MANAGEMENT SCIENCE
Vol. 25, No. 4, April 1979
*Printed in U.S.A.*

# LINEAR PROGRAMMING AND MARKOV DECISION CHAINS*

A. HORDIJK† AND L. C. M. KALLENBERG†

In this paper we show that for a finite Markov decision process an average optimal policy can be found by solving only one linear programming problem. Also the relation between the set of feasible solutions of the linear program and the set of stationary policies is analyzed. (DYNAMIC PROGRAMMING–MARKOV, FINITE STAGE; PROGRAMMING–INFINITE HORIZON)

## 1. Introduction

The use of linear programs to solve dynamic programming problems was introducted by d'Epenoux [6] for the discounted case.

De Ghellinck [7] as well as Manne [11] obtained linear programming formulations for the average reward criterion in the unichain case. The first analysis of linear programs for the multichain case is due to Denardo and Fox [3], [4]. Derman [5] streamlined and slightly improved their results. He shows that, in order to find an optimal policy there have to be solved, in the worst case, two linear programming problems and one search problem. Our interest was raised by Derman's comment [5, p. 84]: "No satisfactory treatment of the dual problem for the multiple class case has been published". We proved that the search procedure can be skipped (by taking arbitrary actions if the corresponding variable is basic with positive value), and that the second linear programming problem can be replaced by a "good" (i.e. polynomial-bounded number of computations) search procedure. Then we found in the paper of Denardo and Fox [4] the personal communication of Bruce Miller that any policy taken as in our Theorem 7 (see §3.2) is an optimal one. One of the authors discussed this point with him in April 1977. At that time Bruce Miller told him that the former statement had to be seen as a conjecture. Inspired by this conjecture, the authors continued their analysis and it turned out to be true indeed. Consequently, in order to find an average optimal policy only one linear programming problem has to be solved.

The second part of the results of this paper is the analysis of the set of feasible solutions of the dual problem.

In §2, we describe the Markov decision model and summarize the main results. Next, in §3, we discuss the average reward criterion. In §3.1 we review some relevant theorems. The main result of §3.2 is that a pure and stationary policy can be obtained from an optimal solution of a linear programming problem by taking in state $i$ an action $a_i$ arbitrarily chosen from the set of actions corresponding to variables which have a positive value in the optimum solution.

Then in §3.3 we study the correspondence between feasible solutions and stationary policies. In contrast with the discounted case it is no longer possible to construct a one-to-one correspondence between the feasible solutions and the randomized stationary policies. As it turns out, we have to use equivalence classes of feasible solutions. We construct a one-to-one correspondence between the stationary policies and the representatives of the equivalence classes. Furthermore, this mapping preserves the optimality property. To be more precise, we prove that optimal solutions

* Accepted by Bennett L. Fox; received September 25, 1978. This paper has been with the authors 3 months for 1 revision.

† University of Leiden, The Netherlands.

0025-1909/79/2504/0352$01.25

are mapped on optimal policies and that optimal policies correspond to representatives which are optimal solutions of the linear program.

Also we prove that pure policies are mapped on extreme points; however, in general, the converse is not true.

Due to space limitations some results are only summarized in this paper. An exhaustive presentation is available in [9].

## 2. Preliminaries

### 2.1 *The Model*

A process is observed at discrete time points $t = 1, 2, \dots$ to be in one of a finite set of possible states. The *state-space* is denoted by $E = \{1, 2, \dots, N\}$.

After observing the state of the process, an *action* must be chosen. Let $A(i)$ denote the set of all possible actions in state $i$. We assume that $A(i)$ has a finite number of elements.

If the system is in state $i$ and action $a \in A(i)$ is chosen, then a *reward* $r_{ia}$ is earned immediately and with probability $p_{iaj}$ the system will be in state $j$ at the next instant.

Let $\{X_t,\ t = 1, 2, \dots\}$ respectively $\{Y_t,\ t = 1, 2, \dots\}$ denote the sequences of observed states respectively chosen actions.

A *decision rule* $\pi^t$ at time $t$ is a function which assigns the probability of taking action $a$ at time $t$; in general, it may depend on all realized states up to and including time $t$ and on all realized actions up to time $t$.

A *policy* $R$ is a sequence of decision rules: $R = (\pi^1, \pi^2, \dots, \pi^t, \dots)$. For a *Markov* policy $R = (\pi^1, \pi^2, \dots)$ we have that $\pi^t$, the decision rule at time $t$, only depends on the state at time $t$. Policy $R = (\pi^1, \pi^2, \dots)$ is a *stationary* policy if all decision rules are identical; we denote the stationary policy $R = (\pi, \pi, \dots)$ also by $\pi^\infty$. A stationary policy is said to be *pure* if its decision rule is nonrandomized. Consequently, a pure and stationary policy is completely described by a mapping $f : E \to \bigcup_{i \in E} A(i)$ such that $f(i) \in A(i)$, $i \in E$. We denote this policy by $f^\infty$.

For any policy $R$ and initial state $i$, we denote by $v_i^\alpha(R)$ respectively $\phi_i(R)$ the total expected discounted reward respectively the average expected reward, where $\alpha$ is the discount factor:

$$v_i^\alpha(R) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_j \sum_a \mathbf{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) \cdot r_{ja},$$

$$\phi_i(R) = \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_j \sum_a \mathbf{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) \cdot r_{ja}.$$

The policy R* is said to be *$\alpha$-discounted optimal* if

$$v_i^\alpha(R^*) = v_i^\alpha, \qquad i \in E, \quad \text{where } v_i^\alpha = \sup_R v_i^\alpha(R).$$

The policy $R^*$ is said to be *average optimal* if

$$\phi_i(R^*) = \phi_i, \qquad i \in E, \quad \text{where } \phi_i = \sup_R \phi_i(R).$$

### 2.2. *Review of the Main Results*

It is well known that a pure and stationary policy which is $\alpha$-discounted optimal can be obtained by solving *one* linear programming problem (e.g. Derman [5]).

In §3.2 we show an analogous result for the average reward case. More precisely, an average optimal pure and stationary policy can be constructed by the following algorithm:

1. Choose, for each $j \in E$, $\beta_j > 0$ such that $\sum_j \beta_j = 1$.

2. Use the simplex method to compute an optimal solution $(x, y)$ of the linear programming problem

$$\begin{array}{llll}
\text{maximize} & \sum_i \sum_a r_{ia} x_{ia} & & \\
\text{subject to} & \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} & = 0, & j \in E. \\
& \sum_a x_{ja} \qquad + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia} & = \beta_j, & j \in E. \\
& x_{ia}, y_{ia} \geqslant 0, \qquad a \in A(i), & & i \in E.
\end{array}$$

(where $\delta_{ij}$ is the Kronecker delta)

3. For each $i \in E$ choose an arbitrary action $a_i$ from the set $\bar{A}(i)$, where

$$\bar{A}(i) = \begin{cases} \{a \mid x_{ia} > 0\}, & i \in E_x = \left\{i \mid \sum_a x_{ia} > 0\right\}. \\ \{a \mid y_{ia} > 0\}, & i \in E \backslash E_x = \left\{i \mid \sum_a x_{ia} = 0\right\}. \end{cases}$$

4. $f^\infty$, where $f(i) = a_i$, $i \in E$, is an average optimal policy.

It is also well known that for the discounted reward case there exists a one-to-one correspondence between the feasible solutions of the linear program and the set of stationary policies.

In §3.3 we derive the following results for the average reward case. For a feasible solution $(x, y)$ we define a corresponding policy $\pi^\infty(x, y)$ by

$$\pi_{ia} = \begin{cases} x_{ia} / \sum_a x_{ia}, & a \in A(i), i \in E_x. \\ y_{ia} / \sum_a y_{ia}, & a \in A(i), i \in E \backslash E_x. \end{cases}$$

We call two feasible solutions $(x^1, y^1)$ and $(x^2, y^2)$ *equivalent* if

$$\pi_{ia}(x^1, y^1) = \pi_{ia}(x^2, y^2) \quad \text{for all } a \in A(i), i \in E.$$

Conversely, let $\pi^\infty$ be a stationary policy. Then we define a corresponding feasible solution by

$$\begin{aligned} x_{ia}(\pi) &= \left[\beta^T P^*(\pi)\right]_i \cdot \pi_{ia}, & a \in A(i), i \in E, \\ y_{ia}(\pi) &= \left[\beta^T D(\pi) + \gamma^T P^*(\pi)\right]_i \cdot \pi_{ia}, & a \in A(i), i \in E, \end{aligned}$$

where $P^*(\pi)$ and $D(\pi)$ are the *stationary* respectively the *deviation matrix* of $P(\pi) = (\sum_a p_{iaj} \pi_{ia})$, and the vector $\gamma$ is defined by formula (6) (see §3.3).

We call $(x(\pi), y(\pi))$ the *representative* of the class of feasible solutions corresponding to policy $\pi^\infty$.

We prove the following properties:

(i) There is a one-to-one correspondence between the stationary policies and the representatives.

(ii) If $\pi^\infty$ is an optimal policy, then $(x(\pi), y(\pi))$ is an optimal solution of the linear program.

(iii) If $(x, y)$ is an optimal solution of the linear program, then the policy $\pi^\infty(x, y)$ is average optimal.

(iv) If $f^\infty$ is a pure and stationary policy, then $(x(f), y(f))$ is an extreme point of the linear program. The converse statement is in general not true.

## 3. Average Expected Reward

### 3.1 *Some Theorems*

THEOREM 1. (BLACKWELL [1]). *There exists a pure and stationary policy $f_0^\infty$ such that $f_0^\infty$ is $\alpha$-discounted optimal for all $\alpha$ near enough to* 1.

For a pure and stationary policy $f^\infty$, let us denote the matrix $(p_{if(i)j})$ by $P(f)$ and the vector $(r_{if(i)})$ by $r(f)$.

THEOREM 2. (KEMENY AND SNELL [10] AND BLACKWELL [1]).
a. *$P^*(f) = \lim_{n\to\infty}(1/n)\sum_{k=1}^n P^{k-1}(f)$ exists and moreover $P(f)P^*(f) = P^*(f)P(f) = P^*(f)P^*(f) = P^*(f)$, $\phi(f^\infty) = P^*(f)r(f)$.*
b. *$D(f) = (I - P(f) + P^*(f))^{-1} - P^*(f)$ exists and $P^*(f)D(f) = 0$.*

THEOREM 3. (BLACKWELL [1]). *For all $i \in E$ and any pure and stationary policy $f^\infty$, we have*

$$\lim_{\alpha\uparrow 1}\left[v_i^\alpha(f^\infty) - \frac{\phi_i(f^\infty)}{1-\alpha}\right] = (D(f)r(f))_i.$$

THEOREM 4. (DERMAN [5]). *The policy $f_0^\infty$ from Theorem 1 is average optimal.*

REMARK. It is also known (e.g. Hordijk [8] and Hordijk and Kallenberg [9]) that the policy $f_0^\infty$ is also average optimal with respect to the stronger criterion

$$\hat{\phi}_i(R) = \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_j\sum_a \mathbf{P}_R(X_t = j, Y_t = a \mid X_1 = i)\cdot r_{ja}, \qquad i \in E.$$

THEOREM 5. *Let $f_0^\infty$ be the policy from Theorem 1, $\phi^0 = \phi(f_0^\infty)$ and $u^0 = D(f_0)r(f_0)$. Then*

$$\phi_i^0 = \max_a \sum_j p_{iaj}\phi_j^0, \qquad i \in E.$$

$$\phi_i^0 + u_i^0 = \max_{a\in A^0(i)}\left\{r_{ia} + \sum_j p_{iaj}u_j^0\right\}, \qquad i \in E.$$

*where $A^0(i) = \{a \in A(i) \mid \phi_i^0 = \sum_j p_{iaj}\phi_j^0\}$, $i \in E$.*

PROOF. From Theorem 1 it follows that there exists a nonnegative real number $\alpha_0 < 1$ such that

$$v_i^\alpha(f_0^\infty) \geqslant r_{ia} + \alpha\sum_j p_{iaj}v_j^\alpha(f_0^\infty), \qquad a \in A(i), i \in E, \alpha \in [\alpha_0, 1).$$

From Theorem 3, we obtain

$$v_i^\alpha(f_0^\infty) = (1-\alpha)^{-1}\phi_i^0 + u_i^0 + \epsilon_i(\alpha), \qquad i \in E, \quad \text{where } \epsilon_i(\alpha) \to 0 \text{ for } \alpha\uparrow 1.$$

Hence,

$$\begin{aligned}
&(1-\alpha)^{-1}\phi_i^0 + u_i^0 + \epsilon_i(\alpha)\\
&\quad\geqslant r_{ia} + \alpha\sum_j p_{iaj}\{(1-\alpha)^{-1}\phi_j^0 + u_j^0 + \epsilon_j(\alpha)\}\\
&\quad= r_{ia} + \{1-(1-\alpha)\}\sum_j p_{iaj}\{(1-\alpha)^{-1}\phi_j^0 + u_j^0 + \epsilon_j(\alpha)\}\\
&\quad= (1-\alpha)^{-1}\sum_j p_{iaj}\phi_j^0 + r_{ia} + \sum_j p_{iaj}u_j^0\\
&\qquad - \sum_j p_{iaj}\phi_j^0 - (1-\alpha)\sum_j p_{iaj}u_j^0 + \alpha\sum_j p_{iaj}\epsilon_j(\alpha).
\end{aligned}$$

Since this inequality holds for all $a \in A(i)$, $i \in E$, $\alpha \in [\alpha_0, 1)$, we have

$$\phi_i^0 \geqslant \sum_j p_{iaj}\phi_j^0, \qquad a \in A(i), i \in E. \tag{1}$$

$$u_i^0 \geqslant r_{ia} + \sum_j p_{iaj}u_j^0 - \phi_i^0, \qquad a \in A^0(i), i \in E. \tag{2}$$

If $a_i = f_0(i)$, $i \in E$, we have equality since

$$\phi^0 = P(f_0)\phi^0 \quad \text{and} \quad (I - P(f_0))u^0 = r(f_0) - \phi^0.$$

Consequently,

$$\phi_i^0 = \max_a \sum_j p_{iaj}\phi_j^0, \qquad i \in E.$$

$$\phi_i^0 + u_i^0 = \max_{a \in A^0(i)} \left\{ r_{ia} + \sum_j p_{iaj}u_j^0 \right\}, \quad i \in E. \quad \text{Q.E.D.}$$

DEFINITION. A pair of functions $(\tilde{\phi}, \tilde{u})$, where $\tilde{\phi} : E \to \mathbf{R}$ and $\tilde{u} : E \to \mathbf{R}$ is *superharmonic* if

$$\tilde{\phi}_i \geqslant \sum_j p_{iaj}\tilde{\phi}_j, \qquad a \in A(i), i \in E.$$

$$\tilde{\phi}_i + \tilde{u}_i \geqslant r_{ia} + \sum_j p_{iaj}\tilde{u}_j, \qquad a \in A(i), i \in E.$$

REMARK. These inequalities have to hold for all actions. Hence, it is a stronger condition than in the assertion of Theorem 5; $(\phi^0, u^0)$ is not necessarily superharmonic.

THEOREM 6. *$\phi$ is the (componentwise) smallest function for which there exists a function $u$ such that $(\phi, u)$ is superharmonic.*

PROOF. From (1) and (2) it follows that there exists a real number $M$ such that $(\phi^0, u^0 + M\phi^0)$ is superharmonic. By Theorem 4, $\phi^0 = \phi$ implying the existence of a function $u$ such that $(\phi, u)$ is superharmonic. Suppose that $(\tilde{\phi}, \tilde{u})$ is also superharmonic. After iterating the definition of superharmonicity it follows that

$$\tilde{\phi} \geqslant P^*(f_0)\tilde{\phi} \geqslant P^*(f_0)r(f_0) + P^*(f_0)(P(f_0) - I)\tilde{u}$$
$$= P^*(f_0)r(f_0) = \phi. \quad \text{Q.E.D.}$$

## 3.2 *Linear Programming*

Since $\phi$ is the smallest function for which there exists a function $u$ such that $(\phi, u)$ is superharmonic, it is plausible to consider the following linear programming problem

$$\begin{aligned} &\text{minimize} && \sum_j \beta_j\tilde{\phi}_j \\ &\text{subject to} && \tilde{\phi}_i \geqslant \sum_j p_{iaj}\tilde{\phi}_j, && a \in A(i), i \in E, \\ &&& \tilde{\phi}_i + \tilde{u}_i \geqslant r_{ia} + \sum_j p_{iaj}\tilde{u}_j, && a \in A(i), i \in E, \end{aligned}$$

where $\beta_j > 0$, $j \in E$, are given numbers with $\sum_j \beta_j = 1$.

From the results in §3.1 we know that $(\phi = \phi(f_0), u = D(f_0)r(f_0) + M\phi(f_0^\infty))$, for $M$ large enough, is an optimal solution.

The dual linear programming problem is

$$\begin{aligned} &\text{maximize} && \sum_i \sum_a r_{ia} x_{ia} && && \\ &\text{subject to} && \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} && = 0, && j \in E. && (3) \\ & && \sum_a x_{ja} + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia} && = \beta_j, && j \in E. && (4) \\ & && x_{ia}, y_{ia} \geqslant 0, \quad a \in A(i), && && i \in E. && (5) \end{aligned}$$

THEOREM 7. *If the simplex method is used to solve the dual problem and an optimal solution* $(x, y)$ *is obtained, then the policy* $f^\infty$, *where*

$$f(i) = a_i \quad \textit{such that} \begin{cases} x_{ia_i} > 0, & i \in E_x = \left\{ i \mid \sum_a x_{ia} > 0 \right\}, \\ y_{ia_i} > 0, & i \notin E_x, \end{cases}$$

*is average optimal.*

REMARK. The above theorem says that an optimal policy is obtained by taking an arbitrary action for which the $x$-variable is positive, if possible; otherwise by taking an arbitrary action for which the $y$-variable is positive. Indeed, it is possible to obtain an optimal solution where in some states there are more than one positive variable (an example can be found in [9]). In that case we can construct different policies. Any of these policies is average optimal.

PROOF. The proof of the theorem is rather long. It has the following structure. After some preliminary statements, mainly to indicate that our rule will always give a policy $f^\infty$, we continue with three separate propositions. After completing the proofs of these propositions we conclude the proof of the theorem by some final conclusions. From (4) and (5) it follows that $\sum_a x_{ja} + \sum_a y_{ja} \geqslant \beta_j > 0, j \in E$. Hence, the policy $f^\infty$ is well defined. Let $(\bar{\phi}, u)$ be an optimal solution of the primal problem. From Theorem 6 it follows that $\bar{\phi} = \phi$.

PROPOSITION 1.

$$\sum_j (\delta_{ij} - p_{ia_ij}) \phi_j = 0, \qquad i \in E.$$

$$\phi_i + \sum_j (\delta_{ij} - p_{ia_ij}) u_j = r_{ia_i}, \qquad i \in E_x.$$

PROOF. Since $x_{ia_i} > 0$ for $i \in E_x$ and $y_{ia_i} > 0$ for $i \notin E_x$, it follows from the complementary slackness property of primal and dual linear programming that

$$\phi_i + \sum_j (\delta_{ij} - p_{ia_ij}) u_j = r_{ia_i}, \qquad i \in E_x.$$

$$\sum_j (\delta_{ij} - p_{ia_ij}) \phi_j = 0, \qquad i \notin E_x.$$

From the primal program we see that $\sum_j (\delta_{ij} - p_{iaj}) \phi_j \geqslant 0$, $a \in A(i)$, $i \in E$. Suppose that $\sum_j (\delta_{kj} - p_{ka_kj}) \phi_j > 0$ for some $k \in E_x$. Since $x_{ka_k} > 0$,

$$\sum_j (\delta_{kj} - p_{ka_kj}) \phi_j x_{ka_k} > 0.$$

Also we have

$$\sum_j (\delta_{ij} - p_{iaj})\phi_j x_{ia} \geqslant 0, \qquad a \in A(i), i \in E.$$

Therefore,

$$\sum_i \sum_a \sum_j (\delta_{ij} - p_{iaj})\phi_j x_{ia} > 0.$$

However, from (3) it follows that

$$\sum_i \sum_a \sum_j (\delta_{ij} - p_{iaj})\phi_j x_{ia} = 0.$$

This contradiction implies that

$$\sum_j (\delta_{ij} - p_{ia_ij})\,\phi_j = 0, \qquad i \in E_x,$$

which completes the proof.

Proposition 2. *$E_x$ is closed, i.e. $p_{ia_ij} = 0$, $i \in E_x$, $j \notin E_x$.*

Proof. Suppose that $p_{ia_ij} > 0$ for some $i \in E_x$, $j \notin E_x$. From (3) it follows that

$$0 = \sum_a x_{ja} = \sum_l \sum_a p_{laj} x_{la} \geqslant p_{ia_ij} x_{ia_i} > 0$$

implying a contradiction.

Proposition 3. *The states of $E \setminus E_x$ are transient in the Markov chain with transition probabilities $p_{ij} = p_{ia_ij}$, $i, j \in E$.*

Proof. Suppose there is a state $j \in E \setminus E_x$ which is nontransient. Since $E_x$ is closed, this implies the existence of a nonempty set $J \subset E \setminus E_x$ which is ergodic. Since $(x, y)$ is an extreme point and $y_{ja_j} > 0$, $j \in J$, we know from the theory of convex polyhedra that the corresponding columns $\{q^j, j \in J\}$, where

$$q_k^j = \begin{cases} 0, & k = 1, 2, \ldots, N, \\ \delta_{jk-N} - p_{ja_jk-N}, & k = N+1, N+2, \ldots, 2N, \end{cases}$$

are linear independent. Let $J = \{j_1, j_2, \ldots, j_m\}$. Since $J$ is an ergodic set we have for $j \in J$, $k - N \notin J$ that $q_k^j = 0$. Hence, the contracted vectors $\{b^1, b^2, \ldots, b^m\}$, where $b_k^l = q_{N+j_k}^{j_l}$, $k, l = 1, 2, \ldots, m$, are also linear independent. However,

$$\sum_{k=1}^m b_k^l = \sum_{k=1}^m q_{N+j_k}^{j_l} = \sum_{k=1}^N q_{N+k}^{j_l} = \sum_{k=1}^N \left(\delta_{j_lk} - p_{j_la_{j_l}k}\right) = 0, \qquad l = 1, 2, \ldots, m,$$

which contradicts the independency.

Now, we can finish the proof of Theorem 7. From Proposition 1, it follows that $P^*(f)\phi = \phi$. Since the states of $E \setminus E_x$ are transient, we have $p_{il}^*(f) = 0$, $l \in E \setminus E_x$. Therefore,

$$\phi(f^\infty) = P^*(f)r(f) = P^*(f)\{\phi + (I - P(f))u\} = P^*(f)\phi = \phi.$$

Hence, $f^\infty$ is an average optimal policy. Q.E.D.

### 3.3 *Relations between Policies and Feasible Solutions*

For a feasible solution $(x, y)$ of the dual problem we define a stationary policy $\pi^\infty(x, y)$ by

$$\pi_{ia}(x,y) = \begin{cases} x_{ia}/\sum_a x_{ia}, & a \in A(i), i \in E_x. \\ y_{ia}/\sum_a y_{ia}, & a \in A(i), i \notin E_x. \end{cases}$$

We call two feasible solutions $(x^1, y^1)$ and $(x^2, y^2)$ *equivalent* if

$$\pi_{ia}(x^1, y^1) = \pi_{ia}(x^2, y^2) \quad \text{for all } a \in A(i), i \in E.$$

This equivalence relation divides the feasible solutions in equivalence classes. Conversely, let $\pi^\infty$ be a stationary policy. Let $P(\pi) = (\sum_a p_{iaj}\pi_{ia})$ and $r(\pi) = (\sum_a r_{ia}\pi_{ia})$. Then, as in §3.1, it holds that

$$P^*(\pi) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P^{k-1}(\pi)$$

exists and $P(\pi)P^*(\pi) = P^*(\pi)P(\pi) = P^*(\pi)P^*(\pi) = P^*(\pi)$. $D(\pi) = (I - P(\pi) + P^*(\pi))^{-1} - P^*(\pi)$ exists and $D(\pi)P^*(\pi) = P^*(\pi)D(\pi) = 0$. $\phi(\pi^\infty) = P^*(\pi)r(\pi)$. Consider the Markov chain induced by the matrix $P(\pi)$. Suppose there are $m$ ergodic sets $E_1, E_2, \ldots, E_m$ and let $T$ be the set of transient states. We define

$$\begin{aligned} x_{ia}(\pi) &= \left[\beta^T P^*(\pi)\right]_i \cdot \pi_{ia}, & a \in A(i), i \in E. \\ y_{ia}(\pi) &= \left[\beta^T D(\pi) + \gamma^T P^*(\pi)\right]_i \cdot \pi_{ia}, & a \in A(i), i \in E. \end{aligned} \tag{6}$$

where

$$\gamma_l = \begin{cases} 0, & l \in T, \\ \max_{i \in E_j} \dfrac{-\sum_k \beta_k d_{ki}(\pi)}{\sum_k p^*_{ki}(\pi)}, & l \in E_j, j = 1, 2, \ldots, m. \end{cases}$$

First, we show that $(x(\pi), y(\pi))$ is a feasible solution.

1. $\sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia}(\pi) = (\beta^T P^*(\pi))_j - (\beta^T P^*(\pi) P(\pi))_j = 0, \qquad j \in E.$

2. $x_{ia}(\pi) \geqslant 0$, $a \in A(i)$, $i \in E$.

3. $\sum_a x_{ja}(\pi) + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia}(\pi)$

$$= (\beta^T P^*(\pi))_j + (\beta^T D(\pi) + \gamma^T P^*(\pi))_j - (\beta^T D(\pi) P(\pi) + \gamma^T P^*(\pi) P(\pi))_j$$

$$= \left[\beta^T \{P^*(\pi) + D(\pi)(I - P(\pi) + P^*(\pi))\}\right]_j = \beta_j, \qquad j \in E.$$

4. $p^*_{ki}(\pi) = 0$, $i \in T$, $k \in E$. Since $p_{ki}(\pi) = p^*_{ki}(\pi) = 0$, $i \in T$, $k \notin T$, we have $d_{ki}(\pi)$

$= 0$, $i \in T$, $k \notin T$. For $i \in T$, $k \in T$ we obtain

$$d_{ki}(\pi) = \left[(I - P(\pi) + P^*(\pi))^{-1}\right]_{ki} - p^*_{ki}(\pi) = \left[(I - P(\pi))^{-1}\right]_{ki}$$

$$= \sum_{j=0}^{\infty} p^j_{ki}(\pi) \geqslant 0.$$

Hence, for $i \in T$:

$$y_{ia}(\pi) = \sum_k \beta_k d_{ki}(\pi) \cdot \pi_{ia} \geqslant 0, \qquad a \in A(i),$$

for $i \in E_j$:

$$y_{ia}(\pi) = \left\{ \sum_k \beta_k d_{ki}(\pi) + \gamma_i \sum_{k \in E_j} p^*_{ki}(\pi) \right\} \cdot \pi_{ia} \geqslant 0, \qquad a \in A(i),$$

Therefore, $y_{ia}(\pi) \geqslant 0$, $a \in A(i)$, $i \in E$.

From the properties 1–4 it follows that $(x(\pi), y(\pi))$ is a feasible solution.

For a stationary policy $\pi^\infty$, let $(X(\pi), Y(\pi))$ be the class of corresponding equivalent feasible solutions. We choose the element $(x(\pi), y(\pi))$ as the *representative* of this equivalence class.

Hence, the mapping defined by (6) is a one-to-one mapping of the stationary policies onto the set of representatives.

THEOREM 8. *The mapping preserves the optimality property, i.e.*

a. *If $\pi^\infty$ is an optimal policy, then $(x(\pi), y(\pi))$ is an optimal solution of the dual linear programming problem.*

b. *If $(x, y)$ is an optimal solution of the dual program, then the policy $\pi^\infty(x, y)$ is average optimal.*

PROOF. a. Since $(\phi = \phi(f_0^\infty)$, $u = D(f_0)r(f_0) + M\phi(f_0^\infty))$, for $M$ large enough, is an optimal solution of the primal problem, it follows from the theory of linear programming (duality theorem) that it is sufficient to show that $\sum_i \sum_a r_{ia} x_{ia}(\pi) = \sum_j \beta_j \phi_j$. We have,

$$\sum_i \sum_a r_{ia} x_{ia}(\pi) = \sum_i \sum_a r_{ia} \big(\beta^T P^*(\pi)\big)_i \cdot \pi_{ia} = \beta^T P^*(\pi) r(\pi)$$

$$= \beta^T \phi(\pi^\infty) = \beta^T \phi.$$

b. If $(x, y)$ is an optimal solution of the dual program, then we can write $(x, y) = \sum_{k=1}^m \lambda_k (x^k, y^k)$, where $\lambda_k > 0$, $k = 1, 2, \ldots, m$, $\sum_k \lambda_k = 1$ and $(x^k, y^k)$ is an extreme optimal point.

Let $(\phi, u)$ be an optimal solution of the primal problem. The proof of this theorem is analogous to the proof of Theorem 7. First, we prove a proposition which is similar to Proposition 1. Then we characterize the set of transient states in the Markov chain with transition matrix $P(\pi(x, y))$. This result looks like Proposition 3 but is different, since, in the Markov chain of Proposition 3, $E_x$ may contain transient states.

PROPOSITION 4.

$$\sum_j (\delta_{ij} - p_{iaj})\phi_j = 0, \qquad a \in \bar{A}(i), i \in E,$$

$$\phi_i + \sum_j (\delta_{ij} - p_{iaj}) u_j = r_{ia}, \qquad a \in \bar{A}(i), i \in E_x,$$

*where $\bar{A}(i) = \{a \mid \pi_{ia}(x, y) > 0\}$, $i \in E$.*

PROOF. If $i \in E_x$ and $a \in \bar{A}(i)$, then $x_{ia} > 0$; if $i \notin E_x$ and $a \in \bar{A}(i)$, then $y_{ia} > 0$. Now we can prove the proposition analogously to Proposition 1.

PROPOSITION 5. *For any feasible solution $(x, y)$ of the dual program $E_x$ is the set of recurrent states in the Markov chain with transition probabilities $p_{ij}(\pi(x, y))$.*

PROOF. For $i \in E$, let $x_i = \sum_a x_{ia}$ and $y_i = \sum_a y_{ia}$. Denote $\pi^\infty(x, y)$ by $\pi^\infty$. Let $T$ be the set of transient states under $P(\pi)$. From (3) it follows that $x^T = x^T P(\pi)$. Since $x_i > 0$, $i \in E_x$, we know from the theory of Markov chains (e.g. [2]) that $T \subset E \backslash E_x$. Similarly to Proposition 2 we can prove that $E_x$ is closed under $P(\pi)$. Suppose that $T \neq E \backslash E_x$. Then we have an ergodic set $E_1 \subset E \backslash E_x$. Hence, it follows from (4) that

$$\begin{aligned} 0 &= \sum_{j \notin E_1} \sum_{i \in E_1} p_{ij}(\pi) = \sum_{j \notin E_1} \sum_{i \in E_1} \sum_a p_{iaj} y_{ia} \\ &= \sum_{j \notin E_1} \sum_i \sum_a p_{iaj} y_{ia} - \sum_{j \notin E_1} \sum_{i \notin E_1} \sum_a p_{iaj} y_{ia} \\ &= \sum_{j \notin E_1} (-\beta_j + x_j + y_j) - \sum_{i \notin E_1} \sum_j \sum_a p_{iaj} y_{ia} \\ &\quad + \sum_{i \notin E_1} \sum_{j \in E_1} \sum_a p_{iaj} y_{ia} \\ &= -\sum_{j \notin E_1} \beta_j + \sum_{j \notin E_1} x_j + \sum_{j \notin E_1} y_j - \sum_{i \notin E_1} y_i \\ &\quad + \sum_{i \notin E_1} \sum_{j \in E_1} \sum_a p_{iaj} y_{ia}. \end{aligned}$$

Since

$$\sum_{i \notin E_1} \sum_{j \in E_1} \sum_a p_{iaj} y_{ia} \geqslant 0$$

and

$$\sum_{j \notin E_1} x_j = \sum_{j \in E_x} x_j = 1 > \sum_{j \notin E_1} \beta_j,$$

the above equation gives a contradiction. Q.E.D.

Now we can finish the proof as follows. From Proposition 4, it follows that $P^*(\pi(x, y))\phi = \phi$. Since $E \backslash E_x$ is the set of transient states, we have also by Proposition 4

$$\begin{aligned} \phi(\pi^\infty(x, y)) &= P^*(\pi(x, y)) r(\pi(x, y)) \\ &= P^*(\pi(x, y))\{\phi + (I - P(\pi(x, y)))u\} \\ &= P^*(\pi(x, y))\phi = \phi. \end{aligned}$$

Hence, $\pi^\infty(x, y)$ is an average optimal policy. Q.E.D.

THEOREM 10. *Let $f^\infty$ be a pure and stationary policy. Then, the corresponding representative $(x(f), y(f))$ is an extreme point of the dual linear programming problem.*

PROOF. Suppose $(x(f), y(f))$ is not an extreme point. Then there exist feasible solutions $(x^1, y^1)$ and $(x^2, y^2)$ such that $x(f) = \lambda x^1 + (1 - \lambda)x^2$ and $y(f) = \lambda y^1 + (1 - \lambda)y^2$ for some $0 < \lambda < 1$. Since $x_{ia}(f) = y_{ia}(f) = 0$, $a \neq a_i$, where $a_i = f(i)$, $i \in E$, we have

$$x_{ia}^1 = x_{ia}^2 = y_{ia}^1 = y_{ia}^2 = 0, \qquad a \neq a_i, i \in E.$$

Let $P = (p_{ia_ij})$, $\bar{x}(f) = (x_{ia_i}(f))$, $\bar{y}(f) = (y_{ia_i}(f))$, $\bar{x}^1 = (x^1_{ia_i})$, $\bar{x}^2 = (x^2_{ia_i})$, $\bar{y}^1 = (y^1_{ia_i})$ and $\bar{y}^2 = (y_{ia_i})$. Then $(\bar{x}(f), \bar{y}(f))$, $(\bar{x}^1, \bar{y}^1)$, $(\bar{x}^2, \bar{y}^2)$ are solutions of the system.

$$\begin{aligned} x^T(I-P) \qquad\qquad &= 0, \\ x^T \qquad + y^T(I-P) &= \beta^T. \end{aligned} \tag{7}$$

Hence,

$$x^T = x^TP = x^TP^2 = \dots = x^TP^* = \beta^TP^* - y^T(I-P)P^* = \beta^TP^*.$$

$$y^T(I - P + P^*) = \beta^T - x^T + y^TP^* = \beta^T(I - P^*) + y^TP^*.$$

So $x^T = \beta^TP^*$ and $y^T = \beta^TD + y^TP^*$. Therefore $\bar{x}(f) = \bar{x}^1 = \bar{x}^2$. Consider the Markov chain induced by the matrix $P$. Suppose there are $m$ ergodic sets $E_1, E_2, \dots, E_m$ and let $T$ be the set of transient states. Then $y_i = (\beta^TD)_i$, $i \in T$. By definition of $\gamma$, there is on each $E_k$ a state $i(k)$ such that $\bar{y}_{i(k)}(f) = 0$; then also $\bar{y}^1_{i(k)} = \bar{y}^2_{i(k)} = 0$. Since $(\bar{x}^1, \bar{y}^1)$ and $(\bar{x}^2, \bar{y}^2)$ are solutions of system (7), and $\bar{y}^1_i - \bar{y}^2_i = 0$, $i \in T$, we have for $k = 1, 2, \dots, m$

$$\bar{y}^1_i - \bar{y}^2_i = \sum_{l \in E_k} (\bar{y}^1_l - \bar{y}^2_l)p_{li}, \qquad i \in E_k,$$

$$\bar{y}^1_{i(k)} - \bar{y}^2_{i(k)} = 0.$$

Then it follows from the theory of Markov chains (e.g. Chung [2, p. 33]) that $\bar{y}^1 = \bar{y}^2$. Hence $(x^1, y^1) = (x^2, y^2) = (x(f), y(f))$, implying that $(x(f), y(f))$ is an extreme point. Q.E.D.

REMARKS. 1. In [9] we have an example which shows that if $(x, y)$ is an extreme point of the dual program, then in general $\pi^\infty(x, y)$ is not pure.

2. In the unichain case (i.e. if there is only one ergodic set $E_1$) $X(\pi)$ has one element, namely $x(\pi)$. Then any $(x, y)$, where $x = x(\pi)$ and $y \in Y^*(\pi) = \{y \mid y_{ia} = [\beta^TD(\pi) + c^TP^*(\pi)]_i \cdot \pi_{ia}$ for some $c \geqslant \gamma\}$, is a feasible solution. However, in general, $Y^*(\pi) \neq Y(\pi)$. Hence, the class $(X(\pi), Y(\pi))$ is even in the unichain case not easy to characterize. In the multichain case $X(\pi)$ can consist of more than one element. Examples showing the above statements can also be found in [9].

## References

1. BLACKWELL, D., "Discrete Dynamic Programming," *Ann. Math. Statist.*, Vol. 33 (1962), pp. 719–726.
2. CHUNG, K. L., *Markov Chains with Stationary Transition Probabilities*, Springer, Berlin, 1960.
3. DENARDO, E. V., "On Linear Programming in a Markov Decision Problem," *Management Sci.*, Vol. 16 (1970), pp. 281–288.
4. ——— AND FOX, B. L., "Multichain Markov Renewal Programs," *SIAM J. Appl. Math.*, Vol. 16 (1968), pp. 468–487.
5. DERMAN, C., *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.
6. D'EPENOUX, F., "Sur un Problème de Production et de Stockage dans l'Aléatoire," *Revue Français de Recherche Opérationelle*, Vol. 14 (1960), pp. 3–16.
7. DE GHELLINCK, G. T., "Les Problèmes de Décisions Sequentielles," *Cahiers du Centre d'Etudes de Recherche Opérationelle*, Vol. 2 (1960), pp. 161–179.
8. HORDIJK, A., *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tract 51, Amsterdam, 1974.
9. ——— AND KALLENBERG, L. C. M., *Linear Programming and Markov Decision Chains* I, II, Reports No. 78-1 and 78-5, Institute of Applied Mathematics and Computer Science, University of Leiden, The Netherlands, 1978.
10. KEMENY, J. G. AND SNELL, J. L., *Finite Markov Chains*, Van Nostrand, New York, 1960.
11. MANNE, A. S., "Linear Programming and Sequential Decisions," *Management Sci.*, Vol. 6 (1960), pp. 259–267.