RAIRO. RECHERCHE OPÉRATIONNELLE

PAUL J. SCHWEITZER

On undiscounted markovian decision processes with compact action spaces

RAIRO. Recherche opérationnelle, tome 19, n° 1 (1985), p. 71-86

<http://www.numdam.org/item?id=RO_1985__19_1_71_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 19, n° 1, février 1985, p. 71 à 86)
rairva$103

ON UNDISCOUNTED MARKOVIAN DECISION PROCESSES WITH COMPACT ACTION SPACES (*)

by Paul J. SCHWEITZER (1)

Abstract. — *We consider a multichain undiscounted, stationary, semi-Markovian decision process with finite state-space, compact action-spaces, and continuous rewards, holding times and transition probabilities. It is shown that the pair of coupled functional equations for the gain and values, which arise in the infinite-horizon formulation, possess a solution if and only if two conditions are met: at least one non-randomized maximal-gain policy exists, and the bias-vectors of all such maximum-gain policies are uniformly bounded above. The uniform-bound restriction is new and absent when the action-spaces are finite. Several sufficient conditions for solvability are discussed as well, plus extensions to higher-order optimality criteria.*

Keywords: Markovian decision processes; dynamic programming.

Résumé. — *Nous considérons un processus de décision semi-markovien, stationnaire, multichaîné, non actualisé, avec espace d'états fini, espace d'actions compact, et bénéfice, durée de service et probabilités de transition continus. Nous montrons que la paire d'équations fonctionnelles couplées pour le gain et la valeur, qui se présentent dans la formulation à horizon infini, possède une solution si et seulement si deux conditions sont satisfaites : au moins une politique de gain maximal non probabilisé existe, et les vecteurs de biais de toutes ces politiques de gain maximal ont une borne supérieure uniforme. La restriction de la borne supérieure uniformément bornée est nouvelle, et absente lorsque les espaces d'action sont finis. Plusieurs conditions suffisantes de résolubilité sont examinées, ainsi que des extensions à des critères d'optimalité d'ordre supérieur.*

1. INTRODUCTION

The functional equations of undiscounted, stationary, infinite-horizon semi-Markovian decision processes (MDPs) are [15]:

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

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

(*) Received March 1983.

(1) The Graduate School of Management, The University of Rochester, Rochester, N.Y. 14627, U.S.A.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/85/01 71 16/$ 3.60

where:

$$L(g, i) \equiv \left[k \in K(i) \left| \sum_{j=1}^{N} P_{ij}^{k} g_j = \max_{a \in K(i)} \sum_{j=1}^{N} P_{ij}^{a} g_j \right. \right].$$

Here N is the finite number of states, $K(i)$ is the non-empty, compact, set of actions in state i, and q_i^k and P_{ij}^k are the expected one-step reward and transition probability to state j if action k is selected in state i. The parameters q_i^k, P_{ij}^k, H_{ij}^k and $T_i^k \equiv \sum_{j=1}^{N} H_{ij}^k$ and sets $K(i)$ are assumed to satisfy:

$$q_i^k,\ P_{ij}^k,\ H_{ij}^k,\ T_i^k \text{ are continuous in } k, \tag{1.3a}$$

$$P_{ij}^k \geqq 0, \qquad \sum_{j=1}^{N} P_{ij}^k = 1, \qquad H_{ij}^k \geqq 0, \tag{1.3b}$$

$$H_{ij}^k = 0 \text{ whenever } P_{ij}^k = 0, \tag{1.3c}$$

$$0 < T_{\min} \leqq T_i^k, \tag{1.3d}$$

$$K(i) \text{ is a compact subset of a metric space.} \tag{1.3e}$$

Here T_i^k represents the mean holding time in state i if action k is chosen. The assumptions ensure that the one-step expected rewards q_i^k and holding times T_i^k are uniformly bounded, hence the gain rate vector for any policy [defined by (2.3), (2.4) below] is bounded. The assumptions also ensure that the maxima on the right-hand-sides of (1.1) and (1.2) (of continuous functions over compact sets) are actually achieved.

The $2N$ unknowns are the gain vector $g=[g_i]$ and relative value vector $v=[v_i]$. It is known [1, 15] that if a solution pair $\{g, v\}$ exists to (1.1), (1.2), then g is unique, g equals the maximal gain rate g^* defined by (2.5) below, and that any non-randomized policy which achieves the $2N$ maxima in (1.1), (1.2) attains the maximum gain rate g^*, i.e., that the set S_{MG} of non-randomized maximal-gain policies is not empty.

Our goal is investigating conditions for the existence of a solution to the functional equations (1.1), (1.2). Establishing existence is more difficult than in the discounted case [4, 8], where one is dealing with the fixed point of a contraction operator, so assumptions (1.3 *a*, *b*, *c*, *d*, *e*) suffice.

The first difficulty arises because the next-to-last paragraph implies that $S_{MG} \neq \varnothing$ is a necessary condition for the existence of a solution to (1.1),

(1.2). However, assumptions (1.3 *a*, *b*, *c*, *d*, *e*) are not sufficient to assure existence of a maximal-gain policy in the multi-chain case. A counter-example with $N=3$ is given in [16], section 4.3:

EXAMPLE 1: $K(1)=\{k \mid 0 \leqq k \leqq 1\}$ and $K(2)$, $K(3)$ are singletons, with parameters:

$$q(k)=[1,\ 7,\ 4], \qquad T(k)=[1,\ 1,\ 1],$$

$$P(k)=\begin{matrix} & 1-k & k-k^2 & k^2 \\ & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{matrix} = H(k),$$

$$P(k)^* \equiv \lim_{m \to \infty} \frac{P(k)+P(k)^2+\ldots+P(k)^m}{m} = \begin{cases} I, & k=0, \\ \begin{matrix} 0 & 1-k & k \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}, & 0<k\leqq 1, \end{cases}$$

$$g(k)=P(k)^* q(k)=\begin{cases} [1,\ 7,\ 4], & k=0, \\ [7-3k,\ 7,\ 4], & 0<k\leqq 1, \end{cases}$$

$$g^*=\sup_{0\leqq k\leqq 1} g(k)=[7,\ 7,\ 4],$$

where $k \in [0, 1]$ is the action chosen in state 1 and $g(k)$ is the gain rate vector corresponding to this choice. There is no k satisfying $g(k)=g^*$.

The second difficulty arises if the discontinuity in the chain structure causes the bias-vectors [defined by (2.6) below] to be unbounded, i.e., assumptions (1.3 *a*, *b*, *c*, *d*, *e*) and $S_{MG} \neq \varnothing$ together are still insufficient for existence of a solution to (1.1), (1.2). A counter-example with $N=2$ adapted from [16], section 4.4, is:

EXAMPLE 2:

$$K(1)=\{k \mid 0 \leqq k \leqq 1\}, \qquad K(2)=\text{singleton},$$

$$q(k)=[k,\ 0], \qquad T(k)=[1,\ 1],$$

$$P(k)=H(k)=\begin{bmatrix} 1-k^2 & k^2 \\ 0 & 1 \end{bmatrix},$$

$$P(k)^* \equiv \lim_{m \to \infty} \frac{P(k)+P(k)^2+\ldots+P(k)^m}{m} = \begin{cases} I, & k=0, \\ \begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix}, & 0<k\leqq 1, \end{cases}$$

$$g(k)=P(k)^* q(k)=[0,\ 0], \qquad 0 \leqq k \leqq 1,$$

where $k \in [0, 1]$ is the action chosen in state 1 and $g(k)$ is the corresponding gain rate vector. Every policy achieves the maximum gain rate $g^* = [0, 0]$, so $S_{MG} = K \equiv \overset{N}{\underset{i=1}{\times}} K(i)$, the full set of policies. However, the functional equations with $g = [0, 0]$ are *unsolvable* because (1.2) with $i = 1$ leads to the impossible condition $0 = \max_{0 \leqq k \leqq 1} [k + k^2 (v_2 - v_1)]$.

More detailed examination shows that the bias vector $w(k)$ (or relative values) of a policy using action k in state 1 satisfies $w(k)_2 - w(k)_1 = 1/k$ hence becomes *unbounded* as $k \downarrow 0$. It therefore appears necessary to require boundedness of bias vectors in order to ensure solvability of the functional equations.

It turns out that these two difficulties are the only ones. Our main result, theorem 1, shows that if (1.3 *a*, *b*, *c*, *d*, *e*) holds, then the functional equations (1.1), (1.2) possess a solution if and only if $S_{MG} \neq \varnothing$, and bias vectors for maximal-gain policies are uniformly bounded above.

This section concludes by describing the relationship between this result and previous published results. First, if each action-space $K(i)$ is *finite* rather than *compact*, solvability of (1.1), (1.2) may be established under (1.3 *b*, *c*, *d*) alone via the multichain version of Howard's policy iteration algorithm (PIA) [1, 7, 17], with the finite action-spaces assuring finite convergence [17]. The technical complications arise only when the action-spaces are not finite: one could envison initiating the PIA with a maximal-gain policy (assuming $S_{MG} \neq \varnothing$) and generating a sequence of maximum-gain policies with ever-increasing bias-vectors. An upper bound on these bias vectors is therefore needed to ensure convergence of the PIA.

Sheu and Farn [16] essentially establish solvability of (1.1), (1.2) under the assumption that $P(f)^*$ [defined by (2.2) below] is continuous for every policy $f \in K$. This works because it ensures (*see* theorem 2) that $S_{MG} \neq \varnothing$ and that bias vectors are uniformly bounded, hence meets our conditions. We also give, in theorem 2, equivalent conditions to this one and a stronger sufficient condition (4.1) as well:

$$\text{the positive } P_{ij}^k \text{ are bounded away from zero.} \tag{1.4}$$

In addition, theorem 3 uses a simple proof by lexicographic optimization to demonstrate, if $P(f)^*$ is continuous, not only that maximal-gain policies exist and maximal-bias polices exist [16] but also higher-order optima as well.

Note in this regard that while the continuity of $P(f)^*$ is a convenient sufficient condition, it is not the weakest one. Theorem 2 shows, for example,

that one may make the weaker assumption that the gain and bias vectors $g(f)$ and $w(f)$ are continuous; this would hold if every $q_i^k = 0$, even when the chain structure is discontinuous.

Note also that for the case where all components of the maximal gain rate are equal, [13] establishes $S_{MG} \neq \varnothing$ under assumption (1.3 *a*, *b*, *c*, *d*, *e*) alone and [12] establishes solvability of the functional equations under the additional assumption (1.4). This again is a specialization of the necessary and sufficient conditions given in theorem 1.

Finally, a treatment with more general state space is given in [10].

2. NOTATION AND PRELIMINARIES

0 denotes a vector or matrix all of whose components are zero. Vector or matrix inequalities $x \geqq y$ hold for every component. $\langle x; y \rangle$ denotes scalar product, and I denotes the identity matrix.

$K \equiv \underset{i=1}{\overset{N}{\times}} K(i)$ denotes the (compact) set of all (stationary, non-randomized) policies. A policy $f = (f(1), f(2), \ldots, f(N)) \in K$ consists of specification of the action $f(i) \in K(i)$ selected in state i for each i. Associated with each $f \in K$ are expected reward and holding time vectors:

$$q(f) \equiv [q(f)_i]_{i=1}^N \equiv [q_i^{f\,(i)}], \qquad T(f) \equiv [T(f)_i]_{i=1}^N \equiv [T_i^{f\,(i)}]$$

and non-negative matrices:

$$H(f) \equiv [H(f)_{ij}]_{i,\,j=1}^N \equiv [H_{ij}^{f\,(i)}],$$
$$P(f) \equiv [P(f)_{ij}]_{i,\,j=1}^N \equiv [P_{ij}^{f\,(i)}],$$

with row sums of $T(f)$ and unity, respectively.

For each $f \in K$, let $n(f)$ denote the number of subchains (closed, irreducible sets of states) of $P(f)$, which we label as $\{C(f, m), 1 \leqq m \leqq n(f)\}$. Let $\pi(f, m)$ be the unique equilibrium distribution of $P(f)$ on $C(f, m)$:

$$\pi(f, m) P(f) = \pi(f, m), \qquad \sum_{j \in C(f, m)} \pi(f, m)_j = 1,$$
$$\pi(f, m)_j > 0 \quad \text{if } j \in C(f, m), \quad = 0 \text{ elsewhere.} \tag{2.1}$$

Let $\varphi(f, m)_i$ be the probability of being ultimately absorbed in $C(f, m)$ if the Markov chain $P(f)$ starts in state i:

$$\varphi(f, m)_i \geqq 0, \qquad \sum_{m=1}^{n(f)} \varphi(f, m)_i = 1,$$

$$\varphi(f, m)_i = 1 \qquad \text{if} \quad i \in C(f, m).$$

For any $f \in K$, define the $N \times N$ (transition probability) matrix:

$$P(f)^* = \lim_{m \to \infty} \frac{1}{m} \sum_{t=1}^{m} P(f)^t, \tag{2.2}$$

It always exists, and satisfies [3], pp. 175-183, [11]:

$$P(f)^*_{ij} = \sum_{m=1}^{n(f)} \varphi(f, m)_i \pi(f, m)_j, \qquad 1 \leqq i, j \leqq N.$$

The *gain-rate vector* for policy $f \in K$ is:

$$g(f)_i \equiv \sum_{m=1}^{n(f)} \varphi(f, m)_i g(f, m), \qquad 1 \leqq i \leqq N, \tag{2.3}$$

$$= \text{expected reward per unit time starting in state } i,$$

where:

$$g(f, m) \equiv \frac{\langle \pi(f, m);\, q(f) \rangle}{\langle \pi(f, m);\, T(f) \rangle}, \qquad 1 \leqq m \leqq n(f), \tag{2.4}$$

$$= \text{gain rate for chain } C(f, m),$$

$$g(f)_i = g(f, m) \quad \text{for all } i \in C(f, m).$$

The *maximal gain-rate vector* g^* is defined by:

$$g^*_i \equiv \sup_{f \in K} g(f)_i, \qquad 1 \leqq i \leqq N. \tag{2.5}$$

Any policy $f \in K$ achieving all N suprema is called *maximal-gain*, and:

$$S_{MG} \equiv \{ f \in K \,|\, g(f) = g^* \},$$

denotes the set of all maximal-gain policies.

For each $f \in K$, define the *fundamental matrix* $Z(f)$ and *bias-vector* $w(f)$ by:

$$Z(f) \equiv [I - P(f) + P(f)^*]^{-1},$$
$$w(f) \equiv Z(f)[q(f) - H(f)g(f)]. \tag{2.6}$$

These always exist [11, 15] and have the properties:

$$w(f) = q(f) - H(f)g(f) + P(f)w(f), \tag{2.7}$$
$$P(f)^* w(f) = \mathbf{0}. \tag{2.8}$$

An alternate characterization of $g(f)$ and $w(f)$ is as the unique solution to (2.7), (2.8) and

$$g(f) = P(f)g(f). \tag{2.9}$$

3. MAIN RESULT

THEOREM 1: *Assume* (1.3 *a, b, c, d, e*) *holds. Then* (1.1), (1.2) *possess a solution pair* $\{g, v\}$ *if and only if* (3.1) *and* (3.2) *hold:*

$$S_{MG} \neq \varnothing, \tag{3.1}$$
$$\{w(f);\ f \in S_{MG}\} \text{ is uniformly bounded above.} \tag{3.2}$$

If a solution pair exists, then any policy achieving all $2N$ *maxima in* (1.1), (1.2) *lies in* S_{MG}.

Proof: Assume first that a solution pair exists. Then the last assertion is in, for example [15], theorem 4.1 *c*, and implies (3.1). In addition, the reasoning for [15], equation (4.1), implies, for any $f \in S_{MG}$, that:

$$v_i \geqq [w(f) + P(f)^* v]_i, \qquad 1 \leqq i \leqq N,$$

or:

$$w(f)_i \leqq v_i - \min_{1 \leqq j \leqq N} v_j, \qquad 1 \leqq i \leqq N,$$

which confirms (3.2).

Conversely, assume (3.1), (3.2) hold and seek a solution pair to (1.1), (1.2). Recall that the PIA, if started at a policy $f^0 \in S_{MG}$, using bias vectors as its choice of relative values, and breaking ties by retaining the previously-used alternative whenever possible, will generate a sequence of policies $\{f^n\}$

satisfying:

$$g(f^n)=g^*, \qquad n=0, 1, 2, 3, \ldots, \tag{3.3}$$

$$w(f^{n+1})\geqq w(f^n), \qquad n=0, 1, 2, 3, \ldots, \tag{3.4}$$

$$q(f^{n+1})-H(f^{n+1})g^*+P(f^{n+1})w(f^n)=Q\,w(f^n)\geqq w(f^n) \tag{3.5}$$

where $Q : E^N \to E^N$ is defined by:

$$Q\,x_i \equiv \max_{k\in L(g^*,\ i)}\left[q_i^k-\sum_{j=1}^{N}H_{ij}^k g_j^*+\sum_{j=1}^{N}P_{ij}^k x_j\right], \qquad 1\leqq i\leqq N.$$

i. e., the PIA will improve bias vectors if it cannot improve gain vectors [17], theorem 6. Insert (3.4) into (3.5) to obtain:

$$w(f^{n+1})=q(f^{n+1})-H(f^{n+1})g^*+P(f^{n+1})w(f^{n+1})$$
$$\geqq Q\,w(f^n)\geqq w(f^n). \tag{3.6}$$

Each $f^n\in S_{MG}$ due to (3.3), hence the monotone sequence $\{w(f^n)\}$ is bounded above via (3.2) and has a limit $\bar{w}$. Then (3.6) has a limit $\bar{w}=Q\,\bar{w}$ as $n\to\infty$, which is (1.2).

To establish (1.1), note f^n lies in both S_{MG} and $\underset{i=1}{\overset{N}{\times}} L(g^*, i)$, so:

$$\max_{k\in K(i)}\sum_{j=1}^{N}P_{ij}^k g_j^*=[P(f^n)g^*]_i=P(f^n)g(f^n)_i=g(f^n)_i=g_i^*,$$

for all i, with the penultimate equality coming from (2.9). □

4. OTHER EXISTENCE CONDITIONS

This section provides other conditions that ensure the solvability of (1.1), (1.2). These conditions are stronger than (3.1), (3.2) but can be easier to verify. The logical relationships among these sufficient conditions are laid out in theorem 2, and the ensuing remarks relate them to previous published results.

THEOREM 2: *Assume* (1.3 *a, b, c, d, e*) *holds. Then,*

1. *The following conditions each ensure the existence of a solution pair to* (1.1), (1.2):

there exists a number $a>0$ *such that for each triple* (i, j, k),

$$\text{with } 1\leqq i, j\leqq N \text{ and } k\in K(i), \text{ either } p_{ij}^k=0 \text{ or } p_{ij}^k\geqq a. \quad (4.1)$$

every transition probability matrix $P(f)$, $f\in K$,

$$\text{has one subchain (closed, communicating set of states)}, \quad (4.2)$$

$$n(f) \text{ is continuous on } f\in K, \quad (4.3)$$

$$P(f)^* \text{ is continuous on } f\in K, \quad (4.4)$$

$$Z(f) \text{ is continuous on } f\in K, \quad (4.5)$$

$$g(f) \text{ and } w(f) \text{ are continuous on } f\in K, \quad (4.6)$$

$$\text{both (3.1), (3.2) hold (recall theorem 1).} \quad (4.7)$$

2. *Conditions* (4.1) *and* (4.2) *each imply* (4.3).

3. *Conditions* (4.3), (4.4) *and* (4.5) *are equivalent and each implies* (4.6).

4. *Condition* (4.6) *implies* (4.7).

Proof: We show (4.1) and (4.2) each imply (4.3); (4.3) and (4.4) are equivalent; (4.4) and (4.5) are equivalent; (4.4) and (4.5) imply (4.6); and finally (4.6) implies (4.7). From theorem 1, (4.7) ensures solvability of (1.1), (1.2).

Conditions (4.1) implies (4.3) because two neighboring policies will either both have $P_{ij}=0$ or both have $P_{ij}\geqq a$, hence their tpm's will have identical chain structure.

Condition (4.2) implies $n(f)=1$ for all f, hence (4.3) holds.

Conditions (4.3) and (4.4) are equivalent, due to [11], theorem 5. They are shown equivalent to (4.5) as follows. If $P(f)^*$ is continuous, so is $I-P(f)+P(f)^*$ and so is the latter's inverse, $Z(f)$. Conversely if $Z(f)=[I-P(f)+P(f)^*]$ is continuous, so is $Z(f)^{-1}-I+P(f)=P(f)^*$.

Conditions (4.3) and (4.4) imply that, in a neighborhood of any f, the chain structure of $P(\ )$ is unchanged, hence the gain rate on each subchain will be continuous and so will be the absorption probabilities from transient states into the subchains. Hence $g(f)$ as defined by (2.3), (2.4), is continuous in f. Recall $w(f)=Z(f)[q(f)-H(f)g(f)]$ where all terms on the right are continuous in f, since $Z(f)$ and $g(f)$ are when (4.3), (4.4) and (4.5) hold. Hence (4.6) holds.

If (4.6) holds, select a policy $\tilde{f}$ which lexicographically maximizes $g(f)_1$, $g(f)_2, \ldots, g(f)_N, w(f)_1, \ldots, w(f)_N$ over $f \in K$, [14], i.e.,

$$g(\tilde{f})_1 \equiv \max\{g(f)_1 \mid f \in K\} \qquad (=g_1^*),$$

$$g(\tilde{f})_i \equiv \max\{g(f)_i \mid f \in K \text{ with } g(f)_j = g(\tilde{f})_j \text{ for } 1 \leqq j \leqq i-1\},$$

$$2 \leqq i \leqq N,$$

$$w(\tilde{f})_1 \equiv \max\{w(f)_1 \mid f \in K \quad \text{with} \quad g(f) = g(\tilde{f})\},$$

$$w(\tilde{f})_i \equiv \max\{w(f)_i \mid f \in K$$

$$\text{with } g(f) = g(\tilde{f}) \text{ and } w(f)_j = w(\tilde{f})_j \text{ for } 1 \leqq j \leqq i-1\},$$

$$2 \leqq i \leqq N.$$

These are all maxima of continuous functions over compact sets, hence the maxima are achieved. Initiate the PIA with policy $\tilde{f}$. As in [14], the PIA cannot improve $g(\tilde{f})$, nor can it improve $w(\tilde{f})$. Hence it terminates with the pair $\{g(\tilde{f}), w(\tilde{f})\}$ solving the functional equations (1.1), (1.2), and consequently $g(\tilde{f}) = g^*$. Thus $\tilde{f} \in S_{MG} \neq \varnothing$. □

REMARKS ON THEOREM 2: 1. If the action-spaces are finite, then K is finite and $n(f)$, $g(f)$ and $w(f)$ are "continuous" on K. Since conditions (4.3) to (4.6) are met, (1.1), (1.2) is solvable.

2. The non-solvability of (1.1), (1.2) for example 2 can now be traced back to the unboundedness of both $w(f)$ and $Z(f)$, the discontinuity in $P(f)^*$, and the change in $P(f)_{12}$ from >0 to $=0$ as $k \downarrow 0$.

3. For the special case where all components of g^* are equal, [12] established solvability of the functional equations under condition (4.1) *via* the Leray-Schauder fixed point theorem. The present proof is simpler. For this special case, $S_{MG} \neq \varnothing$ holds without any assumptions beyond (1.3) [13], and the role of (4.1) in [12] is merely to prove that $Z(f)$ is *bounded uniformly* in $f \in K$. Since $q(f)$, $H(f)$, and $T(f)$ are bounded uniformly in f, so are:

$$|g(f)_i| \leqq \max_{f \in K} \max_{1 \leqq j \leqq N} |q(f)_j| / T_{\min}$$

and:

$$w(f) \equiv Z(f)[q(f) - H(f)g(f)].$$

The solvability of (1.1), (1.2) again can be demonstrated by the PIA argument in the last part of the proof of theorem 1, since only the *boundedness*, not continuity, of $w(f)$ is required.

4. Condition (4.2) requires only $n(f) = 1$, not that $P(f)$ be aperiodic, hence is weaker than the assumption in [16] that every $P(f)$ is ergodic. The

condition $n(f)=1$ forces continuous chain structure, which suffices. Condition (4.2), along with the additional assumption that the maximizing policy in $Q\,x$ is *unique*, was used in [19] to establish convergence of the policy iteration algorithm to a solution of the functional equations.

5. Hordijk [5, 6] was apparently the first to observe that the continuity of $P(f)^*$ suffices to establish existence of a maximal-gain policy. Sheu [16] later showed it suffices to establish existence of a maximal-bias policy. Theorem 3 below shows it in fact suffices to establish existence of an optimal policy for any of the discount-related higher-order optimality criteria.

6. If K is connected (e. g., if convex), then the continuity of $n(f)$ on K reduces to $n(f)=$constant. In general, K decomposes into one or more disjoint connected components, with $n(f)$ constant on each component.

5. HIGHER-ORDER OPTIMALITY CRITERIA

Consider a discounted Markov renewal program whose interest rate $\alpha>0$ goes to zero [2, 9, 17, 18]. The functional equations to be solved are:

$$v(\alpha)_i = \max_{k \in K(i)} \left[q(\alpha)_i^k + \sum_{j=1}^{N} \left[\int_0^\infty dQ(t)_{ij}^k e^{-\alpha t} \right] v(\alpha)_j \right], \qquad 1 \leqq i \leqq N, \tag{5.1}$$

where $q(\alpha)_i^k$ is the expected discounted one-step reward and $Q(t)_{ij}^k$ is the joint probability that the next state is j and that the holding time in state i will not exceed t, given entry into state i and selecting action k. $v(\alpha)_i$ is the maximum expected discounted reward over an infinite horizon, starting from state i.

Heuristically, we might expect that the following two Maclaurin series and Laurent series converge for small $\alpha>0$:

$$q(\alpha)_i^k = \sum_{n=0}^{\infty} q_i^{k(n)} \alpha^n, \tag{5.2}$$

$$\int_0^\infty qQ(t)_{ij}^k e^{-\alpha t} = \sum_{n=0}^{\infty} S_{ij}^{k(n)} \alpha^n, \tag{5.3}$$

$$v(\alpha)_i = \sum_{n=-1}^{\infty} v_i^{(n)} \alpha^n. \tag{5.4}$$

Heuristically, if the maximization over k in (5.1) is replaced by specification of a fixed policy $k=f(i)$, leading to a set of equations for a vector $v(f, \alpha)$, then use of the first $r+1$ terms of (5.2), (5.3) is expected to yield the first r terms of:

$$v(f, \alpha)=\left[I-\int_0^{\infty} dQ(t, f)\, e^{-\alpha t}\right]^{-1} q(f, \alpha)$$

$$=\frac{v(f)^{(-1)}}{\alpha}+v(f)^{(0)}+v(f)^{(1)}\alpha+v(f)^{(2)}\alpha^2+\ldots,$$

where the $v(f)^{(n)}$ are exhibited below. Then the characterization $v(\alpha)_i=\sup_{f\in K} v(f, \alpha)_i$ implies, for small α, that one lexicographically optimizes over $v(f)^{(-1)}$, then over $v(f)^{(0)}$, then over $v(f)^{(1)}$, etc.

The coefficients of α^{-1}, α^0, α^1, ..., α^r in (5.1) are then given, for sufficiently small α, by the following set of $r+2$ *nested functional equations*, obtained by inserting (5.2), (5.3), (5.4) into (5.1):

$$v_i^{(n)}=\max_{k\in L(n-1;\, i)}\left[q_i^{k\,(n)}+\sum_{j=1}^{N}\sum_{l=0}^{n+1} S_{ij}^{k\,(l)} v_j^{(n-l)}\right], \tag{5.5}$$

$$1\leqq i\leqq N, \quad n=-1, 0, 1, 2, \ldots, r,$$

where:

$$L(-2;\, i)\equiv K(i),$$

$$L(n;\, i)\equiv \text{set of maximizing } k\text{'s in (5.5), for } n\geqq -1,$$

$$q_i^{k\,(-1)}\equiv 0.$$

When $r=0$, the two functional equations reduce to (1.1), (1.2) upon making the identifications $v^{(-1)}=g$, $v^{(0)}=v$, $S_{ij}^{k\,(0)}=P_{ij}^k$, $S_{ij}^{k\,(1)}=-H_{ij}^k$, $q^{(0)}=q$.

Using the same techniques as above, we can establish existence of a solution to the $r+2$ nested functional equations, along with the uniqueness characterization:

$$v_i^{(-1)}=\max_{f\in K}[v(f)_i^{(-1)}], \qquad 1\leqq i\leqq N, \tag{5.6a}$$

$$v_i^{(n)}=\max[v(f)_i^{(n)};\, f\in K \text{ with } v(f)^m=v^{(m)} \text{ for } -1\leqq m\leqq n-1], \tag{5.6b}$$

$$0\leqq n\leqq r-1, \quad 1\leqq i\leqq N,$$

[where $v(f)^{(n)}$ will be defined below] as well as existence of a policy $\tilde{f}$ achieving all maxima in $v^{(-1)}, v^{(0)}, \ldots, v^{(r-1)}$. The need for such a policy and such a lexicographic optimization was discussed above in a heuristic way.

We shall assume, in parallel with (1.3), that:

$q_i^{k(n)}$ and $S_{ij}^{k(n)}$ are continuous in k for

$$1 \leqq i, j \leqq N, \quad -1 \leqq n \leqq r+1, \quad k \in K(i), \tag{5.7a}$$

$$P_{ij}^k = S_{ij}^{k(0)} \quad \text{and} \quad H_{ij}^k = -S_{ij}^{k(1)} \text{ satisfy (1.3b, c, d)}, \tag{5.7b}$$

$$\text{(1.3e) holds.} \tag{5.7c}$$

For fixed policy f, we let $q(f)^{(n)}$ and $S(f)^{(n)}$ denote the vector $[q_i^{k=f(i)(n)}]$ and matrix $[S_{ij}^{k=f(i)(n)}]$, and let the vectors $v(f)^{(-1)}, v(f)^{(0)}, v(f)^{(1)}, \ldots, v(f)^{(r)}$ denote the solution to the associated specialization of (5.5):

$$v(f)^{(n)} = q(f)^{(n)} + \sum_{l=0}^{n+1} S(f)^{(l)} v^{(n-l)}, \qquad -1 \leqq n \leqq r. \tag{5.8}$$

LEMMA: *Under the above assumptions,* (5.8) *possesses a solution for any* $f \in K$, *with* $v(f)^{(n)}$ *unique for* $-1 \leqq n \leqq r-1$ *but not unique for* $n=r$. *In addition,* $v(f)^{(n)}$ *for* $0 \leqq n \leqq r-1$ *are* continuous *functions of* f *if* $P(f)^*$ *is. Explicitly, for* $1 \leqq i \leqq N$:

$$v(f)_i^{(-1)} = \sum_{m=1}^{n(f)} \varphi(f, m)_i \frac{\langle \pi(f, m); q(f)^{(0)} \rangle}{\langle \pi(f, m); T(f) \rangle}, \tag{5.9a}$$

$$v(f)_i^{(n)} = Z(f)\left[q(f)^{(n)} + \sum_{l=1}^{n+1} S(f)^l v^{(n-l)} \right]_i + \sum_{m=1}^{n(f)} \varphi(f, m)_i \frac{\langle \pi(f, m); x(f)^{(n)} \rangle}{\langle \pi(f, m); T(f) \rangle}, \tag{5.9b}$$

$$0 \leqq n \leqq r,$$

where:

$$x(f)^{(n)} \equiv \begin{cases} q(f)^{(n+1)} + \sum_{l=2}^{n+2} S(f)^{(l)} v^{(n+1-l)} \\ -H(f) Z(f)\left[q(f)^{(n)} + \sum_{l=1}^{n+1} S(f)^l v(f)^{(n-l)} \right], \\ 0 \leqq n \leqq r-1, \\ \text{arbitrary}, \quad n=r \end{cases}$$

and $Z(f)$ is the fundamental matrix for $P(f)$.

The proof simply involves left multiplication of (5.8) by $Z(f)$, using $Z(f)[I-P(f)]=I-P(f)^*$. The unknows $\langle \pi(f, m); v(f)^{(n)} \rangle$ which appear can be evaluated by taking the scalar production of $\pi(f, m)$ with equation (5.8) for $n+1$, and recalling (1.3*c*).

Equations (5.9) specify $v(f)^{(-1)}$ explicity and the remaining v's *recursively*. The assumptions assure that $n(f)$, $q(f)^{(n)}$, $S(f)^{(n)}$, $Z(f)$, $\varphi(f, m)$, $\pi(f, m)$, $T(f)$ and $x(f)^{(n)}$ are all *continuous* functions of f. Therefore (5.9) implies that the v's are continuous in f. □

THEOREM 3: *Fix $r \geqq 0$. Under assumptions* (5.7) *and* (4.3), (4.4), (4.5), *there exists a solution to the $r+2$ nested functional equations* (5.5), *with the characterization in* (5.6) *of $\{v^{(-1)}, v^{(0)}, \ldots, v^{(r-1)}\}$. Furthermore, there exists a policy $\tilde{f} \in K$ which lexicographically optimizes these $r+1$ vectors:*

$$v^{(-1)} = v(\tilde{f})^{(-1)} = \sup\{v(f)^{(-1)} \mid f \in K\},$$

$$v^{(n)} = v(\tilde{f})^{(n)} = \sup\{v(f)^{(n)} \mid f \in K \text{ with } v(f)^{(m)} = v(\tilde{f})^{(m)} \text{ for } -1 \leqq m \leqq n-1\} \text{ for } 0 \leqq n \leqq r-1.$$

Proof: The extended PIA [2, 9, 17, 18] for the set of $r+2$ nested functional equations attempts to lexicographically optimize the first $r+1$ vectors. By initializing the extended PIA with a policy $\tilde{f}$ that lexicographically optimizes the successive members of the ordered set:

$$\{v(f)_1^{(-1)}, \ldots, v(f)_N^{(-1)}, v(f)_1^{(0)}, \ldots, v(f)_N^{(0)}, v(f)_1^{(1)}, \ldots, v(f)_N^{(1)}, \ldots, v(f)_1^{(r)}, \ldots, v(f)_N^{(r)}\}$$

one obtains one-step convergence of the PIA, just as in theorem 2, to a solution of the $r+2$ functional equations. It is then straightforward to show that $v^{(-1)}, \ldots, v^{(r-1)}$ are maximized lexicographically. □

REMARK: When $r=1$, the 3 functional equations are given in [14] and consist of (1.1), (1.2) plus a third one. The lexicographic optimization was described in [14]. Any solution to these 3 functional equations will satisfy:

$$v^{(-1)} = g^* = \max g(f) \qquad \text{and} \qquad v^{(0)} = w^* = \max\{w(f) \mid f \in S_{MG}\}.$$

Any policy achieving all optima in the 3 functional equations will be bias-optimal (or 1-optimal). This establishes the existence of a 1-optimal policy by a different approach than the one in [16]. Note that, in the Markov renewal case, the bias vector $v(f)^{(0)} = w(f)$ is now more complex than (2.6),

but reduces to (2.6) upon special choices of $q(f)^{(1)}$ and $S(f)^{(2)}$:

$$v(f)^{(0)} = Z(f)[q(f)^{(0)} - H(f)g(f)] + \sum_{m=1}^{n(f)} \varphi(f, m) \langle \pi(f, m);\ q(f)^{(1)} + S(f)^{(2)} g(f) - H(f)Z(f)[q(f)^{(0)} - H(f)g(f)] \rangle / \langle \pi(f, m);\ T(f) \rangle .$$

6. SUMMARY

The main result of this paper is the exhibition of necessary and sufficient conditions (3.1), (3.2) for solvability of the pair of functional equations (1.1), (1.2) under the assumptions of continuous data $\{q(f), H(f), P(f)\}$ and compact action sets. These conditions, especially the upperboundedness of bias vectors, are new, and are weaker than the sufficient condition (4.4) in [16]. We also give equivalent conditions (4.3), (4.5) to (4.4) and new stronger conditions (4.1) and (4.2) as well. The ergodic assumption in [16] is relaxed. It is also pointed out that (4.4) suffices for the existence of maximal-gain policies, maximal-bias policies, *and more*, because it ensures the continuity of $g(f)$, $w(f)$, and more. This provides a painless existence proof for the higher-order optimality criteria.

In addition, the method of proof . . . initiating the (multi-vector) policy iteration algorithm with a lexicographically-optimal policy . . . may be of independent interest.

Finally, the non-existence of solutions in pathological cases is traced to its source . . . a discontinuous chain structure as a transition probability drops from positive level to zero, which in turn causes a discontinuity in $P(f)^*$ and unboundedness in both $Z(f)$ and $w(f)$. In particular, if every $P(f)$ is unichained $[n(f) = 1]$, then $P(f)^*$, $Z(f)$, $g(f)$ and $w(f)$ are continuous for $f \in K$, and solvability is assured.

REFERENCES

1. E. V. Denardo and B. Fox, *Multichain Markov Renewal Programs*, S.I.A.M. J. Appl. Math., Vol. 16, 1968, pp. 468-487.
2. E. V. Denardo, *Markov Renewal Programs with Small Interest Rates*, Ann. Math. Statist., Vol. 42, 1971, pp. 477-496.
3. J. Doob, *Stochastic Processes*, Wiley, New York, 1953.

4. N. FURUKAWA, *Markovian Decision Processes with Compact Action Spaces*, Ann. Math. Statist., Vol. 43, 1972, pp. 1612-1622.
5. A. HORDIJK, *A Sufficient Condition for the Existence of an Optimal Policy with Respect to the Average Cost Criterion in Markovian Decision Processes*, Transactions Sixth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes. Academia, Prague, 1971, pp. 263-274.
6. A. HORDIJK, *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tract, Vol. 51, Amsterdam, 1974.
7. R. A. HOWARD, *Dynamic Programming and Markov Processes*, Wiley, New York, 1960.
8. A. MAITRA, *Discounted Dynamic Programming on Compact Metric Spaces*, Sankhya, Sec. A, Vol. 30, 1968, pp. 211-216.
9. B. L. MILLER and A. F. VEINOTT, Jr., *Discrete Dynamic Programming with a Small Interest Rate*, Ann. Math. Statist., Vol. 40, 1969, pp. 366-370.
10. M. SCHAL, *Stationary Policies in Dynamic Programming Models Under Compactness Assumptions*, Math. of O.R., Vol. 8, 1983, pp. 366-372.
11. P. SCHWEITZER, *Perturbation Theory and Finite Markov Chains*, J. Appl. Prob., Vol. 5, 1968, pp. 401-413.
12. P. J. SCHWEITZER, *On the Solvability of Bellman's Functional Equations for Markov Renewal Programming*, J. Math. Anal. Appl, Vol. 96, 1983, pp. 13-23.
13. P. J. SCHWEITZER, *On the Existence of Relative Values for Undiscounted Markovian Decision Processes with a Scalar Gain Rate*, University of Rochester, Graduate School of Management, Working Paper Series No. QM8225, December 1982. To appear in J. Math. Anal. Appl.
14. P. J. SCHWEITZER, *Solving MDP Functional Equations by Lexicographic Optimization*, Revue Française d'Automatique et de Recherche Operationnelle, Vol. 16, 1982, pp. 91-98.
15. P. J. SCHWEITZER and A. FEDERGRUEN, *The Functional Equations of Undiscounted Markov Renewal Programming*, Math. of Operations Research, Vol. 3, 1978, pp. 308-322.
16. S. S. SHEU and K.-J. FARN, *A Sufficient Condition for the Existence of a Stationary 1-Optimal Plan in Compact Action Markovian Decision Processes*, in *Recent Developments in Markov Decision Processes*, R. HARTLEY, L. C. THOMAS and D. J. WHITE Eds., Academic Press, London, 1980, pp. 111-126.
17. A. F. VEINOTT, Jr., *On Finding Optimal Policies in Discrete Dynamic Programming with no Discounting*, Ann. Math. Statist., Vol. 37, 1966, pp. 1284-1294.
18. A. F. VEINOTT, Jr., *Discrete Dynamic Programming with Sensitive Discount Optimality Criteria*, Ann. Math. Statist., Vol. 40, 1969, pp. 1635-1660.
19. A. HORDIJK and M. L. PUTERMAN, *On the Convergence of Policy Iteration in Finite State Average Reward Markov Decision Processes; the Unichain Case*, Working Paper No. 948, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, British Columbia, May 1983.