stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE
(DEPARTMENT OF OPERATIONS RESEARCH)

BW 104/79 MEI

A. FEDERGRUEN, P.J. SCHWEITZER

A SURVEY OF ASYMPTOTIC VALUE-ITERATION FOR UNDISCOUNTED MARKOVIAN DECISION PROCESSES

Preprint

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

AMS subject classification scheme (1979): 90C47

A SURVEY OF ASYMPTOTIC VALUE-ITERATION FOR UNDISCOUNTED MARKOVIAN DECISION PROCESSES

A. FEDERGRUEN
University of Rochester; Mathematisch Centrum, Amsterdam

P.J. SCHWEITZER
University of Rochester; Mathematisch Centrum, Amsterdam

ABSTRACT

This paper reviews the asymptotic behavior of undiscounted value-iteration in Markov Decision Problems with finite state and action spaces. The asymptotic results concern both the value functions and the sets of optimizing policies as the length of the planning period tends to infinity.

KEY WORDS & PHRASES: *value-iteration, rate of convergence, maximal gain policies, asymptotic behavior, optimality equations, data transformations.*
NOTE: This report has also been published as Working Paper No. 7833 of the Graduate School of Management, University of Rochester, December, 1978. It is not for review but will be submitted for publication in a journal.

1. Introduction

The value-iteration equations for finite, stationary Markov decision processes (MDPs) are

$$(1.1) \qquad v(n)_i = \max_{k \in K(i)} \left[q_i^k + \beta \sum_{j=1}^{N} P_{ij}^k v(n-1)_j \right] ; \; 1 \le i \le N , \quad n = 1,2,3,\ldots$$

[Howard (1960)], where N is the finite number of states, K(i) is the finite, non-empty, set of actions available in state i, β is the one-period discount factor, and $v(0)_i$ denotes the scrap value of state i upon termination. We refer to $0 \le \beta < 1$ and $\beta = 1$ as the discounted and undiscounted cases, respectively. If action $k \in K(i)$ is chosen upon entrance to state i, an immediate expected reward q_i^k is earned (where every q_i^k is finite) and P_{ij}^k denotes the probability that the next state will be j ($P_{ij}^k \ge 0$, $\sum_{j=1}^{N} P_{ij}^k = 1$). Equation (1.1) results from Bellman's principle of optimality [Bellman (1957)] with $v(n)_i$ denoting the maximum expected cumulative return in an n-period process starting from state i.

The following notation will be employed. We let $S_R = \{[f_{ik}] | f_{ik} \ge 0, \sum_{k \in K(i)} f_{ik} = 1\}$ denote the set of all randomized stationary policies. Here f_{ik} is the probability action k is chosen whenever entering state i. A pure (non-randomized) policy has each $f_{ik} = 0$ or 1, since it associates a single action $k = f(i)$ with each $i \in \Omega$. $S_P = X_i K(i)$ will represent the finite subset of pure (stationary) policies.

With each policy $f \in S_R$, we associate an N-component reward vector q(f) and an N×N-transition probability matrix P(f):

$$(1.2) \qquad q(f)_i = \sum_{k\in K(i)} f_{ik} q_i^k \quad ; \quad 1 \le i \le N$$

$$P(f)_{ij} = \sum_{k\in K(i)} f_{ik} P_{ij}^k \quad ; \quad 1 \le i,j \le N \quad .$$

We rewrite (1.1) as

$$(1.3) \qquad v(n) = Tv(n-1) = T^n v(0) \quad ; \quad n = 1,2,\ldots$$

where the operator $T:E^N \to E^N$ is defined by

$$(1.4) \qquad Tx_i = \max_{k\in K(i)} [q_i^k + \beta \sum_j P_{ij}^k x_j] \quad , \quad 1 \le i \le N$$

and where T^n represents the n-fold application of the T-operator.

Finally, for any $\varepsilon \ge 0$, let

$$S(n) = \{f \in S_p | v(n) = q(f) + \beta P(f)v(n-1)\}$$

$$S(n,\varepsilon) = \{f \in S_p | q(f) + \beta P(f)v(n-1) \ge v(n) - \varepsilon \underline{1}\}$$

denote, respectively, the set of pure policies which attain (or come within ε of) $v(n)$, when the remaining planning period consists of n periods. Here $\underline{1}$ represents the N-vector with all components equal to unity.

The present survey describes the undiscounted case, with emphasis on the basic properties of $\{v(n)\}_{n=1}^{\infty}$ (section 2), its dependence upon the scrap value vector $v(0)$ (section 3), and the asymptotic behavior of the sets of

optimal and ε-optimal policies $\{S(n)\}_{n=1}^{\infty}$ and $\{S(n,\varepsilon)\}_{n=1}^{\infty}$ (section 4). Section 5 discusses Lyapunov functions and other techniques for bounding a solution pair to the optimality equations arising in this model, while section 6 describes uses of data-transformations to simplify computations in the infinite horizon case. In each section, differences between the undiscounted and discounted cases are pointed out. Some of the material in sections 2-4 overlaps our previous survey on value-iteration (cf. Federgruen and Schweitzer (1977a)), to which the reader is referred for additional details.

The following notation will be employed: For any $x \in E^N$, let

$$x_{max} = \max_i x_i \; ; \quad x_{min} = \min_i x_i \qquad \text{and}$$

$$|x|_\infty = \max_i |x|_i \quad .$$

In addition, we will use the quasi-norm sp[x], where $sp[x] = x_{max} - x_{min}$ (cf. Bather (1973)).

2. The Asymptotic Behavior of v(n)

The following additional notation is required for studying the asymptotic behavior of v(n). For any pure or randomized policy f, let the N×N-matrix $\Pi(f)$ denote the Cesaro limit of the sequence $\{P^n(f)\}_{n=1}^{\infty}$, and let $R(f) = \{i | \Pi(f)_{ii} > 0\}$ represent the set of recurrent states for P(f). Finally, we associate with each $f \in S_R$ its gain rate vector

$$g(f) = \lim_{M\to\infty} \frac{1}{M+1} \sum_{n=0}^{M} P^n(f)q(f) = \Pi(f)q(f) \quad .$$

Denote the maximal gain rate vector by g^*, where

$$(2.1) \qquad g_i^* = \sup_{f\in S_R} g(f)_i \ , \quad 1 \le i \le N \quad .$$

It is known (cf. Derman (1970)) that a pure policy achieves the N maxima in (2.1) simultaneously. Accordingly, let

$$S_{RMG} = \{f \in S_R | g(f) = g^*\} \quad \text{and} \quad S_{PMG} = \{f \in S_P | g(f) = g^*\}$$

denote the set of all randomized maximal gain policies and the set of all pure maximal gain policies, and define

$$R^* = \{i | i \in R(f) \text{ for some } f \in S_{RMG}\} \quad .$$

It is known that $R^* = \{i | i \in R(f) \text{ for some } f \in S_{PMG}\}$, i.e., any state that is recurrent under a randomized maximal gain policy will also be

recurrent under at least one pure maximal gain policy. It is also known (cf. Schweitzer and Federgruen (1977a)) that the set

$$S^*_{RMG} = \{f \in S_{RMG} | R(f) = R^*\} \neq \phi \quad .$$

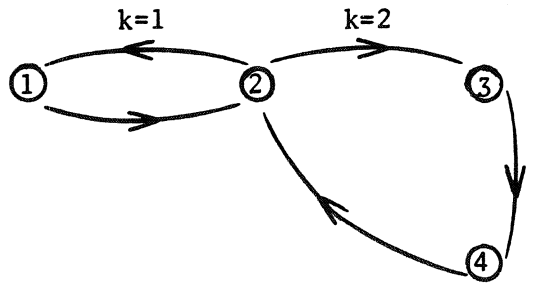
Note that randomization is essential here: there generally does not exist a pure maximal gain policy whose set of recurrent states is R^*. To illustrate this, consider the following 4-state example.

Example 1

$N = 4$; $K(1) = K(3) = K(4) = \{1\}$; $K(2) = \{1,2\}$.

$q_i^k \equiv 0$, $g^* = (0,0,0,0)$, $S_{RMG} = S_R$, $R^* = \{1,2,3,4\}$

i	k	P^k_{i1}	P^k_{i2}	P^k_{i3}	P^k_{i4}
1	1	0	1	0	0
2	1	1	0	0	0
	2	0	0	1	0
3	1	0	0	0	1
4	1	0	1	0	0



in which only state 2 has multiple alternatives.

Both pure policies are maximal gain, with recurrent states {1,2} if k = 1 is chosen and {2,3,4} if k = 2 is chosen. No pure policy has $R^* = \{1,2,3,4\}$ as its set of recurrent states, but any randomized policy which uses both alternatives in state 2 with positive probability will have R^* as its set of recurrent states. Randomization here plays the important role of coalescing the recurrent chains of the pure policies.

The historical account of the literature on the asymptotic behavior of v(n) goes back to Bellman (1957) and Howard (1960). Bellman showed that if every $P_{ij}^k > 0$, then $v(n)_i \sim n\langle g^* \rangle$ where every $g_i^* = \langle g^* \rangle$. Howard gave examples where $\lim_{n\to\infty}[v(n) - ng^*]$ existed, and conjectured that this was always true. The conjecture is false if some of the (maximal gain) policies have tpm's with periodic states. As an illustration, consider the following 2-state example with only one policy.

Example 2

$$q = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad g^* = (0,0)$$

$$v(n) = P^n v(0) = \begin{cases} (v(0)_1,\ v(0)_2) & n \text{ even} \\ (v(0)_2,\ v(0)_1) & n \text{ odd} \end{cases}$$

so that $\lim[v(n) - ng^*]$ exists if and only if $v(0)_1 = v(0)_2$. We remark here that there always exists choices of v(0) such that $\lim[v(n) - ng^*]$ exists.

On the other hand, Brown (1965) showed that $\{v(n) - ng^*\}_{n=0}^{\infty}$ is always bounded in n so that

$$g_i^* = \lim_{n\to\infty} v(n)_i/n \ , \quad 1 \le i \le N$$

(for any v(0)). This justifies the interpretation of g_i^* as the maximum average expected reward per transition when starting in state i. It remained to show when $\lim_{n\to\infty}[v(n) - ng^*]$ exists and the behavior when the limit fails to exist. As mentioned above, non-existence of the limit is associated with the occurrence of maximal gain policies having periodic tpm's.

An elegant result by White [1963] states conditions which exclude periodicities, and ensure existence of the limit. Specifically, if there exists a state s, integer m and number $\alpha > 0$ such that

$$(2.2) \qquad [P(f_1)P(f_2)\ldots P(f_m)]_{is} \ge \alpha > 0 \ , \quad 1 \le i \le N$$

$$\text{all } f_1, f_2, \ldots, f_m \in S_P \ ,$$

then $g_i^* = \langle g^* \rangle$ for all i, $v^* = \lim_{n\to\infty}[v(n) - n\langle g^* \rangle \underline{1}]$ exists for any choice of v(0), and the approach to the limit is geometric.

To avoid the linear divergence of v(n) with n, White proposed working with the relative values, obtained by setting (say) the N^{th} component of v^* equal to (say) zero. Under condition (2.2), the vector

$$v^{*rel} \equiv v^* - (v^*)_N \underline{1} \qquad (\text{with } v_N^{*rel} = 0)$$

is unique, and the relative value iteration scheme

(2.3) $\quad y(n+1) = Ty(n) - [Ty(n)]_{N\underline{1}}$, $\quad n = 0,1,2,\ldots$

(where $y(n) = v(n) - v(n)_{N\underline{1}}$) will converge geometrically to v^{*rel}.

Brown (1965) and Lanery (1967) have shown, albeit with faulty proofs, that there exists an integer $J \geq 1$ such that $\lim_{n\to\infty}[v(nJ) - nJg^*]$ exists for every $v(0) \in E^N$. The present authors (cf. Schweitzer and Federgruen (1978a)) showed, in fact, that in the general case the process has to be observed every J^* steps, where

$$J^* = \min\{J \geq 1 \,|\, \text{there exists } f \in S^*_{RMG} \text{ such that } P(f)^J \text{ is an aperiodic tpm}\}$$

so as to overcome the irregular behavior caused by the periodicities in the tpm's. Note that J^* is <u>independent</u> of the scrap values.

More specifically, we showed that

(2.4) $\quad \lim_{n\to\infty}[v(nJ + r) - (nJ + r)g^*]$ exists for <u>all</u> $v(0) \in E^N$ for some $r = 0,\ldots, J-1$, if and only if $J \geq 1$ is a multiple of J^*.

(2.5) $\quad$ for each $v(0) \in E^N$, there exists an integer J^o, which depends upon $v(0)$ and divides J^*, such that $\lim_{n\to\infty}[v(nJ + r) - (nJ + r)g^*]$ exists for any $r = 0,1,\ldots,J-1$ if and only if J is a multiple of J^o. Thus, in Example 2, $J^* = 2$ and $J^o = 1$ or 2, depending on whether $v(0)_1 - v(0)_2 = 0$ or $\neq 0$.

It follows from (2.4) that $\lim_{n\to\infty}[v(n) - ng^*]$ exists for all $v(0)$ if and only if $J^* = 1$, which holds if and only if there exists a (randomized) policy $f \in S^*_{RMG}$ with $P(f)$ aperiodic. Randomization is essential here because it serves to reduce the periods of the recurrent chains. As an illustration, the two pure policies in Example 1 have periods 2 and 3, while the randomized policy in S^*_{RMG} has period 1, so $J^* = 1$. Coalescing subchains and reducing periods appear to be among the few instances in MDPs where randomized policies play a central role. However, a finite procedure exists to calculate J^* from the periods of the finite set of pure maximal gain policies. [Schweitzer and Federgruen (1978a)]

A sufficient condition for $J^* = 1$, hence for the existence of $\lim_{n\to\infty}[v(n) - ng^*]$ for every $v(0)$, is that every pure maximal gain policy (or every pure policy in S_P) has an aperiodic tpm. The motivation for seeking easily-verified sufficient conditions for $J^* = 1$ is given in Section 4.

The present authors have also shown (cf. Schweitzer and Federgruen (1977b)) that if $v(0)$ is such that $v^* = \lim_{n\to\infty}[v(n) - ng^*]$ exists, then the approach to the limit is ultimately geometric. That is, there exist scalars $c > 0$, and $0 \le \lambda^* < 1$ such that

$$(2.6) \qquad |v(n) - ng^* - v^*|_\infty \le c(\lambda^*)^n \; ; \quad n = 0,1,2,\ldots ,$$

where λ^* represents the ultimate average contraction factor per step, and is independent of $v(0) \in E^N$, and where $c > 0$ does depend upon $v(0)$. The geometric convergence result is achieved by showing the existence of an integer

$M \geq 1$ (independent of $v(0)$) and an integer $n_0 \geq 1$ (dependent upon $v(0)$) as well as an M-step contraction factor $\lambda(v(0))$, such that

$$(2.7) \qquad sp[v(n+M) - (n+M)g^* - v^*] \leq \lambda sp[v(n) - ng^* - v^*]$$

$$\text{for all } n \geq n_0 \ .$$

In fact, whenever $g^* = \langle g^* \rangle \underline{1}$, as happens to be the case in most real life applications, n_0 may be taken to be equal to one.

In the general multi-chain case, however, no good estimates are available for n_0, and likewise no good upper bounds exist for λ^*, except if the problem satisfies a simultaneous scrambling condition (cf. Federgruen, Schweitzer and Tijms (1977), or if one is iterating a single fixed policy. In that case, it may be readily shown (cf. Morton (1977)) that λ can be taken as the modulus of the subdominant eigenvalue of the tpm.

The absence of estimates for λ is unfortunate because it precludes the use of (2.7) to estimate the deviation of $v(n) - ng^*$ from its limit. Consequently, (2.6)-(2.7) are not yet as useful as MacQueen's bounds in the discounted case (cf. MacQueen (1966) and also eq. (5.1)). Additional investigation is needed here to complete our understanding of the rate of contraction.

Finally, a generalization of these results is available for the <u>non-stationary</u> case, where instead of having perfect knowledge of the parameters of the model, only approximations to the latter can be generated, or where it is computationally preferable to generate sequences of approximations

for the parameters (cf. Federgruen and Schweitzer (1978a) for an enumeration of models in which this situation occurs). So in the non-stationary model we assume that we are able to generate sequences

$$\{q_i^k(n)\}_{n=1}^{\infty} \to q_i^k \; ; \quad 1 \le i \le N \quad \text{and} \quad k \in K(i)$$

$$\{P_{ij}^k(n)\}_{n=1}^{\infty} \to P_{ij}^k \; ; \quad 1 \le i,j \le N \quad \text{and} \quad k \in K(i)$$

$$\{K(i,n)\}_{n=1}^{\infty} \to K(i) \; ; \quad 1 \le i \le N \; .$$

Moreover, in most cases <u>geometric</u> convergence for the parameter approximations may be achieved, and the present authors showed (cf. Federgruen and Schweitzer (1978a)) that in this case the quantities of interest in our model can be approximated via the scheme

$$(2.8) \qquad x(n+1)_i = \max_{k \in K(i,n)} [q_i^k(n) + \sum_j P_{ij}^k(n) x(n)_j] \; , \quad 1 \le i \le N \; .$$

In particular, the asymptotic behavior of the sequence $\{x(n)\}_{n=1}^{\infty}$ is similar to that described for the stationary case:

(a) $\{x(n) - ng^*\}_{n=1}^{\infty}$ is bounded

(b) if $\lim_{n\to\infty}[x(n) - ng^*]$ exists, then the limit is approached geometrically

(c) If $J^* = 1$, then $\{x(n) - ng^*\}_{n=1}^{\infty}$ converges for every choice of $x(0) \in E^N$, and if $J^* \geq 2$, then $\lim_{n\to\infty}[x(nJ^* + r) - (nJ^* + r)g^*]$ exists for all $x(0) \in E^N$ and $r = 0,\ldots,J^* - 1$.

However, unlike the stationary case (cf. section 3),

(d) $\lim_{n\to\infty}[x(n) - ng^*]$ may exist for every $x(0)$, even when $J^* \geq 2$

and

(e) $\lim_{n\to\infty}[x(n) - ng^*]$ may fail to exist for every $x(0)$, when $J^* \geq 2$.

Finally, examples in Federgruen (1978) show that, even with all of the tpm's of all of the policies being aperiodic, all kinds of irregular behavior of the sequence $\{x(n)\}_{n=1}^{\infty}$ may be expected when the parameters of the model are approximated at a slower than geometric rate.

3. Dependence of the Asymptotic Behavior upon the Scrap Values

Let W denote the set of starting points for which the value iteration scheme converges, i.e.,

$$W = \{v(0) \in E^N \mid \lim_{n\to\infty}[v(n) - ng^*] \text{ exists}\}$$

$$= \{x \in E^N \mid \lim_{n\to\infty}[T^n x - ng^*] \text{ exists}\} \quad .$$

If $J^* = 1$, then $W = E^N$, while if $J^* \geq 2$, W is a non-empty strict subset of E^N (cf. Schweitzer and Federgruen (1977b)). For $v(0) \in W$, we let

$$L(v(0)) = \lim_{n\to\infty}[v(n) - ng^*] \quad .$$

This section summarizes the relatively few results that are known with respect to both W and $L(\cdot)$; moreover, we state some of our conjectures with respect to their properties.

The differences between the discounted and undiscounted cases deserve some emphasis here. In the discounted case, $\lim_{n\to\infty} v(n)$ always exists, and is independent of v(0). In the undiscounted case, $\lim_{n\to\infty}[v(n) - ng^*]$ doesn't always exist (except when $J^* = 1$), and when it does exist, it will explicitly depend upon v(0). (For instance, adding a constant to every component of v(0) adds the same constant to every component of v(n) and of $L(v(0))$.) These differences motivate our inquiry into the structure of the set W where the limit does exist, and the dependence of $L(\cdot)$ upon the scrap values.

The following notation will be needed. Suppose $v(0) \in W$ and $\lim_{n\to\infty}[v(n) - ng^*] = v^*$. Then $\{g^*, v^*\}$ satisfy the two coupled functional equations (cf. Howard (1960)):

$$(3.1) \qquad g_i = \max_{k\in K(i)} \sum_j P_{ij}^k g_j \quad , \quad 1 \le i \le N$$

$$(3.2) \qquad v_i = \max_{k\in L(i)} \{q_i^k - g_i + \sum_j P_{ij}^k v_j\} \quad , \quad 1 \le i \le N$$

where $L(i) = \{k \in K(i) \mid g_i = \sum_j P_{ij}^k g_j\}$. These equations determine g uniquely as $g = g^*$ in agreement with (2.1). But they determine v only up to certain additive constants. We let

$$(3.3) \qquad V = \{v \in E^N \mid (g^*, v) \text{ is a solution to (3.2)}\}.$$

In general, V may have a complicated structure, a characterization of which is given in Schweitzer and Federgruen (1978b). V is known to be closed, unbounded, connected but generally non-convex. The necessary and sufficient condition for the convexity of V was derived, and each of the following three conditions are easily verified sufficient conditions:

(a) $R^* = \{1,\ldots,N\}$

(b) for each $i \notin R^*$, L(i) is a singleton

(c) $n^* = 1$, where

$$(3.4) \qquad n^* = \min\{n(f) \mid f \in S^*_{RMG}\}$$

with $n(f) \geq 1$ representing the number of subchains (closed, irreducible sets of states) of $P(f)$, $f \in S_R$. $n^* = 1$ is in fact the necessary and sufficient condition for $v \in V$ to be unique up to a multiple of $\underline{1}$, i.e., under $n^* = 1$, V takes the simple structure $V = \{v^o + c\underline{1} \mid -\infty < c < +\infty\}$ and in case $n^* \geq 2$, the set V can be shown to be an n^*-dimensional subset of E^N. A finite algorithm was given to determine the number n^*, as well as a triangular decomposition for the polyhedral set from which we may choose the n^* parameters (degrees of freedom), which determine $v \in V$ (cf. Schweitzer & Federgruen (1977a)). Finally, the condition $n^* = 1$ is trivially met if every pure or every pure maximal gain policy is unichained.

The multichain policy iteration algorithm (Howard (1960)) may be used to find an element of V. If $v(0) \in W$, then $\lim_{n\to\infty}[nv(n-1) - (n-1)v(n)] = \lim_{n\to\infty}[v(n) - ng^*] \in V$ so that value-iteration may be employed to approximate an element of V, since several devices have been proposed to avoid the numerical difficulty of g^* being unknown and of $\{v(n)\}_{n=1}^{\infty}$ diverging linearly with n.

For later use we define for each $v \in V$, the set $S^*(v)$ of maximizing policies in (3.2):

$$(3.5) \qquad S^*(v) = X_{i=1}^{N} L(i,v) \qquad \text{where}$$

$$L(i,v) = \{k \in L(i) \mid v_i = q_i^k - g_i^* + \sum_j P_{ij}^k v_j\} \ , \quad 1 \leq i \leq N \ .$$

We now consider the function $L(x)$, $x \in W$. The following properties are known to hold:

$$(3.6) \qquad L(x + a\underline{1}) = L(x) + a\underline{1} \ , \text{ for any scalar } a, \text{ i.e., } L(\cdot) \text{ is unbounded}$$

(3.7) $L(\cdot)$ is continuous on W with Lipschitz norm of unity:

$$|L(x) - L(y)|_\infty \le |x-y|_\infty \ ;$$

$$sp[L(x) - L(y)] \le sp[x-y] \ ; \quad x,y \in W$$

(3.8) $L(\cdot)$ is a convex function on W:

$$L((1-a)x + ay) \le (1-a)L(x) + aL(y) \ ; \quad x,y \in W$$

for $0 \le a \le 1$ such that $(1-a)x + ay \in W$

(3.9) $L(x) \in V$

(3.10) if $x \in W$, then $Tx \in W$ and $L(Tx) = L(x) + g^*$.

In general, $L(x)$ is very difficult to display in closed form because it involves the (transient-type) decisions when termination of the process is near.

One simple example, patterned after Brown (1965), illustrates some of the structure in $L(\cdot)$:

Example 3: $N = 2$; $q_i^k = 0$ for all $1 \le i \le N$; $k \in K(i)$. Hence $g^* = \underline{0}$.

i	k	P_{i1}^k	P_{i2}^k
1	1	.4	.6
	2	.5	.5
2	1	.6	.4

$J^* = 1$; $W = E^2$, i.e.,

$L(x)$ exists for every $x \in E^2$

Note that

(a) $x_1 \le x_2 \Rightarrow Tx = P(f^1)x$ with $(Tx)_1 \ge (Tx)_2$

(b) $x_1 \ge x_2 \Rightarrow Tx = P(f^2)x$ with $(Tx)_1 \le (Tx)_2$

where $P(f^r)$ uses alternative r in state 1 (r = 1,2). Consequently, value-iteration will alternate between $P(f^1)$ and $P(f^2)$. This implies

$$v(2n) = \begin{cases} [\begin{pmatrix} .5 & .5 \\ .6 & .4 \end{pmatrix} \quad \begin{pmatrix} .4 & .6 \\ .6 & .4 \end{pmatrix}]^n v(0) \text{ if } v(0)_1 \le v(0)_2 \\ [\begin{pmatrix} .4 & .6 \\ .6 & .4 \end{pmatrix} \quad \begin{pmatrix} .5 & .5 \\ .6 & .4 \end{pmatrix}]^n v(0) \text{ if } v(0)_1 \ge v(0)_2 \end{cases}.$$

Letting $n \to \infty$,

$$L(x) = \begin{cases} (\frac{24x_1}{49} + \frac{25x_2}{49})\underline{1} & x_1 \le x_2 \\ (\frac{27x_1}{49} + \frac{22x_2}{49})\underline{1} & x_1 \ge x_2 \end{cases}.$$

The following example illustrates more complex behavior

Example 4: N = 3; K(1) = K(3) = {1}, K(2) = {1,2}

i	k	p^k_{i1}	p^k_{i2}	p^k_{i3}
1	1	0	1	0
2	1	1	0	0
	2	0	0	1
3	1	0	0	1

$$q_i^k \equiv 0 \; ; \quad g^* = (0,0,0)$$

$$V = \{v \in E^3 | v_1 = v_2 \geq v_3\} \text{ is convex} \quad .$$

$$T^{2n}x = T^2x = \begin{bmatrix} \max[x_1, x_3] \\ \max[x_2, x_3] \\ x_3 \end{bmatrix} \qquad n = 1,2,3,\ldots$$

$$T^{2n+1}x = T^3x = \begin{bmatrix} \max[x_2, x_3] \\ \max[x_1, x_3] \\ x_3 \end{bmatrix} \qquad n = 1,2,3,\ldots \; .$$

Note that $\lim_{n\to\infty} T^n x$ exists if and only if $T^2x = T^3x$, hence

$$W = \{x \in E^3 | \max[x_1, x_3] = \max[x_2, x_3]\} \quad .$$

Note that W may be written as the union of two polytopes W_1 and W_2, where

$$W_1 = \{x \in E^3 | x_1 = x_2 \geq x_3\} = V$$

$$W_2 = \{x \in E^3 | x_3 \geq \max[x_1, x_2]\} \quad .$$

In addition,

$$L(x) = \begin{cases} x_1\underline{1} & , \text{ for } x \in W_1 \\ \\ x_3\underline{1} & , \text{ for } x \in W_2 \end{cases} .$$

We know (cf. Schweitzer and Federgruen (1977b)) that W is always closed and unbounded; e.g., if $x \in W$, then $x + a\underline{1} \in W$ for all scalars a. W is connected in all cases examined by the authors, and it is conjectured that this holds in all generality. The above example shows, however, that W does not need to be convex, even if V is convex. The above examples suggest, in addition, that W may always be decomposed into a finite number of polytopes such that $L(\cdot)$ is linear on each of the polytopes and has directional derivatives in any feasible direction wherever two such polytopes join. These polytopes may have the structural form of cones. (See also Theorem 3.2.)

In general, it is very difficult to compute W, which depends sensitively upon the value-iteration decisions when termination is near. It is hard even to give an analytic characterization of W. The following theorem provides a step in that direction, but is not useful at present because the function $L^*(x)$ (defined below) is as poorly understood as $L(x)$ itself.

Define $L^*(x) \equiv \lim_{n\to\infty} [T^{nJ^*} x - nJ^* g^*]$. According to (2.4), $L^*(x)$ exists for every $x \in E^N$. Note that $L(x)$ agrees with $L^*(x)$ for $x \in W$; however, $L(x)$ is undefined for $x \in E^N \backslash W$. Note that the J^*-step operator T^{J^*} may be interpreted as the value iteration operator in a J^*-step MDP, with $\{1,\ldots,N\}$ as its state space, and with action spaces, one step expected rewards and transition probabilities given by:

$$\tilde{K}(i) = \{(f^{J^*},\ldots,f^1) \mid f^r \in S_p,\ 1 \le r \le J^*\} \quad , \quad 1 \le i \le N$$

$$\tilde{q}^{\xi}_i = q(f^{J^*})_i + P(f^{J^*})q(f^{J^*-1})_i + \ldots + [P(f^{J^*})\ldots P(f^2)]q(f^1)_i$$

$$1 \le i \le N \quad \text{and} \quad \xi = (f^{J^*},\ldots,f^1)$$

$$\tilde{P}^{\xi}_{ij} = P(f^{J^*})\ldots P(f^1)_{ij} \ ; \quad 1 \le i,j \le N \ ; \quad \xi = (f^{J^*},\ldots,f^1) \quad .$$

As a consequence, the properties (3.6) and (3.7) carry over to $L^*(x)$, and property (3.8) shows $L^*(x)$ is a convex function everywhere on E^N.

Theorem 3.1: (characterization of W)

$x \in W$ if and only if $L^*(x) \in V$.

Proof: If $x \in W$, combine $L^*(x) = L(x)$ with $L(x) \in V$ to conclude $L^*(x) \in V$. Conversely, assume $L^*(x) \in V$ and define $L^{*,r}(x) = \lim_{n\to\infty} T^{nJ^*+r}x - (nJ^* + r)g^*$ $(r = 1,2,\ldots,J^*)$. Observe that for all $1 \le i \le N$:

$$(3.11) \qquad T^{nJ^*+1}x_i - (nJ^* + 1)g^*_i =$$

$$\max_{k\in K(i)} \{nJ^*[\sum_j P^k_{ij}\, g^*_j - g^*_i] + q^k_i - g^*_i + \sum_j P^k_{ij}[T^{nJ^*}x - nJ^*g^*]\}$$

and note that for n sufficiently large, only $k \in L(i)$ achieve the maximum on the right of (3.11). Use this when letting n tend to infinity, to conclude for all $1 \le i \le N$:

$$L^{*,1}(x)_i = \max_{k\in L(i)} \{q^k_i - g^*_i + \sum_j P^k_{ij} L^*(x)_j\} = L^*(x)_i \ ,$$

where the second inequality follows from $L^*(x) \in V$. Likewise, one proves $L^{*,k}(x) = L^*(x)$ for all $1 \le k \le J^*$, i.e., $\lim_{n\to\infty} T^n x - ng^*$ exists, or $x \in W$.□

Finally, the following theorem given an abstract characterization of the function $L^*(x)$, but lacks utility until a better understanding of the set Γ (as defined below) becomes available.

For any infinite sequence of pure policies $\Phi = \{f^1, f^2, f^3, \ldots\}$, $f^n \in S_p$, where f^n is used when n periods remain before the termination of the planning period, define the n-period rewards and tpm's by:

$$q(\Phi,n) = q(f^n) + P(f^n)q(f^{n-1}) + \ldots + [P(f^n)\ldots P(f^2)]q(f^1)$$

$$P(\Phi,n) = P(f^n)\ldots P(f^1)$$

Let $\Gamma = \{(\alpha,\gamma) \mid \alpha \in E^N$; γ is an N×N stochastic matrix;

(α,γ) is a limit point of $(q(\Phi, nJ^*) - nJ^*g^*, P(\Phi, nJ^*))$

for some infinite policy sequence $\Phi\}$.

Theorem 3.2: Characterization of $L^*(x)$

$$(3.12) \qquad L^*(x)_i = \max_{(\alpha,\gamma)\in\Gamma} [\alpha_i + \gamma x_i] \; ; \quad x \in E^N \; ; \quad 1 \le i \le N \; .$$

In addition, for each x there is a choice $(\alpha^*, \gamma^*) \in \Gamma$ which achieves all N maxima in (3.12).

Proof: Fix $(\alpha,\gamma) \in \Gamma$ and a policy sequence Φ such that $\lim_{k\to\infty}(q(\Phi, n_kJ^*) - n_kJ^*g^*, P(\Phi, n_kJ^*)) = (\alpha,\gamma)$ for some sequence of increasing integers $\{n_k\}_{k=1}^{\infty}$. Next, note that for all $n \geq 1$:

$$(3.13) \qquad v(nJ^*) \geq q(\Phi, nJ^*) + P(\Phi, nJ^*)x$$

where $x = v(0)$. Replace n by n_k, subtract $n_kJ^*g^*$ from both sides in (3.13) and let k tend to infinity, to conclude:

$$(3.14) \qquad L^*(x) \geq \alpha + \gamma x \quad \text{for any } (\alpha,\gamma) \in \Gamma \quad .$$

Finally, let Φ^* be such that $f^r \in S(r)$ for all $r \geq 1$, with $v(0) = x$. Then (3.13) with Φ replaced by Φ^* holds as a strict equality for all n; choose a subsequence $\{n_k\}_{k=1}^{\infty}$ such that the bounded sequence $P(\Phi^*, n_kJ^*) \to$ (say) γ^* and consequently $q(\Phi^*, n_KJ^*) - n_KJ^*g^* = v(n_kJ^*) - n_KJ^*g^* - P(\Phi^*, n_kJ^*)x \to L^*(x) - \gamma^*x \equiv \alpha^*$. Thus, $(\alpha^*, \gamma^*) \in \Gamma$ and (3.14) holds as a strict equality when α and γ are replaced by α^* and γ^*. □

4. The asymptotic behavior of S(n)

In this section we describe the properties of the sets of optimizing policies S(n), as n tends to infinity. The asymptotic behavior differs sharply between the case where $v(0) \in W$ and the case where $\lim_{n\to\infty}[v(n) - ng^*]$ fails to exist. However, even in case $v(0) \in W$, contrived examples show a possibly irregular dependence of S(n) upon n. This suggests that convergence of S(n) cannot be safely relied upon as a termination criterion for value-iteration. This is in surprising contrast with

(i) the fact that for appropriate choices of the sequence $\{\varepsilon_n\}_{n=1}^{\infty} \downarrow 0$ the sequence of ε_n - optimal policies $\{S(n, \varepsilon_n)\}_{n=1}^{\infty}$ will converge whenever $v(0) \in W$.

(ii) the empirical fact that in "real-life" problems, the sets $\{S(n)\}_{n=1}^{\infty}$ converge invariably and unambiguously.

For a more detailed description of the following turnpike properties, cf. Federgruen and Schweitzer (1979a).

(a) if $v^* = \lim_{n\to\infty} [v(n) - ng^*]$ exists, then

$$S(n) \subseteq S^*(v^*) \subseteq S_{RMG} \tag{4.1}$$

for all n exceeding some n_0 which depends upon v(0). (The Cartesian product set $S^*(v^*)$ was defined by (3.5).) Thus S(n) will always settle upon maximal gain policies whenever $\{v(n) - ng^*\}_{n=1}^{\infty}$ has a limit.

An important unsolved problem is the estimation of n_0. Until upper bounds on n_0 are available, one cannot be sure that policies produced by a finite number of value-iteration steps are indeed maximal gain. Examples are known where $\sup_{v(0)\in W} n_0(v(0)) = \infty$, i.e., n_0 may exceed any given integer m by an appropriate bad choice of v(0).

(b) If $v^* = \lim_{n\to\infty} [v(n) - ng^*]$ so that $S(n) \subseteq S^*(v^*)$ for large n, the asymptotic behavior of S(n) may still be erratic [see, e.g., modifications of Shapiro (1968)]. S(n) could be a strict subset of $S^*(v^*)$ for every n, with some members of $S^*(v^*)$ never identified. Also, S(n) could oscillate periodically among members of $S^*(v^*)$ [see Brown (1965)] or could even oscillate in a non-periodic way among members of $S^*(v^*)$ (by modification of an example in Bather (1973)). This potential for irregular behavior (when $S^*(v^*)$ is not a singleton) discourages use of convergence of S(n) as a termination criterion.

It is nevertheless possible to compute $S^*(v^*)$ correctly by means of ϵ-optimal policies, as follows. Let $\{\epsilon_n\}_{n=1}^{\infty}$ denote any sequence of positive numbers which approaches 0 at a slower rate than the geometric rate of convergence of $v(n) - ng^* - v^*$ to $\underline{0}$, e.g., take $\epsilon_n = n^{-1}$ (or the reciprocal of any positive polynomial in n). Then for all n exceeding some n_1,

$$S^*(v^*) = \{f \in S_p \,|\, v(n+1) \leq q(f) + P(f)v(n) + \epsilon_n \underline{1}\} \quad .$$

Once again, no bounds are available for n_1.

Here again, all of the results mentioned under (a) and (b) carry over to the non-stationary model where only geometric approximations for the parameters are available (cf. section 2).

(c) If $\lim_{n\to\infty} [v(n) - ng^*]$ doesn't exist, then S(n) need not lie in S_{RMG} for all large n; it need not even intersect S_{RMG}. The first such example

is given in [Lanery (1967)], where S(n) lies outside S_{RMG} for infinitely many n. Later the authors constructed an example [Federgruen and Schweitzer (1979a)] where S(n) is disjoint from S_{RMG} for every n. These examples contrast sharply with the behavior in the discounted case, where S(n) can contain only optimal policies when n is sufficiently large.

For problems with many thousands of states, value-iteration is the most practical way to locate maximal gain policies. The importance of (a) and (c) is that value-iteration can be relied on to identify maximal gain policies only if $\lim_{n\to\infty}[v(n) - ng^*]$ exists.

This motivates the search for conditions ensuring either $J^* = 1$ (so the limit exists for every v(0)) or $v(0) \in W$ (so the limit exists for this v(0)). Sufficient conditions for $J^* = 1$ were indicated above. If these conditions cannot be verified, and if there is concern about non-existence of $\lim_{n\to\infty}[v(n) - ng^*]$ due to the presence of periodic tpm's, it is suggested that a data-transformation be applied (see below) to ensure that $J^* = 1$.

5. Bounds and Lyapunov Functions

For discounted MDPs, where $\beta < 1$, the unique fixed point $v^* = Tv^*$ of the monotone contraction operator T is sought. The following bounds on v^* were given by MacQueen (1966) and later improved slightly by Porteus (1971).

$$(5.1) \qquad x + \frac{(Tx - x)_{\min}}{1 - \beta} \underline{1} \leq v^* \leq x + \frac{(Tx - x)_{\max}}{1 - \beta} \underline{1} \ , \quad \text{any} \quad x \in E^N \ .$$

These bounds have the following three convenient properties:

invariance: they remain unchanged when x is replaced by $x + a\underline{1}$, $-\infty < a < +\infty$

sharpness: the bounds converge to v^* when x approaches v^*

monotonicity under value iteration: the bounds move monotonically inward when x is replaced by Tx.

The bounds in (5.1) are conveniently used during discounted value-iteration, with $x = v(n)$ and $Tx = v(n+1)$, since the bounds may be computed with almost no additional effort and converge monotonically and geometrically to v^*. The bounds are useful both as a termination criterion (stopping when $|v^* - v(n)|_\infty$ achieves a given precision) and for elimination of suboptimal actions [MacQueen (1967)]. For a recent survey on elimination tests, we refer to White (1978).

Consider now the construction of analogous bounds for undiscounted MDPs, where $\beta = 1$. Differences are immediately apparent because the discounted case has only one set of optimality equations for the N-vector v^*

whereas the undiscounted case has a pair of coupled functional equations for two N-vectors g^* and v^*. Furthermore, in the discounted case v^* is uniquely determined by these equations whereas in the undiscounted case only g^* is uniquely determined, while v^* is determined up to certain additive constants.

We describe first bounds on the gain rate vector g^*. Consider, initially, the special case where all components of g^* are equal: $g^* = \langle g^* \rangle \underline{1}$. This is the most important case in practice, arising in the so-called unichain case where each pure policy f has a tpm P(f) with a single closed irreducible set of states, plus possibly some transient states. Bounds on the scalar $\langle g^* \rangle$ were given first by Odoni (1969)

$$(5.2) \qquad (Tx - x)_{min} \leq \langle g^* \rangle \leq (Tx - x)_{max} \quad \text{any} \quad x \in E_N$$

with the properties

invariance: the bounds remain unchanged when x is replaced by $x + a\underline{1}$

sharpness: the bounds converge to $\langle g^* \rangle$ when x approaches some $v^* \in V$

monotonicity under value iteration: the bounds move monotonically inward (but not necessarily strictly monotonically) when x is replaced by Tx.

These bounds are conveniently used during value-iteration, with $x = v(n)$ and $Tx = v(n+1)$, or (see eq. (2.3)) $x = y(n)$ and $Tx = Ty(n)$). If $\lim_{n\to\infty}(v(n) - ng^*)$ exists, then both upper and lower bounds converge to

$\langle g^* \rangle$; however, if these limits don't exist, then a gap will occur between the asymptotic levels of the upper and lower bounds.

The bounds are useful as a termination criterion (stopping with an estimate of $\langle g^* \rangle$ which achieves a given precision) but so far have not been useful for elimination of suboptimal actions (see further discussion below) due to the absence of estimates of the deviation of $v(n) - n\langle g^* \rangle \underline{1}$ from v^*.

Consider next the general case where the components of g^* can be unequal. The authors obtained the following bounds on the maximal gain rate vector g^* [Schweitzer and Federgruen (1979a)]

$$(5.3) \qquad g + \{ \max_{f \in \Lambda(g)} [q(f) - g + P(f)x - x] \}_{\min} \underline{1} \leq$$

$$\leq g^* \leq g + \{ \max_{f \in \Lambda(g)} [q(f) - g + P(f)x - x] \}_{\max} \underline{1} \quad ,$$

$$g \in G, \; x \in E^N$$

where $G \equiv \{g \in E^N | g = \max_{f \in K} P(f)g\}$ and, for $g \in G$, $\Lambda(g) \equiv \{f \in K | g = P(f)g\}$ and where the expressions within braces in (5.3) are maximized component by component. Note that G is not empty, since $\underline{0}$, $\underline{1}$ and g^* are in G. Note also that when $g^* = \langle g^* \rangle \underline{1}$, the choice $g = \underline{0}$ reduces the above bounds to those of Odoni.

The above bounds have the properties:

<u>invariance</u>: they remain unchanged when replacing (g,x) by $(g + a\underline{1}, x + b\underline{1})$ for any scalars a, b.

sharpness: the bounds converge to g^* when g approaches g^*, and x approaches some $v^* \in V$.

Unfortunately, the monotonicity property has eluded generalization to this case. We lack a value-iteration scheme for generating successive pairs of vectors {g,x} such that the bounds move monotonically inward. The main technical difficulty appears to be the absence of simple characterizations of G and $\Lambda(g)$, $g \in G$, or of simple ways to generate sequences of members of G. However, in the upper bound in (5.3), it is permitted to replace $\max_{f \in \Lambda(g)}$ by $\max_{f \in S_p}$, and the task of minimizing the upper bound is then related to the primal linear program for the gain rate vector [Denardo and Fox (1968)], [Hordijk and Kallenberg (1978)]:

$$\min_{g,x} \sum_i g_i$$

$$g_i \geq \sum_j P_{ij}^k g_j \ , \quad 1 \leq i \leq N \ , \quad k \in K(i)$$

$$x_i \geq q_i^k - g_i + \sum_j P_{ij}^k x_j \ , \quad 1 \leq i \leq N \ , \quad k \in K(i) \ .$$

Finally, we describe bounds on the relative value vector $v^* \in V$. Consider, initially, the case $n^* = 1$, where $g^* = \langle g^* \rangle \underline{1}$ and v^* is unique up to an additive multiple of $\underline{1}$ (cf. (3.4)). It is convenient to measure the deviation of an N-vector x from v^* by $sp[x - v^*]$, which is invariant to the additive constant in v^*. Define

$$\phi(x) = sp[Tx - x]$$

which is non-negative and vanishes if and only if $x \in V$. The following theorem shows that $\phi(x)$ provides a computable measure for the distance between x and v^*:

Theorem 5.1: (cf. Federgruen and Schweitzer (1979b))

Assume $n^* = 1$. Then there exists a constant c_1, $1/2 \leq c_1 < \infty$ such that

$$(5.4) \qquad \frac{\phi(x)}{2} \leq sp[x - v^*] \leq c_1 \Phi(x) \quad \text{for all} \quad x \in E^N$$

if and only if

(5.5) all states in $\hat{R} = \{i | i \in R(f), \text{ for some } f \in S_p\} \supseteq R^*$ can reach each other, i.e., if $i,j \in \hat{R}$ then there exists a policy $f \in S_R$, such that $P(f)^r_{ij} > 0$ for some $1 \leq r \leq N$.

Remark: The condition (5.5) can be expressed in various equivalent ways. For example, (5.5) is equivalent to $\hat{R}$ being a communicating system (cf. Bather (1973)), or to the existence of a (randomized) policy that has $\hat{R}$ as its single subchain. Observe, in addition, that the combination of $n^* = 1$ and (5.5) certainly holds in the unichain case.

The bounds in (5.4) estimate the deviation of x from v^*, in terms of the computable quantity $\Phi(x)$. The upper bound in (5.4) is the direct analogue of (5.1) in the discounted case, which may be rewritten as

$$(5.6) \qquad sp[x - v^*] \leq (1 - \beta)^{-1} sp[Tx - x] \quad .$$

The bounds in (5.4) again have the three desirable properties of (a) invariance (they remain unchanged when replacing x by $x + a\underline{1}$); (b) sharpness (they converge to 0 as x approaches $v^* + a\underline{1}$; (c) monotonicity of the upper bound under value-iteration ($\phi(Tx) \leq \phi(x)$).

In principle, the bounds in (5.4) permit both a termination criterion for achieving a given precision, and elimination of suboptimal actions (those not in $S^*(v^*)$, which set is uniquely determined if $n^* = 1$). Unfortunately, the bounds are not useful as written because c_1 can be enormously large for real problems. Our current upper bound for the unichain case, $c_1 = 4N/[\min\{P^k_{ij} | P^k_{ij} > 0\}]^N$, needs considerable refinement to make these bounds practical.

An alternative method of bounding v^* and eliminating suboptimal actions has been proposed (Federgruen, Schweitzer and Tijms (1977)) in the unichain case where every P(f), $f \in S_p$, is unichained. The method uses the data-transformation (6.1), which converts the original undiscounted MDP into a new one, denoted by a tilde, with the same state and action spaces, S_{RMG} and v^{*rel} left intact, the scalar gain rate multiplied by a scalar τ, $0 < \tau < 1$, and every $\tilde{P}(f)_{ii} \geq 1 - \tau > 0$. White's relative value scheme (2.3) for the transformed model has the form

$$\tilde{y}(n+1) = \tilde{Q}\tilde{y}(n) \quad , \quad \text{where} \tag{5.7}$$

$$\tilde{Q}x = \tilde{T}x - (\tilde{T}x)_N \underline{1} \quad , \text{ and}$$

$$\tilde{T}x = (1-\tau)x + \tau Tx \quad .$$

We recall that the scheme (5.7) has the property $\lim_{n\to\infty} \tilde{y}(n) = v^{*rel}$. If every P(f), $f \in S_p$, is unichained, then $\tilde{Q}$ is an N-step contraction operator on $\tilde{E}^N = \{x \in E^N | x_N = 0\}$ with v^{*rel} as its unique fixed point. This permits the construction of monotonically and geometrically converging upper and lower bounds on v^{*rel}, in terms of $\tilde{y}(n)$, and to use these bounds to eliminate sub-optimal actions in complete analogy with MacQueen's procedures in the discounted case (cf. Federgruen, Schweitzer and Tijms (1977)).

This is believed to be the first published scheme employing value-iteration for monotonically and geometrically converging upper and lower bounds on v^{*rel}, and for permanent action elimination. Computational testing of the scheme is lacking, and it is unknown how to weaken the unichainedness assumption to merely $n^* = 1$. Finally, we refer to Hastings (1976) for a "temporary" elimination scheme of non-optimal actions.

For MDP's with $n^* = 1$, i.e., with $v^* \in V$ unique up to a multiple of $\underline{1}$, we define a continuous function ϕ: $E^N \to E^1$ as a Lyapunov function if it satisfies the following conditions:

(5.8) (a) $\phi(x) \geq 0$, $x \in E^N$; $\phi(x) = 0$ if and only if $x \in V$

(b) $\phi(Tx) \leq \phi(x)$; $x \in E^N$

(c) there exists an integer $M \geq 1$ such that $\phi(T^M x) < \phi(x)$ whenever $\phi(x) > 0$.

One such function is $\phi(x) = sp[Tx - x]$, as used above; if every P(f) is unichained and has a positive diagonal (where the latter can be achieved

via the above discussed data-transformation (6.1)), then (5.8c) holds with $M = N$.

Another possible Lyapunov function is $\phi(x) = sp[x - v^*]$ with (5.8c) following from (2.7) if $\lim_{n\to\infty}[T^n x - ng^*]$ exists. This Lyapunov function, however, cannot be numerically evaluated midway through the value-iteration computations because v^* is unknown. In the discounted case, the right hand side of (5.6) is a Lyapunov function with $M = 1$ and $\phi(Tx) \leq \beta\phi(x)$.

Lyapunov functions have two convenient uses, one theoretical and the other numerical. Theoretically, their existence ensures convergence of the value-iteration scheme $v(n) - ng^*$, i.e., construction of a Lyapunov function is a way of proving algorithm convergence [cf. Zangwill (1970)]. Computationally, their numerical value measures (or bounds) the deviation $v^* - x$ between the limit point and the current guess.

Lastly, we describe difficulties in constructing bounds on $v^* \in V$ in the general multichain case. If $n^* \geq 2$, $v^* \in V$ is <u>not</u> unique up to a multiple of $\underline{1}$; instead, v^* is determined up to n^* additive constants (cf. section 3). Consequently, the expression $sp[x - v^*]$ is not uniquely defined until a particular choice of $v^* \in V$ is made. One natural measure of deviation, $\inf_{v\in V} sp[x-v]$, appears computationally intractable. Another natural choice, to measure the deviation of $x = v(n) - ng^* = T^n v(0) - ng^*$ from $v^* = L(v(0)) = \lim_{n\to\infty}[v(n) - ng^*] \in V$ via $sp[v(n) - ng^* - L(v(0))]$ is again intractable because g^* and $L(v(0))$ are unknown while being midway through the calculations.

The contraction property (2.7) is not helpful because λ and n_0 are usually unknown.

A third choice is to compute bounds on the optimal bias w^* [Denardo (1970)] $\in V$. Here exact variational bounds are available [Schweitzer and Federgruen (1979a)]

$$(5.9) \qquad v + \{ \max_{f \in S^*(v)} [-P(f)v + P(f)y - y] \}_{\min} \underline{1} \leq w^* \leq$$

$$v + \{ \max_{f \in S^*(v)} [-P(f)v + P(f)y - y] \}_{\max} , \quad v \in V, \; y \in E^N$$

where V and $S^*(v)$ were defined in (3.3) and (3.5), respectively. These bounds are both invariant and sharp.

In addition to the absence of a compelling reason to select w^* as the prototype member of V, the bounds in (5.9) are not computationally useful until simple ways are discovered to characterize V and $S^*(v)$, $v \in V$, and to generate sequences from V. These deficiencies in our computational procedures indicate that the multichain undiscounted case is still an open area for investigation.

6. Data Transformations

Data transformations of the parameters of our model are meant to convert one MDP into another (cf. Schweitzer (1971b), Schweitzer (1972), Federgruen and Schweitzer (1979c), Lippman (1975), Porteus (1975), and Porteus and Totten (1974)). Their main use is to create a transformed MDP which is easier either to analyze theoretically or to compute numerically. Data transformations generally destroy the interpretation of $v(n)$ as the vector of maximal n-period rewards. However, they are useful for the infinite-horizon case, provided that the quantities of interest transform in a tractable manner.

A useful data-transformation for the undiscounted MDP, indicated by a tilde, is

$$\tilde{N} = N \; ; \quad \tilde{K}(i) = K(i) \; , \quad 1 \le i \le N \tag{6.1}$$

$$\tilde{q}_i^k = \tau q_i^k \; ; \quad \tilde{P}_{ij}^k = (1-\tau)\delta_{ij} + \tau P_{ij}^k \; ;$$

$$1 \le i,j \le N \; ; \quad k \in K(i)$$

where $0 < \tau < 1$. This may be interpreted as observing the process, and making decisions, at intervals τ rather than at unit intervals. It has the properties

$$\tilde{\Pi}(f) = \Pi(f) \; ; \quad \tilde{g}(f) = \tau g(f) \; , \quad f \in S_p \tag{6.2}$$

$$\tilde{g}^* = \tau g^* \; ; \quad \tilde{S}_{RMG} = S_{RMG} \; ; \quad \tilde{V} = V \; .$$

The important new feature is that every $\tilde{P}(f)_{ii} \geq 1 - \tau > 0$ so that every tpm $\tilde{P}(f)$ is aperiodic. It follows that value-iteration on the transformed problem will be guaranteed to converge for any choice of the scrap value vector, since $\tilde{J}^* = 1$. Thus value-iteration on the transformed problem will be sure to identify maximal gain policies, and is hence preferred over value-iteration in the original model, if the latter has periodic tpm's.

A second use of data transformations is to convert semi-Markovian decision processes (SMDP's) where the transitions are not equally spaced in time, into MDPs, where the transitions are one unit time apart. Consequently, techniques developed for the infinite-horizon MDP may be invoked for the infinite-horizon SMDP.

Consider first the undiscounted infinite-horizon semi-MDP. The functional equations to be solved are (cf. Jewell (1963)):

$$(6.3) \qquad g_i^* = \max_{k \in K(i)} \sum_j P_{ij}^k g_j^* \quad , \quad 1 \leq i \leq N$$

$$v_i^* = \max_{k \in L(i)} \{q_i^k - g_i^* H_i^k + \sum_j P_{ij}^k v_j^*\} \quad , \quad 1 \leq i \leq N$$

where $L(i) = \{k \in K(i) \mid g_i^* = \sum_j P_{ij}^K g_j^*\}$ and where $H_i^k > 0$ is the mean holding time in state i, when action k is chosen.

The appropriate data transformation is (cf. Schweitzer (1971b))

$$(6.4) \qquad \tilde{N} = N \; ; \; \tilde{K}(i) = K(i) \quad , \quad 1 \leq i \leq N$$

(contd. on page 37)

$$\tilde{q}_i^k = \tau q_i^k / H_i^k \; ; \quad \tilde{P}_{ij}^k = (1 - \frac{\tau}{H_i^k})\delta_{ij} + \tau \frac{P_{ij}^k}{H_i^k} \; ;$$

$$1 \leq i,j \leq N \; ; \quad k \in K(i)$$

$$\tilde{H}_i^k = \tau$$

where

$$(6.5) \qquad 0 < \tau \leq \min\{H_i^k/(1 - P_{ii}^k) \,|\, (i,k) \text{ with } P_{ii}^k < 1\} \; .$$

The transformed problem is an undiscounted MDP with decisions at fixed intervals τ, with $\tilde{S}_{RMG} = S_{RMG}$, $\tilde{g}^* = \tau g^*$, $\tilde{L}(i) = L(i)$, and $\tilde{V} = V$. Moreover, by choosing τ strictly less than the upper bound in (6.5), we have $\tilde{P}(f)_{ii} > 0$ for all $1 \leq i \leq N$ and $f \in S_p$. As a consequence, value-iteration as applied to the transformed model will be guaranteed to converge and will yield maximal gain policies as well as a solution $v \in V$. (Schweitzer (1971b).)

In the special case where $g^* = \langle g^* \rangle \underline{1}$, Odoni's bounds (cf. (5.2)) for the scalar gain rate of the transformed problem are just the bounds given by Hastings (1971) and Schweitzer (1971a) for the gain rate of an SMDP with $g^* = \langle g^* \rangle \underline{1}$:

$$(6.6) \qquad \min_i \max_{k \in K(i)} \left[\frac{q_i^k + \sum_j P_{ij}^k x_j - x_i}{H_i^k} \right] \leq \langle g^* \rangle \leq$$

$$\max_i \max_{k \in K(i)} \left[\frac{q_i^k + \sum_j P_{ij}^k x_j - x_i}{H_i^k} \right] \quad , \quad x \in E^N \; .$$

Consider next the discounted infinite horizon SMDP with functional equation:

(6.7) $v^* = Qv^*$, where

$$Qx_i = \max_{k \in K(i)} [q_i^k + \sum_{j=1}^{N} B_{ij}^k x_j] \ , \quad 1 \le i \le N$$

and

$$B_{ij}^k \ge 0, \ B_{i,sum}^k = \sum_j B_{i,j}^k \le \delta < 1 \ .$$

The operator Q is a monotone contraction operator with contraction modulus $\delta < 1$, hence it has a unique fixed point v^*, which can be approximated fast via the successive approximation scheme $v(n) = Qv(n-1) = Q^n v(0)$.

In addition, upper and lower bounds on $|v(n) - v^*|_\infty$ follow from standard contraction operator theory, and are based on the maximal and minimal magnitudes of $B_{i,sum}^k$ (cf. Porteus (1971), Hastings (1971)). Improvements on the latter can, however, be obtained via the following data transformation

(6.8) $\tilde{N} = N \ ; \ \tilde{K}(i) = K(i)$ all i

$$\tilde{q}_i^k = \frac{(1 - \tilde{\beta}) q_i^k}{1 - B_{i,sum}^k} \ ; \quad \tilde{B}_{ij}^k = \delta_{ij} + \frac{(1 - \tilde{\beta})(B_{ij}^k - \delta_{ij})}{1 - B_{i,sum}^k} \ ;$$

$$1 \le i,j \le N \ ; \quad k \in K(i)$$

where $\tilde{\beta}$ is chosen to satisfy

$$0 \le \max_{\substack{1\le i\le N \\ k\in K(i)}} \frac{B^k_{i,sum} - B^k_{ii}}{1 - B^k_{ii}} \le \tilde{\beta} < 1 \quad .$$

This ensures that $\tilde{B}^k_{ij} \ge 0$ and $\tilde{B}^k_{i,sum} = \tilde{\beta} < 1$. The transformed (tilde) process is a discounted MDP with discount factor $\tilde{\beta}$, and with the same fixed point v^* and the same set of optimal policies as the original SMDP. The tilde value-iteration scheme $\tilde{v}(n+1) = \tilde{Q}\tilde{v}(n)$ where $\tilde{Q}x_i = \max_{k\in K(i)} [\tilde{q}^k_i + \sum_j \tilde{B}^k_{ij}x_j]$, $1 \le i \le N$ may converge to v^* quicker than the original scheme $v(n+1) = Qv(n)$.

In addition, applying MacQueen's bounds to the operator $\tilde{Q}$, we obtain new bounds on the fixed point v^*:

$$(6.9) \qquad x_i + \min_r \max_{k\in K(i)} \left[\frac{q^k_r + \sum_{j=1}^N B^k_{rj}x_j - x_r}{1 - B^k_{r,sum}} \right] \le v^*_i$$

$$\le x_i + \max_r \max_{k\in K(i)} \left[\frac{q^k_r + \sum_{j=1}^N B^k_{rj}x_j - x_r}{1 - B^k_{r,sum}} \right] , \quad 1 \le i \le N$$

for any $x \in E^N$. These bounds are <u>invariant</u> when replacing x by $x + a\underline{1}$, <u>sharp</u> when x approaches v^*, and move <u>monotonically</u> inward when x is replaced by $\tilde{Q}x$. They reduce to MacQueen's bounds if every $B^k_{i,sum} = \beta$; e.g., if $B^k_{ij} = \beta P^k_{ij}$. If the row sums are unequal, the bounds in (6.9) appear to be tighter than those due to Porteus and Hastings.

These examples illustrate the usefulness of data-transformations in painlessly extending algorithms and bounds from one model to an "equivalent" one (especially from MDPs to SMDPs).

References

[1] Bather, J., "Optimal decision procedures for finite Markov Chains," Adv. in Appl. Prob. 5 (1973), Parts I, II and III, 328-339, 521-540, 541-553.

[2] Bellman, R. "A Markovian decision process," J. Math. Mech. 6 (1957), 679-684.

[3] Brown, B., "On the iterative method of dynamic programming on a finite space discrete time Markov Process," Ann. Math. Stat. 36 (1965), 1279-1285.

[4] Denardo, E., "Contraction mappings in the theory underlying dynamic programming," SIAM Rev. 9 (1967), 165-177.

[5] Denardo, E., "Computing a bias-optimal policy in a discrete-time Markov Decision Problem," Oprns. Res. 18 (1970), 279-289.

[6] Denardo, E. & B. Fox, "Multichain Markov Renewal Programs," SIAM J. Appl. Math. 16 (1968), 468-487.

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

[8] Federgruen, A., "The rate of convergence for backwards products of a convergent sequence of finite Markov matrices," Graduate School of Management Working Paper Series No. 7827, University of Rochester (1978) (to appear in Stoch. Proc. and their Appl.).

[9] Federgruen, A. & P. J. Schweitzer, "Discounted and undiscounted value iteration in Markov decision problems: A survey," Math. Center Report BW78/77 (1977) (to appear in the Proceedings of the International Conference on Dynamic Programming and its Applications, Vancouver, April 1977, to be published by Academic Press, New York).

[10] Federgruen, A. & P. J. Schweitzer, "Nonstationary Markov Decision Problems with converging parameters," Math. Center Report BW91/78 (1978a).

[11] Federgruen, A. & P. J. Schweitzer, "Turnpike properties in undiscounted Markov Decision Problems," (1979a), (forthcoming).

[12] Federgruen, A. & P. J. Schweitzer, "A Lyapunov function for Markov Renewal Programming," (1979b), (in preparation).

[13] Federgruen, A. & P. J. Schweitzer, "Data transformations for Markov Renewal Programming," (1979c), (forthcoming).

[14] Federgruen, A., P. J. Schweitzer & H. C. Tijms, "Contraction Mappings underlying undiscounted Markov Decision Problems," J. Math Anal. Appl. 65 (1978), 711-730.

[15] Hastings, N., "Bounds on the gain of a Markov Decision Process," Oprns. Res. 19 (1971), 240-244.

[16] Hastings, N., "A test for nonoptimal actions in undiscounted finite Markov Decision Chains," Man. Sci. 23 (1976), 87-92.

[17] Hastings, N. & J. Mello, "Tests for suboptimal actions in discounted Markov Programming," Man. Sci. 19 (1973), 1019-1022.

[18] Hordijk, A. & L. Kallenberg, "Linear Programming and Markov Decision Chains," Report 78-1, Institute of Applied Mathematics and Computer Science, University of Leiden, Netherlands. (January, 1979)

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

[20] Jewell, W., "Markov Renewal Programming," Oprns. Res. 11 (1963), 938-971.

[21] Lanery, E., "Etude asymptotique des systèmes Markoviens à commande," Rev. Inf. Rech. Op. 1 (1967), 3-56.

[22] Lippman, S. (1975), "Applying a new device in the optimization of exponential queueing systems," Oprns. Res. 23 (1975), 687-710.

[23] MacQueen, J., "A Modified Dynamic Programming Method for Markovian Decision Problems," J. Math. Anal. Appl. 14 (1966), 38-43.

[24] MacQueen, J., "A test for suboptimal actions in Markovian Decision Problems," Oprns. Res. 15 (1967), 559-561.

[25] Morton, T. & W. Wecker, "Discounting, Ergodicity and Convergence for Markov Decision Processes," Man. Sci. 23 (1977), 890-900.

[26] Odoni, A., "On finding the maximal gain for Markov Decision Processes," Oprns. Res. 17 (1969), 857-860.

[27] Porteus, E., "Some bounds for discounted sequential decision processes," Man. Sci. 18 (1971), 7-11.

[28] Porteus, E., "Bounds and transformations for discounted finite Markov decision chains," Oprns. Res. 23 (1975), 761-784.

[29] Porteus, E. & J. Totten, "Extrapolations for iterative methods of solving M-matrix equations," GSB Report RT 209, (1974), Stanford University, Stanford, California.

[30] Schweitzer, P. J., "Multiple Policy improvements in undiscounted Markov Renewal Programming," Oprns. Res. 19 (1971a), 784-793.

[31] Schweitzer, P. J., "Iterative Solution of the functional equations for undiscounted Markov Renewal Programming," J. Math. Anal. Appl. 34 (1971b), 495-501.

[32] Schweitzer, P. J., "Data Transformations for Markov Renewal Programming," ORSA National Meeting, Atlantic City, New Jersey, November 10, 1972.

[33] Schweitzer, P. J. & A. Federgruen, "Functional Equations of undiscounted Markov Renewal Programming," Math. Center Report BW60/76, BW71/77 (1977a).

[34] Schweitzer, P. J. & A. Federgruen, "Geometric convergence of value-iteration in multichain Markov Renewal Programming," (1977b), Math. Center Report BW80/77 (to appear in Adv. Appl. Prob.).

[35] Schweitzer, P. J. & A. Federgruen, "The asymptotic behaviour of undiscounted value iteration in Markov Decision Problems," Math. of O.R. 2, (1978a), 360-381.

[36] Schweitzer, P. J. & A. Federgruen, "Functional equations of undiscounted Markov Renewal Programming," (1978b), Math. of O.R. 3, 308-322.

[37] Schweitzer, P. J. & A. Federgruen, "Variational Characteristics in Markov Renewal Programs," (1979a), (forthcoming).

[38] Shapiro, J., "Turnpike planning horizons for a Markovian Decision Model," Man. Sci. 14 (1968), 292-300.

[39] White, D., "Dynamic Programming, Markov Chains, and the method of successive approximations," J. Math. Anal. Appl. 6 (1963), 373-376.

[40] White, D., "Elimination of non-optimal actions in Markov decision processes," in The Proceedings of the International Conference on Dynamic Programming and its Applications, Vancouver, April 1977, to be published by Academic Press, New York.

[41] Zangwill, W., Nonlinear Programming; A Unified Approach, Englewood Cliffs, Prentice-Hall, Inc., (1969).