The Annals of Mathematical Statistics
1968, Vol. 39, No. 4, 1220–1227

AN OPTIMALITY CONDITION FOR DISCRETE DYNAMIC PROGRAMMING WITH NO DISCOUNTING[1]

BY E. V. DENARDO AND B. L. MILLER

The RAND Corporation

0. Summary. In this paper we consider the discrete time finite state Markov decision problem with Veinott's criterion of maximizing the Cesaro mean of the vector of expected returns received in a finite horizon as the horizon tends to infinity. A necessary and sufficient condition for optimality is obtained, and at the same time we verify Veinott's conjecture that there are optimal stationary policies.

1. Introduction. This paper verifies a conjecture of Veinott [8] concerning the discrete-time Markov decision model. To introduce the model, consider a system that is observed sequentially at epochs labeled $1, 2, \cdots$. At each epoch, the system is observed to be in one of N states numbered $1, 2, \cdots, N$. If state i is observed at epoch n, a decision k in a finite set M_i is selected. This yields an immediate expected return $r(i, k)$ and a probability $p(j:i, k)$ that the observed state at epoch $n + 1$ will be state j, with $\sum_{j=1}^{N} p(j:i, k) = 1$. The data $r(i, k)$ and $p(j:i, k)$ are known to the decision-maker and depend only on the current state i and decision k, not on prior states or decisions.

A stationary non-randomized policy δ for this system is a rule that for each state i selects a decision δ_i in M_i. The set Z of all such decision rules is called the policy space and is given by $Z = \times_{i=1}^{N} M_i$. With $\delta \varepsilon Z$, we have $\delta = (\delta_1, \cdots, \delta_i, \cdots, \delta_N)$, with δ_i being the decision (in M_i) that is made at any epoch at which state i is observed. Policy δ has associated with it a vector $r(\delta)$ of immediate returns and a transition matrix $P(\delta)$ with $r(\delta)_i = r(i, \delta_i)$ and $P(\delta)_{ij} = p(j:i, \delta_i)$. A non-randomized transition counting (Markov) policy Δ is an element of $Z^\infty = \times_{i=1}^{\infty} Z$, with $\Delta = (\delta^1, \delta^2, \cdots)$ and δ_i^n being the decision (in δ^n) if state i is observed at epoch n. Let P_Δ^n be the N by N matrix whose ijth element is the probability that state j is observed at epoch $n + 1$, given state i at epoch 1 and given policy Δ. Then, $P_\Delta^1 = P(\delta^1)$ and $P_\Delta^{n+1} = P_\Delta^n P(\delta^{n+1})$. Similarly, let $V(n, \Delta)$ be the vector of expected total returns for epochs 1 through n using policy Δ in Z^∞. Then, $V(n, \Delta) = \sum_{i=1}^{n} P_\Delta^{i-1} r(\delta^i)$.

The average rate of gain for the first n epochs using policy Δ is $n^{-1}V(n, \Delta)$. A standard criterion for the undiscounted problem is to select a policy Δ that maximizes $n^{-1}V(n, \Delta)$ as $n \to \infty$. However, this criterion is rather unselective in that the average depends only on the tail of the income stream and not on the income in the first millennium. Examining $V(n, \Delta)$ rather than $n^{-1}V(n, \Delta)$ one

Received 28 December 1967.

[1] This research is supported by the United States Air Force under Project RAND—Contract No. F44620-67-C-0045—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF.

would say that Π is at least as good as Δ if $V(n, \Pi) \geqq V(n, \Delta)$ for all n sufficiently large or, alternatively and more conservatively, if $\liminf_{n\to\infty} [V(n, \Pi) - V(n, \Delta)] \geqq 0$. Unfortunately, both of these methods of comparison are overly selective; they sometimes preclude the existence of an optimal policy, as the following example attests. There are three states, 1, 2, and 3, actions a and b for state 1 and one action, c, for the others. One has $p(3{:}2, c) = p(2{:}3, c) = 1$, $r(2, c) = 0$, $r(3, c) = 2$, $r(1, a) = 1$ and $p(2{:}1, a) = 1$, $r(1, b) = 0$ and $p(3{:}1, b) = 1$. Choosing a for state 1 yields the cumulative income stream $(1, 1, 3, 3, 5, \cdots)$, while b yields $(0, 2, 2, 4, 4, \cdots)$; action a is better for n odd and worse for n even. In the example and in general, the essential difficulty is that two policies Δ and Π may have $\{V(n, \Pi) - V(n, \Delta)\}$ oscillating around zero with amplitude that remains finite as $n \to \infty$. Veinott [8] uses $(C, 1)$ summation to damp down such oscillations. That is, he writes $\Pi \geqq \Delta$ if

$$\liminf_{n\to\infty} n^{-1} \textstyle\sum_{i=1}^{n} \{V(i, \Pi) - V(i, \Delta)\} \geqq 0. \tag{1}$$

We shall call a policy Π *Veinott-optimal* if $\Pi \geqq \Delta$ for every policy Δ in Z^{∞}. One is tempted to call such a policy Π "optimal." However, we prefer to reserve this term for Blackwell's [1] meaning—namely, for a policy that is optimal for all discount factors near enough to 1. A policy may then be Veinott-optimal but not optimal, as in example 1 of [1].

This paper shows that a Veinott-optimal policy exists and provides a characterization of the Veinott-optimal policies. Rather than selecting an appropriate criterion for the undiscounted problem as Veinott did, one could demand optimal or near-optimal behavior of the system for discount factors near 1, as Blackwell [1] did. We feel that both approaches have merit. Veinott's approach seems related to the turnpike theorems in economics.

Toward reviewing Blackwell's approach [1] and associating it with Veinott's, first note that since $P(\delta)$ is a stochastic matrix, the sequence $\{n^{-1} \sum_{i=1}^{n}[P(\delta)]^{i}\}$ converges to a stochastic matrix P_δ^* such that $P_\delta^* = P_\delta^* P(\delta) = P(\delta)P_\delta^* = P_\delta^* P_\delta^*$. With discount factor $c < 1$ per epoch, the N-vector v_c^{δ} of expected total discounted return using policy δ is given by $v_c^{\delta} = \sum_{n=0}^{\infty} c^n [P(\delta)]^n r(\delta) = [I - cP(\delta)]^{-1} r(\delta)$, since the series converges geometrically. Blackwell examined the limiting behavior of v_c^{δ} as c approaches 1 and obtained the asymptotic expression

$$v_c^{\delta} = g^{\delta}/(1 - c) + w^{\delta} + o(1) \tag{2}$$

where $g^{\delta} = P_\delta^* r(\delta)$ and where w^{δ} is the unique solution of the equations

$$r(\delta) + P(\delta)w^{\delta} = w^{\delta} + g^{\delta}, \qquad P_\delta^* w^{\delta} = 0. \tag{3}$$

For any stationary policy δ in Z, let $\delta^{\infty} = (\delta, \delta, \cdots)$. As one might suspect from Abelian analogies, $V(n, \delta^{\infty})$ is readily associated with g^{δ} and w^{δ}. Using the fact that $P_\delta g^{\delta} = g^{\delta}$, one can readily verify inductively the observation of Veinott [8] (see also Denardo [3]) that

$$V(n, \delta^\infty) = ng^\delta + w^\delta - [P(\delta)]^n w^\delta, \tag{4}$$

$$n^{-1} \sum_{l=1}^n V(l, \delta^\infty) = (\tfrac{1}{2})(n+1)g^\delta + w^\delta - e_n \tag{5}$$

$$\text{with} \quad e_n = n^{-1} \sum_{l=1}^n [P(\delta)]^l w^\delta \to 0,$$

the last since $P_\delta{}^* w^\delta = 0$. Let g^* and w^* be the N-vectors defined by

$$g_i{}^* = \max \{g_i{}^\delta : \delta \varepsilon Z\}, \qquad w_i{}^* = \max \{w_i{}^\delta : \delta \varepsilon Z, g_i{}^\delta = g_i{}^*\}.$$

Policy λ in Z is called *g-optimal* if $g^\lambda = g^*$ and *(g, w)-optimal* if, in addition, $w^\lambda = w^*$. Existence of a (g, w)-optimal policy is a non-trivial question, since N maxima must be attained simultaneously. For a proof, see Blackwell [1] or Veinott [8].

Veinott ([8], Theorem 7) showed that every (g, w)-optimal policy λ has $\lambda^\infty \geqq \delta^\infty$ for every δ in Z; i.e., that for stationary policies (g, w)-optimality implies Veinott-optimality. He further conjectured that $\lambda^\infty \geqq \Pi$ for every Π in Z^∞, which we shall verify.

With $\Delta = (\delta^1, \delta^2, \cdots)$ in Z^∞, define the operator $H_\Delta{}^n$ on Euclidean N-space by

$$H_\Delta{}^n(u) = \sum_{l=0}^{n-1} P_\Delta{}^l r(\delta^{l+1}) + P_\Delta{}^n u$$

for each N-vector u. Then $H_\Delta{}^n(u)$ is the vector of total expected rewards for epochs 1 through n using policy Δ and having terminating reward vector u. The main result of this paper is summarized in

THEOREM 1. *A necessary and sufficient condition for policy* Π *in* Z^∞ *to be Veinott-optimal is that* Π *satisfy the equations*

$$H_\Pi{}^n(w^*) = w^* + ng^* \quad \textit{for every} \quad n,$$

$$\lim_{n\to\infty} \{n^{-1} \sum_{l=1}^n P_\Pi{}^l w^*\} = 0.$$

Every (g, w)-optimal policy λ in Z is Veinott-optimal.

Of course, the conditions for Veinott-optimality in Theorem 1, coupled with equations (4) and (5), imply that every stationary Veinott-optimal policy is (g, w)-optimal.

Theorem 1 is proved in the next section. It is a simple matter to adapt the proof to the case in which Z^∞ is replaced by the (larger) set of randomized transition-counting policies. This, coupled with a result of Derman and Strauch [6], yields the generalization of Theorem 1 in which Z^∞ is replaced by the (still larger) set of all history-remembering randomized decision rules.

We observe that if "lim inf" were replaced in equation (1) by the less demanding "lim sup", the optimality of a stationary policy would be readily verified by an Abelian argument that also works for the more general Markov renewal programming model. We close this section by pointing out that Denardo [3] has recently shown how to obtain (g, w)-optimal policies by solving at most three simpler Markovian decision problems, each of which can be solved by linear programming or policy iteration.

2. The proof. This section is devoted to the proof of Theorem 1. Always, Δ is the policy $(\delta^1, \delta^2, \cdots)$ using the corresponding small Greek letters. If policy Δ is Veinott-optimal, then $\Delta \geqq \lambda^\infty$, where λ is any (g, w)-optimal policy. Then, from equations (1) and (5), we see that a prerequisite for Δ to be Veinott-optimal is that

$$(6) \qquad \liminf_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^n V(l, \Delta) - \frac{1}{2}(n+1)g^* - w^*\} \geqq 0.$$

Our line of attack is to fix state i and policy Π, assume that the pair (i, Π) satisfy

$$(7) \qquad \limsup_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^n V(l, \Pi)_i - \frac{1}{2}(n+1)g_i^* - w_i^*\} \geqq 0,$$

and investigate the consequences. We start with three observations about equation (7). First, it is satisfied by at least one policy; in fact, equation (5) assures that for every (g, w)-optimal policy π in Z the policy $\Pi = \pi^\infty$ satisfies equation (7). Second, we shall eventually conclude that strict inequality in equation (7) is impossible; were it possible, no (g, w)-optimal policy would be Veinott-optimal, invalidating Theorem 1. Third, were "lim sup" replaced in equation (7) by "lim inf," the finding that equality holds in equation (7) would not let us conclude that a (g, w)-optimal policy is Veinott-optimal. This explains why "lim sup" is used here.

For every N-vector u, let $\|u\| = \max_{1 \leqq i \leqq N} |u_i|$ and let $\mathbf{1}$ be the N-vector of 1's. Lemma 1 summarizes results, some well known, about the model that are germane to our argument.

LEMMA 1. (a) *$P(\delta)g^* \leqq g^*$ for every δ in Z. Also, $P_\Delta{}^n g^* \leqq g^*$ for every n and every Δ in Z^∞.*

(b) *There exists an N-vector u^* such that $r(\delta) + P(\delta)u^* \leqq u^* + g^*$ for every δ in Z.*

(c) *$H_\Delta{}^n(u^*) \leqq u^* + ng^*$ for every n and Δ in Z^∞.*

(d) *$n^{-1}V(n, \Delta) \leqq g^* + n^{-1}2\,\|u^*\|\,\mathbf{1}$ for every n and Δ.*

(e) *If $P(\delta)g^* = g^*$, then $r(\delta) + P(\delta)\,w^* \leqq w^* + g^*$.*

PROOF. The first statement in (a), and (e), are prerequisites for Howard's [7] policy iteration routine to terminate at g^*; for proofs see, e.g., [7], [1], [8], or [4]. The second half of (a) is a trivial consequence of the first and the monotonicity of $P(\delta)$ since, for instance, $P_\Delta{}^2(g^*) = P(\delta^1)P(\delta^2)g^* \leqq P(\delta^1)g^* \leqq g^*$. (b) is Lemma 7 of Denardo and Fox [4] and, in slightly different form, Lemma 4.1 in Brown [2]. (c) follows routinely from (a) and (b) by induction; for instance, $H_\Delta{}^2(u^*) = H_\Delta{}^1[r(\delta^2) + P(\delta^2)u^*] \leqq H_\Delta{}^1(u^* + g^*) = H_\Delta{}^1(u^*) + P_\Delta{}^1(g^*) \leqq u^* + 2g^*$. For (d), note that $V(n, \Delta) = H_\Delta{}^n(u^*) - P_\Delta{}^n(u^*) \leqq H_\Delta{}^n(u^*) + \|u^*\|\,\mathbf{1} \leqq ng^* + 2\,\|u^*\|\,\mathbf{1}$ by (c). $\square$

LEMMA 2. *Suppose policy Π and state i satisfy equation (7). Then $(P_\Pi{}^n g^*)_i = g_i^*$ for every n.*

PROOF. (a) of Lemma 1 establishes $P_\Pi{}^n g^* \leqq g^*$. Suppose Lemma 2 is false. Then there exists an integer m and a number $a > 0$ such that $(P_\Pi{}^m g^*)_i = g_i^* - a$. Coupling the fact that $V(j, \Delta) = H_\Delta{}^j(u) - P_\Delta{}^j(u)$ for every j, u and Δ with

(c) of Lemma 1 produces, for $n > m$,

$$
\begin{aligned}
V(n, \Pi) &= H_\Pi^n(u^*) - P_\Pi^n(u^*) \leqq H_\Pi^n(u^*) + \|u^*\| \mathbf{1} \\
&\leqq H_\Pi^m[u^* + (n - m)g^*] + \|u^*\| \mathbf{1} \\
&= H_\Pi^m(u^*) + P_\Pi^m[(n - m)g^*] + \|u^*\| \mathbf{1} \\
&\leqq (n - m)(g^* - ae_i) + H_\Pi^m(u^*) + \|u^*\| \mathbf{1}
\end{aligned} \tag{8}
$$

where e_i is the N-vector with 1 in position i and zeros elsewhere. Then, for some scalar K, $V(n, \Pi) \leqq n(g^* - ae_i) + K\mathbf{1}$ for every n, contradicting equation (7). □

Let

$$
\begin{aligned}
E(j) = \{d \in M_j : \textstyle\sum_l p(l:j, d)g_l^* &= g_j^* \\
\text{and}\quad r(j, d) + \textstyle\sum_l p(l:j, d)w_l^* &= w_j^* + g_j^*\};
\end{aligned}
$$

$$
\begin{aligned}
C_{ij}^k &= 1 \quad \text{if} \quad \pi_j^k \notin E(j) \\
&= 0 \quad \text{otherwise};
\end{aligned}
$$

$$S_i^n = \textstyle\sum_{k=1}^n \sum_{j=1}^N (P_\Pi^{k-1})_{ij} C_{ij}^k, \qquad S_i = \lim_{n\to\infty} S_i^n;$$

$$H(\delta, u) = r(\delta) + P(\delta)u.$$

Lemma 3. *Suppose policy* Π *and state* i *satisfy equation* (7). *Then* $S_i < \infty$ *and there exists a number* $b < 0$ *such that for every* n

$$V(n, \Pi)_i \leqq ng_i^* + w_i^* + P_\Pi^n(-w^*)_i + S_i^n b. \tag{9}$$

Proof. We first obtain the intermediate result that

$$H_\Pi^n(w^*)_i = ng_i^* + w_i^* + \textstyle\sum_{k=1}^n \{P_\Pi^{k-1}[H(\pi^k, w^*) - g^* - w^*]\}_i. \tag{10}$$

Since $(P_\Pi^k g^*)_i = g_i^*$ for every k by Lemma 2,

$$
\begin{aligned}
ng_i^* + w_i^* + \textstyle\sum_{k=1}^n \{P_\Pi^{k-1}[H(\pi^k, w^*) - g^* - w^*]\}_i \\
= w_i^* + \textstyle\sum_{k=1}^n \{P_\Pi^{k-1}[r(\pi^k) + P(\pi^k)w^* - w^*]\}_i \\
= H_\Pi^n(w^*)_i .
\end{aligned}
$$

Note that $[P(\pi^k)g^*]_j = g_j^*$ whenever $(P_\Pi^{k-1})_{ij} > 0$, since otherwise $[P_\Pi^k g^*]_i < g_i^*$ which would contradict Lemma 2. Then, by (e) of Lemma 1, $H(\pi^k, w^*)_j - w_j^* - g_j^* \leqq 0$ whenever $(P_\Pi^{k-1})_{ij} > 0$. Furthermore, if $H(\pi^k, w^*)_j - w_j^* - g_j^*$ is negative it must be bounded away from zero, since Z is finite; let $b < 0$ be this bound. Since $V(n, \Pi) = H_\Pi^n(w^*) + P_\Pi^n(-w^*)$, substitution in equation (10) yields equation (9). If $S_1 = \infty$, then equation (9) contradicts equation (7). □

For Lemma 4 we make a simple preliminary observation. Suppose τ in Z satisfies $g^\tau = g^*$. Then $0 \geqq P_\tau(w^\tau - w^*)$; hence $0 \geqq P_\tau^n(w^\tau - w^*)$, implying $0 \geqq P_\tau^*(w^\tau - w^*) = P_\tau^*(-w^*)$. Let $E = \times_{j=1}^N E(j)$ and consider

LEMMA 4. *Suppose policy* $\Delta = (\delta^1, \delta^2, \cdots)$ *and the integer* M *satisfy* $\delta^n \varepsilon E$ *for all* $n > M$. *Then*

$$\limsup_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^{n} P_\Delta{}^l(-w^*)\} \leqq 0. \tag{11}$$

PROOF. Recall that the average of a series depends only on its tail. Then, with policy $T = (\tau^1, \tau^2, \cdots)$ defined by $\tau^n = \delta^{M+n}$ for each n,

$$\limsup_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^{n} P_\Delta{}^l(-w^*)\} \leqq P_\Delta{}^M\{\limsup_{n\to\infty} n^{-1} \textstyle\sum_{l=0}^{n-1} P_T{}^l(-w^*)\}.$$

Since $P_\Pi{}^M$ has non-negative elements, it suffices for Lemma 4 to show that

$$\limsup_{n\to\infty} \{n^{-1} \textstyle\sum_{l=0}^{n-1} P_T{}^l(-w^*)\} \leqq 0. \tag{12}$$

Note that $n^{-1} \sum_{l=0}^{n-1} P_T{}^l(-w^*)$ can be interpreted as the average rate of gain for epochs 1 through $n - 1$ using policy T of the following discrete time Markov decision process: the states are 1 through N as before, the decision set for state i is $E(i)$, the immediate return for being in state i is $-w_i^*$ (independent of the decision) and the transition probabilities are unchanged. This "new" problem is precisely the one introduced by Veinott [8] and further studied by Denardo [3]. We noted above that $P_\tau{}^*(-w^*) \leqq 0$ for every τ in E. Of course, every (g, w)-optimal policy λ for the original problem satisfies $\lambda \varepsilon E$ and $P_\lambda{}^*(-w^*) = 0$. Hence, λ is g-optimal for the new problem and the maximum gain rate for the new problem is zero. Applying (d) of Lemma 1 to the new problem (for which $g^* = 0$) immediately yields equation (12). □

LEMMA 5. *Suppose policy* Π *and state* i *satisfy equation* (7). *Then,* $S_i = 0$, *expression* (7) *is satisfied as an equality, and*

$$\limsup_{n\to\infty} [n^{-1} \textstyle\sum_{l=1}^{n} P_\Pi{}^l(-w^*)]_i = 0. \tag{13}$$

PROOF. We shall truncate policy Π and use Lemma 4. Let λ be a (g, w)-optimal policy. Deferring momentarily the selection of the truncation integer M, define policy $\Delta = (\delta^1, \delta^2, \cdots)$ by

$$\begin{aligned} \delta_j{}^n &= \lambda_j \qquad \text{if } \pi_j{}^n \not\varepsilon E(j) \quad \text{and} \quad n > M \\ &= \pi_j{}^n \qquad \text{otherwise.} \end{aligned}$$

Lemma 3 assures $S_i < \infty$ and thus allows us to pick M big enough that, given $\epsilon > 0$,

$$|(P_\Pi{}^n)_{ij} - (P_\Delta{}^n)_{ij}| < \epsilon \qquad \text{for all } n \text{ and } j. \tag{14}$$

As defined, Δ satisfies the hypothesis of Lemma 4. It follows from equations (14) and (11) that

$$\limsup_{n\to\infty} [n^{-1} \textstyle\sum_{l=1}^{n} P_\Pi{}^l(-w^*)]_i \leqq \epsilon \, \|w^*\| N,$$

where, we recall, N is the number of states. Since ϵ is arbitrary, this implies

$$\limsup_{n\to\infty} [n^{-1} \textstyle\sum_{l=1}^{n} P_\Pi{}^l(-w^*)]_i \leqq 0. \tag{15}$$

Finally, combining equations (7) and (9),

$$\begin{aligned} 0 &\leqq \limsup_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^{n} [V(l, \Pi)_i - lg_i^* - w_i^*]\} \\ &\leqq \limsup_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^{n} [P_\Pi^l(-w^*)_i + S_i^l b]\} \\ &= \limsup_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^{n} P_\Pi^l(-w^*)_i\} + S_i b \leqq 0, \end{aligned}$$

the last since both terms are non-positive. Thus, equality must hold throughout, verifying that $S_i = 0$, proving that equation (7) is satisfied as an equality and establishing equation (13). □

Note that since $(P_\Pi^n g^*)_i = g_i^*$ and $S_i = 0$, the definition of S_i yields $H_\Pi^n(w^*)_i = ng_i^* + w_i^*$ for each n. This leads us directly to the

PROOF OF THEOREM 1. Consider the conditions on a policy Π:

$$(16)\quad H_\Pi^n(w^*) = w^* + ng^* \quad \text{for each } n, \quad \lim_{n\to\infty} n^{-1} \textstyle\sum_{l=1}^{n} P_\Pi^l(-w^*) = 0.$$

With λ as any (g, w)-optimal policy and $\Pi = \lambda^\infty$, policy Π satisfies (16). We shall show that (16) is a necessary and sufficient condition for a policy to be Veinott-optimal.

First, suppose policy Δ satisfies (16). Then, since $V(l, \Delta) = H_\Delta^l(w^*) + P_\Delta^l(-w^*)$, Δ satisfies (6). For any Π in Z^∞ the inequality of (7) goes the other way by Lemma 5. Hence $\Delta \geq \Pi$ for every Π in Z^∞; i.e., Δ is Veinott-optimal.

Next, suppose policy Π is Veinott-optimal. Then Π satisfies equation (6) and hence equation (7) for every i, allowing us to apply Lemmas 2 through 5 to Π. By Lemma 5, $S_i = 0$ for each i. Hence, equation (10) implies $H_\Pi^n(w^*) = ng^* + w^*$. By equation (6) and the fact that $V(l, \Pi) = H_\Pi^l(w^*) + P_\Pi^l(-w^*)$,

$$\liminf_{n\to\infty} \{n^{-1} \textstyle\sum_{l=1}^{n} P_\Pi^l(-w^*)\} \geqq 0$$

which, when combined with equation (13), shows that $n^{-1} \sum_{l=1}^{n} P_\Pi^l(-w^*) \to 0$ and completes the proof. □

We close the discussion by sketching a line of argument quite different from the above that obtains equation (15) immediately from Lemma 2. Let $\{P\}_i$ denote the ith row of the matrix P. This argument exploits a result of Derman [5], namely that every limit point of $\{n^{-1} \sum_{l=1}^{n} P_\Pi^l\}_i$ is attained by an initial randomization over stationary non-randomized policies, i.e., is equal to $\sum_{\delta \varepsilon Z} c^\delta \{P_\delta^*\}_i$ where $c^\delta \geqq 0$ and $\sum_\delta c^\delta = 1$. One argues that a prerequisite for equation (7) is that $g_i^\delta = g_i^*$ whenever $c^\delta > 0$. Then one uses the fact given after Lemma 3 that $P_\delta^*(-w^*) \leqq 0$ whenever $g^\delta = g^*$ to obtain equation (15).

Acknowledgment. We wish to thank Ben Fox and Ralph Strauch for pointing out several obscurities in an earlier draft.

REFERENCES

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

[2] BROWN, B. (1965). On the iterative method of dynamic programming on a finite space discrete time Markov process. *Ann. Math. Statist.* **36** 1279–1285.

[3] DENARDO, E. V. (1967). Computing (g,w)-optimal policies in discrete and continuous Markov programs. The RAND Corporation, RM-5538-PR.

[4] DENARDO, E. V. and FOX, B. L. (1967). Multichain Markov renewal programs. To appear in *J. SIAM*.

[5] DERMAN, C. (1964). On sequential control processes. *Ann. Math. Statist.* **35** 341–349.

[6] DERMAN, C. and STRAUCH, R. E. (1966). A note on memoryless rules for controlling sequential decision processes. *Ann. Math. Statist.* **37** 276–279.

[7] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes*. Wiley, New York.

[8] VEINOTT, A. F. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.* **37** 1284–1294.