JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 26, 221–232 (1969)

Linear Programming Considerations on Markovian Decision Processes with No Discounting

SHUNJI OSAKI AND HISASHI MINE

Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University, Kyoto, Japan

Submitted by R. Bellman

1. INTRODUCTION

A stochastic sequential control system has been studied as a Markovian Decision Process (M.D.P.) originally discussed by Bellman [1]. In M.D.P.'s two approaches have been studied. One is a Policy Iteration Algorithm (P.I.A.) originally formulated by Howard [2]. This approach has been extended by Blackwell [3], Veinott [4], and others. Another is a Linear Programming (L.P.) Algorithm originally formulated by Manne [5]. Further, Wolfe and Dantzig [6], Derman [7], D'Epenoux [8], De Ghellinck and Eppen [9], and others have also discussed L.P. approach. It is also known that these two approaches are mutually dual in mathematical programming, i.e., these are equivalent. This fact of duality is only known when M.D.P.'s are discounting, completely ergodic with no discounting in the sense of [2], or terminating in the sense of [10]. But no result has been established for a general M.D.P. in which there are some ergodic sets plus a transient set, and these sets may change according to any policy we choose.

In this paper we formulate a general M.D.P. with no discounting by an L.P. problem. And we give a procedure to solve this L.P. problem. We further show that P.I.A. is equivalent to this L.P. problem, i.e., P.I.A. is a special structure algorithm of the revised L.P. in which pivot operations for many variables are performed simultaneously. An example is presented to understand this relation of equivalence. We show that L.P. problems formulated here contain those of completely ergodic, and terminating M.D.P.'s as special cases. Finally we extend this discussion to semi-Markovian Decision processes (semi-M.D.P.'s).

2. PRELIMINARIES

Consider a system whose state space S has a finite set of states labeled by $i = 1, 2,..., N$. We observe periodically one of states at time $n = 1, 2,...$ and

have to make a decision chosen from given actions at each time. In each state $i \in S$ we have a set K_i of actions labeled by $k = 1, 2, \ldots, K_i$. When we choose an action k in state i at any time, the following two things happen:

(i) We receive the return r_i^k.

(ii) The system obeys the probability law p_{ij}^k ($j \in S$) at next time, where p_{ij}^k is the transition probability that the system is in state j at next time, given that the system is in state i at this time.

Here we assume that r_i^k is bounded. Moreover, we give an initial distribution

$$a = (a_1, a_2, \ldots, a_N), \tag{1}$$

where

$$a_i \geqslant 0 \quad (i \in S), \qquad \sum_{i \in S} a_i = 1. \tag{2}$$

Let F be a set of functions f from S to $\mathbf{X}_{i=1}^N K_i$ (Cartesian product). A policy π is defined as a sequence $(f_1, f_2, \ldots, f_n, \ldots)$, where $f_n \in F$. We call $\pi = (f, f, \ldots, f, \ldots)$ a stationary policy and write $\pi = f^\infty$. When we choose an action k in state i at time n, we write $k = f_n(i)$. Specifying $f \in F$, we have an $N \times N$ Markov matrix $Q(f)$ whose i-jth element p_{ij}^k, and an $N \times 1$ column vector $r(f)$ whose ith element r_i^k, where $k = f(i)$.

LEMMA 1 (Blackwell [3]). *Let Q be any $N \times N$ Markov matrix.*

(a) $1/n \sum_{i=0}^{n-1} Q^i$ *converges as* $n \to \infty$ *to a Markov matrix* Q^* *such that*

$$QQ^* = Q^*Q = Q^*Q^* = Q^*. \tag{3}$$

(b) $\operatorname{rank}(I - Q) + \operatorname{rank} Q^* = N.$ (4)

(c) *For every $N \times 1$ column vector c, the system*

$$Qx = x, \qquad Q^*x = Q^*c \tag{5}$$

has a unique solution.

(d) $I - (Q - Q^*)$ *is nonsingular, and*

$$H(\beta) = \sum_{i=0}^{\infty} \beta^i (Q^i - Q^*) \to H = (I - Q + Q^*)^{-1} - Q^* \tag{6}$$

as $\beta \nearrow 1$ $(0 \leqslant \beta < 1)$.

$$H(\beta)Q^* = Q^*H(\beta) = HQ^* = Q^*H = 0 \tag{7}$$

and

$$(I - Q)H = H(I - Q) = I - Q^*. \tag{8}$$

Let β $(0 \leqslant \beta < 1)$ be a discount factor. Then the discounted total expected return vector starting in each state $i \in S$ is given by

$$V_\beta(\pi) = \sum_{i=0}^{\infty} \beta^i Q_i(\pi)\, r(f_{i+1}), \tag{9}$$

where $Q_i(\pi) = Q(f_1) \cdots Q(f_i)$ $(i > 0)$ and $Q_0(\pi) = I$.

THEOREM 2 (Blackwell [3]). *Take any $f \in F$ and denote by $Q^*(f)$ the matrix Q^* associated with $Q(f)$. Then*

$$V_\beta(f^\infty) = \frac{u(f)}{1-\beta} + v(f) + \epsilon(\beta, f), \tag{10}$$

where $u(f)$ is a unique solution of

$$(I - Q(f))\, u = 0, \qquad Q^*(f)\, u = Q^*(f)\, r(f), \tag{11}$$

$v(f)$ is a unique solution of

$$(I - Q(f))\, v = r(f) - u(f), \qquad Q^*(f)\, v = 0, \tag{12}$$

and $\epsilon(\beta, f) \to 0$ as $\beta \nearrow 1$.

We now consider the limit infimum of the average return per unit time starting in an initial distribution a. For any policy π, we define

$$G(\pi) = \liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} a Q_i(\pi)\, r(f_{i+1}). \tag{13}$$

Then π^* is called an optimal policy (under average return criterion) if $G(\pi^*) = \sup_\pi G(\pi)$.

Next theorem is well known. The proof needs a slight modification of Blackwell [3], or Derman [7].

THEOREM 3. *There exists an optimal policy which is stationary.*

As Blackwell [3] has pointed out, a nearly optimal policy, which is in optimal policies defined here, is important if two or more optimal policies exist. But in this paper we restrict our attention to finding only an optimal stationary policy. We do not discuss here a problem of finding a nearly optimal policy. A problem of finding a nearly optimal policy will be discussed in near future.

3. LINEAR PROGRAMMING ALGORITHM

Using Theorem 3, we study a problem of finding an optimal stationary policy. Restricting our attention to stationary policies, we write f instead of f^∞.

From (13) we have for any policy

$$G(f) = aQ^*(f)\, r(f) = \sum_{i\in S} \sum_{j\in S} a_i q_{ij}^*(f)\, r_j(f), \tag{14}$$

where $Q^*(f) = [q_{ij}^*(f)]$. Thus an optimal objective is to find an f such that

$$\max_{f\in F} \sum_{i\in S} \sum_{j\in S} a_i q_{ij}^*(f)\, r_j(f). \tag{15}$$

It is convenient to extend policies to randomized stationary ones. So, let d_j^k denote the probability that we choose an action k in state j. It is evident that

$$d_j^k \geqslant 0 \quad (j \in S, k \in K_j), \qquad \sum_{k\in K_j} d_j^k = 1. \tag{16}$$

As we note in later discussion, we consider any fixed nonrandomized policy.

Setting

$$x_j^k = \sum_{i\in S} a_i q_{ij}^*(f)\, d_j^k \qquad (j \in S, k \in K_j), \tag{17}$$

$$y_j^k = \sum_{i\in S} a_i h_{ij}(f)\, d_j^k \qquad (j \in S, k \in K_j), \tag{18}$$

where

$$[h_{ij}(f)] = H(f) = (I - Q(f) + Q^*(f))^{-1} - Q^*(f), \tag{19}$$

and using the relations $Q^*(I - Q) = 0$ (from (3)), $Q^* + H(I - Q) = I$ (from (8)), we have

$$\sum_{j\in S} \sum_{k\in K_j} (\delta_{jl} - p_{jl}^k)\, x_j^k = 0 \qquad (l \in S), \tag{20}$$

and

$$\sum_{k\in K_l} x_l^k + \sum_{j\in S} \sum_{k\in K_j} (\delta_{jl} - p_{jl}^k)\, y_j^k = a_l \qquad (l \in S), \tag{21}$$

where δ_{jl} is the Kronecker's delta. (See Ref. [11], which is suggestive to calculate the above equations.) It is evident from (1), (16), and (17) that

$$x_j^k = \sum_{i\in S} a_i q_{ij}^*(f)\, d_j^k \geqslant 0 \qquad (j \in S, k \in K_j). \tag{22}$$

While the sign of y_j^k is not clear, and Lemma 6 will give the answer.

Thus we have a following L.P. problem with (N) redundant constraints:

Maximize

$$\sum_{j\in S}\sum_{k\in K_j} r_j^k x_j^k \tag{23}$$

subject to

$$\sum_{j\in S}\sum_{k\in K_j}(\delta_{jl}-p_{jl}^k)\,x_j^k = 0 \qquad (l\in S), \tag{24}$$

$$x_j^k \geqslant 0 \qquad (j\in S,\ k\in K_j), \tag{25}$$

$$\sum_{k\in K_l} x_l^k + \sum_{j\in S}\sum_{k\in K_j}(\delta_{jl}-p_{jl}^k)\,y_j^k = a_l \qquad (l\in S). \tag{26}$$

We shall afterwards show that $y_j^k \geqslant 0$ for any j of a subset of S.

For a fixed policy f, Markov matrix $Q(f)$ has some ergodic sets plus a transient set. Approrpiately relabeling the number of states, we have the following form for Markov matrix $Q(f)$:

$$Q(f) = \left[\begin{array}{cccc|c} Q_{11} & & & 0 & \\ & Q_{22} & & & \\ & & \ddots & & 0 \\ & & & & \\ 0 & & & Q_{\nu\nu} & \\ \hline Q_{\nu+1,1} & Q_{\nu+1,2} & \cdots & Q_{\nu+1,\nu} & Q_{\nu+1,\nu+1} \end{array}\right], \tag{27}$$

where $Q_{11},\ldots,Q_{\nu\nu}$ are submatrices associated with each ergodic set $E\mu$ ($\mu = 1,\ldots,\nu$), respectively, and the remaining states specify a set T of transient states.

Next two lemmas are useful to eliminate redundant constraints in (24) and (26). In the discussion of these lemmas (containing Lemma 6) we have to restrict to any fixed nonrandomized policy since we cannot consider simultaneously state classification for randomized policies.

LEMMA 4. *Take any $f\in F$. For any $\ell\in E\mu$ ($\mu = 1,\ldots,\nu$), constraints* (26) *become*

$$\sum_{l\in E_\mu} x_l^k + \sum_{l\in E_\mu}\sum_{j\in T}(\delta_{jl}-p_{jl}^k)\,y_j^k = \sum_{l\in E_\mu} a_l \qquad (\mu = 1,2,\ldots,\nu). \tag{28}$$

PROOF. Using Lemma 1(b), we combine constraints (26) by summing on E_μ. While,

$$\sum_{j\in S} = \sum_{j\in E_\mu} + \sum_{\substack{\nu=1\\ \nu\neq\mu}}^{\nu}\sum_{j\in E_\nu} + \sum_{j\in T} \qquad \text{and} \qquad \sum_{l\in E_\mu}(\delta_{jl}-p_{jl}^k) = 0 \text{ for } j\notin T$$

(from (27)) imply (28).

LEMMA 5. *For any ergodic set* E_μ $(\mu = 1,..., \nu)$, *one of constraints* (24) *associated with* E_μ *is redundant.*

PROOF. It is obvious from the property of Markov chains and the fact that $Q_{\nu\nu}$ $(\nu = 1,..., \nu)$ is a Markov matrix.

The above two lemmas give the necessary constraints for any ergodic set. Next lemma also gives a constraint for any transient state.

LEMMA 6. *Take any fixed* $f \in F$. *For any state* $l \in T$, *a constraint* (26) *becomes*

$$\sum_{j \in T} (\delta_{jl} - p_{jl}^k) y_j^k = a_l \qquad (l \in T), \tag{29}$$

where

$$y_l^k \geqslant 0 \qquad (l \in T, k = f(l)). \tag{30}$$

PROOF. From $[q_{ij}^*(f)]_{i \in S, j \in T} = [0]_{i \in S, j \in T}$ and (17), we have $x_l^k = 0$ for any $l \in T$ and $k = f(l)$. Thus we have (29). While, from

$$[h_{ij}(f)]_{i,j \in T} = [I - Q_{\nu+1,\nu+1}]^{-1} = \sum_{n=0}^{\infty} Q_{\nu+1,\nu+1}^n \geqslant 0,$$

$$[h_{ij}(f)]_{i \notin T, j \in T} = \left[\sum_{n=0}^{\infty} (Q^n(f) - Q^*(f)) \right]_{i \notin T, j \in T} = [0]_{i \notin T, j \in T}$$

and (18), we have

$$y_l^k = \sum_{i \in S} a_i h_{il}(f)\, d_i^k \geqslant 0 \qquad (l \in T, k = f(l)).$$

From (30), we suppose that $y_j^k \geqslant 0$ for any $j \in S$, $k \in K_j$ since y_j^k disappears for any ergodic state j.

Thus we also consider a dual problem of the L.P. problem (23), (24), (25), (26), and (30). Let $N \times 1$ column vectors $u(f)$ and $v(f)$ be the corresponding dual variables. Then its dual problem is:

Maximize

$$\sum_{i \in S} a_i u_i(f) \tag{31}$$

subject to

$$u_i(f) \geqslant \sum_{j \in S} p_{ij}^k u_j(f) \qquad (i \in S, k \in K_i), \tag{32}$$

$$u_i(f) + v_i(f) \geqslant r_i^k + \sum_{j \in S} p_{ij}^k v_j(f) \qquad (i \in S, k \in K_i), \tag{33}$$

$$u_i(f), v_i(f); \qquad \text{unconstrained in sign } (i \in S), \tag{34}$$

where $u_i(f)$, $v_i(f)$ is ith element of $u(f)$, $v(f)$, respectively. We know that this dual problem corresponds to P.I.A., i.e., this dual problem is immediately derived from P.I.A. We also note that the dual variables $u(f)$ and $v(f)$ are unique solutions of

$$u(f) = Q(f)\,u(f), \qquad u(f) + v(f) = r(f) + Q(f)\,v(f), \tag{35}$$

where from (4) we set the value of one $v_i(f)$ in each ergodic set to zero, which refers to (28). Then $v(f)$ is a relative solution and the difference between the exact solution of (12) and $v(f)$ in (35) is a constant.

Now we have an L.P. Algorithm for a general M.D.P. The algorithm is made by using Lemmas 1, 4, 5, and 6. Further we note that the dual variables $(u(f), v(f))$ are also simplex multipliers. Using these simplex multipliers, we have the simplex criterion, which corresponds to Policy Improvement Routine in P.I.A. The direct proof of increasing the average return by its simplex criterion without L.P. properties has been given by Howard [2] and Veinott [4, pp. 1291-1294]. It is convenient to consider the following set of actions which corresponds to the simplex criterion of the primal problem using simplex multipliers $(u(f), v(f))$:

$$G(i,f) = \Big\{ k \in K_i \mid \sum_{j \in S} p_{ij}^k u_j(f) > u_i(f), \text{ or } \sum_{j \in S} p_{ij}^k u_j(f) = u_i(f) \text{ and } r_i^k + \sum_{j \in S} p_{ij}^k v_j(f) > u_i(f) + v_i(f) \Big\}. \tag{36}$$

Then we have the following proposition of describing the linear programming algorithm without proof.

PROPOSITION 7. *Taking any $f \in F$ and determining the constraint for each state according to the state classification (using Lemmas 4, 5, 6), we can obtain a basic feasible solution which corresponds to a policy f and gives its dual variables $u(f)$ and $v(f)$ (simplex multipliers). Using these simplex multipliers, we have the simplex criterion. That is, $G(i, f)$ is empty for all $i \in S$, then we have an optimal stationary policy. Otherwise, select a new policy g such that $g(i) \in G(i, f)$ and $g(i) = f(i) \notin G(i, f)$. Returning to the first part of this proposition, repeat until an optimal policy is obtained.*

Note that Proposition 7 describes a special structure L.P. algorithm such that pivot operations for many variables are performed simultaneously if there are two or more nonempty sets $G(i, f)$.

Proposition 7 and primal and dual problems imply the following corollary.

COROLLARY 8. P.I.A. *is equivalent to the primal* L.P. *Algorithm.*

PROOF. Finding a basic feasible solution (i.e., dual variables) corresponds to Value Determination Operation, and the simplex criterion of the next step corresponds to Policy Improvement Routine. But in P.I.A. pivot operations are performed simultaneously for many variables. The fact that the objective in L.P. increases for such a new policy is given by Howard [2] and Veinott [4]. Veinott also showed that cycling does not occur. These facts imply the result.

Next corollary is clear if we consider P.I.A.

COROLLARY 9. *An optimal policy is independent of the initial distribution a.*

4. AN EXAMPLE

In practical situations we encounter the problem of a general M.D.P. because state classification is not clear, and/or changes for each policy. Here we shall solve an L.P. problem and its dual of an M.D.P. for Howard's example [2, p. 65]. The data of the problem is given in [2, p. 65]. Let an initial policy denote by f, its associated Markov matrix by $Q(f)$ and the return vector by $r(f)$. Then

$$f = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \qquad Q(f) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \qquad r(f) = \begin{bmatrix} 3 \\ 4 \\ 8 \end{bmatrix}.$$

Primal and dual problems are written by the following tableau (reduced Tucker Diagram), where only the data associated with a basic feasible solution are given. And the superscripts of its chosen policy are omitted.

	x_1	x_2	x_3		
v_1	1	0	-1	$=0$	
u_2	0	1	0	$=a_2$	$(v_2=0)$
$u_1=u_3$	1	0	1	$=a_1+a_3$	$(v_3=0)$
	VI	VI	VI		
	3	4	8		

Dual Variables

$$u(f) = \begin{bmatrix} 11/2 \\ 4 \\ 11/2 \end{bmatrix}, \qquad v(f) = \begin{bmatrix} -5/2 \\ 0 \\ 0 \end{bmatrix}.$$

Using the simplex criterion (or equivalently Policy Improvement Routine), we have a next improved policy g and its data.

$$g = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}, \qquad Q(g) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad r(g) = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}.$$

Then

$$\begin{array}{rc} & \begin{array}{ccc} y_1 & y_2 & x_3 \end{array} \\ \begin{array}{r} v_1 \\ v_2 \\ u_1 = u_2 = u_3 \end{array} & \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{array}\right] \begin{array}{l} = a_1 \\ = a_2 \\ = a_3 \end{array} \qquad (v_3 = 0) \\ & \begin{array}{ccc} \vee & \vee & \vee \\ 3 & 5 & 7 \end{array} \end{array}$$

Dual Variables

$$u(g) = \begin{bmatrix} 7 \\ 7 \\ 7 \end{bmatrix}, \qquad v(g) = \begin{bmatrix} -4 \\ -2 \\ 0 \end{bmatrix}.$$

Using the simplex criterion, we have an optimal policy g since $G(i, g)$ is empty for any $i \in S$.

5. Special Cases

In the preceding sections we have discussed an L.P. solution for a general M.D.P. In this section we restrict our attention to special structure problems, e.g., completely ergodic, terminating, or other special processes.

First, we consider a completely ergodic process in the sense of Howard [2, p. 32]. In this case Markov chains under consideration are always ergodic whatever policies we choose, so the L.P. problem is the following form:

Maximize

$$\sum_{j \in S} \sum_{k \in K_j} r_j^k x_j^k \tag{37}$$

subject to

$$\sum_{j \in S} \sum_{k \in K_j} (\delta_{jl} - p_{jl}^k)\, x_j^k = 0 \qquad (l = 1, \ldots, N-1), \tag{38}$$

$$\sum_{j \in S} \sum_{k \in K_j} x_j^k = 1, \tag{39}$$

$$x_j^k \geqslant 0 \qquad (j \in S, k \in K_j), \tag{40}$$

where a redundant constraint is eliminated (Lemma 5). Using the simplex

multipliers or dual variables ($v(f)$ is a relative value), we have the similar set of simplex criterion

$$G(i,f) = \left\{k \in K_i \mid r_i^k + \sum_{j \in S} p_{ij}^k v_j(f) > u_i(f) + v_i(f)\right\}, \tag{41}$$

because $u(f)$ is the column vector with all elements identical and $Q(g)\,u(f) = u(f)$ for any $g \in F$. While, it is clear from the properties of basic feasible solutions that

$$r_i^{f(i)} + \sum_{j \in S} p_{ij}^{f(i)} v_j(f) = u_i(f) + v_i(f),$$

where $u_i(f) = u_j(f)$ for any i, j. Thus we may write

$$G(i,f) = \left\{k \in K_i \mid r_i^k + \sum_{j \in S} p_{ij}^k v_j(f) > r_i^{f(i)} + \sum_{j \in S} p_{ij}^{f(i)} v_j(f)\right\}, \tag{42}$$

which corresponds to Policy Improvement Routine in [2, p. 38].

Second, we consider a terminating process [10] in which state 1 is absorbing and other states are transient whatever policies we choose. And we consider the (finite) total expected return before absorption as a criterion. So, we consider the following quantity

$$\left[a \sum_{n=0}^{\infty} Q^n(f)\, r(f)\right]_{i,j \in T} = \sum_{i=2}^{N} \sum_{j=2}^{N} a_i h_{ij}(f)\, r_j(f) = \sum_{j=2}^{N} \sum_{k \in K_j} r_j^k y_j^k, \tag{43}$$

where y_j^k is defined in (18). Thus we have the following L.P. problem:

Maximize

$$\sum_{j=2}^{N} \sum_{k \in K_j} r_j^k y_j^k \tag{44}$$

subject to

$$\sum_{j=2}^{N} \sum_{k \in K_j} (\delta_{jl} - p_{jl}^k)\, y_j^k = a_l \qquad (l = 2,\ldots, N), \tag{45}$$

$$y_j^k \geqslant 0 \qquad (j = 2,\ldots, N;\ k \in K_j), \tag{46}$$

where a redundant constraint for $\ell = 1$ is eliminated. And the set of simplex criterion for a terminating process is

$$G(i,f) = \left\{k \in K_i \mid r_i^k + \sum_{j=2}^{N} p_{ij}^k v_j(f) > v_i(f)\right\} \qquad (i = 2,\ldots, N). \tag{47}$$

Generally, we consider also a special process in which every state is the same structure of state classification whatever policies we choose. For this process we have a special L.P. problem derived from a general L.P. problem. And the similar discussion is made for this problem.

6. Extension to Semi-Markovian Decision Processes

In this section we extend the preceding discussions of M.D.P. to semi-M.D.P. with no discounting. Following the notation of Osaki and Mine [11], we have a following L.P. problem for a general semi-M.D.P. under the average criterion:

Maximize

$$\sum_{j\in S}\sum_{k\in K_j} r_j^k x_j^k \tag{48}$$

subject to

$$\sum_{j\in S}\sum_{k\in K_j} (\delta_{jl} - p_{jl}^k)\, x_j^k = 0 \qquad (l \in S), \tag{49}$$

$$x_j^k \geqslant 0 \qquad (j \in S, k \in K_j), \tag{50}$$

$$\sum_{k\in K_l} \eta_l^k x_l^k + \sum_{j\in S}\sum_{k\in K_j} (\delta_{jl} - p_{jl}^k)\, y_j^k = a_l \qquad (l \in S), \tag{51}$$

$$y_j^k \geqslant 0 \qquad (j \in S, k \in K_j). \tag{52}$$

Note that the constraints of this L.P. problem have N redundant constraints, and according to state classification these constraints are eliminated and combined. Also note that primal variables x_j^k, y_j^k have N positive values for any basic feasible solution as we have stated in Section 3.

Special cases discussed in Section 5 are straightforward. These special cases have been given by Osaki and Mine [11].

References

1. R. Bellman. A Markovian decision process. *J. Math. Mech.* **6** (1957), 679–684.
2. R. A. Howard. "Dynamic Programming and Markov Processes". M.I.T. Press, Cambridge, Massachusetts, 1960.
3. D. Blackwell. Discrete dynamic programming. *Ann. Math. Statist.* **33** (1962), 719–726.
4. A. F. Veinott. On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.* **37** (1966), 1284–1294.

5. A. S. MANNE. Linear programming and sequential decisions. *Mangt. Sci.* **6** (1960), 259–267.
6. P. WOLFE AND G. B. DANTZIG. Linear programming in a Markov chain. *Opns Res.* **10** (1962), 702–710.
7. C. DERMAN. On sequential decisions and Markov chains. *Mangt. Sci.* **9** (1962), 16–24.
8. F. D'EPENOUX. A probabilistic production and inventory problem. *Mangt. Sci.* **10** (1963), 98–108.
9. G. T. DE GHELLINCK AND G. D. EPPEN. Linear programming solutions for separable Markovian decision problems. *Mangt. Sci.* **13** (1967), 371–394.
10. S. OSAKI AND H. MINE. Some remarks on a Markovian decision problem with an absorbing state. *J. Math. Anal. Appl.* **23** (1968), 327–333.
11. S. OSAKI AND H. MINE. Linear programming algorithms for semi-Markovian decision processes. *J. Math. Anal. Appl.* **22** (1968), 356–381.