MATHEMATICS OF OPERATIONS RESEARCH
Vol. 3, No. 4, November 1978
Printed in U.S.A.

THE FUNCTIONAL EQUATIONS OF UNDISCOUNTED MARKOV RENEWAL PROGRAMMING*†

P. J. SCHWEITZER** AND A. FEDERGRUEN***

This paper investigates the solutions to the functional equations that arise inter alia in Undiscounted Markov Renewal Programming. We show that the solution set is a connected, though possibily nonconvex set whose members are unique up to n^* constants, characterize n^* and show that some of these n^* degrees of freedom are locally rather than globally independent.

Our results generalize those obtained in Romanovsky [20] where another approach is followed for a special class of discrete time Markov Decision Processes. Basically our methods involve the set of randomized policies. We first study the sets of pure and randomized maximal-gain policies, as well as the set of states that are recurrent under some maximal-gain policy.

I. Introduction. This paper investigates the solution (g, v) to the $2N$ functional equations:

$$g_i = \max_{k \in K(i)} \sum_{j=1}^{N} P_{ij}^k g_j, \qquad i = 1, \ldots, N, \tag{1.1}$$

$$v_i = \max_{k \in L(i)} \left[q_i^k - \sum_{j=1}^{N} H_{ij}^k g_j + \sum_{j=1}^{N} P_{ij}^k v_j \right], \qquad i = 1, \ldots, N, \tag{1.2}$$

where

$$L(i) = \left\{ k \in K(i) \mid g_i = \sum_{j=1}^{N} P_{ij}^k g_j \right\}. \tag{1.3}$$

The $K(i)$ are given finite sets and the q_i^k, P_{ij}^k, H_{ij}^k are given arrays with $P_{ij}^k, H_{ij}^k \geqslant 0$ for all i, j, k; $\sum_{j=1}^N P_{ij}^k = 1$, for all i, k. Also we assume property A to be stated below.

For the special cases $H_{ij}^k = P_{ij}^k \cdot \tau_{ij}^k$ with $\tau_{ij}^k \geqslant 0$ and $H_{ij}^k = \delta_{ij}$, the functional equations arise in Markov Decision Theory with $\Omega = \{1, \ldots, N\}$ as state space, q_i^k as the one-step expected reward, P_{ij}^k the transition probability to state j and $T_i^k = \sum_j H_{ij}^k$ the expected holding time, when alternative k is chosen in state i (cf. Bellman [2], [3], Blackwell [4], Howard [11], [12], De Cani [6], Jewell [13], Denardo and Fox [8], Denardo [7], Derman [9], Schweitzer [21], [22], [23]). The solution to (1.1) and (1.2) is not unique, although g is uniquely determined. The purpose of this paper is to characterize

$$V = \{ v \in E^N \mid v \text{ satisfies } (1.2) \}.$$

We show that V is a connected, though possibly nonconvex, set whose members are unique up to n^* constants, characterize n^*, and show that some of these n^* degrees of freedom are locally rather than globally independent.

* Received February 2, 1976; revised February 24, 1978.
AMS 1970 subject classification. Primary 90C45. Secondary 90C40.
IAOR 1973 subject classification. Main: Markov decision programming. Cross reference: Dynamic programming.
Key words. Markov Renewal Programs, average return optimality, functional equations, fixed points.
† This paper has been registered as Math. Center reports BW 60/76 and 71/77.
** University of Rochester.
*** Mathematisch Centrum.

Our results generalize those obtained in Romanovsky [20] where another approach is followed for a special class of discrete time Markov Decision Processes (MDP's).

Basically our methods involve the set of randomized policies. We first study the sets S_{PMG} and S_{RMG} of pure and randomized maximal-gain policies, and characterize the set R^* of states that are recurrent under some maximal gain policy. In §2 we give the notation and some preliminaries. In §3 we characterize the sets S_{RMG} and R^*. The properties of V are studied in §4, while in §5 the n^* degrees of freedom are characterized.

II. Notation and preliminaries. A *(stationary) randomized policy* f is a tableau $[f_{ik}]$ satisfying $f_{ik} \geqslant 0$ and $\sum_{k \in K(i)} f_{ik} = 1$ for all $i \in \Omega$. In the Markov decision model, f_{ik} denotes the probability that the kth alternative is chosen when entering state i.

We let S_R denote the set of all randomized policies and S_P the subset of all *pure* (nonrandomized) policies, i.e. for $f \in S_P$, each $f_{ik} = 0$ or 1. For $f \in S_P$, we use the notation $f^{\#} = (\beta_1, \ldots, \beta_N)$ where $\beta_i \in K(i)$ denotes the single alternative used in state i.

Associated with each $f \in S_R$ are N-component "reward" vector $q(f)$ and "holding time" vector $T(f)$, and two matrices $P(f)$ and $H(f)$:

$$q(f)_i = \sum_{k \in K(i)} f_{ik} q_i^k; \qquad T(f)_i = \sum_{k \in K(i)} f_{ik} T_i^k;$$

$$P(f)_{ij} = \sum_{k \in K(i)} f_{ik} P_{ij}^k; \qquad H(f)_{ij} = \sum_{k \in K(i)} f_{ik} H_{ij}^k.$$

Note that $P(f)$ is a stochastic matrix. For any $f \in S_R$, define the stochastic matrix $\Pi(f)$ as the Cesaro limit of the sequence $\{P(f)^n\}_{n=1}^{\infty}$ and define the *fundamental* matrix $Z(f)$ as $[I - P(f) + \Pi(f)]^{-1}$. These matrices always exist and have the following properties (cf. [4], [14]):

$$\Pi(f) = P(f)\Pi(f) = \Pi(f)P(f) = \Pi(f)^2 = \Pi(f)Z(f) = Z(f)\Pi(f), \tag{2.1}$$

$$[I - P(f)]Z(f) = Z(f)[I - P(f)] = I - \Pi(f), \tag{2.2}$$

$$Z(f) = I + \lim_{a \uparrow 1} \sum_{n=1}^{\infty} a^n [P(f)^n - \Pi(f)]. \tag{2.3}$$

Denote by $n(f)$ the number of subchains (closed, irreducible sets of states) for $P(f)$. Then:

$$\Pi(f)_{ij} = \sum_{m=1}^{n(f)} \phi_i^m(f) \pi_j^m(f), \qquad 1 \leqslant i, j \leqslant N, \tag{2.4}$$

where the row vector $\pi^m(f)$ is the unique equilibrium distribution of $P(f)$ on the mth subchain $C^m(f)$, and $\phi_i^m(f)$ is the probability of absorption in $C^m(f)$, starting from state i (cf. [7] and [23]). Observe $\sum_i \pi_i^m(f) = 1$ and $\pi^m(f)P(f) = \pi^m(f)$.

Let $R(f) = \{j \mid \Pi(f)_{jj} > 0\}$, i.e. $R(f)$ is the set of recurrent states for $P(f)$. Note that the column vector $\phi^m(f) = P(f)\phi^m(f)$ for all m and that the $\{\phi^m(f) \mid m = 1, \ldots, n(f)\}$ are linearly independent. Since any solution to $P(f)x = x$ satisfies $\Pi(f)x = x$ and the rank of $[I - \Pi(f)]$ is $N - n(f)$, it easily follows that the solution set of $P(f)x = x$ is given by:

$$x = \sum_{m=1}^{n(f)} a_m \phi^m(f) \tag{2.5}$$

with $a_1, \ldots, a_{n(f)}$ arbitrary scalars.

LEMMA 2.1. *Fix $f \in S_R$. Suppose $\Pi(f)b = 0$ and $(I - P(f))x - b = y \geq 0$. Then $(I - \Pi(f))x - Z(f)b = z \geq 0$. Also $\Pi(f)y = \Pi(f)z = 0$, i.e. in both inequalities the equality sign holds for each component $i \in R(f)$.*

PROOF. Multiplying $[I - P(f)]x - b \geq 0$ by $\Pi(f) \geq 0$ yields $0 = \Pi(f)([I - P(f)]x - b)$, implying that the former inequality is a strict equality for components $i \in R(f)$. Using this and the fact that as a result of (2.3), for $j \notin R(f)$, $Z(f)_{ij} \geq 0$ for all i, with $Z(f)_{ij} = 0$ when $i \in R(f)$, we get the desired result by multiplying $[I - P(f)]x \geq b$ by $Z(f)$ and invoking (2.2). ∎

LEMMA 2.2. *For any $f \in S_R$, any $i \in R(f)$ and any k having $f_{ik} > 0$, there exists a pure policy h that has the properties:* (a) $h_{ik} = 1$; (b) $h_{jr} = 0$ *whenever* $f_{jr} = 0$; (c) $i \in R(h)$ *and* (d) *every subchain of $P(h)$ is contained within a subchain of $P(f)$.*

PROOF. Let h meet conditions (a) and (b), and assume i is contained within the subchain C of $P(f)$. In view of (b), we have that every subchain of $P(f)$ is closed under $P(h)$ as well, so that no subchain of $P(h)$ can intersect two subchains of $P(f)$; and as a consequence the proof of part (d) reduces to showing that $R(h) \subseteq R(f)$. The latter trivially holds if $\Omega = R(f)$. Otherwise, let Γ initially be equal to $R(f)$ and define $\bar{\Gamma} = \Omega \backslash \Gamma$. Choose a state $t_0 \in \bar{\Gamma}$ and a path $\{t_0, \ldots, t_n\}$ such that $P(f)_{t_l t_{l+1}} > 0$ for $l = 0, \ldots, n-1$ and $t_n \in \Gamma$. Such a path clearly exists, since t_0 is transient under $P(f)$ and $\Gamma \supseteq R(f)$. Transfer $\{t_0, \ldots, t_{n-1}\}$ from $\bar{\Gamma}$ to Γ and define for $l = 0, \ldots, n-1$ $h_{t_l r} = 1$ for any r with $f_{t_l r} > 0$ and $P^r_{t_l t_{l+1}} > 0$. Repeat this step until $\bar{\Gamma}$ is empty. Finally, to ensure property (c), let Δ initially be equal to $\{i\}$ and define $\bar{\Delta} = C \backslash \Delta$. Next the following step is performed: Choose a state $j \in \bar{\Delta}$ and an alternative r such that $f_{jr} > 0$ and $P^r_{jt} > 0$ for some $t \in \Delta$, transfer j from $\bar{\Delta}$ to Δ, and define $h_{jr} = 1$. Clearly, such a j and r can be found, since all states in C communicate under $P(f)$. Repeat this step for the new Δ and $\bar{\Delta}$, until $\bar{\Delta}$ is empty. This construction shows that under policy h, state i can be reached from any state in $C \backslash \{i\}$. Together this and the fact that C is closed under $P(h)$ implies condition (c). ∎

In the remainder of this paper, we assume that property A holds.

A: If f is any pure policy and $C^m(f)$ is any subchain of $P(f)$, then $i \in C^m(f)$ implies $H(f)_{ij} = 0$ for $j \notin C^m(f)$, and $\sum_{i \in C^m(f)} T(f)_i > 0$.

This property is satisfied for both the Markov Renewal Programs (MRP's) with $H^k_{ij} = P^k_{ij}\tau^k_{ij}$ and the discrete time model with $H^k_{ij} = \delta_{ij}$. Using the previous lemma, one easily verifies that if property A holds for all pure policies, it holds for all randomized policies as well.

LEMMA 2.3. (*Gain and Relative Value Vectors*). *Fix $f \in S_R$. The general solution to the equations*

$$\text{(a)} \quad g = P(f)g, \qquad \text{(b)} \quad v = q(f) - H(f)g + P(f)v \tag{2.6}$$

is given by

$$g_i = g(f)_i = \sum_{m=1}^{n(f)} \phi_i^m(f) g^m(f), \qquad i = 1, \ldots, N, \tag{2.7}$$

with

$$g^m(f) = \langle \pi^m(f), q(f) \rangle / \langle \pi^m(f), T(f) \rangle$$

and

$$v_i = Z(f)[q(f) - H(f)g]_i + \sum_{m=1}^{n(f)} a_m \phi_i^m(f), \qquad i = 1, \ldots, N, \tag{2.8}$$

with $a_1, \ldots, a_{n(f)}$ arbitrary scalars.

PROOF. Note that multiplication of (2.6)(b) by $\Pi(f)$ leads to:

$$\Pi(f)[q(f) - H(f)g] = 0. \tag{2.9}$$

Using property A, it follows from the proof of lemma 1 of [7] that $g(f)$ is the unique solution to (2.6)(a) and (2.9). Hence, any solution (g, v) to (2.6) has $g = g(f)$. Using (2.2) one next verifies by mere insertion that $(g = g(f), v = Z(f)[q(f) - H(f)g(f)])$ satisfy (2.6). Finally (2.8) follows from (2.5), since (2.6)(b) is a linear system of equations with $Z(f)[q(f) - H(f)g(f)]$ as a particular solution. ∎

The unique solution $g(f)$ to (2.6) will be called the *gain rate vector*, and $g^m(f)$ the gain rate of the subchain $C^m(f)$. A solution v to (2.6) will be called a *relative-value vector* and denoted by $v(f)$.

In the remainder, we will refer to the following example:

EXAMPLE 1. $N = 4$, $K(1) = K(2) = \{1\}$; $K(3) = K(4) = \{1, 2\}$; $H_{ij}^k = \delta_{ij}$ for all i, j, k.

i	k	P_{i1}^k	P_{i2}^k	P_{i3}^k	P_{i4}^k	q_i^k
1	1	0	1	0	0	0
2	1	1	0	0	0	0
3	1	1	0	0	0	$q_3^1 < 0$
3	2	0	0	1	0	0
4	1	0.4	0.4	0.2	0	0
4	2	0.8	0.2	0	0	0

Using (3.1) and theorem 3.1 part (c) one verifies that

$$V = \{v^* \in E^4 \mid v_1^* = v_2^*;\ v_3^* \geqslant q_3^1 + v_1^*;\ v_4^* = \max[0.8v_1^* + 0.2v_3^*;\ v_1^*]\}.$$

With $0 = v_1^* = v_2^*$ we get $v_3^* \geqslant q_3^1$ and $v_4^* = \max\{0.2v_3^*; 0\}$; so V is nonconvex. Note furthermore, that for $f \in S_{\text{RMG}}$, if f makes "unwise" decisions in states in $\Omega \backslash R(f)$, then there do not necessarily exist additive constants such that $v(f) \in V$ (cf. theorem 3 of [22], [25] and our theorem 4.1 part (b)). Take the above example and the pure policy $f^\# = (1, 1, 1, 1)$ with $P(f)$ unichained, and $v(f) = (0\ 0\ q_3^1\ 0.2q_3^1) + a(1\ 1\ 1\ 1) \notin V$ for any choice of the additive constant a.

Finally, reference [25] provides examples where the choice of additive constants in $v(f)$ affects the Policy Iteration Algorithm (PIA) (cf. [6], [8], [13]).

III. Properties of maximal gain policies. In this section we give some properties of maximal gain policies; some of the notions and properties presented here are related to results in [15], [16], [17], [18].

First, define the *maximal gain rate*

$$g_i^* = \sup_{f \in S_R} g(f)_i, \qquad i = 1, \ldots, N. \tag{3.1}$$

For any $v \in V$, $k \in K(i)$, and $f \in S_R$, define

$$b(v)_i^k = q_i^k - \sum_j H_{ij}^k g_j^* + \sum_j P_{ij}^k v_j - v_i, \qquad i = 1, \ldots, N,$$

and

$$b(v, f)_i = \sum_{k \in K(i)} f_{ik} b(v)_i^k = [q(f) - H(f)g^* + P(f)v - v]_i; \qquad i = 1, \ldots, N.$$

Since $g(f)$ can be interpreted as the average gain rate vector of f for a MRP with transition probabilities P_{ij}^k, one-step expected rewards q_i^k, and holding times T_i^k, we know from Derman [9] that there exists a pure policy that attains the N suprema in (3.1) simultaneously. Hence $g_i^* = \max_{f \in S_P} g(f)_i$. Accordingly define:

$$S_{\text{PMG}} = \{f \in S_P \mid g(f) = g^*\}$$

and

$$S_{\mathrm{RMG}} = \{f \in S_R \mid g(f) = g^*\}.$$

Finally, let:

$$w_i^* = \max_{f \in S_{\mathrm{PMG}}} Z(f)\big[q(f) - H(f)g^*\big]_i, \qquad i = 1, \ldots, N.$$

THEOREM 3.1 (*Properties of Maximal-Gain Policies*).

(a) *$f \in S_{\mathrm{RMG}}$ if and only if $g^* = P(f)g^*$ and $\Pi(f)[q(f) - H(f)g^*] = 0$.*

(b) *The functional equations* (1.1) *and* (1.2) *always have the solution $g = g^*$, $v = w^*$. Hence V is nonempty. Also, there exists a policy $f \in S_{\mathrm{PMG}}$ such that $w^* = Z(f)[q(f) - H(f)g^*]$.*

(c) *In any solution (g, v) of the functional equations* (1.1) *and* (1.2), *$g = g^*$, hence g and each $L(i)$ is unique.*

(d) *If f is any policy, and if C is any subchain of $P(f)$, then $g_i^* =$ constant, $i \in C$.*

(e) (*Cf.* [15, *p.* 16, *remark* 2]). *If $v \in V$, then $\max_{k \in L(i)} b(v)_i^k = 0$, for every i.*

Let $f \in S_R$.

(1) *Suppose that $k \in L(i)$ for each (i, k) with $f_{ik} > 0$ and that for some $v \in V$, $b(v)_i^k = 0$ for each (i, k) with $i \in R(f)$ and $f_{ik} > 0$. Then $f \in S_{\mathrm{RMG}}$.*

(2) *Conversely, if $f \in S_{\mathrm{RMG}}$, then for each $i = 1, \ldots, N$, $f_{ik} > 0$ implies $k \in L(i)$, and for $i \in R(f)$, $f_{ik} > 0$ implies $b(v)_i^k = 0$ for all $v \in V$.*

PROOF. (a) As noted in the proof of lemma 2.3, $g(f)$ is the unique solution to the equations $g = P(f)g$ and (2.9).

(b) Invoking the above mentioned interpretation of g^*, we know from theorem 1 in Denardo and Fox [8] that $g_i^* = \max_k \sum_j P_{ij}^k g_j^*$. Consider the discrete time decision model with $\bar{K}(i) = L(i) = \{k \mid g_i^* = \sum_j P_{ij}^k g_j^*\}$, $\bar{P}_{ij}^k = P_{ij}^k$ and $\bar{q}_i^k = q_i^k - \sum_j H_{ij}^k g_j^*$.

Note that in this model each policy has $\bar{g}(f) \leqslant 0$. Moreover, it follows from part (a) that $\bar{g}(f) = 0$ if and only if $f \in S_{\mathrm{RMG}}$. Hence the discrete time model has $\bar{g}^* = 0$ and, with $\bar{S}_{\mathrm{PMG}} = \{f \in X_{i=1}^N \bar{K}(i) \mid \bar{g}(f) = \bar{g}^* = 0 = \} S_{\mathrm{PMG}}$, we have:

$$\max_{f \in S_{\mathrm{PMG}}} Z(f)\big[q(f) - H(f)g^*\big]_i = \max_{f \in S_{\mathrm{PMG}}} Z(f)\big[\bar{q}(f) - \bar{g}^*\big]_i, \quad \text{for } i = 1, \ldots, N.$$

Use theorem 4 of [4] in order to prove the existence of a policy $f \in S_{\mathrm{PMG}}$ for which $w^* = Z(f)[q(f) - H(f)g^*]$, as well as the fact that w^* satisfies (1.2).

(c) Fix a solution (g, v) to (1.1) and (1.2). Using property A, a minor modification of the proof of lemma 4 of [8], shows that $g \geqslant g(f)$ for all $f \in S_P$ with equality for any f^0 such that $f_{ik}^0 = 1$ for some k maximizing (1.1) and (1.2). Hence $g = g^*$.

(d) Since g^* satisfies (1.1), we have $P(f)g^* \leqslant g^*$ for all $f \in S_R$. The assertion then follows from lemma 2-a in [8].

(e) The first result follows from the very definition of $b(v)_i^k$

(1) From the definition of $b(v)_i^k$, we have $v_i - \sum_j P(f)_{ij} v_j = q(f)_i - \sum_j H(f)_{ij} g_j^*$ for $i \in R(f)$. Multiplying this equation with $\Pi(f)_{ki}$ and summing over i, we obtain $\Pi(f)[q(f) - H(f)g^*] = 0$. Use this and $g^* = P(f)g^*$ in order to apply part (a).

(2) If $f \in S_{\mathrm{RMG}}$, $g^* = P(f)g^*$ follows from part (a). Hence $f_{ik} > 0$ implies $k \in L(i)$ and $b(v)_i^k \leqslant 0$. So $b(v, f) \leqslant 0$, for any $v \in V$. Since we know from part (a) that $\Pi(f)b(v, f) = 0$ for $f \in S_{\mathrm{RMG}}$, it follows that for $j \in R(f)$, $b(v, f)_j = 0$, i.e. $f_{jk} > 0$ implies $b(v)_j^k = 0$. ∎

Next define

$$R^* = \{i \mid i \in R(f) \text{ for some policy } f \in S_{\mathrm{PMG}}\}. \tag{3.2}$$

Next we define, for any $i \in R^*$, the set $K^*(i)$ as the set of actions which a pure maximal gain policy that has i among its set of recurrent states could prescribe:

$$K^*(i) = \{k \in K(i) \mid \text{there exists a } f \in S_{\mathrm{PMG}} \text{ with } i \in R(f) \text{ and } f_{ik} = 1\}, \qquad i \in R^*. \tag{3.3}$$

Finally, select a randomized policy f^* with

$$\{k \mid f^*_{ik} > 0\} = \begin{cases} K^*(i), & i \in R^*, \\ L(i), & i \in \Omega \backslash R^*. \end{cases} \tag{3.4}$$

Note that the chain- and periodicity structure of a stochastic matrix P merely depends upon the index set $I = \{(i, j) \mid P_{ij} > 0\}$ of *positive* entries, rather than upon the numerical values of the probabilities $[P_{ij}]$ themselves. As a consequence, the chain- and periodicity structure of a randomized policy f is completely determined by the sets of alternatives the policy uses (i.e. attributes positive weight to) in each of the states of Ω, rather than by specifying the entire tableau of numerical values $[f_{ik}]$. Hence, let $n^* = n(f^*)$ and let $\{R^{*\alpha} \mid \alpha = 1, \ldots, n^*\}$ denote the set of subchains of $P(f^*)$. The following theorem gives a characterization of the sets R^*, $\{R^{*\alpha} \mid \alpha = 1, \ldots, n^*\}$, the action sets $K^*(i)$, $i \in \Omega$, the integer n^*, and the policy f^*:

First note that $f^* \in S_{\mathrm{RMG}}$, in view of theorem 3.1 part (e).

THEOREM 3.2. (a)

$$K^*(i) = \{k \in L(i) \mid \textit{there exists a } f \in S_{\mathrm{RMG}} \textit{ with } i \in R(f) \textit{ and } f_{ik} > 0\}, \qquad i \in R^*, \tag{3.5}$$

$$R^* = \{i \in \Omega \mid i \in R(f), \textit{for some } f \in S_{\mathrm{RMG}}\}. \tag{3.6}$$

(b) $R(f^*) = R^*$, *i.e. the set* $\{f \in S_{\mathrm{RMG}} \mid R(f) = R^*\}$ *is nonempty.*

(c) *Any subchain of any* $f \in S_{\mathrm{RMG}}$ *is contained within a subchain of* $P(f^*)$, *i.e.*

$$n^* = \min\{n(f) \mid f \in S_{\mathrm{RMG}}, \textit{with } R(f) = R^*\}. \tag{3.7}$$

(d) *Let* $S^*_{\mathrm{RMG}} = \{f \in S_{\mathrm{RMG}} \mid R(f) = R^*, n(f) = n^*\}$. *All* $f \in S^*_{\mathrm{RMG}}$ *have the same collection of subchains* $\{R^{*\alpha} \mid \alpha = 1, \ldots, n^*\}$.

(e) *For any* α, $1 \leqslant \alpha \leqslant n^*$, $g^*_i = g^{*\alpha}$ *(say) for all* $i \in R^{*\alpha}$.

(f) *Let* $R^{(1)}, \ldots, R^{(m)}$ *be disjoint sets of states such that*

(1) *if* C *is a subchain of some* $f \in S_{\mathrm{RMG}}$ *then* $C \subseteq R^{(k)}$ *for some* k, $1 \leqslant k \leqslant m$,

(2) *there exists a* $f \in S_{\mathrm{RMG}}$ *with* $\{R^{(k)} \mid k = 1, \ldots, m\}$ *as its set of subchains.*

Then, $m = n^*$ *and, after (possible) renumbering,* $R^{(\alpha)} = R^{*\alpha}$ *for* $\alpha = 1, \ldots, n^*$.

(g) *For any* $v \in V$,

$$K^*(i) = \left\{k \in L(i) \mid b(v)^k_i = 0 \textit{ and } \sum_{j \in R^{*\alpha}} P^k_{ij} = 1\right\}, \qquad i \in R^{*\alpha};\ \alpha = 1, \ldots, n^*. \tag{3.8}$$

PROOF. (a) Fix a policy $f \in S_{\mathrm{RMG}}$ and a state $i \in R(f)$, as well as an alternative $k \in L(i)$ such that $f_{ik} > 0$. Consider a policy h satisfying the conditions (a), (b), (c) and (d) of lemma 2.2. Then, $i \in R(h)$ and $k \in K^*(i)$, whereas $h \in S_{\mathrm{PMG}}$ is verified by theorem 3.1, part (e). Thus the right-hand side of (3.6) is contained within R^*, whereas the reversed inclusion is immediate. Thus having shown (3.6), it follows that the right-hand sides of (3.5) are contained within the sets $K^*(i)$, $i \in R^*$ (whereas the reversed inclusion is immediate).

(b) We show that all states in R^* are recurrent under $P(f^*)$, i.e. $R(f^*) \supseteq R^*$ whereas the reversed inclusion is immediate from the definition of R^*. Let $i \in R^*$, and assume that state j can be reached from i under $P(f^*)$, i.e. there exists $(i_0 = i, \ldots, i_n = j)$ with $P(f^*)_{i_l i_{l+1}} > 0$ for $l = 0, \ldots, n-1$. Verify by complete induction that for all $l = 0, \ldots, n-1$, i_l and i_{l+1} belong to the same subchain of some maximal gain policy, hence i_l can be reached from i_{l+1} under $P(f^*)$. Conclude that state i can be reached from state j, under $P(f^*)$, so that $i \in R(f^*)$.

(c) Assume $P(f)$, for $f \in S_{\mathrm{RMG}}$, has a subchain $C^m(f)$ that intersects say the subchains R^{*1} and R^{*2} of $P(f^*)$. Then a policy f^{**} with $\{k \mid f^{**}_{ik} > 0\} = \{k \mid f^*_{ik} > 0\} \cup \{k \mid f_{ik} > 0\}$ for all $i \in C^m(f)$, and $\{k \mid f^{**}_{ik} > 0\} = \{k \mid f^*_{ik} > 0\}$ otherwise, is maximal gain, has $R(f^{**}) = R^*$, and its number of subchains is at most $n^* - 1$, since the states of R^{*1} and R^{*2} communicate with each other under $P(f^{**})$. On the other hand, $\{k \mid f^{**}_{ik} > 0\} = \{k \mid f^*_{ik} > 0\}$, for all $i \in \Omega$ in view of part (a), so that $P(f^{**})$ and $P(f^*)$ must have the same chain structure, i.e. $n(f^{**}) = n^*$ which contradicts $n(f^{**}) \leqslant n^* - 1$.

(d) Note that for all $f \in S^*_{\mathrm{RMG}}$, $\bigcup_{m=1}^{n^*} C^m(f) = R^*$ while each $C^m(f)$ $(1 \leqslant m \leqslant n^*)$ is contained within some set $R^{*\alpha}$ $(1 \leqslant \alpha \leqslant n^*)$.

(e) Use the fact that f^* is maximal gain, as well as part (d) of theorem 3.1.

(f) Apply property (1) to conclude $R^{*\alpha} \subseteq R^{(k(\alpha))}$. Apply part (c) and property (2) to conclude $R^{(k(\alpha))} \subseteq R^{*\alpha}$ $(1 \leqslant \alpha \leqslant n^*)$.

(g) Fix $\alpha \in \{1, \ldots, n^*\}$, $i_0 \in R^{*\alpha}$. First, let $k \in K^*(i)$ and $f \in S_{\mathrm{PMG}}$, with $i \in R(f)$ and $f_{ik} = 1$ and apply part (e) of theorem 3.1 and part (d) of this theorem, in order to prove that $K^*(i)$ is contained within the set on the right hand side of the equality. Next, take $k_0 \in L(i_0)$ such that $b(v)^{k_0}_{i_0} = 0$ and $\sum_{j \in R^{*\alpha}} P^{k_0}_{ij} = 1$. Define f^{**} such that $f^{**}_{i_0 k_0} = 1$ and $f^{**}_{jk} = f^*_{jk}$, for all $j \neq i_0$, $k \in K(j)$. Obviously, all states in $R^{*\alpha} \backslash \{i_0\}$ can reach state i_0 under $P(f^{**})$, whereas state i_0 can only reach states within $R^{*\alpha}$. We conclude that $i_0 \in R(f^{**})$ while $f^{**} \in S_{\mathrm{RMG}}$, as can be verified using part (e) of theorem 3.1., hence $k_0 \in K^*(i)$, thus proving the reversed inclusion. ∎

REMARK 1. A policy f^* as defined by (3.4) may be constructed in the following way: Fix an enumeration $f^1, \ldots, f^M$ of S_{PMG}. For any $i \in R^*$, let $A_i = \{r \mid i \in R(f^r)\}$. Consider the following equivalence relation on $\mathcal{C} = \{C^m(f^r) \mid 1 \leqslant r \leqslant M;\ 1 \leqslant m \leqslant n(f^r)\}$: Let $C \sim C'$ if there exists $\{C^{(1)} = C, C^{(2)}, \ldots, C^{(n)} = C'\}$ with $C^{(i)} \in \mathcal{C}$ and $C^{(i)} \cap C^{(i+1)} \neq \emptyset$ for $i = 1, \ldots, n-1$. Let f^* satisfy: (1) $\{k \mid f^*_{ik} > 0\} = \bigcup_{r \in A_i} \{k \mid f^r_{ik} > 0\} = K^*(i)$ for $i \in R^*$; (2) $\{k \mid f^*_{ik} > 0\} = L(i)$ for $i \in \Omega \backslash R^*$. The equivalence classes generated by the above defined relation constitute the subchains of $P(f^*)$ since they are closed under $P(f^*)$ and since the states belonging to the same equivalence class communicate with each other. Note that randomization, by coalescing subchains, is essential for the recurrency properties: in general, there may fail to exist a pure maximal gain policy f with $R(f) = R^*$, or which achieves the minimal number n^* of subchains.

A finite procedure for calculating R^*, n^*, the $R^{*\alpha}$ and a $f^* \in S^*_{\mathrm{RMG}}$ is therefore as follows: use the PIA to find g^* and a $v \in V$. Compute $S_P(v) = X_{i=1}^N \{k \in L(i) \mid b(v)^k_i = 0\} = \{f \in S_P \mid f$ achieves the $2N$ minima in (1.1) and (1.2)$\} \subseteq S_{\mathrm{PMG}}$. Note from part (e) of theorem 3.1 that for all $f \in S_{\mathrm{PMG}}$ there exists a policy $h \in S_P(v)$, such that both policies coincide on $R(f)$. Conclude that $R^* = \{i \mid i \in R(f)$, for some $f \in S_P(v)\}$ (cf. also [17, algorithm on p. 353–359]). Determine $\{R^{*\alpha} \mid \alpha = 1, \ldots, n^*\}$ as the equivalence classes of the above defined relation, with respect to the set of subchains of policies belonging to $S_P(v)$ (cf. theorem 3.2 part (g)). Finally, select a policy f^* satisfying (3.4), where the sets $K^*(i)$, $i \in R^*$, are determined using theorem 3.2 part (g).

IV. Properties of V. Some basic properties of V are given by:

THEOREM 4.1. (*Basic Properties of V*). (a) *V is closed and unbounded, as $v \in V$ implies $v + a_1\mathbf{1} + a_2 g^* \in V$, for any scalars a_1, a_2 (where $\mathbf{1}$ is the N-vector with all coordinates unity).*

(b) (*Maximality of relative values*). *For any $v^* \in V$ and $f \in S_{\mathrm{RMG}}$, it is possible to choose the $n(f)$ additive constants in $v(f)$ such that $v^* \geqslant v(f)$ with equality for components in $R(f)$.*

(c) (*Cf.* [3], [15], [16], [21].) *$v \in V$ if and only if*

$$v_i = \max_{f \in S_{\mathrm{PMG}}} \{Z(f)[q(f) - H(f)g^*]_i + \Pi(f)v_i\}, \qquad i = 1, \dots, N. \tag{4.1}$$

In addition, if $v \in V$, then a policy $f \in S_{\mathrm{PMG}}$ achieves all N maxima in (4.1) *if and only if it achieves the $2N$ maxima in* (1.1) *and* (1.2).

PROOF. (a) Immediate to verify.

(b) Choose in (2.8) $a_m = \langle \pi^m(f), v^* \rangle$. From part (e) of theorem 3.1, it follows that $\{k \mid f_{ik} > 0\} \subseteq L(i)$ for each i, hence $v^* \geqslant q(f) - H(f)g^* + P(f)v^*$, which implies, using theorem 3.1 part (a), lemma 2.1, (2.4) and (2.8):

$$\begin{aligned} v^* &\geqslant Z(f)[q(f) - H(f)g^*] + \Pi(f)v^* \\ &= Z(f)[q(f) - H(f)g^*] + \sum_{m=1}^{n(f)} a_m \phi^m(f) = v(f) \end{aligned}$$

with equality for components in $R(f)$.

(c) First assume $v \in V$. In part (b) we proved that for any $f \in S_{\mathrm{PMG}}$, $v \geqslant Z(f)[q(f) - H(f)g^*] + \Pi(f)v$, with strict equality for $f \in S_P(v)$. Hence, $v \in V$ implies (4.1) and any policy achieving the $2N$ maxima in (1.1) and (1.2) achieves all N maxima in (4.1).

Conversely, if v satisfies (4.1), we define

$$\tilde{v} = \max_{k \in L(i)} \left[q_i^k - \sum_j H_{ij}^k g_j^* + \sum_j P_{ij}^k v_j \right], \qquad i = 1, \dots, N, \tag{4.2}$$

and show both $\tilde{v} \geqslant v$ and $\tilde{v} \leqslant v$, hence $\tilde{v} = v \in V$.

For any $f \in S_{\mathrm{PMG}}$, $f_{ik} = 1$ implies $k \in L(i)$ by theorem 3.1 part (e); hence using (4.1), (2.2) and theorem 3.1 part (a):

$$\begin{aligned} \tilde{v} &\geqslant q(f) - H(f)g^* + P(f)v \geqslant [I + P(f)Z(f)][q(f) - H(f)g^*] + \Pi(f)v \\ &= Z(f)[q(f) - H(f)g^*] + \Pi(f)v, \qquad f \in S_{\mathrm{PMG}}. \end{aligned}$$

This implies $\tilde{v} \geqslant v$. Let h denote a pure policy in $X_{i=1}^N L(i)$, achieving all maxima in (4.2). Then:

$$v_i \leqslant \tilde{v}_i = [q(h) - H(h)g^* + P(h)v]_i; \qquad i = 1, \dots, N. \tag{4.3}$$

Multiply (4.3) with $\Pi(h) \geqslant 0$ in order to get $0 \leqslant \Pi(h)[q(h) - H(h)g^*] \leqslant 0$, the latter inequality following from (2.9) and $g(h) \leqslant g^*$. Hence $h \in S_{\mathrm{PMG}}$, by part (a) of theorem 3.1.

Using lemma 2.1, (4.3) implies $v \leqslant Z(h)[q(h) - H(h)g^*] + \Pi(h)v$. Insert this on the right-hand side of (4.2) and use $\Pi(h)[q(h) - H(h)g^*] = 0$, to obtain:

$$\begin{aligned} \tilde{v} &\leqslant [I + P(h)Z(h)][q(h) - H(h)g^*] + \Pi(h)v \\ &= Z(h)[q(h) - H(h)g^*] + \Pi(h)v \\ &\leqslant \max_{f \in S_{\mathrm{PMG}}} \{Z(f)[q(f) - H(f)g^*] + \Pi(f)v\} = v. \end{aligned}$$

Finally, if $f \in S_{\mathrm{PMG}}$ achieves the N maxima in (4.1), multiply the resulting equality in (4.1) with $Z(f)^{-1}$ to show that it achieves the N maxima in (1.2), as well as the N maxima in (1.1), since $f_{ik} = 1$ implies $k \in L(i)$. This completes the proof. ∎

Since for $f \in S_{\mathrm{RMG}}$, $\Pi(f)_{ij} = 0$ if $j \notin R^*$, we have by part (c) of theorem 4.1 that $v \in V$ if and only if

$$v_i = \max_{f \in S_{\mathrm{PMG}}} \left\{ Z(f)[q(f) - H(f)g^*]_i + \sum_{j \in R^*} \Pi(f)_{ij} v_j \right\}, \qquad i \in R^*, \tag{4.4}$$

$$v_i = \max_{f \in S_{\mathrm{PMG}}} \left\{ Z(f)[q(f) - H(f)g^*]_i + \sum_{j \in R^*} \Pi(f)_{ij} v_j \right\}, \qquad i \in \Omega \backslash R^*. \tag{4.5}$$

Observe that (4.4) involves only $(v_i \mid i \in R^*)$ and can be studied in isolation. The $(v_i \mid i \in \Omega \backslash R^*)$ are uniquely determined via (4.5), for any $(v_i \mid i \in R^*)$. Define now

$$V^R = \{(v_i \mid i \in R^*);\ v_i \text{ satisfy } (4.4)\}. \tag{4.6}$$

THEOREM 4.2. (a)

$$V^R = \left\{ (v_i \mid i \in R^*);\ v_i \geqslant Z(f)[q(f) - H(f)g^*]_i + \sum_{j \in R^*} \Pi(f)_{ij} v_j, \right.$$
$$\left. \textit{for all } i \in R^*, f \in S_{\mathrm{PMG}} \right\}. \tag{4.7}$$

Hence, V^R is a closed, convex, unbounded, polyhedral set.

(b) *V is connected.*

PROOF. (a) Clearly, V^R is contained within the polyhedron that is defined in the right side of (4.7). Conversely fix $i \in R^*$ and $h \in S_{\mathrm{PMG}}$ with $i \in R(h)$. Then, by multiplying the inequalities in (4.7) with $\Pi(h) \geqslant 0$, we obtain $v_i = Z(h)[q(h) - H(h) g^*]_i + \sum_{j \in R^*} \Pi(h)_{ij} v_j$; hence (4.4) holds. The unboundedness of V is proved as in theorem 4.1.

(b) The assertion follows by showing that for any $v, \tilde{v} \in V$, the curve $\{v(\lambda) \mid \lambda \in [0, 1]\}$ with parameter representation: $v(\lambda)_i = \lambda v_i + (1 - \lambda)\tilde{v}_i$, $i \in R^*$ and

$$v(\lambda)_i = \max_{f \in S_{\mathrm{PMG}}} \left\{ Z(f)[q(f) - H(f)g^*]_i + \sum_{j \in R^*} \Pi(f)_{ij} v(\lambda)_j \right\},$$

for $i \notin R^*$, connects v with $\tilde{v}$, lies within V as a consequence of (4.5) and part (a), and is continuous, since all its components are continuous functions of λ. ∎

We already saw that V may not be convex. The following theorem gives a necessary and sufficient condition for the convexity of V.

This property is especially important when considering MRPs, where for several quantities of interest (e.g. the optimal bias vector) variational characterizations may be obtained of the nature: $\max_{v \in V}[c + Bv]$ (where c and B are expressions in q_i^k, P_{ij}^k and H_{ij}^k) and the latter is a linear program if and only if V is convex.

THEOREM 4.3. *V is convex if and only if for each $i \in \Omega - R^*$ there exists an alternative $k(i) \in L(i)$, such that for all $v \in V$:*

$$v_i = q_i^{k(i)} - \sum_j H_{ij}^{k(i)} g_j^* + \sum_j P_{ij}^{k(i)} v_j. \tag{4.8}$$

Moreover, V is convex if and only if it is a polyhedron.

PROOF. We first observe that for any $i \in R^*$, there is a $h \in S_{\mathrm{PMG}}$, with $i \in R(h)$, hence by part (e) of theorem 3.1 there exists an alternative $k(i) \in L(i)$ with $b(v)_i^{k(i)}$

$= 0$, for any $v \in V$. Thus (4.8) always holds for $i \in R^*$. Suppose it holds for $i \in \Omega \backslash R^*$ as well. Then the functional equations (1.2) are equivalent to the linear (in)equalities $b(v)_i^{k(i)} = 0$ for $i = 1, \ldots, N$ and $b(v)_i^k \leqslant 0$ for $k \in L(i) \backslash \{k(i)\}$ and $i = 1, \ldots, N$. Hence V is a convex polyhedron.

Conversely, suppose V is convex. Assume to the contrary that there exists a state $i \in \Omega \backslash R^*$ and a finite set of $v^{(m)}$'s in V, such that no $k \in L(i)$ achieves the maximum in (1.2) for all $v^{(m)}$. However, since V is convex, it is immediate to verify that a $k \in L(i)$ achieving the maximum in (1.2) for a positive convex combination $\bar{v}$ of the $v^{(m)}$'s, achieves the maximum in (1.2) for each $v^{(m)}$. ∎

REMARK 2. Condition (4.8), hence convexity of V, holds trivially if (1) $R^* = \Omega$, or (2) $L(i)$ is a singleton for each $i \in \Omega \backslash R^*$, or (3) there is only one maximal gain policy or (4) $n^* = 1$, since in this case $v \in V$ is unique up to a multiple of $\mathbf{1}$ (cf. remark 3).

For discrete time Markovian decision processes, where $H_{ij}^k = \delta_{ij}$, the value iteration equations take the form:

$$v(n+1)_i = \max_{k \in K(i)} \left\{ q_i^k + \sum_j P_{ij}^k v(n)_j \right\}, \tag{4.9}$$

with $v(0)$ a given vector.

It is well known that $\{v(n) - ng^*\}_{n=1}^{\infty}$ may fail to converge. In a forthcoming paper [24] it will be shown that there exists an integer J such that

$$u_i^{(r)} = \lim_{n \to \infty} \{ v(nJ + r)_i - (nJ + r) g_i^* \}$$

exists for all i, with $u^{(r+J)} = u_i^{(r)}$ (previous proofs in [5] and [15] are both incorrect; cf. [24]).

Accordingly, define $\bar{v}$ as the Cesaro-limit of the sequence $\{v(n) - ng^*\}_{n=1}^{\infty}$. Example 1 with $v(0) = [1\ 0\ 1\ 0.6]$ shows that in general $\bar{v} \notin V$ ($v(2n)_1 = 1$; $v(2n+1)_1 = 0$; $v(2n)_2 = 0$; $v(2n+1)_2 = 1$; $v(n)_3 = 1$; $v(2n)_4 = 0.6$; $v(2n+1)_1 = 0.8$; $\bar{v} = [0.5\ 0.5\ 1\ 0.7] \notin V$).

The relation between $\bar{v}$ and V is as follows:

THEOREM 4.4. (a) $\{\bar{v}_i \mid i \in R^*\} \in V^R$.

(b) *There exists a vector* $v \in V$, *such that* $v \leqslant \bar{v}$ *with equality for components in* R^*.

PROOF. Note that for all $i \in \Omega$:

$$u_i^{(r+1)} = \max_{k \in L(i)} \left\{ q_i^k - g_i^* + \sum_j P_{ij}^k u_j^{(r)} \right\},$$

since for all n sufficiently large the maximizing alternatives in (4.9) belong to $L(i)$ as observed in [5] and [15].

Since $\bar{v} = (1/J)\sum_{r=0}^{J-1} u^{(r)}$, we obtain by averaging over $r = 0, \ldots, J-1$:

$$\bar{v}_i \geqslant q_i^k - g_i^* + \sum_j P_{ij}^k \bar{v}_j, \qquad i = 1, \ldots, N \text{ and } k \in L(i).$$

Take any $f \in S_{\mathrm{PMG}}$ to obtain: $\bar{v} \geqslant q(f) - g^* + P(f)\bar{v}$, and hence, using lemma 2.1: $\bar{v} \geqslant Z(f)[q(f) - g^*] + \Pi(f)\bar{v}$, with equality for $i \in R(f)$. This implies: $\bar{v} \geqslant \max_{f \in S_{\mathrm{PMG}}}\{Z(f)[q(f) - g^*] + \Pi(f)\bar{v}\}$ with equality for components in R^*. Using (4.4) and (4.5) we obtain that the vector v defined by (1) $v_i = \bar{v}_i$, $i \in R^*$ and (2) $v_i = \max_{f \in S_{\mathrm{PMG}}} \{Z(f)[q(f) - g^*]_i + \sum_{j \in R^*} \Pi(f)_{ij} v_j\}$ for $i \in \Omega \backslash R^*$, belongs to V with $v \leqslant \bar{v}$ and equality for components in R^*. ∎

V. The n^* degrees of freedom in V. In this section we show that the convex polyhedral set V^R has dimension n^* and that its elements, and hence V, are fully determined by n^* parameters $(y_1, \ldots, y_{n^*})$.

Romanovsky [20] obtained the same result for the functional equations that arise in discrete time Markov models with $g^* = \langle g^* \rangle \mathbf{1}$. In addition, as our methods involve the chain structure, a fuller characterization of the parameter space is possible.

The key observation is that any two vectors $v^0, \tilde{v} \in V$ have the property: $\tilde{v}_i - v_i^0$ = constant $= y_\alpha$ for $i \in R^{*\alpha}$, $\alpha = 1, \ldots, n^*$. By fixing $v^0 \in V$ and picking these n^* constants, one thus determines $(\tilde{v}_i \mid i \in R^*)$ and hence $\tilde{v}$ by (4.5) in terms of v^0. Hence, by fixing v^0, and sweeping out all permitted values of y, we sweep out all vectors $\tilde{v}$ in V. In particular (5.1) below shows that $\tilde{v}$ is a convex piecewise linear function in y.

THEOREM 5.1. *Let $v \in V$. The following are equivalent*:

(a) $v + x \in V$,

(b) $x_i = \max_{k \in L(i)} [b(v)_i^k + \sum_j P_{ij}^k x_j], \quad i = 1, \ldots, N,$

(c) $x_i = \max_{f \in S_{PMG}} [Z(f)b(v,f) + \Pi(f)x]_i, \quad i = 1, \ldots, N,$

(d) *there are n^* constants $y = (y_1, \ldots, y_{n^*})$ satisfying*

$$x_i = \begin{cases} y_\alpha, & i \in R^{*\alpha}, \alpha = 1, \ldots, n^*, \\ \max\limits_{f \in S_{PMG}} \left[Z(f)b(v,f)_i + \sum\limits_{\beta=1}^{n^*} \left(\sum\limits_{j \in R^{*\beta}} \Pi(f)_{ij} \right) y_\beta \right], & i \in \Omega \backslash R^*, \end{cases} \tag{5.1}$$

$$y_\alpha \geq Z(f)b(v,f)_i + \sum_{\beta=1}^{n^*} \left(\sum_{j \in R^{*\beta}} \Pi(f)_{ij} \right) y_\beta,$$
$$\alpha = 1, \ldots, n^*; \quad i \in R^{*\alpha}, f \in S_{PMG}. \tag{5.2}$$

PROOF. (a)$\Leftrightarrow$(b): (b) is the requirement that $v + x \in V$.

(a)$\Leftrightarrow$(c): Cf. (4.1) and the definition of $b(v,f)$.

(a)$\Rightarrow$(d): Take $\hat{f} \in S^*_{RMG}$. As $v, v + x \in V$, we have from part (e) of theorem 3.1: $v_i = [q(\hat{f}) - H(\hat{f})g^* + P(\hat{f})v]_i$ and $(v + x)_i = [q(\hat{f}) - H(\hat{f})g^* + P(\hat{f})(v + x)]_i$ for all $i \in R^* = R(\hat{f})$. Subtraction yields: $x_i = [P(\hat{f})x]_i = [\Pi(\hat{f})x]_i = \langle \pi^\alpha(\hat{f}), x \rangle$ for $i \in R^{*\alpha}$, which proves the first part of (5.1). Moreover, this implies the remainder of (d), using (4.4) and (4.5) and the definition of $b(v,f)$.

(d)$\Rightarrow$(a): Use (4.4), (4.5) and the definition of $b(v,f)$. ∎

Fix $v \in V$. Define the set of allowed constants

$$Y(v) = \{ y \in E^{n^*} \mid y \text{ satisfies } (5.2) \}.$$

Note that,

$$Z(f)b(v,f) \leq 0 \quad \text{for all } f \in S_{PMG}. \tag{5.3}$$

(5.3) follows from lemma 2.1, with $x = 0$, using $b(v,f) \leq 0$ and $\Pi(f)b(v,f) = 0$ (cf. theorem 3.1 parts (d) and (e)).

Clearly, by (5.3), (5.2) is automatically satisfied for (α, i, f) with $\sum_{j \in R^{*\alpha}} \Pi(f)_{ij} = 1$. We accordingly define:

$$\tilde{K}(\alpha) = \left\{ (i,f) \mid i \in R^{*\alpha}, f \in S_{PMG}, \sum_{j \in R^{*\alpha}} \Pi(f)_{ij} < 1 \right\}, \qquad \alpha = 1, \ldots, n^*,$$

and make the partition $\{1, 2, \ldots, n^*\} = E \cup F$, where $E = \{\alpha \mid \tilde{K}(\alpha) = \emptyset\}$, $F = \{\alpha \mid \tilde{K}(\alpha) \neq \emptyset\}$,

For $\xi = (i,f) \in \tilde{K}(\alpha)$, define

$$\tilde{q}_\alpha^\xi = [Z(f)b(v,f)]_i \quad \text{and} \quad \tilde{P}_{\alpha\beta}^\xi = \sum_{j \in R^{*\beta}} \Pi(f)_{ij}.$$

Note that $\tilde{q}_\alpha^\xi \leq 0$, $\tilde{P}_{\alpha\beta}^\xi \geq 0$, $\sum_{\beta=1}^{n^*} \tilde{P}_{\alpha\beta}^\xi = 1$, $\tilde{P}_{\alpha\alpha}^\xi < 1$ for all $\alpha \in F$, and $\xi \in \tilde{K}(\alpha)$. Then $Y(v)$ consists of all $y \in E^{n^*}$ satisfying

$$y_\alpha \geq \tilde{q}_\alpha^\xi + \sum_{\beta=1}^{n^*} \tilde{P}_{\alpha\beta}^\xi y_\beta, \qquad \alpha \in F, \xi \in \tilde{K}(\alpha). \tag{5.4}$$

In order to show that $Y(v)$ is an n^*-dimensional polyhedral set, we need the following discrete time Markovian model with state space $\{1, \ldots, n^*\}$: For $\alpha \in F$, let $\tilde{K}(\alpha)$ be the set of feasible decisions. For $\xi \in \tilde{K}(\alpha)$, let $\tilde{q}_\alpha^\xi$ and $\tilde{P}_{\alpha\beta}^\xi$ denote the associated one-step reward and transition probabilities (we already noted that $\tilde{P}_{\alpha\beta}^\xi \geq 0$, $\sum_\beta \tilde{P}_{\alpha\beta}^\xi = 1$).

For $\alpha \in E$, add a decision ξ_0 to the empty $\tilde{K}(\alpha)$ with $\tilde{q}_\alpha^{\xi_0} = -1$ and $\tilde{P}_{\alpha\beta}^{\xi_0} = \delta_{\alpha\beta}$. Let Φ denote the set of pure policies. For $\varphi \in \Phi$, the quantities $\tilde{q}(\varphi)$, $\tilde{P}(\varphi)$, $\tilde{\Pi}(\varphi)$ and $\tilde{Z}(\varphi)$ are defined analogously to $q(f)$, $P(f)$, $\Pi(f)$ and $Z(f)$ for $f \in S_P$. Also let $\{\tilde{g}_\alpha^*\}$ be the maximal gain vector for the new process. Note that $\tilde{q}(\varphi) \leq 0$ for any $\varphi \in \Phi$, so $\tilde{g}_\alpha^* \leq 0$ for all α. Also $\tilde{g}_\alpha^* = -1$ for $\alpha \in E$, since each state $\alpha \in E$ is a trapping state for $\tilde{P}(\varphi)$, for all $\varphi \in \Phi$. The following lemma characterizes the subchains of $\tilde{P}(\varphi)$ on F:

LEMMA 5.2 (*Properties of subchains of $\tilde{P}(\varphi)$ on F.*) *Fix $v \in V$. Assume $F \neq \emptyset$. Suppose for some policy $\varphi \in \Phi$, $\tilde{P}(\varphi)$ has a subchain $C \subseteq F$. Then*
(a) *C has at least two members,*
(b) *$\tilde{q}(\varphi)_\alpha$ is strictly negative for at least one $\alpha \in C$.*

PROOF. (a) Part (a) follows from $\tilde{P}_{\alpha\alpha}^\xi < 1$ for any $\alpha \in F$ and $\xi \in \tilde{K}(\alpha)$.

(b) Let policy φ use action $(i(\alpha), f(\alpha)) \in \tilde{K}(\alpha)$ for each $\alpha \in C$. For $\alpha \in C$, define $S(\alpha) = \{j \mid P(f(\alpha))_{i(\alpha)j}^n > 0$, for some $n = 0, 1, 2, \ldots\}$. Note that $i(\alpha) \in S(\alpha)$ and that:

$$\alpha \in C, i \in S(\alpha) \text{ imply } P(f(\alpha))_{ij} > 0 \text{ only if } j \in S(\alpha). \tag{5.5}$$

Now assume to the contrary that for each $\alpha \in C$, $0 = \tilde{q}(\varphi)_\alpha = Z(f(\alpha))b(v, f(\alpha))_{i(\alpha)}$. Since $f(\alpha) \in S_{\mathrm{PMG}}$, $b(v, f(\alpha)) \leq 0$ with equality for components in $R(f(\alpha))$. Hence, using (2.3),

$$\begin{aligned} 0 &= \tilde{q}(\varphi)_\alpha \\ &= \sum_{j \notin R(f(\alpha))} Z(f(\alpha))_{i(\alpha)j}\, b(v, f(\alpha))_j \\ &= \sum_{j \notin R(f(\alpha))} \sum_{n=0}^{\infty} [P(f(\alpha))]_{i(\alpha)j}^n\, b(v, f(\alpha))_j \end{aligned}$$

where the interchange of $\sum_n$ and $\lim_{a\uparrow 1}$ is justified by the monotone convergence theorem. Hence:

$$b(v, f(\alpha))_j = 0 \quad \text{for } j \in S(\alpha), \alpha \in C. \tag{5.6}$$

We now exhibit a policy $f^0 \in S_{\mathrm{RMG}}$ with the contradictory properties that $R^0 = \bigcup_{\alpha \in C} [R^{*\alpha} \cup S(\alpha)]$ is closed under $P(f^0)$ while every state in R^0 is transient for $P(f^0)$.

Consider a policy $f^* \in S_{\mathrm{RMG}}$. Define f^0 as follows:

Initially, for $i \in R^*$ set $\{k \mid f_{ik}^0 > 0\} = \{k \mid f_{ik}^* > 0\}$. Then for $i \in S(\alpha)$ add $\{k \mid f(\alpha)_{ik} > 0\}$ to $\{k \mid f_{ik}^0 > 0\}$. Finally, for $i \in \Omega \backslash R^0$, set $\{k \mid f_{ik}^0 > 0\} = \{k \in L(i) \mid b(v)_i^k = 0\}$.

From (5.6), the definition of f^* in combination with theorem 3.1 part (e), and the definition of f^0 on $\Omega \backslash R^0$ it follows that $f_{ik}^0 > 0$ implies $b(v)_i^k = 0$, for all i, hence $f^0 \in S_{\mathrm{RMG}}$.

For $i \in R^0$, (5.5) and the fact that $f^* \in S^*_{\mathrm{RMG}}$ imply that $P(f^0)_{ij} > 0$ only for $j \in R^0$; hence, R^0 is closed under $P(f^0)$.

As $\sum_{j \notin R^{*\alpha}} \Pi(f(\alpha))_{i(\alpha)j} > 0$, there exists a $j \notin R^{*\alpha}$, and an integer $n \geqslant 1$, with $P(f(\alpha))^n_{i(\alpha)j} > 0$ and so $P(f^0)^n_{i(\alpha)j} > 0$. Hence $i(\alpha) \in R^{*\alpha}$ is transient under $P(f^0)$, since the subchains of a maximal gain policy are all contained within a single $R^{*\beta}$ (cf. theorem 3.2 part (c)).

Now, observe that for each $\alpha \in C$, all states in $R^{*\alpha}$ communicate with $i(\alpha) \in R^{*\alpha}$ for $P(f^0)$, since they communicate with $i(\alpha)$ for $P(f^*)$. However, this implies that each state in $\bigcup_{\alpha \in C} R^{*\alpha}$ is transient, since a transient state cannot be reached from a recurrent state.

It remains to be proved that each $j \in S(\alpha)$ $(\alpha \in C)$ is transient for $P(f^0)$: Fix $j \in S(\alpha)$, $\alpha \in C$. Since $f(\alpha)$ is maximal gain, there is a state $r \in R^{*\beta}$, for some β, such that $P(f(\alpha))^m_{jr} > 0$, for some $m \geqslant 1$. Hence $P(f^0)^m_{jr} > 0$. Let n be such that $P(f(\alpha))^n_{i(\alpha)j} > 0$. Finally $\beta \in C$ follows from

$$\tilde{P}(\varphi)_{\alpha\beta} \geqslant \Pi(f(\alpha))_{i(\alpha)r} = \left[P(f(\alpha))^n \Pi(f(\alpha))\right]_{i(\alpha)r}$$

$$\geqslant P(f(\alpha))^n_{i(\alpha)j} \Pi(f(\alpha))_{jr} > 0$$

and the fact that C is a subchain if $\tilde{P}(\varphi)$. This implies that r is transient for $P(f^0)$ and so is j, since a transient state cannot be reached from a recurrent state. ∎

Together part (b) of lemma 5.2 and the choice of $\tilde{q}^{\xi_0}_\alpha = -1$ for $\alpha \in E$ imply:

$$\tilde{g}^*_\alpha < 0 \quad \text{for } \alpha = 1, \ldots, n^*. \tag{5.7}$$

THEOREM 5.3 (*Cf. theorem* 3 *of* [20]). *Fix $v \in V$. Given any $\{y_\alpha \mid \alpha \in E\}$ there exist $\{y_\alpha \mid \alpha \in F\}$ such that the following strict inequalities hold*:

$$y_\alpha > \tilde{q}^\xi_\alpha + \sum_{\beta=1}^{n^*} \tilde{P}^\xi_{\alpha\beta} y_\beta \quad \text{for all } \alpha \in F, \xi \in \tilde{K}(\alpha). \tag{5.8}$$

PROOF. It suffices to show that there exists a solution y^0 to (5.8) for some $\{y^0_\alpha \mid \alpha \in E\}$ since a solution for any $\{y_\alpha \mid \alpha \in E\}$ is then obtained by first adding a large positive constant to every y_α, and then reducing $\{y_\alpha \mid \alpha \in E\}$ to the desired magnitudes, thereby strengthening the inequalities (5.8).

Since $\tilde{q}^{\xi_0}_\alpha = -1$ and $\tilde{P}^{\xi_0}_{\alpha\alpha} = 1$, for $\alpha \in E$, the solution set to (5.8) is not altered by adding the inequalities $y_\alpha \geqslant \tilde{q}^{\xi_0}_\alpha + \sum_{\beta=1}^{n^*} \tilde{P}^{\xi_0}_{\alpha\beta} y_\beta$, $\alpha \in E$. Now assume to the contrary, that the solution set of (5.8) is empty. Then for the LP-problem:

$$\min Z \quad \text{subject to}$$

$$y_\alpha + Z \geqslant \tilde{q}^\xi_\alpha + \sum_{\beta=1}^{n^*} \tilde{P}^\xi_{\alpha\beta} y_\beta, \qquad \alpha = 1, \ldots, n^*; \xi \in \tilde{K}(\alpha),$$

we have $\min Z \geqslant 0$, which according to theorem 2 of [19], implies $\max_{\alpha=1,\ldots,n^*} \tilde{g}^*_\alpha \geqslant 0$. This contradicts (5.7). ∎

Since the solution set to (5.8) is open, for any y satisfying (5.8), there exists a $\delta > 0$, so that $|y - y'| < \delta$ implies $y' \in Y(v)$. Hence the n^* parameters $(y_1, \ldots, y_{n^*})$ may be chosen independently over some (finite) region. V and V^R have exactly $n^* = |E \cup F|$ degrees of freedom of which $|E|$ are globally independent and $|F|$ are only locally independent. Examples can be constructed where E (or F) can be empty; e.g. F is empty if $n^* = 1$. Finally note:

REMARK 3. $n^* = 1 \Leftrightarrow v \in V$ is unique up to a multiple of $\mathbf{1}$.

VI. Acknowledgement. We wish to express our sincere thanks to Dr. Henk Tijms, for his useful comments and careful reading of this and previous versions of this paper.

References

[1] Bather, J. (1973). Optimal Decision Procedures for Finite Markov Chains, Part III. *Advances in Appl. Probability* **5** 541–554.

[2] Bellman, R. (1957). A Markovian Decision Process. *J. Math. Mech.* **6** 679–684.

[3] ———. (1955). Functional Equations in the Theory of Dynamic Programming, V. Positivity and Quasi-Linearity. *Proc. Nat. Acad. Sci. U.S.A.* **41** 743–746.

[4] Blackwell, D. (1962). Discrete Dynamic Programming. *Ann. Math. Statist.* **33** 719–726.

[5] Brown, B. (1965). On the Iterative Method of Dynamic Programming on a Finite State Space Discrete Time Markov Process. *Ann. Math. Statist.* **36** 1279–1285.

[6] DeCani, J. (1964). A Dynamic Programming Algorithm for Embedded Markov Chains when the Planning Horizon Is at Infinity. *Management Sci.* **10** 716–733.

[7] Denardo, E. (1971) Markov Renewal Programs with Small Interest Rates. *Ann. Math. Statist.* **42** 447–496.

[8] ——— and Fox, B. (1965). Multichain Markov Renewal Programs. *SIAM J. Appl. Math.* **16** 468–487.

[9] Derman, C. (1970). *Finite State Markovian Decision Processes.* Academic Press, New York.

[10] Galperin, A. (1976). The General Solution of a Finite System of Linear Inequalities. *Math. Oper. Res.* **1** 185–196.

[11] Howard, R. (1960). *Dynamic Programming and Markov Processes.* Wiley, New York.

[12] ———. (1963). Semi Markovian Decision Processes. *Bull. Inst. Internat. Statist.* **40** 625–652.

[13] Jewell, W. (1963). Markov Renewal Programming. *Operations Res.* **11** 938–971.

[14] Kemeny, J. and Snell, J. (1961). *Finite Markov Chains.* Van Nostrand, Princeton, N.J.

[15] Lanery, E. (1967). Etude Asymptotiques des Systèmes Markoviens à Commande. *R.I.R.O.* **1** 3–56.

[16] ———. (1968). *Compléments à l'étude Asymptotique des Systemes Markoviens à Commande.* I.R.I.A., Rocquencourt, France.

[17] Lembersky, M. (1974). Preferred Rules in Continuous Time Markov Decision Processes. *Management Sci.* **21** 348–357.

[18] Miller, B. (1968). Finite State Continuous Time Markov Decision Processes with an Infinite Planning Horizon. *J. Math. Anal. Appl.* **22** 552–569.

[19] Romanovskii, I. V. (1972). The Turnpike Theorem for Semi-Markov Decision Processes. In: Linnik, Yu.V. *Theoretical Problems in Math Statistics.* American Math. Soc., Providence, R.I., 249–267. Translated from the Proc. Steklov Inst. Math. **111**.

[20] Romanovsky, I. (1973). On the Solvability of Bellman's Functional Equation for a Markovian Decision Process. *J. Math. Anal. Appl.* **42** 485–498.

[21] Schweitzer, P. J. (1965). *Perturbation Theory and Markovian Decision Processes.* Ph.D. dissertation, MIT; MIT Operations Research Center Report 15.

[22] ———. (1969). Perturbation Theory and Undiscounted Markov Renewal Programming. *Operations Res.* **17** 716–727.

[23] ———. (1968). Perturbation Theory and Finite Markov Chains. *J. Appl. Probability* **5** 401–413.

[24] ——— and Federgruen, A. (1976). *Asymptotic Value Iteration for Undiscounted Markov Decision Problems.* Math. Center Report BW 44/76 (To appear in *Math. Oper. Res.*)

[25] ——— and ———. (1976). *Foolproof Convergence in Multichain Policy Iteration.* I.B.M. Thomas J. Watson Research Center Report RC 5894 (To appear in *Math. Anal. Appl.*)

[26] Williams, A. (1970). Complementary Theorems for Linear Programming. *SIAM Rev.* **12** 135–137.

GRADUATE SCHOOL OF MANAGEMENT, UNIVERSITY OF ROCHESTER, ROCHESTER, NEW YORK 14627

MATHEMATISCH CENTRUM, 2DE BOERHAAVESTRAAT 49 A, AMSTERDAM, THE NETHERLANDS