stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BW 78/77 AUGUSTUS
(DEPARTMENT OF OPERATIONS RESEARCH)

A. FEDERGRUEN & P.J. SCHWEITZER

DISCOUNTED AND UNDISCOUNTED VALUE-ITERATION IN MARKOV DECISION PROBLEMS: A SURVEY

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(MOS) subject classification scheme (1970): 90C45

Discounted and undiscounted value-iteration in Markov Decision Problems: a survey[**]

by

A. Federgruen & P.J. Schweitzer[*]

ABSTRACT

A survey is given of the present state of the art of value-iteration and related successive approximation methods, as well as of resulting turnpike properties in both the *discounted* and *undiscounted* version of finite Markov Decision Problems.

KEY WORDS & PHRASES: *value-iteration, turnpike results, asymptotic behaviour, geometric convergence, data-transformations*

[*] I.B.M. Thomas J. Watson Research Center, Yorktown Heights, N.Y. 10598 U.S.A.

[**] This report will be submitted for publication elsewhere.

1. INTRODUCTION

This paper surveys both older and recent results on the asymptotic behaviour of the value-iteration scheme

$$(1.1)\qquad v(n+1)_i = \max_{k\in K(i)} [q_i^k+\beta \sum_{j=1}^{N} P_{ij}^k v(n)_j], \qquad 1 \le i \le N \qquad n = 0,1,2,\ldots$$

which arises in finite-state ($N<\infty$) and finite action ($1\le|K(i)|<\infty$) Markov Decision Processes (cf.[3],[22]). Here q_i^k and $P_{ij}^k \ge 0$ denote, respectively, the one-step expected reward and transition probability to state j when action k is chosen in state i ($\sum_j P_{ij}^k=1$; $i = 1,\ldots,N$). The starting point v(0) (scrap value vector) is arbitrary and $v(n)_i$ denotes the maximum possible expected n-period reward starting from state i.

Asymptotic results are of interest because they show the relation between the finite-horizon and infinite-horizon models where use of the latter case is justified if the planning horizon is large, although possibly not (exactly) known. Two types of asymptotic results are presented. One type involves the asymptotic behaviour of the value function, i.e.

(1) v(n) if the discount factor β satisfies $0 \le \beta < 1$, or

(2) $v(n) - ng^*$ where g^* is the maximal gain rate vector in the *undiscounted* case where $\beta = 1$.

The other type of asymptotic result concerns the behaviour of the sequence of the sets of optimizing policies S(n), where

$$(1.2)\qquad S(n) = X_{i=1}^{N} K(n,i); \qquad n = 1,2,3,\ldots$$

with

$$K(n,i) = \{k \in K(i) \mid v(n)_i = q_i^k + \beta \sum_{j=1}^{N} P_{ij}^k\, v(n-1)_j\}; \quad i = 1,\ldots,N$$

as well as the existence of so-called initially stationary or periodic optimal or ε-optimal strategies (see below).

The following notation will be employed. We let $S = X_{i=1}^{N} K(i)$ denote the finite set of *policies*.

A strategy $\pi = (\ldots,A^{(\ell)},\ldots,A^{(1)})$ is an infinite sequence of policies where applying strategy π means using policy $A^{(\ell)}$ when there are ℓ periods to go.

A strategy is said to be *stationary* if it uses the same policy at each period, i.e. if $A^{(\ell)} = A$ for all $\ell = 1,2,\ldots$. Note that each policy specifies a stationary strategy. Likewise, a strategy $\pi = (\ldots,A^{(\ell)},\ldots,A^{(1)})$ is called *initially* stationary if there exists an integer $n_0 \geq 1$ and a policy A such that $A^{(\ell)} = A$ for all $\ell \geq n_0$.

Finally, a strategy is *optimal* (or ε-optimal for any $\varepsilon > 0$) if for each $n = 1,2,\ldots$ each component of its expected n-period reward vector equals (comes within ε of) the maximal vector $v(n)$.

Observe that a strategy $\pi = (\ldots,A^{(\ell)},\ldots,A^{(1)})$ is optimal if and only if $A^{(\ell)} \in S(\ell)$ for all $\ell = 1,2,\ldots$. For each $\varepsilon > 0$, and $n = 1,2,\ldots$ we define $S(n,\varepsilon)$:

$$S(n,\varepsilon) = \{A \in S \mid q_i^A + \beta[P^A v(n-1)]_i \geq v(n)_i - \varepsilon, \quad i = 1,\ldots,N\}.$$

Associated with each policy $A = (A(1),A(2),\ldots,A(N)) \in S$ are the reward vector $q^A = [q_i^{A(i)}]$ and transition probability matrix (<u>tpm</u>) $P^A = [P_{ij}^{A(i)}]$. Thus (1.1) and (1.2) may be rewritten as:

$$(1.3) \qquad v(n+1) = Tv(n) = T^{n+1}v(0) \qquad n = 0,1,2,3,\ldots$$

where

$$(1.4) \qquad Tx = \max_{A\in S}[q^A+\beta P^A x] \text{ (component by component maximization).}$$

Separate treatment will be given for the discounted and undiscounted cases. In both models, the geometric rate of convergence of the value function (i.e. of $v(n)$ or of $v(n) - ng^*$) plays a central role.

Finally, in section 6 we show that elementary data-transformations turn both discounted and undiscounted Markov Renewal Programs (cf. [7], [23]) into (discrete-time) Markov Decision Problems which are equivalent in the sense that they have the same state- and policy space as well as the same total discounted return or gain rate vector for each policy.

2. DISCOUNTED CASE: ASYMPTOTIC BEHAVIOUR OF v(n)

The discounted case possesses an elegant treatment because the T operator defined by (1.4) is a contraction operator with contraction modulus less than or equal to $\beta < 1$ when the L_∞-norm is used:

(2.1) $\|Tx-Ty\| \le \beta\|x-y\|$ all $x,y \in E^N$.

The classical theory of contraction operators summarized for example in DENARDO [6], may be brought to bear, with the following immediate results:

(2.2) T has a unique fixed point $v^* = Tv^*$

(2.3) for any starting point x, $T^n x$ converges geometrically to the fixed point:

(2.4) $\|T^n x-v^*\| \le \beta^n\|x-v^*\|$; $n = 1,2,3,\ldots$

(2.5) an upperbound on the distance between $T^n x$ and v^* can be computed after just one iteration of T via

(2.6) $\|T^n x-v^*\| \le \beta^n\|Tx-x\|/(1-\beta)$; $n = 1,2,\ldots$

which is fairly sharp provided β is not too close to unity.

Additional properties follow from the fact that T is a monotone operator ($x \ge y$ implies $Tx \ge Ty$). E.g.

(2.7) $v_i^* = \max_{A\in S} v_i^A$, $1 \le i \le N$

where v^A is the total expected discounted return vector associated with policy A:

(2.8) $v^A = [I-\beta P^A]^{-1} q^A$

Observe that both v^* and v^A, $A \in S$, are independent of the scrap value vector $v(0) \in E^N$.

As a consequence the unique fixed point of the T-operator coincides with the maximal total discounted return vector. Moreover,

$$(2.9) \qquad x \geq v^*(x \leq v^*) \text{ implies } T^n x \downarrow v^* (T^n x \uparrow v^*)$$

Some of these results have been modified by using instead of (2.1):

$$(2.10) \qquad \beta(x-y)_{\min} \leq (Tx-Ty)_{\min} \leq (Tx-Ty)_{\max} \leq \beta(x-y)_{\max}$$

where $x_{\min} = \min_i x_i$ and $x_{\max} = \max_i x_i$.
Thus (2.6) is replaced by

$$(2.11) \qquad T^n x_i + \frac{\beta^n}{(1-\beta)} (Tx-x)_{\min} \leq v_i^{A(n)} \leq v_i^* \leq T^n x_i + \frac{\beta^n}{(1-\beta)} (Tx-x)_{\max}$$

where $A(n) \in S(n)$ and $v^{A(n)}$ is the associated total return vector. Note that the bounds in (2.11) are invariant to adding a constant c to each component of x. This bound was originally derived for n = 0 by MAC QUEEN [29] and PORTEUS [32] later pointed out that Tx is a better approximation to v^* than x because the n = 1 bound is tighter.

Additional imporvements on the bounds as well as on the rate of convergence can be based upon data transformations ([33], [39], [11], [40]) or Gauss-Seidel variants of the iterative scheme ([16],[24],[39]), extrapolation and over-relaxation techniques ([35],[36],[48]), as well as by removal of self-transitions. These transformations obviously destroy the interpretation of v(n).

In terms of the original value-iteration scheme $v(n) = T^n x$ where $x = v(0)$, the above results have been useful in at least four ways:

(a) v(n) is shown to approach v^* geometrically fast

(b) the n = 0 or 1 versions of (2.6) or (2.11) get computable bounds on the error between the fixed point v^* and the current best guess (x or Tx)

(c) elimination via the bounds of alternatives which are not optimal for the ∞-horizon problem, cf. MAC QUEEN [29], HASTINGS [19] and GRINOLD [15]

(d) prior estimation of how many *additional* iterations n(x) are required given that the current estimate of v^* is x, until the new estimate $T^n x$ lies within ε of v^* or until a policy $A(n) \in S(n)$ found at the end of these n iterations has a return vector $v^{A(n)}$ which lies within ε of v^*.

Bounds on n(x) are obtained by setting (cf. (2.6))

$$(2.12) \qquad \|T^n x - v^*\| \le \frac{\beta^n \|Tx-x\|}{1-\beta} \le \varepsilon$$

or cf. FINKBEINER and RUNGALDIER [14]:

$$(2.13) \qquad 0 \le v^* - v^{A(n)} \le \frac{2\beta^n \|Tx-x\|}{1-\beta} \le \varepsilon$$

with the result that at most

$$(2.14) \qquad n(x) \le \ln\left[\frac{(1 \text{ or } 2)\|Tx-x\|}{\varepsilon(1-\beta)}\right] / |\ln(\beta)|$$

additional iterations are required. This has the property $n(Tx) \le n(x)-1$, so that the number of remaining iterations to get accuracy ε decreases by at least unity with each iteration; hence the termination criterion will be met after a finite number of steps.

Unfortunately n(x) can be large if β is close to unity or if the initial guess x is far from v^*. An encouraging feature is that n(x) varies only logarithmically with ε so that it is practical to achieve high precisions as long as β is not too close to unity.

We finally note that using (2.11) the upperbound for n(x) in (2.14) may be replaced by:

$$(2.14') \qquad n(x) \le \ln\left[\frac{sp[Tx-x]}{\varepsilon(1-\beta)}\right] / |\ln(\beta)|$$

where $sp[x] = x_{\max} - x_{\min}$ denotes the *span* of x (cf. BATHER [2]).

3. DISCOUNTED CASE; ASYMPTOTIC BEHAVIOUR OF S(n) AND THE EXISTENCE OF INITIALLY STATIONARY ε-OPTIMAL STRATEGIES

The main question of interest is the relation of the sets S(n) to the set S^* of policies which are optimal for the infinite horizon problem.

$$(3.1) \qquad S^* = \{A \in S \mid v^* = q^A + \beta P^A v^*\}.$$

Note that S^* is uniquely determined and has a Cartesian product structure.

It follows directly from (2.13) and (2.14) that for each starting point $x = v(0)$, $S(n) \subseteq S^*$, for sufficiently large n, say $n \geq n_1(x)$. As a choice of $n_1(x)$ one may evaluate (2.14) with

$$(3.2) \qquad \varepsilon < \varepsilon_0 = \begin{cases} \min\{v_i^*-v_i^A \mid A \in S \text{ and } 1 \leq i \leq N \text{ such that} \\ \qquad\qquad v_i^* > v_i^A\}, \text{ if } S^* \neq S \\ \infty \qquad\qquad\qquad , \text{ if } S^* = S. \end{cases}$$

Thus value-iteration eventually settles upon optimal policies. Unfortunately this result can not be used in general while performing calculations because the lack of prior knowledge about v^* - and the resulting inability to evaluate ε_0 - makes it impossible to calculate $n_1(x)$ a priori. Estimation of $n_1(x)$ remains an outstanding problem. Until the problem is resolved, no ways are available to deduce whether a policy in S(n) lies in S^*, except by elimination of suboptimal actions. That is, a policy can appear during the first (say) 50 iterative steps yet fail to be optimal for the infinite horizon-model. Furthermore a policy from S^* might appear in say S(1), *not* appear in S(2) and reappear in S(4) (or never reappear); so that a policy which has "dropped out" of S(n) cannot be eliminated as suboptimal (cf.[45]).

In the special case where v^* is known, ε_0 may be estimated (cf. SHAPIRO [45]) from:

$$(v^*-v^A)_i = [I-\beta P^A]^{-1}\Gamma_i^A = \sum_{n=0}^{\infty} \sum_{j=1}^{N} (\beta P^A)^n_{ij}\Gamma_j^A$$

where

$$\Gamma^A = [v^* - q^A - \beta P^A v^*] \geq 0.$$

Namely assuming that S^* is a proper subset of S, i.e. $\varepsilon_0 < \infty$ we can pick a pair (i,A) which achieve ε_0 in (3.2) and a state j and an integer $n \leq N$, such that $(\beta P^A)^n_{ij} > 0$ and $\Gamma^A_j > 0$. We thus find:

$$\varepsilon_0 \geq (\beta\alpha)^n \delta_0$$

where

$$\alpha = \min\{P^k_{rs} \mid \text{all } 1 \leq r,s \leq N \text{ and } k \in K(r) \text{ with } P^k_{rs} > 0\}$$

$$\delta_0 = \min\{\Gamma^A_j \mid \text{all } A \in S, 1 \leq j \leq N \text{ with } \Gamma^A_j > 0\} > 0$$

the last inequality following from the assumption $S^* \neq S$. Hence it suffices to take $\varepsilon_0 = (\beta\alpha)^N \delta_0$ when computing $n_1(x)$ via (2.14).

The following properties are known regarding convergence of S(n) for large n

(a) if S^* is a singleton, S(n) must reduce to S^* for large enough n (i.e. for $n \geq n_1(x)$)

(b) if S^* is not a singleton, S(n) does not need to possess a limit as n tends to infinity. SHAPIRO [45] has constructed a 2-state example where S(n) oscillates with period 2 between the two members of S^*.
Both his example, and an example in BROWN [5] suggest that the set S(n) exhibits at least an ultimately *periodic* behaviour.
However, an example which is similar to the one given in BATHER [2] for the *undiscounted* case (see below) shows that the worst behaviour of S(n) will be *non-periodic* oscillations.

(c) since S(n), for large n, may oscillate or contain only a *proper* subset of S^*, the individual S(n)'s do not by themselves determine S^*. However, one may find the entire set S^* from ε-optimal policies, i.e. from

(3.3) $$S^* = \lim_{n\to\infty} S(n, \varepsilon_n)$$

where $\{\varepsilon_n\}_{n=1}^{\infty}$ may be taken as an arbitrary sequence of positive numbers approaching 0, provided that the rate of convergence of $\{\varepsilon_n\}_{n=1}^{\infty}$ is slower than the one $\{v(n)\}_{n=1}^{\infty}$ exhibits, i.e. whenever $\varepsilon_n \beta^{-n} \to \infty$, as $n \to \infty$. One choice is $\varepsilon_n = n^{-1}$ and more generally, take ε_n^{-1} as a positive polynomial in n. To confirm (3.3) note that

$$S(n,\varepsilon_n) = \{A \in S \mid v(n) - q^A - \beta P^A v(n-1) \leq + \varepsilon_n \underline{1}\}$$

where $\underline{1}$ is the N-vector with all components unity. Insert $v(n) = = v^* + O(\beta^n)$ to obtain

$$S(n,\varepsilon_n) = \{A \in S \mid v^* - q^A - \beta P^A v^* \leq - \varepsilon_n \underline{1} + O(\beta^n)\}$$

$+ \varepsilon_n \underline{1} + O(\beta^n)$ approaches zero; hence for all n large enough, $0 < + \varepsilon_n \underline{1} + O(\beta^n) < \varepsilon_0 \underline{1}$ which proves the limit result in (3.3).

We finally turn to the issue of determining initially stationary optimal strategies. We observed before that an optimal strategy must lie in $X_{n=1}^{\infty} S(n)$. SHAPIRO's example (cf. [45]) shows that in general there may be no (optimal) policy which is contained within all of the sets S(n) for all n large enough. That is, $\liminf_{n\to\infty} S(n)$ may be empty, or, none of the sequences of policies that may be generated by value-iteration needs to converge. So in general, no initially stationary optimal strategy may exist and the adaptation of example 1 in BATHER [2], mentioned above, shows that in general no initially *periodic* optimal strategy needs to exist either.

Only in the case where S^* is a singleton ($S^*=\{B\}$), do we know that $S(n) = \{B\}$ for all $n \geq n_1(x)$, so that in this case every optimal strategy is initially stationary. Or, in other words, B is the best choice of current policy if the planning horizon is *at least* $n_1(x)$ *additional* periods and this choice is optimal without knowing the exact length of the planning horizon.

We observe however that every policy in S^* comes closer and closer to being optimal at the n^{th} stage, as n tends to infinity. This may be verified from

$$\|v(n)-q^A-\beta P^A v(n-1)\| = \|v(n)-v^*-\beta P^A[v(n-1)-v^*]\|$$

$$\leq 2\beta^n\|Tx-x\|/(1-\beta), \qquad A \in S^*$$

using (2.6).

This in turn implies for every $\varepsilon > 0$, the existence of an initially stationary strategy that is ε-*optimal*. In addition we point out the following two properties:

(1) Any policy in S^* may be used in the initially stationary part of the ε-optimal strategy; i.e. the initially stationary part does not depend upon the scrap-value vector $v(0)$.

(2) An upperbound for the length of the non-stationary tail of the ε-optimal strategy is given by

$$m(x) \leq \ln\left[\frac{2\|Tx-x\|}{\varepsilon(1-\beta)^2}\right] / |\ln(\beta)|$$

which varies again logarithmically with the precision ε.

4. UNDISCOUNTED CASE: ASYMPTOTIC BEHAVIOUR OF $v(n)-ng^*$

In the undiscounted case, $\beta = 1$ and T is a non-expansive operator:

$$(4.1) \qquad (x-y)_{\min} \leq (Tx-Ty)_{\min} \leq (Tx-Ty)_{\max} \leq (x-y)_{\max}; \text{ all } x,y \in E^N.$$

In addition the T operator has the property

$$(4.2) \qquad T(x+c\underline{1}) = Tx + c\underline{1} \quad \text{for all } x \in E^N \text{ and scalars } c.$$

Note as a consequence of (4.1) that the T operator never has a unique fixed point and hence is never a contraction operator on E^N (and neither is any of its powers). Both (4.1) and (4.2) suggest choosing

$$(4.3) \qquad sp[x] = x_{\max} - x_{\min}$$

as a quasi-norm (cf. BATHER [2]). However, example 1 in [13] shows that T (or any of its powers) is not necessarily *contracting* with respect to the sp-norm either. That is, only under special conditions with respect to the (chain- and periodicity) structure of the problem, (cf. [13]) does there exist a number $0 \leq \alpha < 1$, and an integer $n \geq 1$ such that

$$sp[T^n x - T^n y] \leq \alpha \; sp[x-y]; \text{ for all } x,y \in E^N.$$

As a consequence the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$ requires an entirely different and more complicated analysis in the undiscounted case.

Define the gain rate vector g^A of policy $A \in S$ by

$$g^A = \lim_{m\to\infty} m^{-1}[I+P^A+(P^A)^2+\dots(P^A)^{m-1}]q^A$$

and define the maximal gain rate vector g^* by

$$(4.4) \qquad g_i^* = \max_{A\in S} g_i^A; \qquad 1 \leq i \leq N.$$

HOWARD [22] and DERMAN [8] have shown that policies exist which attain the N maxima simultaneously, so the set

$$S_{MG} = \{A \in S \mid g^A = g^*\}$$

of maximal gain policies is non-empty.

In contrast with the discounted case, a *pair* of optimality equations is needed in order to characterize the set of optimal (maximal gain) policies

$$(4.5) \qquad g_i = \max_{k\in K(i)} \sum_j P_{ij}^k \, g_j, \qquad i = 1,\dots,N$$

$$(4.6) \qquad v_i = \max_{k\in L(i)} [q_i^k - g_i + \sum_j P_{ij}^k \, g_j], \qquad i = 1,\dots,N$$

where

$$L(i) = \{k \in K(i) \mid g_i = \sum_j P_{ij}^k \, g_j\}, \qquad i = 1,\dots,N$$

(4.5) and (4.6) always have a solution pair (cf. HOWARD [22]) and each solution pair (g,v) has $g = g^*$, so that the sets $L(i)$ are uniquely determined as well. However, unlike the discounted case, v is *not* uniquely determined by (4.6). Note e.g. that if v satisfies (4.6) then so does $v + c\underline{1}$ for any scalar c. A characterization of the set $V = \{v \mid v \text{ satisfies } (4.6)\}$ is given in [41], and is rather complex. For each $v \in V$, we define

$$S^*(v) = \{A \in S \mid v = q^A - g^* + P^A v\}$$

i.e. $S^*(v)$ is the Cartesian product set of policies achieving the maxima in (4.6) for the particular solution $v \in V$.

A policy A is *maximal gain if* for some $v \in E^N$ satisfying (4.6), $A(i)$ attains the maximum in (4.5) for all $i = 1,\dots,N$ as well as in (4.6) for every state that is recurrent under P^A. *Conversely*, if A is maximal gain, then $A(i)$ satisfies (4.5) for all $i = 1,\dots,N$, as well as (4.6) in the recurrent states for any solution to (4.6).

In the case where each $P^A > 0$, BELLMAN [3] showed that $v(n)$ has the asymptotic behaviour $v(n) \sim ng^*$ for any $v(0) \in E^N$. BROWN [5] showed in all generality that $\{v(n) - ng^*\}_{n=1}^{\infty}$ is bounded in n, permitting the interpretation of $g_i^* = \lim_{n\to\infty} v(n)_i/n$ as the maximal expected return per unit time starting from state i.

Two cases can be distinguished.

In the first case $\{v(n) - ng^*\}_{n=1}^{\infty}$ has a limit for *any* choice of $v(0)$. This corresponds roughly to the situation in the discounted process. In the second case, $\{v(n) - ng^*\}_{n=1}^{\infty}$ has a limit for some, but not all choices of $v(0)$. It is possible to show that for each Markov Decision Process there exist $v(0) \in E^N$ such that $\lim_{n\to\infty} [v(n)-ng^*]$ exists, namely $v(0) = v^* + ag^*$ where $a >> 0$ and v^* satisfies the optimality equation (4.6) above.

It is also possible to construct MDP's in which case 2 holds, namely when certain tpm's have periodic states. For example consider a four-state MDP with only one policy A having

$$q^A = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad P^A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad g^* = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

(4.7) $\lim_{n\to\infty} v(n)$ exists if and only if $v(0) = (b,b,b,b)$

(4.8) $\lim_{n\to\infty} v(2n)$ exists, whereas $\{v(n)\}_{n=1}^{\infty}$ has two distinct limit points, if $v(0) = (b,c,b,c)$ with $b \neq c$

(4.9) $\lim_{n\to\infty} v(4n)$ always exists, whereas $\{v(n)\}_{n=1}^{\infty}$ has four distinct limit points, if $v(0) = (b,c,d,e)$ with b,c,d,e distinct.

Conditions determining the existence of $\lim_{n\to\infty} [v(n)-ng^*]$ are of importance for at least the following reasons:

(1) If $v(0)$ is such that $\lim_{n\to\infty} [v(n)-ng^*]$ exists, then $v(n)-v(n-1)$ converges to g^*, and $nv(n-1)-(n-1)v(n)$ converges to a solution $v \in V$. That is, both the maximal gain rate vector and a solution to the optimality equation (4.6) can be computed.

(2) Convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ guarantees that $S(n) \subseteq S_{MG}$ for all n large enough (cf. ODONI [31]), hence value-iteration may be used to identify maximal gain policies. However if $v(0)$ is such that $\lim_{n\to\infty}[v(n)-ng^*]$ does not exist then $S(n) \subseteq S_{MG}$ is *not* guaranteed to hold for all large n: LANERY [25] has given an example where $S(n) \subseteq S\backslash S_{MG}$ for infinitely many n, and the authors have given an example (cf. [10]) where $S(n) \subseteq S\backslash S_{MG}$ for every n. In such cases value-iteration will not settle on maximal gain policies. In section 5 a more detailed analysis of the asymptotic behaviour of $\{S(n)\}_{n=1}^{\infty}$ will be given, both for the case where $\{v(n)-ng^*\}_{n=1}^{\infty}$ converges and for the one where it fails to converge.

Since value-iteration is the only practical computational method for finding maximal gain policies when $N >> 1$, it is desirable to check whether $\lim_{n\to\infty} [v(n)-ng^*]$ is guaranteed to exist, or whether a data transformation should be performed (cf. section 6) on the original data so as to enforce convergence.

(3) Convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ guarantees the existence of initially stationary ε-optimal strategies for any positive ε.

Conversely, MDP's may be constructed in which for some choices of the scrap value vector $v(0)$ for which $\{v(n)-ng^*\}_{n=1}^{\infty}$ fails to converge, no initially stationary strategy can be found which is ε-optimal for ε sufficiently small (see section 5 below).

Sufficient conditions for the convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ were obtained by WHITE [49], SCHWEITZER [37] and others. BROWN [5] and LANERY [25] both obtained, albeit with faulty proofs, that there exists a positive integer J^*, such that

$$\lim_{n\to\infty} [v(nJ^*+r)-(nJ^*+r)g^*] \text{ exists for any } v(0) \text{ and any } r = 0,\dots,J^*-1.$$

A new proof was provided by the present authors who obtained the following generalizations (cf. [42]):

(a) there exists an integer $J^* \geq 1$ such that $\lim_{n\to\infty}[v(nJ+r)-(nJ+r)g^*]$ exists for every $v(0) \in E^N$ and $r = 0,\dots,J-1$ *if and only if* J is a multiple of J^*

(b) for any given $v(0) \in E^N$, there exists an integer $J^0 \geq 1$ which depends upon $v(0)$ such that

$$\lim_{n\to\infty} [v(nJ+r)-(nJ+r)g^*] \text{ exists for } \textit{some } r = 0,\dots,J-1$$

if and only if

(4.10) $\quad J$ is a multiple of J^0.

In addition, if (4.10) holds then $\lim_{n\to\infty} [v(nJ+r)-(nJ+r)g^*]$ exists for *all* $r = 0,\dots,J-1$.

As an illustration of part (b), (4.7) - (4.9) show $J^0 = 1,2,4$ depending upon $v(0)$. Note also that J^0 divides J^*, which is 4 in this example, and that for some $v(0)$, J^0 equals J^*.

The above results require a detailed investigation of the chain- and periodicity structure of the set of maximal gain policies, including the *randomized* ones. In fact J^* can be computed using a *finite* algorithm, and can be expressed as a function of the periods (and the chain structure) of the policies in S_{MG}.

The consequence of (a) is 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

(I) there exists an aperiodic *randomized* maximal gain policy A, the set of recurrent states of which equals $R^* = \{i \mid i \text{ is recurrent under } P^A, \text{ for some } A \in S_{MG}\}$

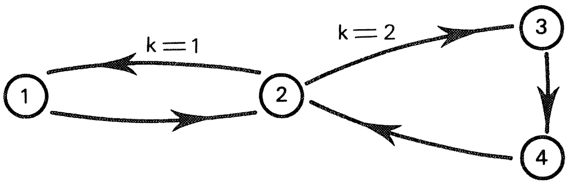
or if and only if

(II) each state $i \in R^*$ lies within an *aperiodic* subchain of some randomized maximal gain policy.

Randomization is essential in the analysis and is e.g. indispensible in the formulation of conditions (I) and (II), as may be illustrated by the following example:

EXAMPLE 1.

$N = 4$; $K(1) = K(3) = K(4) = \{1\} = K(2) = \{1,2\}$;
$P^1_{12} = P^1_{34} = P^1_{42} = P^1_{21} = P^2_{23} = 1$; all $q^k_i = 0$, i.e.



Note that the two policies in S (and S_{MG}) are both periodic with periods 2 and 3; however a *randomized* policy which uses both actions in state 2 is aperiodic and as a consequence $J^* = 1$ in this example.

Whenever $\{v(n)-ng^*\}^{\infty}_{n=1}$ converges, it can be shown (cf. [43]) that the approach to the limit is ultimately geometric in the sense that there exist numbers c and λ, with $0 \leq \lambda < 1$ such that

$$(4.11) \qquad sp[v(n)-ng^*-L(v(0))] \leq C\lambda^n, \qquad \text{for all } n = 1,2,\ldots$$

where $L(v(0)) = \lim_{n\to\infty} v(n)-ng^* \in V$.

A characterization of the ultimate convergence rate was obtained in [43] which is independent of the starting point v(0) and which in the special case where a unique maximal gain policy A exists reduces to the subdominant eigenvalue of the matrix P^A (cf. also MORTON and WECKER [30]). (4.11) is derived by showing that there exists an integer $M \geq 1$ which is independent of v(0), and an integer $n_0(v(0)) \geq 1$ such that for all $n \geq n_0$:

$$(4.12) \qquad sp[v(n+M)-(n+M)g^*-L(v(0))] \leq \mu(v(0))\ sp[v(n)-ng^*-L(v(0))]$$

where $0 \leq \mu(v(0)) < 1$.

In (4.12) $n_0(v(0)) < \infty$ indicates the number of steps after which $\{v(n)-ng^*-L(v(0))\}_{n=1}^{\infty}$ is monotonously non-increasing in sp[x]-norm. The latter is guaranteed as soon as the T operator selects policies $A \in X_i L(i)$ exclusively which is known to occur after a finite number of steps. The first n_0 steps of value-iteration thus constitute a *first* phase of the convergence process, during which the behaviour of $\{sp[v(n)-ng^*-L(v(0))]\}_{n=1}^{\infty}$ may be very irregular (it may e.g. be alternatively increasing and decreasing). The first phase is obviously *non-existent* whenever L(i) = K(i) for all i = 1,...,N which in turn is guaranteed to hold whenever the maximal gain rate is independent of the initial state of the system (g^* is a multiple of $\underline{1}$). M indicates the number of steps needed for strict contraction in the *second* phase and is uniformly bounded in v(0). Clearly one would like to obtain an upperbound for M, as a function of N. Using a combinatorial proof the present authors obtained the bound $M \leq N^2 - 2N + 2$ (cf. [43]) under the special condition

(H): $v \in V$ is unique up to a multiple of $\underline{1}$

which is equivalent to the existence of a *randomized* maximal gain policy which has R^* as its single subchain (closed, irreducible set of states). An example shows that the bound is at least sharp up to a term of O(N). The upperbound $N^2 - 2N + 2$ for the number of iterations needed for contraction represents the worst case behaviour and is enormously high compared with the empirical fact that in most cases M = 1 or M = 2. For example, SU[46] and TIJMS [47]have solved up to 1000-state with good convergence after 10-100 value iterations.

The geometric convergence result is surprising since it was noted above that T is not necessarily contracting with respect to the sp[x]-"norm". In fact it can be shown that no *uniform* m-step contraction factor needs to exist, for any m = 1,2,..., i.e. we may have for all m = 1,2,... and $n \geq n_0$:

$$(4.13)\qquad \sup\left\{\frac{sp[T^{n+m}x-(n+m)g^*-L(x)]}{sp[T^n x-ng^*-L\ (x)]}\ \middle|\ \text{for all } x\text{, for which } L(x) \text{ exists, and } T^n\ x \notin V\right\} = 1$$

Under (H), the present authors obtained the necessary and sufficient condition for the existence of a uniform m-step contraction factor for some $m = 1,2,...$ (cf. [43]).

An open question is obtaining a computationally tractible estimate of the size of λ. Nothing is known with the exception of the above mentioned case where S_{MG} is a singleton, and the cases studied by WHITE [49] and ANTHONISSE and TIJMS [1] where a n-step generalization of the ergodic- (or scrambling-) coefficient provides an upperbound for λ. Until a better understanding of λ is available, it appears unlikely that bounds on $v(n)-ng^*-L(v(0))$ can be developed similar to (2.11) in the discounted case (cf. also FEDERGRUEN, SCHWEITZER and TIJMS [13], section 4).

Applying the same analysis to so-called multi-step policies (cf. section 5), we obtain in general that for all $r = 1,2,...,J^0$

$v(nJ^0+r)-(nJ^0+r)g^*$ approaches its limit *geometrically* fast

where J^0 was defined above.

By contrast, bounds on the maximal gain rate vector g^* are available. In the case where $g^* = \langle g^*\rangle \underline{1}$ these were obtained by ODONI [31], and HASTINGS [17] namely:

$$(4.14)\qquad [Tx-x]_{min} \le g_i^A \le \langle g^*\rangle \le [Tx-x]_{max}$$

for all $x \in E^N$ and $i = 1,...,N$ and A achieving Tx.

Moreover, both bounds are sharp when $x \in V$. In the context of value-iteration this becomes

$$(4.15)\qquad [v(n+1)-v(n)]_{min} \le \langle g^*\rangle \le [v(n+1)-v(n)]_{max}.$$

The bounds move inward (monotonously) as n increases and if $\lim_{n\to\infty} v(n)-ng^*$ exists, the bounds both converge geometrically fast to $\langle g^*\rangle$. In the case where $g^* = \langle g^* \rangle \underline{1}$ it is common to avoid the linear divergence of $v(n)$ with n by using instead the variables $y(n) = v(n) - v(n)_N \underline{1}$ introduced by WHITE [49] and employing the iterative scheme

$$(4.16) \qquad y(n+1) = Ty(n) - [Ty(n)]_N \underline{1}$$

with the property that when $v^* = \lim_{n\to\infty} v(n) - ng^*$ exists

$$(4.17) \qquad \begin{aligned} & y(n) \to v^* - v_N^* \underline{1} \in V \\ & [Ty(n)]_N \to g^* \\ & [Ty(n)-y(n)]_{\max} \downarrow g^* \\ & [Ty(n)-y(n)]_{\min} \uparrow g^* \end{aligned}$$

where all four limits are approached geometrically fast. The bounds in (4.14) have been generalized to semi-Markov Decision Processes where $g^* = \langle g^* \rangle \underline{1}$, see HASTINGS [17] and SCHWEITZER [40]. They have recently been generalized by the present authors [44] to the multichain case where the components of g^* are <u>un</u>equal.

Under (H) the bounds on the scalar gain rate $\langle g^* \rangle$ have been unaccompanied by corresponding bounds on the deviation of the current vector x from $v^* \in V$ which is unique up to a multiple of $\underline{1}$. In view of the latter, this bound should be invariant to a replacement of x by $x + a\underline{1}$ for some scalar a. The existence of such bounds is also useful for demonstrating *convergence* of this or related types of value-iteration schemes. Specifically, ZANGWILL [50] has shown that an iterative scheme $x(n+1) = Qx(n)$ will converge to x^* if the continuous operator Q and a continuous Lyapunov function $\phi(x)$ satisfy:

(4.18) (a) $\phi(x) \geq 0 \quad$ all $x \in E^N$

(b) $\phi(x) = 0 \quad$ if and only if $x = x^*$

(c) $\phi(Qx) \leq \phi(x) \quad$ all $x \in E^N$

(d) for some integer $m \geq 1$, $(Q^m x) < \phi(x)$, for all x with $\phi(x) > 0$.

One choice of a Lyapunov function, not computable until v^* is known, is

$$\phi_1(x) = sp[x-v^*], \text{ with}$$

(4.19) $$Qx = Tx-(Tx)_N\underline{1} \qquad (\text{cf. } (4.16))$$

$$x^* = v^* - v_N^*\underline{1}$$

which confirms (4.16); condition (4.18d) may be verified as the scalar gain rate version of (4.12), with $M = N^2 - 2N + 2$ (see above), assuming that $J^* = 1$.

Another choice of Lyapunov function which may be computed while in the midst of the value-iteration process is

(4.20) $$\phi_2(x) = sp[Tx-x]$$

with the same choice of Q and x^*. The conditions (4.18) (a)-(c) are easily verified while (4.18) (d) holds e.g. when every policy in S_{MG} is unichained, and assuming that a data-transformation has been applied so as to ensure that $J^* = 1$ (cf. section 6 and [12]).

The important *new* property is that the deviation of v^* from x may be deduced from $\phi_2(x)$ just as (2.6) and (2.11) were used in the *discounted* case. Specifically, under (H) there exists a constant $\rho \geq 0$ such that

(4.21) $$\tfrac{1}{2}\phi_2(x) \leq sp[x-v^*] \leq \rho\phi_2(x) \qquad \text{for all } x \text{ if and only if}$$

(H2): there exists a *randomized* policy which has $\hat{R} = \{i \mid i \text{ is recurrent under some policy } A \in S\} \supseteq R^*$ as its single subchain.

Under (H) a unique representation v^* of $v \in V$ can be obtained by requiring that $v_N^* = 0$. So far, bounds for each of the components of v^* have only been obtained for the case where every policy is unichained (cf. FEDERGRUEN, SCHWEITZER and TIJMS [13]). The bounds arise by showing that the MDP can be transformed into an equivalent one (cf. section 6) in which the operator $\hat{T}$:

(4.22) $$\hat{T}x = Tx-[Tx]_N\underline{1}$$

is a N-step contraction operator on $\hat{E}^N = \{x \in E^N \mid x_N = 0\}$. The bounds are of the same type as in (2.11) and allow for the derivation (although not for the actual computation *prior* to solving the MDP) of upperbounds on the number of iterations needed to have

(1) $\hat{T}^n x$ within ε of v^*, (2) $S(n) \subseteq S_{MG}$,

(3) v^* as a relative value vector for any policy in $S(n)$, i.e. $S(n) \subseteq S^*(v^*)$. as well as on

(4) the length of the tail of an initially stationary (periodic) ε-optimal strategy.

All of the bounds in (1)-(4) vary logarithmically with $\varepsilon^{-1}sp[Tx-x]$ where in (2), ε has to be taken $\leq \min\{g^*-g^A \mid A \in S, g^* > g^A\}$ and in (3), ε has to be taken $\leq \min\{sp[v^*-v^A] \mid A \in S,\ sp[v^*-v^A] > 0\}$. Except for the case where every policy is unichained (cf. [13]) and due to the lack of bounds on $v \in V$, no tests have been proposed for *permanent* elimination of non-optimal actions. However, a device for *temporary* elimination was recently obtained in HASTINGS [18].

5. UNDISCOUNTED CASE; ASYMPTOTIC BEHAVIOUR OF S(n) AND THE EXISTENCE OF INITIALLY STATIONARY OR PERIODIC ε-OPTIMAL STRATEGIES

As discussed earlier, separate treatment is given to the cases where $\lim_{n\to\infty} [v(n)-ng^*]$ exists and where the sequence fails to converge.

We mentioned earlier having constructed in the latter case an example in which $S(n)$ lies outside S_{MG} for every n. In this case one merely knows (cf. BROWN [5]) that for large n, $S(n) \subseteq X_i L(i)$.

In the case where $v^* = \lim_{n\to\infty} [v(n)-ng^*]$ exists then for large n:

$$S(n) \subseteq S^*(v^*) \subseteq S_{MG} \subseteq X_i L(i). \tag{5.1}$$

Thus (5.1) shows that value-iteration settles upon maximal gain policies provided that convergence is guaranteed.

The explanation of this discrepancy with respect to the behaviour of $S(n)$ between the case where $[v(n)-ng^*]$ converges and the one where it fails to converge, requires the notion of multistep policies and periodic strategies.

For each integer $J \geq 1$, a J-step policy is a J-tuple of policies $(A^{*(1)},\ldots,A^{*(J)})$ and specifies a J-periodic strategy

$$\pi = (\ldots,A^{(\ell)}\ldots,A^{(1)})$$

with

$$A^{(nJ+r)} = A^{*(r)} \qquad \text{for all } n = 0,1,\ldots \text{ and } r=1,\ldots,J,$$

so, a J-step policy is called maximal gain, if the long run average return vector of the associated J-periodic strategy equals g^*.

(5.1) holds for the special case where $v(n)-ng^*$ exists, i.e. where $J^0(v(0)) = 1$, and the following generalization for $J^0 \geq 2$ may be obtained (cf. [10]):

(5.2) For all n large enough, each J^0-tuple of policies in $S(n+1) \times\ldots\times S(n+J^0)$ is maximal gain as a J^0-step policy.

Apparently a multistep-policy may be maximal gain, with each of the component-policies being *non*-maximal gain. Indeed, this phenomenon is explained by the fact that the actions prescribed by the component policies need to satisfy the optimality equations (4.5) and (4.6) only in a very special subset of $\{1,\ldots,N\}$ (cf. [10]).

The aforementioned example in [2] shows that even in the case where $v(n)-ng^*$ converges (in fact even in the case where each policy is unichained and aperiodic), S(n) may have a very irregular behaviour, the worst case of which exhibits nonperiodic oscillations.

As a consequence we are only guaranteed to have an initially stationary (or periodic) strategy if $S^*(v^*)$ is a singleton where $v^* = \lim_{n\to\infty} v(n)-ng^*$. Using the geometric convergence result as discussed in section 4, we obtain however (cf. [10]) that for all $\varepsilon > 0$, there exists an initially periodic strategy which is ε-optimal. In fact, the (initial) period of this strategy may be taken to be equal to $J^0(v(0))$.

In particular, we see that in case $J^0 = 1$, i.e. when $v(n)-ng^*$ exists, an initially stationary ε-optimal strategy exists for all $\varepsilon > 0$, and in addition $S^*(v^*)$ represents the set of policies which can be used in the initially stationary part of the strategy. This generalizes LANERY [26] who established the above result for all $\varepsilon \geq$ (some) ε^*. When $J^0 \geq 2$, a similar characterization may be given for the set of J^0-step policies which can be used in the initially periodic part of any ε-optimal strategy. In addition, MDP's can be constructed in which there exist choices of $v(0)$ for which every initially J-periodic ε-optimal strategy (with ε small enough) has J as a multiple of J^0 (this result obviously doesn't hold for every MDP with $J^* \geq 2$ as is illustrated by the case where S is a singleton). Observe that unless condition (H) is met, and unlike the discounted case the best (or ε-best) choice of current policy depends upon the terminal reward vector $v(0)$, whatever the length of the planning horizon.

Since this terminal reward vector may not be known (exactly) in advance, and since $S^*(v^*)$ may depend discontinuously upon $v(0)$, it would be desirable to choose a policy which lies in the intersection of the sets $\{S^*(v^*) \mid v^* \in V\}$. However, $\bigcap_{v \in V} S^*(v)$, which may be written as a *finite* intersection, *may be empty*.

In [41] we showed that *convexity of* V is the necessary and sufficient condition for $\bigcap_{v \in V} S^*(v) \neq \emptyset$, i.e. for the existence of a policy which can be used in the initially stationary part of the ε-optimal strategy, *in complete independence of* $v(0)$. Moreover an example was provided in which convexity of V fails to hold. Sufficient conditions for convexity of V are given by:

(1) $R^* = \Omega$; (2) $K(i)$ is a singleton for all $i \in \Omega \backslash R^*$; (3) (H).

It is worthwhile observing that in some cases a (Blackwell-) optimal policy, i.e. a policy which is optimal in the discounted model for all β sufficiently close to 1 (cf. BLACKWELL [4]) cannot be used in the initially stationary part of the ε-optimal policy (cf. [10]).

In the unichained case, i.e. when all policies are unichained, an explicit upperbound may be derived for $m(v(0))$ the length of the non-stationary (or non-periodic) tail of the ε-optimal strategy; the latter being due to the existence of bounds for the distance between v^* and the relative value vector of a policy in $S(n)$ (cf. section 4).

However, in the general multichain case and unlike the discounted model no bounds have been obtained as yet for $m(v(0))$. In analogy with the discounted case, $m(v(0))$ can however *in all generality* be shown to vary logarithmically with the precision ε. For the case of continuous time Markov Decision Problems, in which no periodicity problems arise, some of the above results were obtained by LEMBERSKY [27], and [28].

Finally, several difficulties appear when trying to find the set S_{MG}. First for all $v \in V, S^*(v)$ can be a *strict* subset of S_{MG} so that value-iteration fails to yield *all* maximal gain policies. Indeed even $\cup_{v^* \in V} S^*(v^*)$ can be a strict subset of S_{MG} so that varying the starting point $v(0)$ of value-iteration will fail to identify all maximal-gain policies. The explanation is that a maximal gain policy A is merely required to choose actions within $L(i)$, in states that are transient under P^A (cf. section 4).

The second difficulty is provided by the irregular behaviour of the sets $\{S(n)\}_{n=1}^{\infty}$ as described above. This difficulty can however be overcome by noting as in the discounted case that whenever $v^* = \lim_{n\to\infty} v(n) - ng^*$ exists:

$$(5.3) \qquad S(n,\varepsilon_n) = S^*(v^*) \qquad \text{for all } n \text{ sufficiently large}$$

provided that $\{\varepsilon_n\}_{n=1}^{\infty}$ is taken as a sequence of positive numbers approaching 0, at a slower rate than the (geometric) convergence rate of $[v(n)-ng^*]$, i.e. whenever $\lim_{n\to\infty} \varepsilon_n/\lambda^n = \infty$, e.g. when taking $\varepsilon_n = n^{-1}$.

6. DATA-TRANSFORMATIONS

In section 4 we observed that only if $[v(n)-ng^*]$ converges, will value-iteration be guaranteed to ultimately settle upon maximal gain policies and only then can sequences be derived from $\{v(n)\}_{n=1}^{\infty}$ which converge to g^* and some $v \in V$.

In case $J^* \geq 2$, i.e. in case $v(n)-ng^*$ may fail to converge, the following two alternatives have been proposed:

(A) Use the discounted value-iteration scheme with discountfactor β depending upon the index of the iteration stage, i.e.

$$(6.1) \qquad w(n+1)_i = \max_{k\in K(i)} [q_i^k + \beta_n \sum_{j=1}^{N} P_{ij}^k w(n)_j], \qquad i = 1,\ldots,N$$

where $\beta_n \to 1$ (cf. HORDIJK and TIJMS [20]). *In case* $g^* = \langle g^* \rangle \underline{1}$ then,

$$(6.2) \qquad w(n) - \gamma_n . g^* \to w^* \in V, \qquad \text{as } n \to \infty$$

where $\{\gamma_n\}_{n=1}^{\infty}$ is obtained recursively by $\gamma_{n+1} = 1 + \beta_n\gamma_n$ and $\gamma_0 = 0$ provided that

(a) $\beta_n . \beta_{n-1} \ldots \beta_1 \to 0$

(b) $\sum_{j=2}^{n} \beta_n \ldots \beta_{j+1} |\beta_j - \beta_{j-1}| \to 0$

(a) and (b) are guaranteed to hold when choosing $\beta_n = 1 - n^{-b}$, $0 \le b \le 1$. The numerical difficulty of divergence of w(n) is again avoided by using in stead the variables $\tilde{w}(n) = w(n) - w(n)_N \underline{1}$; related sequences have been derived which converge to g^* and w^*, and at each iteration step upper and lower bounds for g^* can be computed.

The convergence rate is $O(n^{-b} \ln n)$ which is considerably slower than the geometric rate ordinary undiscounted value-iteration exhibits. The nice properties of this scheme are:

(1) convergence is guaranteed regardless of the periodicity structure of the problem

(2) $\{w(n) - \gamma_n g^*\}_{n=1}^{\infty}$ converges to the optimal bias-vector (cf. BLACKWELL [4]) rather than to an arbitrary solution $v \in V$.

(B) The following data-transformation was proposed in SCHWEITZER [40]:

$$(6.4) \qquad P_{ij}^{k} = \tau(P_{ij}^{k} - \delta_{ij}) + \delta_{ij}; \qquad 1 \le i,j \le N \quad \text{and} \quad k \in K(i)$$

where $0 < \tau < 1$.

This transformation turns the MDP into on *equivalent* MDP in the sense that it has the same state- and policy-space and that the gain rate vector of each policy is identical in the original and in the transformed MDP.

The transformed problem has the nice property of *aperiodicity* for all of the policies since all of the diagonal elements of the transition probability matrices are positive. That is, the transformed problem has $J^* = 1$ and convergence of $\{v(n) - ng^*\}_{n=1}^{\infty}$ is guaranteed for any v(0). In addition the following relation exists between V and $\tilde{V}$, the set of solutions to the optimality equation (4.6) in the transformed model:

$$(6.5) \qquad \tilde{V} = \{v \in E^N \mid \tau v \in V\}.$$

A generalization of this data-transformation (cf. SCHWEITZER [40] and [11]) turns every undiscounted Markov Renewal Program (cf. JEWELL [23], DENARDO and FOX [7]) into an equivalent undiscounted MDP, in which each policy is aperiodic.

Hence, undiscounted value-iteration when applied to the transformed model, is guaranteed to settle upon maximal gain policies, and sequences may be derived which converge to the maximal gain rate vector and a solution to the optimality equation for Markov Renewal Programs.

A similar transformation may be applied in order to turn every *discounted* Markov Renewal Program into a discounted MDP which is equivalent in the sense that it has the same state- and policy space and each policy has the same total discounted return vector v^A in both models. This transformation enables us to apply value-iteration to the transformed model thus frequently improving the convergence rate, as well as to use all of the bounds and tests for non-optimal actions as discussed in section 2. Some of these bounds turn out to be sharper than the ones that have been derived by *direct* analysis of the Markov Renewal Program.

Concerning the case of a denumerable state space, the data-transformation may still prove useful for reducing the Markov Renewal Program to a discrete-time process (cf. FEDERGRUEN and TIJMS [9]). However the convergence analysis given in section 4, now needs obvious care due to the possibly more complex chain structure (cf. HORDIJK, SCHWEITZER and TIJMS [21]).

REFERENCES

[1] ANTHONISSE, J. & H.C. TIJMS, *Exponential convergence of products of stochastic matrices*, (1975) (to appear in J.M.A.A.).

[2] 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.

[3] BELLMAN, R., *A Markovian decision process*, J. Math. Mech. 6 (1957), 679-684.

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

[5] BROWN, B., *On the iterative method of dynamic programming on a finite space discrete time Markov Process*, Ann. Math. Statist. 36 (1965), 1279-1285.

[6] DENARDO, E., *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev. 9 (1967), 165-177.

[7] _________ & B. FOX, *Multichain Markov Renewal Programs*, SIAM J. Appl. Math. 16 (1968), 468-487.

[8] DERMAN, C., *Finite state Markovian Decision Processes*, Academic Press, New York (1970).

[9] FEDERGRUEN, A. & H.C. TIJMS, *The optimality equation in average cost denumerable state semi-Markov Decision Problems, recurrency conditions and algorithms*, Math. Center Report BW 74/77 (1977).

[10] __________ & P.J. SCHWEITZER, *Turnpike properties in undiscounted Markov Decision Problems* (forthcoming).

[11] __________ & _______________, *Data transformations for Markov Renewal Programming* (forthcoming).

[12] __________ & _______________, *A Lyapunov function for Markov Renewal Programming* (in preparation).

[13] __________ & _______________ & H.C. TIJMS, *Contraction Mappings underlying undiscounted Markov Decision Problems*, Math. Center report BW 72/77 (1977) (to appear in J. Math. Anal. Appl).

[14] FINKBEINER, B. & W. RUNGALDIER, *A value iteration algorithm for Markov Renewal Programming*, in *Computing methods in optimization problems 2*, ed. by L. Zadeh. New York, Academic Press (1969), 95-104.

[15] GRINOLD, R., *Elimination of suboptimal actions in Markov decision problems*, Op. Res. 21 (1973), 848-851.

[16] HASTINGS, N., *Optimization of discounted Markov Decision Problems*, Op. Res. Quart 20 (1969), 499-500.

[17] ___________, *Bounds on the gain of a Markov Decision Process*, Op.Res., 19 (1971), 240-244.

[18] ___________, *A test for nonoptimal actions in undiscounted finite Markov Decision Chains*, Man. Sci. 23 (1976), 87-92.

[19] ___________ & J. MELLO, *Tests for suboptimal actions in discounted Markov Programming*, Man.Sci. 19 (1973), 1019-1022.

[20] HORDIJK, A. & H.C. TIJMS, *A modified form of the iterative method of dynamic programming*, Ann. of Stat.3 (1975), 203-208.

[21] ___________ & P.J. SCHWEITZER & H.C. TIJMS, *The asymptotic behaviour of the minimal total expected cost for the denumerable state Markov decision model*, J. Appl.Prob. 12 (1975), 298-305.

[22] HOWARD, R., *Dynamic programming and Markov Processes*, John Wiley New York (1960).

[23] JEWELL, W., *Markov Renewal Programming*, Op.Res. 11(1963), 938-971.

[24] KUSHNER, H. & A. KLEINMAN, *Accelerated Procedures for the solution of discrete Markov control problems*, IEEE Trans. Automatic Control AC-16 (1971), 147-152.

[25] LANERY, E., *Etude asymptotique des systèmes Markoviens à commande*, Rev. Inf.Rech.Op. 1 (1967), 3-56.

[26] __________, *Compléments à l'étude asymptotique des systèmes Markoviens à commande*, I.R.I.A. Rocquencourt, France (1968).

[27] LEMBERSKY, M., *On maximal rewards and ε-optimal policies in continuous time Markov Decision Chains*, Ann. of Stat. 2 (1974), 159-169.

[28] ___________ , *Preferred rules in continuous time Markov Decision Processes*, Man.Sci. 21 (1974), 348-357.

[29] MAC QUEEN, J., *A test for suboptimal actions in Markovian Decision Problems*, Op. Res. 15 (1967), 559-561.

[30] MORTON, T. & W. WECKER, *Discounting, Ergodicity and Convergence for Markov Decision Processes*, Man.Sci. 23(1977), 890-900.

[31] ODONI, A., *On finding the maximal gain for Markov Decision Processes*, O.R. 17 (1969), 857-860.

[32] PORTEUS, E., *Some bounds for discounted sequential decision processes*, Man. Sci. 18 (1971), 7-11.

[33] __________, *Bounds and transformations for discounted finite Markov decision chains*, Op.Res. 23 (1975), 761-784.

[34] __________ & J. TOTTEN, *Extrapolations for iterative methods of solving* M-*matrix equations*, GSB report RT 209, (1974), Stanford University.

[35] REETZ, D., *Solution of a Markovian decision problem by successive overrelaxation*, Zeitschr. f. Op.Res. 21 (1973), 29-32.

[36] SCHELLHAAS, H., *Zur Extrapolation in Markoffschen entscheidungsmodellen mit Diskontierung*, Zeitschr. f. Op.Res. 18 (1974), 91-104.

[37] SCHWEITZER, P.J., *Perturbation Theory and Markovian Decision Processes*, Ph.D. dissertation, M.I.T. Operations Research Center Report 15 (1965).

[38] ______________ , *A turnpike theorem for undiscounted Markovian Decision Processes*, presented at ORSA/TIMS, National Meeting, May 1968.

[39] ______________, *Multiple Policy improvements in undiscounted Markov Renewal Programming*, Op.Res. 19 (1971), 784-793.

[40] ______________, *Iterative Solution of the functional equations for undiscounted Markov Renewal Programming*, J.M.A.A. 34 (1971), 495-501.

[41] ______________ & A. FEDERGRUEN, *Functional Equations of undiscounted Markov Renewal Programming*, Math. Center Report BW 60/76, BW 71/77 (1976) (to appear in Math. of O.R.).

[42] _______________ & ___________, *The asymptotic behaviour of undiscounted value iteration in Markov Decision Problems*, Math. Center Report BW 44/76 (1976) (to appear in Math. of O.R.).

[43] _______________ & ___________, *Geometric Convergence of value-iteration in multichain Markov Decision Problems*, (forthcoming).

[44]_______________ & ___________, *Variational Characterizations in Markov Renewal Programs*, (forthcoming).

[45] SHAPIRO, J., *Turnpike planning horizons for a Markovian Decision Model*, Man.Sci. 14 (1968), 292-300.

[46] SU, Y. & R. DEININGER, *Generalization of White's method of successive approximations to Periodic Markovian Decision Processes*, O.R. 20 (1972) 318-326.

[47] TIJMS, H., *An iterative method of approximating average cost optimal* (s,S) *inventory policies*, Zeitschr. f. Op.Res. 18 (1974), 215-223.

[48] VAN NUNEN, J., *A set of successive approximation methods for discounted Markovian Decision Problems*, Zeitschr. f. Op.Res. 20 (1976), 203-209.

[49] WHITE, D., *Dynamic Programming, Markov Chains, and the method of successive approximations*, J.M.A.A. 6 (1963), 373-376.

[50] ZANGWILL, W., *Nonlinear Programming; a unified approach*, Englewood Cliffs, N.J. Prentice Hall, Inc., (1969).