_j^R\}$, $j \varepsilon I$ satisfying condition (D). The numbers $\{g^R, v_j^R\}$, are bounded uniformly over $j \varepsilon I$, $R \varepsilon C''$.

We then have the following existence:

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

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

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

Let g^* be the greatest lower bound of all g^R, $R \varepsilon C''$. Let $\{R_n\}$, $n = 1, \cdots$, 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 $\{R_{n_\nu}\}$, $\nu = 1, 2, \cdots$, such that $\lim_{\nu\to\infty} v_j^{R_{n_\nu}} = v_j^*$, $j \varepsilon I$, 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

$$\begin{aligned} g^* + v_i^* &= \lim_{\nu\to\infty} \{g^{R_{n_\nu}} + v_i^{R_{n_\nu}}\} \\ &= \lim_{\nu\to\infty}\{w_{iR_{n_\nu}} + \textstyle\sum_{j\varepsilon I} q_{ij}(R_{n_\nu})v_j^{R_{n_\nu}}\} \\ &= w_{iR^*} + \textstyle\sum_{j\varepsilon I} q_{ij}(R^*)v_j^*, \qquad i \varepsilon I. \end{aligned} \tag{11}$$

(The fact that $\lim_{\nu\to\infty} \sum_{j\varepsilon I} q_{ij}(R_{n_\nu})v_j^{R_{n_\nu}} = \sum_{j\varepsilon I} q_{ij}(R^*)v_j^*$ is easily shown.) Thus $\{g^*, v_j^*\}$, $j \varepsilon I$, is a bounded set of numbers satisfying (2) (or (10)) for $R = R^* \varepsilon C''$. That is, $g^* = g^{R^*}$, $v_j^* = v_j^{R^*}$, $j \varepsilon I$. 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, $R \varepsilon C''$. 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, $i \varepsilon I$) be any rule in C''. We define an iteration of the *policy improvement procedure for denumerable states* as the transformation from R to R' where the decisions $\{k_i'\}$ of R' are decisions for which $\{w_{ik} + \sum_{j\varepsilon I} q_{ij}(k)v_j^R\}$ 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 $R \varepsilon C''$. Before stating conditions under which a sequence of policy improvement iterations converges to an optimal rule we prove another lemma.

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

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

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

$$\begin{aligned} \epsilon_i^R &= (w_{ik_i} + \textstyle\sum_{j\varepsilon I} q_{ij}(k_i)v_j^R) - (w_{ik_i'} + \textstyle\sum_{j\varepsilon I} q_{ij}(k_i')v_j^R) \\ &= g^R + v_i^R - (w_{ik_i'} + \textstyle\sum q_{ij}(k_i')v_j^R), \qquad i \varepsilon I, \end{aligned}$$

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

LEMMA 2. *Assume conditions* (A), (B), (C), (D), (E), *and* (F). *Let*

$R_1 = R \varepsilon C''$ be arbitrary, and $\{R_n\}$ be a sequence of policy improvement iterations; then, for each $i \varepsilon I$, $\lim_{n\to\infty} \epsilon_i^{R_n} = 0$.

PROOF. Under assumption (C), for each $n = 1, 2, \cdots, \pi_{ij}(R_n) = \pi_j(R_n)$, the steady-state probability of state j under rule R_n. From the remark following Lemma 1 we can write, for each n,

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

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 $\epsilon_i^{R_n} \geqq 0$, it follows from condition (F) that $\lim_{n\to\infty} \epsilon_i^{R_n} = 0$, $i \varepsilon I$.

We can now state

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

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

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

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 $v_i^* = v_i^{R^*}$, $i \varepsilon I$. Clearly,

$$g^* + v_i^* \geqq \limsup_{\nu\to\infty} \min_k \{w_{ik} + \sum q_{ij}(k) v_j^{R_{n_\nu}}\}, \qquad i \varepsilon I. \tag{13}$$

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

$$g^{R_{n_\nu}} + v_i^{R_{n_\nu}} \leqq \min_k \{w_{ik} + \sum q_{ij}(k) v_j^{R_{n_\nu}}\} + \epsilon_i^{R_{n_\nu}}, \qquad i \varepsilon I. \tag{14}$$

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

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

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

$$\lim_{\nu\to\infty} \min_k \{w_{ik} + \sum_{j\varepsilon I} q_{ij}(k) v_j^{R_{n_\nu}}\} \leqq \lim_{\nu\to\infty} \{w_{ik} + \sum_{j\varepsilon I} q_{ij}(k) v_j^{R_{n_\nu}}\},$$

so that from (15)

$$\begin{aligned} g^* + v_i^* &\leqq \min_k \lim_{\nu\to\infty} \{w_{ik} + \sum_{j\varepsilon I} q_{ij}(k) v_j^{R_{n_\nu}}\} \\ &= \min_k \{w_{ik} + \sum_{j\varepsilon I} q_{ij}(k) v_j^*\}. \end{aligned}$$

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

5. Remarks. Conditions (D) and (E) require solutions to the equations (2). In a forthcoming paper by Derman and Veinott conditions for the existence and the form of the solutions will be given.

The conditions given in this paper are too strong to apply to special cases, such as certain inventory and replacement problems. It is not clear that they can be weakened sufficiently to cover these cases. It will probably be necessary to exploit the special structure of the processes under consideration, as was done in [10] and [14], in order to obtain comparable results.

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. (1966). A Markovian decision model for a joint replacement and stocking problem. Technical report number 93, O.N.R. contract number (Nonr-225(53)-(NR-042-002)), Stanford Univ. (To be submitted to *Management Sci.*)

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

[9] Howard, Ronald (1960). *Dynamic Programming and Markov Processes.* 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 Univ. Press.

[11] Maitra, Ashok (1964). Dynamic programming for countable state systems. Doctoral thesis, Univ. 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, Univ. of California, Berkeley.

[14] Taylor, Howard (1965). Markovian sequential replacement processes. *Ann. Math. Statist.* **36** 1677–1694.